

XDATA Overview



What is DARPA?

- The Defense Advanced Research Projects Agency (DARPA) was established in 1958 to
 - prevent strategic surprise from negatively impacting U.S. national security and
 - create strategic surprise for U.S. adversaries by maintaining the technological superiority of the U.S. military
- DARPA undertakes projects that are finite in duration but that create lasting revolutionary change.
 - Short duration
 - High impact
- Time limited Program Managers

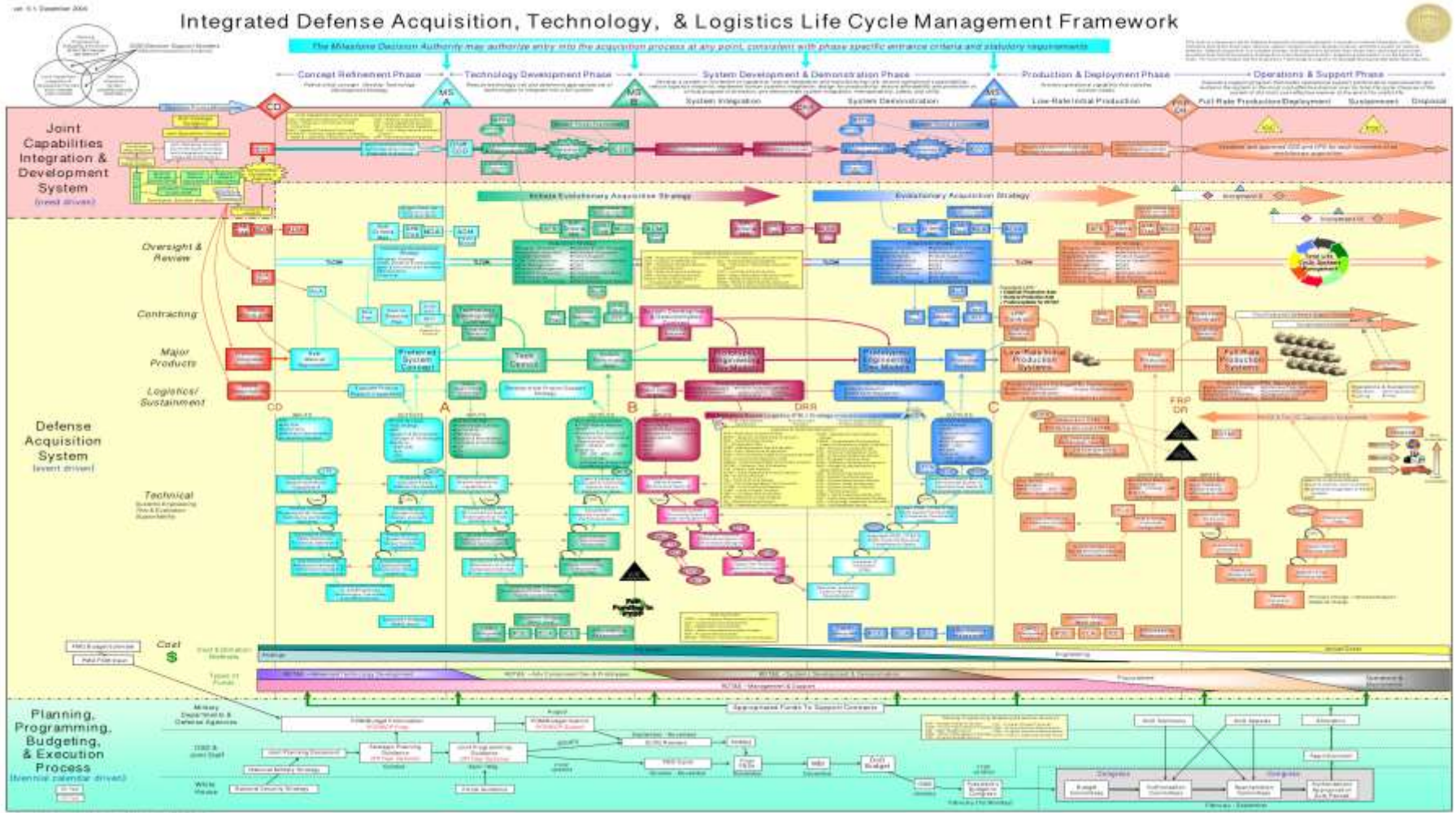
Outline

- DARPA Tech Transfer
 - Government Acquisition
- The XDATA Stack
- XDATA Transition
- DARPA Support -> Community Support

How do DARPA programs transition?

- During a DARPA program
 - Technology development
 - Prototype developed with transition partner
 - Technology demonstration
 - Declare success
- After DARPA program
 - Services start acquisition “Program of record”
 - Then...

DoD Acquisition



This can take between 5-20 years to deliver

Open Source Alternative

- Open source software
 - “Try before you buy”
 - Prototyping of complex systems from DARPA components
- XDATA Transition Model
 - DARPA resources applied, developers embedded at receiving site
 - Tools identified, gaining org resources, DARPA provides limited expertise
 - Org contracts with XDATA Performer under own contract to act as conduit for tool integration
 - Tools identified, no DARPA funding or expertise applied
- As a data science research effort, we don't build systems/applications, rather modules that enable systems/applications

Outline

- DARPA Tech Transfer
 - Government Acquisition
- **The XDATA Stack**
- XDATA Transition
- DARPA Support -> Community Support

The XDATA Program

- XDATA: 26 developers of big-data components
 - Languages: Julia, SciDB, SciPy, Numba, Delite
 - Platform: Spark, BayesDB, Blaze, Gunrock
 - Math/ML: CVX, igraph, skylark, elemental, SNAP
 - Analytics/NLP: MITIE, topic, GraphQuBE
 - Viz: Tangelo, D3, Vega, Bokeh, Ozone, Aperture
- Transition investment up front: \$8m/year investment to build customer-facing systems
- All open source, non-GPL (i.e. liberally licensed)

