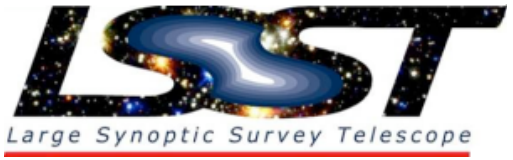


From Physics to Photons and Back:
Scalable Science in the LSST Era

Andrew Connolly

Department of Astronomy

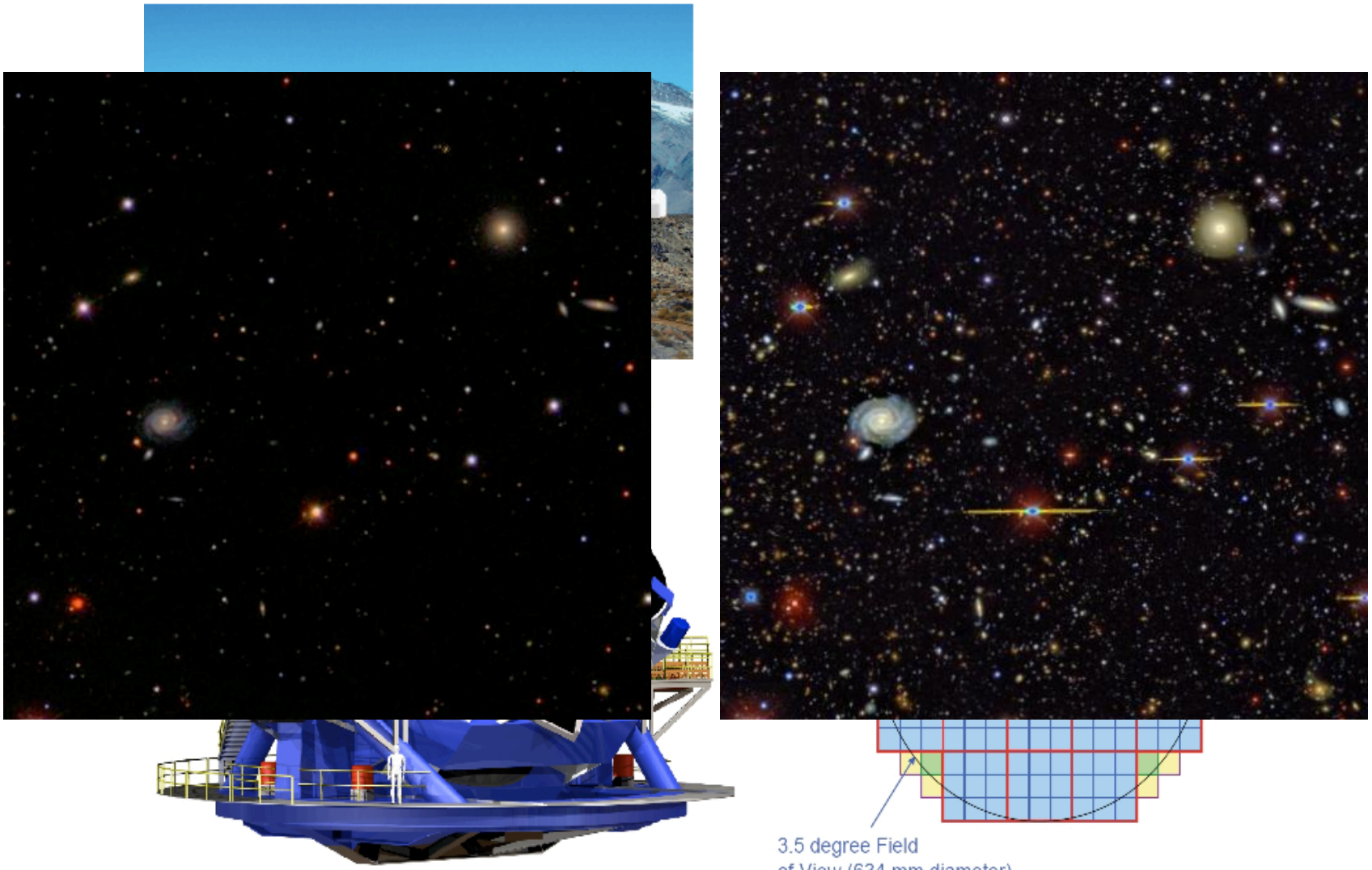
University of Washington

Collaborators: **Garret Jernigan**, **John Peterson**, Jeff Schneider, Andrew Moore, **Simon Krughoff**, Ryan Scranton, Ching-Wa Yip, Chris Genovese, **Jeremy Kubica**, Larry Wasserman, Andrew Moore, Dan Pelleg, Jeff Schneider, **Jake vanderPlas**

Looking for Needles

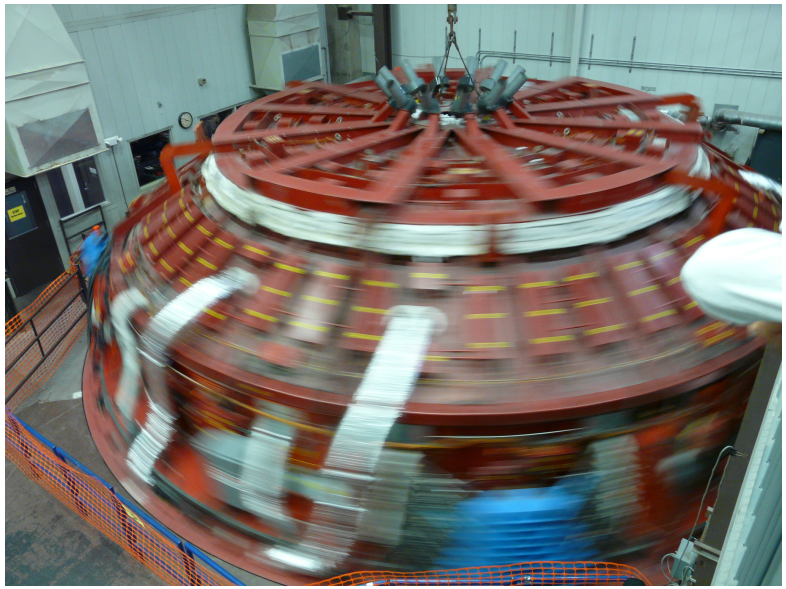
- **Outline**
  - Data flow from astronomy in the next decade
  - Simulating the sky at high fidelity
  - Data driven compression for astronomy
  - Anomalies in high dimensional space
  - It moves….
  - Scaling the science

The Large Synoptic Survey Telescope



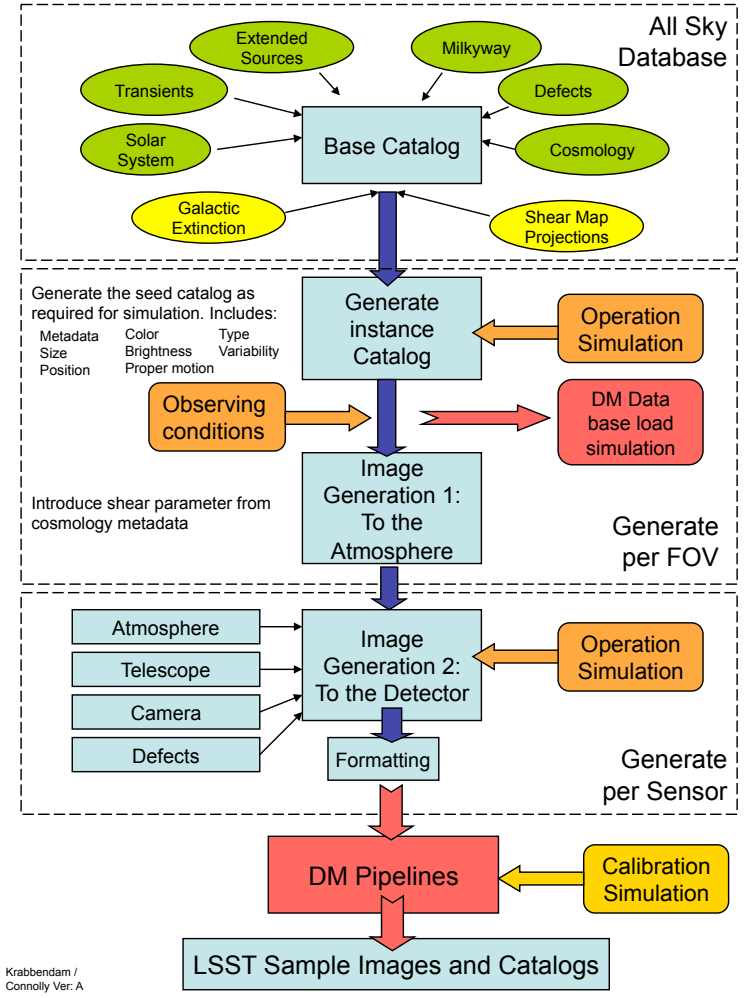
Streaming the Sky

- **Survey Characteristics**
  - 20,000+ degrees^2
  - 2000 visits per field
  - 0.2 arcsec/pixel
  - 320–1050 nm
  - 25 mag (AB) per visit
  - 27.7 mag (AB) total depth
  - 20-40 TB per night
- **Broad science goals**
  - Weak lensing
  - Supernovae (100K/yr)
  - Asteroids (10^6)
  - Variable sky
  - Dark energy
  - Dark matter
  - …



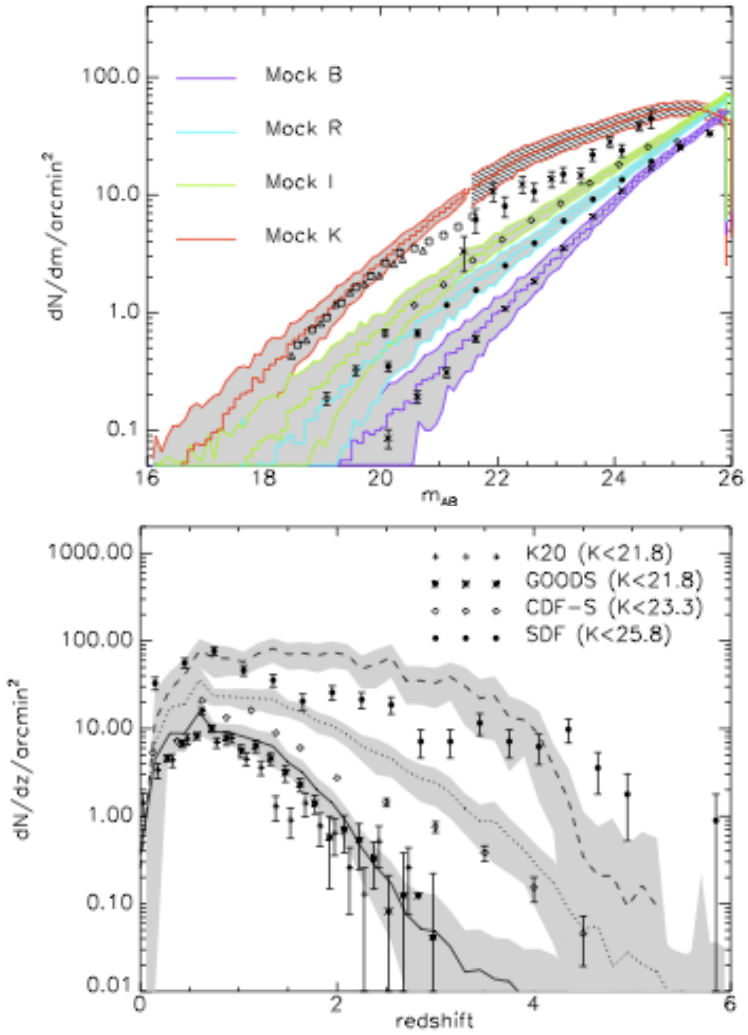
Simulating a Petabyte Data Stream

- **High fidelity simulations**
  - **Test algorithms and science**
    - Data management design
    - Algorithm optimization
    - Systematic limitations
  - **Atomic representation of cosmology**
    - Input cosmology
    - Milky-way model
    - Extinction, shear
  - **Base catalog database**
    - Variable sources
    - AGN
    - Moving sources



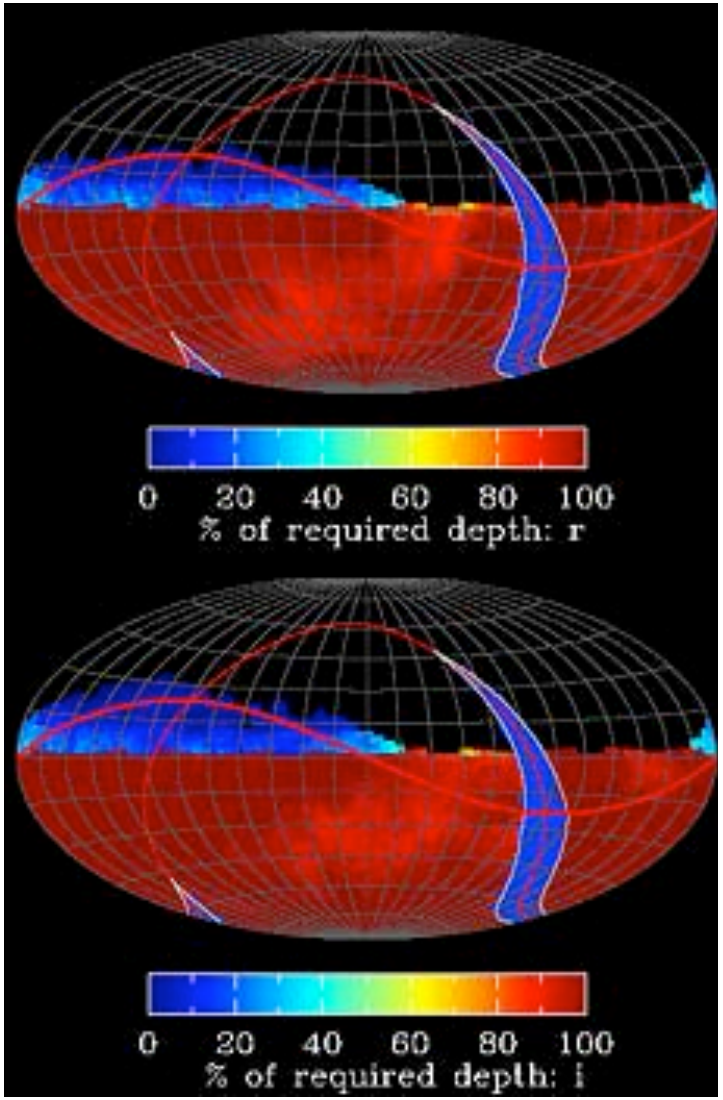
Current models

- **Millennium Simulations**
  - **Kitzbichler and White (2006)**
    - 6 fields, 1.4x1.4 deg per field
    - $6x10^6$ source per catalog
    - Based on Croton et al (2006) and De Lucia and Blaizot (2006) models
    - r<26 magnitude limit
    - z>4 redshift limit
    - BVRIK Johnson and griz SDSS
    - Extended to fit LSST u,g,r,i,z,y
    - Derived SED for all sources



