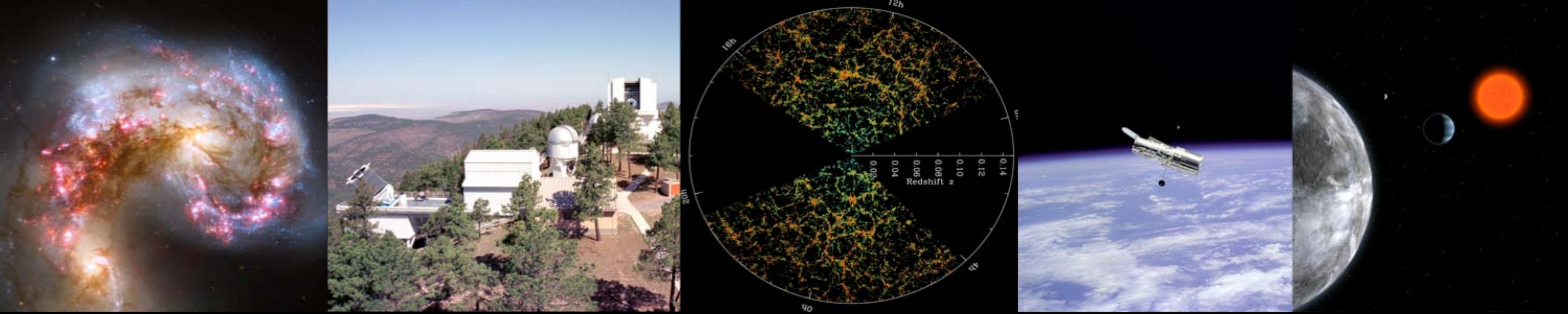
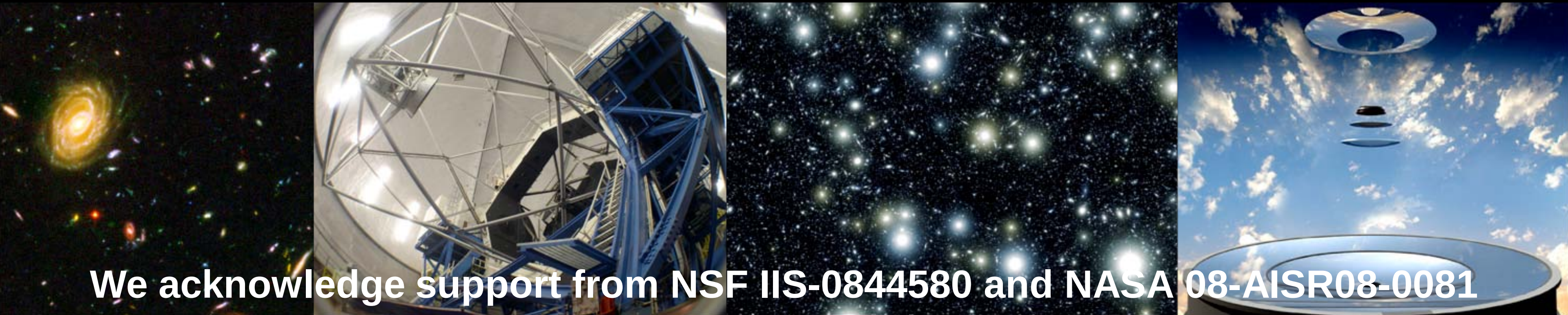


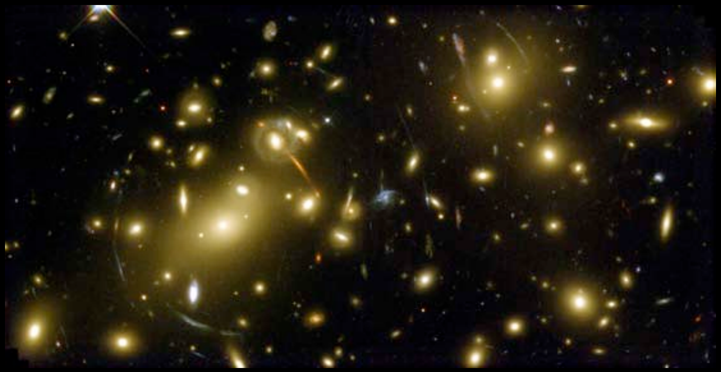
The Challenge of Data in an Era of Petabyte Surveys

Andrew Connolly
University of Washington



We acknowledge support from NSF IIS-0844580 and NASA 08-AISR08-0081

The science of big data sets



Big Questions

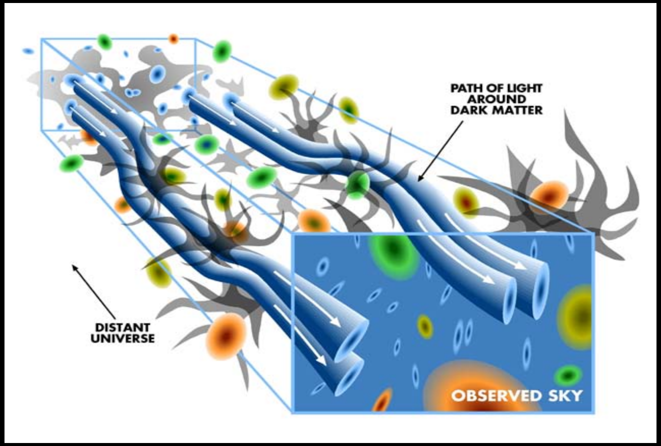
Nature of Dark Energy

Nature of Dark Matter

Small Effects

Requires large volumes

Systematics are important



Large projects, small science teams

Collaborative

Distributed ideas

What is the science we want to do?

- **Finding the unusual**
  - Billion sources a night
  - Nova, supernova, GRBs
  - Instantaneous discovery
- **Finding moving sources**
  - Asteroids and comets
  - Proper motions of stars
- **Mapping the Milky Way**
  - Tidal streams
  - Galactic structure
- **Dark energy and dark matter**
  - Gravitational lensing
  - Slight distortion in shape
  - Trace the nature of dark energy

What are the operations we want to do?

- Finding the unusual
  - Anomaly detection
  - Dimensionality reduction
  - Cross-matching data
- Finding moving sources
  - Tracking algorithms
  - Kalman filters
- Mapping the Milky Way
  - Density estimation
  - Clustering (n-tuples)
- Dark energy and dark matter
  - Computer vision
  - Weak Classifiers
  - High-D Model fitting

Science is driven by precision we need to tackle issues of complexity:

1. **Complex models of the universe**
   What is the density distribution and how does it evolve
   What processes describe star formation and evolution

2. **Complex data streams**
   Observations provide a noisy representation of the sky

3. **Complex scaling of the science**
   Scaling science to the petabyte era
   Learning how to do science without needing a CS major

The challenge of big surveys

2000 - 2014

Sloan Digital Sky Survey (SDSS)

120 Mpixel camera, (0.08 PB in 10 yrs)

300 Million unique sources (4 TB)

PanSTARRS (PS1)

1.4 Gpixel camera (0.4 PB per year)

2018 –

Large Synoptic Survey Telescope (LSST)

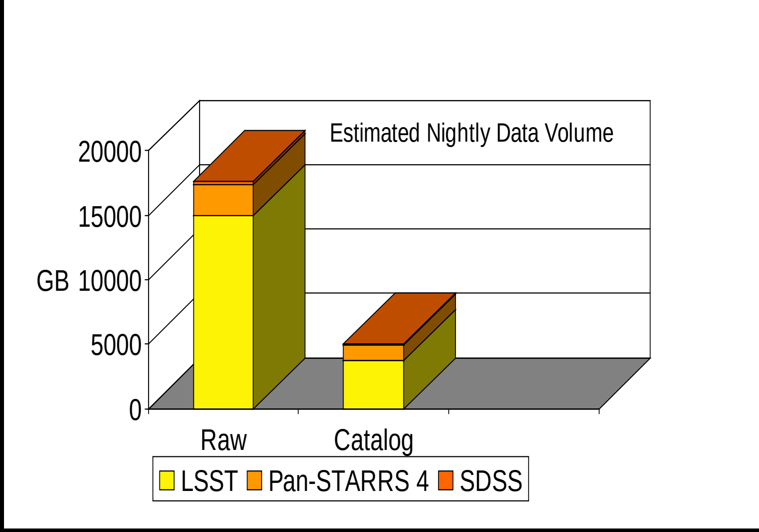
3.2 Gpixel camera (6 PB per year)

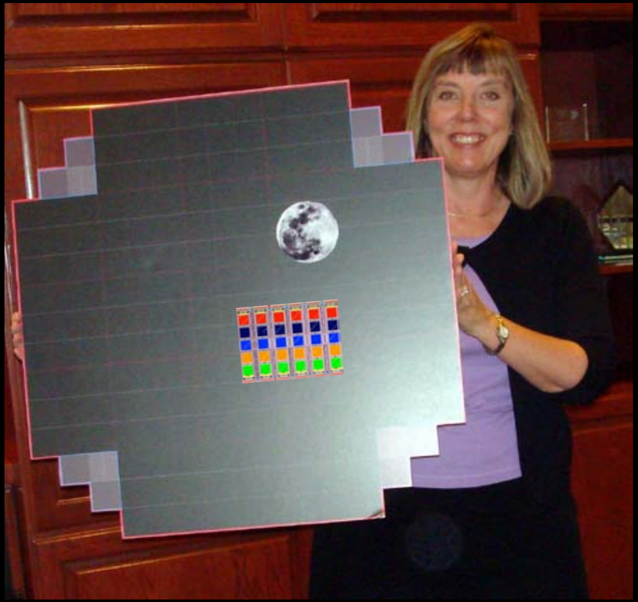
1000 observations of every source

Simulations (gorilla in the room)

