Detecting the Unexpected

Discovery in the Era of Astronomically Big Data

Insights from Space Telescope Science Institute's first *Big Data* conference

Josh Peek, Chair

Sarah Kendrew, Co-Chair

SOC: Erik Tollerud, Molly Peeples, Támas Budavari, Rick White, Mario Jurić, Tod Lauer, Mike Fall, Armin Rest, Alyssa Goodman, Coryn Bailer-Jones

Big data according to STScI

"Big Data" is a term borrowed from industry, which collects data on large scale from users and sensors, and is typically used to serve advertisements with a higher return on investment than would otherwise be possible.

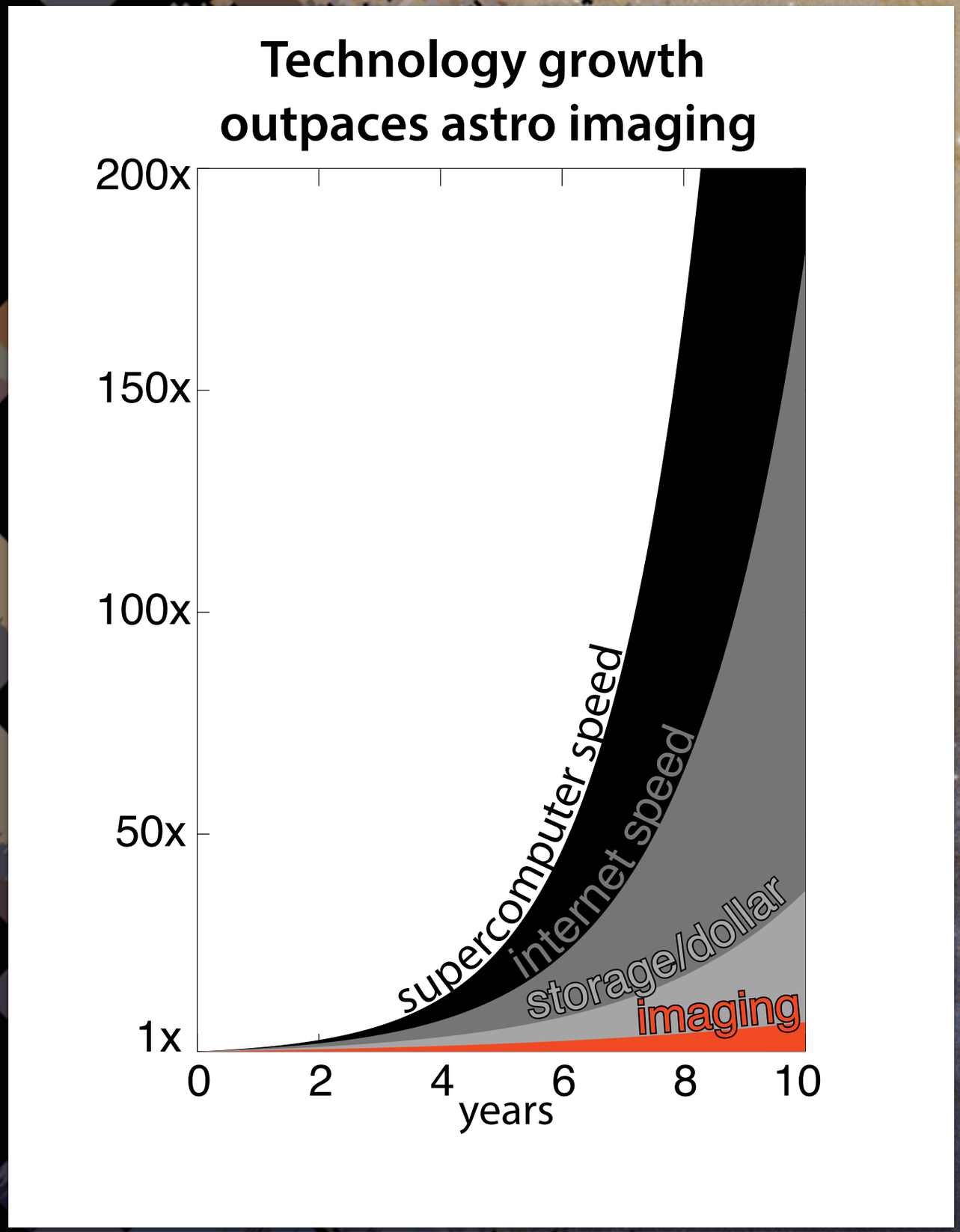
Our definition:

- Data whose raw form is so large we must qualitatively change the way in which we reduce, store, and access it.

