New Approach to Science Data Discovery in Petascale Systems

Tamara Sipes[1,2]
[1]SciberQuest, Inc.
[2]UCSD

Collaborators:
H. Karimabadi
B. Loring
A. Roberts
Y. Wang
B. Lavraud
J. Merka
J. Gosling
V. Roytershteyn
W. Daughton
H. X. Vu
Y. Omelchenko
J. Raeder
J. Dorelli
A. Yilmaz

The 2011 Earth Science Technology Forum, June 21-23, 2011, Pasadena, California

Problem description

- Only about 10% of spacecraft data collected from all missions have been explored
  - Stale data
  - Low ROI
  - Biased and limited event examples make statistical studies very difficult

- New 3D kinetic codes can generate over **200 TB** of data from a **single run** with expected increase to **20 PB** by 2012!

Sheer size and complexity mandate new approach to more fully exploit data archives.

Current Approach

- Dominant form of data analysis - *visual inspection of data*
- Standard data mining techniques (ANN, SVM etc.) are not directly applicable to analysis of ***spacecraft data***:
  - multi-variate, multi-source, multi-scale time series data
  - multi-variate time series classification – very challenging
  - need for increased speed through parallelization or use of graphics processing units (GPUs)
  - lack of expertise to use/develop data mining techniques
- Not surprisingly, 90% of spacecraft data remains unexplored
- Our solution → Physics Mining

Current Approach

- Scientific visualization is the dominant form of analysis of ***simulation data***, in space sciences and other fields of science
- while useful, scientific visualization lacks quantitative analysis capability
- existing visualization systems support knowledge discovery poorly, in fact, they often just *present graphics*
- typically the visualizations are not deeply integrated with methods and tools that would enable scientific insight
- Our solution → SciVis

Definition: Physics Mining

Common Terms:

- Machine Learning
- Data Mining
- Artificial Intelligence
- Computer Vision

Physics Mining: Algorithmic approach to data analysis and derivation of analytical models from the data

MineTool Algorithm Formula

(Karimabadi et al., 2007, 2009)

$$y_i = \mathbf{X}_i'\boldsymbol{\alpha} + \underbrace{\sum_{p=1}^{P} \zeta(\mathbf{X}_i)'\boldsymbol{\delta}_p}_{Linear\ Transforms} + \underbrace{\sum_{q=1}^{Q} \boldsymbol{\Psi}(\mathbf{X}_i, \boldsymbol{\gamma}_q)'\boldsymbol{\beta}_q}_{Non-linear\ Transformations}$$

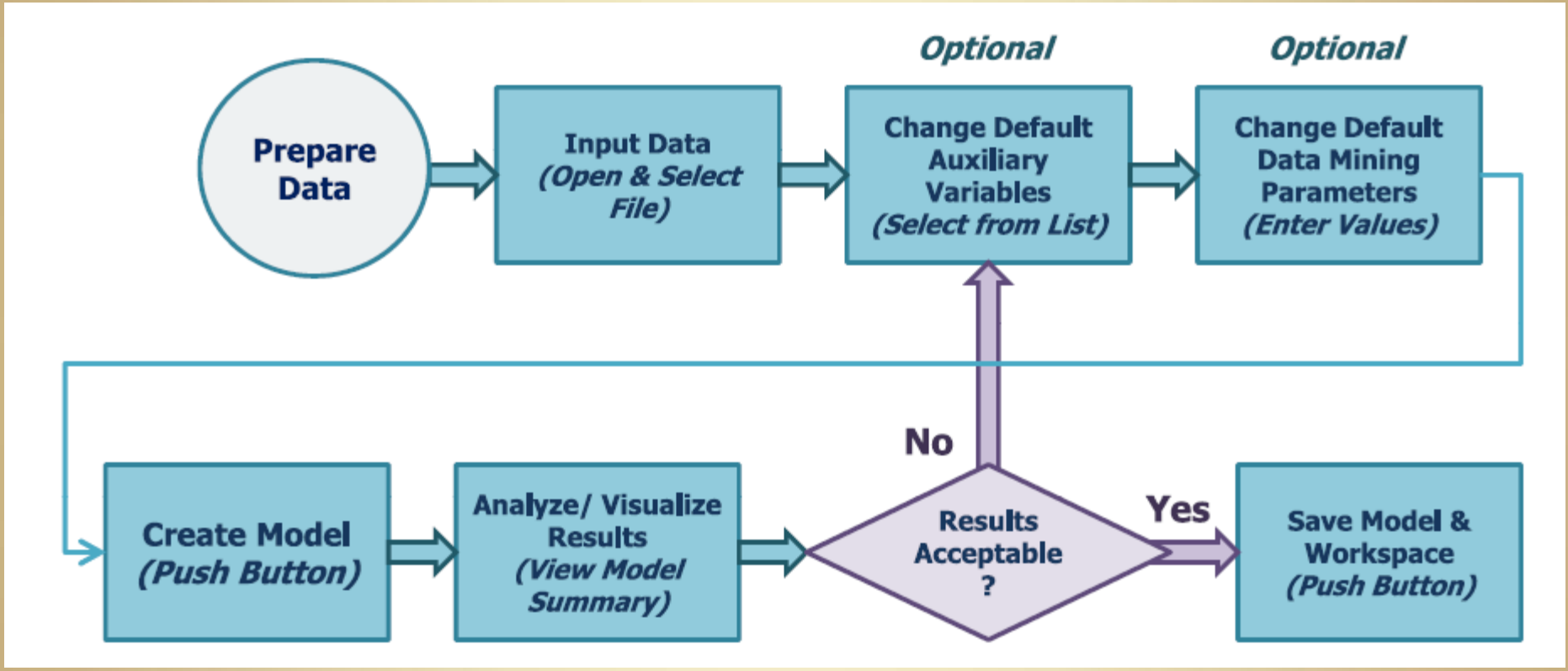
Reduces the problem to least squares fit rather than the difficult nonlinear optimization

Features of Physics Mining

Innovations in data mining include the following features:

- Applicable for both multi-variate static and time series/image data/simulation data
- Near real time processing of large data sets through use of GPUs/parallelization
- Capable of handling time series of unequal length and multi-scale events
- Unsupervised algorithms for both static and time series data
- Anomaly detection

Process for applying MineTool



MINETOOL USER INTERFACE DESIGN

MineTool

SciberQuest

Open Dataset | Choose Auxiliary Variables | Set Data Mining Parameters | Visualize | Save

Select Input Dataset: Filename [Browse]

Accepted formats: .txt and .csv

A variable is either an input, illustrative or output variable

Dataset Variables

Number of instances: 10000

Number of variables: 4

List of Variables:

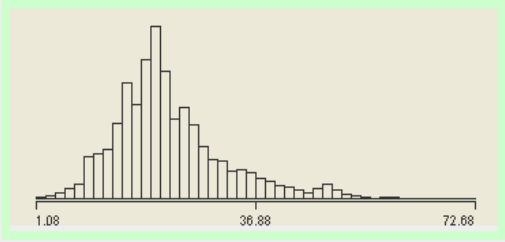
Index	✓	Variable Name	Input	Illustrative
1	☐	ID	☐	☒
2	☐	x	☒	☐
3	☐	y	☒	☐
4	☐	z	☒	☐
5	☐	goal	☐	☐

Output Variable: Select output variable ▽

Selected Variable Characteristics

Name	bmag
Type	Numeric
Missing	0 (0%)
Distinct	2752
Unique	816 (9%)

Statistic	Value
Minimum	1.08
Maximum	72.68
Mean	22.791
StdDev	9.215



Variable Preprocessing

Preprocess Selected Variables

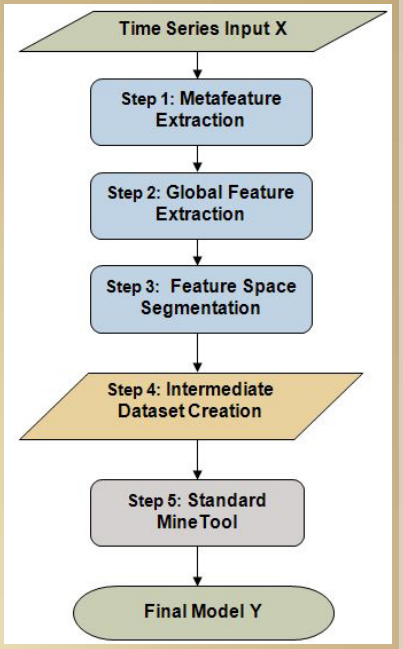
- Normalize
- Interpolate
- Smooth
- Remove Outliers
- Nominal to Numeric

Undo | Save File

Computer Vision

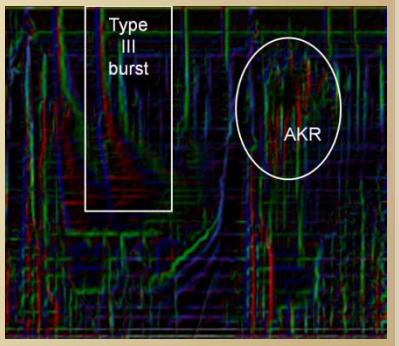
- Physics Mining can also combine advanced techniques in image analysis and computer vision
- Label images – drag and draw an ellipse bounding an event, select label
- Feature extraction from images and videos - detection of pixels and their surroundings, both spatial and spatiotemporal, that contains a variation in the observations
- Variations can be quantified by computing the changes in the locality of a pixel from its surrounding in the operator-selected event boundary

Physics Mining Techniques…



Gee Whiz?

Or Practical Technology?