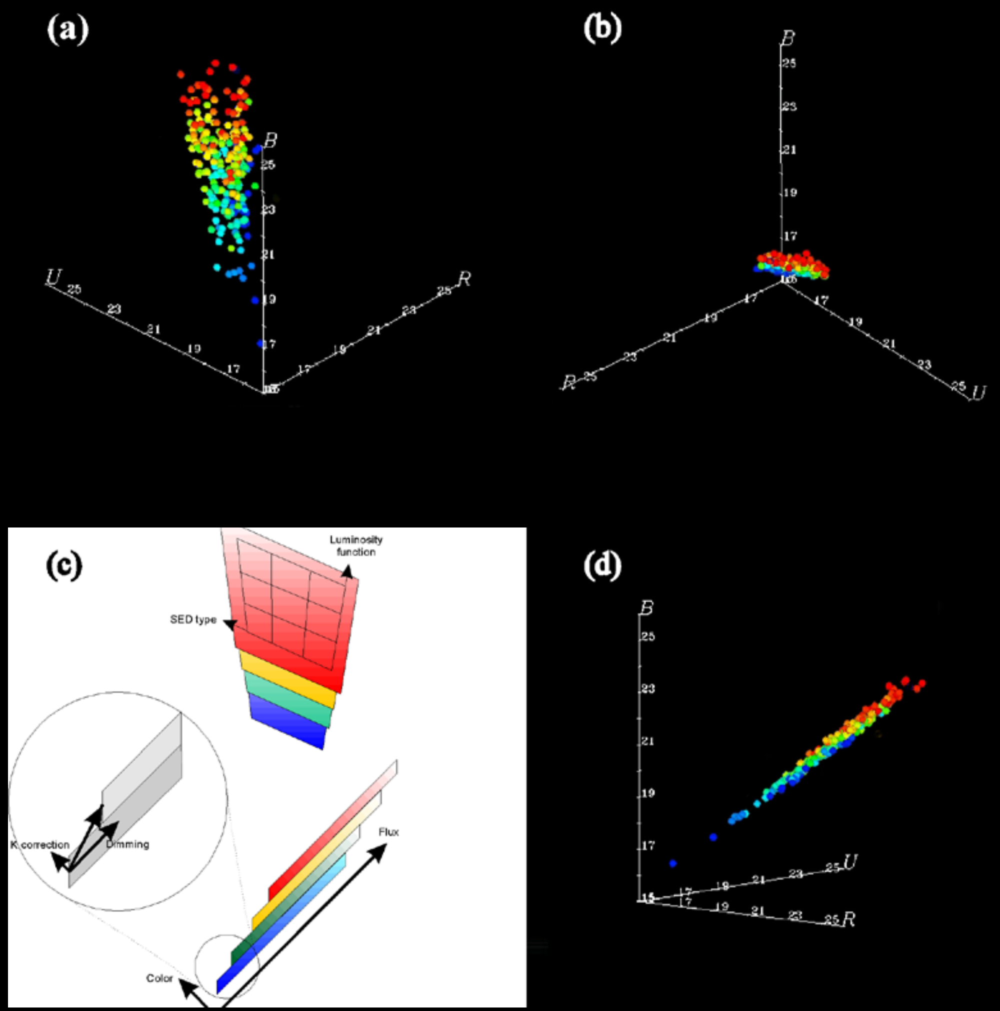
TBs per run generated today

TBs per hour in the next 5 years





Case Study: Complexity and simplifying data



We can measure many attributes about sources we detect…

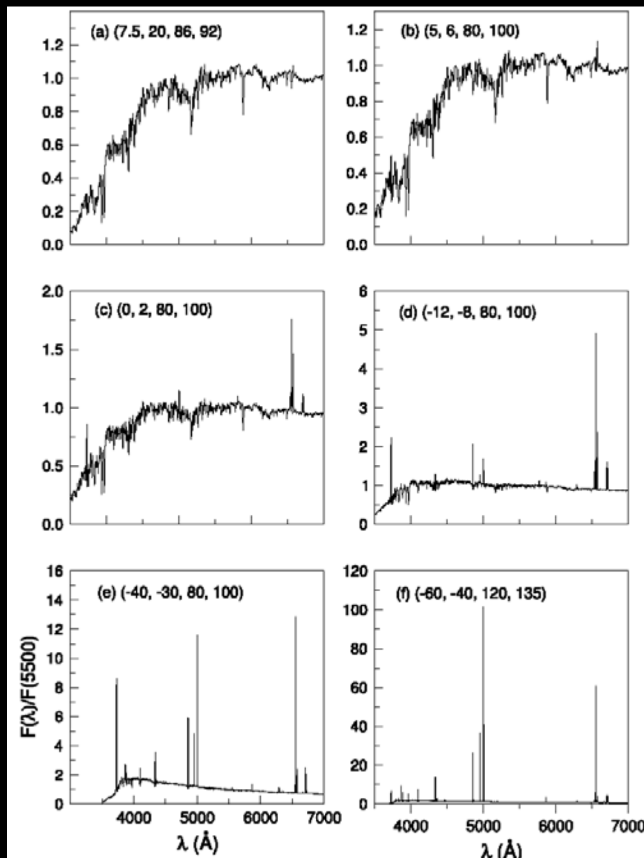
… which ones are important and why (what is the dimensionality of the data and the physics)

