Scalable Data Analysis (CIS 602-02)

Data Sources

Dr. David Koop
Technical Papers

• A document that describes scientific research

• Two general categories:
  - Survey: What has been done in a specific area
  - Research: a problem, related work, solution, and results

• Writing helps clarify your own thinking and communicate it to others [N. Feamster]

• "The purpose of research is to increase the store of human knowledge, and so even the very best work is useless if you cannot effectively communicate it to the rest of the world." — M. Ernst

• Research papers are primary sources, textbooks are secondary sources

• Most recent research is not in a textbook

• Technical Reports vs. Journal Articles/Conference Proceedings
Paper Structure

- Title & Author List
- Abstract
- Introduction
- [Background/Preliminaries]
- Contribution (Approach/Theory/Specification/Implementation)
- Evaluation (Experiments, case studies)
- [Discussion]
- Related Work (here or after introduction)
- Conclusion [& Future Work]
- [Appendices]
Citations and References

I've been reading the latest draft of your paper, Cecilia.

And?

You need to include more references.

Really?

Yes, in academic writing every single statement must be supported by data or credited to a reliable source.

Says who?

Me, et al.

[Piled Higher and Deeper, J. Cham, 9/11/2015]
Reading Papers

• "How to Read and Evaluate Technical Papers", B. Griswold modified by G. Murphy

• Sometimes useful to read the paper "out of order"

• Five questions you should answer when reading a paper:
  1. What are the motivations for this work?
     • People problem
     • Technical problem
  2. What is the proposed solution?
  3. What is the evaluation of the proposed solution?
  4. What are the contributions?
  5. What are future directions for this research?
Reading Responses

• Read the white paper: The Digital Universe of Opportunities

• Write a reading response and turn in via myCourses
  - Summarize the key contributions (1 paragraph)
    • Highlight key contributions of the paper
  - Offer your critique of the paper (1-2 paragraphs)
    • Describe why the paper is important, what concepts you agree or disagree with, and how the work might be extended or integrated with other work.
    • May include questions in the critique, but they must be questions that are not trivial
  - The response must be your own writing; you may not copy text from any other source including the paper itself, other students, online resources, etc.
Assignment 1

• Start working with data in Python
• United Nations High Commissioner for Refugees maintains statistics about population of concern around the world
• Includes data about refugees and asylum seekers in different countries
• Use the provided template
• You must complete the provided functions and not rename them
• Turn in a **single** .py file via myCourses
Reading Presentations

- Present paper and critique of it
- Goal: help the class better understand a specific paper topic
- Two students per reading
  - Both summarize key points
  - One student focuses on positive points, associated ideas, and past work
  - One student focuses on criticisms, improvements, later work
- After presentations, everyone in the class should be involved in discussing items further
- Let me know of your interest in specific topics
- Volunteers for Tuesday's paper?
The Digital Universe

V. Taylor, D. Reinsel, J. F. Gantz, S. Minton
History

• EMC and IDC have been examining the growth of the "Digital Universe" since 2007
• Yearly reports that focus on key trends and data
• Focused on supporting business, not necessarily academic pursuits
• Collect some useful and provoking statistics
Growth of Data

The Digital Universe: 50-fold Growth from the Beginning of 2010 to the End of 2020

Geography of Data


- United States: 32%
- Western Europe: 13%
- China: 4%
- India: 19%
- Rest of the World: 32%

Total: 2,837 EB

Classification of Data

Candidates for Big Data

- Surveillance
- Embedded and Medical
- Data Processing
- Entertainment and Social Media
- Consumer Images and Voice

(Share of Data That is Useful If Tagged and Analyzed)

Usefulness of Data

Opportunity for Big Data

- Digital Universe
- Useful If Tagged and Analyzed

Analyzed Data

The Untapped Big Data Gap (2012)

- Useful if Tagged and Analyzed: 23%
- Tagged: 3%
- Analyzed: 0.5%

Security and Privacy

The Need for Information Security

- Privacy
- Compliance
- Custodial
- Confidential
- Lockdown

2010

2015

2020

(10% 20% 30% 40%)

(% of the Digital Universe by Security Category)

Data in the Cloud

The Digital Universe and the Cloud (2020)

- 24% Not “Touched” by the Cloud (25,030 EB)
- 13% Stored in the Cloud (5,208 EB)
- 63% Additional “Touched” by the Cloud (9,788 EB)*

* Processed or transmitted by the cloud, but not stored

What will be stored in the Cloud?