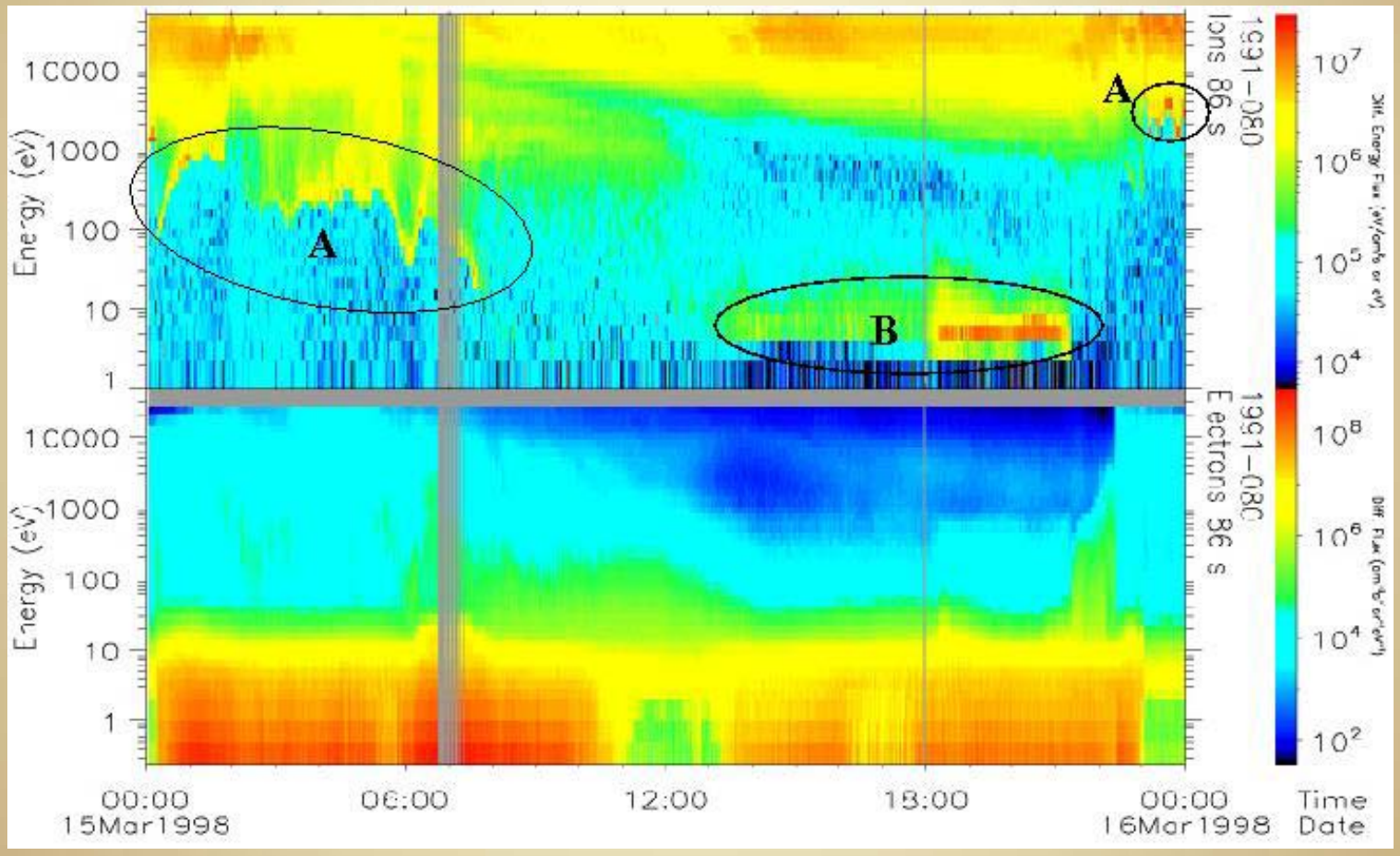


Fun, But Impressive Application

From this picture

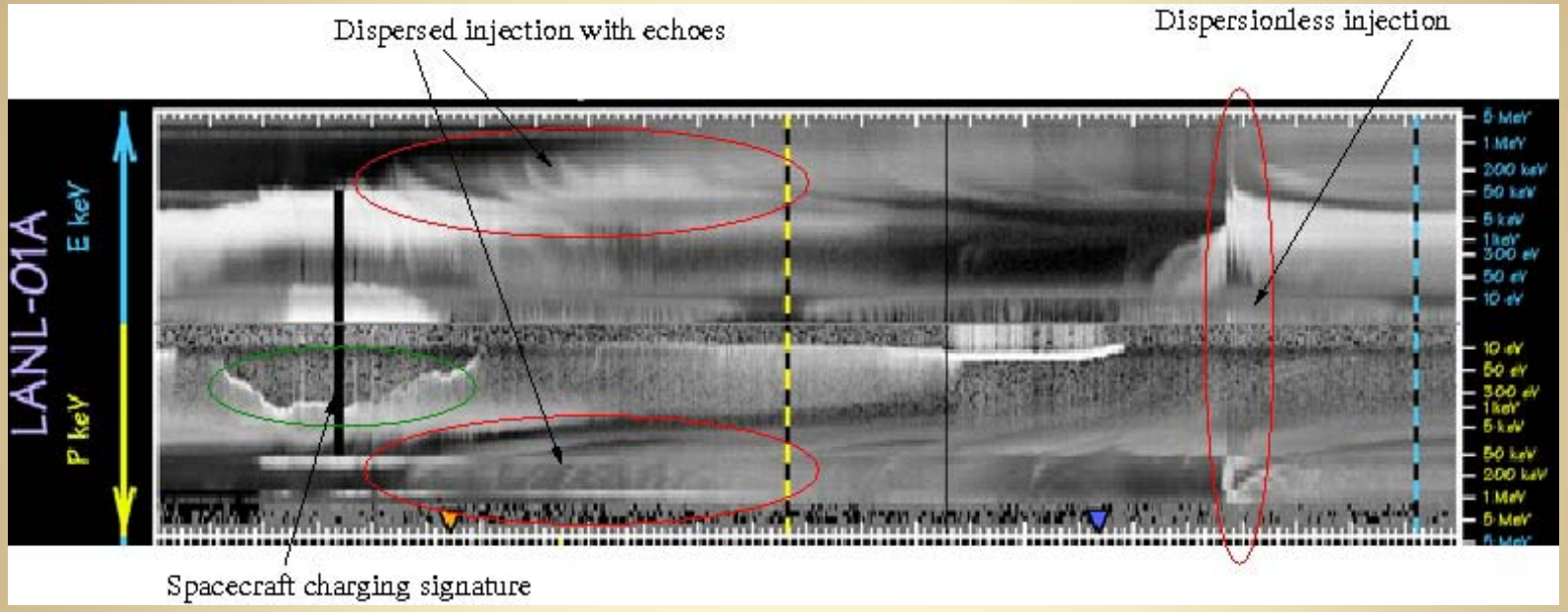
Found “Dr. K” in this picture

Example: Physics Mining Applied to Spacecraft Charging



Plasma data from MPA instrument (individual spectra stacked in time) on board spacecraft 1991-080 for 15 March, 1998 - Courtesy: R. Friedel

Substorm injection



Substorm Injection Signatures: electrons and ions plotted together with the zero energy in the center and energy increasing upwards for electrons, downwards for ions. Courtesy: R. Friedel

Challenges in Global Simulations

- Multi-Scale Coupled System
  - Spatial scales vary from centimeters to 200 RE
  - Temporal scales vary from less than milliseconds to days
  - Kinetic effects have global consequences
- Multi-physics
  - electron physics: e.g., controls reconnection rate
  - ion physics: e.g., dominates formation of boundaries and transport
  - global features and dynamics: e.g., magnetotail/energetic particles
  - coupling to the ionosphere

Three Approaches to Global Simulations

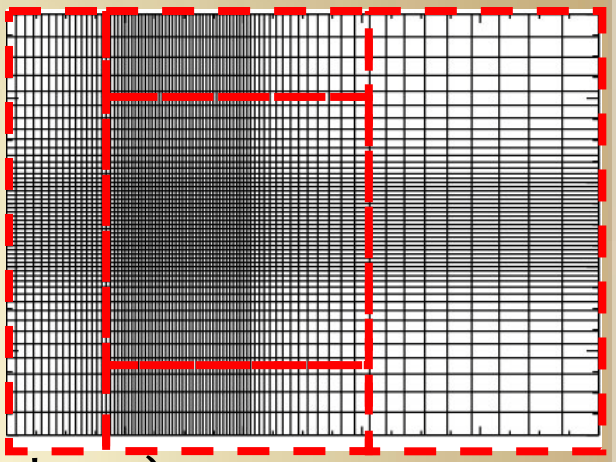
- MHD (single fluid)
  - used extensively
  - does not resolve important ion physics
  - does not correctly capture the physics of reconnection
  - not suitable for studies of boundaries & discontinuities

Petascale computing has enabled:

- **Hybrid** (fluid electrons, kinetic ions)
  - fluid electrons, kinetic ions
  - it is the next stage of advance in 3D global simulations
  - resolves ion spatial scales (ion inertial length) and ion temporal scales (gyroperiod)
- **Full Particle** (kinetic electrons, kinetic ions)
  - it is the most complete description
  - only 2D is possible

Our Approach

Hybrid Code



- Nonuniform Mesh + Multi-time zones

- Discrete Event Technique (Omelchenko et al 2011)
  - robust (stable) and efficient (no idle computation)
  - works for arbitrary meshes
  - Predict local Δt for each variable f ("state") based on its estimated trajectory, $f=f_E(t)$ and a given Δf_E (selected based on local CFL/reaction conditions)

Fully Kinetic Code

- 2D global for capturing the microphysics of reconnection in a global setting

The Hybrid Approximation

- Ions: kinetic particles
- Electrons: massless, quasi-neutral ($en_e = q_i n_i$) fluid
- Electromagnetic fields:

  - Faraday's law

  $(\partial/\partial t)\,\mathbf{B} = -c\ \nabla \times \mathbf{E}$

  - Ampere's law

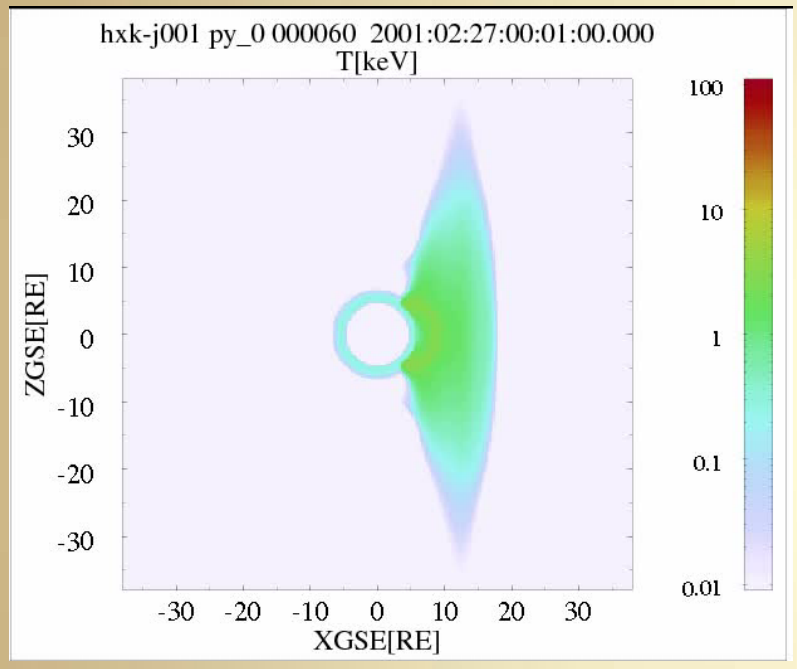
  $\nabla \times \mathbf{B} = 4\pi\,\mathbf{J}/c = 4\pi\, q_i n_i\,(\mathbf{v}_i - \mathbf{v}_e)/c$

  - Electric field from electron momentum equation

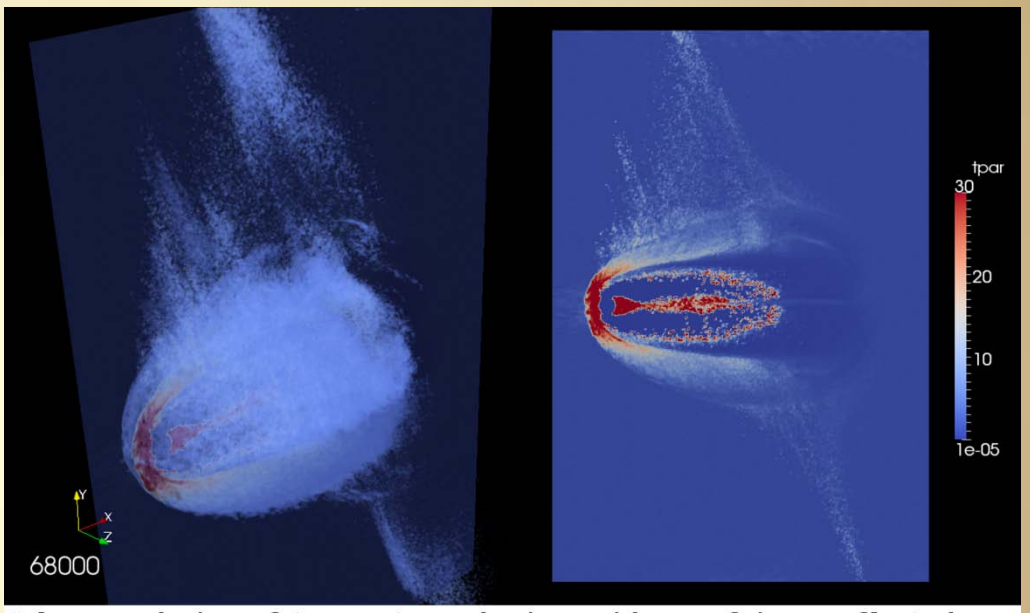
  $\mathbf{E} = -\mathbf{v}_i \times \mathbf{B}/c - \nabla \mathbf{p}_e/(q_i n_i) - \mathbf{B} \times (\nabla \times \mathbf{B})/(4\pi\, q_i n_i) - \eta \mathbf{J}$

