Astronomy Data Landscape and Observable Parameter Spaces

S. George Djorgovski, Caltech

KISS Short Course

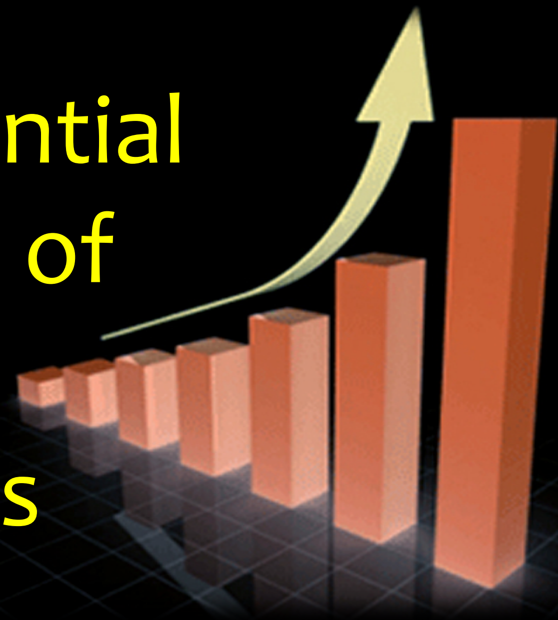
Data-Driven Approaches to Searches for Technosignatures

May 2019

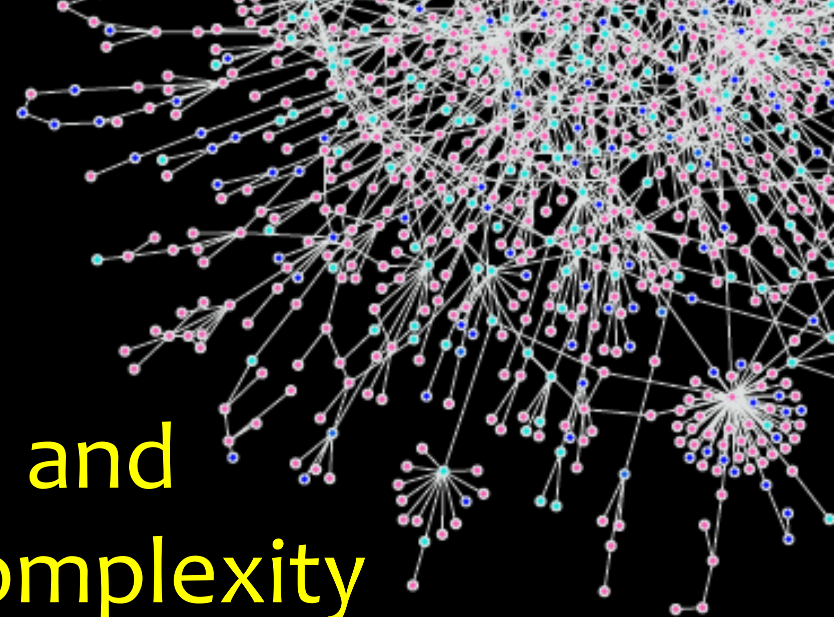
Caltech

Keck INSTITUTE FOR SPACE STUDIES

Exponential Growth of Data Volumes



... and Complexity



on Moore's law time scales

Understanding of complex phenomena requires complex data!

From data poverty to data glut
From data sets to data streams
From static to dynamic, evolving data
From anytime to real-time analysis and discovery
From centralized to distributed resources
From ownership of data to ownership of expertise

What is Fundamentally New Here?

- The ***information volumes and rates*** grow exponentially

  ⇨ ***Most data will never be seen by humans***

- A great increase in the data ***information content***

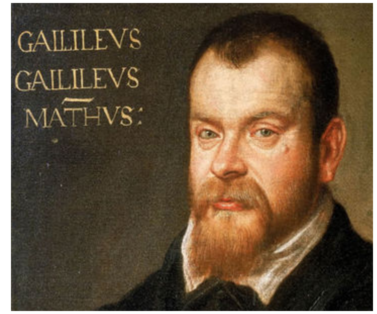
  ⇨ ***Data driven vs. hypothesis driven science***

- A great increase in the ***information complexity***

  ⇨ ***There are patterns in the data that cannot be comprehended by humans directly***

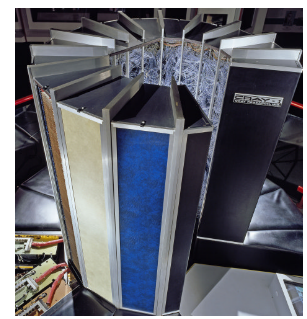
The Evolving Paths to Knowledge

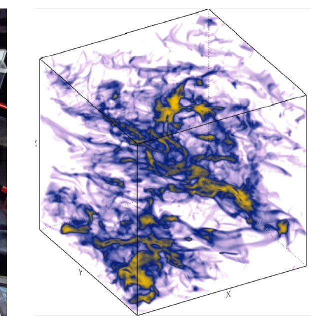
- The First Paradigm:
  Experiment/Measurement



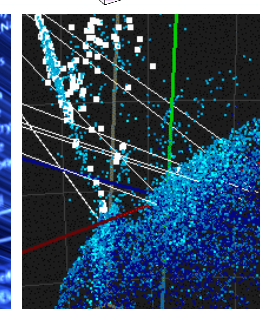
- The Second Paradigm:
  Analytical Theory

- The Third Paradigm:
  Numerical Simulations

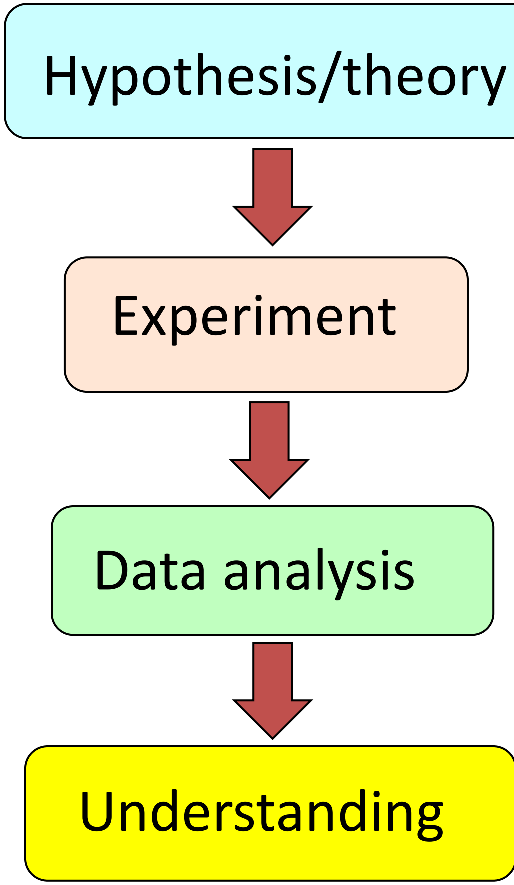




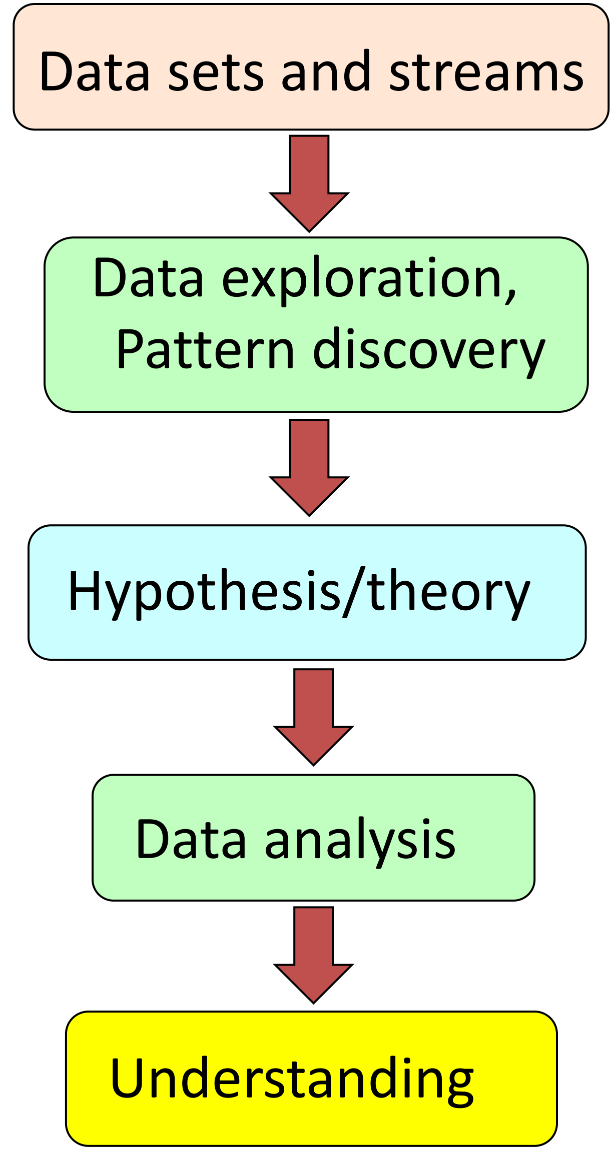
- The Fourth Paradigm:
  Data-Driven Science



Hypothesis-driven science



Data-driven science



The two approaches are complementary

The Evolving Data-Rich Astronomy

An example of a "Big Data" science driven by the advances in computing/information technology

1980 | 1990 | 2000 | 2010 | 2020

MB | GB | TB | PB | EB

CCDs

Image Proc.

Surveys

Pipelines

Databases

VO

Machine Learning

AstroInfo

LSST, SKA…

AI

Key challenges: data heterogeneity and complexity

How Much Data* is There in Astronomy?

*** Archived, curated, accessible**

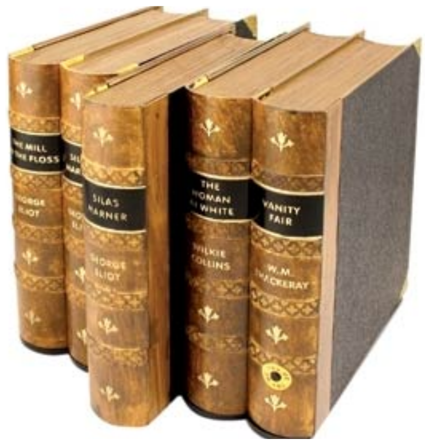
- My best guesstimate (early/mid 2019): $\boxed{\sim 200 \text{ PB} \times 2^{\pm 1}}$
  - Estimated data rate > 100 TB/day
- Most data come from sky surveys
- Both data volumes and data rates grow exponentially, with a ***doubling time ~ 1.5 years***
- Even more important is the growth of ***data complexity*** and ***data quality*** (information content)
- For comparison:

  Human Genome < 1 GB

  Human Memory < 1 GB (?)

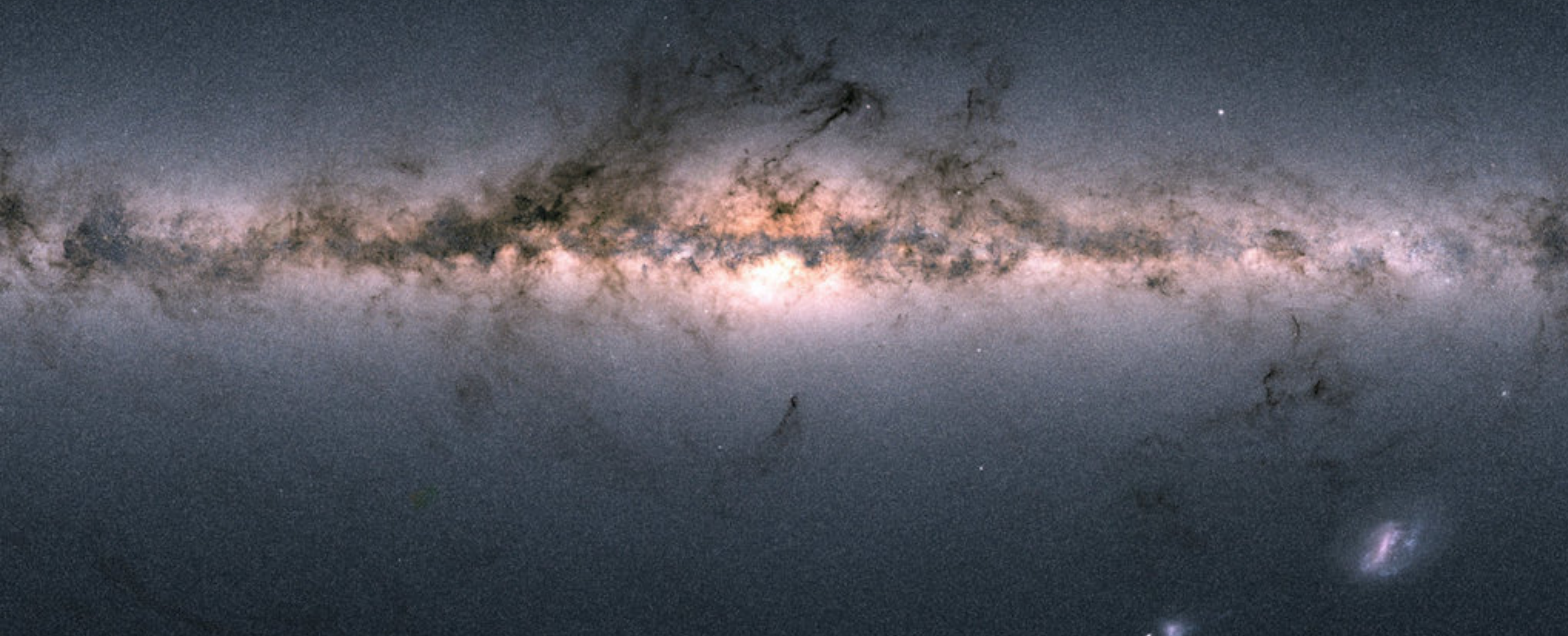
  1 TB ~ 2 million books

  Human Bandwidth ~ 1 TB / year (±)



