Big Data Analytics: What is Big Data?

H. Andrew Schwartz

CSE545
Spring 2020
Big Data, what is it?

data that will not fit in main memory.

traditional computer science
Big **Data**, what is it?

*data that will not fit in main memory.*

**SSD Sequential Read:**

~500 MB/s

For example...

- busy web server access logs
- graph of the entire Web
- all of Wikipedia
- daily satellite imagery over a year

traditional computer science
Big **Data**, what is it?

- data that will not fit in main memory.
- data with a *large* number of observations and/or features.
Big *Data*, what is it?

**Tall data:**
- edge list of a large graph
- rgb values per pixel location in large images

**Wide data:** mobile app usage statistics of 100 people
Big **Data**, what is it?

- data that will not fit in main memory.
- data with a *large* number of observations and/or features.
Big **Data**, what is it?

- **traditional computer science**
  - data that will not fit in main memory.
  - data with a *large* number of observations and/or features.
- **other fields**
  - non-traditional sample size (i.e. > 100 subjects); can’t analyze in stats tools (Excel).
- **statistics**
1. Survey of SDG-related Big Data projects

<table>
<thead>
<tr>
<th>Type of data source(s)</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobile phone data</td>
<td>23</td>
</tr>
<tr>
<td>Satellite imagery</td>
<td>20</td>
</tr>
<tr>
<td>Web data</td>
<td>12</td>
</tr>
<tr>
<td>Twitter data</td>
<td>12</td>
</tr>
<tr>
<td>Other social networks</td>
<td></td>
</tr>
<tr>
<td>Financial transaction data</td>
<td></td>
</tr>
<tr>
<td>Scanner data</td>
<td></td>
</tr>
<tr>
<td>Facebook data</td>
<td>11</td>
</tr>
<tr>
<td>Sensor data</td>
<td>11</td>
</tr>
<tr>
<td>Smart meter data</td>
<td>8</td>
</tr>
<tr>
<td>Health records</td>
<td>6</td>
</tr>
<tr>
<td>Ships identification data</td>
<td>5</td>
</tr>
<tr>
<td>Public transport usage data</td>
<td>2</td>
</tr>
<tr>
<td>Credit card data</td>
<td></td>
</tr>
</tbody>
</table>

- Mobile (23), Satellite imagery (20) and social media (12+12+6) are the most prominent sources.
Big Data, what is it? Industry View

**Figure 2: Sources of big data**
Which of the following do you consider part of big data (regardless of whether your company uses each)?

- Large data files (20 terabytes or larger) - 65%
- Advanced analytics or analysis - 63%
- Data from visualization tools - 50%
- Data from social networks - 48%
- Unstructured data (e.g., video, open text, voice) - 43%
- Geospatial/location information - 38%
- Social media/monitoring/mapping - 37%
- Telematics - 34%
- Unstructured data/lcg files/free text - 25%

Source: Accenture Big Success with Big Data Survey, April 2014
### Big Data, what is it? Industry View

**Figure 2: Sources of big data**

Which of the following do you consider part of big data (regardless of whether your company uses each)?

<table>
<thead>
<tr>
<th>Source</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large data files (20 terabytes or larger)</td>
<td>65%</td>
</tr>
<tr>
<td>Advanced analytics or analysis</td>
<td>60%</td>
</tr>
<tr>
<td>Data from visualization tools</td>
<td>50%</td>
</tr>
<tr>
<td>Data from social networks</td>
<td>48%</td>
</tr>
<tr>
<td>Unstructured data (e.g., video, open text, voice)</td>
<td>43%</td>
</tr>
<tr>
<td>Geospatial/location information</td>
<td>38%</td>
</tr>
<tr>
<td>Social media/monitoring/mapping</td>
<td>37%</td>
</tr>
<tr>
<td>Telematics</td>
<td>34%</td>
</tr>
<tr>
<td>Unstructured data/log files/free text</td>
<td>25%</td>
</tr>
</tbody>
</table>

Source: Accenture Big Success with Big Data Survey, April 2014
**Big Data, a type of analytics**

**Figure 2: Sources of big data**

Which of the following do you consider part of big data (regardless of whether your company uses each)?