Disk / RAM / GPU

Disk / RAM / GPU

RAM / Cache / GPU

- Sources
- Transforms
- Batch (size)
- Stream (rate)
- Radar
- Tracklets
- Speech
- Speech to text
- Video
- Content extraction
- Text
- Entity Extraction
- Topic Modelling
- Spreadsheets
- Metadata
- PCAP
- Pcap unpacking
- IP Networks
- Geocoding
- Bio
- Etc...
- Other basic transforms

Data Ingest API

Tikka, OODT-Wings Workflow (MDA/JPL), Zephyr (Sotera), OptiWrangler (Stanford O), Other ETL

Platform

HDFS / Hadoop
Map / Reduce
Giraph *Purdue*
Oozie *USC*
Hama *Sotera*
RHIFE *MITLL*
Hive *Data*
GoFS *Tactics*
USC

Accumulo
BlinkDB *UCB*
Storm
Spark Streaming
Mesos
Tachyon
Spark
Shark *UCB*
Bagel

MPI *Stanford*
Elemental *GA Tech*
BLAS *IBM*
OpenMPI *Continuum*

Single Machine
R *JHU*
Matlab *Stanford*
Stata *Phronesis*
Riposte *CMU*
SNAP *Purdue*
Stanford O

HPC TBD

GPU *Royal*
gpuGraph *Caliper*
UC Davis
SYSTAP


Abstractions

Java
HiveQL
Python
Anaconda
Blaze
Numba
Continuum
Clojure
SNAP
Delite
OptiCVX
OptiGraph
OptiML
Stanford O
g++
CUDA
Bigdata DB

Data Analytics API

Lincoln Labs API, HiveQL API, OODT-Wings Workflow (MDA/JPL)

Analytics



Graph / Matrix
IBM Skylark(RNLA)
Sotera Louvain
USC BSP Optimization
GA Tech NNMF
JHU VNOM
Stanford O SNAP

Time Series
Sotera ARIMA
GA Tech Time Series
SSCI/MIT/UL BayesDB

Machine Learning
SSCI/MIT/UL BayesDB
MITLL/BBN NLP
Boeing/Upitt SMILE
MITLL TopicID
DT/UMD/MSR
Classification
UCB BlinkDB

Distance Measures
TBD e.g. Cosine Sim,
Page rank
GA Tech / Stanford B /
Continuum
Convex Optimization
Sotera
Correlation
Approximation

Ad-Hoc Retrieval API

Mongo, Impala, Shark (UCB), OODT-Wings Workflow (MDA/JPL)

Visualization

HTML5
d3.js
Vega
Stanford / Kitware

CherryPy
Tangelo
Kitware

Aperture *Oculus*

OWF / Neon *Next Century*

Bokeh *Continuum*

3d
Virtual Environments
Immersive *ICT*

New School
Defaults and design
recommendations
Boeing / U Pitt
Genie
Continuum Stencil
DSL

Logging API

Attention Tracking / Eye Tracking / User Interface Action Tracking via keystroke logging and mouse clicking (Draper)

Interaction

Geo/Time-Like
Oculus / Kitware

SQL-Like
Oculus / Kitware

Graph-Like
Oculus / Kitware

Interactive Coding (Wakari, Julia)
MITLL / Continuum

Physical Devices (Multi-touch)
ICT

Interactive Querying

Search / Query / Subselect / Summarize / Link-Brush / Pivot / Zoom / Export / Integrate / Intersect

Queries

“Find me everything on this number”

“Find me a pattern of behavior on this entity”

“Where will this person be tomorrow”

“What groups does this group interact with”

Quality of Service
Latency from click
RAID and redundancy; #GPUs, RAM, SSD, CPU
Adjustable tradeoffs: Volume, Time, Accuracy / Uncertainty

Workflow Management / System Architecture
MDA/JPL, UCB, Sotera, Draper, Data-Tactics

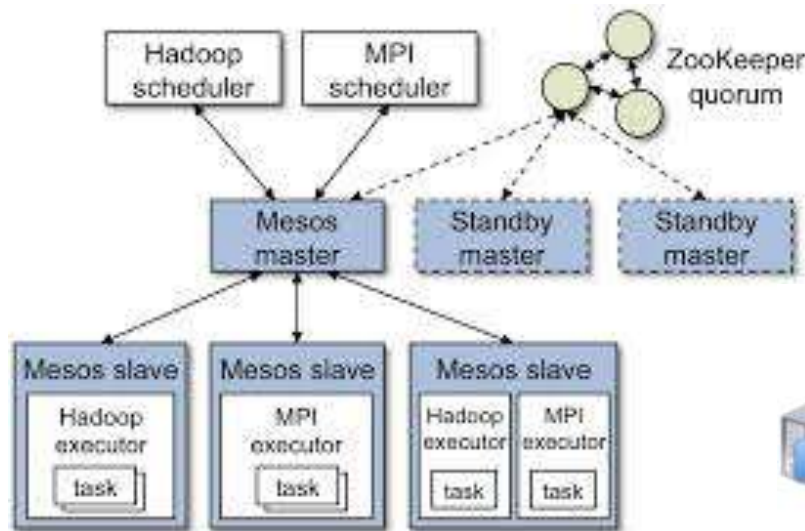
Resource Management/Planning
Dependencies; Scheduler
Performance and load estimates
User Interface with workflow planning



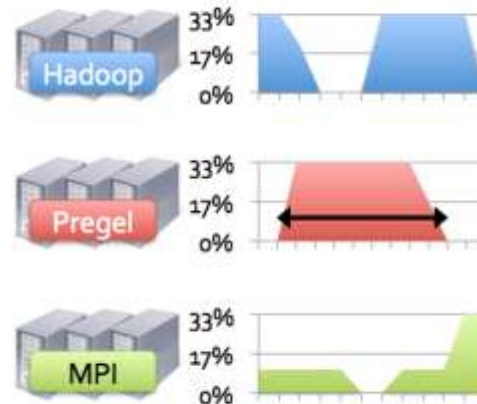
Mesos



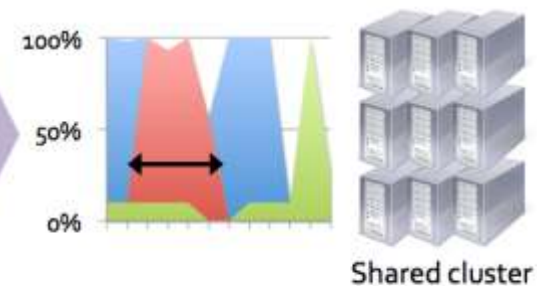
- Cloud scheduling via linux containers
- Plays nice with Hadoop/Spark/MPI clusters



