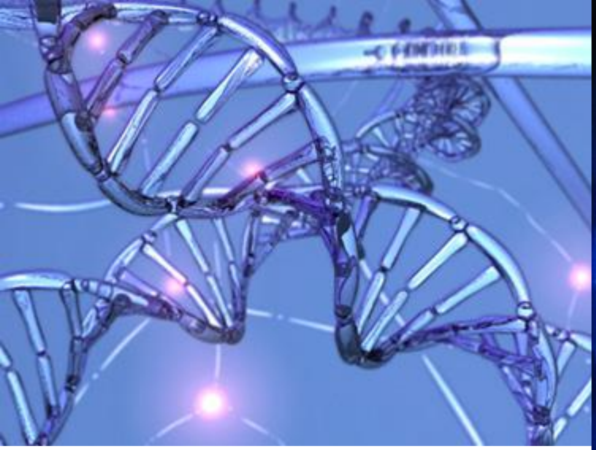


(Too Much) **Data Everywhere**

Alex Szalay

Institute for Data-Intensive Engineering and Science

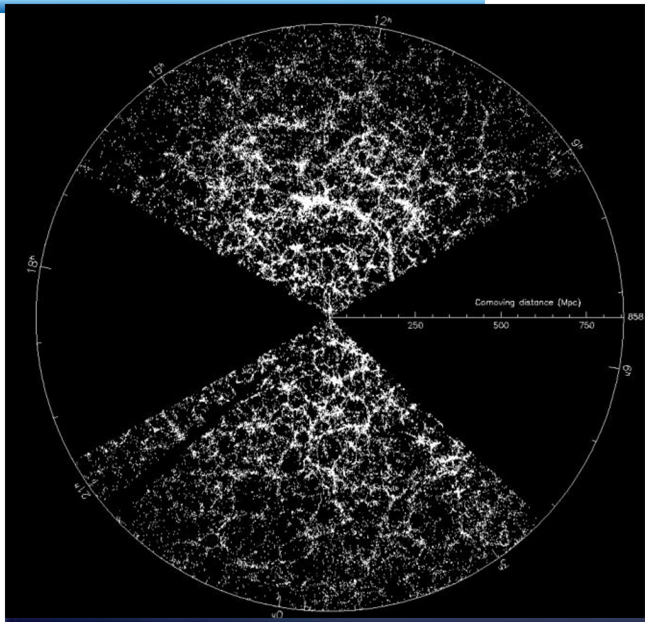
The Johns Hopkins University

Sloan Digital Sky Survey

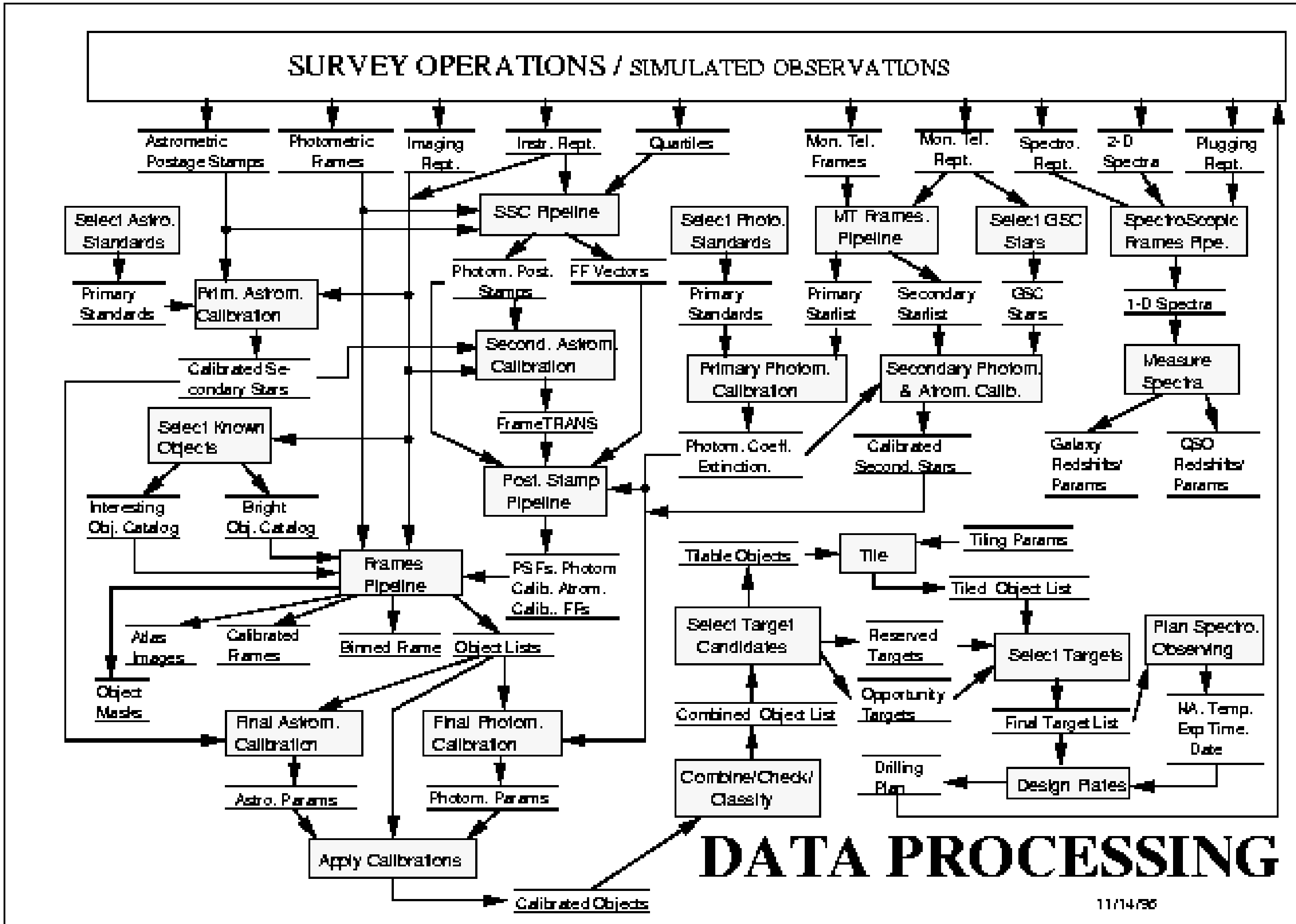
“The Cosmic Genome Project”

- Started in 1992, finished in 2008
- Data is public
  - 2.5 Terapixels of images => 5 Tpx of sky
  - 10 TB of raw data => 100TB processed
  - 0.5 TB catalogs => 35TB in the end

- Database and spectrograph
  built at JHU (SkyServer)
- Now SDSS-3/4 data served from JHU



Data Processing Pipelines



Wide Range of Science

- 5,000 publications, 200,000 citations
- More papers from outside the collaboration
- From cosmology/LSS to galaxy evolution, quasars, stellar evolution, even time-domain
- Combination of 5-band photometry and matching spectroscopy provided unique synergy
- Overall, seeing not as good as originally hoped for, but systematic errors extremely well understood
- Very uniform, statistically complete data sets

