Introduction to Big Data Science

01st Period

An Overview of Big Data

Contents

- What is Big Data?
- Why Big Data?
- Potential Applications
- Examples
- Data Science Process

What is Big Data?

Dataset which is too large for traditional data processing systems.

4Vs: three main differences from `data` or `data analytics`.

- Volume

  As of 2012, about 2.5 exabytes (2.5 billion gigabytes) of data are created each day. The number is doubling about every 40 years. (Harvard Business Review Oct. 2012)

  90% of the data in the world today has been created in the last two years. (IBM ``What is big data?'')

- Velocity

  Real-time information makes it possible for a company to be more agile than its competitors.

- Variety

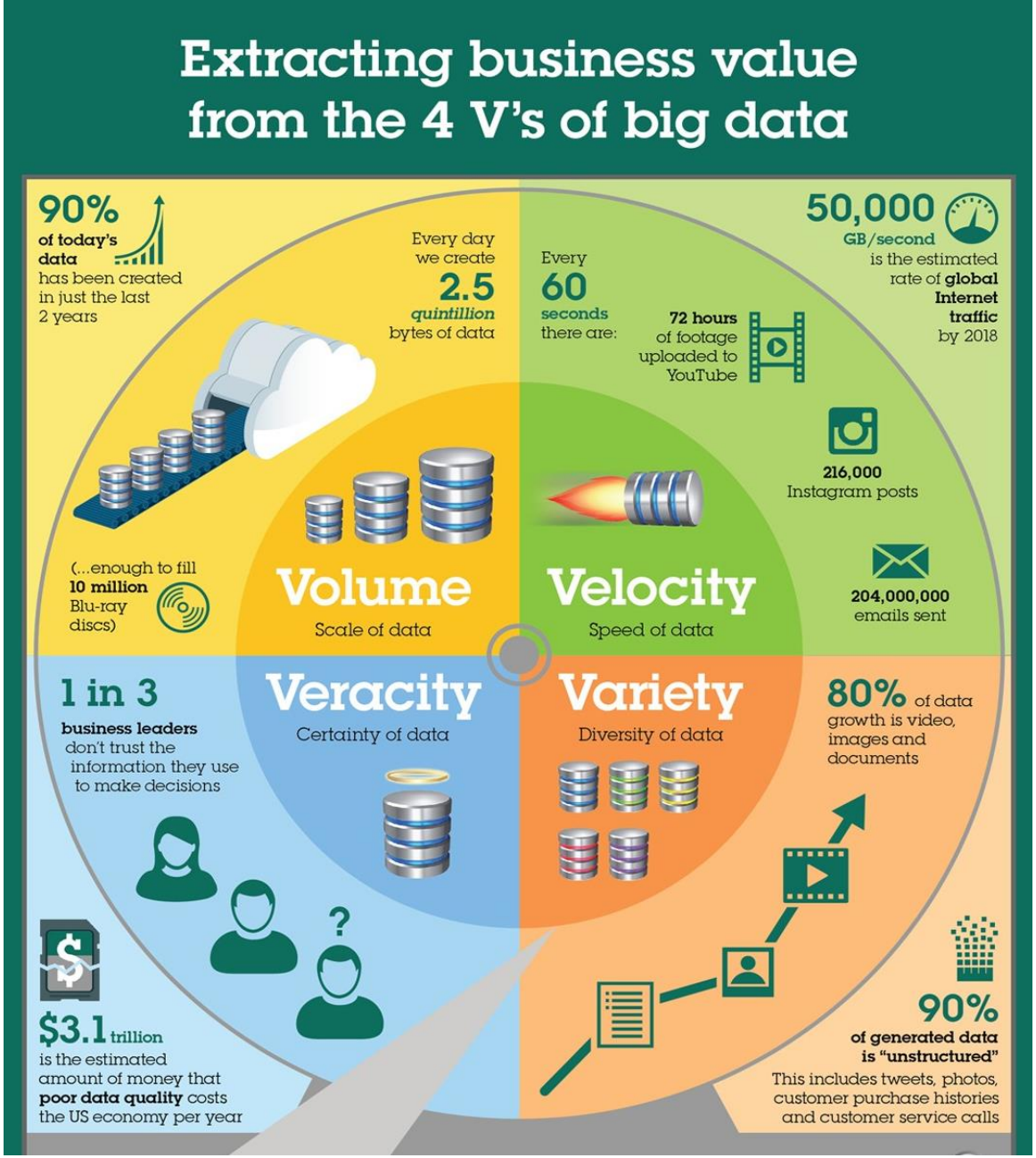
  Big data takes the form of:

  - Messages, updates and images posted to social networks,
  - Readings from sensors,

- **Veracity**

  Certainty of Data

  Truth

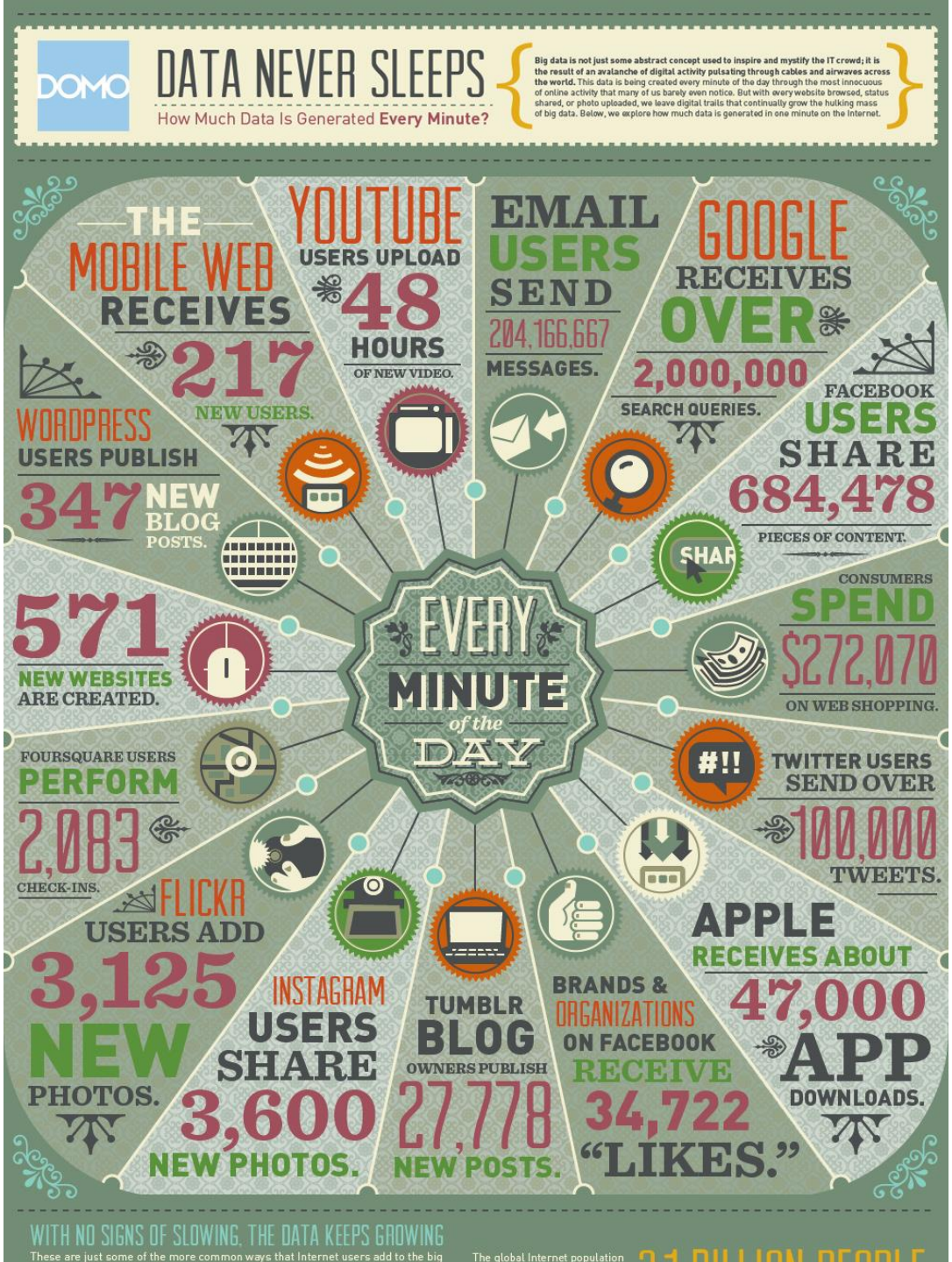


http://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data

How much data
is created?

An estimation by DOMO.
(http://www.domo.com/blog/2012/06/how-much-data-is-created-every-minute/)

Data variety
Text: google, twitter, facebook
Images: Facebook, instagram
Movies: youtube



◆ Why Big Data?

Make use of Big Data to make better decisions.

The analysis of data will enable us to

- Predict whether something will happen.
- Predict how much something will happen. ⇒ **Decisions**
- Group items by their similarity.

If data helps us, Big Data should help us more.

Some evidence that it actually helps.