<table>
<thead>
<tr>
<th>Source</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large data files (20 terabytes or larger)</td>
<td>65%</td>
</tr>
<tr>
<td>Advanced analytics or analysis</td>
<td>60%</td>
</tr>
<tr>
<td>Data from visualization tools</td>
<td>50%</td>
</tr>
<tr>
<td>Data from social networks</td>
<td>48%</td>
</tr>
<tr>
<td>Unstructured data (e.g., video, open text, voice)</td>
<td>43%</td>
</tr>
<tr>
<td>Geospatial/location information</td>
<td>38%</td>
</tr>
<tr>
<td>Social media/monitoring/mapping</td>
<td>37%</td>
</tr>
<tr>
<td>Telematics</td>
<td>34%</td>
</tr>
<tr>
<td>Unstructured data/log files/free text</td>
<td>25%</td>
</tr>
</tbody>
</table>

*Source: Accenture Big Success with Big Data Survey, April 2014*
Big Data, a type of analytics

Analyses which can handle the “3 Vs”:

1. Volume - large quantity
2. Velocity - arriving quickly
3. Variety - [un]structured, multi-modal
Big Data, a type of analytics

Analyses which can handle the “3Vs”:

- Volume
- Velocity
- Variety
- Veracity
Big Data, a type of analytics
Big Data, a type of analytics
The Big Data Challenge

We have now collected 250 terabytes of data about our customers and the software has analyzed the data.

Great! Big Data! What does the software tell us?

It says we have 250 terabytes of data.
Big Data, a buzz word?

(Gartner Hype Cycle)
Big Data, a buzz word?
Big Data, a buzz word?

(Gartner Hype Cycle)
Big Data, a buzz word?

(Gartner Hype Cycle)
Big Data, a buzz word?

- Realization of what subfields are really doing “big data” (i.e. data mining, ML, Statistics, computational social sciences).
- Best practices being established.

(Gartner Hype Cycle)
Big Data, in demand?

Figure 3: Main challenges with big data projects
What are the main challenges to implementing big data in your company?

- Security: 51%
- Budget: 47%
- Lack of talent to implement big data: 41%
- Lack of talent to run big data and analytics on an ongoing basis: 37%
- Integration with existing systems: 35%
- Procurement limitations on big data vendors: 33%
- Enterprise not ready for big data: 27%

Source: Accenture Big Success with Big Data Survey, April 2014
Big Data, in demand?

**Figure 6: Big data’s competitive significance**

- Big data will revolutionize the way we do business to a degree similar to the advent of the Internet in the 1990s:
  - Strongly Agree: 51%
  - Agree: 38%
  - Neither Agree nor Disagree: 10%
  - Disagree: 1%

- Big data will dramatically change the way we do business in the future:
  - Strongly Agree: 39%
  - Agree: 46%
  - Neither Agree nor Disagree: 13%
  - Disagree: 2%

- Companies that do not embrace big data will lose their competitive position and may even face extinction:
  - Strongly Agree: 37%
  - Agree: 42%
  - Neither Agree nor Disagree: 19%
  - Disagree: 2%

- We feel we are ahead of our peers in using big data and this creates a competitive advantage for us:
  - Strongly Agree: 37%
  - Agree: 46%
  - Neither Agree nor Disagree: 12%
  - Disagree: 4%

Source: Accenture Big Success with Big Data Survey, April 2014
Big Data, in demand?

Adoption of Big Data 2015-2018
(Copyright 2018 – Dresner Advisory Services)

Big Data, in demand?

By the requirements in job ads.
(Muenchen, 2019)
Big Data, What is it?

**Short Answer:**

*Big Data ≈ Data Mining ≈ Predictive Analytics ≈ Data Science*

(Leskovec et al., 2014)
Big Data, What is it?

Short Answer:

Big Data $\approx$ Data Mining $\approx$ Predictive Analytics $\approx$ Data Science

(Leskovec et al., 2014)

CSE545 focuses on:

- How to analyze data that is mostly too large for main memory.
- Analyses only possible with a large number of observations or features.
Big Data, What is it?

Goal: Generalizations
A *model* or *summarization* of the data.

- How to analyze data that is mostly too large for main memory.
- Analyses only possible with a *large* number of observations or features.
Big Data, What is it?

**Goal:** Generalizations
A *model* or *summarization* of the data.

E.g.
- Google’s PageRank: *summarizes* web pages by a single number.
- Twitter financial market predictions: *Models* the stock market according to shifts in sentiment in Twitter.
- Distinguish tissue type in medical images: *Summarizes* millions of pixels into clusters.
- Mental health diagnosis in social media: *Models* presence of diagnosis as a distribution (a summary) of linguistic patterns.
- Frequent co-occurring purchases: *Summarize* billions of purchases as items that frequently are bought together.
Big Data, What is it?

