Introduction to Big Data

Professor Kartik Hosanagar

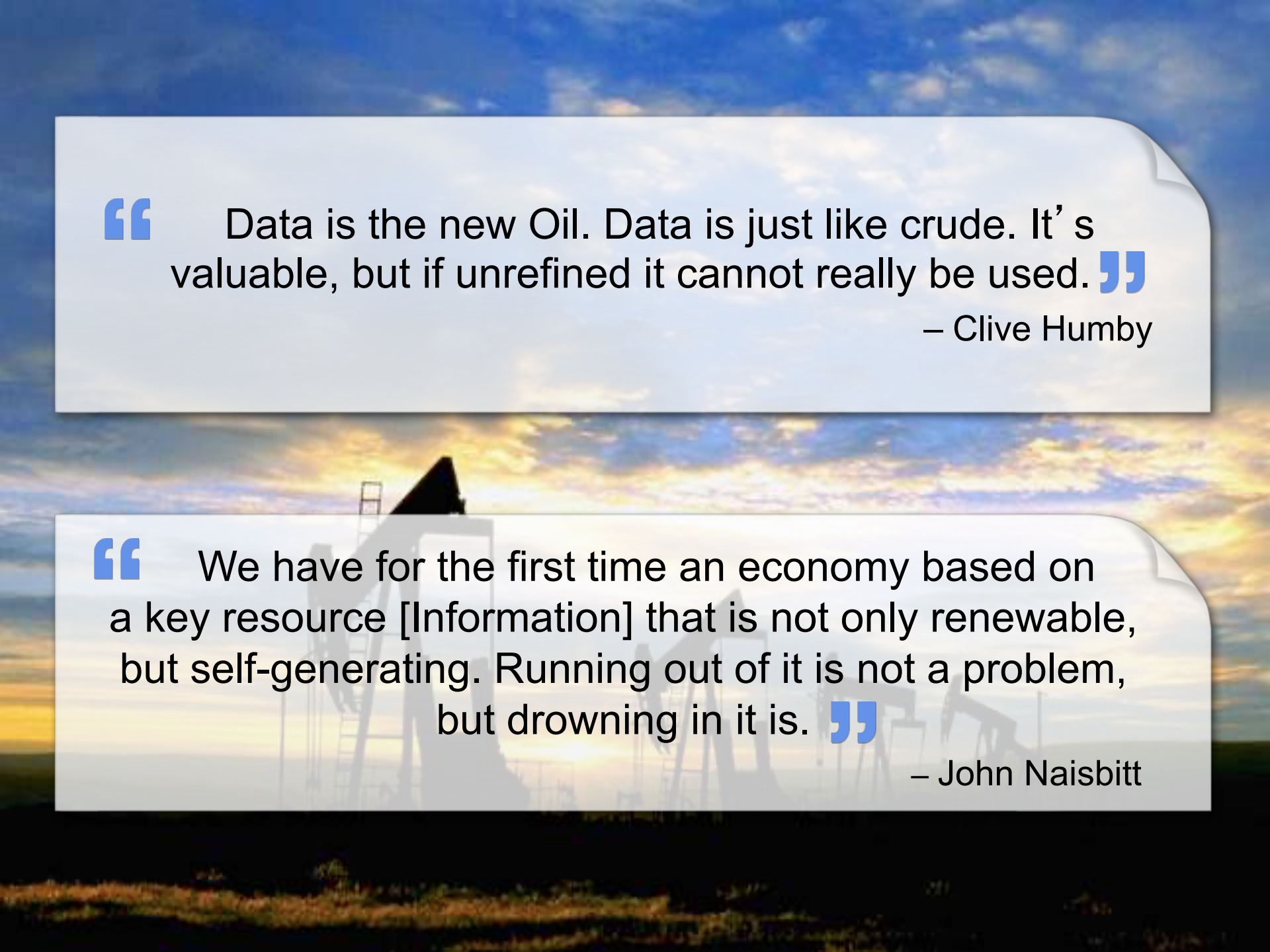
Agenda for Module 1

- **Big data overview**
  - What big data is, how it is being generated, and why it matters
- **Big data skills**
  - Approach to analysis, analytics competencies, and broad skillset needs
- **Big data tools (interview with guest speaker)**
  - Data management tools and data analysis tools
- **Extracting intelligence from big data**
  - Predictive analytics and implications for business strategy

Wharton
University of Pennsylvania

Big Data Overview

- What is big data?
- How is big data being generated?
- Why does big data matter?


