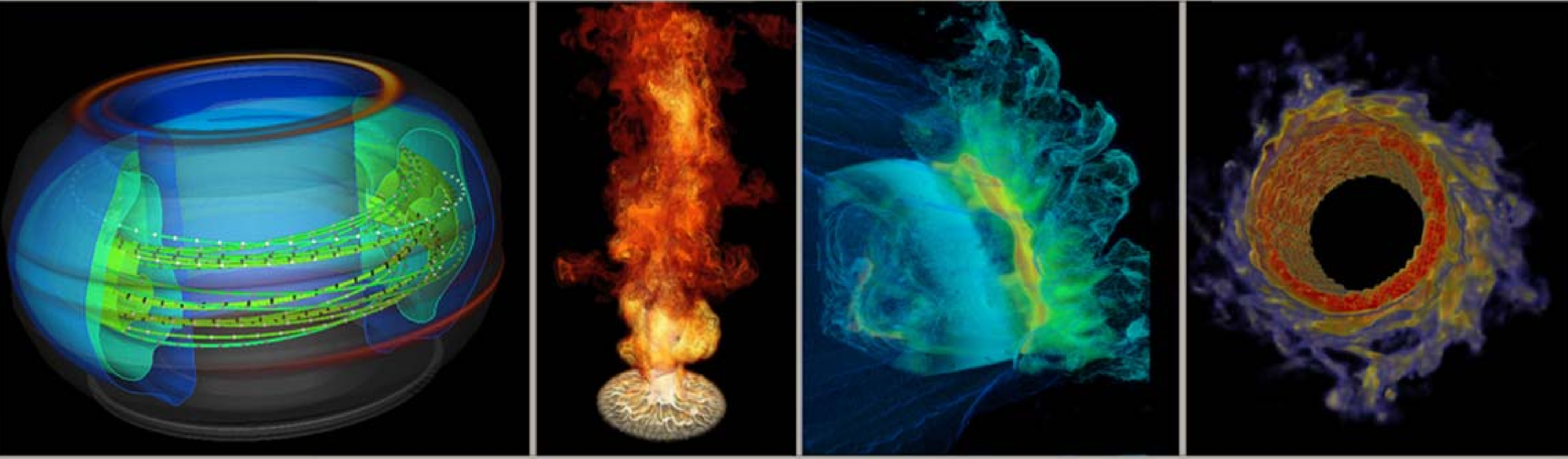
VACET

DOE SciDAC VISUALIZATION AND ANALYTICS CENTER FOR ENABLING TECHNOLOGIES



- IDAV/UC Davis
- LBNL
- LLNL
- ORNL
- SCI/Utah

Occam's Razor and Petascale Visual Data Analysis

E. Wes Bethel

Lawrence Berkeley National Lab

26 Sept 2007

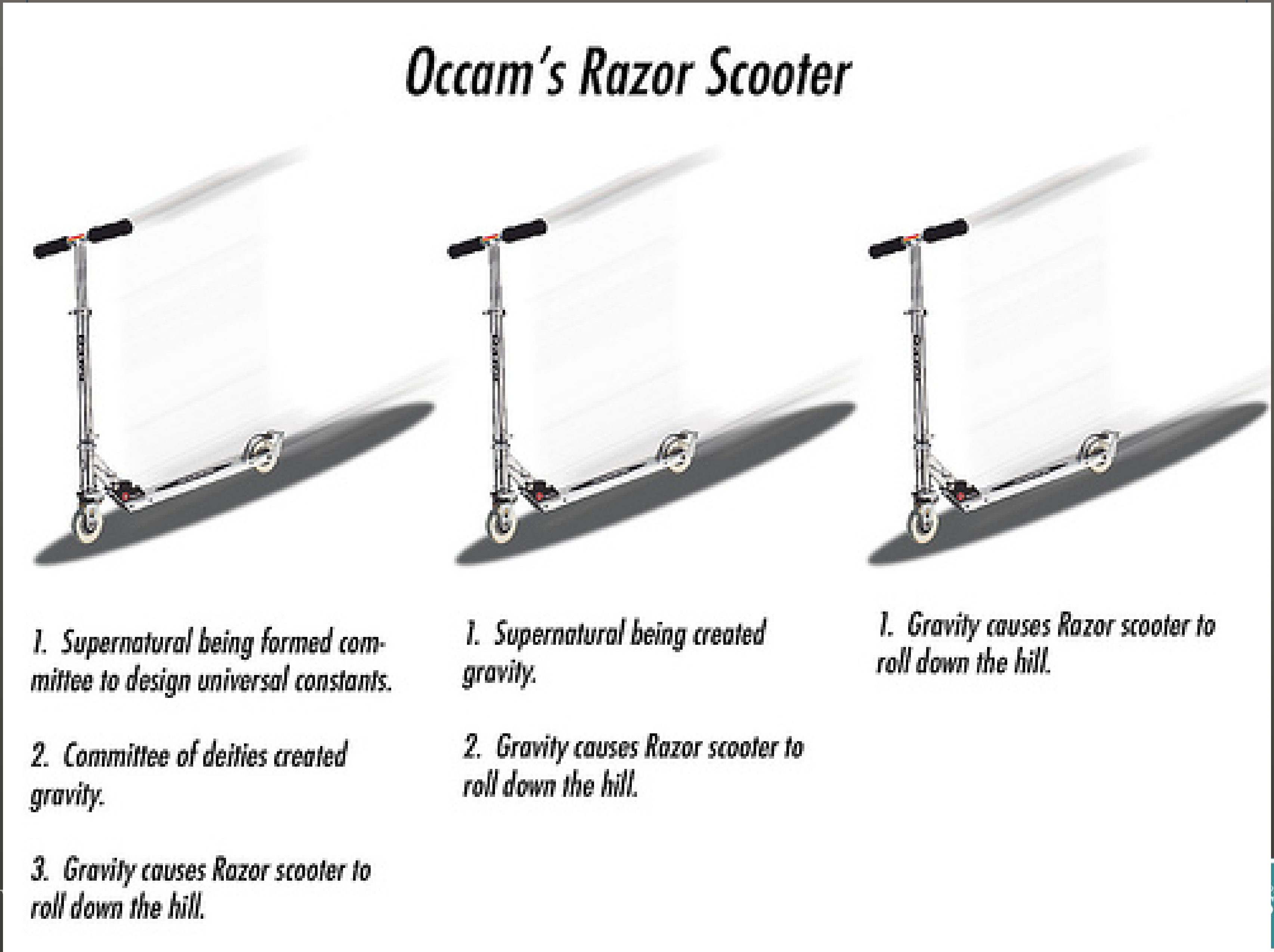
Fall Creek Falls
www.ccs.ornl.gov/workshops/FallCreek/
Center for Computational Sciences
Conference

Outline

- Occam’s Razor and Petascale Visual Data Analysis?
- Challenges:
  - Limited cognitive bandwidth
  - Too much data
  - Not enough time
  - Production deployment
- Conclusion

Occam’s Razor

- Principle: “less is more”
  - Explanation of phenomena should make as few assumptions as possible, eliminate assumptions that make no difference in observable predictions of the explanatory hypothesis or theory.
- Attributed to 14th century English logician Franciscian friar William of Ockham
- Foundation of scientific method
  - “We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances.” – *Sir Isaac Newton*
  - “Everything should be made as simple as possible, but not simpler.” – *Albert Einstein*



Occam's Razor Scooter

1. Supernatural being formed committee to design universal constants.

2. Committee of deities created gravity.

3. Gravity causes Razor scooter to roll down the hill.

1. Supernatural being created gravity.

2. Gravity causes Razor scooter to roll down the hill.

1. Gravity causes Razor scooter to roll down the hill.

Occam’s Razor

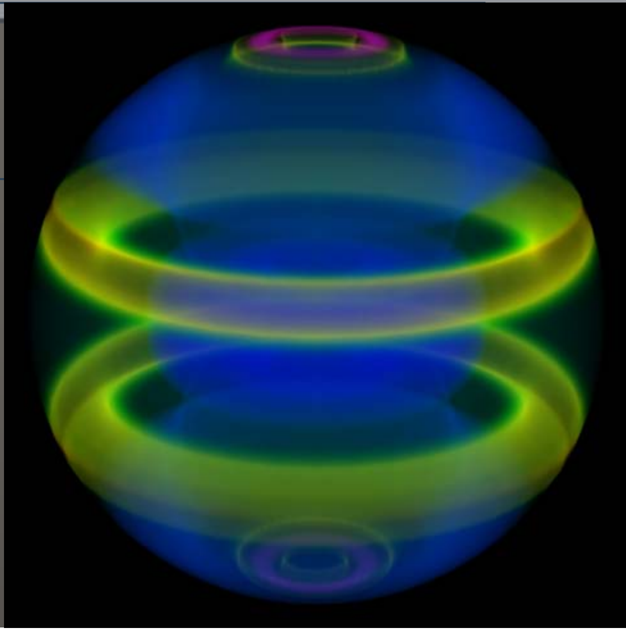
- Physics
  - – Einstein’s theory of special relativity vs. Lorentz’s theory that rulers contract and clocks slow down when in motion through the Ether.
    - Equations are “the same.”
    - Ether not detectable with Lorentz’s equations.
  - – Justification of Heisenberg’s Uncertainty Principle in quantum mechanics
    - Impossible to know exact position and momentum of a particle at the same time.
    - Why?
    - The process of observation influences the observed.

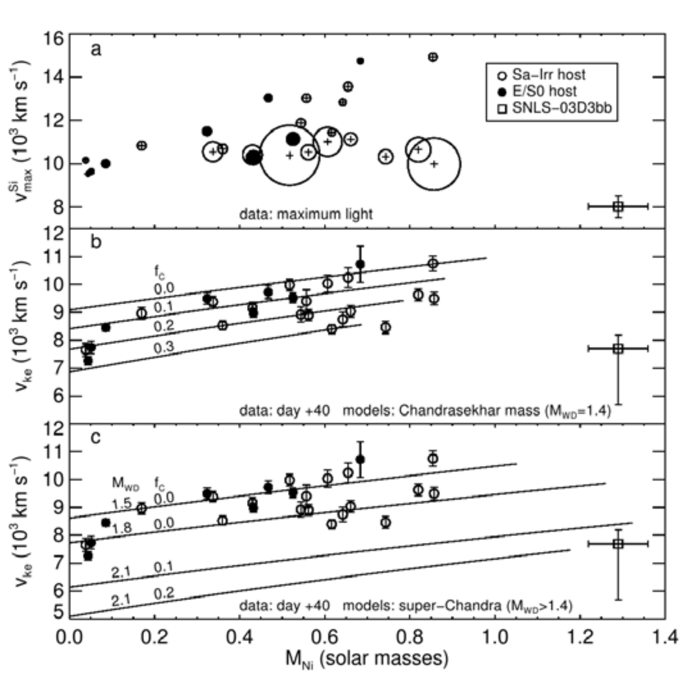
Occam’s Razor and Petascale Visual Data Analysis?

