X-Informatics Introduction: What is Big Data, Data Analytics and X-Informatics? (Continued) Physics Use Case (Start)

January 16 2013

Geoffrey Fox
gcf@indiana.edu

http://www.infomall.org/X-InformaticsSpring2013/index.html

Associate Dean for Research and Graduate Studies, School of Informatics and Computing
Indiana University Bloomington

2013

ONE SIZE DOESN'T FIT ALL

- There isn't one solution for driving an organization with big data:
  - **Hadoop** is for:
    - Engineers, batch (asynchronous), map reduce (divide and conquer), unstructured, flexible problems
  - **HBase** is for:
    - Engineers, real-time, large data blob, unstructured, key lookup, flexible problems
  - **Teradata** (or another data warehousing solution) is for:
    - Analysts, real-time or batch, structured, flexible problems
  - **Cassandra** (or **MongoDB** or …) is for:
    - Engineers, real-time, smaller data blob, unstructured, key lookup, flexible problems
  - Some problems warrant specialized solutions

Data Science Process (Continued)

The Rise of the Data Scientist

Hybrids	• Half analytical, with modeling, statistics, and experimentation skills • Half focused on data management – extraction, filtering, sampling, structuring • Lots of programming skills – Python, Ruby, Hadoop, Pig, Hive
Scientific	• Experimental physicists • Computational biologists • Statisticians with dirty hands • Ecologists, anthropologists, psychologists, etc.
Impatient	• Try something and iterate • Don't wait for a data person to get your data • "We're a pain in the ass" • Job tenure is short
Ground-breaking	• "Nobody's ever done this before" • "If we wanted to deal with structured data, we'd be on Wall Street" • "Being a consultant is the dead zone – too hard to get things implemented" • "The output should be a product or a demo – not a report"

Tom Davenport Harvard Business School http://fisheritcenter.haas.berkeley.edu/Big_Data/index.html Nov 2012

Is Davenport Correct?

- There are the 1.5 million decision makers/ managers of McKinsey report
- Up to 190,000 “nerds”
- Davenport appears to describe nerds not the larger 1.5M body of “generalists”

Jeff Hammerbacher's Process

1) Identify problem
2) Instrument data sources
3) Collect data
4) Prepare data (integrate, transform, clean, impute, filter, aggregate)
5) Build model
6) Evaluate model
7) Communicate results

20120117berkeley1.pdf Jeff Hammerbacher

Another Jeff Hammerbacher Process

1) Obtain
2) Scrub
3) Explore
4) Model
5) Interpret

20120117berkeley1.pdf Jeff Hammerbacher

Statistician Colin Mallows

1) Identify data to collect and its relevance to your problem
2) Statistical specification of the problem
3) Method selection
4) Analysis of method
5) Interpret results for non-statisticians

Ben Fry Data Visualization
http://en.wikipedia.org/wiki/Benjamin_Fry

1) Acquire
2) Parse
3) Filter
4) Mine
5) Represent
6) Refine
7) Interact