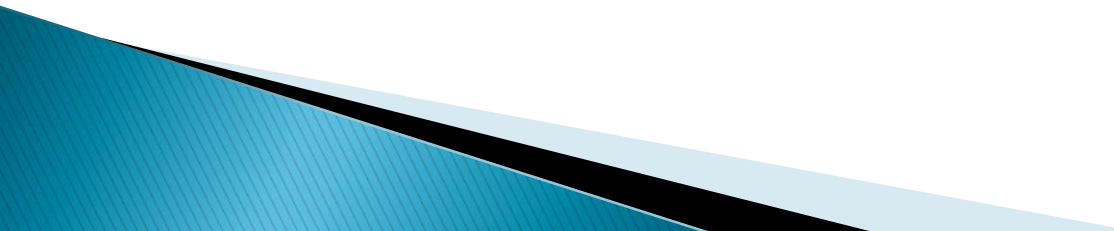


Big Data and Data Mining



Big Data

Big data -

- ▶ Extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations
 - ▶ Data sets larger than commonly used software tools can capture, store, manage and process
 - ▶ Set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale
- 

Big Data

In 2010, then-Google CEO Eric Schmidt said people currently create as much data very two days as was previously created in all of history up to 2003.

Eric Schmidt
speaking at Techonomy Conference

Big Data

Big data is at the foundation of all of the megatrends that are happening today, from social to mobile to the cloud to gaming.”

Chris Lynch, ex-Vertica CEO

Quotes

“Data really powers everything that we do.”

Jeff Weiner, LinkedIn

“Data is the new oil!”

Clive Humby, dunnhumby

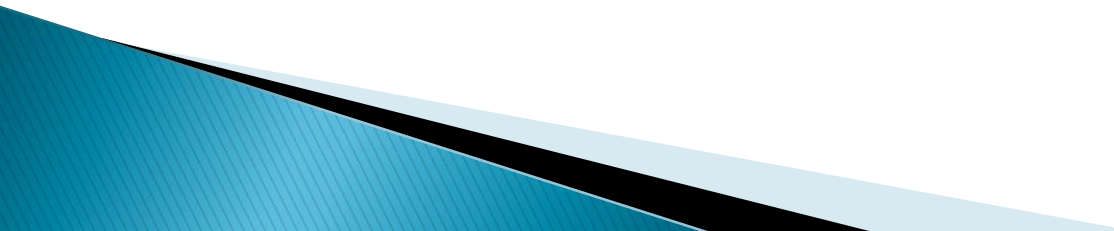
“Data is the new oil? No: Data is the new soil.”

David McCandless



Characteristics of Big Data

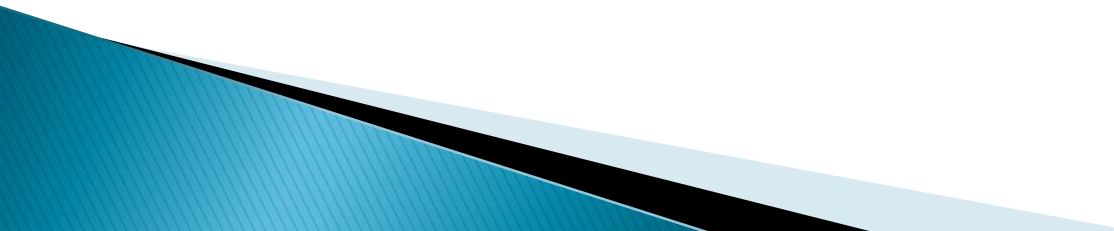
Characteristics:

- ▶ Volume – quantity
 - ▶ Variety – may come from multiple sources
 - ▶ Velocity – speed of generation
 - ▶ Variability – inconsistency of data
 - ▶ Veracity – data quality can vary greatly
 - ▶ Complexity – due to all the above
- 