Today: static partitioning



Mesos: dynamic sharing

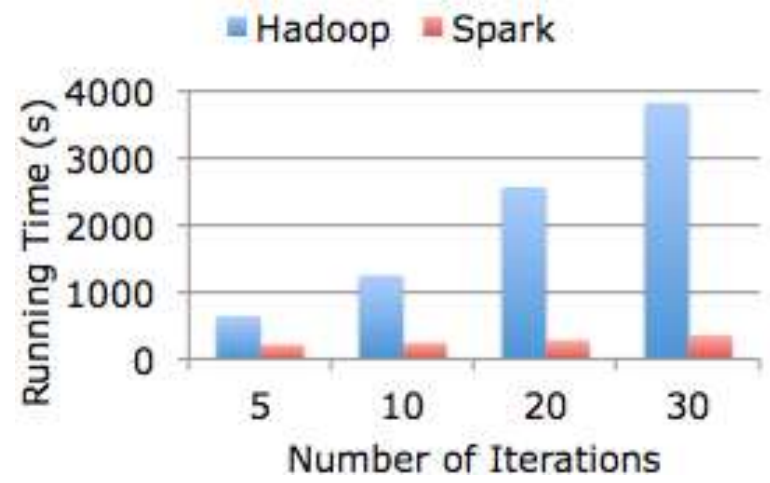
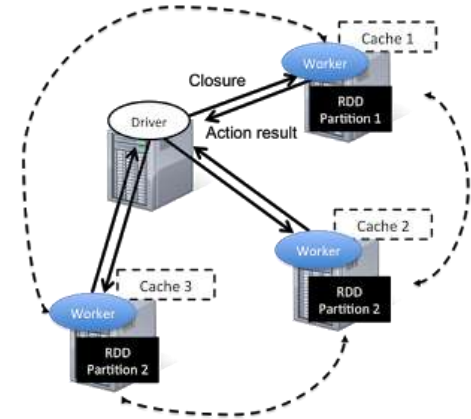
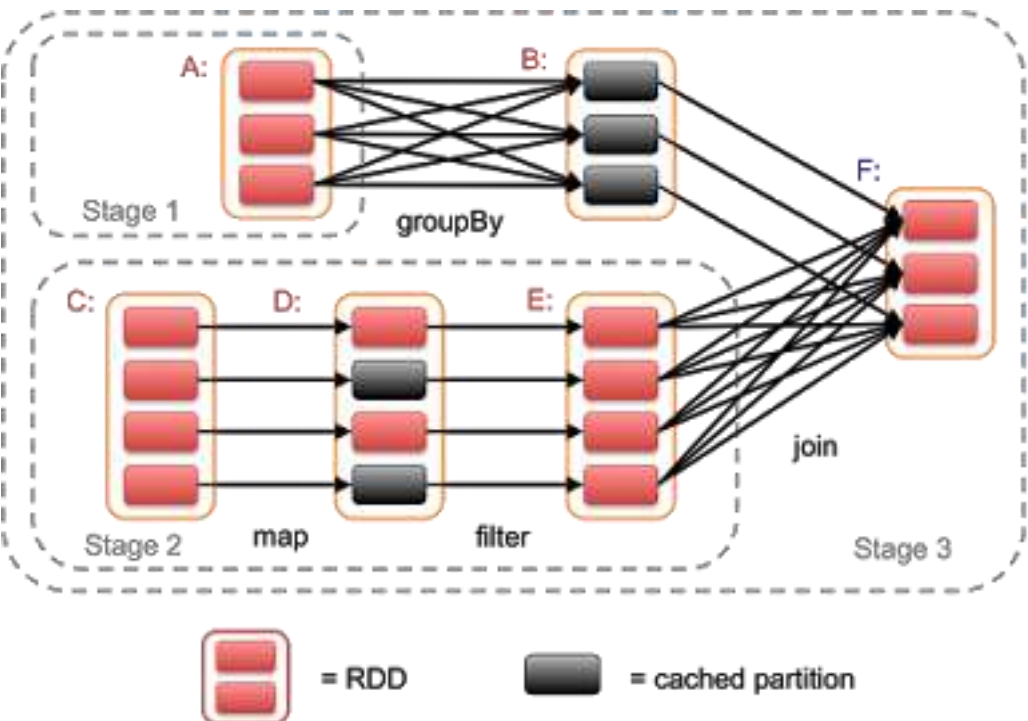




Spark



- In-memory distributed database/computation
- Interfaces to HIVE/HDFS/Mesos
- Apache top-level





