Big Data Lessons from the ZOONIVERSE

Lucy Fortson
University of Minnesota

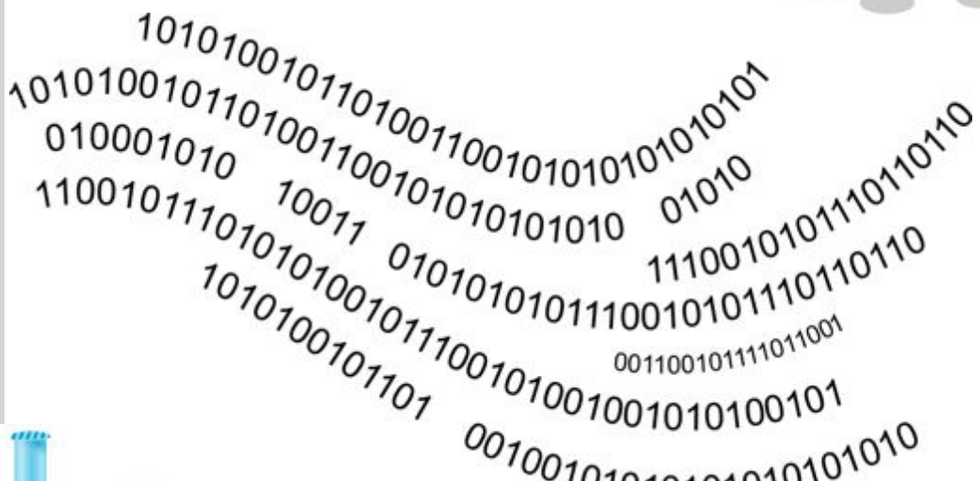
NSF Workshop: Towards Big Steps Enabled by “Big Data Science”

January 29-30, 2015

The Big Data Status Quo

Data Consumers

Data Producers



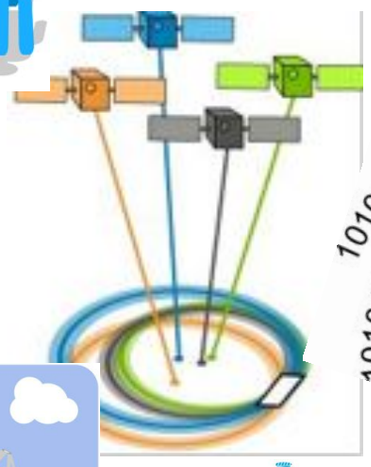
Where do the citizen scientists come in?

Building a seamless knowledge discovery system that optimizes the human-machine partnership…

Human-Machine Partnership

Data Producers and Consumers

Data Producers and Consumers



Data Producers and Consumers

Data Producers and Consumers

…integrating across disciplines, knowledge systems and data types.

The Galaxy Zoo Story

Amazingly, only two basic galaxy shapes – but very complex, no two are the same.

Spirals

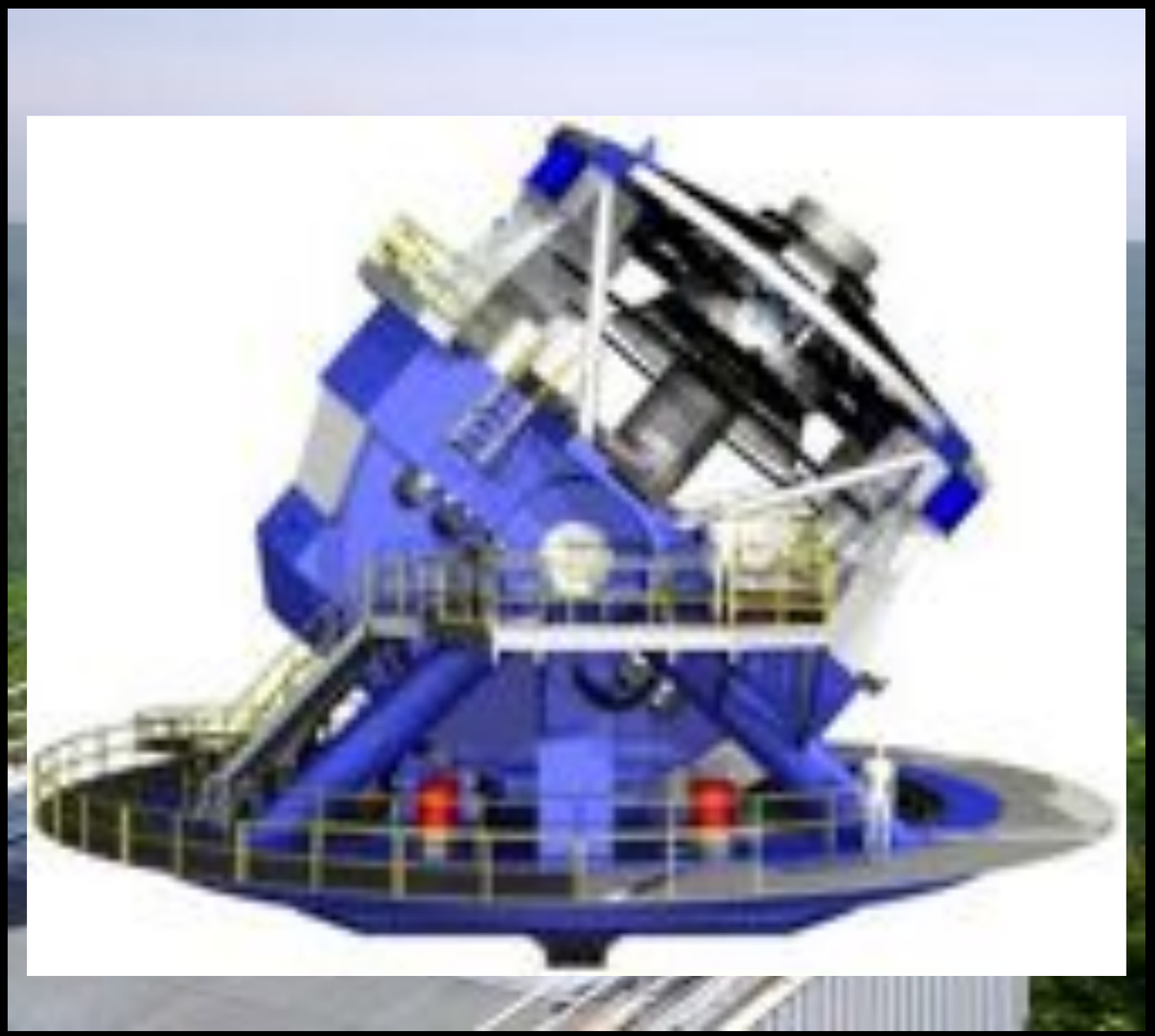
Lots of star formation so mostly (but not all!) blue-ish.

Ellipticals

Older star populations so mostly (but not all!) red-ish.

Historically, astronomers would classify galaxies “by eye” even when 10,000 images! But employ machine algorithms in big data era.

Astronomy as example of data flood



2.0 meter Sloan Digital Sky Survey Telescope
Apache Point, New Mexico

Large Synoptic Survey Telescope design

1980 – Palomar Sky Survey:
10,000 galaxies
One expert can classify galaxies visually.

2000s - Sloan Digital Sky Survey:

~10^6 galaxies

2020s - Large Synoptic Survey Telescope (LSST)

~10^{10} galaxies

Machine algorithms use proxies: ~80-90% efficient

But how can this algorithm be “computerized?

The human brain can be processed in parallel fashion...

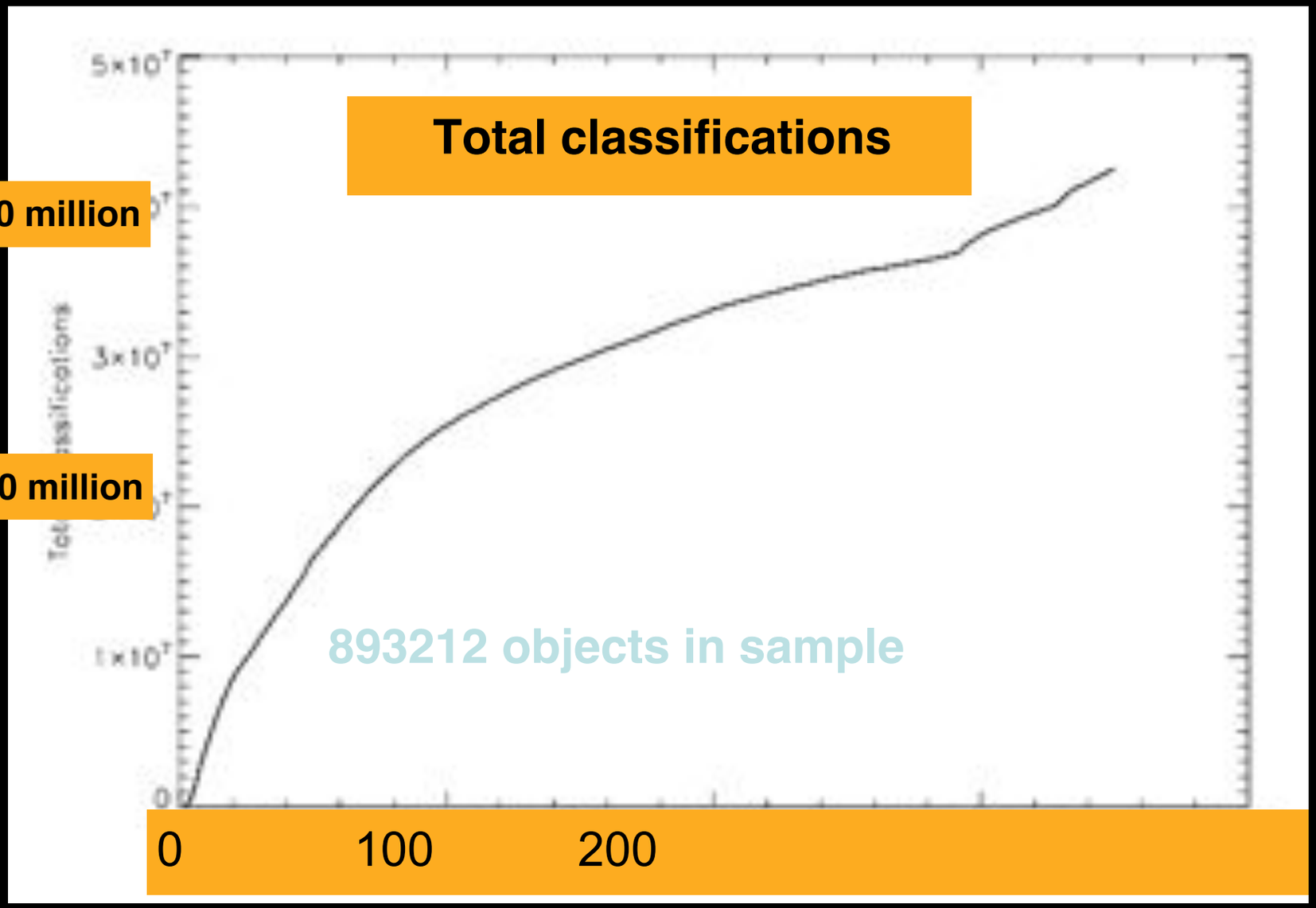
Crowdsourcing!

Galaxy Zoo launched in 2007 inviting public to classify galaxi

In 1.5 years, 35 million classifications by ~150,000 users

Roughly 3.3 continuous person-years!

First six months



After ‘cleaning’ raw clicks:

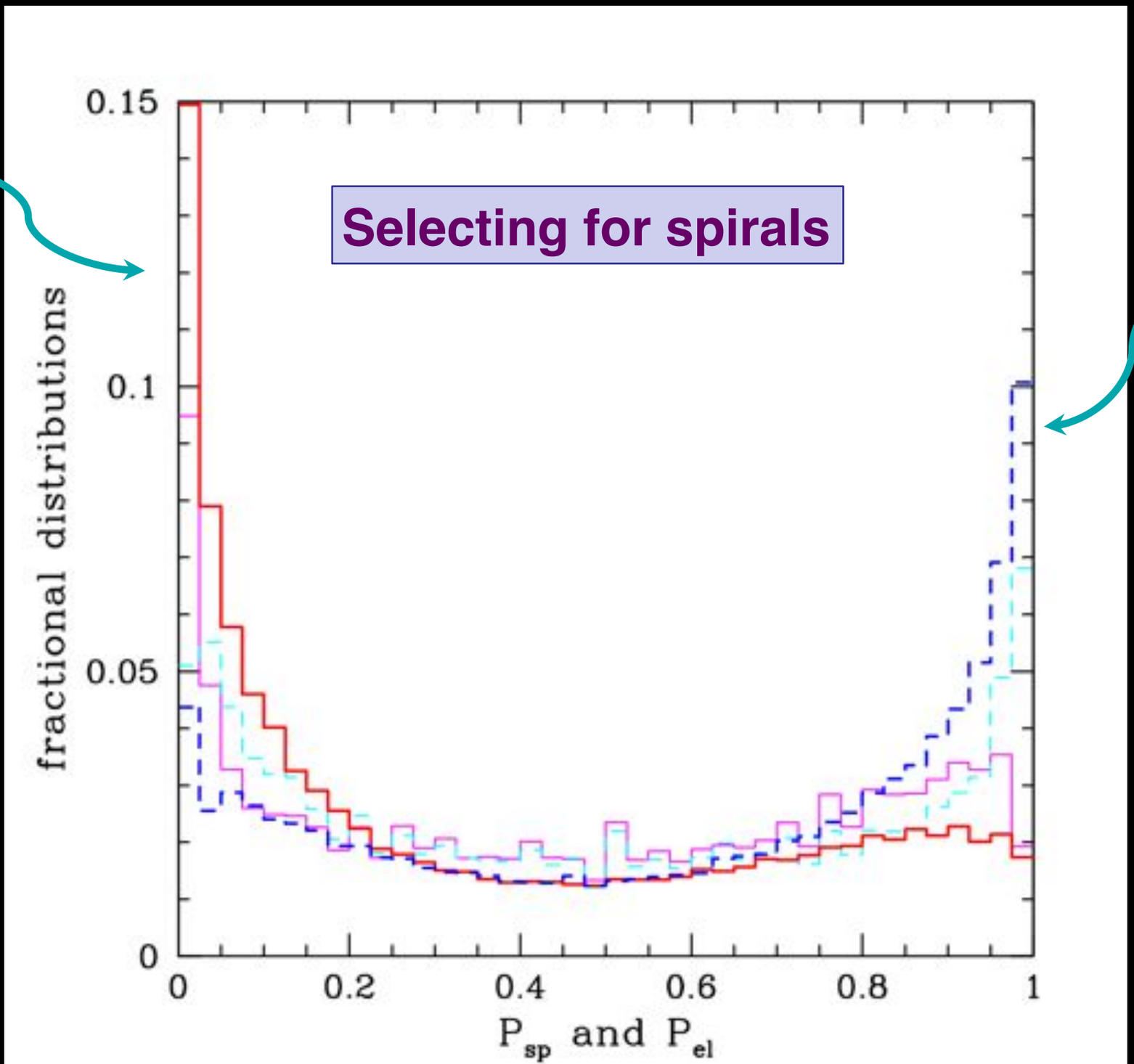
34,617,406 classifications by 82,931 users

Roughly 3.3 continuous person-years!