Goal: Generalizations
A model or summarization of the data.

1. Descriptive analytics
Describe (generalizes) the data itself

2. Predictive analytics
Create something generalizeable to new data
Big Data Analytics, The Class

Core Data Science Courses

CSE 519: Data Science Fundamentals
CSE 544: Prob/Stat for Data Scientists
**CSE 545: Big Data Analytics**
CSE 512: Machine Learning
CSE 537: Artificial Intelligence
CSE 548: Analysis of Algorithms
CSE 564: Visualization

Applications of Data Science

CSE 527:
  - Computer Vision
CSE 538:
  - Natural Language Processing
CSE 549:
  - Computational Biology
...
Big Data Analytics, The Class

Core Data Science Courses
CSE 519: Data Science Fundamentals
CSE 544: Prob/Stat for Data Scientists
**CSE 545: Big Data Analytics**
CSE 512: Machine Learning
CSE 537: Artificial Intelligence
CSE 548: Analysis of Algorithms
CSE 564: Visualization

Applications of Data Science
CSE 527:
Computer Vision
CSE 538:
Natural Language Processing
CSE 549:
Computational Biology
...

Key Distinction:
Focus on scalability and algorithms / analyses not possible without large data.
**Goal:** Generalizations

A *model* or *summarization* of the data.

- Data Frameworks
- Algorithms and Analyses
Big Data Analytics, The Class

**Goal:** Generalizations
A model or summarization of the data.

- Data Frameworks
  - Hadoop File System
  - Spark
  - MapReduce
  - Streaming
  - Tensorflow

- Algorithms and Analyses
Big Data Analytics, The Class

Goal: Generalizations
A model or summarization of the data.

Data Frameworks
- Hadoop File System
- Spark
- MapReduce
- Streaming
- Tensorflow

Algorithms and Analyses
- Similarity Search
- Linear Modeling
- Recommendation Systems
- Graph Analysis
- Deep Learning
Preliminaries

Ideas and methods that will repeatedly appear:

- Bonferroni's Principle
- Normalization (TF.IDF)
- Power Laws
- Hash functions
- IO Bounded (Secondary Storage)
- Unstructured Data

- Parallelism
- Functional Programming
Statistical Limits. Goal: Generalization

Bonferroni's Principle

A to consider goal of generalization:
   Find events that didn’t just happen by chance.
Statistical Limits.

Bonferroni's Principle; an example:
Statistical Limits.

Bonferroni's Principle; an example:
Statistical Limits.

Bonferroni's Principle

Goal: Generalization (i.e. not by chance)
Bonferroni’s Principle; Task Example

snazzyphones.com wants to know which case to eliminate.

6 total cases:

Red
Green
Blue
Teal
Purple
Yellow
Bonferroni’s Principle; Task Example

snazzyphones.com wants to know which case to eliminate.

6 total cases:  

<table>
<thead>
<tr>
<th>Color</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Green</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blue</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Purple</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yellow</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

first day, 17 sales:

What is the data telling you?
Bonferroni’s Principle; Task Example

snazzyphones.com wants to know which case to eliminate.

6 total cases:  

Red  
Green  
Blue  
Teal  
Purple  
Yellow  

first day, 17 sales:

１ ２ ３ ４ ５ ６ ７ ８ ９ 10 11 12 13 14 15 16 17

What is the data telling you?

The blue isn’t selling?
Statistical Limits.  

Goal: Generalization

Bonferroni's Principle

Roughly, calculating the probability of any of \( n \) findings being true requires \( n \) times the probability as testing for 1 finding.

https://xkcd.com/882/

In brief, one can only look for so many patterns (i.e. features) in the data before one finds something just by chance (i.e. finding something that does not generalize).

“Data mining” is a bad word in some communities!
**Statistical Limits.**  

**Goal:** **Generalization**

Note: *Bonferroni’s principle* is simply an abstract idea inspired by a precisely defined method of hypothesis testing called “Bonferroni correction”.

We will go over this correction method later. The *principle* is the more important idea to understand as a big data practitioner.

In brief, one can only look for so many patterns (i.e. features) in the data before one finds something just by chance (i.e. finding something that does **not** generalize).

“Data mining” is a bad word in some communities!
Normalizing

Count data often need *normalizing* -- putting the numbers on the same “scale”.

Prototypical example: TF.IDF
Normalizing

Count data often need \textit{normalizing} -- putting the numbers on the same “scale”.