Type of Information in the Cloud in 2020

- Surveillance: 35.1%
- Entertainment: 46.7%
- Computers, Phones, Consumer Electronics: 9.7%
- Embedded and Medical: 8.5%

Total: 14,996 EB

Data Sources

• Analog to digital transitions
  - Photos
  - Voice Data
  - Video
  - Sensors

• Transient data
  - "Normal" data is often discarded
  - Data may be summarized or aggregated
"Internet of Things"

- Many devices are now digital, connected to the internet
- Wired or wireless
- 14 billion devices now (out of ~200 billion), 32 billion by 2020
- Often real-time, no need to wait for data
- Containers (aka "Files"):
  - Sensors produce small bits of data often
  - Why would this data not be aggregated before being stored?
Example Data Sources
Example Data Sources

- Radio Telescopes
- Twitter
- Wind Turbine Sensors
- Surveillance Cameras
- Cars & Airplanes
- Dog Collars
- Dishwashers
- Traffic Lights
- MRI Scanners
- NFL Football Players
- Farming
High-Value Data

**Easy to access.**
Can you obtain the data, or is it hopelessly locked away on end-user PCs, shuttling about on closed-end data processing systems, or trapped in proprietary embedded systems?

**Transformative.**
Could this kind of data, properly analyzed and acted upon, actually change a company or society in a meaningful way?

**Real-time.**
Is the data available in real-time, or does much of it come too late to drive real-time decisions and actions?

**Intersection synergy.**
Could this kind of data have more than one of the above attributes?

**Footprint.**
Could top-notch analysis of this data affect a lot of people, major parts of the organization, or lots of customers?
Messy Data

- Unstructured
- Diversely formatted
- Uncertain accuracy
- Unknown origin
- Unpredictable value
- Demands real-time attention

"Data Lakes"
- Flexible storage
- Processed and classified later
The human element

• "Increased investment in human capital and the new skills required today is the first order of business for all organizations"
• How do we decide what data is useful?
• What questions do we ask about the data to process it?
Critique

- Very focused on how data affects business
- Defines very real problems and suggests solutions
- No description of how the data was collected
  - Were surveys of businesses done?
  - Where do the estimates for the amount of data come from?
- No citations (white paper)
  - Hard to evaluate the importance of the problems
  - Hard to evaluate the utility of the solutions
Discussion

• Other sources of data?
• What makes data valuable?
• Cloud storage?
• Security & privacy
### TABLE 1.1 Scientific and Engineering Fields Impacted by Massive Data

<table>
<thead>
<tr>
<th>Area Affected in 1995</th>
<th>Area Affected in 2012</th>
<th>Noteworthy Use Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical sciences</td>
<td>Physical sciences</td>
<td>Astronomy, particle physics</td>
</tr>
<tr>
<td>Climatology</td>
<td>Climatology</td>
<td></td>
</tr>
<tr>
<td>Signal processing</td>
<td>Signal processing</td>
<td></td>
</tr>
<tr>
<td>Medicine</td>
<td>Medicine</td>
<td>Imaging, medical records</td>
</tr>
<tr>
<td>Artificial intelligence</td>
<td>Artificial intelligence</td>
<td>Natural language processing, computer vision</td>
</tr>
<tr>
<td>Marketing</td>
<td>Marketing</td>
<td>Internet advertising, corporate loyalty programs</td>
</tr>
<tr>
<td>N/A</td>
<td>Political science</td>
<td>Agent-based modeling of regime change</td>
</tr>
<tr>
<td>N/A</td>
<td>Forensics</td>
<td>Fraud detection, drug/human/ CBRNe trafficking</td>
</tr>
<tr>
<td>N/A</td>
<td>Cultural studies</td>
<td>Human terrain assessment, land use, cultural geography</td>
</tr>
<tr>
<td>N/A</td>
<td>Sociology</td>
<td>Comparative sociology, social networks, demography, belief and information diffusion</td>
</tr>
<tr>
<td>N/A</td>
<td>Biology</td>
<td>Genomics, proteomics, ecology</td>
</tr>
<tr>
<td>N/A</td>
<td>Neuroscience</td>
<td>fMRI, multi-electrode recordings</td>
</tr>
<tr>
<td>N/A</td>
<td>Psychology</td>
<td>Social psychology</td>
</tr>
</tbody>
</table>