- What is the minimum needed to do “the job?”
- What is “the job?”
  - Visualization: a flexible, powerful knowledge discovery technology/set of methodologies.
  - Can do basically “anything”
  - “Any sufficiently advanced technology is indistinguishable from magic.” – *Arthur C. Clarke*
- With infinite flexibility, must choose carefully

Visualization Use Models

- Presentation visualization
  - You know what’s there and want to show it to someone else
- Analytical Visualization
  - You know what you are looking for
- Discovery Visualization
  - You have no idea what you’re looking for

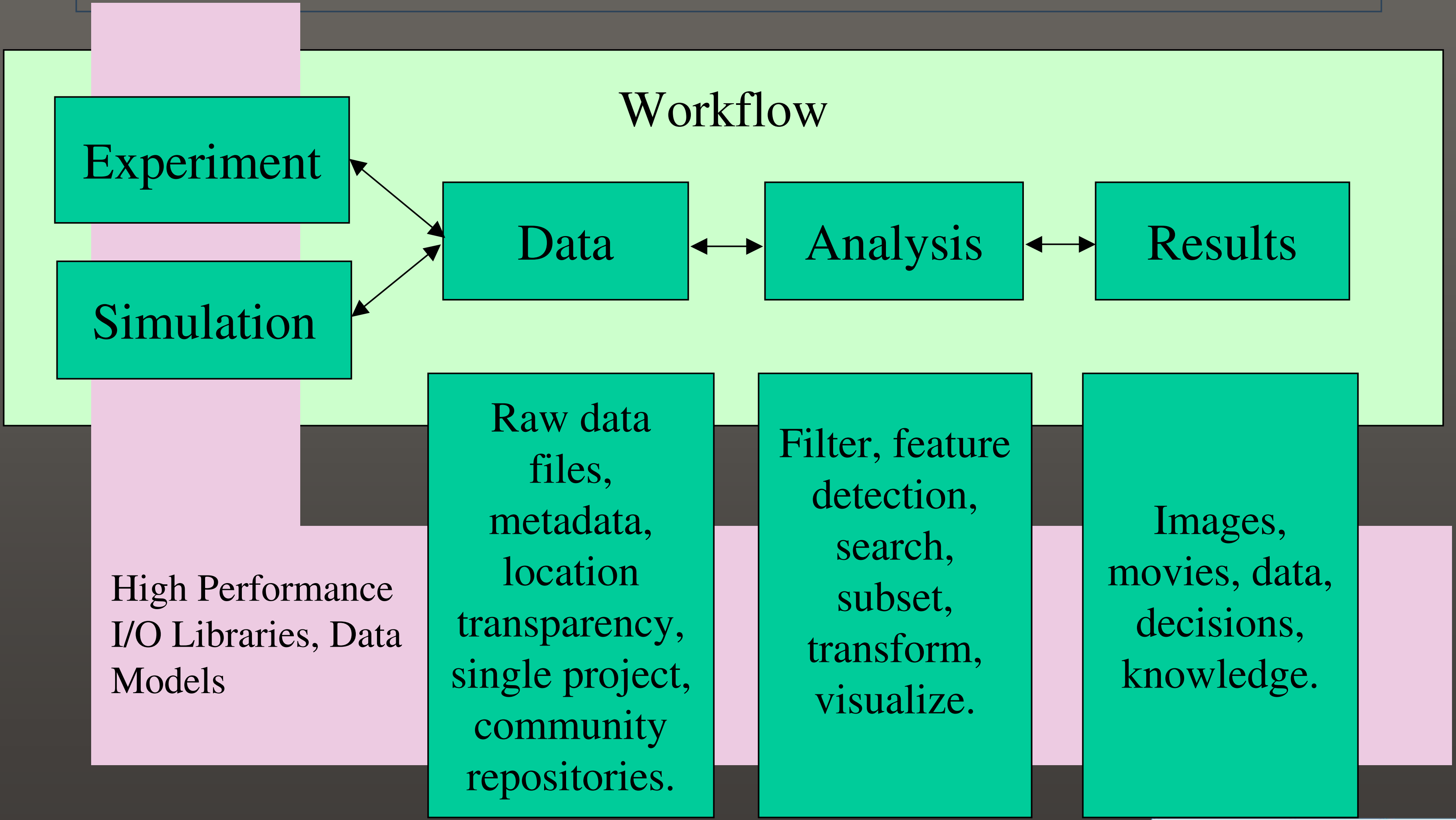



a
b
c
○ Sa–Irr host
● E/S0 host
□ SNLS–03D3bb
data: maximum light
data: day +40 models: Chandrasekhar mass (M_{WD}=1.4)
data: day +40 models: super–Chandra (M_{WD}>1.4)
v^{Si}_{max} (10^3 km s^{-1})
v_{ke} (10^3 km s^{-1})
v_{ke} (10^3 km s^{-1})
f_c
0.0
0.1
0.2
0.3
M_{WD} f_c
1.5 0.0
1.8 0.0
2.1 0.1
2.1 0.2
16
14
12
10
8
12
11
10
9
8
7
6
12
11
10
9
8
7
6
5
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
M_{Ni} (solar masses)

Brief Digression – Theory of Education

- Three stages of learning:
  - Romance
    - “Play and wonder,” which are initial steps towards
  - Precision
    - Deeper understanding leading to
  - Generalization
    - Broad insights

Map of the Problem Space



Workflow

Experiment

Simulation

Data

Analysis

Results

High Performance I/O Libraries, Data Models

Raw data files, metadata, location transparency, single project, community repositories.

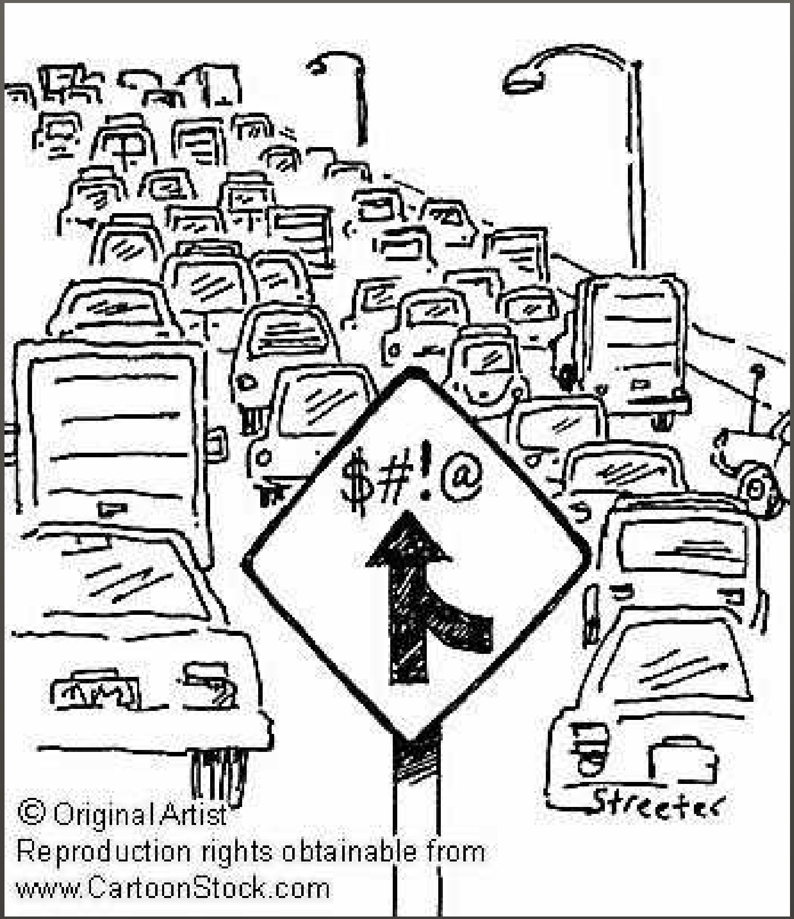
Filter, feature detection, search, subset, transform, visualize.

Images, movies, data, decisions, knowledge.

The Big Challenge

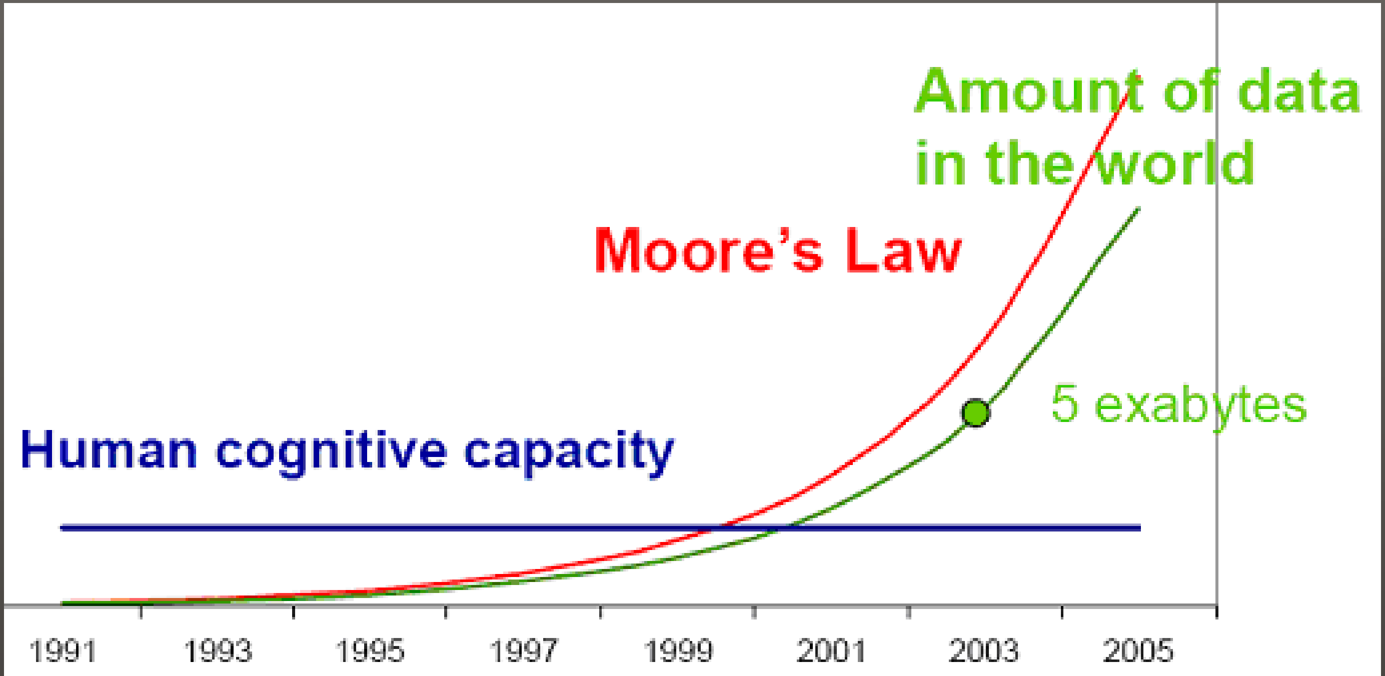
- How to effectively enable knowledge discovery at the petascale given:
  - Limited cognitive bandwidth
  - Too much data
  - Not enough time
  - Deploy *production quality* tools for scientific communities
- Rest of discussion examines these topics in more detail.