Combining clicks

"Pure" elliptical sample

"Pure" spiral sample



Galaxy Classification – now as probabilities!

Beyond Galaxy Zoo … to the Zooniverse

Types of Citizen Science

1. **Passive** – distributed computing so no involvement from volunteer other than their computer or the “internet of things” (SETI@Home, Airegg)

2. **Data Collection** – distributed sample collection/observation (**bird counts**, variable star observations, weather data, **participatory monitoring**)
   - “Classic" citizen science solving the need for distributed sensors as well as current internet-enabled data collection projects.

Zooniverse developed to exploit this niche !

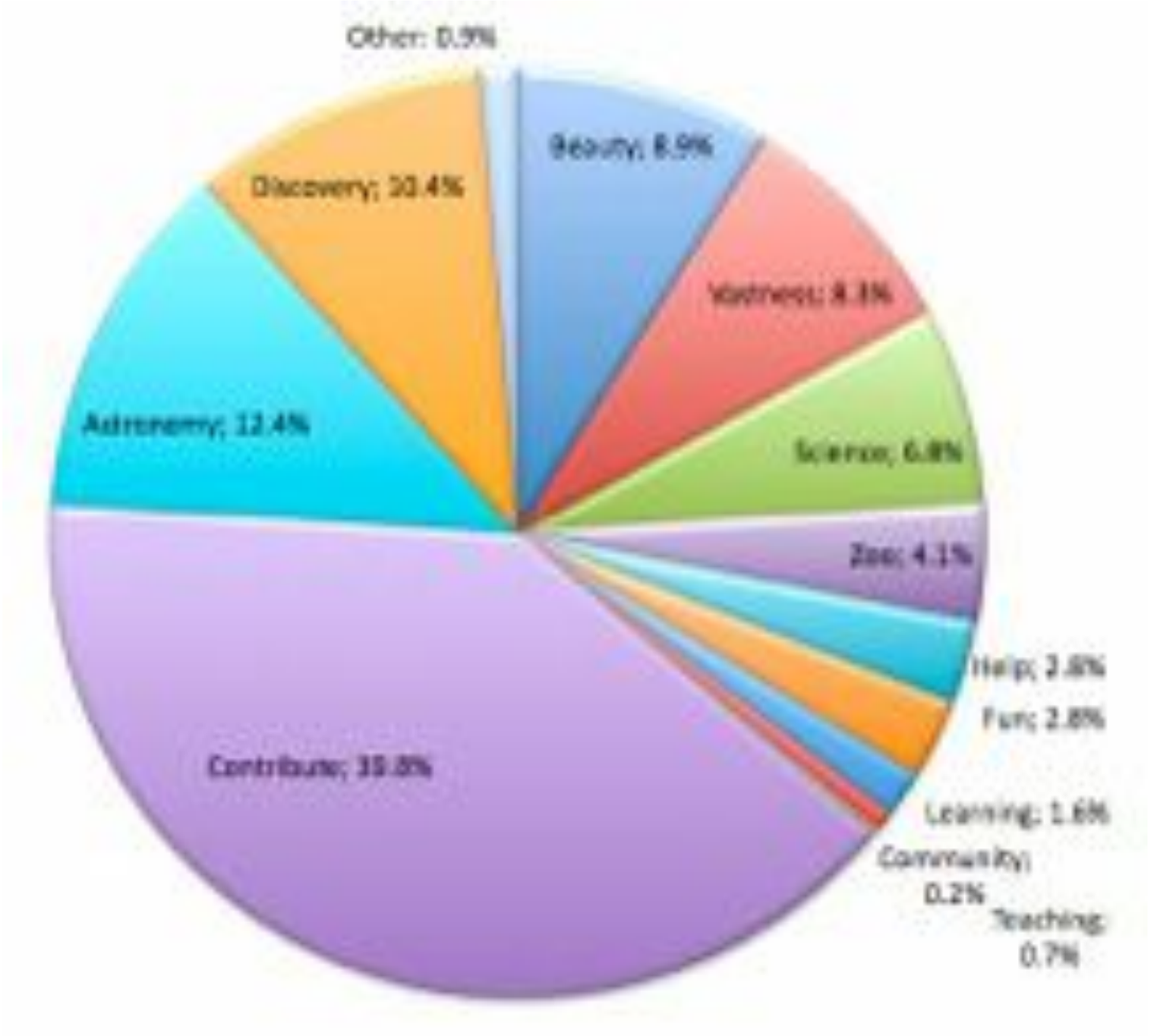
3. **Data Analysis** - distributed data analysis: enabled by the internet (**Galaxy Zoo**, FoldIt!)
   - Solution to large amounts of complex data where volunteers are part of data-processing pipeline for complex systems e.g. pattern matching

But first need to answer:

1. **Why would the crowd be interested?**
2. **How much time do they have?**

Beyond Galaxy Zoo … to the Zooniverse

Motivation to Participate in Galaxy Zoo



"**Contribute to Research**" most common motivation to participate in Galaxy Zoo. Wow!

Raddick et al. Astronomy Education Review (2010) vol 9, 1

Beyond Galaxy Zoo … to the Zooniverse

Cognitive Surplus!

Goggle Boxes

Hours spent…

200 billion hours

A year spent watching TV by US adults

US Adults: 200 billion hours a year watching TV

16 years

every day!

100 million hours to create Wikipedia

Cognitive Surplus by Clay Chirky // InformationIsBeautiful.net

ZOONIVERSE CITIZEN SCIENCE PORTAL

Over 1 million volunteers worldwide contributing to real research through online crowdsourcing of data analysis.

 Solution to "Big Data" problem for complex data (images, simulations, texts, videos, sound clips…)

 30 projects and growing: astrophysics, climate science, biology, humanities, nature **generating over 500 million classifications**

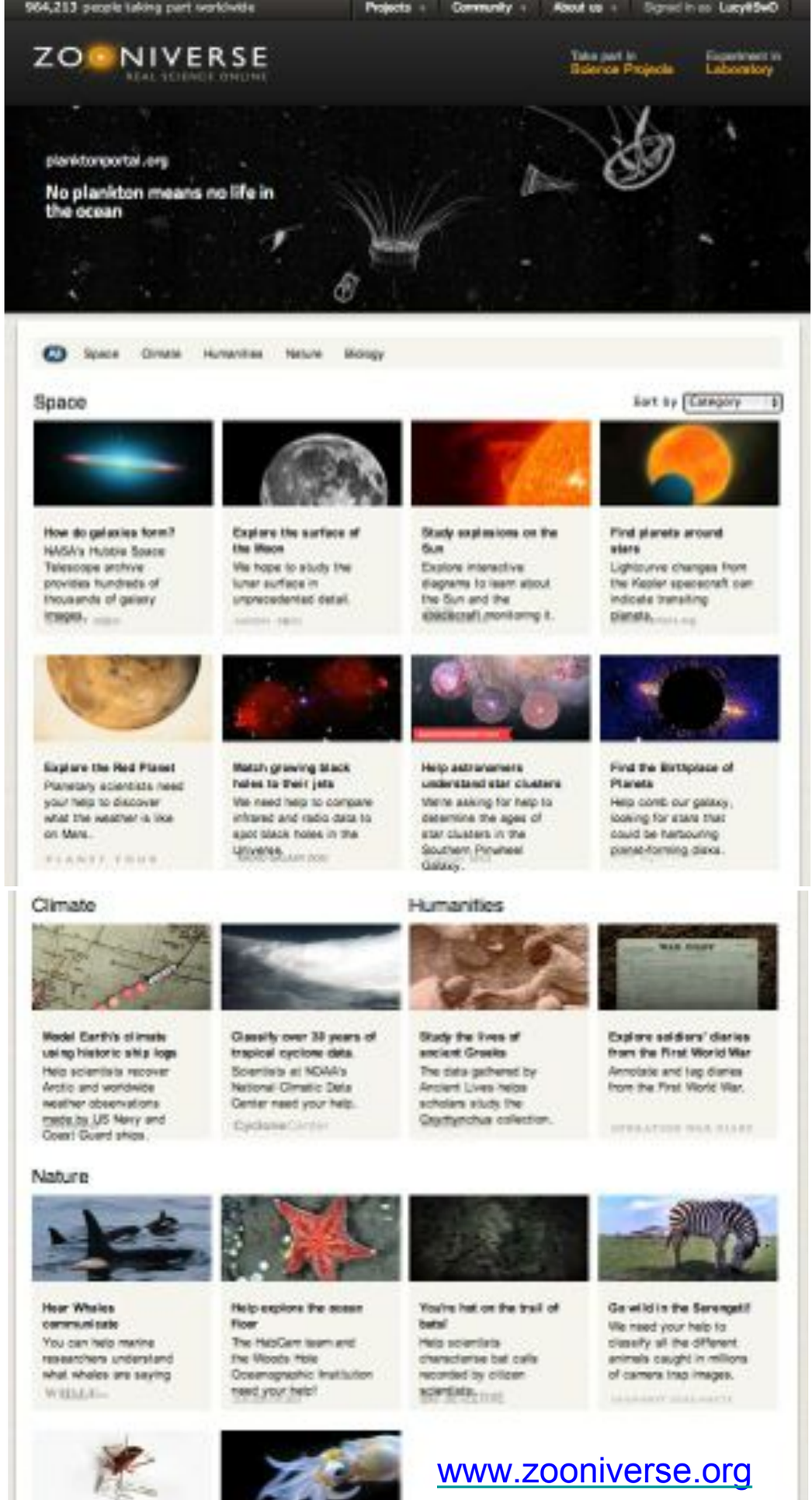
Over 60 peer-reviewed publications including several discoveries made by members of the public.

Hosted on Amazon Web Services – instantly scalable to 100,000's users

Funded by: National Science Foundation, Sloan Foundation, Google Global Impact, Leverhulme Trust

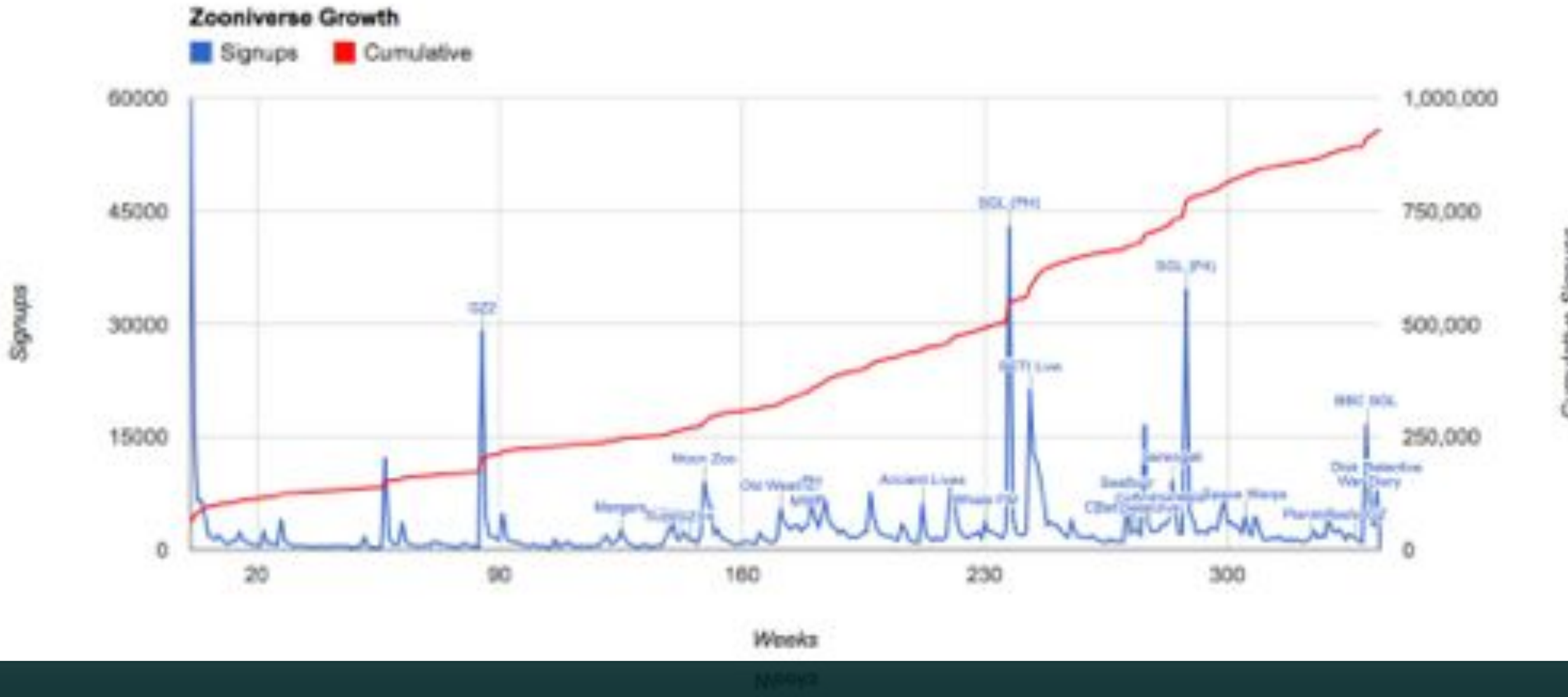
Operated by: The Citizen Science Alliance
www.citizensciencealliance.org

Member Institutions: Oxford, Portsmouth, Nottingham, ETH Zurich, Carto DB, Johns Hopkins, ASIAA, Adler Planetarium, U Minnesota



www.zooniverse.org

Growth of Zooniverse



Volunteer contributions to Old Weather: each box = 1 volunteer; size = number of transcriptions