There Are *Lots* Of Stars In The Sky…

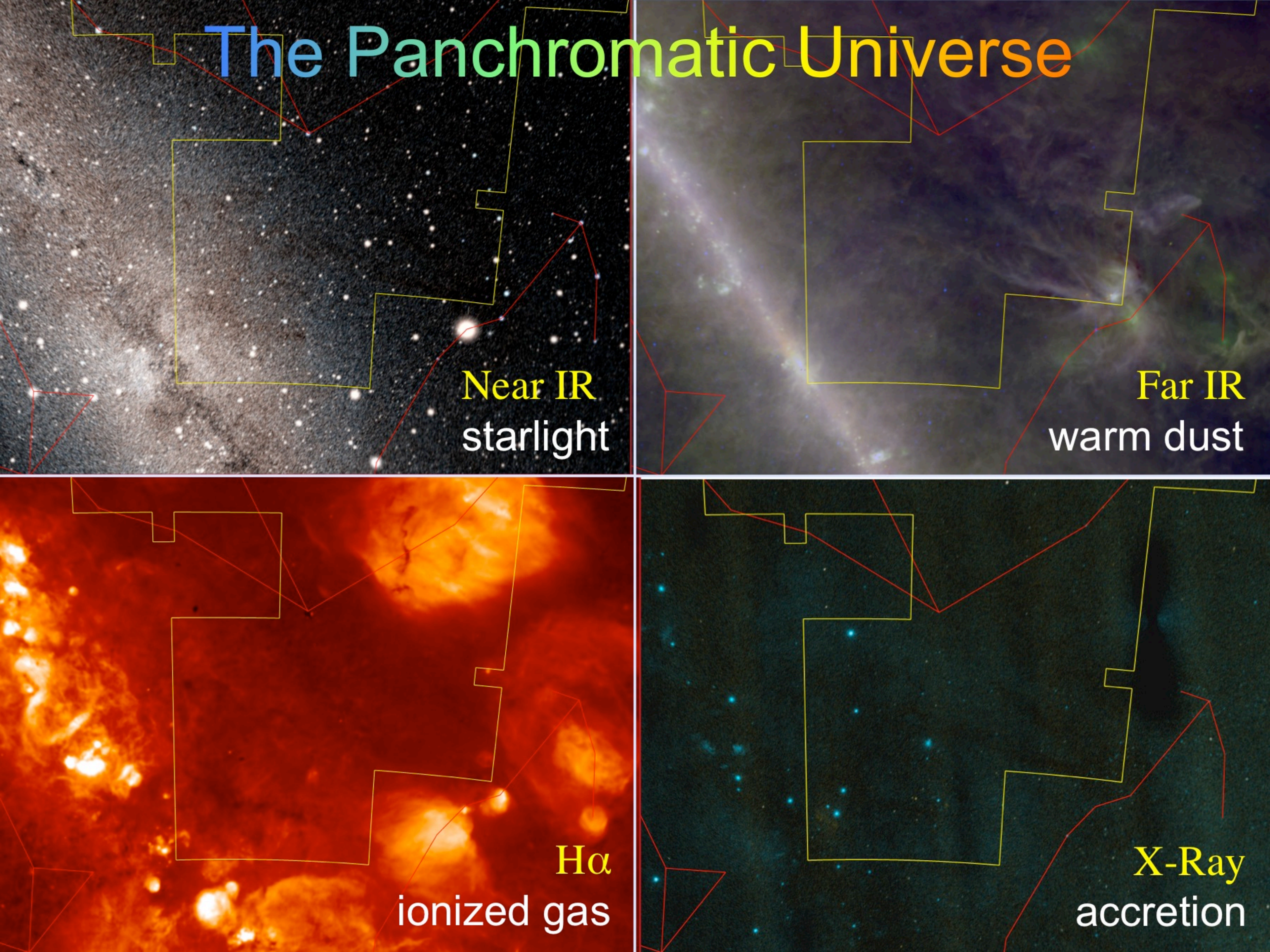
Modern sky surveys obtain ~ 10^{15} – 10^{16} bytes of images, catalog ~ 10^{9} objects (stars, galaxies, etc.), and measure ~ 10^{2} – 10^{3} numbers for each



… and then do it again, and again, …

Gaia DR2 Catalog Image

The Panchromatic Universe



Sky Surveys: Data Volumes

Sky Survey Projects	Data Volume
DPOSS (The Palomar Digital Sky Survey)	3 TB
2MASS (The Two Micron All-Sky Survey)	10 TB
GBT (Green Bank Telescope)	20 PB
GALEX (The Galaxy Evolution Explorer)	30 TB
SDSS (The Sloan Digital Sky Survey)	170 TB (DR15) ~~40 TB~~
SkyMapper Southern Sky Survey	500 TB
PanSTARRS (The Panoramic Survey Telescope and Rapid Response System)	~ 40 PB expected
LSST (The Large Synoptic Survey Telescope)	~ 200 PB expected
SKA (The Square Kilometer Array)	~ 4.6 EB expected

1990s

2000s

2010s

ZTF: ~ 1 PB/yr

2020s

(from Zhang 2015)

Some "Local" Producers:

- CRTS (all surveys, per A. Drake):
  - ~ 100 TB total to date
  - Current data rate ~ 25 TB/yr
- ZTF (3 year survey, per F. Masci):
  - ~ 3.2 PB total archived
  - Current data rate ~ 1 TB/night (images), real-time data products ~ 4 TB/night
- OVRO (per G. Hallinan):
  - LWA: Raw data rate ~ 12 PB/day, archived ~ 50 TB/day ~ 18 PB/yr
    - MWA: ~ Raw data rate ~ 0.65 PB/day, archived ~ 27 PB/yr
  - DSA: Raw data rate ~ 7 PB/day, much less archived

Some space missions:

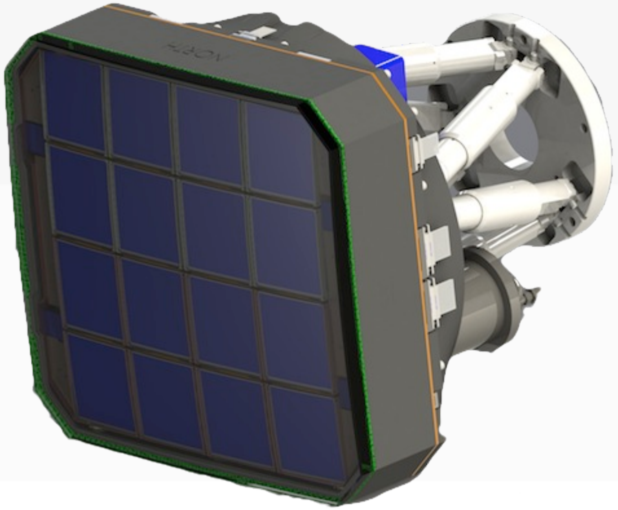
✧ Kepler ~ 20 TB

✧ GALEX ~ 30 TB

✧ Gaia, 5-yr mission: ~ 200 TB

Zwicky Transient Facility (2017-)

- New camera on Palomar Oschin 48” with 47 deg^2 field of view
- 3750 deg^2 / hr to 20.5-21 mag (1.2 TB / night)
- Full northern sky (~12,000 deg^2) every three nights
- Galactic Plane every night
- Over 3 years: 3 PB, 750 billion detections, ~1000 detections / src
- First megaevent survey: 10^6 alerts per night (Apr 2018)

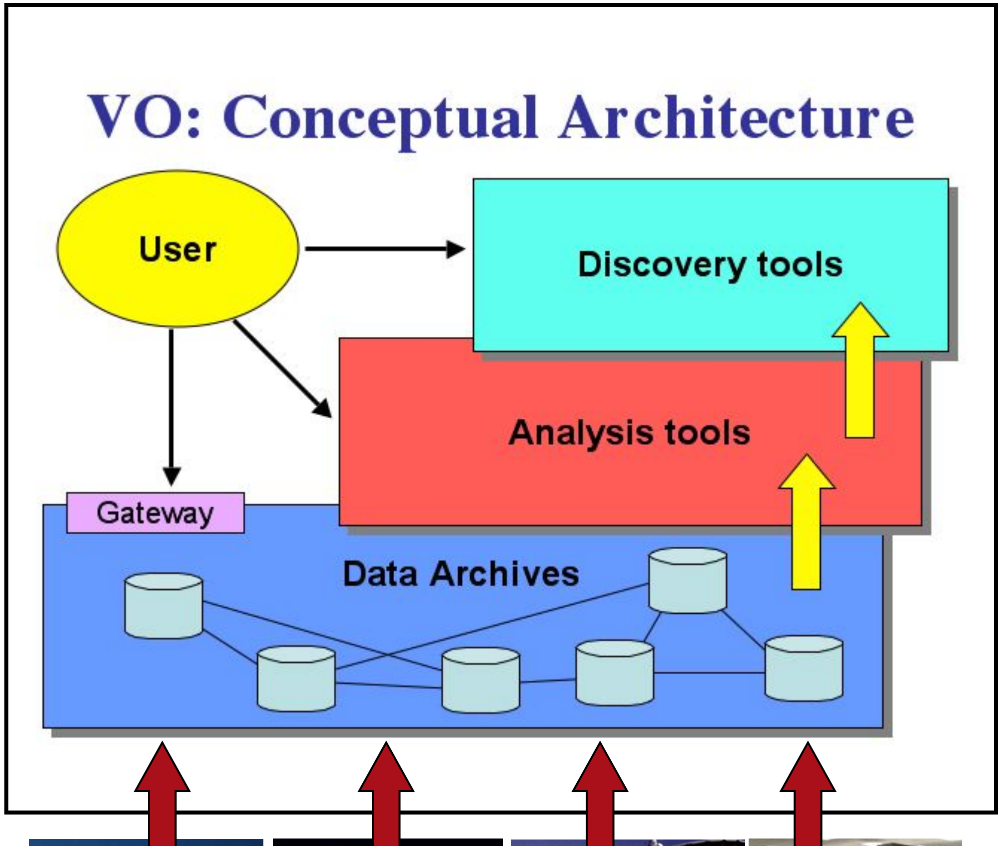


ZTF = 0.1 LSST

	ZTF	LSST
No. of sources	1 billion	37 billion
No. of detections	1 trillion	37 trillion
Annual visits per source	1000 (2+1 filters)	100 (6 filters)
No. of pixels	600 million (1320 cm^2 CCDs)	3.2 billion (3200 cm^2 CCDs)
Field of view	47 deg^2	9 deg^2
Hourly survey rate	3750 deg^2	1000 deg^2
Nightly alert rate	1 million	10 million
Nightly data rate	1.4 TB	15 TB

The Virtual Observatory Concept

- Envisioned as a complete, dynamical, distributed, open ***research environment for the new astronomy with massive and complex data sets***
  - Provide and federate content (data, metadata) services, standards, and analysis/compute services
  - Develop and provide data exploration and discovery tools (…)
  - Today it is the global data grid of astronomy
  - A successful example of a science Cyber-Infrastructure



Djorgovski

IVOA: The Virtual Observatory Reified

- Formed in 2002 to facilitate the international collaborative effort needed to enable integrated access to astronomical archives
- 21 international members
- Working Groups and Interest Groups overseen by Technical Coordination Group reporting to Executive Committee:
  - Applications
  - Data Access Layer
  - Data Models
  - Grid and Web Services
  - Registry
  - Semantics
  - Data Curation and Preservation
  - Knowledge Discovery in Databases
  - Education
  - Operations
  - Solar System
  - Theory
  - Time Domain
- Committee for Science Priorities
- Engage with big projects

IVOA.net

Resources at http://ivoa.net

INTERNATIONAL VIRTUAL OBSERVATORY ALLIANCE

Home | Astronomers | Deployers | Members | About

VO Applications for Astronomers

In this section, scientists can find available VO-compatible applications for their immediate use to do science. The level of maturity of the applications depends on a high degree on the level of maturity of the corresponding IVOA protocols and standards.. As a consequence of the flexibility of the standards, several of the applications might overlap in functionality. **The IVOA does not manage or guarantee these services/tools.**

Applications (in alphabetical order)

Aladin
AppLauncher
CASSIS
CDS Xmatch Service
Data Discovery Tool
Filter Profile Service
Iris
Montage
Octet
SkyView
Specview
SPLAT
TAPHandle
. . .