The Broad Impact of SDSS

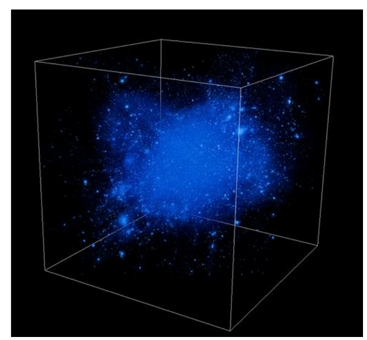
- Changed the way we do astronomy
- Remarkably fast transition seen for the community
- Speeded up the first phase of exploration
- Wide-area statistical queries easy
- Multi-wavelength astronomy is now the norm
- SDSS earned the TRUST of the community
- Enormous number of projects, way beyond original vision and expectation
- Many other surveys now follow
- Established expectations for data delivery
- Serves as a model for other communities of science

Science is Changing

THOUSAND YEARS AGO
science was empirical
describing natural phenomena

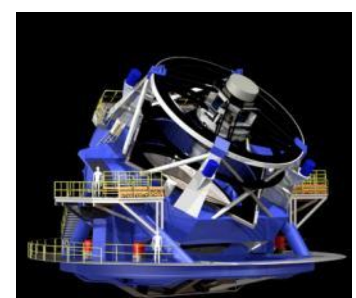
LAST FEW HUNDRED YEARS
theoretical branch using models, generalizations

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$



LAST FEW DECADES
a computational branch simulating complex phenomena

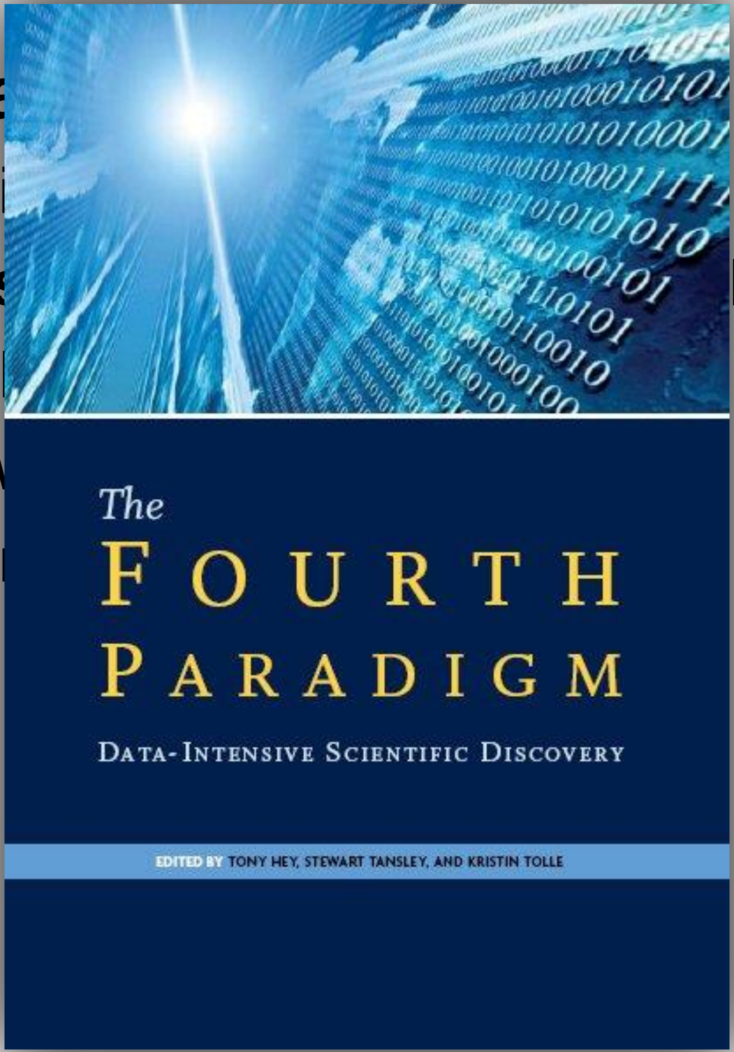
TODAY
data intensive science, synthesizing theory, experiment and computation with statistics
►new way of thinking required!



Gray's Laws of Data Engineering

Jim Gra

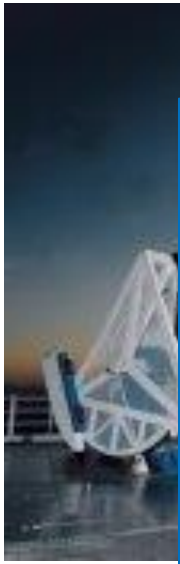
- Scienti   around **data**
- Need s   lysis
- Take t
- Start w
- Go fro



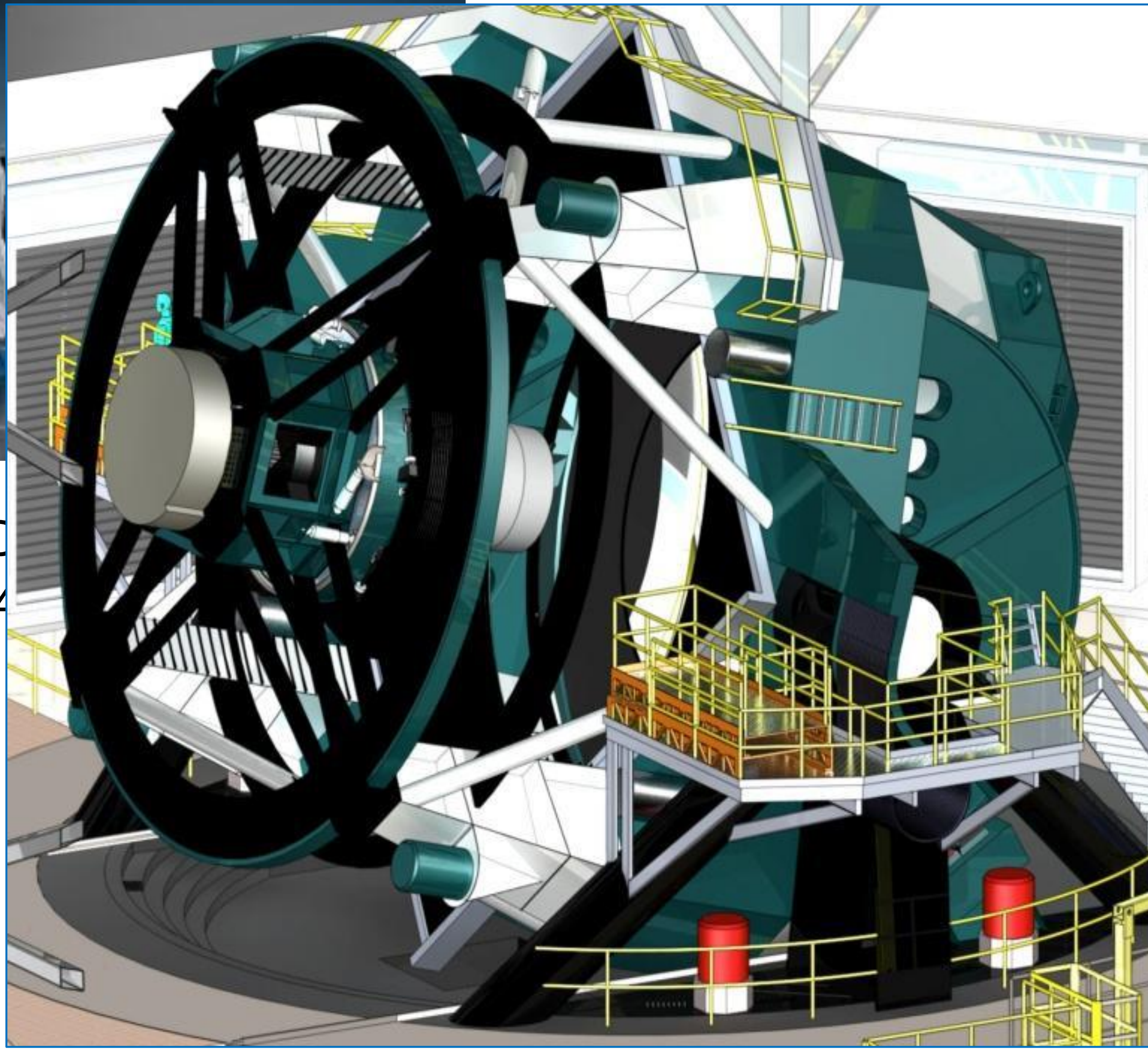
How Do We Prioritize?

