Appendix 3 to *Narratives of Epistemic Agency in Citizen Science Classification Projects: Ideals of Science and Roles of Citizens*.
Submitted to *AI & Society*.

Corresponding Author: Marisa Ponti, Department of Applied Information Technology, University of Gothenburg, Gothenburg, Sweden, marisa.ponti@ait.gu.se, ORCID ID: http://orcid.org/0000-0003-4708-4048

Legenda: When inserting a reference to a specific source, we used a code for each classification project in the metasummaries in Appendix 1, e.g., IN for I-Naturalist, then the initials of the first author followed by the year of publication. If this was the first paper in a certain year by the author, we leaved it at that. If it was the second (third, etc.) one, we appended a "b" ("c"). A whole paper is then referred to as [IN-].

In the detailed responses in the coding, we wanted statements to be supported by concrete quotes in papers. These were added by adding "-x", where x is the number of previous quotes. Example [GZA-KA19-1].

Web resources are named in a similar manner (WEB-GalaxyOrg-1) and added to the specific section of the classification project if it is only relevant to the one application.

Entries in alphabetical order:

1. Galazy Zoo AI
2. Human Protein Atlas
3. iNaturalist
4. MAIA
5. Virus Spot
6. Milky Way
7. Mindcontrol
8. Multiple Sclerosis
9. Observation
10. Plantsnap
11. Snapshot Serengeti
12. Twitter Suicide

Galaxy Zoo AI [GZA]

Journals

- **[GZA-KA19]:** https://www.sciencedirect.com/science/article/pii/S0370269319303879

  Khan, A., Huerta, E. A., Wang, S., Gruendl, R., Jennings, E., Zheng, H. (2019). Deep learning at scale for the construction of galaxy catalogs in the Dark Energy Survey, Physics Letters B, Volume 795, pp. 248-258. ISSN 0370-2693, https://doi.org/10.1016/j.physletb.2019.06.009.

  - [GZA-KA19-1]: under “Introduction”, quote: “we also need to demonstrate the applicability of this approach for DES galaxies that have not yet been observed in previous surveys. This can only be accomplished once more DES galaxies are labelled.”
  - [GZA-KA19-2]: under “Introduction”, quote: “To streamline and accelerate this method, we introduce the first application of deep transfer learning and distributed training in cosmology, reducing the training stage of the Xception model with galaxy image datasets from five hours to just eight minutes, using 64 K80 GPUs in the Cooley supercomputer.”
  - [GZA-KA19-3]: quote: “we use a pre-trained model for real-world object recognition, and then transfer its knowledge to classify SDSS and DES galaxies.”
  - [GZA-KA19-4]: quote: “To streamline and accelerate this method, we introduce the first application of deep transfer learning and distributed training in cosmology, reducing the training stage of the Xception model with galaxy image datasets from five hours to just eight minutes, using 64 K80 GPUs in the Cooley supercomputer.”
  - [GZA-KA19-5]: quote: “We show that our neural network model trained by transfer learning achieves state-of-the-art accuracy, 99.6%, to classify DES galaxies that overlap the footprint of the SDSS survey.”
  - [GZA-KA19-6]: quote: “We use our neural network classifier to label over ten thousand unlabelled DES galaxies that have not been observed in previous surveys.”
  - [GZA-KA19-7]: quote: “We then turn our neural network model into a feature extractor to show that these unlabelled datasets can be clustered according to their morphology, forming two distinct datasets.”
  - [GZA-KA19-8]: quote: “Finally, we use the newly labelled DES images and do unsupervised recursive training to retrain our deep transfer learning model, boosting its accuracy to classify unlabelled DES galaxies in bulk in new regions of parameter space.”

- [GZA-KA19-9]: quote: “The combination of all the aforementioned deep learning methods lays the foundations to exploit deep transfer learning at scale, data clustering and recursive training to produce large-scale galaxy catalogs in the LSST era.”
- [GZA-KA19-10]: quote: “We transfer knowledge from the state-of-the-art neural network model for image classification, Xception [15], trained with the ImageNet dataset [17], to classify SDSS galaxy images, achieving state-of-the-art accuracies 99.8%.”

Web

- **[WEB-article]:** https://www.journals.elsevier.com/physics-letters-b/featured-articles/classifying-galaxies-with-ai-and-people-power
  - [WEB-article-1]: "We trained our algorithm using data on over 32,000 galaxies from the labelled Galaxy Zoo dataset, so the whole project was driven by the general public."
  - [WEB-article-2]: "Khan's algorithm uses deep transfer learning, which applies the knowledge of neural networks trained with large, carefully curated datasets like the ImageNet dataset to classifying other types of images. This approach helps researchers to design and train neural network models in an optimal manner, achieving state-of-the-art results. A deep convolutional neural network called Xception, which was pre-trained with the ImageNet dataset, was tuned to recognise spiral and elliptical galaxies using the Galaxy Zoo data and then exposed to unlabelled images of galaxies from the Dark Energy Survey."

- **[WEB-Elsevier]:** https://www.journals.elsevier.com/physics-letters-b/featured-articles/classifying-galaxies-with-ai-and-people-power

  - [WEB-Elsevier-1]: quote: “The general public's interest in astronomy has been harnessed by citizen science since the original SETI@home project recruited 5 million volunteers to the ‘Search for Extraterrestrial intelligence’.”
  - [WEB-Elsevier-2]: quote: “The SDSS Galaxy Zoo project, launched in 2007, used hundreds of thousands of volunteers to classify over 50 million galaxies in a year. However, the data mountain is now growing so fast that no such project could ever classify the number of galaxies that can, or will, be observed. Enter artificial intelligence algorithms, which have become much more powerful in the decade since Galaxy Zoo started.”

  - [WEB-Elsevier-3]: quote: "Asad Khan and his co-workers at the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign, USA, worked with colleagues at the Argonne National Laboratory to develop a machine-learning algorithm that classifies galaxies much faster than the most expert volunteers."
  - [WEB-Elsevier-4]: quote: "they reduced the classification period from about 5 hours to less than 8 minutes."
  - [WEB-Elsevier-5]: quote: "Khan's algorithm uses deep transfer learning, which applies the knowledge of neural networks trained with large, carefully curated datasets like the ImageNet dataset to classifying other types of images."
  - [WEB-Elsevier-6]: quote:" 'It is an exciting time to be at the intersection of AI and astrophysics; the convergence of deep learning and high-performance computing can address big-data challenges in our field,' says Khan. 'We are uniquely poised to combine the power of these technologies for data analysis.' This analysis, at least, would not have been possible without the people power of the Galaxy Zoo."

- **[WEB-GalaxyOrg]:** https://blog.galaxyzoo.org/tag/machine-learning/

  - [WEB-GalaxyOrg-1]: quote: "The AI can guess which challenging galaxies, if classified by you, would best help it to learn. Each morning, we upload around 100 of these extra-helpful galaxies. The next day, we collect the classifications and use them to teach our AI. Thanks to your classifications, our AI should improve over time. We also upload thousands of random galaxies and show each to 3 humans, to check our AI is working and to keep an eye out for anything exciting. With this approach, we combine human skill with AI speed to classify far more galaxies and do better science. For each new survey:

    40 humans classify the most challenging and helpful galaxies

    Each galaxy is seen by 3 humans

    The AI learns to predict well on all the simple galaxies not yet classified"

  - [WEB-GalaxyOrg-2]: quote: "To keep up, Galaxy Zoo needs an automatic classifier. Other researchers have used responses that we've already collected from volunteers to train classifiers. The best performing of these are convolutional neural networks (CNNs) – a type of deep learning model tailored for image recognition. But CNNs have a drawback. They don't easily handle uncertainty. When learning, they implicitly assume that all labels are equally confident – which is definitely not the case for Galaxy Zoo (more in the section below). And when making (regression) predictions, they only give a 'best guess' answer with no error bars."

- [WEB-GalaxyOrg-3]: quote: "In our paper, we use Bayesian CNNs for morphology classification. Our Bayesian CNNs provide two key improvements:
    1. They account for varying uncertainty when learning from volunteer responses
    2. They predict full posteriors over the morphology of each galaxy

  Using our Bayesian CNN, we can learn from noisy labels and make reliable predictions (with error bars) for hundreds of millions of galaxies."

- [WEB-GalaxyOrg-4]: under the heading "How Bayesian Convolutional Neural Networks Work", quote: "There's two key steps to creating Bayesian CNNs.

  1. Predict the parameters of a probability distribution, not the label itself

  Training neural networks is much like any other fitting problem: you tweak the model to match the observations. If all the labels are equally uncertain, you can just minimise the difference between your predictions and the observed values. **But for Galaxy Zoo, many labels are more confident than others.** If I observe that, for some galaxy, 30% of volunteers say "barred", my confidence in that 30% massively depends on how many people replied – was it 4 or 40?

  Instead, we predict the probability that a typical volunteer will say "Bar", and minimise how surprised we should be **given the total number of volunteers who replied.** This way, our model understands that errors on galaxies where many volunteers replied are worse than errors on galaxies where few volunteers replied – letting it learn from every galaxy.

  2. Use Dropout to Pretend to Train Many Networks

  Our model now makes probabilistic predictions. But what if we had trained a different model? It would make slightly different probabilistic predictions. We need to **marginalise over the possible models we might have trained**. To do this, we use dropout. Dropout turns off many random neurons in our model, permuting our network into a new one each time we make predictions.

Below, you can see our Bayesian CNN in action. Each row is a galaxy (shown to the left). In the central column, our CNN makes a single probabilistic prediction (the probability that a typical volunteer would say "Bar"). We can interpret that as a posterior for the probability that k of N volunteers would say "Bar" – shown in black. On the right, we marginalise over many CNN using dropout. Each CNN posterior (grey) is different, but we can marginalise over them to get the posterior over many CNN (green) – our Bayesian prediction."

○ [WEB-GalaxyOrg-5]: under the heading of "Active learning", quote: "Modern surveys will image hundreds of millions of galaxies – more than we can show to volunteers. Given that, which galaxies should we classify with volunteers, and which by our Bayesian CNN?

Ideally we would **only show volunteers the images that the model would find most informative.** The model should be able to ask – hey, these galaxies would be really helpful to learn from– can you label them for me please? Then the humans would label them and the model would retrain. This is active learning.

In our experiments, applying active learning reduces the number of galaxies needed to reach a given performance level by up to 35-60% (See the paper).

We can use our posteriors to work out which galaxies are most informative. Remember that we use dropout to approximate training many models (see above). We show in the paper that **informative galaxies are galaxies where those models confidently disagree.**

Informative galaxies are galaxies where the each model is confident (entropy H in the posterior from each model is low) but the average prediction over all the models is uncertain (entropy across all averaged posteriors is high). See the paper for more.

This is only possible because we think about labels probabilistically and approximate training many models.

What galaxies are informative? Exactly the galaxies you would intuitively expect.

The model strongly prefers diverse featured galaxies over ellipticals

For identifying bars, the model prefers galaxies which are better resolved (lower redshift)

This selection is completely automatic. Indeed, I didn't realise the lower redshift preference until I looked at the images!

I'm excited to see what science can be done as we move from morphology catalogs of hundreds of thousands of galaxies to hundreds of millions. If you'd like to know more or you have any questions, get in touch in the comments or on Twitter (@mike_w_ai, @chrislintott, @yaringal).

- **[WEB-HPC]:** https://www.hpcwire.com/2019/07/08/scientists-leverage-hpc-and-ai-to-wrangle-the-galaxy-zoo/
  - [WEB-HPC-1]: quote: "Using the millions of classifications carried out by the public in the Galaxy Zoo project to train a neural network is an inspiring use of the citizens science program."
  - [WEB-HPC-2]: quote: "The researchers extracted the overlapping images from the two datasets using the NCSA's Blue Waters supercomputer, then taught their deep learning model on the Pittsburgh Supercomputing Center's Bridges supercomputer. The team also used the K80 Nvidia GPUs in the Cooley supercomputer at ALCF to reduce the training stage for the Xception model from five hours to eight minutes."

- **[WEB-The Atlantic]:** https://www.theatlantic.com/science/archive/2016/05/a-constellation-of-computers/481839/
  - [WEB-The Atlantic-1]: quote: "This means trying to train computers to recognize patterns as well as we can, one of the thorniest problems in computer science. Computers are still second to humans on this and they have much longer learning curves, says Matias Carrasco Kind, an astronomer at the University of Illinois at Urbana-Champaign."
  - [WEB-The Atlantic-2]: quote: "We can recognize faces in a big crowd, blurry objects in a picture, and notice people from behind or from the way they walk," he says. "By just looking at a few examples, we can extrapolate much better than computers, which need a much larger training set, and more time to process. [And computers have] a harder time [thinking] 'outside the box.' This is especially true when it comes to characterizing galaxies. You

> have to account for brightness, which is different across pixels; the galaxies' shape and symmetry; and their orientation, including whether we're looking at them face-on or sideways. Humans can do this very quickly, which is why astronomers created something called the Galaxy Zoo."

Human Protein Atlas [HPA]

Journals

[HPA-SD18]: https://www.nature.com/articles/nbt.4225#citeas

Sullivan, D., Winsnes, C., Åkesson, L. *et al.* Deep learning is combined with massive-scale citizen science to improve large-scale image classification. *Nat Biotechnol* **36,** 820–828 (2018). https://doi.org/10.1038/nbt.4225

- ○ [HPA-SD18-1]: under "Abstract", quote: "Pattern recognition and classification of images are key challenges throughout the life sciences. We combined two approaches for large-scale classification of fluorescence microscopy images. First, using the publicly available data set from the Cell Atlas of the Human Protein Atlas (HPA), we integrated an image-classification task into a mainstream video game (EVE Online) as a mini-game, named Project Discovery. Participation by 322,006 gamers over 1 year provided nearly 33 million classifications of subcellular localization patterns, including patterns that were not previously annotated by the HPA. Second, we used deep learning to build an automated Localization Cellular Annotation Tool (Loc-CAT). This tool classifies proteins into 29 subcellular localization patterns and can deal efficiently with multi-localization proteins, performing robustly across different cell types. Combining the annotations of gamers and deep learning, we applied transfer learning to create a boosted learner that can characterize subcellular protein distribution with F1 score of 0.72."
- ○ [HPA-SD18-2]: on page 820, quote: "... the large amounts of data that are generated as automated fluorescence microscopy systems become ever more widely used in quantitative biology create new challenges for automated image analysis."
- ○ [HPA-SD18-3]: on page 820, quote: "We found that engaging players of commercial computer games provided data that augmented deep learning and enabled scalable and readily improved image classification."
- ○ [HPA-SD18-4]: quote: "Crowd-sourced citizen science offers an alternative for large-scale image classification."

- [HPA-SD18-5]: on page 820, quote: "...large numbers of non-expert volunteers have contributed valuable scientific information".
- [HPA-SD18-6]: on page 820, quote: "The major drawback of this approach is that implementing an engaging citizen science project requires resources, knowledge and time that most laboratories lack. Furthermore, creating and maintaining an engaged user base is difficult in one-off citizen science projects. One method of dealing with this is paying for citizen science efforts, as in Amazon's mechanical turk (mturk); however, this method is prone to exploitation and low data quality."
- [HPA-SD18-7]: on page 820, quote: "Here we demonstrate two complementary and successful approaches for large-scale classification of protein localization patterns in microscopy images from the HPA Cell Atlas. The first utilizes the power of massive multiplayer online (MMO) games to create a new approach to citizen science and was a collaborative effort between the HPA, Massive Multiplayer Online Science (MMOS) and the video game developer CCP Games. This partnership substantially reduced the effort to the lab by allowing CCP Games to develop the interface and MMOS to handle data management and serving. The result was the scientific project of image classification seamlessly integrated into the EVE Online universe, an MMO science fiction game with ~500,000 active players each month. The resulting mini-game, Project Discovery (PD), was successful in terms of participation, player retention, number of images classified and accuracy."
- [HPA-SD18-8]: on page 820, quote: "In the second approach, we present Loc-CAT, a model for automated image classification of subcellular protein distribution patterns using deep neural networks (DNNs). To the best of our knowledge, this method represents the first tool for classifying protein distribution in human cells in microscope images capable of predicting robustly across cell types for proteins with an unknown number of locations. Furthermore, we compared the performance of the respective approaches and found that the gamer output could be used to improve deep learning models. Altogether, both approaches provide a refinement of the biological details in the HPA Cell Atlas. We believe that integration of scientific tasks into established computer games can be a valuable approach in the future with the power of rapidly leveraging the output of large-scale science efforts."
- [HPA-SD18-9]: on page 820, quote: "Each sample in the HPA Cell Atlas consists of human cells that are immunofluorescently labeled for one protein of interest and three reference markers: DAPI for the nucleus and antibody-based labeling of microtubules and the endoplasmic reticulum."
- [HPA-SD18-10]: on page 821, quote: "In PD, players in EVE Online performed the aforementioned protein image classification. This project represents the first time

a scientific task has been directly and seamlessly integrated into a mainstream video game narrative."