Functionality

Search for Images:
Aladin, Datascope, SkyView, VODesktop, Data Discovery Tool

Search for Spectra:
Aladin, CASSIS, Datascope, SPLAT, Specview, VOServices, VOSpec, Data Discovery Tool

Search for Catalogues:
Aladin, Datascope, TOPCAT, VODesktop, Data Discovery Tool

Search for Time Series
. . .

VO-compliant Tools & Services

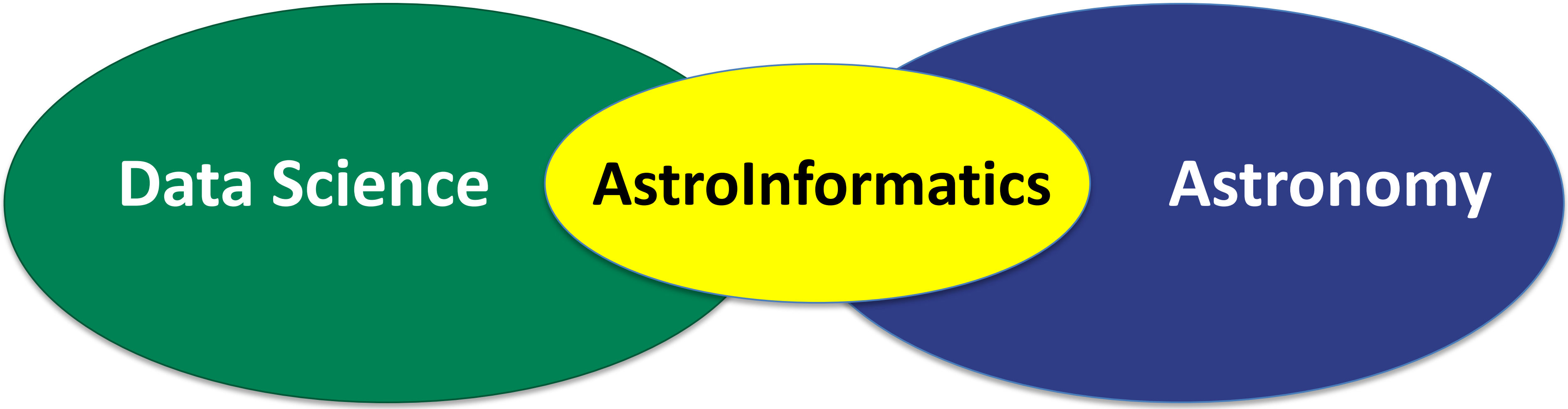
DS9: Image visualiasation
GOSSIP: SED fitting
VirGO: Search for Images and Spectra
IRAF: Image Reduction & Analysis
World Wide Telescope
Gaia - Graphical Astronomy and Image Analysis
SIMBAD
TESELA
VizieR
. . .

A compilation of tools and services

IVOA is now mainly a standards coordination body

AstroInformatics

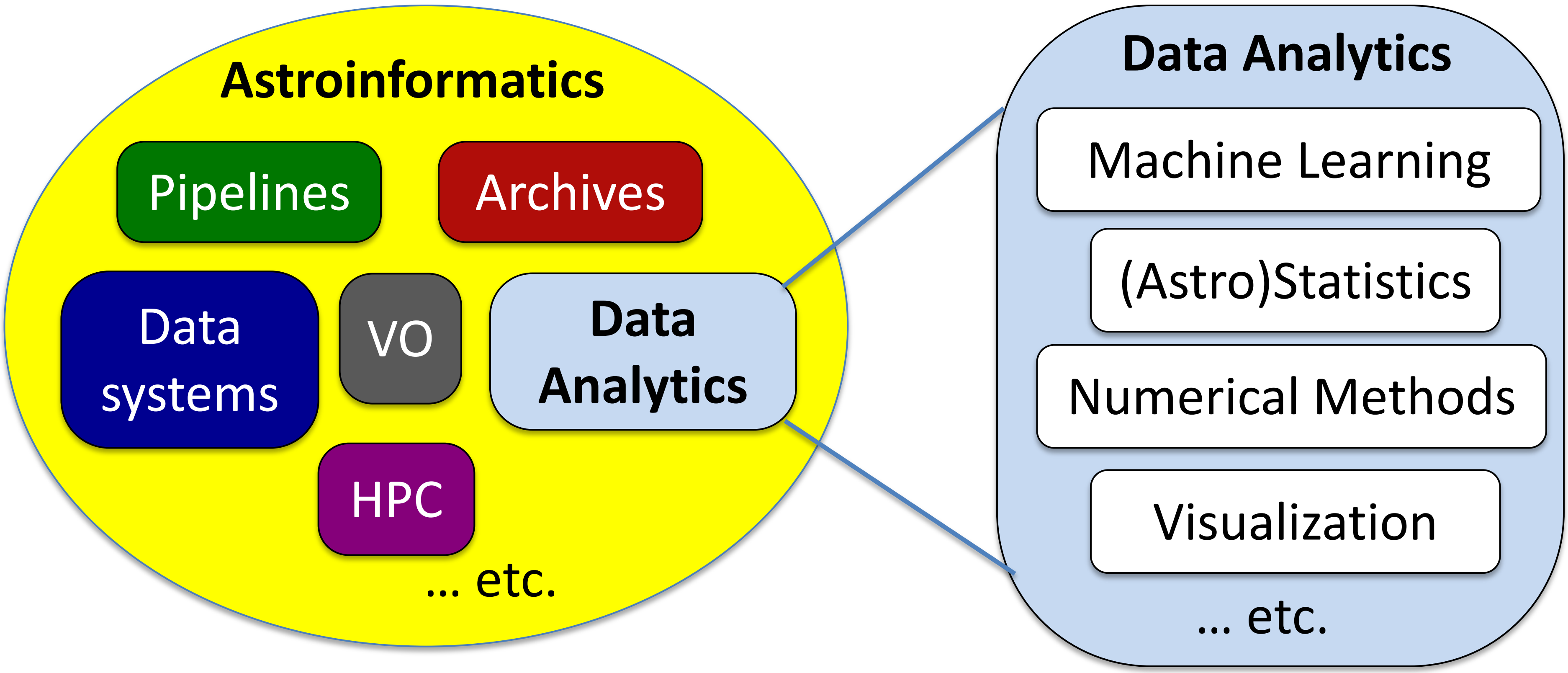
is essentially astronomical applications of Data Science



- While VO became a global data grid of astronomy, astroinformatics focuses of the **knowledge discovery tools**
- It includes a growing community of scientists, both as contributors and as users
- Like other X-Informatics (X = bio, geo, …) it is a bridge between astronomy and data science, and for the methodology sharing with other fields.

AstroInformatics

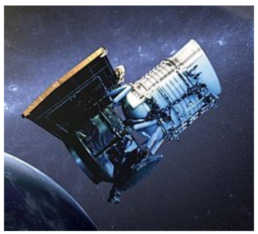
It contains all of the components of Data Science, in their astronomical applications, and their interconnections



The 10th international conference,
astroinformatics2019.org, at Caltech, June 24-27, 2019

Survey-Based Astronomy

Survey Telescopes



Data Reduction Pipeline

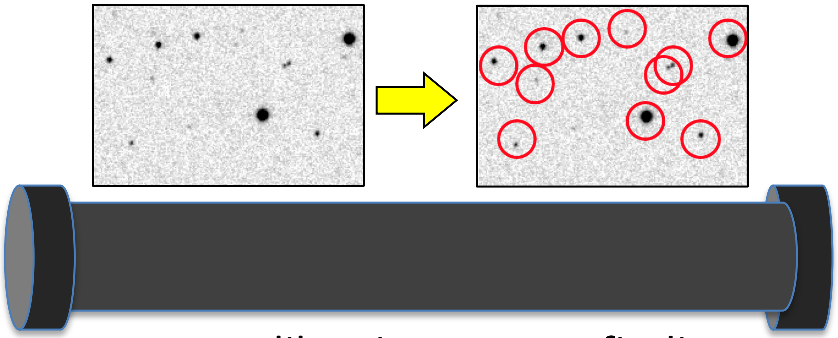


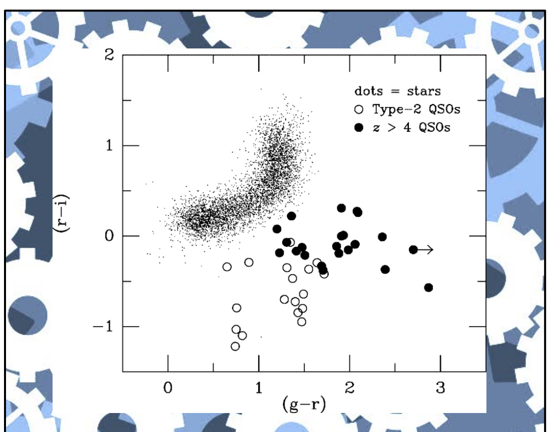
Image calibration, source finding and parametrization

Archive Database

CSS171012:095944+641448	149.93362	64.24678	20171012.47	19.36	yes	2018-06-08	16377
CSS171012:075222+652857	118.09216	65.48244	20171012.41	18.49	yes	2018-05-17	19128
CSS171012:072652+630200	111.71466	63.03323	20171012.39	18.77	no	2018-05-09	14639
CSS171012:084704+593309	131.76754	59.55244	20171012.43	19.37	yes	2018-05-25	15507
CSS171012:172358+530024	260.99323	53.00658	20171012.12	18.73	yes	2018-06-19	11012
CSS171012:084412+531251	131.04870	53.21413	20171012.43	17.29	yes	2018-05-25	11414
CSS171012:164452+443946	251.21538	44.66291	20171012.10	17.80	yes	2018-06-19	12124
CSS171012:235710+401134	359.29246	40.19289	20171012.23	18.44	no	2018-06-19	12739
CSS171012:235703+395916	359.26183	39.98778	20171012.23	14.90	no	2018-06-19	12034
CSS171012:010541+431030	16.42287	43.17505	20171012.28	19.03	yes	2018-06-19	23148
CSS171012:003812+401052	9.55193	40.18108	20171012.25	19.76	yes	2018-06-19	12674
CSS171012:001439+425157	3.66175	42.86579	20171012.26	18.25	no	2018-06-19	21801

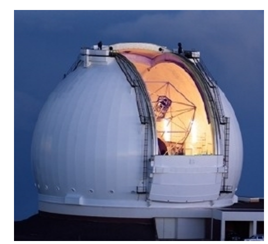
Source catalogs define feature spaces

Data Analysis, Target Selection



Modeling, Machine Learning...