Companies in the data-driven decision making:

5% more productive, and 6% more profitable than competitors.

(HBR Oct. 2012)

How it helps?

- Provide accurate information. → We can make an optimal choice.
- Grouping similar items. → We can use the same strategy for similar items.

Example (Predicting time of arrival):
Accurate information about flight arrival times matters. The ground staff needs to be ready for a plane landing. Inaccurate information is very costly.

Their duties include the handling of the baggage, stocking the plane with the refreshments, and cleaning the plane between flights.

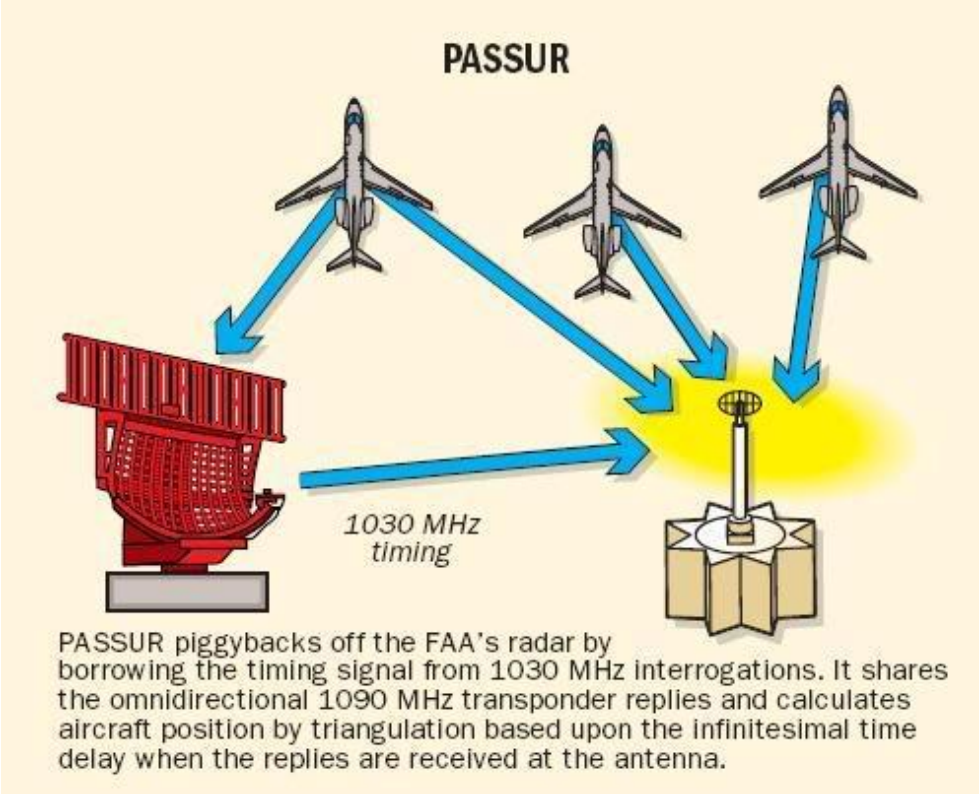
Ground staff scheduling is important. It can be a topic of a Ph.D. thesis…

Inaccurate estimations on arrival times will waste your perfect scheduling.

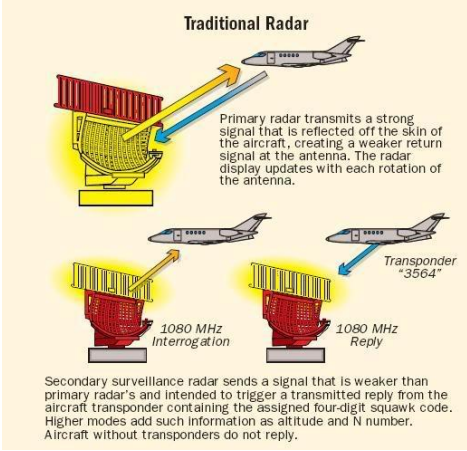
Example (Predicting time of arrival):
Accurate information about flight arrival times matters. The ground staff needs to be ready for a plane landing. Inaccurate information is very costly.

PASSUR (www.passur.com) offers its own arrival estimates as a service.
They use …

- Public data, e.g. weather,
- Proprietary sensor data (from a network of passive radar stations).