3D MHD vs 3D Hybrid

Temperature



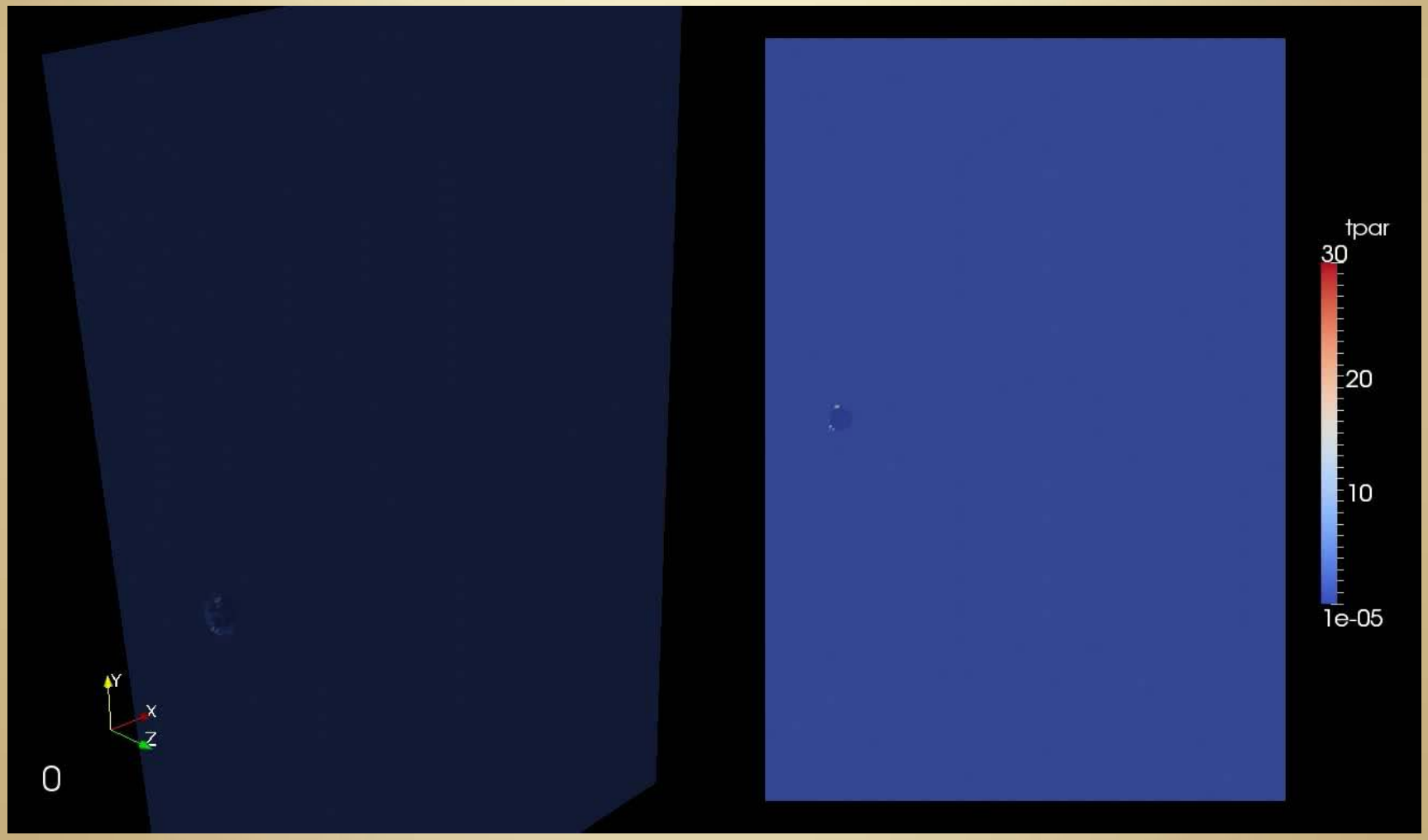
Temperature



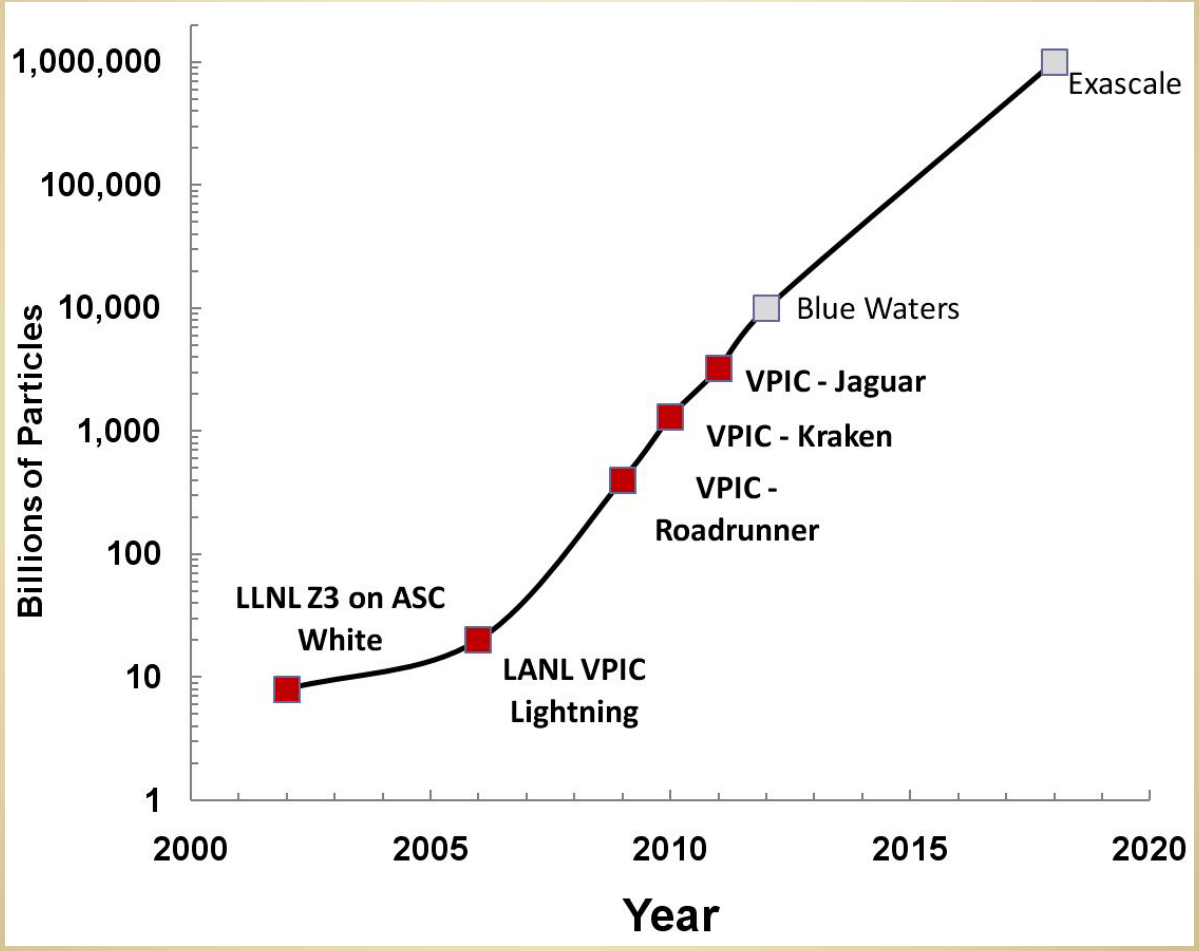
Volume rendering of temperature showing evidence of ions reflected off the bow shock.

- The physics responsible for modeling backstreaming ions reflected at the bow shock are present in the hybrid simulations (right), but missing from MHD simulations (left).
- The enhanced physics of the hybrid simulation advances understanding of the real world processes.

3D Hybrid



Number of Particles as a Measure of Size of Kinetic Simulations

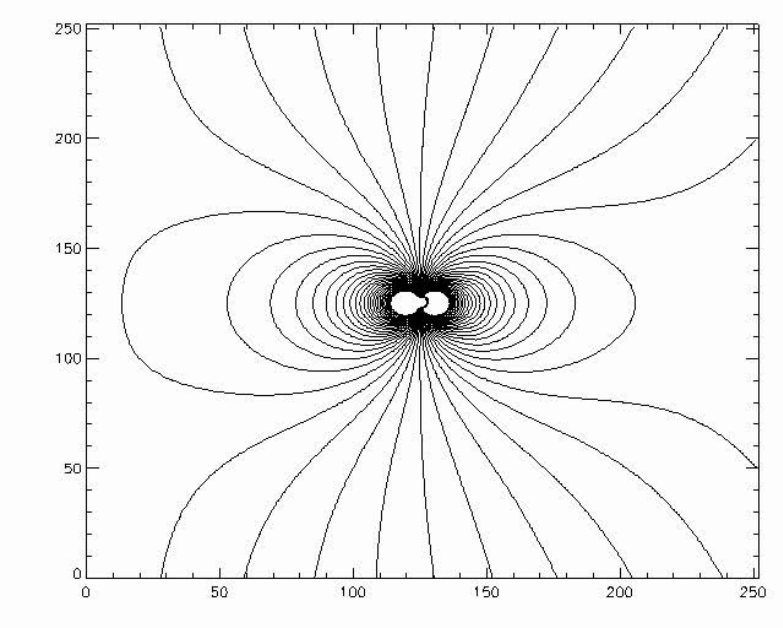


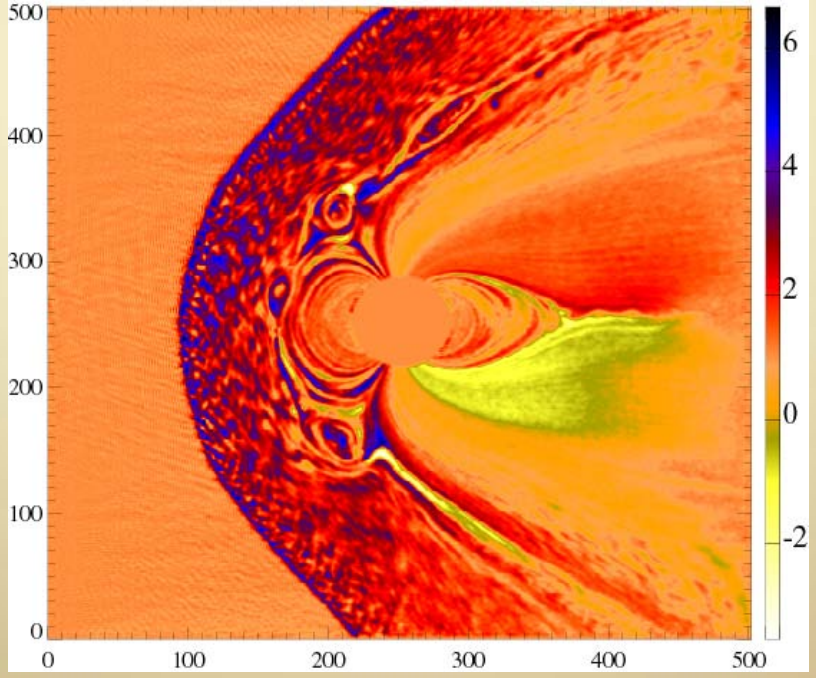
Size of largest PIC simulations in units of billion particles

Simulation Example: Feature Extraction and Tracking Flux Ropes in Global Hybrid Simulations