“ Data is the new Oil. Data is just like crude. It’s valuable, but if unrefined it cannot really be used. ”
– Clive Humby
“ We have for the first time an economy based on a key resource [Information] that is not only renewable, but self-generating. Running out of it is not a problem, but drowning in it is. ”
– John Naisbitt

What is Big Data?

- Data that “exceeds the capacity or capability of…conventional methods and systems” (National Institute of Standards and Technology).

- But big data is not only about volume of data, it is also about:
  - the structure of the data set
  - the speed at which it is created
  - the tools you need to analyze it
  - what you can do with the data set

Quote from: https://www.technologyreview.com/s/519851/the-big-data-conundrum-how-to-define-it/
Additional content from: “Big Data” by Victor Mayer- Schonberger and Kenneth Cukier and https://www.sas.com/en_us/insights/big-data/what-is-big-data.html

Wharton
University of Pennsylvania

Big Data Characteristics

- New data characteristics created by today's digitized marketplace:

"Terabytes to petabytes of existing data to process."

"Structured, unstructured, text, multimedia."

"Streaming data, milliseconds to seconds to respond."

"Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations."

Image from: https://www.cbpr.me/big-data-and-public-relations/4vs/
Additional quotes from another diagram found at https://www.datasciencecentral.com/profiles/blogs/data-veracity

Drivers of Big Data

- **Computing capacity**: the capacity to store data has increased and the associated cost has decreased.
- **Data generation**: the world is going digital- more people and things are connected than ever before.

Content/quotes from: “Big Data” by Victor Mayer- Schonberger and Kenneth Cukier

What does Big Data Change?

- Big data allows you to:

1. Ask new questions

2. Answer same questions better

- This can be done across industries
  - Healthcare, Education, Transportation, and more

Wharton
University of Pennsylvania

Big Data in Healthcare

Big Data

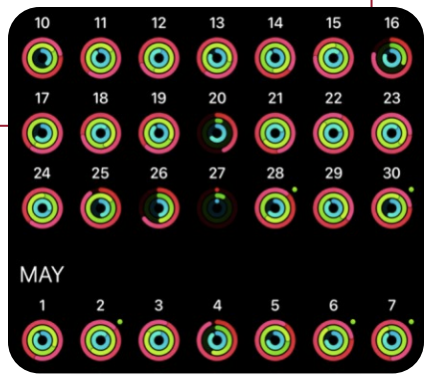
- Data about heart rate, sleep quality, exercise and more

- Gathered continuously and available in real time

Availability

Result

- Improves tracking of health patterns

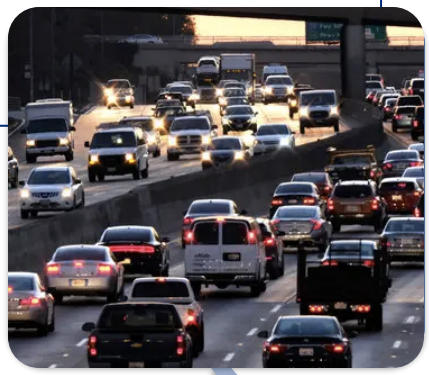


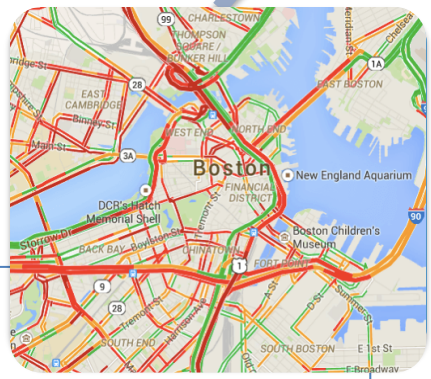
Content/quotes from: https://intellipaat.com/blog/7-big-data-examples-application-of-big-data-in-real-life/

Big Data in Transportation

Big Data

- Data about traffic, road closures, accidents, etc.