Anaconda/Numba

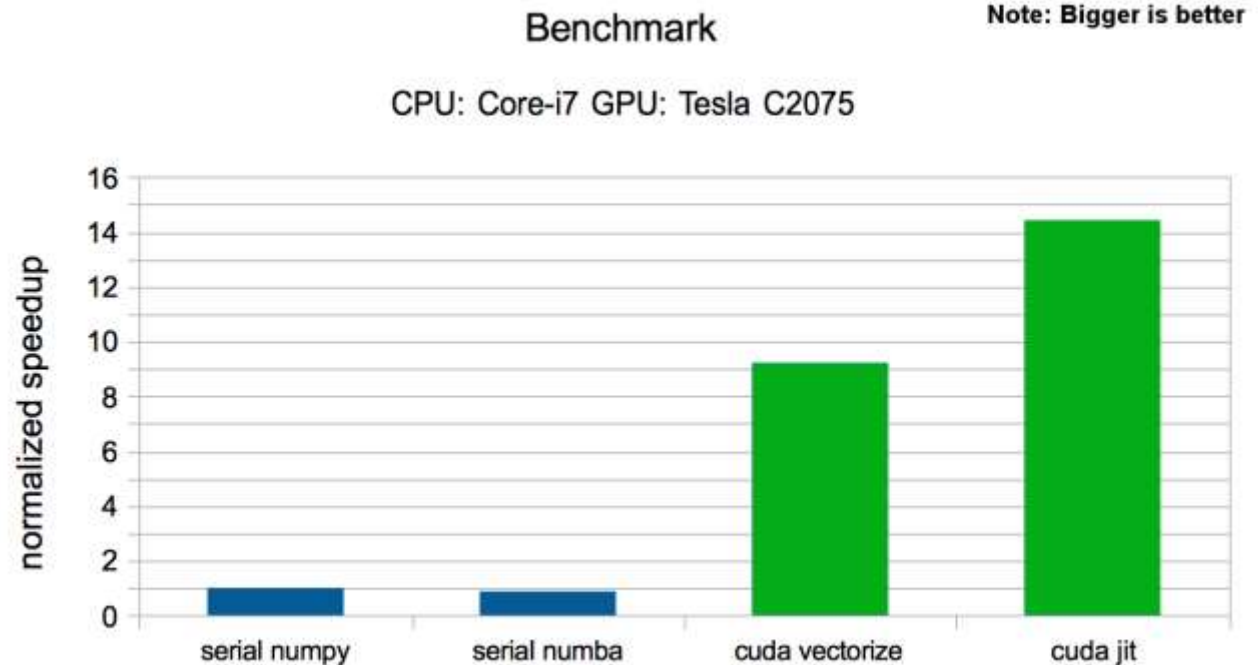


1. Python distribution for large-scale data processing, predictive analytics, and scientific computing
2. Python compiled to CPU/GPUs with autovectorization

```
from numba import jit
from numpy import arange

# jit decorator tells Numba to compile
# The argument types will be inferred
@jit
def sum2d(arr):
    M, N = arr.shape
    result = 0.0
    for i in range(M):
        for j in range(N):
            result += arr[i,j]
    return result

a = arange(9).reshape(3,3)
print(sum2d(a))
```





Julia



- High-level, high-performance dynamic programming language for technical computing, with syntax that is familiar to users
 - Python/java call interfaces
 - Simple distributed and concurrent programming

Simple Code

```
function mandel(z)
  c = z
  maxiter = 80
  for n = 1:maxiter
    if abs(z) > 2
      return n-1
    end
    z = z^2 + c
  end
  return maxiter
end
```

Parallelization

```
nheads = @parallel (+) for i=1:1000
  int(randbool())
end
```

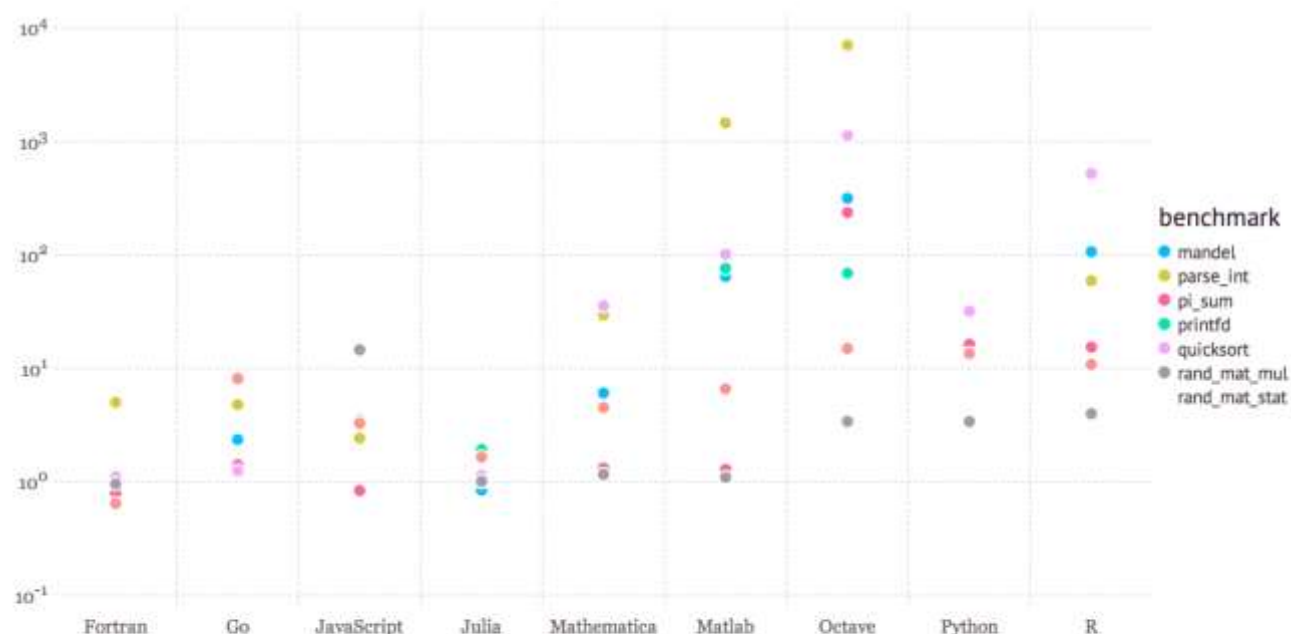
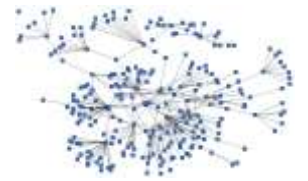


Figure: benchmark times relative to C (smaller is better, C performance = 1.0).

