

# DS 555 Data Science and Business Strategy

---

## *INTRODUCTION*

– O.Örsan Özener

# Quote of the day

---

*“The purpose of scientific computing is insight, not numbers”.*

– Richard Hamming

# Big Data & Business

---

## Data:

- Who collects it?
- Who owns/manages it?
- Who uses it?



The days when business stakeholders could relinquish control of data and analytics to IT are over.

# Big Data & Business

---

## Data & Business Stakeholders:

- Leverage Big Data
- Optimize key business process
- Uncover new monetization opportunities
- Create (new) competitive differentiation angles

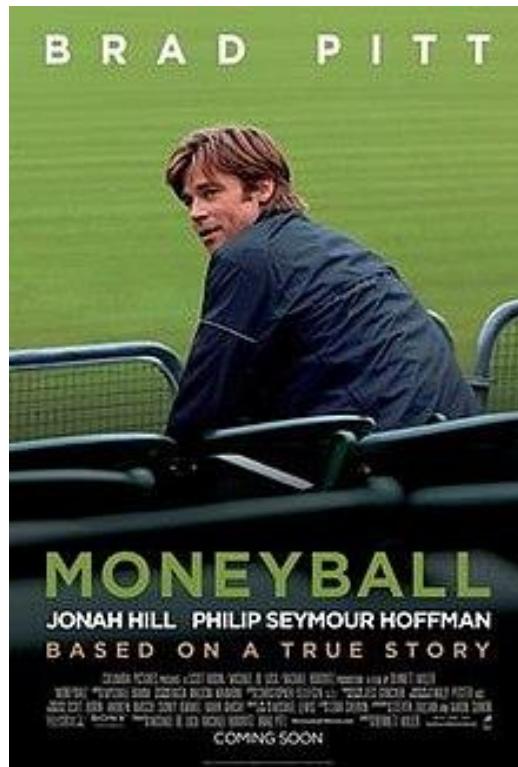
**Who will do this? Data Scientist?  
Managers?**

# Moneyball

---

“People in both fields operate with beliefs and biases. To the extent you can eliminate both and replace them with data, you gain a clear advantage.”

Michael Lewis, Moneyball: The Art of Winning an Unfair Game



# Circulation Estimation

## Newspaper Sales Estimation:



If you stare into the abyss, the abyss stares back at you

- Friedrich Nietzsche

# Circulation Estimation

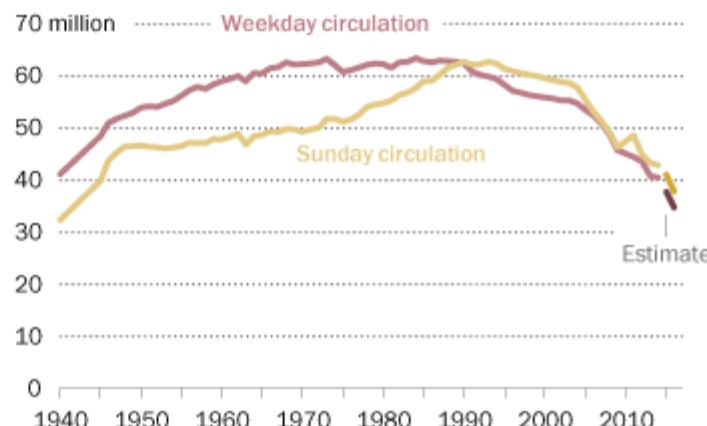
## Newspaper Sales Estimation: Humans vs Machines

### Newspaper circulation continues to fall

*Total circulation for U.S. daily newspapers*

Note: Line break indicates switch to estimates. No data for 1941-44 and 2010. See full post for details on how estimate was calculated.

Source: Editor & Publisher (through 2014); estimation based on Pew Research Center analysis of Alliance for Audited Media data (2015-2016.)



PEW RESEARCH CENTER

# Demand Estimation

---

**Durable Goods Demand Estimation:** Which data? From which source?



# Internet Service Provider

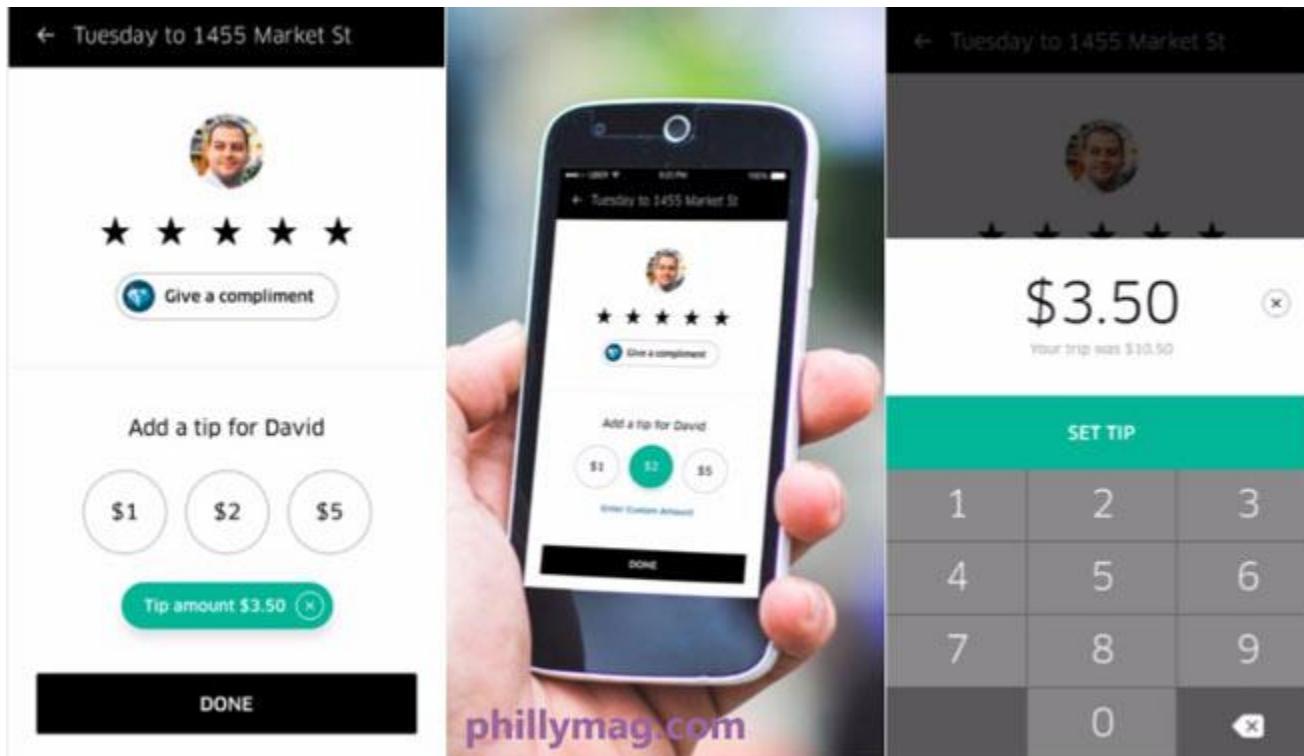
---

**Customer Data:** Lots of Data? What to do?



# Tipping

## Tipping: Does it work?



# Data Lies ?

## Elections and Polls: Why polls failed?



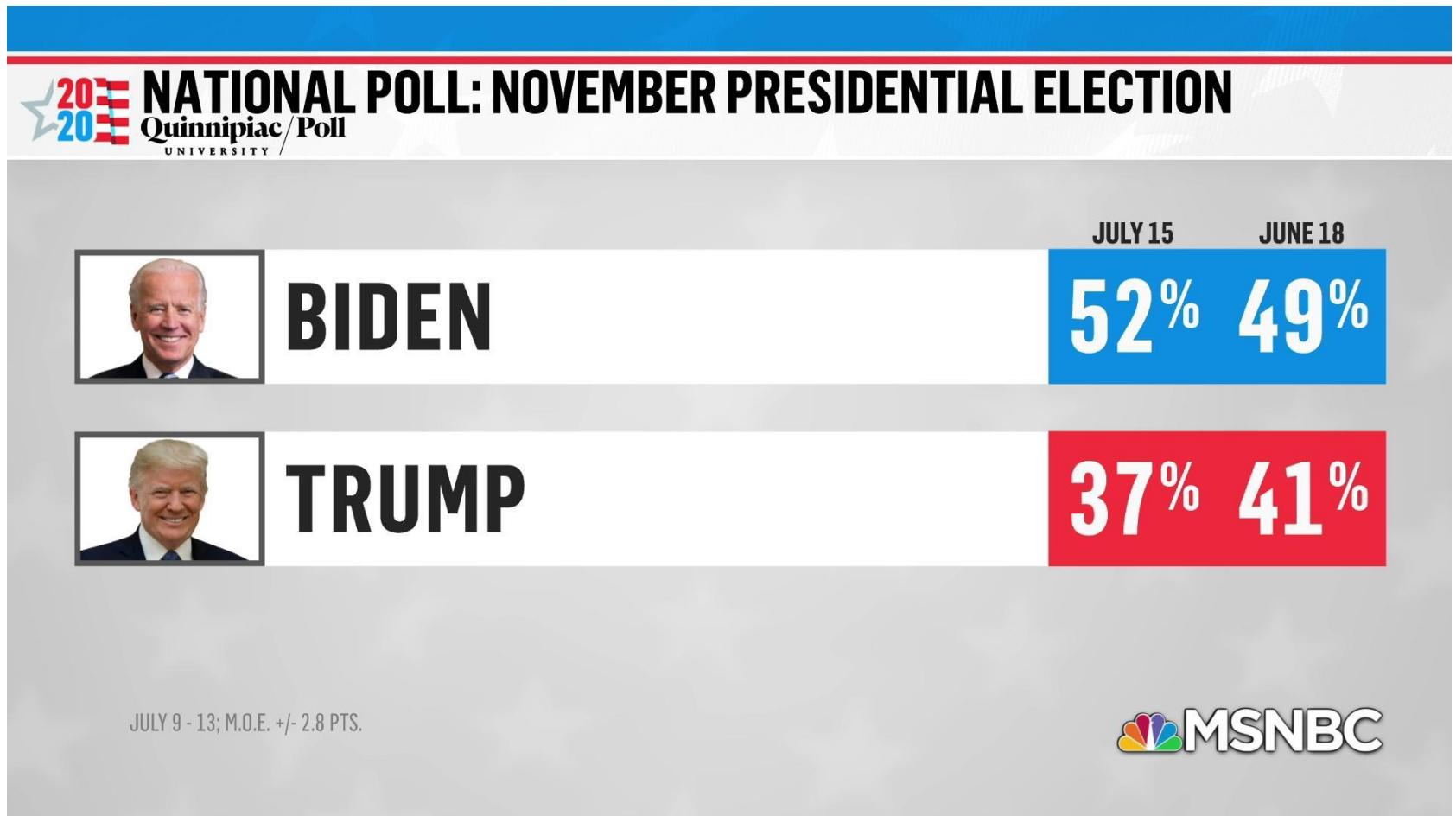
# Data Lies ?

Elections and Polls: Different estimation strategies? Different Data Sources?



# Data Lies ?

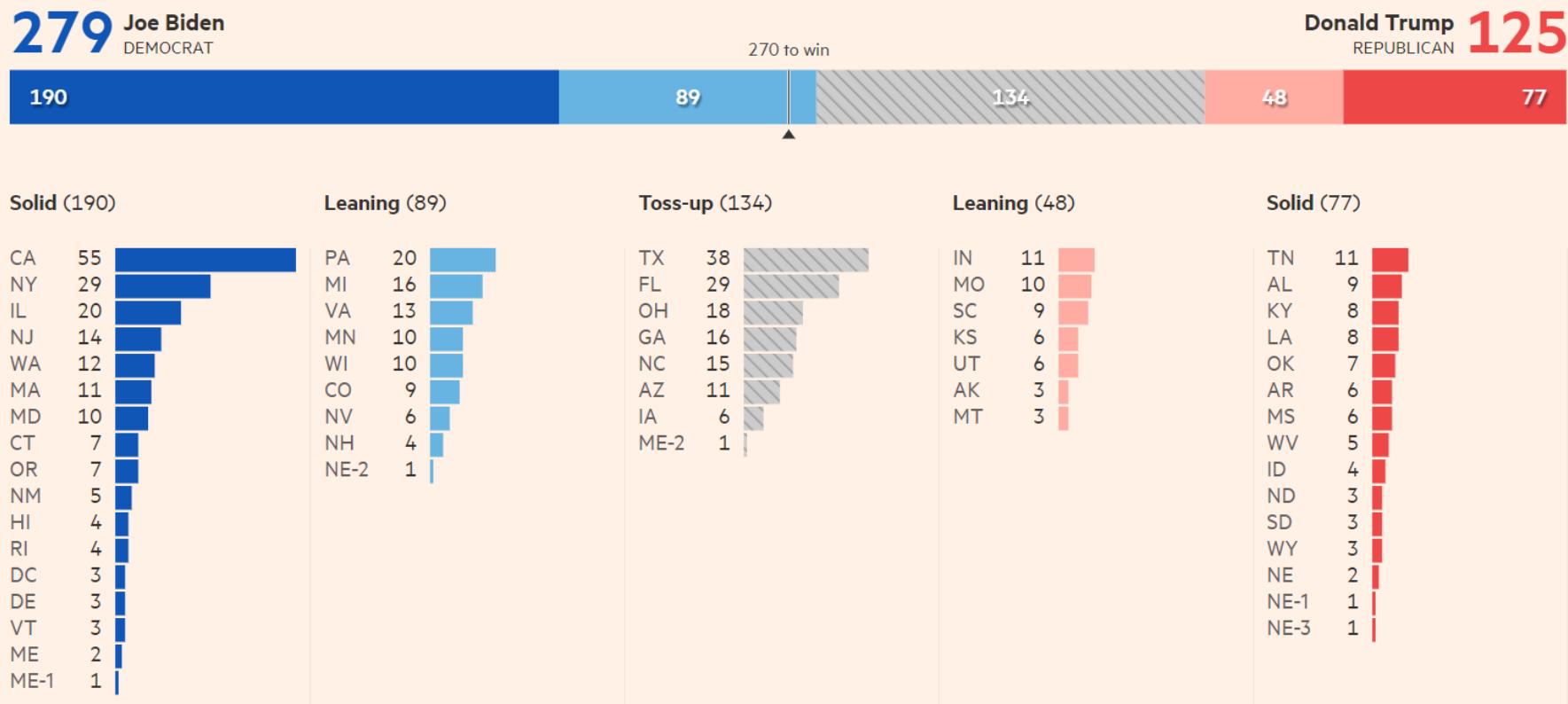
## Elections and Polls: Dynamic Environment?



# Data Lies ?

## Elections and Polls: Current Situation

If the election were held today, the latest polls suggest this outcome in the electoral college:



# Data Lies ?

---

## Elections and Polls: How Elections Work?

	Party A	Party B	Party C	Party D	Party E	Party F
Votes	470000	160000	158000	120000	61000	31000

	Party A	Party B	Party C	Party D	Party E	Party F
Votes	470000	160000	158000	120000	61000	31000
1	470000	160000	158000	120000	61000	31000
2	235000	80000	79000	60000	30500	15500
3	156667	53333	52667	40000	20333	10333
4	117500	40000	39500	30000	15250	7750
5	94000	32000	31600	24000	12200	6200
6	78333	26667	26333	20000	10167	5167
7	67143	22857	22571	17143	8714	4429
8	58750	20000	19750	15000	7625	3875
9	52222	17778	17556	13333	6778	3444
10	47000	16000	15800	12000	6100	3100

# Data Lies ?

---

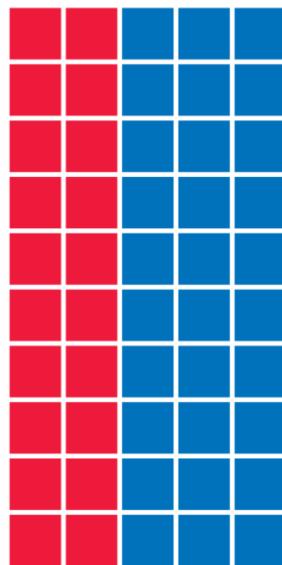
## Elections and Polls: How Elections Work?

	Party A	Party B	Party C	Party D	Party E	Party F
Votes	470000	160000	158000	120000	61000	31000
1	470000	160000	158000	120000	61000	31000
2	235000	80000	79000	60000	30500	15500
3	156667	53333	52667	40000	20333	10333
4	117500	40000	39500	30000	15250	7750
5	94000	32000	31600	24000	12200	6200
6	78333	26667	26333	20000	10167	5167
7	67143	22857	22571	17143	8714	4429
8	58750	20000	19750	15000	7625	3875
9	52222	17778	17556	13333	6778	3444
10	47000	16000	15800	12000	6100	3100
	Party A	Party B	Party C	Party D	Party E	Party F
1	1	3	4	6	13	25
2	2	9	10	14	26	42
3	5	16	17	20	33	48
4	7	20	22	27	43	51
5	8	23	24	30	46	54
6	11	28	29	34	49	56
7	12	31	32	39	50	57
8	15	34	36	44	52	58
9	18	37	38	45	53	59
10	19	40	41	47	55	60
Repre.	5	2	2	1	0	0

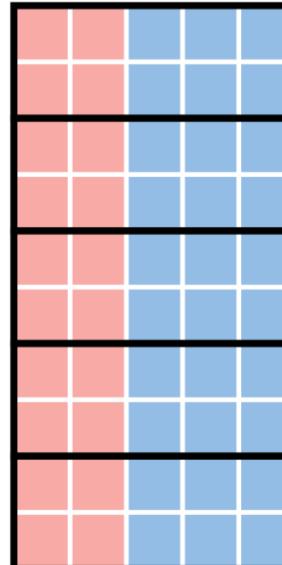
# Data Lies ?

## Elections and Polls: Manipulating the Outcome

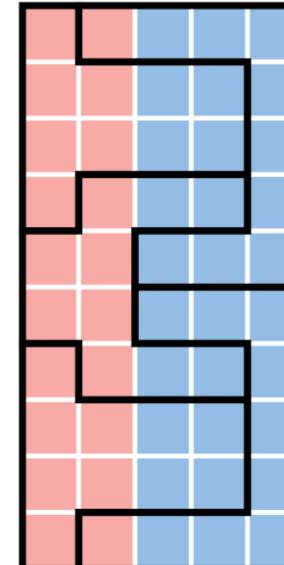
### HOW TO STEAL AN ELECTION



50 PRECINCTS  
60% BLUE  
40% RED



5 DISTRICTS  
5 BLUE  
0 RED  
BLUE WINS

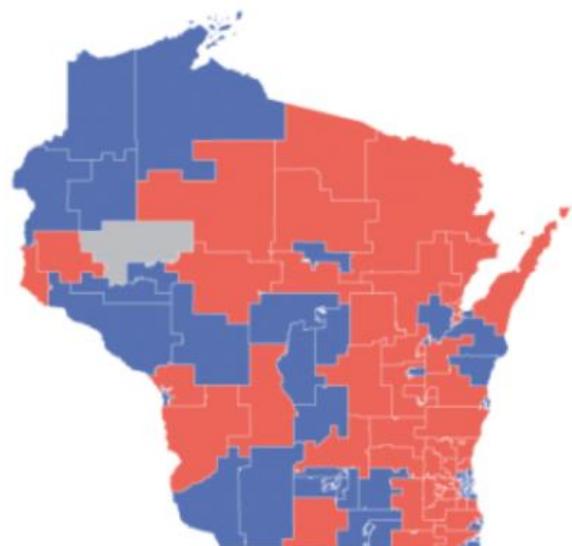


5 DISTRICTS  
3 RED  
2 BLUE  
RED WINS

# Data Lies ?

## Elections and Polls: Manipulating the Outcome

Wisconsin District Plan  
2008 Election Results

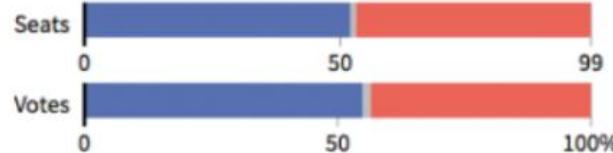


**DEMOCRATS**

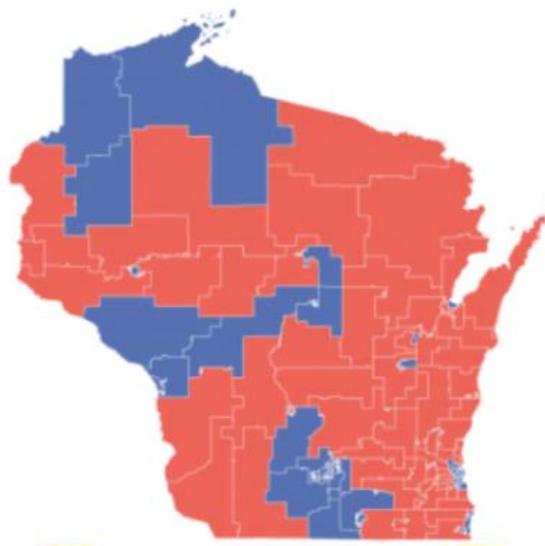
Seats: 52  
Votes: 1,478,830

**REPUBLICANS**

Seats: 46  
Votes: 1,172,877



Wisconsin District Plan  
2012 Election Results

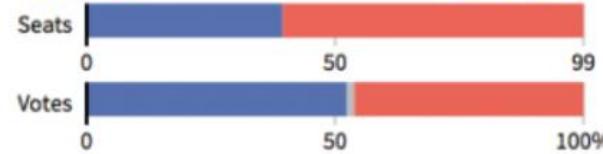


**DEMOCRATS**

Seats: 39  
Votes: 1,417,359

**REPUBLICANS**

Seats: 60  
Votes: 1,249,562



# Data Lies ?

## Elections and Polls: Manipulating the Outcome



Princeton Gerrymandering Project

Practical bug fixes for democracy

Tests

State-Level Reforms

Resources

We bridge the gap between mathematics and the law to achieve fair representation through redistricting reform.