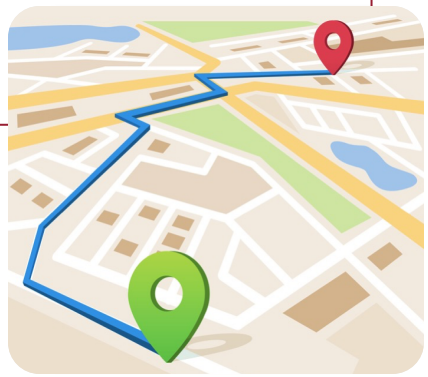


Availability

- Available in a visual format in real time

Result

- Allows for better route planning and scheduling



Content/quotes from: https://intellipaat.com/blog/7-big-data-examples-application-of-big-data-in-real-life/

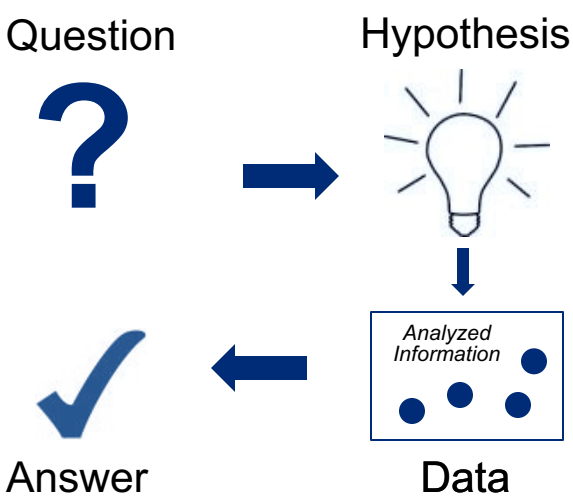
Big Data Skills

- Big data approach to analysis
- Big data skillset
- Choosing a big data tool

Big Data Approach to Analysis

Traditional Analytics

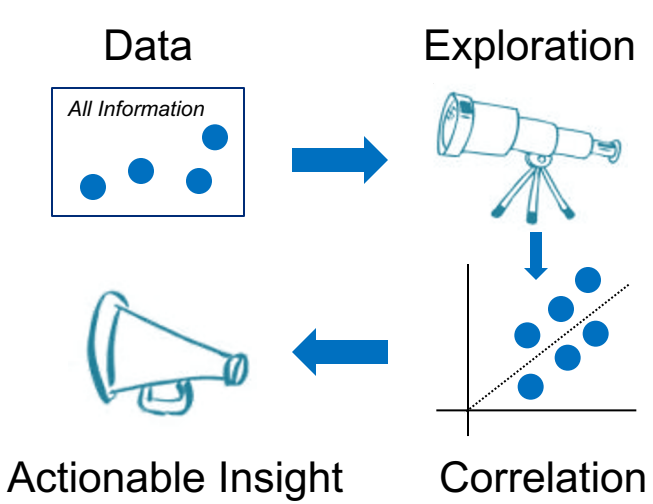
Structured & Repeatable



Start with hypothesis
Test against selected data

Big Data Analytics

Iterative & Exploratory



Data leads the way
Explore *all* data, identify correlations

