Machine Learning in the Big Data Era: Are We There Yet?
Experiences with Data-Parallel Frameworks

Rangan Sukumar (Email: sukumarsr@ornl.gov)

ORNL Contributors: Seung-Hwan Lim, Matt Lee (post-doc), Regina Ferrell, Arjun Shankar, Robert Patton and Chris Symons

Interns:

Tyler Brown (HERE), Keela Ainsworth (HERE), Larry Roberts (SULI), Yifu Zhao (ORSS), Nathaniel Bond (SULI), Carlos del-Castillo-Negrette (SULI), Katie Senter (RAMS) and Gautam Ganesh (HERE).

1. University of Tennessee, Knoxville, 2. Tennessee Tech University, Cookeville, 3. Denison University, Ohio, 4. Cedarville University, 5. Yale University, 6. Penn State University, 7. University of Texas, Dallas
A thought experiment...

What is common to (airplane, rainbow, star, thunder)?
What did we just do?

**Intelligence**

Discovery by “interrogation”
A storage and retrieval problem
Example: Memorizing a vocabulary

Discovery by “modeling/simulation”
A pattern-discovery/modeling problem
Example: Creating models for recognition

Discovery by “association”
A disparate data fusion problem
Example: Mapping visual and semantic similarity
Today’s Talk

Is our ability to learn keeping pace with the ability to store and compute?

• If Yes: How so?
• If Not: Why not?
National Scale Big Data: Healthcare

Data

- Patients
  - Eligibility
  - Enrollment/Buy-in
  - Demographics

- Provider
  - Enrollment
  - Feedback

- Health
  - Medicare vs. Medicaid
  - Operations
  - Coverage / Eligibility
  - Geography

- Cost
  - Standardization
  - Coverage / Eligibility

- Policy
  - Legislation
  - Guidelines

Analytics

- Descriptive Analytics
- Diagnostic Analytics
- Predictive Analytics
- Prescriptive Analytics

Insights

- What happened?
- Why did it happen?
- What will happen?
- How can we make it happen?

Concept adapted from Gartner’s Webinar on Big Data

Motivation: Need insights for better integrity, quality and policy.
National Scale Big Data: Health Data

Transactions
Images
Sequences
Text
Sensors

Data Sources

The Grand Challenge
Enable Discovery

Knowledge Discovery Infrastructure

Outcomes

Motivation: Need insights for the “science” of personalized medicine
National Scale Big Data: National Security

Motivation: Automatically tag and triage images for law enforcement

MIRFLICKR
~25K – 1 M images, ~8 tags per image
- Nicaragua, Landscape, Outdoor, Mountains, Hill
- Barn, Mountains, Sky, Wyoming
- Tanzania, Mountains, Grass Plains, Landscape
- U.S Capitol, Capitol, Hill, Capital, Senate

WordNet
{30K adjectives, 146 K nouns, ~25 K verbs} in ~117K synsets
- mound, hill
- ascent, acclivity, rise, raise, climb, upgrade

ImageNet
~14M images, ~22k synsets
- mountain
- cabin
- valley
- has_property
- Defined As
- some_home

ConceptNet
~3.4 M words, ~57 K types of relationships, ~10 million assertions
- mt. pisgah
- is a
- located near
- located near
- natural place
- forest dwelling

~25K – 1 M images, ~8 tags per image

~14M images, ~22k synsets

~3.4 M words, ~57 K types of relationships, ~10 million assertions
What is common to national security and healthcare?

Healthcare

- $O(10^6)$ of Patients
- $O(10^9)$ of Transactions / year
- $O(10^3)$ of Features
- $O(10^4)$ of Diagnoses and Procedures
- ~100 TBs of Data

Image Triage

- $O(10^6)$ of Users
- $O(10^9)$ of Pictures
- $O(10^3)$ of Features
- $O(10^4)$ of Words (tags)
- ~100 TBs of Data

The Lifecycle of Data-Intensive Discovery

- Querying and Retrieval e.g. Google, Databases
- Interrogation
- Better Data Collection
- Association
- Data-fusion e.g. Mashups
- Modeling, Simulation, & Validation
- Predictive Modeling e.g. Fraud Detection

Can we scale up all three aspects of data-driven discovery?
More specifically...

