Theoretical Basics of Big Data Analytics and Real Time Computation Algorithms

Spring 2016

What is Big Data?
– Why is it new and important?
– What tools do we need to deal with it?

Big collections of data:
– Retail store inventory and transactions
– Cell company records
– Weather records

Databases: store and manage data
• Digitized data - routinely collected
  – Shopping transactions
  – Search queries
  – Internet traffic
  – Readings from sensors

Often tossed away or stored and never used

• Recently: something special
  – New kinds of information
  – Was not anticipated
  – Emerges when BIG

• Valuable hidden information
  – not visible
  – has to be extracted,
  – processed

• Big Data is about:
  – New information
  – Specific tools

Such definition is
  – is too vague
  – relies on the notion of information
What is Iron?
– Go to Iron age
– Ask Ironsmith
– Examples: raw, processing, things

Result: unsatisfactory
– No frame of reference
– Need to know: molecular structure...

Now we are in Information Age
– Know it exists
– Have examples
– Definitions in special cases

Entering Big Data (sub) Age
Examples of Big Data

“BIG DATA: A Revolution That Will Transform How We Live, Work, and Think”

Viktor Mayer-Schönberger and Kenneth Cukier

• **Target** - Detecting: a woman is pregnant
  – Two dozen products used as proxies
  – Estimate pregnancy stages, due date
  – Send relevant coupons

• **Correlation-based techniques**
  – Predict mechanical failures
  – Things break down *gradually*
  – Sensors + correlation analysis:
    * Whirling motor
    * Excessive heat
    * ...
• UPS

– Replacing parts: 2-3 years
  * Inefficient
– Predictive analysis
  * Monitoring individual parts
– Predictions made automatically
– Based on:
  * Great number of cases
  * Correlation analysis
  * No complex models

Modern cars
– Lots of sensors
  * Temperature, Vibrations, Voltages...
– Use complex models of prediction
– Information tossed away - no learning
– Imagine: transmitted, collected, and analyzed...
• H1N1 Virus 2009

– Only hope: to slow its spread
– Need to know: where it is
– US Center for Diseases Control (CDC):
  * Doctors: to inform of new flu cases
  * Week or two out of date
  * ⇒ Delays blinded health agencies

Few weeks before H1B1:

* Google: paper in “Nature”
* Predict spread if winter flu by looking what people were searching
* 3 Bil. queries a day
• Google “learning” technique
  – CDC data for 2003-2008
  – Correlations:
    * Search queries (50M most common)
    * Flu spread
  – Result:
    * Combination of 45 search terms
    * + Math model
    * = Strong correlation with official figures
  – So, in 2009 Google - more timely indicator

No need in:
* mouth swabs
* contacting doctors...
Instead: huge amount of data
* Too Big
* Too Noisy
• Unexpected data in existing collections:
  – Too Big
  – Too Noisy

• Arranging new studies
  Aspirin and orange juice vs Cancer

  – Standard way:
    * Specific tests
    * Time
    * Low confidence (small amount of data)

  – Big Data way:
    * Digitized med. records
    * Shopping transactions
    * Search queries
    * ... (lots of other data)
Other Big Data challenges

- **LHC Large Hadron Collider**
  - 150 Mil. sensors
  - 40 Mil. times per second
  - Only 0.001% saved
  - 25 PB in 2012 (1PB=1000TB)
  - If all recorded:
    * 150 Mil. PB/year
    * = 200 x all other sources in the world

- **Modern Aircraft**
  - 100,000 sensors
  - Only 3 GB in an hour flight
Seems not Big, but
- Monitoring in real time
- Combinations of readings
- In dynamics
- Need to make very fast predictions
⇒ Big Data challenge

- **Digitized Media Streaming**
  - Large volumes
  - But: Nothing is hidden
⇒ Not considered as Big Data

- **Information in Big Data**
  - Hidden
  - Requires special tools

Analogy:
- Rare mineral
- Nuclear fusion energy
Big Data Manipulations: Basic Steps

- **Extract** pieces of information (probably from distributed sources)
- **Unify** - transform to "canonical" form
  - Compact
  - Easy to handle
  - Contains sufficient information
- **Combine** pieces
- **Update** when new info arrives
- **Utilize** - Decision making
Simplest Example

$x$ - object of interest (unknown value)

Observations:

\[ y_i = x + \epsilon_i, \quad i = 1, \ldots, n \]

\(\epsilon_i\) - i.i.d. random values, \(\mathbb{E}\epsilon_i = 0\).