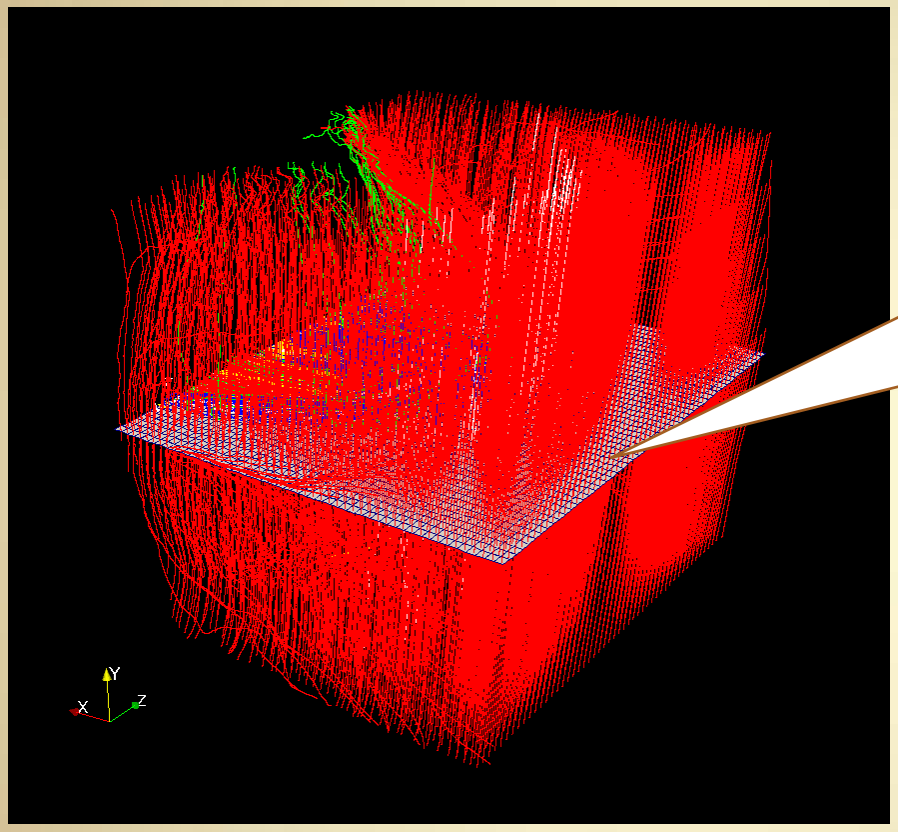
Problem: How to find flux ropes in a highly turbulent medium in a very large data set

2-D Tracking is Straight-forward





Information Overload : Looking for FTEs

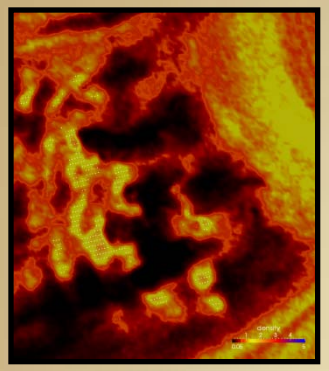


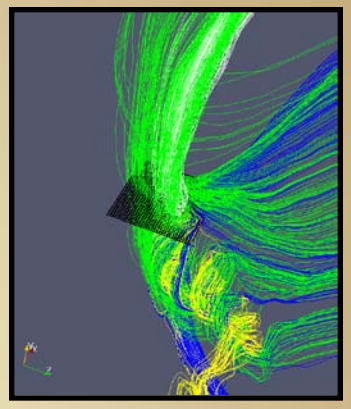
3-D Is More Complex

Challenge:

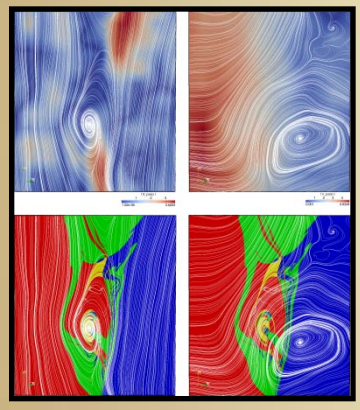
Very course resolution trace in a large dataset, ~100 cells per line. It is very likely that a feature will be overlooked or missed altogether.

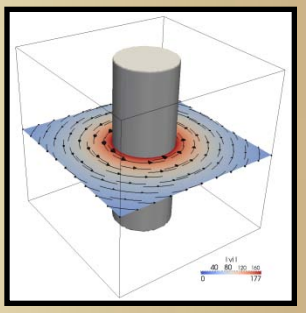
SciVis: Intelligent Scientific Visualization

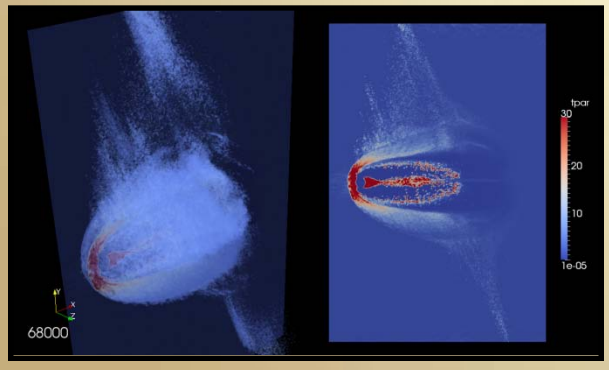


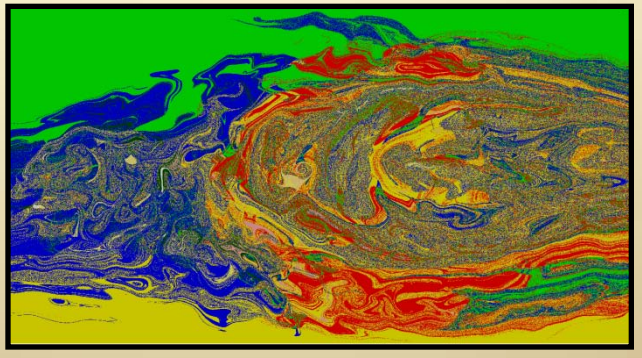


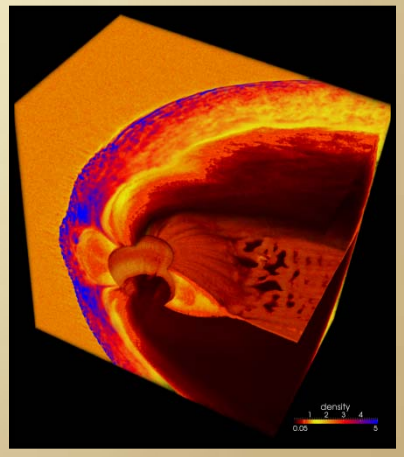
- Domain Specific Customizations for ParaView
- Incorporate Physics into Visualization Pipeline
- Autonomous Processing of Large Datasets
- Feature detection and Tracking
- Query Driven Visualization
- Incorporate Techniques from Data Mining and Computer Vision



