Data Science with Visual Analytics

30 Min Teaser Talk

Klaus Mueller
Stony Brook University and SUNY Korea

SUNY Korea 2015 HOT–T–CS
Hot Topics in Computer Science

July 13 – 17, 2015
SUNY Korea, Songdo, Korea
DATA SCIENCE –
WHY ALL THE EXCITEMENT?
Dr. John Snow’s London Cholera Map (1854)
- data collection
- data assimilation
- statistical testing
- visualization
- computational analysis (brain)
- domain knowledge

Very early example of data science
MODERN DATA SCIENTIST

MATH & STATISTICS
- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression

PROGRAMMING & DATABASE
- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages, e.g., R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query

DOMAIN KNOWLEDGE & SOFT SKILLS
- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION
- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau
Google Flu Trends

Predict emerging flu from search terms in specific regions

Could predict regional outbreaks of flu up to 10 days before reported by the CDC
Nate Silver’s Election Predictions

Takes a big-picture approach

- use multiple sources of unique data
- combine with historical data
- apply principles of sound statistical analysis
Opportunities Galore

Government achieves significant cost savings and ability to react to potential threats quickly

Government cuts acoustic analysis from hours to 70 Milliseconds

Utility provider improves prediction of power outages

Utility avoids power failures by analyzing 10 PB of data in minutes

Hospital detects and intervenes in potentially life-threatening conditions

Hospital analyzes streaming vitals to intervene 24 hours earlier

Retailer optimizes inventory levels and product mix

Retailer reduces time to run queries by 80%

Stock exchange reduces time to insights to achieve optimal buying / selling strategies

Stock Exchange cuts queries from 26 hours to 2 minutes on 2 PB

Telco provider improves ability to quickly address network issues / opportunities

Telco analyses streaming network data to reduce hardware costs by 90%
Million Server Data Center

Not always needed
Data Scientists

The U.S. will need 140,000-190,000 predictive analysts and 1.5 million managers/analysts by 2018

- McKinsey Global Institute’s June 2011

New Data Science institutes being created or repurposed – NYU, Columbia, Washington, UCB,...

New degree programs, courses, boot-camps:

- e.g., at Berkeley: Stats, I-School, CS, Astronomy...

Stony Brook as well as SUNY Korea will offer an MS Specialization in Data Science starting now (Fall 2015)
Characteristics of Big Data

**Volume**
- Data at scale
- Terabytes to petabytes of data

**Variety**
- Data in many forms
- Structured, unstructured, text, multimedia

**Velocity**
- Data in motion
- Analysis of streaming data to enable decisions within fractions of a second

**Veracity**
- Data uncertainty
- Managing the reliability and predictability of inherently imprecise data types
**Big Data Approach to Science**

### Traditional Analytics
- Structured & Repeatable
- Structure built to store data

1. **Hypothesis**
2. **Question**
3. **Data**
4. **Answer**

Start with hypothesis
Test against selected data

Analyze after landing…

### Big Data Analytics
- Iterative & Exploratory
- Data is the structure

1. **Data**
2. **Exploration**
3. **Actionable Insight**
4. **Correlation**

Data leads the way
Explore *all* data, identify correlations

Analyze in motion…
## Databases vs. Data Science

<table>
<thead>
<tr>
<th></th>
<th>Databases</th>
<th>Data Science</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Value</td>
<td>“Precious”</td>
<td>“Cheap”</td>
</tr>
<tr>
<td>Data Volume</td>
<td>Modest</td>
<td>Massive</td>
</tr>
<tr>
<td>Examples</td>
<td>Bank records, Personnel records, Census, Medical records</td>
<td>Online clicks, GPS logs, Tweets, Building sensor readings</td>
</tr>
<tr>
<td>Priorities</td>
<td>Consistency, Error recovery, Auditability</td>
<td>Speed, Availability, Query richness</td>
</tr>
<tr>
<td>Structured</td>
<td>Strongly (Schema)</td>
<td>Weakly or none (Text)</td>
</tr>
<tr>
<td>Properties</td>
<td>Transactions, ACID*</td>
<td>CAP* theorem (2/3), eventual consistency</td>
</tr>
<tr>
<td>Realizations</td>
<td>SQL</td>
<td>NoSQL: Riak, Memcached, Apache River, CouchDB. etc.</td>
</tr>
<tr>
<td>Approach</td>
<td>Query the past</td>
<td>Query the future</td>
</tr>
</tbody>
</table>