A good estimate of \(x\):

\[ \hat{x} = \frac{1}{n} \sum_{i=1}^{n} y_i \]

\[ \begin{align*}
    y_1 & \rightarrow \hat{x}_1 = y_1, \\
    y_2 & \rightarrow \hat{x}_2 = \frac{y_1 + y_2}{2}, \\
    \vdots & \rightarrow \hat{x}_n = \frac{y_1 + \cdots + y_n}{n}
\end{align*} \]
**Updating \( \hat{x} \)**

\[ n : \text{have } \hat{x}_n, \text{ get } y_{n+1} \]

\[
\hat{x}_{n+1} = \frac{1}{n+1} \sum_{i=1}^{n+1} y_i = \frac{1}{n+1} \left( \sum_{i=1}^{n} y_i + y_{n+1} \right)
\]

\[ = \frac{1}{n+1} (n\hat{x}_n + y_{n+1}) \]

or

\[
\hat{x}_{n+1} = \hat{x}_n + \frac{1}{n+1} (y_{n+1} - \hat{x}_n)
\]

In addition to \( \hat{x}_n \) need to keep \( n \).

"Explicit" form of information: \((n, \hat{x}_n)\).
Updating "Canonical" Information

\[(n, S) \Rightarrow \hat{x} = \frac{S}{n}\]

\[n - \text{number of readings}, \quad S = \sum_{i=1}^{n} y_i\]

\[y_1 \rightarrow (1, y_1)\]
\[y_2 \rightarrow \oplus \rightarrow (1+1, \frac{y_2+y_1}{n_2})\]
\[y_3 \rightarrow \oplus \rightarrow (2+1, \frac{s+y_3}{n_2})\]
\[\vdots\]
\[y_n \rightarrow \oplus \rightarrow (n, S_n)\]
\[y_{n+1} \rightarrow \oplus \rightarrow (n+1, S_n+y_{n+1})\]

\[\begin{align*}
\begin{bmatrix}
  y_1 \\
  \vdots \\
  y_n \\
  \vdots \\
  y_{2k}
\end{bmatrix}
\rightarrow (n, S_1) \quad \text{combine} \quad (n+k, S_1+S_2)
\end{align*}\]
Concurrent Combining

\[ y_1 \rightarrow (1, y_1) \rightarrow (2, s_1^{(1)}) \]
\[ y_2 \rightarrow (1, y_2) \rightarrow (2, s_2^{(1)}) \]
\[ y_3 \rightarrow (1, y_3) \rightarrow (2, s_3^{(1)}) \]
\[ y_4 \rightarrow (1, y_4) \rightarrow (2, s_4^{(1)}) \]
\[ \vdots \]
\[ y_n \rightarrow (1, y_n) \rightarrow (2, s_n^{(1)}) \]

\[ n = 2^k \]
\[ \kappa = \log_2 n \]

If \( n = 1000 \)
\[ \kappa = 10 \]

If \( n = 1M \)
\[ \kappa = 20 \]

If \( n = 1B \)
\[ \kappa = 30 \]
Precision of $\hat{x}$

\[ y_i = x + \varepsilon_i \]

$\varepsilon_i$ - i.i.d., $E\varepsilon_i = 0$

\[ \text{Var}(\varepsilon_i) = E(\varepsilon_i - E\varepsilon_i)^2 = E\varepsilon_i^2 = \sigma^2 \]

\[ \hat{x} = \frac{1}{n} \sum_{i=1}^{n} y_i \text{ - unbiased est. of } x, \text{ i.e. } E\hat{x} = x \]

\[ E\hat{x} = E \left( \frac{1}{n} \sum_{i=1}^{n} y_i \right) = \frac{1}{n} E \sum_{i=1}^{n} (x + \varepsilon_i) \]

\[ = \frac{1}{n} \sum_{i=1}^{n} x + \frac{1}{n} E \sum_{i=1}^{n} \varepsilon_i = x \]

\[ \sum_{i=1}^{n} \varepsilon_i = 0 \]

\[ \text{Var}(\hat{x}) = E(\hat{x} - x)^2 = E \left( \frac{1}{n} \sum_{i=1}^{n} (x + \varepsilon_i) - x \right)^2 \]

\[ = E \frac{1}{n^2} \left( \sum_{i=1}^{n} (x + \varepsilon_i - x) \right)^2 \]

\[ = \frac{1}{n^2} \sum_{i,j=1}^{n} \text{Var}(\varepsilon_i \varepsilon_j) \]

\[ = \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n} \]

\[ \sum_{i=1}^{n} \varepsilon_i = 0 \]

\[ \sum_{i=1}^{n} \varepsilon_i^2 = \text{ind} \Rightarrow E\varepsilon_i \varepsilon_j = \frac{E\varepsilon_i \cdot E\varepsilon_j}{\sqrt{n}} \quad \text{if } i \neq j \]

\[ E\varepsilon_i \varepsilon_j = 0 \text{ if } i \neq j \]
\[ \text{Var}(\hat{x}) = \frac{\sigma^2}{n} \rightarrow 0 \text{ as } n \rightarrow \infty. \]