- Data Explosion: science is becoming data driven
- It is becoming “too easy” to collect even more data
- Robotic telescopes, next generation sequencers, complex simulations
- **How long can this go on?**

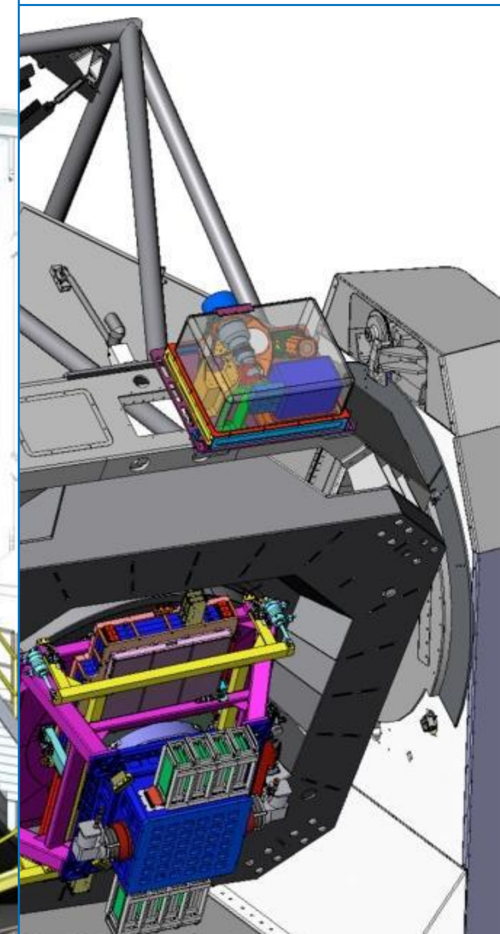
- “Do I have enough data or would I like to have more?”
- No scientist ever wanted less data….
- But: Big Data is synonymous with Dirty Data
- How can we decide how to collect data that is ***more relevant*** ?



SD
2.4

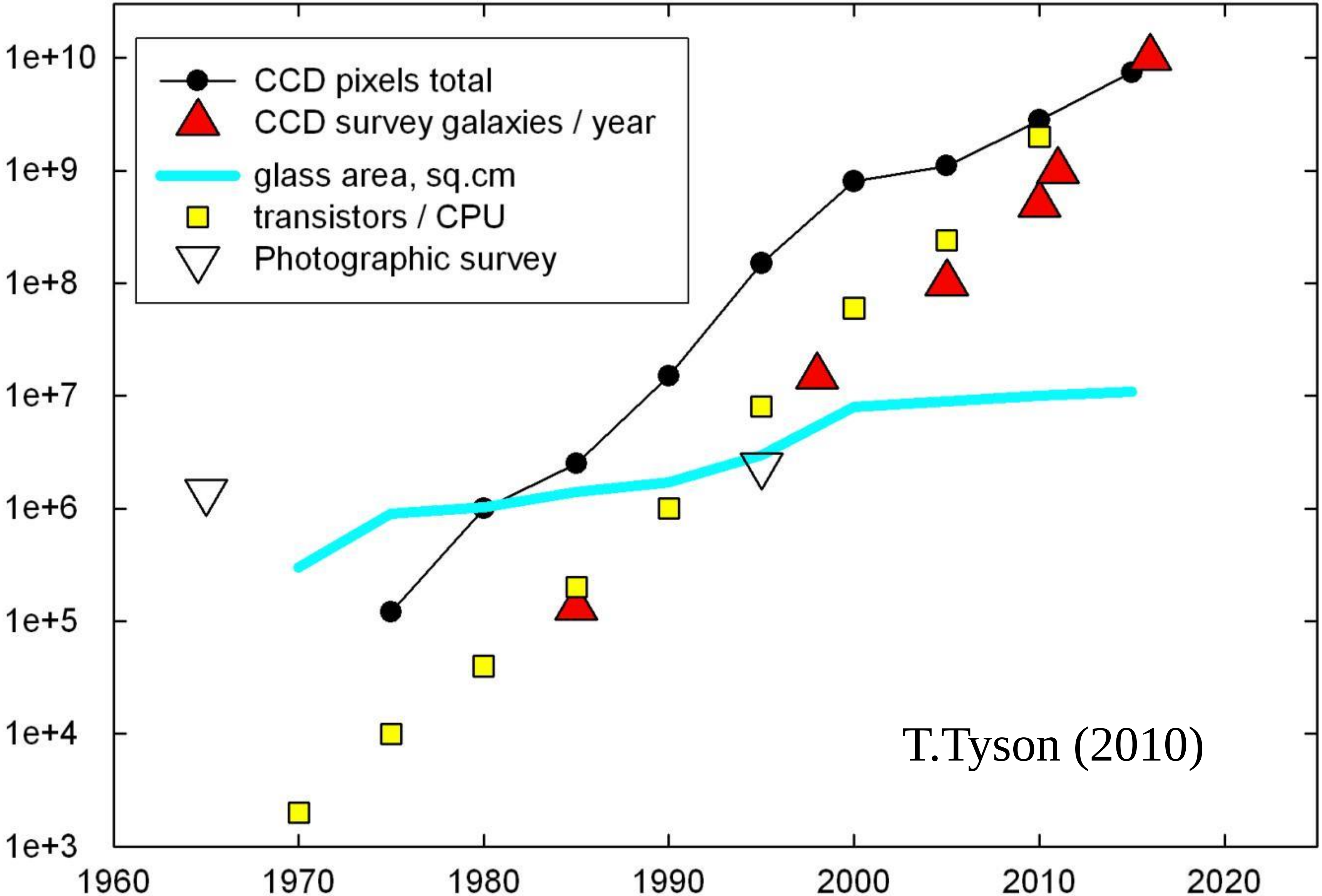


LSST
8.4m 3.2Gpixel



PanSTARRS
1.8m 1.4Gpixel

Survey Trends



Decision Making in Science

- Traditionally: human scientists decide what experiments to do next
- SDSS Example: the Black Book
  - *Optimization and tradeoffs were done by committee*
  - *In the end >5000 publications, many outside the team*
  - *Many science projects were never thought of*
- Given the huge amounts of data, the possible number of new experiments and analyses explodes
- But: we cannot do it all, we cannot foresee it all!
- Need to involve intelligent tools aiding the scientist

What Will the 5th Paradigm Be?

- Next step: not just discovery but experiment design!!!

- Probabilistic approach to everything
- Accelerated design cycles
- Clear cost function driving tradeoffs
- How to collect **more relevant** data?