ZOONIVERSE

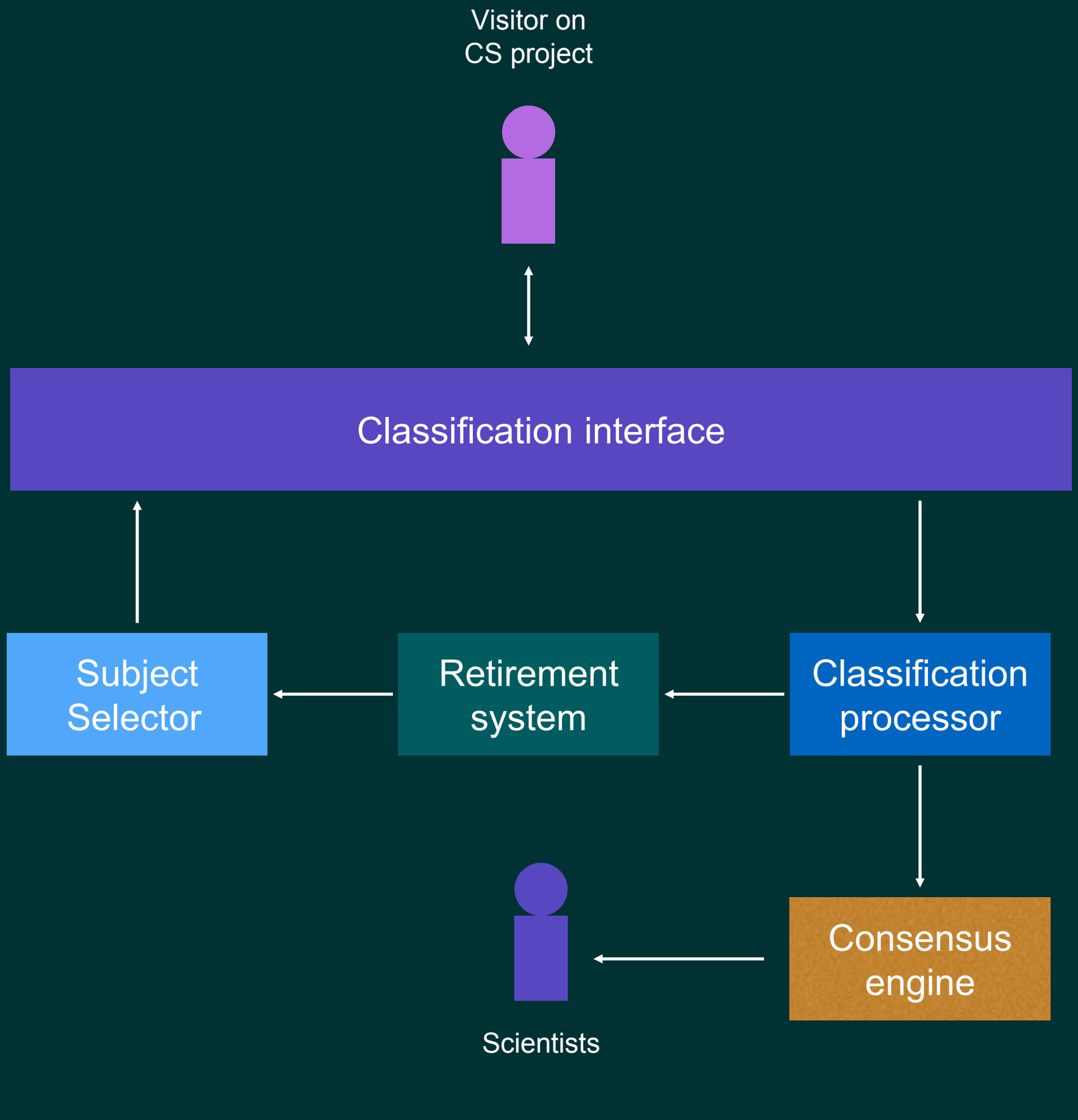
What data do our consumers want?

Consumer = research team, producer = citizen scientist

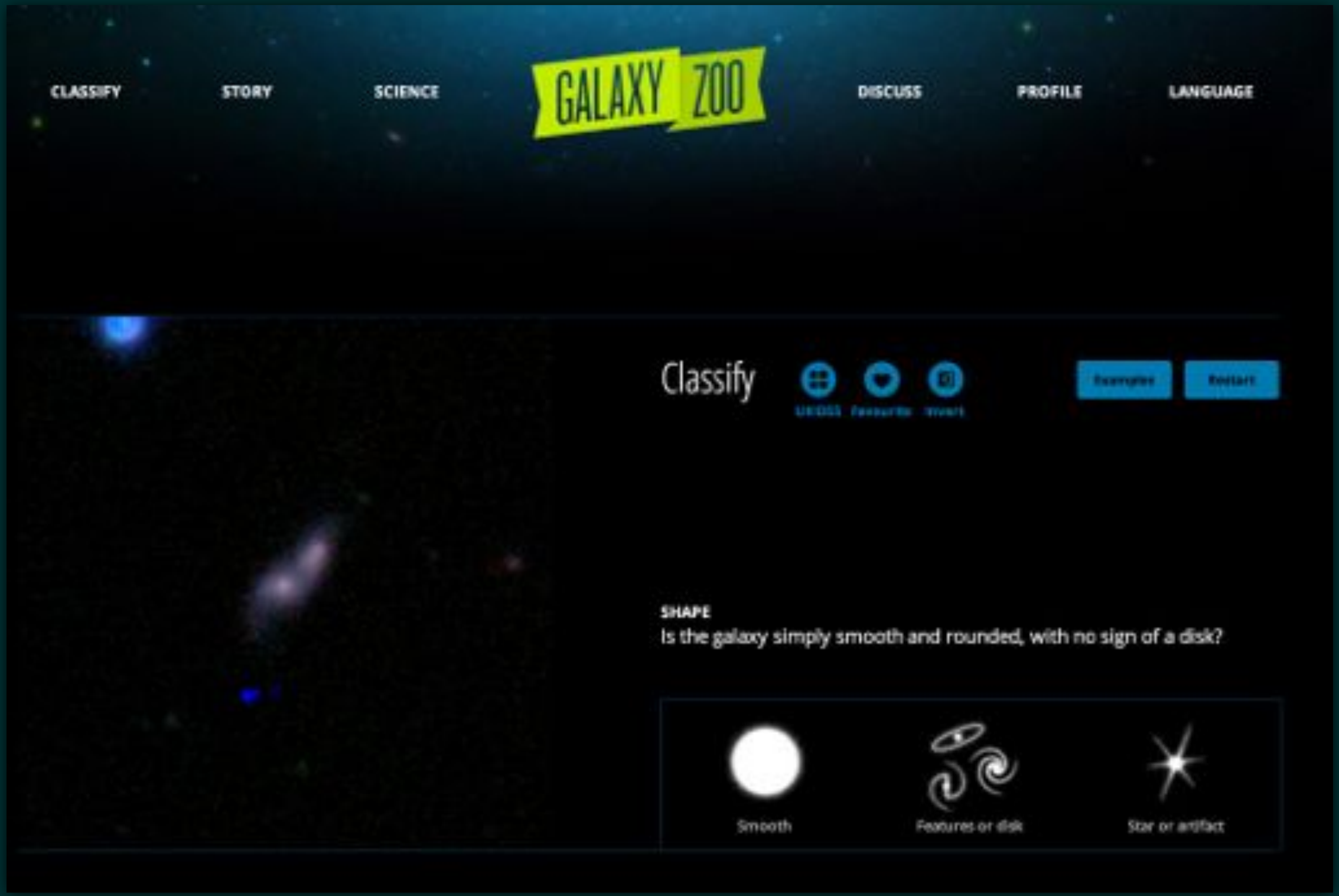
Data products derived from applying HPU heuristic processing on input data

- **variety of tasks**
- **variety of output data types**
- **final products require application of consensus algorithms (simple or complex)**
- **training sets for improving machine algorithms**
- **data that allows the study of the process of citizen science**

Consensus outputs on complex data inputs.

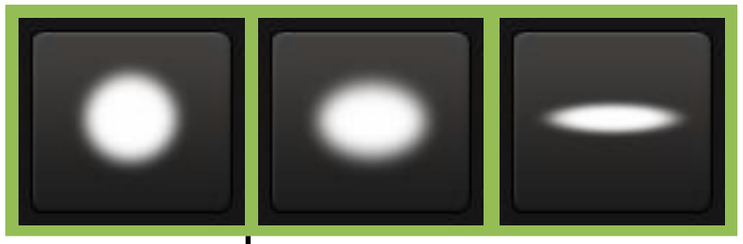

Visitor on
CS project
Classification interface
Subject
Selector
Retirement
system
Classification
processor
Consensus
engine
Scientists

Decision tree

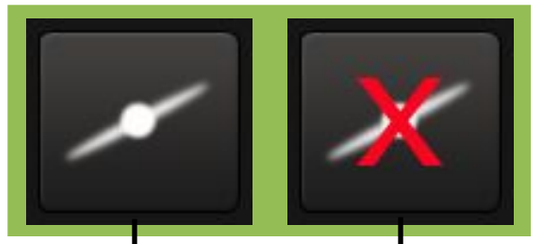


Is the galaxy simply smooth and rounded, with no sign of a disk?

How rounded is it?

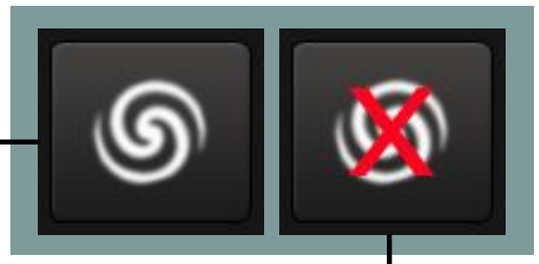


Could this be a disk viewed edge-on?



Is there a sign of a bar feature through the centre of the galaxy?

Is there any sign of a spiral arm pattern?



Does the galaxy have a bulge at its centre? If so, what shape?

Is there anything odd?

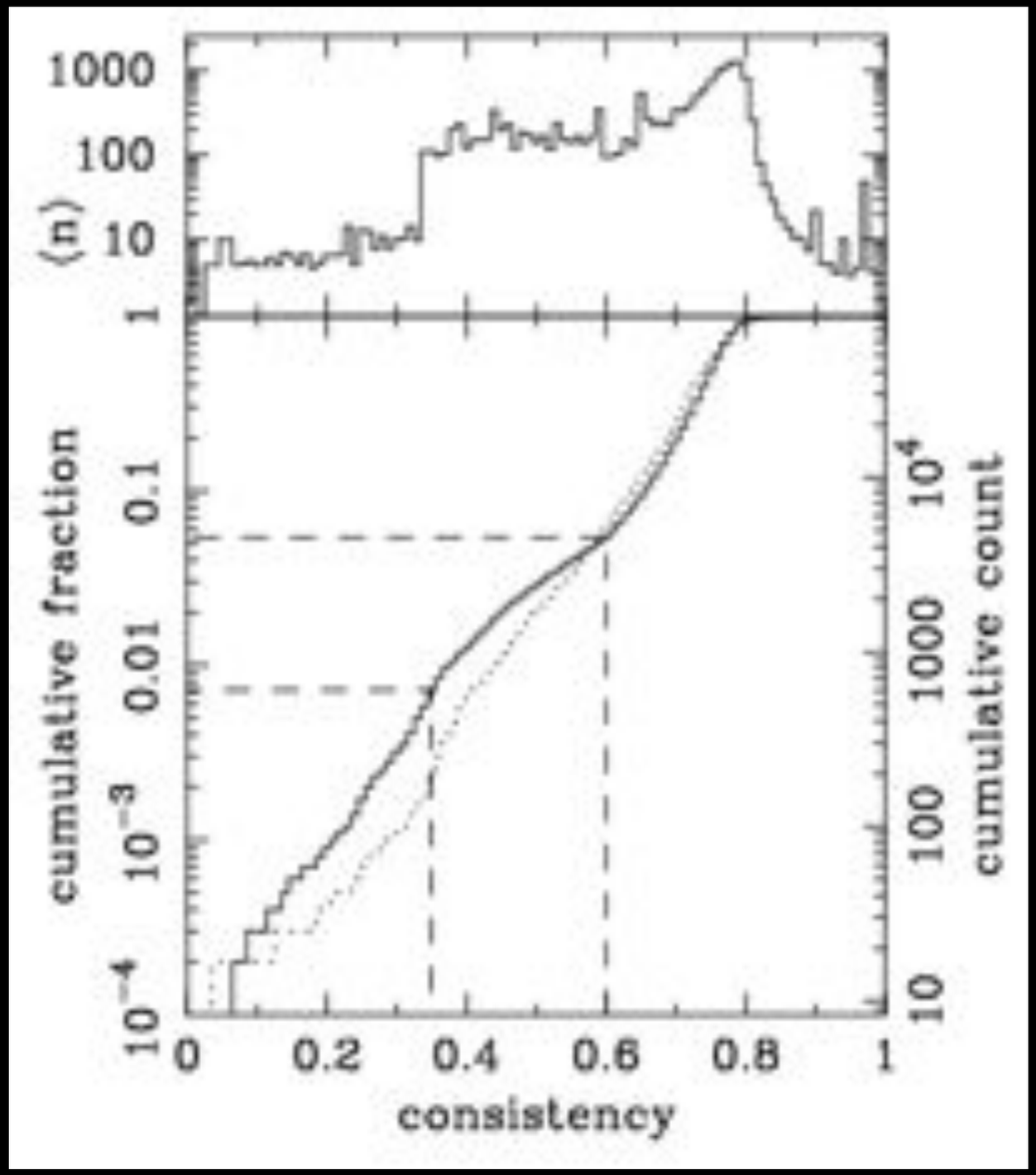
Is the odd feature a ring, or is the galaxy disturbed or irregular?

How tightly wound do the spiral arms appear?

How many spiral arms are there?

How prominent is the central bulge, compared to the rest of the galaxy?