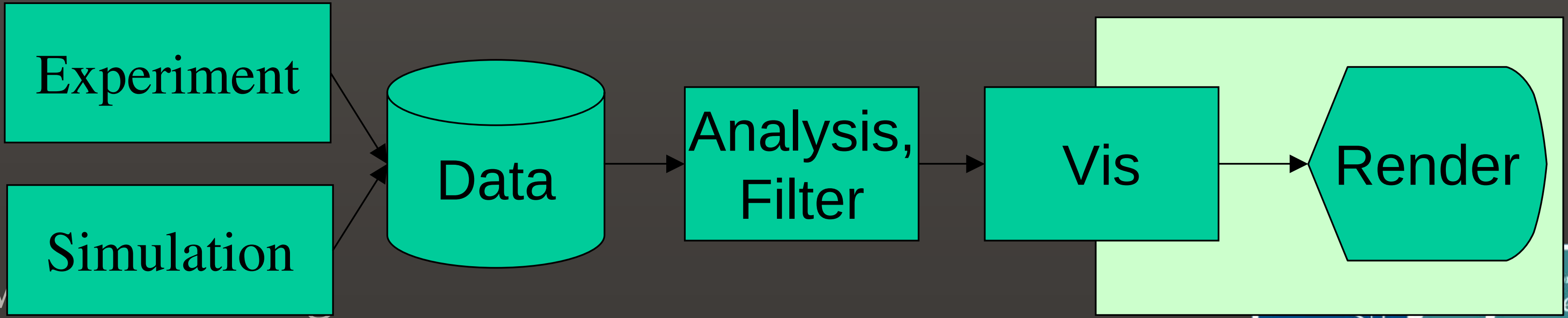
Limited Cognitive Bandwidth



© Original Artist
Reproduction rights obtainable from
www.CartoonStock.com

The Cognitive Bandwidth Challenge





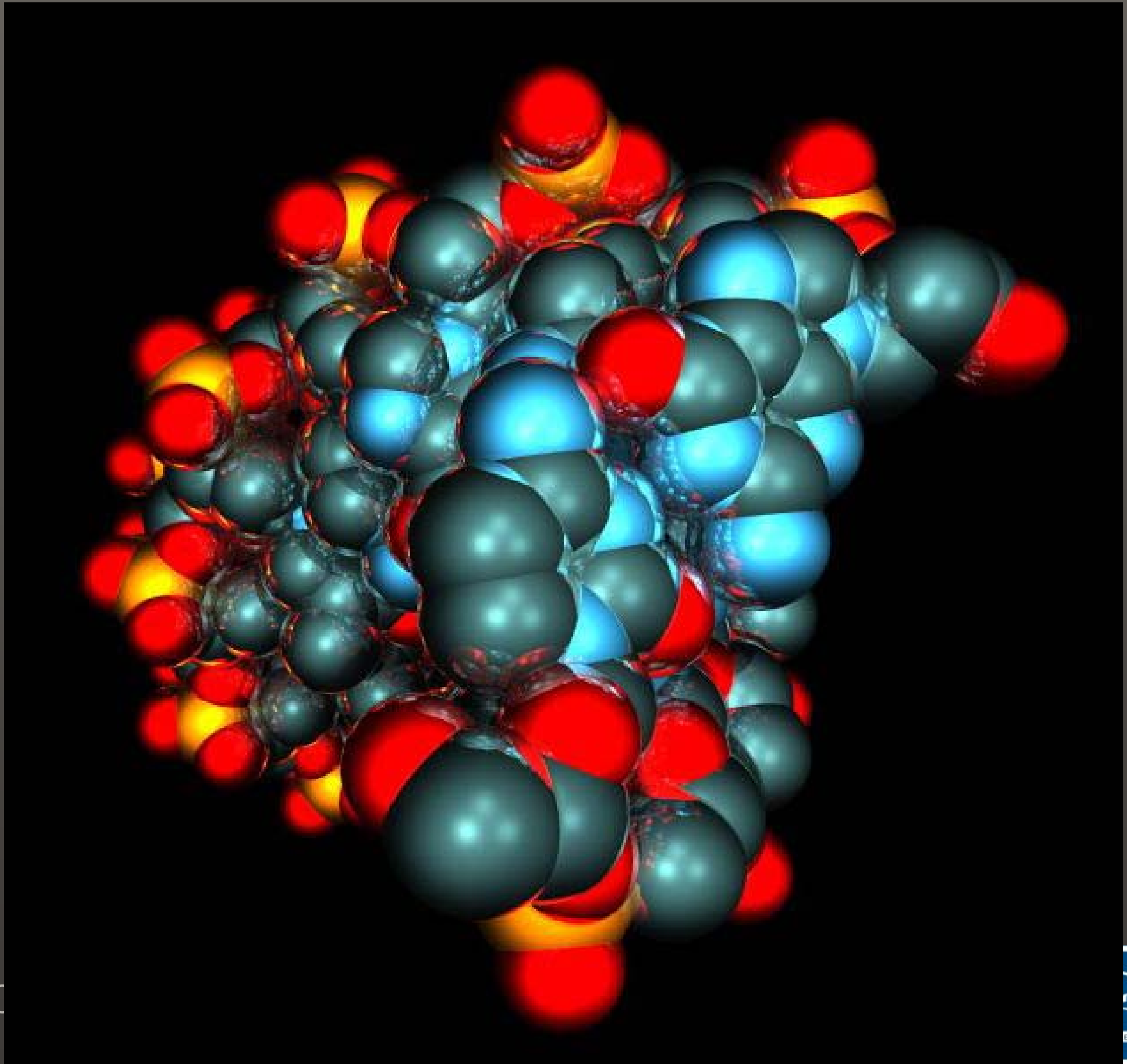
Cognitive Bandwidth Challenge

- “Picture worth a 1000 words.” F. Barnard, 1921, N. Bonaparte c. 1800.
  - At a glance, humans are great at perceiving trends, anomalies, deriving meaning and understanding from seemingly “random” data.
- “One equation is worth 1000 pictures.” J. Bell, 2006.

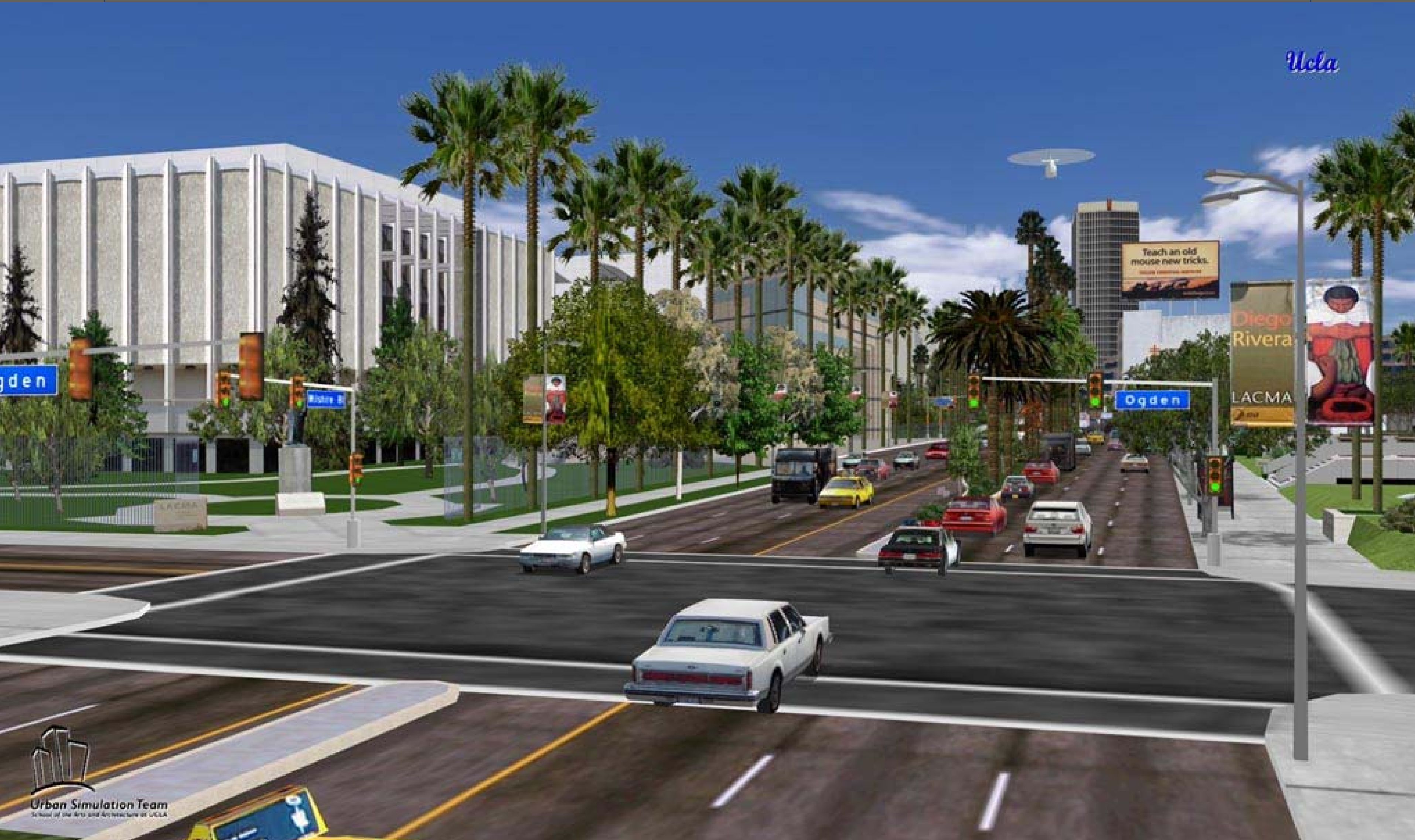
Cognitive Bandwidth Challenge

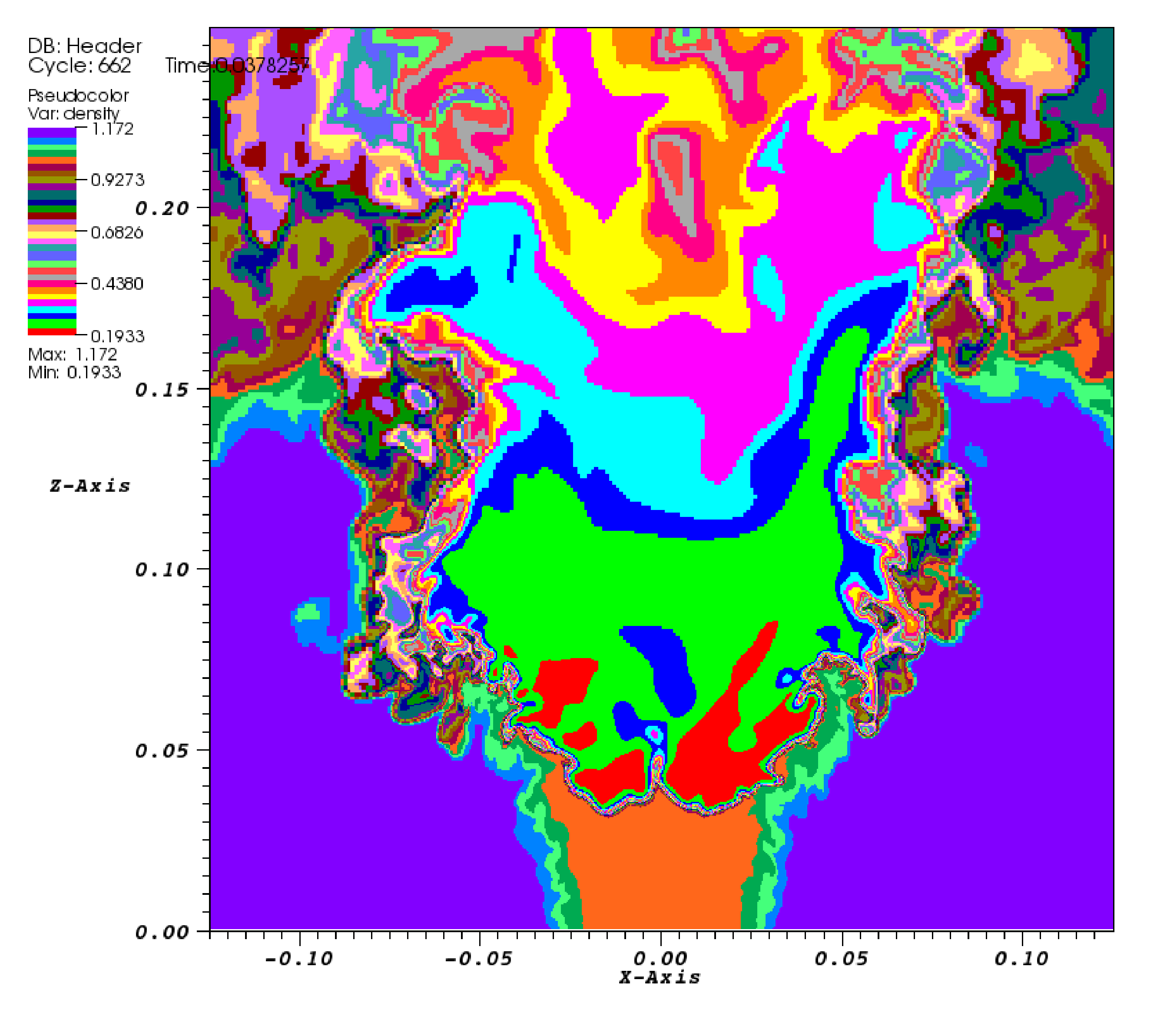
- Can’t “see” a TB, much less a PB
  - – Mesh cells map to subpixels
  - – Depth complexity
  - – Spatial complexity
- Occam might say:
  - – Show only what is important to more closely match cognitive capacity.
- Counterexample:
  - – Visual simulation and ultra high resolution displays