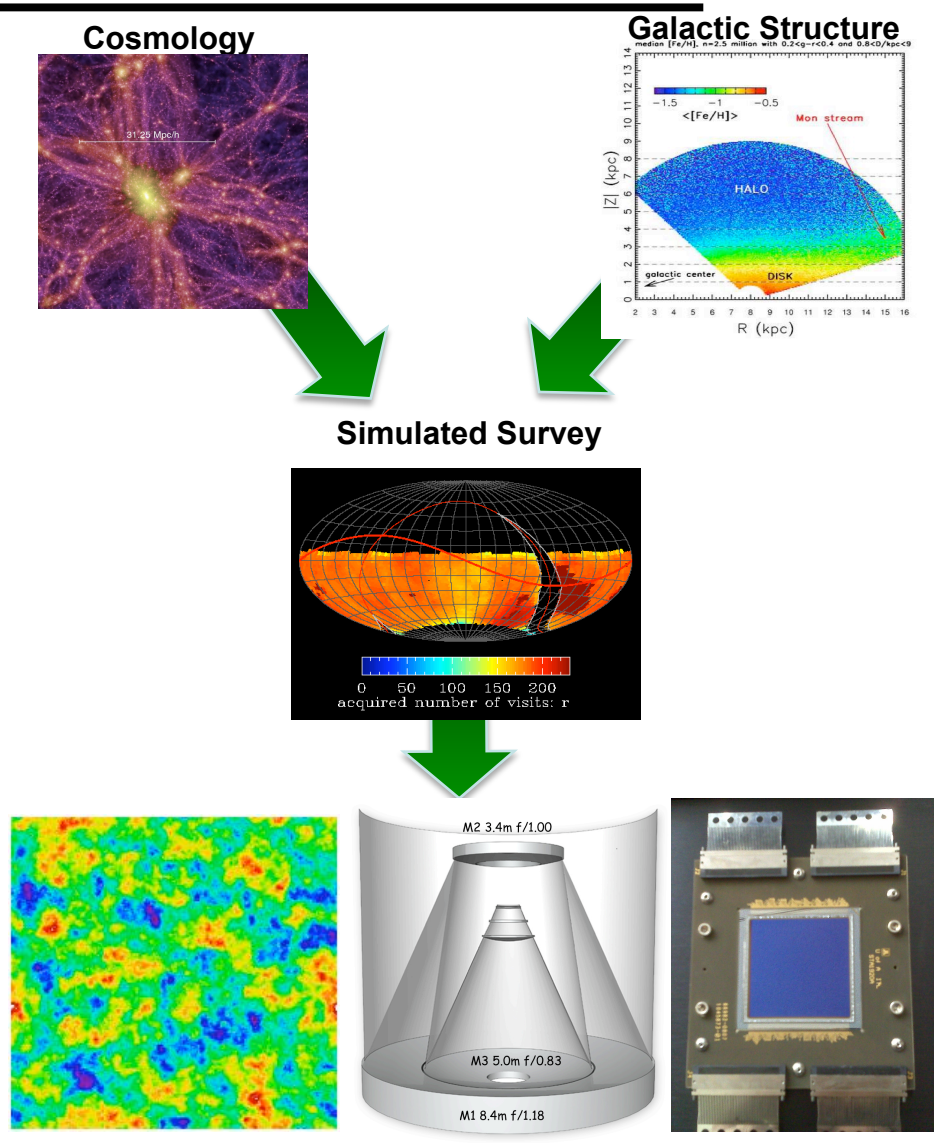
“Observing” the LSST Simulation

- **An instance catalog**
  - **Sampling the base catalog**
  - **Input from Operational Simulator**
    - Position, atmosphere, time, filter
    - Cloud models for Cerro Pachon
    - Airmass and sky backgrounds
    - Sample light curves for variables
  - **Derived catalogs of sources**
    - Calibration samples
    - Large areal coverage
    - Input to image simulations
    - r<28



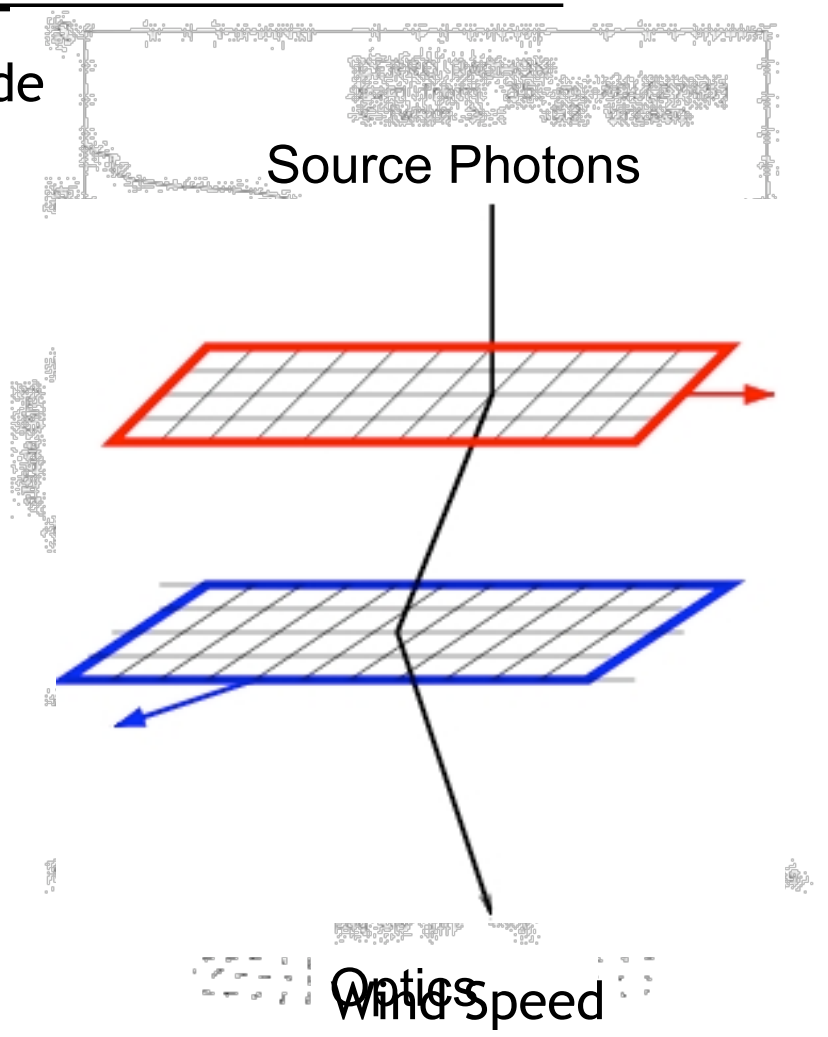
Streaming the sky

- **Simulating the Sky**
  - **High fidelity simulations**
  - **In each focal plane**
    - 40 million stars and galaxies
    - 189 CCDs
    - 3.2 billion pixels
    - 10^{11} photons
    - 12.8 GB image
    - 15s of LSST data
  - **Ray trace every photon**
  - **Multilayer atmosphere**
  - **Multi-component telescope**
  - **Conversion of photons to e^-**
  - **2000 CPU hrs per focal plane**
  - **Run on 100s/1000s CPUs**



Atmosphere Models & turbulence

- **Turbulent screens**
  - **Data from Cerro Pachon**
  - **Six layers (0.1 - 16 km)**
  - **Correlated wind vectors**
  - **Photons traced and shifted**
  - **Vector Screen:**
    - 2048 squared
    - 0.1m/pixel
    - 0.1 to 30m
  - **Mie Scattering**
  - **Atmospheric transmission and dispersion**

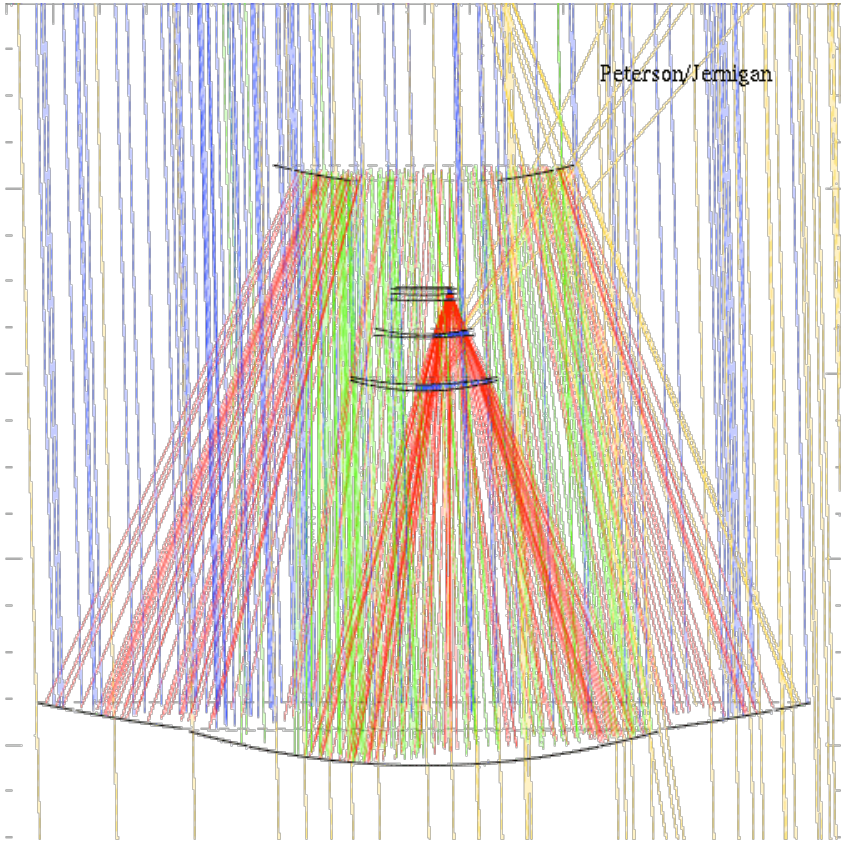


Jernigan and Peterson

Vernin et al., Gemini RPT-A0-G0094

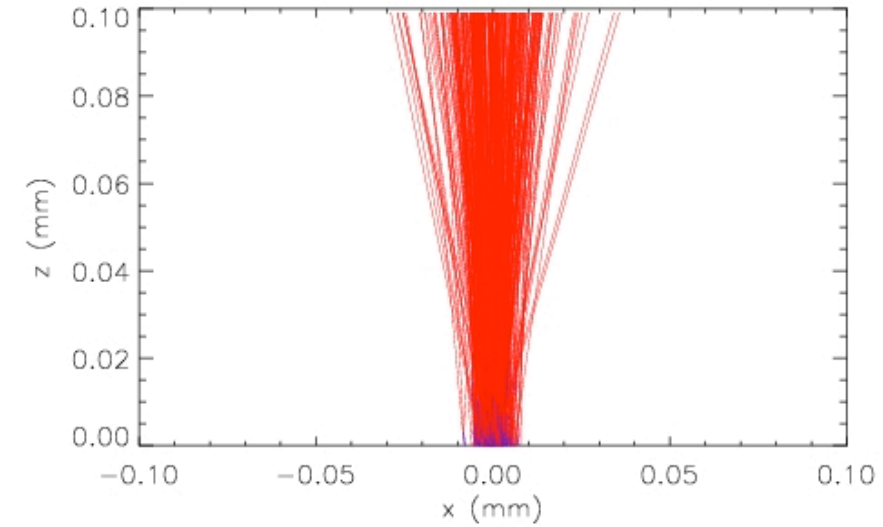
Telescope Optics

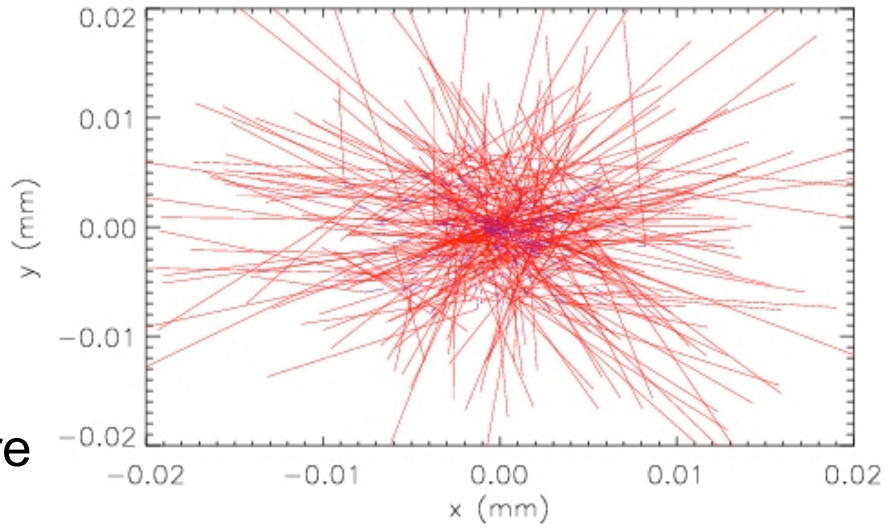
- **Telescope model**
  - LSST 3 mirror design
  - Fast ray reflection and refraction algorithms
  - Diffraction from spider
  - Wavelength-dependent effects
  - Trace beam at appropriate angle through the filter
  - Can do sequential ray-trace



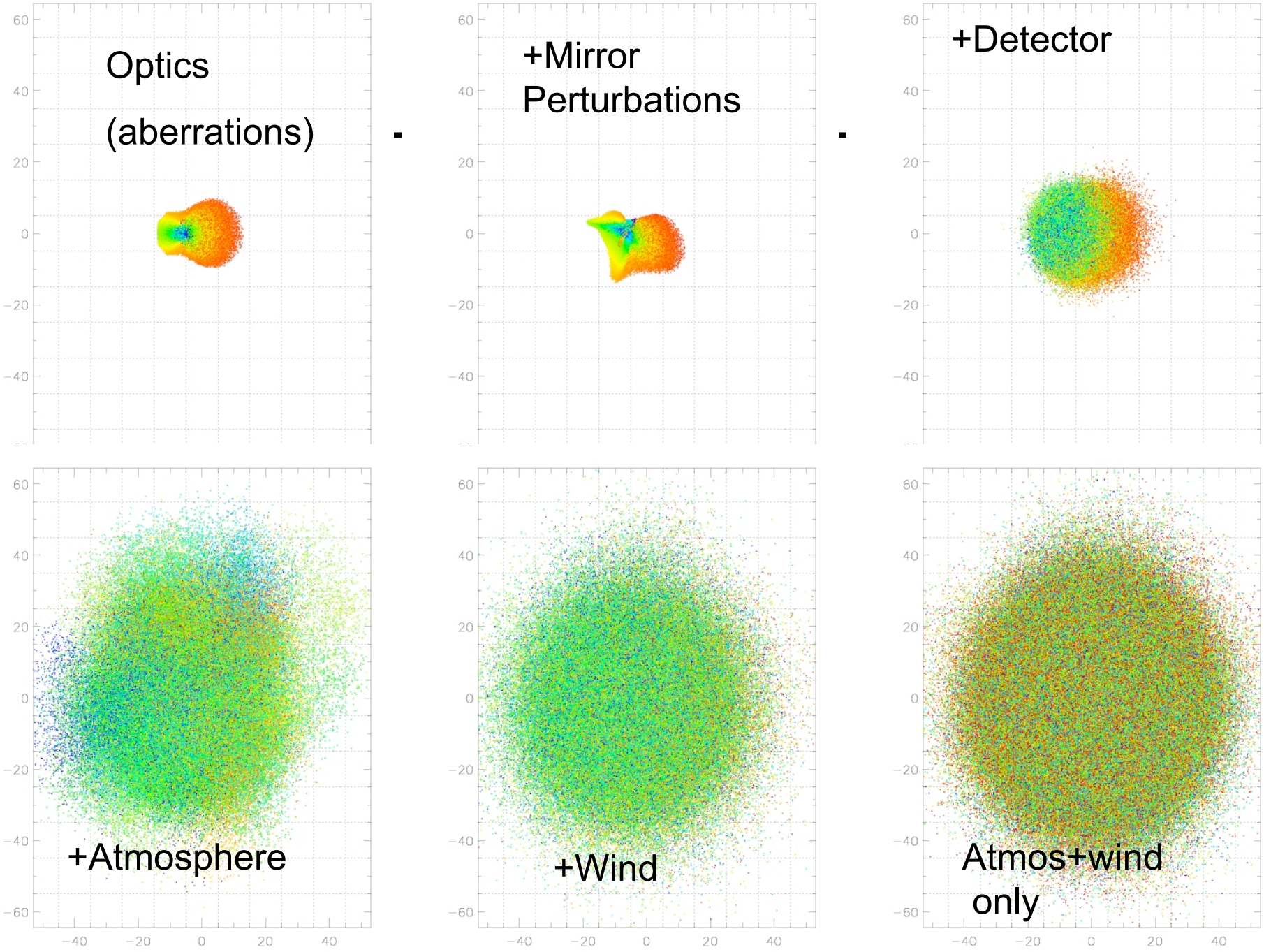
Camera and Detector model

- **Focal plane model**
  - **Modeling for 189 CCDs in focal plane**
  - **Can incorporate chip tilts, heights, structure, pixel -to-pixel effects**
- **Detector model**
  - **Refraction for light entering the Si surface**
  - **Photon interaction (λ and T dependent)**
  - **Lateral charge diffusion**