Simple Consensus: user weighting

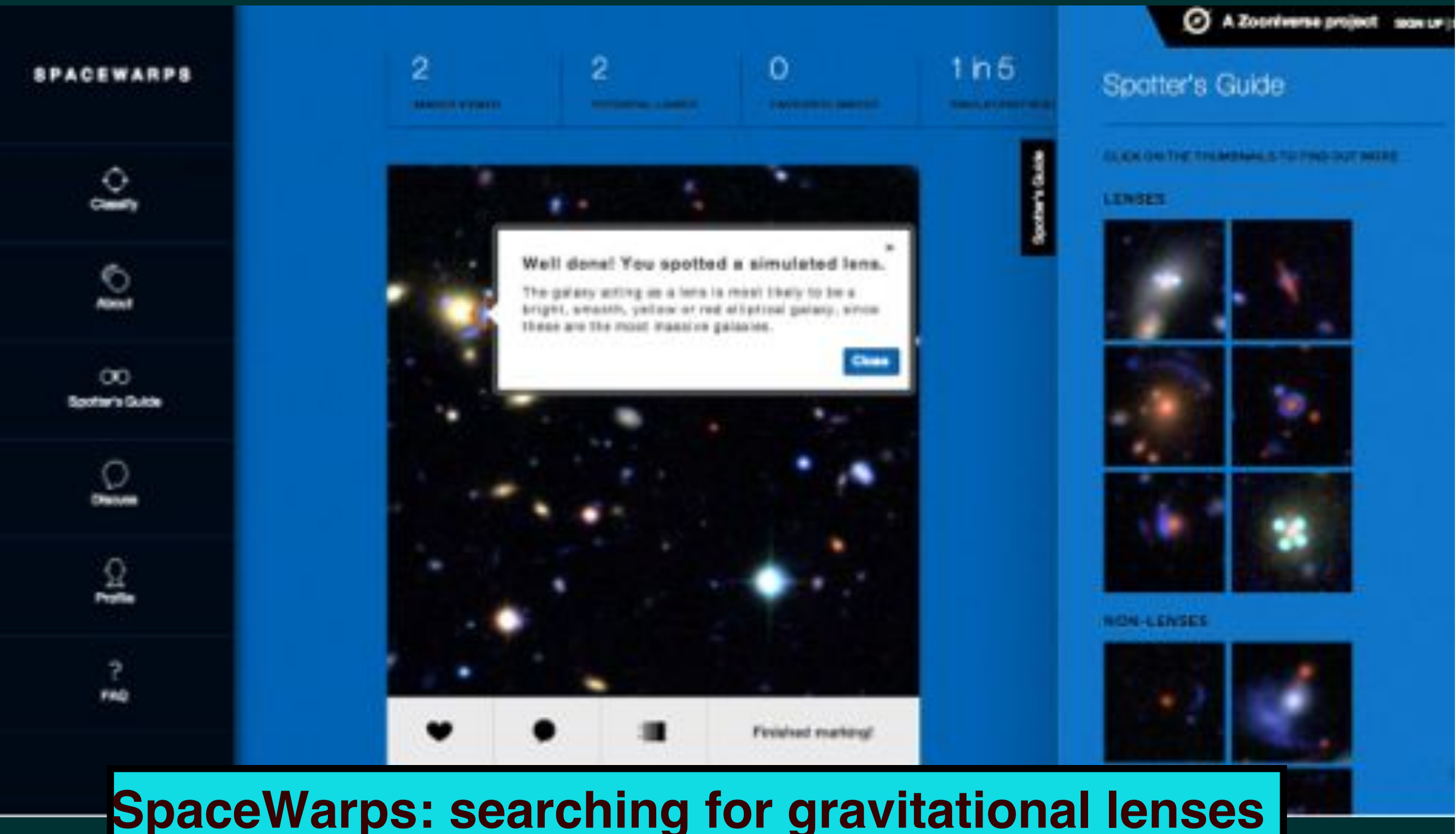


- Should we weight a user based on whether they agree with the majority?
- Or should we weight a user based on whether they agree with experts?
- What about bots?

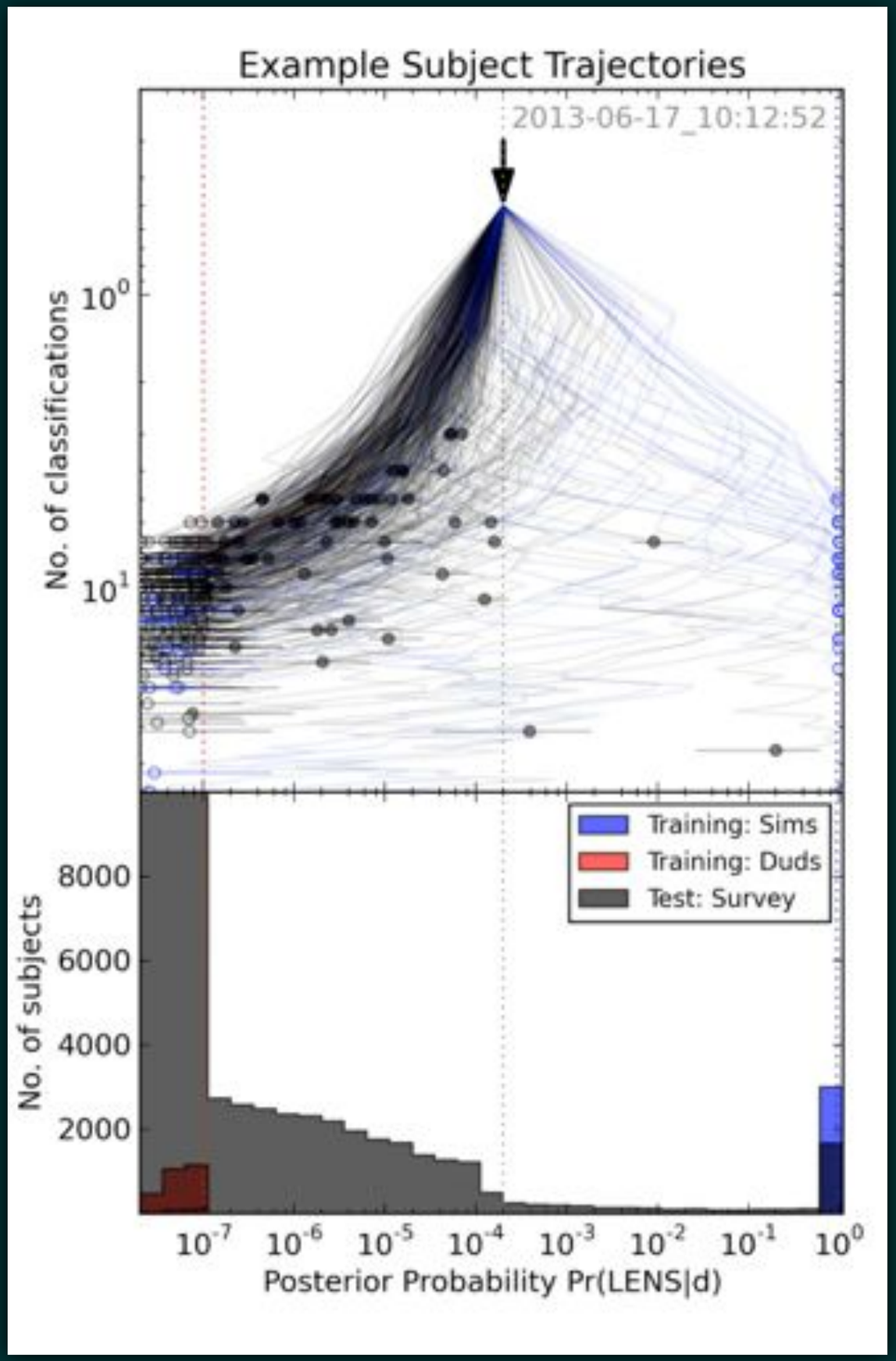
Galaxy Zoo 2 weighting process based on user consistency.

Willett, K. W., et al., Galaxy Zoo 2: detailed morphological classifications for 304 122 galaxies from the Sloan Digital Sky Survey, MNRAS, 435, 2835 (2013)

Marking
Real & Simulated data



SpaceWarps: searching for gravitational lenses



Complex consensus: dynamic subject retirement

- Initial probability for each subject prior to first classification: Pr(Lens) ~ 2×10^{-4}

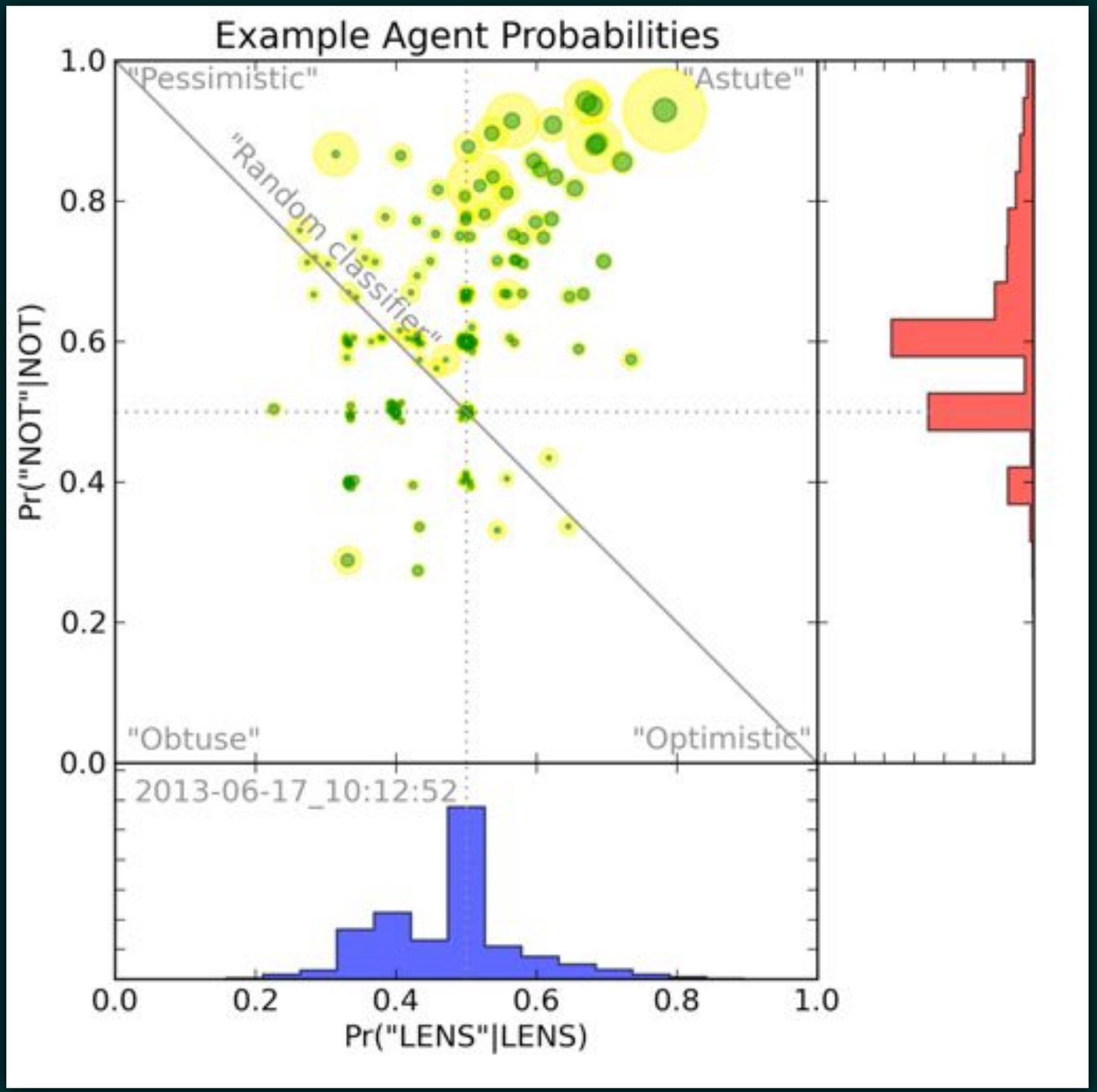
- Rejection threshold: Pr(Lens) ~10^{-7}
- Detection threshold: Pr(Lens) ~ 0.95

- As soon as a subject crossed the rejection threshold it was retired from the site.

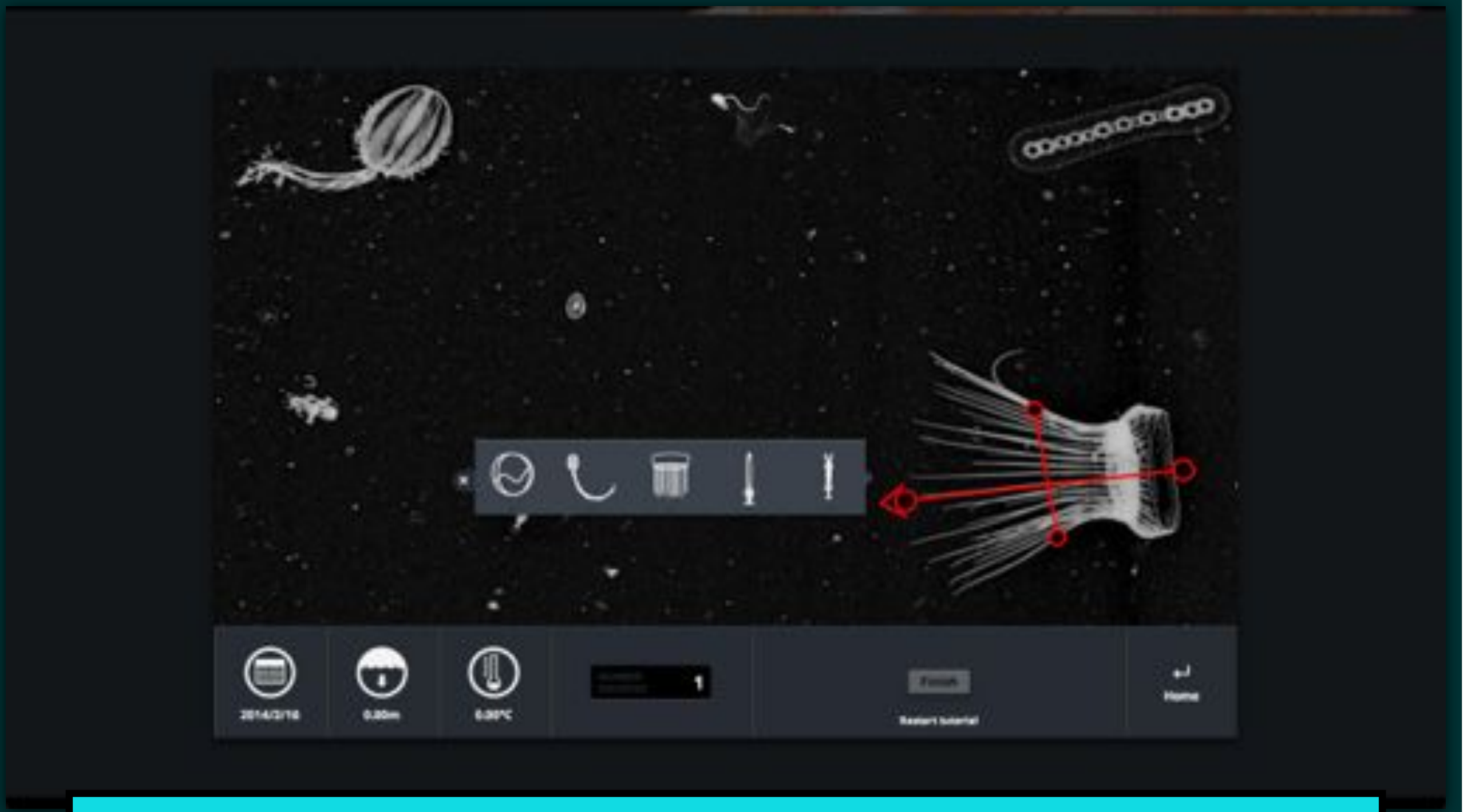
- About 10 classifications are required for a subject to reach the retirement threshold.