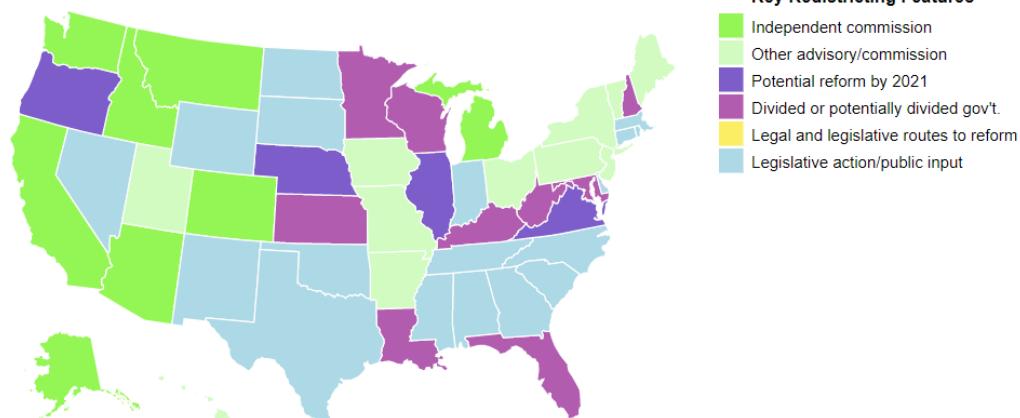
Sat Oct 3rd, 2020



From the blog: [Sam on The New Abnormal \(again\)](#)

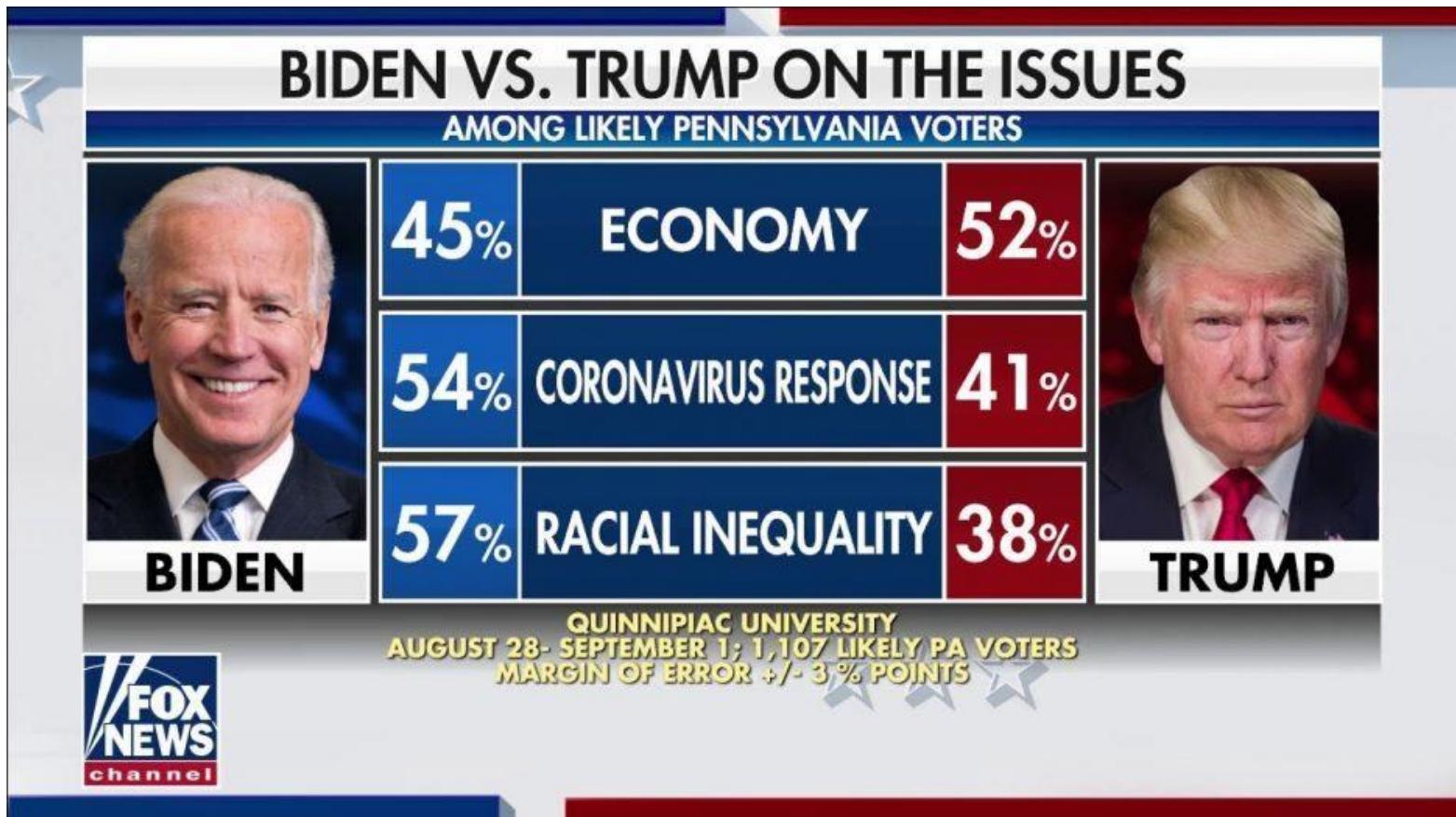
Molly Jong-Fast and Rick Wilson had me back on their podcast, The New Abnormal! So pleased that they had me back. Molly is total fun, and so disrespectful. We talked elections of course, with a focus on Moneyball 2020. [\[Read more →\]](#)

### State-by-State Redistricting Reform: The Local Routes



# Data Lies ?

## Elections and Polls: Ulterior Motive?



# Alcohol and Accidents

---

**Hypothesis:** Everything starts with a hypothesis

**Data:** A year in a U.S. state

- 48 U.S. states, so  $n = \# \text{ of entities} = 48$
- 7 years (1982,..., 1988), so  $T = \# \text{ of time periods} = 7$
- Balanced panel, so total # observations =  $7 \times 48 = 336$

**Variables:**

- Traffic fatality rate (# traffic deaths in that state in that year, per 10,000 state residents)
- Tax on a case of beer
- Other (legal driving age, drunk driving laws, etc.)

# Alcohol and Accidents

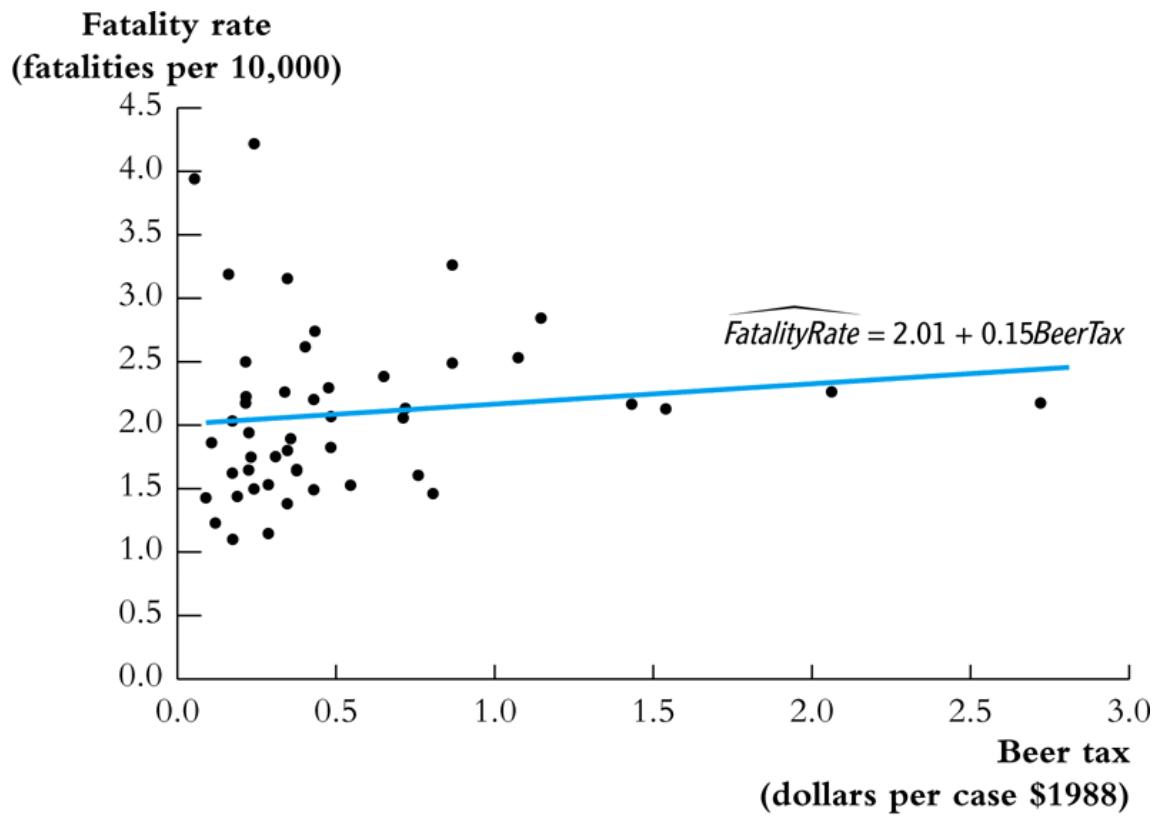
---

Standard Regression:

$$FatalityRate_p = \beta_0 + \beta_1 BeerTax_p + u_p$$

$$FatalityRate_p = 1.85 + 0.36 BeerTax_p + u_p$$

# Alcohol and Accidents



(a) 1982 data

Higher alcohol taxes, more traffic deaths?

# Alcohol and Accidents

---

## Other Factors:

- Other factors that determine traffic fatality rate:
- Quality (age) of automobiles
- Quality of roads
- “Culture” around drinking and driving
- Density of cars on the road

# Alcohol and Accidents

---

Regression with Fixed Effects:

$$FatalityRate_{it} = \beta_0 + \beta_1 BeerTax_{it} + \beta_2 Z_i + u_{it}$$

$$FatalityRate_{it} = \beta_0 - 0.66 BeerTax_{it} + \beta_2 Z_i + u_{it}$$

# By the way... how much do beer taxes vary?

## Beer Taxes in 2005

Source: Federation of Tax Administrators

<http://www.taxadmin.org/fta/rate/beer.html>

	<b>EXCISE TAX RATES (\$ per gallon)</b>	<b>SALES TAXES APPLIED</b>	<b>OTHER TAXES</b>
Alabama	\$0.53	Yes	\$0.52/gallon local tax
Alaska	1.07	n.a.	\$0.35/gallon small breweries
Arizona	0.16	Yes	
Arkansas	0.23	Yes	under 3.2% - \$0.16/gallon; \$0.008/gallon and 3% off-10% on-premise tax
California	0.20	Yes	
Colorado	0.08	Yes	
Connecticut	0.19	Yes	
Delaware	0.16	n.a.	
Florida	0.48	Yes	2.67¢/12 ounces on-premise retail tax

Georgia	0.48	Yes	\$0.53/gallon local tax
Hawaii	0.93	Yes	\$0.54/gallon draft beer
Idaho	0.15	Yes	over 4% - \$0.45/gallon
Illinois	0.185	Yes	\$0.16/gallon in Chicago and \$0.06/gallon in Cook County
Indiana	0.115	Yes	
Iowa	0.19	Yes	
Kansas	0.18	--	over 3.2% - {8% off- and 10% on-premise}, under 3.2% - 4.25% sales tax.
Kentucky	0.08	Yes*	9% wholesale tax
Louisiana	0.32	Yes	\$0.048/gallon local tax
Maine	0.35	Yes	additional 5% on-premise tax

Maryland	0.09	Yes	\$0.2333/gallon in Garrett County
Massachusetts	0.11	Yes*	0.57% on private club sales
Michigan	0.20	Yes	
Minnesota	0.15	--	under 3.2% - \$0.077/gallon. 9% sales tax
Mississippi	0.43	Yes	
Missouri	0.06	Yes	
Montana	0.14	n.a.	
Nebraska	0.31	Yes	
Nevada	0.16	Yes	
New Hampshire	0.30	n.a.	
New Jersey	0.12	Yes	
New Mexico	0.41	Yes	

New York	0.11	Yes	\$0.12/gallon in New York City
North Carolina	0.53	Yes	\$0.48/gallon bulk beer
North Dakota	0.16	--	7% state sales tax, bulk beer \$0.08/gal.
Ohio	0.18	Yes	
Oklahoma	0.40	Yes	under 3.2% - \$0.36/gallon; 13.5% on-premise
Oregon	0.08	n.a.	
Pennsylvania	0.08	Yes	
Rhode Island	0.10	Yes	\$0.04/case wholesale tax
South Carolina	0.77	Yes	
South Dakota	0.28	Yes	
Tennessee	0.14	Yes	17% wholesale tax
Texas	0.19	Yes	over 4% - \$0.198/gallon, 14% on-premise and \$0.05/drink on airline sales

Utah	0.41	Yes	over 3.2% - sold through state store
Vermont	0.265	no	6% to 8% alcohol - \$0.55; 10% on-premise sales tax
Virginia	0.26	Yes	
Washington	0.261	Yes	
West Virginia	0.18	Yes	
Wisconsin	0.06	Yes	
Wyoming	0.02	Yes	
Dist. of Columbia	0.09	Yes	8% off- and 10% on-premise sales tax
U.S. Median	\$0.188		

# Alcohol and Accidents

---

Regression with Fixed Effects:

$$FatalityRate_{it} = \beta_0 + \beta_1 BeerTax_{it} + \beta_2 Z_i + \beta_3 S_t + u_{it}$$

$$FatalityRate_{it} = \beta_0 - 0.64 BeerTax_{it} + \beta_2 Z_i + \beta_3 S_t + u_{it}$$

# NBA and Poverty

---

Hypothesis: But not too much!!!



NBA player ~ grow up  
poor or middle-class?

# NBA and Poverty

---

**Data:** No real/reliable data

- Birthplace ~ Average income level of the county
- Racial factors
- Family Backgrounds
- Name ???

**Result:** Not True? Why?

- Nutrition and other privileges
- Social skills

# Outline of the Course

---

## Subjects:

- Big Data Business Cycles
- Big Data Strategy Document
- Big Data Organization
- When Data Lies
- Data Science and Strategic Thinking
- Data Science and Business Intelligence
- Sources of Business Data
- Customer Oriented Thinking
- Data Science and Strategic Decision Making – Optimization

# Your Expectations

---



# Administrative Details

---

**Grading:** Exams, ???

Homework ???

Project ???

**Late Submission:** No Late Submission

**Cheating:**

# **DS 555 Data Science and Business Strategy**

---

## *BIG DATA STRATEGY*

– O.Örsan Özener

# Big Data & Business

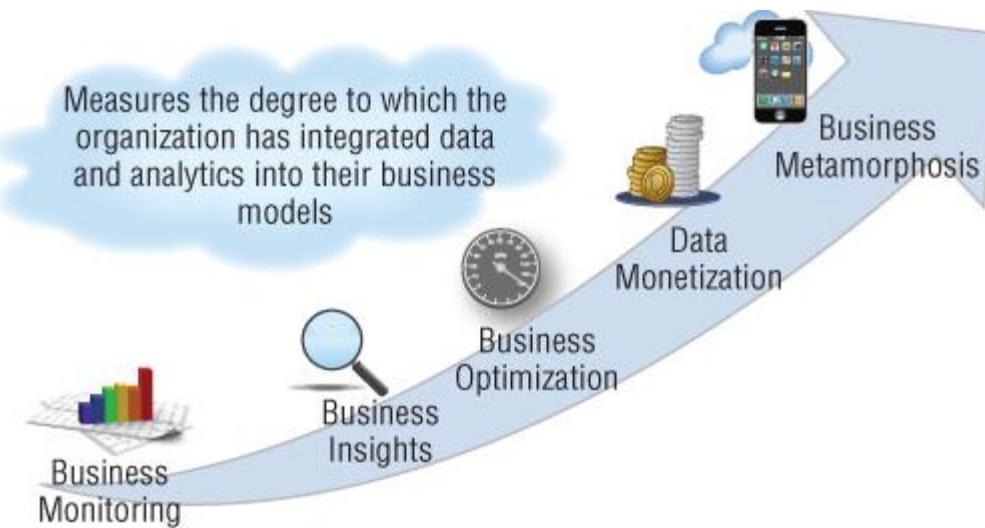
## Big Data & Business:

**How effective is an organization at integrating data and analytics into the business models?**



# Big Data & Business

## Big Data Business Model Maturity Index:



# Phase I: Business Monitoring

---

## Business Monitoring:

- Deploying Business Intelligence (BI) and data warehousing solutions to monitor ongoing business performance
- Create reports and dashboards
- Investment on key operational data sources
- Limited Monetary Gains
- Examples?

# Phase II: Business Insights

---

## Business Insights:

- Using internal and external structured and unstructured data
- Predictive analytics to uncover customer, product, and operational insights buried in the data.
- Uncovering the **hidden truth** in the data
- Need to exploit the economies of scale in data operations
  - Gathering
  - Integrating
  - Storing/Managing
  - Analyzing
  - Acting on Data
- Examples?

# Phase II: Business Insights

---

## Sub-Phases:

- Access to All of the Organization's Transactional and Operational Data
- Access to Internal and External Unstructured Data
- Exploiting Real-Time Analytics
- Integrating Predictive Analytics

## Business Insights:

- Strategic
- Actionable
- Material

# Phase II: Business Insights

---

## Challenges:

- Most Difficult Phase
- Motto: **Think Differently**
- Traditional Approach vs Innovative Approach

# Phase III: Business Optimization

---

## Business Optimization:

- Predictive Analytics vs Prescriptive Analytics
- Motto: **Tell me what to do!**
- Delivers actionable recommendations
- Influence/Manipulate
  - Customers
  - Market
  - Suppliers
- **Examples?**

# Phase IV: Data Monetization

---

## Data Monetization:

- Leverage insights and optimization to create revenue opportunities
- NOT about IT or BI
- Motto: **Show me the money!**
- Examples?

# Phase V: Business Metamorphosis

---

## Business Metamorphosis:

- Ultimate goal
- Transform Organization's Business Model
- Motto: **A new World!**
- Might require technology/legislation etc.
- Examples?



# Phase IV & V: DM & BM

---

## Business Models for DM and BM

- Return On Advantage Model
  - Customer Targeting: Loyalty, Cross-sell, Up-sell
  - Risk Mitigation & Fraud Detection
- Premium Service Model
- Differentiator Model
- Syndication Model

# Phase IV & V: DM & BM

## Critical Discovery Concepts

### AGGREGATION

Analyses of aggregated data are most often central to the identification of key data analytic signals. Patterns, shifts, pivots, and anomalies become crystallized and observable where data is aggregated by various dimensions and explored from different perspectives or frames of reference.

### TRIANGULATION

Powerful technique that facilitates validation of data and verifying insights through cross-verification from two or more sources. New data sources may be created with this method, including a "1+1=3" greater sum phenomenon that is achieved when data is combined and correlated in unique ways.

### FRAME OF REFERENCE

The insight that can be extracted is a direct function of the perspective of the comparisons and exploration that is applied to the situation. Look at a coffee cup from a direction and you may see the logo of the store. Turn it or change your frame of reference, and you may see the name of the individual who ordered the cup on the other side. Changing your frame of reference adds to the depth of the insight that can be considered.

### PRIVACY PRESERVATION

There needs to be a balance between risk control and value preservation. Often tactical data monetization initiatives struggle when data privatization factors are considered. This dilemma is amplified when data monetization is treated as only a "transactional" driven opportunity. Broadening your view may help discover high value opportunities with reduced risks.

# Phase IV & V: DM & BM

---

## Value Contributors

- Performance Contributors
  - How well we are doing?
- Predictive Contributors
  - What to do now?

# Phase IV & V: DM & BM

---

## Delivery Focus

1

*Vertical value delivery*, where value is delivered and data solutions are tailored to specific industries (e.g., prescription data for the pharmaceutical companies).

