SAND2010-6734P

Network Algorithms for Information Analysis using the Titan Toolkit

October 6, 2010

William C. McLendon III
Sandia National Laboratories

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company,
for the United States Department of Energy's National Nuclear Security Administration
under contract DE-AC04-94AL85000. This document is SAND Number: 2009-6318 P

Outline

- **Introduction to the toolkit**
  - Toolkit provisions & data pipelines

- **Applications**
  - Text Corpora Analysis
  - Fun with Web Crawls
  - Network Packet Streaming

- **Coding Example**
  - Using Python to load a graph from disk

What is Titan?

Data Structures
- Table
- Tree
- DAG
- Directed Graph
- Undirected Graph
- Sparse N-way Array
- Dense N-way Array
- Unicode Text

Database Drivers
- MySQL
- Postgres
- Oracle
- SQLite
- ODBC
- Netezza

Readers
- Dimacs
- DOT
- GXL
- Chaco
- XML
- Tulip
- DelimitedText
- Unicode Delimited Text
- FixedWidth
- ISI, RIS
- Palantir XML

Multidimensional Analysis
- TPP / PARAFAC
- Trilinos Integration

Graph Algorithms
- Network Communities
- ST Search
- CSG Search
- Temporal Search
- Breadth First Search
- Connected Components
- Biconnected Components
- Brandes Centrality
- Subgraph Isomorphism

MATLAB Integration

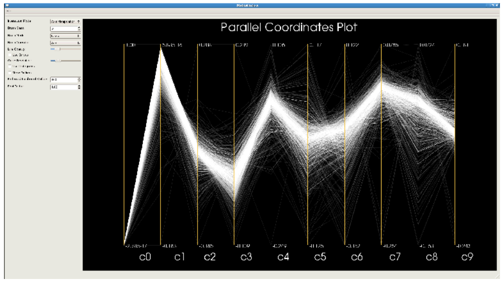
R Integration

Parallel Statistics Algorithms
- Descriptive
- Order
- Correlative
- Contingency

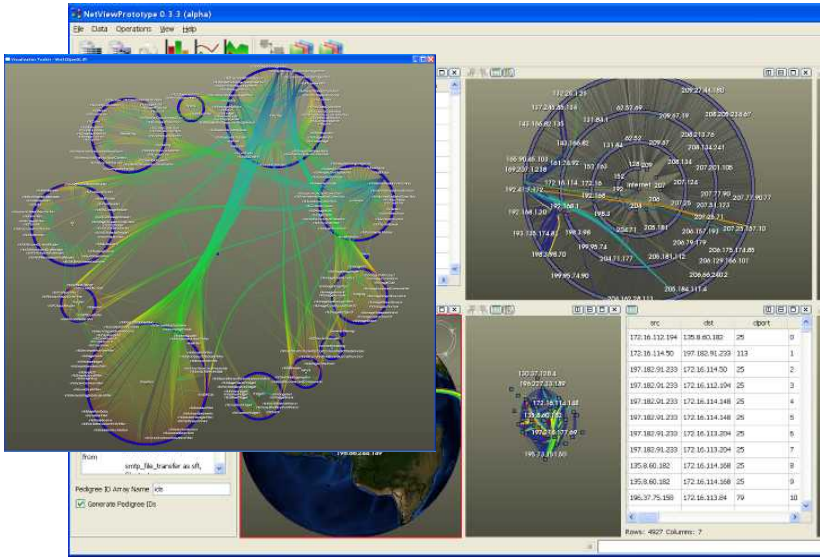
Linear-Time Graph Layouts
- GSpace
- Hierarchical
- Clustered
- Three tree-based variants

Text Analysis
- Latent Semantic Analysis (LSA)
- Latent Dirichlet Allocation (LDA)

Multiple View Types
- Render (3D)
- Graph
- Hierarchical Graph
- Tree
- Treemap
- Georeferenced
- Parallel Coordinates
- Radial Tree Ring

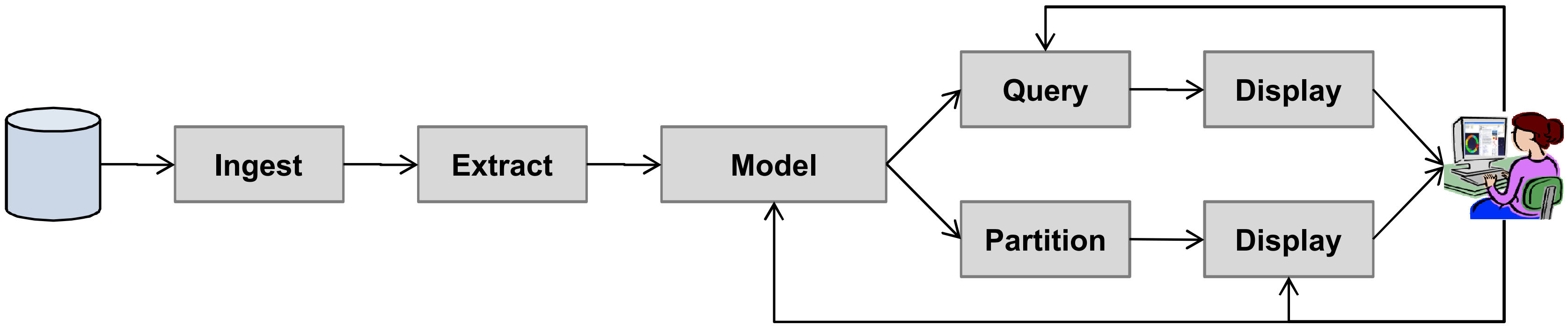


Multiple Platforms / Languages
- Windows, Linux, OSX, HPC
- Write components in C++
- Use with C++, TCL, Python, Java, .NET, COM
- Use as OverView "plugins"



Building Applications Using Pipelines

Applications are built by composing pipelines of filters.



Provides a direct mapping between whiteboard and code.

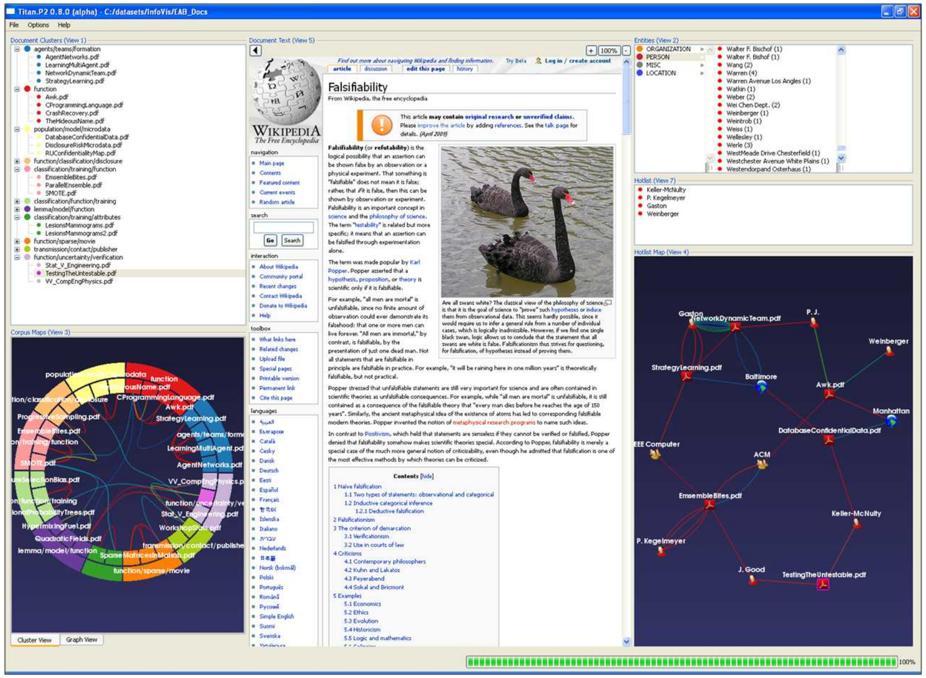
Provides flexibility / exploration (for developers!)

Manages execution:

- Manages the “flow” of data through components.
- In what order are components executed?
- What needs to be recomputed when parameters change?

Application: Text Corpora Analysis

- **Sense-making of a corpus of documents.**
- **Document-Document Similarity**
  - Computed using clustering algorithms (LDA, LSA, etc.)
- **Entity-Document Similarity**
  - Named Entity Recognition
  - Merged DD and DE graphs
  - Connection Subgraphs Searches



Titan.P2 0.8.0 (alpha) - C:/datasets/InfoVis/EAB_Docs

- agents/teams/formation
  - AgentNetworks.pdf
  - LearningMultiAgent.pdf
  - NetworkDynamicTeam.pdf
  - StrategyLearning.pdf
- function
  - Awk.pdf
  - CProgrammingLanguage.pdf
  - CrashRecovery.pdf
  - TheHideousName.pdf
- population/model/microdata
  - DatabaseConfidentialData.pdf
  - DisclosureRiskMicrodata.pdf
  - RUConfidentialityMap.pdf
- function/classification/disclosure
- classification/training/function
  - EnsembleBites.pdf
  - ParallelEnsemble.pdf
  - SMOTE.pdf
- classification/function/training
- lemma/model/function
- classification/training/attributes
  - LesionsMammograms.pdf
  - LesionsMammograms2.pdf
- function/sparse/movie
- transmission/contact/publisher
- function/uncertainty/verification
  - Stat_V_Engineering.pdf
  - TestingTheUntestable.pdf
  - VV_CompEngPhysics.pdf

Falsifiability

From Wikipedia, the free encyclopedia

This article **may contain original research or unverified claims**. Please improve the article by adding references. See the talk page for details. *(April 2009)*

Falsifiability (or **refutability**) is the logical possibility that an assertion can be shown false by an observation or a physical experiment. That something is "falsifiable" does not mean it is false; rather, that *if* it is false, then this can be shown by observation or experiment. Falsifiability is an important concept in science and the philosophy of science. The term "testability" is related but more specific; it means that an assertion can be falsified through experimentation alone.

Are all swans white? The classical view of the philosophy of science is that it is the goal of science to "prove" such hypotheses or induce them from observational data. This seems hardly possible, since it would require us to infer a general rule from a number of individual cases, which is logically inadmissible. However, if we find one single black swan, logic allows us to conclude that the statement that all swans are white is false. Falsificationism thus strives for questioning, for falsification, of hypotheses instead of proving them.

The term was made popular by Karl Popper. Popper asserted that a hypothesis, proposition, or theory is scientific only if it is falsifiable.

For example, "all men are mortal" is unfalsifiable, since no finite amount of observation could ever demonstrate its falsehood: that one or more men can live forever. "All men are immortal," by contrast, is falsifiable, by the presentation of just one dead man. Not all statements that are falsifiable in principle are falsifiable in practice. For example, "it will be raining here in one million years" is theoretically falsifiable, but not practical.

Popper stressed that unfalsifiable statements are still very important for science and are often contained in scientific theories as unfalsifiable consequences. For example, while "all men are mortal" is unfalsifiable, it is still contained as a consequence of the falsifiable theory that "every man dies before he reaches the age of 150 years". Similarly, the ancient metaphysical idea of the existence of atoms has led to corresponding falsifiable modern theories. Popper invented the notion of metaphysical research programs to name such ideas.

In contrast to Positivism, which held that statements are senseless if they cannot be verified or falsified, Popper denied that falsifiability somehow makes scientific theories special. According to Popper, falsifiability is merely a special case of the much more general notion of criticizability, even though he admitted that falsification is one of the most effective methods by which theories can be criticized.

