Data Science for Business
Lecture #1
The promise and limitations of data science for business

Prof. Alan L. Montgomery

The University of Hong Kong & Carnegie Mellon University, Tepper School of Business

email: alanmontgomery@cmu.edu

Lecture Outline

What is Data Science?

Promise of Big Data

Limitations of Data Science

Data Science

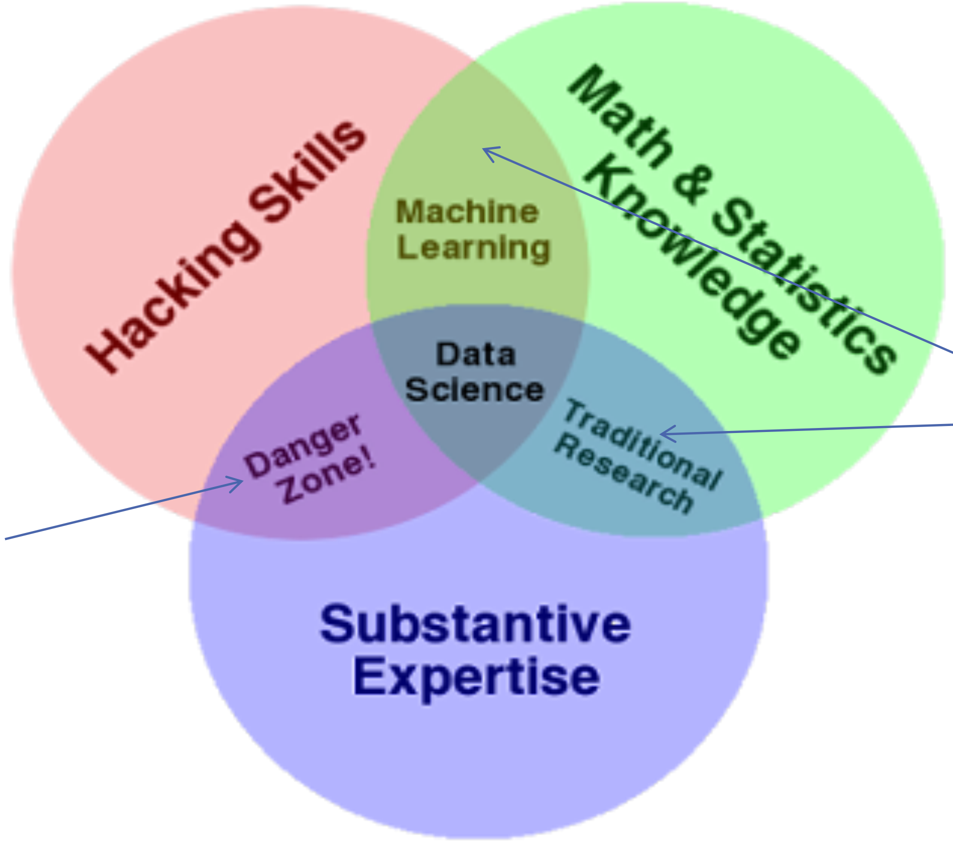
The Science of extract knowledge from data

What is Data Science?

Extracting knowledge from data to make better decisions in business.

- Data can be structured (transactions) or unstructured (emails, videos, photos, social media, …).
- Data are facts, while knowledge is generalizable.
- The extraction process uses data mining techniques.
- Data scientists need to have both technical skills for data mining, but domain knowledge to know how to structure an analysis to have an impact.
- Notice that Data Science is an empirical science. It is based upon experimentation or observation (evidence), as opposed to case studies or theoretical research methods.

What is data science?



Why “Danger”?

What is the Difference Between “Data Mining” and “Traditional Research”?

Empiricism versus Rationalism

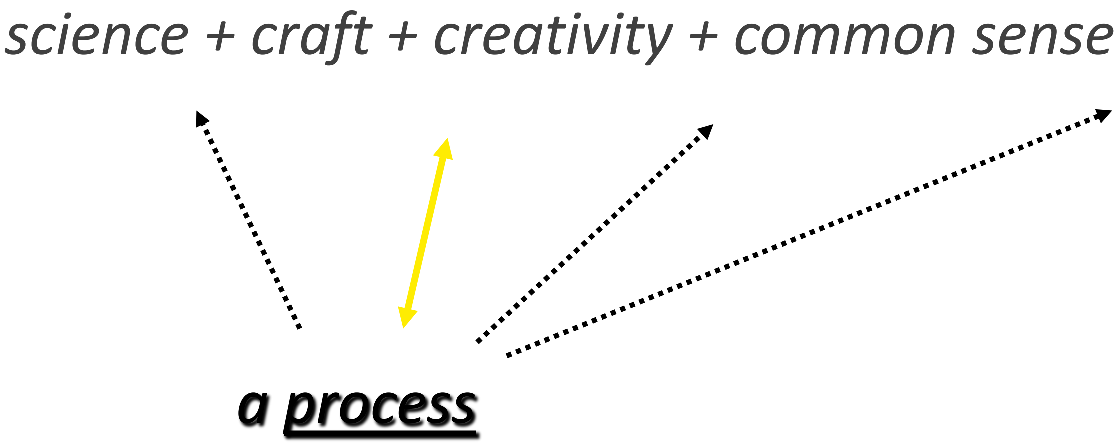
Empiricists believe that we learn about our world through previous experience. Rationalists believe reason is the basis for understanding anything.

- *Strength of empiricism*: Objectivity, measurability, replicability, …
- *Weaknesses of empiricism:* Inefficient, Potential for bias in data due to (bad) prior research or incorrect collection, Lack of information (e.g., predict future), …

Example: Why is our customer loyalty dropping?

- Empiricists collect data, rationalists argue we know that loyalty is driven by consumer heuristic to “keep doing it if it works”

Data science is a process



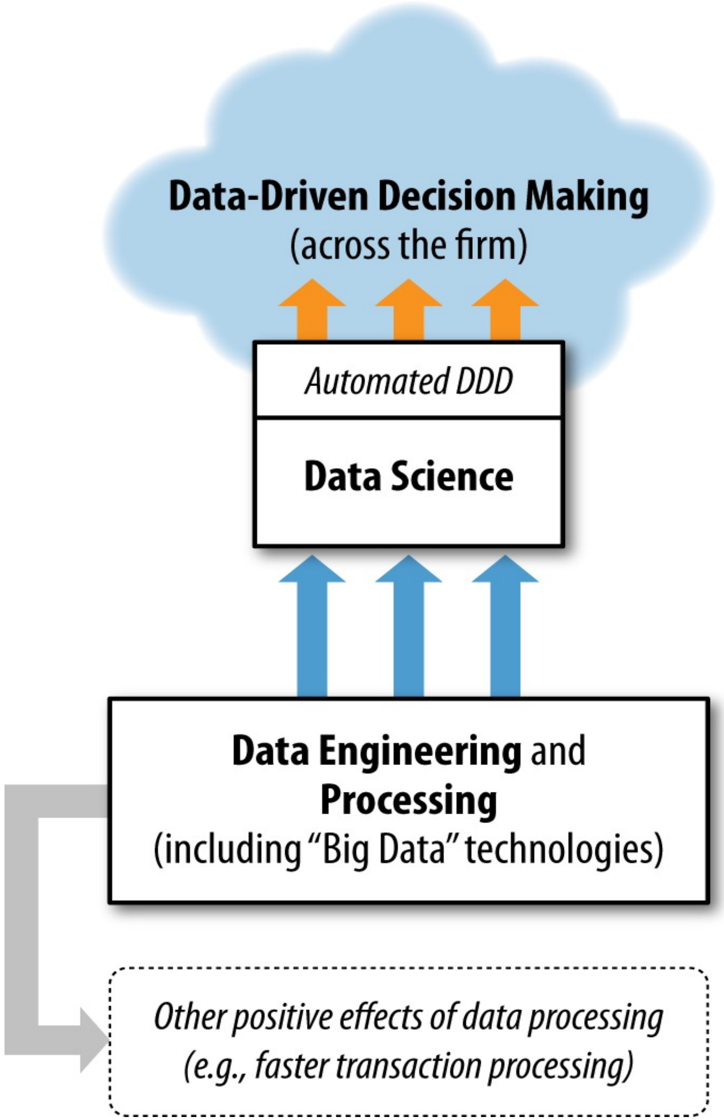
science + craft + creativity + common sense

a <u>process</u>

Source: Foster Provost

Relationship of Data Mining to

- Data Science
- Decision Making
- Big Data
- Business Analytics



Data-Driven Decision Making
(across the firm)
Automated DDD
Data Science
Data Engineering and Processing
(including "Big Data" technologies)
Other positive effects of data processing
(e.g., faster transaction processing)

Source: Foster Provost

Connection with Other Related Areas

Machine Learning

Artificial Intelligence

Optimization

Operations Research

Exploratory Data Analysis

Visualization

Statistical Modeling

Statistical Computing

Hadoop/MapReduce/GraphLab/Parallel Computing

What are the most important qualities that recruiters want in a data scientist?

Working in Teams

Statistical Analysis

Software Engineering

Signal Processing

Programming

Pattern Recognition

Optimization

Machine Learning

Data Visualization

Database Design

Clear and Effective Communication

Business Acumen

What are the most important qualities that recruiters want in a data scientist?