2

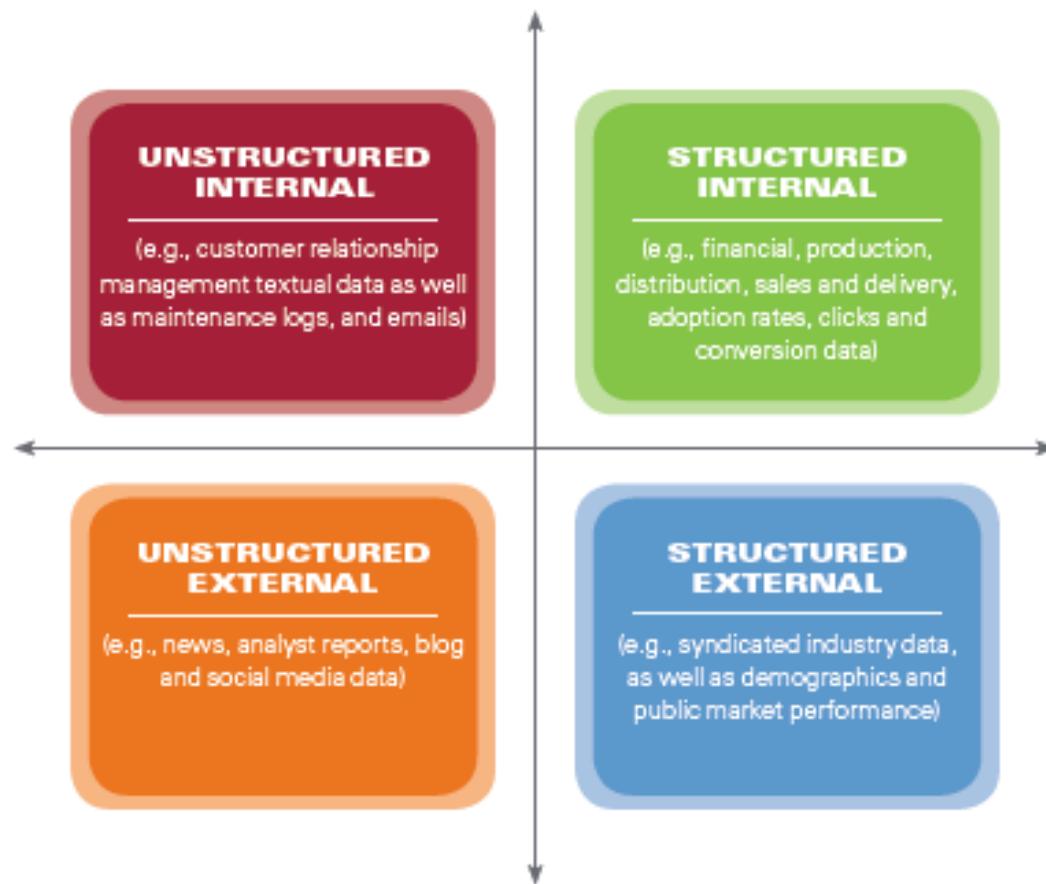
*Horizontal value delivery*, where the same set of data products may be valuable in similar form and format to various industries with similar needs (e.g., economic indices and indicators are used to project retail sales also used for real estate pricing).

3

*Cross-market value delivery*, where data that is collected for one purpose in one industry is valuable for another purpose in another, often adjacent, industry (e.g., good driver signals captured from vehicles and used for insurance pricing is also valuable for pricing car lease rates).

# Phase IV & V: DM & BM

## Data



# Phase IV & V: DM & BM

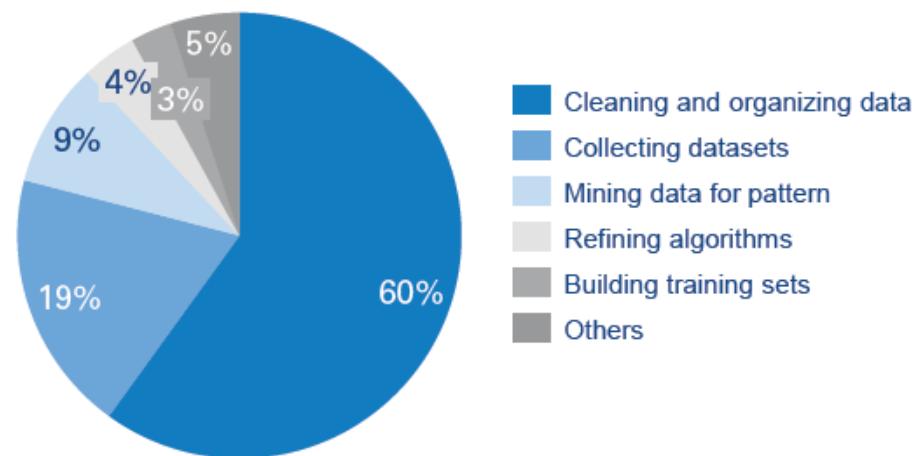
---

## Roadblocks:

- Data Overload
- Data Access
- Data Cleansing
- Data Scalability

Time spent by data scientists

---



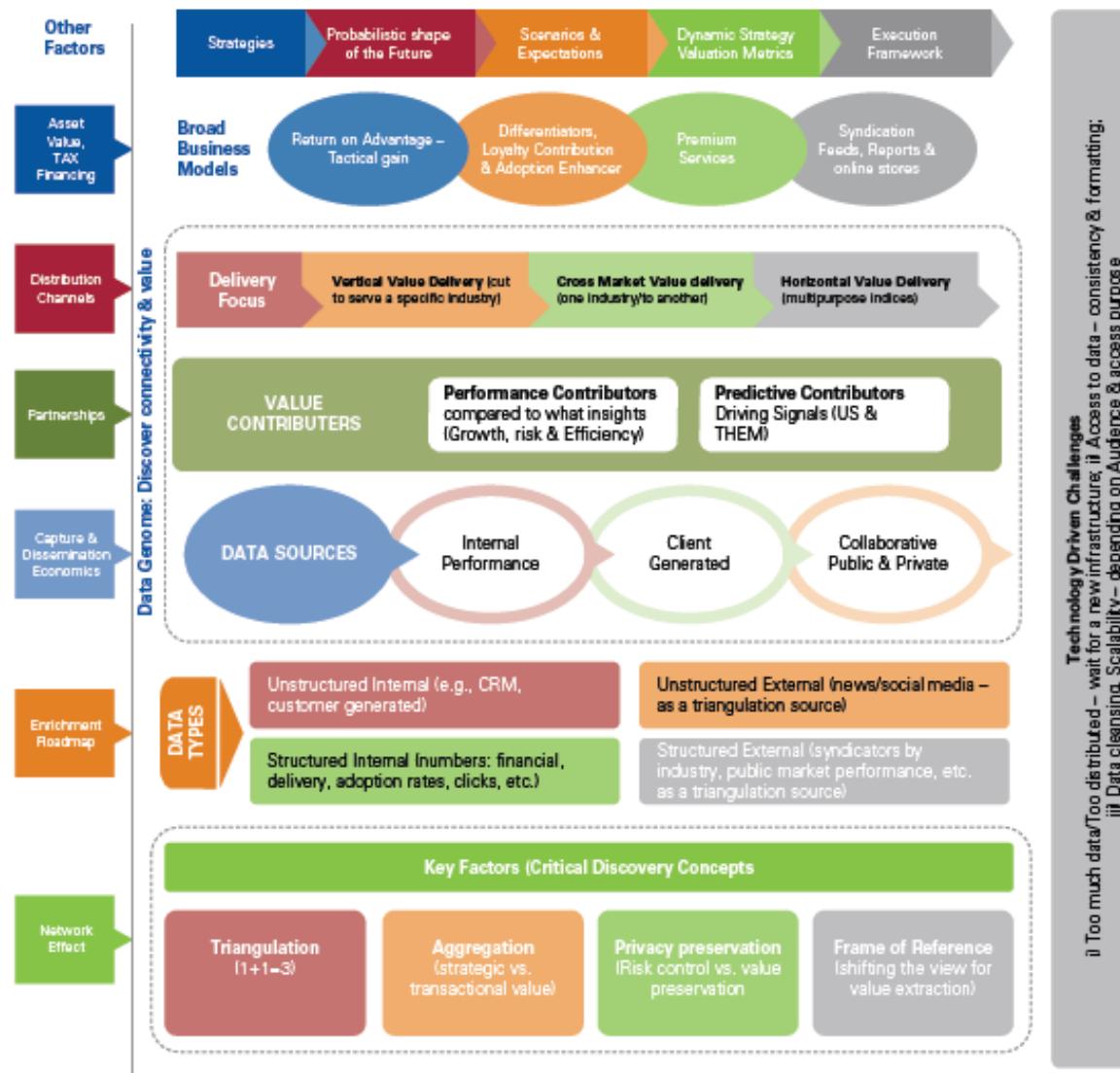
# Phase IV & V: DM & BM

---

## Non IT Roadblocks:

- Enrichment road map
- Capture and dissemination economics
- Network effect & natural barriers
- Asset value determination/tax and financing implications
- Partnerships & organization structures

# Phase IV & V: DM & BM



# Phase IV: Data Monetization

Example: Fitness tracker



# Phase IV: Data Monetization

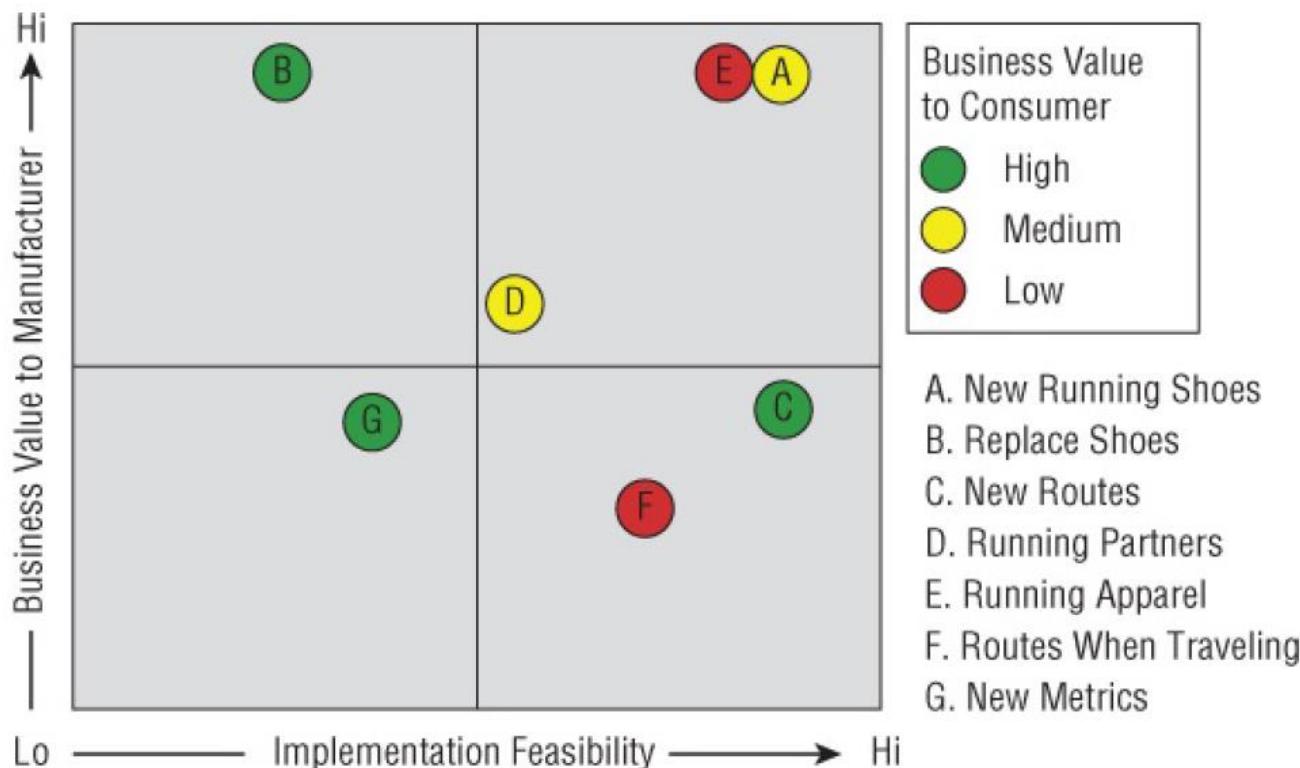
---

**Example:** Fitness tracker

<b>Recommendation</b>	<b>Consumer Value</b>	<b>Manufacturer Value</b>	<b>Feasibility</b>
A. Optimal new running shoes <sup>1</sup>	Medium	<b>High</b>	<b>High</b>
B. When to replace running shoes	<b>High</b>	<b>High</b>	Low
C. New local routes	<b>High</b>	Low	Medium
D. Running partners	Medium	Low	Low
E. Running apparel	Medium	<b>High</b>	<b>High</b>
F. Routes when traveling	Low	Low	Medium
G. New running metrics	<b>High</b>	Low	Medium

# Phase IV: Data Monetization

Example: Fitness tracker



# Phase V: Business Metamorphosis

---

**Example:** Boeing selling Air-miles instead of Airplanes



# Case Study

## Data Monetization & Business Metamorphosis:

- Telecommunications



# Case Study

---

## Data Monetization & Business Metamorphosis:

- Identifying value pools
- Making technology decisions
- Choosing the optimal operating model
- Bringing monetization strategy to action

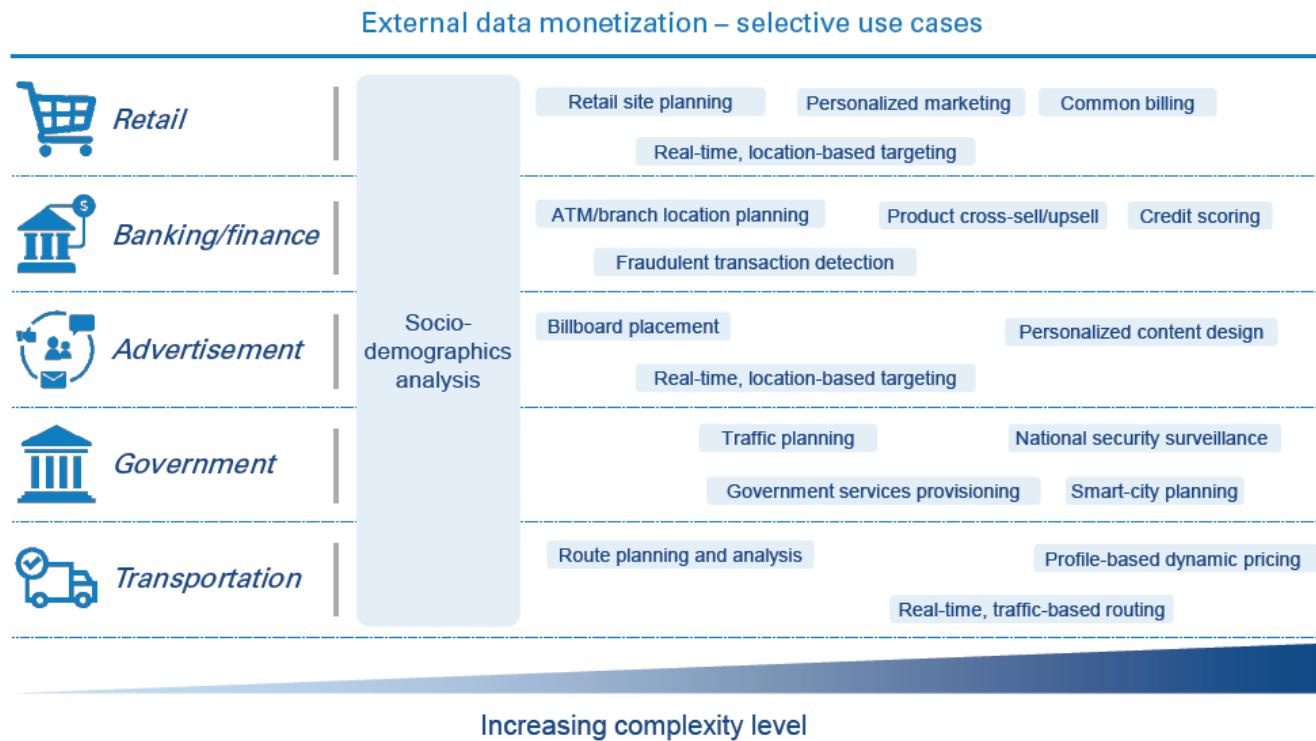
# Case Study

## Identifying value pools:

Focus big data/advanced analytics building block					
Marketing & sales	Customer service	Product	Enabling	Network	Steering & support
Personalized marketing	Customer activation and provisioning	Automated tariff optimization	Automated IT security threat detection	Predictive maintenance of network services	HR process automation
Real-time sales process monitoring	Real-time field force scheduling	Real-time product creation and management	Real-time IT fault detection and prediction	Real-time network fault detection and prediction	Field force analytics
Real-time media performance tracking	Boosted failure analysis for complaints	Product renewal enhancement	IT development quality assurance	End-to-end service quality monitoring	Smart procurement
Customer Onboarding and ARPU growth	Customer journey and satisfaction	Credit scoring as service	Application performance monitoring	Network roll-out optimization	Inventory optimization
Heavy roaming abuse detection	Automated natural conversation	Product convergence	Service activation optimization	Network monetization	Receivables optimization
Outdoor advertising optimization	Contact center productivity	Recommendation engine for external clients	IPTC service assurance	Network performance optimizations	Revenue assurance and fraud management