Prototypical example: TF.IDF of word $i$ in document $j$:

Term Frequency:

$$t_f_{ij} = \frac{\text{count}_{ij}}{\max_k \text{count}_{kj}}$$

Inverse Document Frequency:

$$idf_i = \log_2\left(\frac{\text{docs}_*}{\text{docs}_i}\right) \propto \frac{1}{\text{docs}_i / \text{docs}_*}$$

$$t_f.idf_{ij} = t_f_{ij} \times idf_i$$

where docs is the number of documents containing word $i$. 
Normalizing

Count data often need *normalizing* -- putting the numbers on the same "scale".

Prototypical example: TF.IDF of word *i* in document *j*:

Term Frequency:

\[ t_f_{ij} = \frac{\text{count}_{ij}}{\max_k \text{count}_{kj}} \]

Inverse Document Frequency:

\[ \text{idf}_i = \log_2\left(\frac{\text{docs}_*}{\text{docs}_i}\right) \propto \frac{1}{\frac{\text{docs}_i}{\text{docs}_*}} \]

\[ t_f \cdot \text{idf}_{ij} = t_f_{ij} \times \text{idf}_i \]

where docs is the number of documents containing word *i*. 
Normalizing

**Standardize:** puts different sets of data (typically vectors or random variables) on the same scale with the same center.

- Subtract the mean (i.e. “mean center”)
- Divide by standard deviation

\[
Z_i = \frac{x_i - \bar{x}}{s_x}
\]
Power Law

Characterized many frequency patterns when ordered from most to least:

**County Populations** [r-bloggers.com]

# links into webpages [Broader et al., 2000]

Sales of products [see book]

**Frequency of words** [Wikipedia, “Zipf’s Law”]

(“popularity” based statistics, especially without limits)
Power Law

\[ \log y = b + a \log x \]

raising to the natural log:

\[ y = e^b e^{a \log x} = e^b x^a = cx^a \]

where \( c \) is just a constant

Characterizes “the Matthew Effect” -- the rich get richer
Power Law

message-level  user-level  county-level

Hash Functions and Indexes

Review:

\[ h: \text{hash-key} \rightarrow \text{bucket-number} \]

Objective: uniformly distribute hash-keys across buckets.

Example: storing word counts.
Hash Functions and Indexes

Review:

$h: \text{hash-key} \rightarrow \text{bucket-number}$

Objective: uniformly distribute hash-keys across buckets.

Example: storing word counts.

\[
h(\text{word}) = \left( \sum_{\text{char} \in \text{word}} \text{ascii(char)} \right) \% \#\text{buckets}
\]
Hash Functions and Indexes

Review:

\[ h: \text{hash-key} \rightarrow \text{bucket-number} \]

Objective: uniformly distribute hash-keys across buckets.

Example: storing word counts.

\[
h(word) = \left( \sum_{\text{char} \in \text{word}} \text{ascii(char)} \right) \mod \#\text{buckets}
\]

Data structures utilizing hash-tables (i.e. O(1) lookup; dictionaries, sets in python) are a friend of big data algorithms! Review further if needed.
Hash Functions and Indexes

Review:

\[ h: \text{hash-key} \rightarrow \text{bucket-number} \]

Objective: uniformly distribute hash-keys across buckets.

Example: storing word counts.

**Database Indexes:** Retrieve all records with a given value. (also review if unfamiliar / forgot)

Data structures utilizing hash-tables (i.e. \(O(1)\) lookup; dictionaries, sets in python) are a friend of big data algorithms! Review further if needed.
IO Bounded

Reading a word from disk versus main memory: $10^5$ slower!

Reading many contiguously stored words is faster per word, but fast modern disks still only reach 150MB/s for sequential reads.
IO Bounded

Reading a word from disk versus main memory: $10^5$ slower!

Reading many contiguously stored words is faster per word, but fast modern disks still only reach 150MB/s for sequential reads.

IO Bound: biggest performance bottleneck is reading / writing to disk.

(starts around 100 GBs; ~10 minutes just to read).
Data

Structured  Unstructured

- Unstructured ≈ requires processing to get what is of interest
- Feature extraction used to turn unstructured into structured
- Near infinite amounts of potential features in unstructured data
Data

Structured

mysql table  email header  satellite imagery  images
vectors matrices  facebook likes

Unstructured

• Unstructured ≈ requires processing to get what is of interest
• Feature extraction used to turn unstructured into structured
• Near infinite amounts of potential features in unstructured data