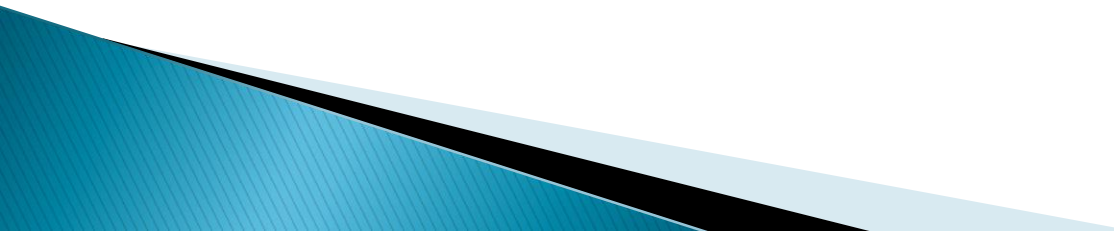
H1N1 Flu

- ▶ 2009 H1N1 flu virus outbreak
- ▶ Center for Disease Control and Prevention (CDC) requires doctors to report new flu cases
- ▶ Useful information but 2 week delay
- ▶ Google took 50 million most common search terms, compared with a CDC list of terms, and processed 450 million different mathematical models to test search terms, comparing their predictions against actual 2007 & 2008 flu cases.
- ▶ Found a combination of 45 search terms that had a strong correlation with incidences of the flu to identify the spread of flu in real time.

Amazon

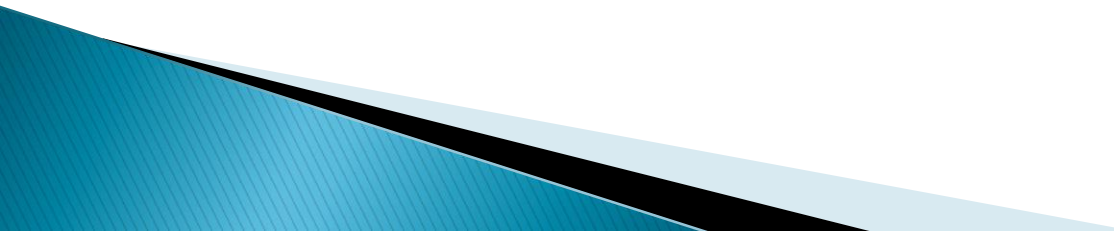
- ▶ CEO & founder Jeff Bezos wanted the company to make recommendations of books to customers
 - ▶ Amazon had lots of data, who bought what, who looked at what and didn't buy, etc.
 - ▶ Employees analyzed data to find similarities in customers and make recommendations
 - ▶ Editorial staff wrote reviews, recommendations, created bestseller lists
- 

Amazon – different way

- ▶ Greg Linden suggests, don't compare customers, find associations between products using item-to-item collaborative filtering (finding correlations)
 - ▶ Generated significantly more sales
 - ▶ Don't know **why** people buy, just know **what** (don't know **causation**, just know **correlation**)
- 

WalMart

- ▶ LOTS of data
 - ▶ Can hypothesize that prior to a hurricane sales of flashlights increase
 - ▶ Found also sales of Pop-Tarts increase

 - ▶ As storm approaches, stock boxes of Pop-Tarts at the front
- 

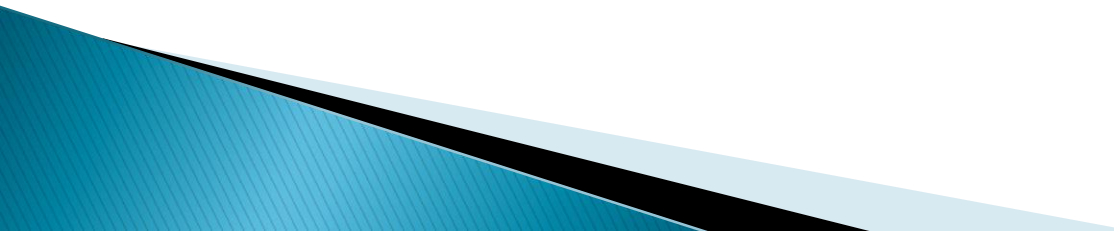
Aviva, insurance firm

- ▶ Want to identify those who might be at higher risk of illnesses such as high blood pressure, diabetes or depression
- ▶ Collecting blood and urine samples predicts fairly well and costs around \$125/person
- ▶ Found can predict from lifestyle data including hobbies, websites visited, amount of TV watched, income level for around \$5/person