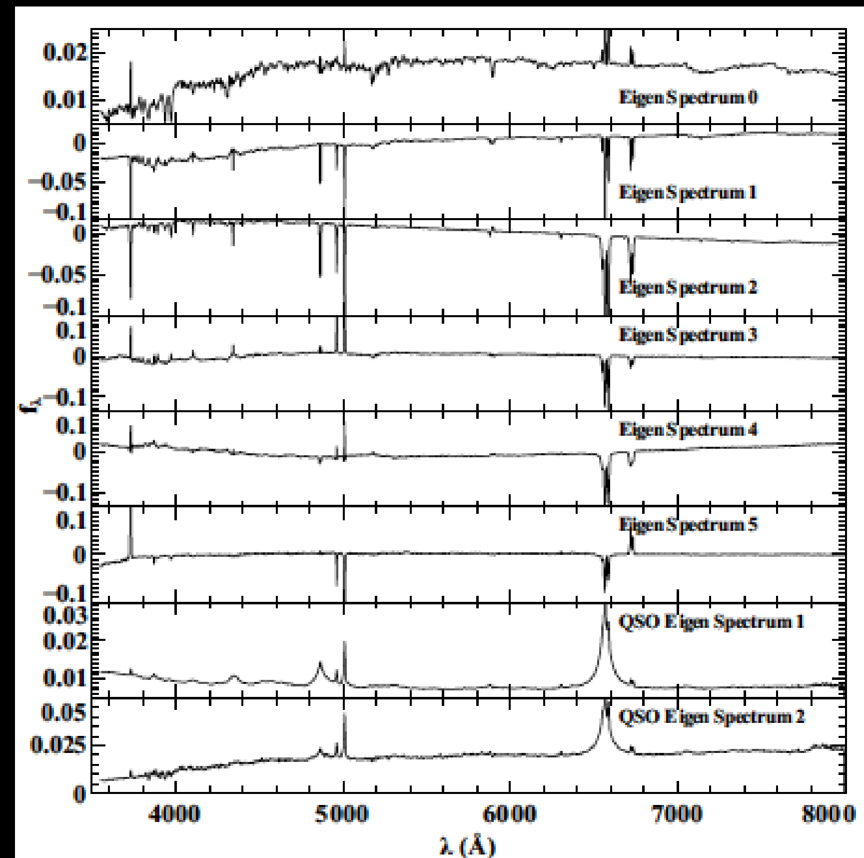
Connolly et al 1995

Low dimensionality even with complex data

Old

Young



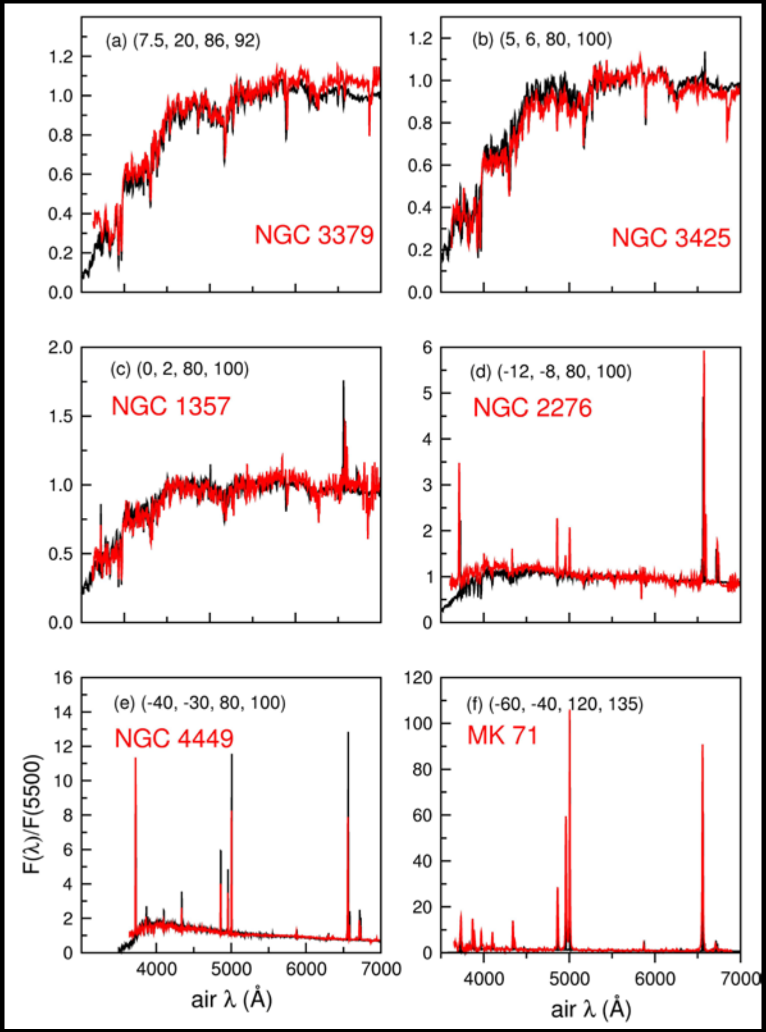


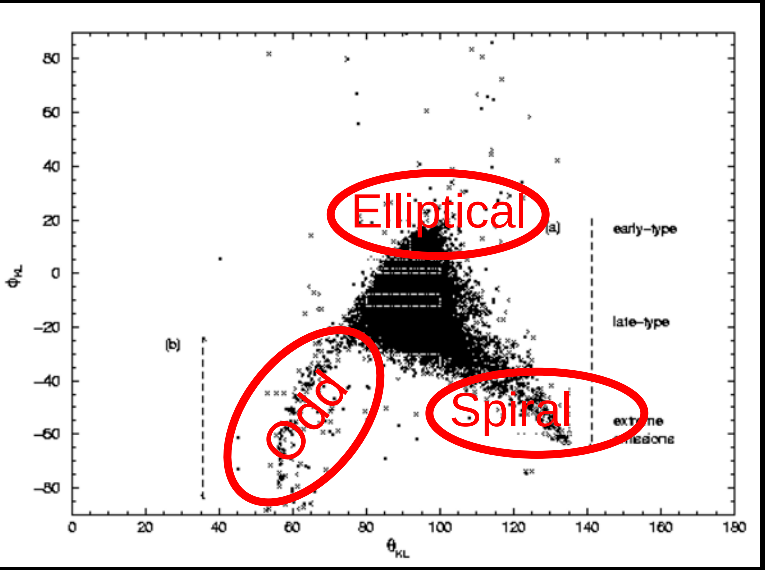
4000-dimensional (λ’s)

$$f(\lambda) = \sum_{i<N} a_i e_i(\lambda)$$

10 components Ξ >99% of variance

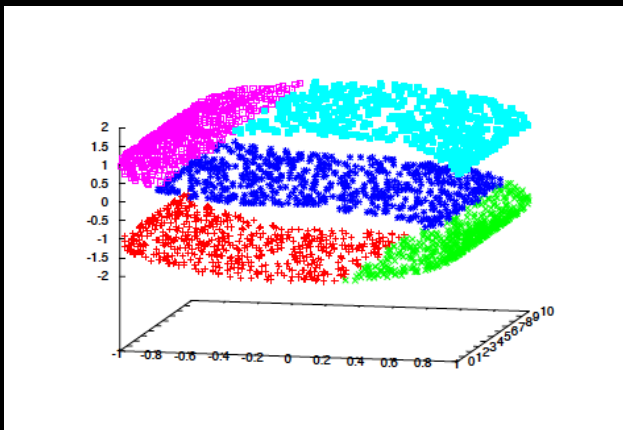
Dimensionality relates to physics

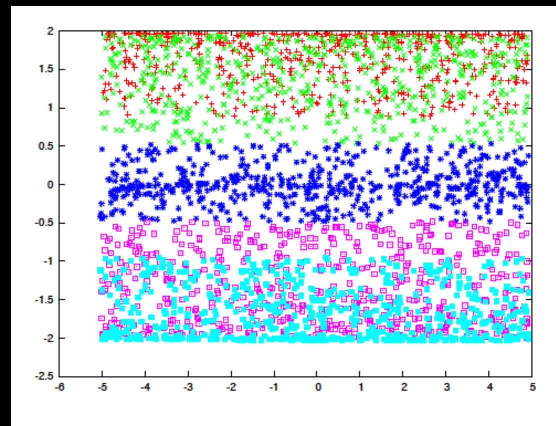




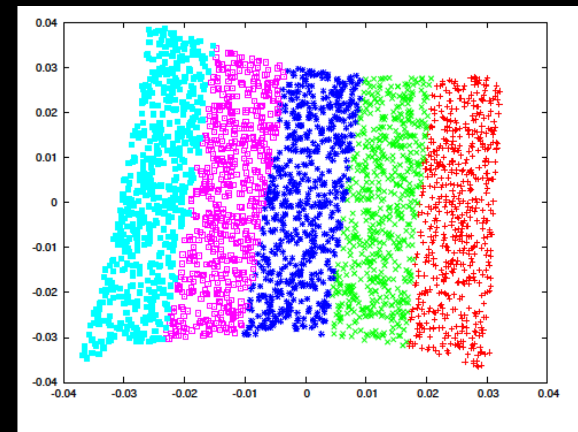
400-fold compression
Signal-to-noise weighted
Accounts for gaps and noise
Compression contains physics

Responding to non-linear processes





PCA



LLE

Local Linear Embedding (Roweis and Saul, 2000)

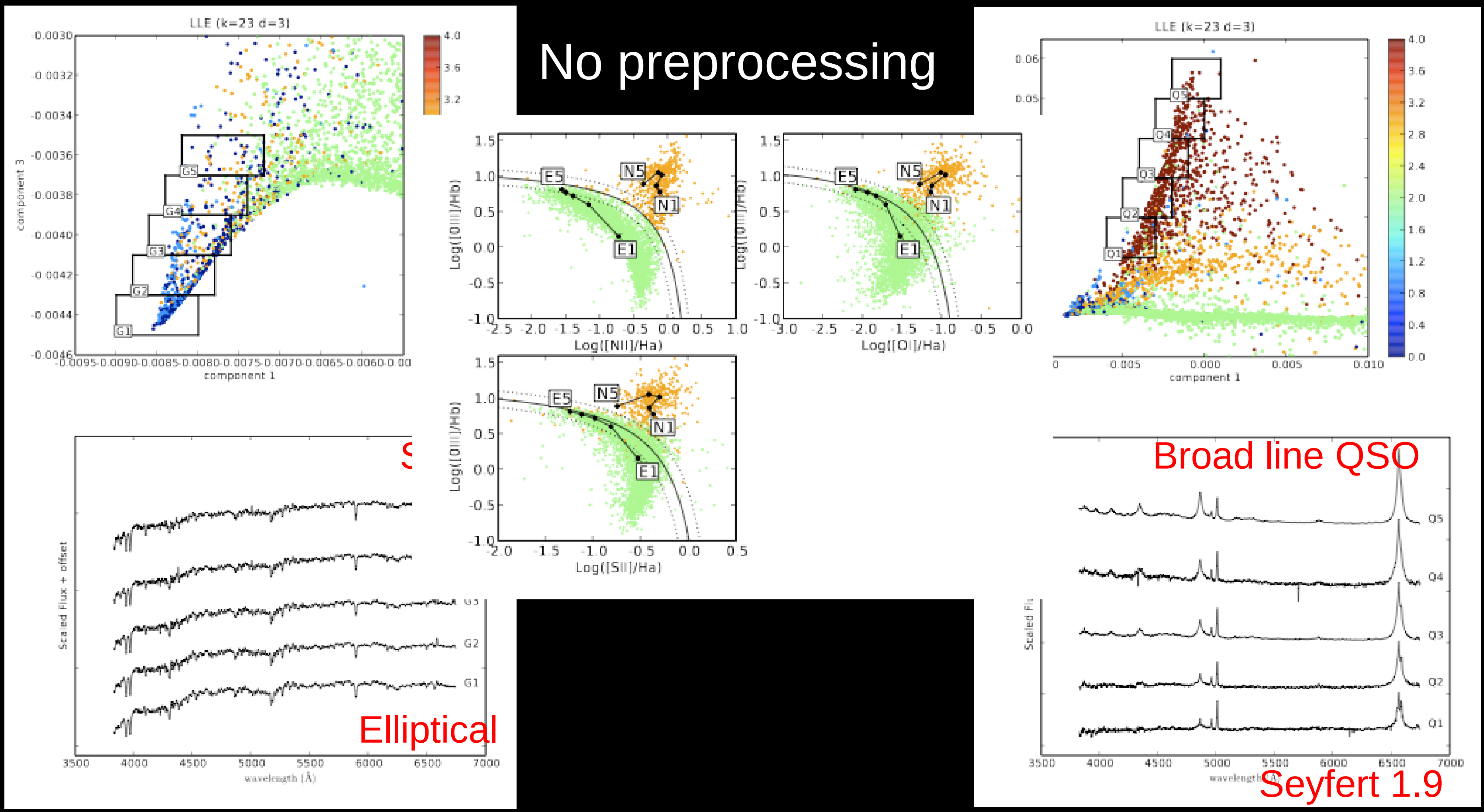
$$\mathcal{E}_1^{(i)}(\mathbf{w}^{(i)}) = \left| \mathbf{x_i} - \sum_{j=1}^{K} w_j^{(i)} \mathbf{x}_{n_j^{(i)}} \right|^2$$

$$\mathcal{E}_2(\mathbf{Y}) = \sum_{i=1}^{N} \left| \mathbf{y_i} - \sum_{j=1}^{K} w_j^{(i)} \mathbf{y}_{n_j^{(i)}} \right|$$