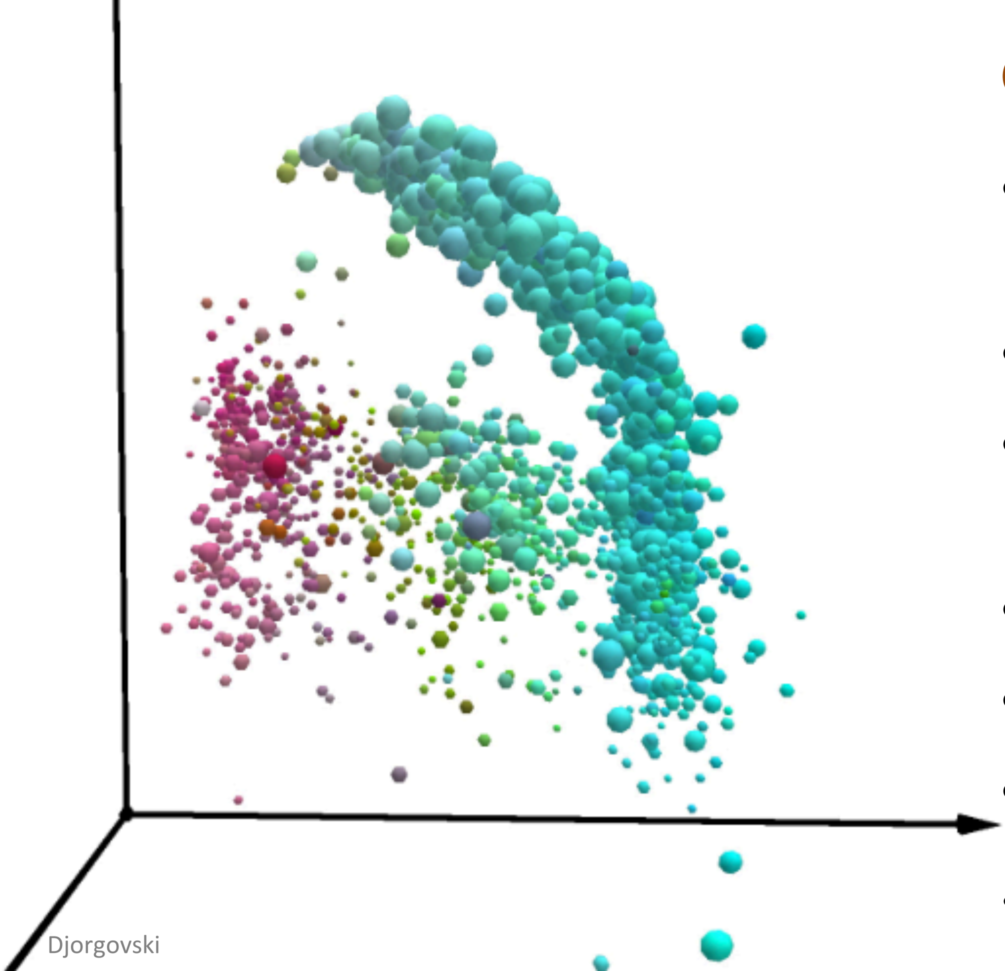
Follow-up Telescopes



Exploration of Parameter Spaces is a Central Problem of Data Science

Clustering, classification, correlation and outlier searches, …

Machine Learning Is the Key Methodology



Djorgovski

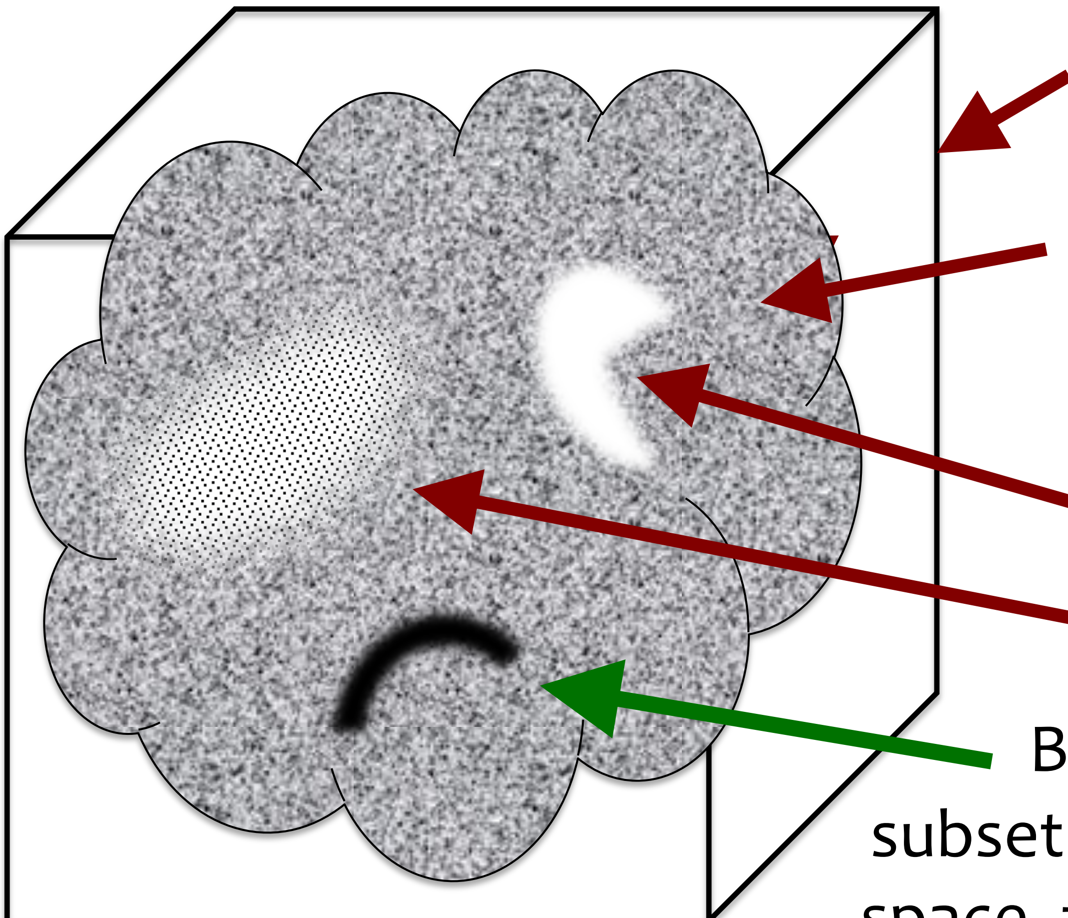
Challenges:

- Algorithm and data model choices
- Data incompleteness
- Feature selection and dimensionality reduction
- Uncertainty estimation
- Scalability
- Visualization

… etc.

} Especially with the data dimensionality

Pattern or structure (Correlations, Clustering, Outliers, etc.) Discovery in High-Dimensional Parameter Spaces



D >> 3 parameter/feature space hypercube

High-D data cloud: mostly noise, with an arbitrary PDF distribution

Missing data

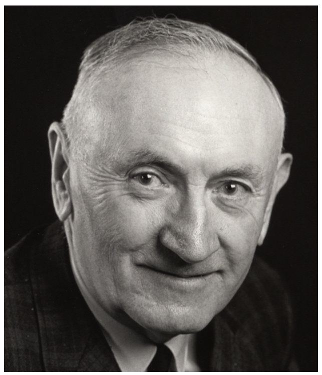
Data heterogeneity

But in some corner of some subset of dimensions of this data space, there is ***something ≠ noise***, i.e., a statistically significant structure with an unknown form

Mapping the Entropy of Large Data Spaces?

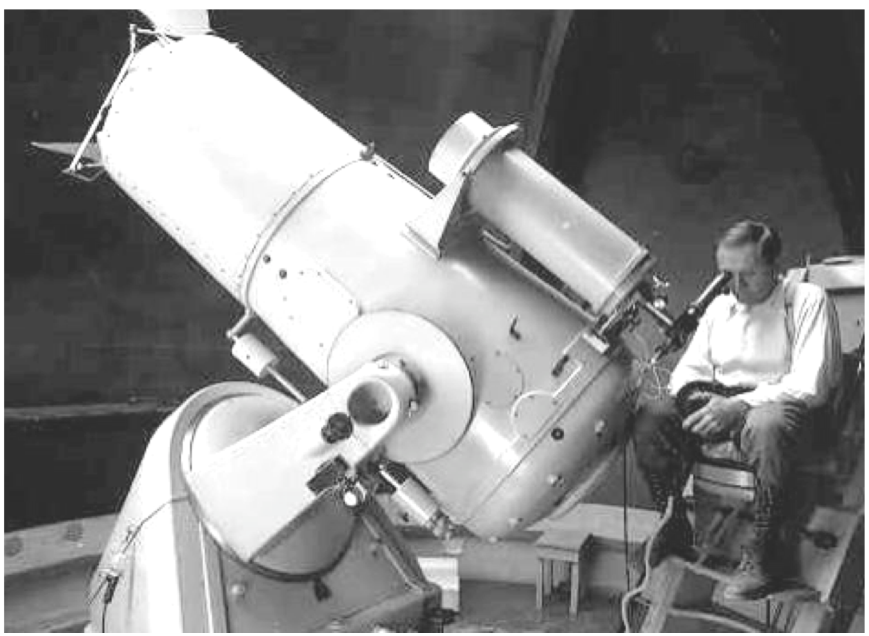
Djorgovski

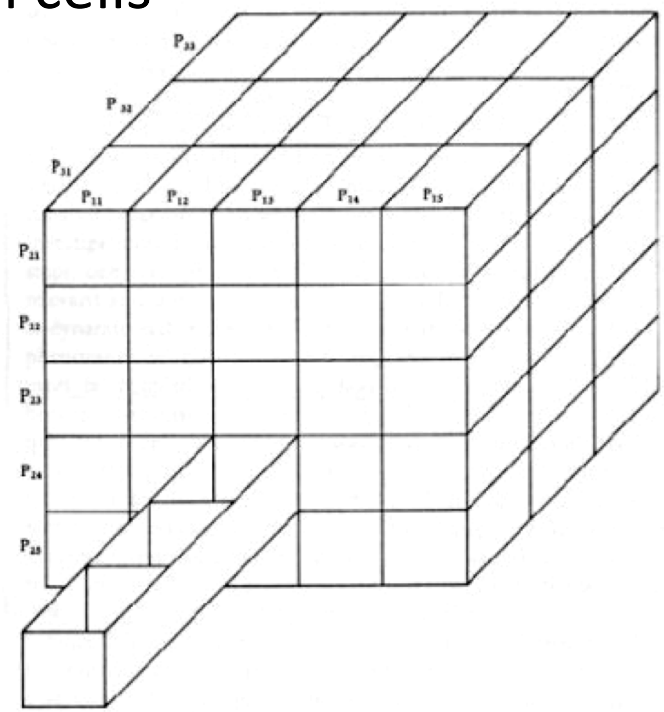
From “Morphological Box” to the Observable Parameter Spaces



Fritz Zwicky

Zwicky’s concept: explore all possible combinations of the relevant parameters in a given problem; these correspond to the individual cells in a “Morphological Box”



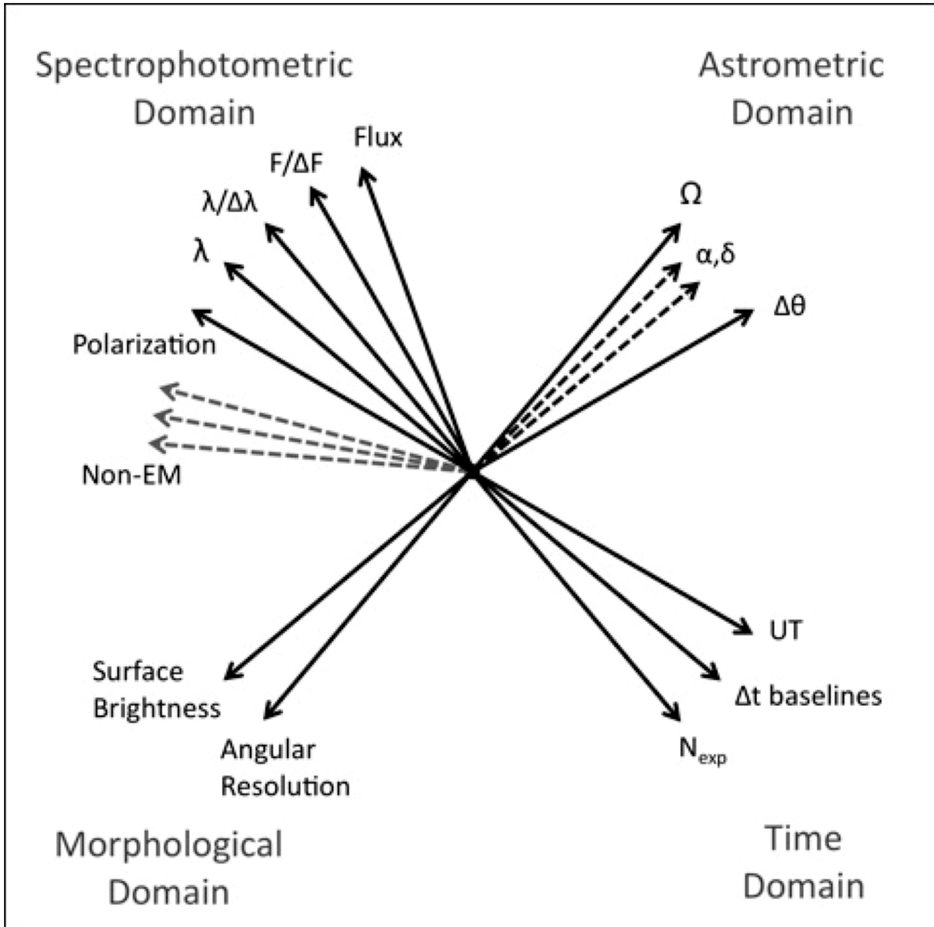


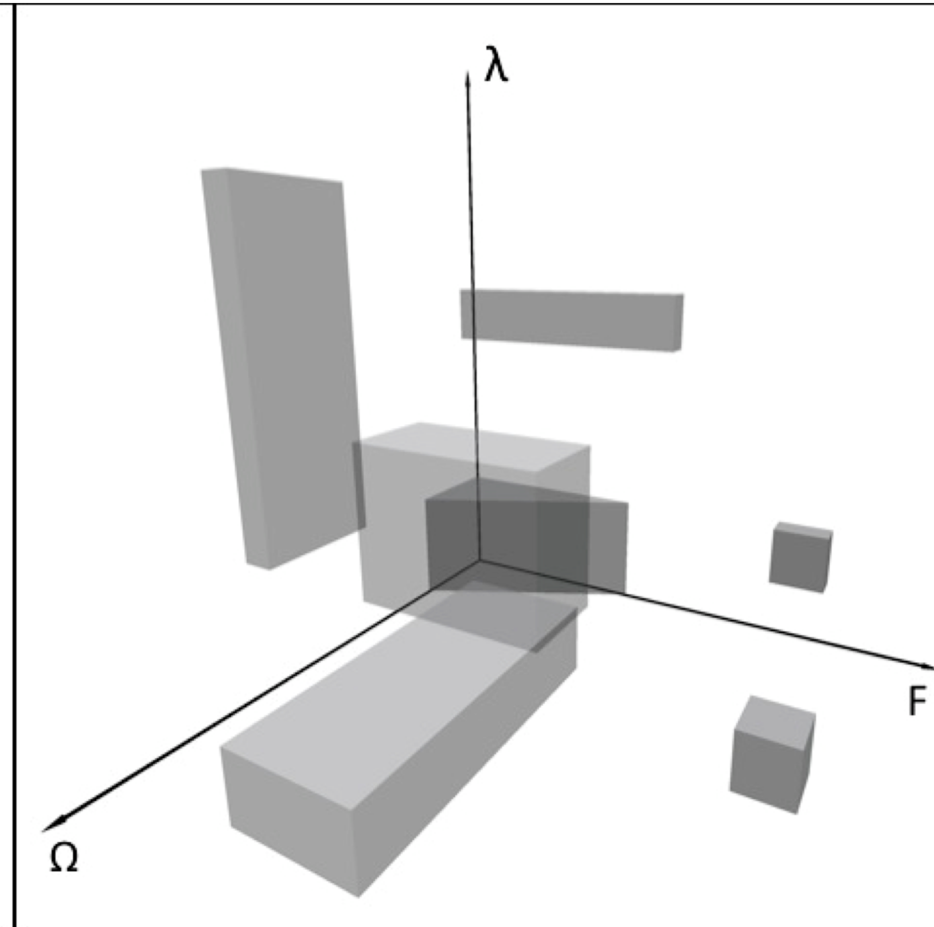
Example: Zwicky’s discovery of the compact star-forming dwarfs