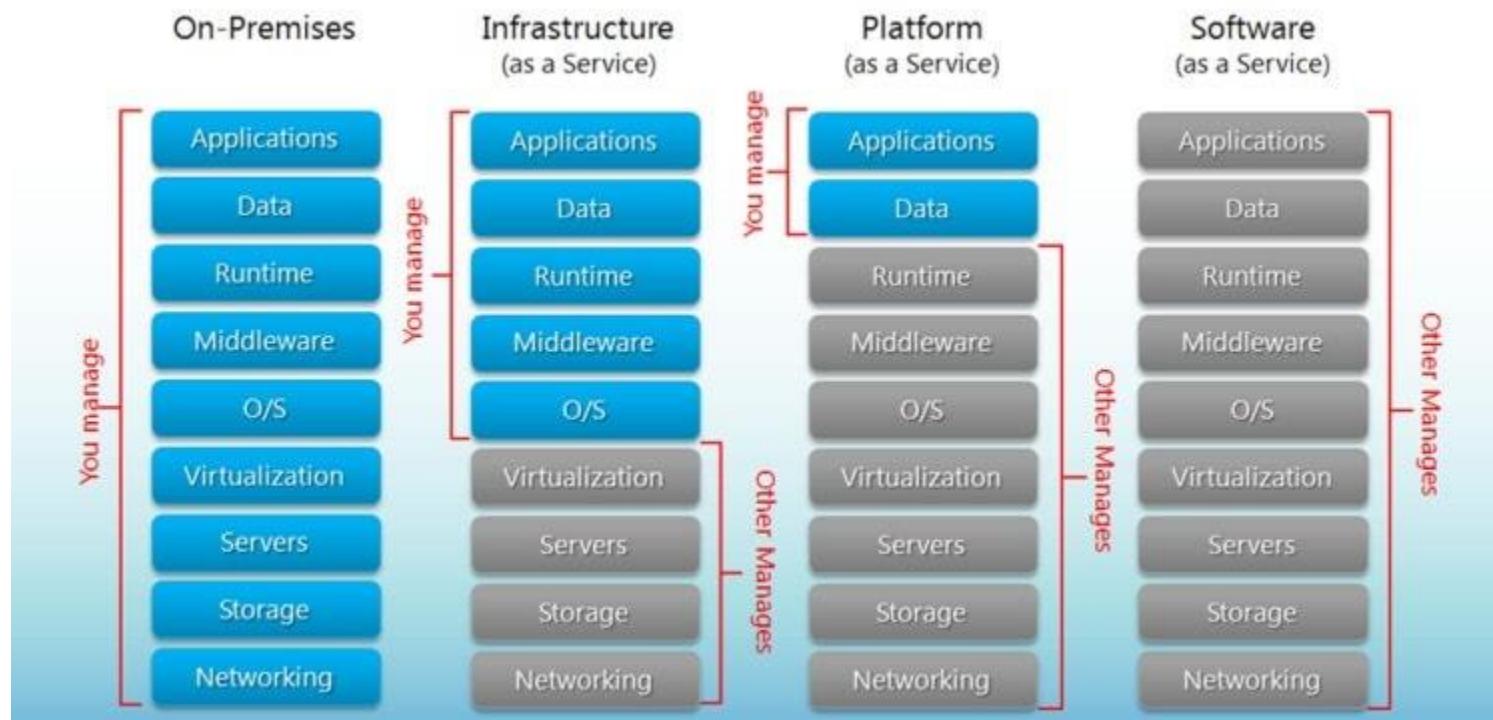
# Case Study

## Identifying value pools:



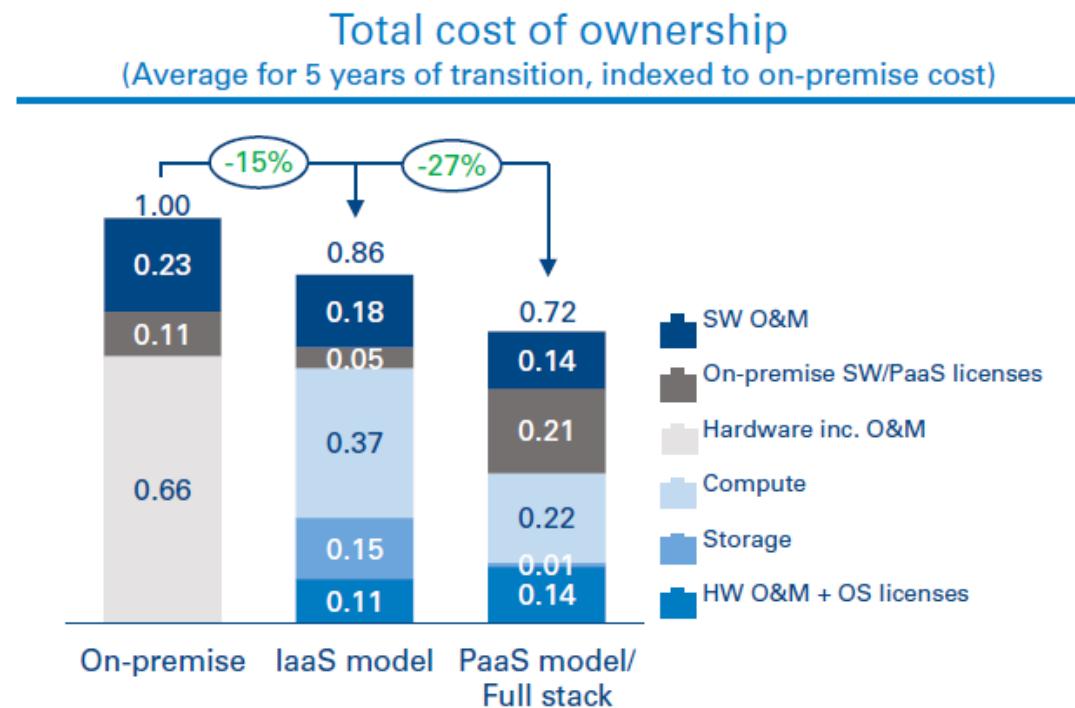
# Case Study

## Making technology decisions:



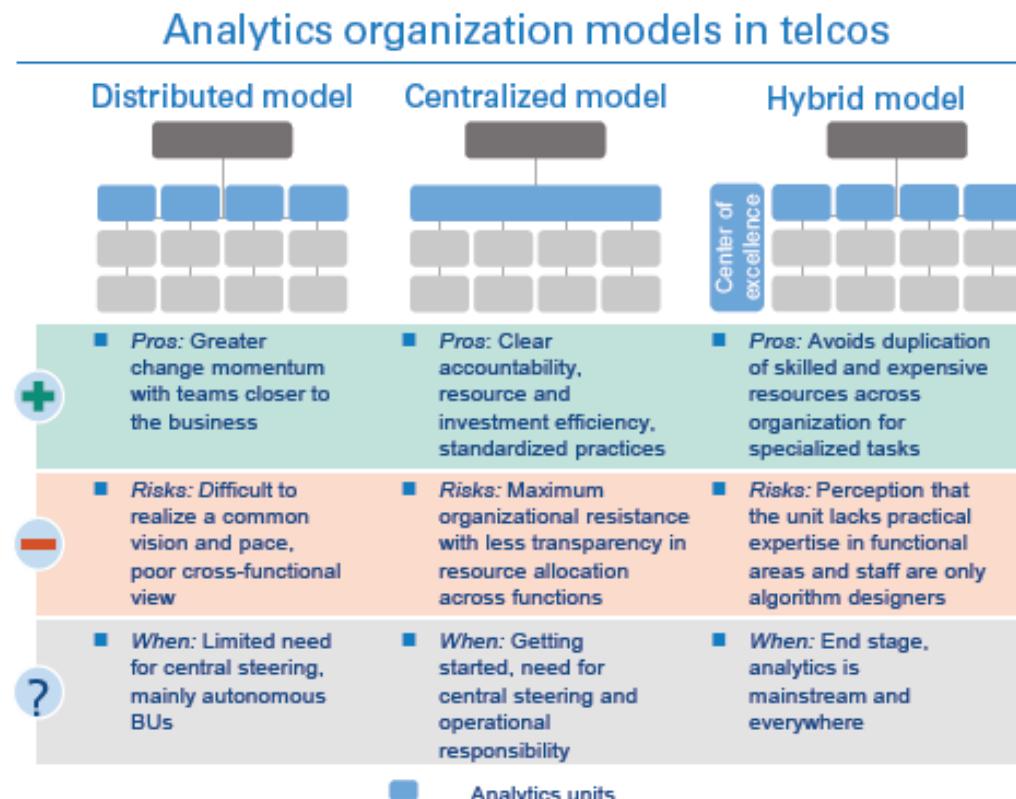
# Case Study

## Making technology decisions:



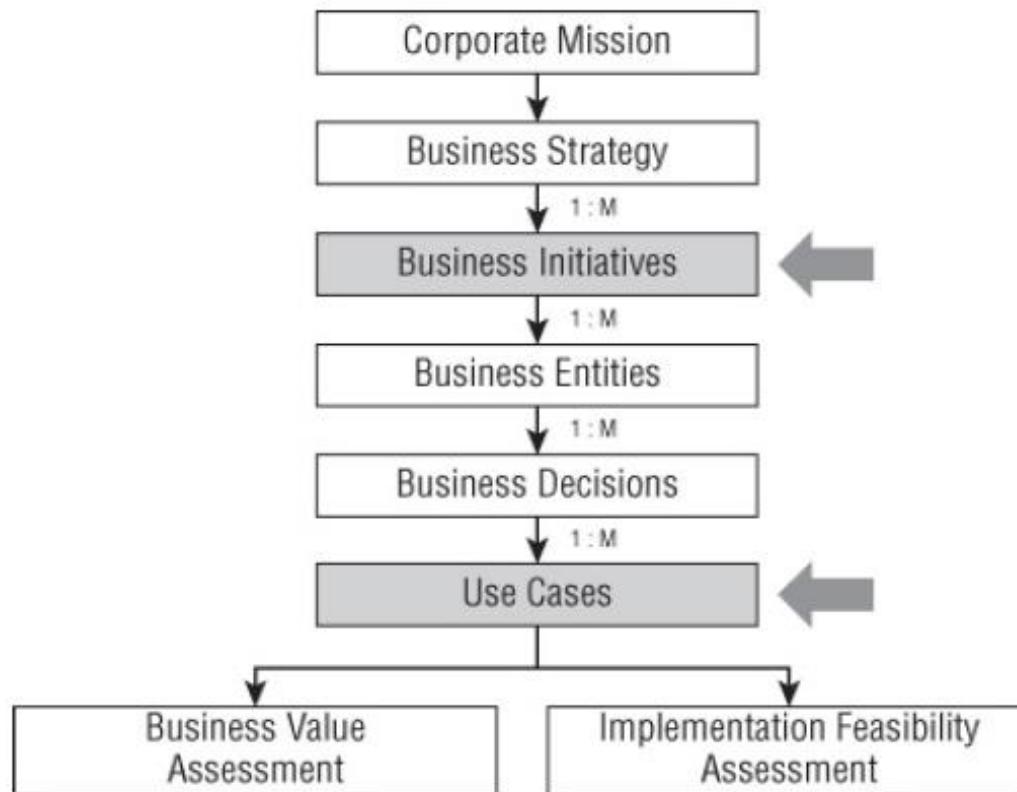
# Case Study

## Choosing the optimal operating model:



<https://hbr.org/2017/06/how-to-integrate-data-and-analytics-into-every-part-of-your-organization>

# Big Data Strategy



# Competitive Advantage

---

**Competition:** Today's in most sectors competition is on a global scale

- Direct competition
- Substitute competition
- Budget competition

**Competitive advantage:** Defined as the strategic advantage one business entity has over its rival entities within its competitive industry

**What are the competitive advantages of the following companies?**

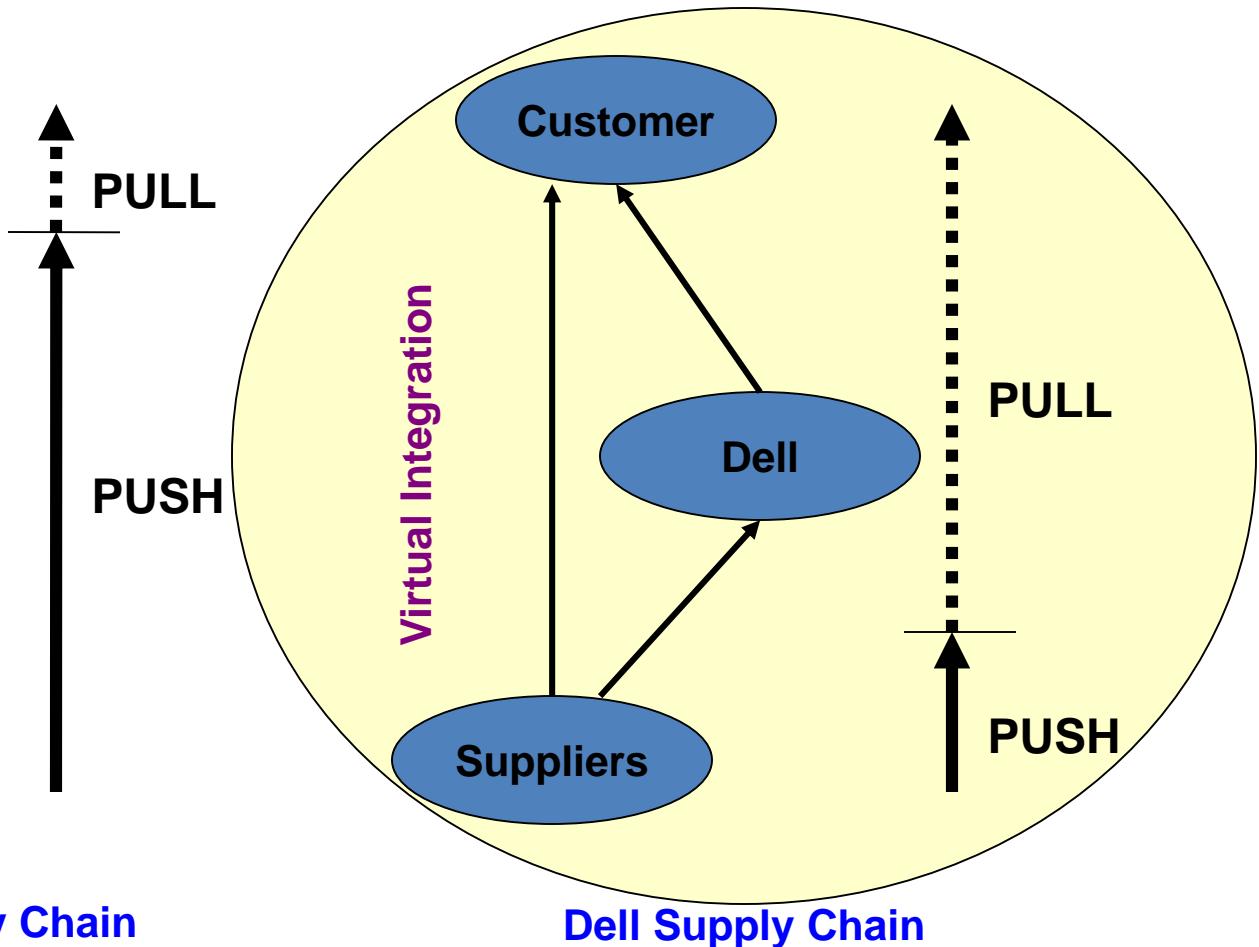
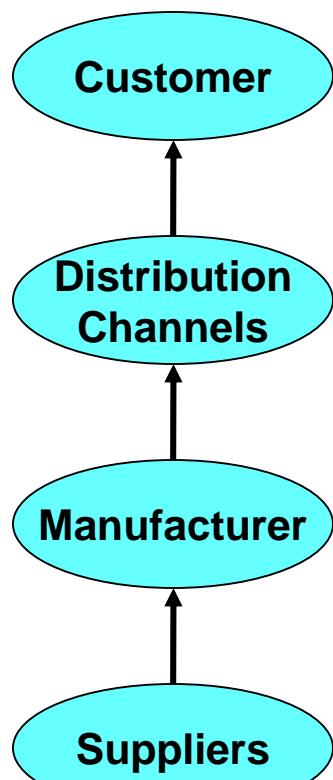
- Apple
- Pegasus Airlines
- Vakko
- Is Bankasi
- Turkcell

# Competitive Strategy

---



# Competitive Strategy



# Innovator's Dilemma

---

Innovator's Dilemma:

[https://www.youtube.com/watch?v=NrC\\_tR8hxjQ](https://www.youtube.com/watch?v=NrC_tR8hxjQ)

# Innovator's Dilemma

---

**Innovator's Dilemma:** Christensen's book suggests that successful companies can put too much emphasis on customers' current needs, and fail to adopt new technology or business models that will meet their customers' unstated or future needs. Christensen calls the anticipation of future needs "disruptive innovation".

**Henry Ford:** "If I had asked people what they wanted, they would have said faster horses"



# Sustaining vs Disruptive Innovation

---

**Sustaining Innovation:** Company improves the product's (service) performance based on the feedback from the main customers. Usually, it is about improving the product or adding features.

*Philosophy: Customer is always right!*

**Disruptive Innovation:** Lower performance of the product (service) in many key features valued by the market. From the mainstream market's perspective, it may look like a failure.

*Target: Niche Market, unsatisfied by the incumbent product*

*Sustaining Innovation ~ Market's Current Needs*

*vs*

*Disruptive Innovation ~ Market's Future Needs*

# Innovator's Dilemma

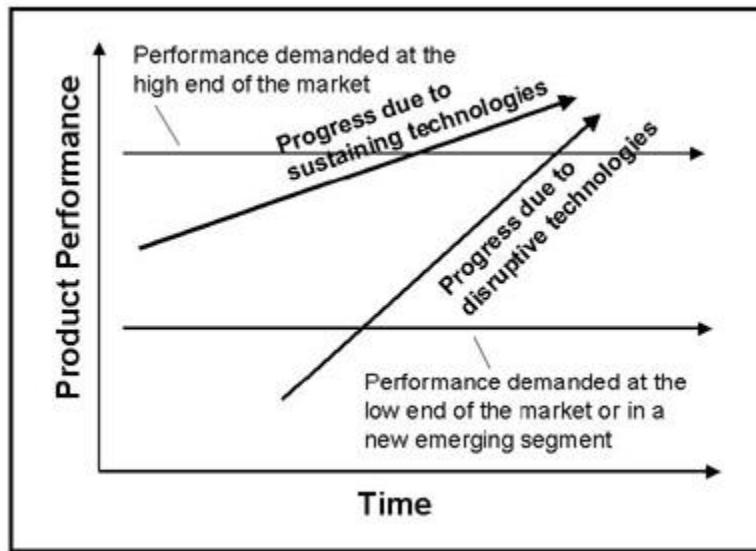
---

**Innovator's Dilemma:** Sustaining innovation is the safe path, but may doom the company. Disruptive innovation may have potential but dedicating resources to an unproven market may be a risk.



*Why may be more difficult for a market leader to attempt disruptive innovation?*

# Sustaining vs Disruptive Innovation



# Examples

---



# Examples

---



# Examples

---



**WIKIPEDIA**  
The Free Encyclopedia



**NETFLIX**

Google Maps

**amazon**



# Innovator's Dilemma

---

*Why may be more difficult for a market leader to attempt disruptive innovation?*

## Market Leader:

- Use Spin-off companies
- Market Size Match, project sizes small enough for the niche market
- Fail early and inexpensively
- Use resources only, not the processes and the values
- Understand that it is a marketing challenge, not a technical challenge

# Big Data and Disruptive Innovation

---

## Disruptive Innovation:

- Digitization alone isn't the key to sustainable competitive advantage
- Application of digital intelligence, fuelled by the monetization of data and insight
- Only a few organizations are immune to digital disruption.

# Big Data and Disruptive Innovation

---

## Disruptive Innovation:

FIGURE 1

### MOST FACE DIGITAL MARKET DISRUPTIONS

Percentage rating how susceptible their organization is to market disruption in the next three years from a new competitor based on their use of insight, data, or analytics.

(SCALE OF 1-10)



**29%**  
EXTREMELY SUSCEPTIBLE (8-10)

**43%**  
MODERATELY SUSCEPTIBLE (5-7)

**27%**  
NOT AT ALL SUSCEPTIBLE (1-4)

**1%**  
DON'T KNOW

SOURCE HARVARD BUSINESS REVIEW ANALYTIC SERVICES SURVEY, JUNE 2016

# Big Data and Disruptive Innovation

---

## Disruptive Innovation:

### EFFECTIVENESS IN NEW DIGITAL BUSINESS MODEL INNOVATION

Percentage rating how effective their organization is at innovating new digital business models—either to respond to new threats or to capitalize on new opportunities. (SCALE OF 1-10)



**23%**  
EXTREMELY EFFECTIVE (8-10)

**43%**  
MODERATELY EFFECTIVE (5-7)

**33%**  
COMPLETELY INEFFECTIVE (1-4)

**1%**  
DON'T KNOW

SOURCE HARVARD BUSINESS REVIEW ANALYTIC SERVICES SURVEY, JUNE 2016

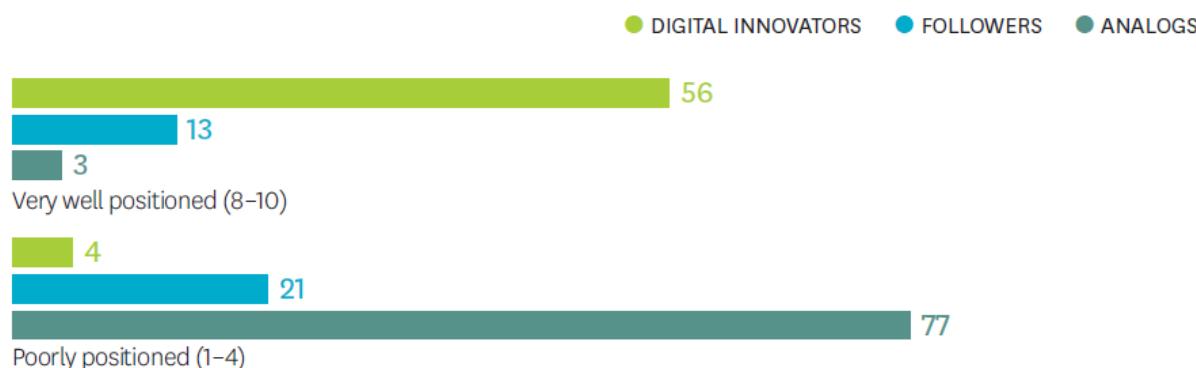
# Big Data and Disruptive Innovation

## Examples:

- Finance sector
- Construction
- Hospitality
- Pharmaceutical

### DIGITAL INTELLIGENCE IS A COMPETITIVE ADVANTAGE

Percentage indicating how well positioned their organization is to use digital intelligence capabilities as a competitive advantage.



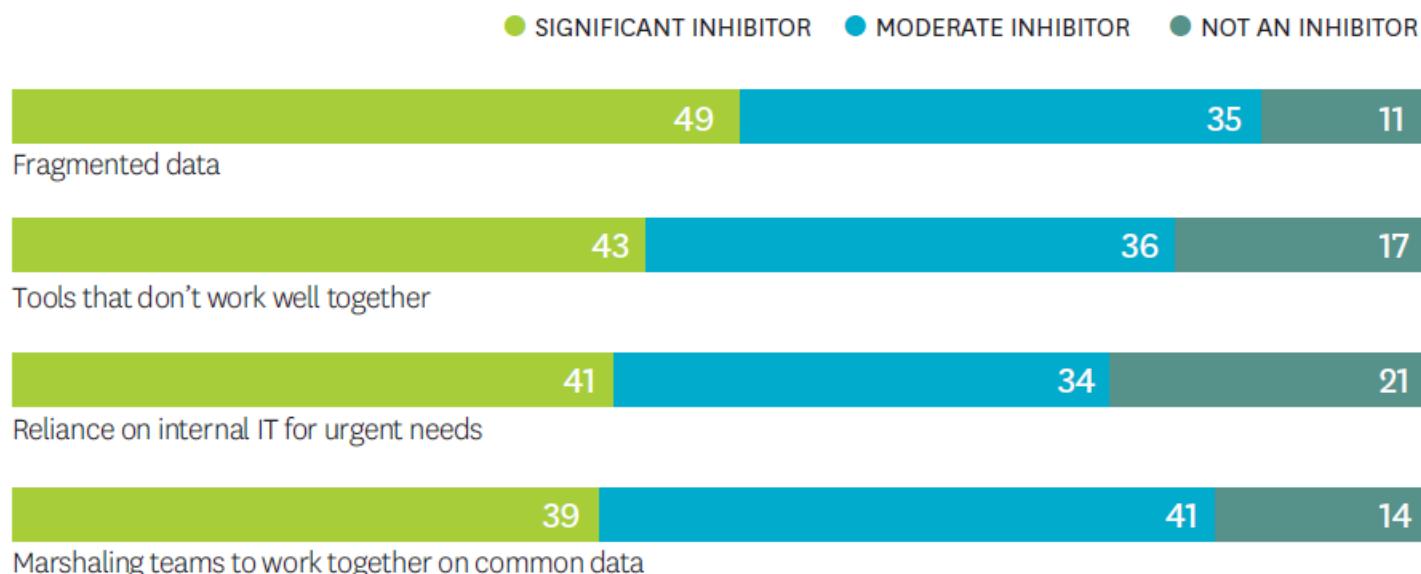
SOURCE HARVARD BUSINESS REVIEW ANALYTIC SERVICES SURVEY, JUNE 2016

# Big Data and Disruptive Innovation

## Barriers:

### FRAGMENTED DATA AND RESOURCES SLOW THINGS DOWN

Percentage indicating the extent to which the following inhibit your organization's ability to achieve business objectives on a timely basis.