Contents [hide]

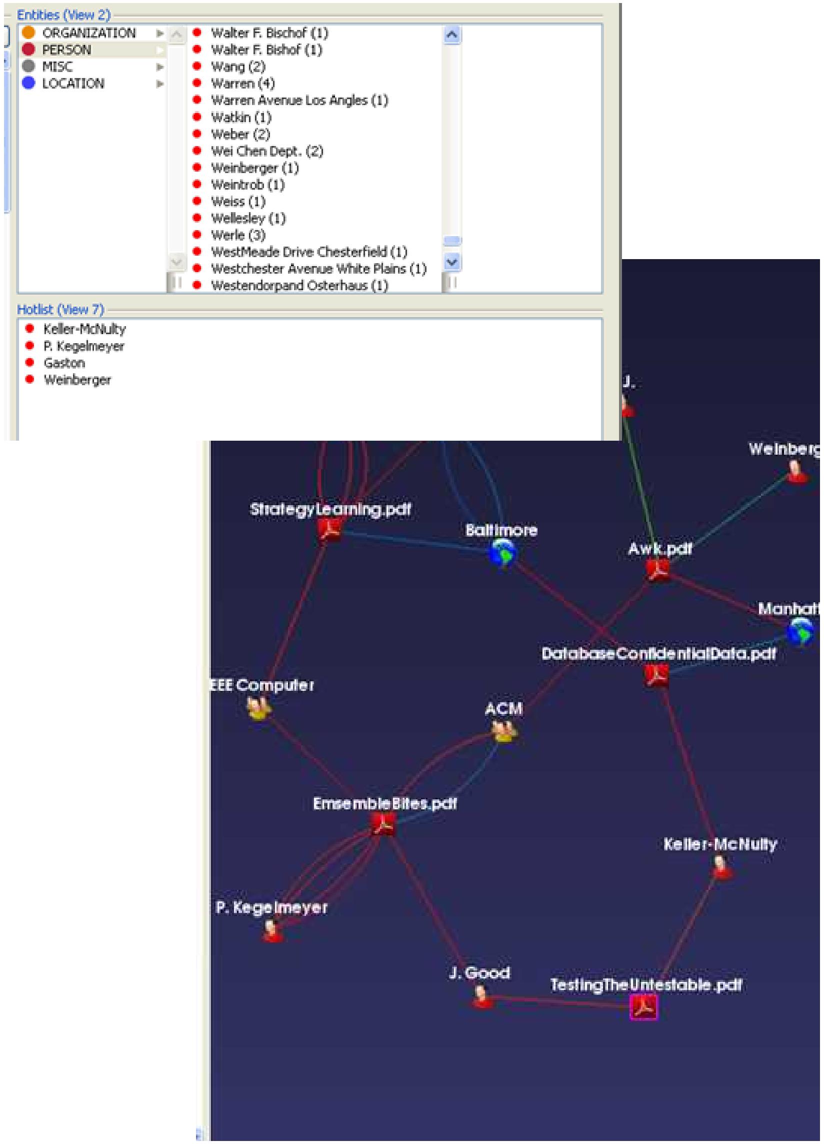
- Walter F. Bischof (1)
- Walter F. Bishof (1)
- Wang (2)
- Warren (4)
- Warren Avenue Los Angles (1)
- Watkin (1)
- Weber (2)
- Wei Chen Dept. (2)
- Weinberger (1)
- Weintrob (1)
- Weiss (1)
- Wellesley (1)
- Werle (3)
- WestMeade Drive Chesterfield (1)
- Westchester Avenue White Plains (1)
- Westendorpand Osterhaus (1)

Hotlist (View 7)

- Keller-McNulty
- P. Kegelmeyer
- Gaston
- Weinberger

Following a Soap Opera

- The “Hot List” is a subset of named-entities that are “interesting”…
- Generate a community graph that connects selected entities
- Why is this interesting?
  - Shows how entities are related and which documents the info came from



Entities (View 2)

- ● ORGANIZATION
- ● PERSON
- ● MISC
- ● LOCATION

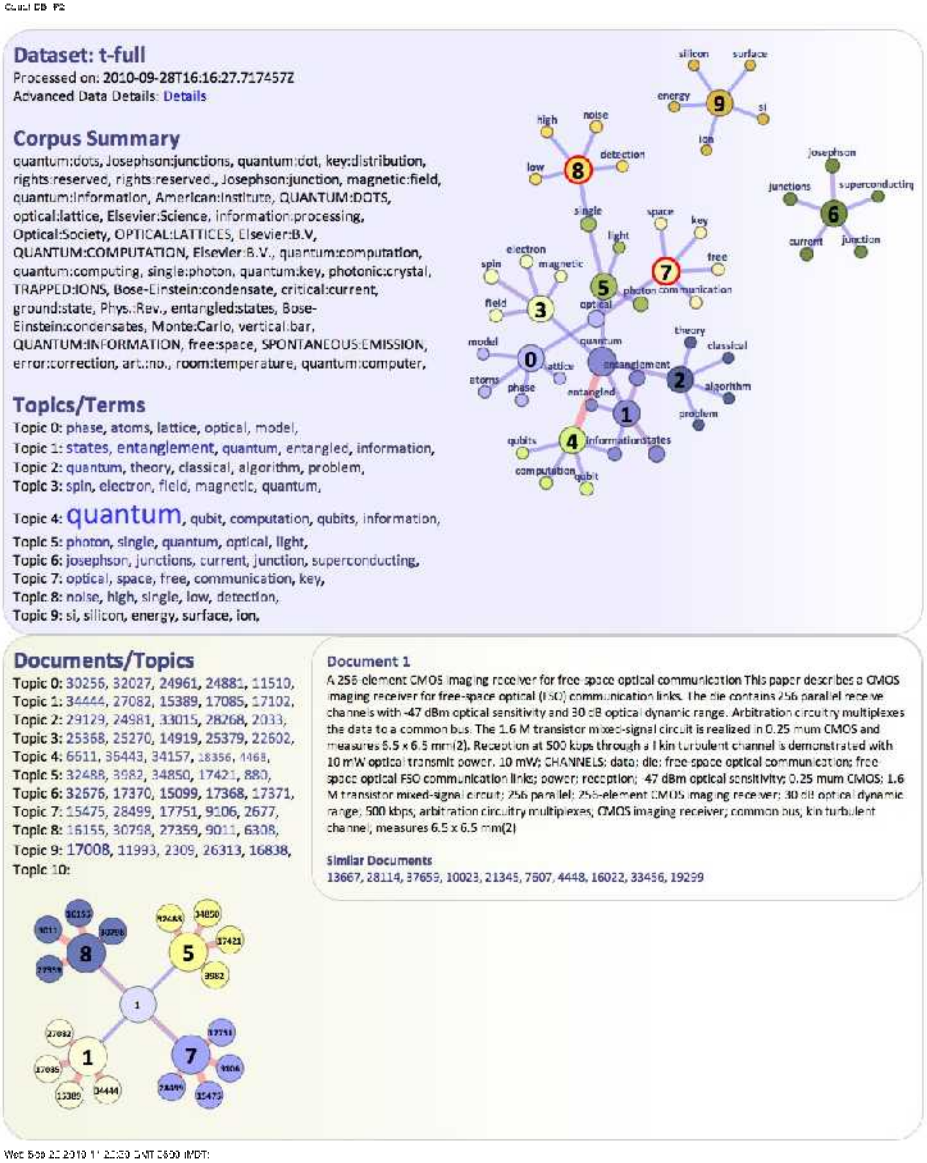
- Walter F. Bischof (1)
- Walter F. Bishof (1)
- Wang (2)
- Warren (4)
- Warren Avenue Los Angles (1)
- Watkin (1)
- Weber (2)
- Wei Chen Dept. (2)
- Weinberger (1)
- Weintrob (1)
- Weiss (1)
- Wellesley (1)
- Werle (3)
- WestMeade Drive Chesterfield (1)
- Westchester Avenue White Plains (1)
- Westendorpand Osterhaus (1)

Hotlist (View 7)

- Keller-McNulty
- P. Kegelmeyer
- Gaston
- Weinberger

More Corpora Analysis

- **Problem: Determine the topicality of a set of documents.**
- **Reuses many elements of the previous application data pipeline.**
- **Key differences:**
  - Implements a subset
  - Uses a web-delivery model (CouchDB)
  - 100% browser delivery
  - Developed by two people using 10% time over a couple weeks.



CouchDB P2

Dataset: t-full
Processed on: 2010-09-28T16:16:27.717457Z
Advanced Data Details: Details

Corpus Summary
quantum:dots, Josephson:junctions, quantum:dot, key:distribution, rights:reserved, rights:reserved., Josephson:junction, magnetic:field, quantum:Information, American:Institute, QUANTUM:DOTS, optical:lattice, Elsevier:Science, information:processing, Optical:Society, OPTICAL:LATTICES, Elsevier:B.V, QUANTUM:COMPUTATION, Elsevier:B.V., quantum:computation, quantum:computing, single:photon, quantum:key, photonic:crystal, TRAPPED:IONS, Bose-Einstein:condensate, critical:current, ground:state, Phys.:Rev., entangled:states, Bose-Einstein:condensates, Monte:Carlo, vertical:bar, QUANTUM:INFORMATION, free:space, SPONTANEOUS:EMISSION, error:correction, art.:no., room:temperature, quantum:computer,

Topics/Terms
Topic 0: phase, atoms, lattice, optical, model,
Topic 1: states, entanglement, quantum, entangled, information,
Topic 2: quantum, theory, classical, algorithm, problem,
Topic 3: spin, electron, field, magnetic, quantum,
Topic 4: quantum, qubit, computation, qubits, information,
Topic 5: photon, single, quantum, optical, light,
Topic 6: josephson, junctions, current, junction, superconducting,
Topic 7: optical, space, free, communication, key,
Topic 8: noise, high, single, low, detection,
Topic 9: si, silicon, energy, surface, ion,

Documents/Topics
Topic 0: 30256, 32027, 24961, 24881, 11510,
Topic 1: 34444, 27082, 15389, 17085, 17102,
Topic 2: 29129, 24981, 33015, 28268, 2033,
Topic 3: 25368, 25270, 14919, 25379, 22602,
Topic 4: 6511, 35443, 34157, 18356, 4468,
Topic 5: 32488, 3982, 34850, 17421, 880,
Topic 6: 32676, 17370, 15099, 17368, 17371,
Topic 7: 15475, 28499, 17751, 9106, 2677,
Topic 8: 16155, 30798, 27359, 9011, 6308,
Topic 9: 17008, 11993, 2309, 26313, 16838,
Topic 10:

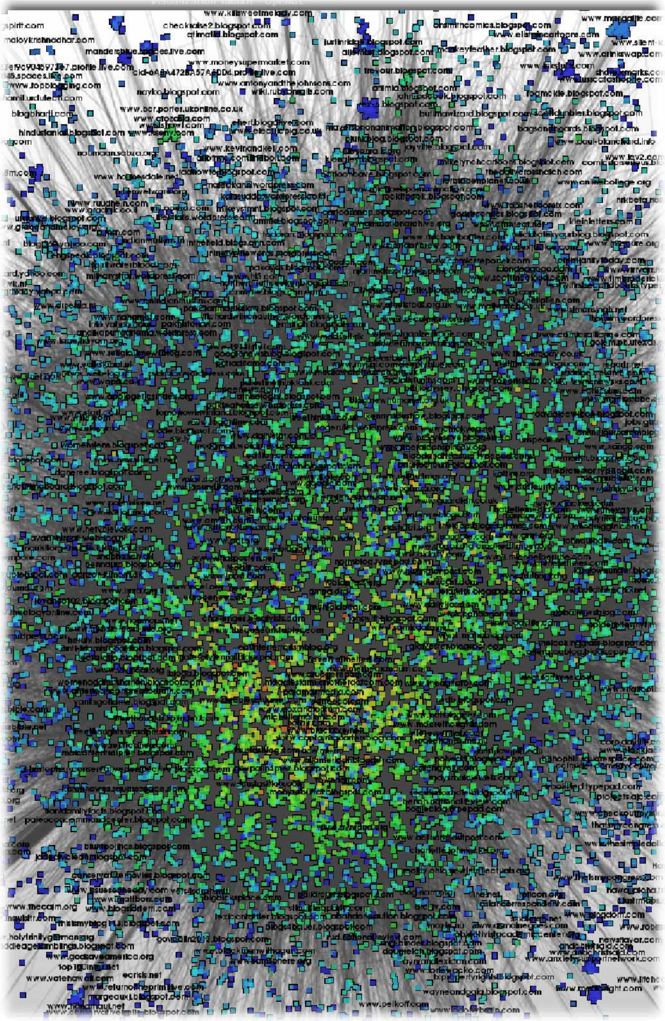
Document 1
A 256-element CMOS imaging receiver for free-space optical communication This paper describes a CMOS imaging receiver for free-space optical (FSO) communication links. The die contains 256 parallel receive channels with -47 dBm optical sensitivity and 30 dB optical dynamic range. Arbitration circuitry multiplexes the data to a common bus. The 1.6 M transistor mixed-signal circuit is realized in 0.25 mum CMOS and measures 6.5 x 6.5 mm(2). Reception at 500 kbps through a I km turbulent channel is demonstrated with 10 mW optical transmit power. 10 mW; CHANNELS; data; die; free-space optical communication; free space optical FSO communication links; power; reception; -47 dBm optical sensitivity; 0.25 mum CMOS; 1.6 M transistor mixed-signal circuit; 256 parallel; 256-element CMOS imaging receiver; 30 dB optical dynamic range; 500 kbps; arbitration circuitry multiplexes; CMOS imaging receiver; common bus; km turbulent channel; measures 6.5 x 6.5 mm(2)