- [HPA-SD18-11]: on page 821, quote: "Participants were trained using a small set of preselected images gradually increasing in difficulty".
- [HPA-SD18-12]: on page 821, quote: "Participants in PD were motivated with leveled badges and ingame currency with which they could purchase exclusive items. This approach was able to easily gather and maintain participants, something other citizen science projects have struggled with."
- [HPA-SD18-13]: on page 821, quote: "players passed the training and tutorial phases and had above threshold performance, leading to 23.7 million high-quality image classifications."
- [HPA-SD18-14]: on page 821, quote: "To assess data quality, we used the F1 score, a measure of accuracy suitable for multi-label data, with the HPA Cell Atlas v14 image labels as ground truth."
- [HPA-SD18-15]: on page 821, quote: "This resulted in a rapid improvement of accuracy (Fig. 2d). On the basis of player feedback, we created and implemented a larger set of more difficult control images including multi-localizing proteins and image artifacts. This led to a significant increase in data quality (day 50, P < 4 × 10−70, day 0–50 versus 50+, two-tailed t test; Fig. 2d)."
- [HPA-SD18-16]: on page 821, quote: "To guard against erroneous annotations, we required a minimum of 12 votes per image before evaluating each task for a consensus using a hypergeometric test. Consensus was considered to be reached only if the number of votes for at least one class was significantly greater than would be expected at random (P < 0.01) and no other classes were near the decision threshold (P < 0.1). If consensus was not reached, the task was kept open and more votes were acquired. On average, each task required 15 player votes (median = 13) to reach a consensus."
- [HPA-SD18-17]: on page 824, quote: "A major contribution of the participants in PD was to refine the classifications in the Cell Atlas."
- [HPA-SD18-18]: on page 824, quote: "Another approach for classification of image patterns is machine learning. Toward this end, we used a deep neural network to create LocCAT".
- [HPA-SD18-19]: on page 824 to 825, quote: "Evaluation of Loc-CAT and citizen science performance - Despite the high performance of Loc-CAT, players in PD (average per-class F1 = 0.53) outperformed Loc-CAT (average per-class F1 = 0.47), particularly in many of the less common classes, for example, microtubule ends, which has only 32 images. Loc-CAT outperformed PD in most other classes, particularly on classes with large amounts of training data and endoplasmic reticulum (ER) where Loc-CAT has access to an additional reference channel players in PD did not (Fig. 5a). This makes the two methods closer in performance

when comparing overall F1 score (Loc-CAT = 0.65, PD = 0.68). PD continued to outperform Loc-CAT when examining the middle layer of resolution in the organelle hierarchy (Figs. 4b and 5a). Notably, gamers appeared to be more accurate at identifying nucleoli-related patterns and continued to outperform Loc-CAT in the cytoskeleton and microtubule organization meta-classes."

- [HPA-SD18-20]: on page 826, quote: "Gamer augmented transfer learning improves Loc-CAT accuracy … To leverage this information, we applied a transfer-learning approach in which we fed gamer annotations as a set of additional input features to Loc-CAT, resulting in increased performance (GA Loc-CAT; Fig. 6c)."
- [HPA-SD18-21]: on page 827, quote: "we speculate that the integration of scientific tasks into established computer games will be a commonly used approach in the future to harness the brain processing power of humans and that intricate designs of citizen science games feeding directly into machine learning models through techniques such as reinforcement learning have the power of rapidly leveraging the output of large-scale science efforts."
- [HPA-SD18-22]: on page 827, quote: "This approach reduces development costs to labs for citizen science and demonstrates that players in MMO games can produce high-quality data despite potentially being motivated by alternative in-game dynamics or fun, rather than connection to a cause."
- [HPA-SD18-23]: on page 827, quote: "we showed how gamers and DNNs excel at different types of classifications and that gamer output can be used to augment and improve deep learning models."
- [HPA-SD18-24]: on page 827, quote: "we demonstrated two alternative approaches for large-scale classification of protein distribution patterns in microscopy images."

[HPA-TP-2018]: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5734309/

Thul, P. J., & Lindskog, C. (2018). The human protein atlas: A spatial map of the human proteome. *Protein science : a publication of the Protein Society*, *27*(1), 233–244. https://doi.org/10.1002/pro.3307

- [HPA-TP-2018-1]: page 233, quote: "The article summarizes recent updates and current status of the Human Protein Atlas, www.proteinatlas.org, which is the largest and most comprehensive database for spatial distribution of proteins in human tissues and cells. An overview of the publicly available database is provided, and its functions and potential implications for use as well as the future path of spatial proteomics are discussed."

- [HPA-TP-2018-2]: page 243, quote: “The HPA represents the largest and most comprehensive database for spatial distribution of proteins in tissues and cells, providing an invaluable resource for exploration of expression patterns at a single-cell resolution.”
- [HPA-TP-2018-3]: page 244, quote: “Moreover, both IHC scoring parameters and sub-cellular localization classifications will be refined to add more cells types, more organelles, and provide intra-organellar locations.”

iNaturalist [IN]

Journals

[IN-HJ18]: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6240452/
Heberling, J. M., & Isaac, B. L. (2018). iNaturalist as a tool to expand the research value of museum specimens. *Applications in plant sciences*, *6*(11), e01193. https://doi.org/10.1002/aps3.1193

- [IN-HJ18-1]: under “Abstract”, quote: “Innovative approaches to specimen collection and curation are needed to maximize the utility of natural history collections in a new era of data use.”
- [IN-HJ18-2]: under “Abstract”, quote: “We leveraged the widely used citizen science platform, iNaturalist, to permanently associate field-collected data to herbarium specimens, including information not well preserved in traditional specimens. This protocol improves the efficiency and accuracy of all steps from the collecting event to specimen curation and enhances the potential uses of specimens.”
- [IN-HJ18-3]: under “Abstract”, quote: “iNaturalist provides a standardized and cost-efficient enhancement to specimen collection and curation that can be easily adapted for specific research goals or other collection types beyond herbaria.”
- [IN-HJ18-4]: under “Abstract”, quote: “Here, we introduce a practical method that utilizes the popular biodiversity-based citizen science platform iNaturalist (iNaturalist, 2018) to facilitate plant specimen collecting and curate field images alongside physical specimens, thereby augmenting their research value. Previous software applications have been developed to facilitate data capture in the field (e.g., Maya-Lastra, 2016), but none have yet been widely adopted. Although an important improvement, digital data capture in the field streamlines traditional field collection protocols but does not fundamentally improve the research value of new collections. iNaturalist provides several potential advantages as a tool for plant collectors, herbarium curators, and downstream researchers alike. The most notable of these include: (1) it is widely available (free, online resource) and externally supported (i.e., independent of herbarium- or project-specific funds); (2)

it could permanently link images and other metadata collected in the field with specimen records, which is of critical practical importance, as most herbaria do not have the infrastructure to store and curate associated field images and other data beyond the physical specimen and label metadata; (3) it could connect associated observation records (which may or may not be plant taxa or physical specimen-based) to physical specimens; (4) it provides a flexible platform for an editable taxonomy and specimen identifications; and (5) it holds the potential to engage a wider community of citizen scientists in natural history collection practices."

- [IN-HJ18-5]: under "Methods and Results", quote: "Users record biodiversity observations, including date, time, location, taxonomic identification, images, audio recordings, and a countless number of other user-defined data fields."
- [IN-HJ18-6]: under "Methods and Results", quote: "iNaturalist is a joint initiative of the California Academy of Sciences and the National Geographic Society, maintained by a dedicated staff and a community of citizen scientists (iNaturalist, 2018)."
- [IN-HJ18-7]: under "Methods and Results", quote: "Two powerful benefits of iNaturalist are, first, the permanent integration of field images and an array of metadata linked to an observation and, second, the community-driven process for taxonomic identification and record validation."
- [IN-HJ18-8]: under "Methods and Results", quote: "iNaturalist provides algorithmic identification suggestions to its users based on visual characteristics of the uploaded images, proximity of similar records, and identification history of the taxa in question, although all identifiers are free to choose any taxon they think the image depicts."
- [IN-HJ18-10]: under "Methods and Results", quote: "iNaturalist follows a set of established taxonomic authorities, which are updated by expert users ("Curators")."
- [IN-HJ18-11]: under "Methods and Results", quote: "There are many parallels between these digital collections that are recorded and curated by the iNaturalist community to physical collections in herbaria curated by botanical researchers. These photographic records are essentially digital specimens that lack physical voucher material to reference. Given the striking overlap in the data associated with iNaturalist observation-based records and those of specimen-based records, iNaturalist is well-suited as a data capture tool for plant collectors in the field."
- [IN-HJ18-12]: under "Methods and Results", quote: "The process (Appendix 1) starts with careful field documentation of a given individual(s) to be collected. Ideally, representative iNaturalist observations for each species in the entire community would be taken, regardless of whether physical vouchers are taken. Digital images can be taken using any camera, but the use of a GPS-enabled

device (e.g., smartphone) is most efficient because geolocational information can be automatically entered when uploaded to iNaturalist.

- [IN-HJ18-13]: under “Methods and Results”, quote: “Images stored on mobile devices can be directly uploaded in the field (or later) using the iNaturalist app. Images can also be uploaded or added to existing observations at any time through the app or online. At least one image should be taken of each specimen prior to collection, but additional images from different perspectives and focusing on different plant structures are preferred and sometimes necessary.
- [IN-HJ18-14]: under “Methods and Results”, quote: “Including objects or rulers in images for scale can be helpful for reference. Traditional specimens are collected following standard methods (Bridson and Forman, 1998), which may also include tissue samples for genomic studies (Funk et al., 2017). iNaturalist observations with associated physical vouchers are added to a “Project” in iNaturalist to facilitate necessary data entry and downstream curatorial tasks of printing specimen labels and exporting data to relevant specimen databases. iNaturalist “Projects” are easily set up online (https://www.inaturalist.org/).
- [IN-HJ18-15]: under “Methods and Results”, quote: “We recommend each herbarium (or plant collector) design its own Project to suit its needs, being sure to include necessary user-defined fields that are not already part of the core iNaturalist data fields (e.g., collector number; see Appendix 1). Data fields in iNaturalist can easily be adapted to follow Darwin Core data standards for biodiversity data (Wieczorek et al., 2012), which facilitate data exported from iNaturalist to local or online collections databases (Table 1)... Through the use of “Projects” in iNaturalist, metadata for specific observations (Table 1) are exported to print herbarium labels and merged into relevant collections databases.”
- [IN-HJ18-16]: under “Methods and Results”, quote: “In addition to permanently archiving the corresponding iNaturalist record number (URL; Table 1) in the collections database (via Darwin Core field: “associatedMedia”), we also include this web link on the physical specimen label in the form of a Quick Response Code (QR code)... We use QR codes on specimen labels to store the URL for the associated iNaturalist observation record. In this way, herbarium users examining a specimen can scan the label to be instantly directed to additional information, including field images (Fig. 1).”
- [IN-HJ18-17]: under “Methods and Results”, quote: “Similarly, herbarium users can search online specimen data portals for specimens associated with iNaturalist images through the Darwin Core field “associatedMedia.” We recommend using this data field to archive the iNaturalist URLs as a standard component of specimen metadata. An example data set (Appendix S1) and a Microsoft Word template (Appendix S2) for making herbarium labels from iNaturalist data are available in the Supporting Information.”

- [IN-HJ18-18]: under “Methods and Results”, quote: “approach leverages the existing infrastructure of iNaturalist to connect specimens to images from the field.”
- [IN-HJ18-19]: under “Methods and Results”, quote: “In addition to plant traits, iNaturalist also provides ecological and environmental context, which, to date, is infrequently and inconsistently recorded with existing specimens (e.g., habitat and/or associated species on herbarium labels).”
- [IN-HJ18-20]: under “Methods and Results”, quote: “Second, this approach provides a platform to search for related regional observations as well as other observations that were recorded in the same locality and/or on the same collection date.”
- [IN-HJ18-21]: under “Methods and Results”, quote: “Third, the integration of physical specimens with iNaturalist observations engages a community of citizen scientists for the curation of specimen metadata, including taxonomic identification and phenological scoring.”
- [IN-HJ18-22]: under “Methods and Results”, quote: “If using a GPS-enabled camera (e.g., smartphone), latitude and longitude information is automatically included with uploaded image(s). To date, however, elevation is not automatically recorded by iNaturalist, but can easily be included using the built-in or other mobile device apps.”
- [IN-HJ18-23]: under “Methods and Results”, quote: “Taxonomic identification is facilitated by artificial intelligence features in iNaturalist, the iNaturalist community, and a community-curated taxonomic nomenclature. iNaturalist improves efficiency and accuracy for botanists relative to field guides or memory alone by providing a list of identification suggestions and a set of pre-defined taxonomic names from which to choose (e.g., avoid misspellings, taxonomic synonyms).”
- [IN-HJ18-24]: under “Methods and Results”, quote: “Accuracy is also improved with identification suggestions or verifications from other iNaturalist users.”
- [IN-HJ18-25]: under “Methods and Results”, quote: “Second, once observations are complete, these data can be easily exported and directly converted into herbarium labels and merged into collections databases. The use of iNaturalist “Projects” permits plant collectors to directly share data with herbarium staff (including localities of rare species that may be censored to the general public).”
- [IN-HJ18-26]: under “Methods and Results”, quote: “Data quality is also improved by providing a standardized set of required or suggested data fields for all new vouchers being deposited, which is especially useful for outside consultants or amateur botanists who are new to plant collecting or infrequently deposit specimens.”
- [IN-HJ18-27]: under “Methods and Results”, quote: “this method effectively expands the collection event to include both specimen-based and observation-based records across and within taxonomic groups.”

Web

[WEB-iNaturalist org]: https://www.inaturalist.org/pages/help#general1

- [WEB-iNaturalist org-1]: Under “General”, Quote: “3. What technologies and data sources does the project use? iNaturalist is built using Ruby on Rails, MySQL, jQuery and Google Maps, and Flickr. It also utilizes the Catalogue of Life, uBio, and a variety of other data sources for taxonomic data.”
- [WEB-iNaturalist org-2]: Under “General”, Quote: “4. What can I do to help iNaturalist? First and foremost, you can be an active member of the community by adding your observations and helping other community members identify their unidentified observations. You can also help by sharing your ideas and feedback. Join our Community Forum to interact with other users, report bugs, and request new features. If you know how to code and want to help work on some features, fork us on GitHub! You can donate to support iNaturalist. There's even more ways to help out, explained on the iNaturalist Community Forum.”
- [WEB-iNaturalist org-3]: Under “Observations”, Quote: “What is the data quality assessment and how do observations qualify to become "Research Grade"? The Data Quality Assessment is a summary of an observation's accuracy, completeness, and suitability for sharing with data partners. The building block of iNaturalist is the verifiable observation. A verifiable observation is an observation that:

    1. has a date
    2. is georeferenced (i.e. has lat/lon coordinates)
    3. has photos or sounds
    4. isn't of a captive or cultivated organism

    Verifiable observations are labeled "Needs ID" until they either attain Research Grade status, or are voted to Casual via the Data Quality Assessment. Observations become "Research Grade" when

    5. the community agrees on species-level ID or lower, i.e. when more than 2/3 of identifiers agree on a taxon

    Observations will revert to "Casual" if the conditions for Verifiable aren't met or

    6. the community agrees the date doesn't look accurate
    7. the community agrees the location doesn't look accurate (e.g. monkeys in the middle of the ocean, captive/collected organisms observed inside a building but unlikely to have been found there naturally, etc.)

8. the community agrees the organism isn't wild/naturalized (e.g. captive or cultivated by humans or intelligent space aliens)
9. the community agrees the observation doesn't present evidence of an organism, e.g. images of landscapes, water features, rocks, etc.
10. the community agrees the observation doesn't present recent (~100 years) evidence of the organism (e.g. fossils, but tracks, scat, and dead leaves are ok)
11. the community agrees the observation no longer needs an ID *and* the community ID is above family
12. the observer has opted out of the community ID and the community ID taxon is not an ancestor or descendant of the taxon associated with the observer's ID

And of course there are even more caveats and exceptions:

13. "Research Grade" observations will become "Needs ID" if the community ID shifts above the species-level
14. "Research Grade" observations will become "Needs ID" if the community votes that it needs more IDs
15. Observations can be "Research Grade" at the genus level if the community agrees on a genus-level ID and votes that the observation does not need more IDs
16. The system will vote that the observation is not wild/naturalized if there are at least 10 other observations of a genus or lower in the smallest county-, state-, or country-equivalent place that contains this observation and 80% or more of those observations have been marked as not wild/naturalized.

- [WEB-iNaturalist org-4]: Under “How can I get help identifying what I saw?”, quote: “Just make observations of wild organisms that have photos, locations, and dates. Every observation with those things gets automatically placed in the "Needs ID" category so people who are looking for observations to identify will find them. Observations without those three things are not eligible for "Research Grade" status and thus get placed in the "Casual" category, since identifiers probably won't be able to help if there's no photo or location.
- [WEB-iNaturalist org-5]: Under “Identifications”, quote: “There are several types of IDs:
    1. Leading: Taxon descends from the community taxon. This identification could be leading toward the right answer.
    2. Improving: First suggestion of this taxon that the community subsequently agreed with. This identification helped refine the community taxon.
    3. Supporting: Taxon is the same as the community taxon. This identification supports the community ID.