Marshall et al “Space Warps: Crowd-sourcing the Discovery of Gravitational Lenses”, in preparation

Provides information about each classifier



Marking and measuring



Plankton Portal: understanding species and their distributions

Transcription: structured

Operation War Diary: transcription of unit diaries from Western Front

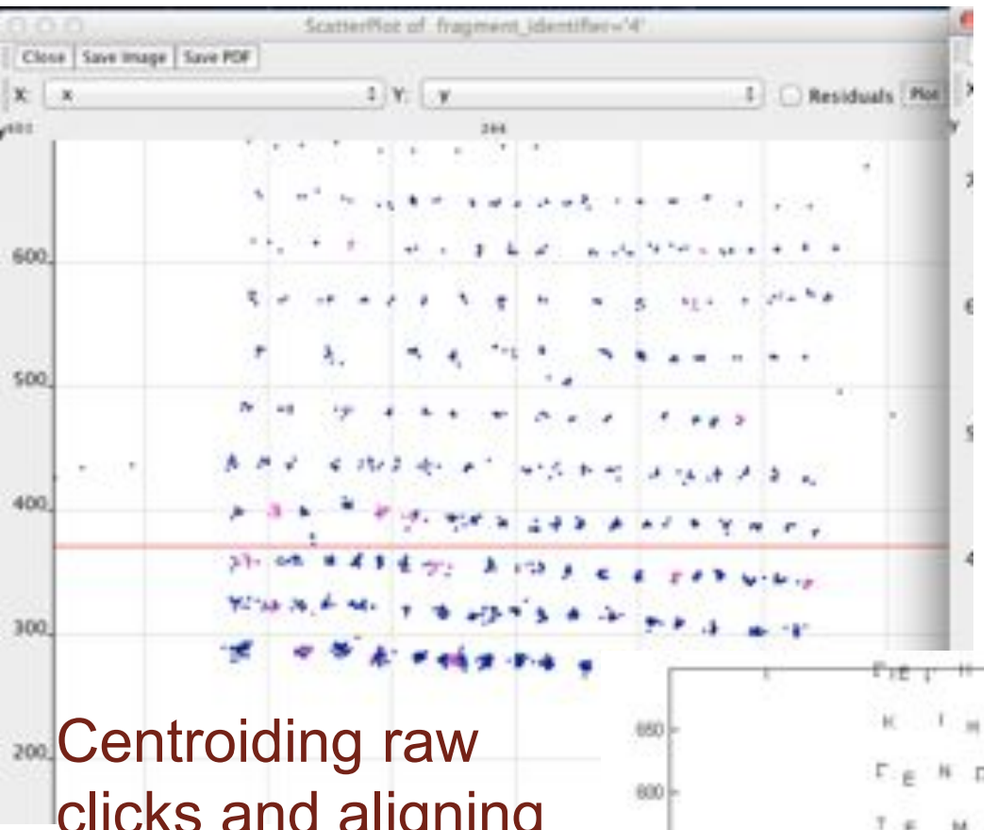
Transcription: full text

Ancient Lives: Searchable text-based cultural studies via transcription of 2000-year-old papyri fragments from Oxyrhynchus, Egypt.

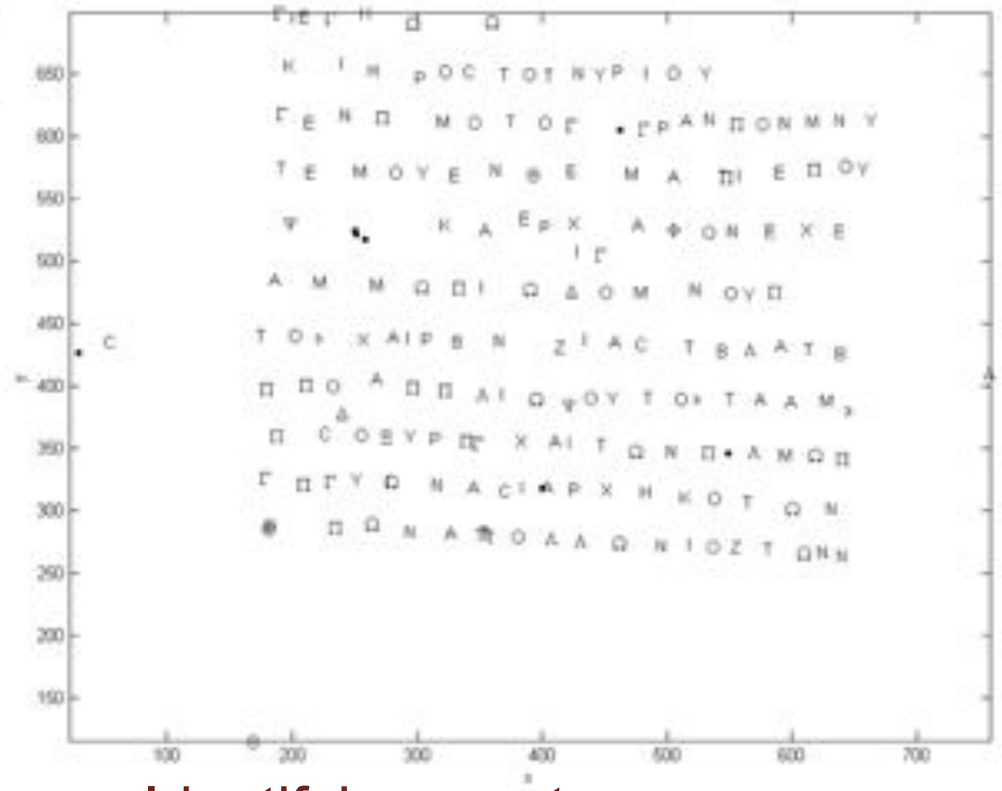
300,000
volunteers

classifications
100m

Ancient Lives: “Minority report” consensus pipeline allows transcribed text to be edited, annotated and curated.



Centroiding raw clicks and aligning centroids to factor out scribe’s slant.



Identifying most probable letter with each centroid.

Annotation and curation tool displaying original image and consensus transcription that can be edited or augmented with metadata, matched to known texts.

Selection with filters

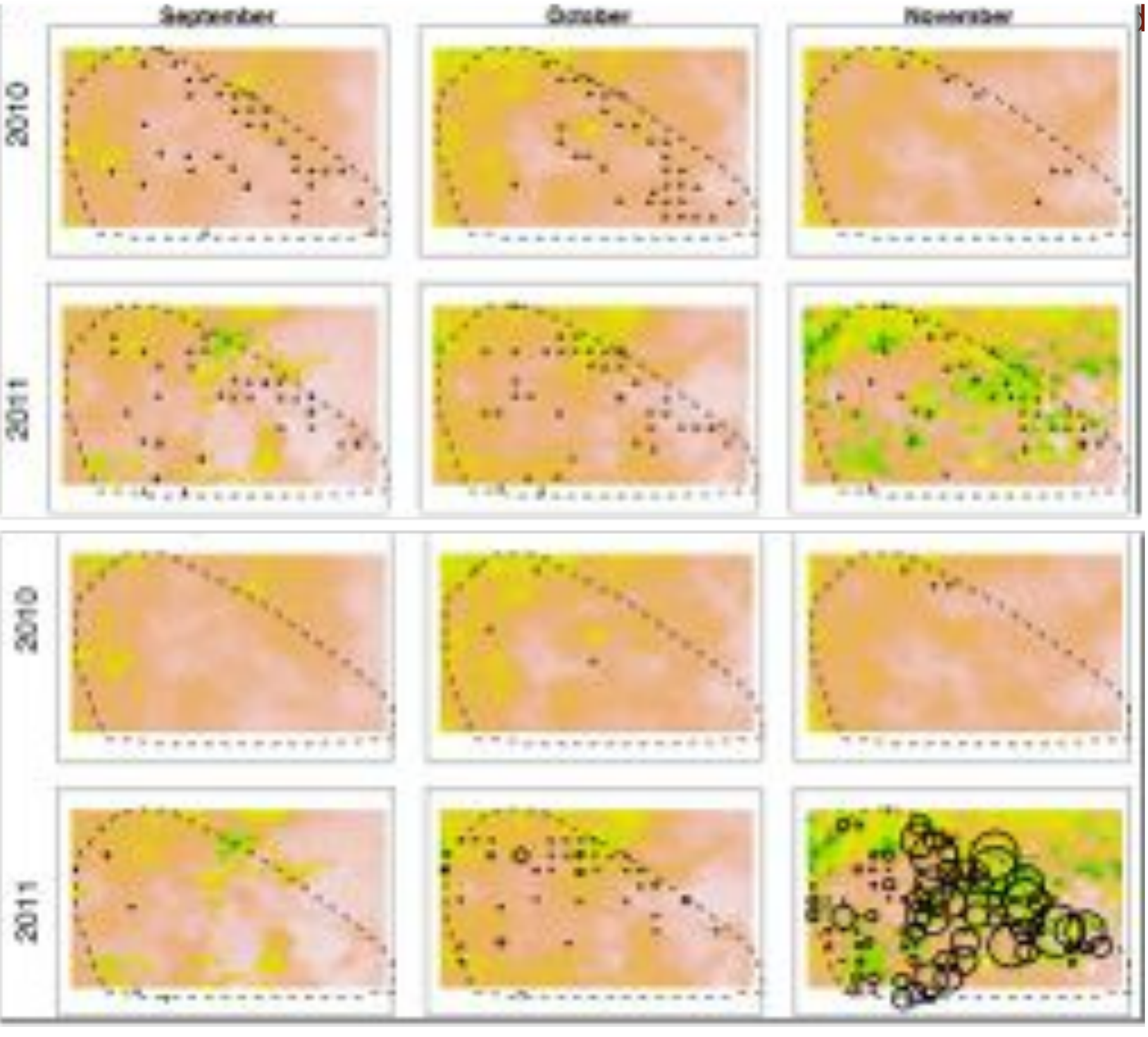
few

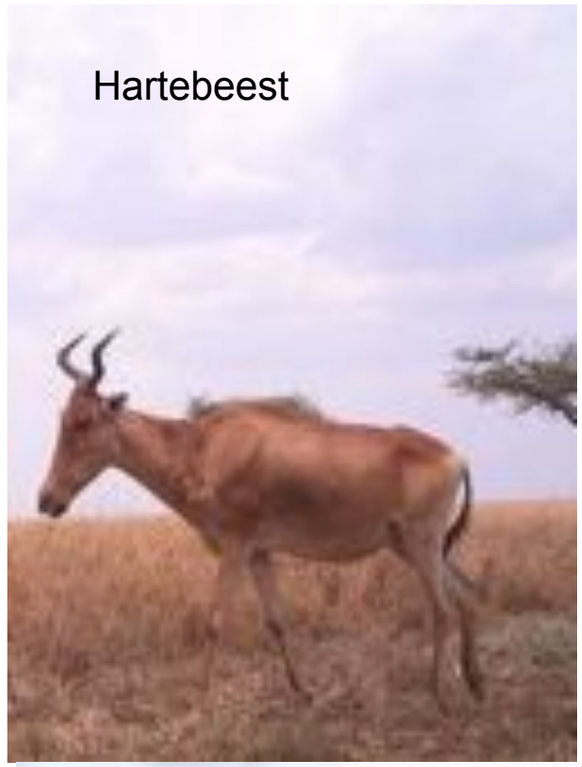
Many

Dry

Wet

UMN grad students developed idea, and are analyzing five seasons' data.