Similar Documents
13667, 28114, 37659, 10023, 21345, 7507, 4448, 16022, 33456, 19299

Wed Sep 22 2010 11:22:30 GMT-0600 (MDT)

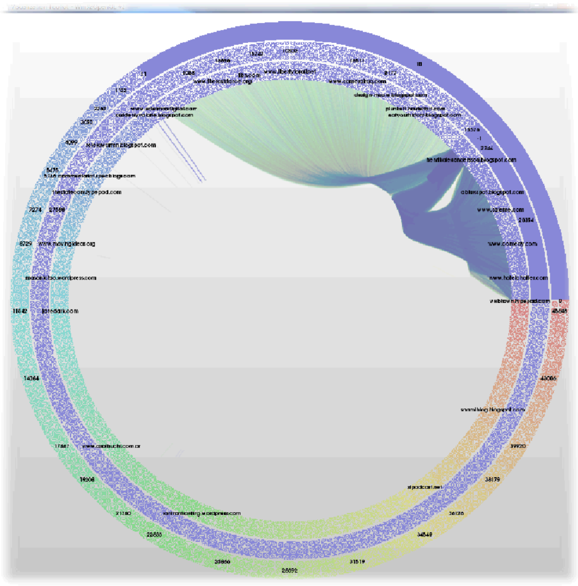
Web-Crawl Data

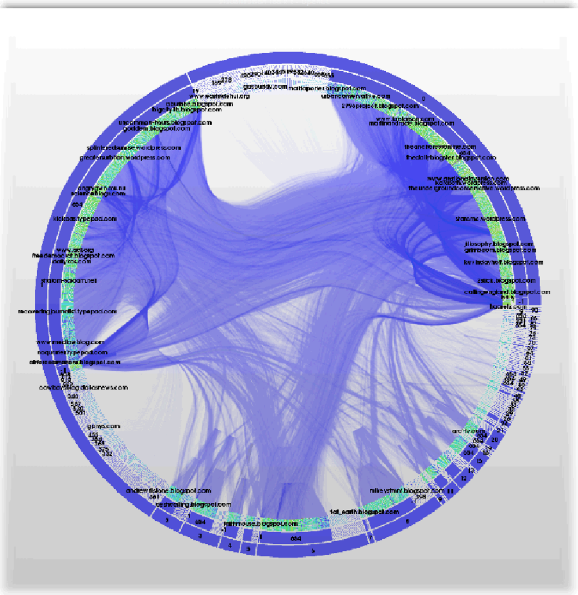
- **Web-crawl data project**
- **Analyze a targeted web-crawl to:**
  - Identify topological 'communities' within the web network.
  - Perform text analysis on the page content (LSA, LDA) to pull topics of interest.
  - Cluster the pages around topics
  - Track the evolution of topics within communities over time.
  - Present results as an interactive set of charts and plots accessible through a common web browser.



Validating a Graph

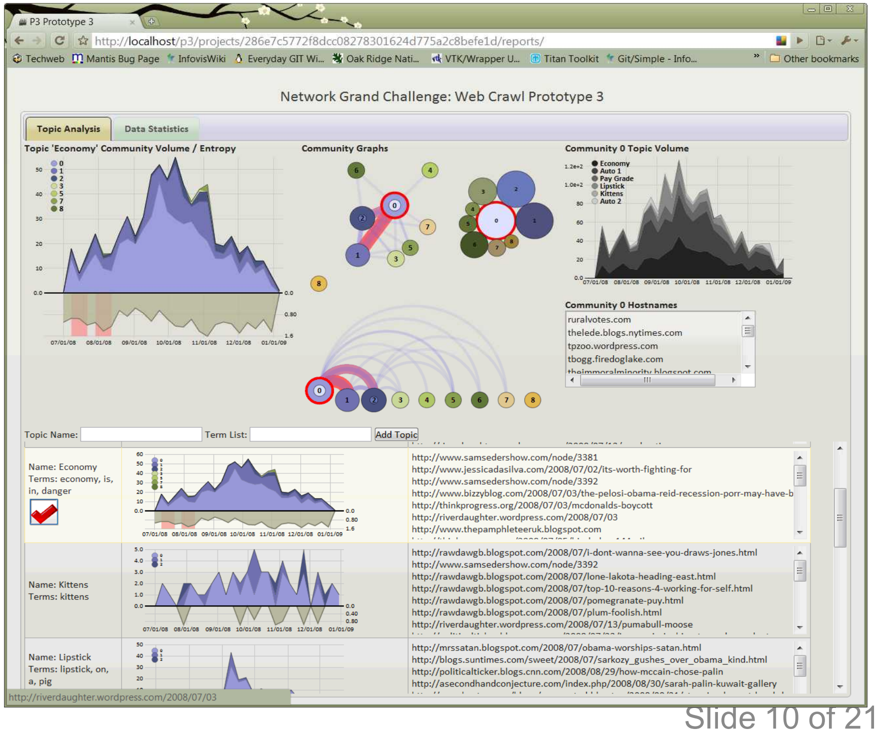
- **Often we receive data that we did not collect ourselves.**
- **We must validate the data**
  - Real world data can be messy and/or incomplete.
  - If the data does not conform to expectations, we must discover why.
- **In our web-crawl example, the initial data set was incorrect.**
  - Turned out to be a bug in the post processing scripts that pushed the crawl into a database.





The Web Analysis Application

- **Toolkit breadth leveraged**
  - – Database Drivers
  - – Web Server / Client tiers in Titan
  - – Latent Semantic Analysis (LSA)
  - – Document Clustering
  - – Statistics
  - – Graph Algorithms
  - – Protovis for charts
- **Results delivered using a browser.**
- **App developed in a short time period.**



Network Grand Challenge: Web Crawl Prototype 3

Community 0 Hostnames

ruralvotes.com
thelede.blogs.nytimes.com
tpzoo.wordpress.com
tbogg.firedoglake.com
theimmoralminority.blogspot.com

Name: Economy
Terms: economy, is, in, danger

http://www.samsedershow.com/node/3381
http://www.jessicadasilva.com/2008/07/02/its-worth-fighting-for
http://www.samsedershow.com/node/3392
http://www.bizzyblog.com/2008/07/03/the-pelosi-obama-reid-recession-porr-may-have-b
http://thinkprogress.org/2008/07/03/mcdonalds-boycott
http://riverdaughter.wordpress.com/2008/07/03
http://www.thepamphleteeruk.blogspot.com

Name: Kittens
Terms: kittens

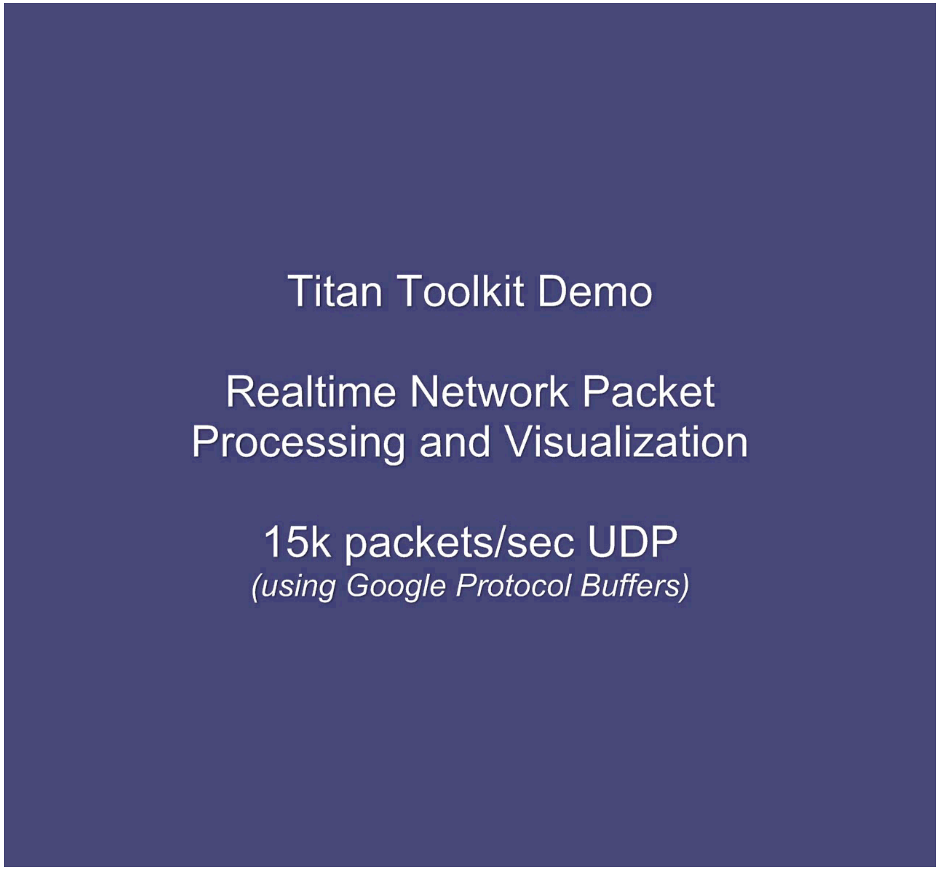
http://rawdawgb.blogspot.com/2008/07/i-dont-wanna-see-you-draws-jones.html
http://www.samsedershow.com/node/3392
http://rawdawgb.blogspot.com/2008/07/lone-lakota-heading-east.html
http://rawdawgb.blogspot.com/2008/07/top-10-reasons-4-working-for-self.html
http://rawdawgb.blogspot.com/2008/07/pomegranate-puy.html
http://rawdawgb.blogspot.com/2008/07/plum-foolish.html
http://riverdaughter.wordpress.com/2008/07/13/pumabull-moose

Name: Lipstick
Terms: lipstick, on, a, pig

http://mrssatan.blogspot.com/2008/07/obama-worships-satan.html
http://blogs.suntimes.com/sweet/2008/07/sarkozy_gushes_over_obama_kind.html
http://politicalticker.blogs.cnn.com/2008/08/29/how-mccain-chose-palin
http://asecondhandconjecture.com/index.php/2008/08/30/sarah-palin-kuwait-gallery

http://riverdaughter.wordpress.com/2008/07/03

Network Packet Streaming



Titan Toolkit Demo

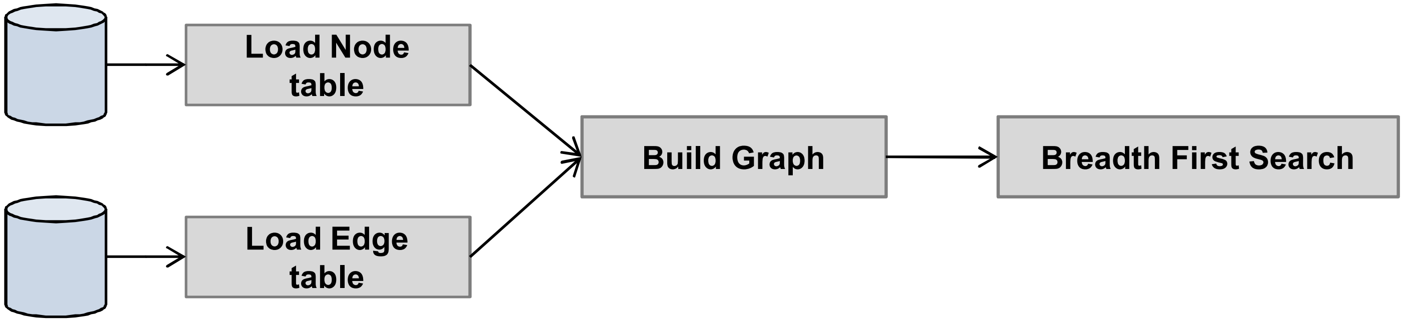
Realtime Network Packet
Processing and Visualization

15k packets/sec UDP
(using Google Protocol Buffers)

- **Real-Time Streaming App**
- **15k packets/second**
- **Combines monolithic app model with ‘web’ tech**
- **Real-Time modification of views by editing underlying javascript**

Code Example: Loading a Network

- **The following few slides walks through a quick example that loads a graph using Titan via Python.**
  - Load a graph from two ‘CSV’ files
  - Run an algorithm on it, Breadth-First Search



Load the Nodes Table

csv_to_graph.py

```
from vtk import *

nodes = vtkDelimitedTextReader()
nodes.SetFieldDelimiterCharacters(“,”)nodes.Se
tHaveHeaders(True)
nodes.SetDetectNumericColumns(True)
nodes.SetFileName(“nodes.csv”)
```

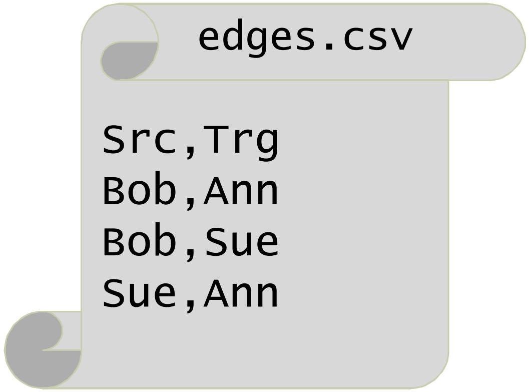
nodes.csv

```
ID,Name,Age
1,Bob,12
2,Ann,25
3,Sue,72
```

- **Explanation**
  - Loads the VTK libraries
  - Load the file *nodes.csv* into the *nodes* pipeline object using *vtkDelimitedTextReader*.