4. Maverick: Taxon is not a descendant or ancestor of the community taxon. The community does not agree with this identification.”

- [WEB-iNaturalist org-6]: Under “Identifications”, quote: “I identified my observation after someone else added a higher-level ID, so why is the observation stuck with the higher-level ID? That's the way the community ID system works: iNat chooses the taxon with > 2/3 agreement, and if that's impossible, it walks up the taxonomic tree and chooses a taxon everyone agrees with, so if I say it's *Canis* and you say it's *Canis familiaris*, 2/2 identifications agree it's in *Canis* but only 1/2 think it's *Canis familiaris* so iNat goes with *Canis*. If you don't like this and want your ID to take priority for your observation, just reject the community ID by clicking the "Reject" link under the community ID. You can also opt-out of community IDs entirely by editing your settings. You don't need to ask people to remove their higher-level ID, especially if it's accurate (but not precise). This doesn't affect an observation's potential to reach Research Grade status, it just gives the observer control over what taxon the observation is associated with.”
- [WEB-iNaturalist org-7]: Under “How can I download data from iNaturalist? Anyone with an account can export data from iNaturalist as a spreadsheet in csv format. You can start from the Explore page and click download in the lower right of the filters box. Or you can go directly to the export page (https://www.inaturalist.org/observations/export). If you plan to publish a paper using iNaturalist data, we recommend downloading iNaturalist data from the Global Biodiversity Information Facility because they will issue a citable DOI (see below for more details).”

[WEB-i Programmer]: https://www.i-programmer.info/news/105-artificial-intelligence/10848-inaturalist.html/

- [WEB-i Programmer-1]: quote: “iNaturalist.org is an established and popular website. Its mission is to connect experts and amateur "citizen scientists," encouraging people to get interested and involved with the natural world while using the data gathered to potentially help professional scientists monitor changes in biodiversity or even discover new species. Founded in 2008 by students at University of California, Berkeley and recently acquired by the California Academy of Sciences, it used to rely on crowdsourcing. When users posted a photo of a plant or animal, its community of scientists and naturalists will identify it.”
- [WEB-i Programmer-2]: quote: “Another issue is that as the site becomes more popular the number of observers (people posting photos) far exceeds that of identifiers (people telling you what they are of) which threatens to overwhelm the volunteer experts. To help take the burden off the volunteer experts, the iNaturalist team collaborated with the Cornell Lab of Ornithology, developers of the Merlin

bird identification app, and Visipedia to use machine learning to deliver higher quality identifications faster as the number of observers continues to grow."

- [WEB-i Programmer-3]: quote: "Visipedia, short for "Visual Encyclopedia," is a joint project between Caltech and Cornell Tech, is a network of people and machines designed to harvest and organize visual information and make it accessible to anyone who has a visual query. Using TensorFlow deep learning framework with NVIDIA hardware the Visipedia team trained the neural networks on the iNaturalist database of images that have been labeled by the site's community of experts. Currently, iNaturalist has around 4,000,000 'verifiable' observations, i.e. observations that have all the necessary data quality attributes (eg. photos, locations, not pets) and have been vetted by experts and can be considered 'research grade'. These represent 100,000 species."
- [WEB-i Programmer-4]: quote: "The crowdsourced model generally works well according to Scott Loarie, iNaturalist's co-director. Half of users' mystery observations are identified within 2 days, even quicker if like Laurie your posts originate in California, where an identification can be made within an hour."
- [WEB-i Programmer-5]: quote: "iNaturalist determined that having at least 20 research grade observations was necessary to include a species in its model. While the above chart indicates there are 13,730 species that qualify, this number is probably closer to 10,000 species as steps were taken to ensure that each species had at least 20 distinct observers to control for observer effects."
- [WEB-i Programmer-6]: quote: "The new app uses the research grade observations to give a confident response about an animal's genus plus a more tentative suggestion of its species with its top 10 options. Initially it was correct with regard to genus 86% of the time and gave the correct species in its top 10 results 77% of the time. These numbers should improve as the model continues to be trained and, of course, the app itself contributes new observations and new confirmed identifications, leading to new species being added to the model at at rate of 1 every 1.7 hours."
- [WEB-i Programmer-7]: quote: "The iNaturalist app seems a very worthwhile addition to the range of software for identifying pants and animals. It provides a quick and easy way to record observations with photos and GPS locations and then access other people's observations from around the world and become part of the citizen science movement and the growing community of iNaturalist members of observers and experts."

MAIA [MA]

Journals

[MA-ZM18]:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6239313/pdf/pone.0207498.pdf

Zurowietz M, Langenkämper D, Hosking B, Ruhl HA, Nattkemper TW (2018) MAIA—A machine learning assisted image annotation method for environmental monitoring and exploration.PLoS ONE 13(11):e0207498.https:// doi.org/10.1371/journal.pone.0207498

- [MA-ZM18-1]: Under "Abstract", quote: "In the case of the marine environment, mobile platforms such as autonomous underwater vehicles (AUVs) are now equipped with high-resolution cameras to capture huge collections of images from the seabed. However, the timely evaluation of all these images presents a bottleneck problem as tens of thousands or more images can be collected during a single dive. This makes computational support for marine image analysis essential."
- [MA-ZM18-2]: Under "Abstract", quote: "Computer-aided analysis of environmental images (and marine images in particular) with machine learning algorithms is promising, but challenging and different to other imaging domains because training data and class labels cannot be collected as efficiently and comprehensively as in other areas. In this paper, we present Machine learning Assisted Image Annotation (MAIA), a new image annotation method for environmental monitoring and exploration that overcomes the obstacle of missing training data."
- [MA-ZM18-3]: Under "Abstract", quote: "The method uses a combination of autoencoder networks and Mask Region-based Convolutional Neural Network (Mask R-CNN), which allows human observers to annotate large image collections much faster than before."
- [MA-ZM18-4]: Under "Introduction", quote: "However, automatic object detection is still below expert performance in most contexts."
- [MA-ZM18-5]: Under "Introduction", quote: "While in computer science research areas such as media informatics or image databases, image annotation refers to the assignment of semantics to whole images, describing the content on a high level, image annotation in this context refers to the assignment of meaning (i.e. a class label selected from a given taxonomy) to a region of an image. Even with dedicated tools, the purely manual approach to this kind of image annotation is still a time consuming and error-prone task. In addition, owing to the complexity and diversity of organisms found in marine imagery, only domain experts are usually able to provide object detection and class labels with sufficient quality (i.e. sufficient inter-/intraobserver agreement and accuracy)."

- [MA-ZM18-6]: Under “Introduction”, quote: “During manual image annotation, most of the time is spent on locating objects of interest (OOI) rather than assigning a correct class label for the object.”
- [MA-ZM18-7]: Under “Introduction”, quote: “An object detection method that efficiently and effectively generates annotation candidates (i.e. regions showing potential OOI) for marine images to increase the speed and volume of manual annotations, has yet to be developed.”
- [MA-ZM18-8]: Under “Introduction”, quote: “In this paper, we present the Machine learning Assisted Image Annotation method (MAIA). The method aims to speed up manual image annotation by automating the process of object detection and instance segmentation of OOI in the images. Liberated from the task of object detection, human observers can concentrate purely on the classification of OOI. In the context of marine biology and environmental sciences, OOI could be particular species of interest or general megafauna like starfishes, holothurians or sponges. MAIA is based on a combination of two machine learning methods: Autoencoder networks (AEN) [9] and fully convolutional networks (FCN).”
- [MA-ZM18-9]: Under “Introduction”, quote: “Since the collection of training data in marine sciences is considerably expensive as it requires substantial experience and academic education, we developed MAIA to reduce time and effort on this end.”
- [MA-ZM18-10]: Under “Introduction”, quote: “Fully convolutional networks were introduced by Long et al. [10], who adapted a neural network for image classification into a fully convolutional network that performs semantic segmentation of an image. One of the more recent variants of FCNs is the Mask Region-based Convolutional Neural Network (Mask R-CNN) [11], which is capable of instance segmentation. Instance segmentation generates a bounding box, a class label and a pixel-accurate segmentation mask for each OOI in an image. Mask R-CNN can achieve impressive results, condition to the availability of a sufficient number of annotated training data samples, i.e. images with marked objects. Since the collection of training data in marine sciences is considerably expensive as it requires substantial experience and academic education, we developed MAIA to reduce time and effort on this end. In MAIA, Mask R-CNN is employed for instance segmentation and the resulting segmentation masks can be used as annotation candidates. To the best of our knowledge, we are the first to employ Mask R-CNN in the context of marine environmental monitoring and exploration.”
- [MA-ZM18-11]: Under “Introduction”, quote: “As the supervised Mask R-CNN needs a sufficient amount of training data, we apply unsupervised AEN for novelty detection to efficiently and effectively generate training proposals, which are points of potential OOI in the images that could be considered for training a Mask R-CNN

model. Based on the assumption that OOI are rare in the images, “background” pixels of the seabed can be regarded as common patterns and “interesting” pixels of objects are regarded as novel patterns. The concept of AEN was first presented by Baldi and Hornik in 1989 [9] and has been used in various contexts like dimensionality reduction [12, 13], human pose recovery [14] or cell nuclei detection [15]. AEN have been previously used for novelty detection as well [16, 17] but not in the context of FCN training data collection or environmental imaging.”

- [MA-ZM18-12]: Under “Introduction”, quote: “crowdsourcing can be employed to generate large amounts of image annotations for everyday objects.
- [MA-ZM18-13]: Under “Methods”, quote: “MAIA consists of four consecutive stages (see Fig 1), all of which are described in the following sections. In Stage I, unsupervised novelty detection generates an initial set of training proposalsTp, which are image patches (i.e. regions of an image) showing patterns different from the background. These training proposals contain potential OOI that could be used in a training dataset for the instance segmentation.”
- [MA-ZM18-14]: Under “Methods”, quote: “In Stage II,Tp is manually filtered to keep only those training proposals that actually show OOI relevant to the application context. In addition to that, the training proposals are manually refined with regard to their centroids and size, resulting in the dataset of training samples Ts⊆Tp.”
- [MA-ZM18-15]: Under “Methods”, quote: “In Stage III, the setTs is used to train a Mask R-CNN model for instance segmentation which is subsequently applied to produce a setAc of annotation candidates for a whole image dataset.”
- [MA-ZM18-16]: Under “Methods”, quote: “In Stage IV, these candidates are manually reviewed to remove false positives for the final setA⊆Ac of detected objects and their bounding boxes.”
- [MA-ZM18-17]: Under “Stage I”, quote: “Each resulting clusterUk contains images featuring global similarities, mostly dependent on the background sediment. For each cluster, one AEN is trained, all sharing the same architecture.”
- [MA-ZM18-18]: Under “Stage I”, quote: “As motivated above, we consider image patches that show the pure seabed or background as common patterns and patches that show an interesting object (like a shell or a starfish) as novelties. This definition is based on the assumption that interesting objects are rare in deep sea image datasets.”
- [MA-ZM18-19]: Under “Stage I”, quote: “The binary segmentation of all novelty mapsNi is used to extract the set of training proposalsTp from all images.”
- [MA-ZM18-20]: Under “Stage II”, quote: “Stage II: Manual filtering and refinement of training proposals”.
- [MA-ZM18-21]: Under “Stage II”, quote: “A considerable amount of the training proposals inTp may show patterns that do not represent anything of interest for the domain experts. Thus, we apply a quick manual filtering step, where a human

observer selects only those training proposals that contain OOI… For each training proposal, the human observer has to determine if it contains (part of) an OOI or not."

- [MA-ZM18-22]: Under "Stage II", quote: "The filtering has been implemented as single patch classification as defined in the RecoMIA guidelines [3], so only the isolated image regions of the training proposals are displayed to the human observer instead of the complete images. This saves the time the human observer would need to screen the complete images for multiple regions of interest. For each training proposal, the human observer has to determine if it contains (part of) an OOI or not. To accelerate manual filtering, the label review grid overview tool Largo of BIIGLE 2.0 is used (see S1 Fig in the supporting information). Training proposalsyj are displayed in a regular grid in descending order of their novelty scores η(yj). This allows human observers to spot OOI very quickly and to review a large number of training proposals in a very short time (see Fig 1e)."
- [MA-ZM18-23]: Under "Stage II", quote: "The sorting by novelty score is a similar technique than the saliency ranking described by [7]. Starting from the training proposals with the highest novelty score, a human observer can stop reviewing once a sufficient amount of training proposals have been selected. In this work we define 600 as the limit for the required number of selected training proposals for each object class."
- [MA-ZM18-24]: Under "Stage II", quote: "The performance of an FCN-based instance segmentation method like Mask R-CNN is crucially dependent on the quality of the training dataset. If the samples in the training dataset are of low quality, i.e. with many discrepancies between interesting and non-interesting image regions, the performance of instance segmentation may be very poor. To obtain a setTs with training samples of appropriate quality from the training proposals Tp, a manual refinement step is performed after the filtering. The filtered training proposals are shown to a human observer, each with a suggested centroid and size (i.e. a circle) that marks the OOI. The observer can modify the circle position or size so it closely fits the position and size of the OOI (see Fig 1f). To further accelerate the refinement, we use the volume label review tool Volare of BIIGLE 2.0 (see S2 Fig in the supporting information). With Volare, the viewport of the annotation tool jumps directly from one circle to the next, saving the time a human observer would need to look for and zoom in to each circle on an image."
- [MA-ZM18-25]: Under "Stage III", quote: "The filtered and refined training proposals are used to build a dataset of training samples."
- [MA-ZM18-26]: Under "Stage III", quote: "Due to the fixed number of images that are considered to generate a training dataset, there may be only a few hundred or less samples of a particular class of OOI. For comparison, in datasets like MS COCO [6] there are many thousands of instances for each object class. To

increase the number of object instances that are available for training, we boost the training dataset (see Fig 1h). Details on the boosting we apply can be found in the supporting information (see S2 Text)."

- [MA-ZM18-27]: Under "Stage III", quote: "In MAIA, the training is performed with the set of boosted training samples (see Fig 1i) and the default configuration of the Mask R-CNN implementation."
- [MA-ZM18-28]: Under "Stage IV", quote: "Stage IV: Annotation candidate review To eliminate the false positive detections, which mark regions of the images that show no OOI, the annotation candidates Ac are manually reviewed in the last step. This is done analogously to the manual filtering of training proposals in Stage II. Each annotation candidate is shown as an image patch in a regular grid. A human observer then selects all candidates that are true positives, i.e. those that mark an OOI. This yields the final setAof annotations."
- [MA-ZM18-29]: Under "3 Datasets", quote: "3 Datasets We evaluated MAIA on three marine image datasets that were collected in different research projects. From each datasetΓ2{JC77, PAP, SO242}, we extracted 500 random images as training subsetTΓ. For each subset, MAIA was performed to train a Mask R-CNN model. The detection performance of the model was evaluated on another 50 random images as validation subsetVΓ for each dataset. The images of the validation subset have been fully annotated using "traditional" methods."
- [MA-ZM18-30]: Under "Conclusion", quote: " Based on these results, we conclude that MAIA is a promising method for image annotation in all environmental monitoring and exploration scenarios with large image collections."

Virus Spot

Web

[WEB-Microbiology]: https://naturemicrobiologycommunity.nature.com/posts/44573-citizen-scientists-enlisted-to-spot-viruses

- [WEB-Microbiology-1]: quote: "It's a big data cliché that one experiment can give more information than one scientist can analyse in a lifetime. Not only is computing power and storage an issue, but even the simplest tasks can take an extremely long time on large datasets."
- [WEB-Microbiology-2]: quote: "The idea is simple. Citizen scientists are shown cryo-EM images of viruses and then asked to find them in a series of images from real experiments. This input will help to train an AI algorithm which will be able to sort cry-EM data in the future and streamline the data analysis pipeline."

- [WEB-Microbiology-3]: quote: “The ultimate goal is to completely automate segmentation using advances in deep learning. Such methods require significant quantities of already segmented data to train the systems we use.”
- [WEB-Microbiology-4]: quote: “To build segmented data for this development, Zooniverse will offer members of the public across the globe the chance to partake in segmenting datasets to help researchers.”
- [WEB-Microbiology-5]: quote: “Artificial Intelligence has begun to have a massive impact on the world in the last few years, from beating humans in games such as Go, to the amazing advances in self-driving cars. These dramatic developments have been aided by the availability of vast quantities of data with which AI systems can be trained with.”
- [WEB-Microbiology-6]: quote: “Researchers spend much of their time manually processing their data and this is an area where AI could be heavily used.”
- [WEB-Microbiology-7]: quote: “for machine learning to be possible, we need human input to guide the process and this is where members of the public can make a huge difference to our work.”
- [WEB-Microbiology-8]: quote: “This project aims to address these issues … in a standardised way that can be used to automate the process in the future, thereby helping fasten the analysis process from weeks to days or less.”