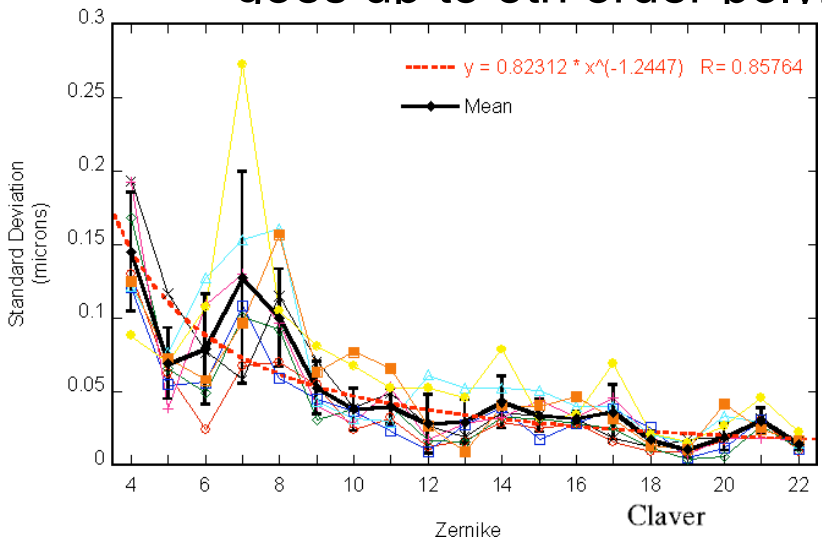


Rasmussen and Gilmore

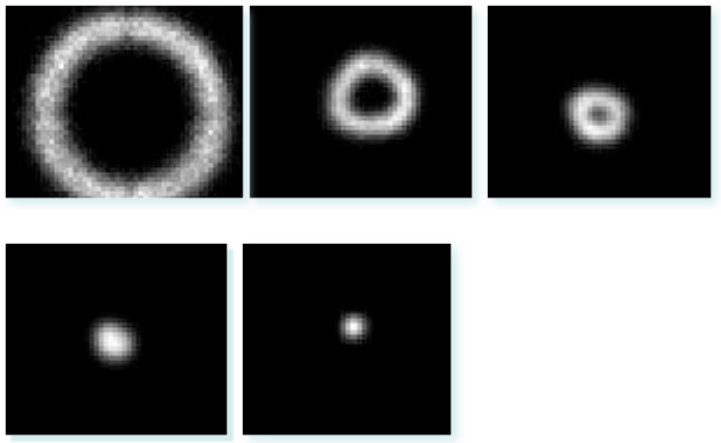

Optics
(aberrations)
-
+Mirror
Perturbations
-
+Detector
+Atmosphere
+Wind
Atmos+wind
only

Examples: Thermal and mechanical distortions

- **Simulating perturbations**
  - **Each optic has 6 dof (decenter, defocus, three euler angles)**
    - Perturbations are placed on the three mirrors using a Zernike expansion to simulate the possible residual control system errors each mirror can have an arbitrary amplitude code goes up to 5th order polynomials

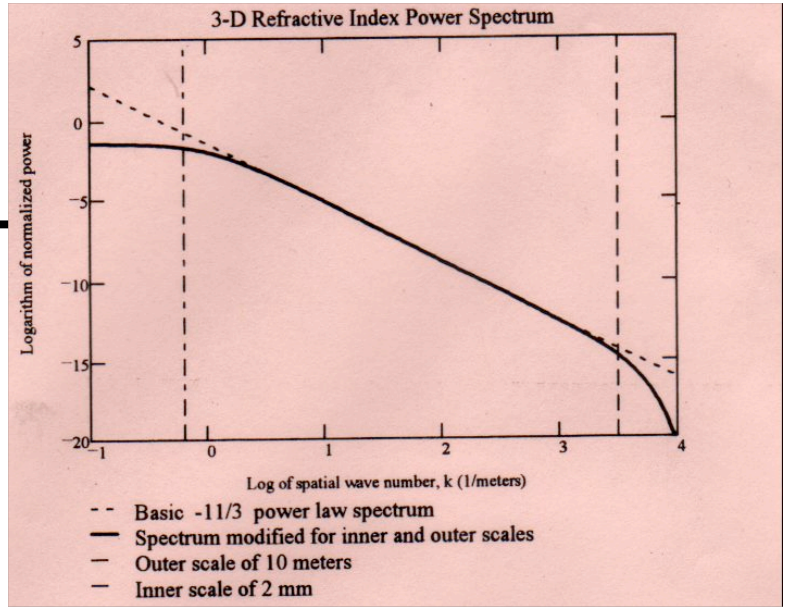


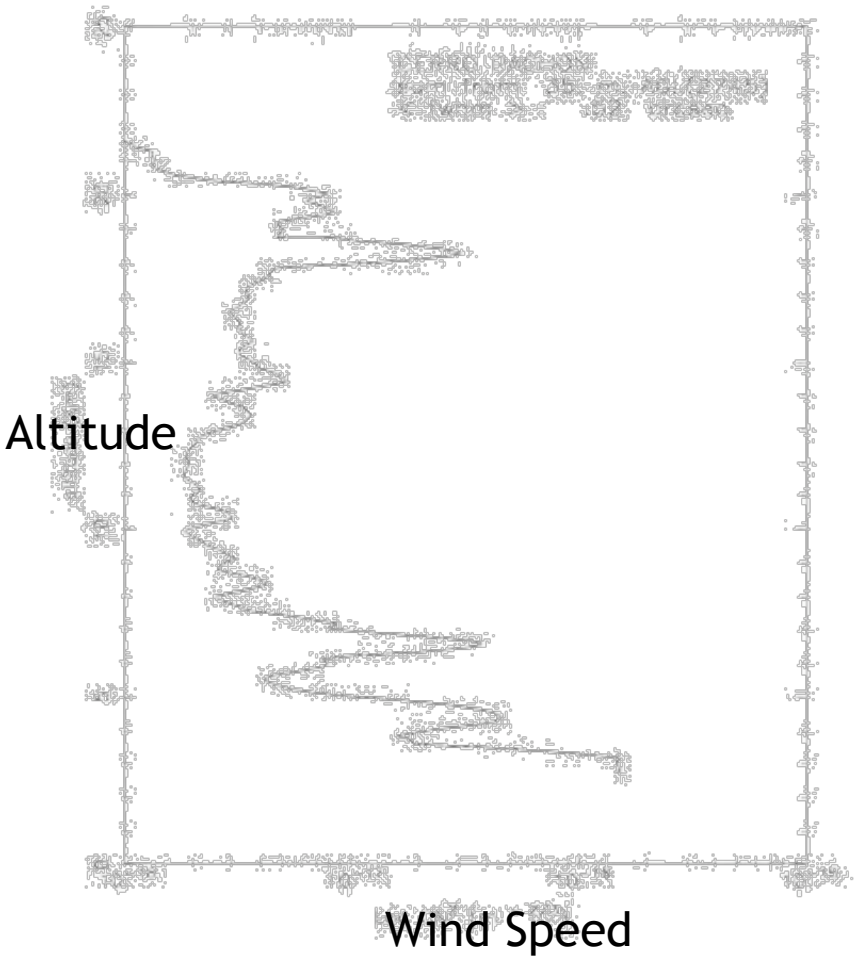
Perturbation spectrum
From Claver



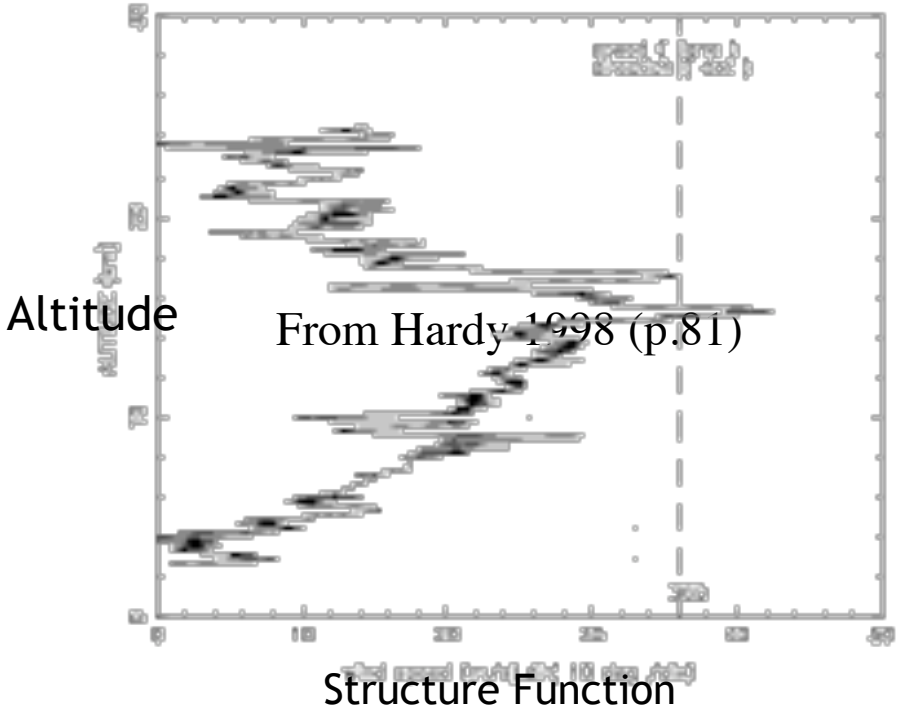
e.g. Mirror Defocus

Atmosphere Models & Kolmogorov turbulence

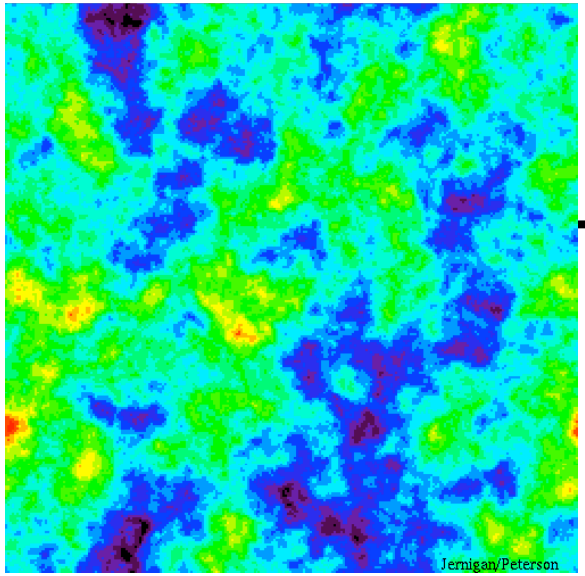




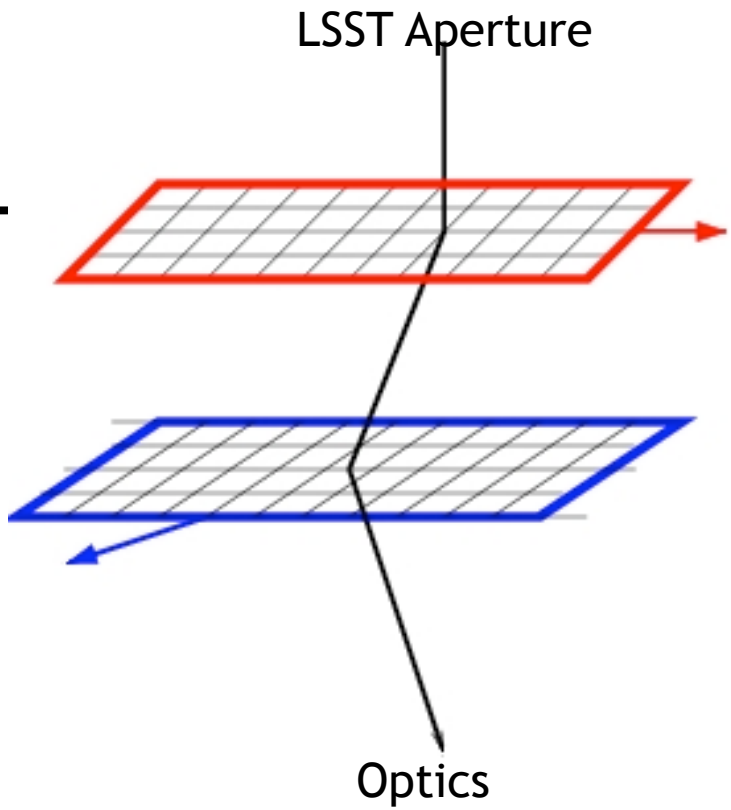
Vernin et al., Gemini RPT-A0-G0094

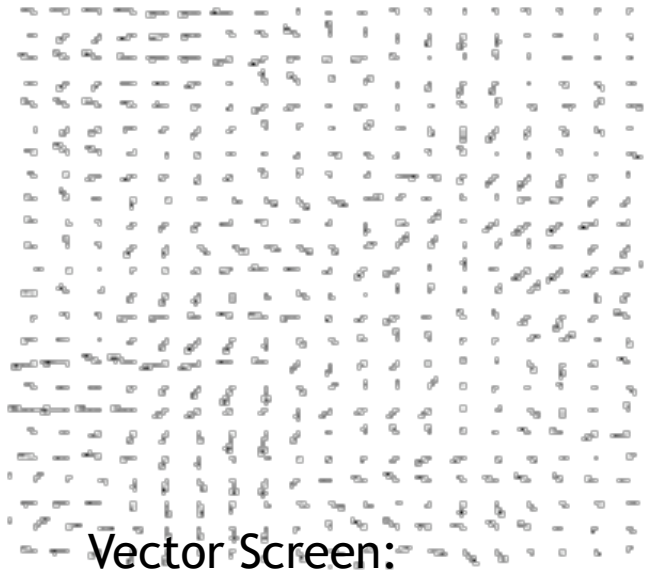


From Hardy 1998 (p.81)