Preserves local structure

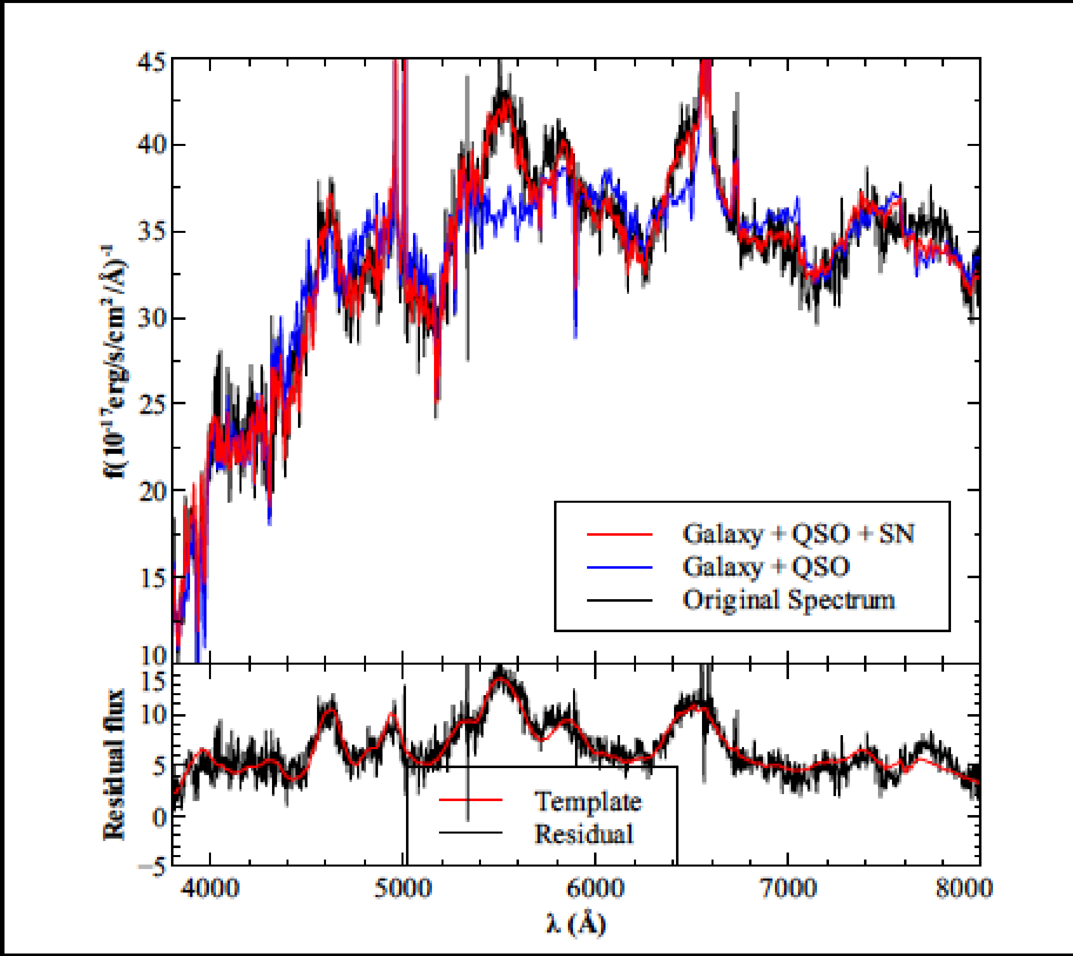
Slow and not always robust to outliers

A compact representation accounting for broad lines



VanderPlas and Connolly 2009

Case Study: Learning structure to find the unusual



Type Ia supernovae
0.01% contamination
to SDSS spectra

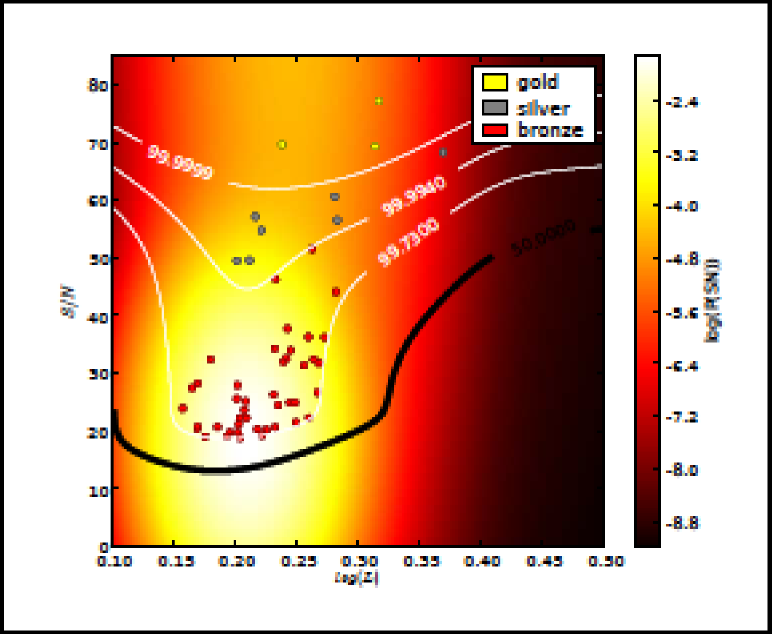
Type Ia supernovae
Visible for long
(-15 to 40 days)

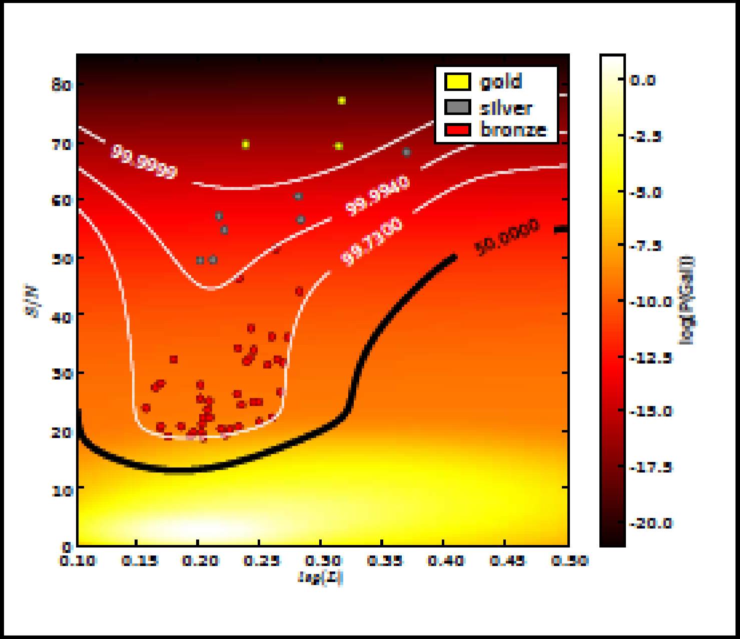
Well defined spectral
signatures

Magwick et al 2003

$$SN(\lambda) = f(\lambda) - \sum_{i<N} a_i e_{g_i}(\lambda) - \sum_{i<N} q_i e_{q_i}(\lambda)$$