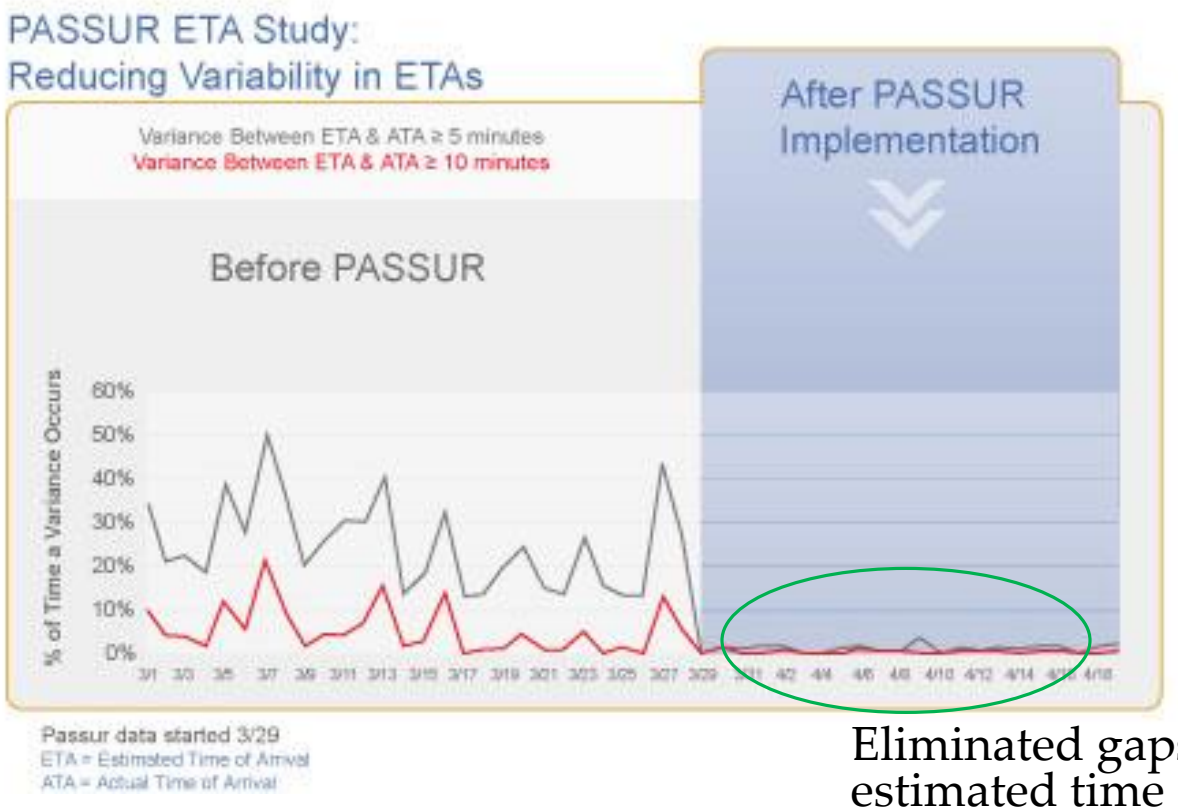


PASSUR piggybacks off the FAA's radar by borrowing the timing signal from 1030 MHz interrogations. It shares the omnidirectional 1090 MHz transponder replies and calculates aircraft position by triangulation based upon the infinitesimal time delay when the replies are received at the antenna.



Diagrams from Aviation week

% of time misestimating occurs



Eliminated gaps between estimated time and actual time

Example (Grouping customers):

Ira Haimowitz and Henry Schwartz (1997) show an example of how clustering was used to improve decisions about how to set credit lines for new credit customers.

Data: existing GE Capital customers' use of their cards, payment of their bills, and profitability to the company.

GE Capital is a financial service unit of General Electric.
One of their services is offering credit cards.
In Japan, it is known as a former owner of Lake. (It already sold Lake.)

Haimowitz and Schwartz clustered those GE Capital customers based on similarity.

They settled on five clusters that represented very different consumer credit behavior (e.g., those who spend a lot but pay off their cards in full each month versus those who spend a lot and keep their balance near their credit limit).

These different sorts of customers can tolerate very different credit lines (in the two examples, extra care must be taken with the latter to avoid default).

Five clusters (groups of customers) they found;