Atmospheric Screens and Raytrace

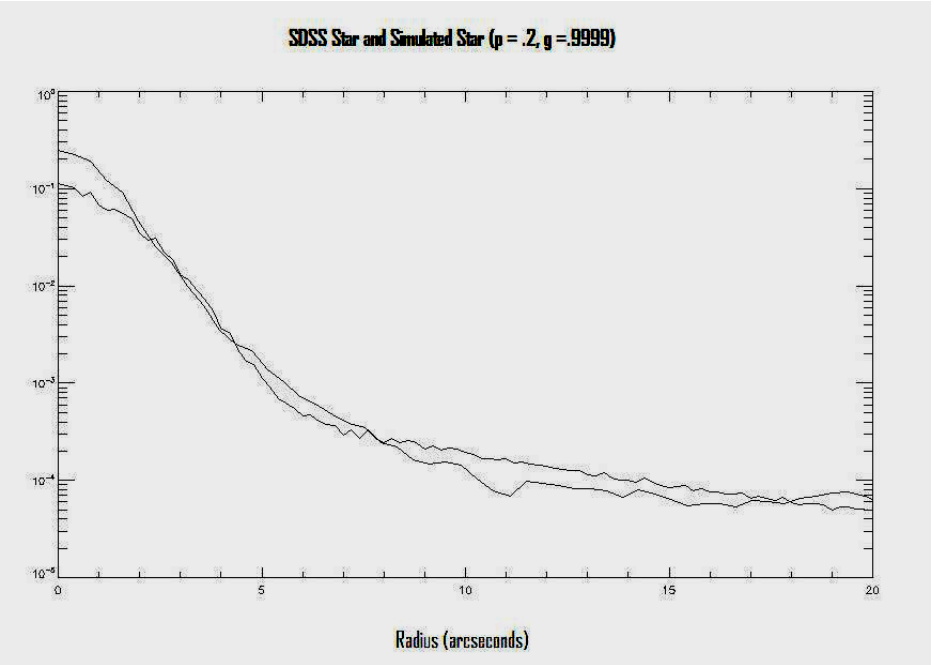




Vector Screen:
2048 squared
0.1m/pixel

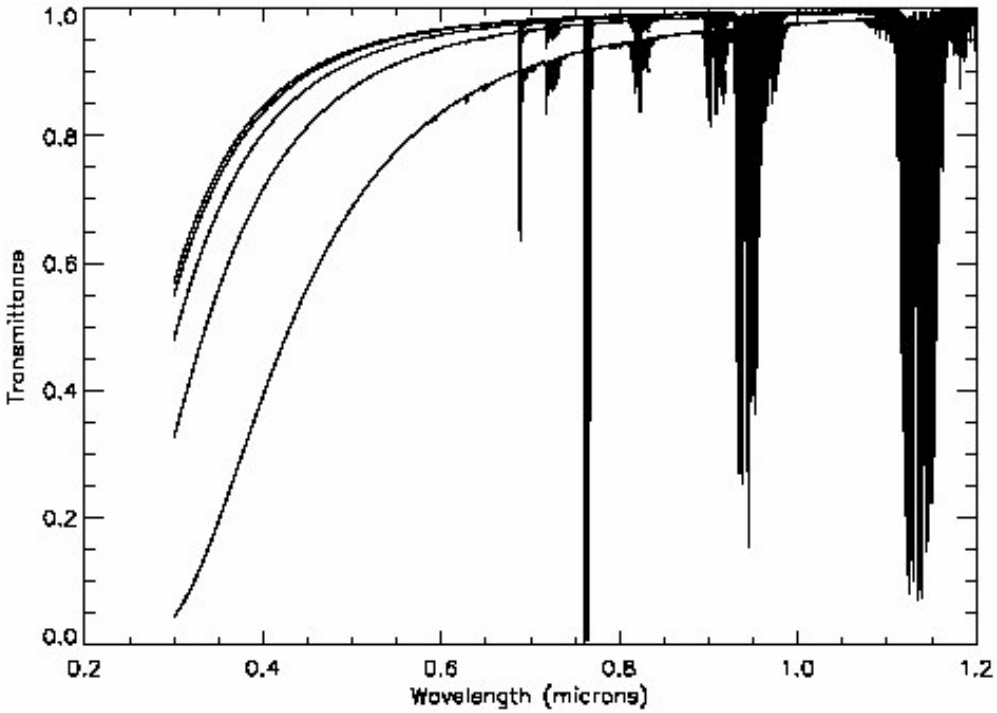
DC 3 Parameters:
(6 layers x 6 = 36 parameters)
6 layers (enough to range of corr. scales) at height~0.1,1,2,4,8,16 km
Total seeing specified by DM
Effective inner cutoff ~ 0.1 m
Outer scale ~ 30 m
Wind speed: 0-20 m/s uniform dist.
Wind dir: Random but each layer uniform dist +/- 20 deg.
Relative seeing in each layer:
uniform number*exp(-h/(7km))

Additional Atmospheric ~~Physics:~~



Mie Scattering:

Based on Henyey-Greenstein model; verified with SDSS



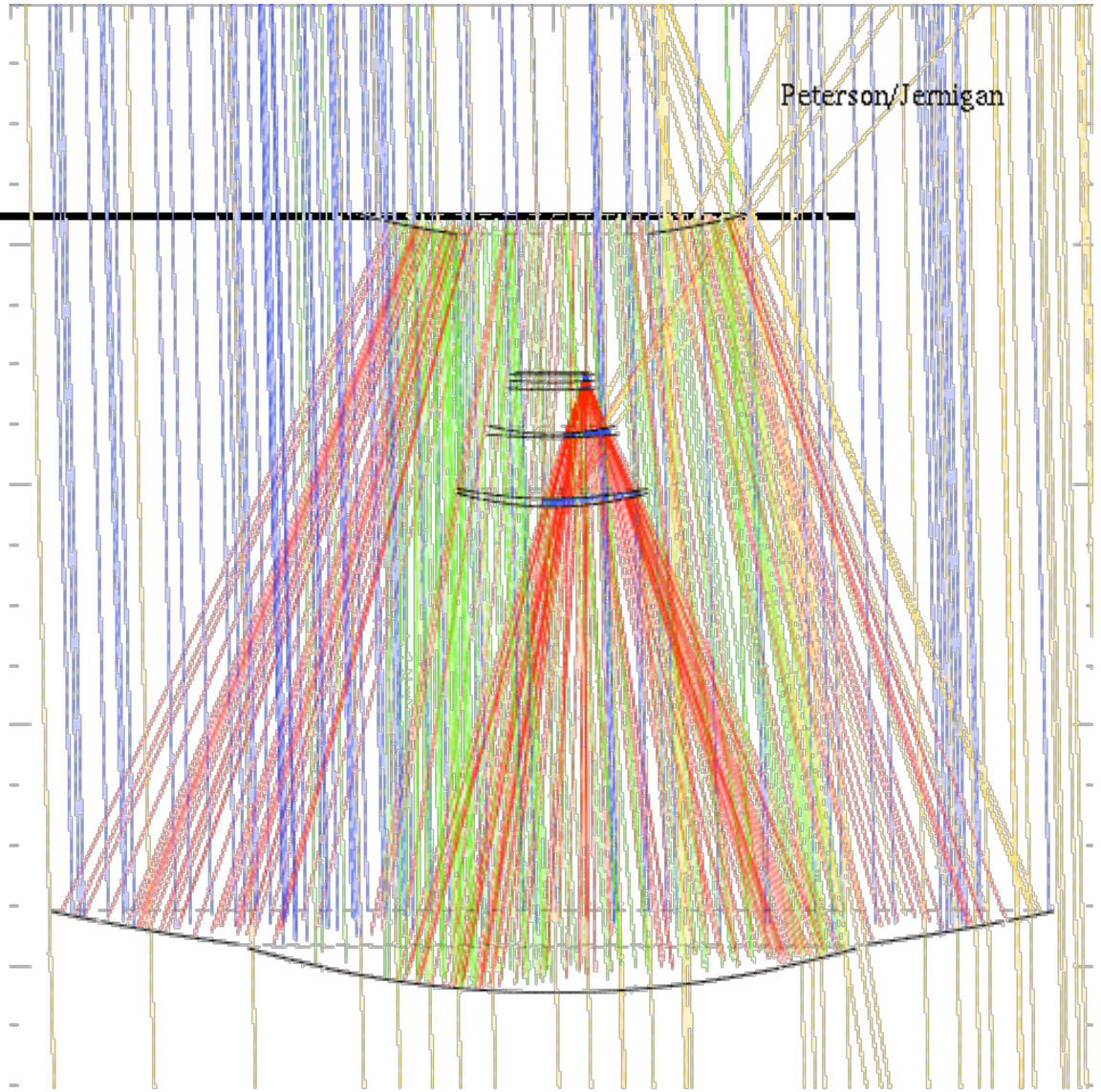
Atmospheric Transmission:

Zenith-dependent transmission

Atmospheric Dispersion:

Standard literature formulae

Unperturbed optics



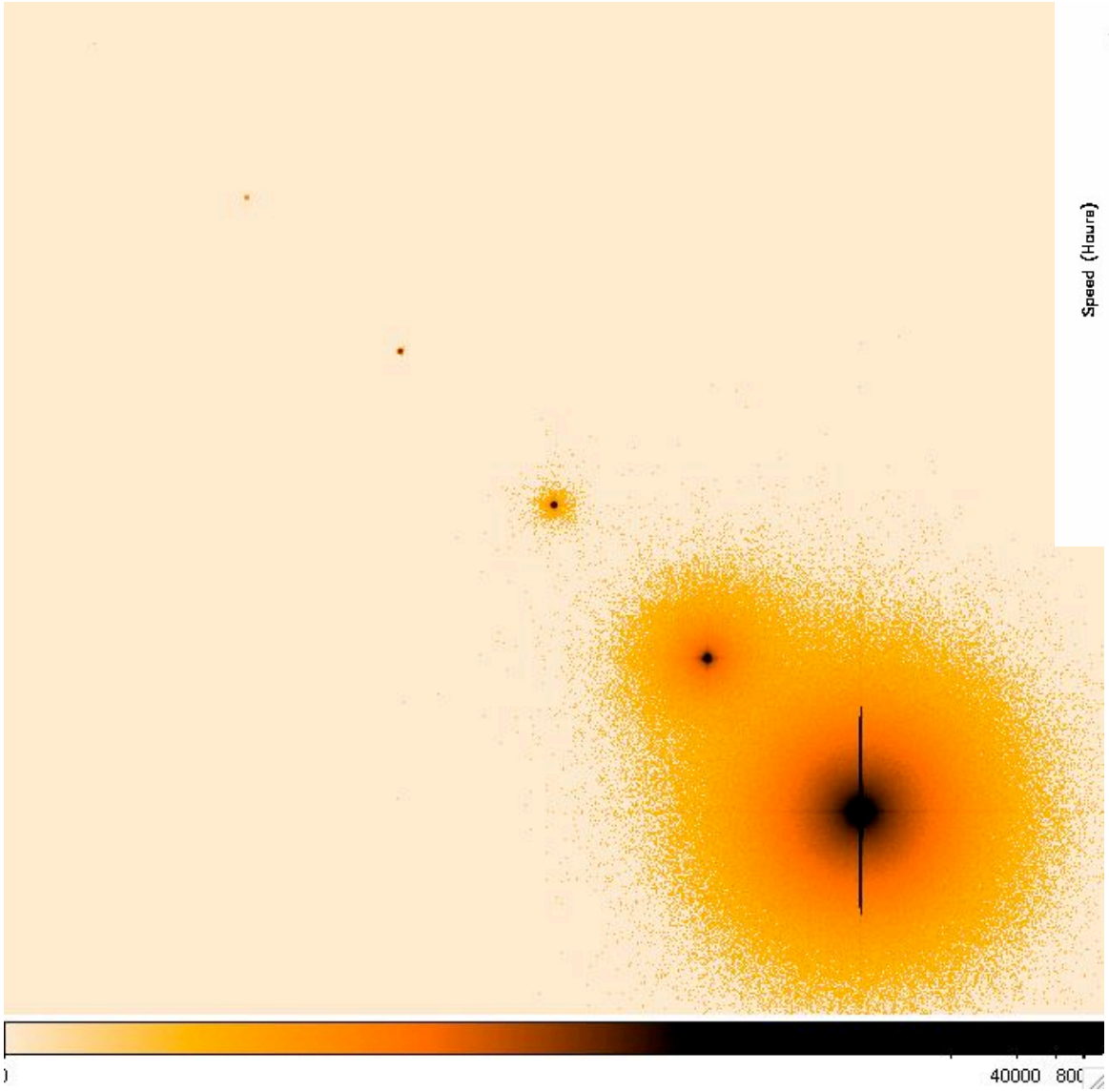
Fast ray trace
Calculates ray intercepts
Uses fast reflection
and refraction algorithms
Wavelength-dep index of
refractions
Carefully compared with Zemax

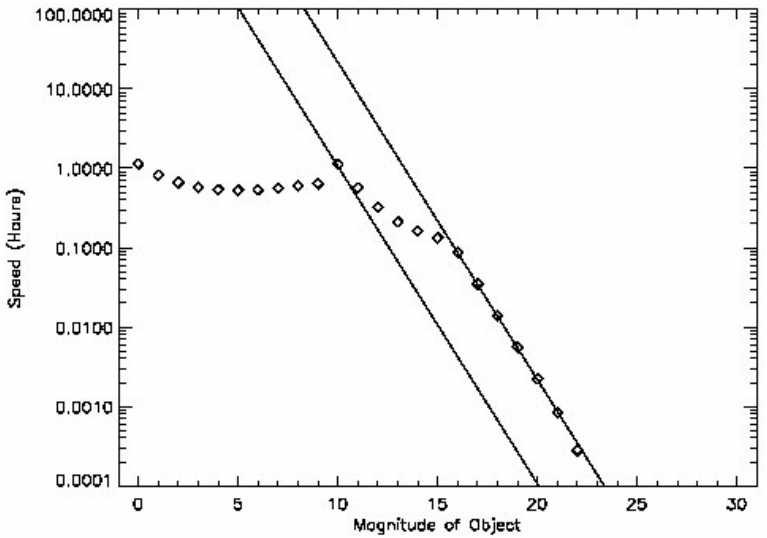
Optics Design: based on
collection 457 (all filter
configurations)

Filter: Monte Carlo use of
transmission curves
based on on-axis shift of curves
from document 1089

Diffraction: based on
spider in collection 457
$\tan \alpha \sim \lambda/(4\pi \min(\text{dist in pupil}))$

Has non-sequential raytrace switch:
Usually turn this off





Bright object effects:

Saturation: stop at 100,000 e^- and bloom

Blooming: shift up or down to halfway point (2048)

Diffraction: rotation of spider w.r.t camera

Adaptive trick: “learns” where photons are going to saturate and distributes them probablistically (factor of 20 speed up for 10-17 mag)

