**PS3**

sexier title?

Phil Simon, 8/2/2013
Hi. I’m Phil.
Who am I?

• Award-winning author of six books, most recently *The Visual Organization*

• Speaker, consultant, and technology expert

• Huge *Breaking Bad* fan (more on that later)
Doctors vs. Pilots

Why do pilots always listen to data but doctors so often ignore it?
Just What is Big Data, Anyway?

“Success begins with a common understanding of terms.”

—Probably Winston Churchill or someone else really smart
Big Data: Major Characteristics

- The 3 v’s
- Mostly external to the enterprise
- Mostly unstructured data
Some Major Types of Data

- Structured
- Semi-Structured
- Unstructured
- Metadata
# Structured Data

<table>
<thead>
<tr>
<th>EmployeeID</th>
<th>Last_Name</th>
<th>First_Name</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>1234</td>
<td>Spacey</td>
<td>Kevin</td>
<td>$123,123.00</td>
</tr>
<tr>
<td>23411</td>
<td>Peart</td>
<td>Neil</td>
<td>$121,111.00</td>
</tr>
<tr>
<td>778</td>
<td>Lee</td>
<td>Geddy</td>
<td>$78,000.00</td>
</tr>
<tr>
<td>445</td>
<td>Lifeson</td>
<td>Alex</td>
<td>$89,000.00</td>
</tr>
</tbody>
</table>
Semi-Structured Data
Unstructured Data
The information about the package is just as important as the package itself.

--Fred Smith, Founder and CEO of FedEx, 1978
Big Data: Major Characteristics

- Increasingly generated by machines, not people
- Inherently “unmanageable” in the traditional sense
- Doesn’t play nice with SQL
The era of Big Data does not mean that Small Data has become irrelevant.

Rather, Big Data extends the power of Small Data.
The Ability to Know Why

• Determine why things happen:
  • For example, why do library patrons rent certain types of books, CDs, DVDs, periodicals, etc.?
  • Why do they ignore others?
  • What about segmentation
New data sources create new opportunities and exponential value

We can ask better questions of our data
allen - see here.
Phil Simon, 8/2/2013
Big Data Tools

- Different statistical packages (R)
- Hadoop
- NoSQL/NewSQL
- Sentiment Analysis
- Text Analytics/Text Mining
Big Data Tools

• Natural Language Processing
• A/B Testing
• Predictive Analytics

Source: The Visual Thesaurus
Big Data Tools

- Data Visualization
  - Netflix
  - Twitter
  - Autodesk
Netflix Stats

- Roughly 40M customers
- Nearly $27B market cap
- Responsible for nearly one-third of all US weeknight Internet traffic

Data as of March 3, 2014
Netflix Facts

- Single biggest AWS customer
  - 2012 Christmas day outage
- In September of 2013, Netflix became the first non-TV network to win an Emmy for *House of Cards*
The Netflix Data Credo - 1

• Data should be accessible, easy to discover, and easy to process for everyone.

Source: Netflix - tinyurl.com/tvo-netflix
Computers and algorithms only get you so far...
Augmenting Existing Data
Move Well Beyond Simple Genres

- Comedies
- Dramas
- Westerns
- Documentaries
77,000 Subgenres of Movies

- Dark Suspenseful Sci-Fi Horror Movies
- Gritty Suspenseful Revenge Westerns
- Romantic Indian Crime Dramas
- Evil Kid Horror Movies
- Visually-Striking Goofy Action & Adventure Films
- Violent Suspenseful Action & Adventure Films from the 1980s

Source: The Hollywood Reporter
The Netflix Data Credo - 2

- The longer you take to find the data, the less valuable it becomes.
Stats

50,000 Netflix subscribers watched all 13 episodes of Season 4 of *Breaking Bad* the day before Season 5 premiered.

Source: *The Hollywood Reporter*
The Netflix Data Credo - 3

• Whether a dataset is large or small, being able to visualize it makes it easier to explain.
Figure 3.1 Detailed Color Comparison of House of Cards and Macbeth
Source: Netflix Technology Blog (techblog.netflix.com)
Figure 3.2 Detailed Color Comparison of *Hemlock Grove*, *House of Cards*, and *Arrested Development*

Source: Netflix Technology Blog (techblog.netflix.com)
What does Netflix know about each of its 40M streaming customers?

- What they watch
- When they watch
- The device on which they’re watching
- When they pause and/or resume watching
Twitter Facts

- 241M users; roughly 20% live in the U.S.
- Collectively send more than 400M tweets/day
- Market cap: $27B
Twitter Facts

• Helped cause the overthrow longstanding dictatorships (Arab Spring)
• Revolutionized global communications
Each Tweet Contains

- 31 pieces of metadata associated with it
  - Time
  - Date
  - Location
  - Hashtags
Each Tweet Contains

- 31 pieces of metadata associated with it
  - withheld_copyright
  - withheld_in_countries
  - possibly_sensitive
Global Ukraine-Based Tweets

Source: @TwitterData 02.25.2014
Current Record: Highest Number of Retweets
Most Followed on Twitter?
Interesting Twitter Facts

• Responsible for 30% of total global social sharing (Source: Media Bistro, 12/13)
• 30% of airtime tweets are sent during commercials (Source: Media Bistro, 10/13)
• New infrastructure can support 18 quintillion accounts (Source: VentureBeat, 10/13)

For more, see:  http://tinyurl.com/phil-tw2
A Week in the Life of Twitter

Source: Santiago Ortiz - @moebio
Recommendations on Getting Started
Big Data: Human Considerations

- Threatens the status quo and takes people out of their comfort zones
- Often meets with resistance—and lots of it
- Essential to the future of business
Other Considerations

- Privacy
- Security
- General creepiness
Recommendations on Getting Started

- Embrace data discovery
- Think incrementally and add new types of data
- Don’t try to boil the ocean
- Think description first, prediction later
Recommendations on Getting Started

- Manage expectations, under-promise, and over-deliver
- Build internal momentum
  - Aim for little victories—and the communicate them
  - Make the skeptics and dataphiles come to you
Recommendations on Getting Started

- Reject “set it and forget it.”
- Look outside of the organization
- Lead, follow, or get out of the way
- Be transparent
Recommendations on Getting Started

- Describe first
- Predict later
Recommendations on Getting Started

- It’s a marathon...
- not a sprint
Connect with me

www.philsimon.com
@philsimon
Questions?