Healthcare
- $O(10^6)$ of Patients
- $O(10^9)$ of Transactions / year
- $O(10^3)$ of Features
- $O(10^4)$ of Diagnoses and Procedures
- ~ 100 TBs of Data

Image Triage
- $O(10^6)$ of Users
- $O(10^9)$ of Pictures
- $O(10^3)$ of Features
- $O(10^4)$ of Words (tags)
- ~ 100 TBs of Data

The Machine Learning Problem at Scale

Given examples of a function $(x, f(x))$, find function $f$ that can predict for new samples $x$.

<table>
<thead>
<tr>
<th>$f_1$</th>
<th>$f_2$</th>
<th>...</th>
<th>...</th>
<th>...</th>
<th>$f_d$</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>features</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$c_1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$c_2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$c_3$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$c_4$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$c_k$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>categories</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$c_1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$c_2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$c_3$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$c_4$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$c_k$</td>
</tr>
</tbody>
</table>
## The opportunity

### Machine Learning in the Big Data Era: Literature Survey

<table>
<thead>
<tr>
<th></th>
<th>1990 – 2000s</th>
<th>2010-Present</th>
<th>Insight</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Assumption</strong></td>
<td>A model exists. Better data will reveal the beautiful model. (Knowing “why” is important)</td>
<td>A model may not exist, but find a model anyway. (“Why” is not as important)</td>
<td>Dilemma: Better data or better algorithms.</td>
</tr>
<tr>
<td><strong>Complexity of data</strong></td>
<td>$N \sim O(10^2)$, $d \sim O(10^4)$ (e.g. IRIS data) $k \sim O(1)$</td>
<td>$N \sim O(10^6)$, $d \sim O(10^4)$ (e.g. ImageNet) $k \sim O(10^4)$</td>
<td>Volume, Velocity, Variety and Veracity have all increased several orders of magnitude.</td>
</tr>
<tr>
<td><strong>Data – Model Relationship</strong></td>
<td>Model abstracts data</td>
<td>Data is the model</td>
<td>Models aggregated data. It is not anymore about the average. It is about every individual data point.</td>
</tr>
<tr>
<td>$\hat{p}(X_j</td>
<td>C = c_i) = \frac{1}{\sqrt{2\pi\sigma_j}} \exp \left( -\frac{(X_j - \mu_j)^2}{2\sigma_j^2} \right)$</td>
<td>$f(x) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{h_i} G(\frac{x - x_i}{h_i})$</td>
<td></td>
</tr>
<tr>
<td><strong>Model Parameter Complexity</strong> (e.g. Size of Neural Network)</td>
<td>$O(10^3)$</td>
<td>$O(10^{10})$</td>
<td>10-billion parameter network learned to recognize cats from videos.</td>
</tr>
<tr>
<td><strong>Accuracy, Precision, Recall</strong> (e.g. Face Recognition, Scene Recognition)</td>
<td>~ 70% was accepted. Not possible</td>
<td>~95% is the norm. ~10% is the best result to date.</td>
<td>Big Data also means Big Expectations.</td>
</tr>
<tr>
<td><strong>Computing Capability Personal Computing High Performance Computing</strong></td>
<td>1 core, 256MB RAM, 8GB disk 1000 cores, 1 teraflops</td>
<td>16 cores, 64 GB RAM, 2TB disk 3 million cores, 34 petaflops</td>
<td>Commercial tools are keeping pace with the PC market and not HPC market.</td>
</tr>
</tbody>
</table>
## The opportunity at ORNL

Leadership computing resources (along with Other Test-beds)