Cheat at billion photons: put down more than 1 photon at a time

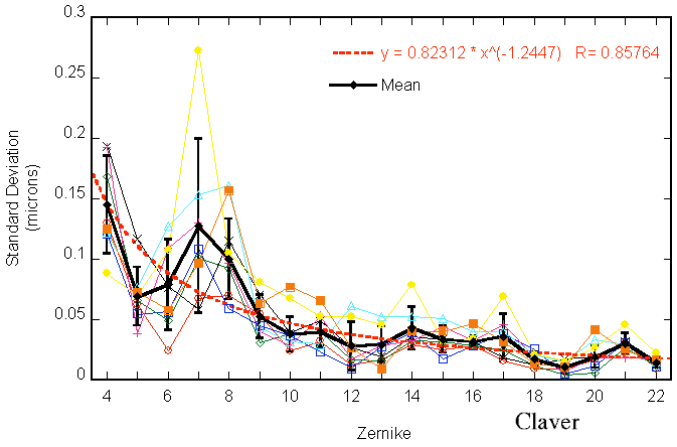
For DC 3 run:

Measured response to each parameter

Simplified the list

Found overall scale so PSF_{optics}=0.25” FWHM using gaussian distribution of pars

Assumption is whatever you see in the PSF, an ideal control system will be able to control



Perturbation spectrum from Claver

	Angles	Translation
Mirrors	1”, 3”, 2”	5,5,10 μm
Lens/Filter	20”	80 μm
Detector	20”	5 μm
2nd order & higher perturbations	0.3, 0.3,0.5 μm * (zernike order)$^{-1.2}$	

Will need to add dynamics:
could be simple to adjust few pars for guiding & refocus
or could be a nightly dynamic model

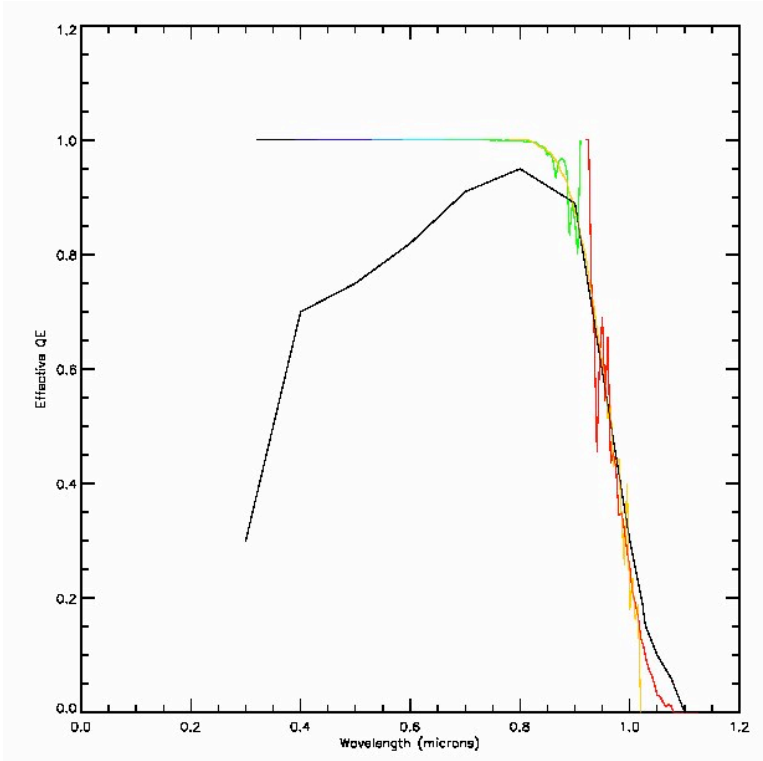
Detector

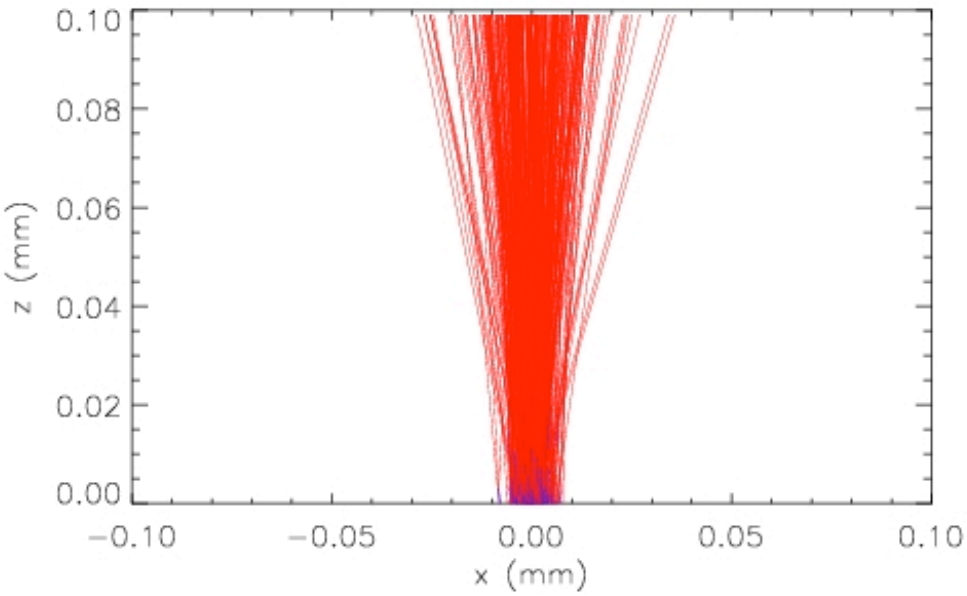
Refraction for light entering the Si surface

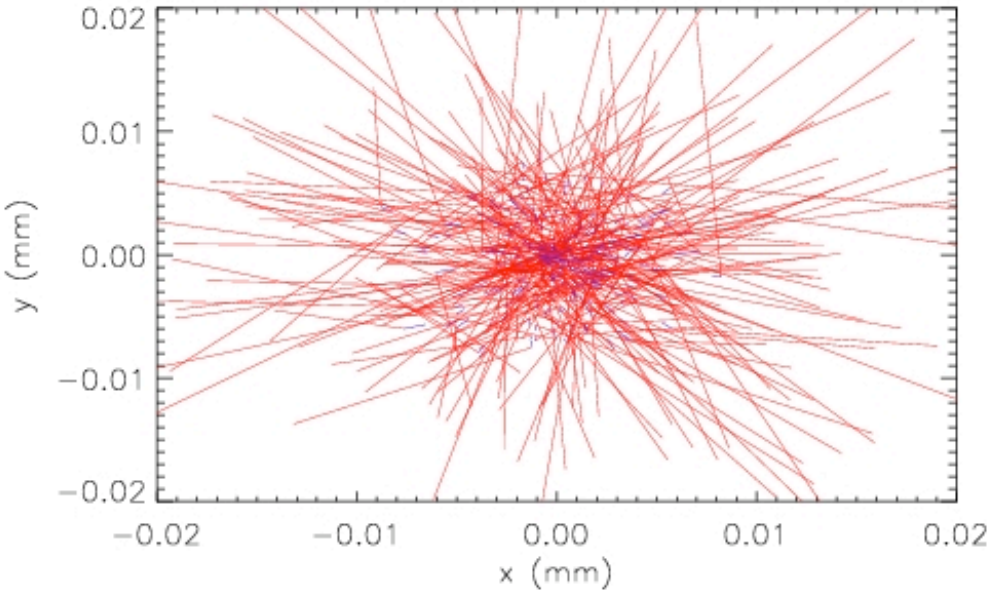
Photon interaction (wavelength and temperature dependent)

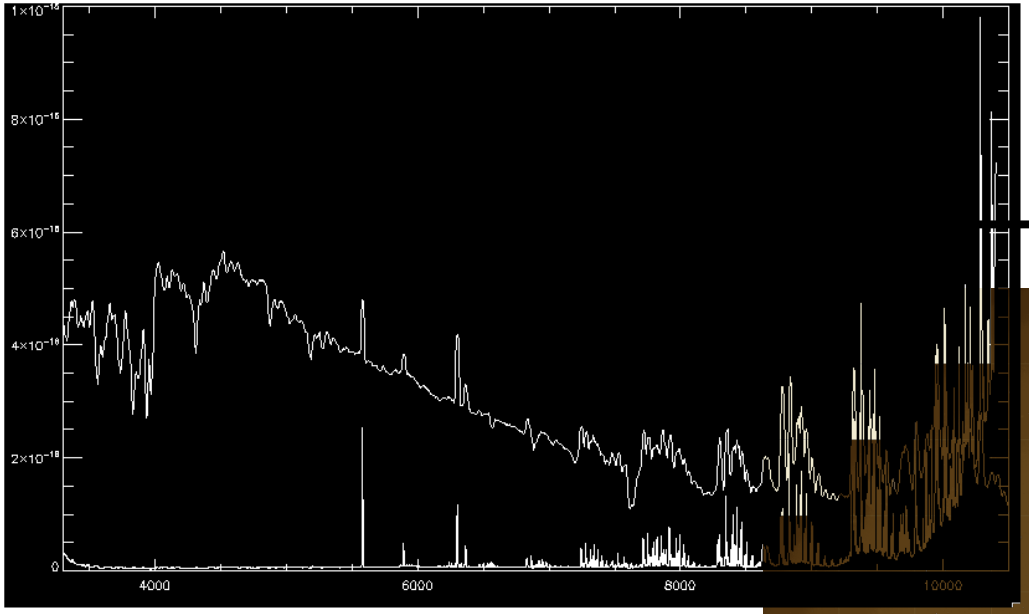
Lateral diffusion due to finite electric field

Physical model so diffusion & QE done simultaneously









Background Model

Post processing model after doing some test sims
Moon spectrum & dark sky spectrum
Moons angular variation according to Rayleigh scattering
Include varation of moon
Across FOV (few %)
Wavelength dep.
Vignetting derived from sim

Agrees with mag model in ETC to 10%

DC 3a statistics:

Lot of Initial Planning & Code dev.: 89 version changes from July to Feb ; 8 documents