The systematic involvement of computational statistics and optimizations in the design of the next generation of "experiments":

prediction/Inference/UQ + design/synthesis/fabrication

How to Do with Less Data?

- Collect less but more relevant data
  - *Use active learning*
  - *Compressive sensing: nature is sparse*
  - *Random sampling of long tails: stratified sampling*
- Streaming, sublinear randomized algorithms
  - *Streaming look at simulations as well*
  - *Not just sequence of snapshots, but world-lines*
- Automation, machine learning to find relevant data

Probabilistic Approach

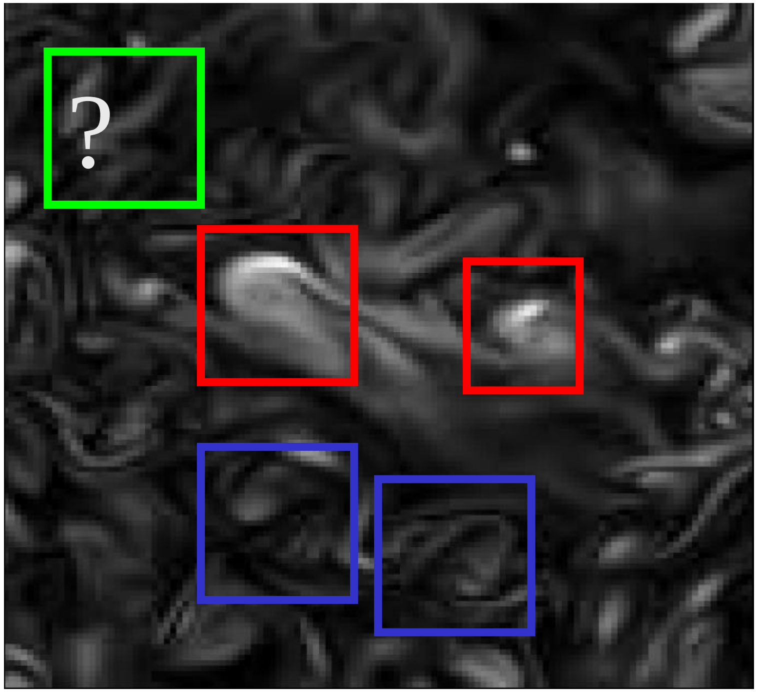
- Time-to-result: how to trade speed for accuracy
  - *statistical and algorithmic challenges*
  - *statistical vs systematic errors*
  - *best result in 1 min, 1 hour, 1 day*
  - *Cost of computing is becoming a significant factor*
- Simulations: how to do better UQ
  - *from single large realization to ensembles (Coyote Universe, INDRA)*
  - *sparsely sampled outputs*
- Experiments
  - *from driven by "feeling" (and experience) to objective design based on statistics, automated choice of parameters*
  - *Ensembles of experiments optimally sampling parameters*

Active Learning

- Given our existing data, of all possible experiments which would yield the most **new** information?
- Ross King (2004) drug design study:
  - *Adam, the Robot Scientist*
- Personalized Medicine
- Finding patterns in large scale simulations

Applications of ML to Turbulence

Renyi divergence



Vorticity

Similarity between regions

- ❑clustering,
- ❑classification,
- ❑anomaly detection

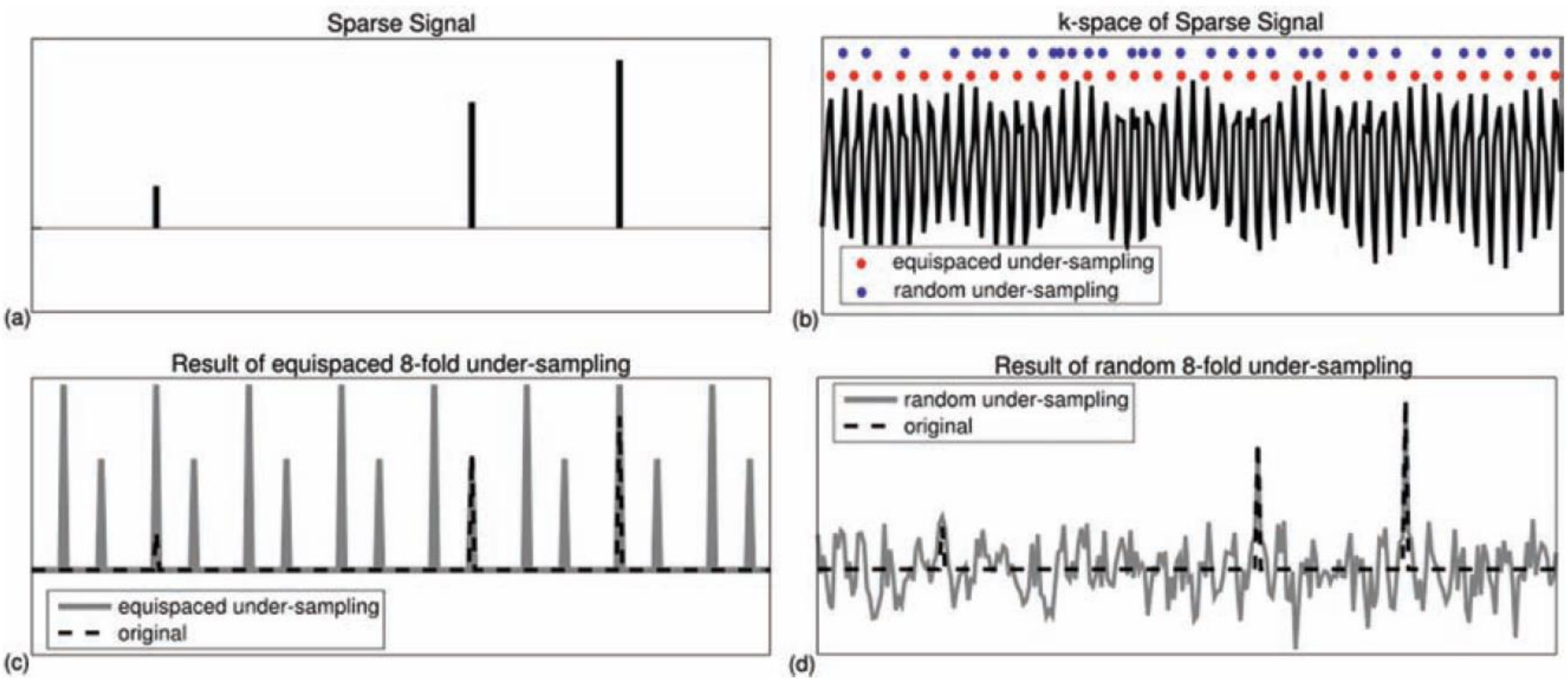
J. Schneider, B. Poczos, CMU

Nature is Sparse

- Many natural processes are dominated by a few processes and described by a sparse set of parameters
- Compressed Sensing has emerged to identify in high dimensional data sets the underlying sparse representation (Candes, Donoho, Tao, et al)
- This enables signal reconstruction with much less data!
- The resolution depends not on the pixel count but on the information content of an image…