So successful that other insurance agencies s let clients opt-in to sharing their lifestyle information

Data Mining

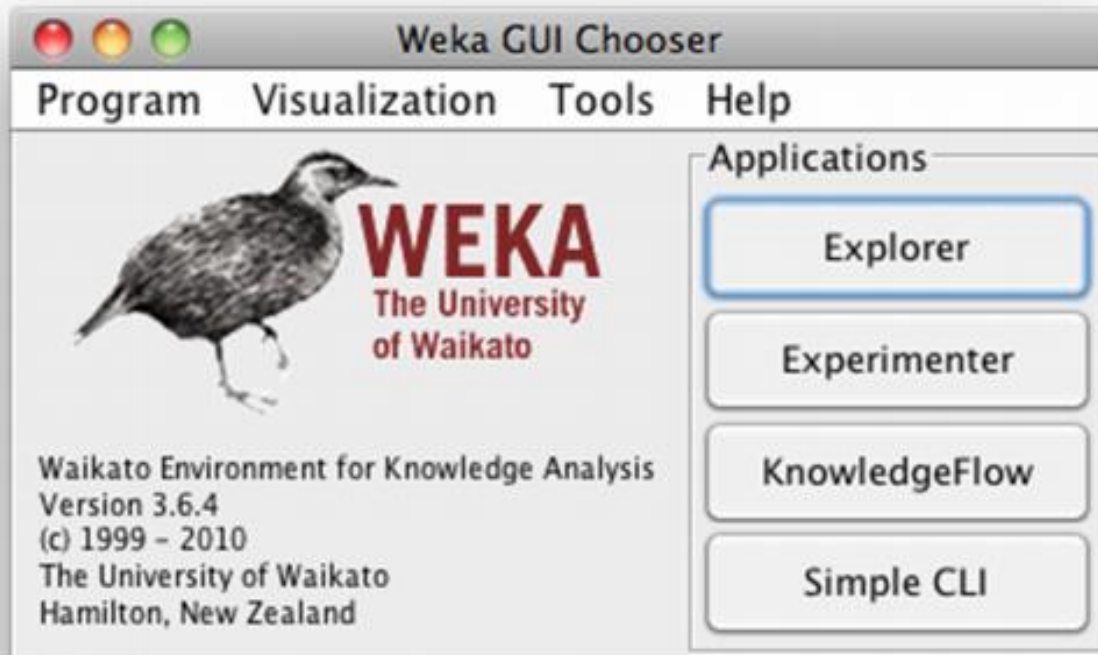
Data Mining

- ▶ process of discovering patterns in data
 - ▶ Techniques for finding and describing structural patterns in data, as a tool for helping to explain that data and make predictions from it
 - ▶ Going from data to information
- 

Correlation not Causation

Historical	Data Mining
Theorize about causation Hypothesize Verify hypothesis via statistical tests	Find correlations to learn Test that the correlations hold
Don't search for correlations since data sets are small and they'll be no way to validate the data	Aren't finding causations, finding correlations

Weka



Performance comparisons

Graphical interface

Command-line interface

Weka

- ▶ Weka contains machine learning algorithms for data mining tasks:
 - 100+ algorithms for classification
 - 75 for data preprocessing
 - 25 to assist with feature selection
 - 20 for clustering, finding association rules, etc.

Weka

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