Bayesian classification of outliers

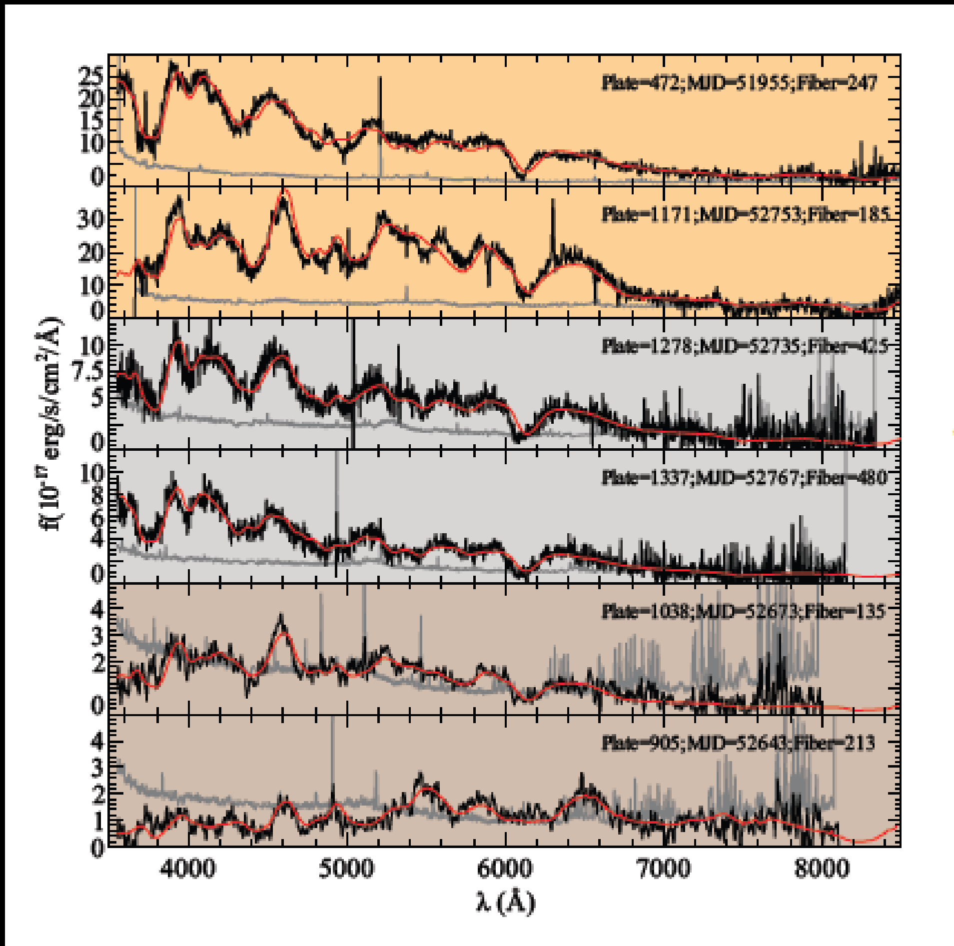


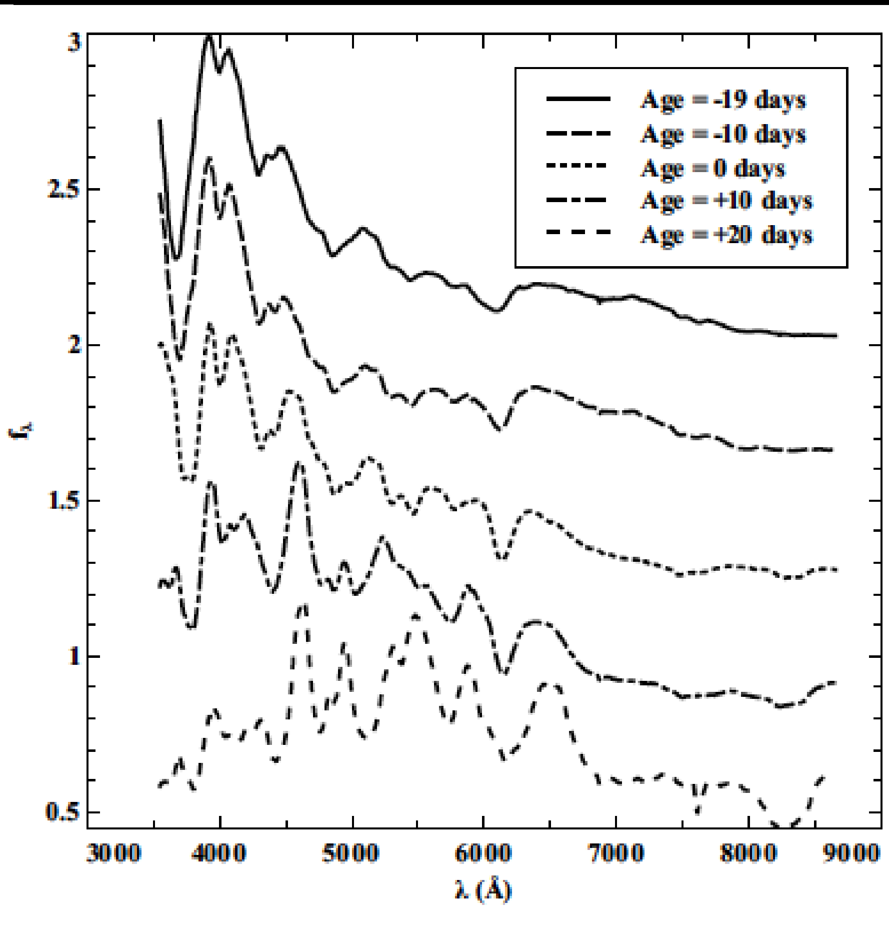


$$P(C_1|x) = \frac{p(x|C_1)P(C_1)}{p(x|C_1)P(C_1) + p(x|C_2)P(C_2)}$$

Density estimation using a mixture of Gaussians gives P(x|C): likelihood vs signal-to-noise of anomaly