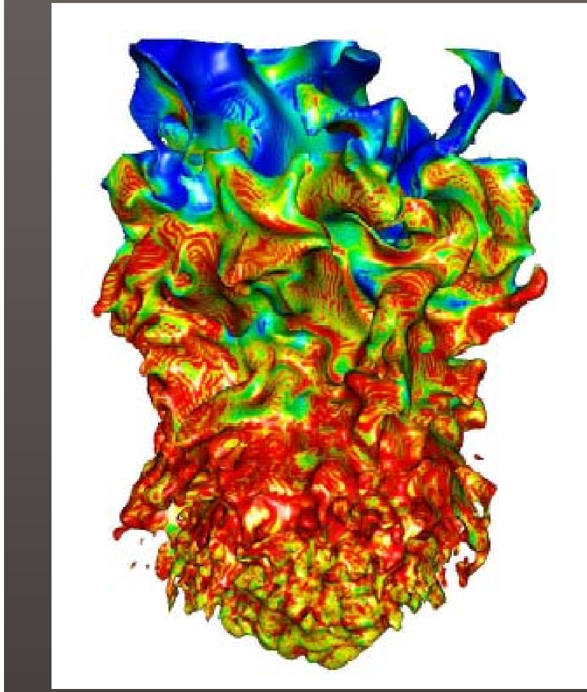
Photorealistic Rendering of Molecules



Visual Simulation – UCLA’s UST


Ucla
gden
Wilshire Bl
Ogden
Teach an old mouse new tricks.
Diego Rivera
LACMA
Urban Simulation Team
School of the Arts and Architecture at UCLA


DB: Header
Cycle: 662
Time:0.0378257
Pseudocolor
Var: density
1.172
0.9273
0.6826
0.4380
0.1933
Max: 1.172
Min: 0.1933
Z-Axis
0.20
0.15
0.10
0.05
0.00
-0.10
-0.05
0.00
0.05
0.10
X-Axis

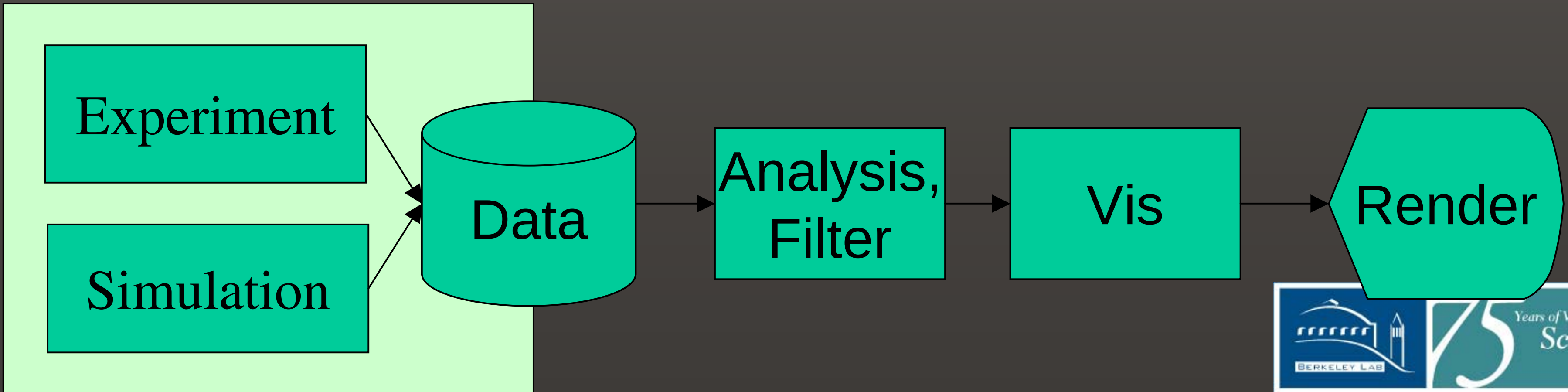


Data Tsunami Challenge

- Science bottleneck: management of and knowledge discovery from large collections of scientific data

Reducing I/O (and Compute) Load

- Adaptive Mesh Refinement
- Compression
- Subsampling, statistical models
- Other issues: storage, access, high performance I/O, movement, sharing, provenance, ontology, …

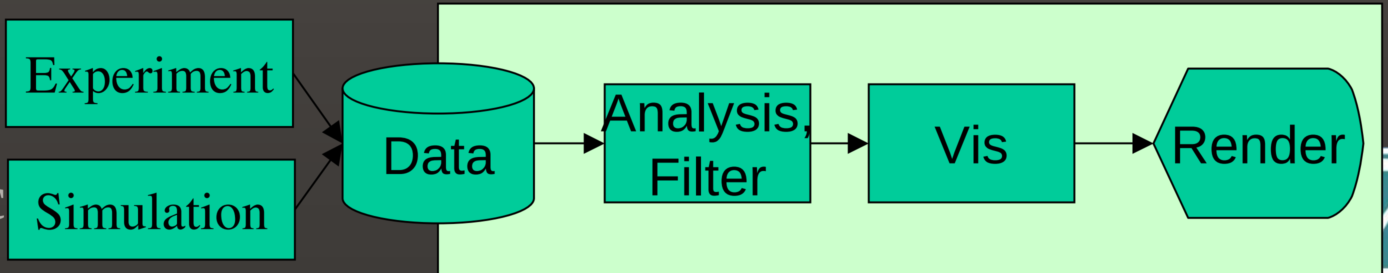


Another Approach – In-place Analysis

- Idea: do visualization and analysis in the simulation, save the analysis results rather than the simulation data.
  - Pro:
    - Avoids I/O and storage issue by having the simulation perform analysis and save only analysis results.
  - Cons:
    - No possibility of unexpected discovery
    - Assumes image is final analysis result
    - Simulation code “feature creep”

Another Approach – Focus on Interesting

- Restrict visual analysis and associated I/O, processing to “interesting data.”
  - Pros:
    - Applicable to many problem domains, including both experimental and computational data
    - Applicable to all visualization use modalities
  - Cons:
    - Still have the I/O and storage problem
- Comment
  - Good fit for human learning



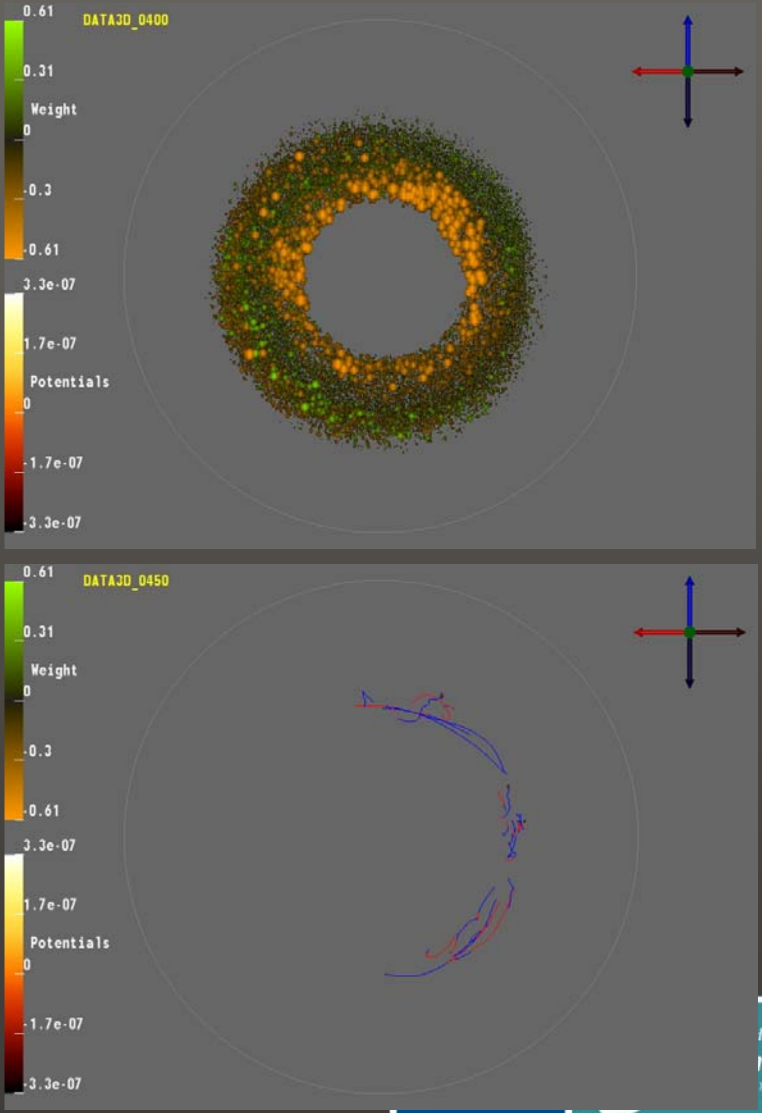
Query-Driven Visualization

- Focus visualization processing on subsets of data deemed to be “interesting.”
  - “Interesting” is something the user needs to define.

- Challenges
  - How to define “interesting.”
    - Formulation of definition (domain-specific).
    - Expression of definition (semantic).
  - Find interesting data quickly (data management).
  - Effective visual presentation of “interesting data” (visualization).
  - Architectures/deployment that complements existing visualization algorithms and applications (computer science).

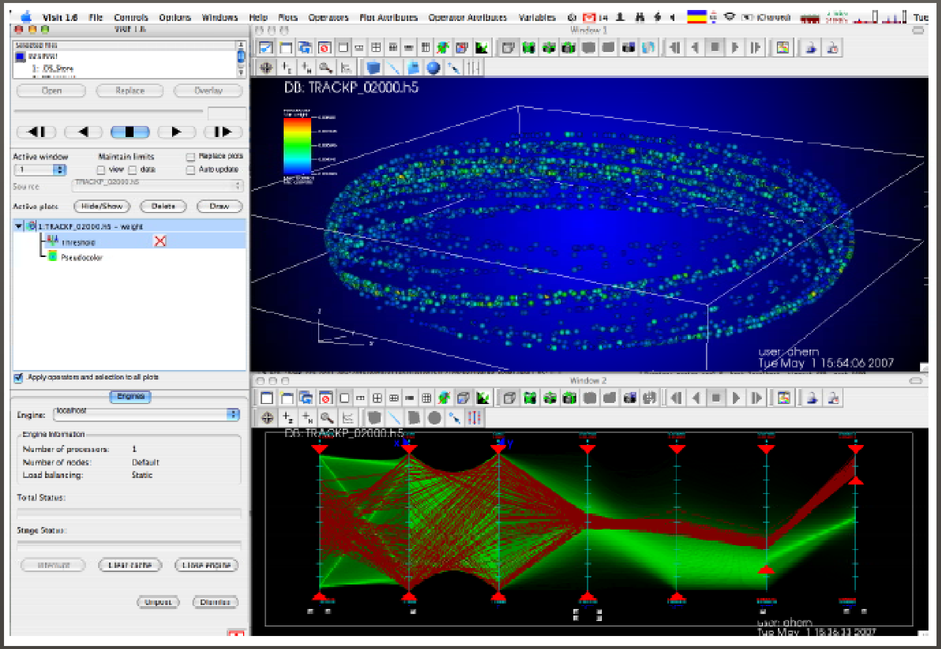
QDV and Fusion

- GTC is a PIC code for modeling microturbulence
  - Top: all particles from a single timestep
  - Bottom: particles that undergo “trapping” at least 20 times.
  - *New VACET project: statistical downsampling to reduce I/O load.*



QDV and Fusion

- Application: fusion microturbulence
- Objective: interactive multidimensional filtering to locate and analyze interesting phenomena.
- Result: capability in production software (VisIt).