Compressed Sensing

- Example: sparse signal sampled randomly in Fourier space



Donoho, Candes, Tao…

Principal component pursuit

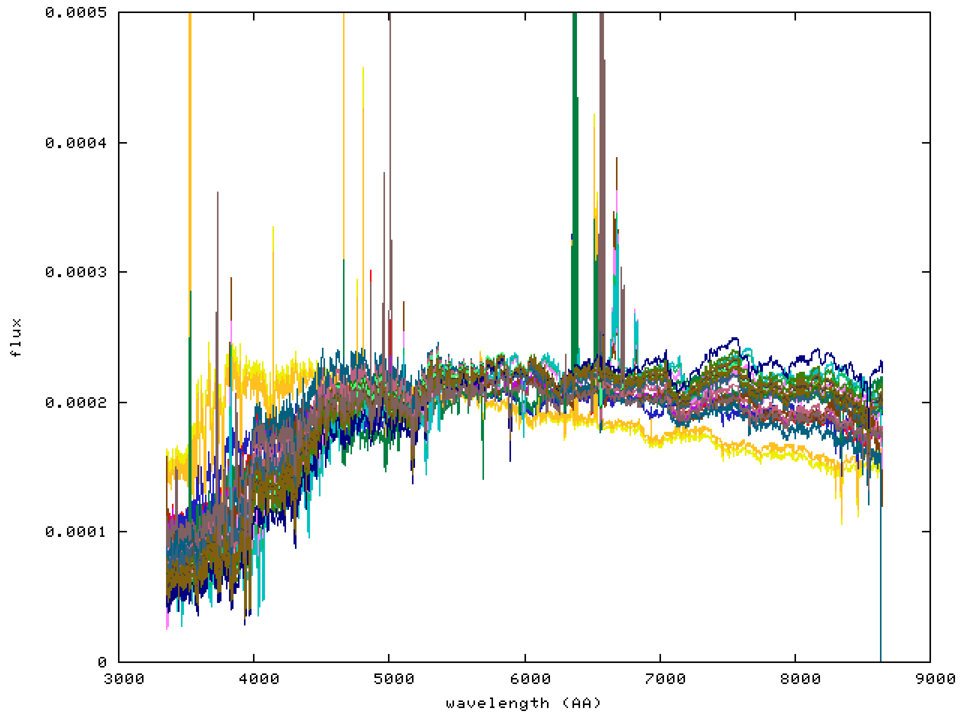
- Low rank approximation of data matrix: X
- Standard PCA: $$\min \|X - E\|_2 \quad \textit{subject to} \;\; rank(E) \le k$$
  - *works well if the noise distribution is Gaussian*
  - *outliers can cause bias*
- Principal component pursuit $$\min \|A\|_0 \quad \textit{subject to} \;\; X = N + A, \;\; rank(N) \le k$$
  - *"sparse" spiky noise/outliers: try to minimize the number of outliers while keeping the rank low*
  - *NP-hard problem*
- The L1 trick: $$\min_{N,A}\left(\|N\|_* + \lambda\|A\|_1\right) \;\; \textit{subject to} \;\; X = N + A$$
  - *numerically feasible convex problem (Augmented Lagrange Multiplier)*

$$\min_{N,A}\left(\|N\|_* + \lambda\|A\|_1\right) \quad \textit{subject to} \;\; \|X - (N + A)\|_2 < \varepsilon$$

* E. Candes, et al. "Robust Principal Component Analysis". preprint, 2009.
Abdelkefi et al. ACM CoNEXT Workshop (traffic anomaly detection)

Testing on Galaxy Spectra

- Slowly varying continuum + absorption lines
- Highly variable “sparse” emission lines
- This is the simple version of PCP: the position of the lines are known
  - but there are many of them, automatic detection can be useful
  - spiky noise can bias standard PCA



DATA:
Streaming robust PCA implementation for galaxy spectrum catalog (L. Dobos et al.)

SDSS 1M galaxy spectra
Morphological subclasses
Robust averages + first few PCA directions

Streaming PCA

- Initialization
  - *Eigensystem of a small, random subset*
  - *Truncate at p largest eigenvalues*

$$\boldsymbol{C} \approx \boldsymbol{E}_p \boldsymbol{\Lambda}_p \boldsymbol{E}_p^{\mathrm{T}}$$

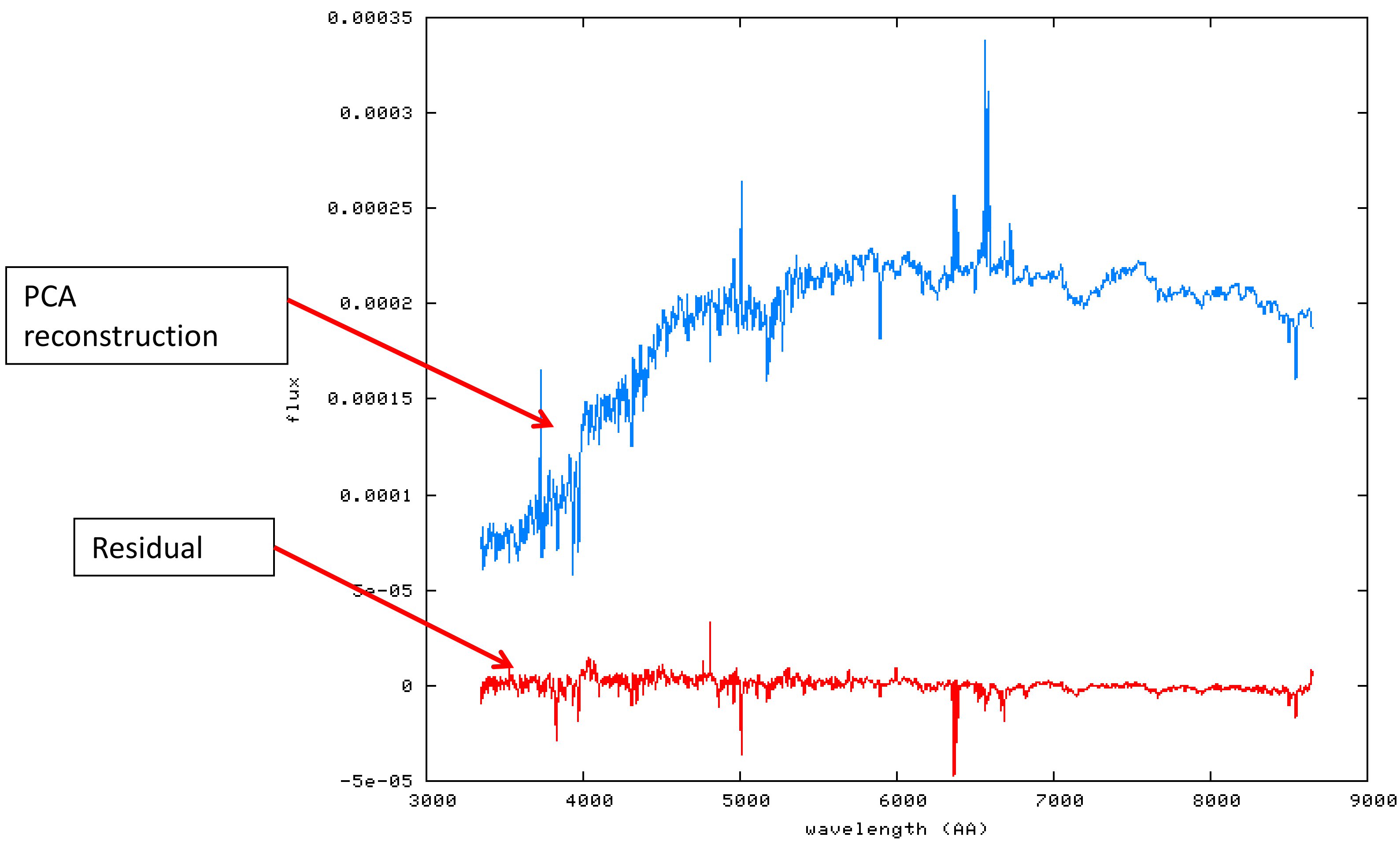
- Incremental updates
  - *Mean and the low-rank A matrix*
  - *SVD of A yields new eigensystem*

$$\boldsymbol{C} \approx \gamma \boldsymbol{E}_p \boldsymbol{\Lambda}_p \boldsymbol{E}_p^{\mathrm{T}} + (1-\gamma) \boldsymbol{y} \boldsymbol{y}^{\mathrm{T}}$$