[WEB-Science Scribbler]: https://www.zooniverse.org/projects/markbasham/science-scribbler-virus-factory/about/research

- [WEB-Science Scribbler-1]: quote: “... developing novel software to reduce the difficulty of analysing the 3D image data we're interested in.”
- [WEB-Science Scribbler-2]: https://www.zooniverse.org/about/faq quote: “**Why do researchers need your help? Why can't computers do these tasks?** Humans are better than computers at many tasks. For most Zooniverse projects, computers just aren’t good enough to do the required task, or they may miss interesting features that a human would spot - this is why we need your help. Some Zooniverse projects are also using human classifications to help train computers to do better at these research tasks in the future. When you participate in a Zooniverse project, you are contributing to real research.”
- [WEB-Science Scribbler-3]: https://www.zooniverse.org/about/faq quote: “Human beings are really good at pattern recognition tasks, so generally your first guess is likely the right one.”
- [WEB-Science Scribbler-4]: https://www.zooniverse.org/about quote: “With the help of Zooniverse volunteers, researchers can analyze their information more quickly and accurately than would otherwise be possible, saving time and resources, advancing the ability of computers to do the same tasks, and leading to faster progress and understanding of the world, getting to exciting results more quickly.”

- [WEB-Science Scribbler-5]: https://www.zooniverse.org/about quote: “Our projects combine contributions from many individual volunteers, relying on a version of the ‘wisdom of crowds’ to produce reliable and accurate data. By having many people look at the data we often can also estimate how likely we are to make an error. The product of a Zooniverse projects is often exactly what’s needed to make progress in many fields of research.”
- [WEB-Science Scribbler-6]: https://www.zooniverse.org/about quote: “A significant amount of this research takes place on the Zooniverse discussion boards, where volunteers can work together with each other and with the research teams. These boards are integrated with each project to allow for everything from quick hashtagging to in-depth collaborative analysis. There is also a central Zooniverse board for general chat and discussion about Zooniverse-wide matters. Many of the most interesting discoveries from Zooniverse projects have come from discussion between volunteers and researchers. We encourage all users to join the conversation on the discussion boards for more in-depth participation.”
- [WEB-Science Scribbler-7]: https://www.zooniverse.org/get-involved quote: “Volunteers also help us test projects before they are launched to check that they work properly. This involves working through some classifications on the beta project to check that it works, looking for any bugs, and filling out a questionnaire at the end. This helps us find any issues in the project that need resolving and also assess how suitable the project is for the Zooniverse.”

Milky Way

Journals

[MM-BC]: https://arxiv.org/pdf/1406.2692v1.pdf

Beaumont, C. N., Goodman, A. A., Kendrew, S., Williams, J. P., & Simpson, R. (2014). The Milky Way Project: Leveraging Citizen Science and Machine Learning to Detect Interstellar Bubbles. *The Astrophysical Journal Supplement Series*. Vol. 214, No. 1, pp. 1-18. doi:10.1088/0067-0049/214/1/3

- [MM-BC-1]: Under “Abstract”, quote: “We present Brut, an algorithm to identify bubbles in infrared images of the Galactic midplane. Brut is based on the Random Forest algorithm, and uses bubbles identified by > 35,000 citizen scientists from the Milky Way Project to discover the identifying characteristics of bubbles in images from the Spitzer Space Telescope.”
- [MM-BC-2]: Under “Abstract”, quote: “We demonstrate that Brut’s ability to identify bubbles is comparable to expert astronomers. We use Brut to reassess the

bubbles in the Milky Way Project catalog, and find that 10−30% of the objects in this catalog are non-bubble interlopers. Relative to these interlopers, high-reliability bubbles are more confined to the mid plane, and display a stronger excess of Young Stellar Objects along and within bubble rims. Furthermore, Brut is able to discover bubbles missed by previous searches – particularly bubbles near bright sources which have low contrast relative to their surroundings. Brut demonstrates the synergies that exist between citizen scientists, professional scientists, and machine learning techniques. In cases where "untrained" citizens can identify patterns that machines cannot detect without training, machine learning algorithms like Brut can use the output of citizen science projects as input training sets, offering tremendous opportunities to speed the pace of scientific discovery.

- [MM-BC-3]: Under "Abstract", quote: "A hybrid model of machine learning combined with crowdsourced training data from citizen scientists can not only classify large quantities of data, but also address the weakness of each approach if deployed alone."
- [MM-BC-4]: Under "Introduction", quote: "Unfortunately, due to their complex morphologies, bubbles—like many features of the ISM—are difficult to identify and analyze. Existing catalogs of spatially extended bubbles have typically been identified visually. This has two main disadvantages. First, it is cumbersome and increasingly infeasible as data sets grow ever larger. Second, manual classification is inherently subjective and non-repeatable; humans are susceptible to fatigue, boredom, and subtle selection biases whose impact on the resulting catalog is difficult to calibrate. The problems associated with manual bubble detection are germane to many analyses with a subjective component.Machine learning techniques represent a promising solution to these problems."
- [MM-BC-5]: Under "Introduction", quote: "Our goal in this work is to apply machine learning techniques to the task of bubble detection, and to evaluate the potential of this approach."
- [MM-BC-6]: Under "Introduction", quote: "Using a catalog of known bubbles identified by the citizen scientists of the Milky Way Project (Simpson et al. 2012), we "teach" an algorithm to identify bubbles in image data from the *Spitzer Space Telescope*."
- [MM-BC-7]: Under "Introduction", quote: "we use a set of expert classifications to measure Brut's performance at bubble detection."
- [MM-BC-8]: Under "Introduction", quote: "we demonstrate that this detector (which is Brut) produces useful probabilistic estimates for whether any particular image contains a bubble—these probabilities correlate well with how astronomers classify similar regions."

- [MM-BC-9]: Under "Introduction", quote: "We use this detector (Brut) to look for biases and incompleteness in existing bubble catalogs."
- [MM-BC-10]: Under "Introduction", quote: "This analysis yields a new catalog of high-probability bubbles, and we explore how the ensemble properties of this catalog differ from the old catalog."
- [MM-BC-11]: Under "Introduction", quote: "we apply Brut to the task of discovering bubbles missing from current catalogs."
- [MM-BC-12]: Under "Introduction", quote: "The citizen science effort produced a dramatically larger catalog of ~5000 objects, nearly 10 times the number in the catalog of ~600 shells cataloged by the four astronomers of the Churchwell et al. (2006, 2007) surveys."
- [MM-BC-13]: Under "Introduction", quote: "In terms of accuracy, humans still outperform computers in most image-based pattern recognition tasks (e.g., Zhang & Zhang 2010). Because of this, morphologically complex structures in the ISM (including supernova remnants, outflows, bubbles, H ii regions, molecular and infrared dark clouds, and planetary nebulae) are still traditionally cataloged manually."
- [MM-BC-14]: Under "Introduction", quote: "Human classification has several disadvantages, however. First, human classification is time-consuming, and people-hours are a limited resource. Even by enlisting large communities of citizen scientists, data from next generation surveys will be too large to search exhaustively. For example, the >35,000 citizen scientists of the MWP classified roughly 45 GB of image data from *Spitzer*. Current and next-generation astronomical data sets are many thousands of times larger than this, suggesting tens of millions of citizen scientists would be needed for similar exhaustive searches through tera- and petabyte data sets. Second, many scientifically important tasks are not suitable for enlisting the public. Part of the appeal of the MWP is due to the fact that the *Spitzer* images are beautiful, contain many bubbles, and are compelling to browse through. Searches for very rare objects, or tasks where the scientific justification is less apparent to a citizen scientist, may be less likely to entice large volunteer communities. Raddick et al. (2013) considers the motivations of citizen scientists in greater detail. Finally, manual classification is not easily repeatable, and hard to calibrate statistically. For example, it is unknown how well the consensus opinion among citizen scientists corresponds to consensus among astronomers. The MWP catalog does not include any estimate of the probability that each object is a real bubble, as opposed to another structure in the ISM. Automatic classifications driven by machine learning techniques nicely complement human classification."
- [MM-BC-15]: Under "Introduction", quote: "Such an approach easily scales to large data volumes and is immune to some of the factors that affect humans, like

boredom and fatigue. Furthermore, because algorithmic classifications are systematic and repeatable, they are easier to interpret and statistically characterize. Despite the structural complexity of the ISM, Beaumont et al. (2011) demonstrated that automatic classification algorithms can discriminate between different ISM structures based upon morphology."

- [MM-BC-16]: Under "Introduction", quote: "It is difficult to robustly classify ISM structures using templates, due to the heterogeneity and irregularity of the ISM—simple shapes like expanding spheres are often poor approximations to the true ISM."
- [MM-BC-17]: Under "Introduction", quote: "The automatic classifier is capable of producing quantitative reliability estimates for each bubble in the MWP catalog, potentially flagging non-bubble interlopers and leading to a cleaner catalog."
- [MM-BC-18]: Under "Introduction", quote: "We can search for bubbles not detected by MWP citizen scientists."
- [MM-BC-19]: Under "Introduction", quote: "We can treat this task as a case study for complex classification tasks in future data sets, where exhaustive manual classification will not be feasible."
- [MM-BC-20]: Under "Classification method", quote: "Our goal is to use the set of citizen-scientist-identified bubbles to build an automatic detector that, when presented with a region of the sky in the *Spitzer* glimpse and mipsgal surveys, accurately determines whether or not the image contains a bubble. Our approach, which we name Brut,[4] is an example of a supervised learning problem. Here is a brief overview of the task. 1. Build a representative training set of examples of bubble and non-bubble images. This will be derived from the MWP data set. 2. Convert each example to a numerical *feature vector* that describes each object, and captures the difference between bubbles and non-bubbles. 3. Feed the training data to a learning algorithm to build a model. 4. Use a subset of the examples not used during training to optimize the tunable parameters (so-called *hyper-parameters*) of the learning algorithm."
- [MM-BC-21]: Under "Classification method", quote: "Brut uses the Random Forest classification algorithm (Breiman 2001) to discriminate between images of bubbles and non-bubbles. Random Forests are aggregates of a simpler learning algorithm called a decision tree. A decision tree is a data structure which classifies feature vectors by computing a series of constraints, and propagating vectors down the tree based on whether these constraints are satisfied."
- [MM-BC-22]: Under "Introduction", quote: "Decision trees are constructed using an input training set of pre-classified feature vectors. During tree construction, a quality heuristic is used to rate the tree. A few heuristics are common, which consider both the classification accuracy and the complexity of the tree itself—highly complex trees are more prone to over-fitting, and thus disfavored... Decision

trees are constructed one node at a time, in a "greedy" fashion. That is, at each step in the learning process, a new boolean constraint is added to the tree to maximally increase the score of the quality heuristic. This process repeats until the quality heuristic reaches a local maximum."

- [MM-BC-23]: Under "Classification method", quote: "The individual "objects" that Brut classifies are square regions of the sky, and the goal of the classifier is to determine whether each region is filled by a bubble. Each field is identified by three numbers: the latitude and longitude of the center of the field, and the size of the field. We decompose the glimpse survey coverage into fields of 18 different sizes, logarithmically spaced from $0\overset{\circ}{.}02$ to $0\overset{\circ}{.}95$. At each size scale, we define and classify an overlapping grid of fields. Neighboring tiles in this grid are separated by one-fifth of the tile size."
- [MM-BC-24]: Under "Classification method", quote: "As a preprocessing step, we extracted two-color postage stamps for each field (at 8 μm and 24 μm), and resampled these postage stamps to (40 × 40) pixels. Following a scheme similar to Simpson et al. (2012), these images were intensity clipped at the 1st and 97th percentiles, normalized to maximum intensity of 1, and passed through a square root transfer function. The intensity scaling tends to do a good job of emphasizing ISM structure, making bubbles more visible to the eye. Likewise, the (40 × 40) pixel resolution was chosen because it is reasonably small, yet has enough resolution that postage stamps of known bubbles are still recognizable as such by humans. Figure 4 shows four preprocessed fields toward known bubbles.[5] The goal of preprocessing is to standardize the appearance of bubbles as much as possible, across different size scales and ambient intensities. All subsequent stages of Brut work exclusively with these images, as opposed to the unscaled pixel data."
- [MM-BC-25]: Under "Classification method", quote: "Building a Random Forest requires providing a set of pre-classified feature vectors… we manually curated a list of 468 objects in the MWP catalog which were clear examples of bubbles."
- [MM-BC-26]: Under "Classification method", quote: "A better set of negative examples includes more fields containing structure (Figure 6). We built such a collection in a bootstrapped fashion. We began with a random set of negative fields, distributed uniformly in latitude and longitude, with sizes drawn from the size distribution of the positive training set. We trained a classifier with these examples and used it to scan 20,000 bubble-free regions. We then discarded half of the initial negative examples (those classified most confidently as not containing bubbles), and replaced them with a random sample of the misclassified examples. We repeated this process several times, but found that one iteration was usually sufficient to build a good set of training data."

- [MM-BC-27]: Under “Classification method”, quote: “Our final training set consists of 468 examples of bubbles identified by the Milky Way Project, and 2289 examples of non-bubble fields.”
- [MM-BC-28]: Under “Classification method”, quote: “Instead of building a single Random Forest classifier, we trained three forests on different subsets of the sky. Each forest was trained using examples from two-thirds of the glimpse survey area, and used to classify the remaining one-third. The motivation for doing this is to minimize the chance of over-fitting, by ensuring that the regions of the sky used to train each classifier do not overlap the regions of sky used for final classification.”
- [MM-BC-29]: Under “Classification method”, quote: “We explored the impact of these hyperparameter settings via cross validation. During cross validation, we split the training examples into two groups: a primary set with 154 bubble examples and 1462 non-bubble examples, and a validation set of 157 bubble and 10,000 non-bubble examples.”
- [MM-BC-30]: Under “Classification method”, quote: “We trained a Random Forest with a particular choice of hyper parameters, and measured the accuracy and false-positive rate of the classifier on the validation set...Once we converged on an optimal set of hyperparameters, we trained the final three Random Forest classifiers using the scheme discussed above.”
- [MM-BC-31]: Under “Classification method”, quote: “To classify a region after training, we compute the feature vector and dispatch it to one of the three Random Forests (depending on longitude, as discussed in Section 2.5). The forest produces a classification score between −1 and 1. This number is equal to the fraction of trees in the forest which predict the feature vector is a bubble, minus the fraction of trees which predict it is not. This score provides more information than a simple binary classification, as it gives some sense of the confidence of the classification. Furthermore, one can adjust the threshold that defines when an object is considered to be a bubble.”
- [MM-BC-32]: Under “Expert validation”, quote: “To evaluate the performance of Brut, we need a set of "ground truth" classifications. However, to some extent identifying bubbles in *Spitzer* images is inherently subjective… To better measure the level of expert consensus in bubble identification, we conducted a small online survey. The astronomer participants of this survey were presented with a sequence of 92 *Spitzer* images at three zoom levels and two contrast settings. They were asked to assign each image to one of three categories: clear examples of bubbles or H ii regions, ambiguous or irregular bubbles, or non-bubbles.”
- [MM-BC-33]: Under “Expert validation”, quote: “Of the 92 images in the expert survey, 45 were a random subset of objects in the MWP catalog (the remaining fields are discussed in the next section)...The expert reclassifications of this sample of MWP objects can be used to convert a raw score like the joint score to

a calibrated probability that an expert would classify an object as a bubble, given that score. To achieve this, we perform a logistic regression against each of the three scores (hit rate, Brut score, and joint score). The logistic regression only considers whether the plurality category for each object is a bubble—the consensus information is not used."

- [MM-BC-34]: Under "Expert validation", quote: "The expert reclassifications of this sample of MWP objects can be used to convert a raw score like the joint score to a calibrated probability that an expert would classify an object as a bubble, given that score. To achieve this, we perform a logistic regression against each of the three scores (hit rate, Brut score, and joint score). The logistic regression only considers whether the plurality category for each object is a bubble—the consensus information is not used."
- [MM-BC-35]: Under "Expert validation", quote: "The remaining 47 regions in the expert survey were selected randomly from the full set of fields analyzed during Brut's full scan of the *Spitzer* images. A fully random sample from this collection is uninteresting, since the majority of fields are blank areas of the sky. Instead, these 47 images are approximately uniformly distributed in the Brut confidence score. We refer to these images as the "uniform" sample."
- [MM-BC-36]: Under "Blind search", quote: "The previous section focused on using Brut to reassess bubbles previously identified by citizen scientists. We have demonstrated that Brut is successful at identifying the high-reliability subset of the MWP catalog and, conversely, at flagging probable interlopers in the catalog. The result is a purer statistical sample of bubbles in the Milky Way."
- [MM-BC-37]: Under "Blind search", quote: "Discovering bubbles without knowing the citizen-science hit rate at each location is a harder task; Brut does not benefit from complementary information about how citizen scientists classify a particular region. However, this task is relevant to future projects where machine learning techniques assist manual search. For applications where exhaustive human search is infeasible, machine learning algorithms can conduct exhaustive searches and flag interesting candidate objects for human attention or followup observation."
- [MM-BC-38]: Under "Next steps", quote: "The success of Brut demonstrates the potential synergies that exist between machine learning, professional scientists, and citizen scientists. Note the complementary strengths and weaknesses of each resource. 1. Professional scientists are best-suited to perform nuanced classification tasks that require domain-specific knowledge. They are also the most resource-limited. 2. Citizen scientists outnumber professional scientists by orders of magnitude (in the case of Bubble detection, the factor is nearly 10,000:1). They are equally capable with the generic aspects of pattern recognition, but do not possess the domain expertise of professionals. Furthermore, curious citizen

scientists are well-situated for serendipitous discovery of unusual objects (Lintott et al. 2009; Cardamone et al. 2009). 3. Supervised machine learning algorithms have no a priori pattern recognition ability, and require external training. However, once supplied with this information, computer-driven analyses are reproducible and extremely scalable.