☑ Apply operators and selection to all plots

Engine: localhost

Engine information

Number of processors: 1

Number of nodes: Default

Load balancing: Static

Total Status:

Stage Status:

QDV at the Petascale

- Problem: want to do better than O(*n*) for data analysis (where *n* is the size of the data).
- Solution(s):
  - Use state-of-the-art index/query. VACET and SDM Center integrating such technology at the data I/O layer in a way that is transparent to simulation code developers.
    - Note: tree-based index/query systems suffer from "*The Curse of Dimensionality*"; compressed bitmap indices do not:
    - O(N**D) vs. O(N*D): Trees vs. CBI
  - Leverage this capability in visual analysis tools, can provide assistance for use in other types of tools.

Not Enough Time Challenge

- Topics
  - Make it go faster
    - Software architecture for production quality high performance visualization on DOE’s leading computational platforms (Hank’s talk).
  - Make it do more, make it do what I want.
    - Analysis and visualization
    - Production analytics pipelines
    - Community-centric

Visualization and Analysis

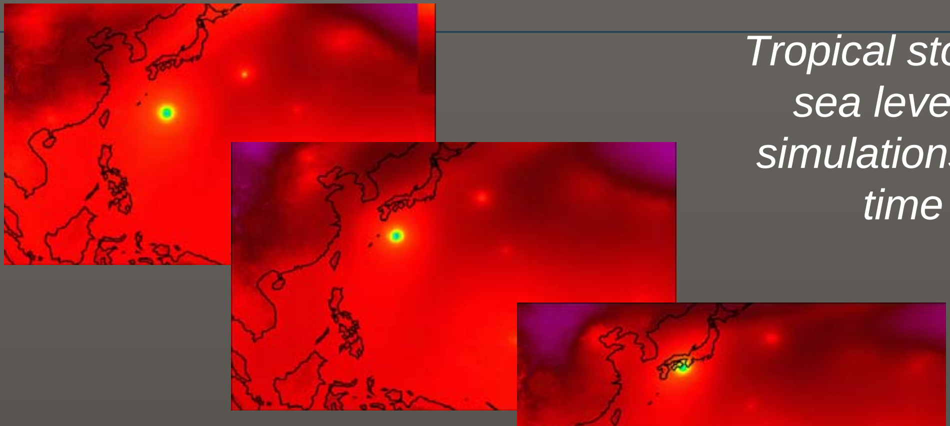
- Objective: convey deeper meaning than possible with "data exploration"
  - Relationships between variables
  - Characteristics of data
  - Compute, display and analyze "features"

Analysis and Visualization

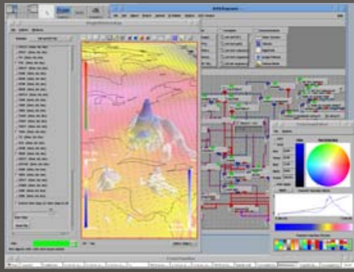
- Premise: find and analyze “features” in data.
- Methods of location and analysis vary:
  - PCA, ICA, Machine Learning, Support Vector Machines, etc.
  - Topological analysis.
- Impact:
  - Quantitative analysis
  - Traction on “big data problems”
  - One potential basis for comparative analysis (rather than “chi-by-eye”)
- Examples

Analysis and Visualization – Climate

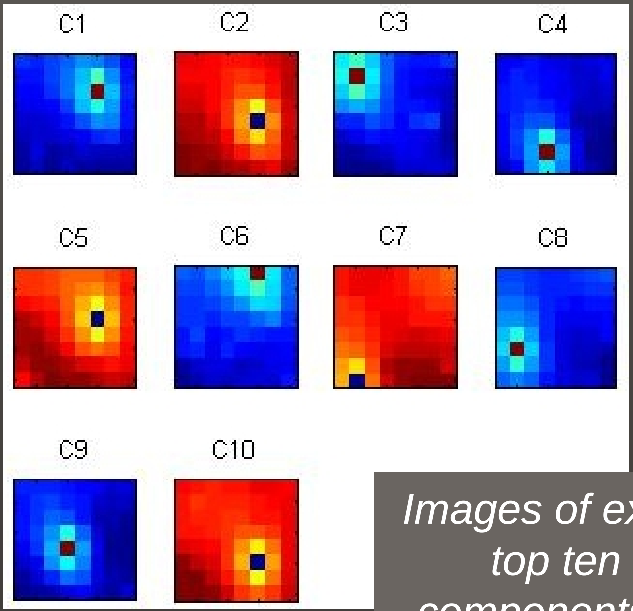
Extracted features can be used as templates for finding similar features.

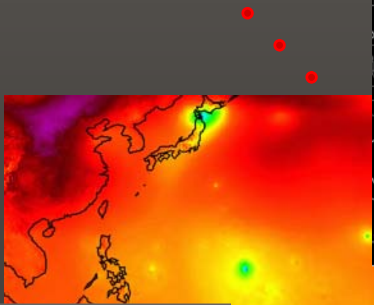


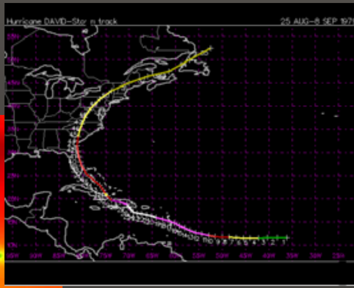
Tropical storm visible in sea level pressure simulations at multiple time steps.



In this case, the features were variations on rotating low-pressure systems. This was not assumed a priori.



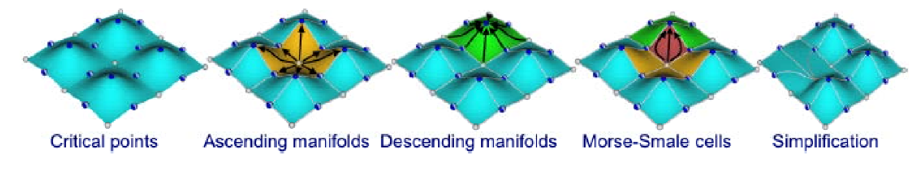


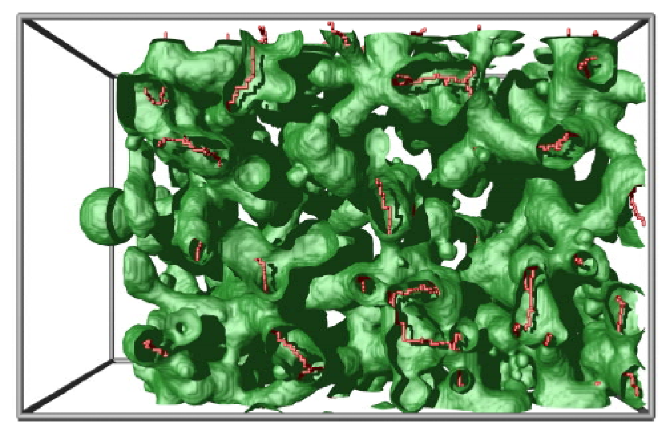


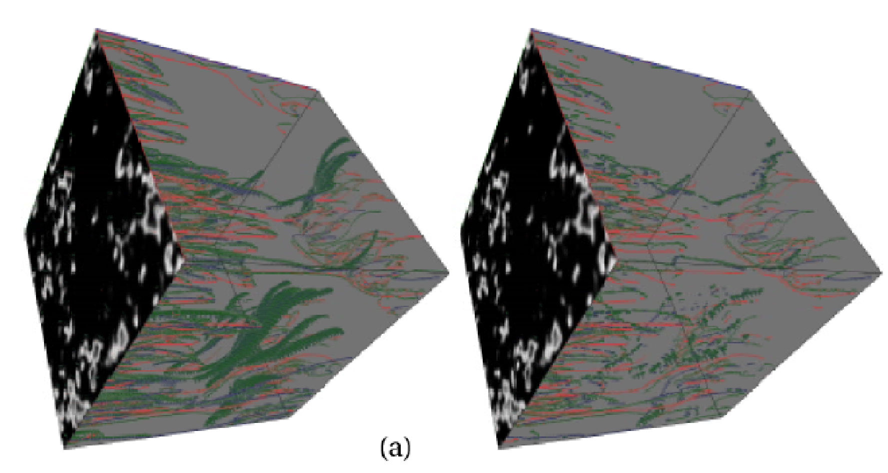
Images of extracted features: top ten independent components were extracted from set of all 8x8 subimages.

Analysis and Visualization – Topology

- Channel structures in porous media : green isosurface separates solid material and empty space; red curves, connecting maxima and 2-saddles, represent channels.
- Combustion kernel feature identification, tracking and analysis. Kernels appear, merge, and extinguish over time. Why? How many?