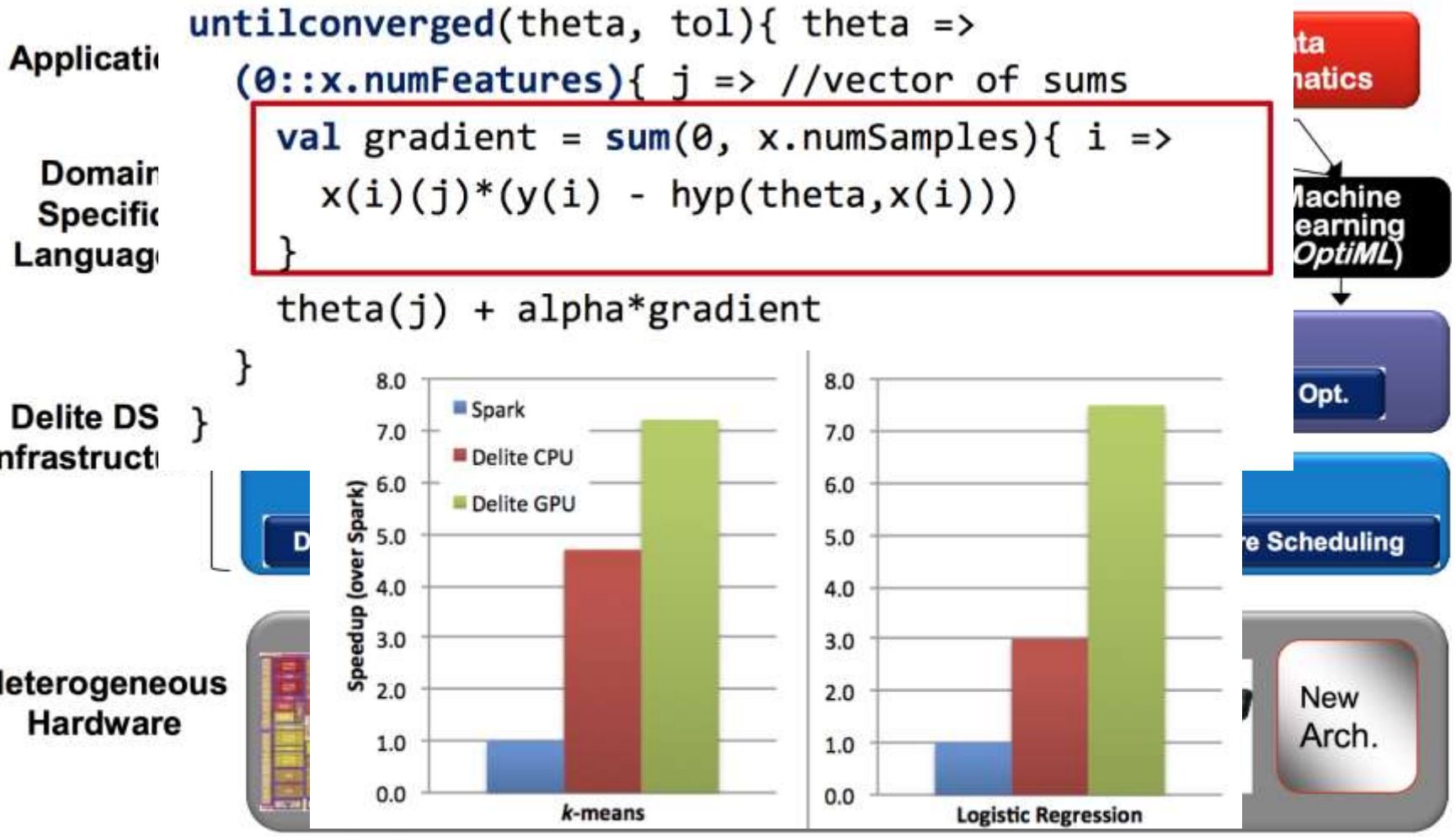


DELITE



OptiGraph, OptiCVX, OptiML

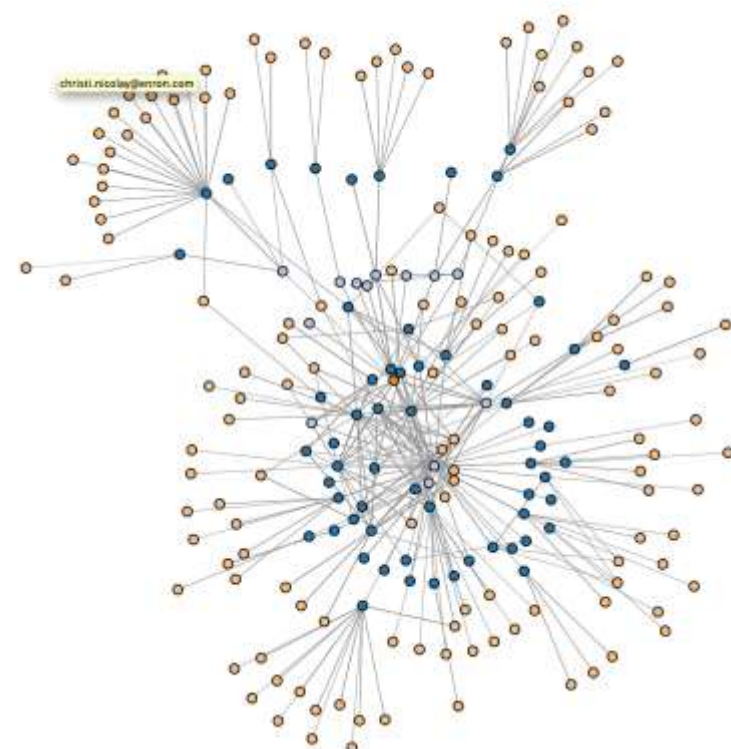
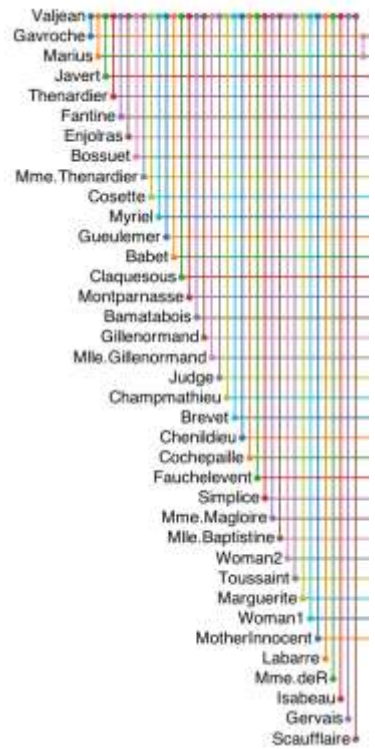
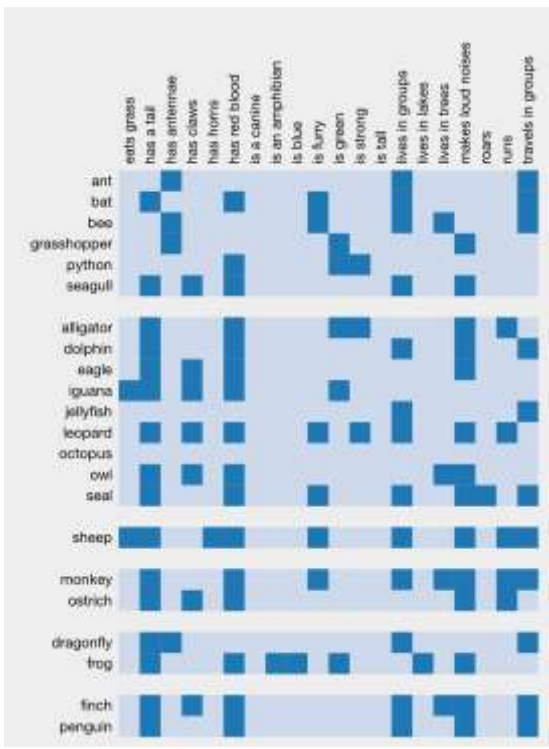
- Compiler framework for parallel domain specific languages