Working in Teams

Statistical Analysis

Software Engineering

Signal Processing

Programming

Pattern Recognition

Optimization

Machine Learning

Data Visualization

Database Design

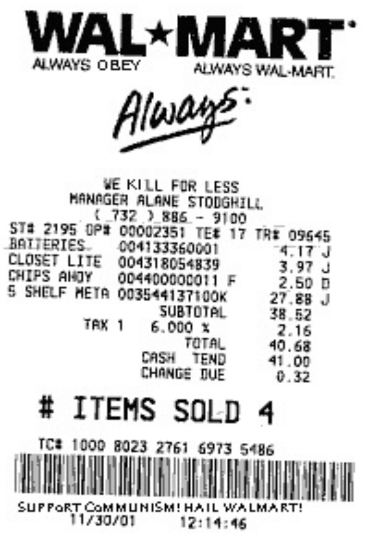
Clear and Effective Communication

Business Acumen

The Promise of Big Data

Motivating the need for “Data Science”

Consumers generate Big Data

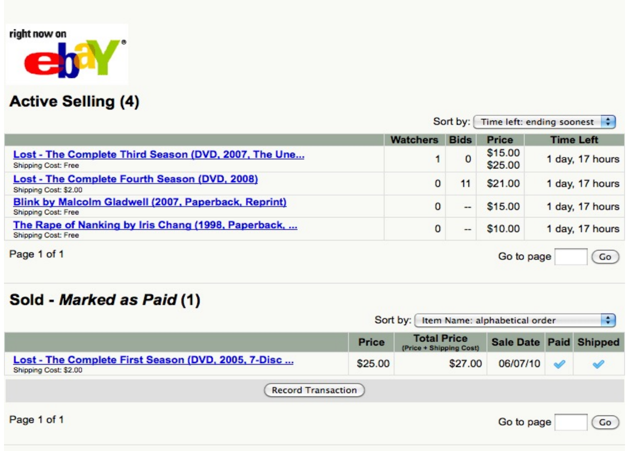


WAL★MART
ALWAYS OBEY ALWAYS WAL-MART.
Always:

```
        WE KILL FOR LESS
   MANAGER ALANE STODGHILL
      ( 732 ) 886 - 9100
ST# 2195 OP# 00002351 TE# 17 TR# 09645
BATTERIES    004133360001      4.17 J
CLOSET LITE  004318054839      3.97 J
CHIPS AHOY   004400000011 F    2.50 D
5 SHELF META 0035441371OOK    27.88 J
                 SUBTOTAL     38.52
       TAX 1   6.000 %         2.16
                    TOTAL     40.68
               CASH  TEND     41.00
               CHANGE DUE      0.32

     # ITEMS SOLD 4

   TC# 1000 8023 2761 6973 5486
SUPPORT COMMUNISM! HAIL WALMART!
     11/30/01        12:14:46
```



right now on ebay

Active Selling (4)

Sort by: Time left: ending soonest

	Watchers	Bids	Price	Time Left
Lost - The Complete Third Season (DVD, 2007, The Une... Shipping Cost: Free	1	0	$15.00 $25.00	1 day, 17 hours
Lost - The Complete Fourth Season (DVD, 2008) Shipping Cost: $2.00	0	11	$21.00	1 day, 17 hours
Blink by Malcolm Gladwell (2007, Paperback, Reprint) Shipping Cost: Free	0	--	$15.00	1 day, 17 hours
The Rape of Nanking by Iris Chang (1998, Paperback, ... Shipping Cost: Free	0	--	$10.00	1 day, 17 hours

Page 1 of 1 Go to page Go

Sold - *Marked as Paid* (1)

Sort by: Item Name: alphabetical order

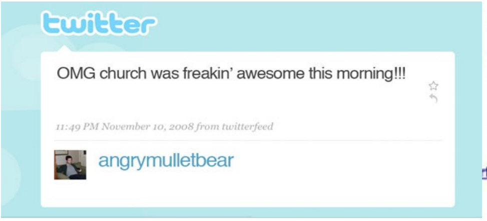
	Price	Total Price (Price + Shipping Cost)	Sale Date	Paid	Shipped
Lost - The Complete First Season (DVD, 2005, 7-Disc ... Shipping Cost: $2.00	$25.00	$27.00	06/07/10	✓	✓

Record Transaction

Page 1 of 1 Go to page Go

☒ משתמשים ☒ תאונות ☒ שוטרים ☒ דיווחי עומסים ☒ דיווחי משתמשים
user-117478842
15 קמ"ש
Scale = 1 : 27K
34.80917, 32.08712
Permalink

Google UK

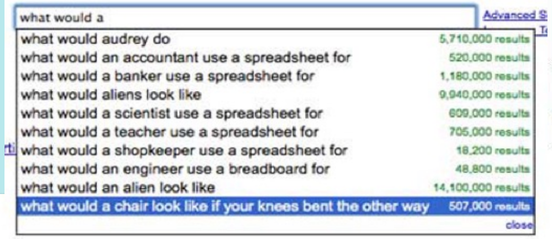


twitter

OMG church was freakin' awesome this morning!!!

11:49 PM November 10, 2008 from twitterfeed

angrymulletbear



what would a

what would audrey do	5,710,000 results
what would an accountant use a spreadsheet for	520,000 results
what would a banker use a spreadsheet for	1,180,000 results
what would aliens look like	9,940,000 results
what would a scientist use a spreadsheet for	609,000 results
what would a teacher use a spreadsheet for	705,000 results
what would a shopkeeper use a spreadsheet for	18,200 results
what would an engineer use a breadboard for	48,800 results
what would an alien look like	14,100,000 results
what would a chair look like if your knees bent the other way	507,000 results

close

What is Big Data?

Four basic components:

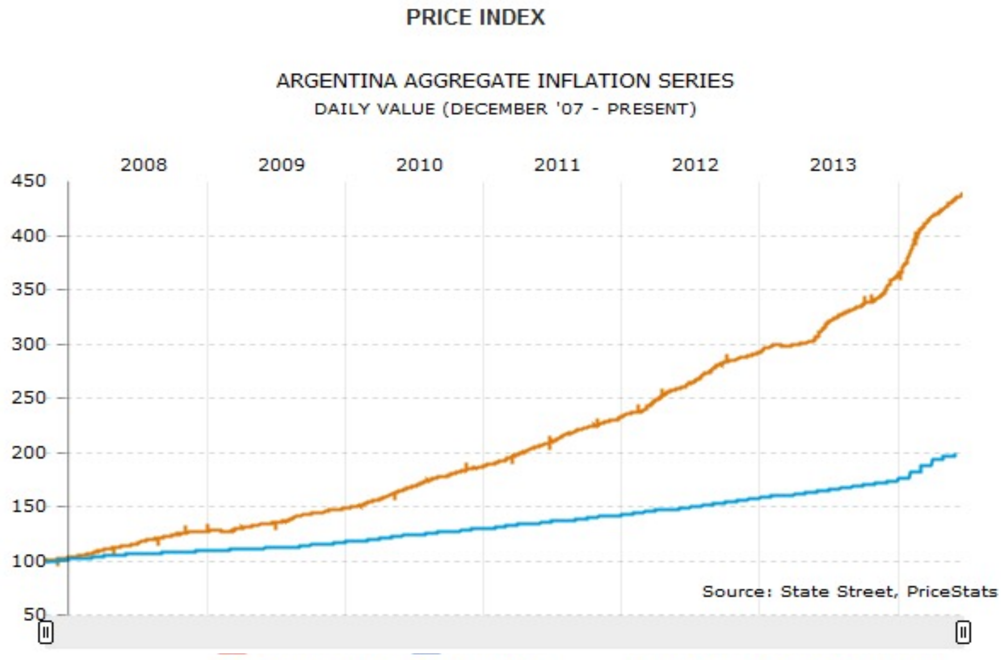
- Massive datasets
- Unstructured data
- Collected as a by-product from transactions (not for decision making)
- Populations not samples

Related to Business Analytics, Data Analysis, Data Mining, Data Science, Machine Learning, Statistics

Big Data Example:
Billion Prices Project

MIT project which aggregates price information from daily price fluctuations of ~5 million items sold by ~300 online retailers in more than 70 countries

Contrast with CPI that focuses on a basket of items that are monitored periodically



Source: State Street, PriceStats

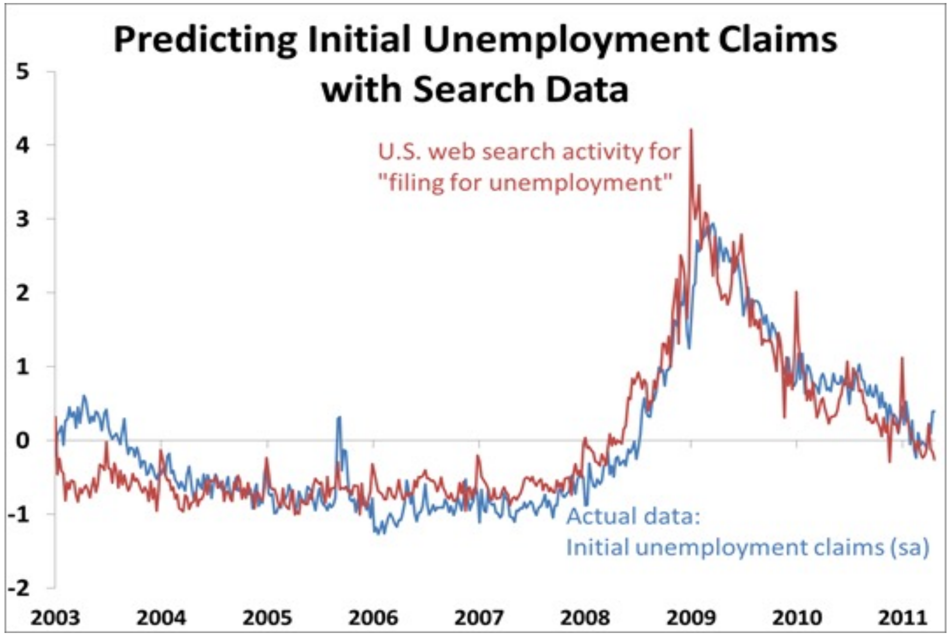
Big Data Example: Google Correlates

Provide Google with your weekly time series and it will tell you which search terms are most closely correlated with your data

Question: What predicts Initial Unemployment Claims?

