The Zooniverse at your Fingertips

A crowdsourcing solution for producing research with the data deluge

Dr. Lucy Fortson

University of Minnesota
School of Physics and Astronomy

Graduate and Professional Education Assembly

April 2, 2014
What are these?

And what do they have to do with these?
Or in fact these?

Or these?
Or even these?

Or these?
These are all images from research projects that rely on analysis of complex images and pattern recognition capabilities that are currently beyond computer algorithms!
Which is better at spotting the tiger?
Classifying: another human evolutionary tic

Amazingly, only two basic galaxy shapes – or morphologies

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

Ellipticals
Older star populations so mostly (but not all!) red-ish.
Astronomy as example of data flood

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

2000s - Sloan Digital Sky Survey:
~1000000 galaxies

Need computers to process that much data!
But how can we “computerize” the best algorithm that can process complex data?
Galaxy Zoo launched in 2007 inviting public to classify galaxies

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

Roughly 3.3 continuous person-years!
Data Deluge: Crowdsourcing to the Rescue

- **Today – Data-Intensive Research**

Researchers overwhelmed with data sets from many different sources

- Data captured by instruments
- Data generated by simulations
- Data generated by sensor networks
- Data generated through digital recovery methods

- **eScience is the set of tools and technologies to support data federation and collaboration**

  - For analysis and data mining
  - For data visualization and exploration
  - For scholarly communication and dissemination

---

**eScience is also the technological backbone that enables today’s Citizen Science**
Motivation to Participate in Galaxy Zoo

“Contribute to Research” most common motivation to participate in Galaxy Zoo
Cognitive Surplus

Goggle Boxes

Hours spent...

200 billion hours
a year spent watching TV by US adults

16 years every day!
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).

Serendipity – finding the “unknown unknowns”
- The ability to ask ‘what’s that weird thing’ is both useful and (nearly?) impossible to program

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.

Machine Learning – creating a partnership between human and machine so each does what they’re good at
- Improved accuracy, easier to eliminate false positives.
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…)
- 25 projects and growing: astrophysics, climate science, biology, humanities, nature generating over 500 million classifications
- Over 55 peer-reviewed publications including several discoveries made by members of the public.

Zooniverse has embodied trans-disciplinary work at UMN bringing together researchers from CLA, CBS, CEHD, CSE and MSI.

In addition to enabling excellent research at UMN, faculty are integrating Zooniverse into UMN undergraduate classes providing authentic research experiences in entry-level courses.

www.zooniverse.org

Wildebeest 300,000 volunteers
classifications
100m

τὰ ἀρχαία βιβλία
Ancient Lives Project: UMN developed the pipeline to turn clicks into consensus into transcribed text that can be edited and curated.

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

Identifying most probable letter with each centroid.

Funded by UMN Futures grant!

Annotation and curation tool displaying original image and consensus transcription that can be edited or augmented with metadata.
Snapshot Serengeti: UMN led project to classify millions of camera trap photos taken in the Serengeti.

160,000 volunteers
10 million classifications
UMN grad students developed idea, and are analyzing five seasons’ data.
Galaxy Zoo: help astronomers classify galaxies
Galaxy Zoo: UMN post-doc and grad student – developed consensus algorithm for GZ2 and writing science papers with data on galaxy evolution.

Melanie Galloway et al.

School of Physics and Astronomy, University of Minnesota, 116 Church St. SE, Minneapolis, MN 55455, USA

ABSTRACT

We study the influence of the presence of large scale stellar bars in galaxies which host active galactic nuclei (AGNs). Using data from the Sloan Digital Sky Survey and morphological classifications from the Galaxy Zoo 2 project, we create a volume-limited sample of 19,765 disk galaxies at $0.01 < z < 0.05$ which have been visually

Accepted 2013 August 2. Received 2013 July 31; in original form 2013 June 17
Anatomy of a Zooniverse Project

“Talk” – where volunteers discuss their findings.

“Blogs” – where volunteers hear from the research team.
Anatomy of a Zooniverse Project

Data analysis tools - enabling volunteers to explore the data:

tools.zooniverse.org
Zooniverse in the Classroom

Using Galaxy Zoo in high school or undergrad laboratory exercises.

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

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

- Large volume of complex images – big data challenge
  - Crowdsourcing - Human processing in combination with machines – is one solution

- Zooniverse: largest crowdsourcing research platform
  - Research across multiple domains and data formats
    - UMN: astrophysics, ecology, digital humanities research
    - Any others out there from UMN ?? Please let me know!
  - At the same time we do research on the science of crowdsourcing (how to optimize human-machine partnership -> optimizing human attention and cycle of machine learning)
    - UMN: computer science research
  - Provides education opportunity for millions of members of the public as well as in formal education
    - UMN: education research and implementation of Zooniverse in Introductory Biology and Astronomy classes.

- Future: the Zooniverse is developing a “Do it Yourself” API that will allow anyone with a good chunk of data to put forth a Zooniverse project as easy as putting out a blog. Stay tuned!