Tangelo

- Python/web-based visualization of big data
- Examples
 - <http://xdata.kitware.com/examples/>

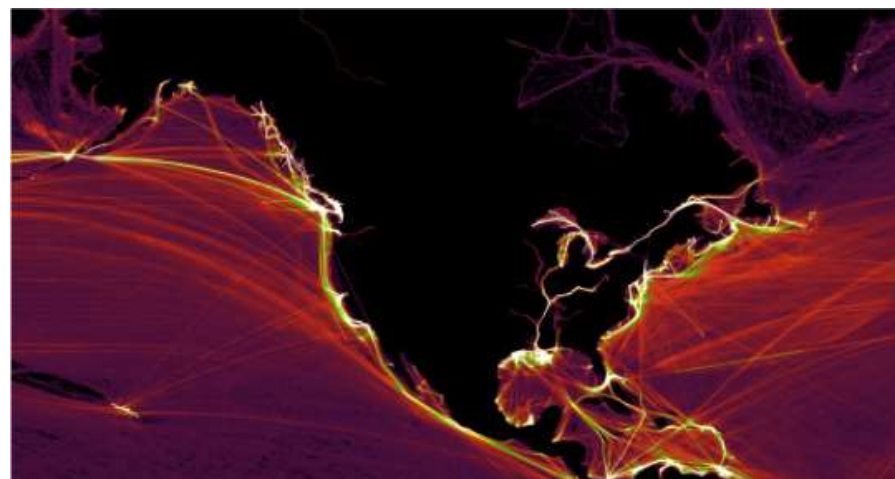
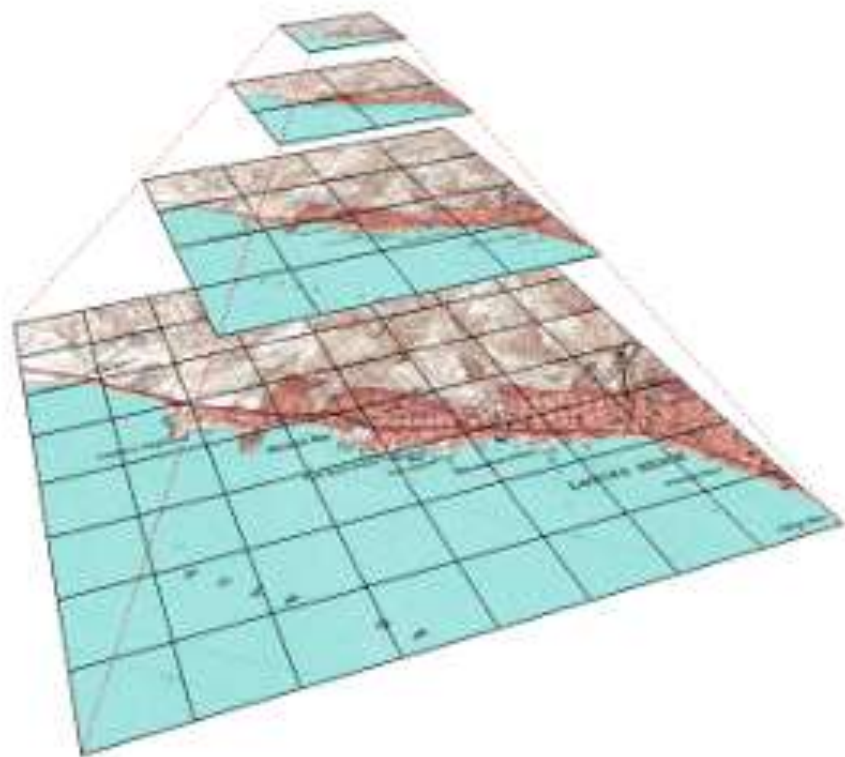




Aperture and Aperture Tiles



- Scalable visualization of huge datasets from hadoop/spark/impala



Outline

- DARPA Tech Transfer
 - Government Acquisition
- The XDATA Stack
- **XDATA Transition**
- DARPA Support -> Community Support



XDATA Challenge Datasets

TWITTER



Tweets: 292.7M+
 # Unique Users: 7.6M
 Total Size: 232 GB

Tweets: 1B+
 # Unique Users: 94M+
 # Geolocated: 31M+
 Total Size: 146 GB

GISR

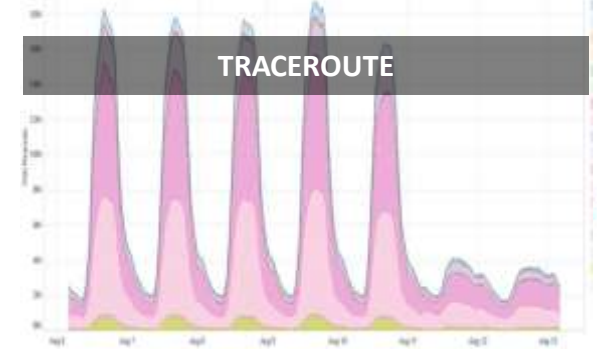


VADER: 5 GB
 SR Hawk: 60 MB
 MAGIC: 1.2 GB
 LYNX II: 11 GB

HYDRA: 44 GB
 GMBIT: 2.2 TB
 GEOnet: 3.2 TB

Categorized Internet Usage Across the U.S.