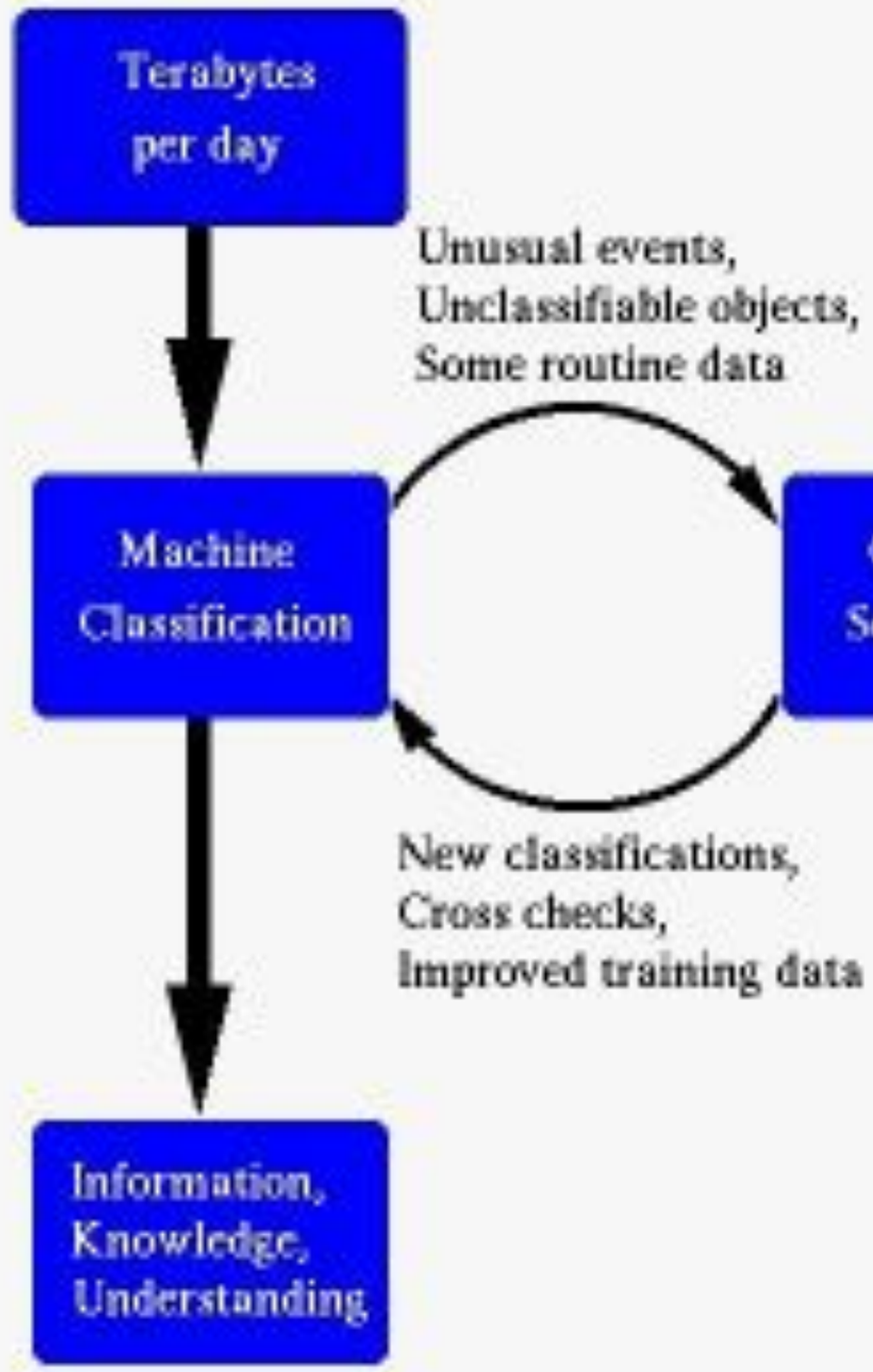




Comparison


WHALE FM
Welcome to the Whale Song Project
You can help marine researchers understand what whales are saying. Listen to the large sound and find the small one that matches it best. Click 'Help' below for an interactive guide.
Home
About
How to Take Part
Account
Help
16 NOVEMBER 2008 08:49
NAME: SCAR
LOG IN TO ENABLE TRACKING
CONFIRMATION
Once selected, you can view both the calls. If you think these represent the same type of whale call, then click match.
NEXT
MATCH

Human-Machine Partnership



Machine-learning cycle optimizing CPU+HPU

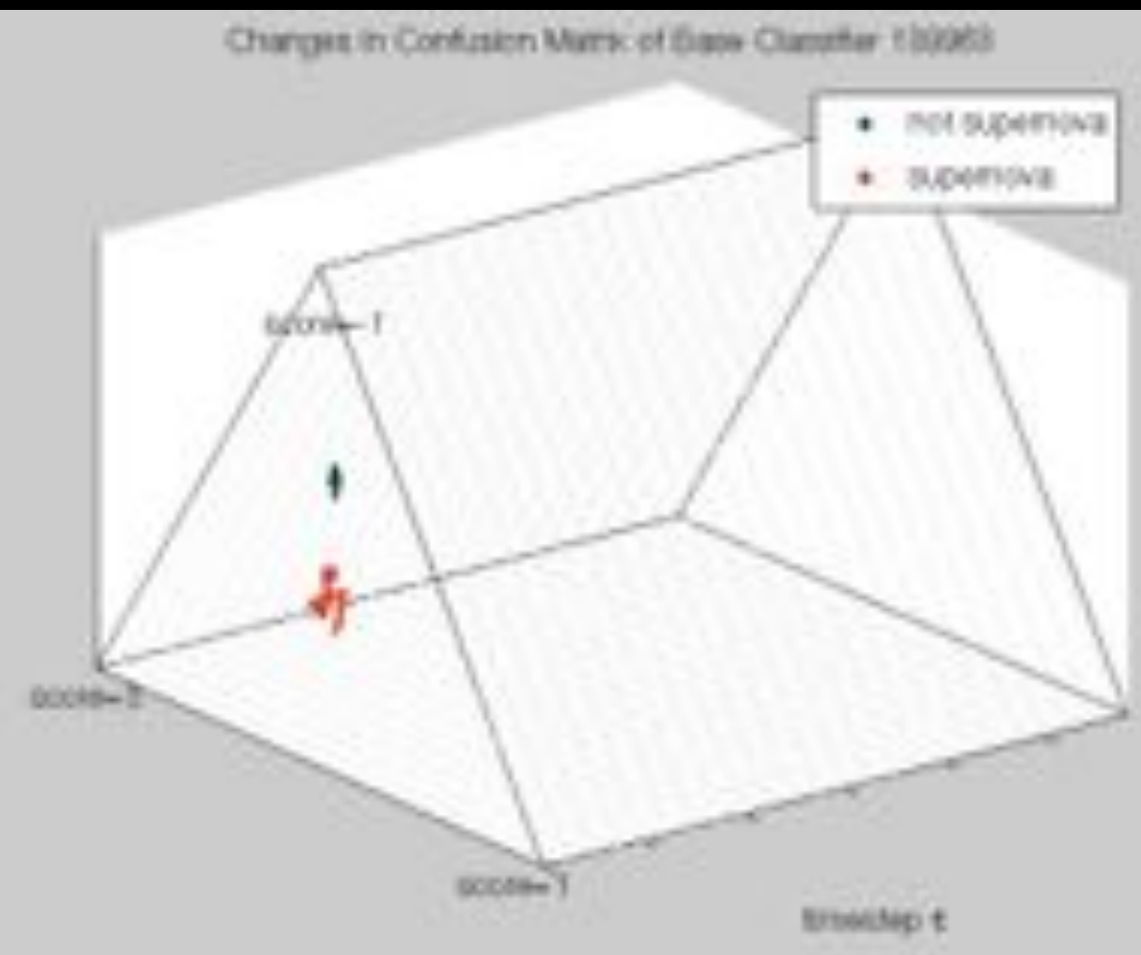
Supernova Zoo: implemented cycle in real time!

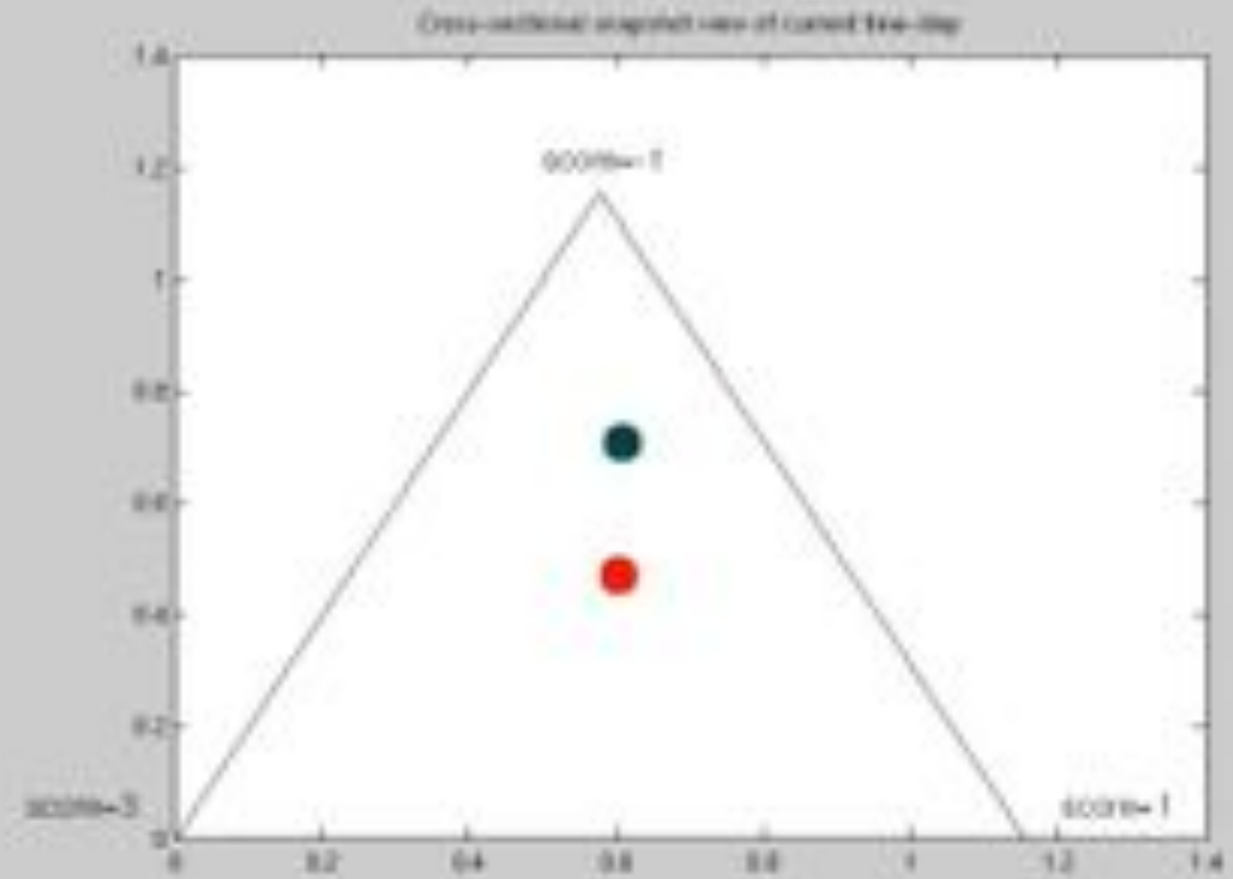
Smith et al, 2012 MNRAS, Vol 412, Issue 2, pp. 1309

Also enables “social system” research to optimize human attention.

Simpson et al, 2013 Studies in Computational Intelligence, Vol 474, pp. 1

Citizen Scientists and Machines: Optimizing Human Attention





"Dynamic Bayesian Combination of Multiple Imperfect Classifiers"

Simpson et al, 2013 Studies in Computational Intelligence, Vol 474, pp. 1

Learn when to give which volunteer which classification task!

ZOONIVERSE

What data do our consumers want?

Consumer = citizen scientist, producer = research team

Data that they know requires human processing at the same time must be engaged by project through compelling research coupled with engaged research team

Opportunity to make discoveries

Entertainment and education opportunities

Ethics of Citizen Science

Must be useful to researchers (a solid, well-defined research case drives the best project designs)

Must respect time and effort of volunteers

Therefore:

Must have a science case that *requires* the citizen science method of data processing

(Machine algorithms must currently fail in some manner for the science case)

Must acknowledge volunteers in publication

Must keep volunteers informed of progress and process

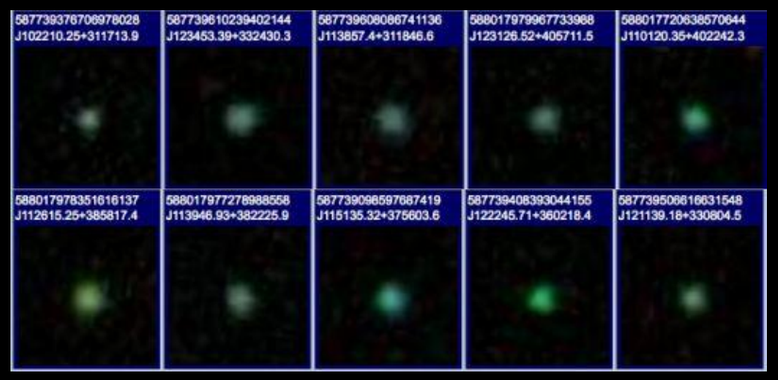
Do volunteers have the right to delete all the data they have contributed?

Classes of Knowledge Discovery:

Known knowns : Primary task. Data reduction by science team.

Known unknowns : Related to primary task. Results funneled to specific researchers.

Unknown unknowns : Serendipity. Currently rely on forum moderators to filter.



Need tools to accelerate knowledge discovery!

What has the Zooniverse taught us?

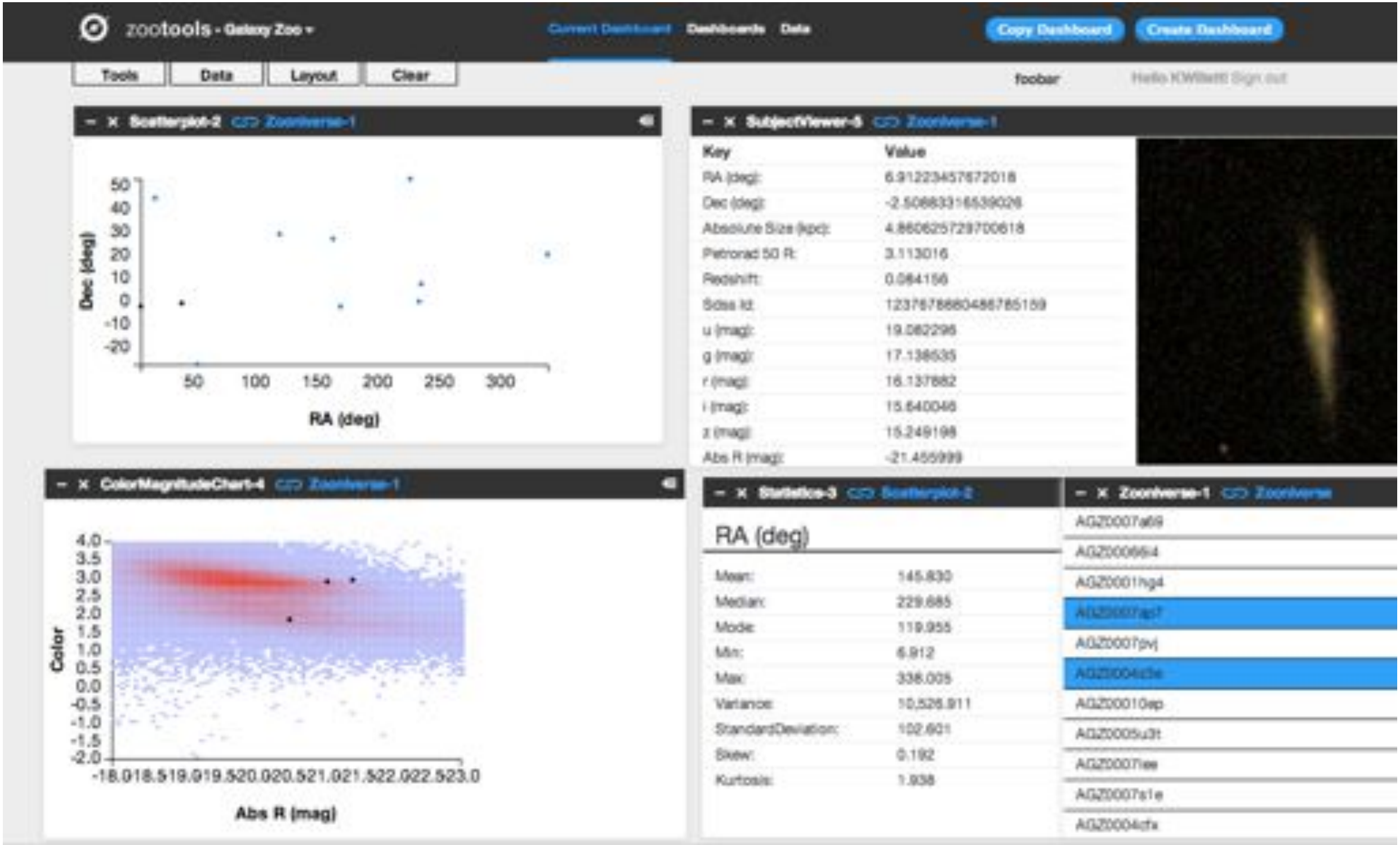
"Talk" – object oriented discussion tool where volunteers ask for help, create collections, and discuss their findings with each other and the science team.