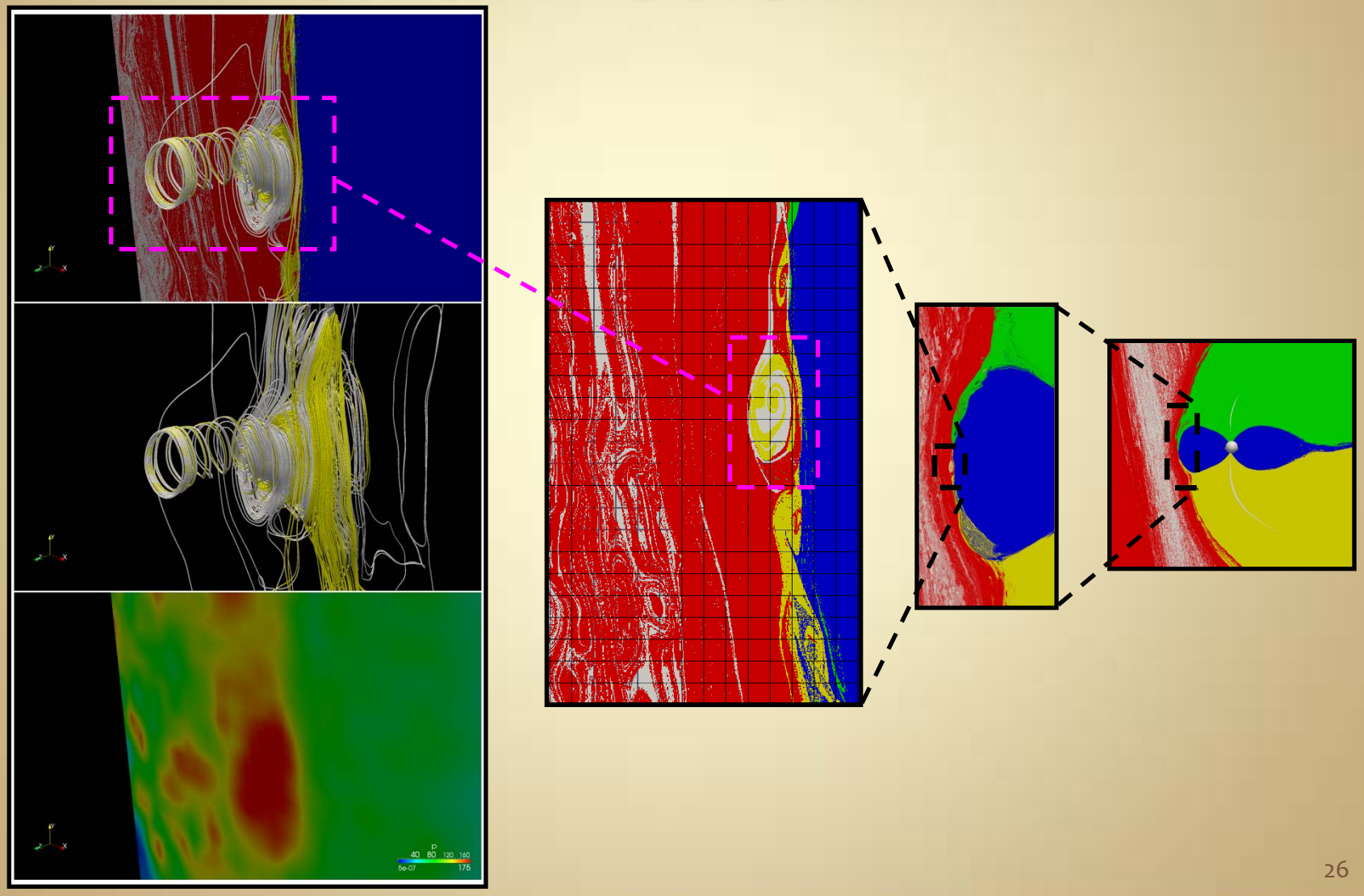




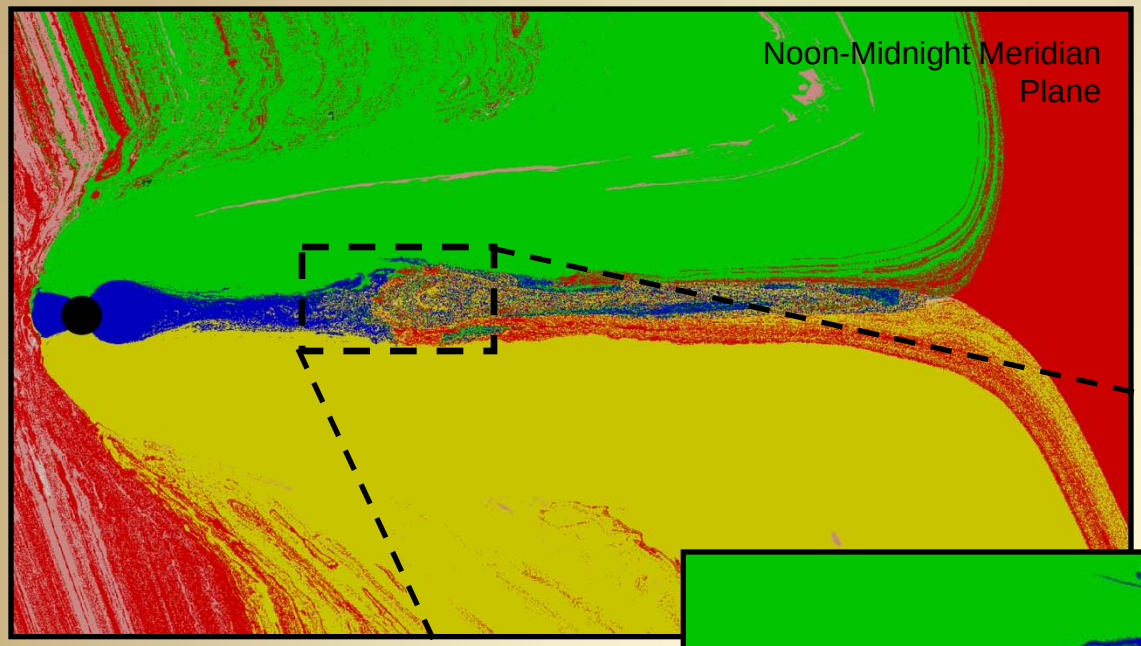




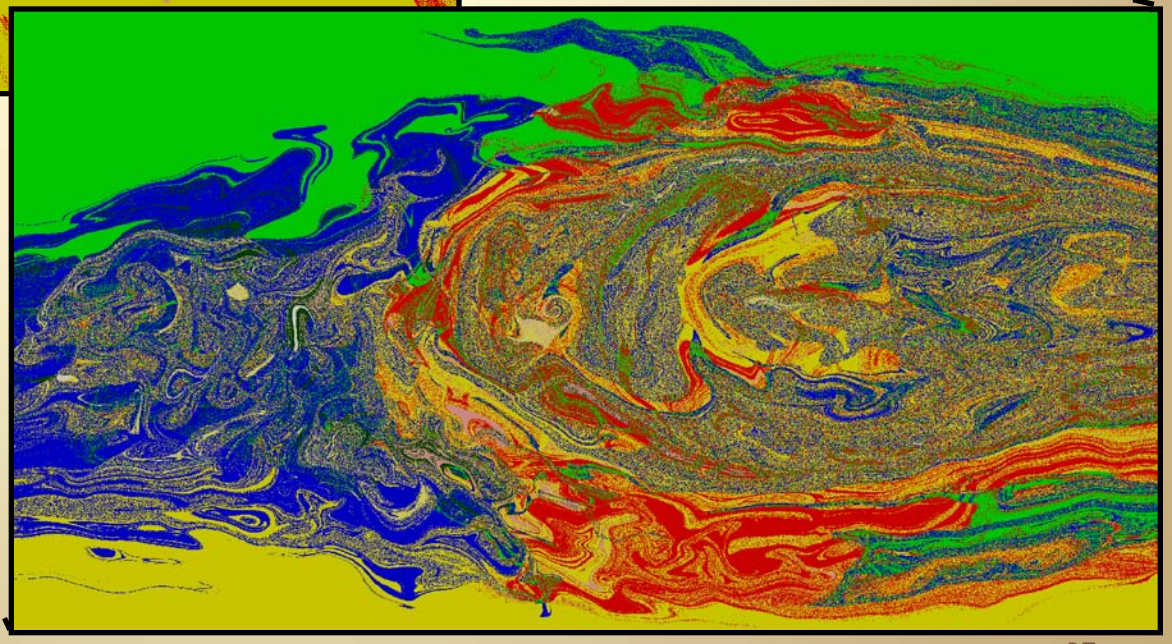
Detection of FTE: Needle in Hay-stack Problem



Turbulent Magnetosphere

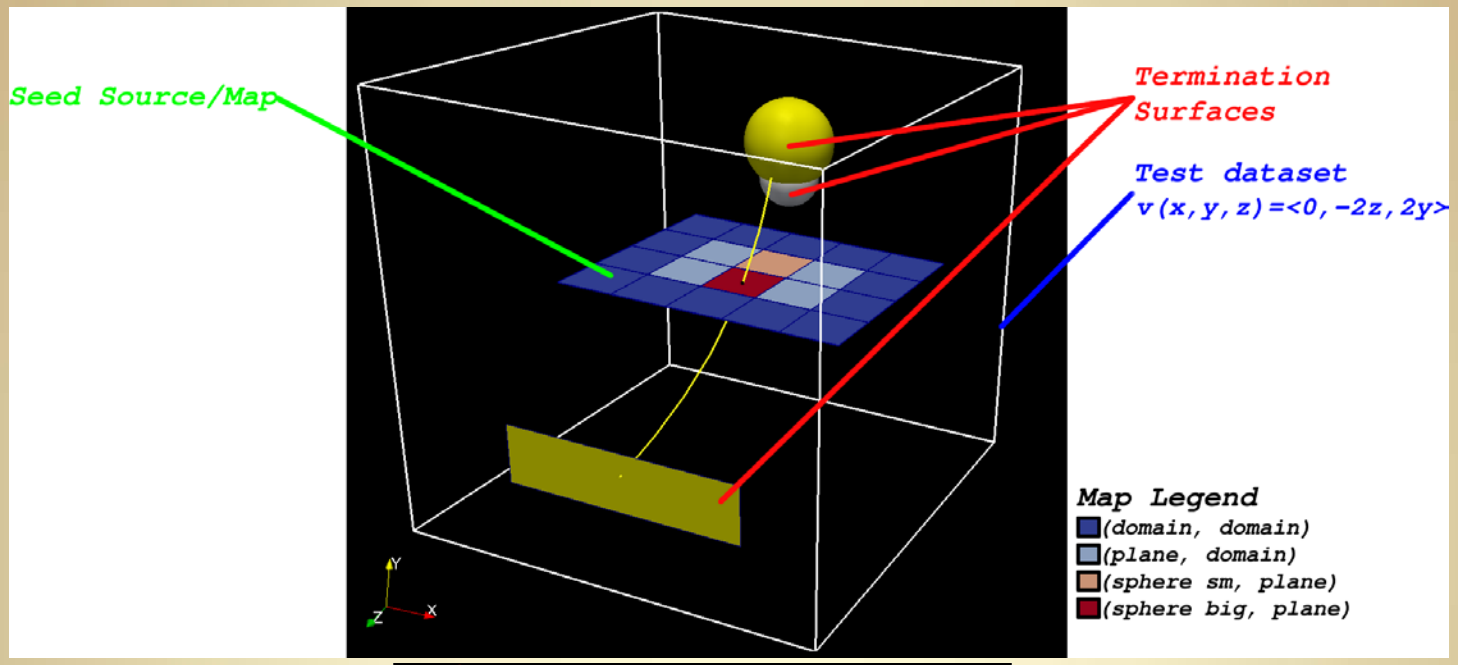


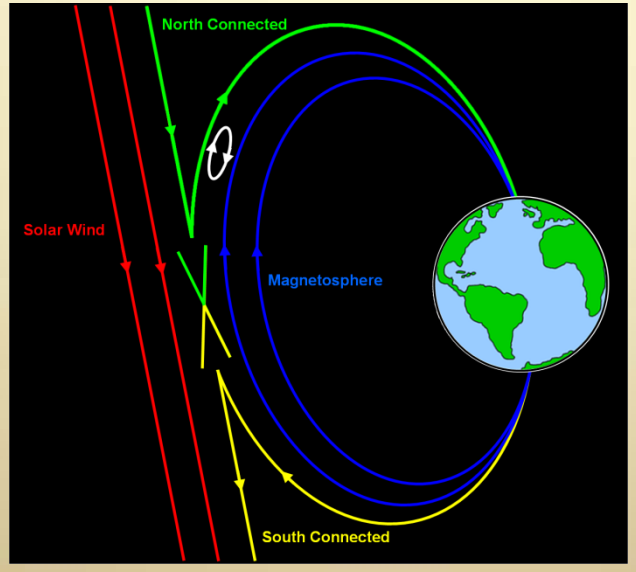
Interactive physics-based magentic topology visualization tool (plug in for ParaView):
-visually spot interesting regions
-highly complex, 2D slices, tracking difficult



Legend		
Class	Color	Description
d ⇔ d		solar wind
ø ⇔ d		
i ⇔ d		
n ⇔ s		closed magnetosphere
n ⇔ d		northern hemisphere connected
n ⇔ ø		
n ⇔ i		
n ⇔ n		
s ⇔ d		southern hemisphere connected
s ⇔ ø		
s ⇔ i		
s ⇔ s		
ø ⇔ ø		field null
i ⇔ ø		
i ⇔ i		

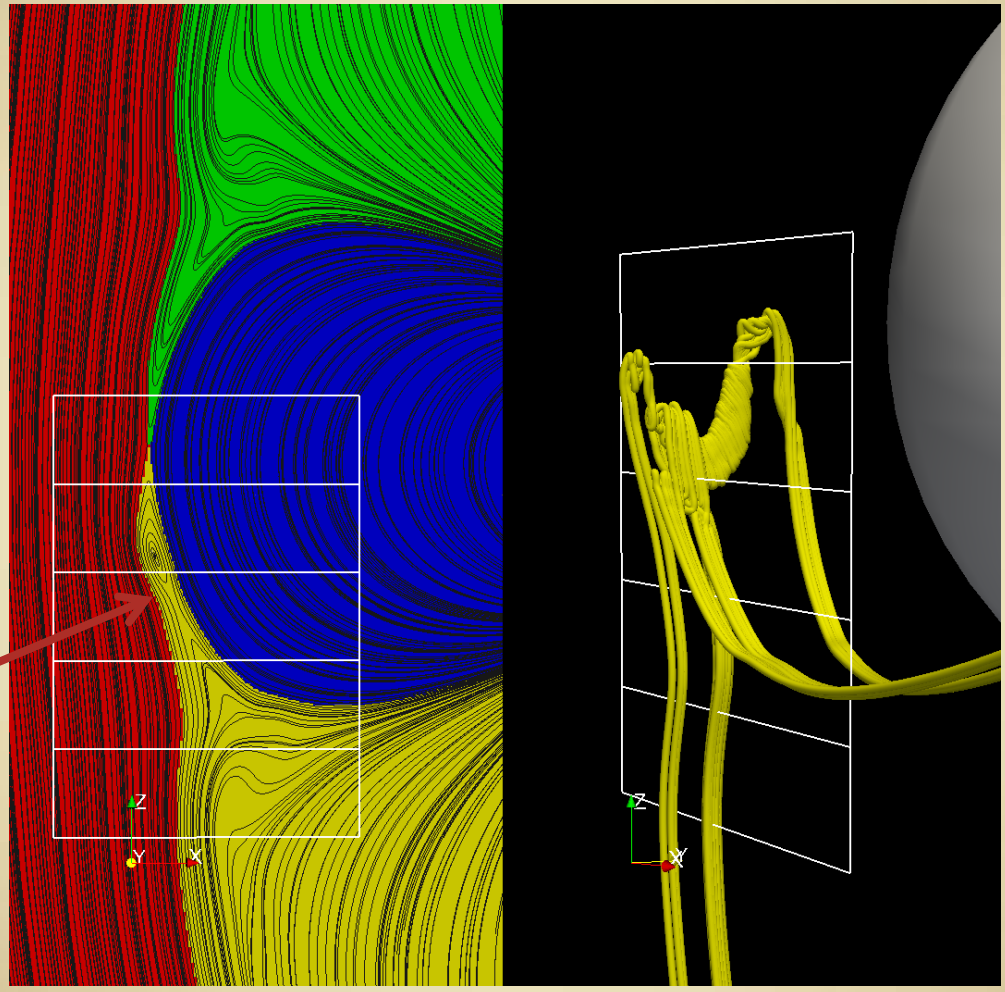
Topomap – Key Diagnostic





5 topological classes of magnetic field lines:
- solar wind lines
- magnetosphere lines
- northern/southern ionosphere field lines
- lines that end at a field null