Answer: “filing for unemployment”

Provides a leading indicator based upon search

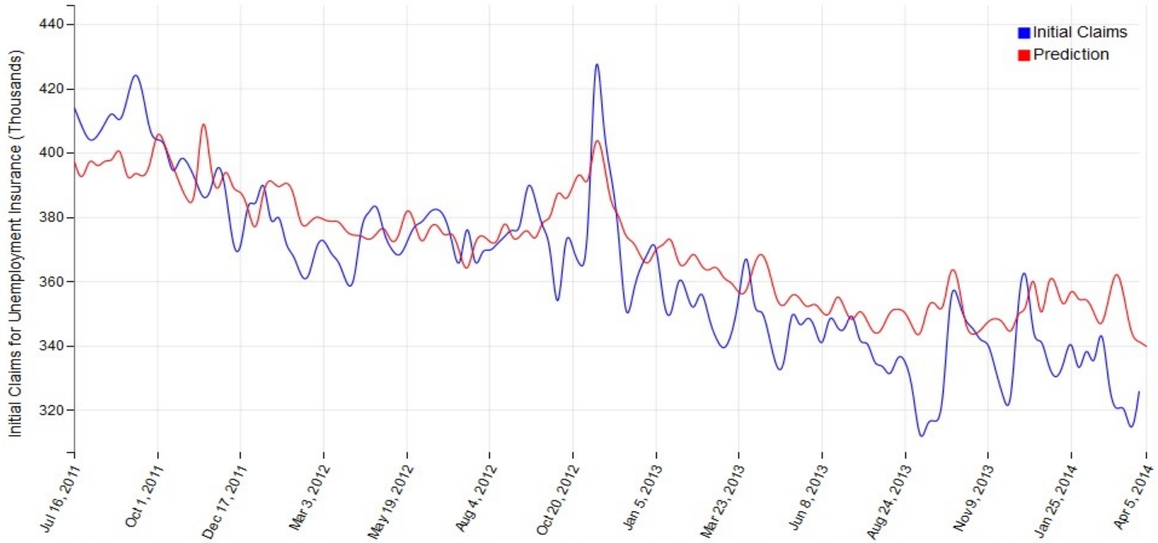


Big Data Example:
Tweeting about Unemployment

Tweets like “I just lost my job. Who’s buying my drinks tonight?” can be used to predict unemployment.

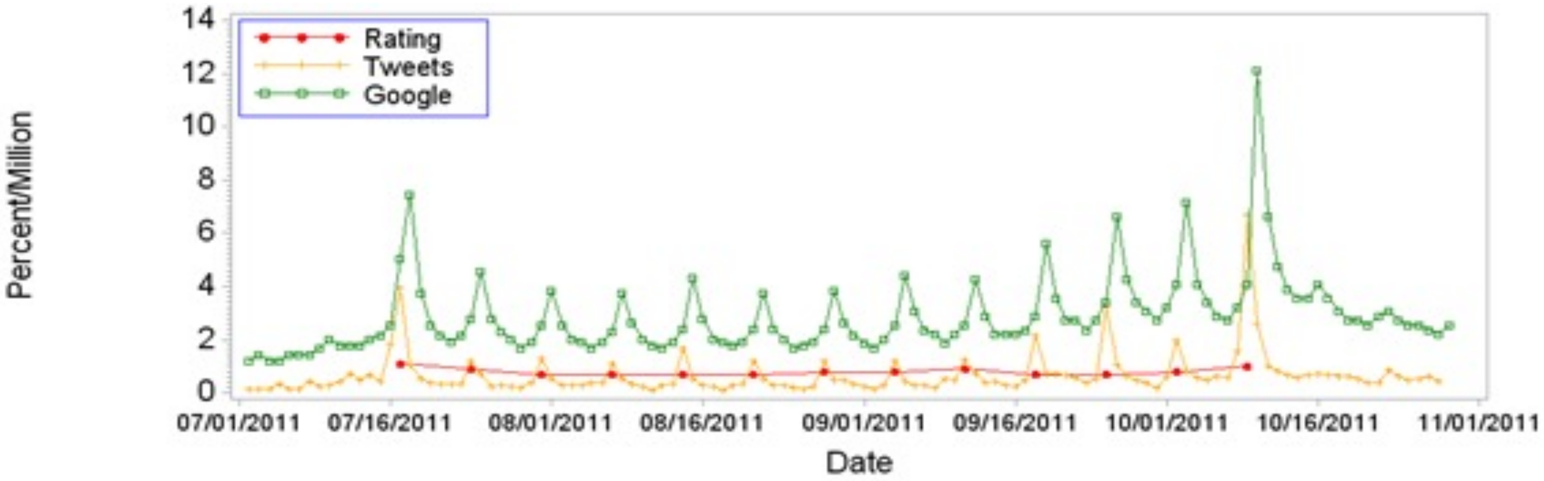
Predicts 15-20% of the variance of the prediction error of the consensus forecast for initial claims.

Source: Antenucci et al, NBER Working Paper 20010



Sources: Initial Claims for Unemployment Insurance (seasonally adjusted), U.S. Department of Labor; Prediction, University of Michigan Social Media Job Loss Index.

Big Data Example: Predicting TV Ratings



Data for "Breaking Bad" in 2011 shows that using Tweets can accurately predict TV ratings

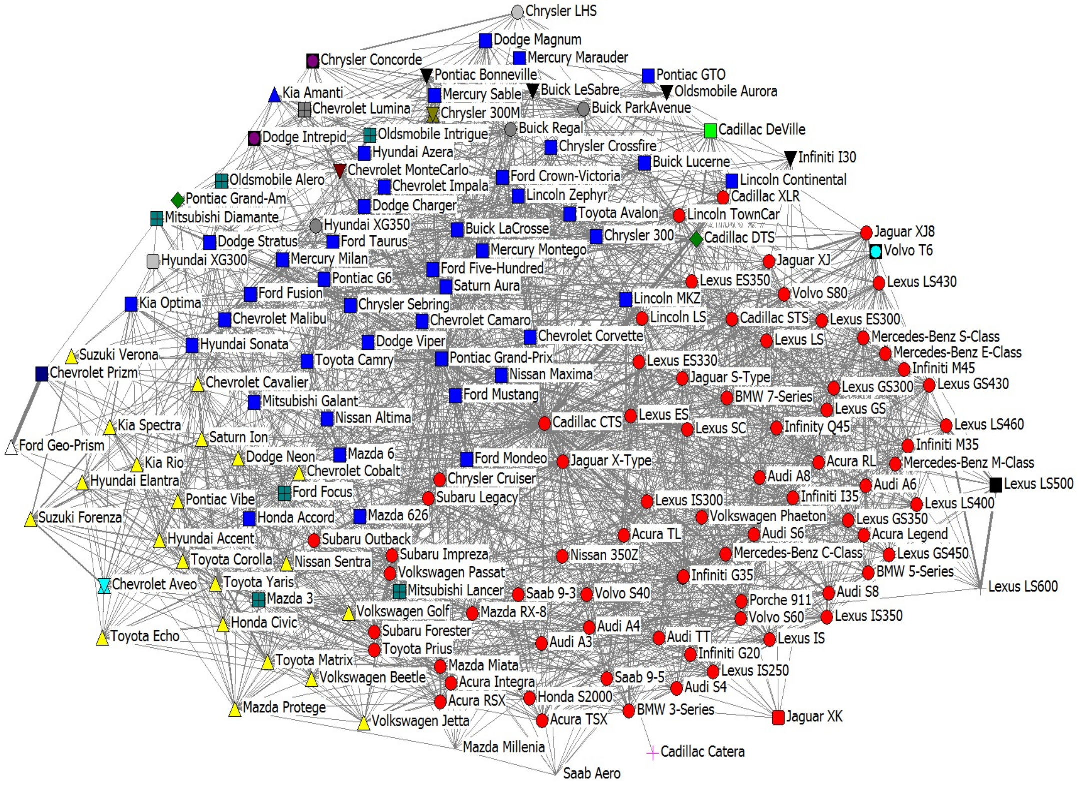
Online search is a very weak predictor of TV show demand, but using sentiment of Tweets can explain up to 90% of variation