<table>
<thead>
<tr>
<th></th>
<th>Titan</th>
<th>Apollo</th>
<th>CADES (Cloud)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Discovery Approach</strong></td>
<td>Modeling and Simulation</td>
<td>Association</td>
<td>Querying, Prediction</td>
</tr>
<tr>
<td><strong>Architecture</strong></td>
<td>Shared-compute</td>
<td>Shared-memory</td>
<td>Shared-storage</td>
</tr>
<tr>
<td><strong>Scalability</strong></td>
<td>Compute (# of cores)</td>
<td>Horizontal (# of datasets)</td>
<td>Vertical (# of rows)</td>
</tr>
<tr>
<td><strong>Algebra</strong></td>
<td>Linear</td>
<td>Relationship</td>
<td>Set-theoretic</td>
</tr>
<tr>
<td><strong>Challenge (Pros)</strong></td>
<td>Resolution</td>
<td>Heterogeneity</td>
<td>Cost</td>
</tr>
<tr>
<td><strong>Challenge (Cons)</strong></td>
<td>Dimensionality</td>
<td>Custom Solution</td>
<td>Flexibility</td>
</tr>
<tr>
<td><strong>Leadership</strong></td>
<td>#2 in the world (2013)</td>
<td>1 of 15 installs (2013)</td>
<td>--</td>
</tr>
<tr>
<td><strong>User-interface</strong></td>
<td>OpenMP, MPI, CUDA</td>
<td>SPARQL</td>
<td>SQL</td>
</tr>
</tbody>
</table>
What did we learn about machine learning in the Big Data Era?

• Big Data is only going to generate “Bigger Data” because discovery is a never-ending thirst for more.
  • Unfortunately, algorithms today do not evolve with evolving knowledge.

Three Big Gaps specific to the Big Data Era

1. The Data Science
2. The Science of Data
3. The Science of Scalable Predictive Functions
**Gap #1: Data Science**

Need: Infrastructure-aware algorithm design requires a diverse skill set.

<table>
<thead>
<tr>
<th>Systems</th>
<th>Data</th>
<th>Compute</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infrastructure</td>
<td>Management</td>
<td>Structure</td>
<td>HPC</td>
</tr>
<tr>
<td>Design</td>
<td>Quality</td>
<td>Matrix/Table</td>
<td>TITAN</td>
</tr>
<tr>
<td>Operations</td>
<td>Privacy</td>
<td>Text, Image, Video</td>
<td>CADES</td>
</tr>
<tr>
<td>Management</td>
<td>Provenance</td>
<td>Graphs</td>
<td>Cloud</td>
</tr>
<tr>
<td></td>
<td>Governance</td>
<td>Sequences</td>
<td>Urika</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spatiotemporal</td>
<td>Hadoop</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Schema</td>
<td></td>
</tr>
</tbody>
</table>

- Performance of “algorithm” dependent on architecture.
  - Most data scientists/algorithm specialists are used to in-memory tools such as R, MATLAB etc.
  - Existing cloud-based solutions are designed for high performance storage and not high-performance compute or in-memory operations.
  - Steep learning curve towards programming “new” innovative algorithms. Too many options without guiding benchmarks.
Gap #1: Data science (contd.)

Algorithm performance is a function of time, space and i/o requirements and not all aspects of are parallelizable.

• **Staging for Predictive Modeling**
  - Extract, Transform, Load
  - Data Pre-processing
  - Feature engineering

• **Predictive Modeling**
  - Rule-base extraction
  - Pairwise-similarity (Distance Computation)
  - Model-parameter estimation
  - Optimization/Regularization

• **Inference/ Model Deployment**
  - Cross-validation
  - Data is model ? Model is data ?
  - Adaptive model ? Reinforcement ?

  **Disk Intensive**
  File processing and repeated retrieval best done in massively parallel file systems or databases

  **Memory and Compute Intensive**
  Typically computing an aggregate measure, vector product, a kernel function etc.

  **Memory + Compute Intensive**
  Real-time low-latency requirements
Gap #1: Data Science (contd.)

How does choice of architecture affect analytical algorithms on contemporary Big Data infrastructures?