CAP = Consistency, Availability, Partition Tolerance
ACID = Atomicity, Consistency, Isolation and Durability

John Canny, Berkeley
**Scientific Modeling vs. Data-Driven Modeling**

<table>
<thead>
<tr>
<th>Scientific Modeling</th>
<th>Data-Driven Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physics-based models</td>
<td>General inference engine replaces model</td>
</tr>
<tr>
<td>Problem-Structured</td>
<td>Structure not related to problem</td>
</tr>
<tr>
<td>Mostly deterministic, precise</td>
<td>Statistical models handle true randomness, and unmodeled complexity</td>
</tr>
<tr>
<td>Run on Supercomputer or High-end Computing Cluster</td>
<td>Run on cheaper computer Clusters</td>
</tr>
</tbody>
</table>

John Canny, Berkeley
Visual Data Science or Visual Analytics
AN INTRODUCTORY EXAMPLE

The Georgia Tech Jigsaw System
Next a more Conceptual View
Visual Analytics

Visual Interface

Computer

Data

Human
VISUAL ANALYTICS

Visual Interface

Computer
- computing hardware
- algorithms

Data
manage

Human
Visual Analytics

Visual Interface

Computer
- computing hardware
- algorithms

Data

Human
- pattern recognition
- creative thought

manage
Visual Analytics

Visual Interface

Computer
- computing hardware
- algorithms

Data

Human
- pattern recognition
- creative thought
- mental model
- abstracted knowledge

manage
Visual Analytics

Visual Interface

Computer
- computing hardware
- algorithms
- formal model
- formatted knowledge

Data

Human
- pattern recognition
- creative thought
- mental model
- abstracted knowledge

manage

formalized insight
**VISUAL ANALYTICS**

**Computer**
- computing hardware
- algorithms
- apply/update
- formal model
- formatted knowledge

**Data**

**Human**
- pattern recognition
- creative thought
- mental model
- abstracted knowledge

**Visual Interface**

- interact
- learn
- manage
**Visual Analytics**

**Computer**
- Computing hardware
- Algorithms
- Formal model
- Formatted knowledge

**Human**
- Pattern recognition
- Creative thought
- Apply/update
- Mental model
- Abstracted knowledge

**Visual Interface**

- Update
- Manage
- Visualize

Data

---

**Formal Model**

- Algorithms

**Pattern Recognition**

- Abstracted knowledge

---

**Computing Hardware**

- Algorithms

**Visualize**

- Creativity

---

**Format Knowledge**

- Algorithms

**Creative Thought**

- Apply/update

---

**Knowledge Abstraction**

- Algorithms

**Apply/Update**

- Abstracted knowledge

---

**Formalized Knowledge**

- Algorithms

**Manage**

- Creativity

---

**Formal Model**

- Algorithms

**Abstracted Knowledge**

- Algorithms
**Visual Analytics**

**Computer**
- computing hardware
- algorithms
- apply/update
- formal model
- formatted knowledge

**Human**
- pattern recognition
- creative thought
- apply/update
- mental model
- abstracted knowledge

**Visual Interface**

- update
- learn
- interact
- visualize
- manage
MODEL–LEARNING EXAMPLE
Model-Learning Example
webpage_load(X) :-
same_src_ips(X), same_dest_ips(X),
same_src_port(X,80),
timeframe_upper(X,10)
Model-Learning Example
webpage_load(X) :-
same_src_ips(X), same_dest_ips(X), same_src_port(X, 80),
timeframe_upper(X, 10), length(X, L), greaterthan(L, 8).
Project Suggestions
Mayhem at DinoFun World

Find out what happened when a peaceful celebration in a small town turns into crime and mayhem perpetrated by a poor, misguided and disgruntled figure from the past
For VAST 2015

- select among two mini challenges and a grand challenge
- solving the grand challenge gives more prestige for the award

Kaggle competitions [https://www.kaggle.com/competitions](https://www.kaggle.com/competitions)

- San Francisco Crime Classification
- Titanic: Machine Learning from Disaster
- Identify hand motions from EEG recordings (award $10k)
- Predict if context ads will earn a user's click (award $20k)
- Model quoted prices for industrial tube assemblies (award $30k)
- Identify signs of diabetic retinopathy in eye images (award $100k)
- and others on that site
PRESENTATION FINISHED

ANY QUESTIONS...