Source unknown

Big Data Requires a Broad Skillset

Manage the data

Tool Developers:
Data Experts: Data architects, governance, policy

Understand the data

Data Science: statistics, computer science

Visualization Expertise: Interpret data, graph them in meaningful ways

Act on the data

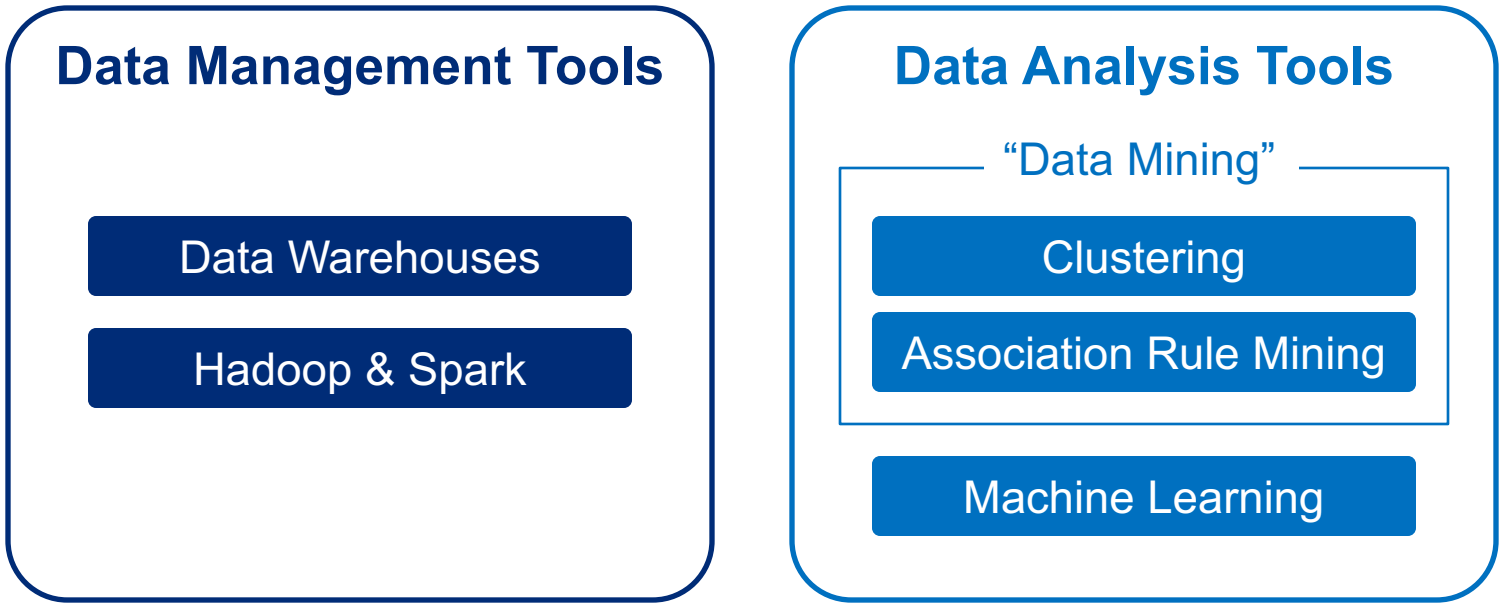
Decision Making- Exec. & Management: Apply data to solve business issues

Industry Vertical Domain Expertise: Identify relevant business issues, ask the right questions

Wharton
University of Pennsylvania

Choosing a Big Data Tool

- Two broad categories of big data tools to choose from, depending on whether you are trying to manage the data or analyze it.