3D Global MHD



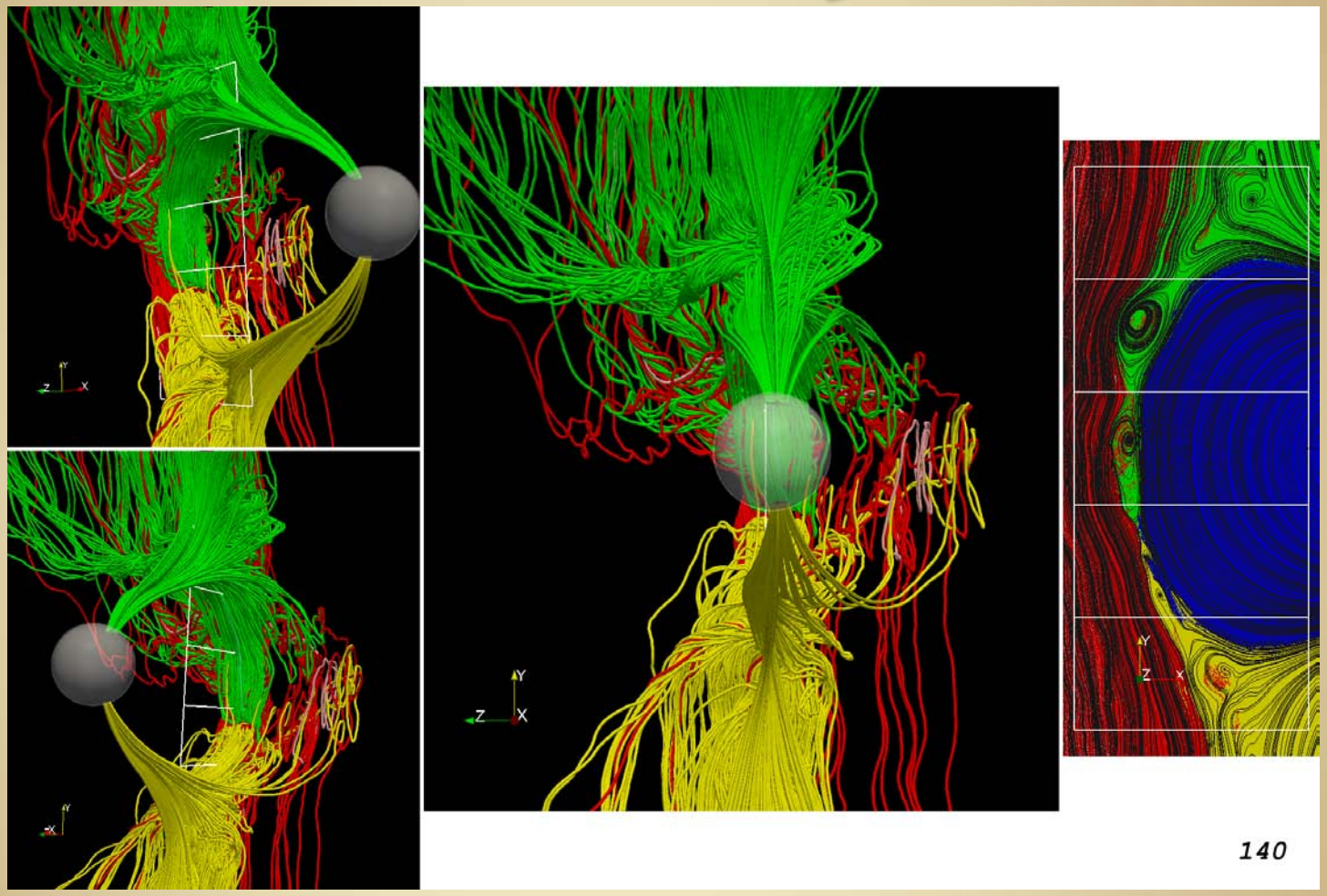
Purely Southward

No Dipole Tilt

False X-line which may be misinterpreted as MXR

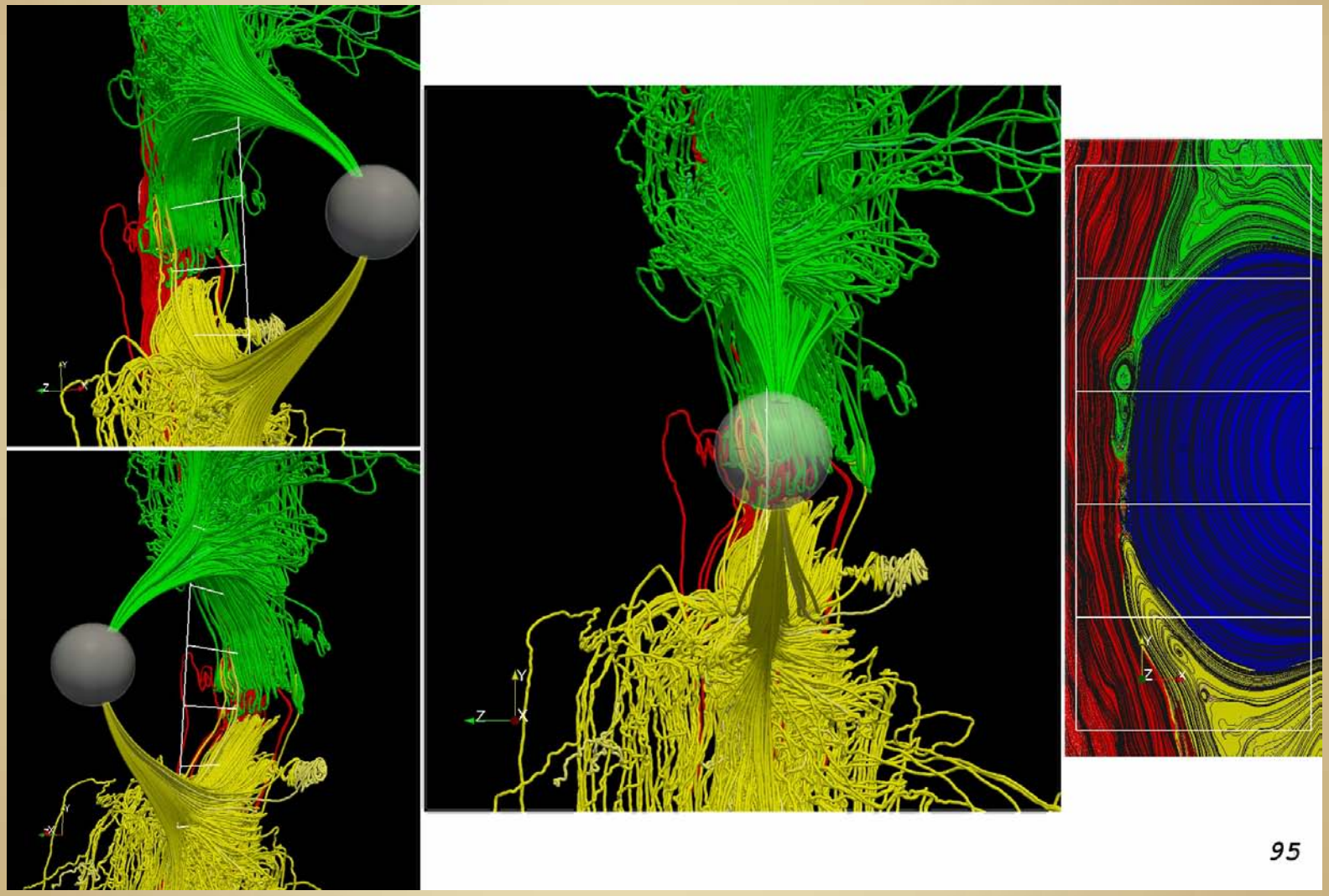
Very Little Mixing

3D Global Hybrid



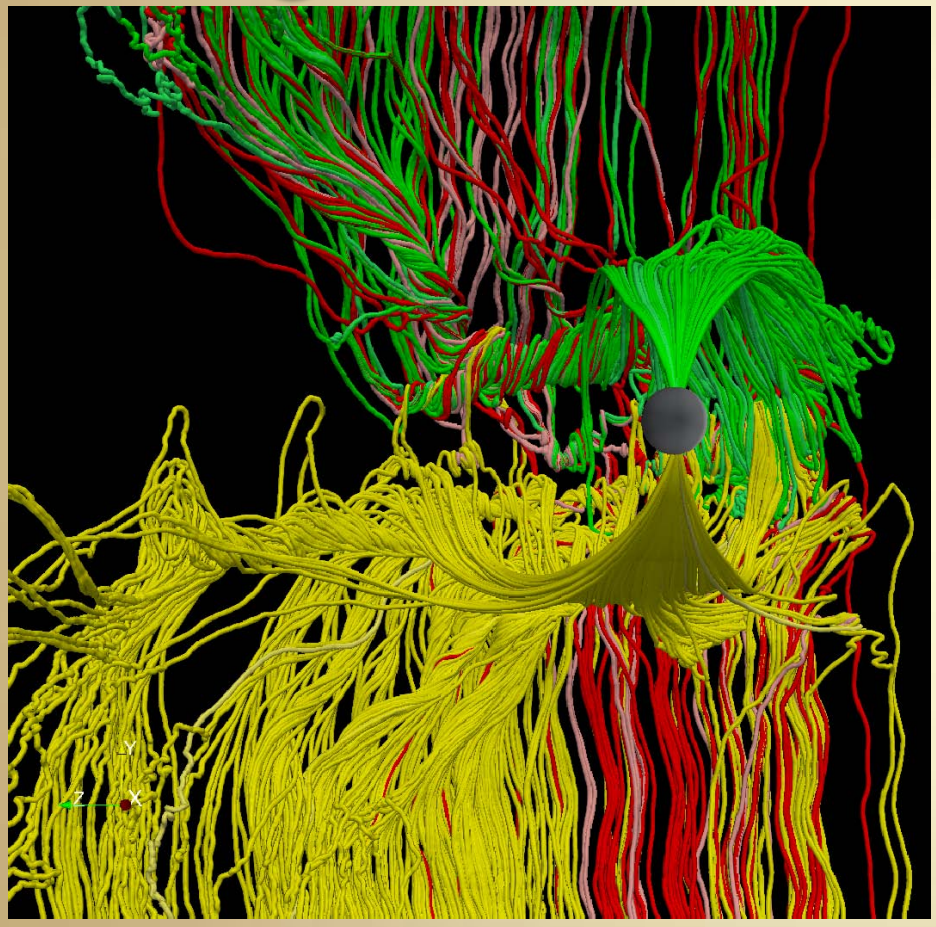
Run on 98 K cores on Kraken

3D Global Hybrid

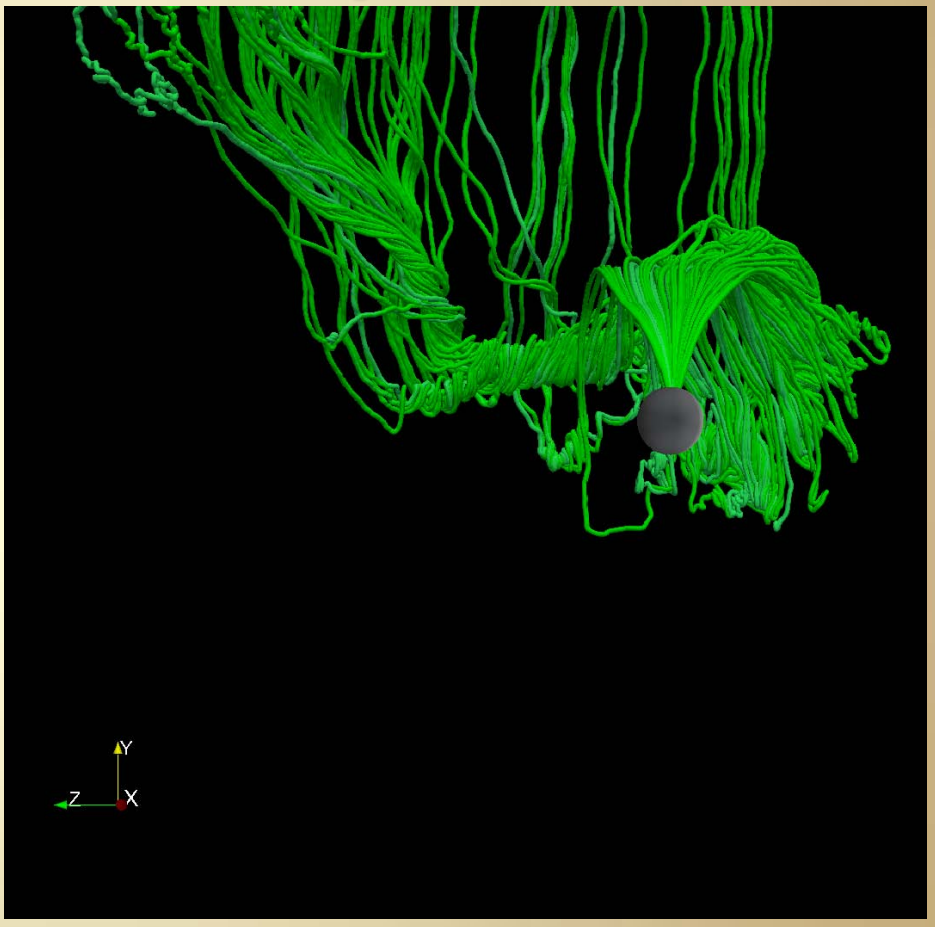


Successful Tracking of Flux Ropes in Time

Mixing of Field Lines



Shape of FTE



Case study: Automated Detection of FTE in the Cluster Data

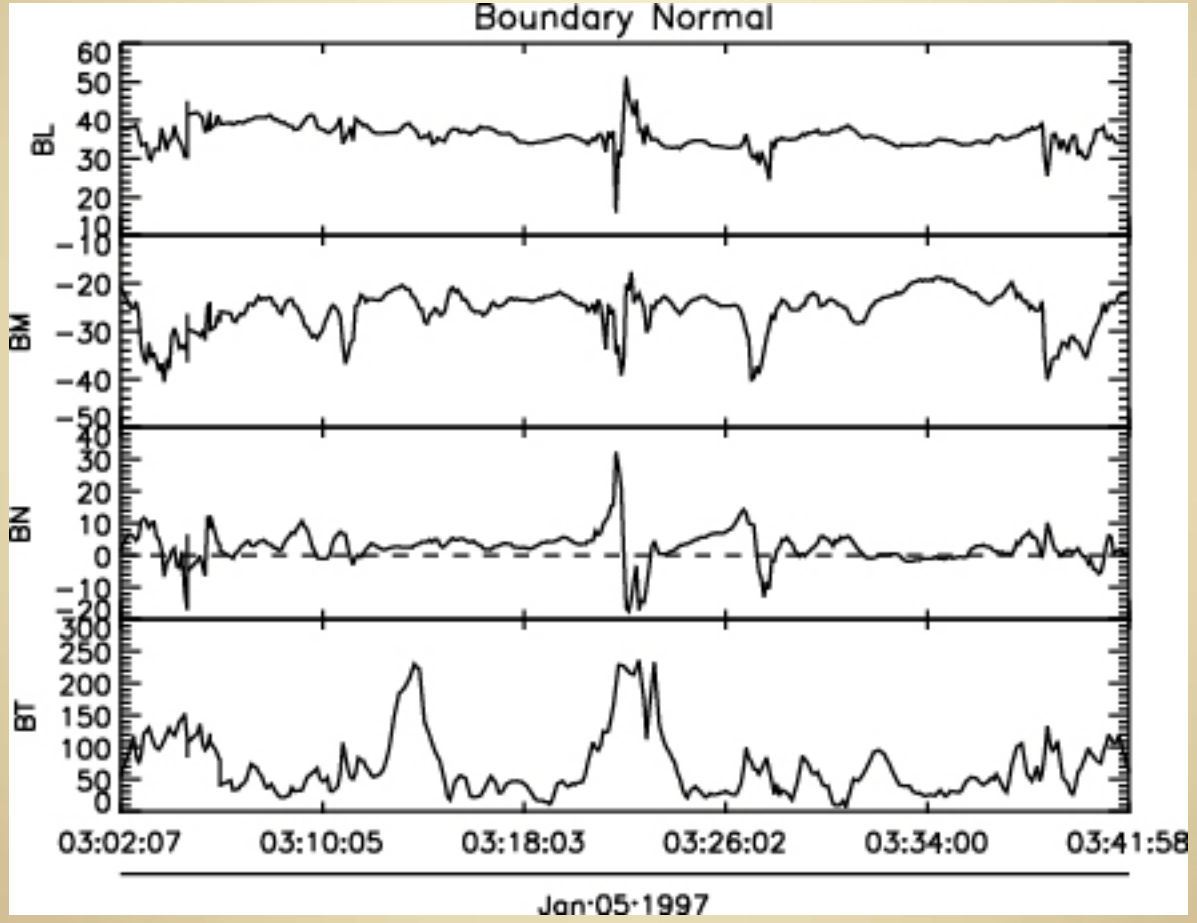
Problem: Traditional approach of discovering FTEs in cluster data is subjective, tedious, and limited.

Karimabadi, H., T. Sipes, Y. Wang, B. Lavraud, A. Roberts, JGR 2009

Comparison of traditional and automated approaches

Traditional Approach Using Visual Data Analysis	Query Driven Analysis Using MineTool
Transform to normal boundary coordinates using a model of MP	Automate the manual processes to place focus on answering questions: • Can FTEs be found in GSM coordinates? • Can FTEs be found based on plasma data alone? • What