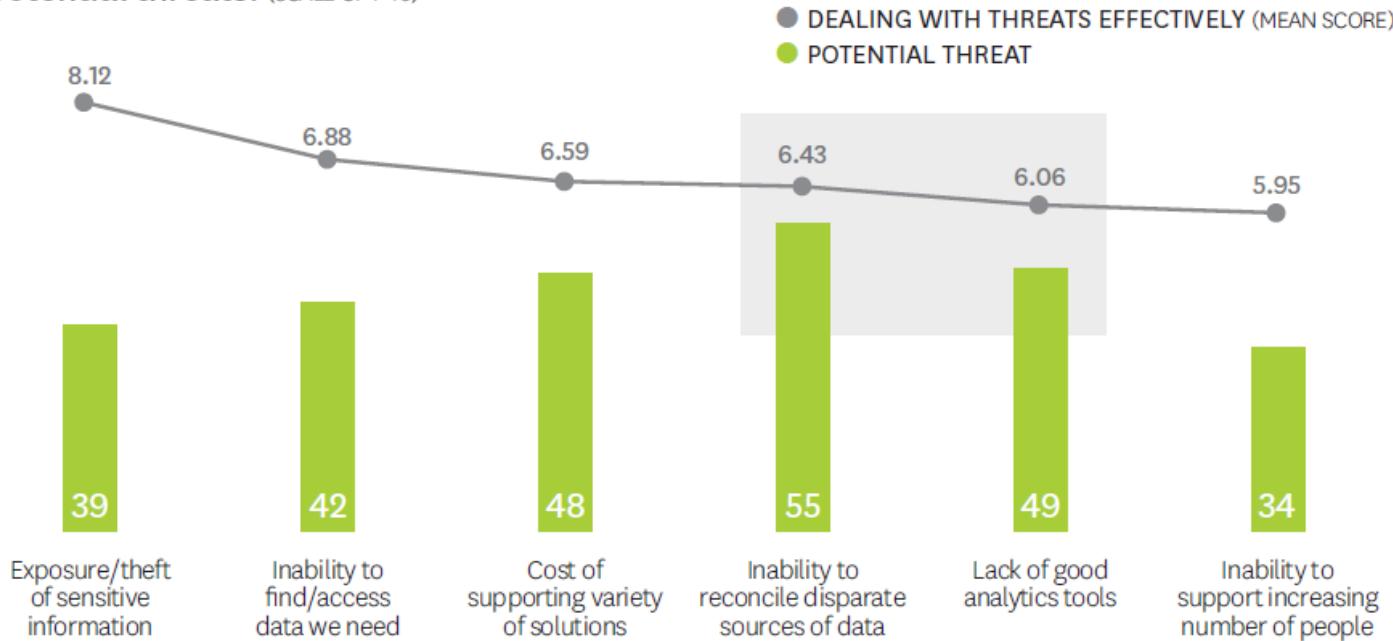
SOURCE HARVARD BUSINESS REVIEW ANALYTIC SERVICES SURVEY, JUNE 2016

# Big Data and Disruptive Innovation

## Barriers:

### DIGITAL INNOVATORS MANAGING KEY THREATS ADEQUATELY

Percentage indicating which of the following are potential threats to their organization's ability to make progress and how effectively their organization is addressing the following potential threats. (SCALE OF 1-10)



SOURCE HARVARD BUSINESS REVIEW ANALYTIC SERVICES SURVEY, JUNE 2016

# Big Data and Disruptive Innovation

## Barriers:

### FOLLOWERS/ANALOGS EXPOSED TO KEY THREATS

Percentage indicating which of the following are potential threats to their organization's ability to make progress and how effectively their organization is addressing the following potential threats. (SCALE OF 1-10)



SOURCE HARVARD BUSINESS REVIEW ANALYTIC SERVICES SURVEY, JUNE 2016

# Big Data and Disruptive Innovation

---

Examples:



# Big Data and Disruptive Innovation

---

Examples: <https://medium.com/make-innovation-work/disney-bad-for-disney-bad-for-consumers-but-their-only-choice-left-72c36b14b567>



# Big Data Strategy

## Phase I:

- Identify the Organization's Key Business Initiatives



## Chipotle Business Strategy

- Continue to build a people culture that attracts and empowers top performers
- Continue to grow revenues (up 20.3% in 2012) by opening new stores (opened 183 in 2012)...
  - ...and increase comparable restaurant sales growth (7.1% in 2012)
- Marketing focused on building the Chipotle brand and engaging with our customers in ways that create stronger, deeper bonds

# Big Data Strategy

---

## Phase II:

- Identify Key Business Entities and Key Decisions

**What are the key business entities for Chipotle?**

**What are the key business decisions for Chipotle?**

# Big Data Strategy

## Phase III:

- Identify Financial Drivers
  - “how do we make more money” opportunities

Chipotle Use Cases	Potential Analytic Models
Increase Store Traffic	Store Marketing Effectiveness Store Layout Flow Analysis Store Remodeling Lift Analysis Store Customer Targeting
Increase Shopping Bag Revenue and Margin	In-store Merchandising Effectiveness Pricing Optimization Up-sell/Cross-sell Effectiveness Market Basket Analysis
Increase Number of Corporate Events	Campaign Effectiveness Pipeline and Sales Effectiveness Pricing Optimization Customer Lifetime Value Score Likelihood to Recommend Score
Improve Promotional Effectiveness	Promotional Effectiveness Pricing Optimization Market Basket Analysis Up-sell/Cross-sell Effectiveness
Improve New Product Introductions	Pricing Optimization New Product Introductions Effectiveness Up-sell/Cross-sell Effectiveness

# Big Data Strategy

## Phase IV:

- Identify and Prioritize Data Sources

Key  
Worst... Best

Data Source	Increase Store Traffic	Increase Shopping Bag Revenue	Increase # Corporate Events	Increase Promotional Effectiveness	Improve NPI Effectiveness
Point of Sales Transactions	●	●	●	●	●
Market Baskets	●	●	○	●	●
Store Demographics (Zip Code)	●	●	●	●	●
Local Competitive Stores	○	○	○	○	○
Store Manager Demographics	○	○	●	○	○
Consumer Comments	●	●	●	●	○
Social Media	○	○	○	●	●
Weather	●	○	○	○	○
Local Events	●	○	○	○	○
Traffic	●	○	○	○	○
Zillow	○	●	○	○	○

# Big Data Strategy

## Phase IV:

- Identify and Prioritize Data Sources

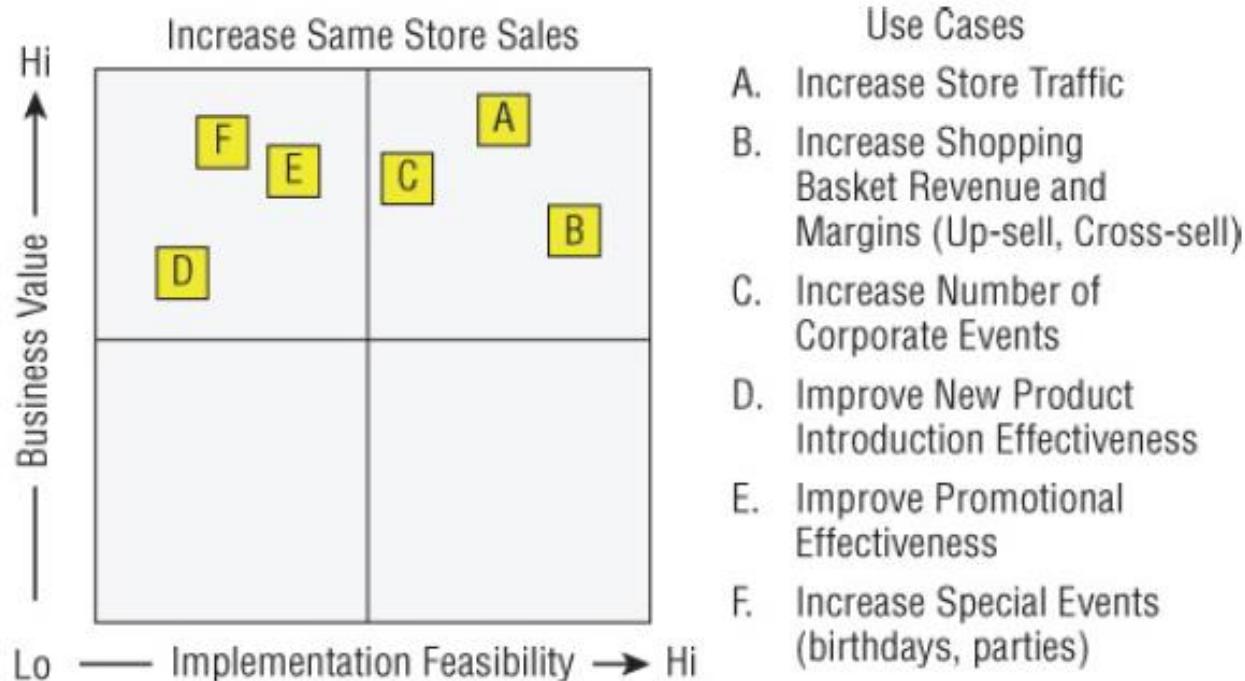


Data Source	Ease of Acquiring	Cleanliness	Accuracy	Granularity	Cost
Point of Sales Transactions	●	●	●	●	●
Market Baskets	●	●	●	●	●
Store Demographics (Zip Code)	●	●	●	●	●
Competitive Stores Sales	○	○	○	○	○
Store Manager Demographics	●	●	●	●	●
Consumer Comments	○	○	○	○	○
Social Media	○	○	○	○	○
Weather	●	○	○	○	○
Local Events	○	○	○	○	○
Traffic	○	○	○	●	○
Zillow	○	○	○	○	○

# Big Data Strategy

## Phase V:

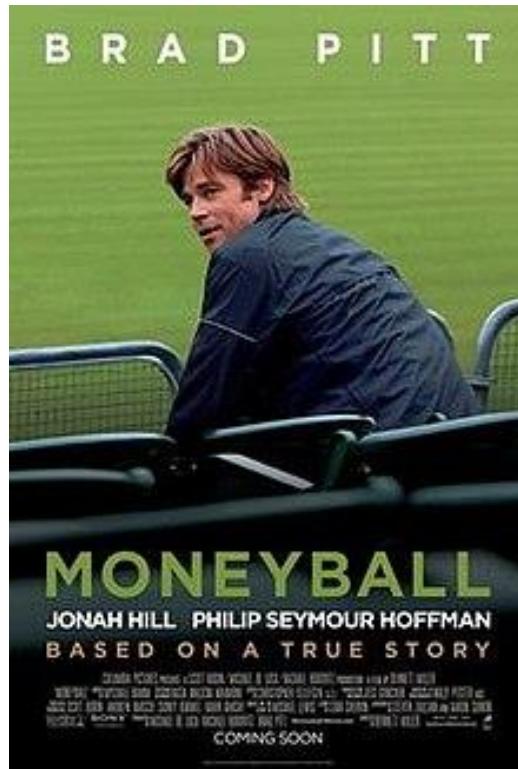
- Prioritization Matrix



# Big Data Strategy

---

Homework:



# DS 555 Data Science and Business Strategy

---

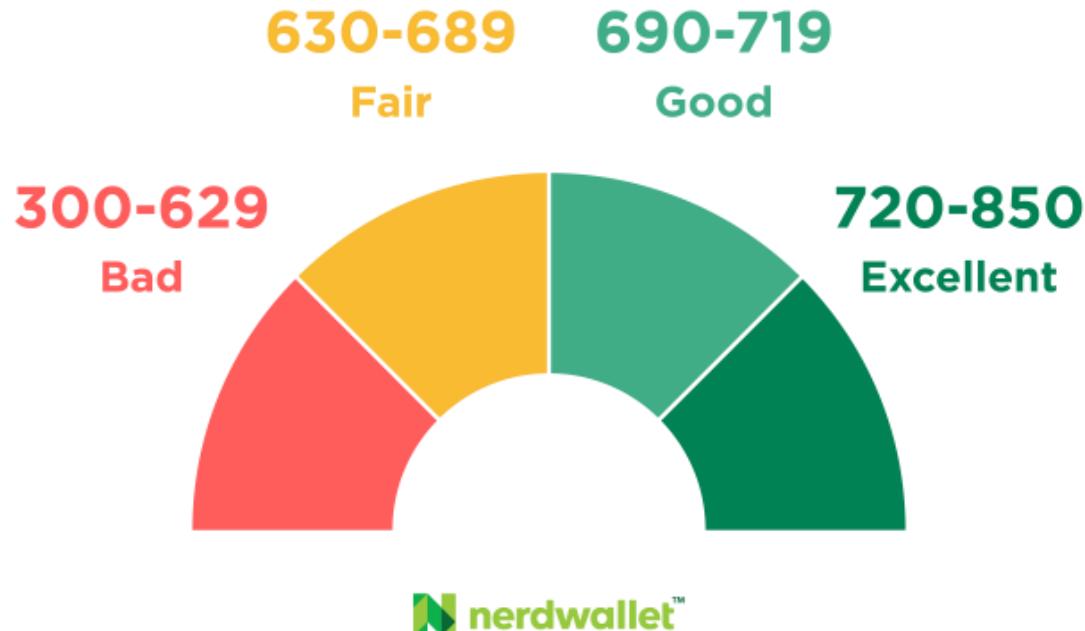
## *SCORE DEVELOPMENT*

– O.Örsan Özener

# Score Development

## Scores:

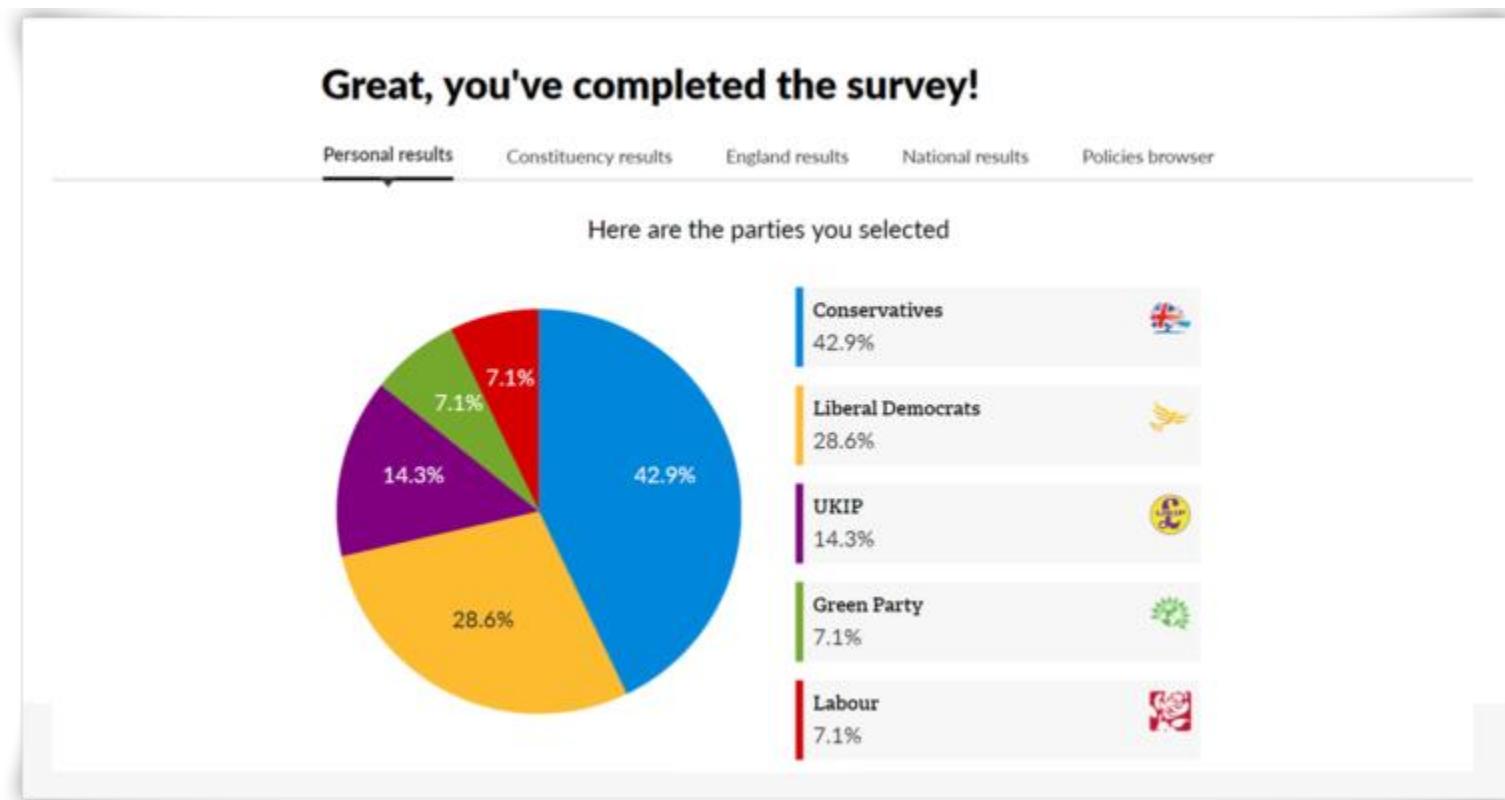
- An aggregate measurement
- Helps in decision making
- groupings of metrics and variables combined to an **actionable score**



Why do we focus on “scores” in making key decisions? 2

# Score Development

## Voting:



# Score Development

## Scoring:



# Score Development

## Scoring:

Type of event?	How many attendees?	Is alcohol being served?
<input checked="" type="radio"/> Neighborhood party <input type="radio"/> Sports event <input type="radio"/> House party <input type="radio"/> Visit by a head of state	<input type="radio"/> 0 - 10 <input checked="" type="radio"/> 11 - 50 <input type="radio"/> 51 - 250 <input type="radio"/> 251 <	<input type="radio"/> Yes <input checked="" type="radio"/> Possibly <input type="radio"/> No <input type="radio"/> Unknown
<a href="#">Next</a>	<a href="#">Next</a>	<a href="#">Next</a>

Risk assessment



The event is deemed **LOW RISK**

[Continue to advice for emergency services](#)

# Score Development

## Scoring:

Questions	
Type	Neighborhood 0
	Sports event 2
	House party 4
	Visit head of state 5
Visitors	0 – 10 0
	11 – 50 1
	51 – 250 2
	251 > 4
Alcohol	Yes 3
	Possibly 2
	No 0
	Unknown 2
Season	Spring 0
	Summer 1
	Autumn 0
	Winter 0,5
Risk assessment	
< 3	Low risk
4 – 7	Medium risk
8 >	High risk

# Score Development

---

**Balanced Scorecard:**

<https://www.youtube.com/watch?v=biyGxEix5Zs>

# Score Development

---

## Advantages:

- A single score instead of complicated analysis, intervals, charts
- Groups and assesses the relative importance of the metrics
  - Especially from different perspectives
- Makes the evaluation process standard (e.g. free of human based, current conditions etc.)
- Makes the knowledge sharing easier
- Specific and customizable
- Efficient

# Score Development

---

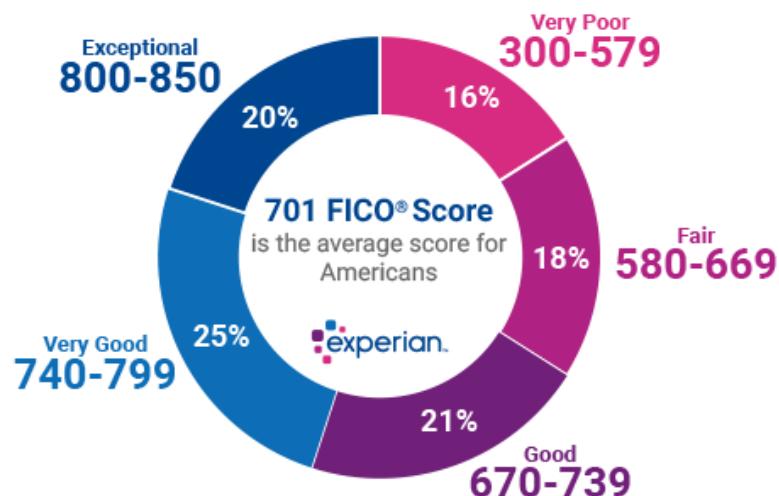
## Disadvantages:

- Some information is lost in translation, calculation etc.
  - Once the “so-called” score is calculated no one rechecks the data
  - Examples???
- Too much aggregation
  - Outlier might spoil the process
  - Not great in evaluating non standard
- OTHERS ???