Source: Liu, Singh and Srinivasan (2014)

Big Data Example:
Understanding Market Structure

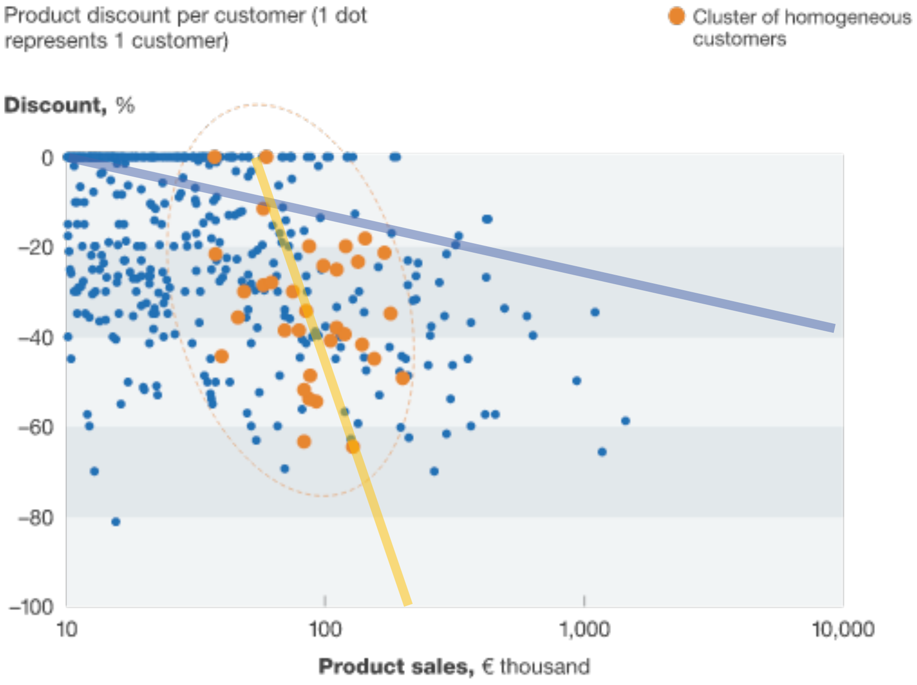
Message #1199 **Civic vs. Corolla** by mcmanus *Jul 21, 2007 (4:05 pm)*
Yes DrFill, the Honda car model is sporty, reliable, and economical vs the **Corolla** that is just reliable and economical. Ironically its Toyota that is supplying 1.8L turbo ... Neon to his 16 year old brother. I drove it about 130 miles today. Boy does that put all this **Civic** vs. **Corolla** back in perspective! The Neon is very crudely designed and built, with no low ...

Audi A6	Honda Civic	252
Audi A6	Toyota Corolla	101
Honda Civic	Audi 6	252
Honda Civic	Toyota Corolla	2762
Toyota Corolla	Audi A6	101
Toyota Corolla	Honda Civic	2762



Source: Netzer (2011), “Mine Your Own Business”

Making Better Pricing Decisions



Source: multinational energy company (disguised example); McKinsey analysis

Price is a highly leveraged decision tool

- a 1% price increase translates into an 8.7% increase in operating profits (assuming no loss of volume)

The problem is that many managers rely on rules of thumb, “market prices”, or outdated strategies to set prices

This example illustrates how leverage granular data (customer/invoice level) to infer price sensitivity

- Control for factors that affect response, otherwise you want not be able to see response

https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/using-big-data-to-make-better-pricing-decisions#

Advantages of Big Data

Can detect patterns by leveraging these qualities:

- Massive
- Immediate and timely
- Predictive
- Free (or inexpensive)

Limitations of Data Science

Why companies fail to leverage Big Data

Some illustrations

1. Assessing Model Uncertainty in Financial Risk Management
2. Retail Price Optimization using Business Rules
3. Predicting Flu Trends
4. Problems with Big Data
5. Mangers and culture

1. Assessing Model Uncertainty in Financial Risk Management

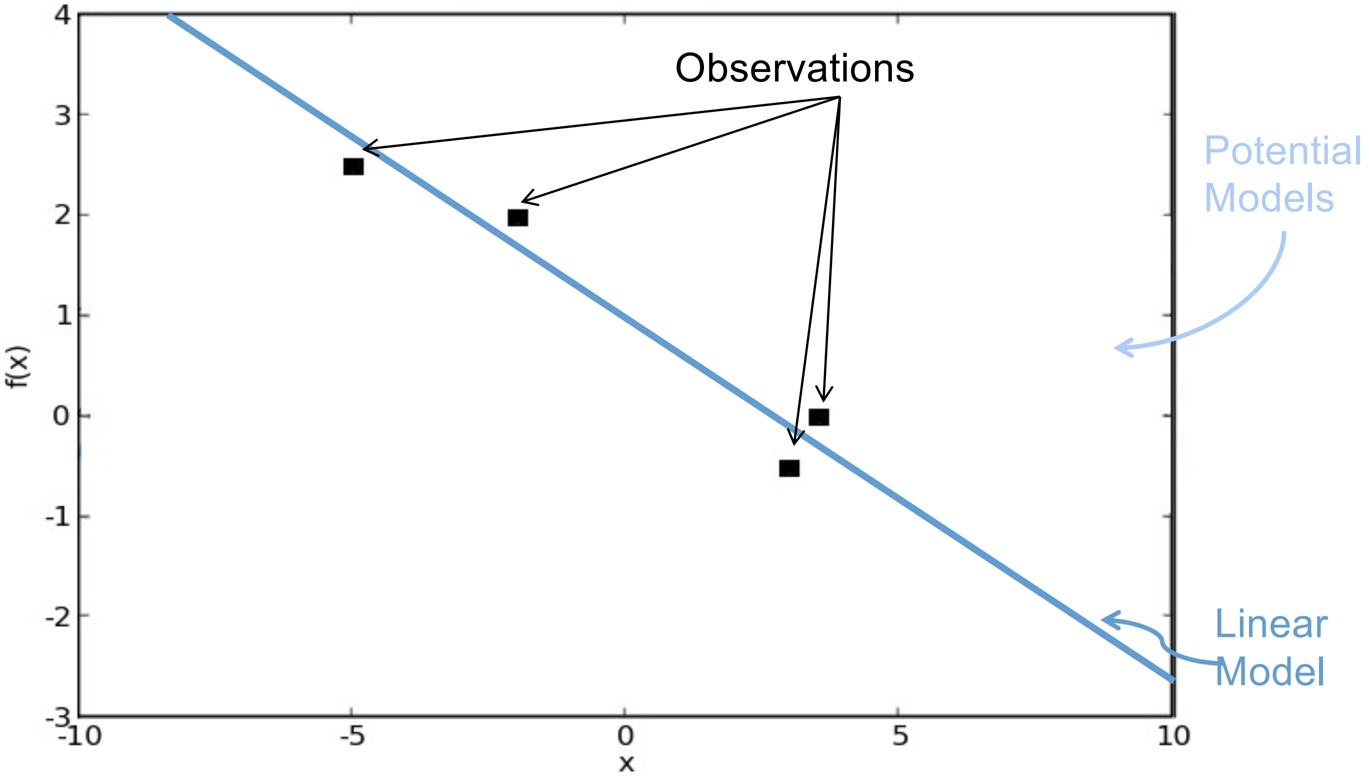
Making Decisions with Big Data

Data mining models look for patterns in data that can be used to make better predictions and decisions

The problem is “that all models are wrong; the practical question is how wrong do they have to be to not be useful”? (Box and Draper 1987)