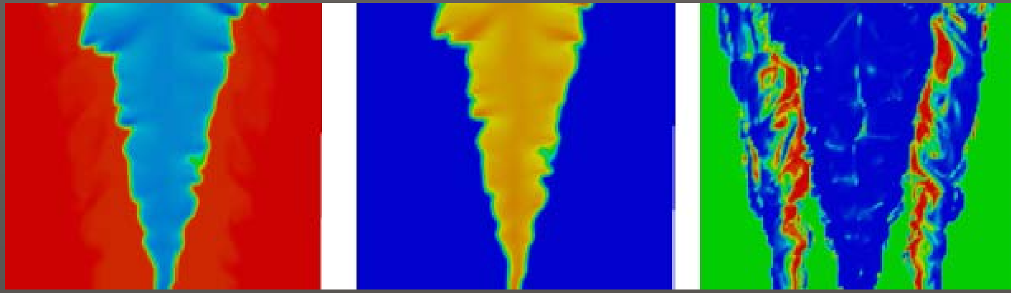


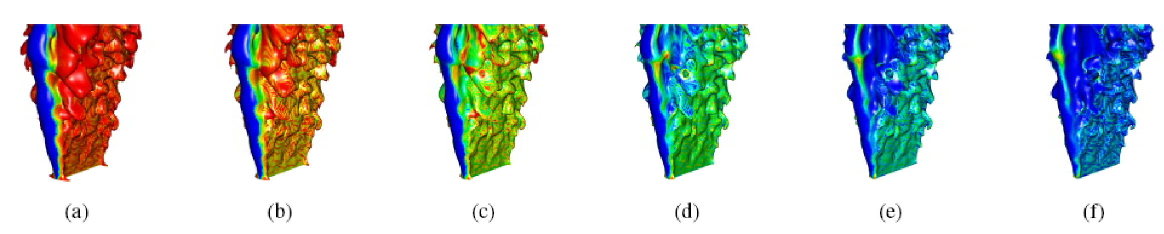




Discovering Relationships

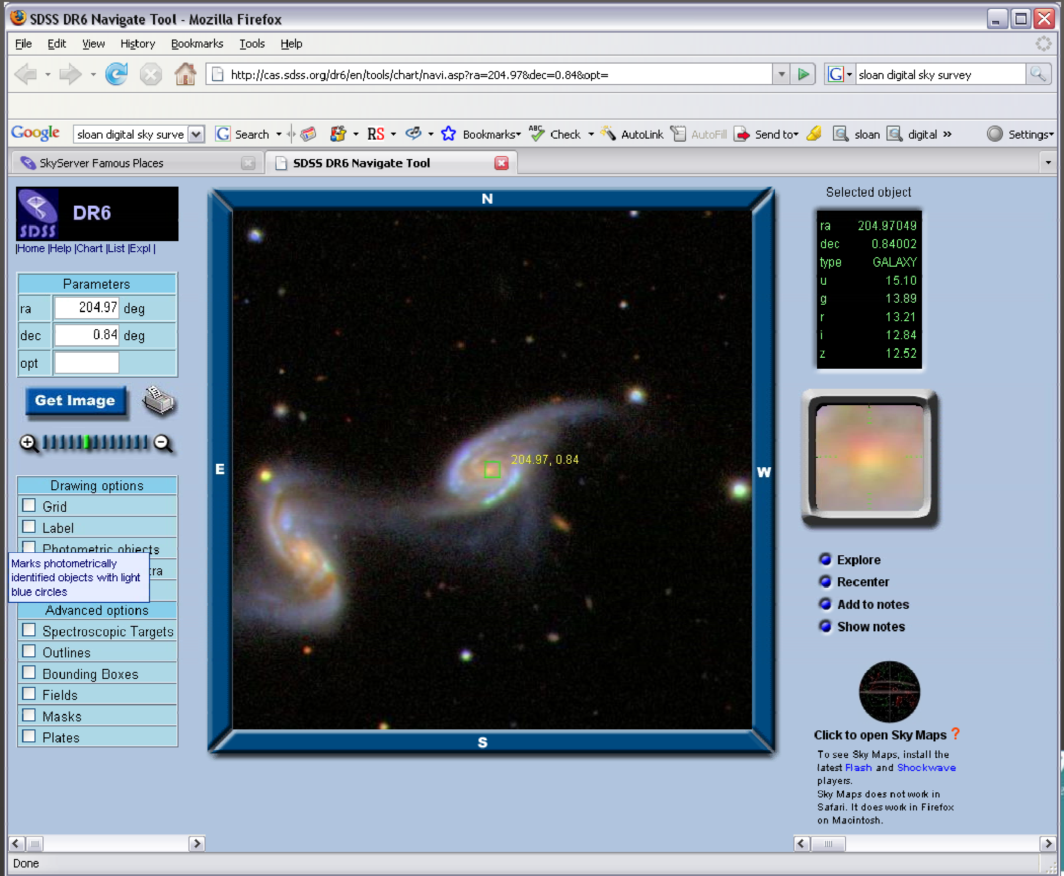
- How are variables related to one another?
- O_2 and CO_2 concentration (left and middle), correlation field (right)
- Same idea, but in 3D and mapped onto varying isotherms.





Production Analytics

- Visual Data Cartography
  - Google Earth, Google Sky, Sloan Digital Sky Survey, etc.



http://cas.sdss.org/dr6/en/tools/chart/navi.asp?ra=204.97&dec=0.84&opt=

sloan digital sky survey

Parameters

ra	204.97	deg
dec	0.84	deg
opt		

Drawing options

- □ Grid
- □ Label
- □ Photometric objects

Marks photometrically identified objects with light blue circles

Advanced options

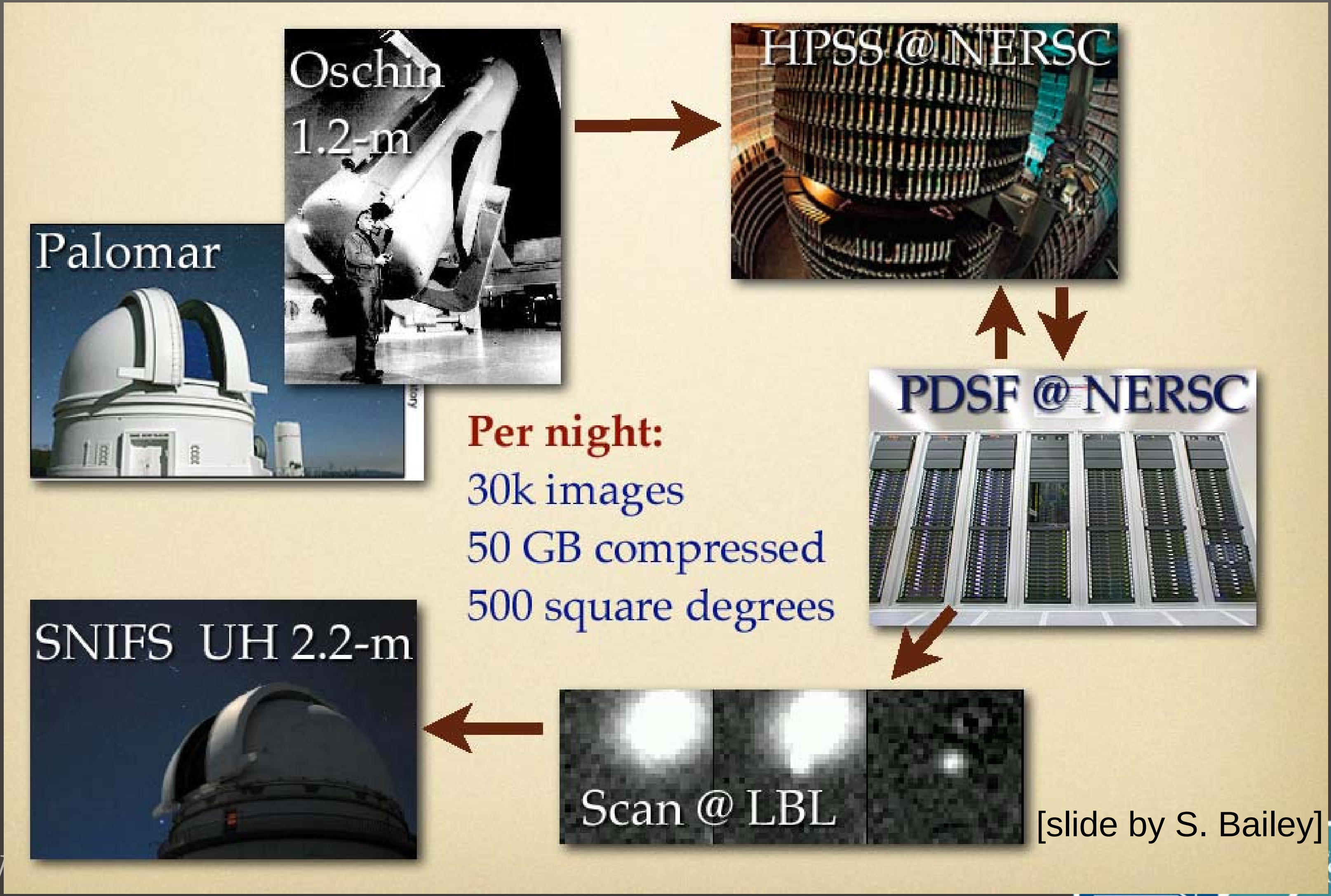
- □ Spectroscopic Targets
- □ Outlines
- □ Bounding Boxes
- □ Fields
- □ Masks
- □ Plates

Selected object

ra	204.97049
dec	0.84002
type	GALAXY
u	15.10
g	13.89
r	13.21
i	12.84
z	12.52

To see Sky Maps, install the latest Flash and Shockwave players.
Sky Maps does not work in Safari. It does work in Firefox on Macintosh.

Nearby Supernova Factory: Production Analytics



Per night:

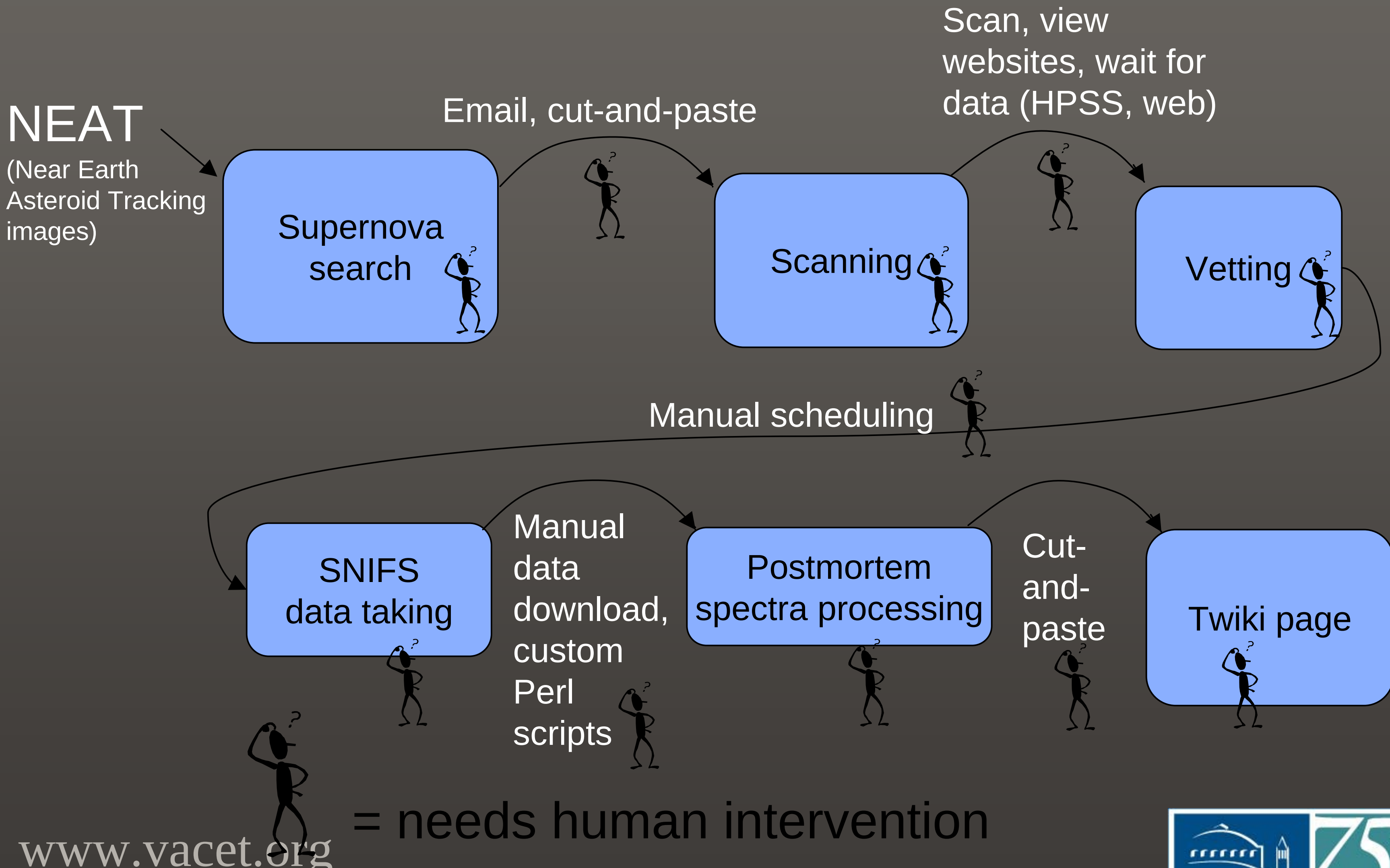
30k images

50 GB compressed

500 square degrees

[slide by S. Bailey]