# Score Development

## Example - FICO:

- FICO score is used to predict the likelihood of a borrower to repay a loan.
- combines multiple metrics (financial, credit, and payment history) to create a singular score



**FICO® Score 8**  
300 to 850

**FICO® Mortgage Score**  
300 to 850

**FICO® Auto Score**  
250 to 900

**FICO® Bankcard Score**  
250 to 900

# Score Development

## Example - FICO:

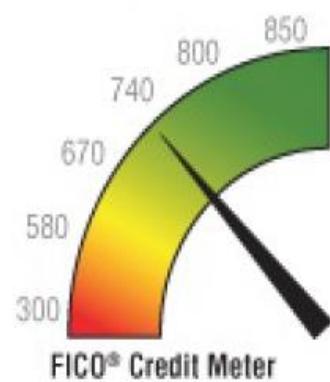
### DESCRIPTIVE ANALYTICS

- What are your credit card balances?
- What is your credit card payment history?
- How long have you had the credit cards?
- What is your credit utilization?
- How many car loans do you have?
- What is your home mortgage payment?
- What are your student loan payments?
- What is your checking balance?
- What is your savings balance?

### PREDICTIVE ANALYTICS

FICO score is used by lenders to predict your ability to repay a loan:

- Your credit worthiness in applying for credit or a loan
- The interest rate and loan terms that you receive for a home mortgage or car loan



# Scoring - LeBron

---

Example - LeBron:



# Scoring - LeBron

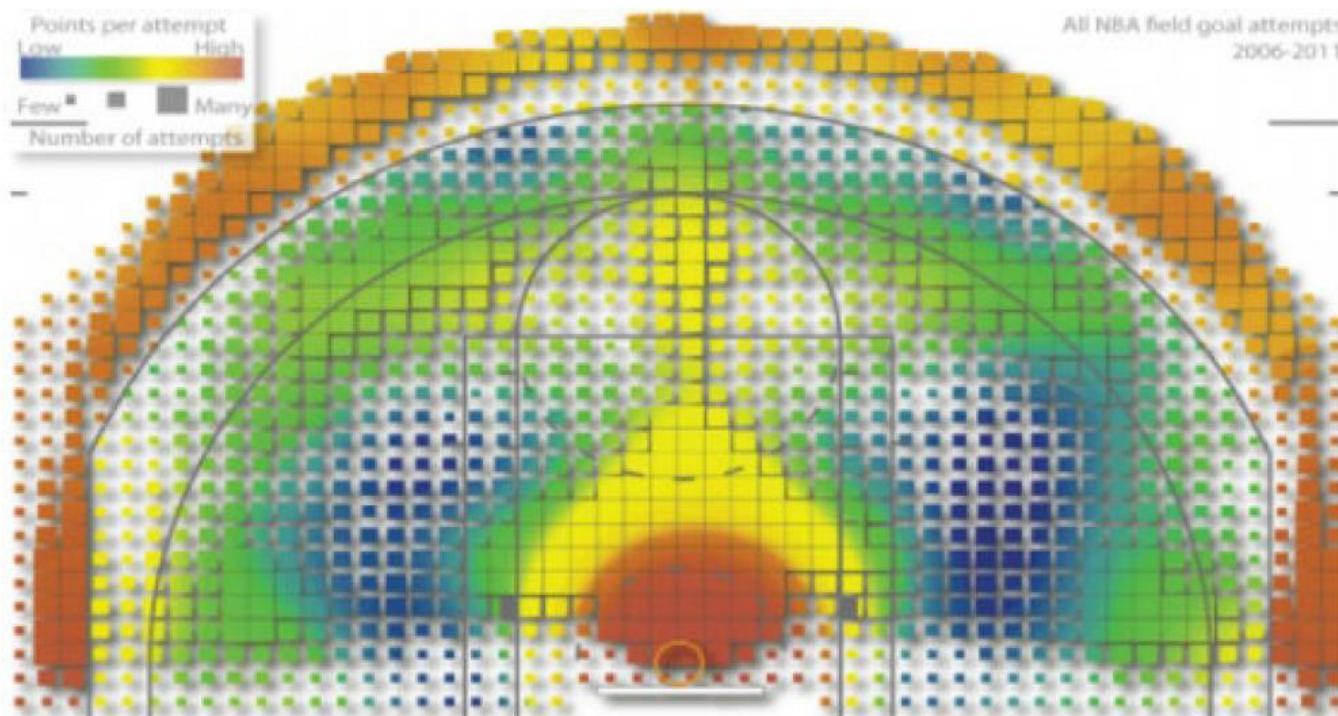
---

## “By” Analysis:

- exploits a business user's natural “question and answer” enquiry process to identify new data sources, dimensional characteristics, variables, and metrics
- Used in building the predictive and prescriptive analytic models to help predict business performance
- leverages a business stakeholder's natural curiosity to brainstorm and fuel the group innovative thinking process (remember the “ugly-girlfriend” from Moneyball”)

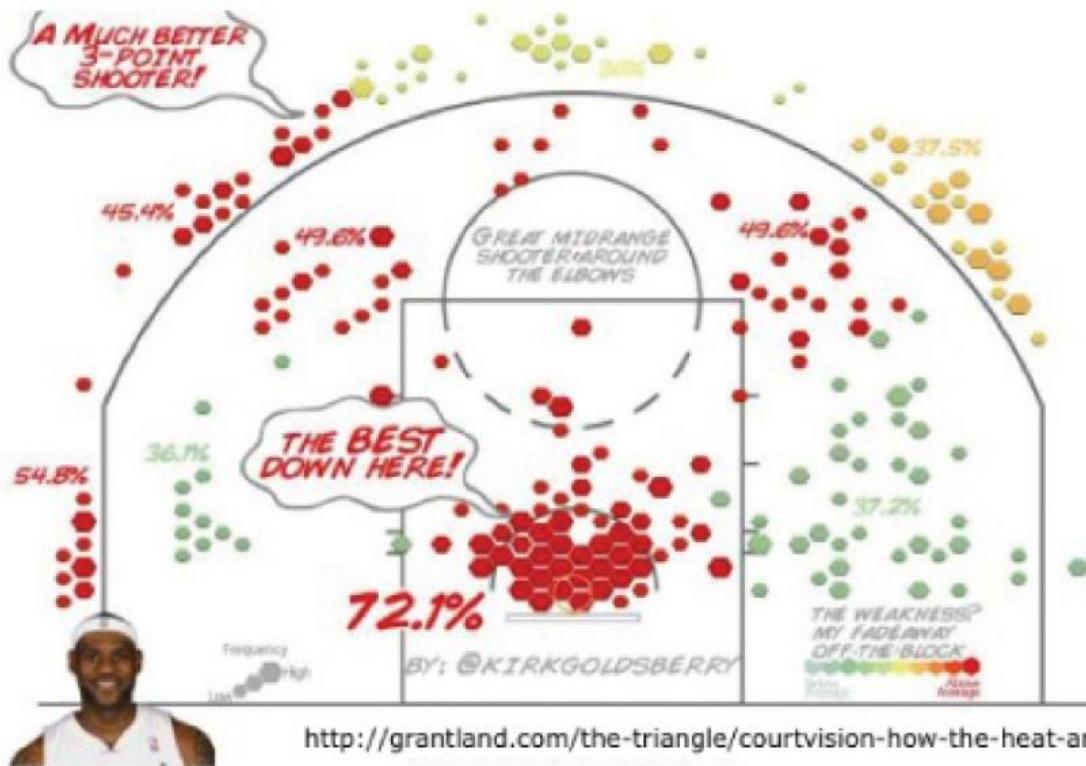
# Scoring - LeBron

LeBron “By” Analysis:



# Scoring - LeBron

LeBron “By” Analysis:



# Scoring - LeBron

## Potential Metrics:

At home versus on the road	Game location elevation	Total minutes played in game	Number of assists
Number of days of rest	Game time weather	Number of shots attempted	Playing a former team
Shot area	Game time temperature	Number of shots made	Record of opponent
Opposing team	Time (hours) since last game	Location of shots attempted	Performance in last game
Defender	Average time of ball possession	Location of shots made	Time of day
Game location	Time left in game	Number of fouls	????

# Scoring - LeBron

## Stats:

2014–2015	Overall Shooting Percentage	Overall Shooting Index	3-point Shooting Percentage	3-point Shooting Index
Regular season	48.8	100.0	35.4	100.0
Home	47.3	96.9	35.6	100.6
Road	50.2	102.9	35.3	99.7
0 days rest	49.8	102.0	38.0	107.3
1 day rest	46.3	94.9	32.3	91.2
2 days rest	51.3	105.1	37.3	105.4
3 days rest	52.7	108.0	42.9	121.2
4 days rest	57.1	117.0	60.0	169.5
6+ days rest	48.5	99.4	30.8	87.0

# Scoring - LeBron

## Decisions and Scores:

Business Initiative: <i>Mitigate LeBron James's offensive effectiveness</i>		
Persona: Golden State Coaching Staff		
Decisions	Potential Recommendations	Potential Scores / Metrics
Who is going to guard LeBron James?	<ul style="list-style-type: none"><li>• Which defender?</li><li>• Which defender at which times of the game?</li><li>• Which defender in which game situations?</li></ul>	<p>Fatigue Score</p> <ul style="list-style-type: none"><li>• Hours since last game</li><li>• How many games played in the season</li><li>• Average number of minutes played per game</li><li>• Minutes played in the current game</li><li>• Minutes handling the ball in the current game</li><li>• Number of shots taken in the current game</li><li>• Time remaining in the current game</li><li>• Away or home game</li></ul>
What is the best individual defensive approach?	<ul style="list-style-type: none"><li>• Do we deny LeBron the ball?</li><li>• Do we deny the 3-point jumper?</li><li>• Do we deny the 2-point jumper?</li><li>• Do we deny the drive into the lane?</li></ul>	<p>Motivation Score</p> <ul style="list-style-type: none"><li>• In-game performance</li><li>• Record of opponent</li><li>• Defender guarding him</li><li>• Volume of boos</li><li>• Playing against a former team</li><li>• Number of LeBron jerseys in the stands</li></ul>
What is the best team defensive approach?	<ul style="list-style-type: none"><li>• When do we guard LeBron straight up?</li><li>• When do we double-team LeBron?</li><li>• When do we hedge?</li><li>• When do we help?</li></ul>	

# Scoring - Housing

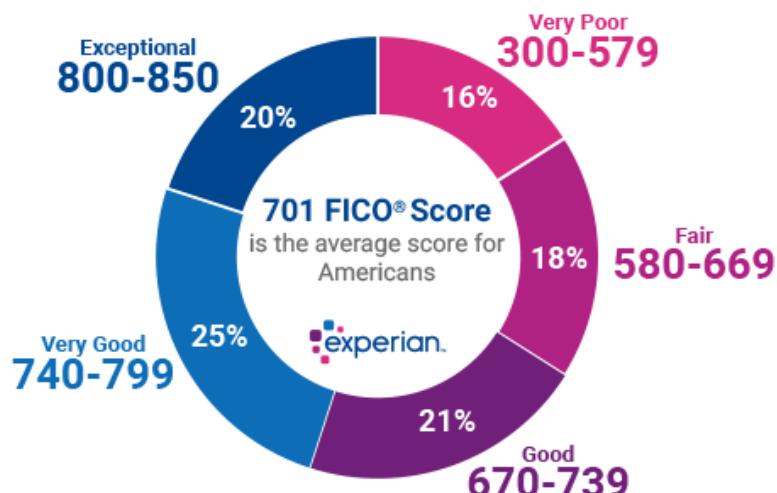
---

Example:



# Scoring - Credit

**Example:** Given the data, find a good scoring model.



**FICO® Score 8**  
300 to 850

**FICO® Mortgage Score**  
300 to 850

**FICO® Auto Score**  
250 to 900

**FICO® Bankcard Score**  
250 to 900

# DS 555 Data Science and Business Strategy

---

*WHEN DATA LIES*

– O.Örsan Özener

# Quote of the day

---

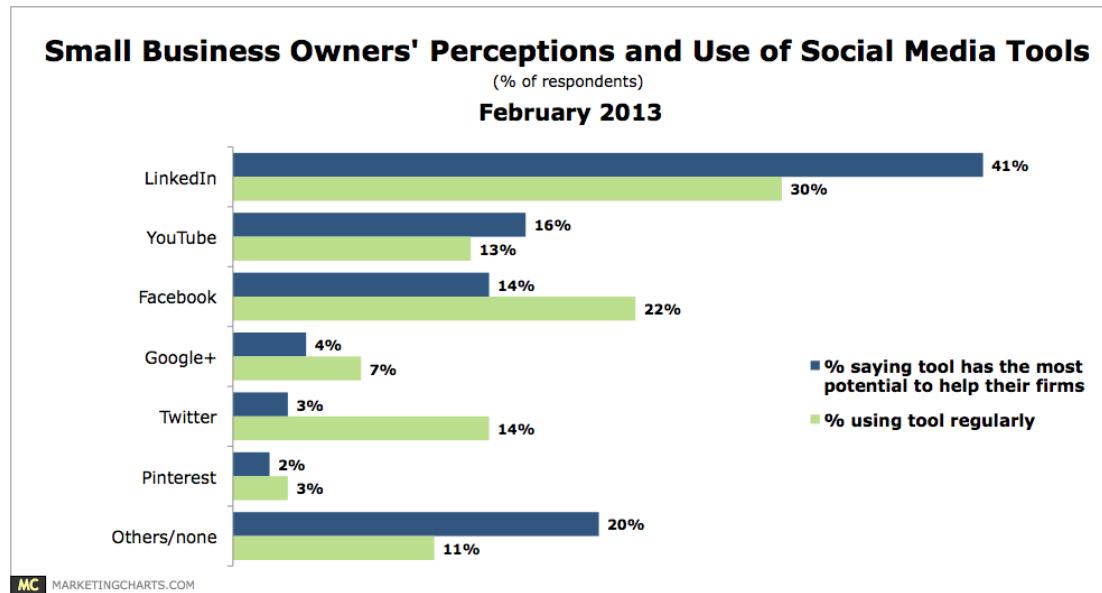
*“There are three kinds of lies: lies, damned lies, and statistics.”*

– Mark Twain ???

# Data Lies - Surveys

## Surveys:

- The Wall Street Journal recently reported that 70% of small business owners think social media is important.
- Citibank survey said only 36% of small businesses use social media and a mere 24% have found social media useful for finding leads or generating revenue.

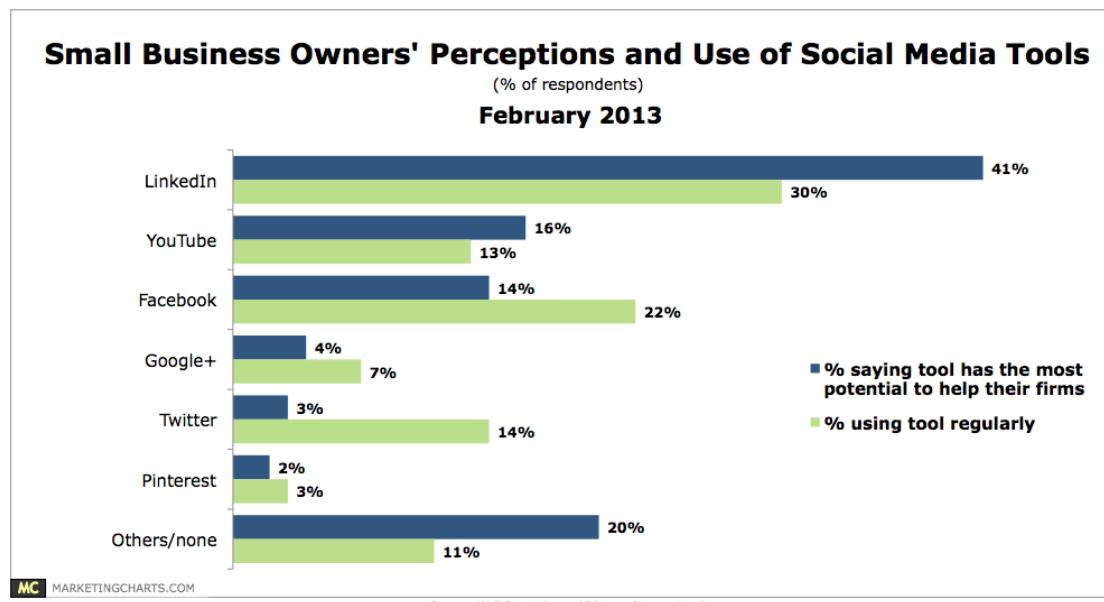


Why such a discrepancy in the results?

# Data Lies - Surveys

## Surveys:

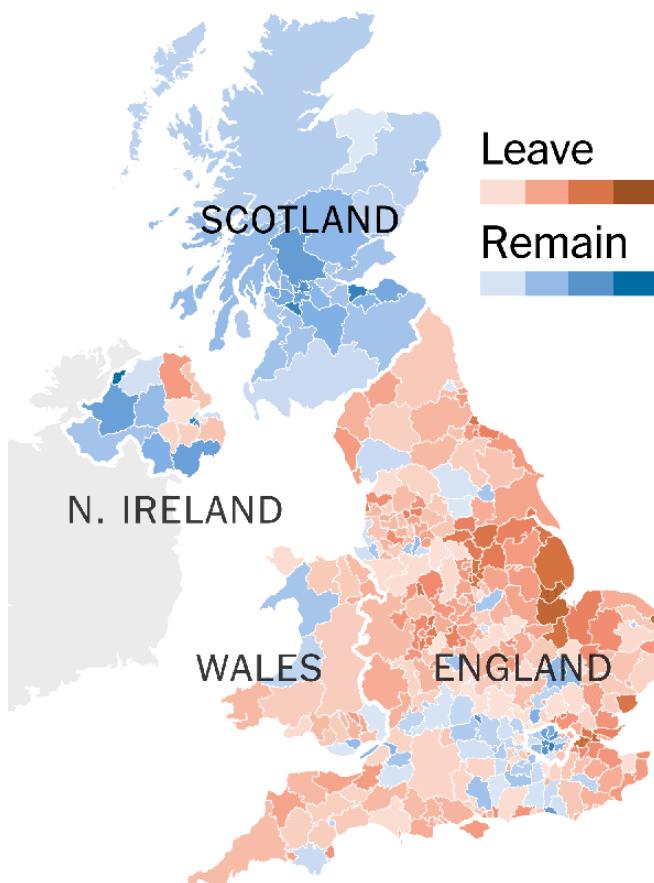
- Methodology and sampling techniques.
- Both surveys are valid in their specific context, both were done professionally but they tell a different story



# Data Lies - Surveys

## Surveys:

- Questions



### Results by location

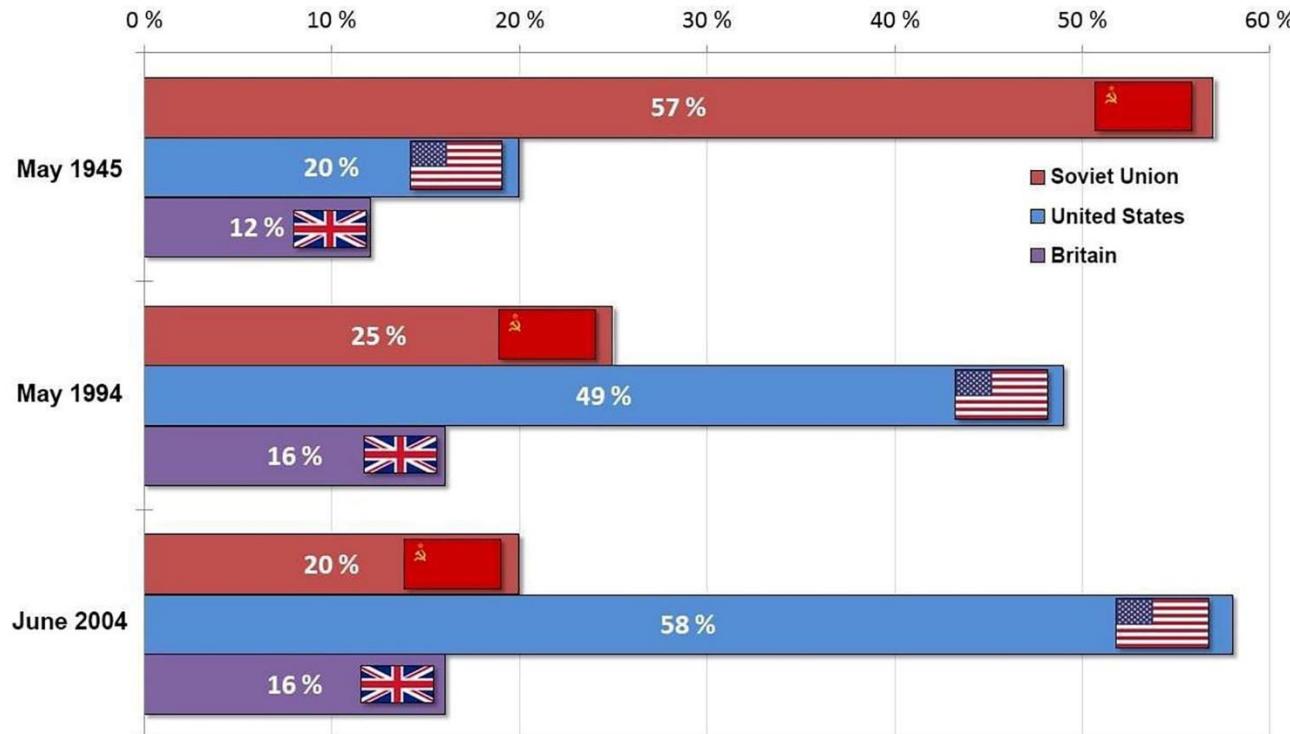
	Remain	Leave	
Britain	48%	52%	
England	47%	53%	
London	60%	40%	
Scotland	62%	38%	
Wales	48%	53%	
N. Ireland	56%	44%	

# Data Lies - Surveys

## Surveys:

- Manipulation

Poll in France: “In your view, which is the country that contributed the most to the defeat of Nazi Germany in 1945?” (Source : sondages IFOP 1945, 1994, 2004)



© Olivier Berruyer, [www.les-crises.fr](http://www.les-crises.fr) (Translated by Ben Norton)

# Ted Talk

---

**Ted Talk:** <https://www.youtube.com/watch?v=1Totz8aa2Gg>

<https://www.youtube.com/watch?v=gIXoPT-uWu0>

# Data Lies - Surveys

---

## Surveys:

- Do people lie in surveys?
- "A third of people in the UK will not give truthful answers about themselves when asked questions by pollsters, according to a new survey."
- Example: 1992 election, when people did not want to admit voting for John Major.