combination of variables are good indicators of FTEs? • Are there FTEs with no enhancement in |B|? • Does B_M have a well defined signature in all FTEs? • How complete is our taxonomy of FTEs? ***Huge return with limited investment.***
Choose a specific criterion for FTE D. Sibeck: clear bipolar signatures and well defined B_M component Y. Wang: clear bipolar signature in B_N and |B| enhancements	
Visually inspect the data … … … ***Huge investment with limited return.***	

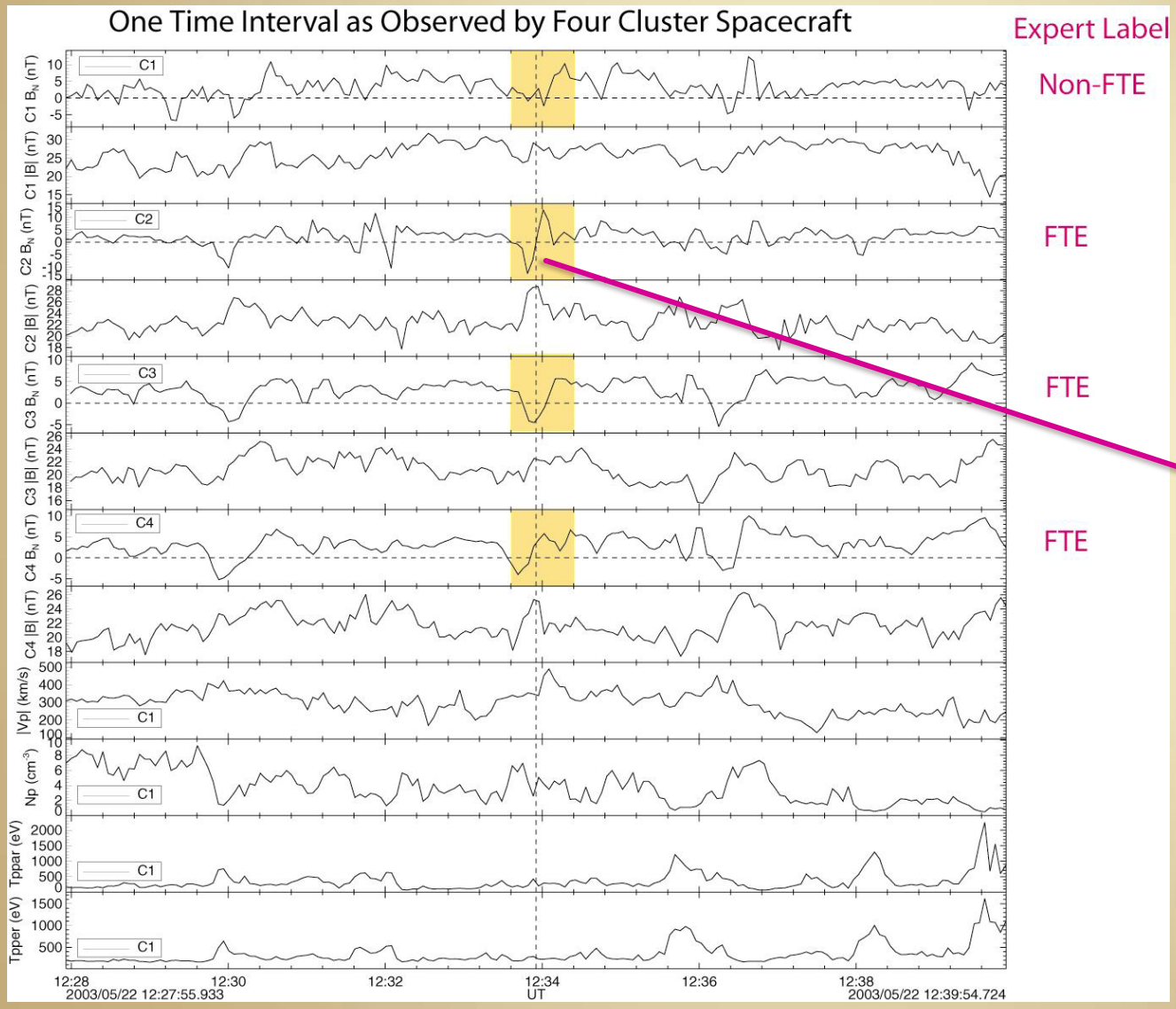
Example of an FTE



FTE Cluster Data Characteristics

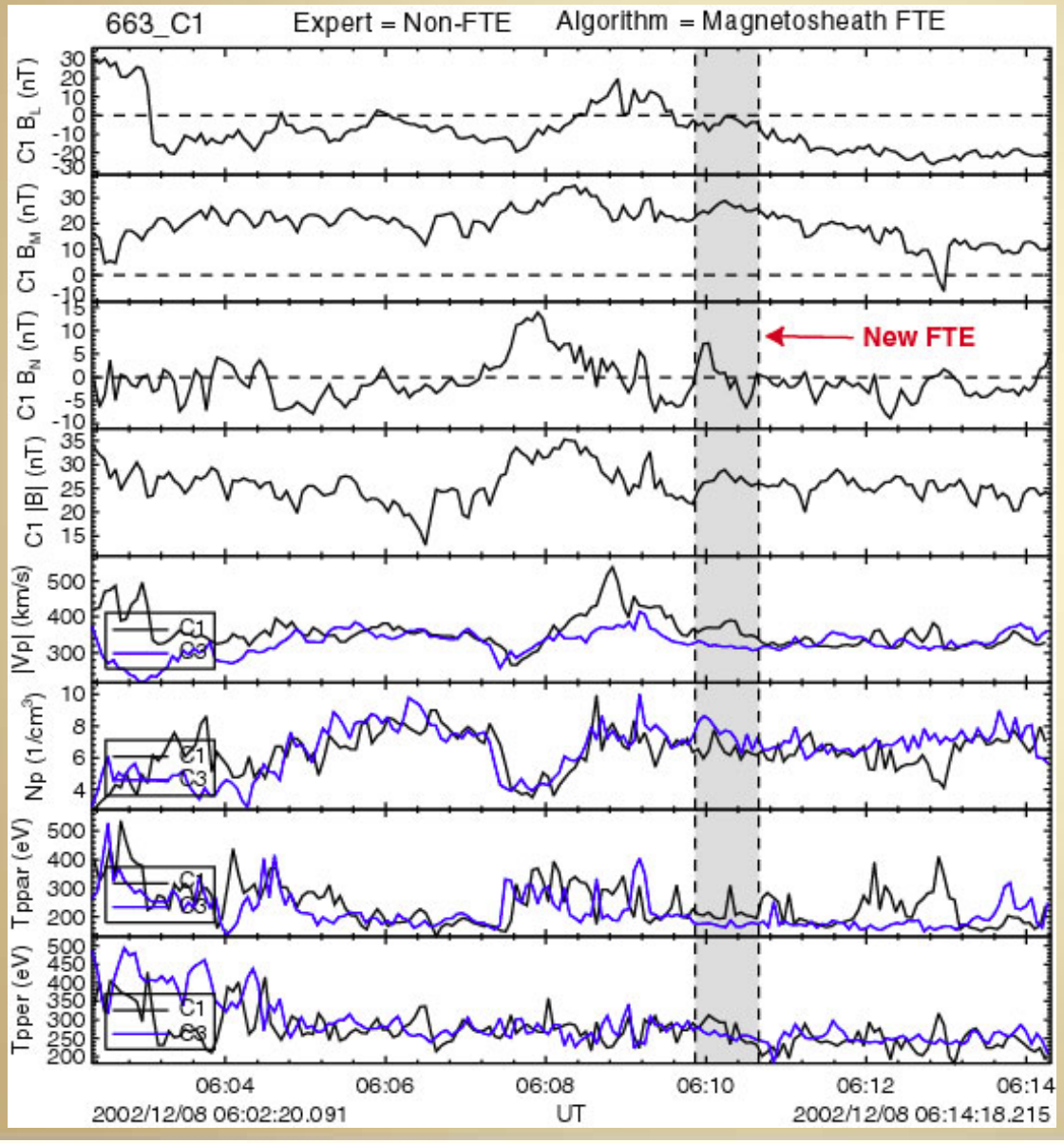
- Dataset containing magnetospheric and IMF condition measurements:
  - Bx, By, Bz, BL, BM, BN, Bmag, Np, Vx, Vy, Vz, Vmag, Tpar, Tper
- 615 events (~30% nonFTEs)
- Labeled data:
  - magnetosheath FTEs, magnetospheric FTEs and nonFTEs
- Each time series was 12 minutes in duration

Efficiency Gains with Physics Mining



Physics mining only asks that you indicate the existence of an event - no need to provide start and end times, as in traditional approach.