Mindcontrol [MC]

Journals

- **[MC-KA19]:** https://doi.org/10.3389/fninf.2019.00029
  Keshavan, A., Yeatman, J. D., & Rokem, A. (2019). Combining Citizen Science and Deep Learning to Amplify Expertise in Neuroimaging. Frontiers in Neuroinformatics, 13, 29, pp. 1-13. ISSN: 1662-5196. DOI: 10.3389/fninf.2019.00029.
  - [MC-KA19-1]: Under “Abstract”, quote: “Big Data promises to advance science through data-driven discovery. However, many standard lab protocols rely on manual examination, which is not feasible for large-scale datasets. Meanwhile, automated approaches lack the accuracy of expert examination. We propose to (1) start with expertly labeled data, (2) amplify labels through web applications that engage citizen scientists, and (3) train machine learning on amplified labels, to emulate the experts. Demonstrating this, we developed a system to quality control brain magnetic resonance images. Expert-labeled data were amplified by citizen scientists through a simple web interface. A deep learning algorithm was then trained to predict data quality, based on citizen scientist labels. Deep learning performed as well as specialized algorithms for quality control (AUC = 0.99). Combining citizen science and deep learning can generalize and scale expert decision making; this is particularly important in disciplines where specialized, automated tools do not yet exist.”
  - [MC-KA19-2]: Under the headline of “Overview”, quote: “Figure 1 shows an overview of the procedure and provides a summary of our results. At the outset, a group of neuroimaging experts created a gold-standard quality control dataset on a small subset of the data (*n* = 200), through extensive visual examination of the full 3D volumes of the data.
  - [MC-KA19-3]: Under the headline of “Overview”, quote: “In parallel, citizen scientists were asked to “pass” or “fail” two-dimensional axial slices from the full dataset (*n* = 722; five slices from each brain) through a web application called braindr that could be accessed through a desktop, tablet

or mobile phone (https://braindr.us). Amplified labels, that range from 0 (fail) to 1 (pass), were generated from citizen scientist ratings."

- [MC-KA19-4]: Under the headline of "Overview", quote: "Two different receiver operating characteristic (ROC) curves were generated to assess the performance of citizen scientists. The first used simply the averaged ratings for each brain across the citizen scientists that rated this brain. The other used the labels that were generated by a classifier that weights ratings more heavily for citizen scientists who more closely matched the experts in the subset rated by both (gold-standard)."
- [MC-KA19-5]: Under the headline of "Overview", quote: "Next, a neural network was trained to predict the weighted labels."
- [MC-KA19-6]: Under the headline of "Aggregating Citizen Scientist Ratings to Emulate Expert Labels", quote: "Citizen scientists were given a brief explanation of how to look at MRI images, and then saw six examples in the braindr interface demonstrating images that should pass and fail quality control based on experts' ratings."
- [MC-KA19-7]: Under the headline of "Aggregating Citizen Scientist Ratings to Emulate Expert Labels", quote: "Given that training was very brief (<1 min) it is no surprise that citizen scientists who rated images through the braindr web application differed substantially in terms of how well their ratings matched the experts' ratings on the full gold-standard subset."
- [MC-KA19-8]: Under the headline of "Aggregating Citizen Scientist Ratings to Emulate Expert Labels", quote: "In order to capitalize on citizen scientists to amplify expert ratings to new data, a weighting of each citizen scientist was learned based on an accurate match to expert agreement in slices from the gold-standard set."
- [MC-KA19-9]: Under the headline of "Aggregating Citizen Scientist Ratings to Emulate Expert Labels", quote: "We used the XGBoost algorithm (Chen and Guestrin, 2016), an ensemble method that combines a set of weak learners (decision trees) to fit the gold-standard labels based on a set of features. This algorithm was chosen because it is able to handle missing data. Each image was rated 18.9 times on average and the features were the average rating of the slice image from each citizen scientist. Since some images were viewed and rated more than once, the image ratings could vary between 1 (always "pass") and 0 (always "fail"). We then used the weights to combine the ratings of the citizen scientists and predict the left out test set. Figure 2A shows ROC curves of classification on the left-out test set for different training set sizes, compared to the ROC curve of a baseline model in which equal weights were assigned to each citizen scientist. We see an improvement in the AUC of the XGBoosted labels

(0.97) compared to the AUC of the equi-weighted labels (0.95). Using the model trained on two-thirds of the gold standard data (*n* = 670 slices), we extracted the probability scores of the classifier on all slices (see Figure 2B). The distribution of probability scores in Figure 2B matches our expectations of the data; a bimodal distribution with peaks at 0 and 1, reflecting that images are mostly perceived as "passing" or "failing." The XGBoost model also calculates a feature importance score (F). F is the number of times that a feature (in our case, an individual citizen scientist) has split the branches of a tree, summed over all boosted trees. Figure 2C shows the feature importance for each citizen scientist, and Figure 2D shows the relationship between a citizen scientist's importance compared to the number of images they rated. In general, the more images a citizen scientist rates, the more important they are to the model. However, there are still exceptions where a citizen scientist rated many images and their ratings were incorrect or unreliable, so the model gave them less weight during aggregation."

- [MC-KA19-10]: Under the headline of "2.3. Training Deep Learning to Automate Image Labeling", quote: "Thus, we trained a deep learning model to predict the XGBoosted labels that were based on aggregated citizen scientist ratings. A VGG16 neural network (Simonyan and Zisserman, 2014) pretrained on the ImageNet challenge dataset (Russakovsky et al., 2015) was used: we removed the top layer of the network, and then trained a final fully-connected layer followed by a single node output layer. The training of the final layer was run for 50 epochs and the best model on the validation set was saved. To estimate the variability of training, the model was separately trained through 10 different training courses, each time with a different random initialization seed. Typically, training and validation loss scores were equal at around 10 epochs, after which the model usually began to overfit (training error decreased, while validation error increased, see Figure 3A). In each of the 10 training courses, we used the model with the lowest validation error for inference on the held out test set, and calculated the ROC AUC. AUC may be a problematic statistic when the test-set is imbalanced (Saito and Rehmsmeier, 2015), but in this case, the test-set is almost perfectly balanced (see Methods). Thus, we found that a deep learning network trained on citizen scientist generated labels was a better match to expert ratings than citizen scientist generated labels alone: the deep learning model had an AUC of 0.99 (+/− standard deviation of 0.12, see Figure 3B)."
- [MC-KA19-11]: Under the headline of "2.4. Crowd Amplification and Deep Learning Strategy Performs as Well as a Specialized QC Algorithm", quote: "We validated our generalized approach of crowd-amplification and deep

learning by comparing classification results against an existing, specialized algorithm for QC of T1 weighted images, called MRIQC (Esteban et al., 2017). The features extracted by MRIQC are guided by the physics of MR image acquisition and by the statistical properties of images. An XGBoost model was trained on the features extracted by MRIQC on a training subset of gold-standard images, and evaluated on a previously unseen test subset. The AUC was also 0.99, matching the performance of our crowd-trained deep learning model."

○ [MC-KA19-12]: Under the headline of "4.1. The Healthy Brain Network Dataset", quote: "Mindcontrol raters, who were all neuroimaging researchers with substantial experience in similar tasks, provided informed consent, including consent to publicly release these ratings. Mindcontrol raters were asked to pass or fail images after inspecting the full 3D volume, and provide a score of their confidence on a 5 point Likert scale, where 1 was the least confident and 5 was the most confident. Mindcontrol raters received a point for each new volume they rated, and a leaderboard on the homepage displayed rater rankings. The ratings of the top 4 expert raters (including the lead author) were used to create a gold-standard subset of the data."

○ [MC-KA19-13]: Under the headline of "4.1. The Healthy Brain Network Dataset", quote: "The gold-standard subset of the data was created by selecting images that were confidently passed or confidently failed (confidence equal or larger than 4) by the 4 expert raters. In order to measure reliability between expert raters, the ratings of the second, third, and fourth expert rater were recoded to a scale of –5 to 5 (where –5 is confidently failed, and 5 is confidently passed). An ROC analysis was performed against the binary ratings of the lead author on the commonly rated images, and the area under the curve (AUC) was computed for each pair. An average AUC, weighted by the number of commonly rated images between the pair, was 0.97, showing good agreement between expert raters. The resulting gold-standard dataset consisted of 200 images. Figure 5 shows example axial slices from the gold-standard dataset. The gold-standard dataset set contains 100 images that were failed by experts, and 100 images that were passed by experts."

○ [MC-KA19-14]: Under the headline of "introduction", quote: "We've seen a shift from desktop computers to cyberinfrastructure (Van Horn and Toga, 2013), from small studies siloed in individual labs to an explosion of data sharing initiatives (Ferguson et al., 2014; Poldrack and Gorgolewski, 2014), from idiosyncratic data organization and analysis scripts to standardized file structures and workflows (Gorgolewski et al., 2016, 2017b), and an overall shift in statistical thinking and computational methods (Fan et al., 2014) that can accommodate large datasets. But one often overlooked aspect of our protocols in neuroimaging has not yet evolved to meet the needs of Big Data: expert decision making. Specifically, decisions made by scientists with expertise in neuroanatomy and MRI methods (i.e., neuroimaging experts) through visual inspection of imaging data cannot be accurately scaled to large datasets."

- [MC-KA19-15]: Under the headline of "introduction", quote: "On large datasets, especially longitudinal multisite consortium studies, these expert decisions cannot be reliably replicated because the timeframe of these studies is long, individual experts get fatigued, and training teams of experts is time consuming, difficult and costly. As datasets grow to hundreds of thousands of brains it is no longer feasible to depend on manual interventions."
- [MC-KA19-16]: Under the headline of "introduction", quote: "One solution to this problem is to train machines to emulate expert decisions. However, there are many cases in which automated algorithms exist, but expert decision-making is still required for optimal results. For example, a variety of image segmentation algorithms have been developed to replace manual ROI editing, with Freesurfer (Fischl, 2012), FSL (Patenaude et al., 2011), ANTS (Avants et al., 2011), and SPM (Ashburner and Friston, 2005) all offering automated segmentation tools for standard brain structures. But these algorithms were developed on a specific type of image (T1-weighted) and on a specific type of brain (those of healthy controls). Pathological brains, or those of children or the elderly may violate the assumptions of these algorithms, and their outputs often still require manual expert editing."
- [MC-KA19-17]: Under the headline of "introduction", quote: "Another fundamental step in brain image processing that still requires expert examination is quality control. There are several automated methods to quantify image quality, based on MRI physics and the statistical properties of images, and these methods have been collected under one umbrella in an algorithm called MRIQC (Esteban et al., 2017). However, these methods are designed for specific types of MR images, and cannot generalize to other types of image acquisitions, let alone data from other scientific domains. To address all of these cases, and scale to new, unforeseen challenges, we need a general-purpose framework that can train machines to emulate experts for any purpose, allowing scientists to fully realize the potential of Big Data."

- **[MC-KA18]:** https://www.sciencedirect.com/science/article/pii/S1053811917302707
Keshavan, A., Datta, E., McDonough, I. M., Madan, C. R., Jordan, K., & Henry, R. G. (2018). Mindcontrol: A web application for brain segmentation quality control. *NeuroImage*, 170, pp. 365-372. DOI: 10.1016/j.neuroimage.2017.03.055

    - [MC-KA18-1]: quote: "We propose an open source web-based brain quality control application called Mindcontrol: a dashboard to organize, QC, annotate, edit, and collaborate on neuroimaging processing. Mindcontrol provides an intuitive interface for examining distributions of descriptive measures from neuroimaging pipelines (e.g., surface area of the right insula), and viewing the results of segmentation analyses using the Papaya.js volume viewer (https://github.com/rii-mango/Papaya)."
    - [MC-KA18-2]: quote: "Researchers must be able to QC outputs from any type of neuroimaging software package, so Mindcontrol was specified to flexibly accommodate any file organization structure, with configurable "modules" that can contain any type of descriptive statistics and 3D images."
    - [MC-KA18-3]: quote: "Finally, changes to the database (like the addition of new images), changes in descriptive measures, and new edits/annotations, should be reflected in the application in real-time to foster collaboration."
    - [MC-KA18-4]: under the section "2.3. Client-side features", quote: "The user interface consists of a dashboard view and an imaging view, as shown in Fig. 2, Fig. 3, respectively. The primary dashboard view consists of processing module sections, a query controller, data tables, and descriptive statistic visualizations. Each entry in the table is a link that, when clicked, filters all tables on the page. The filters or queries can be saved, edited, and loaded in the query controller section, as shown in Fig. 4."
    - [MC-KA18-5]: under "Abstract", quote: "Assessment with the human eye is vital to correct various errors inherent to all currently available segmentation algorithms. Manual quality assurance becomes methodologically difficult at a large scale - a problem of increasing importance as the number of data sets is on the rise. To make this process more efficient, we have developed Mindcontrol, an open-source web application for the collaborative quality control of neuroimaging processing outputs."

- **[MC-EO19]:** https://doi.org/10.1038/s41597-019-0035-4
Esteban, O., Blair, R.W., Nielson, D.M. *et al.* (2019). Crowdsourced MRI quality metrics and expert quality annotations for training of humans and machines. *Sci Data* 6, *Article* 30 (2019).

  - [MC-EO19-1]: under the section of “Background and Summary”, quote: “As described previously, rating the quality of every image in large databases is an arduous, unreliable, and costly task. The convergence of limited size of samples annotated for quality and the labels noise preclude the definition of normative, standard values for the IQMs that work well for any dataset, and also, the generalization of machine learning solutions.”
  - [MC-EO19-2]: under the section of “Background and Summary”, quote: “By collecting several ratings per screened entity, they were able to effectively minimize the labels noise problem with the averaging of expert ratings...In sum, automating QC requires large datasets collected across sites, and rated by many individuals in order to ensure generalizability.”

WEB

- **[WEB-MC-Omictools]**: https://omictools.com/mindcontrol-tool?t=tab-tool-variant-1
  - [WEB-MC-Omictools-1]: quote: “Provides an open-source web application for the collaborative quality control of neuroimaging processing outputs. The Mindcontrol platform consists of a dashboard to organize data, descriptive visualizations to explore the data, an imaging viewer, and an in-browser annotation and editing toolbox for data curation and quality control. Mindcontrol is flexible and can be configured for the outputs of any software package in any data organization structure. Example configurations for three large, open-source datasets are tested: the 1000 Functional Connectomes Project (FCP), the Consortium for Reliability and Reproducibility (CoRR), and the Autism Brain Imaging Data Exchange (ABIDE) Collection. These demo applications link descriptive quality control metrics, regional brain volumes, and thickness scalars to a 3D imaging viewer and editing module, resulting in an easy-to-implement quality control protocol that can be scaled for any size and complexity of study.”

Multiple sclerosis [MSC]

Journals

- **[MSC-TA-17]:** https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5990125/

  Tacchella, A., Romano, S., Ferraldeschi, M., Salvetti, M., Zaccaria, A., Crisanti, A., & Grassi, F. (2017). Collaboration between a human group and artificial intelligence can improve prediction of multiple sclerosis course: a proof-of-principle study. *F1000Research*, *6*, 2172. https://doi.org/10.12688/f1000research.13114.2

- [MSC-TA-17-1]: under “Abstract”, quote: “Multiple sclerosis has an extremely variable natural course. In most patients, disease starts with a relapsing-remitting (RR) phase, which proceeds to a secondary progressive (SP) form. The duration of the RR phase is hard to predict, and to date predictions on the rate of disease progression remain suboptimal. This limits the opportunity to tailor therapy on an individual patient's prognosis, in spite of the choice of several therapeutic options. Approaches to improve clinical decisions, such as collective intelligence of human groups and machine learning algorithms are widely investigated.”
- [MSC-TA-17-2]: under “Abstract”, quote: “Medical students and a machine learning algorithm predicted the course of disease on the basis of randomly chosen clinical records of patients that attended at the Multiple Sclerosis service of Sant'Andrea hospital in Rome.”
- [MSC-TA-17-3]: under “Abstract”, quote: “In this work we present proof-of-principle that human-machine hybrid predictions yield better prognoses than machine learning algorithms or groups of humans alone.”
- [MSC-TA-17-4]: under “Abstract”, quote: “To strengthen and generalize this preliminary result, we propose a crowdsourcing initiative to collect prognoses by physicians on an expanded set of patients.”
- [MSC-TA-17-5]: under “Introduction”, quote: “In the clinics, as in any other fields of human knowledge, innovative approaches based on machine learning and collective reasoning methods are used in an attempt to succeed where traditional methods of forecasting failed.”
- [MSC-TA-17-6]: under “Introduction”, quote: “Machine learning algorithms catch complex relations among existing data to an extent beyond standard regression models. Good performances have been obtained for the diagnosis of Parkinson's disease and the prognosis of disease progression in amyotrophic lateral sclerosis (Dinov et al., 2016; Küffner et al., 2015). For MS, machine learning algorithms can correctly classify disease course in about 70% of cases of both clinically definite MS and of clinically isolated syndrome (Fiorini et al., 2015; Wottschel et al., 2014; Zhao et al., 2017), a good result that still requires improvement to become of clinical value.”
- [MSC-TA-17-7]: under “Introduction”, quote: “Through collective reasoning, or collective intelligence, groups of lay people may perform as well as experts. In principle, the larger the group, the higher the prediction accuracy, which led to the development of several crowdsourcing initiatives. Possibly, the forerunner was FOLDIT study on protein folding, but crowdsourcing has been exploited also for diagnostic purposes in pathologies, such as breast cancer, skin cancer or ophtalmology. However, when expert people are involved, even small groups can outperform the

best among them, at least when a yes/no answer to well-defined diagnostic questions is requested based on radiographic/histological images. Studies with medical students show that working in pairs, either interacting while responding or aggregating responses *ex post*, ameliorates diagnostic ability, with further improvements when group size increases, in line with the core idea of Collective intelligence. Similar results have been obtained also for prognoses on critically ill patients."