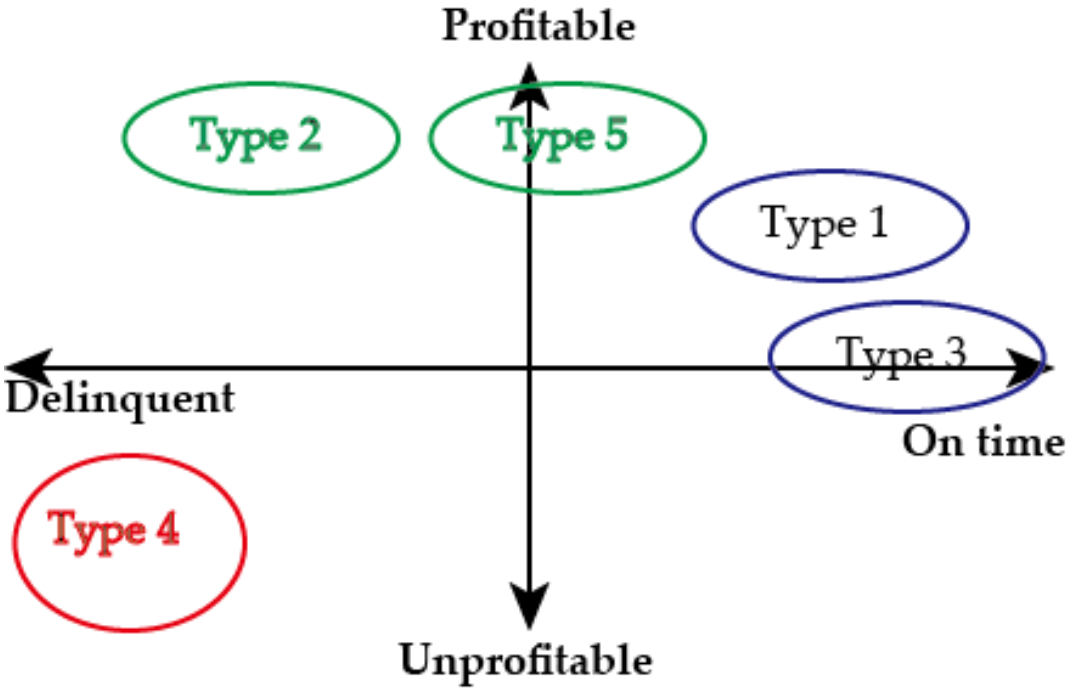
1. Usually on time with payments, pay most of their monthly balance, use some of their credit line, fairly high sales, and fairly profitable.

2. Fairly delinquent accounts, pay some of their monthly balance, high sales, and very profitable. Should be treated with caution in times of recession.

3. On time with payments, but very little sales activity. Not very profitable.

4. Very delinquent; all of these are write-offs. Generate fairly high sales but are unprofitable. Creditors lose money on these.

5. Mixture of on-time and delinquent accounts, generate high sales, and are very profitable, especially at lower credit lines.



The problem with using this clustering immediately for decision making is that the data are not available when the initial credit line is set. Haimowitz and Schwarz took this new knowledge and cycled back to the beginning of the data mining process. They used the knowledge to define a precise predictive modeling problem: using data that are available at the time of credit approval, predict the probability that a customer will fall into each of these clusters. This predictive model then can be used to improve initial credit line decisions.

(``Data Science for Business'' by Provost&Fawcett)

What is Data Science?

Data analysis: help improving decision making.

Data Science: provide principles, processes, techniques for understanding phenomena via data analysis.

Principles: Incorporated into processes/techniques in the course.
Processes: CRISP-DM (Lec. #2), SEMMA and etc.
Techniques: Regression, Classification, Association and etc.

What is Data Science?

Data analysis: help improving decision making.

Data Science: provide principles, processes, techniques for understanding phenomena via data analysis.

Principles: Incorporated into processes/techniques in the course.
Processes: CRISP-DM (Lec. #2), SEMMA and etc.
Techniques: Regression, Classification, Association and etc.

An example of principles:

Extracting useful knowledge from data to solve problems should be treated systematically. → Incorporated into CRISP-DM.

What is Data Science?

Data analysis: help improving decision making.

Data Science: provide principles, processes, techniques for understanding phenomena via data analysis.

Principles: Incorporated into processes/techniques in the course.
Processes: CRISP-DM (Lec. #2), SEMMA and etc.
Techniques: Regression, Classification, Association and etc.

An example of principles:

Extracting useful knowledge from data to solve problems should be treated systematically. → Incorporated into CRISP-DM.

You will find something from any set of data – but it might not generalize.
→ Evaluation of each technique (overfitting)

What is Data Science?

Data analysis: help improving decision making.