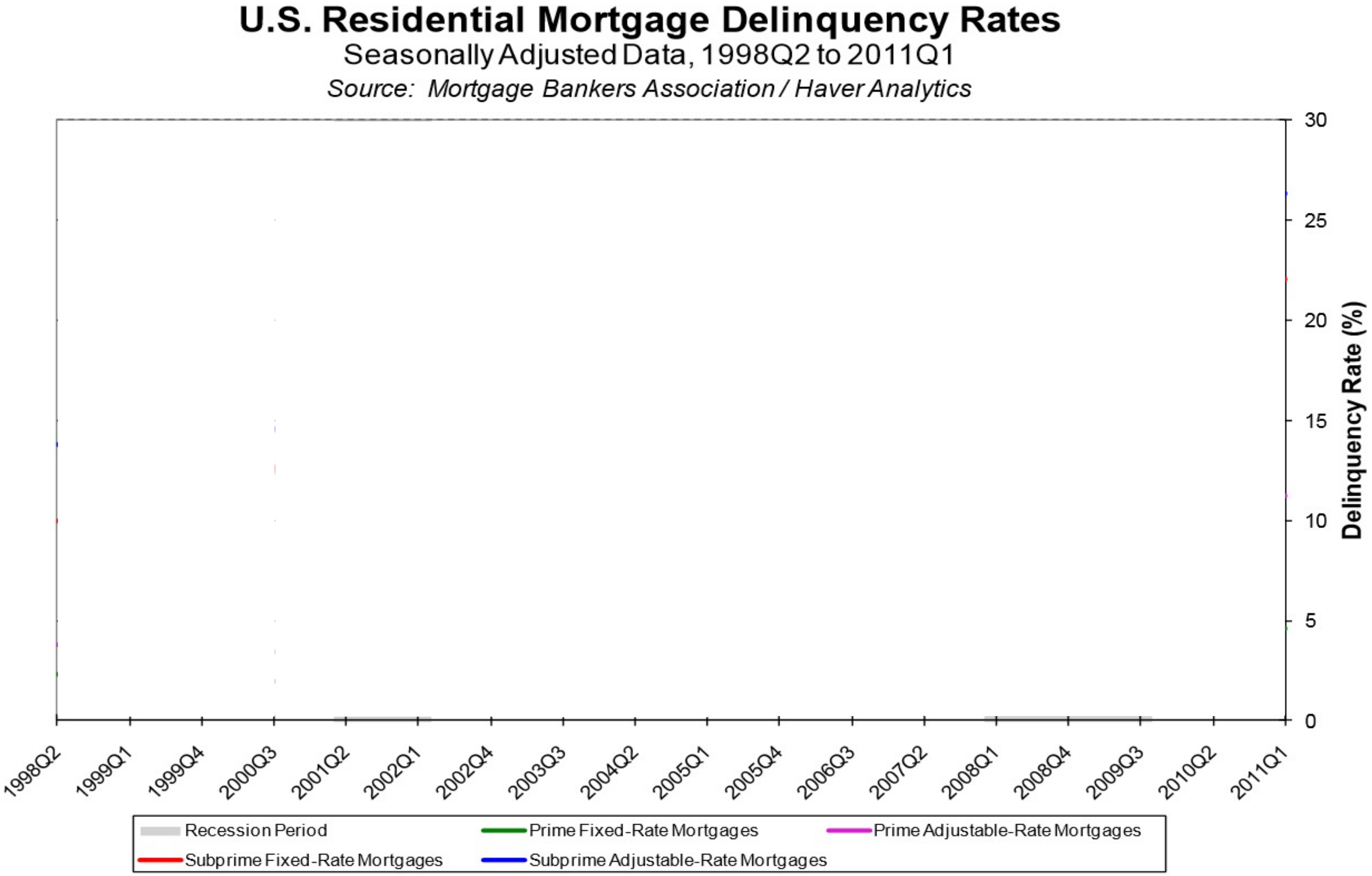
The Federal Reserve Bank has mandated banks to evaluate their risk exposure to quantitative models. But industry surveys (PWC 2013) show no commonly accepted standard for evaluating risk.

Illustration: Uncertain Form



Models have a huge impact on interpolations and extrapolations

Illustration: Insufficient Data



Overconfidence Problem

If we identify and estimate the model using the same data as we are for making predictions then we are prone to be *overconfident* in our models.

To compensate for this overconfidence we need to consider…

- How much are we learning about functional form observed data
- That many errors are correlated (unemployment, hurricanes, …) but most models assume independence, especially if the model is trained in a good economic cycle and we want to forecast in a bad economic cycle
- Relationships may not be stable over time, economic relationships may be impacted by business cycles (which tend to be slow, infrequent)

2. Retail Price Optimization using Business Rules

Price Optimization in Practice

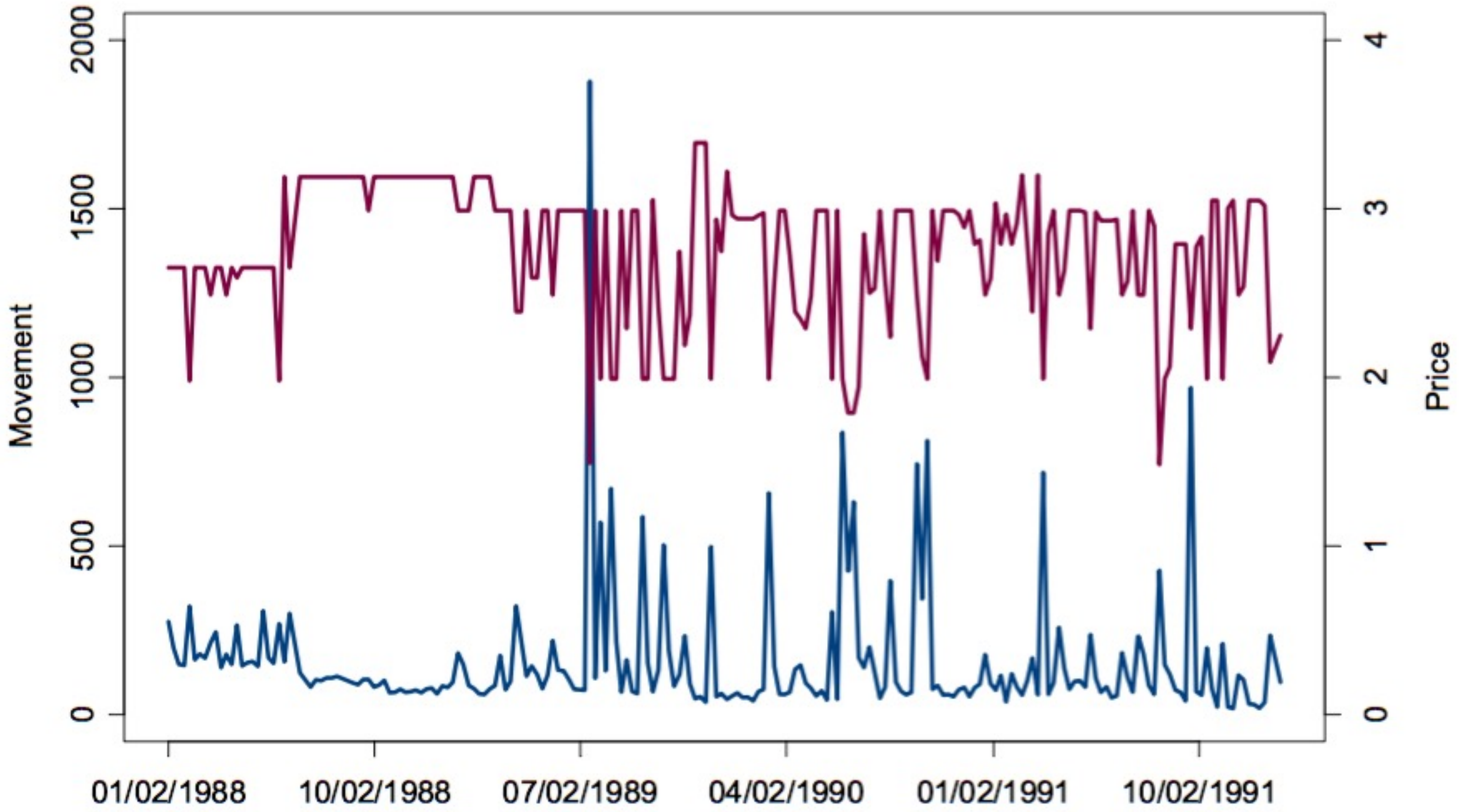
Huge growth of price optimization in practice for both retailers and manufacturers

Gartner Marketscope states that “through 2010, price optimization technology will have a more direct impact on increasing revenue or margins than any other CRM technology”

The Yankee Group estimated that more than one billion dollars would be spent on these systems in 2007

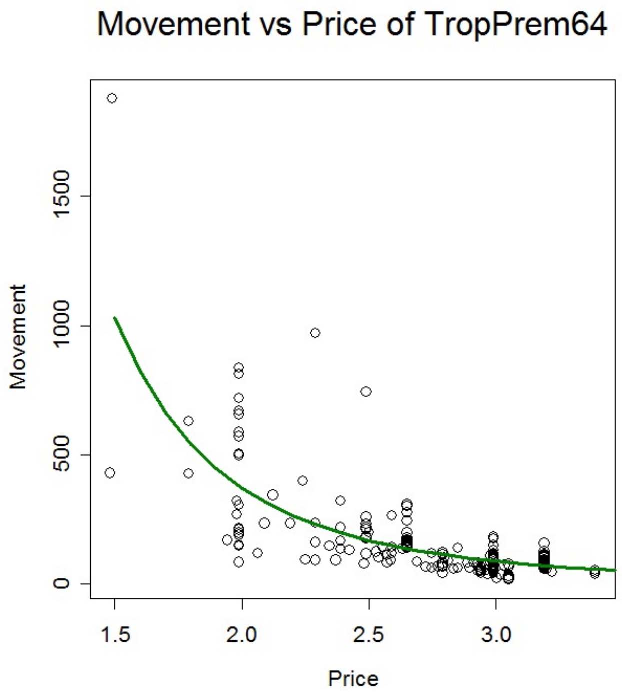
Anecdotal reports suggest increases in gross margins in the range of 2-8%, retailers typically have gross margins of about 25% and have annual revenues of $2.5 trillion. Suggests benefit for retailers alone would be between $12.5b and $50b annually.