**Shared-storage (Cloud)**

**Shared-nothing HPC**

**Experiment:** Benchmark Apache Hadoop (Big Data Stack) on Cloud and HPC architectures

---

**TestDFSIO Read 1GB files**

<table>
<thead>
<tr>
<th></th>
<th>Rhea-4</th>
<th>Rhea-8</th>
<th>Rhea-12</th>
<th>Rhea-15</th>
<th>m2.4xlarge-4</th>
<th>m2.4xlarge-8</th>
<th>m2.4xlarge-12</th>
<th>m2.4xlarge-15</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Avg. individual throughput (MB/sec)</strong></td>
<td><img src="image1.png" alt="Graph" /></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Number of files</strong></td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
</tr>
</tbody>
</table>

**TestDFSIO Write 1GB Files**

<table>
<thead>
<tr>
<th></th>
<th>Rhea-4</th>
<th>Rhea-8</th>
<th>Rhea-12</th>
<th>Rhea-15</th>
<th>m2.4xlarge-4</th>
<th>m2.4xlarge-8</th>
<th>m2.4xlarge-12</th>
<th>m2.4xlarge-15</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Avg. Test Exec Time (sec)</strong></td>
<td><img src="image2.png" alt="Graph" /></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Number of files</strong></td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
</tr>
</tbody>
</table>
Gap #2: Science of Data

Need: Data-aware algorithm design

• Data-science is not the same as “science of data”
  – Is the process of understanding characteristics of data before applying/designing a machine-learning algorithm.

  Data Sources

  - Data characterization – (i.e., avoid using machine learning as a black box)
    • Signal-noise-ratio, bound on noise
    • i.i.d sampling assumptions
    • stationarity, randomness, ergodicity, periodicity
    • Generating models behind data
Gap #2: Science of Data (contd.)

Need: Data-aware algorithm design

• Today’s Big Data: Matrices and Databases
  • Feature engineering requires automated methods to deal with unstructured data in future.
  • Traditional (pain-staking) pipeline of SMEs creating features from the data will have to be augmented by automated methods.
  • Tomorrow is unstructured.

• Dealing with hierarchies and latent relationships in data
  • Potential ideas:
    • Deep learning, Multi-task learning, etc.
  • The ability to predict the performance of an algorithm on a dataset.
    • Quantify the curse of dimensionality, the informative and discriminatory power of the feature matrix, etc.
Gap #3: Scalable Predictive Functions

How are learning algorithms handling increasing samples (N), dimensionality (d) and categories (k) ?

Experiment: Benchmark libSVM on NUS-WIDE Scene recognition challenge dataset
Ongoing Work:

Designing knowledge-nurturing architectures and scalable learning algorithms
Requirements

Key Requirements: Scalability, Flexibility, Linked Structures, Relationship Analytics

Potential Solution: Scalable Graph Computing
State-of-the-art with graph computing

In-memory graph analysis

NetworkX

iGraph

Pajek

Distributed-storage graph analysis

Neo4j

GraphLab

Project Pegasus

CytoScape

Different programming models for scalable graph computing:

- Asynchronous batch processing
- Map-Reduce (distributed storage)
- Pregel (distributed memory)
- SPARQL (shared-memory and shared storage)
- In-database massively-parallel SQL
- In-memory databases

When: Datasets > memory in desktop + Real-time reasoning and inference
The Data Science

How do contemporary infrastructures scale for graph computing?

Benchmarked popular graph mining algorithms: Degree distribution, PageRank, Connected components, Radius, Triangle counting and Shortest path

Shared-memory architecture wins:

- Compared to CMU Pegasus (2010) – ten times speed-up.
- Compared to Berkeley GraphX (2014) – 2 to 5 times speed-up.