Systematic Exploration of the Observable Parameter Spaces (OPS)

Its axes are defined by the observable quantities

Every observation, surveys included, carves out a hypervolume in the OPS

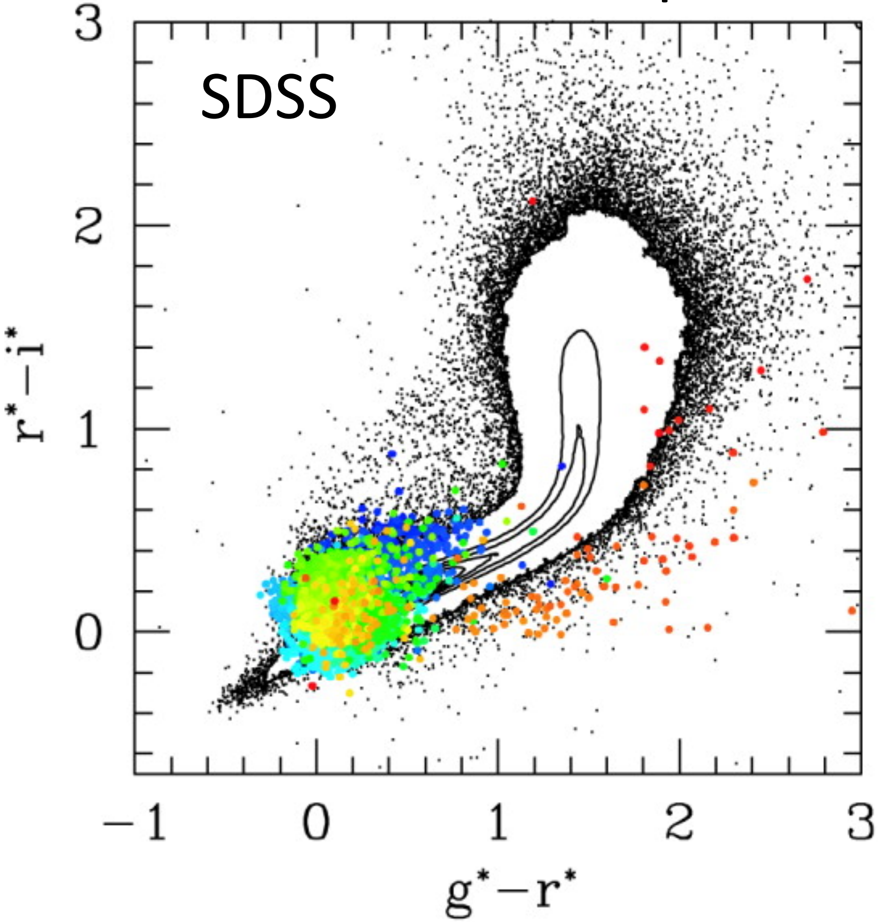




Technology opens new domains of the OPS → New discoveries

Measurements Parameter Space

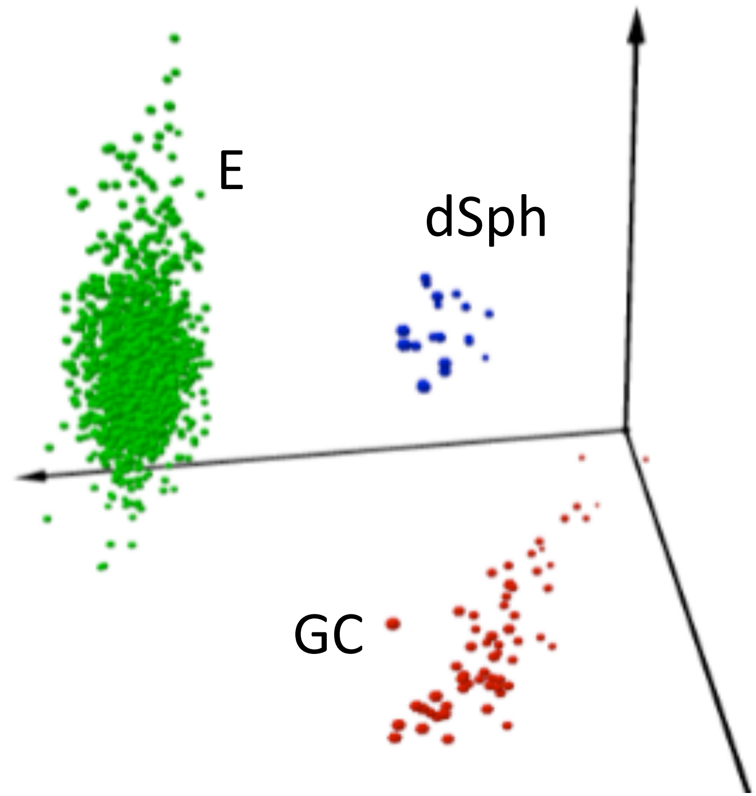
Colors of stars and quasars



Dimensionality ≤ the number of observed quantities

Physical Parameter Space

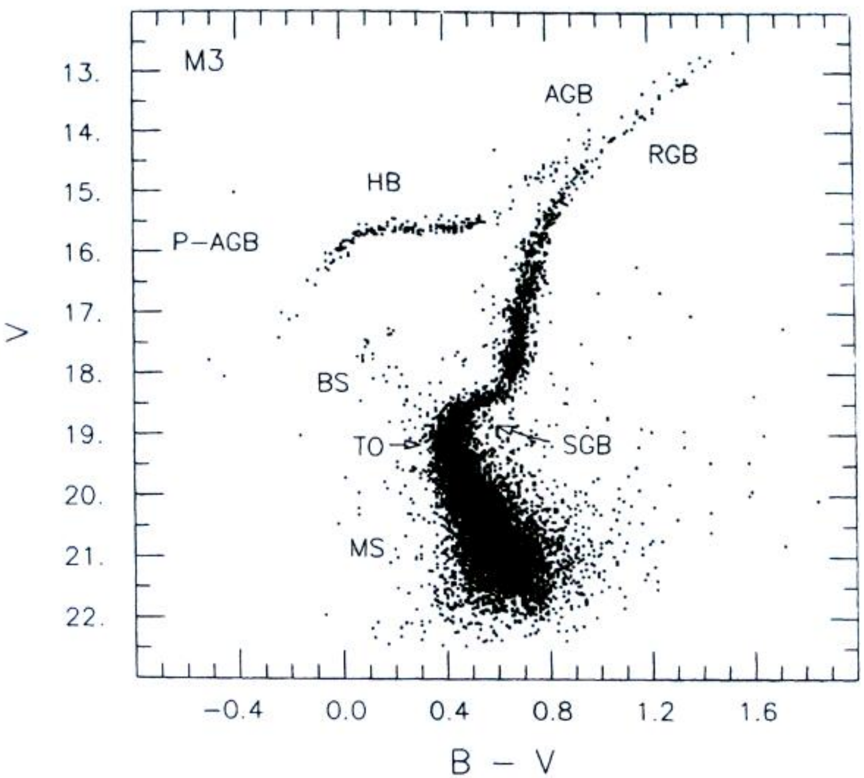
Fundamental Plane of hot stellar systems



Both are populated by objects or events

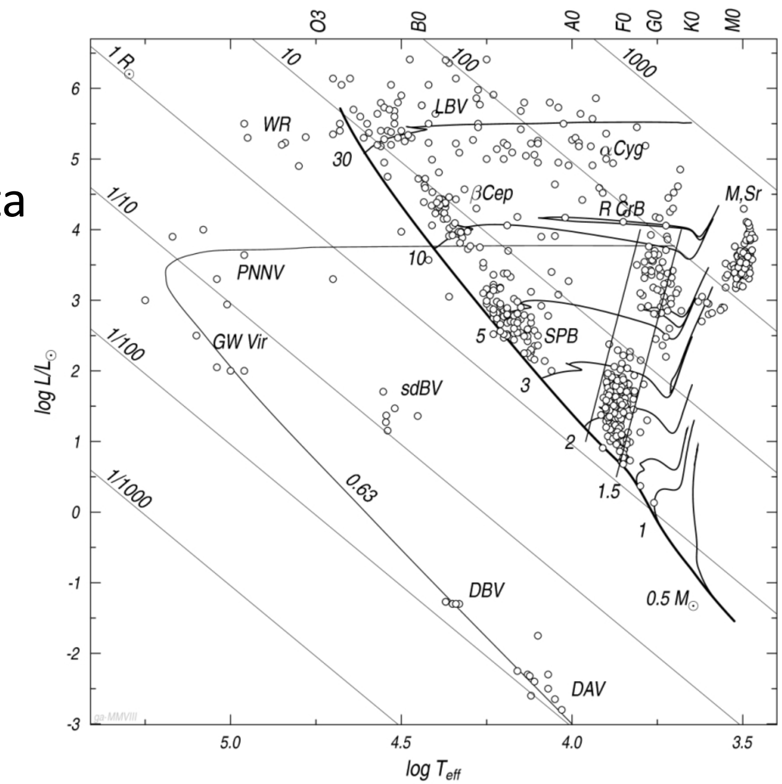
A Familiar Example: HR Diagram

Observable
Color-magnitude diagram



Theory
+
Other data

Theoretical
Temperature-Luminosity Space

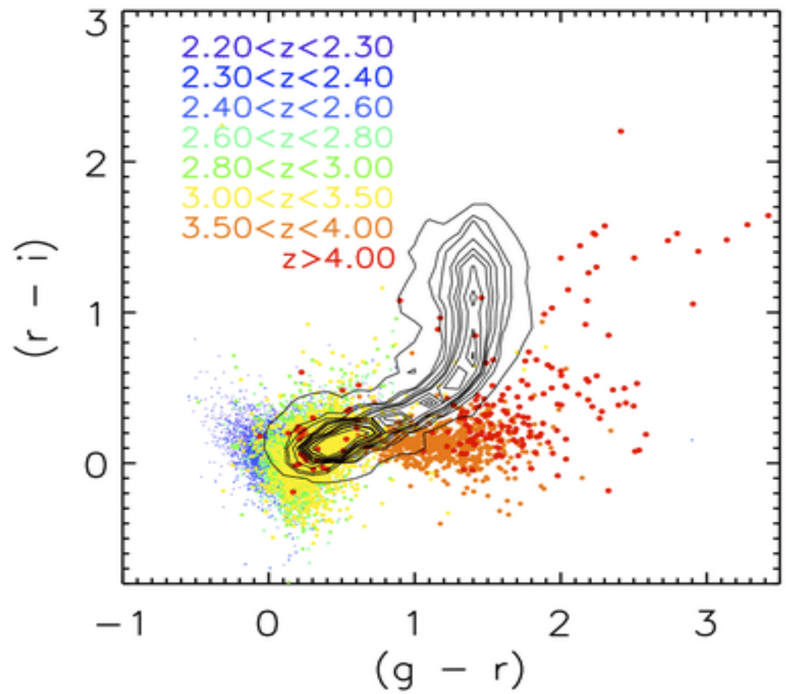


- Not filled uniformly: clustering indicates different families
- Empty regions may be due to selection effects or physics
- Clustering + dimensionality reduction = correlations

Mapping the Data Parameter Spaces

- Objects of a particular type (e.g., stars, galaxies, SNe, Quasars, …) may occupy only specific regions of a parameter space, and form clusters