TRACEROUTE



Hourly Summaries: 630 Million+
 # Observed Hits: 734 Trillion+
 # Bytes Transacted: 31 Exabytes+

KIVA MICROLOANS



Loans: 524,514
 # Lenders: 1.1M
 # Partners: 238

Transactions: 4,069,217
 # Total \$ Loaned: \$418M (USD)
 # Journal Entries: 307,831

BITCOIN



Transactions: 15.8M+
 # Edges: 37.4M+

Senders: 5.4M+
 # Receivers: 6.3M+
 # Bitcoins Xted: 1.4M+

Text/Unstructured Data



Game, Player, Team statistics
 # 6 Independent text types for each game

#2.1 Million unique employment ads in Spanish from South America



XDATA – Summer Workshop 2014

- **Summer 2014 – 8 Weeks**

- 30+ teams; average of 45 performers/day onsite
- **Phase 1** – Finalized documentation and publication of 75 open source software components and 135 academic papers on the DARPA Open Catalog
- **Phase 2** – User testing, benchmarking, and new development



- **In 2014, 4 new datasets added**

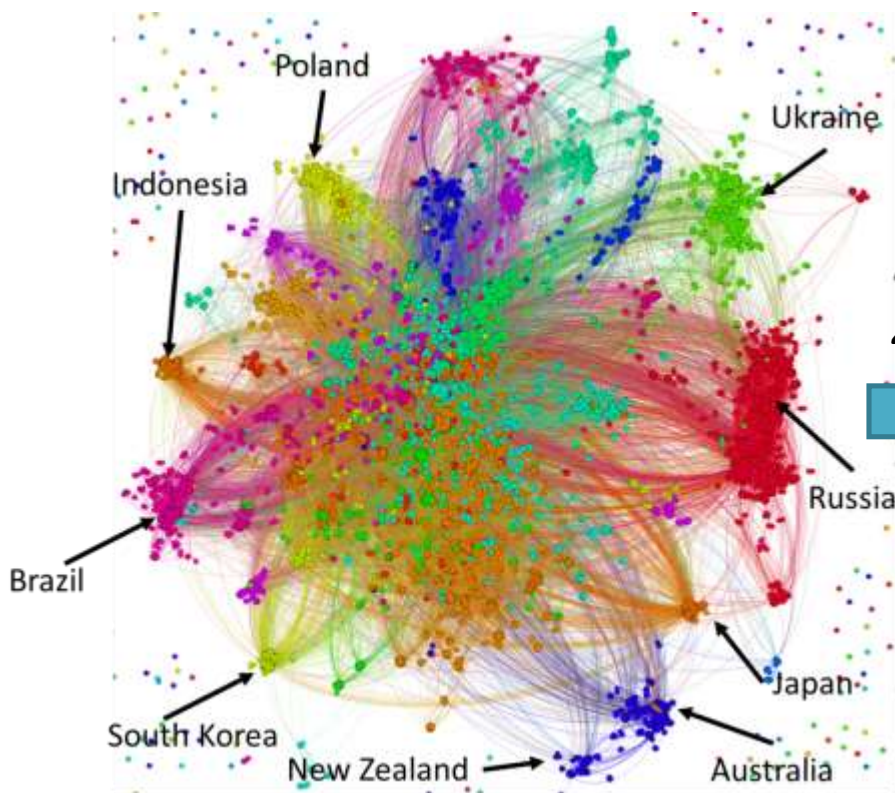
- Cyber (Web Data Commons Hyperlinks - *3.5 billion pages, 128 billion edges*)
- Cyber (Distributed Net Scans - *6 TB*)
- Unstructured Text (Foreign language employment advertisements – *40 GB with 2.1 million unique jobs*)
- Structured, unstructured (NBA statistical, news reports and crowd-sourced - *6 independent text types for each game*)





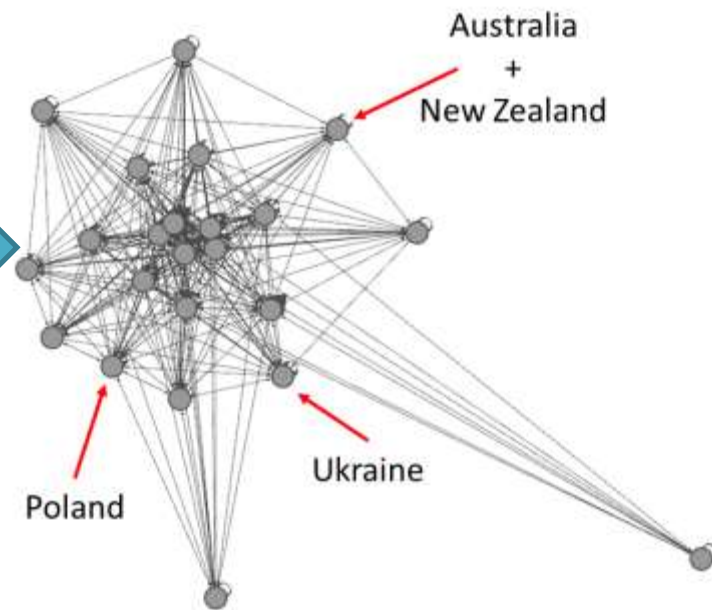
XDATA Application to **Cyber**: Reveal Organizations

- **Automated community detection based on structural characteristics**
 - Processing 7B+ traceroute-hops map the Internet as traveled
 - Detects hierarchical groups of IP addresses with high density connections



$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

Zoom Out!