Load the Edges Table

```
edges = vtkDelimitedTextReader()
edges.SetFieldDelimiterCharacters(“,”)edges.Se
tHaveHeaders(True)
edges.SetDetectNumericColumns(True)
edges.SetFileName(“edges.csv”)
```



edges.csv

```
Src,Trg
Bob,Ann
Bob,Sue
Sue,Ann
```

- **Explanation**
  - Essentially same as previous slide, but we’re loading the edges file this time.

Construct Graph from Nodes & Edges

```
Graph = vtkTableToGraph()
Graph.SetDirected(True)
Graph.AddInputConnection(edges,GetOutputPort())
Graph.SetVertexTableConnection(nodes.GetOutputPort())
Graph.AddLinkVertex(“src”,”Name”,False)
Graph.AddLinkVertx(“trg”,”Name”,False)
Graph.AddLinkEdge(“src”, “trg”)
```

- **Explanation**
  - Uses vtkTableToGraph, which takes two tables (vertices, edges) as inputs.
  - Vertices are created from the nodes table
  - Edges are created by linking the “src” and “trg” entries in rows against nodes using the “Name” field.

Run the BFS Algorithm

```
bfs = vtkBoostBreadthFirstSearch()
bfs.AddInputConnection(Graph.GetOutputPort())
bfs.SetOriginVertex(“Bob”)

bfs.Update()
```

- **Explanation**
  - Attaches vtkBoostBreadthFirstSearch to the pipeline.
  - Computes a BFS ordering on the graph.
- **Pipeline Construction**
  - The ‘output port’ from one filter is connected to the ‘input port’ of the next.
  - Once we’ve got the pipeline started, we can add additional filters to operate on the data.

Conclusion

- Titan is a toolkit which unifies many powerful toolkits around a flexible pipeline model.
- This flexibility allows development of applications to solve complex informatics problems with a small developer team.
- Multiple language support (Python, Java, etc.)
- The toolkit is fully open-source and available via Git.

http://titan.sandia.gov

Questions / Discussion

Related Work

Prefuse Visualization Toolkit

A Java-based toolkit for building interactive information visualization applications.

Tulip Toolkit

A C++ toolkit for building interactive information visualization applications.

GraphViz

A set of libraries specifically for the visualization of many different types of graphs.

InfoVis Toolkit

A Java-based toolkit for development of Information Visualization applications and components.

InfoVis Cyberinfrastructure

It is a set of libraries that provide a simple and uniform programming-interface to algorithms using the Eclipse Rich Client Platform (RCP).

Piccolo Toolkit

Piccolo is a layer built on top of a lower level graphics API. Currently supports Java and C# (Piccolo.Java Piccolo.NET)

GeoVista Studio

GeoVISTA Studio is an open software development environment designed for geospatial data.

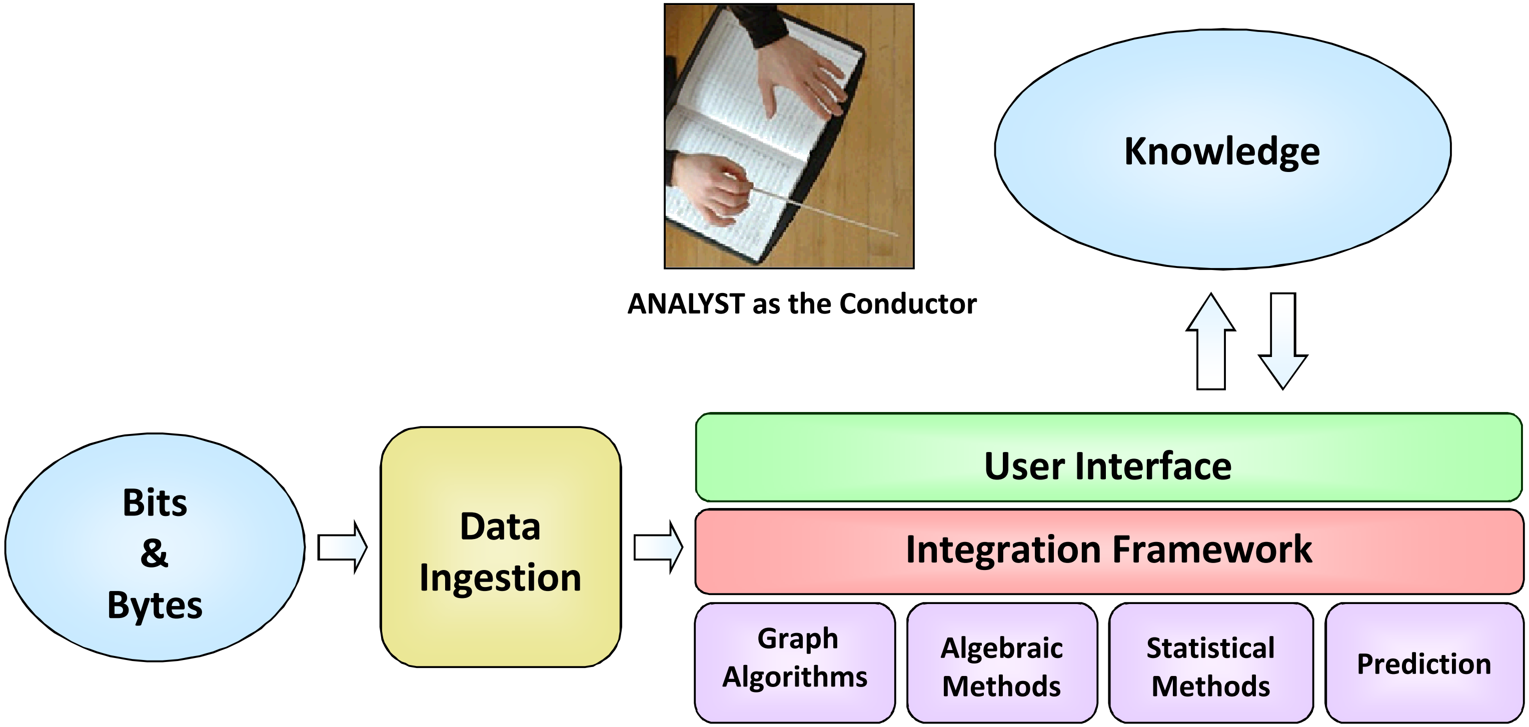
Improvise

A coordinated multi-view, Java-based, open source (GPL) information visualization application for MacOS X, Linux, and Windows.

Protovis

Open-source library using JavaScript and SVG for web-native visualizations; works with any modern web browser, no plugins required.

Titan Informatics Toolkit



ANALYST as the Conductor

Knowledge

User Interface

Integration Framework

Bits & Bytes

Data Ingestion

Graph Algorithms

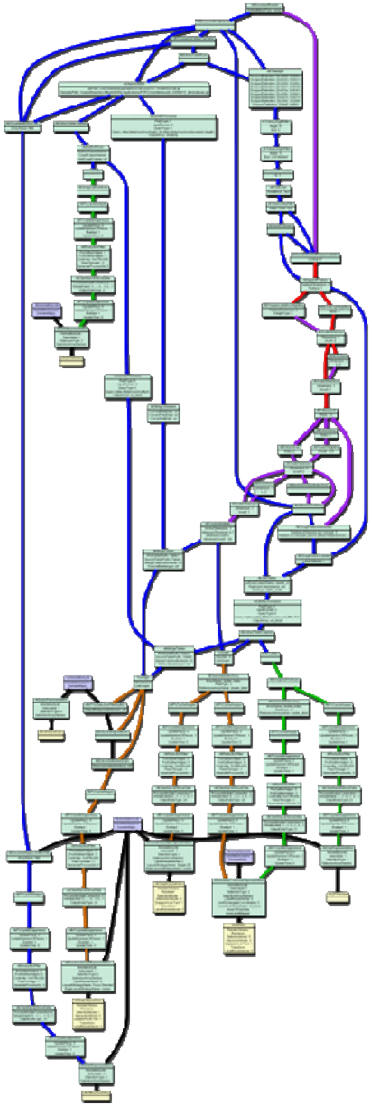
Algebraic Methods

Statistical Methods

Prediction

What is Titan?

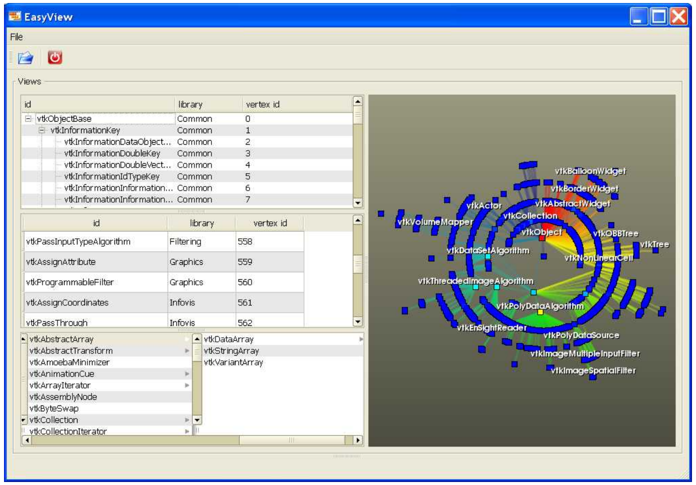
A flexible parallel pipeline architecture, written in C++



- 🟦 Table
- 🟩 Tree
- 🟧 Graph
- 🟥 Sparse Matrix
- 🟪 Dense Matrix
- ⬛ Selection Link

Titan Language Bindings

C++ on Windows XP



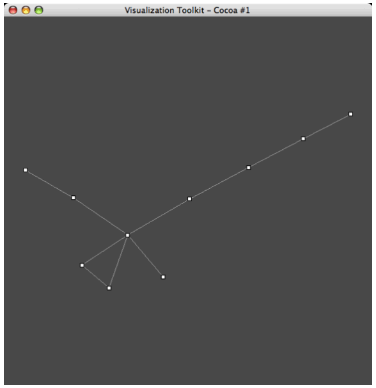
id	library	vertex id
vtkObjectBase	Common	0
vtkInformationKey	Common	1
vtkInformationDataObject...	Common	2
vtkInformationDoubleKey	Common	3
vtkInformationDoubleVect...	Common	4
vtkInformationIdTypeKey	Common	5
vtkInformationInformation...	Common	6
vtkInformationInformation...	Common	7

id	library	vertex id
vtkPassInputTypeAlgorithm	Filtering	558
vtkAssignAttribute	Graphics	559
vtkProgrammableFilter	Graphics	560
vtkAssignCoordinates	Infovis	561
vtkPassThrough	Infovis	562