Successful Discovery of FTE in Cluster Data



Physics mining successfully identified new FTE that was missed by visual inspection

Physics Mining – Confidence Study

Three-Tier Classification		Classification of FTE vs Non-FTE					Classification of Msheath vs Msphere FTE		
MODEL	Correctly classified (P_3)	Correctly classified (FTE vs non-FTE) (P_1)	# of false non-FTEs		# of false FTEs		Correctly classified (msth vs msph FTE) (P_2)	# of false msth FTEs	# of false msph FTEs
			msth	msph	msth	msph			
$T_\perp/T_\parallel$	59%	83%	7	0	23	4	67%	50	0
N_p	77%	92%	2	0	13	1	79%	30	1
B_{tot}	91%	95%	0	0	8	2	94%	0	9
B_z	92%	93%	9	0	14	0	98%	0	3
T_p	92%	97%	0	0	6	0	93%	0	10
B_L	92%	96%	0	0	7	1	95%	3	5
$T_\parallel$	93%	97%	0	0	5	1	94%	0	9
$T_\perp$	93%	99%	0	0	2	0	91%	11	2
V_x	93%	98%	0	0	4	0	93%	7	4
B_x	93%	94%	0	0	11	1	99%	0	2
V_z	94%	97%	0	0	6	0	95%	0	7
B_y	94%	98%	0	0	4	1	95%	0	8
B_N	95%	99%	0	0	2	0	94%	3	6
V_y	96%	98%	0	0	3	2	98%	0	3
B_M	96%	99%	0	0	2	1	97%	0	5

Physics Mining – Confidence Study

Three-Tier Classification		Classification of FTE vs Non-FTE					Classification of Msheath vs Msphere FTE		
MODEL	Correctly classified	Correctly classified(FTE vs non-FTE)	# of false non-FTEs		# of false FTEs		Correctly classified (msth vs msph FTE)	# of false msth FTEs	# of false msph FTEs
			msth	msph	msth	msph			
B_x,B_{tot}	93%	95%	0	0	10	0	97%	0	4
B_x,B_y,B_z, B_{tot}	95%	95%	0	0	10	0	100%	0	0
B_x,B_y,B_z, B_{tot},T_p,N_p	96%	96%	0	0	8	0	99%	0	1
B_N,B_{tot},N_p	96%	97%	0	0	6	0	99%	0	2
B_N,B_{tot}	96%	98%	0	0	5	0	98%	0	3
B_N,B_{tot},N_p	96%	97%	0	0	6	0	99%	0	2
B_x,B_y,B_z, B_{tot},V_x,V_y, V_z, T_p,N_p	96%	96%	0	0	8	0	100%	0	0
$T_\perp$,$T_\parallel$	97%	99%	1	0	1	0	97%	0	5

Physics Mining – Confidence Study

Three-Tier Classification		Classification of FTE vs Non-FTE					Classification of Msheath vs Msphere FTE		
MODEL	Correctly classified (P_3)	Correctly classified (FTE vs non-FTE) (P_1)	# of false non-FTEs		# of false FTEs		Correctly classified (msth vs msph FTE) (P_2)	# of false msth FTEs	# of false msph FTEs
			msth	msph	msth	msph			
B_L,B_M,B_N,B_{tot}, T_p,N_p	97%	99%	0	0	3	0	97%	0	4
B_L,B_M,B_N, B_{tot}	98%	98%	0	0	4	0	99%	0	1
B_x,B_y,B_z, B_{tot},V_x,V_y, V_z, N_p,	98%	99%	0	0	3	0	99%	0	2
B_L,B_M,B_N, B_{tot}, N_p, V_x,V_y,V_z,T_p	98%	98%	0	0	4	0	100%	0	0
B_L,B_M,B_N, B_{tot},N_p	99%	99%	0	0	3	0	100%	0	0
B_x,B_{tot}	93%	95%	0	0	10	0	97%	0	4
B_x,B_y,B_z, B_{tot}	95%	95%	0	0	10	0	100%	0	0

Questions answered using query driven analysis approach

Traditional Approach Using Visual Data Analysis	Query Driven Analysis Using MineTool
Transform to normal boundary coordinates using a model of MP	• Can FTEs be found in GSM coordinates? YES • Can FTEs be found based on plasma data alone? YES • What combination of variables are good indicators of FTEs? Surprising results. • Are there FTEs with no enhancement in |B|? YES • Does B_M have a well defined signature in all FTEs? NO. Signatures can be visually indiscernible to human eye. • How complete is our taxonomy of FTEs? Not diverse enough for statistical study. ***1 hour later, science advances***
Choose a specific criterion for FTE D. Sibeck: clear bipolar signatures and well defined B_M component Y. Wang: clear bipolar signature in B_N and |B| enhancements	
Visually inspect the data ***3 years later…***	

What Problems Can It Solve?

- Feature extraction and tracking
  - Substorm injection detection, FTEs, MP crossings
- Classification
  - Flux ropes in the SW
- Dependency analysis
  - What triggers FTEs?
- Modeling
  - Developing 3D model of magnetopause from s/c data
- Anomaly detection
  - Spacecraft charging

Advantage of Intelligent Archiving

- Current Archive Search
  - Search general description of data set for type of data.
  - Requires analyzing data to discover events of interest.

- Intelligent Archive Search
  - Automatically generate metadata about events in the data set using data mining techniques.
  - Search for data ***content*** - what the data set contains.
  - Events of interest are already identified and accessible to the user through a query.

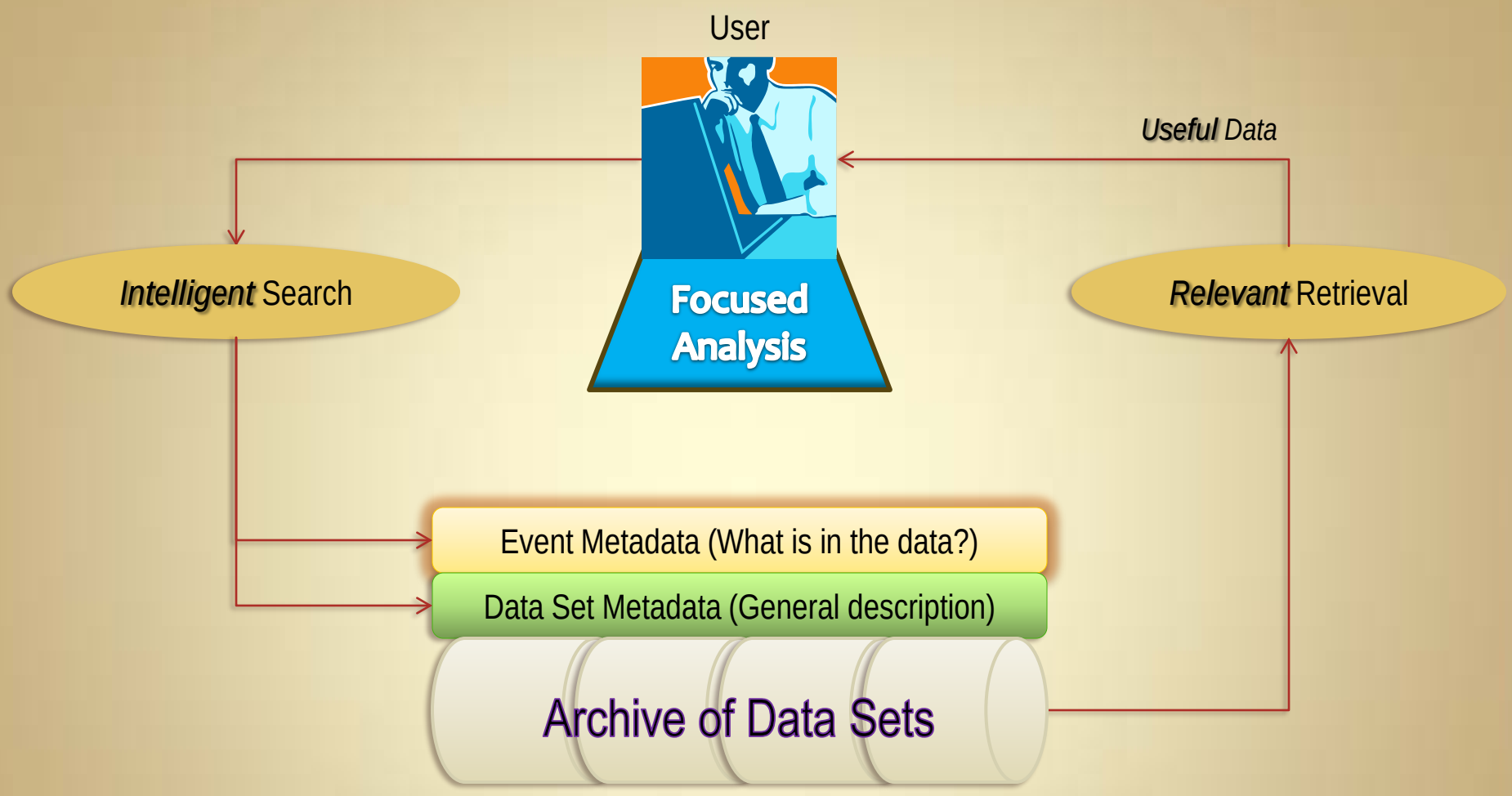
Knowledge is extracted directly from the data about events, promoting intelligent search on archives.

Summary

- Knowledge discovery from increasingly complex and large data sets is a major bottleneck to progress in space sciences today.
- Physics mining techniques are powerful tools that optimize knowledge extraction from large data sets.
- MineTool has been successfully applied to several space physics problems.
- Intelligent archiving results in significant increase in ROI by extracting knowledge about the data content.

Expected Scientific Breakthroughs Using Global Simulations on Exascale

Tera	Peta	Exa	*Too many cores to count*
2D Hybrid	3D Hybrid	3D Hybrid - Real time	
Breakthroughs	*Breakthroughs*	*Expected Breakthroughs*	
Physics of ion fore shock / bow shock	Transport, formation of boundaries such as PSBL, FTE generation, PDL, cusp physics, …	Modeling of events over MHD scales	
	2D full PIC – reduced mass ratio	3D full PIC – reduced mass ratio	3D full PIC Simulation - high mass ratio
	Breakthroughs	*Expected Breakthroughs*	*Expected Breakthroughs*
	The first study of reconnection under realistic configuration	Closure of long standing issues related to onset and time dependence of reconnection in the tail and dayside	Proper capture of instabilities such as LHDI or electron kinetic effects associated with bow shock



The creation of event metadata associated with the data sets is essential to extend reach and richness of data, and to drive collaboration by making events visible to the community of users. Enhancing the metadata allows Intelligent Search, significantly increasing value of the archive.