Probabilistic identification with no visual inspection

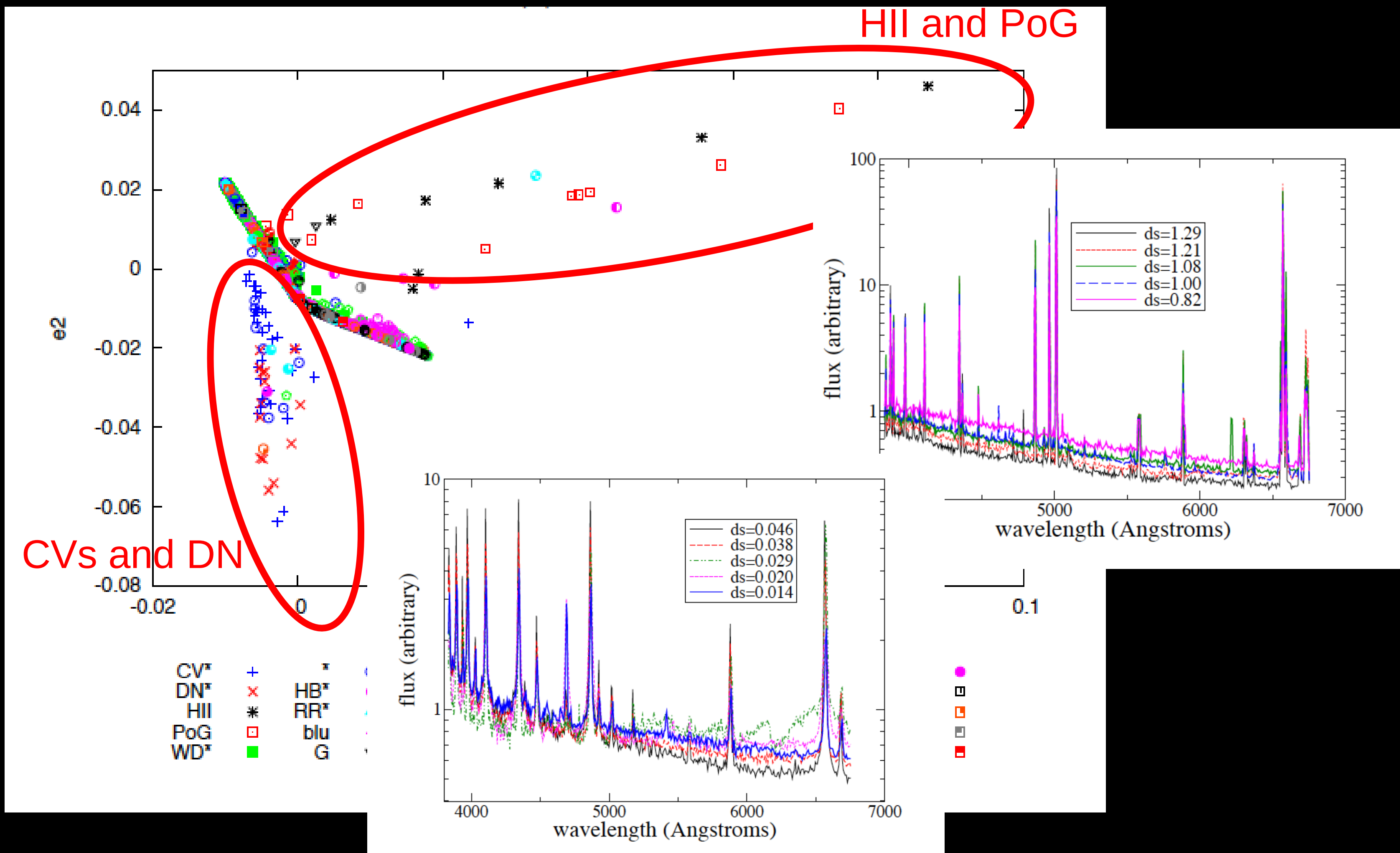




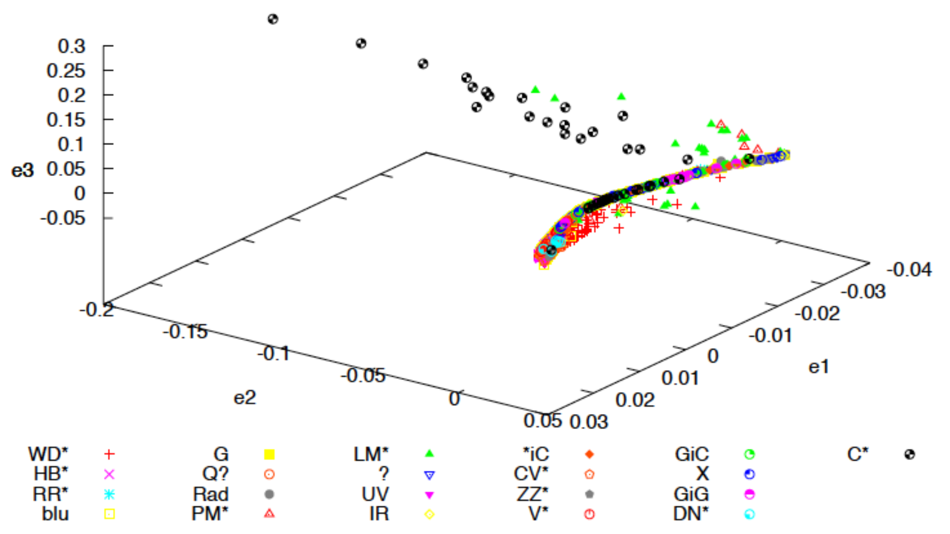
Krughoff et al 2011

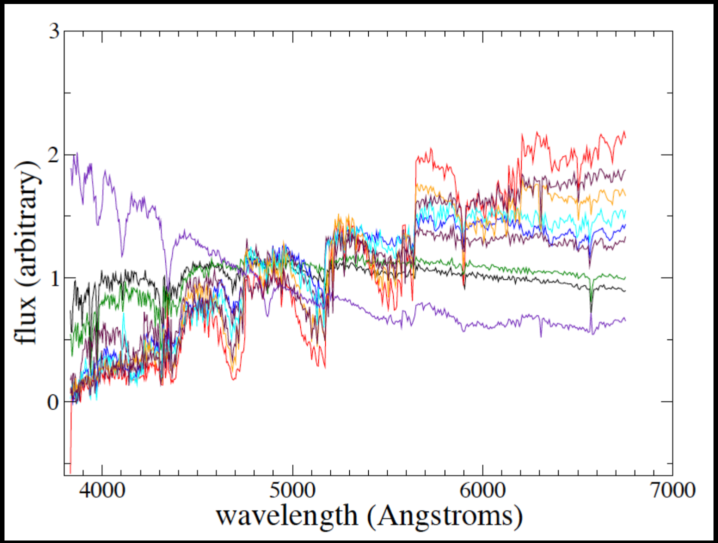
Nugent et al 1994

Case Study: How to find anomalies when we don’t have a model for them



Anomaly discovery from a progressive refinement of the subspace

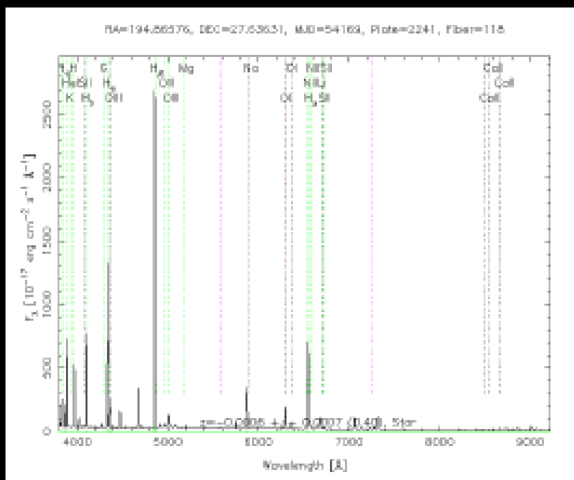




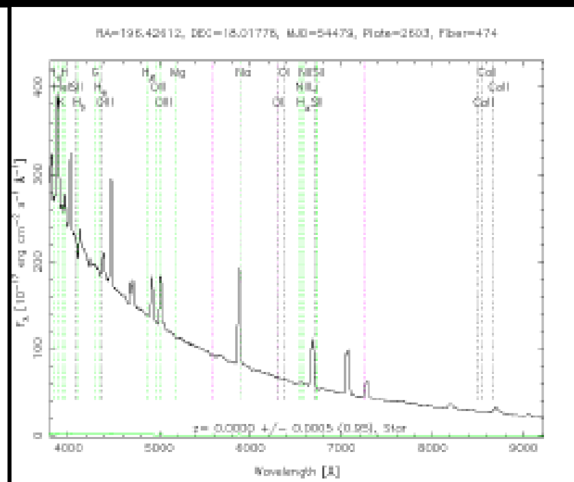
Outliers impact the local subspace determination (dependent on number on nearest neighbors). Progressive pruning identifies new components (e.g. Carbon stars).

Need to decouple anomalies from overall subspace

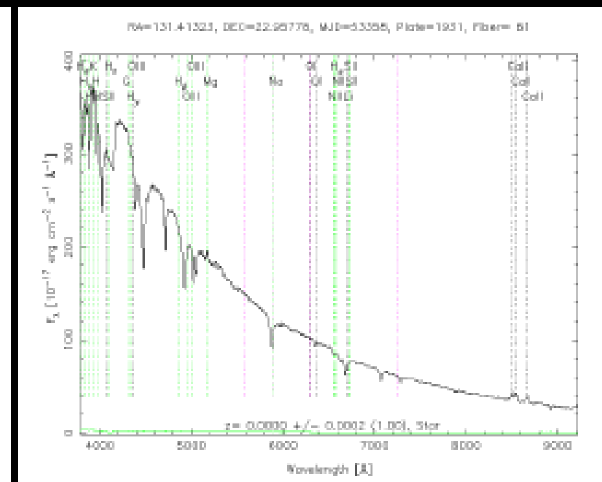
Anomalies within the SDSS spectral data



PN G049.3+88.1
Ranked first
Expect 1-3 PNE
Found 2



CV-AM
2 orbiting WDs
Ranked top 10



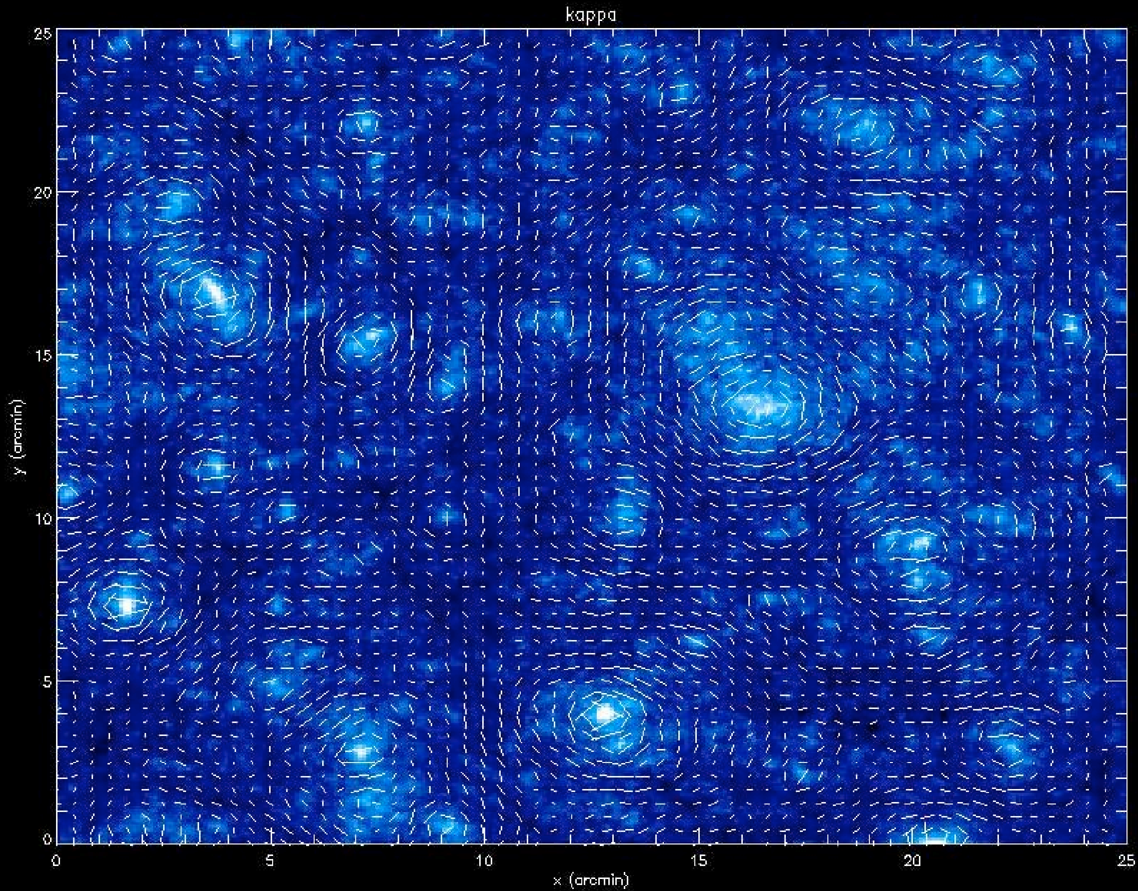
WD with debris disk
Ranked top 30
Only 3 known in SDSS