SNFactory Pipeline circa 2005



NEAT

(Near Earth Asteroid Tracking images)

Supernova search

Email, cut-and-paste

Scanning

Scan, view websites, wait for data (HPSS, web)

Vetting

Manual scheduling

SNIFS data taking

Manual data download, custom Perl scripts

Postmortem spectra processing

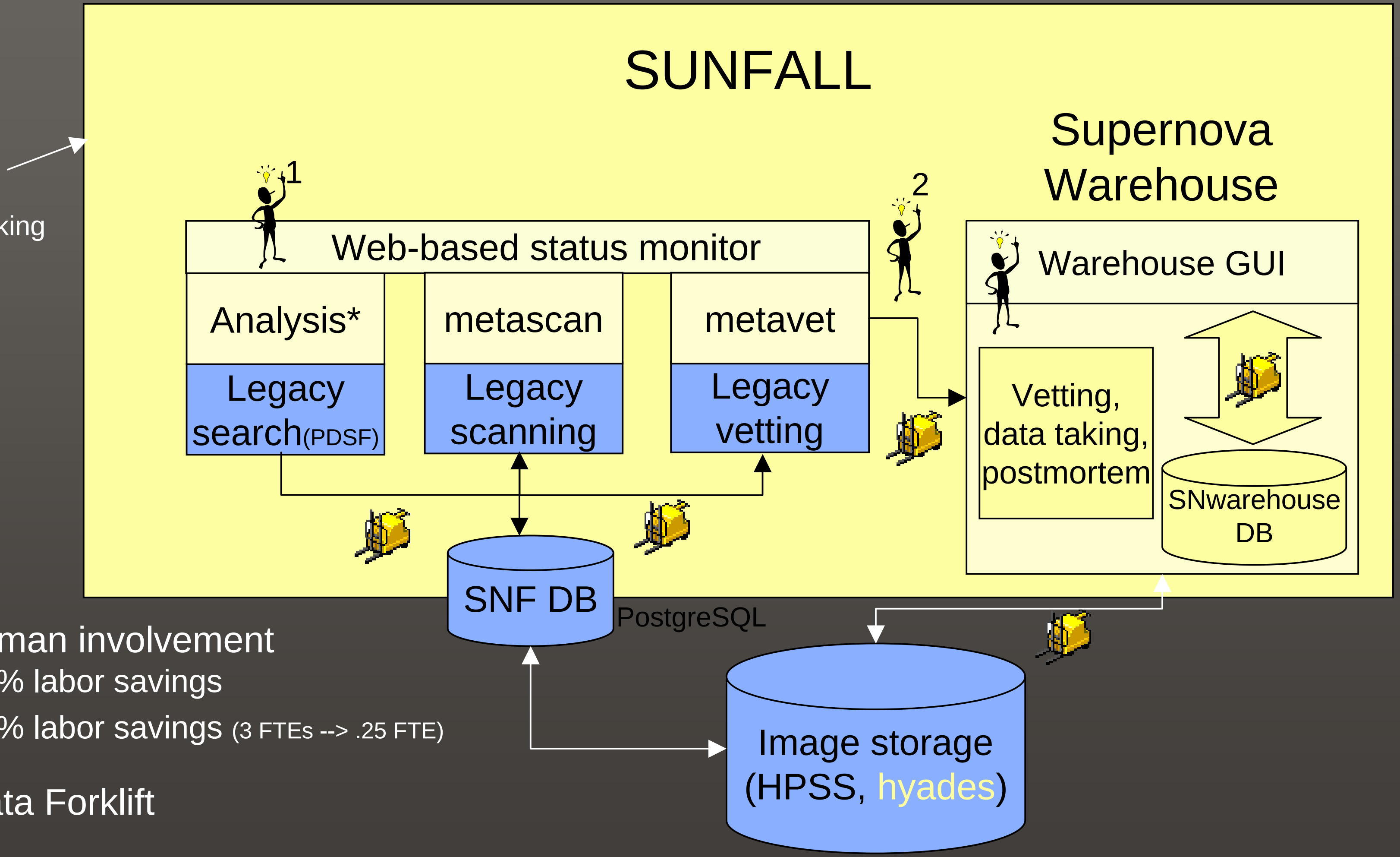
Cut-and-paste

Twiki page

= needs human intervention

SNFactory Pipeline – Present Day

NEAT
(Near Earth Asteroid Tracking images)



= Human involvement

[1]60% labor savings

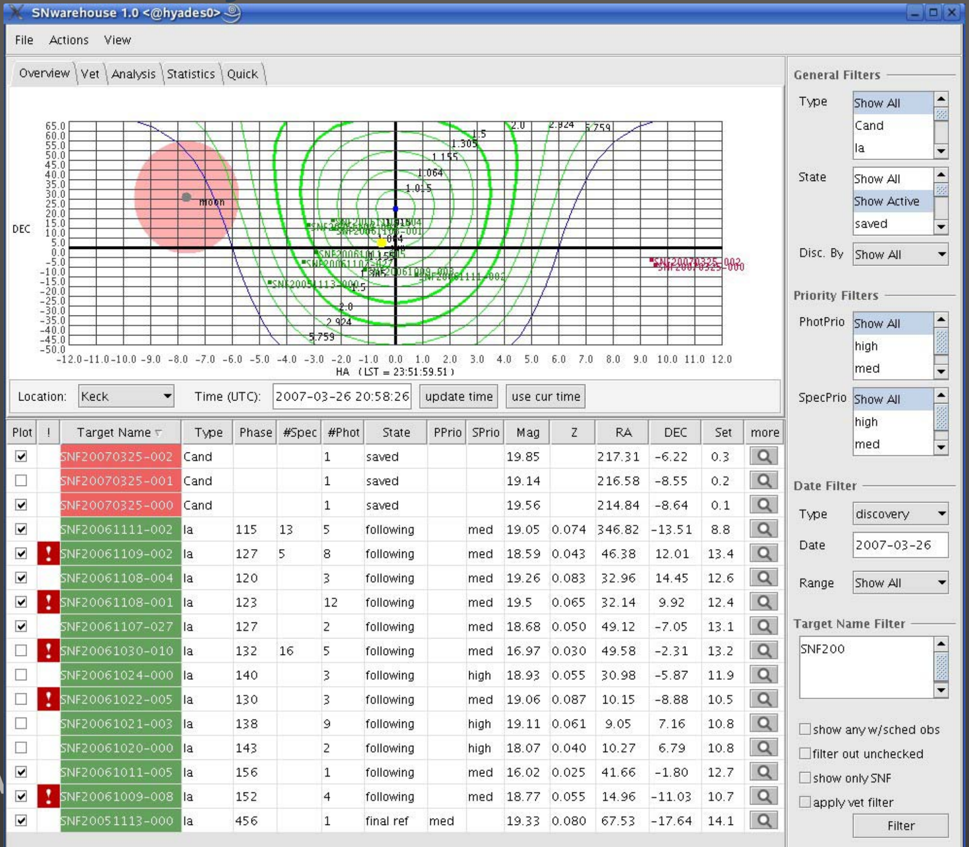
[2]90% labor savings (3 FTEs --> .25 FTE)

= Data Forklift

*new image and machine learning algorithms

Sunfall: Supernova Warehouse

- A comprehensive supernova data management, workflow visualization, collaborative scientific analysis tool



Location: Keck

Time (UTC): 2007-03-26 20:58:26

Plot	!	Target Name	Type	Phase	#Spec	#Phot	State	PPrio	SPrio	Mag	Z	RA	DEC	Set	more
☑		SNF20070325-002	Cand			1	saved			19.85		217.31	-6.22	0.3	🔍
☐		SNF20070325-001	Cand			1	saved			19.14		216.58	-8.55	0.2	🔍
☑		SNF20070325-000	Cand			1	saved			19.56		214.84	-8.64	0.1	🔍
☑		SNF20061111-002	Ia	115	13	5	following		med	19.05	0.074	346.82	-13.51	8.8	🔍
☑	!	SNF20061109-002	Ia	127	5	8	following		med	18.59	0.043	46.38	12.01	13.4	🔍
☑		SNF20061108-004	Ia	120		3	following		med	19.26	0.083	32.96	14.45	12.6	🔍
☑	!	SNF20061108-001	Ia	123		12	following		med	19.5	0.065	32.14	9.92	12.4	🔍
☑		SNF20061107-027	Ia	127		2	following		med	18.68	0.050	49.12	-7.05	13.1	🔍
☐	!	SNF20061030-010	Ia	132	16	5	following		med	16.97	0.030	49.58	-2.31	13.2	🔍
☐		SNF20061024-000	Ia	140		3	following		high	18.93	0.055	30.98	-5.87	11.9	🔍
☐	!	SNF20061022-005	Ia	130		3	following		med	19.06	0.087	10.15	-8.88	10.5	🔍
☐		SNF20061021-003	Ia	138		9	following		high	19.11	0.061	9.05	7.16	10.8	🔍
☐		SNF20061020-000	Ia	143		2	following		high	18.07	0.040	10.27	6.79	10.8	🔍
☑		SNF20061011-005	Ia	156		1	following		med	16.02	0.025	41.66	-1.80	12.7	🔍
☑	!	SNF20061009-008	Ia	152		4	following		med	18.77	0.055	14.96	-11.03	10.7	🔍
☑		SNF20051113-000	Ia	456		1	final ref	med		19.33	0.080	67.53	-17.64	14.1	🔍

General Filters

Type: Show All / Cand / Ia

State: Show All / Show Active / saved

Disc. By: Show All

Priority Filters

PhotPrio: Show All / high / med

SpecPrio: Show All / high / med

Date Filter

Type: discovery

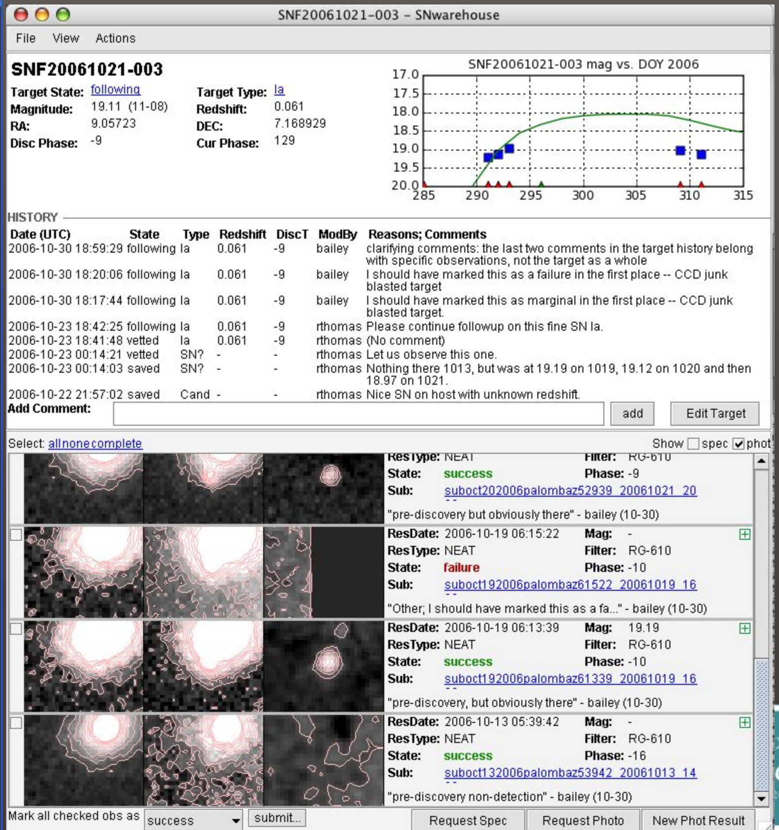
Date: 2007-03-26

Range: Show All

Target Name Filter

SNF200

☐ show any w/sched obs
☐ filter out unchecked
☐ show only SNF
☐ apply vet filter



SNF20061021-003

Target State: following **Target Type:** Ia
Magnitude: 19.11 (11-08) **Redshift:** 0.061
RA: 9.05723 **DEC:** 7.168929
Disc Phase: -9 **Cur Phase:** 129

HISTORY

Date (UTC)	State	Type	Redshift	DiscT	ModBy	Reasons; Comments
2006-10-30 18:59:29	following	Ia	0.061	-9	bailey	clarifying comments: the last two comments in the target history belong with specific observations, not the target as a whole
2006-10-30 18:20:06	following	Ia	0.061	-9	bailey	I should have marked this as a failure in the first place -- CCD junk blasted target
2006-10-30 18:17:44	following	Ia	0.061	-9	bailey	I should have marked this as marginal in the first place -- CCD junk blasted target.
2006-10-23 18:42:25	following	Ia	0.061	-9	rthomas	Please continue followup on this fine SN Ia.
2006-10-23 18:41:48	vetted	Ia	0.061	-9	rthomas	(No comment)
2006-10-23 00:14:21	vetted	SN?	-	-	rthomas	Let us observe this one.
2006-10-23 00:14:03	saved	SN?	-	-	rthomas	Nothing there 1013, but was at 19.19 on 1019, 19.12 on 1020 and then 18.97 on 1021.
2006-10-22 21:57:02	saved	Cand	-	-	rthomas	Nice SN on host with unknown redshift.