If we collect canonical info \((n, S)\):

\[(n, S) \Rightarrow \hat{x} = \frac{S}{n}, \quad \text{Var}(\hat{x}) = \frac{\sigma^2}{n}.\]

\((n, S)\) is sufficient to obtain \(\hat{x}\) and its variance, but only when \(\sigma^2\) is known.

**Suppose \(\sigma^2\) is not known**

\[\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \hat{x})^2\]

- Unbiased estimate of \(\sigma^2\).

\[\sum_{i=1}^{n} (y_i - \hat{x})^2 = \sum_{i=1}^{n} y_i^2 - 2 \sum_{i=1}^{n} y_i \cdot \hat{x} + n\hat{x}^2\]

\[= T - 2S\frac{S}{n} + n \left(\frac{S}{n}\right)^2 = T - \frac{S^2}{n}\]

\[T = \sum_{i=1}^{n} y_i^2\]
New canonical information $(n, S, T)$:

\[ n = \sum_{i=1}^{n} y_i^0, \quad S = \sum_{i=1}^{n} y_i^1, \quad T = \sum_{i=1}^{n} y_i^2 \]

\[ \hat{\sigma}^2 = \frac{1}{n-1} \left( T - \frac{S^2}{n} \right) \]

Estimate of the variance of $\hat{x}$:

\[ V = \text{Var}(\hat{x}) = \frac{\hat{\sigma}^2}{n} = \frac{1}{n(n-1)} \left( T - \frac{S^2}{n} \right) \]

$(n, S, T) \Rightarrow \hat{x} = \frac{S}{n'}$, \quad $V = \frac{1}{n(n-1)} \left( T - \frac{S^2}{n} \right)$

Updating can. info:

$(n, S, T) \xrightarrow{y} (n+1, S+y, T+y^2)$

Combining can. info:

$(n_1, S_1, T_1) \xrightarrow{y} (n_1+n_2, S_1+S_2, T_1+T_2)$

$(n_2, S_2, T_2)$
Info in explicit form \((n, \hat{x}, V)\):

Have \((n, \hat{x}_n, V_n)\), receive \(y_{n+1}\)

\[
\hat{x}_{n+1} = \hat{x}_n + \frac{y_{n+1} - \hat{x}_n}{n + 1}
\]

\[
\sigma^2_{n+1} = \frac{n-1}{n}\sigma^2_n + \frac{(y_{n+1} - \hat{x}_n)^2}{n + 1} V_n
\]

\[
V_{n+1} = \frac{\sigma^2_{n+1}}{n + 1} = \frac{n-1}{n + 1} V_n + \left(\frac{y_{n+1} - \hat{x}_n}{n + 1}\right)^2
\]

Updating explicit info:

\((n, \hat{x}, V)\) \rightarrow \( (n+1, \hat{x}_n + \frac{y_{n+1} - \hat{x}_n}{n + 1}, V_{n+1}) \)

Explicit form for single observation:

\[
y \rightarrow (1, y, \frac{?}{?})
\]

\[
V = \frac{1}{n(n-1)} \sum_{i=1}^{n} (y_i - \hat{x})^2
\]

⇒ Information in explicit form may not exist
Center of a set of points $x_1, \ldots, x_n$

- Center

\[
\sum_{i=1}^{n} |x - x_i| \sim \min_x \\
\sum_{i=1}^{n} (x - x_i)^2 \sim \min_x \\
\max_{i=1, \ldots, n} (x - x_i) \sim \min_x
\]
a) median

\[ \text{even} \quad \text{not unique} \]

\[ \sum_{i=1}^{5} (x - x_i)^2 = \sum_{i=1}^{5} (x - \bar{x} + (x - x_i))^2 = \]

\[ = n (x - \bar{x})^2 + 2 \sum_{i=1}^{5} (x - \bar{x})(x - x_i) \]

\[ + \sum_{i=1}^{5} (x - x_i)^2 \]

\[ = \text{const} \]

\[ = n (x - \bar{x})^2 + \sum_{i=1}^{5} (x - x_i)^2 \]

\[ \implies \text{center is mean} \]
\[
\min_{i} x_i \leq x \leq \max_{i} x_i
\]

\[
\min_{i} \max_{i} |x - x_i|
\]

Call it "middle"