- If enough known, training examples are known, this can be used for an automated, ***supervised classification***, or the searches for the rare, but known objects (e.g., quasars)

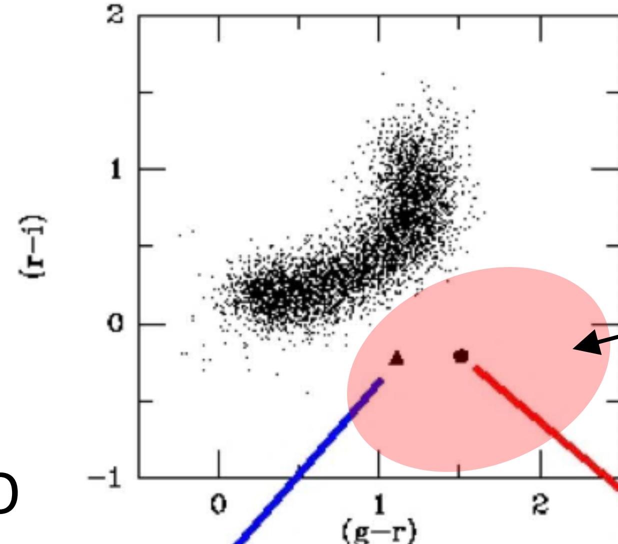


- **Unsupervised clustering** (let the data tell you what clusters are present) may reveal previously unknown types of objects, as outliers from the known clusters

Model-Based Outlier Search and Surprises

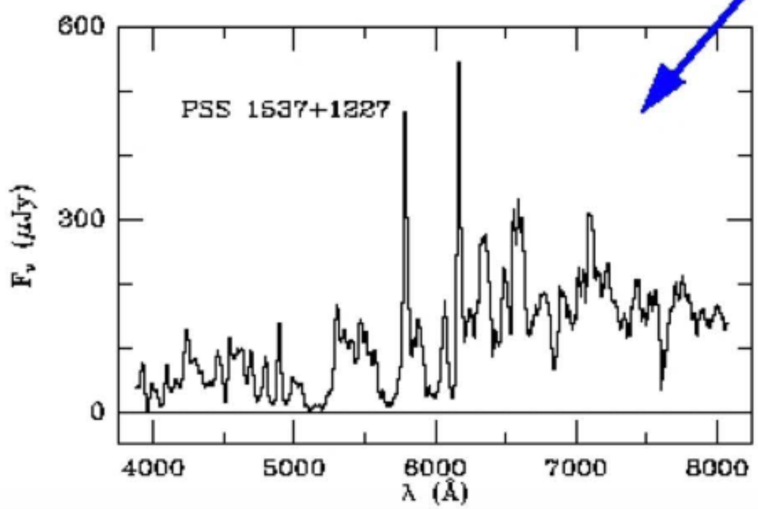
Sometimes we know where to look for outliers on the basis of a prior knowledge, e.g., quasars or brown dwarfs in a color space

... but sometimes you find something unexpected:

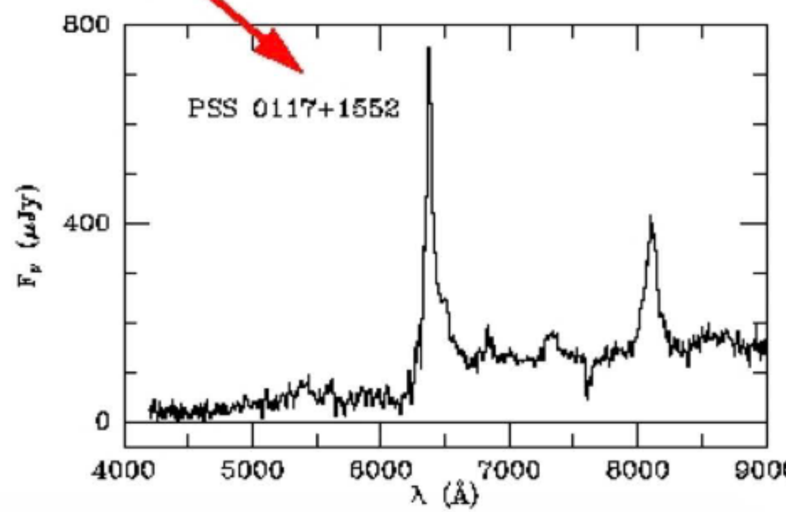


Expected high-z quasar region

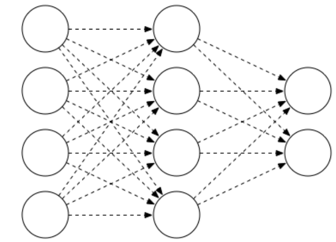
Peculiar Lo-BAL QSO



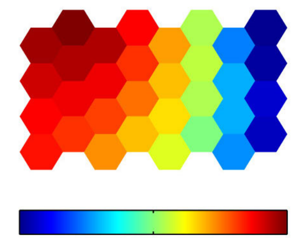
Typical z > 4 Quasar



Classification, Clustering, and Outliers



- **Supervised learning (classification):** use a known set of objects to train a classifier
  - Hard to find previously unknown things



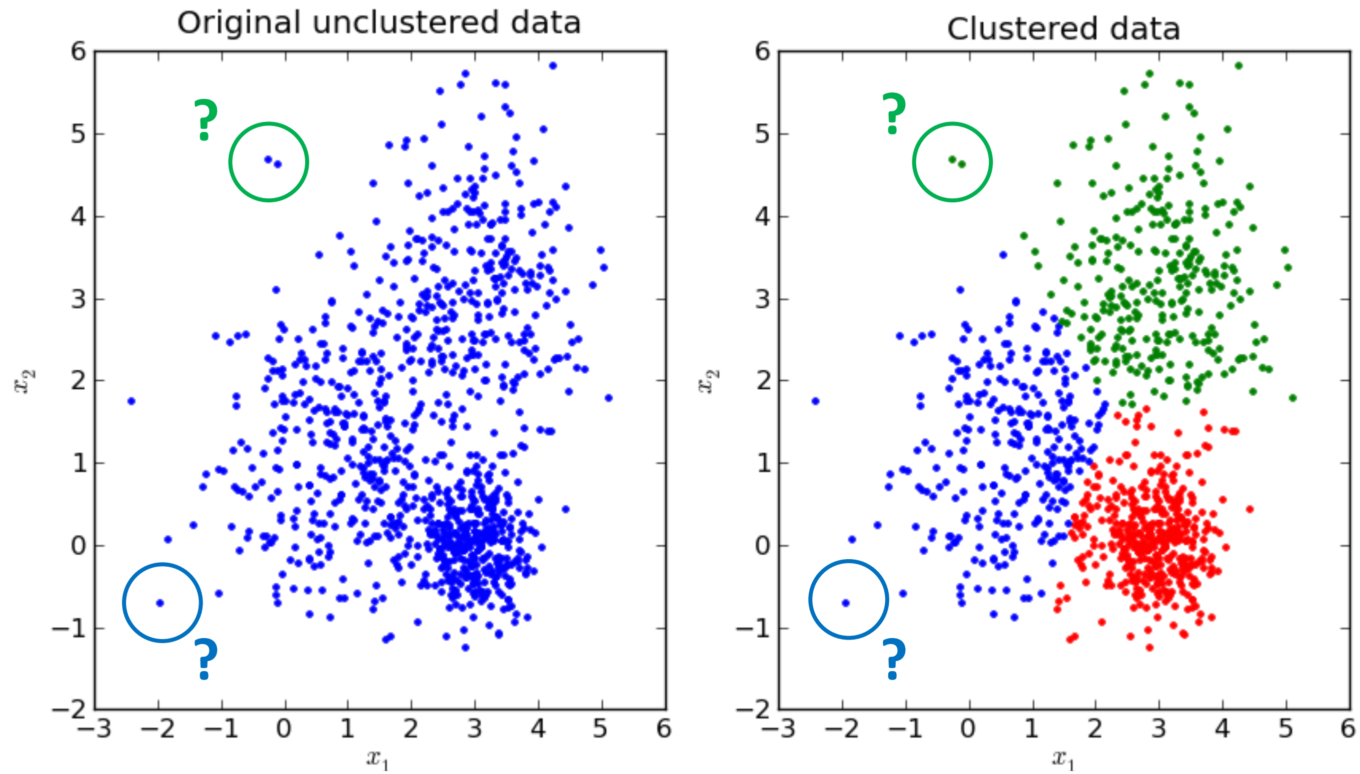
- **Unsupervised learning (clustering):** let the data tell you how many different kinds of things are there
  - Could find previously unknown types as outliers

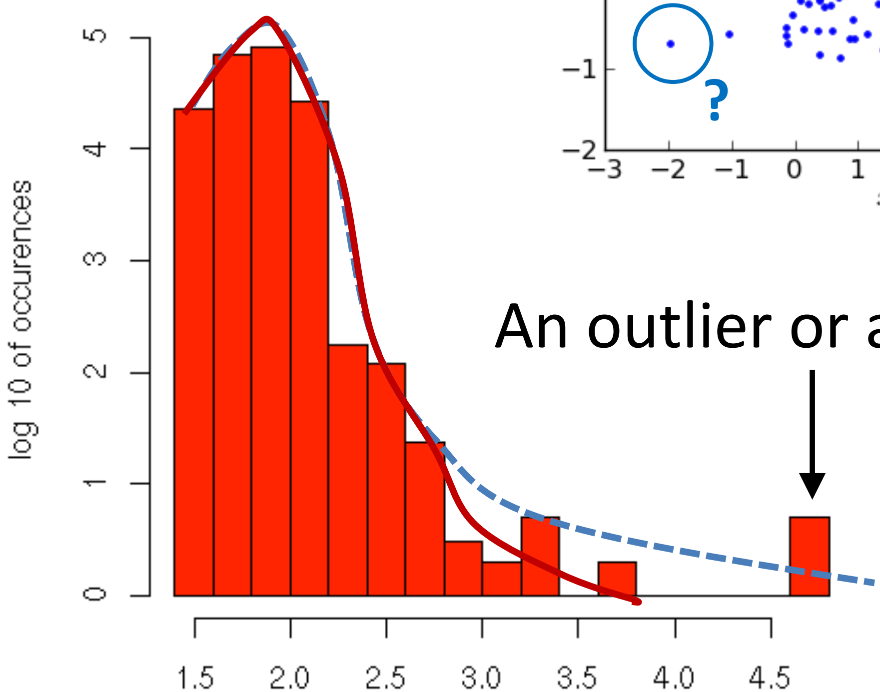
Supervised Algorithms
Neural Networks (MLP)
Boltzmann Machines
RBM
Decision Trees
Nearest Neighbor
Naive Bayes Classifiers
Bayesian Networks
Gaussian Processes
Regression
...

There is **no** "one size fits all": different choices for different problems

Unsupervised Algorithms
K-Means
Self-Organizing Maps
RDF
Fuzzy Clustering
CURE
ROCK
Vector Quantization
Probabilistic Principal Surfaces
...

What is an Outlier?