Add Comment:

Select: all none complete Show ☐ spec ☑ phot

ResType: NEAT **Filter:** RG-610
State: success **Phase:** -9
Sub: suboct202006palombaz52939_20061021_20
"pre-discovery but obviously there" - bailey (10-30)

ResDate: 2006-10-19 06:15:22 **Mag:** -
ResType: NEAT **Filter:** RG-610
State: failure **Phase:** -10
Sub: suboct192006palombaz61522_20061019_16
"Other; I should have marked this as a fa..." - bailey (10-30)

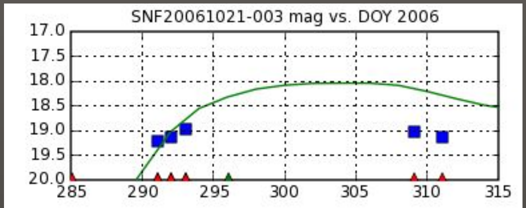
ResDate: 2006-10-19 06:13:39 **Mag:** 19.19
ResType: NEAT **Filter:** RG-610
State: success **Phase:** -10
Sub: suboct192006palombaz61339_20061019_16
"pre-discovery, but obviously there" - bailey (10-30)

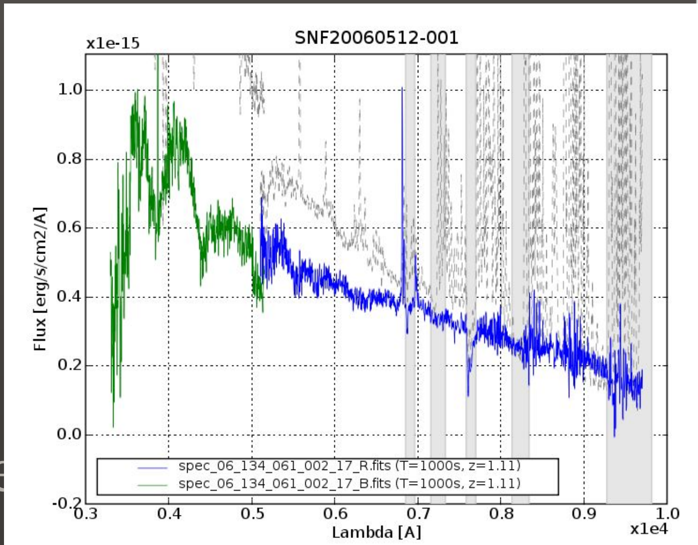
ResDate: 2006-10-13 05:39:42 **Mag:** -
ResType: NEAT **Filter:** RG-610
State: success **Phase:** -16
Sub: suboct132006palombaz53942_20061013_14
"pre-discovery non-detection" - bailey (10-30)

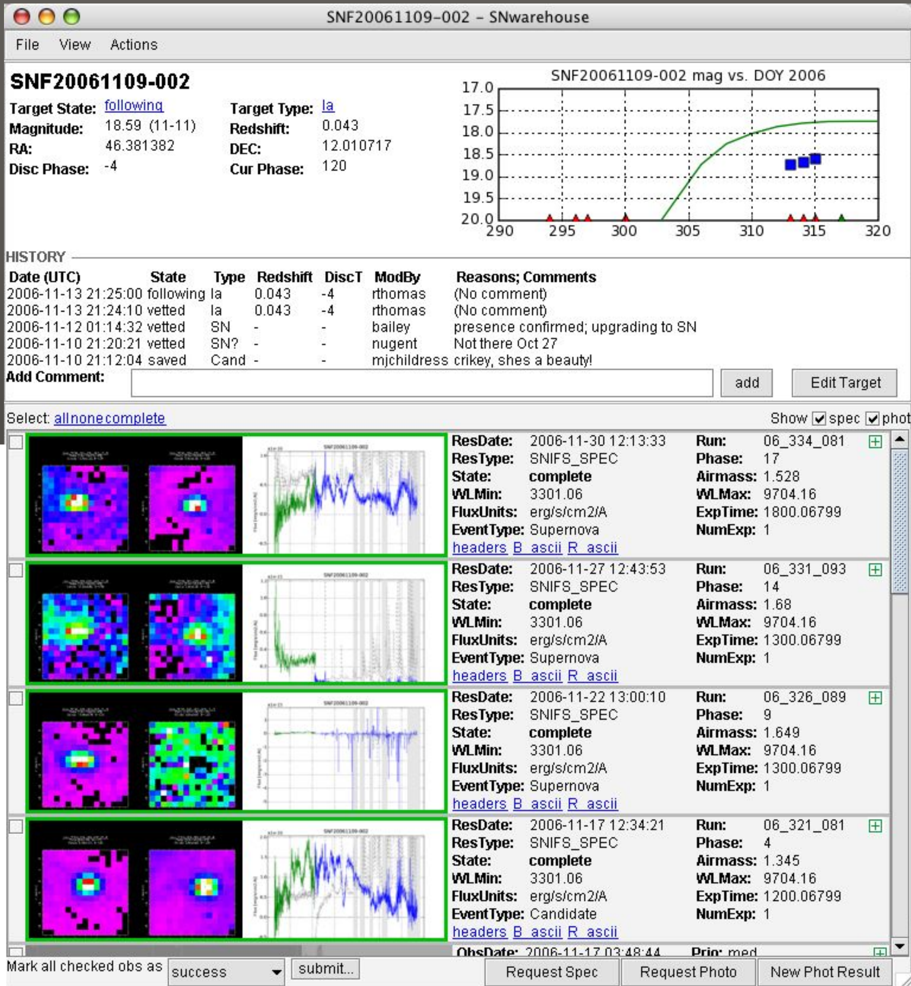
Mark all checked obs as success

Sunfall: Supernova Warehouse

- Improved situational awareness, decreased repetitive labor, more science!







SNF20061109-002

Target State: following **Target Type:** Ia
Magnitude: 18.59 (11-11) **Redshift:** 0.043
RA: 46.381382 **DEC:** 12.010717
Disc Phase: -4 **Cur Phase:** 120

HISTORY

Date (UTC)	State	Type	Redshift	DiscT	ModBy	Reasons; Comments
2006-11-13 21:25:00	following	Ia	0.043	-4	rthomas	(No comment)
2006-11-13 21:24:10	vetted	Ia	0.043	-4	rthomas	(No comment)
2006-11-12 01:14:32	vetted	SN	-	-	bailey	presence confirmed; upgrading to SN
2006-11-10 21:20:21	vetted	SN?	-	-	nugent	Not there Oct 27
2006-11-10 21:12:04	saved	Cand	-	-	mjchildress	crikey, shes a beauty!

Add Comment:

Select: all none complete Show ☑ spec ☑ phot

ResDate: 2006-11-30 12:13:33 **Run:** 06_334_081
ResType: SNIFS_SPEC **Phase:** 17
State: **complete** **Airmass:** 1.528
WLMin: 3301.06 **WLMax:** 9704.16
FluxUnits: erg/s/cm2/A **ExpTime:** 1800.06799
EventType: Supernova **NumExp:** 1
headers B ascii R ascii

ResDate: 2006-11-27 12:43:53 **Run:** 06_331_093
ResType: SNIFS_SPEC **Phase:** 14
State: **complete** **Airmass:** 1.68
WLMin: 3301.06 **WLMax:** 9704.16
FluxUnits: erg/s/cm2/A **ExpTime:** 1300.06799
EventType: Supernova **NumExp:** 1
headers B ascii R ascii

ResDate: 2006-11-22 13:00:10 **Run:** 06_326_089
ResType: SNIFS_SPEC **Phase:** 9
State: **complete** **Airmass:** 1.649
WLMin: 3301.06 **WLMax:** 9704.16
FluxUnits: erg/s/cm2/A **ExpTime:** 1300.06799
EventType: Supernova **NumExp:** 1
headers B ascii R ascii

ResDate: 2006-11-17 12:34:21 **Run:** 06_321_081
ResType: SNIFS_SPEC **Phase:** 4
State: **complete** **Airmass:** 1.345
WLMin: 3301.06 **WLMax:** 9704.16
FluxUnits: erg/s/cm2/A **ExpTime:** 1200.06799
EventType: Candidate **NumExp:** 1
headers B ascii R ascii

ObsDate: 2006-11-17 03:48:44 **Prio:** med

Mark all checked obs as success submit...

Production Analytics Impact on Science

- Significant, measurable effect on a major science project, with impact on physics grand challenge (dark energy).
  - Reduced number of false positive candidate supernovae by 80%
  - Up to 90% reduction in labor costs in areas of SNfactory pipeline
  - This freed up more time for science
    - SNF pubs 2005: 0 refereed
    - SNF pubs 2006 & 1st 3 mos 2007: 3 refereed (1 submitted)

Production Analytics Impact on Science

- From the researchers:
  - “We’re awash in Supernova”
  - More SN discovered in past year than in all previous history
- Observation
  - This system the product of a multidisciplinary team of dedicated scientific staff
  - Mission focus guides R&D

Production Analytics - Challenges

- Coupled code projects:
  - Fusion: FACETS
  - Accelerator: COMPASS
  - Climate: Earth System Model
- Data from one code is input to another
- V&V at each stage
- Model and result is multimodal, multivariate (and multiscale)

Challenge: Software Architecture and Engineering

- Objective: production-quality, petascale-capable visual data analysis software (Hank’s talk)
- Challenges:
  - – Who is the user/stakeholder?
  - – What platforms?
  - – Use models?
  - – What is effective balance between R&D?

Conclusion

- Petascale visual data analysis is a problem-rich environment.
- Given limited bandwidth (cognitive, network, processing, etc.) and time, let Occam’s Razor help as a guide to eliminate unnecessary motion.
- We’ve discussed some, but not nearly all, of the challenges in this space.

Next Up

- Hank Childs, LLNL
  - *Large-scale viz (GNEP connection)*
- Kwan-Liu Ma, UC Davis
  - *Parallel viz. pipeline and insitu Feature detection/tracking*