"Blogs" – where volunteers hear from the research team about project science, papers in progress and other accomplishments.

Each project must have mechanisms for the volunteers to communicate with each other AND the science team.

An engaged science team is critical to project success

Data analysis tools



Enables volunteers to explore and visualize the data: tools.zooniverse.org

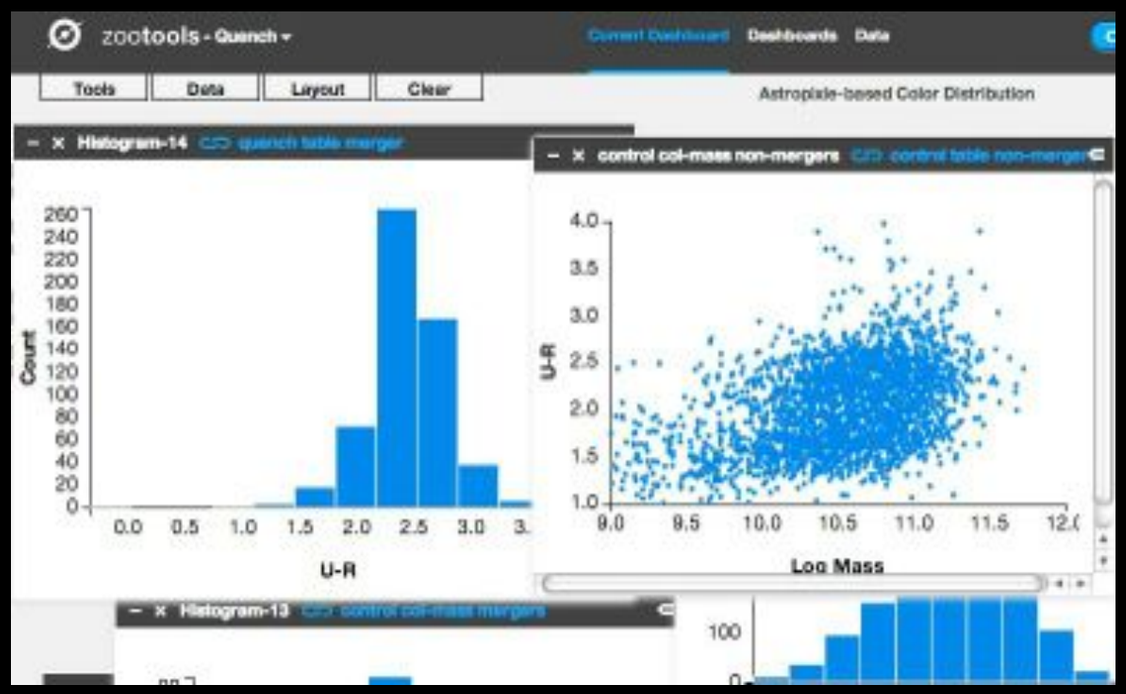
SCIENCE PROJECT GALAXY ZOO QUENCH TALK PAPER

Experience Science from Beginning to End

Stage 1:
Classification
(2500 citizen scientists)

Stage II:
Data Analysis & Discussion
(~250 citizen scientists)

Stage III:
Article Writing
(3 citizen scientists)

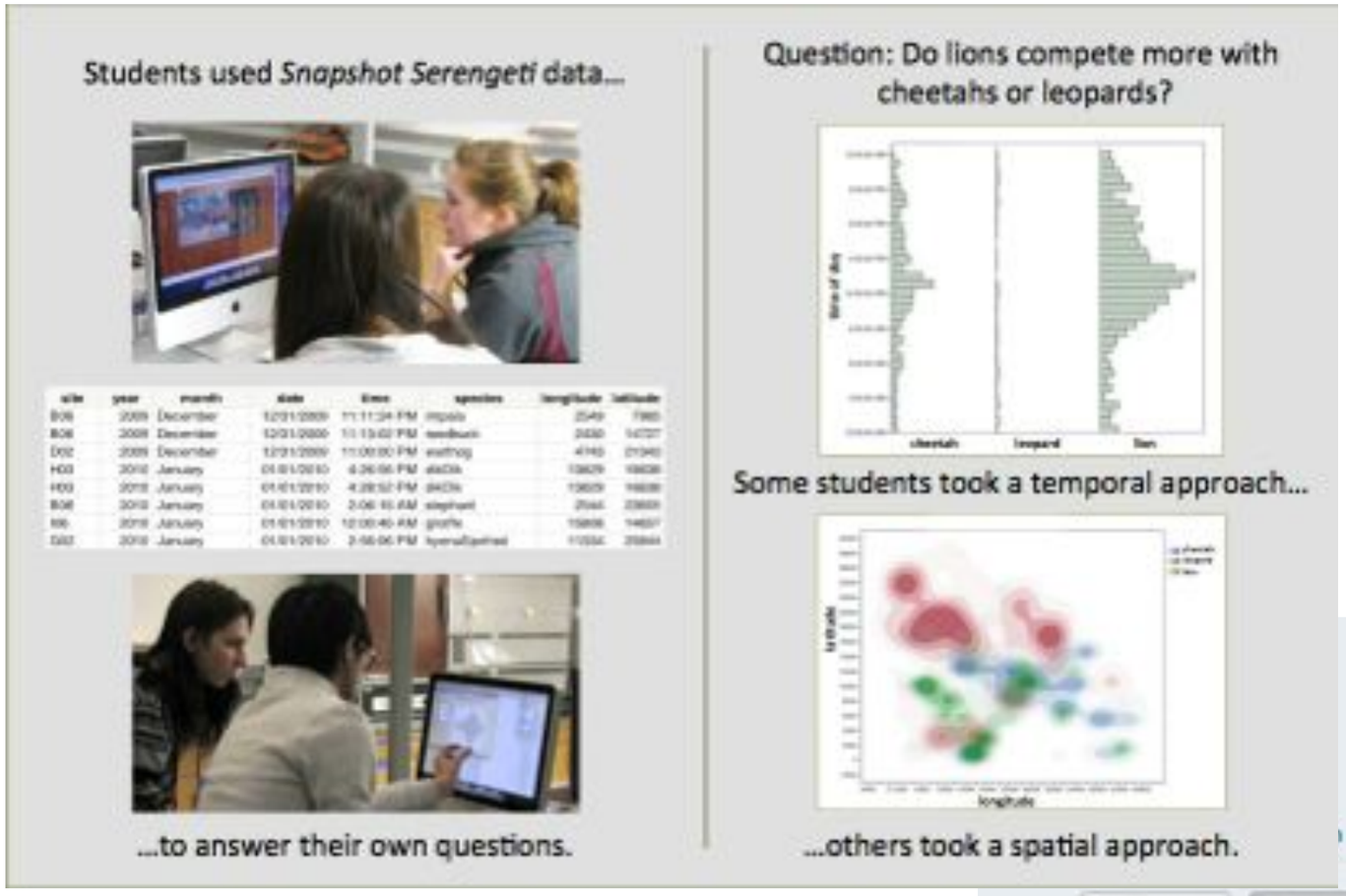


Authorea

lead scientist Dr. Laura Trouille NWU+Adler

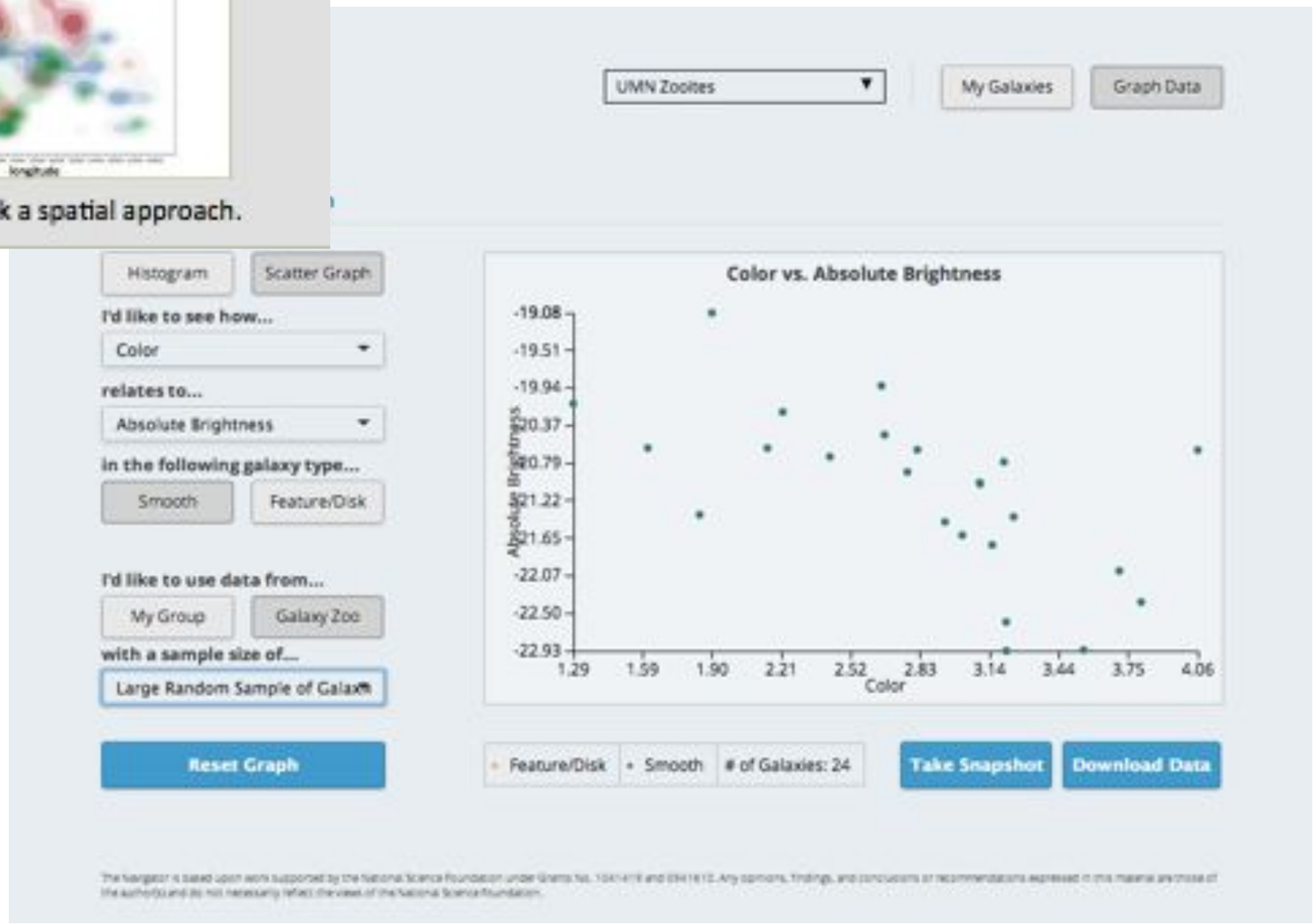
And of course…putting Zooniverse projects in the classroom.

Using Galaxy Zoo in high school or undergrad laboratory exercises.



From Annika Moe, Craig Packer

Using Snapshot Serengeti in non-bio major classes to enable authentic research experiences.



