Crowdsourcing and the VO

Matthew J. Graham (Caltech, NVO)
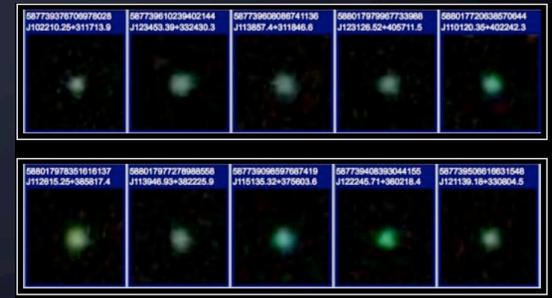et
Roy Williams, Andrew Drake, George Djorgovski
Ashish Mahabal, Ciro Donalek

Humans as CPUs

- Unique juncture in history of science:
  - technological capability exists to network large numbers of people
  - data volumes and complexity are still desktop manageable
  - class of problems that are resistant to present machine learning solutions
- Crowdsourcing/human computation/citizen science projects exploit efforts of volunteers to attack particular areas, e.g. image analysis
- Axes:
  - Sweat shop vs. GWAP
  - Idiot vs. savant

Tonight a galaxy, tomorrow the Zoo

- Initial questions:
  - Are galaxies elliptical or spiral?
  - If spiral, rotating clockwise or anticlockwise?
- 34617406 clicks done by 82931 users
- Main result:
  - Spiral galaxies which share a neighbourhood (a region defined as 65 million light years across) are likely to rotate in the same direction – but only if they formed the vast majority of their stars more than 10 billion years ago.
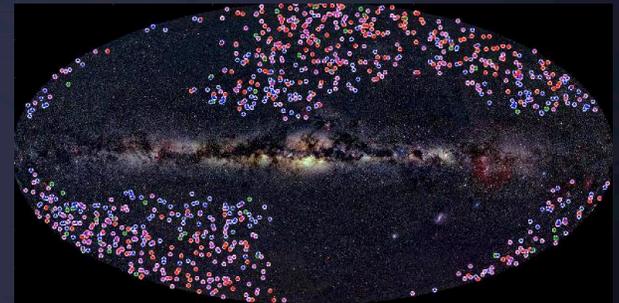- Other results:
  - Hanny's Voorweerp
  - Green Peas

Galaxy Zoo 2

Things that go BANG! in the night

- Catalina Real–Time Transient Survey (http://crts.caltech.edu)
  - Repeatedly surveys ~26000 deg^2
  - 3 telescopes: MLS (1.5m), CSS (0.7m), SSS (0.5m)
  - 1067 new discoveries to date
  - Only completely public transient survey
- SkyAlert (http://www.skyalert.org)
  - enables users to perform complex queries about discoveries in order to receive personally tailored and filtered event streams.
- The VO is useful for:
  - data discovery
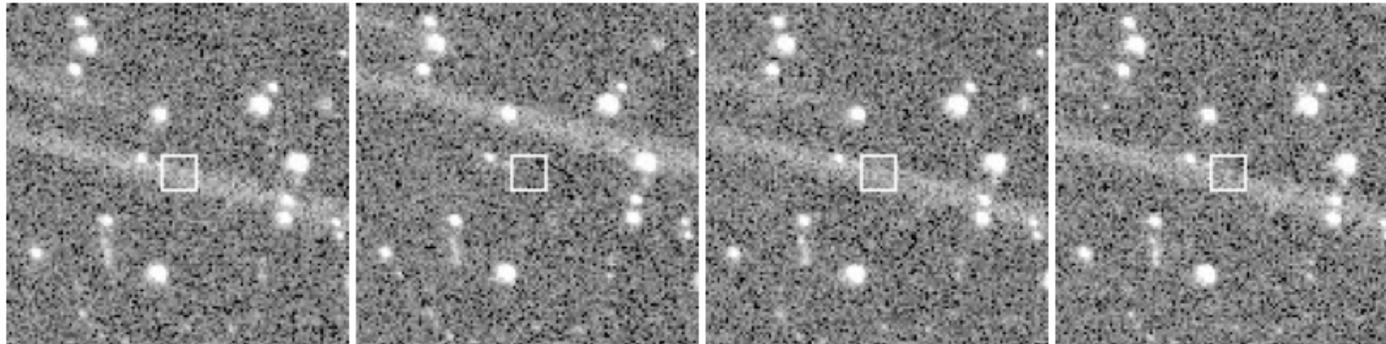 - semantics
  - data mining

Citizen science with CRTS

AstroCollation – I

- Next generation collaborative science venture
- Data mining algorithms applied to transient event data to produce conceptual models describing them
- Models presented to citizen scientists for value judgements, deciding which of a set of models provides the best description

- Citizen scientists can also provide contextual information to aid the classification process
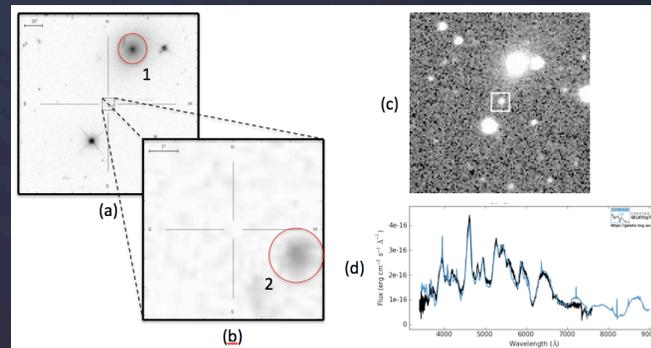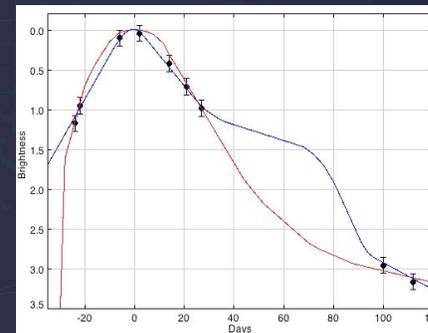
AstroCollation – II

- Decisions and information factored back into the system and consolidated to produce a *consensus* description of an event that can always be retrieved (and reused)
- Produce better (ideal) training sets
- Built upon semantic technologies, CRTS and SkyAlert
- Issues to address:
  - How to formally represent uncertainty in data and *description* in a machine-processible fashion
  - Optimal method to achieve consensus opinion