2 10 sq. degree catalogs constructed at U Wash

Image sims done at Purdue by 3 undergrads on 7000 CPU CONDOR system

Typically ~150 chips running at a time

Not CPU limited (spent only 3 weeks out of 9 months) so could do a lot more

Have learned what needs to be automated

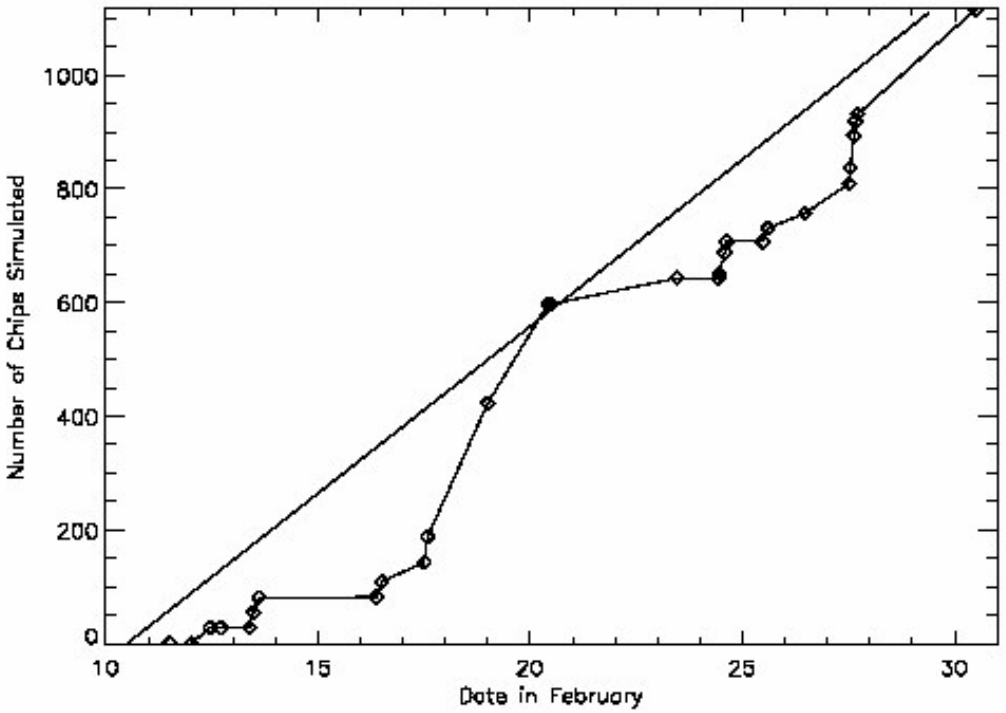
1116 chips

20 billion pixels

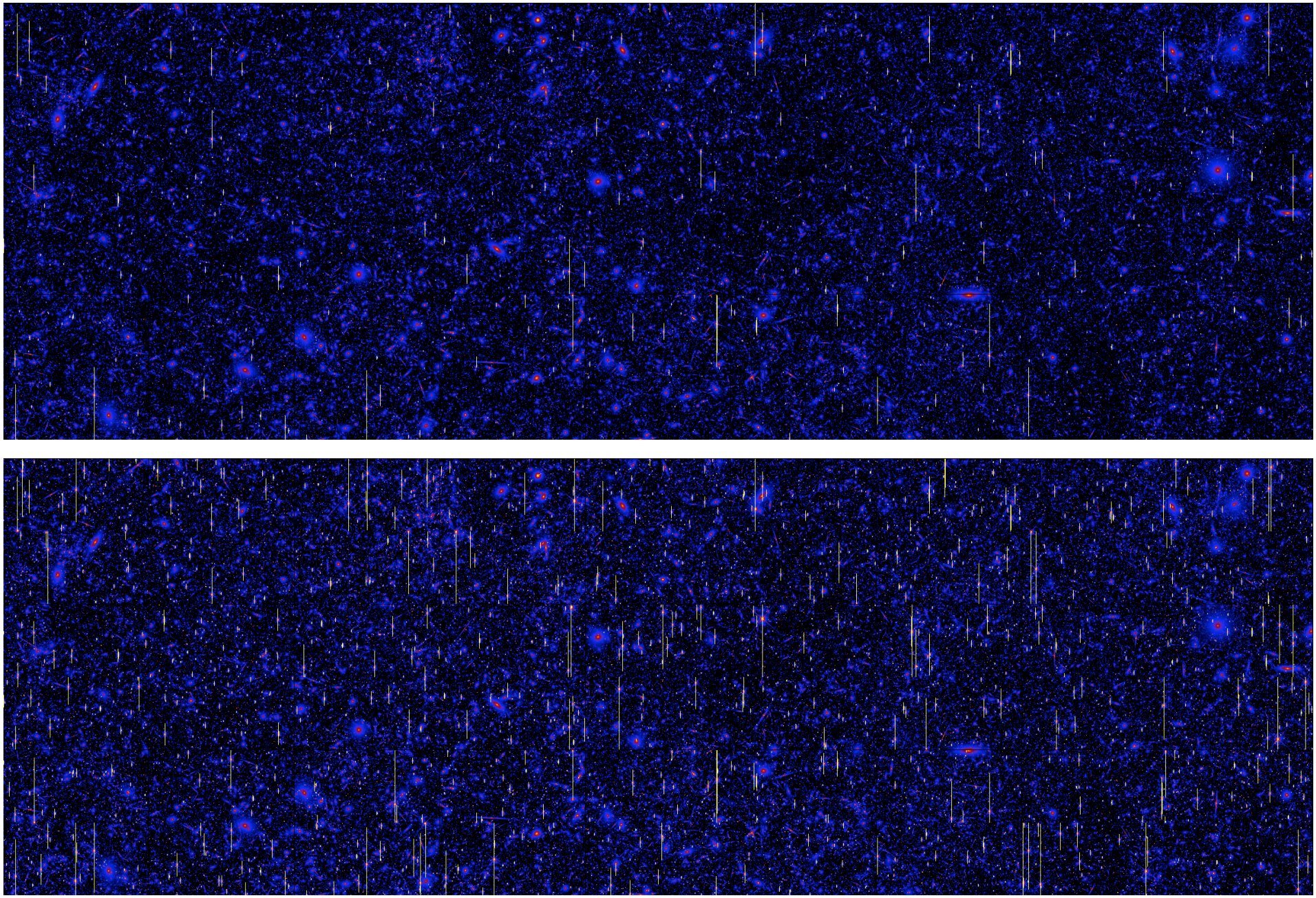
300 million objects

1 trillion photons

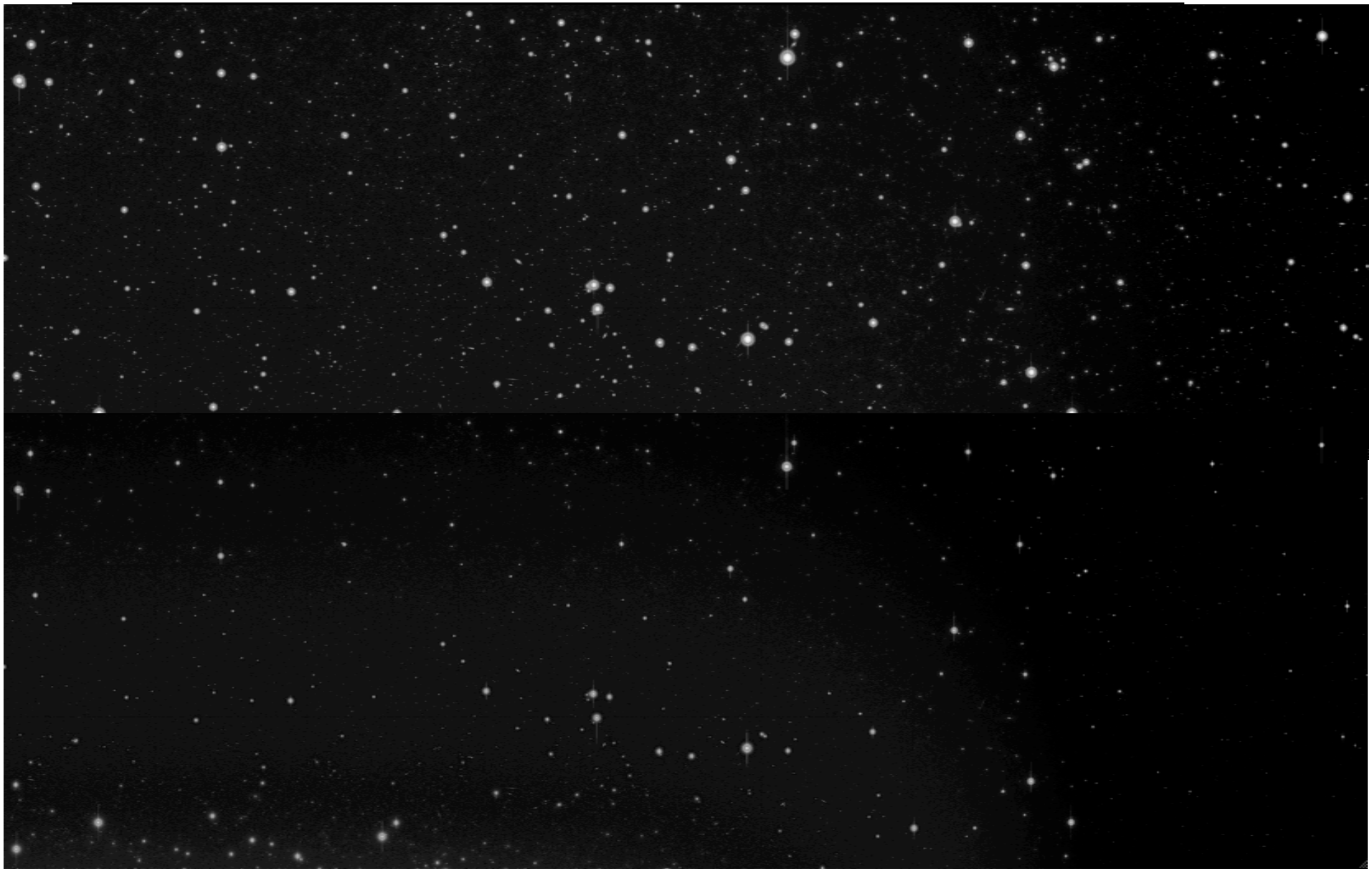
[100 s of LSST]



Ideal:



Variable Sky Background



The LSST focal plane

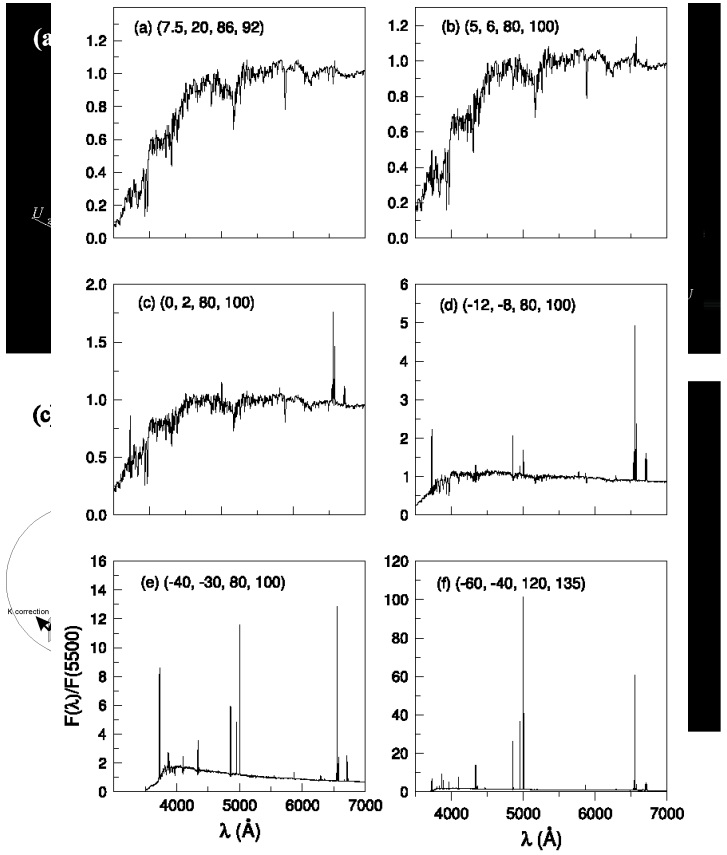
10^{11} photons
12.8 GB image
2000 CPU hrs

Working with Petascale Data Streams

- **Scaling the science**
  - Petabyte per year of catalogs
  - Few hundred parameters per source
  - 1000 time stamps after 10 years
- **Curse of dimensionality**
  - Many attributes but which ones are important
  - Algorithms don’t scale well with dimension
  - Physics usually simpler than just throwing all dimensions at a problem
  - Must account for noise, errors and missing data

Working with Petascale Data Streams

- **What is the dimensionality?**
  - **We measure thousands of parameters**
    - Which parameters are important?
    - These measures are correlated
    - Can we invert the observables
  - **Photometric properties**
    - Multiple colors and passbands
    - Type, luminosity and redshift are natural coordinates?
  - **Spectroscopic properties**
    - Many wavelengths (1000s of dimensions)
    - Apparently much lower dimensional system

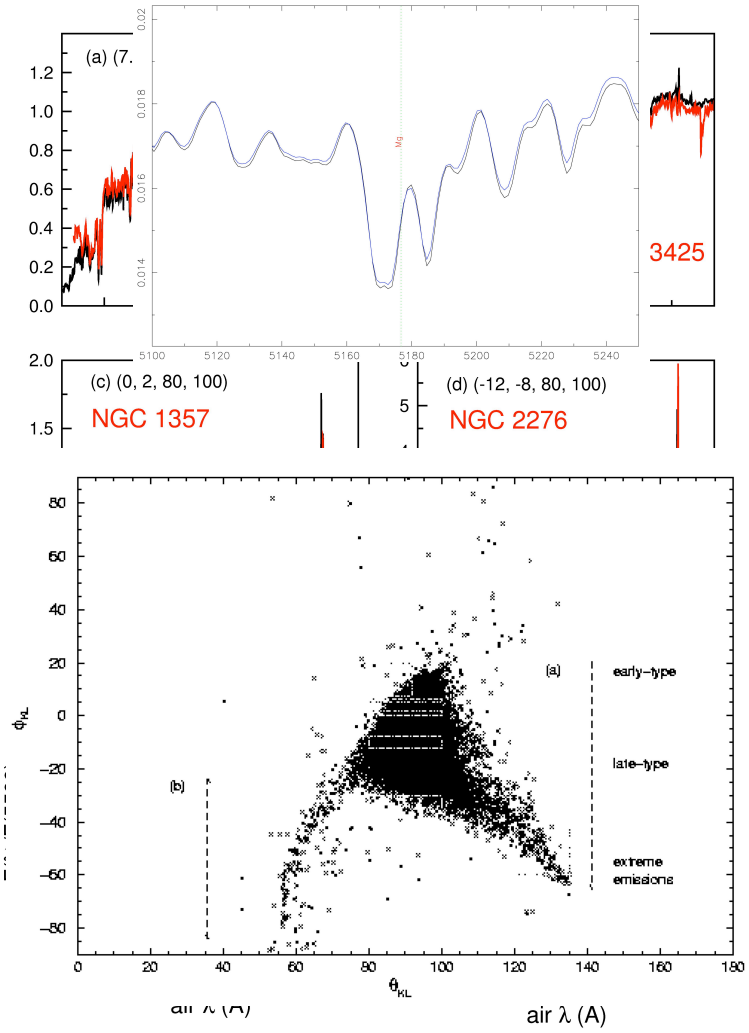


Reducing the dimensions

- **Non-parametric approaches**
  - **Karhunen-Loeve (KL)**

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

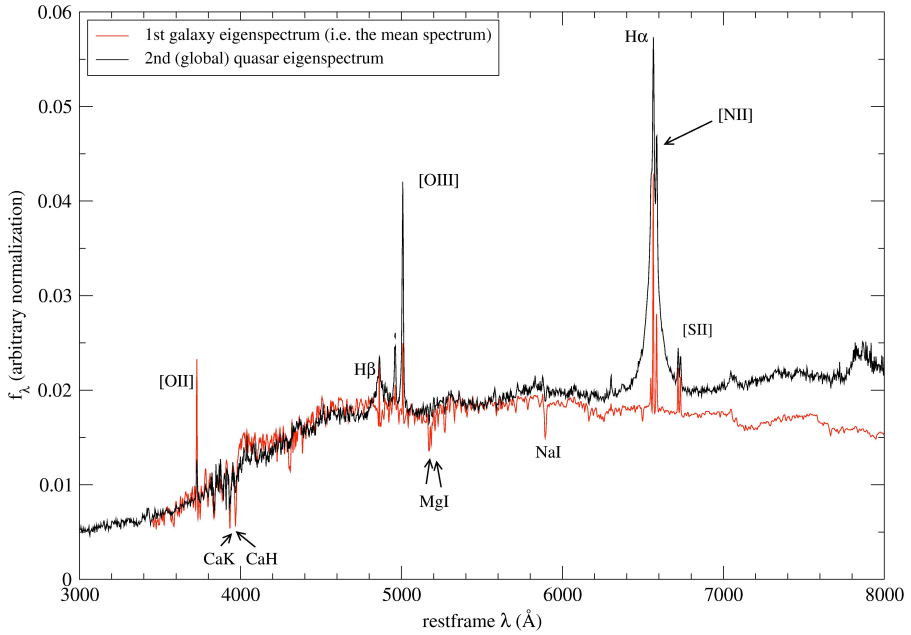
  - **Truncated expansion**
    - Reduced dimensions
    - Signal to noise weighted
    - Compression (400:1)
    - Noise reduction
  - **Classification**
    - Galaxies described by small # of dimensions (10)
    - Correlations are physical
    - 600K spectra



Yip et al 2004

Physical Correlations

- **Extracting QSO correlations**
  - **Correct for coverage**
    - Broad wavelength range
    - Gaps within the data
  - **Physical relations**
    - Increase in iron with decreasing redshift
    - Tracing SN Ia rates?
  - **Host galaxy in QSO spectra**
    - eigenspectra contain the host galaxy component
    - QSO are anomalies within galaxy spectra



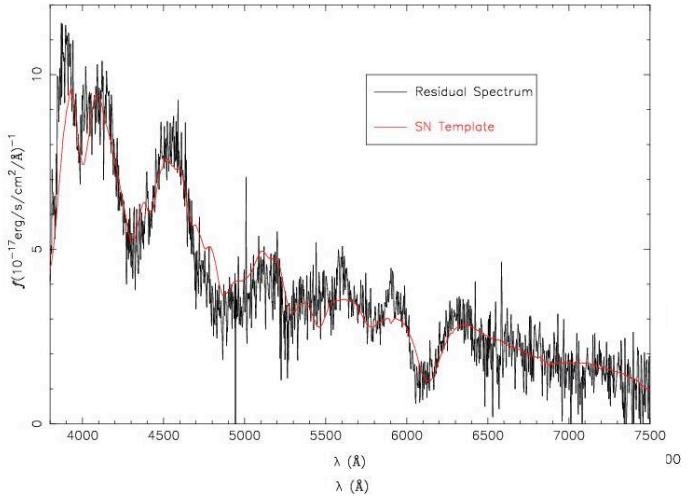
Yip et al 2004

Vanden Berk et al 2005

Deviants in astronomy

- **Anomalies in spectra**
  - **Identifying components**
    - Model the underlying spectrum
    - Identify additional components
    - Joint fit for KL and additional components
    - Spectral contamination by AGNs
  - **Supernovae**
    - 1 per galaxy per 100 years
    - 600K spectra observed
    - 116 Type 1a SNe (1:5200)
    - Visible for 30-60 days

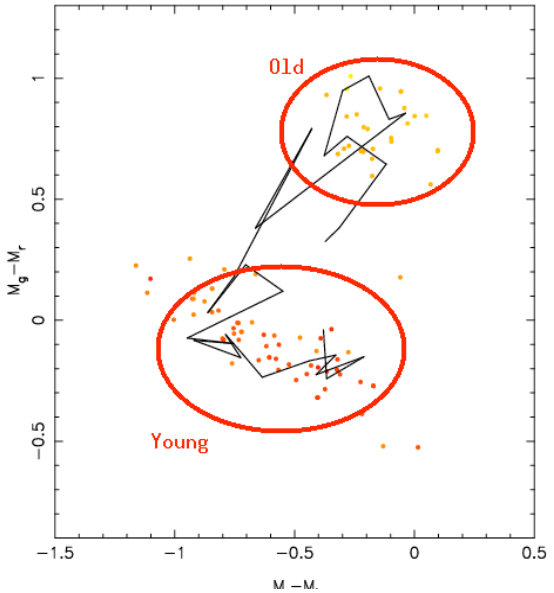
$$f(\lambda) = \sum_{i<N} a_i e_i(\lambda) + b_j SN_j(\lambda) + c_k AGN_k(\lambda)$$