Weekly Movement and Price for Tropicana Premium 64 in one Chicago Supermarket

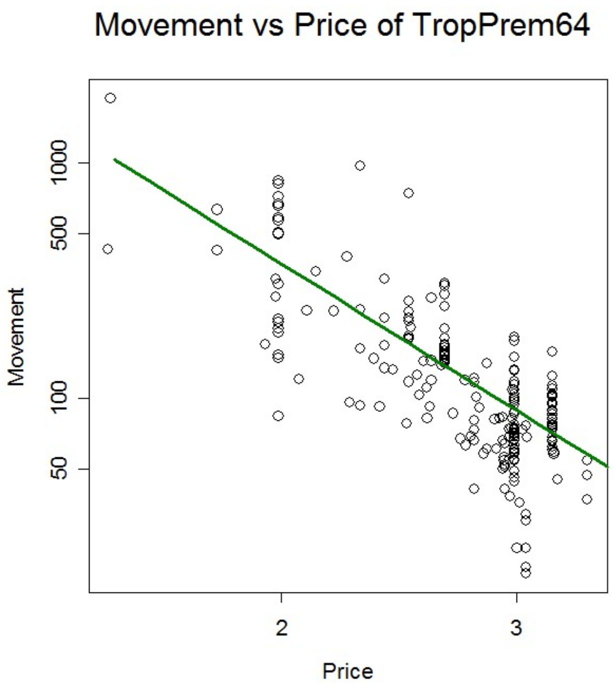


Predictive Modeling with Regression

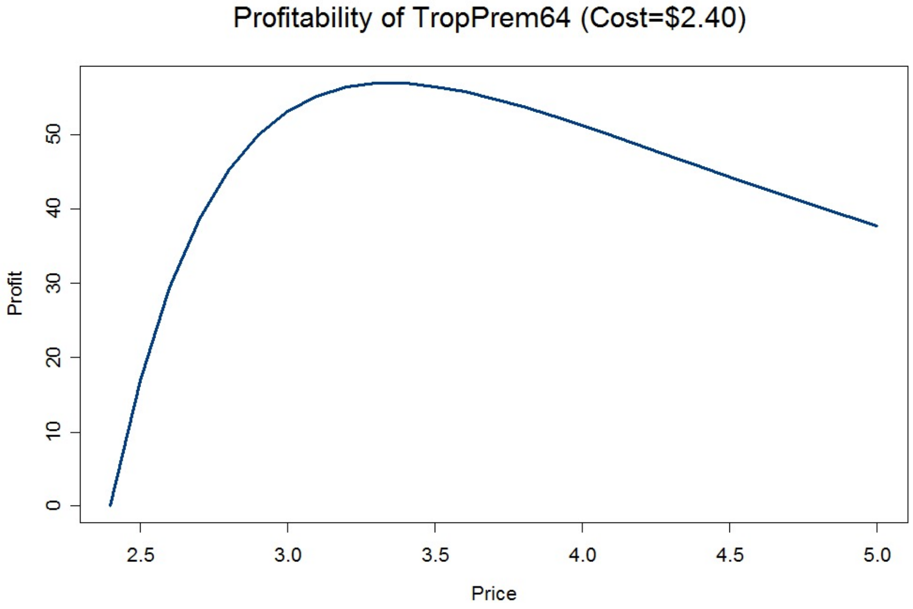
Linear Units gives Nonlinear relationship



Log Units gives Linear Relationship



Optimal Product Pricing



Profits:

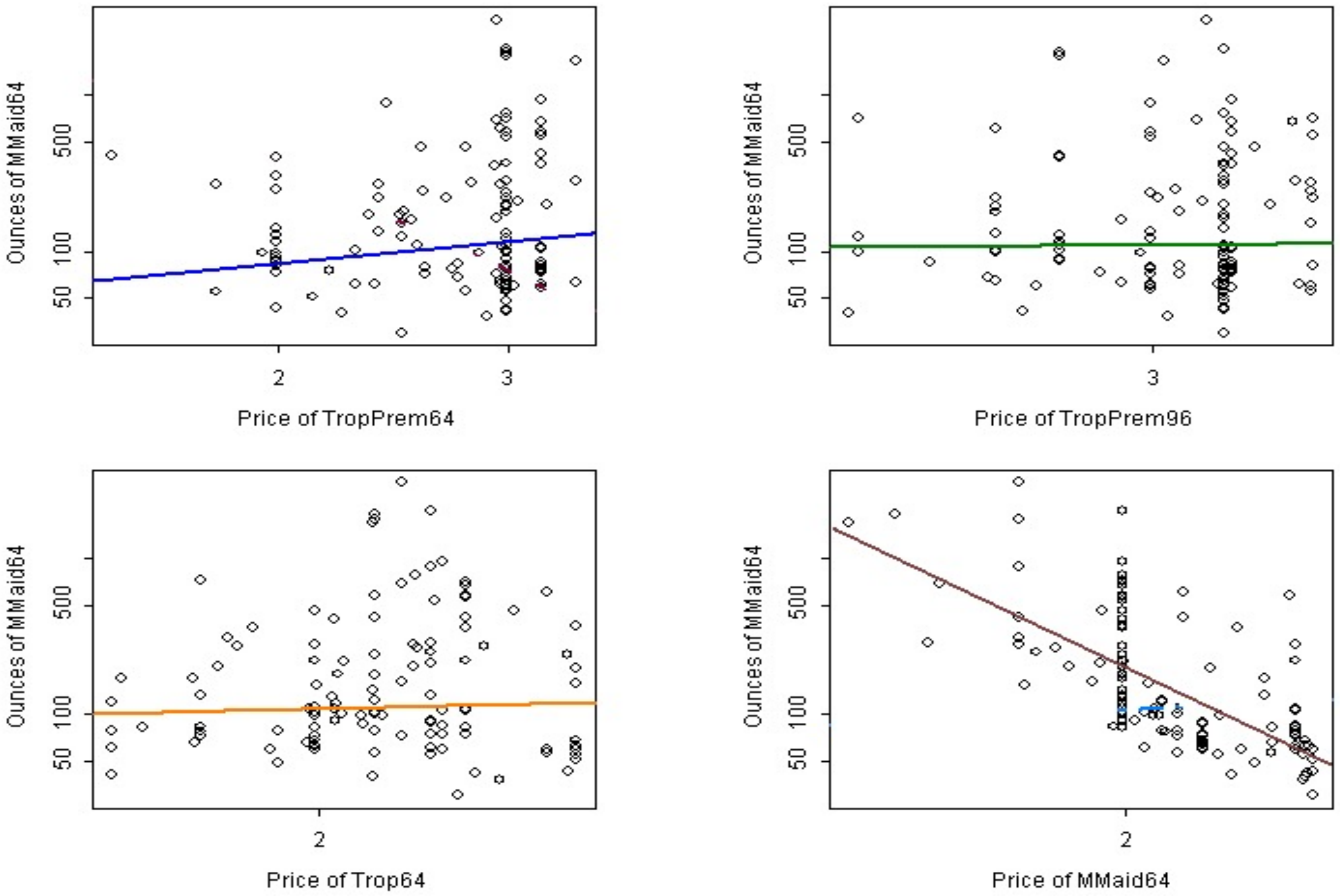
$$\Pi = (p - c)q$$

Optimal pricing rule:

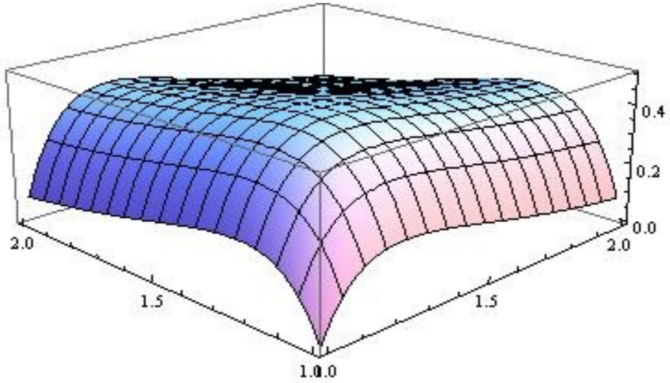
$$p^* = \frac{\beta}{\beta + 1} c$$

Where price elasticity measures demand responsiveness to price changes:

$$\beta = \frac{\partial q}{\partial p} \frac{p}{q} \approx \frac{\%\Delta q}{\%\Delta p}$$


Ounces of MMaid64
500
100
50
2
3
Price of TropPrem64
Ounces of MMaid64
500
100
50
3
Price of TropPrem96
Ounces of MMaid64
500
100
50
2
Price of Trop64
Ounces of MMaid64
500
100
50
2
Price of MMaid64

Optimal Product Line Pricing



Total Profits:

$$\Pi = \sum_{i=1}^{M} (p_i - c_i) q_i$$

Optimal pricing rule:

$$p_i^* = \frac{\beta_{ii}}{\beta_{ii} + 1 + \sum_{j \neq i} \mu_j \beta_{ji} \, s_j / s_i} c_i$$

Where cross price elasticities measure competitive effects:

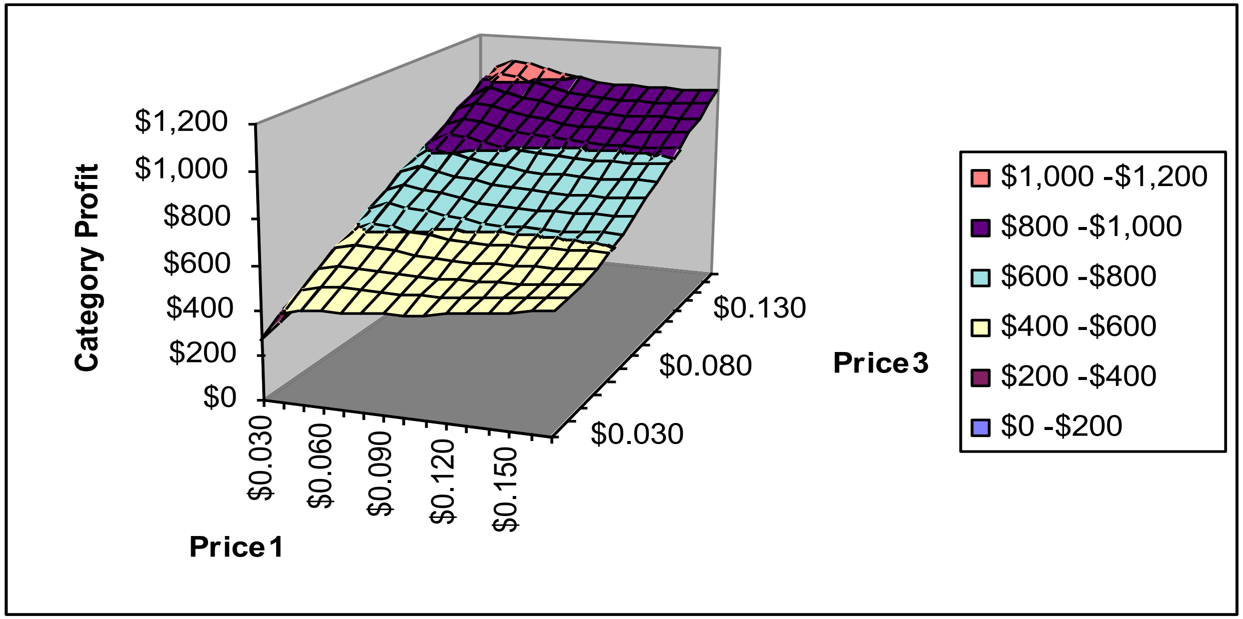
$$\beta_{ij} = \frac{\partial q_i}{\partial p_i} \frac{p_i}{q_i} \approx \frac{\%\Delta q_i}{\%\Delta p_j},$$

$$\mu_i = \frac{p_i - c_i}{p_i}, \ s_i = \frac{p_i q_i}{\sum_j p_j q_j}$$

Joint Price Optimization

Problem: Category Profits (the sum of products from each of the products) rises if all prices go up

Intuition: Model predicts substitution is constant, but will substitution really be the same for a $10 carton juice versus one at $1,000?



Business Rules

Current pricing solutions frequently implement constraints that reflect “business rules”, which codify manager knowledge:

1. Allowed number and frequency of markdowns (e.g., at least a week between two consecutive markdowns)
2. Min-max discount levels or maximum lifetime discount
3. Minimum number of weeks before an initial markdown can occur
4. Types of markdowns allowed (e.g., 10%, 25%, …) or the permissible set of prices
5. The “family” of items that must be marked down together

Source: Elmaghraby and Keskinocak (2003; Management Science)

Conclusion

Current optimal pricing practitioners and researchers are introducing information in an ad hoc manner by relying upon “business rules” or constraints.

An appropriate Bayesian data mining methods avoids ad hoc “corrections” to the predictions (or posterior) and says that the knowledge should be brought in a priori

Leads to better decision support systems that reflect “expert” knowledge efficiently.

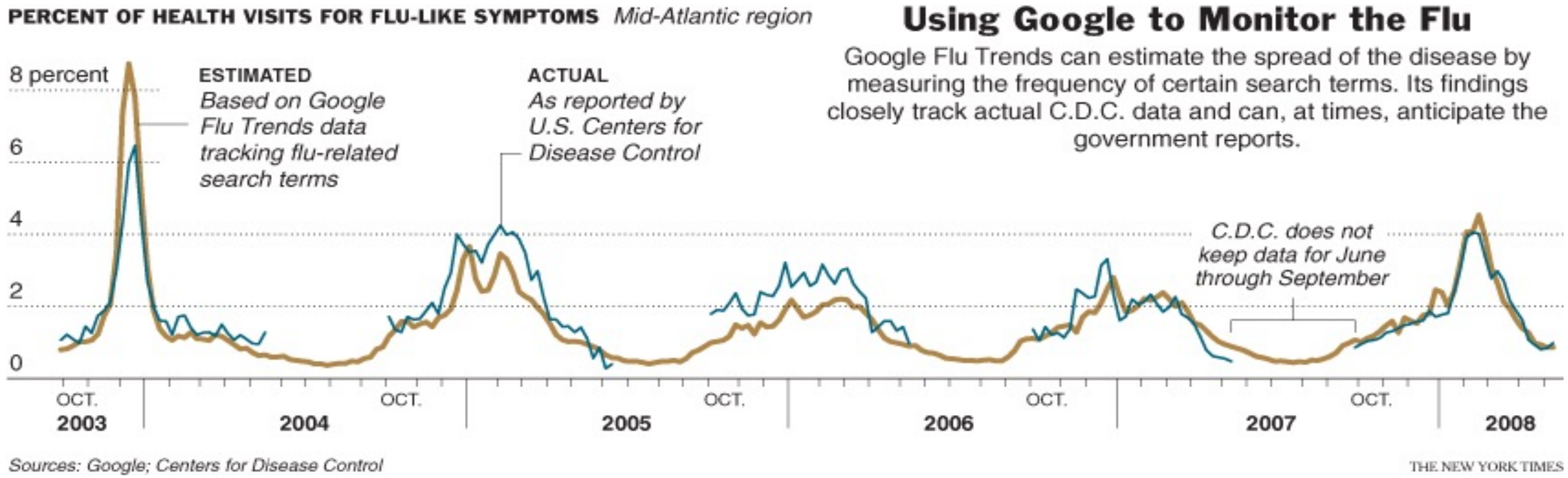
3. Predicting Flu Trends

Source: Lazer, Kennedy, King and Vespignani (2014), “The Parable of Google Flu: Traps in Big Data Analysis”, Science

Google Flu Trends

In 2008 Google released an experiment called Flu Trends to predict the number of flu cases (as reported by the CDC) using searches from about 40 flu-related queries

“The earlier the warning, the earlier prevention and control measures can be put in place, and this could prevent cases of influenza,” said Dr. Lyn Finelli, lead for surveillance at the influenza division of the C.D.C. Source: NYT 12-11-2008