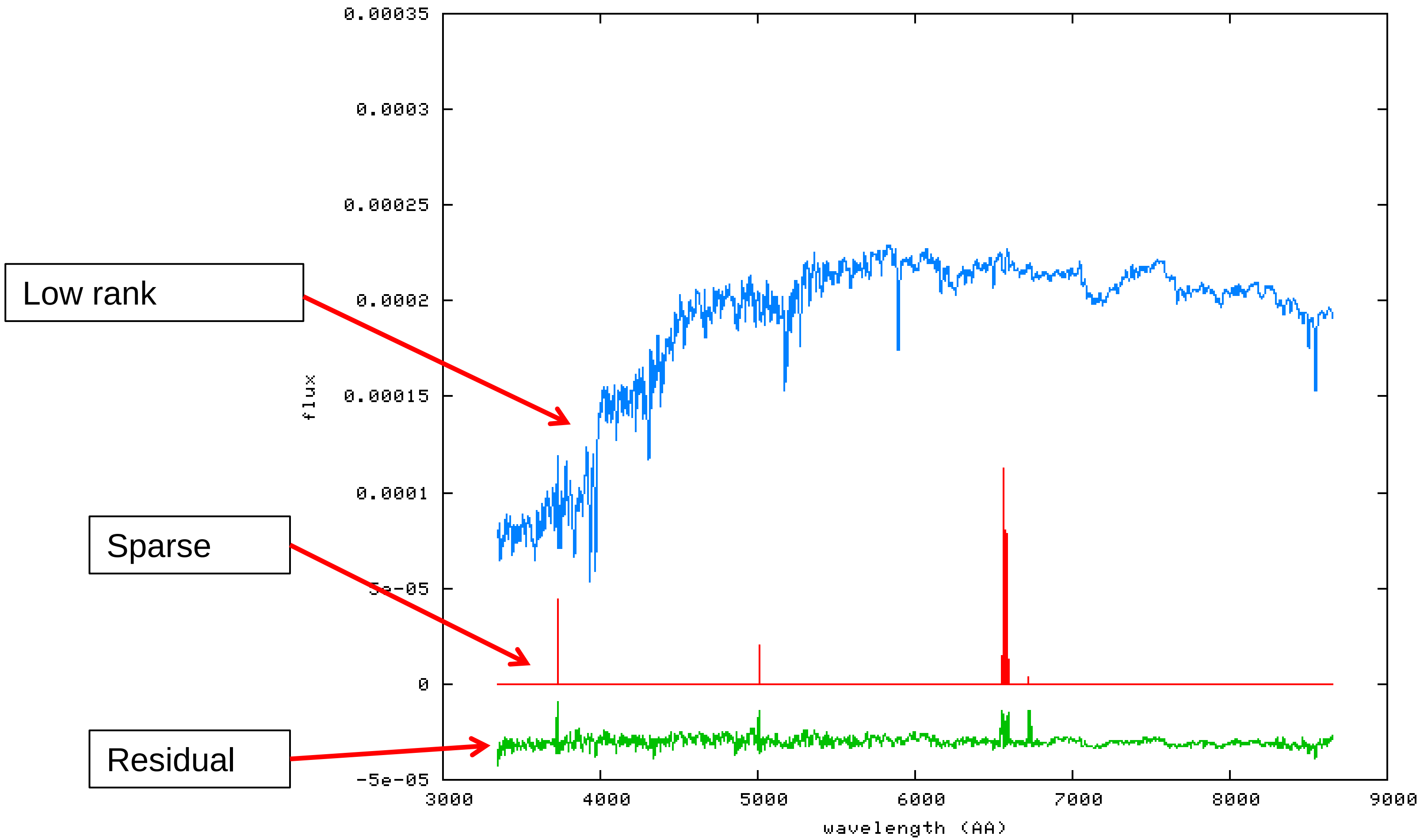
- Randomized sublinear algorithm!

Mishin, Budavari, Ahmad and Szalay (2012)

PCA


PCA reconstruction
Residual
flux
wavelength (AA)
0.00035
0.0003
0.00025
0.0002
0.00015
0.0001
5e-05
0
-5e-05
3000
4000
5000
6000
7000
8000
9000

Principal component pursuit



$\lambda = 0.6/\mathrm{sqrt}(n)$, $\varepsilon = 0.03$

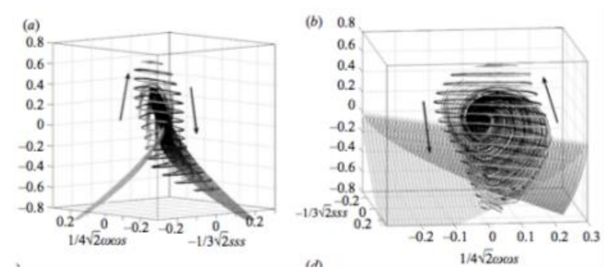
Numerical Simulations

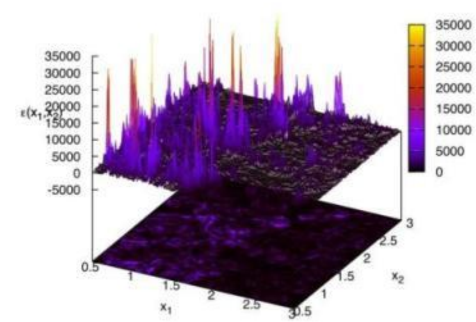
- HPC is an instrument in its own right
  - *Soon largest simulations exceed several petabytes*
  - *Directly compare to the experiments*
- Need public access to the best and latest
  - *Cannot just do in-situ analyses*
- Also need ensembles of simulations for UQ
- Creates new challenges
  - *How to access the data?*
  - *What is the data lifecycle?*
  - *What are the analysis patterns?*
  - *What architectures can support these?*