- [MSC-TA-17-8]: under "Introduction", quote: "Combination of human and machine predictions into hybrid forecasts exploits human intuitive reasoning and computer classification capabilities, potentially boosting both. Indeed, at least in the case of predicting the course of actions in American football games within the frame of prediction markets, hybrid groups performed better than either humans or computers."
- [MSC-TA-17-9]: under "Introduction", quote: "Machine learning and collective intelligence performed almost equally well, but their combination yielded a small, yet statistically significant, improvement in the reliability of the forecasts on disease evolution over different time periods."
- [MSC-TA-17-10]: under "Classification with machine learning", quote: "we used a modified *leave-one-out* approach, training the algorithm with the following rules:

  1. We excluded all visits from one patient from the dataset 2. We built 50 training sets, each composed by 83 records, taking care to include only one clinical record (randomly chosen) for every remaining patient 3. We trained 50 Random Forest models, one for each training set 4. We computed the probability of the transition from RR to SP by averaging the predictions of the 50 models on all the visits of the excluded patient. Predictions consisted in scores from 0 (Extremely unlikely) to 1 (Highly probable)."
- [MSC-TA-17-11]: under "Human predictions", quote: "For adequate comparison with computer predictions, students evaluated 50 medical records, collected in a questionnaire, randomly extracted from the same dataset used for machine learning and estimated the probability that the patient would progress to the SP phase within 180, 360 and 720 days."
- [MSC-TA-17-12]: under "Human predictions", quote: "Forty-two medical students in the final two years of their course (Sapienza University, Rome Italy, based within Sant'Andrea hospital), volunteered to participate in the task."
- [MSC-TA-17-13]: under "Human predictions", quote: "Scores were from 0 (Extremely unlikely) to 5 (Highly probable). Predictions (see Supplementary file Student_Predictions.xlsx) were analysed, using the AUC."

- [MSC-TA-17-14]: under “Hybrid predictions”, quote: “We next integrated human and computer predictions into a hybrid prediction, which combines human clinical reasoning with the classification approach of machine learning algorithms… The simplest approach to aggregate forecast is performing a linear or weighted average of the predictions released by humans or computer.”
- [MSC-TA-17-15]: under “Hybrid predictions”, quote: “Then, a normalized ranking was assigned, ranging from 1 for the most consistent predictions to 0 for the most scattered and ranks were squared to emphasize the contribution of the most consistent agent.”
- [MSC-TA-17-16]: under “Discussion”, quote: “In this work we present proof-of-principle that human-machine hybrid predictions attain prognostic ability above that of machine learning algorithms and groups of humans alone.”
- [MSC-TA-17-17]: under “Discussion”, quote: “In spite of the relatively basic machine learning technique used, the small number of students involved and their limited clinical knowledge, this work suggests that hybrid predictions can be useful to improve the prognosis of MS course.”
- [MSC-TA-17-18]: under “Discussion”, quote: “In the long run, it is possible that further developments in our ability to combine collective reasoning and machine predictions will have a profound impact also on the organization and management of medical care, particularly in hospital settings.”

Observation [OB]

Journals

- **[OB-LH-18]:** https://biss.pensoft.net/article/39229/list/9/

  Hogeweg L, Schermer M, Pieterse S, Roeke T, Gerritsen W (2019) Machine Learning Model for Identifying Dutch/Belgian Biodiversity. Biodiversity Information Science and Standards 3: e39229. https://doi.org/10.3897/biss.3.39229

  - [OB-LH-18-1]**:** Under “Abstract”, quote: “These online platforms, and many scientific studies as well, suffer from a *taxonomic bias*: the effect that certain species groups are overrepresented in the data (Troudet et al. 2017). One of the reasons for this bias is that the accurate identification of species, by non-experts and experts, has been limited by the large number of species that exist.”

- [OB-LH-18-2]**:** Under “Abstract”, quote: “Most of the observations with photos were validated by human experts at Observation.org, creating a unique database suitable for machine learning.”
- [OB-LH-18-3]**:** Under “Abstract”, quote: “We have developed a deep learning-based species identification model using this database containing 13,767 species, 1,530 species-groups, 734 subspecies and 117 hybrids. The model is made available to the public through a web service (https://identify.biodiversityanalysis.nl) and through a set of mobile apps (ObsIdentify). In this talk we will discuss our technical approach for dealing with the large number of species in a deep learning model.”
- [OB-LH-18-4]**:** Under “Abstract”, quote: “We will evaluate the results in terms of performance for different species groups and what this could mean to address part of the taxonomic bias. We will also consider limitations of (image-based) automated species identification and determine venues to further improve identification. We will illustrate how the web service and mobile apps are applied to support citizen scientists and the observation validation workflows at Observation.org. Finally, we will examine the potential of these methods to provide large scale automated analysis of biodiversity data.”

- **[OB-SM-18]:** https://www.researchgate.net/publication/325209899_Supporting_citizen_scientists_with_automatic_species_identification_using_deep_learning_image_recognition_models

  Schermer, M. & Hogeweg, L. (2018). Supporting citizen scientists with automatic species identification using deep learning image recognition models. Biodiversity Information Science and Standards 2: e25268. https://doi.org/10.3897/biss.2.25268

  - [OB-SM-18-1]: Under “Abstract“, quote: “Volunteers, researchers and citizen scientists are important contributors to observation and monitoring databases. Their contributions thus become part of a global digital data pool, that forms the basis for important and powerful tools for conservation, research, education and policy.”
  - [OB-SM-18-2]: Under “Abstract”, quote: “With the data contributed by citizen scientists also come concerns about data completeness and quality. For data generated by citizen scientists taxonomic bias effects, where certain species (groups) are underrepresented in observations, are even stronger

than for professionally collected data. Identification tools that help citizen scientists to access more difficult, underrepresented groups, can help to close this gap."

- [OB-SM-18-3]: Under "Abstract", quote: "We are exploring the possibilities of using artificial intelligence for automatic species identification as a tool to support the registration of field observations. Our aim is to offer nature enthusiasts the possibility of automatically identifying species, based on photos they have taken as part of an observation."
- [OB-SM-18-4]: Under "Abstract", quote: "Furthermore, by allowing them to register these identifications as part of the observation, we aim to enhance the completeness and quality of the observation database. We will demonstrate the use of automatic species recognition as part of the process of observation registration, using a recognition model that is based on deep learning techniques."
- [OB-SM-18-5]: Under "Abstract", quote: "We investigated the automatic species recognition using deep learning models trained with observation data of the popular website Observation.org (https://observation.org/)."
- [OB-SM-18-6]: Under "Abstract", quote: "At Observation.org data quality is ensured by a review process of all observations by experts. Using the pictures and corresponding validated metadata from their database, models were developed covering several species groups."
- [OB-SM-18-7]: Under "Abstract", quote: "These techniques were based on earlier work that culminated in ObsIdentify, an free offline mobile app for identifying species based on pictures taken in the field. The models are also made available as an API web service, which allows for identification by offering a photo through common HTTP-communication essentially like uploading it through a webpage. This web service was implemented in the observation entry workflows of Observation.org."
- [OB-SM-18-8]: Under "Abstract", quote: "By providing an automatically generated taxonomic identification with each image, we expect to stimulate existing citizen scientists to generate a larger quantity of and more biodiverse observations. Additionally we hope to motivate new citizen scientists to start contributing."

WEB

- **[WEB-OB-ObservationOrg]:**
  - [WEB-OB-ObservationOrg-1]: https://observation.org/pages/mission/ Under" Mission", quote: "Observation.org aims to accommodate all its users to register and share nature observations, in order to document the natural richness of the world, for now and into the future. To this end, the platform

collaborates with thousands of volunteers, directly but also through over 300 regional and national working groups. We collect and disseminate data but we do not interpret the data. We leave that task to other organisations. We pursue solid collaboration with all research organizations. Data about vulnerable species and locations may be obfuscated or hidden to prevent their abuse."

- [WEB-OB-ObservationOrg-2]: https://observation.org/pages/getting-started/ Under "Enter: Observation", quote: "Adding an observation can be done in one of two ways: via the website or our mobile applications...Click top left on 'Enter' and choose 'Observation'... Yes, every observation has added value. We are not just a platform for rare species. All observations are welcome and contribute to a better picture of our biodiversity."
- [WEB-OB-ObservationOrg-3]: https://observation.org/pages/getting-started/ Under "Enter: Observation", quote: "You can see at a glance the fields needed to add an observation. These include the date, time, species, number of individuals and notes. The *'Date'* and *'species'* fields are mandatory. These two fields are the minimum required for a basic observation. However, you can choose to add more information. This is not necessary but gives the observation more context. The following fields are accessible when you check the *'Details'* check box:
    1. Activity
    2. Stage of life
    3. Method
    4. On / in
    5. Notes
    6. Substrate
    7. Counting method
    8. Escaped
    9. External reference
    10. Obscure
    11. Hidden until"
- [WEB-OB-ObservationOrg-4]: https://observation.org/pages/getting-started/ Under "Upload your photo or observation", quote: "More and more people take photos of their sightings. if you want to add a photo, click on the blue button *Upload* and select 1 or more photos (maximum 5). When the upload is complete, you will receive a suggestion as to what speccies it could be. This suggestion comes from the automatic image recognition *ObsIdentify*. By clicking on 'Accept', the species name is entered automatically. However, please look critically at the suggestion as ObsIdentify can be wrong. If necessary, please crop the photo so that the species is clearly visible. If photos contain information about time, date and GPS (metadata),

Observation.org will automatically extract that data and use it. In principle, you can add an observation by merely uploading a photo. Editing photos can remove the metadata from photos. A photo larger than 1000 * 1000 pixels will be reduced in size by the server, possibly with a loss of quality. We do this to guarantee the speed of the site. You can always add media (photo or sound) to an observation at a later time or date. Photos apply to all types of redords. Audio recordings are often added too, for example, birds, grasshoppers and crickets and amphibians. Open the observation by going to " Observations " and clicking on the date of the relevant observation. In the observation screen on the right, click the blue button *'Options'* and choose *'Edit'* . You will now see a number of options, namely: Edit - *change the information in the fields or add information* Add photo - *Add a photo to the observation* Add sound - *add a sound to the observation* Delete - *delete an observation.”*

- [WEB-OB-ObservationOrg-5]: https://observation.org/pages/getting-started/ Under “What others have seen in your neighborhood”, quote: “You can discover which species have been seen there via Observation.org. Click on 'Discover' on the homepage and choose 'Locations'. Here you can search for locations where you want to know what has been seen there. You can search by both city and area name. In the location overview you will find useful information about the area such as: Name, area, municipality and province. In addition, you can see at a glance how many observations have been entered in the area, how many users have been there, how many photos and sounds we have in that area and the total number of species that have been observed there. You can see the boundaries of the specific area on the map. There are blue circles that symbolize the latest observations in the area. These are clickable. In the tabs you have the following options:
  1. Details - *as described above - starting point*
  2. Observations - *an overview of the last twenty observations*
  3. Photos - *an overview of photos*
  4. Sounds - *an overview of sounds”*
- [WEB-OB-ObservationOrg-6]: https://observation.org/pages/getting-started/ Under “Obscure”, quote: “In a number of cases it may happen that Observation.org places your observation under embargo for a certain period. Data from vulnerable species and locations can be obscured to prevent abuse. This can be based on: combination species / area (for example, cranes), but also the combination species / behaviour (for example, owls). This is to prevent disturbance during the breeding season and / or hibernation. Embargos are usually requested by site managers and / or local working groups. What options does Observation.org offer to

achieve these two goals? Vulnerable situations and species can be protected on 4 levels.

1) At species level - Observation.org can determine that all observations of a specific species are not visible in detail. In this case, location data is invisible, except for the observer himself and our administrators.

2) At location level - Observation.org can decide on its own initiative or at the request of site owners to automatically hide observations of a specific species in a specific area. Details of the observation will then not be visible, except for the observer himself and our administrators.

3) At the observation level - The person entering the observation can at all times ensure that his or her observation does not become public. The observation is then entered under *obscure* until a date to be entered. In this case too, no details of the observation are visible, except for the observer himself and our administrators. The *obscure option* can be found in the entry screen for new observations and can also be used via mobile entries.

4) At observation level - The person entering the observation can at all times ensure that the location details of his or her observation are not made public. The observation is then *obscured* to km2 level. This is actually a mild variant of placing an observation under "embargo". The exact location is not shown, but it is represented as a square of 1 km2. The observer himself and our administrators can see the exact location. The *fade option* can be found in the entry screen for new observations and can also be used via mobile entries. If a species falls under method 1 or 2, it is not necessary for the observer to include method 3 or 4."

- [WEB-OB-ObservationOrg-7]: https://observation.org/help/icons/ Under "Validation statuses: "icon code short description long description
  i. O unknown Observation has not yet been validated.
  ii. J accepted (with evidence) Observation is convincingly documented with image or sound, or has been approved by an appointed rarities committee.
  iii. P accepted (by admin) Observation is accepted based on expert's knowledge (distribution, experience, previous observations) or other available information, but without documentation with image or sound.
  iv. A accepted (automatic validation) Accepted by automated rules based on validated observations, or by image recognition.

v. I pending Pending validation. Only visible for validators and for the observer.

vi. N rejected Observation does not meet criteria for validation, has been rejected by an appointed rarities committee, or documentation shows different species. Only visible for validators and for the observer.

vii. U cannot be validated (yet) Observation cannot be validated (yet) because of insufficient documentation, because validators could not agree, because observer has expressed uncertainty, or because a rarities committee has yet to reach a decision.

PlantSnap [PS]

WEB

- **[WEB-Steemit]:** https://steemit.com/steemhunt/@autofreak/plantsnap-uses-ai-and-machine-learning-algorithm-to-identify-plants
    - ○ [WEB-Steemit-1]: Under “Features“, quote: “It helps you identify plants and trees through your camera. Identify trees, flowers and many other plants by uploading pictures to Plant Database.”
    - ○ [WEB-Steemit-2]: Under “Learning Algorithm & Plant Database“, quote: “PlantSnap machine-learning algorithm can recognise not less than 2,000 new plant species on monthly basis.”
    - ○ [WEB-Steemit-3]: Under “Learning Algorithm & Plant Database“, quote: “Automatic update when new plant species are added to PlantSnap database, no extra fee is required to receive updates.
    - ○ [WEB-Steemit-4]: Under “Website“, quote: “You can upload plant picture on the web, then view it on mobile application and vice versa. PlantSnap website has simple UI that everyone can use.”
    - ○ [WEB-Steemit-5]: Under “Learning Algorithm & Plant Database“, quote: “PlantSnap can identify mushrooms, cacti, succulent, flowers, trees and many more.”
    - ○ [WEB-Steemit-6]: Under “Discover The World Through Your Environment“, quote: “Rediscover nature with the help of pocket botanist. Add fun to your hikes with loved ones! PlantSnap allows users to build plant library. Bridging gap between nature and technology, go out and have fun with fresh air and improve your mood.”