Same latency on graph retrieval as a desktop but on data 1000 times larger.
Discoveries by interrogation

Urika: Basic Graph Pattern Search and Retrieval

Graph with different node types and edge types

Input graph  +  SME interest pattern  =  Interest pattern pattern retrieval

A query graph pattern

Rule of thumb: Any query that takes longer than 45 seconds (on ~ TBs) is bad code!
Discoveries from association

Advanced pathway patterns

What are pathway patterns?

New longer paths

New triangles and polygons
Discoveries from association

Similarity-based prediction

What is the probability of a particular edge to occur?

Predictive inference using recommender strategies.
Pattern Discovery Algorithms

Find the “largest” recurring sub-graph

Repeating Graph Pattern

Anomalous Graph Pattern
What are we doing with Urika?

Converting the retrieval engine to a discovery engine

Scaling for Volume

Scaling for Heterogeneity

Graph-theory meets Semantic Graphs

EAGLE: ‘Is A’ Algorithmic Graph Library for Exploratory-Analysis
_components_of_EAGLE

PLUS – Programmatic-Python Login for Urika-like SPARQL-Endpoints
This is the debug/test environment that you can build on a local machine using Apache-Jena. Once tested on Jena, it is then a matter of changing a couple of parameters to run on Urika.

FELT – Flexible Extract, Load and Transform Toolkit
Both Jena and Urika process data as RDF-triples. Unfortunately, most real world datasets live in databases or as flat files. These set of scripts loads data in common formats and when specified with a graph-model converts relational data to triples that can be analyzed using SPARQL-end points.

GraphIC – (pronounced as Graph-i-see or graphic) – is the Graph Interaction Console.
This is the light weight graph-browser that retrieves graphs from end-points accessible through PLUS. Tested to work on display devices from IPads to ORNL EVEREST.
Use case example: Visualization of massive graphs

EAGLE – Command line
This is the set of “lego block” python scripts of popular graph-theoretic algorithms (graph summarization, centrality, Page Rank, clustering, pattern discovery etc.).
Use case example: Fraud Detection using Guilt-by association

PAUSE: Predictive Analysis using SPARQL-Endpoints
A toolkit to analyze multi-structure data (numeric – data integrated with domain knowledge). Has built in capability for similarity analysis, link prediction, simultaneous feature sub-setting and feature matching.
Use case example: Disease Prediction, Identity Disambiguation

KENODES: Knowledge Extraction using Network-Oriented Discovery Enabling System
This set of scripts exploits the semantic power of a machine like Urika to reason and interpret knowledge bases and is a first step in automated hypothesis generation.
Use case example: NLM work on literature-based discovery.
Use Cases: Entity Resolution/ Active MDM

Problem: Entities hiding across multiple systems

Can we discover the “golden-rule” when two entities are potentially the same?

Different names associated with the ID in TX-provider file

HEALTHCARE MEDICAL ASSOCIATES
HEALTHCARE MEDICAL ASSOCIATES
MEDISTAT GROUP ASSOCIATES

19 distinct people using the same address and phone in TX.
Use Case: Social Network of Fraud

Given a few examples of fraud (important activity), can we
(i) Discover patterns typically associated with suspicious activity?
(ii) Extrapolate such high-risk patterns for investigation and fraud prevention?

V. Chandola, S.R. Sukumar and J. Schryver, "Knowledge Discovery from Massive Healthcare Claims Data", in the Proc. of the 19th ACM SIGKDD Conference on Knowledge Discovery, 2013 (Acceptance Ratio: 17.4%)
Use Case: Pattern Discovery (contd.)

Understanding relationship patterns using mathematical models

Extrapolating for new hypothesis/investigation
Insight from Patterns

Affiliations to multiple hospitals, owning private and group practice are strong indicators of potential suspicious activity.

High risk “referral” patterns are Y networks and triangles (above), low risk patterns are star-shaped (below).
Use Case: Literature-based Discovery

Given data and meta-data as knowledgebase from different domains, can we
(i) Discover new relationships of entities between domains?
(ii) Automatically extract and prioritize discovered relationships for clinical or subject matter expert validation?