Immersive Turbulence

"… the last unsolved problem of classical physics…" Feynman

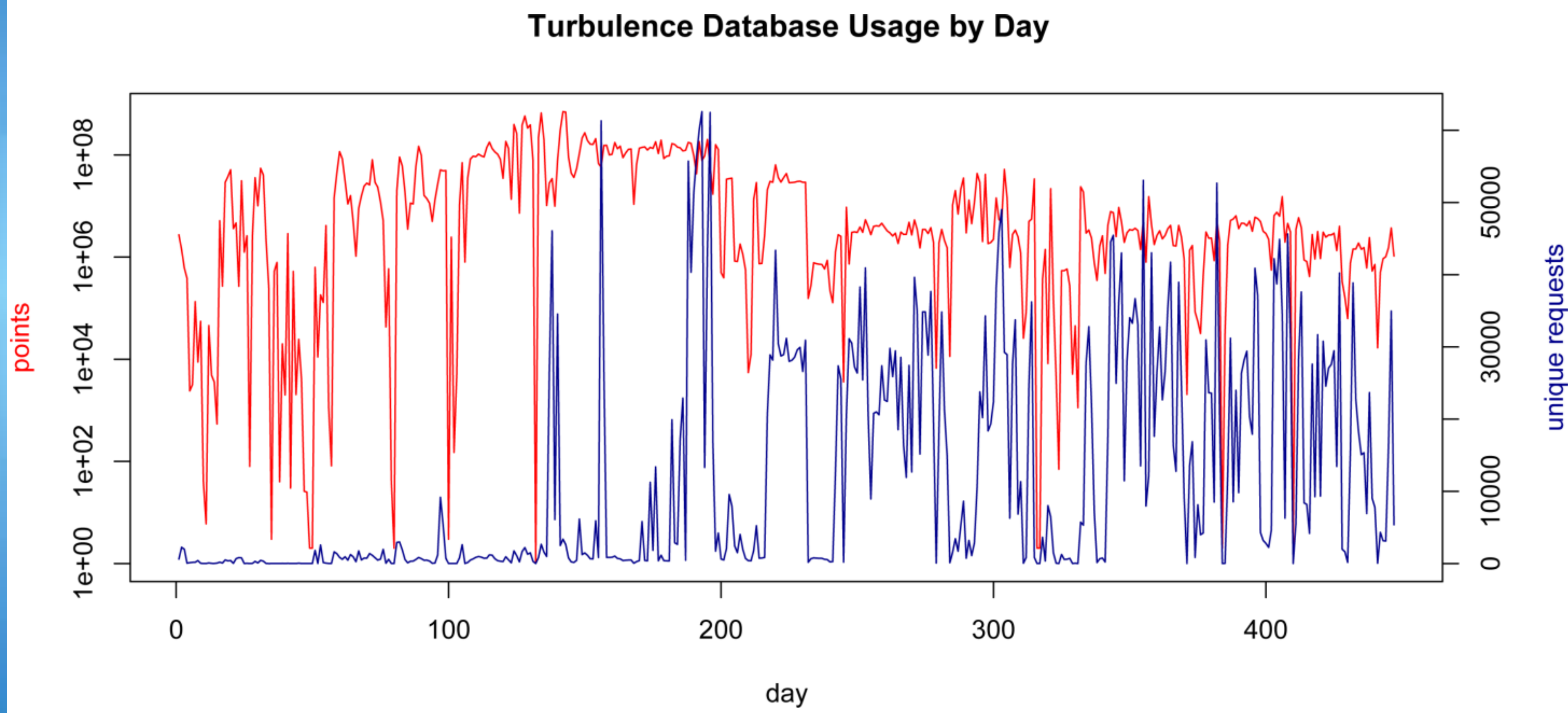
- **Understand the nature of turbulence**
  - *Consecutive snapshots of a large simulation of turbulence: 30TB*
  - *Treat it as an experiment, **play** with the database!*
  - ***Shoot test particles** (sensors) from your laptop into the simulation, like in the movie Twister*
  - *50TB MHD simulation*
  - *Channel flow 100TB, MHD 256TB*
- **New paradigm** for analyzing simulations!





with C. Meneveau (Mech. E), G. Eyink (Applied Math), R. Burns (CS)

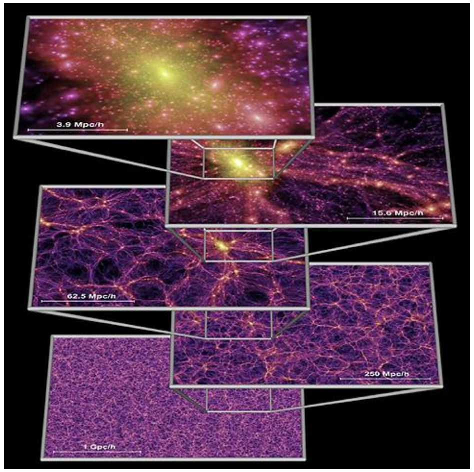
Daily Usage



2015: exceeded 14T points, delivered publicly

Cosmological Simulations

In 2005 cosmological simulations had 10^{10} particles and produced over 30TB of data (Millennium)



http://gavo.mpa-garching.mpg.de/Millennium/

- Build up dark matter halos
- Track merging history of halos
- Use it to assign star formation history
- Combination with spectral synthesis
- Realistic distribution of galaxy types

Today: simulations with $\sim 10^{12}$ particles and almost PB of output are under way (MillenniumXXL, DEUS, Silver River, etc)

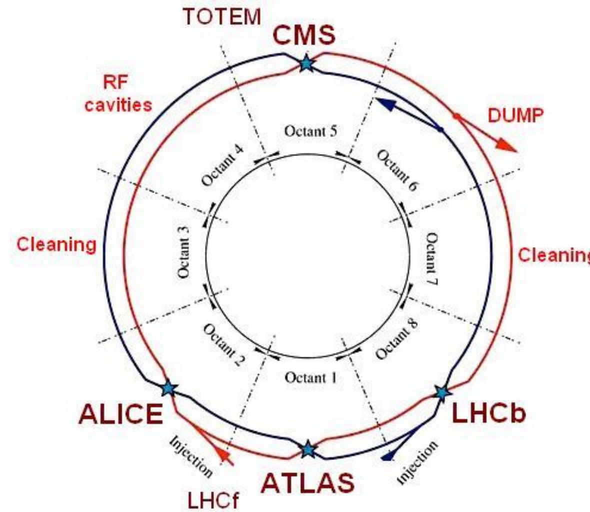
- Hard to analyze the data afterwards -> need DB
- What is the best way to compare to real data?

Numerical Laboratories

- Similarities between Turbulence/CFD, N-body, ocean circulation and materials science
- Differences as well in the underlying data structures
  - *Particle clouds / Regular mesh / Irregular mesh*
- Innovative access patterns appearing
  - *Immersive virtual sensors/Lagrangian tracking*
  - *Posterior feature tagging and localized resimulations*
  - *Machine learning on HPC data*
  - *Joins with user derived subsets, even across snapshots*
  - *Data driven simulations/feedback loop/active control of sims*
- On Exascale everything will be a Big Data problem
- Memory footprint will be >2PB
- With 5M timesteps => 10,000 Exabytes/simulation

LHC Lesson

- LHC has a single data source, $$$$$
- Multiple experiments tap into the beamlines
- They each use **in-situ** hardware triggers to filter data
  - *Only 1 in 10M events are stored*
  - *Not that the rest is garbage, just sparsely sampled*



- Resulting “small subset” analyzed many times **off-line**
  - *This is still 10-100 PBs*
- Keeps a whole community busy for a decade or more

Exascale Simulation Analogy

- Exascale computer running a community simulation
- Many groups plugging their own “triggers” (in-situ), the equivalents of “beamlines”
  - *Keep very small subsets of the data*
  - *Plus random samples from the field*
  - *Immersive sensors following world lines or light cones*
  - *Burst Buffer of timesteps: save precursor of events*
- Sparse output analyzed offline by broader community
- Cover more parameter space and extract more realizations (UQ) using the saved resources