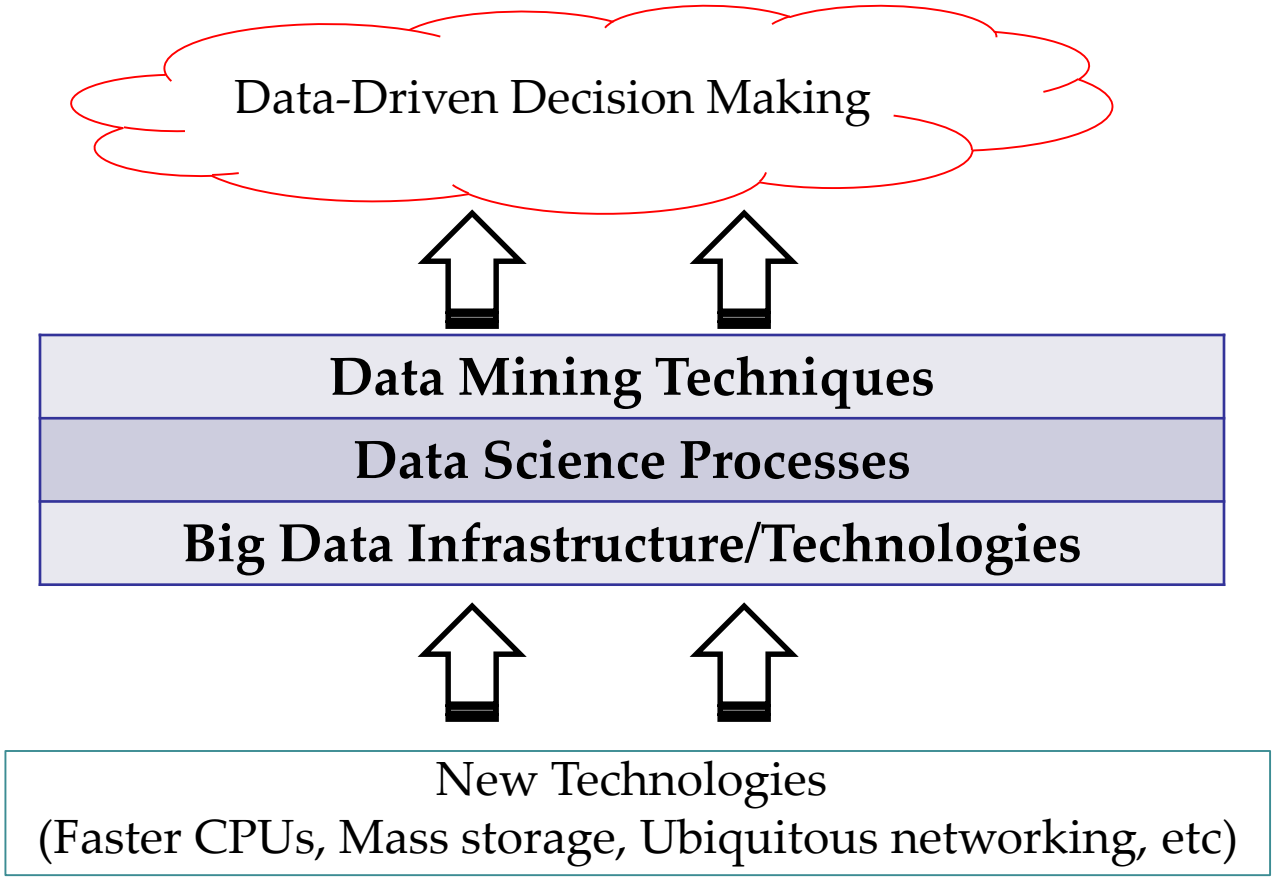
Data Science: provide principles, processes, techniques for understanding phenomena via data analysis.

Principles: Incorporated into processes/techniques in the course.
Processes: CRISP-DM (Lec. #2), SEMMA and etc.
Techniques: Regression, Classification, Association and etc.

Big Data Science: provide principles, processes, techniques for understanding phenomena via big data analysis.

+ Big Data Infrastructure (Hadoop etc.)

Big Data Science



Potential Applications (1/2)

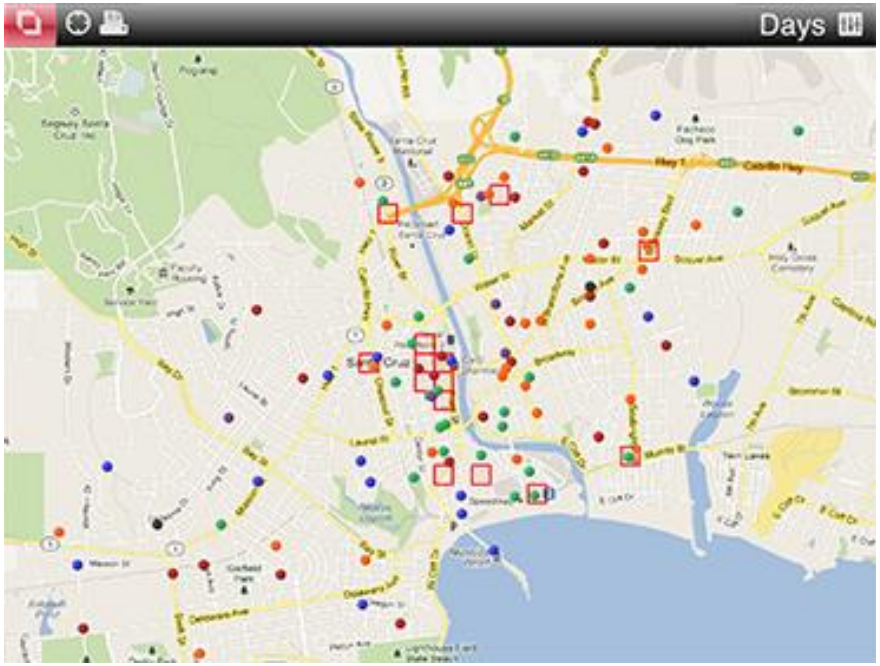
Many Big Data applications are in marketing.