Use Case: Literature-based Discovery
Use Case: Literature-based Discovery

Can Cloroplexithane cause leprosy?
## Use Case: Diabetic Retinopathy (Dr. Ed Chaum, UTHSC)

### Sample Measurements

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Pregnant</th>
<th>Race</th>
<th>OD Condition</th>
<th>Age</th>
<th>HgbA1c</th>
<th>Cholest.</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>N</td>
<td>African</td>
<td>NPDR Severe + CSME</td>
<td>37</td>
<td>Null</td>
<td>185</td>
</tr>
<tr>
<td>43</td>
<td>N</td>
<td>Caucasian</td>
<td>No diabetic retinopathy</td>
<td>64</td>
<td>11.7</td>
<td>161</td>
</tr>
<tr>
<td>58</td>
<td>Y</td>
<td>Unknown</td>
<td>Null</td>
<td>33</td>
<td>10.2</td>
<td>220</td>
</tr>
<tr>
<td>104</td>
<td>N</td>
<td>Hispanic</td>
<td>Other</td>
<td>53</td>
<td>9.7</td>
<td>Null</td>
</tr>
<tr>
<td>135</td>
<td>N</td>
<td>African</td>
<td>NPDR Mild/Minimal - CSME</td>
<td>62</td>
<td>8.2</td>
<td>148</td>
</tr>
</tbody>
</table>

### Images

- **Comments**
  - Poor quality images but adequate for diagnosis
  - Vascular tortuosity (congenital). No retinopathy
  - Mild ischemia (cotton wool spots) No hemorrhages or edema
  - possible mild drusen No DR evident
  - rare microaneurysms only f/u 12 months

### Dataset

- 7600 patients, 31 clinical lab measurements, at least 1 image and report per patient, about a 100 meta-data variables over 3 year period.
Knowledgebase + New Data

Convert data into RDF

Integrate Knowledgebase

Find similar patients

Link Prediction

What did we find?

Simultaneous Feature-subset and link prediction

**Attributes common with top 100 similar patients without DR (> 80% support)**

DM_Status: Normal routine history

DM_Problem: Hypertension

DM_Drug: Lisinopril

**Attributes common with top 100 patients with DR ( > 80% support)**

DM_Medication : "INSULIN + PILL"

Condition_EE : "NPDR Mild/Minimal + CSME"

DM_Drug : "Glyburide"

Image courtesy: Katie Senter
Results

Simultaneous feature-subset and link prediction

Diabetic Retinopathy (DR) and Beta Blockers (BB) in Cross-Sectional Data: Hypothesis Testing

Incidence of DR by Population Characteristic

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Incidence of DR (%)</th>
<th>Total (n)</th>
<th>With DR (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB</td>
<td>46.5</td>
<td>88</td>
<td>41</td>
</tr>
<tr>
<td>No BB</td>
<td>28.6</td>
<td>412</td>
<td>118</td>
</tr>
<tr>
<td>No BB + hypertension (HTN)</td>
<td>29.0</td>
<td>338</td>
<td>98</td>
</tr>
<tr>
<td>Diabetes mellitus (DM) onset &gt; 5 years</td>
<td>39.1</td>
<td>233</td>
<td>91</td>
</tr>
<tr>
<td>DM onset ≤ 5 yr.</td>
<td>7.8</td>
<td>115</td>
<td>9</td>
</tr>
<tr>
<td>DM onset &gt; 5 yr. + BB</td>
<td>54.0</td>
<td>50</td>
<td>27</td>
</tr>
<tr>
<td>DM onset ≤ 5 yr. + BB</td>
<td>14.3</td>
<td>14</td>
<td>2</td>
</tr>
<tr>
<td>DM onset &gt; 5 yr. no BB</td>
<td>35.7</td>
<td>207</td>
<td>74</td>
</tr>
<tr>
<td>DM onset ≤ 5 yr. no BB</td>
<td>6.9</td>
<td>101</td>
<td>7</td>
</tr>
</tbody>
</table>