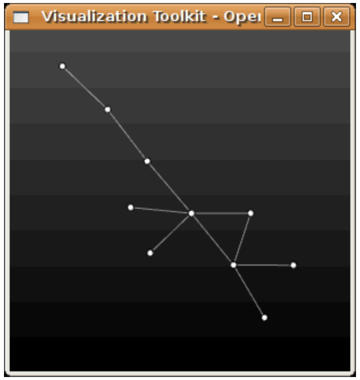
vtkAbstractArray
vtkAbstractTransform
vtkAmoebaMinimizer
vtkAnimationCue
vtkArrayIterator
vtkAssemblyNode
vtkByteSwap
vtkCollection
vtkCollectionIterator

vtkDataArray
vtkStringArray
vtkVariantArray

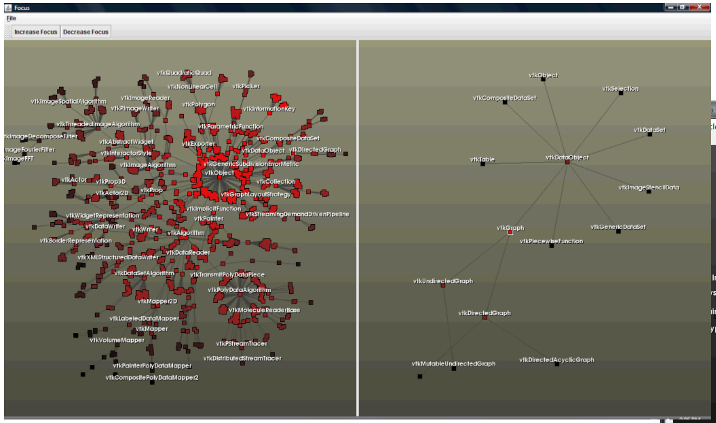
Python on OSX



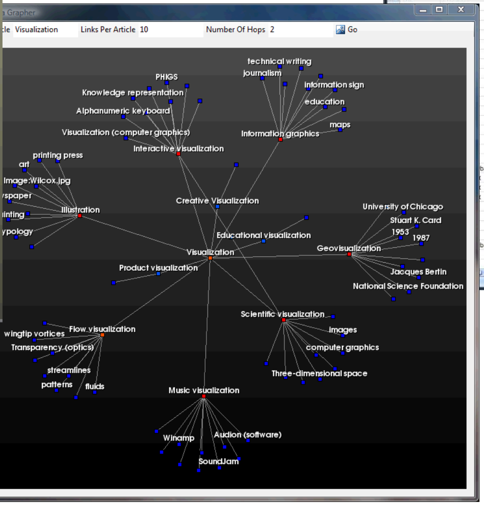
Tcl/Tk on Linux



Java on Vista

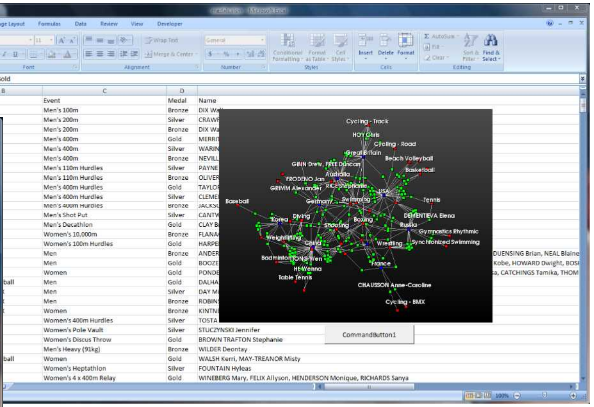


.NET on Vista



cle Visualization   Links Per Article 10   Number Of Hops 2   Go

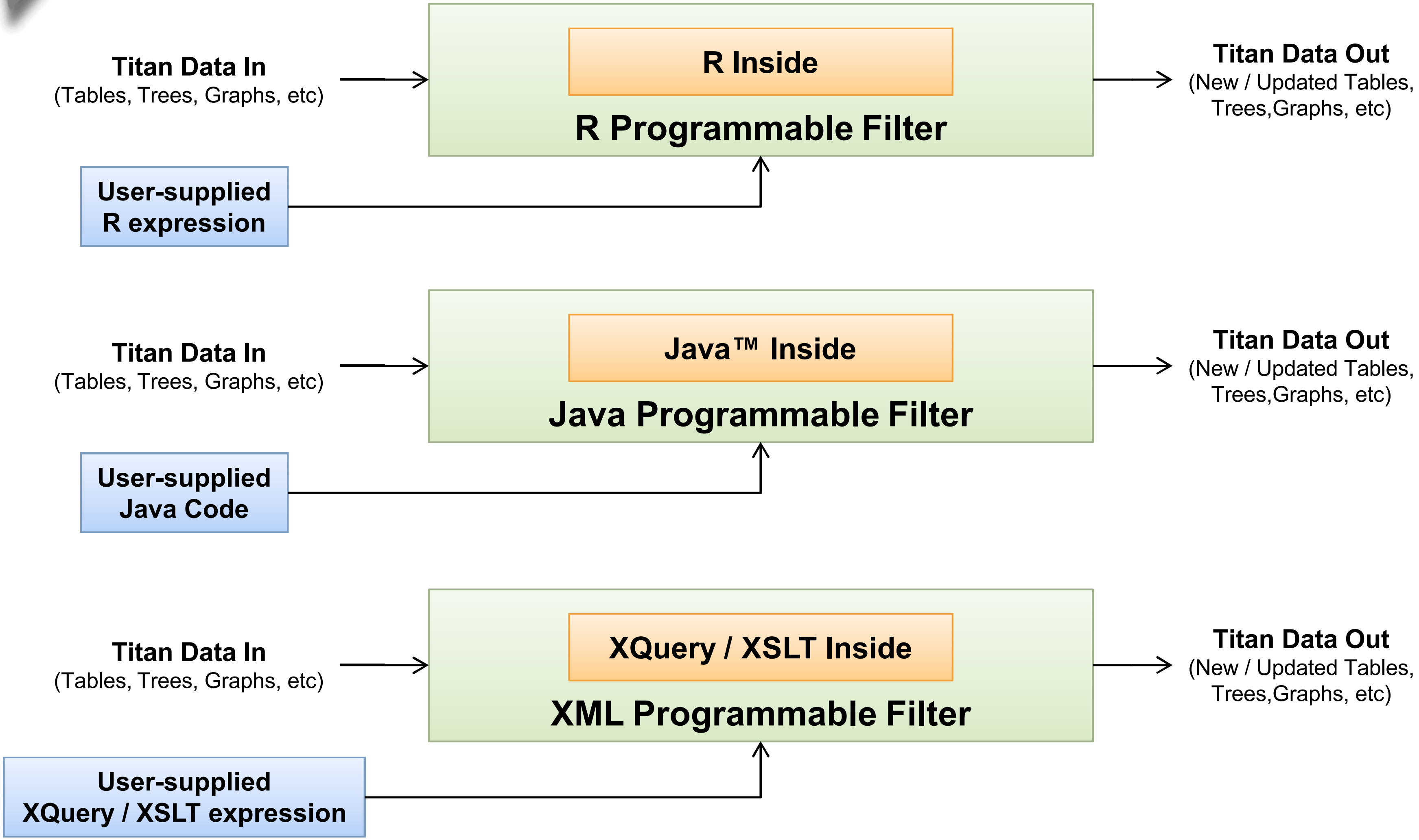
COM and Excel, on Vista



Gold

	A	B	C	D	
1	Country	Discipline	Event	Medal	Name
2	USA	Athletics	Men's 100m	Bronze	DIX Wa
			Men's 200m	Silver	CRAWF
			Men's 200m	Bronze	DIX Wa
			Men's 400m	Gold	MERRIT
			Men's 400m	Silver	WARIN
			Men's 400m	Bronze	NEVILL
			Men's 110m Hurdles	Silver	PAYNE
			Men's 110m Hurdles	Bronze	OLIVER
			Men's 400m Hurdles	Gold	TAYLOR
			Men's 400m Hurdles	Silver	CLEMEI
			Men's 400m Hurdles	Bronze	JACKSC
			Men's Shot Put	Silver	CANTW
			Men's Decathlon	Gold	CLAY B
			Women's 10,000m	Bronze	FLANA
			Women's 100m Hurdles	Gold	HARPE
			Men	Bronze	ANDER DUENSING Brian, NEAL Blaine
			Men	Gold	BOOZE Kobe, HOWARD Dwight, BOSH
			Women	Gold	PONDE a, CATCHINGS Tamika, THOM
		ball	Men	Gold	DALHA
			Men	Silver	DAY M
			Men	Bronze	ROBINS
			Women	Bronze	KINTNE
			Women's 400m Hurdles	Silver	TOSTA
			Women's Pole Vault	Silver	STUCZYNSKI Jennifer
			Women's Discus Throw	Gold	BROWN TRAFTON Stephanie
			Men's Heavy (91kg)	Bronze	WILDER Deontay
		ball	Women	Gold	WALSH Kerri, MAY-TREANOR Misty
			Women's Heptathlon	Silver	FOUNTAIN Hyleas
			Women's 4 x 400m Relay	Gold	WINEBERG Mary, FELIX Allyson, HENDERSON Monique, RICHARDS Sanya

Titan Programmable Filters



Titan Data In
(Tables, Trees, Graphs, etc)

R Inside

R Programmable Filter

Titan Data Out
(New / Updated Tables, Trees,Graphs, etc)

User-supplied
R expression

Titan Data In
(Tables, Trees, Graphs, etc)

Java™ Inside

Java Programmable Filter

Titan Data Out
(New / Updated Tables, Trees,Graphs, etc)

User-supplied
Java Code

Titan Data In
(Tables, Trees, Graphs, etc)

XQuery / XSLT Inside

XML Programmable Filter

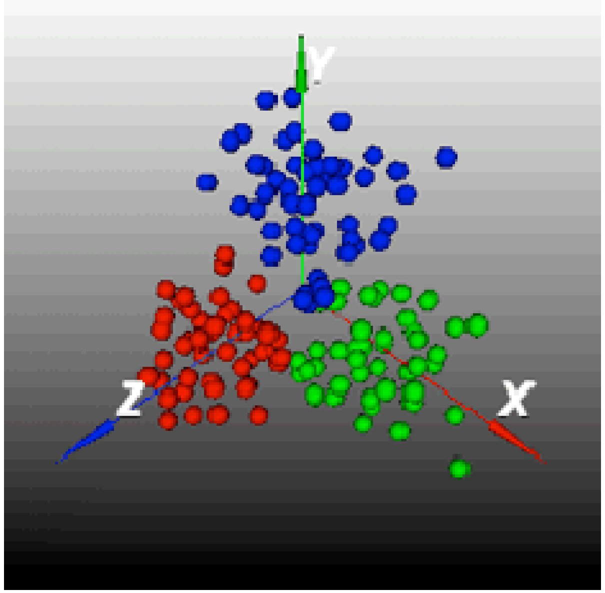
Titan Data Out
(New / Updated Tables, Trees,Graphs, etc)