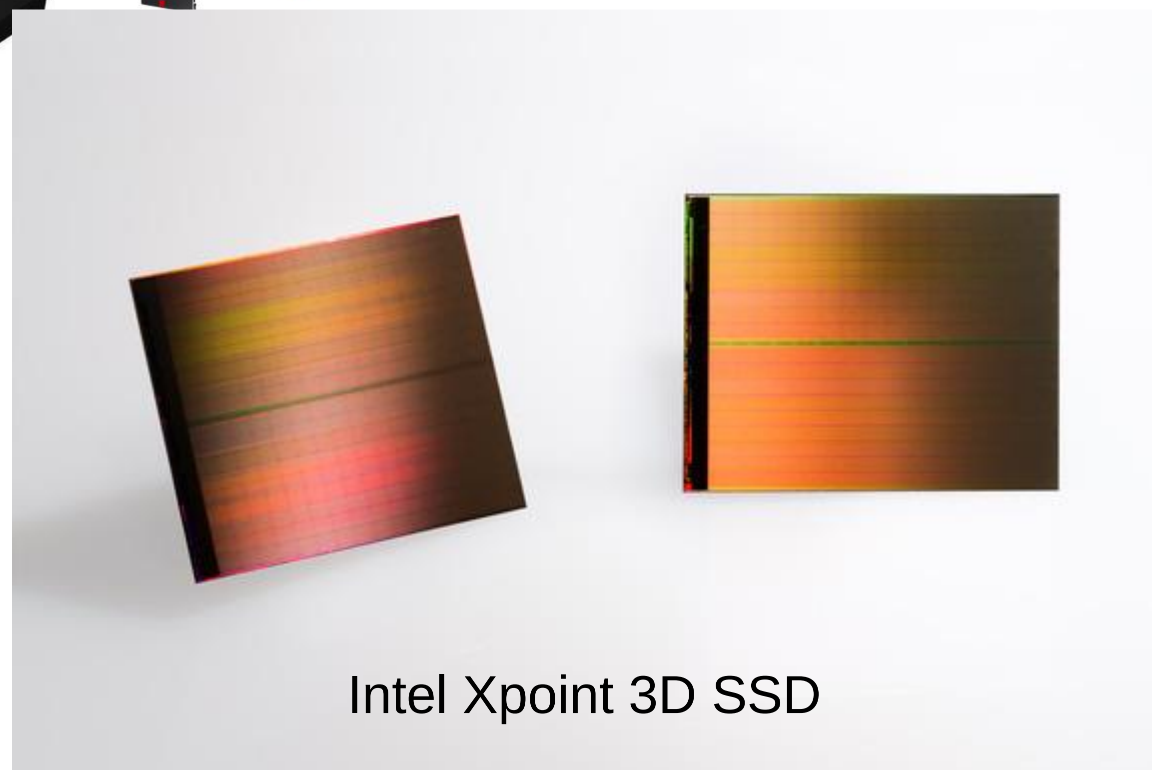
Disruptive Technologies

NEWS

Samsung unveils 15TB SSD based on densest flash memory

MORE LIKE THIS

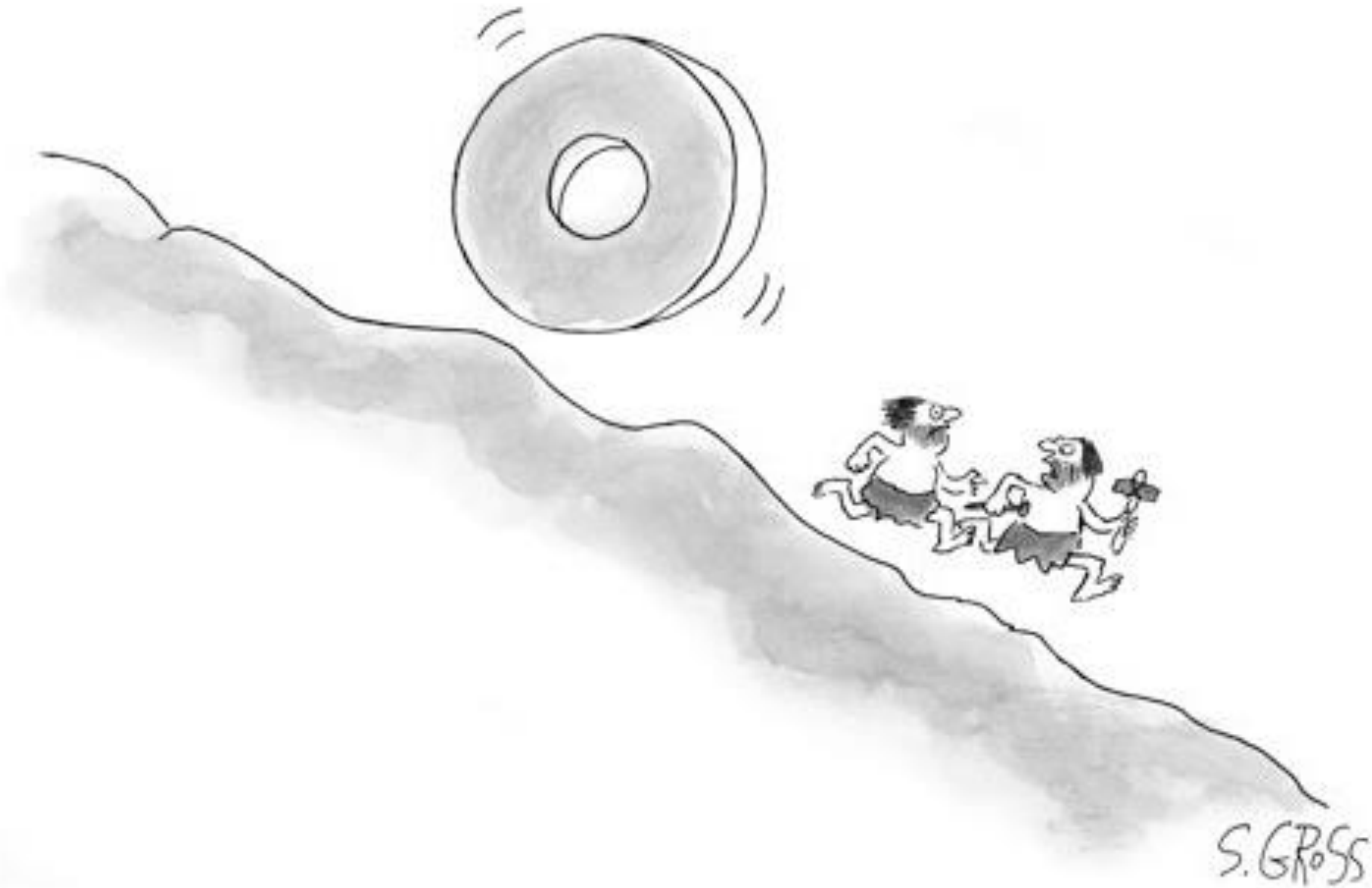
Samsung releases world's first 2TB consumer SSDs



Intel Xpoint 3D SSD

Summary

- Computations even closer to the data
- Cannot afford to store all the incoming data
- Razor sharp tradeoffs, based on algorithms
- Sharp awareness of systematic errors
- Active learning, compressed sensing
- What comes after Data Driven Discoveries (the 4^{th} Paradigm)?
- Exascale simulations become a challenge
- Human aided machine learning becomes part of the scientific process
- Data deluge still getting bigger…



"My next big project is brakes."