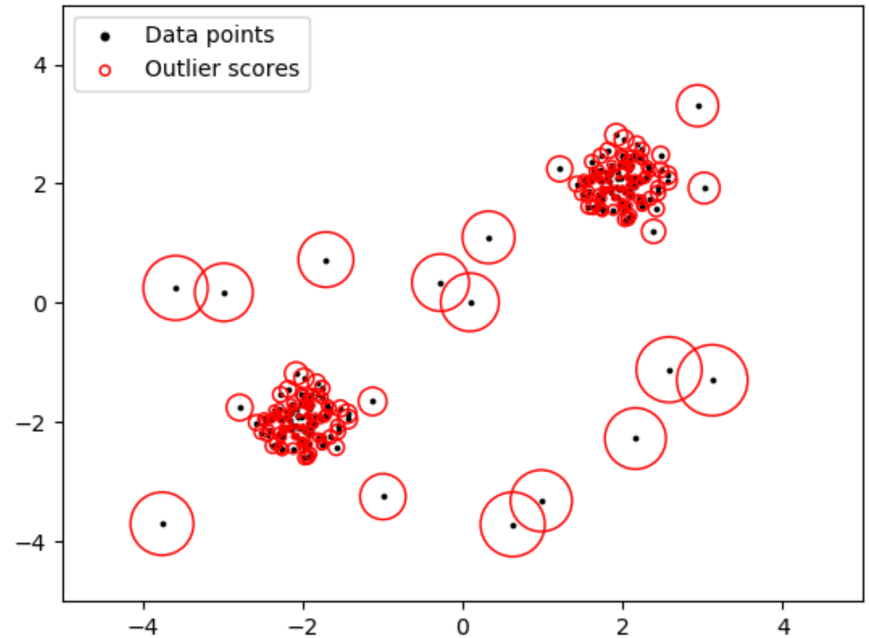


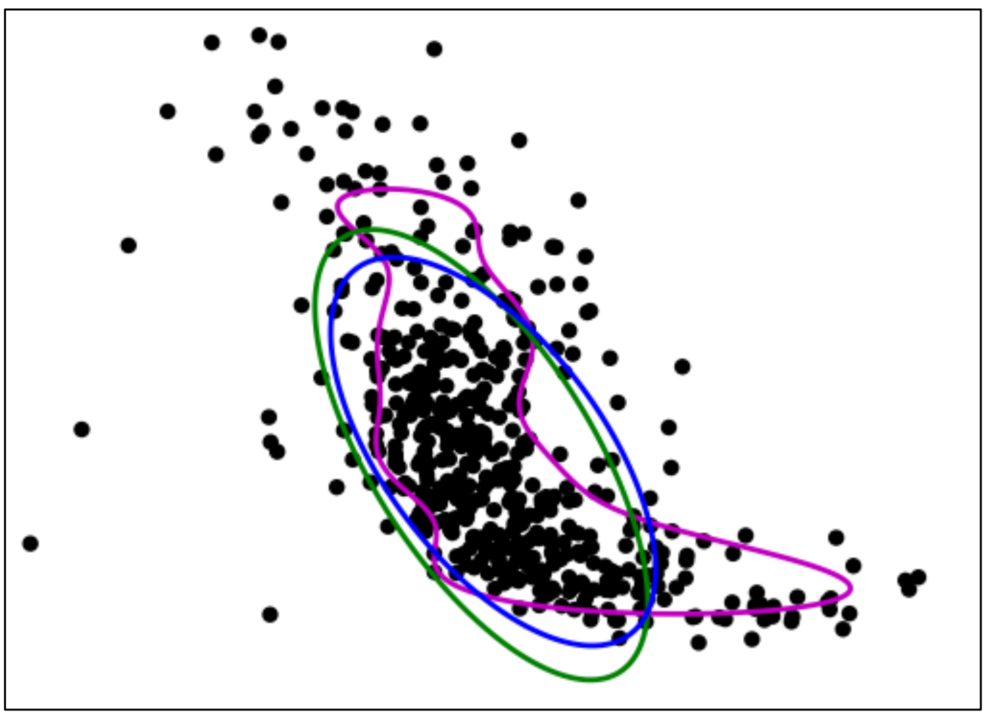
An outlier or a fat tail?

It depends on the underlying probability distribution, and they are seldom Gaussian

Clustering and Searches for Outliers

Sometimes this is easy, not critically dependent on the assumed probability density distributions of the clusters





But sometimes it isn’t

Having the right cluster descriptors, number of clusters, and metric of this feature space is crucial

Parameter Spaces for the Time Domain

(in addition to everything else: flux, wavelength, etc.)

- For ***surveys***:
  - Total exposure per pointing
  - Number of exposures per pointing
  - How to characterize the cadence?

⮑ Window function(s)

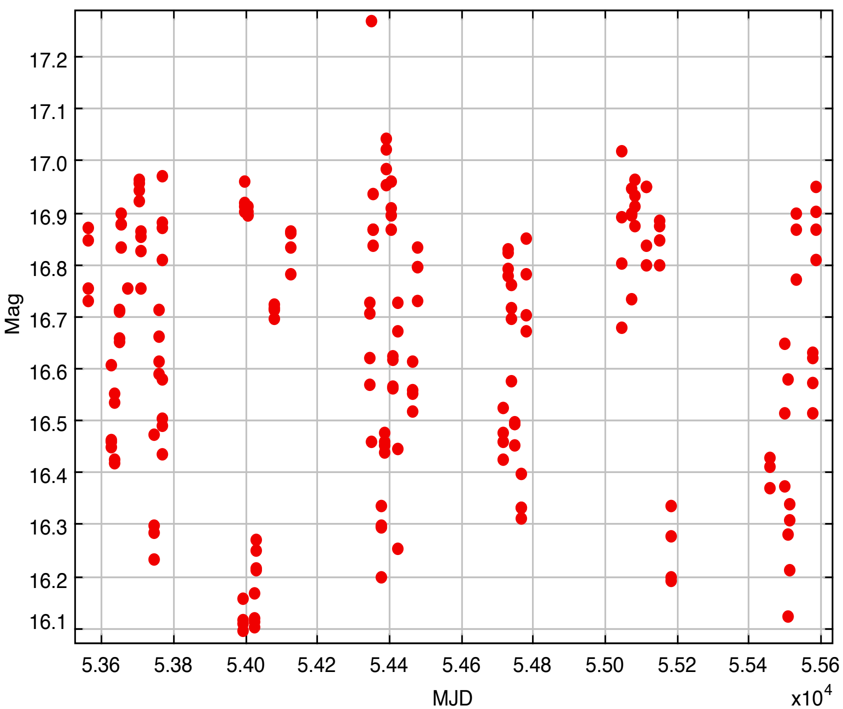
⮑ Inevitable biases

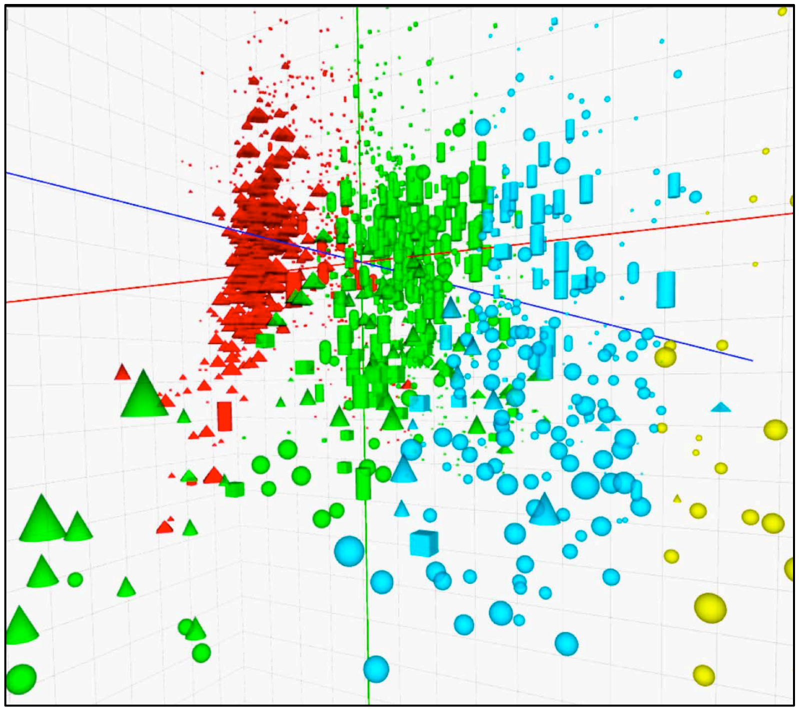
- For ***objects/events*** ~ light curves:
  - Significance of periodicity, periods
  - Descriptors of the power spectrum (e.g., power law)
  - Amplitudes and their statistical descriptors

… etc. – over 70 parameters defined so far, but which ones are the minimum / optimal set?

From Light Curves to Feature Vectors

- We compute ~ 70 parameters and statistical measures for each light curve: amplitudes, moments, periodicity, etc.
- This turns *heterogeneous* light curves into *homogeneous* ***feature vectors*** in the parameter space
- Apply a variety of automated classification methods





Variability Feature Space

- Generate homogeneous representation of time series by defining a number of **descriptive parameters**:
  - Morphology (shape): skew, kurtosis
  - Scale: Median absolute deviation, biweight midvar.
  - Variability: Stetson, Abbe, von Neumann
  - Timescale: periodicity, coherence, characteristic
  - Trends: Thiel-Sen
  - Autocorrelation: Durbin-Watson
  - Long-term memory: Hurst exponent
  - Nonlinearity: Teraesvirta
  - Chaos: Lyapunov exponent
  - Models: HMM, CAR, Fourier decomposition, wavelets
- Defines a **high-dimensional feature space** to characterize the temporal behavior

Feature Selection Algorithms

Most clustering and classification algorithms scale poorly with the dimensionality of the feature spaces. Feature selection is one set of **dimensionality reduction** techniques.

- **Filter methods** apply a statistical measure to assign a scoring to each feature, usually independently (univariate). The features are ranked by the score.
- **Wrapper methods** look for a set of features where different feature combinations are evaluated and compared to other combinations.
- **Embedded methods** learn which features best contribute to the accuracy of the model while the model is being created.
- The **scoring criterion** depends on the goal, e.g.:
  – Accurate predictions for the regression searches
  – Classification discrimination power for clustering

Feature Selection Algorithms

Optimal sets of features may be different for

- Different regression target variables:

  e.g., $y_1 = f_1(x_i, x_j, x_k, \ldots)$, $y_2 = f_2(x_p, x_q, x_r, \ldots)$, etc.

- Different classification tasks:

  e.g., $Class\ (A, B) = f(x_a, x_b, x_c, \ldots)$, $Class\ (A, B, C) = f(x_d, x_e, x_f, \ldots)$

- Different regression or classification algorithms:

  e.g., ANN, DT, RF, SVM, …

. . . so they have to be optimized in each individual case

See:

Donalek et al., IEEE BigData 2013, p. 35 = arxiv/1310.1976

D'Isanto et al. 2016, MNRAS, 457, 3119

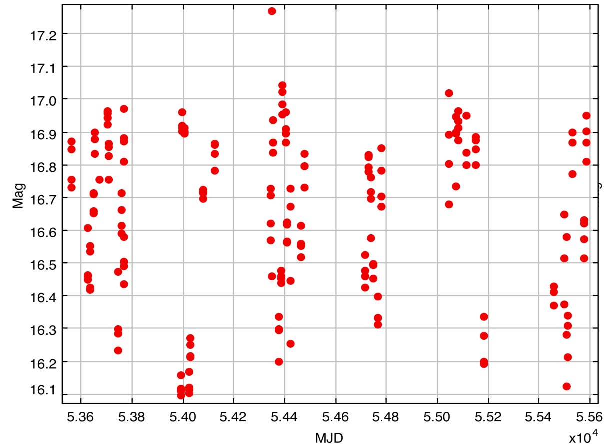
Feature Selection Algorithms: Examples

- **Fast Relief Algorithm** (aka ReliefF) ranks features according to how well their values distinguish between instances.
- **Fisher Discriminant Ratio** (FDR) ranks features according to their classification discriminatory power. It can be applied only to binary classification problems.
- **Correlation-based Feature Selection** (CFS) is a wrapper method which selects features that have low redundancy (i.e., not correlated with each other) and is strongly predictive of a class.
- **Fast Correlation Based Filter** (FCBF) is a supervised filter algorithm, similar to the CFS. Searches for features that have predominant correlation with the class . Can be computationally efficient with very high dimensional data.
- **Multi Class Feature Selection** (MCFS) is an unsupervised method based on the spectral analysis of the data.

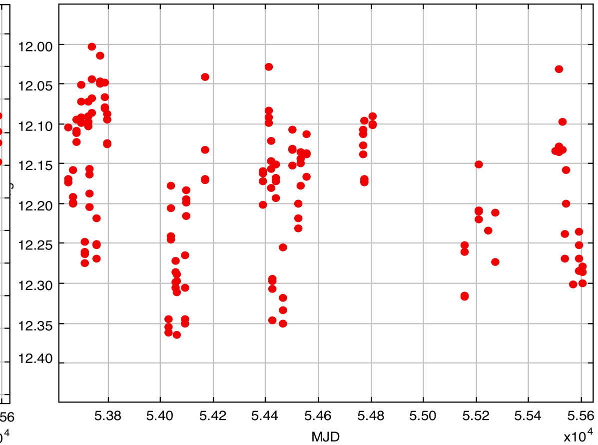
... etc.