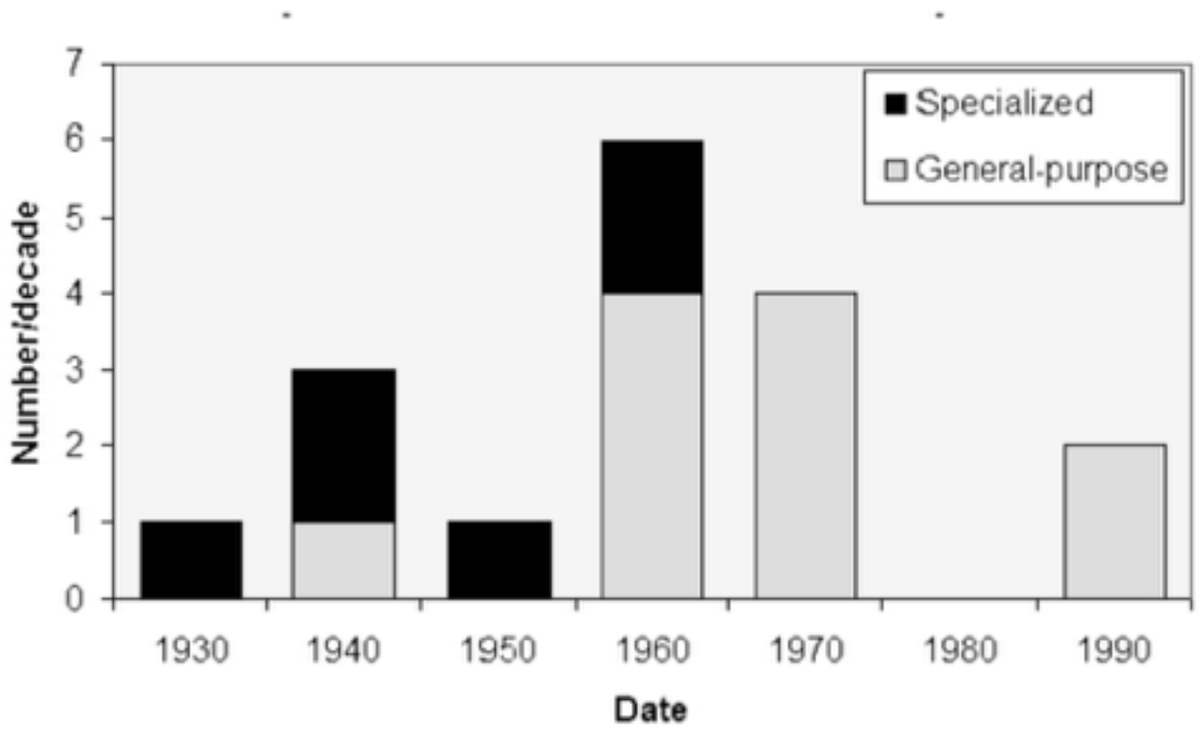
- Data whose reduced form is so large that we must qualitatively change the way in which we interact with and explore it.

- Data whose structure is so complex that our current tools cannot efficiently extract the scientific information we seek.

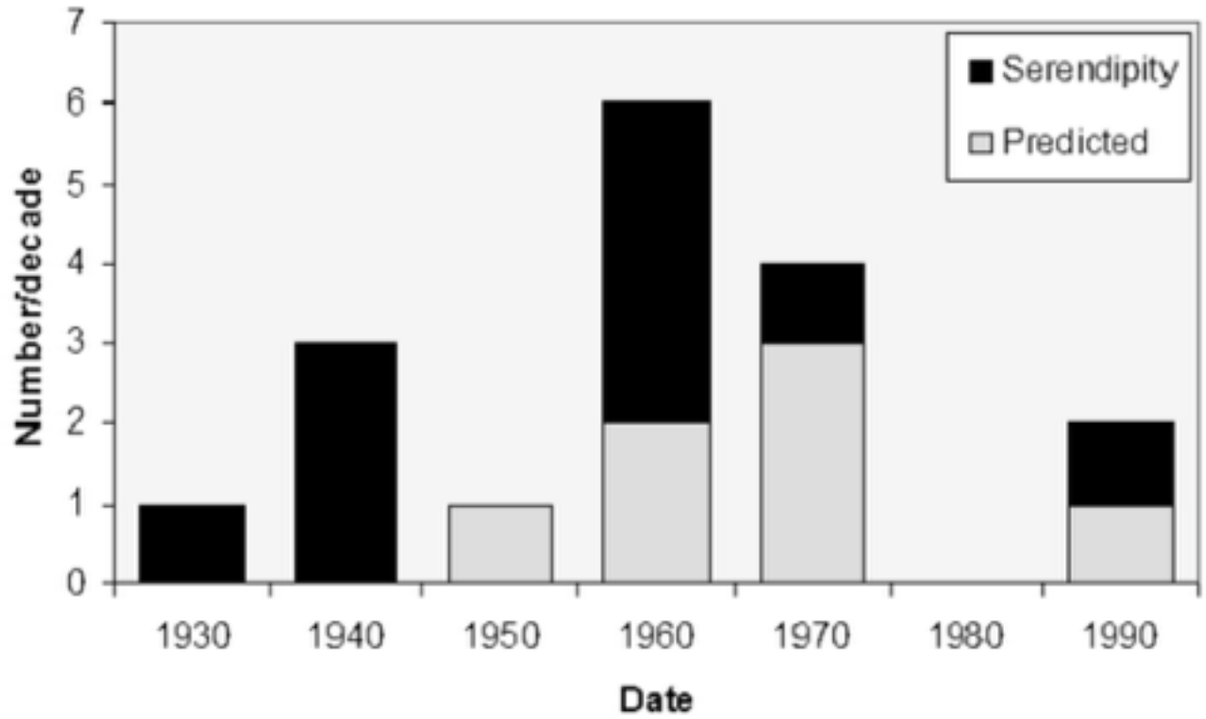
Our conference was not concerned with *data volume* so much as *the scientific method*