Data Analysis: Extracting Intelligence from Big Data

- Data Mining
- Predictive Analytics
- Implications for Business Strategy

Data Mining

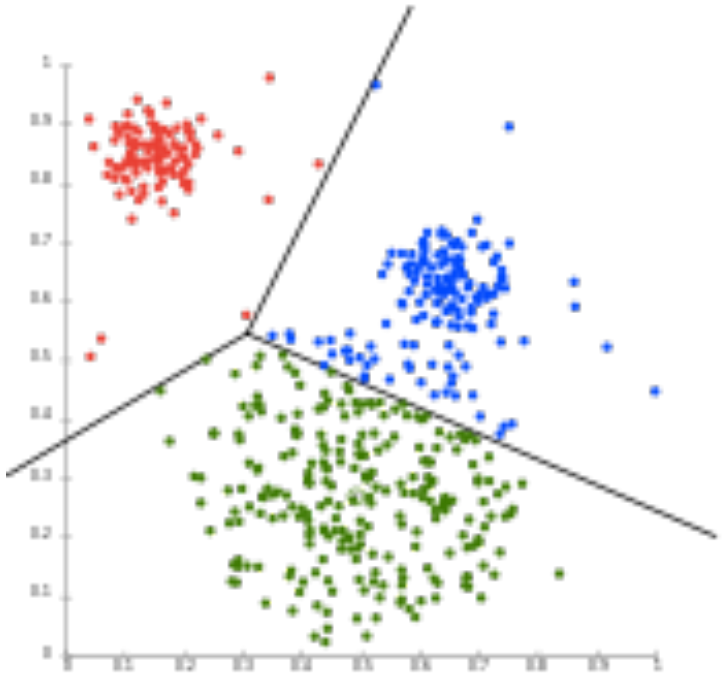
- Term encompassing tools for discovering patterns in large datasets
- Data mining vs. regressions?
  - Representation
    - Risk = 0.93*prior_default + 0.23*nun_cards – 1.3*employed – 0.734
- How did the technique find the relevant variables and coefficients?
  - Regression: Analyst had a hypothesis
  - Datamining: data-driven exploration
- Data mining techniques/ tools
  - Clustering
  - Association rule mining

Wharton
University of Pennsylvania

Clustering

- Clustering: Grouping data such that intra-group similarity is maximized while inter-group similarity is minimized

- Application: Data-driven customer segmentation



Wharton
University of Pennsylvania

Association Rule Mining

- Association rule mining: Finding common co-occurrences in data
  - Basket analysis:
    - {bread, butter} → {milk}
  - Medical diagnosis: Fish oil and Raynaud's disease
    - "Local increase of blood viscosity during cold-induced Raynaud's phenomenon"
    - "Reduction in blood viscosity by eicosapentaenoic acid"

Wharton
University of Pennsylvania

From Description to Prediction

- Data mining (including clustering and association rule mining) is ultimately about discovering & describing what patterns exist in data
- The next step, and the core business opportunity, involves using data to make predictions about the future, called predictive analytics.

> ***"Prediction is the process of filling in missing information. Prediction takes information you have, often called "data," and uses it to generate information you don't have."***

- Predictive analytics can significantly impact a range of industries and be used for a variety purposes within a single industry or company
  - Ex: Used for both recommendations & fraud detection on Amazon.com

Wharton
University of Pennsylvania

Content/quotes from: "Prediction Machines" by Ajay Agrawal, Joshua Gans, and Avi Goldfarb

Predictive Analytics: Amazon Recommendations

Visit website → **Shop and place in cart** → **Pay for items** → **Items shipped**

↑

Amazon's recommendations

amazon.com

Recommended for You

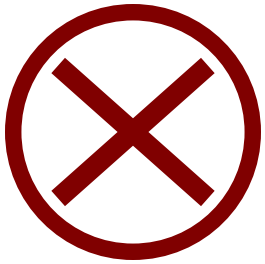
Amazon.com has new recommendations for you based on items you purchased or told us you own.

Content/quotes from: "Prediction Machines" by Ajay Agrawal, Joshua Gans, and Avi Goldfarb

Predictive Analytics: Amazon Payments

Items purchased*

Algorithm predicts fraud has occurred



Payment method rejected

* Past data about customer behavior on the site, coupled with knowledge of which previous purchases were fraudulent, helps train the system to make real-time fraud predictions based on customer behavior. This will be discussed further in module 3.

Content/quotes from: https://stripe.com/radar/guide

Implications of Big Data & Predictive Analytics

- Big data & predictive analytics have direct benefits for Amazon
  - Helpful recommendations → more shoppers and purchases
  - Accurate fraud predictions → fewer losses & happier customers
- Perhaps most importantly, prediction can be a virtuous cycle
  - Better predictions → more shoppers → more data → better predictions

- Next Class: Deep dive into the machine learning techniques for predictive analytics

Content/quotes from: "Prediction Machines" by Ajay Agrawal, Joshua Gans, and Avi Goldfarb


Wharton
University of Pennsylvania