Xiong et al 2011

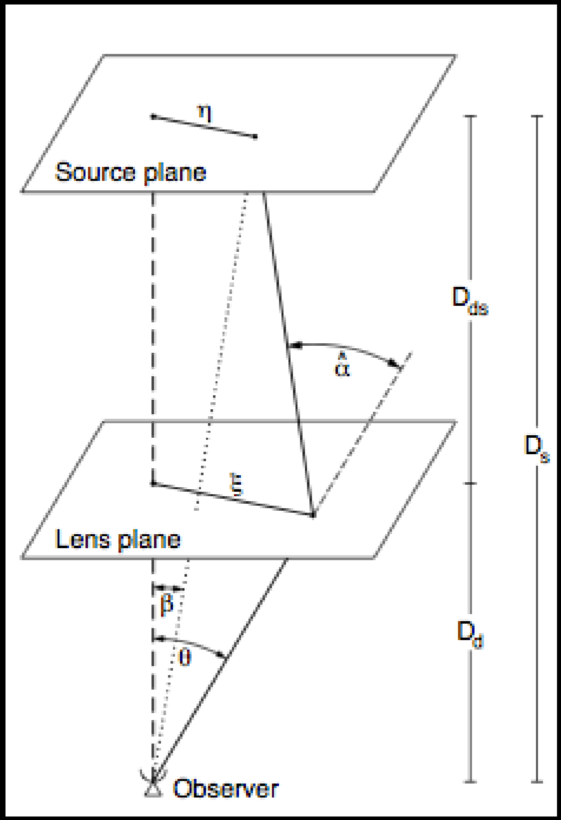
Expert user tagging (http://autonlab.org/sdss)

Xiong et al 2011

Case Study: From high dimension to low signal-to-noise



Jain, Seljak, White



Bartelmann and Schneider

Case Study: How to develop scalable algorithms?

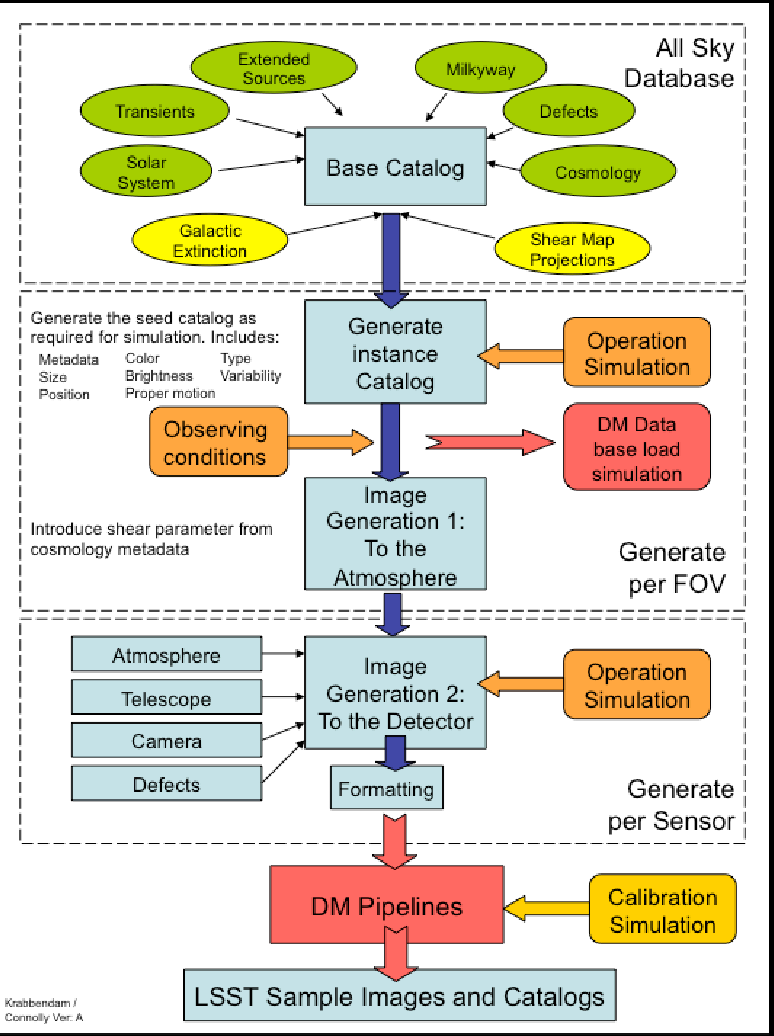
New philosophy of development through high fidelity simulations

Components:

- Survey strategy
- Source catalogs
- Images
- Processing
- End-to-end processing

Algorithms:

- Source detection and image subtraction
- Classification
- Linkage of moving sources
- Scalability



Broad range of astronomical sources

Galaxies

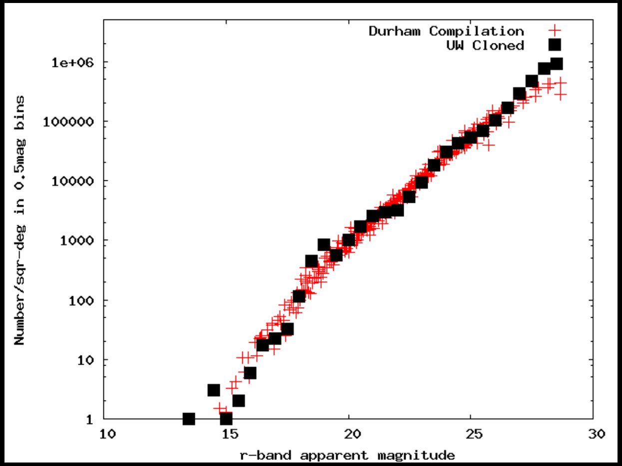
Cosmology from n-body simulations
10^6 sources/ sq deg (r<28)
Morphology, AGN, lenses, variability

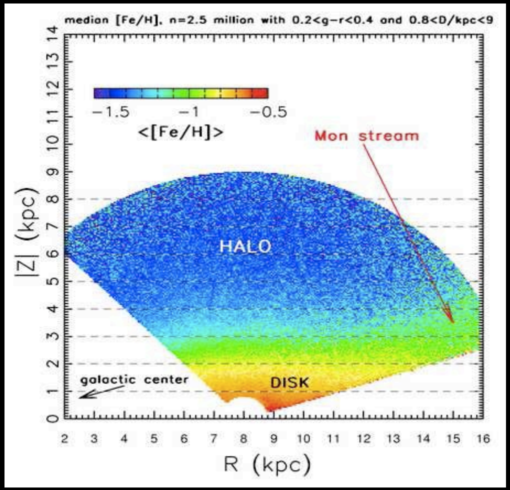
Stars

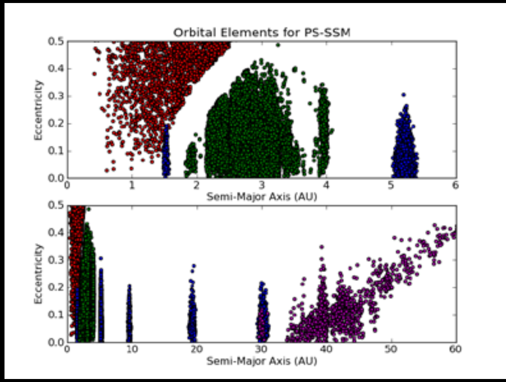
Galactic structure model
Main sequence, giants, dwarfs
Cepheids, flare stars, micro-lensing
Proper motion, parallax, differential effects

Asteroids

Solar system model
10 million main belt
KBO, TNO, Trojans….







Simulating the flow of photons through the atmosphere

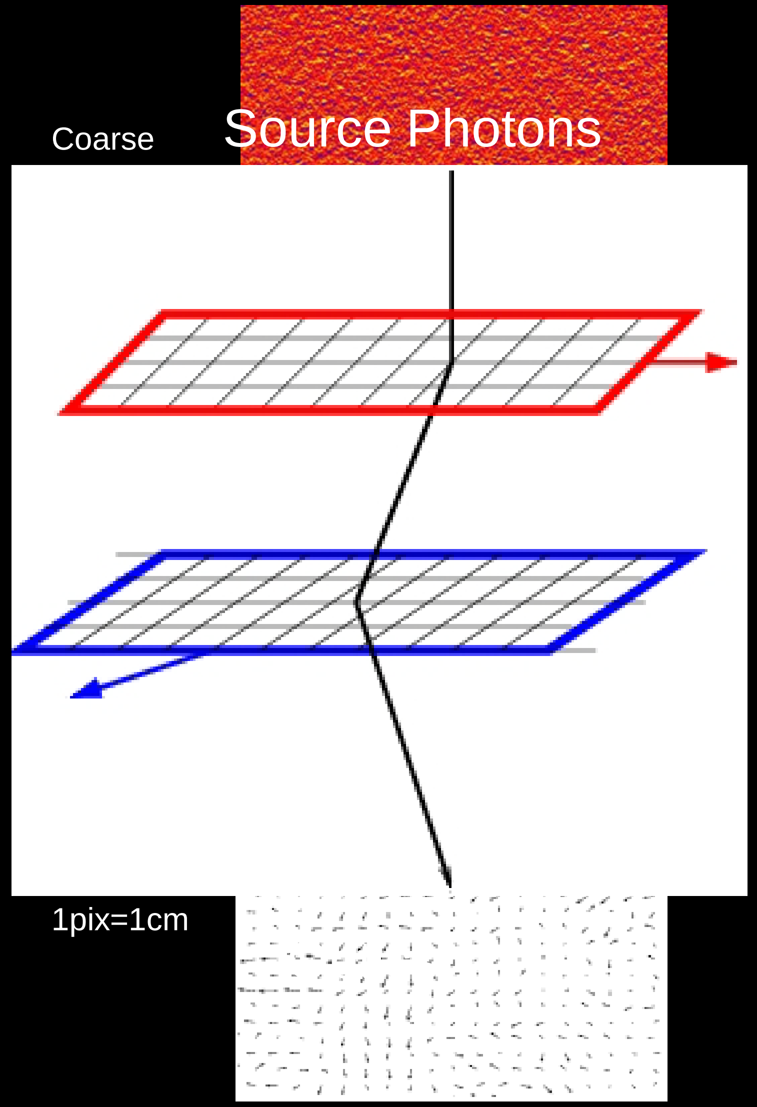
Parameterized a view above the atmosphere