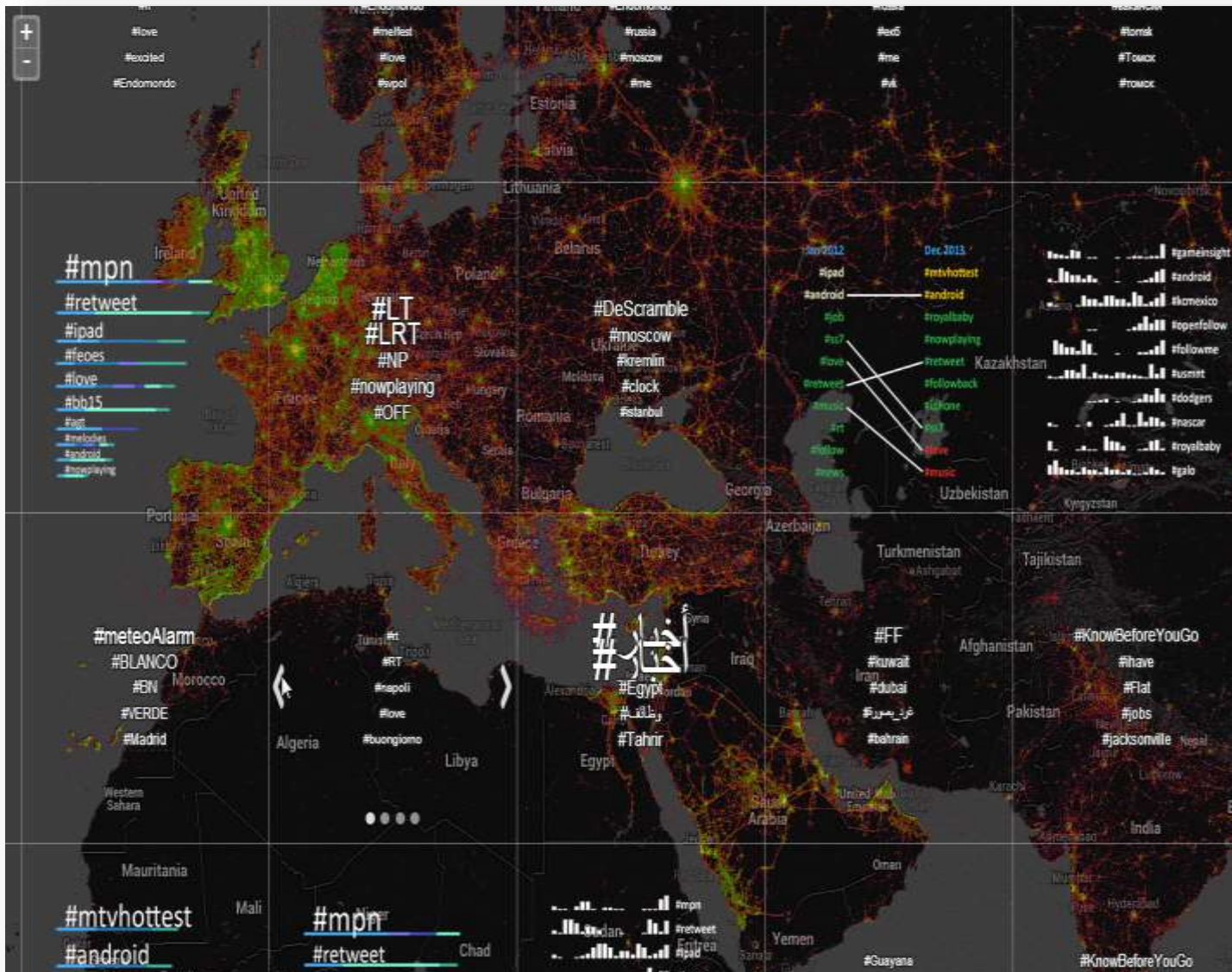
Each dot represents a group of IP addresses; each color shows a detected community, often aligned with political boundaries

Ability to summarize data at an organizational level and provide multiple levels of abstraction / characterization



XDATA Application to Social Media: Interactive Tile Apps

Multi-level interactive geo-tiles with embedded analytics



Tile features:

- Word clouds
- Time series
- Rank changes
- Histograms
- Heat maps
- ...

Interactive using 1B+ Tweets



Outline

- DARPA Tech Transfer
 - Government Acquisition
- The XDATA Stack
- XDATA Transition
- **DARPA Support -> Community Support**

XDATA Transition + Open Catalog

- <http://www.darpa.mil/opencatalog/>

Software	Publications	Data	Search			
<input type="text" value="Search"/> <input type="button" value="Clear"/>						
Team	Project	Description	Instructional Material	Category	Code	
Aptima, Inc.	Network Query by Example	Hadoop MapReduce-over-Hive based implementation of network query by example utilizing attributed network pattern matching. (Java)	Not Available	Analytics	https://github.com/Aptima/pattern-matching.git	
Boeing, University of Pittsburgh	SMILE-WIDE: A scalable Bayesian network library	SMILE-WIDE is a scalable Bayesian network library. Initially, it is a version of the SMILE library, as in SMILE With Integrated Distributed Execution. The general approach has been to provide an API similar to the existing API SMILE developers use to build "local," single-threaded applications. However, we provide "vectorized" operations that hide a Hadoop-distributed implementation. Apart from invoking a few idioms like generic Hadoop command line argument parsing, these appear to the developer as if they were executed locally. (Java)		Analytics	https://github.com/SmileWide/main.git	
Carnegie Mellon University (publications)	Active Search	ActiveSearch takes a collection of emails (or any dataset where a similarity can be generated between elements) and recommends related messages based on user feedback. The user provides an initial seed email then enters into a cycle where ActiveSearch provides a similar email and the user reports whether or not the email was interesting. ActiveSearch is useful for anyone navigating a large set of emails and looking for related messages on a specific topic. As it considers the similarities between emails as well as user feedback, it is an improvement in accuracy, time, and effort over basic text search or a brute force search. (Java, Perl)		Analytics	https://github.com/AutonlabCMU/ActiveSearch	

DARPA's open source office

- DARPA recognized that just open sourcing isn't enough
 - Need for maintenance, improvement, support
- JPL (an FFRDC) is DARPA's open source center of excellence
 - JPL has been part of the Apache foundation board
 - Currently supporting 5 large open source projects
- **Mission: build community, continued support for software maintenance**



Mr. Wade Shen, PM
wade.shen@darpa.mil



www.darpa.mil

Doug Gelbach, SETA
douglas.gelbach.ctr@darpa.mil

<http://www.darpa.mil/opencatalog>