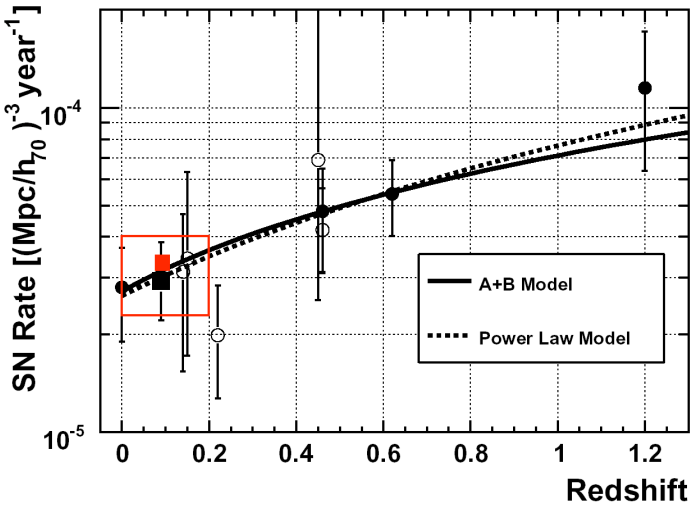


Magewick et al 2004
Krughoff etal 2009

Rates of Supernovae

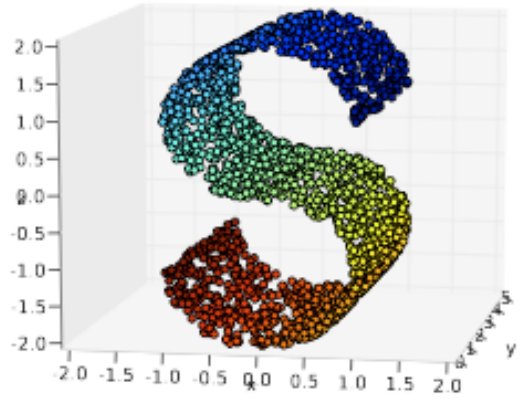
- **Classification is a natural result**
  - Derived from compression
  - Solve for galaxy spectral type
  - Solve for SNe type, luminosity and age
  - Quantify efficiencies
- **Local supernova rates**
  - From SDSS 600K spectra
  - <z> = 0.1011
  - Rate = 0.240 +/- 0.02 SNu
  - 1 SNu = 10^{10} $L_{\odot}$ per century
  - Serendipitous science

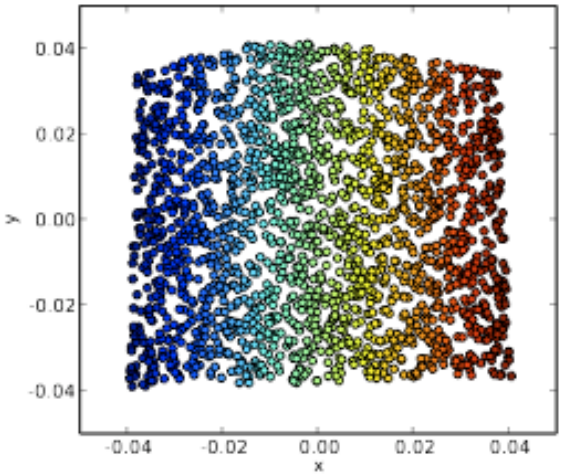




Not just linear

- **Local Linear Embedding**
  - **Roweis and Saul (2000)**
    - KL a global statistic
    - Not compact for non-linear structure
  - **Local Embedding**
    - Local structure
    - Calculate local structure (weights)
    - Find projection that preserves weights
    - Finds lower dimensional manifolds
    - Slow and not always robust to outliers



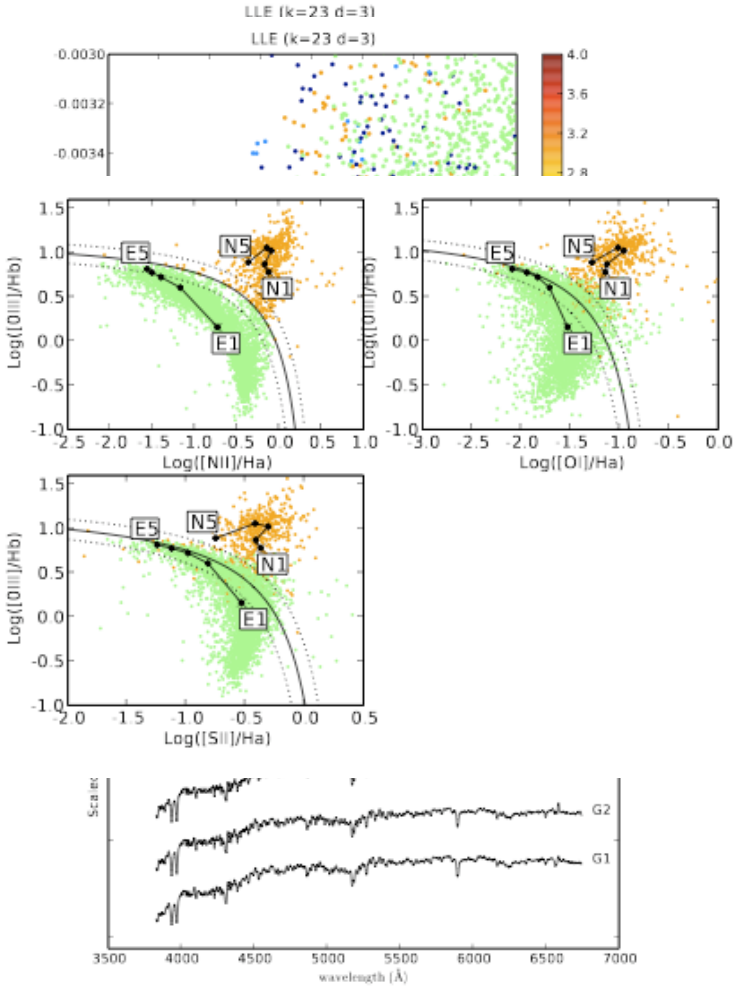


$$\mathcal{E}_1^{(i)}(\mathbf{w}^{(i)}) = \left|\mathbf{x_i} - \sum_{j=1}^{K} w_j^{(i)} \mathbf{x}_{n_j^{(i)}}\right|^2$$

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

LLE for Spectra

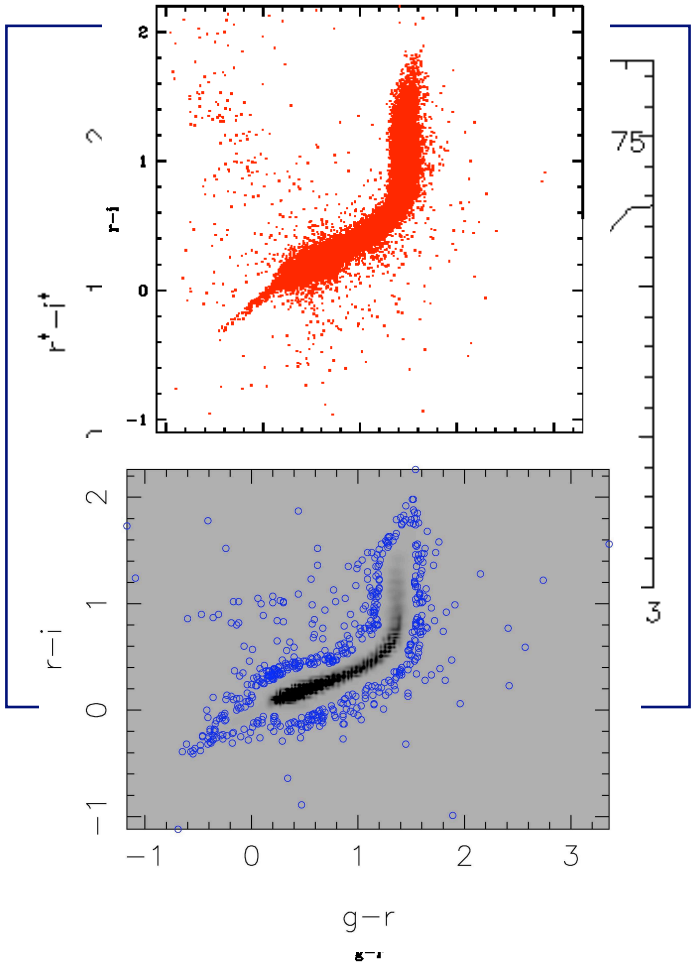
- **Projecting Spectra using LLE**
  - **Define three dimensions**
  - **Sources close in high dimensions are close in low dimensions**
  - **Classification is simpler**
  - **Fits emission lines, broad line, normal galaxies in one classification**
    - More robust than SDSS pipeline
    - No derived attributes
  - **Reproduces Kewley Diagrams**
  - **Used to learn distributions of sources**
    - Training photo-z samples



VanderPlas, ajc 09

Attributes not images/spectra

- **Anomalies are common in astronomy**
  - **Color outliers lead to discovery of QSOs**
- **Density estimation**
  - **Gaussian mixtures as density estimators**
  - **EM algorithm with pruning of mixtures**
  - **Predict likelihood of a source as a fn of attribute**
  - **Dependent on penalized likelihood form**
- **Rank order sources based on density**
  - **Classify the sources**
  - **Estimating the density (probability) of a sources**

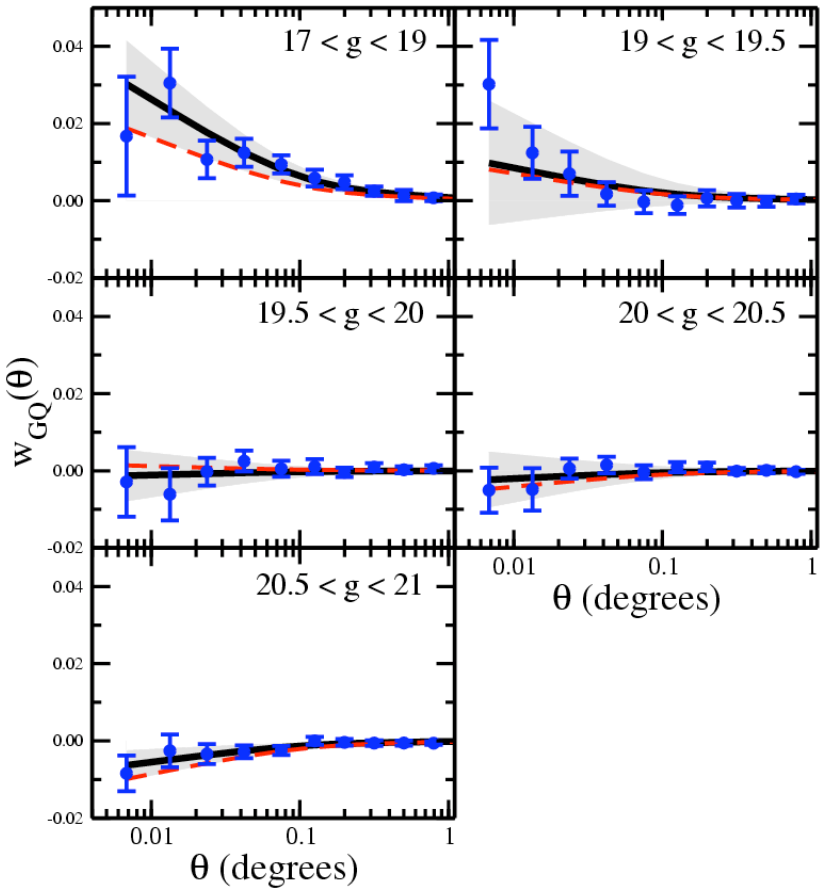


Connolly et al 2000

Richards et al 2003

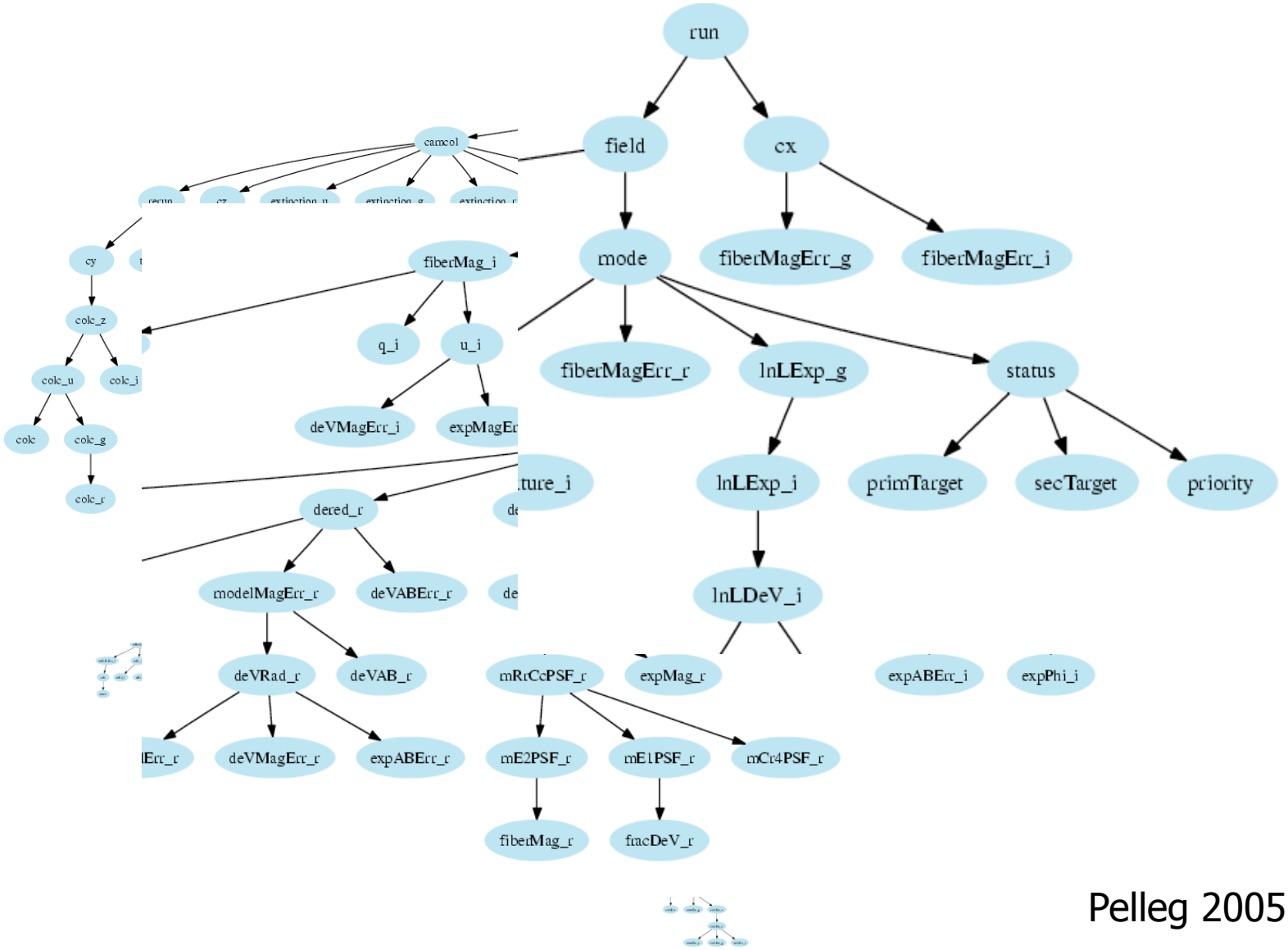
Why Anomalies: Magnification Bias

- **Select QSOs photometrically**
- **Correlate with foreground**
  - **Foreground galaxies lens QSOs**
  - **Lensing magnifies and dilutes QSO distribution**
  - **Magnitude of effect depends on slope of number counts**
    - Slope dN/dm of 0.4 is the crossover point