Turbulent atmosphere

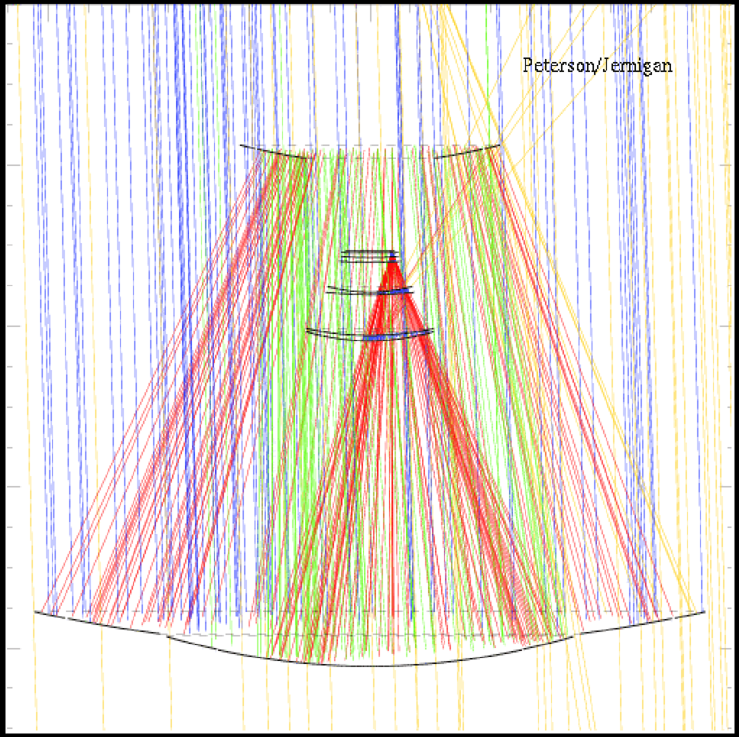
Frozen screens (six layers)
Based on observations

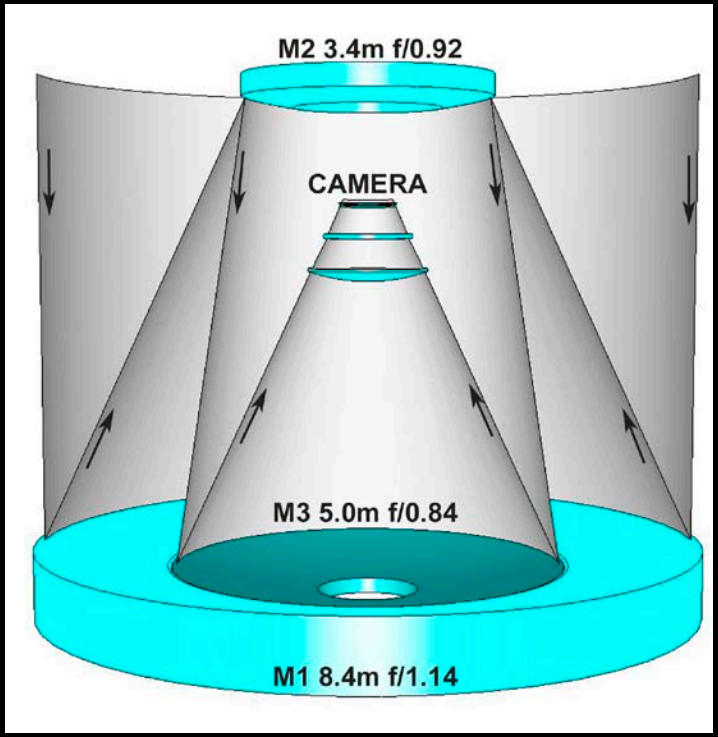
Wavelength dependent

Refraction, Cloud, Scattering



The impact of optics

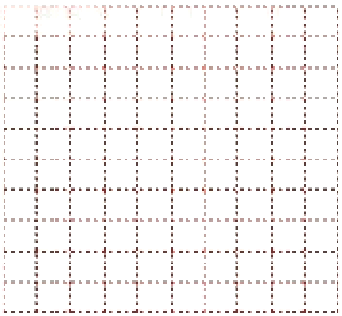




Telescope model

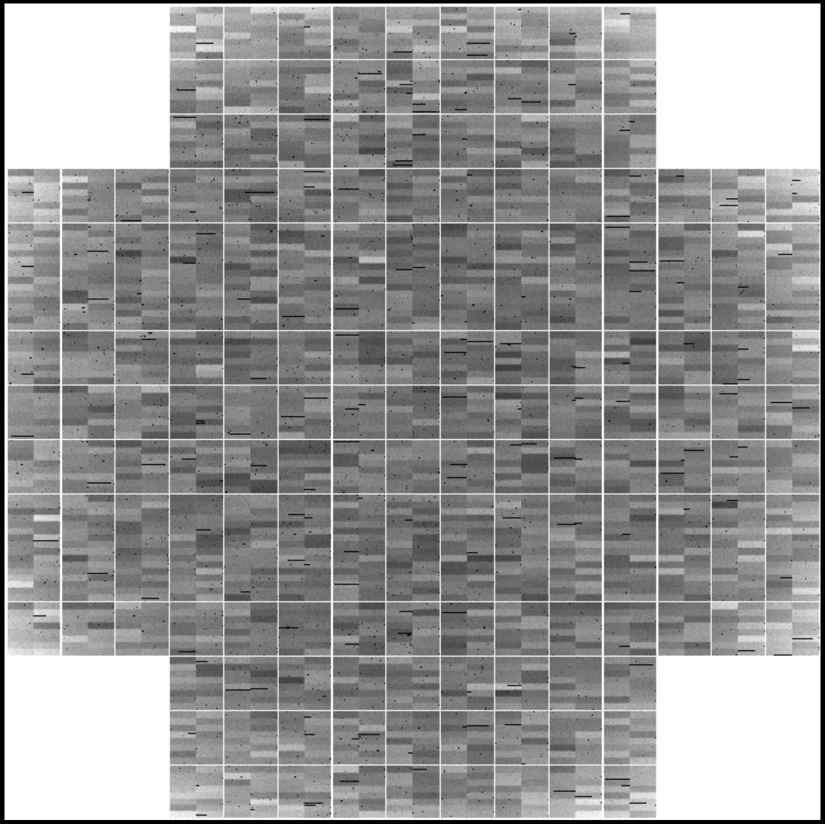
Three mirror modified Paul-Baker design
Fast ray-trace algorithm
Perturb the surfaces (1300) to determine the impact of control system
Conversion of photons to electrons

Following the photon flow…




Optics
+Tracking
+Diffraction
+Detector Misalignments & Perturbations
+Lens Misalignments
+Mirror Misalignments Perturbations, & Micro-roughness
+Detector
+High Altitude Atmosphere
+Mid Altitude Atmosphere
+Low Altitude Atmosphere
+Pixelization
+Saturation & Blooming

The full system



189 CCDs
16 amplifiers per CCD
10^9 photons

Science at the scale of the LSST
With the same cadence and similar systematics
Catalogs, images and scalable science

How do we make the new generation science happen?

Science at the petascale still requires a scientist

- Broad range of abilities and requirements
- Mathematically sophisticated (but not necessarily computationally)
- Good at scripting (IDL, Python)
- Code is often throw away (but this is changing)
- Good at learning new approaches (e.g. SQL, AWS)
  - But needs to see fast returns if an early adopter
  - Community driven
  - Pretty tolerant…

Summary: how do we scale our science?

Collecting data is not the challenge

Storage is not an issue (other than cost)

Not just a question of more CPUs

Need new ways of understanding what information is contained within our data and how we can efficiently extract it

With thanks to:

Scott Daniel (Astro)
Chris Genovese (Statistics)
Garret Jernigan (Astro)
Simon Krughoff (Astro)
Rob Gibson (Astro)
Bhuvnesh Jain (Astro)
Mike Jarvis (Astro)
John Peterson (Astro)
Jeff Schneider (CS)
Ian Smith (Astro)
Liang Xiong (CS)
Jake VanderPlas (Astro)
Ching-Wa Yip (Astro)
LSST Collaboration