	REPORTED ON SURVEY	OFFICIAL COUNT
Registered to vote	83%	69%
Voted in last presidential election	73%	61%
Voted in last mayoral election	63%	36%
Have a library card	20%	13%
Gave to a recent Community Chest charitable drive	67%	33%

# Data Lies - Surveys

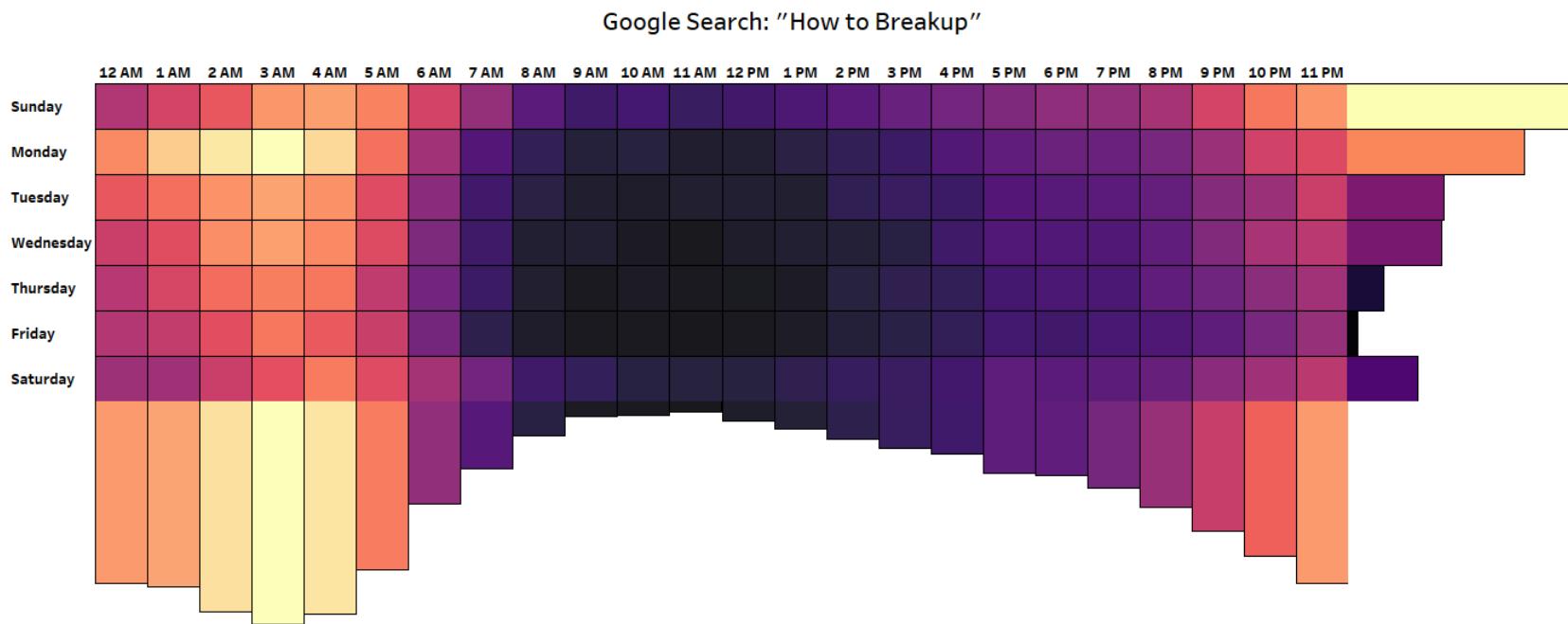
## Google Trends:

- Does anonymity help?



# Data Lies - Breakups

**Google Trends:** When to break-up with your significant other?



# Data Lies - Elections

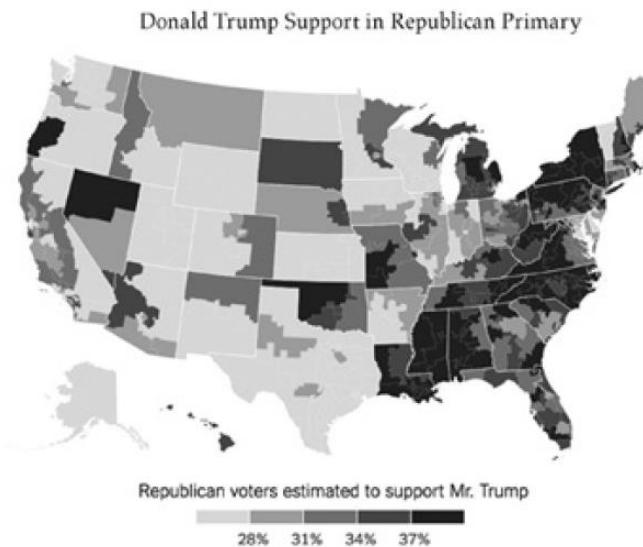
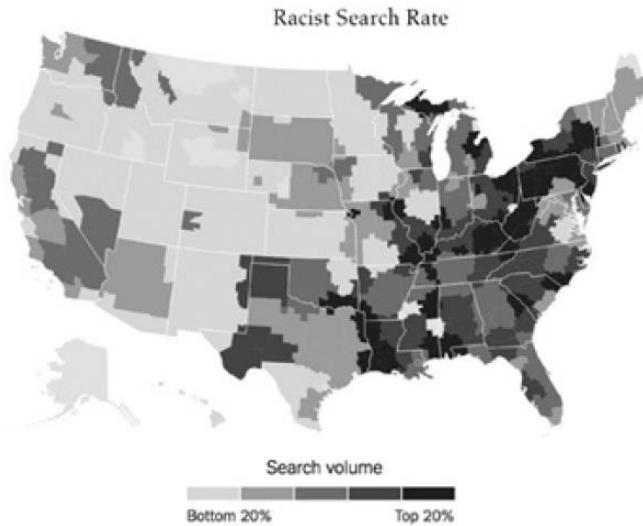
---

Google Trends:



# Data Lies - Elections

## Google Trends:



# Data Lies - Investment

---

Google Trends:



# Data Lies - Investment

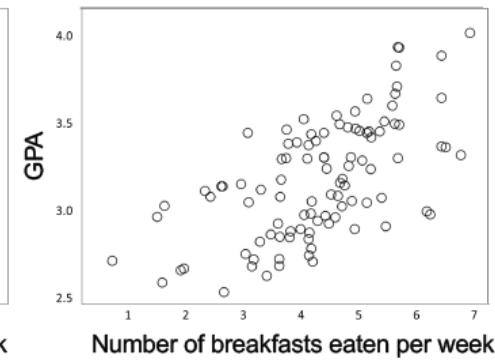
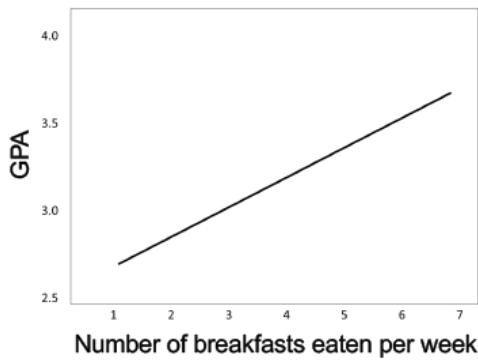
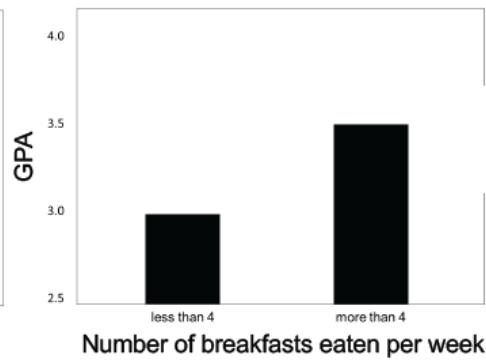
## Google Trends:



# Data Does not Lie – You Do

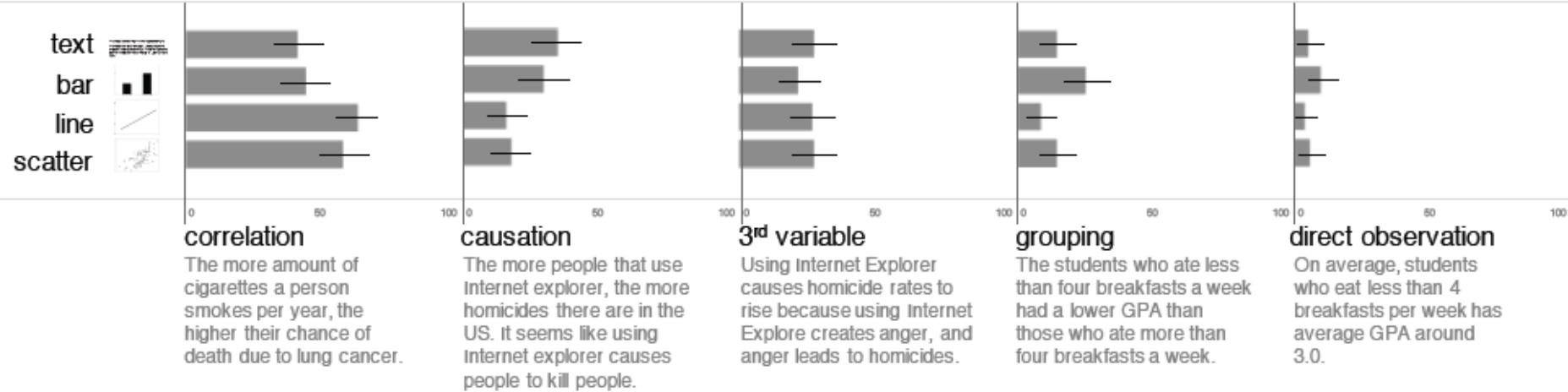
## Graphs and Manipulation:

When students eat breakfast very often (more than 4 times a week), their GPA is around 3.5; while when students eat breakfast not very often (less than 4 times a week), their GPA is around 3.0.



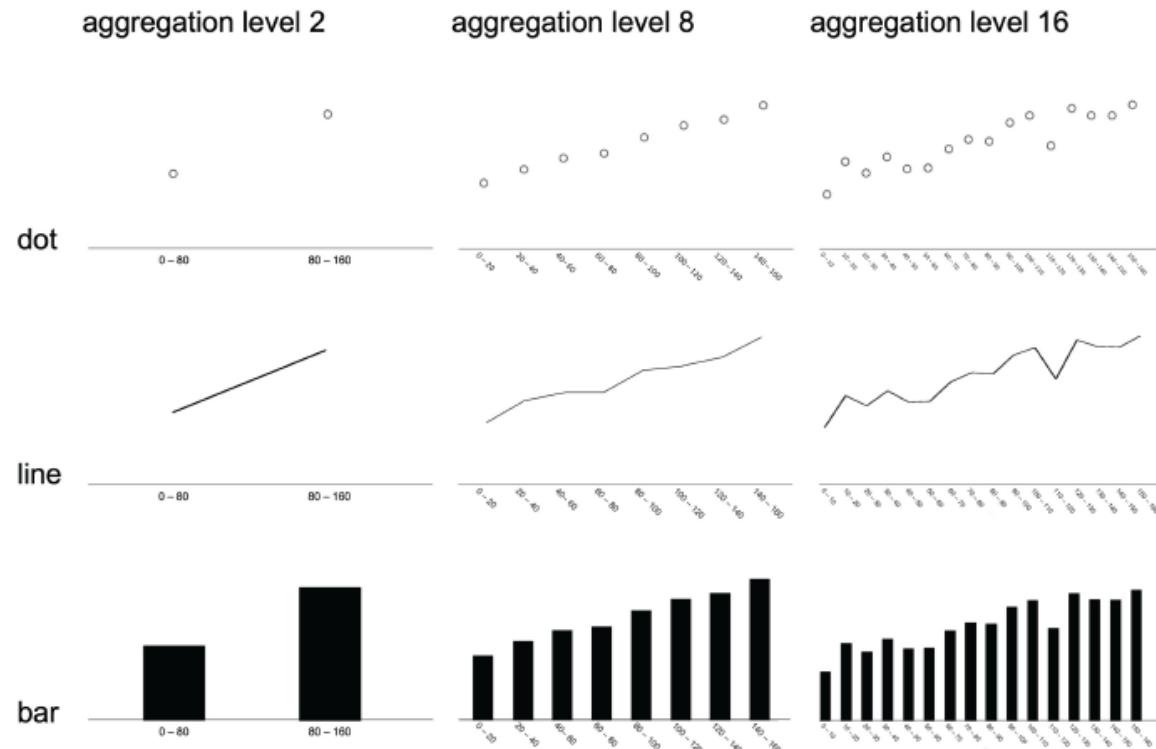
# Data Does not Lie – You Do

## Graphs and Manipulation:



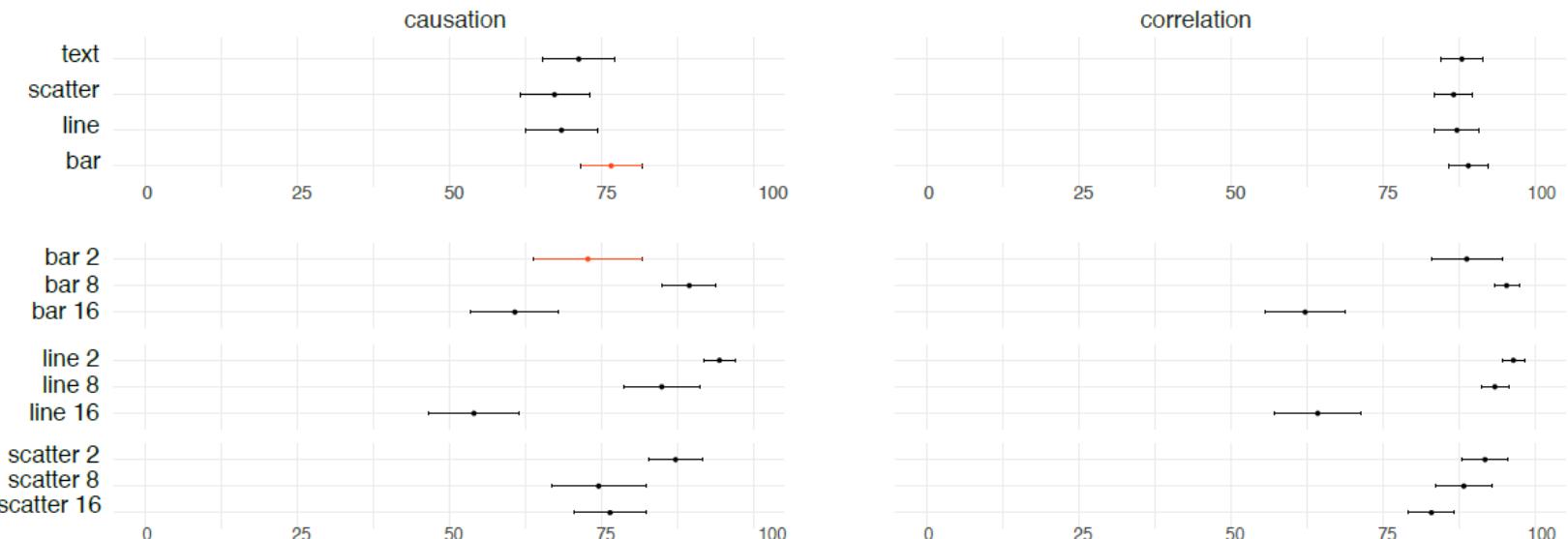
# Data Does not Lie – You Do

## Graphs and Manipulation:



# Data Does not Lie – You Do

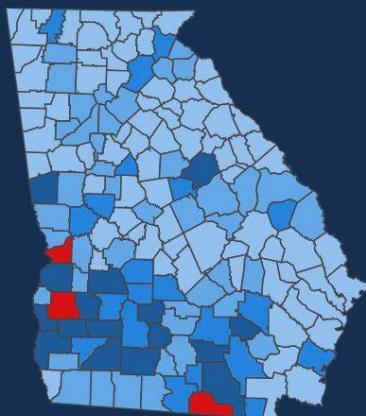
## Graphs and Manipulation:



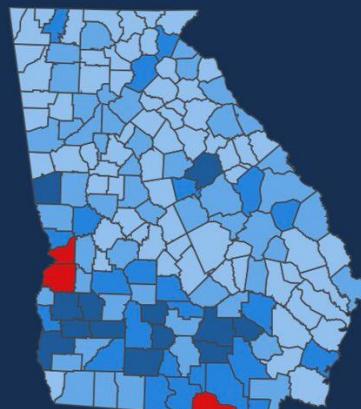
# Data Does not Lie – You Do

## Graphs and Manipulation:

Cases per 100K▼

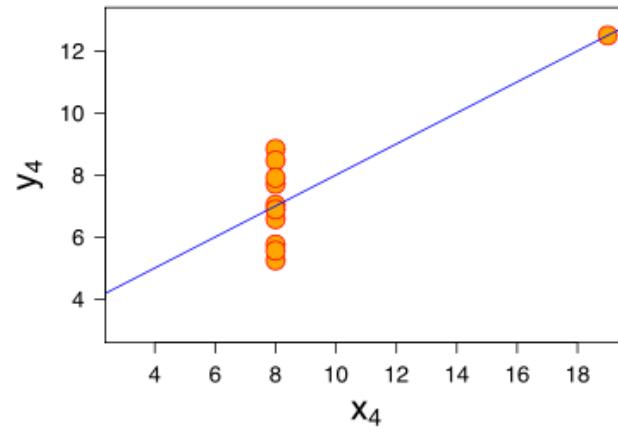
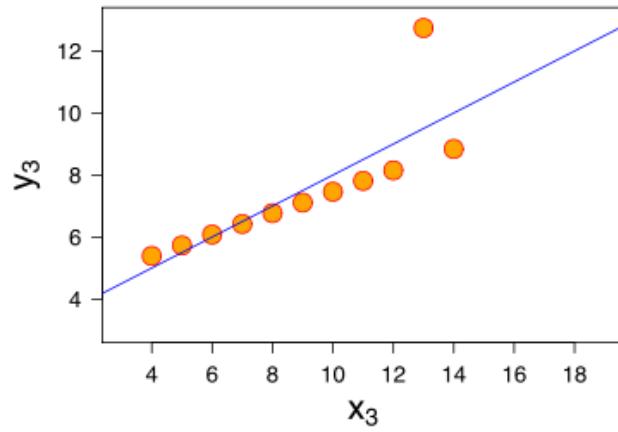
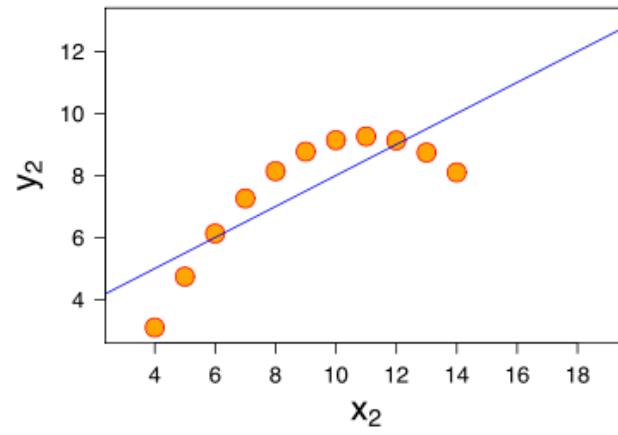
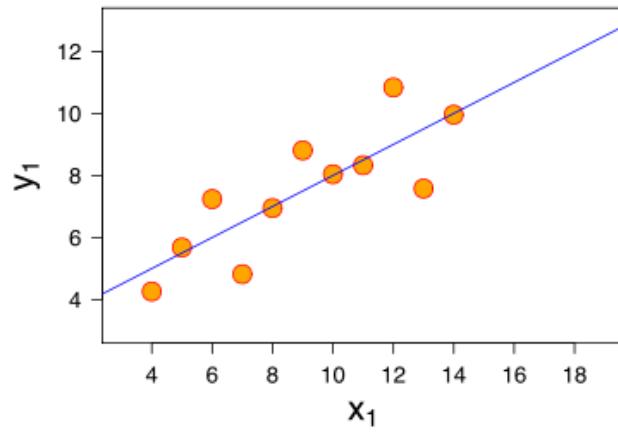