Incidence of DR: 2 × 2 Fisher’s Exact Tests

BB vs. No BB

<table>
<thead>
<tr>
<th></th>
<th>DR</th>
<th>No DR</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB</td>
<td>41</td>
<td>47</td>
</tr>
<tr>
<td>No BB</td>
<td>118</td>
<td>294</td>
</tr>
</tbody>
</table>

p = 0.0010  (extremely statistically significant)
**Other Use Cases: Semantic Pattern Analysis**

A verbal-fluency exam:
Neuro-psychiatrist: *Name as many animals as you can in 60 seconds?*
PTSD Patient: racoon, lion, elk,.....
Normal Patient: cat, tiger, dog, horse,....

<table>
<thead>
<tr>
<th>Participant Number</th>
<th>WAISVocab_R32</th>
<th>WAISVocab_R33</th>
<th>Verbal_Fluency_F</th>
<th>Verbal_Fluency_A</th>
<th>Verbal_Fluency_S</th>
<th>Verbal_Fluency_Anim</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>approach</td>
<td>tantrum</td>
<td>field, fly, fluke, flute,</td>
<td>alert, alarm, aloof, and,</td>
<td>suck, supplement,</td>
<td>whale, fish, kangaroo,</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td>football, fast, follow,</td>
<td>fault, fail, frog,</td>
<td>asterisk, afford, after,</td>
<td>scallops, song, streetcar,</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td>fire, fill, fan,</td>
<td>almost, any, act, action,</td>
<td>sure, shade, shout, siri,</td>
<td>square, squirrel, sun,</td>
</tr>
<tr>
<td>8</td>
<td>don't know</td>
<td>Go on a rampage</td>
<td>fire, fart, flair, fire,</td>
<td>ass, apricot, angle,</td>
<td>smart, sniff, sniff, snort,</td>
<td>possum, raccoon, fish,</td>
</tr>
<tr>
<td>9</td>
<td>don't know</td>
<td>don't know</td>
<td>farm, fart, flair, fire,</td>
<td>although, after,</td>
<td>southern, south, sunrise,</td>
<td>weasel, rabbit, snake,</td>
</tr>
<tr>
<td>10</td>
<td>don't know</td>
<td>don't know</td>
<td>flounder, found,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>carry, not</td>
<td>spazmonic, fit,</td>
<td>fix, fight, fabric,</td>
<td>apparent, able, acetone,</td>
<td>sample, sign, salt, sunny,</td>
<td>cow, horse, buffelo,</td>
</tr>
<tr>
<td>13</td>
<td>don't know</td>
<td>idealism, don't</td>
<td>front, far, fear, lack,</td>
<td>area, ate, and,</td>
<td>single, sit, situation, short,</td>
<td>wallaby, wolverine,</td>
</tr>
<tr>
<td>14</td>
<td>weigh someone</td>
<td>raving</td>
<td>famine, fun, frank,</td>
<td>art, are, and, an, arbore,</td>
<td>submerge, single, sample,</td>
<td>alligator, ape, cow,</td>
</tr>
<tr>
<td>15</td>
<td>responsibility on</td>
<td>express something</td>
<td>fake, five, fifteen,</td>
<td>aplomb, askew, age,</td>
<td>surround, simple, skew,</td>
<td>hippo, rhino, cow,</td>
</tr>
<tr>
<td>16</td>
<td>of real property</td>
<td>into a violent rage</td>
<td>further, future,</td>
<td>advertisement, age,</td>
<td>summation, something,</td>
<td>tiger, leopards, lizards,</td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td>field, follow, funny,</td>
<td>apprehend, ax, ashes,</td>
<td>smell, scent, sway, start,</td>
<td>giraffe, tiger, zebra,</td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>aside</td>
<td>screaming fit</td>
<td>fat, funny, fart, flat,</td>
<td>always, accommodate,</td>
<td>song, sasafras, segway,</td>
<td>snake, whale, porpoise,</td>
</tr>
<tr>
<td>20</td>
<td>burden causing</td>
<td>tirade</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>21</td>
<td>someone down</td>
<td>response- like sad:</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>22</td>
<td>bumps or hurdle to upside down, a fit</td>
<td>fudge, fur, flag,</td>
<td>alphabet, alien, airplane</td>
<td>sensitive silly sound size,</td>
<td>orangutan, monkey,</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>something, hinders</td>
<td>tangent, rant, not a (season), fall (verb)</td>
<td>asterisk, above, after,</td>
<td>store (noun), store (verb),</td>
<td>monkey, zebra,</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>DK</td>
<td>IDX, no idea</td>
<td>fortunate, fetal,</td>
<td>algae, alert, abuse,</td>
<td>seductress, sorry, salute,</td>
<td>leopard, hyena, lion,</td>
</tr>
</tbody>
</table>