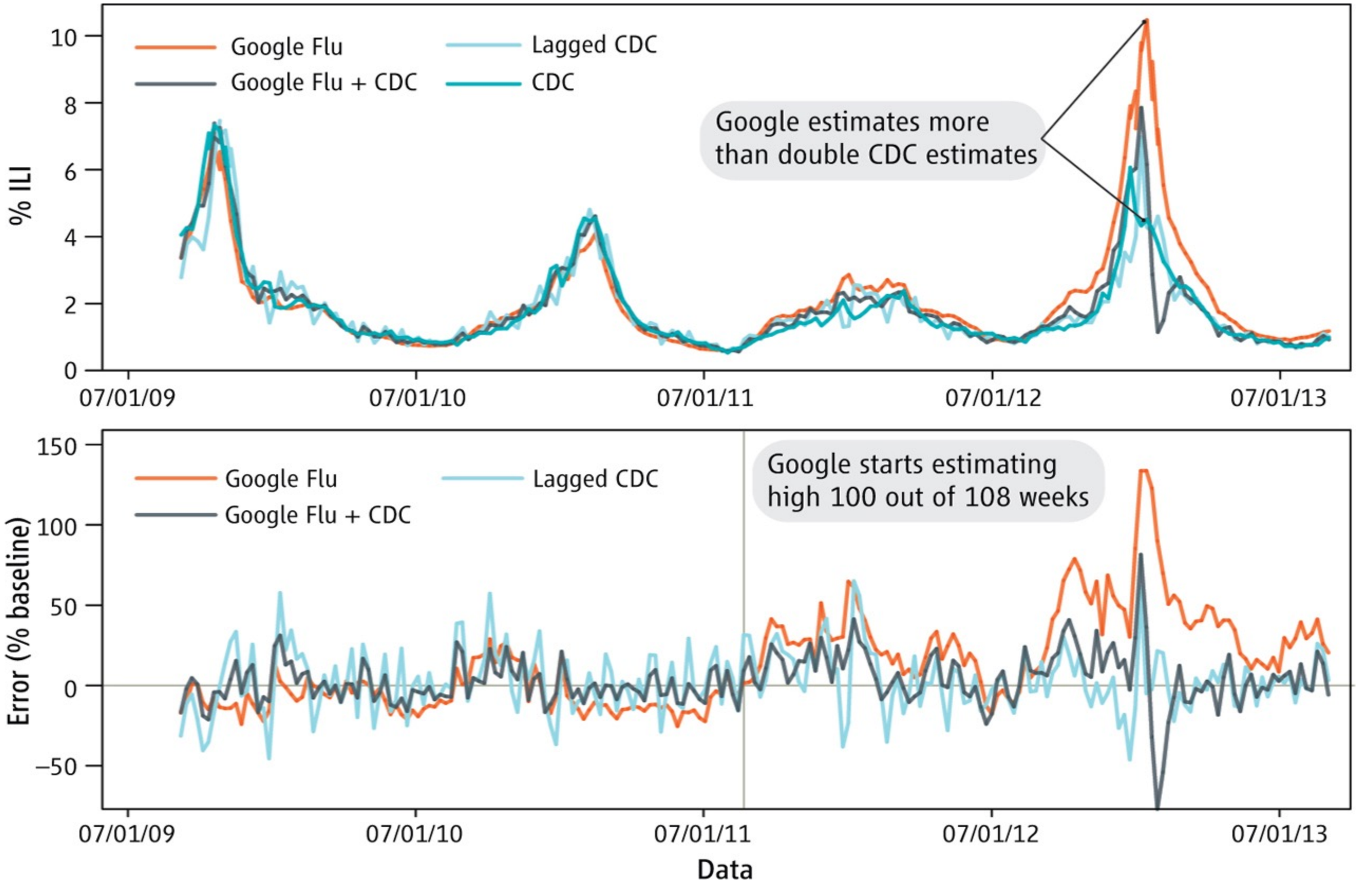
Using Google to Monitor the Flu

Google Flu Trends can estimate the spread of the disease by measuring the frequency of certain search terms. Its findings closely track actual C.D.C. data and can, at times, anticipate the government reports.

Sources: Google; Centers for Disease Control

THE NEW YORK TIMES

GFT overestimation.GFT overestimated the prevalence of flu in the 2012–2013 season and overshot the actual level in 2011–2012 by more than 50%.



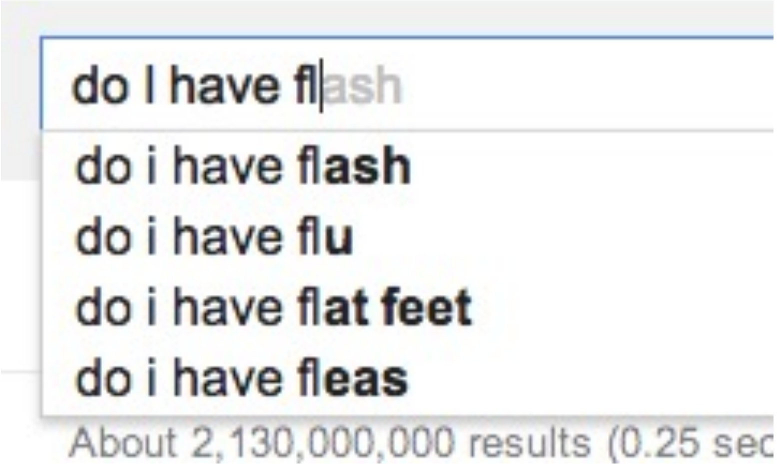
D Lazer et al. Science 2014;343:1203-1205

Published by AAAS

Big Data Headache

Google’s own autosuggest feature may have driven more people to make flu-related searches during the 2013 flu season

Which misled its forecasting system and overstate the number of cases



do I have flash

do i have flash

do i have flu

do i have flat feet

do i have fleas

About 2,130,000,000 results (0.25 sec

Conclusion

“GFT was like the bathroom scale where the spring slowly loosens up and no one ever recalibrated”, David Lazer (Northeastern University)

Google constantly makes tweaks to its general search algorithm averaging more than one a day, and the introduction of its “autosuggest” feature may led to more searches on influenza

Lesson: the underlying patterns within social media and online behavior change, need to recalibrate their accuracy.

4. Problems with Big Data

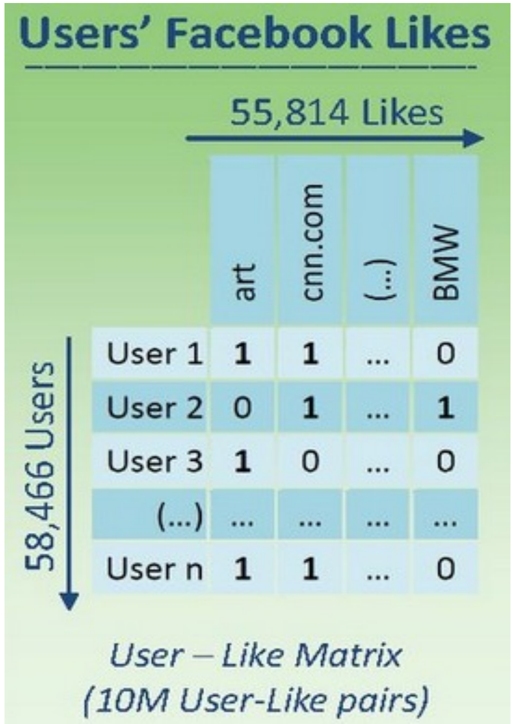
Selection and Privacy

Porsche's Cayenne Introduction

Example Responses from Rennlist (Porsche enthusiast discussion board)

- "It makes me embarrassed of owning a Porsche … kinda like that relative you don't want to admit sharing the same bloodline. Ugh!"
- "There just aint nothing porsche in SUV"
- "I just felt really sad. Now 'soccer moms' can drive their kids around in a Porsche. Pretty sad. Look at the level that Porsche has been brought down to."

What Facebook likes tell about you…



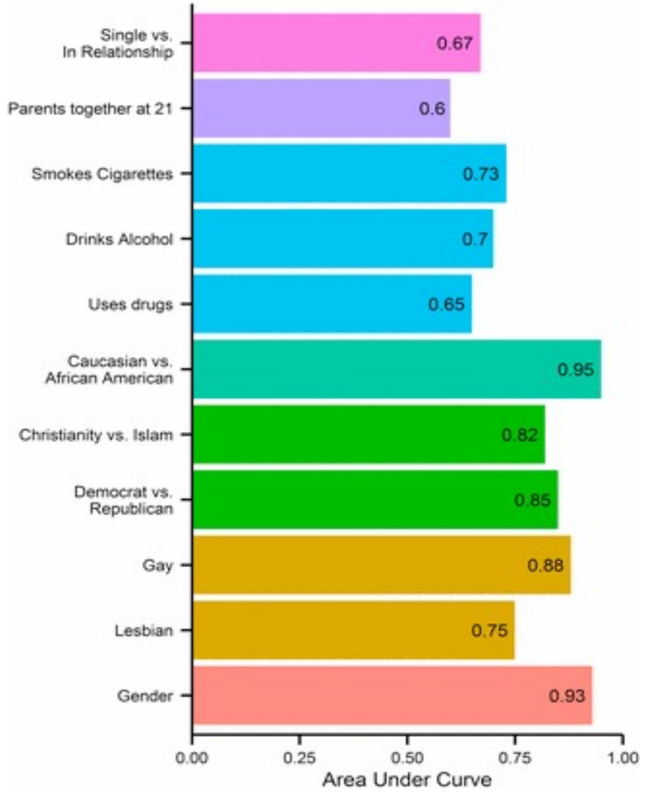
Users' Facebook Likes

55,814 Likes

58,466 Users

	art	cnn.com	(…)	BMW
User 1	**1**	**1**	…	0
User 2	0	**1**	…	**1**
User 3	**1**	0	…	0
(…)	…	…	…	…
User n	**1**	**1**	…	0

User – Like Matrix
(10M User-Like pairs)



Source: Kosinski, Stillwell and Graepel (2013), PNAS

5. Managers and Culture

New Vantage Partners’ 2019 Big Data and AI Executive Survey

64 c-level tech and business executives (eg., American Express, Ford Motor, General Electric, General Motors, and Johnson & Johnson):

- 72% of survey participants report that they have yet to forge a data culture
- 69% report that they have not created a data-driven organization
- 53% state that they are not yet treating data as a business asset
- 52% admit that they are not competing on data and analytics

Percentage of self-identified “data-driven” firms has declined in each of the past 3 years: from 37% in 2017 to 32% in 2018 to 31% in 2019

https://hbr.org/2019/02/companies-are-failing-in-their-efforts-to-become-data-driven

Conclusions

Summary

Big Data has a huge potential to shape our lives through changes in business, government, and science, or society in general

- It is a by-product of our electronic lives and generally the reason it is collected has nothing to do with analysis or learning
- The goal of data science is to tap into this potential of data (big or small) to make decisions

But remember analytical approaches fail because

- Bad/poor/biased data
- Bad/poor/biased models
- Bad/poor/biased managers who interface with the models

Limitations of Big Data and Data Science

The promise of Big Data is that we can solve problems faster, cheaper and better.

The problem is that Big Data is still just data, and we need to know its biases

- Historical data may not be representative of future data
- Participants in social media data may not be representative of society
- The collection and use of “Big Data” changes through time

Knowledge (managerial or theoretical) is still useful

- The data we observe is influenced by our past decisions, which is a function of our “models”, need to consider this feedback relationship

Disadvantages of Data Based Approaches

It may not be representative

- Who writes reviews? Really excited customers and really disappointed ones

Data quality may be poor

- Consumers generate Big Data for themselves not for data miners

Privacy and confidentiality issues

- How can we protect consumers?

Difficult to assess accuracy and uncertainty

The past may not be representative of the future