Scranton et al 2005

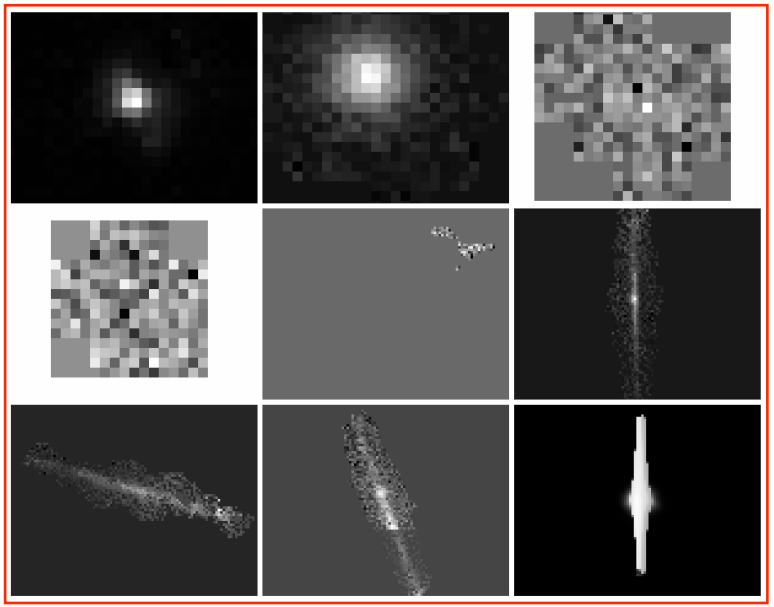
Graphs: acyclic, dependency trees



Pelleg 2005

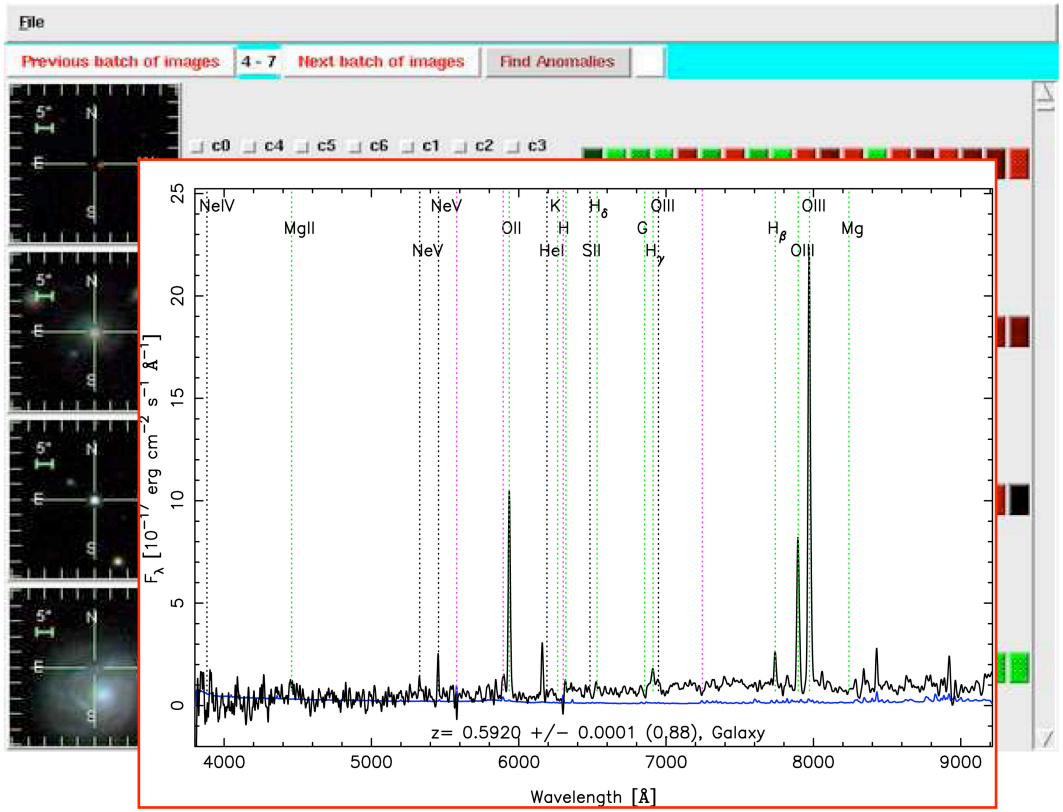
Anomalies from a tree

- **Applying the tree**
  - Test each point against the tree
  - Determine how well a source is drawn from the graph
  - Rank order sources in the SDSS
  - Say why it is anomalous
- **What is a one in a million source?**
  - Anomalies are also artefacts
  - Diffraction spikes
  - Cosmic Rays
  - Bad deblends
  - Real sources



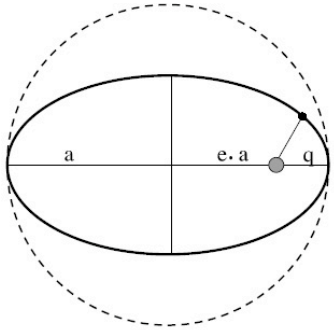
Adaptive learning

- **Blind classification**
  - No expert user input
  - Many anomalies have similar attributes
  - Diffraction spike
- **Iterative anomalies**
  - Classify anomalies
  - Simple Bayesian classifier (single Gaussians)
  - Learn with EM or mixture models
  - Apply classification and relearn the tree
  - Iterate to a solution

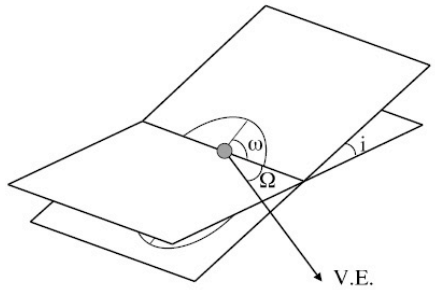


One in a million to a million moving sources

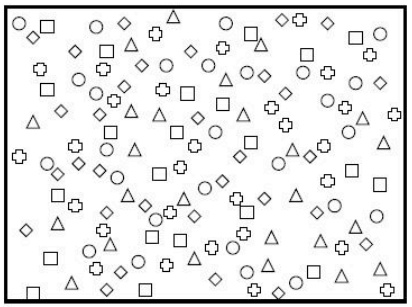
- **From positions to orbits**
  - Series of observations
  - Six parameters define an orbit (ellipse, orientation)
  - ≥3 observations required for an orbit
  - Sampling in time must extend over weeks
- **Computational/operational issues**
  - 10^6 sources (MBA plus PHAs)
  - 15 observations per month
  - Requires fast turn around and robust classification
  - Trade-off between repeated observations vs computational complexity



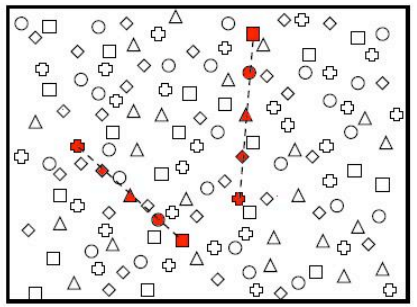
(A)



(B)



(A)

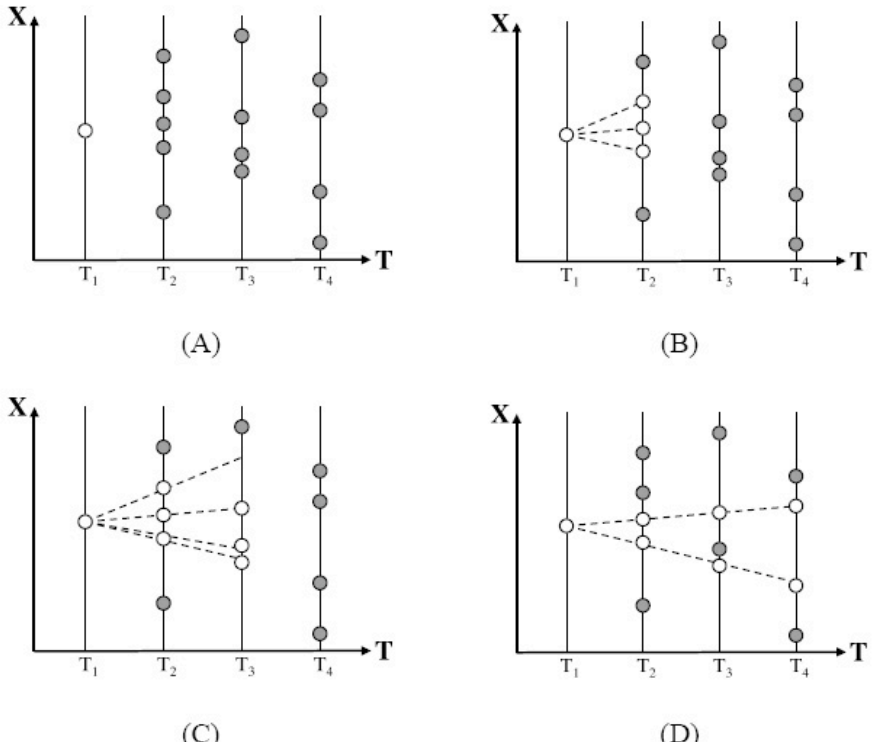


(B)

Kubica et al 2006

Multi-Hypothesis Tracking

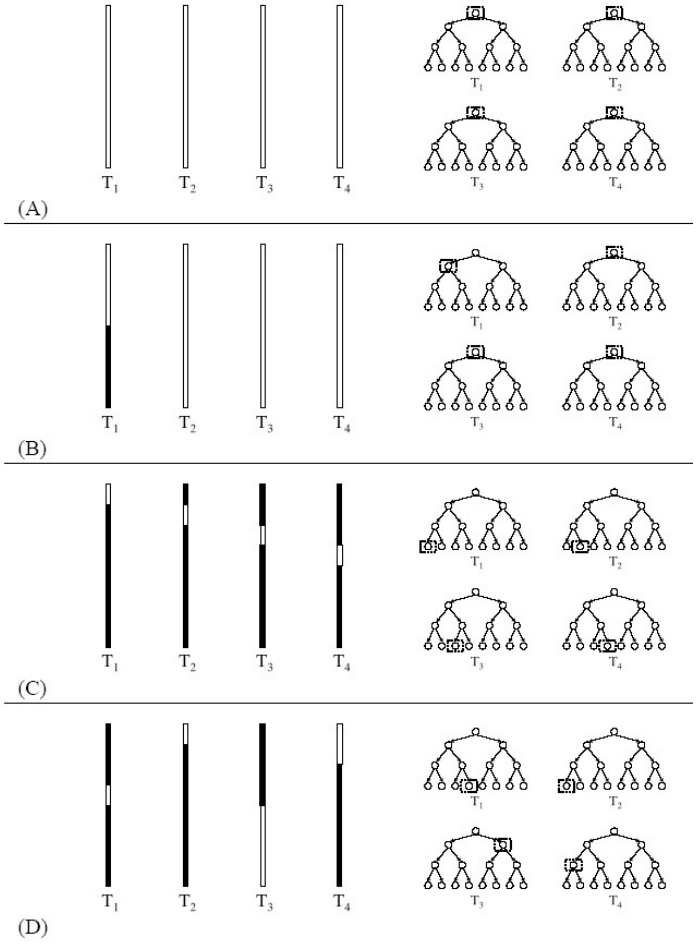
- **Simplified progressive search**
  - **Predict positions at next time step**
  - **Exhaustively search all possibilities**
  - **Prune false tracks by orbit fitting**
  - **Very slow, requires close spacing**
  - **Most assumed linear prediction**
  - **Problem for noisy data (combinatorics)**
  - **Missing data is an issue**



Approximation	kd-Trees?	Time (sec)	P_F	P_C	WI
Linear	NO	93	0.9622	0.0206	832369763
Linear	YES	6	0.9622	0.0206	832369763
Quadratic	NO	59	0.9638	0.8867	55038
Quadratic	YES	3	0.9638	0.8867	55038

Variable Tree Approach

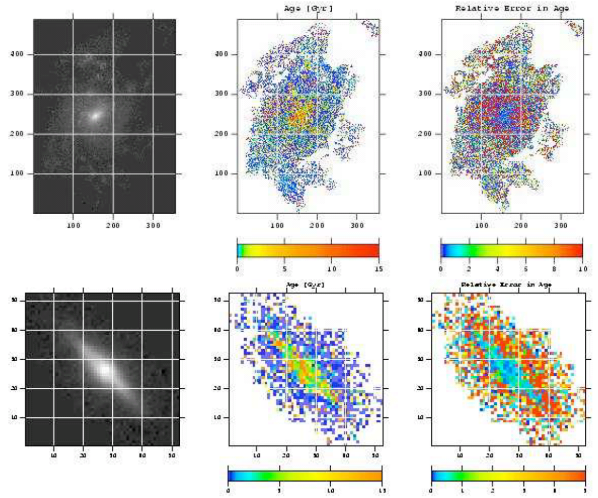
- **Search all trees jointly**
  - **Minimize the number of tree searches**
  - **When a quadratic fit is found look for support in other time steps**
  - **Currently used for PanSTARRS, SDSS-II and LSST asteroid searches**
  - **160,000 asteroids found in the SDSS repeat scans (by 4 undergraduate students)**
  - **Challenges remain for the LSST density and cadences.**

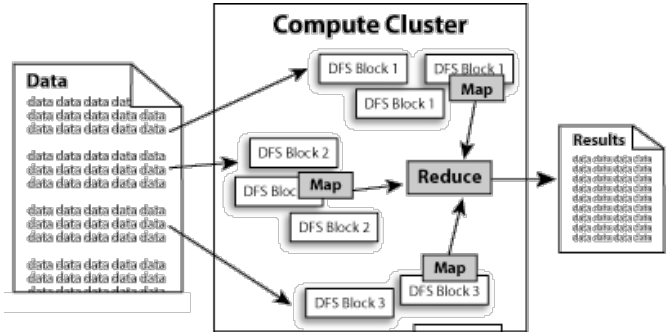


Kubica et al 2006

LSST: Scaling the Science

- **Petascale Science**
  - **Scalable Algorithms**
    - What works for 1000 sources does not necessarily scale to 100 million
    - Creating a framework that naturally scales to the size of LSST data
  - **Distributing the load**
    - Google/IBM/NSF - CluE Cluster
    - Map-reduce framework
    - Distribute the data – distribute the work
  - **Images to events**
    - Simple scaling to thousands of machines
    - Scalable image simulations
    - Cross-matching data
    - Hunting for moving sources at 100 AU





The Future

- **New Surveys = New Science**
  - Dark energy, dark matter, nature of the Solar System
  - High precision cosmology
  - Open data and wide range of application
- **Changing the way we do science**
  - **Science in the Petabyte era will not be the same as today**
    - Neither should the scientist
  - **Science needs to scale**
    - Data is not just the challenge
    - Not just a question of cpus
    - We need to change the way we view data (and the information it contains)