XDATA Overview
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...
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)
Mesos

- Cloud scheduling via Linux containers
- Plays nice with Hadoop/Spark/MPI clusters
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 auto-vectorization

```python
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

```plaintext
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

```plaintext
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

```scala
untilConverged(theta, tol) { theta =>
  x.numFeatures { j => // vector of sums
    val gradient = sum(0, x.numSamples) { i =>
      x(i)(j) * (y(i) - hyp(theta, x(i)))
    }
  }
  theta(j) + alpha * gradient
}
```
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

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

**Kiva Microloans**
- # Loans: 524,514
- # Lenders: 1.1M
- # Partners: 238
- # Transactions: 4,069,217
- # Journal Entries: 307,831
- # Total $ Loaned: $418M (USD)

**Bitcoin**
- # Transactions: 15.8M+
- # Edges: 37.4M+
- # Senders: 5.4M+
- # Receivers: 6.3M+
- # Bitcoins Xted: 1.4M+

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

**Text/Unstructured Data**
- #2.1 Million unique employment ads in Spanish from South America

**VADER**
- Size: 5 GB

**SR Hawk**
- Size: 60 MB

**MAGIC**
- Size: 1.2 GB

**LYNX II**
- Size: 11 GB

**HYDRA**
- Size: 44 GB

**GEOnet**
- Size: 3.2 TB

**GMBIT**
- Size: 2.2 TB

**GEOnet**
- Size: 3.2 TB

**GISR**
- # Senders: 5.4M+
- # Receivers: 6.3M+
- # Bitcoins Xted: 1.4M+

**Bitcoins**
- # Transactions: 15.8M+
- # Edges: 37.4M+

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

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


<table>
<thead>
<tr>
<th>Team</th>
<th>Project</th>
<th>Description</th>
<th>Instructional Material</th>
<th>Category</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aptima, Inc.</td>
<td>Network Query by Example</td>
<td>Hadoop MapReduce-over-Hive based implementation of network query by example utilizing attributed network pattern matching. (Java)</td>
<td>Not Available</td>
<td>Analytics</td>
<td><a href="https://github.com/Aptima/pattern-matching.git">https://github.com/Aptima/pattern-matching.git</a></td>
</tr>
<tr>
<td>Boeing, University of Pittsburgh</td>
<td>SMILE-WIDE: A scalable Bayesian network library</td>
<td>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 &quot;local,&quot; single-threaded applications. However, we provide &quot;vectorized&quot; 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)</td>
<td>Analytics</td>
<td><a href="https://github.com/SmileWide/main.git">https://github.com/SmileWide/main.git</a></td>
<td></td>
</tr>
<tr>
<td>Carnegie Mellon University (publications)</td>
<td>Active Search</td>
<td>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)</td>
<td>Analytics</td>
<td><a href="https://github.com/AutoNlabCMU/ActiveSearch">https://github.com/AutoNlabCMU/ActiveSearch</a></td>
<td></td>
</tr>
</tbody>
</table>
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