User-supplied
XQuery / XSLT expression

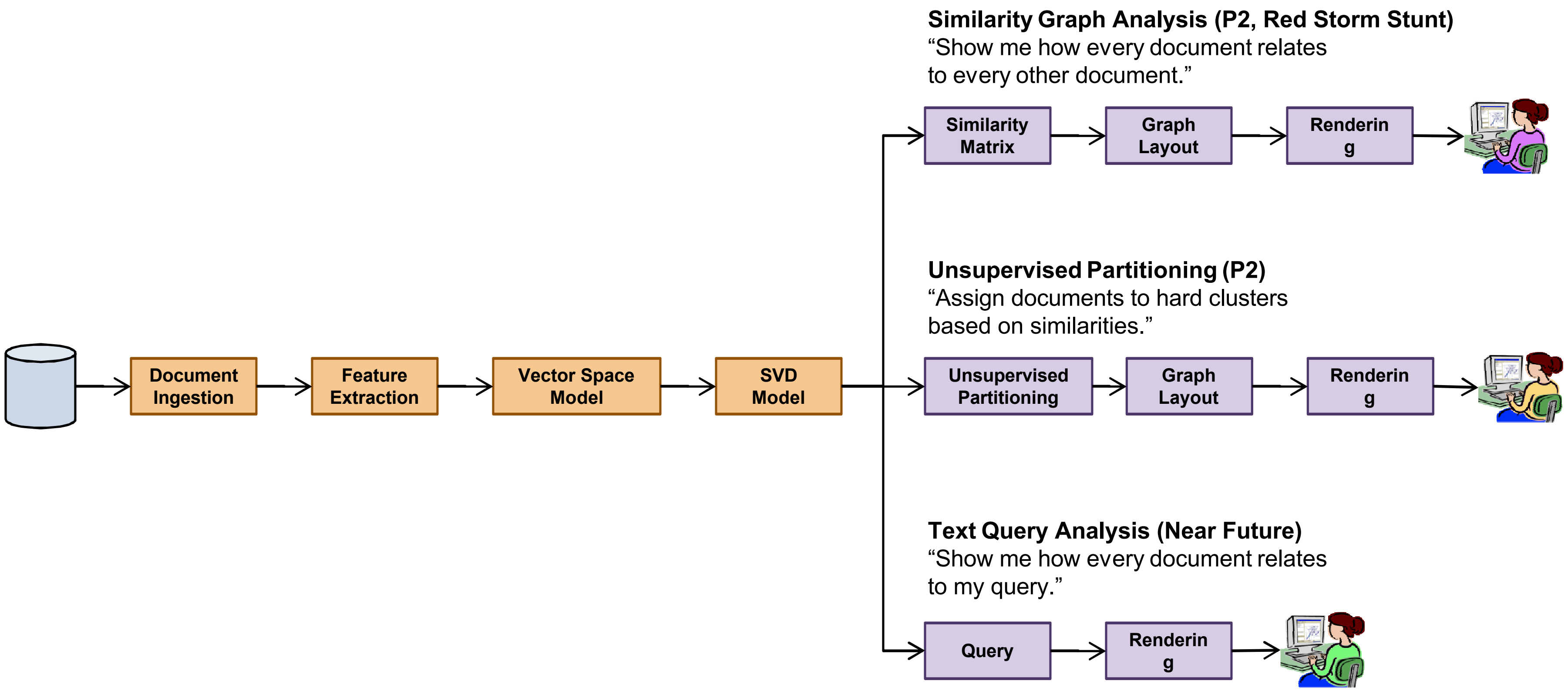
Not shown: Python & Matlab™ programmable filters …

R Programmable Filter Example

```
vtkRcalculatorFilter* rcf = vtkRcalculatorFilter::New();

rcf->SetInput(source->GetOutput());
rcf->PutTable("x");

rcf->SetRscript("m = do.call(cbind,x)\n \
                 cl <- kmeans(m,3)\n \
                 m = cbind(m,cl$cluster)\n \
                 colnames(m)[4] = \"cluster\"\n");

rcf->GetTable("m");

sink->SetInput(rcf->GetOutput());
```



Representative Titan Modeling + Analysis Pipelines



Document Ingestion → Feature Extraction → Vector Space Model → SVD Model

Similarity Graph Analysis (P2, Red Storm Stunt)
"Show me how every document relates
to every other document."

Similarity Matrix → Graph Layout → Renderin g

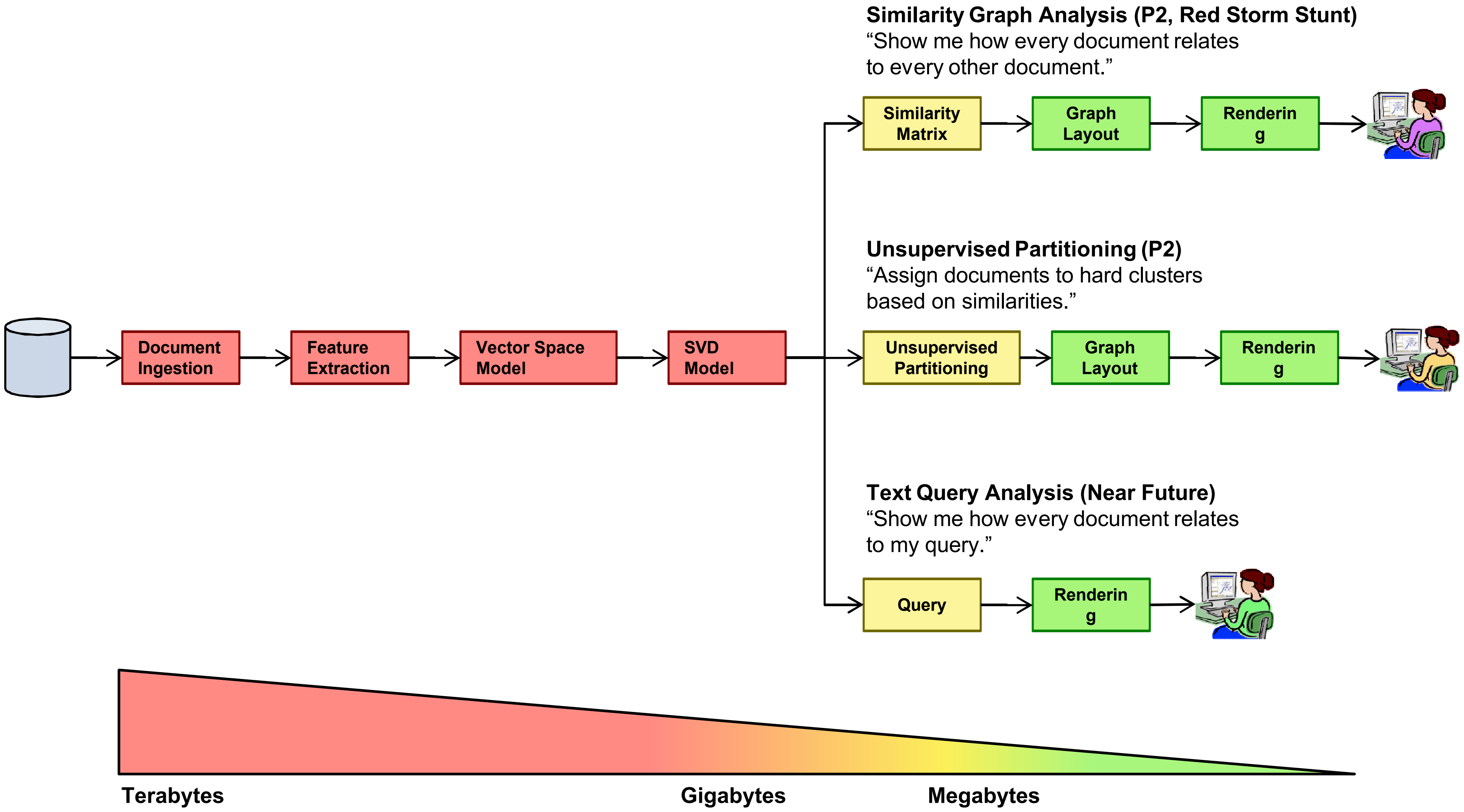
Unsupervised Partitioning (P2)
"Assign documents to hard clusters
based on similarities."

Unsupervised Partitioning → Graph Layout → Renderin g

Text Query Analysis (Near Future)
"Show me how every document relates
to my query."

Query → Renderin g

Representative Titan Modeling + Analysis Pipelines



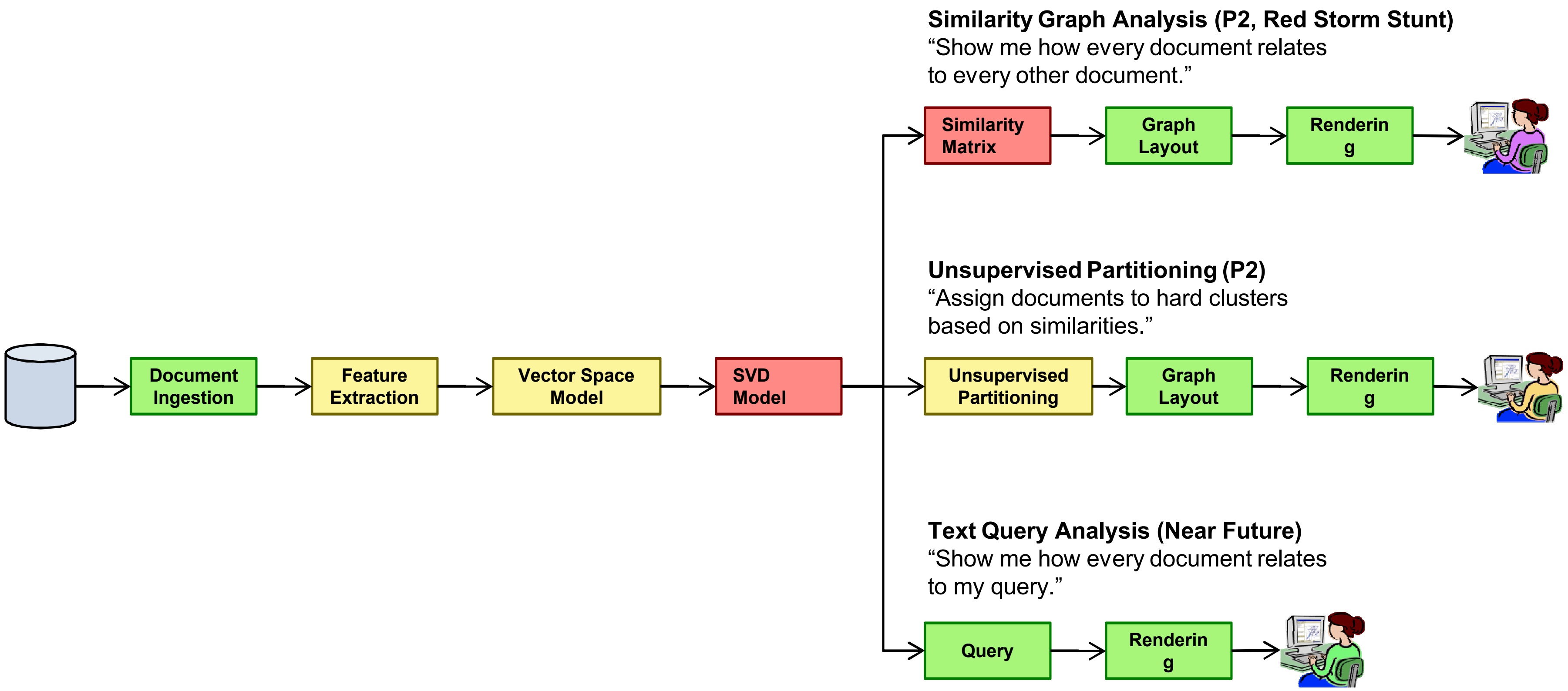
Similarity Graph Analysis (P2, Red Storm Stunt)
"Show me how every document relates to every other document."

Unsupervised Partitioning (P2)
"Assign documents to hard clusters based on similarities."

Text Query Analysis (Near Future)
"Show me how every document relates to my query."