Can we host all of human common sense in-memory (of a computer) to evaluate patterns of thought progression during verbal fluency exams?
Use Cases: Semantic Pattern Analysis

Phonetic, phonemic and conceptual features for semantic pattern similarity in recall and recitation

10 such features evaluated in this study (3 sequence, 4 phonetic and phonemic, 3 conceptual similarity features).

Do words sound or spell similar? Are words conceptually similar?

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Sequence Features (%)</th>
<th>Phonetic Features (%)</th>
<th>Concept Features (%)</th>
<th>All Features (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words: ‘a’</td>
<td>63 (57 &amp; 71)</td>
<td>63 (54 &amp; 70)</td>
<td>64 (56 &amp; 70)</td>
<td>67 (59 &amp; 72)</td>
</tr>
<tr>
<td>Words: ‘s’</td>
<td>61 (53 &amp; 71)</td>
<td>64 (56 &amp; 71)</td>
<td>64 (54 &amp; 75)</td>
<td>61 (52 &amp; 67)</td>
</tr>
<tr>
<td>Words: ‘f’</td>
<td>64 (55 &amp; 74)</td>
<td>59 (49 &amp; 69)</td>
<td>59 (50 &amp; 70)</td>
<td>58 (48 &amp; 67)</td>
</tr>
<tr>
<td>Animals</td>
<td>57 (49 &amp; 63)</td>
<td>62 (53 &amp; 71)</td>
<td>58 (50 &amp; 64)</td>
<td>60 (51 &amp; 67)</td>
</tr>
</tbody>
</table>

So, are we there yet?

• The Big Data community has done well when “analysis” is retrieval.
  – There is tremendous opportunity to do better beyond that.

• The Big Data buzz is not an opportunity for applying old techniques to newer data.
  – Let’s develop smarter algorithms that are infrastructure-aware, data-aware and scale-friendly.
So, are we there yet?

Dimensions of Big Data

**Compute**
- **Storage**
- **Memory**
- **Cores**

**I/O ? Network ?**

**Dimensions of Big Data Software**

- **Volume**
  - Hadoop, MPP, Spider
- **Velocity**
  - Streaming, Batch
- **Variety**
  - SQL, NoSQL, Graph

**Analytical Requirements Algorithms**

- **Programming**
  - MapReduce, MPI, Threads
- **Data-Parallel**
- **Complexity of Algorithms**
  - Linear, Iterative
  - > O(N^2)
- **Compute on Data**
  - Retrieval, Machine Learning
- **Speed of Execution**
  - Real-time, Feasibility

**What can be scaled up ?**
**What aspect of data that needs scale up ?**

- **What aspect of algorithm that needs scale up ?**

• Associative memory architectures are showing promise.
What is the concept behind words – *(airplane, rainbow, star, thunder)*?

- 'music group'
- 'work'
- 'organization'
- 'band'
- 'album'
- 'sky'
- 'musical work'
- 'creative work'
- 'music album'
- 'single'