- **[WEB-PlantSnap]:** https://www.plantsnap.com/?ref=steemhunt
  - [WEB-PlantSnap-1]: **“**PlantSnap is the most high-tech, comprehensive and accurate plant identification app ever created.”
  - [WEB-PlantSnap-2]: Under “How it works”, quote: “To identify a plant you simply need to simply snap a photo of the plant, and the app will tell you what it is in a matter of seconds! PlantSnap can currently recognize 90% of all known species of plants and trees, which covers most of the species you will encounter in every country on Earth.”
  - [WEB-PlantSnap-3]: Under “Community Voices”, quote: “The videos and tutorials have helped me make sure my success rate with the algorithm stays high-big thanks for that.”
  - [WEB-PlantSnap-4]: Under “What’s inside”, quote: “
    1. Snap any plant, mushroom or cactae!
    2. See info about the plant
    3. Experience augumented reality
    4. Explore snaps around the world
    5. Create collections of your favourite plants”
  - [WEB-PlantSnap-5]: Under “Join our amazing community!”, quote: “We are continuously working to improve PlantSnap and one of the most important aspects is creating a better database, so you are just as much a part of our team as the developers are!”
  - [WEB-PlantSnap-6]: https://www.plantsnap.com/who-we-are/ “With over 600,000 plants and 250 million+ images in our database, PlantSnap is currently using Machine Learning technology and artificial intelligence to help anyone, anywhere, identify any plant or tree on planet Earth!”
  - [WEB-PlantSnap-7]: https://www.plantsnap.com/who-we-are/ Under “Instantaneous results”, quote: “Most plant identifying apps can identify no more than 2000 species, at most. They also use crowdsourcing to generate data and that means it takes a lot of time, days, maybe more, to get an opinion, this might prove to be challenging or even frustrating. PlantSnap has over 600,000 plants in our searchable database, and it is translated into 37 languages. This means that PlantSnap will work in any country on Earth, for 95% of the global population.”
  - [WEB-PlantSnap-8]: https://www.plantsnap.com/who-we-are/ Under “no guess work involved”, quote: “Here at PlantSnap we’ve managed to create a system that allows you to upload a photo and instantly get detailed information of the snapped plant with no guesswork or human interaction involved. In addition, the iOS version of PlantSnap uses new technology called auto-detect and augmented reality. Auto-detect actually tells you when to snap the photo so that you get the perfect picture every time.

  - Augmented reality adds and entire new level of immersion and education into the PlantSnap experience."
  - [WEB-PlantSnap-9]: Under "Our Mission", quote: "We want to recreate the connection between people and the amazing natural world around us. Because we believe technology is the answer, we created PlantSnap as the digital interface to bring people and nature together."

- **[WEB-Startengine]:** https://www.startengine.com/plantsnap-inc?utm_source=ct&utm_medium=website&utm_campaign=interstitial
  - [WEB-Startengine-1]: Under "PlatSnap reconnects people with the world around them", quote: "It's important for us to understand that we are a part of nature, not apart from it, and that it's our responsibility to protect our world for future generations. So, we created a tool that encourages people to stop and explore the beauty and wonder of the world they live in. We hope to find like-minded individuals to join us on this journey to complete our growing global plant database in a timely manner, and further our goals of stirring up excitement about getting re-acquainted with the great outdoors. By bringing people back to nature and helping them feel like they are an integral part of this amazing planet, together we can help them understand that we are stewards of the Earth, not owners."
  - [WEB-Startengine-2]: Under "PlanSnap uses AI to identify plants with just a picture", quote: "PlantSnap not only reconnects people to nature but also creates a database of plants that provides useful information to hobbyists and professionals alike. PlantSnap is translated into 37 languages and currently used in over 200 countries every day. With over 620,000 plants now in our database and 32 million installs so far, PlantSnap has become THE go-to app for gardeners, hikers, landscape designers, teachers, students, foragers, and anyone who enjoys nature. At PlantSnap, we're reigniting interest in the beauty and wonder of nature that surrounds us every day, while empowering scientists and nature enthusiasts with the technology to catalog and share their discoveries."
  - [WEB-Startengine-3]: Under "Secured some incredible partnerships", quote: "On top of our existing botanical partnerships, we have also completed our first major sponsorship campaign with Mrs. Meyers Cleaning Products for Earth Month in April of this year. Using PlantSnap, people could snap a photo of a plant and discover whether it was endangered or not in real time. If it was, we asked them to take a picture instead of picking the flower. With each purchase of affiliated products, they would plant the endangered purple coneflower in its preservation garden. To date, they've

distributed and planted more than 270,000 purple coneflowers. The campaign was so successful that the agency created this case study and video, which was selected as an entrant in the Cannes Lions International Festival of Creativity."

- [WEB-Startengine-4]: Under "How it Works", quote: "
    1. Take a Photo - Snap a pic of any plant, flower, tree, cacti, succulent or mushroom and receive results in about 5 seconds. In the wilderness with no internet? No problem! Just save the photo to your phone and analyze with PlantSnap once you have a data connection.
    2. Learn and Discover - Upload, and instantly identify your plant. Learn facts such as where else you can find the plant, its growth habit, its edibility and more.
    3. Keep Track of It All - Compile your discoveries in one place—never forget where you found a plant or what it's called. Be prepared to find all the plants in your area!
    4. Be a part of our Global Citizen Science initiative -- Simply by snapping photos of plants wherever you go, you will play an integral role in this massive, global initiative to map, catalog and evaluate every known plant species on Earth. This data will then be used by scientific organizations and universities around the world to conserve and save plant species, thereby protecting our environment and the global ecosystem."
- [WEB-Startengine-5]: Under "PlantSnap is the digital interface to bring people and nature together", quote: "We believe technology is the answer. With PlantSnap, we hope to bring people back to nature by putting an application on their phones that will allow them to instantly identify any plant, flower or tree simply by snapping a photo. We want to take this a step further by helping people to reconnect with not only nature but also each other. We are launching PlantSnap 3.0 in late September which will incorporate a social component called PlantSnappers. This will allow people to "Friend" other PlantSnappers around the world, share photos, share gardening tips, etc."
- [WEB-Startengine-6]: Under "Calling all PlantSnappers! A way to connect and share your findings", quote: "The goal of PlantSnappers is to connect you to other plant lovers and plant professionals. Our intention is not to replace other social media platforms but offer one that is only for plants. We hope that PlantSnappers will give you an outlet in which to share your enthusiasm for plants and your love of nature and our environment with like-minded people all over the world!

1. Share your favorite findings and discoveries with friends. You can share plant photos straight from your camera roll and PlantSnap will automatically identify the plant for you when it posts
2. Comment and interact with your favorite posts made by your friends
3. Share tips and ask advice on gardening and plant care
4. Connect with a global community of plant lovers all over the world who use PlantSnap in over 200 countries”

- ○ [WEB-Startengine-7]: Under “Together, we can build a bridge to nature that everyone from the everyday nature lover to the lifelong career scientist can use to protect our home.”

Snapshot Serengeti

Journals

- [SnS-SA-15]: https://www.nature.com/articles/sdata201526

  Swanson, A., Kosmala, M., Lintott, C. *et al.* Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. *Sci Data* 2, 150026 (2015). https://doi.org/10.1038/sdata.2015.26

  - [SnS-SA-15-1]: Under “Abstract”, quote: “Multiple users viewed each image and recorded the species, number of individuals, associated behaviours, and presence of young.”
  - [SnS-SA-15-2: Under “Abstract”, quote: “We applied a simple algorithm to aggregate these individual classifications into a final ‘consensus’ dataset, yielding a final classification for each image and a measure of agreement among individual answers.”
  - [SnS-SA-15-3]: Under “Abstract”, quote: “The consensus classifications and raw imagery provide an unparalleled opportunity to investigate multi-species dynamics in an intact ecosystem and a valuable resource for machine-learning and computer-vision research.”
  - [SnS-SA-15-4]: Under “Background and Summary”, quote: “advances in digital technology have increased capacity while lowering prices, resulting in a dramatic increase in the number and diversity of camera trap studies. While traditional analytical approaches for camera trapping data require individually identifiable animals, recent developments have allowed the expansion of camera trap inference to multiple ‘unmarked’ species.”
  - [SnS-SA-15-5]: Under “Background and Summary”, quote: “Although camera-trap surveys are increasing in popularity and scope, they can

produce overwhelming amounts of data, highlighting the need for efficient image processing techniques. Here we describe the datasets generated by *Snapshot Serengeti*, a large-scale survey that (1) deployed 225 camera traps across a 1,125 km^2 area in the Serengeti National Park, Tanzania from 2010–2013, (2) used a citizen science website (www.snapshotserengeti.org) to process millions of images, and (3) used a simple algorithm to ensure high reliability of the resultant species classifications."

- [SnS-SA-15-6]: Under "Background and Summary", quote: "Our camera survey expands upon historical monitoring by providing the first continuous systematic data on all of the larger predator and prey species, day and night, across several years. We set out 225 cameras within a 1,125 km^2 grid inside the long-term lion study area that covers the intersection of open plains and savannah woodlands (Fig. 1 and Fig. 2), and spans a 1.67-fold rainfall gradient and 1.44-fold productivity gradient (T.M. Anderson, unpublished data.). The camera-trap grid offers systematic coverage of the entire study area (as per O'Brien *et al.*[20]) and ensures at least two cameras per home range for each medium to large mammalian species."
- [SnS-SA-15-7]: Under "Background and Summary", quote: "In collaboration with The Zooniverse (www.zooniverse.org), the world's most popular citizen science platform, we developed the website www.snapshotserengeti.org that allowed members of the general public to view and classify each image, identifying species, counting the number of individuals, and characterizing behaviours (Fig. 3)."
- [SnS-SA-15-8]: Under "Background and Summary", quote: "Every image set was circulated to multiple users to improve data accuracy."
- [SnS-SA-15-9]: Under "Background and Summary", quote: "We applied a simple plurality algorithm to produce a 'consensus dataset' of final classifications for each image set."
- [SnS-SA-15-10]: Under "Background and Summary", quote: "The consensus classifications were validated against 4,149 'gold-standard' image-sets that had been classified by experts, revealing 96.6% accuracy for species identifications and 90% accuracy for species counts."
- [SnS-SA-15-11]: Under "Background and Summary", quote: "In this report we describe the field methods, citizen science interface, and consensus algorithm used to produce the following datasets:

  > (1) Images: Full-resolution images produced by the survey. (2) Raw classification data: All individual classifications made by all users on all image sets. (3) Consensus data: Single classification per image set produced by applying the consensus algorithm to raw

classifications, along with image metadata (date, time, location). (4) Operation Dates: Metadata of when each camera was operational. (5) Gold-standard data: Expert classifications for a subset of 4,149 image sets.

- [SnS-SA-15-12]: Under "Background and Summary", quote: "Computer science and informatics researchers can use the raw (un-aggregated) citizen-science answers to develop more complex aggregation algorithms and to test their performance against the gold-standard dataset."
- [SnS-SA-15-13]: Under "Background and Summary", quote: "Additionally, computer-vision researchers need large human-annotated sets of imagery as training sets in machine-learning algorithms."
- [SnS-SA-15-14]: Under "Background and Summary", quote: "Our collaborators are currently using this dataset to automate species detection, classification and similar-species differentiation, as well as to develop combined human-machine learning systems and imaging systems for searchable colour. Subsets of the consensus dataset have also been used in classrooms to engage students in authentic research that spans ecology, animal behaviour, and computer science (see *Usage Notes* for examples)."
- [SnS-SA-15-15]: Under "Methods", quote: "We set all cameras to take 3 photos per trigger in the daytime. At night, infrared-flash cameras took 3 photos per trigger, but incandescent-flash cameras could only take 1 image per trigger due to flash limitations (and occasional camera malfunction created a small number of image sets with varying numbers of images). We refer to each trigger as a 'capture event' and the resulting 1–3 images as an 'image set'; capture events are the units of analysis for ecological studies and comprise the results presented here. We set cameras to ensure at least 1-minute delay between capture events to prevent the memory card being filled to capacity by a single individual or herd."
- [SnS-SA-15-16]: Under "Methods", quote: "We checked each camera every 6–8 weeks. Except in cases of camera malfunction or damage, this schedule was sufficient to replace batteries and SD cards and ensure continuous operation. We labelled SD cards with the Site ID and the date retrieved and reviewed images in the field to ensure that the camera had functioned properly. We then installed new SD cards and triggered cameras to photograph placards that indicated Site ID, date, and time."
- [SnS-SA-15-17]: Under "Data management", quote: "We wrote Python scripts to extract date/time from the image files and season, site, and card information from the directory structure. Common errors that arose from camera malfunction (typically due to animal or weather damage) included: the recording of videos instead of still images, incorrect time-stamps for a

portion of images, and only 1–2 photos per capture event instead of three. We wrote code in Python, MySQL, and R to flag and correct these errors in the metadata."

- [SnS-SA-15-18]: Under "Data processing", quote: "We partnered with the online citizen science platform The Zooniverse (www.zooniverse.org) to develop the *Snapshot Serengeti* website (www.snapshotserengeti.org), an online interface where the general public helps process camera trap data. The *Snapshot Serengeti* website utilizes the Zooniverse's platform *Ouroboros*, written in Ruby on Rails (https://github.com/zooniverse/serengeti)."
- [SnS-SA-15-19]: Under "Data processing", quote: "Volunteer classifiers interact with a custom-built JavaScript front-end to classify image sets and results are saved in a MongoDB datastore. Each classification is recorded alongside the time of classification and the identity of the classifier in the form of either a unique identifier assigned by the Zooniverse (for logged in users) or an IP address (for users who have not logged in). *Ouroboros* also allows for custom rules for image-set retirement, as discussed below, and the system can scale rapidly to cope with the demands of a popular site. The interface and images are hosted on Amazon Web Services via Amazon's Simple Storage Service (S3)."
- [SnS-SA-15-20]: Under "Data processing", quote: "On the *Snapshot Serengeti* interface (Fig. 3), volunteers identify species in each image set, count the number of individuals, classify behaviour, and indicate the presence/absence of young."
- [SnS-SA-15-21]: Under "Data processing", quote: " For image sets that contain more than one image, volunteers initially see the second image in the set and can toggle between images or use the 'play' feature to animate the images. We designed the task flow to help guide people with no background knowledge through the process of identifying the animal(s) in question from 48 possible species and species groups while still providing a rapid route to classification for more knowledgeable participants."
- [SnS-SA-15-22]: Under "Data processing", quote: "Image difficulty (and probability of being correct) can instead be assessed by measuring variance across individual volunteer answers (see Technical Validation)."
- [SnS-SA-15-23]: Under "Data processing", quote: "Users filter potential species matches by morphological characteristics such as horn shape, body shape, colour, pattern, and tail shape or jump straight to selecting from a list of all species. A 'nothing here' button allows users to classify image sets without any animals present. We do not offer an 'impossible' or 'I don't know' option because previous testing on a small-scale prototype indicated

that such answers were overused and provided no information on the actual species classification, thus wasting volunteer effort. “

- [SnS-SA-15-24]: Under “Data processing”, quote: “Sample images from image sets retired from Snapshot Serengeti as (**a**) *blank*: receiving five consecutive ‘nothing here’ classifications, (**b**-**c**) *consensus:* receiving 10 matching species classifications, and (**d**) *complete*: receiving 25 classifications regardless of agreement. Note that the plurality algorithm correctly arrived at ‘giraffe,’ ‘spotted hyena,’ and ‘impala’ for images **b**-**d**, respectively (see Tables 2 and 3 for individual classifications). *Blank*: the first 5 classifications are ‘nothing here’. *Blank_Consensus*: 10 ‘nothing here’ classifications, not necessarily consecutive. *Consensus*: 10 matching classifications of species or species combination (e.g., 10 identifications of ‘lion’ or 10 identifications of ‘lion-zebra’); these classifications do not have to be consecutive. *Complete*: 25 total non-‘nothing here’ classifications (does not require consensus for any single species). Note that volunteers classified Snapshot Serengeti data faster than images were produced, and images were re-circulated for classroom use and testing the value of additional classifications. As a result, the number of classifications (11–57 for images containing animals) generally exceeded the number needed for retirement under the above rules.”
- [SnS-SA-15-25]: Under “Data aggregation”, quote: “We implemented a simple plurality algorithm to transform the volunteer classifications for each image set into a single aggregated species identification.”
- [SnS-SA-15-26]: Under “Data aggregation”, quote: “First, we calculated the number of different species present in an image set as the median number of different species identified across all users for that image set. For all image sets, we assigned the one (or more) species with the most ‘votes’ as the aggregated answer. We calculated the number of individuals present for each identified species as the median number reported for that image set for that species by all volunteers. We also calculated the proportion of users who chose each behavioural activity or presence of young. To assess the accuracy of aggregated classifications, we calculated an evenness index, using all non-blank classifications for each image set. When all classifications were in agreement, we assigned the value zero, indicating high accuracy. Otherwise, we used Pielou’s evenness index (Pielou 1966). The Pielou evenness index ranges from 0 to 1, with 0 indicating low evenness and high accuracy and 1 indicating high evenness and low accuracy. Note that the Pielou evenness index is expected to be high for image sets with multiple species and therefore is not a useful gauge of accuracy in these cases.”