"Data Reduction / Aggregation"

Representative Titan Modeling + Analysis Pipelines



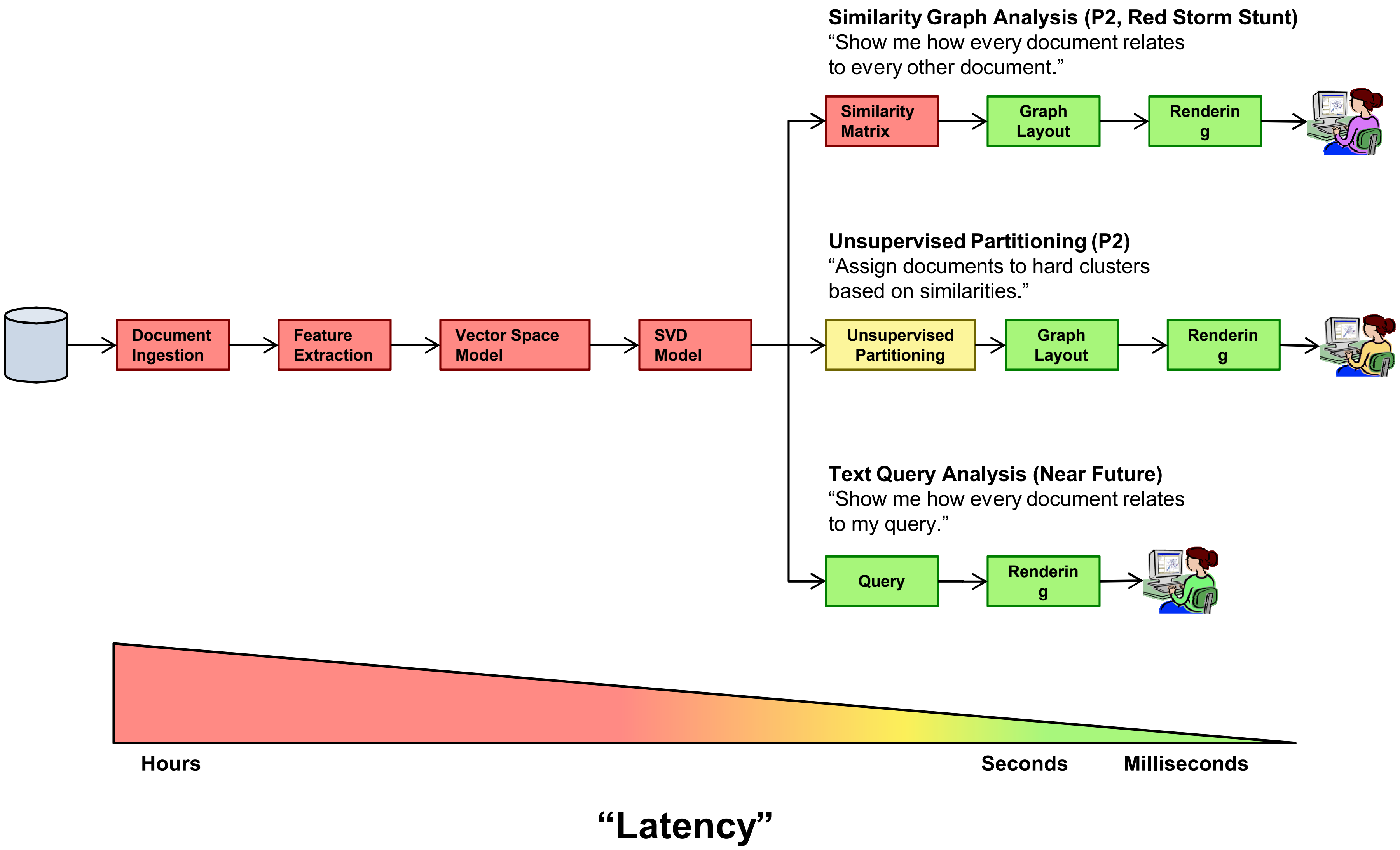
Similarity Graph Analysis (P2, Red Storm Stunt)
"Show me how every document relates
to every other document."

Unsupervised Partitioning (P2)
"Assign documents to hard clusters
based on similarities."

Text Query Analysis (Near Future)
"Show me how every document relates
to my query."

"Computational Complexity"

Representative Titan Modeling + Analysis Pipelines



Similarity Graph Analysis (P2, Red Storm Stunt)
"Show me how every document relates
to every other document."

Similarity Matrix → Graph Layout → Renderin g

Unsupervised Partitioning (P2)
"Assign documents to hard clusters
based on similarities."

Document Ingestion → Feature Extraction → Vector Space Model → SVD Model → Unsupervised Partitioning → Graph Layout → Renderin g

Text Query Analysis (Near Future)
"Show me how every document relates
to my query."

Query → Renderin g

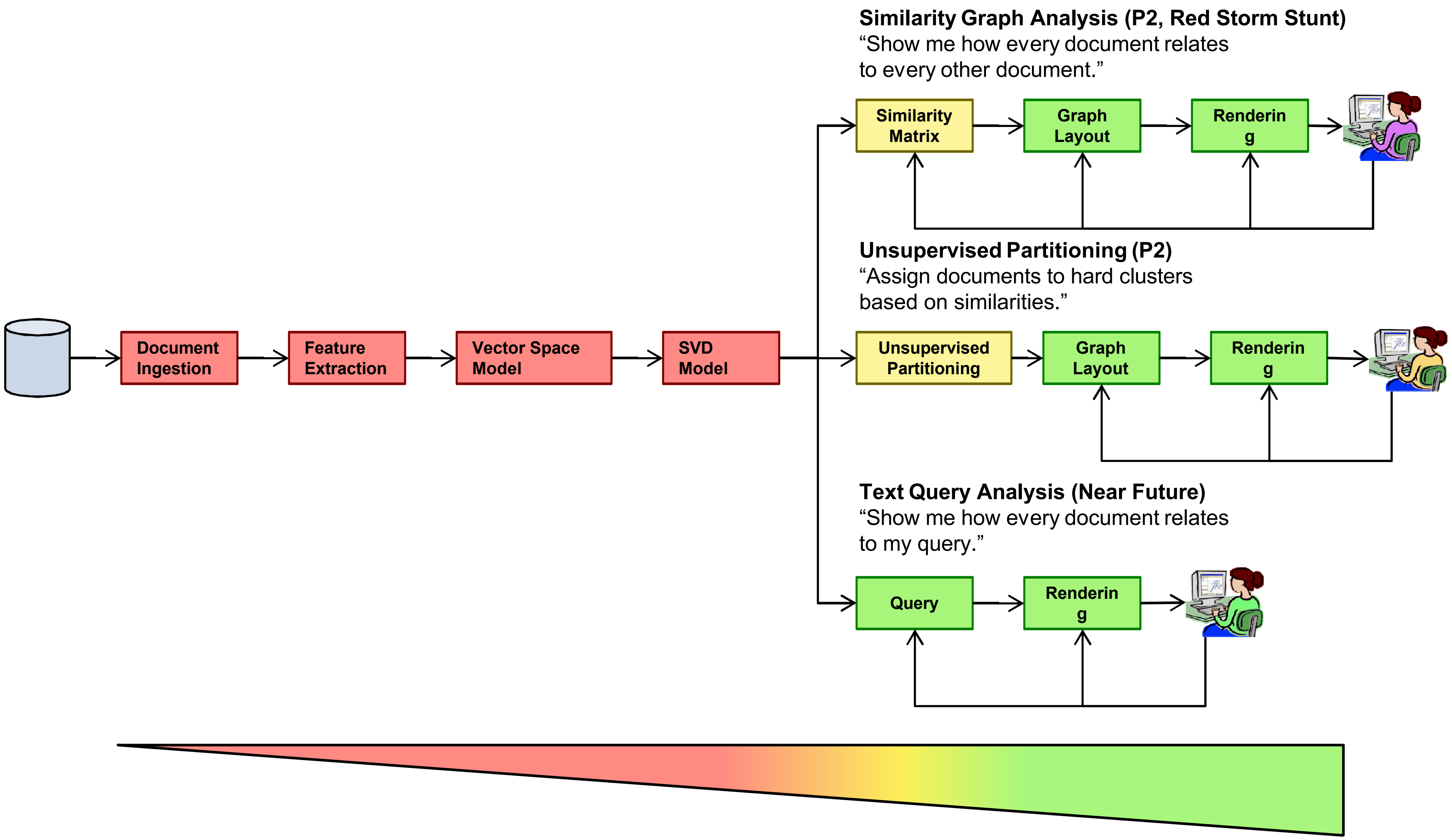
Hours

Seconds

Milliseconds

"Latency"

Representative Titan Modeling + Analysis Pipelines



Document Ingestion → Feature Extraction → Vector Space Model → SVD Model

Similarity Graph Analysis (P2, Red Storm Stunt)
"Show me how every document relates
to every other document."

Similarity Matrix → Graph Layout → Renderin g

Unsupervised Partitioning (P2)
"Assign documents to hard clusters
based on similarities."

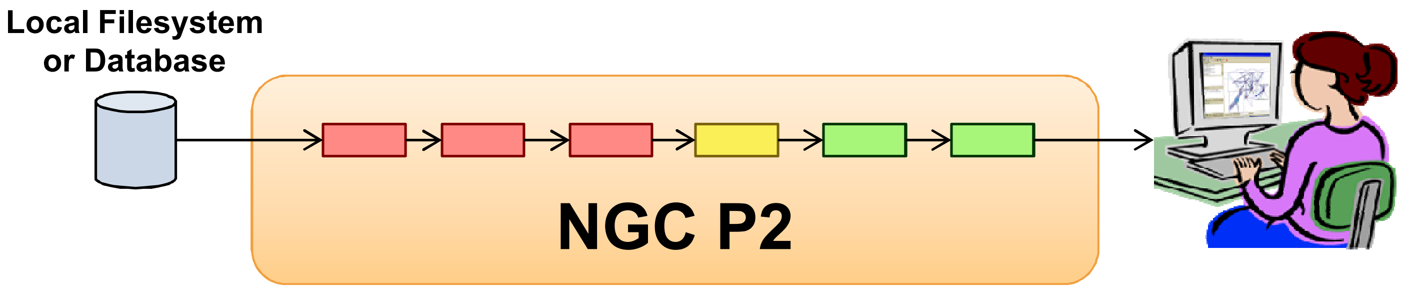
Unsupervised Partitioning → Graph Layout → Renderin g

Text Query Analysis (Near Future)
"Show me how every document relates
to my query."

Query → Renderin g

"Interaction & Feedback"

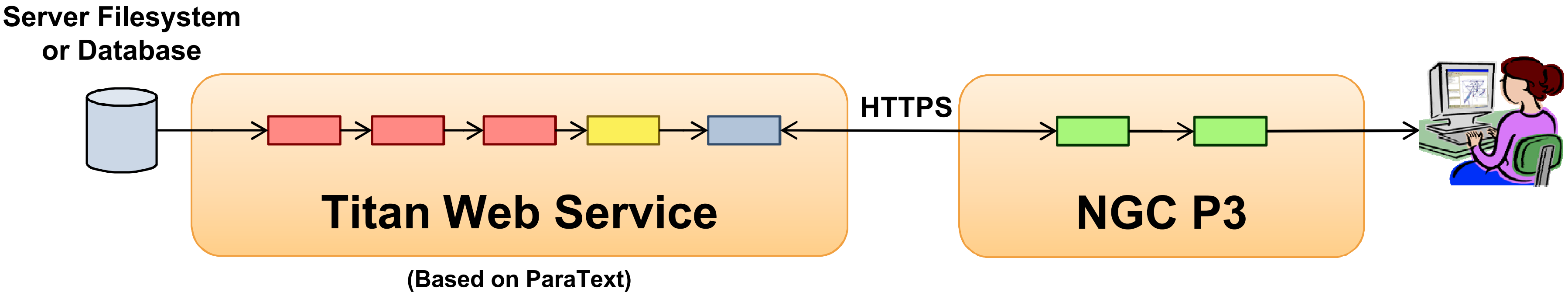
Where We Are Today



Serial Option: put the entire pipeline in one process (P2)

Zero administration, easy to use, online computation, small data.

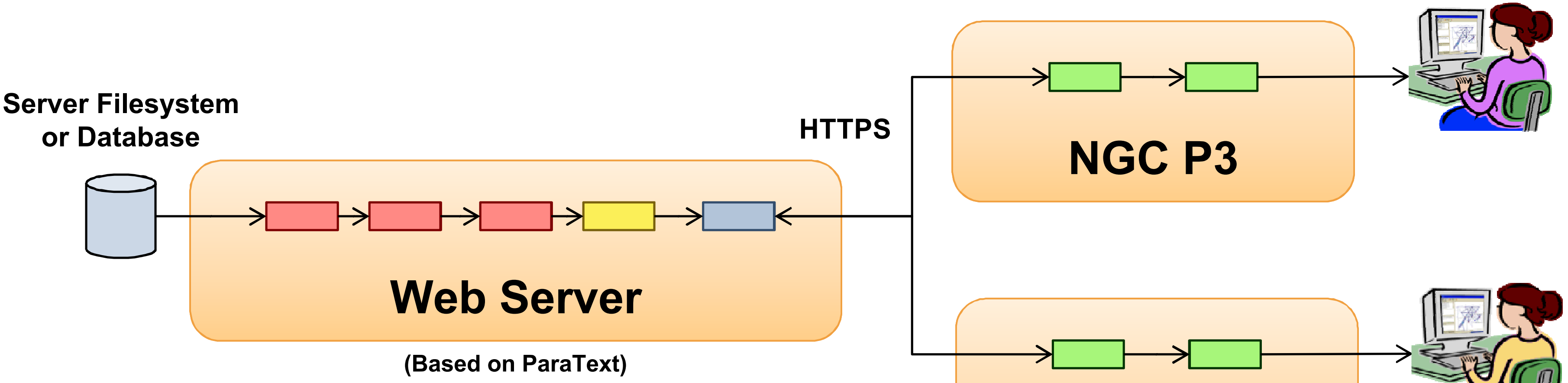
Modest Client / Server Capability for Organizations without HPC



Client / Server Option: model generation and some analysis on server, remaining analysis in client (P3)

Modest administration, slightly more complex, some computation offline, modest data sizes.