Serendipitous discovery is a foundation of astronomy



(a) Type of instrument,



(b) Predicted v Serendipity

Figure 4: Key Discoveries in Radio Astronomy from [12]

Ekers 09

Is that foundation secure in the era of Big Data?

What do we need to do technically and culturally to continue to detect the unexpected?

Detecting the Unexpected

Discovery in the Era of Astronomically Big Data

Themes

Machine Learning: *Have a computer look at it*

Citizen Science: *Have a lot of people look at it*

Visualization: *Look at it better*

Machine Learning (non-parametric classification of high-D data)

- Many outlier detection methods were discussed: DEMUD, RF methods tSNE, various eigenbasis methods
- Many finds including galaxies and stars with unexplained spectra
- Often outliers are “close” in these spaces; extreme outliers are known/artifacts
- Classifiers are becoming *portable* from theory to data, and from data to data
- Models (theory) can quantify the expected look for unexpected in *residuals*

Deep Learning
(data classification *directly from pixel data)*

- Deep Learning / Artificial Neural Nets are seeing a fast expansion in astronomy, especially Convolutional Neural Nets (CNNs)
- CNNs are mostly used for classification but can also be used to hone intuition
- Topics include strong lenses, galaxy morphology, dark matter, dark energy
- Deep Learning requires astronomers to think about their *models* as big data, which is a wholly new skill set

Citizen Science
(contributions from laypeople over web)

- Zooniverse is the dominant platform for citizen science, with many big unexpected discoveries to date (Boyajian's star, Hanny's voorwerp, green peas, etc.)
- CS can be used in synergy with machines in the big data era: different weaknesses
- Building CS tools is extremely easy — a presenter built one *during* her 15 minute talk
- CS requires a *new skill set* for astronomers to use it effectively: community management

Visualization

- Multidimensional visualization is key for understanding and exploring big data
- Hackable interfaces allow users to explore data easily but also deeply
- Fast visualization of very large data sets is a *real* engineering challenge
- Federation of big data systems will be needed if we expect to visualize them with a range of tools

Philosophy & Culture

- If we want junior people doing high risk data exploration we need better ways to protect them in case of failure
- Our culture needs to think more clearly about the term “fishing expedition”
- We need to reaffirm our commitment to software as “soft instrumentation”
- Detecting the Unexpected requires teamwork between domain specialists and methods experts, which requires breaking the lone genius myth

Detecting the Unexpected: Data and Methods Bazaar

http://bit.ly/dtu17-dmb

Session 1 (4:55-5:15 / 5:15-5:35):

- (1) Graham: "Letting data describe itself"
- (2) Snyder: "Observing Virtual Universes: Synthetic HST & JWST Images from the Illustris Project"
- (3) Teuben: "ALMA Data Mining Toolkit (or Astro Data Mining Toolkit)"
- (4) Price-Whelan: "A custom Monte Carlo sampler for sparse or noisy radial velocity data: how to sample from highly multi-modal distributions over Keplerian orbital parameters"
- (CafeCon) Bot/Bosh: "Building Hierarchical Progressive Surveys (HIPS) from a set of images and using Multi Order Coverage maps (MOC) to explore large catalogs"

Session 2 (5:35-5:55 / 5:55-6:15):

- (1) Gordon: "The Unexpected BEAST?: PHAT catalogs of stellar and dust extinction properties of millions of stars in M31"
- (2) Sanderson: "Galaxia on FIRE: Mock surveys of high-resolution cosmological, hydrodynamic Milky Way simulations"
- (3) Loredo: "CUDAHM: A C++ framework for GPU-accelerated hierarchical Bayesian inference with simple graphical models"
- (CafeCon) Weaver/Olsen: "Exploring Large Datasets with the NOAO Data Lab"

Hack Day Projects

- Exploiting variation in survey filters to find extreme emission lines
- Developing new outlier detection methods based on topology/homology
- Deploying a citizen science tool on very complex ISM data
- Convolutional Neural Nets for merger classification and spectral typing
- Developing a database of outlier detection methods
- Developing new tech for serving histograms of big data over networks
- Extreme Deconvolution on stellar density catalogs
- Using DEMUD to determine how computers and human find outlier supernovae differently