From Kyle Willett, Tommy Nelson

Formatting real data for the classroom:
http://www.galaxyzoo.org/#/navigator/home

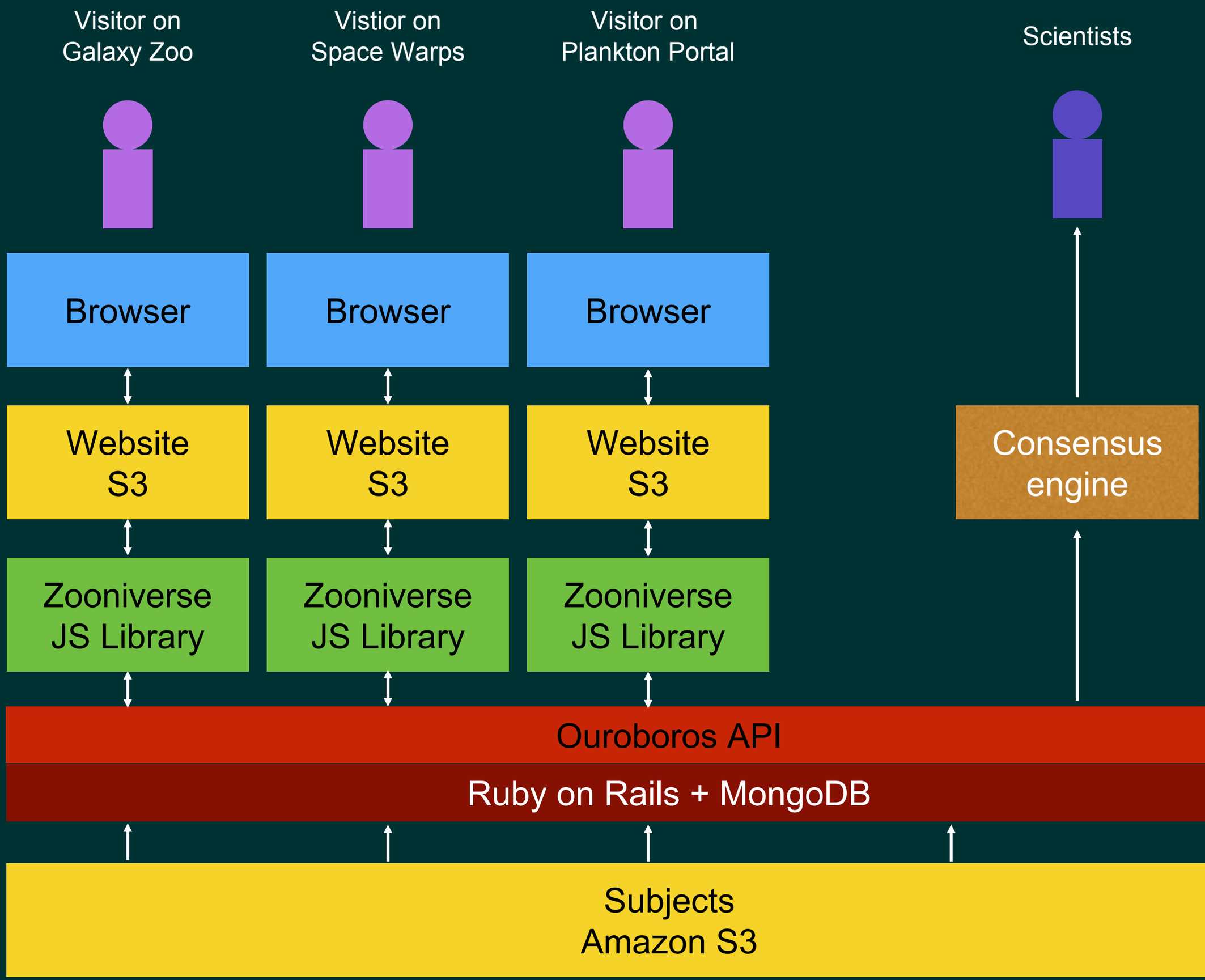
ZOONIVERSE

Infrastructure, funding and policies?

Centralized portal allows for public "brand", controlled innovation and consolidation of lessons learned. Difficulty in funding basic running costs - constant need for innovation-based funding can be destabilizing.

New "Panoptes" platform will allow anyone to spin up a zooniverse-style project through online templates (a la wordpress and blogs):

- **can be small projects, or behind University firewalls (health data privacy issues or other proprietary data) - may not want or need Zooniverse brand**
- **if hosted on Zooniverse, who pays for I/O and data storage (warm and cold)? which projects get selected through what process?**


Visitor on
Galaxy Zoo
Vistior on
Space Warps
Visitor on
Plankton Portal
Scientists
Browser
Browser
Browser
Website
S3
Website
S3
Website
S3
Consensus
engine
Zooniverse
JS Library
Zooniverse
JS Library
Zooniverse
JS Library
Ouroboros API
Ruby on Rails + MongoDB
Subjects
Amazon S3

Visitor on Galaxy Zoo

Vistior on Space Warps

Visitor on Plankton Portal

Third parties

Browser

Browser

Browser

Machine Learning

Dynamic programatic project front end

Processed event stream

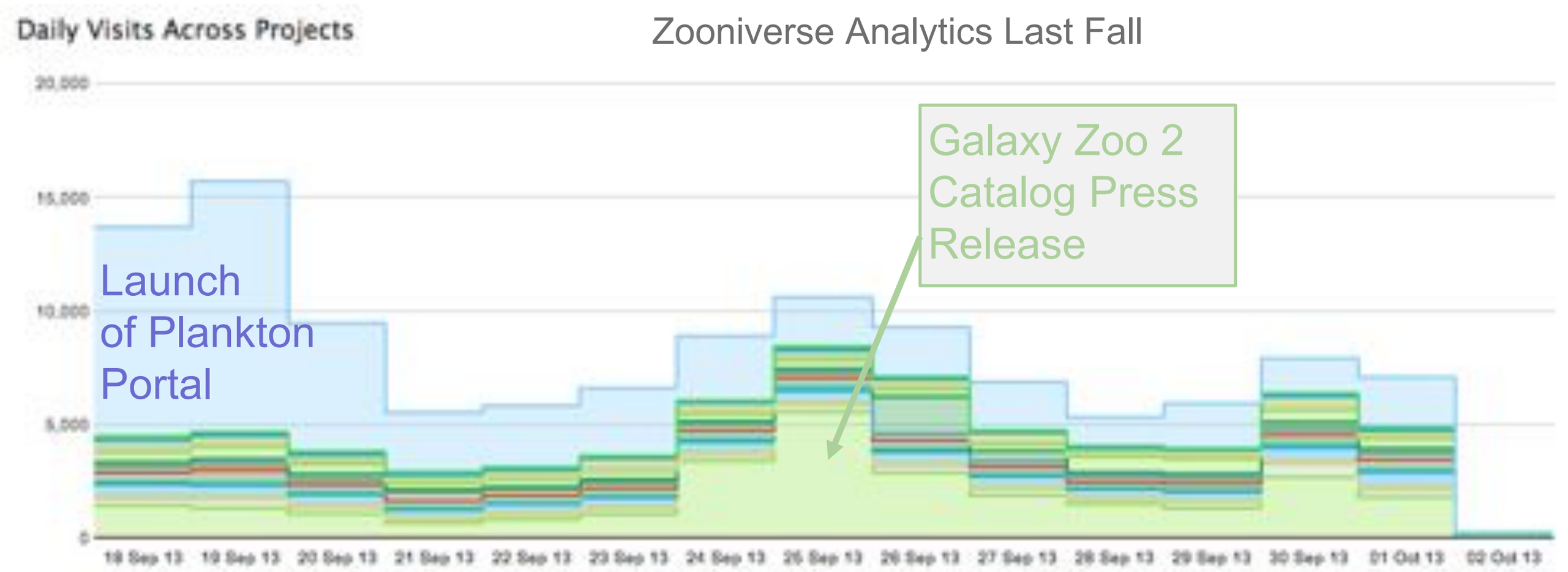
Request API

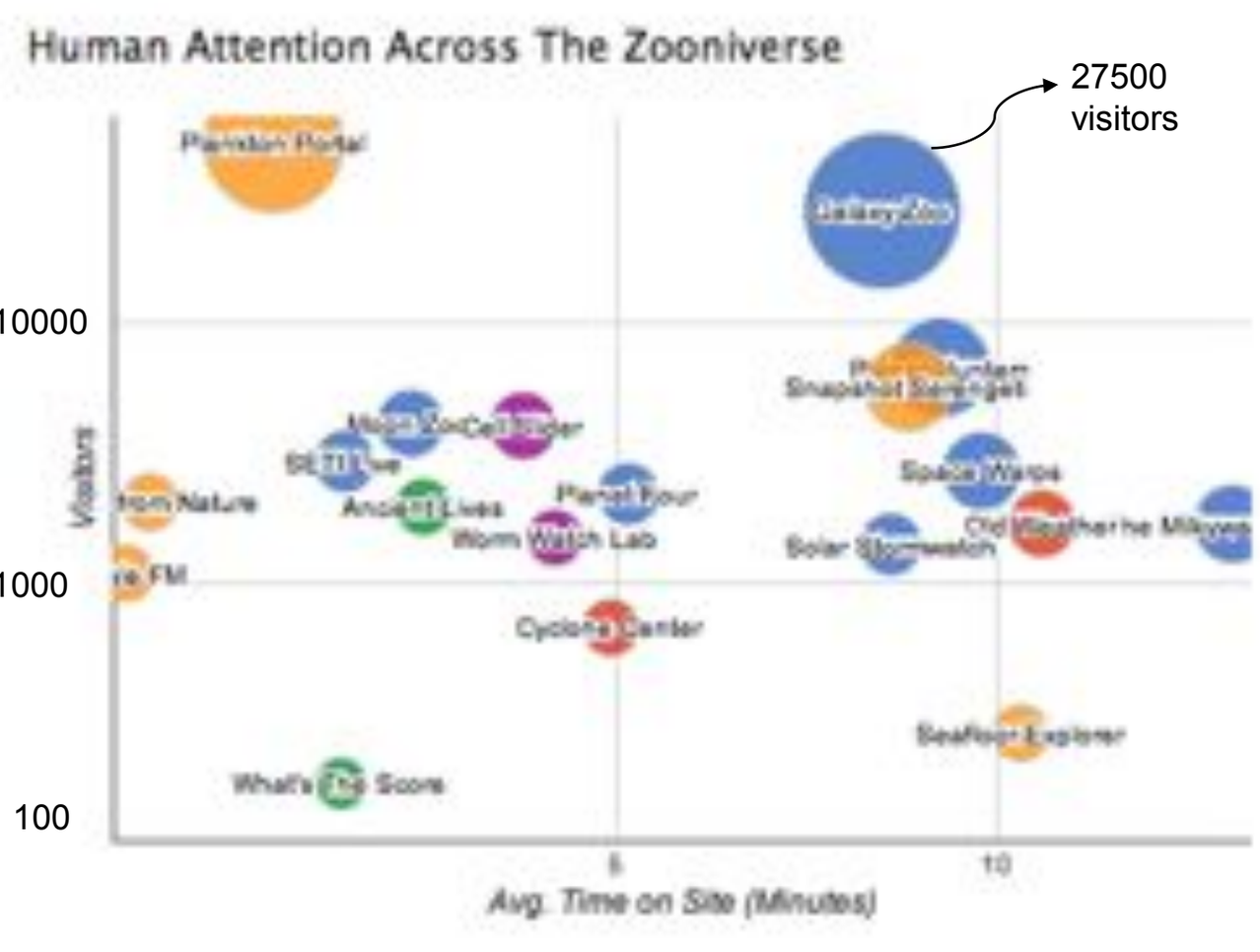
PANOPTES API

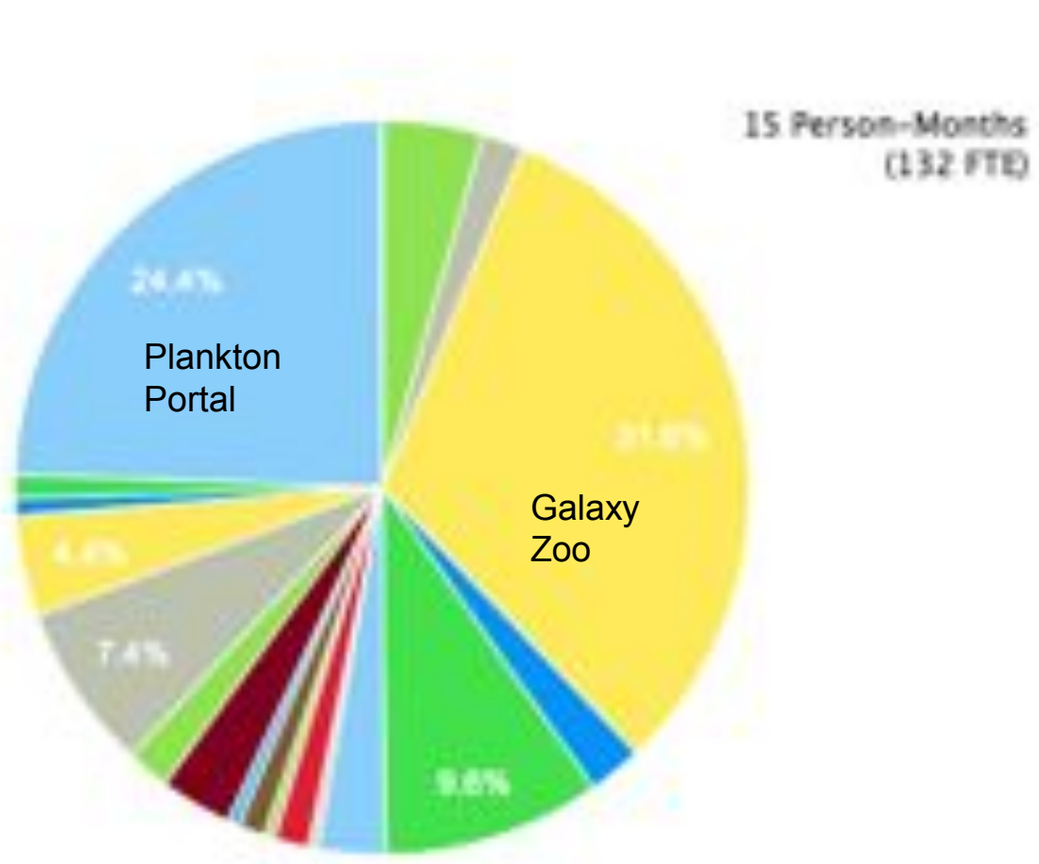
Ruby on Rails + Postgres

Subjects
Amazon S3

Zooniverse Analytics Last Fall







ZOONIVERSE

Changed the field?

- Enabled specific research questions to be tackled at a scale not available before (passing the numbers of big data past a variety/complexity choke-point in the pipeline)
- Enabled association of labels across large corpora of newly transcribed structured (and unstructured) texts.
- Enabled the study of human-machine systems on a real-world platform
- Accelerated discovery of unknown unknowns

Academic Legitimacy

- What difficulties could citizen science projects face within the realm of academia?

- How can project leaders ensure acceptance by academia of results based on data analyzed via the citizen science method?

- How does engagement with or “managing” the volunteers affect the legitimacy of the data sets?

The case for crowdsourcing your research!

The scale of the problem – what do we do with 40 Tbytes a day?

- Largest professional classification is only ~5% of SDSS

  Galaxy Zoo provided 3.3 person years in first 6 months

(Measurable) Accuracy – wisdom of the crowds

- Multiple independent classifications give us an estimate of error

- For a set of noisy data, enough inexpert classifiers will produce more accurate classifications than an expert classifier (where 'enough' depends on how noisy the data, and how inexpert and expert the classifiers).

Machine Learning – creating a partnership between human and machine so each does what they're good at

- Improved accuracy, easier to eliminate false positives.

Education – engaging the public in the process of research

- Citizen science projects directly engage large numbers of people in doing science, not just learning about it.

Serendipity – finding the "unknown unknowns"

-The ability to ask 'huh, what's that?' is both useful and (nearly?) impossible to program

- Development of tools to enhance discovery leads to greater engagement