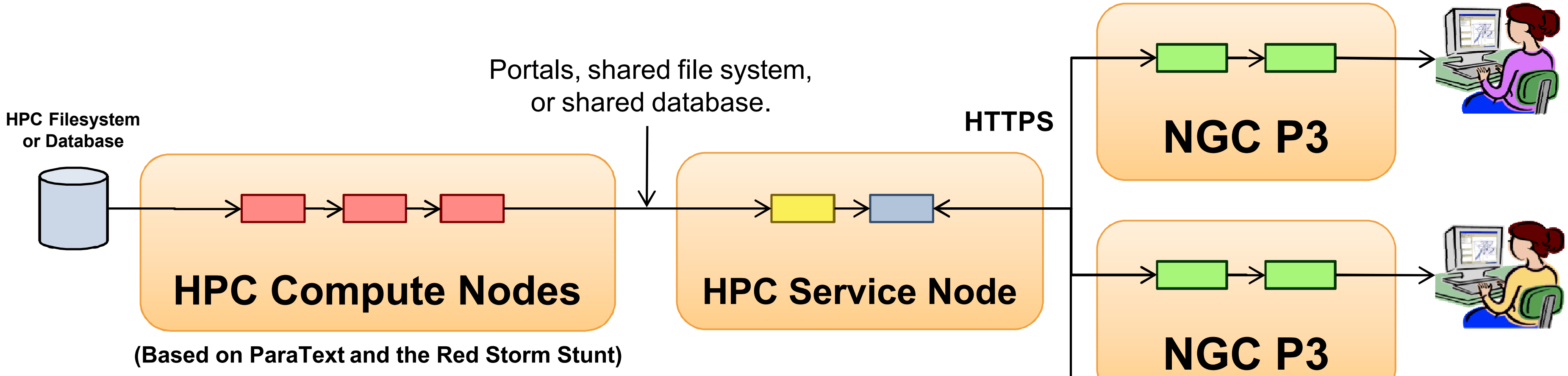
Modest Client / Server Capability for Organizations without HPC



Client / Server Option: model generation and some analysis on server, remaining analysis in client (P3)

Modest administration, slightly more complex, some computation offline, modest data sizes.

How the NGC will deliver “HPC informatics capabilities that are both *usable* and *useful* to analysts."

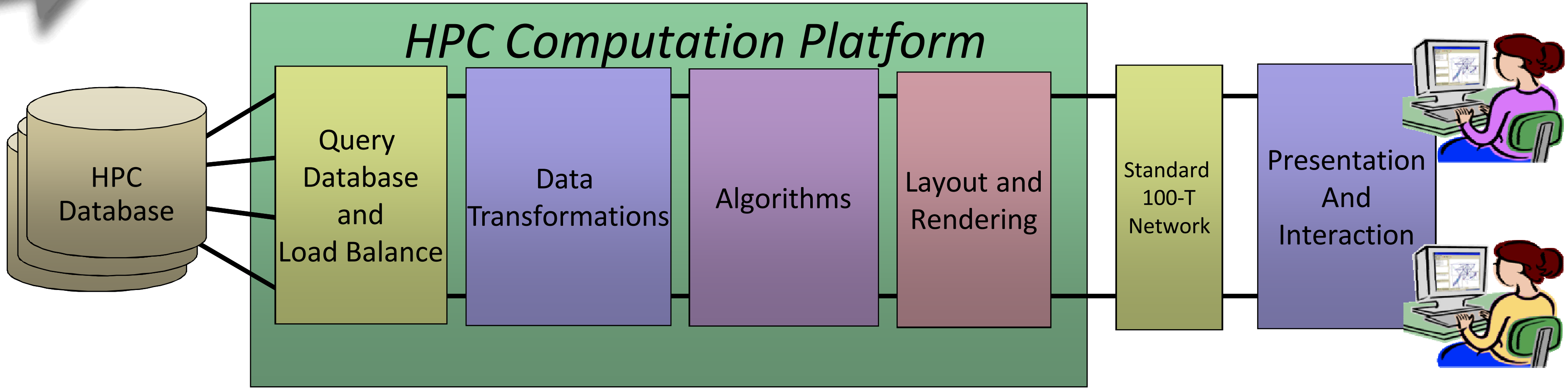

HPC Filesystem or Database
Portals, shared file system, or shared database.
HTTPS
HPC Compute Nodes
HPC Service Node
NGC P3
NGC P3
Client X

(Based on ParaText and the Red Storm Stunt)

HPC Option: model generation on HPC compute nodes, some analysis on a service node, remaining analysis in client (P3)

More administration, more offline computation, largest data sizes.

HPC Informatics for Intelligence Analysis



Query

- Select * from network_packet_table where date > 7/1/2008 and date < 7/8/2008

Data Transformations

- ***vtkTableToTree***
- ***vtkTableToGraph***
- ***vtkTableToSparseArray***
- ***vtkTableToDenseArray***

Algorithms

- ***Statistics***
- ***Linear Algebra***
- ***Tensor Methods***
- ***Graph Algorithms***
- ***Matlab and "R"***
- ***MapReduce***

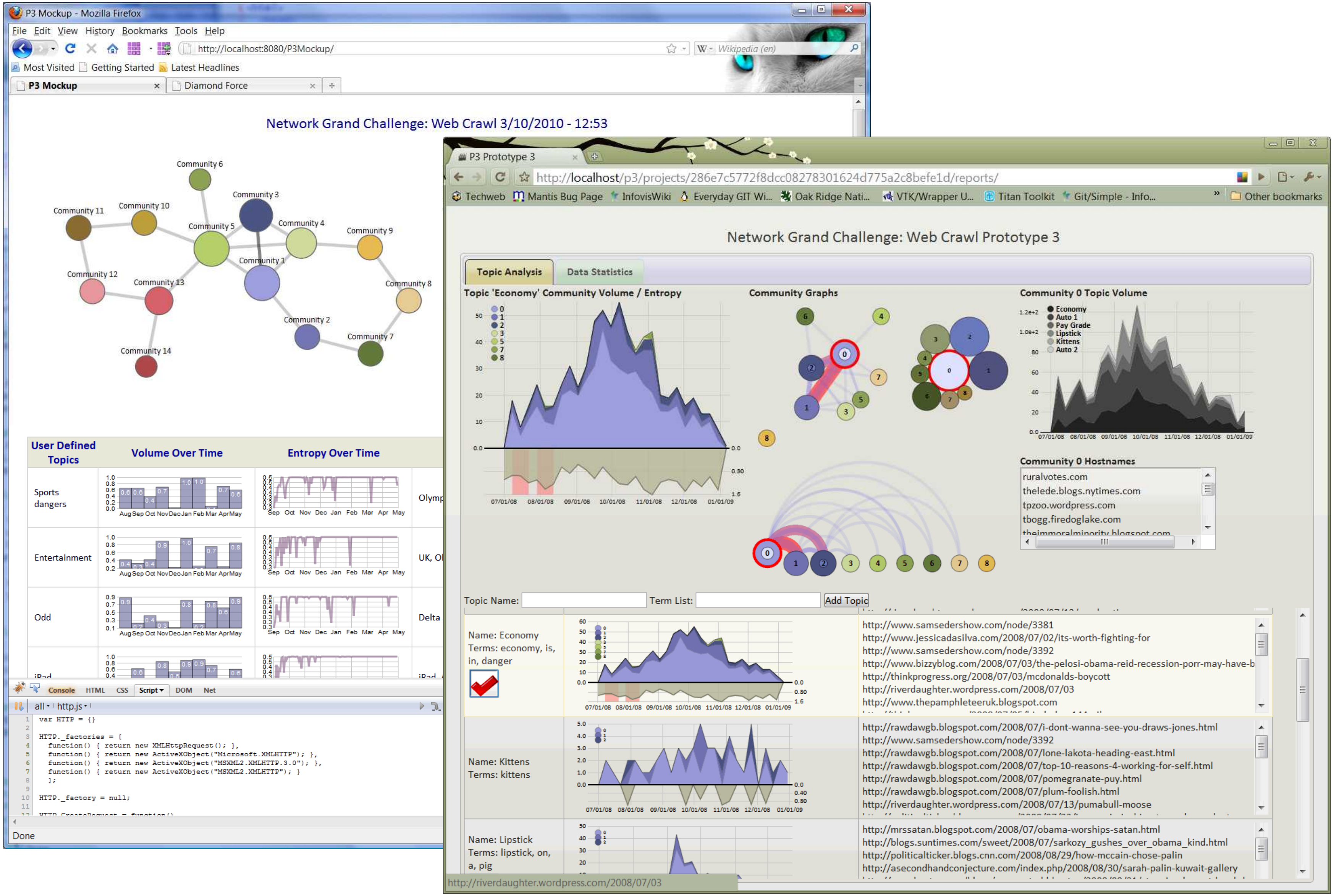
Layout and Rendering

- ***Tree Layout***
- ***Tree Map***
- ***Graph Layout***
- ***Hierarchical***
- ***Geodesic***

Presentation and Interaction

- ***Client/Server***
- ***Geometry and Image Delivery***
- ***Windows/Unix/MAC cross platform UI***
- ***Linked Selection***

Web Interfaces



P3 Mockup - Mozilla Firefox

http://localhost:8080/P3Mockup/

Network Grand Challenge: Web Crawl 3/10/2010 - 12:53

User Defined Topics	Volume Over Time	Entropy Over Time	
Sports dangers			Olymp
Entertainment			UK, Ol
Odd			Delta
iPad			iPad

Aug Sep Oct Nov Dec Jan Feb Mar Apr May
Sep Oct Nov Dec Jan Feb Mar Apr May
Console
HTML
CSS
Script
DOM
Net
all
http.js

```
var HTTP = {}

HTTP._factories = [
  function() { return new XMLHttpRequest(); },
  function() { return new ActiveXObject("Microsoft.XMLHTTP"); },
  function() { return new ActiveXObject("MSXML2.XMLHTTP.3.0"); },
  function() { return new ActiveXObject("MSXML2.XMLHTTP"); }
  ];

HTTP._factory = null;
```

Done

P3 Prototype 3

http://localhost/p3/projects/286e7c5772f8dcc08278301624d775a2c8befe1d/reports/

Network Grand Challenge: Web Crawl Prototype 3

Community 0 Hostnames

ruralvotes.com
thelede.blogs.nytimes.com
tpzoo.wordpress.com
tbogg.firedoglake.com
theimmoralminority.blogspot.com

Topic Name: Term List: Add Topic

Name: Economy
Terms: economy, is, in, danger

http://www.samsedershow.com/node/3381
http://www.jessicadasilva.com/2008/07/02/its-worth-fighting-for
http://www.samsedershow.com/node/3392
http://www.bizzyblog.com/2008/07/03/the-pelosi-obama-reid-recession-porr-may-have-b
http://thinkprogress.org/2008/07/03/mcdonalds-boycott
http://riverdaughter.wordpress.com/2008/07/03
http://www.thepamphleteeruk.blogspot.com

Name: Kittens
Terms: kittens

http://rawdawgb.blogspot.com/2008/07/i-dont-wanna-see-you-draws-jones.html
http://www.samsedershow.com/node/3392
http://rawdawgb.blogspot.com/2008/07/lone-lakota-heading-east.html
http://rawdawgb.blogspot.com/2008/07/top-10-reasons-4-working-for-self.html
http://rawdawgb.blogspot.com/2008/07/pomegranate-puy.html
http://rawdawgb.blogspot.com/2008/07/plum-foolish.html
http://riverdaughter.wordpress.com/2008/07/13/pumabull-moose

Name: Lipstick
Terms: lipstick, on, a, pig

http://mrssatan.blogspot.com/2008/07/obama-worships-satan.html
http://blogs.suntimes.com/sweet/2008/07/sarkozy_gushes_over_obama_kind.html
http://politicalticker.blogs.cnn.com/2008/08/29/how-mccain-chose-palin
http://asecondhandconjecture.com/index.php/2008/08/30/sarah-palin-kuwait-gallery

http://riverdaughter.wordpress.com/2008/07/03