- [SnS-SA-15-27]: Under "Data Records", quote: "Consensus classification data and metadata: (consensus_data.csv; 334,671 data rows) Applying the plurality algorithm to the raw classification data yielded a single classification per capture event, accompanied by measures of uncertainty and difficulty."
- [SnS-SA-15-28]: Under "Technical validation", quote: "We asked five researchers with extensive wildlife identification experience to classify 4,149 randomly selected image sets containing animals using the *Snapshot Serengeti* interface; 263 image sets received two expert classifications and 8 image sets received three, for a total of 4,428 classifications. The experts noted whether any image sets were especially difficult or whether they thought the image was identifiable at all. In cases where experts disagreed with the results of the plurality algorithm or had marked an image set as particularly difficult or impossible, AS and CP made the final authoritative identification. Thus, the gold standard dataset included a small number of images that were agreed by multiple experts to be 'impossible' to identify. Because the Snapshot Serengeti interface does not allow 'impossible' as an option, the consensus answers for these images are incorrect by definition. We compared citizen-science classifications derived from the plurality algorithm with the expert-classified 'gold standard' dataset to assess accuracy of species identifications and counts of individuals."
- [SnS-SA-15-29]: Under "Citizen science and informatics analyses", quote: "Crowdsourcing and citizen science are being used increasingly often to produce science datasets[22–24], but they require robust methods to measure and validate data quality. While our consensus dataset derives from a simple plurality algorithm, more complex algorithms can improve upon these results. For example, Hines *et al.*[25] weighted raw classifications by individual accuracy, raising overall accuracy to 98%. Our raw classification dataset could be used to develop and test algorithms that employ user-weighting or even apply a Bayesian framework to incorporate information about species likelihood based on previous or subsequent images."
- [SnS-SA-15-30]: Under "Computer vision", quote: "Object search-and-recognition research requires large data sets of labelled imagery.
- [SnS-SA-15-31]: Under "Computer vision", quote: "Reliable data sets of wild animals are rare, due to the enormous task of hand-annotating large numbers of images. By using the raw images together with the consensus dataset, machine-learning algorithms could be developed to automatically detect and identify species, using part of the dataset for training the image-recognition algorithm and the rest for testing the algorithm. Raw images could be used separately, or in conjunction with the consensus data set, to

research automatic detection of textures, patterns, and other characteristics of outdoor scenes."

- [SnS-SA-15-32]: Under "Data Records", quote: "Gold standard data: (gold_standard_data.csv; 4,432 data rows) Expert classifications for 4,149 capture events. Note that gold-standard answers are more accurate than answers provided by a single expert because multiple experts reviewed all images for which any single expert expressed uncertainty."

WEB

- **[WEB-SnS-ZooniversOrg]:** https://www.zooniverse.org/projects/zooniverse/snapshot-serengeti/about/research
  - [WEB-SnS-ZooniversOrg-1]: Under "The Serengeti Lion Project", quote: "This daily monitoring has produced one of the most extensive datasets on any mammalian species anywhere in the world."
  - [WEB-SnS-ZooniversOrg-2]: Under "The Serengeti Lion Project", quote: "The daily records include information on the lions' location, group size, diet, food intake, health and reproduction. You can find out more on the Lion Project website."
  - [WEB-SnS-ZooniversOrg-3]: Under "Observing animals in the wild", quote: "Over the last 45 years, the University of Minnesota Lion Project has discovered a lot about lions – everything from why they have manes to why they live in groups. Now we're turning our sights to understanding how an entire community of large animals interacts. With photographs from these cameras, we're able to study how over 30 species are distributed across the landscape – and how they interact with lions and one another."
  - [WEB-SnS-ZooniversOrg-4]: under "What we do", quote: "We check on the camera traps in the course of daily lion monitoring. We change batteries, exchange the SD cards, and cut tall grass in front of the camera so that grass waving in the wind doesn't accidentally trigger the sensor. 225 cameras are a lot of work! When things run smoothly, a camera can last about two months before needing maintenance. But that's not always the case – sometimes we return to a camera only to find it chewed on by hyenas or torn down by elephants, waterlogged from a heavy rain or infested by ants."
  - [WEB-SnS-ZooniversOrg-5]: under "How the cameras work", quote: "The cameras use passive infrared sensors that are triggered when an object warmer than the ambient temperature moves in front of the sensor. This is

usually an animal…but tall sunlit grass can also trigger the camera when it blows in the wind. We currently use the Scoutguard 565 and DLC Covert Reveal models – these are incandescent flash cameras (with a white flash). Some people worry that incandescent flashes startle the animals, but in our study area the same individuals often come back to the same camera site night after night!”

- ○ [WEB-SnS-ZooniversOrg-6]: under “Our scientific questions”, quote: “Understanding how competing species coexist is a fundamental theme in ecology, with important implications for food webs, biodiversity, and the sustainability of life on Earth. Much of our current research focuses on how carnivores coexist with carnivores, herbivores with herbivores, and the joint dynamics of predators and their prey. These insights will guide strategies for species reintroduction, conservation, and ecosystem management around the world.

Twitter Suicide [TS]

Journals

- **[TS-KD17]:** https://doi.org/10.3389/fninf.2019.00029
  Karamshuk, D., Shaw, F., Brownlie, J., & Sastry, N. (2017). Bridging big data and qualitative methods in the social sciences: A case study of Twitter responses to high profile deaths by suicide. *Online Social Networks and Media*, 1, 33-43. DOI: 10.1016/j.osnem.2017.01.002.

  - ○ [TS-KD17-1]: under “Abstract”, quote: “With the rise of social media, a vast amount of new primary research material has become available to social scientists, but the sheer volume and variety of this make it difficult to access through the traditional approaches: close reading and nuanced interpretations of manual qualitative coding and analysis. This paper sets out to bridge the gap by developing semi-automated replacements for manual coding through a mixture of crowdsourcing and machine learning, seeded by the development of a careful manual coding scheme from a small sample of data. To show the promise of this approach, we attempt to create a nuanced categorisation of responses on Twitter to several recent high-profile deaths by suicide. Through these, we show that it is possible to code automatically across a large dataset to a high degree of accuracy (71%), and discuss the broader possibilities and pitfalls of using Big Data methods for Social Science.”
  - ○ [TS-KD17-2]: under “introduction”, quote: “Social science has always had to find ways of moving between the small-scale, interpretative concerns of qualitative research and the large-scale, often predictive concerns of the quantitative. The quantitative end of that spectrum has traditionally had two inter-related features: active collection of data and creating a suitable sub-sample of the wider population.

To the extent that such methods have also captured open-ended or qualitative data, the solution has been to apply manual coding, using a frame developed on the back of intensive qualitative analysis or an exhaustive coding of a smaller sample of responses. Although labour-intensive, manual coding has been critical for obtaining a nuanced understanding of complex social issues."

- [TS-KD17-3]: under "introduction", quote: "With social media, we now have so much information that it is impossible to process everything using either the detailed analysis methods of qualitative research or the application of manual coding approaches of the kind used in survey research."
- [TS-KD17-4]: under "introduction", quote: "And yet the application of traditional methods from qualitative social science, such as the close analysis of a small-scale sample of tweets relating to a public death, or the manual application of a coding frame to a larger volume of responses, are likely to miss crucial insights relating to the volume, patterning or dynamics. We therefore need a mechanism to train the social scientists' close lens on unmanageably large datasets – to bridge the gap between close readings and large scale patterning."
- [TS-KD17-5]: under "introduction", quote: "We argue that this approach has particular potential for the study of emotions at scale. Emotions have a mutable quality [1] and this is especially true in the context of social media. Thus, intensive manual coding over a small-scale sample may miss some of the temporal and volume dynamics that would be critical for a full sociological understanding of public expressions of emotion, in contrast to the semi-automated coding we propose here, which captures the entire dataset and its dynamics."
- [TS-KD17-6]: under "introduction", quote: "This paper develops a possible approach, that we term semi-automated coding: Our three-step method first manually bootstraps a coding scheme from a micro-scale sample of data, then uses a crowdsourcing platform to achieve a meso-scale model, and finally applies machine learning to build a macro-scale model. The bootstrapping is carefully done by trained researchers, creating the nuanced coding scheme necessary for answering social science questions, and providing an initial 'golden set' of labelled data. Crowdsourcing expands the labels to a larger dataset using untrained workers. The quality of crowd-generated labels is ensured by checking agreement among crowdworkers and between the crowd workers' labels and the golden set. This larger labelled dataset is then used to train a supervised machine learning model that automatically labels the entire dataset." (for sequence)
- [TS-KD17-7]: under "introduction", quote: "A key issue, both within the case study, and more generally, for the success of semi-automated coding as an approach, is the accuracy of the automatically generated labels. One source of error is the quality of crowd-generated labels. As mentioned above, we control for this using different forms of agreement, among crowd workers, and with a curated golden set."
- [TS-KD17-8]: under "3.1 Datasets", quote: "To analyse public discourses on social media relating to high-profile death by suicides, we chose five such deaths which

were highly publicised, either because the person was famous before their death or because of the circumstances of their death."

- [TS-KD17-9]: under "3.1 Datasets", quote: "We collected five datasets of related Twitter posts for 20 days following each death."
- [TS-KD17-10]: under "3.2 analysis approach: semi-automated coding", quote: "The social scientist's typical alternative would be to select and focus on a small sample of the dataset. Unfortunately, this is not a fully satisfactory solution for two reasons: First, it is not a priori clear which parts of the dataset would be most interesting and should be selected for intensive analysis. Second, focusing on a small sample misses aggregate characteristics, such as the relative volumes and temporal dynamics of different classes of responses, which can provide a new dimension to many social science questions, including ours, as it focuses on public, or aggregate, expressions of empathy. We argue therefore that manual coding needs to be adapted using computational methods, to scale up the volume of data created by social media platforms."
- [TS-KD17-11]: under "3.2 analysis approach: semi-automated coding", quote: "we start by noting that manual inspection cannot be avoided, because a) social scientists need to come up with a coding frame that makes sense for the research questions that are of interest and b) given that the classes of interest encompass nuanced, higher-order social-interaction concepts, it is easiest to define these by example rather than develop complicated rules or heuristics that can identify tweets belonging to the class."
- [TS-KD17-12]: under "3.2 analysis approach: semi-automated coding", quote: "Therefore, as a first step, researchers can identify the concepts/classes of interest, and provide examples. Subsequently, our goal was to build a machine learning model that can learn these concepts based on the examples given."
- [TS-KD17-13]: under "3.2 analysis approach: semi-automated coding", quote: "we adopted a two step approach to generate examples: First, trained researchers created a coding frame and a carefully curated set of example tweets. Next, an untrained set of workers on a crowd-sourcing platform were used to label a larger set of tweets. We controlled for the quality of labelling using agreement between crowd-workers, and agreement between crowd-workers' labels and the labels associated by researchers for the curated set of tweets."
- [TS-KD17-14]: under "3.2 analysis approach: semi-automated coding", quote: "as with any application of machine learning, the automatically generated set of labels is bound to have a few errors."
- [TS-KD17-15]: under "4. Bootstrapping coding using manual effort", quote: "To tackle scalability issues, we designed a *hybrid methodology* in which our coding typology was applied manually and gradually on different data scales. We began by manually coding a few hundred tweets which were subsequently used to guide the execution of a large-scale labelling experiment on Crowdflower, a crowd-sourcing platform. We then used twelve thousand labelled tweets obtained from the results of the CrowdFlower experiment to train a state-of-the-art machine

learning algorithm for short text analysis and to automatically label the full dataset (discussed in Section 5).”

- [TS-KD17-16]: under “4.1. Coding typology using trained researchers”, quote: “found ways to differentiate appropriately between tweet types based on emotional content (blame vs. grief, for example) and whom the tweet was directed at (other Twitter users, the deceased, certain people in particular, or society in general).”
- [TS-KD17-17]: under “4.2. Scaling the coding using crowd-sourcing”, quote: “Next, we created jobs on the Crowdflower crowdsourcing platform to expand the list of human labelled tweets.”
- [TS-KD17-18]: under “4.2. Scaling the coding using crowd-sourcing”, quote: “Tweets are often ambiguous, containing multiple communicative acts, and might be coded ’correctly’ in several ways...In order to help with this, we require each Tweet to be coded with exactly one label, and created a *decision tree* to help coders make decisions about how to code a particular tweet (Fig. 1).”
- [TS-KD17-19]: under “4.3. Fine-tuning execution parameters”, quote: “We chose CrowdFlower as a platform for executing our experiments because it provided enough flexibility to fine-tune our experiment and coders from specific countries – a requirement imposed by our ethics board. More specifically, we employed workers from the 15 (a limit imposed by Crowdflower) European Union countries with the largest populations.”
- [TS-KD17-20]: under “4.3. Fine-tuning execution parameters”, quote: “CrowdFlower provided several mechanisms to control the quality of coders for the experiment, of which the coders’ agreement with a short scale *golden set* of pre-coded answers proved the most effective.”
- [TS-KD17-21]: under “4.3. Fine-tuning execution parameters”, quote: “Two researchers labelled a sample of 200 tweets (40 from each use case) for the golden set experiment and refined this after three iterations of test runs on CrowdFlower. Further, we removed ambiguous tweets to ensure every possible chance for crowd workers to agree with the golden set. Finally, we followed the CrowdFlower’s recommendations[7] and balanced the number of tweets in each class ending up with a golden set of 64 tweets, with 8 tweets from each class.”
- [TS-KD17-22]: under “4.3. Fine-tuning execution parameters”, quote: “These tweets, rather than being representative of the entire dataset, functioned as a benchmark to test the accuracy and agreement among coders in the experiment, and allowed us to ensure that tweets were coded by Crowdflower workers who had the best understanding of the appropriateness of a particular code for a particular tweet. Coding by those who showed an accuracy of less than 65% and 66% in relation to the golden set was excluded from the results of the first and the second experiments, respectively.”
- [TS-KD17-23]: under “4.3. Fine-tuning execution parameters”, quote: “To encourage participation of high-quality coders we doubled the default pay for the job and noted in the description that the job required extra attention and that a good performance would be rewarded with bonuses. We then ran two experiments trading off between speed and quality (i.e. level of conservatism in selecting new

coders) and labelled an overall sample of around 12k tweets, with each tweet coded by two CrowdFlower workers. We opted to collect more data points at the cost of having fewer judgments for each label; at the same time, we were conservative in selecting only consensus votes for the next – machine learning – step of our analysis (in Section 5)."

- [TS-KD17-24]: under "4.3. Fine-tuning execution parameters", quote: "A few factors contributed to this decision. On the one hand, we had already imposed several measures to control the quality of the labelling process – by choosing only high-quality coders and opting for consensus votes from two coders. On the other hand, we expected our machine learning algorithm to benefit more from a diversity of data points rather than from a diversity of judgments. Since we were interested in analysing the temporal evolution of the discourse in our datasets, we sampled an equal number of tweets from each of the first twenty days in each considered use case.
- [TS-KD17-25]: under "4.3. Fine-tuning execution parameters", quote: "4.4. Validation of crowdsourced labels", quote: "We next validated the crowdsourced labels by analysing the sentiments of the tweets for which the labels were generated. Most sentiment analysis tools typically attach a positive or negative 'sentiment score', and therefore are less specific and nuanced than the coding frames typically used in social science. However, understanding the general sentiment scores of different classes that the crowd has identified provides us with a coarse-grained assurance in the validity of the results. To this end, we used the SentiStrength library [29], considered to be one of the best tools for short texts [30], and associated each tweet with a score between 1 and 5 for positive and negative sentiments."
- [TS-KD17-25]: under "5. Machine learning approach to understanding online mourning", quote: "In order to scale up our analysis from twelve thousand to a million tweets, we used a supervised machine learning algorithm for processing short texts."
- [TS-KD17-26]: under "5. Machine learning approach to understanding online mourning", quote: "The goal of the machine learning model is to mimic the human researcher who codes (i.e., classifies) tweets based on their content. To recreate this effect, we exploited and adapted a state-of-the-art deep convolutional *neural network* architecture *CharSCNN* for short text classification proposed in [16] that was designed to operate at a word-level to capture syntactic and semantic information, and at a character-level to capture morphological and shape information… CharSCNN showed significant improvement over alternative – recursive deep neural networks [31] and traditional bag-of-words models – when applied for fine-grained classification of Tweets."
- [TS-KD17-27]: under "5. Machine learning approach to understanding online mourning", quote: "The goal of the machine learning model is to mimic the human researchIn order to scale up our analysis from twelve thousand to a million tweets, we used a supervised machine learning algorithm for processing short texts.

- [TS-KD17-28]: under “5.2. Cross-validation”, quote: “We validated the performance of the algorithm over the dataset of tweets labelled by the CrowdFlower workers as described in the previous section. Specifically, we used all labels with agreement between the coders which resulted in a dataset of 7.1k tweets.”
- [TS-KD17-29]: under “5.3. Manual validation”, quote: “However there were also several instances where it did less well, and the repetition of similar tweets or claims might then lead to inaccuracies in the overall volume of tweets in each category. In relation to the question of add-ons to quoted tweets, this proved problematic in some cases… Despite these distortions, it is clear that the machine learning correctly identifies lack of empathy to be more prevalent in this case, and that this changes over time. However, in a multi-case study, we should be aware that individual circumstances surrounding events may have an impact on the accuracy of comparisons between cases, and any large-scale analysis would need to take into consideration that the machine learning model would have some erroneously labelled tweets.”
- [TS-KD17-30]: under “Discussion, conclusions and lessons”, quote: “The analysis presented here suggests that the combination of qualitative analysis with machine learning can offer both a big picture view of public events and close analysis of particular turning points or key moments in discussions of such events. As such, it can potentially yield new insights not easily achievable through traditional qualitative social science methods.”