But, its applications are not just in marketing…

- Predict crimes

LAPD (L.A. Police Dept.) uses Big data to predict crimes.

PredPol (http://www.predpol.com/gun-violence/)



Red boxes indicate where the patrols should watch.

Jeff Brantingham [creator of PredPol], an anthropologist at UCLA, wanted to see if computers could model future crime *the same way they model earthquake aftershocks*.

You may be thinking about the sci-fi movie Minority Report. But this is different. No psychics sleeping in bathtubs […] this doesn't predict ***who*** will commit a future crime.

The software uses past statistics to project ***where*** crime is moving.

("Can Software That Predicts Crime Pass Constitutional Muster?", National Public Radio, 2013/07)

Criminal offences, like infectious disease, form patterns in time and space. A burglary in a placid neighborhood represents a heightened risk to surrounding properties; the threat shrinks swiftly if no further offences take place.

("Don't even think about it", Economist, 2013/07)

- Potential Applications (2/2)
  - Weather Forecast

WeatherNews has made a weather communication community among users of its phone application. It collects users' weather observations and use them to forecast the weather together with publicly available data.

Some users' weather reports (weathernews.jp)

Weathernews collects weather reports from their community.

Reports around Fukushima

In some cases, they provided more accurate predictions than the public weather forecast, JMA (Japan Meteorological Agency).

For a forecast for Feb. 6 2013, JMA predicted (relatively heavy) snow, but WNews predicted rain. And it rained. (http://diamond.jp/articles/32435)

- Example 1

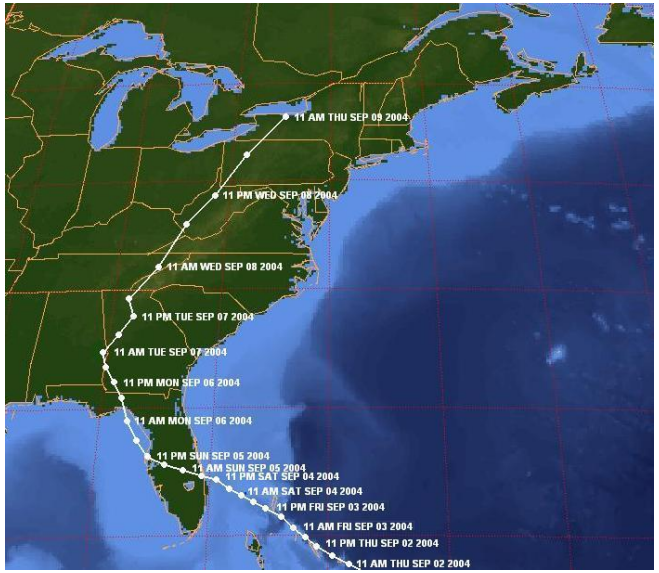
A *New York Times* story from 2004:

"Hurricane Frances was on its way, barreling across the Caribbean, threatening a direct hit on Florida's Atlantic coast. Residents made for higher ground, but far away, in Bentonville, Ark., executives at Wal-Mart Stores decided that the situation offered a great opportunity for one of their newest data-driven weapons … predictive technology.

A week ahead of the storm's landfill, Linda M. Dillman, Wal-Mart's chief information officer, pressed her staff to come up with forecasts based on what had happened when Hurricane Charley struck several weeks earlier. Backed by the trillions of bytes' worth of shopper history that is stored in Wal-Mart's data warehouse, she felt that the company could 'start predicting what's going to happen, instead of waiting for it to happen,' as she put it. (Hays, 2004)"

See http://www.nytimes.com/2004/11/14/business/yourmoney/14wal.html?_r=0

Hurricane Frances



Would data analysis be useful in this scenario?