Optimizing Feature Selection

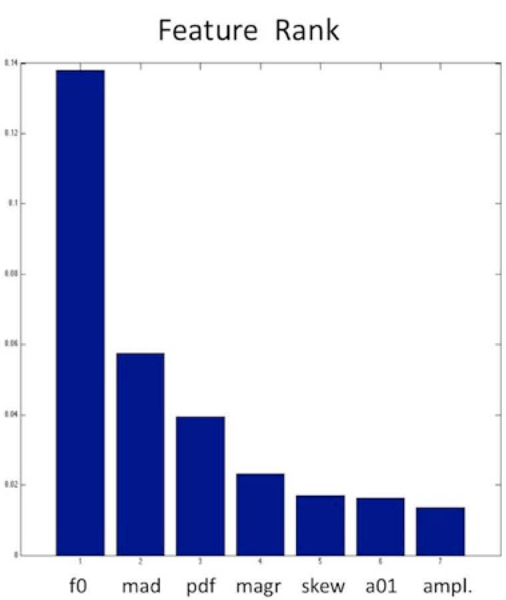
Rank features in the order of classification quality for a given classification problem, e.g., RR Lyrae vs. WUMa

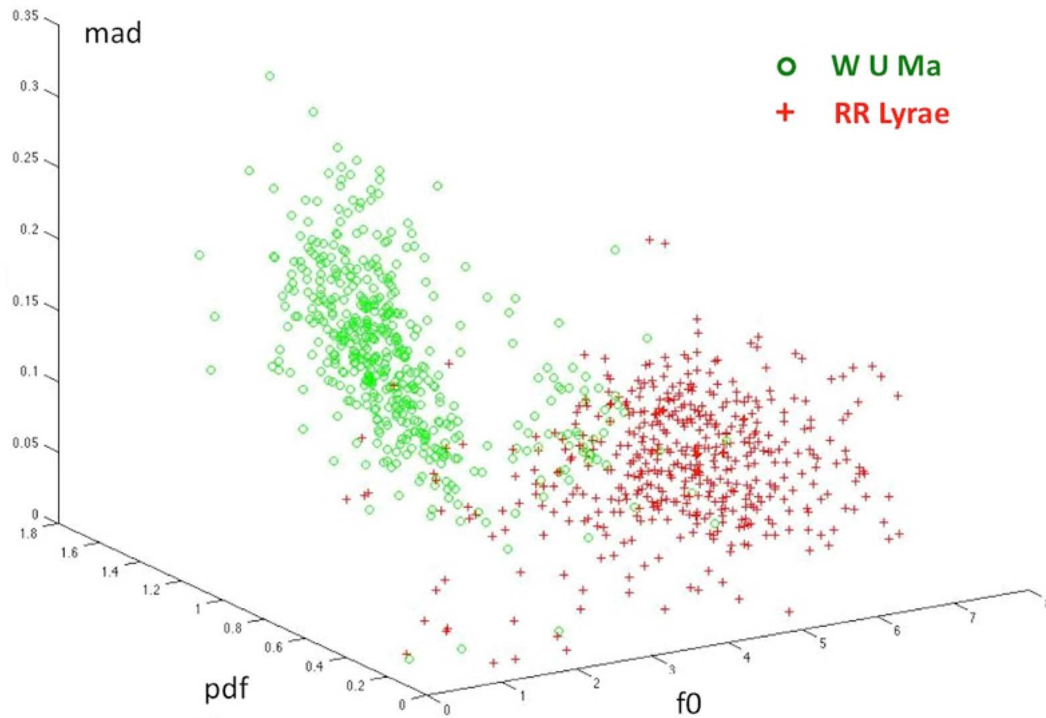


RR Lyrae



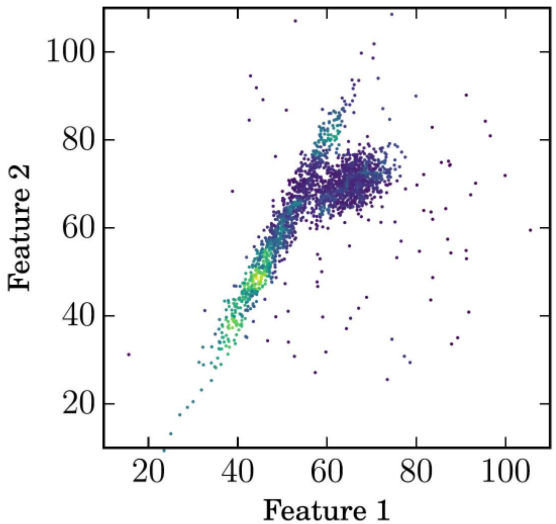
Eclipsing binary (W U Ma)



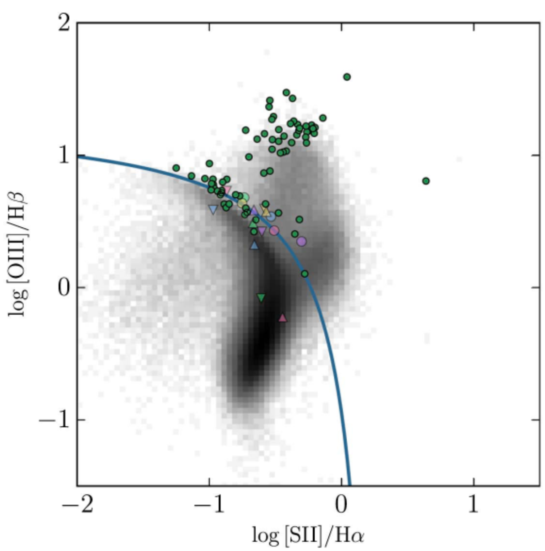


Examples from Astronomy:

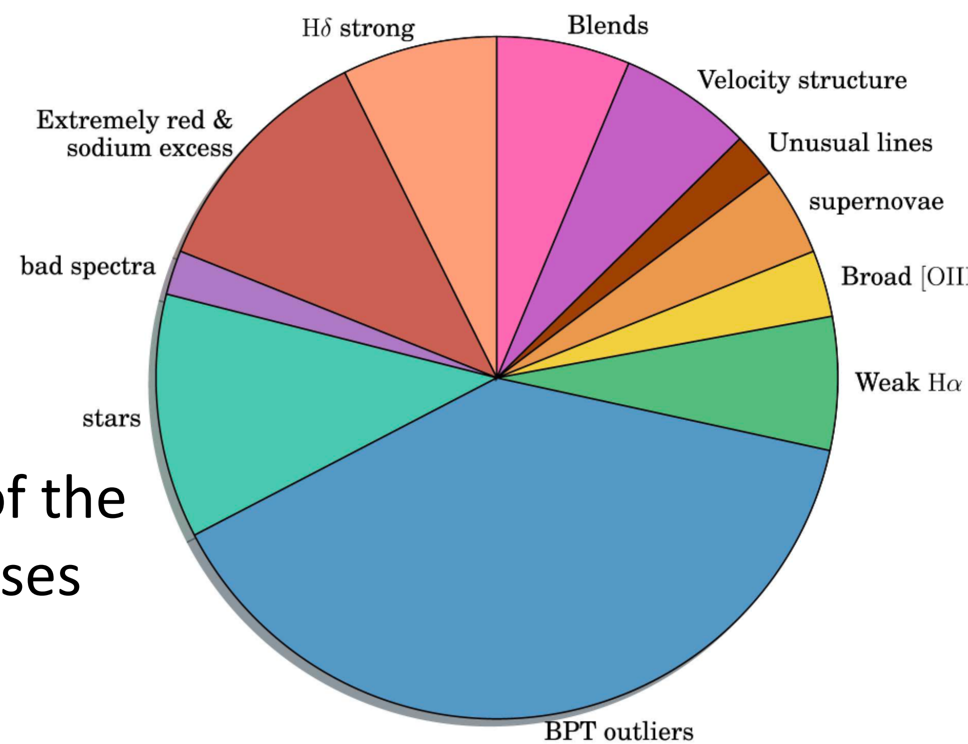
"The weirdest SDSS galaxies: results from an outlier detection algorithm", D. Baron & D. Poznanski 2017, MNRAS 465, 4530



Used Random Forests algorithm to classify SDSS galaxies using spectroscopic properties. Defined a "Weirdness" parameter to quantify the outliers.

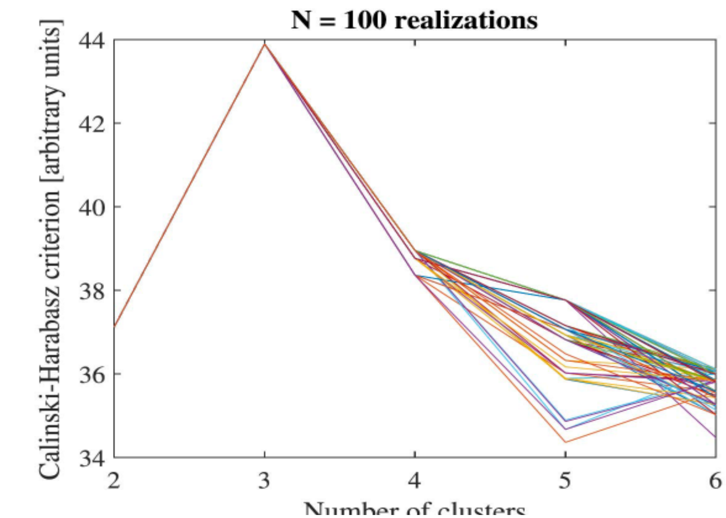


All outliers found are members of the known classes of objects.

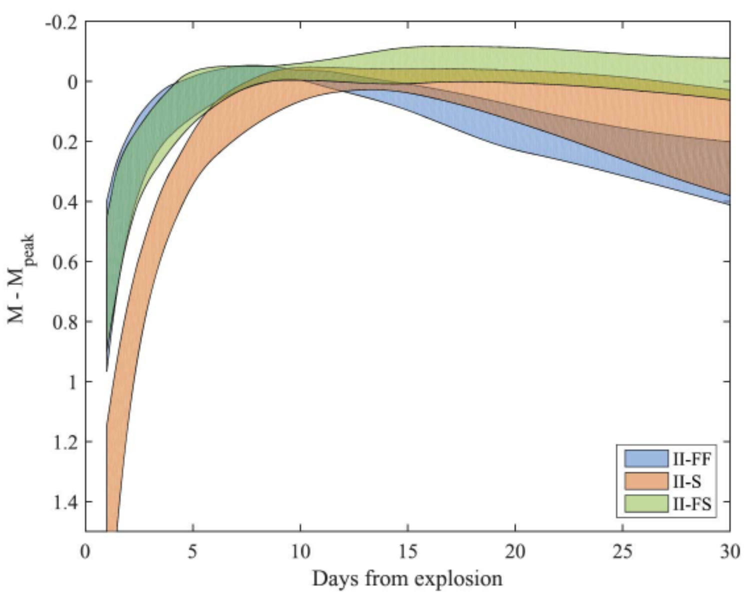


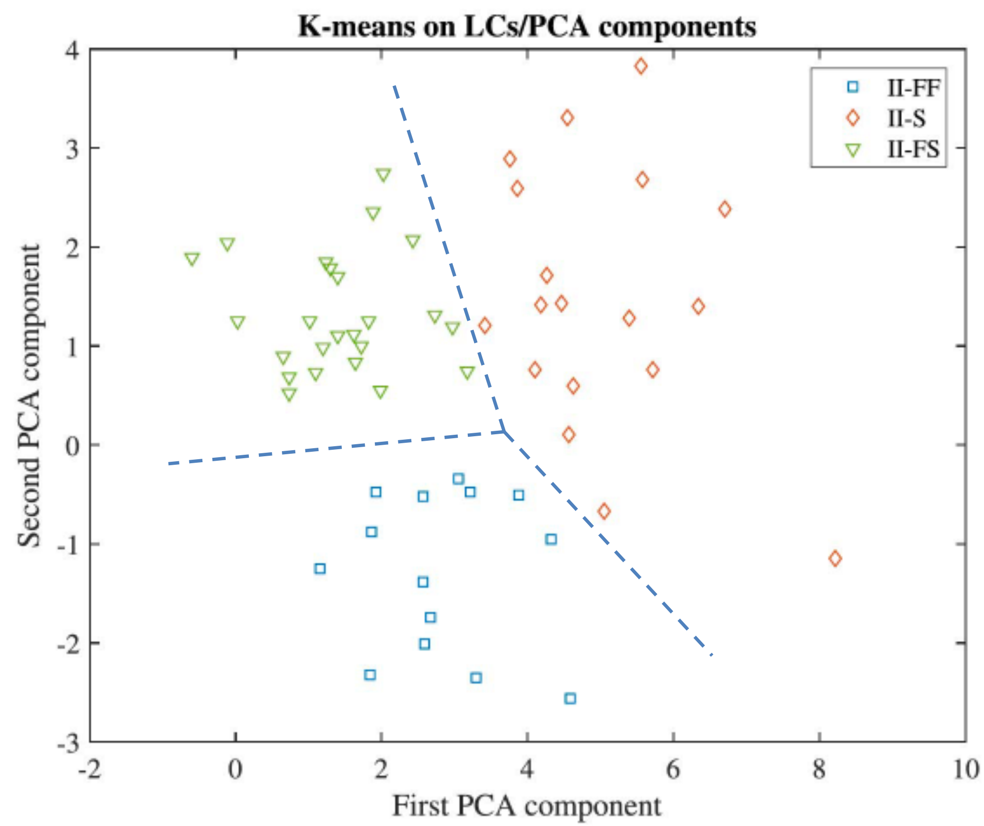
Examples from Astronomy:

"Unsupervised Clustering of Type II Supernova Light Curves", A. Rubin & A. Gal-Yam 2016, ApJ, 828, 111



Used the K-Means algorithm to identify 3 principal clusters: slow rise, fast rise – fast decay, and fast rise – slow decay





To Recap:

- Astronomy is now well into the **Petascale data regime,** and data volumes and rates grow exponentially according to Moore's Law
  - Most data come from the large surveys
  - The biggest growth now is in the time domain
  - This is true across all wavelengths
  - Growth of **data complexity** and **information content**
- Derived source catalogs typically contain **$\sim 10^9$ objects**, with **$\sim 10^2$ - 10^3 parameters** (features) each
  - Data fusion of different surveys increases the data complexity and discovery potential
  - We use **Machine Learning to** process and analyze the data, including source classification and selection of **interesting targets** for the follow-up studies