NOTE: CBRNe, chemical, biological, radiological, nuclear, enhanced improvised explosive devices; fMRI, functional magnetic resonance imaging; N/A, not applicable.
Examples of Massive Data

- Earth and Planetary Observations
- Astronomy
- Biological and Medical Research
- Large Numerical Simulations
- Telecommunications and Networking
- Social Networks
- National Security

[Frontiers in Massive Data Analysis]
Square Kilometer Array (SKA) 1 Mid

Total raw data output:
2 terabytes per second
62 exabytes per year

x340,000

340,000 average laptops with content every day

Compared to the JVLA, the current best similar instrument in the world:
4x the resolution
5x more sensitive
60x the survey speed

[www.skatelescope.org]
SKA1-Low

Total raw data output:
- 157 terabytes per second
- 4.9 zettabytes per year

Enough to fill up 35,000 DVDs every second.

5x the estimated global internet traffic in 2015 (source: Cisco)

Compared to LOFAR Netherlands, the current best similar instrument in the world:
- 25% better resolution
- 8x more sensitive
- 135x the survey speed

[www.skatelescope.org]
Large Synoptic Survey Telescope (LSST)

- Image every 15 seconds
- 100PB over 10 years
Biology and Medical Research

- MRIs for Neuroscience (10s of TB per session)
- Genomics: Short Read Archive approaching 1PB
- Brain Imaging
- High-resolution microscopy
Large Numerical Simulations

- Millennium simulation: dark matter, 30TB raw data

[V. Springel et al., 2005]
Social Networks

- Facebook storing petabytes of data per year
- Twitter: 500,000,000 tweets (8 TB of data) per day
- Millions of users, different links
- Huge graphs
More Big Data Use Cases

- Government Operation (Census, Archives, Surveys)
- Commercial (Financial, Netflix, Web Search, Cargo Shipping, Materials Data)
- Defense (Geospatial Analysis, Object Identification)
- Health (Electronic Medical Records, Digital Pathology, Epidemiology, Biodiversity)
- Energy (Smart Grids)
Big Data or Small Data?

- http://smalldatatagroup.com
Jobs on a Large Analytics Cluster

![Graph showing the distribution of input job sizes for a large analytics cluster.](image)

[R. Appuswamy et al., 2013]
Big Data or Small Data?

- Many companies feel the need to overclaim the amount of data
- "when you take a normal tech company and sprinkle on data, you get the next Google" — [C. O'Neil]
- Many large datasets are not useful
- Twitter processes 8TB, but the tweets only take about 30GB…
- Wikipedia can be downloaded onto a USB drive
- All MP3s can be stored on a moderately sized disk array
- Can learn a lot from a "small" dataset, e.g. sensors from a single turbine, grocery store, Apple Watch
- Small data focused on end-user, more timely insights?
Big Data or Small Data

- Projects with big data concerns that require strategies and infrastructure
- Lots of small data waiting for analysis
Next Class

• Data Cleaning
  - Research paper
  - Volunteers to present?
• Reading Response: by 12pm noon before class
• Assignment 1 due by 11:59pm (after class)
• Topic interests
• Be prepared for quizzes about the readings