- Find that people in the path of the hurricane would buy more bottled water.

Maybe, but this point seems obvious.

- Discover patterns due to the hurricane that were not obvious.

(More valuable)

The New York Times (Hays, 2004) reported that:

"… the experts mined the data and found that the stores would indeed need certain products —and not just the usual flashlights. 'We didn't know in the past that **strawberry Pop-Tarts** increase in sales, like **seven times** their normal sales rate, ahead of a hurricane,' Ms. Dillman said in a recent interview."

Remark:

Your data can tell you that the sales will increase.

Your data cannot tell you why the sales will increase.

You often need to infer the implication of your analysis if you need to convince people to follow your suggestions.

Saying "Hey, according to data, your store should stock up pop-tarts" may not be enough.

So, let's infer the implication…

What will people stock up on before hurricanes?

One of such items is non-perishable food that can be eaten easily and without heat.

Why? After hurricanes electricity will often be out for multiple days.

Pop-Tarts require **no heating** to eat.

They do **not need a fridge or freezer** to be stored.

Hurricanes often hit the Southeast US during summer where it is very hot and humid. A great fruit to cool off with, next to watermelon, is the strawberry

→ Strawberry pop-tarts seem to perfectly fit.

- Example 2

The second-largest discount retailer in the United States

Target's 'Pregnancy Prediction Score'

Nowadays, retailers care about consumers' shopping habits.

what drives them, and what can influence them?

Finding: Consumers have inertia in their habits and getting them to change is very difficult. However, that the arrival of a new baby is one point where people do change their shopping habits significantly.

"As soon as we get them buying diapers from us,
they're going to start buying everything else too."
by a Target analyst

Difficulty: Most birth records are public.

There is a fierce competition: most retailers obtain information on births and send out special offers to the new parents.

Target explored whether they could predict that people are expecting a baby.

- If they could, they would gain an advantage by making offers before their competitors!
- Target were able to extract information that could predict which consumers were pregnant.
- (analyzed historical data on customers who later were revealed to have been pregnant.)
- Pregnant mothers often change: their diets, their wardrobes, their vitamin regimens, and so on.

http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=all

Sneak preview: Data Science Process

In this course, you will…

- Learn how to approach business/research problems data-analytically.
- Be able to assess whether and how data can solve problems.

How to approach problems data-analytically??

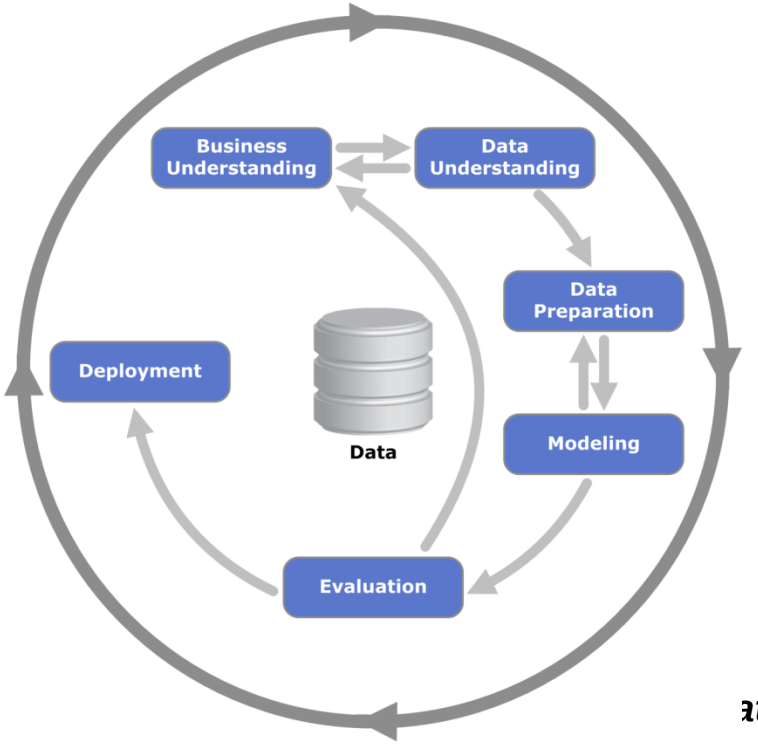
Sneak preview: Data Science Process

In this course, you will…

- Learn how to approach business/research problems data-analytically.
- Be able to assess whether and how data can solve problems.

How to approach problems data-analytically??

There exists a standard process, CRISP-DM.



Data Science Process consists of ..

1. Business understanding
2. Data understanding
3. Data preparation
4. Modeling
5. Evaluation
6. Deployment