Cases per 100K▼



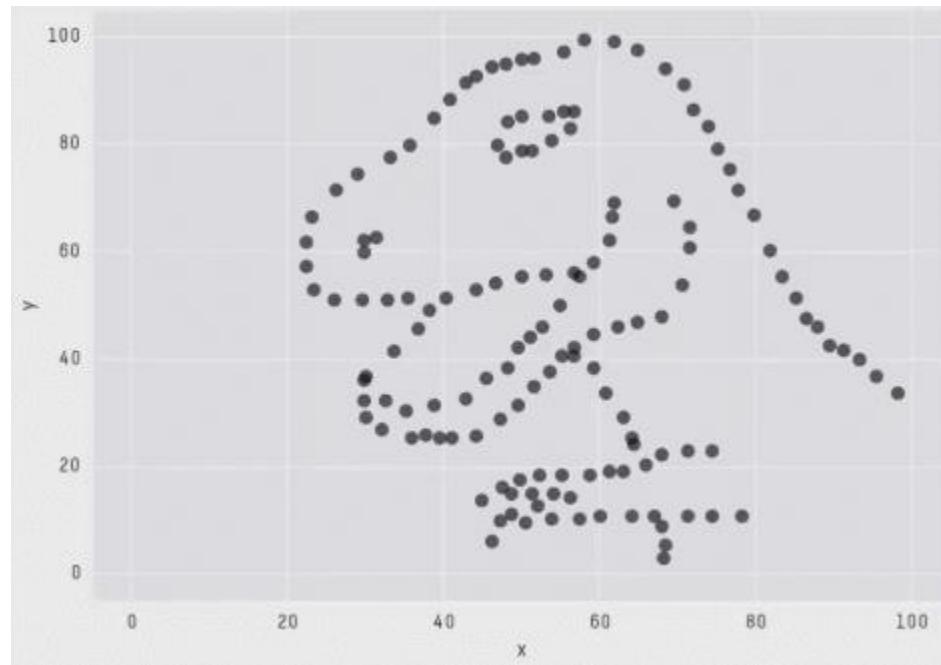
# Data Does not Lie – You Do

## Graphs and Manipulation:



# Data Does not Lie – You Do

## Graphs and Manipulation:



X Mean: 54.2659224  
Y Mean: 47.8313999  
X SD : 16.7649829  
Y SD : 26.9342120  
Corr. : -0.0642526

# Everybody Lies or Do They?

## Incompatibility :

PLOS MEDICINE

advanced search

OPEN ACCESS

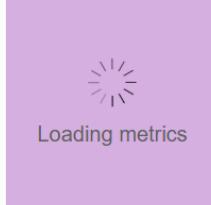
ESSAY

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • <https://doi.org/10.1371/journal.pmed.0020124>

Article	Authors	Metrics	Comments	Media Coverage
▼				

 Loading metrics

[Download PDF](#) ▾  
[Print](#) [Share](#)

 Check for updates

**Related PLOS Articles**

**has COMPANIONS**  
Why Current Publication Practices May Distort Science  
[View Page](#) [PDF](#)

*Why Most Published*

# Everybody Lies or Do They?

---

## Incompatibility :



# Shortages

---

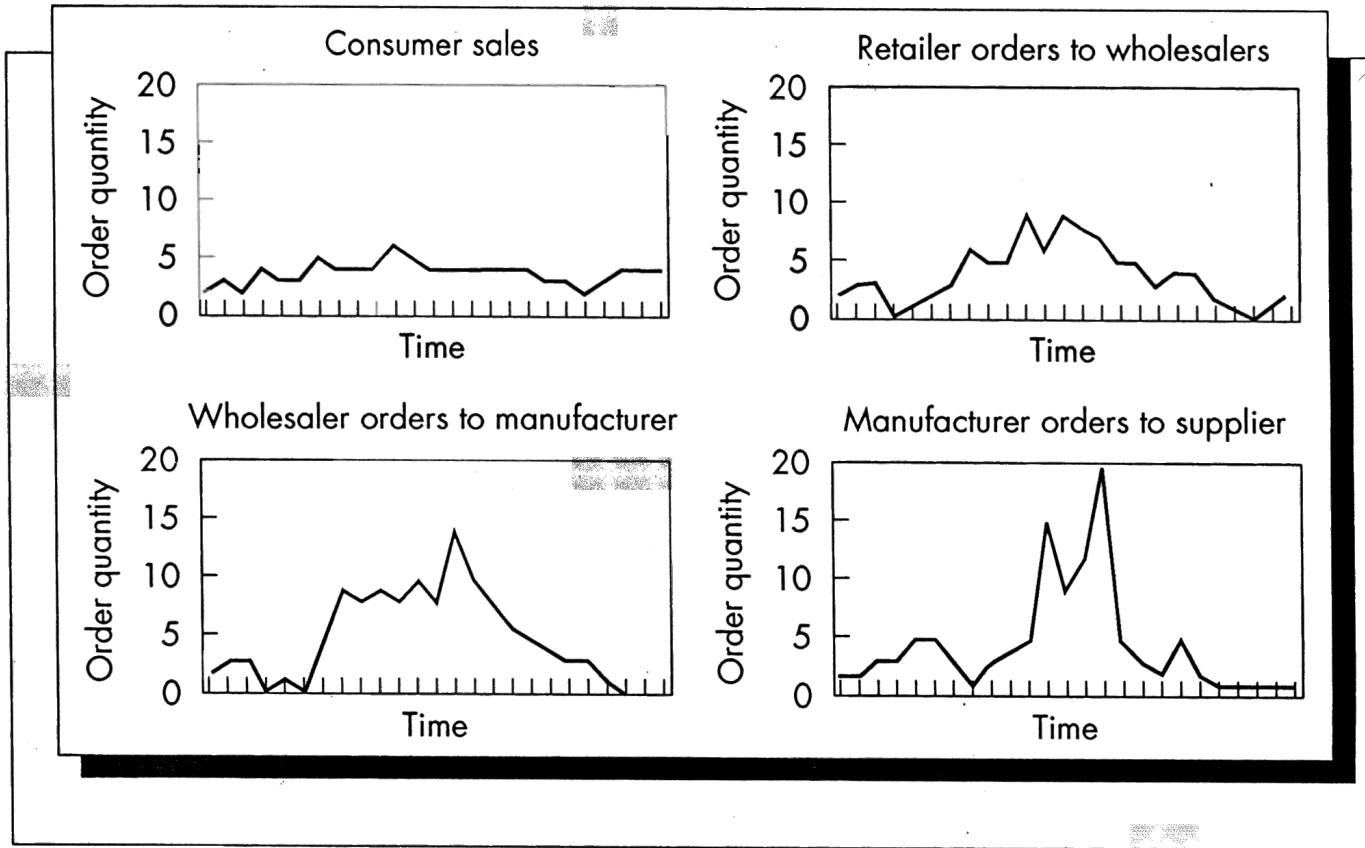
## Covid-19 and Toilet Paper:

- <https://www.tampabay.com/news/business/2020/04/02/why-panic-buying-toilet-paper-is-even-worse-than-you-think/>
- <https://marker.medium.com/what-everyones-getting-wrong-about-the-toilet-paper-shortage-c812e1358fe0>



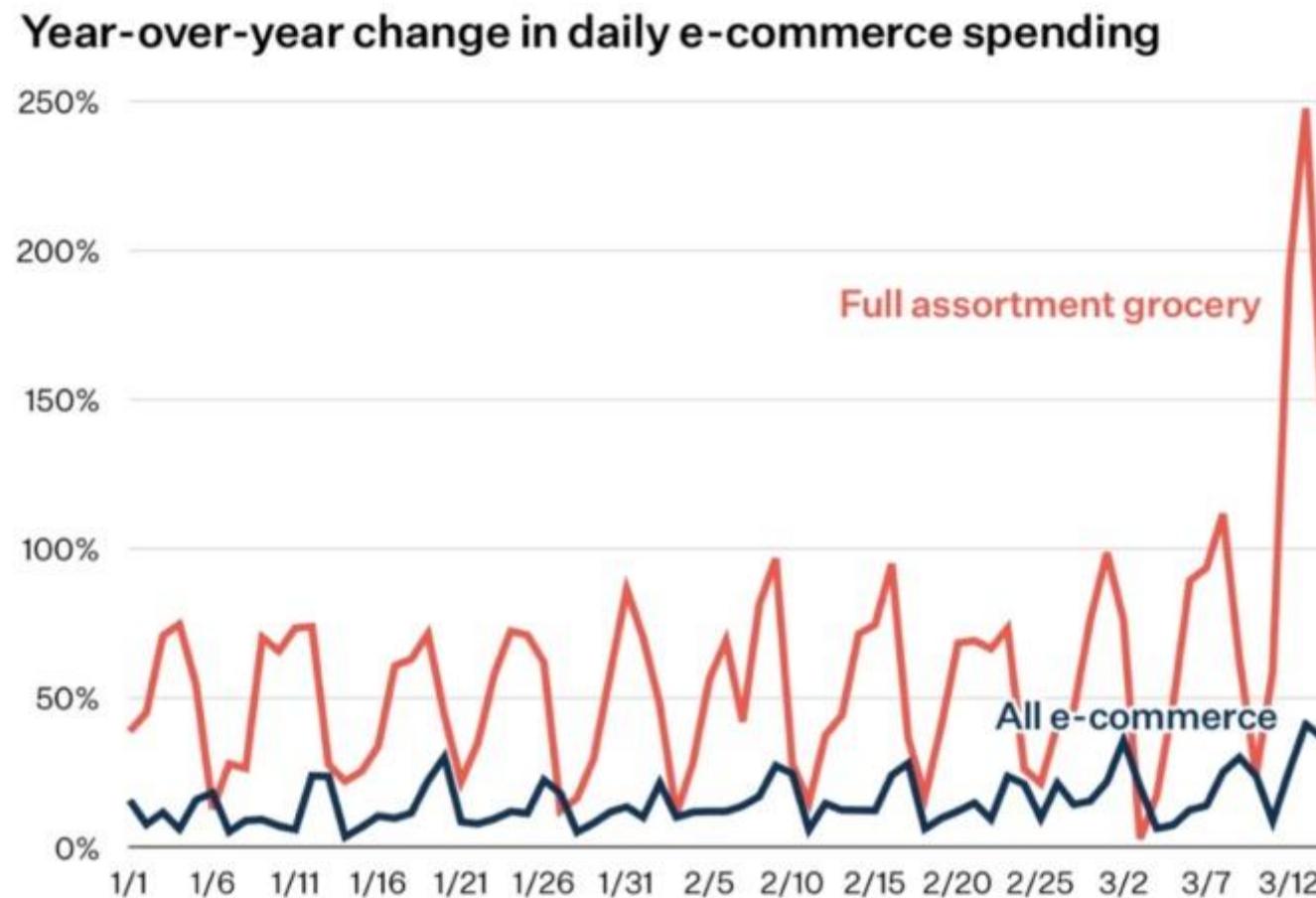
# Bullwhip Effect

TM12–2  
Caption to Come



# Everybody Lies or Do They?

E-commerce:



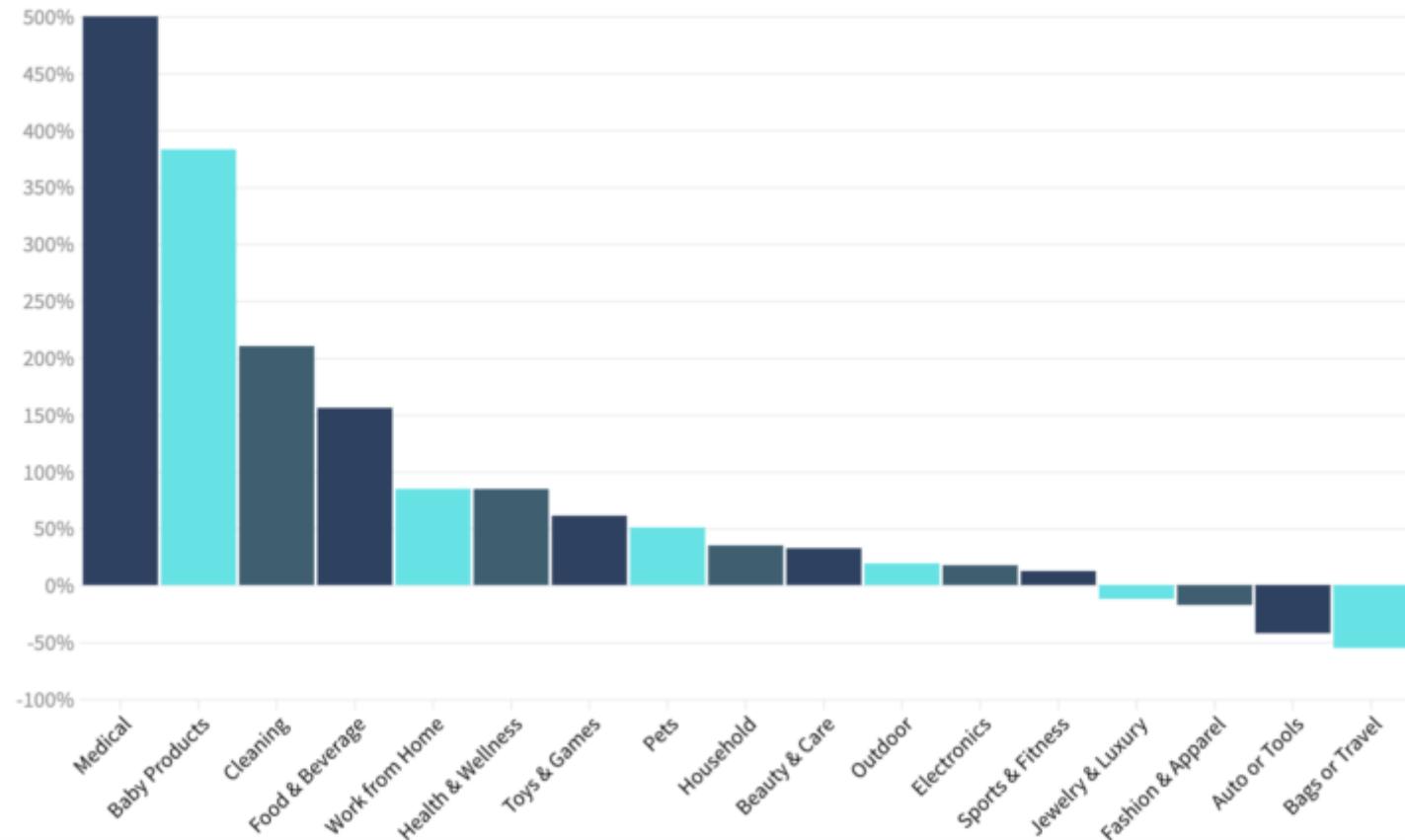
Data: Rakuten Intelligence

# Everybody Lies or Do They?

## E-commerce:

**Ecommerce Consumer Sales (COVID-19) +28.48%**

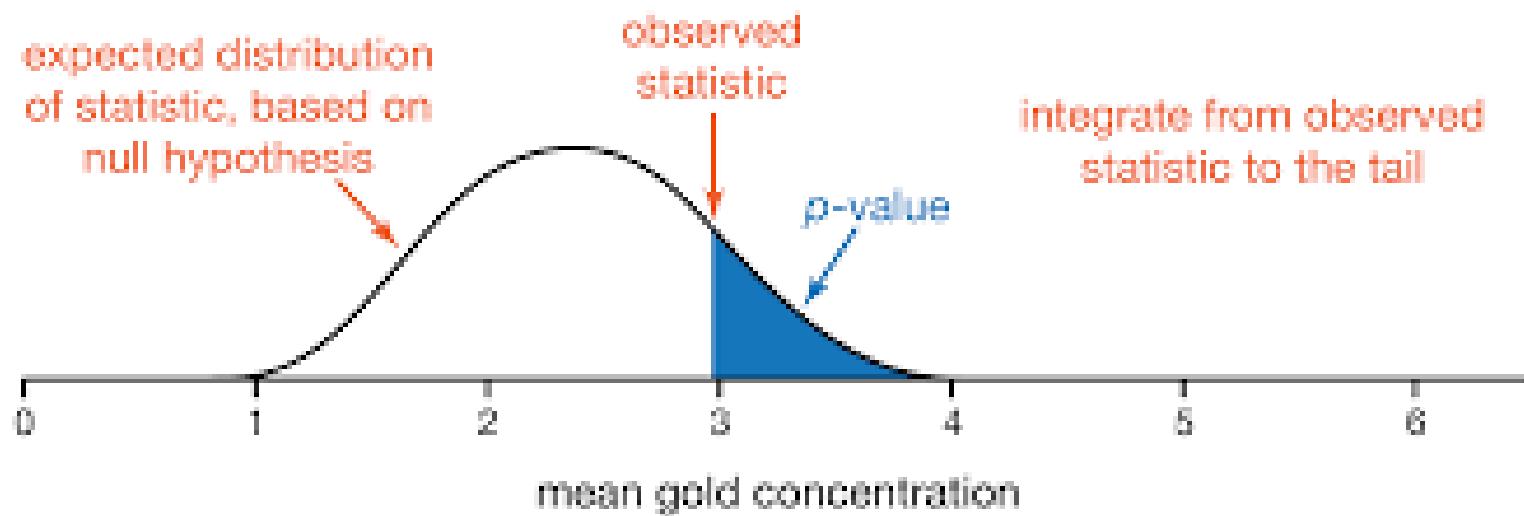
Aggregated via ShipBob, ShipHero, Attentive, Stackline, CTC, Klaviyo & Adobe



# Everybody Lies or Do They?

## P-Value:

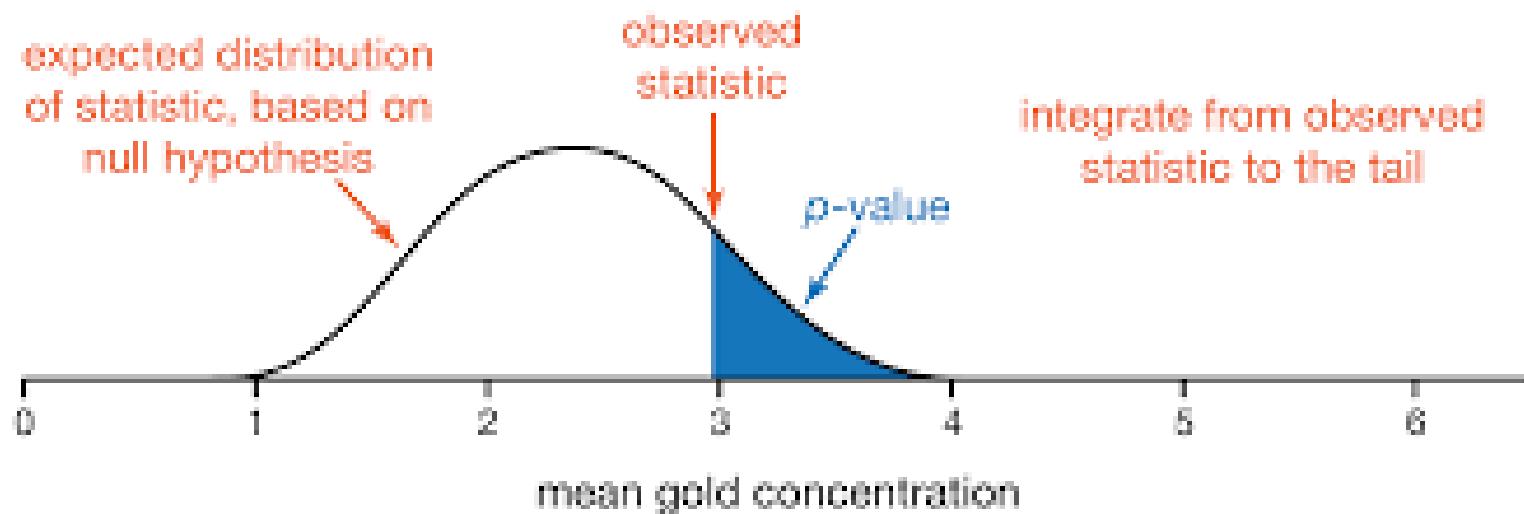
reject null hypothesis if  $p\text{-value} < \alpha$   
accept null hypothesis if  $p\text{-value} > \alpha$



# Everybody Lies or Do They?

## Publication Bias:

reject null hypothesis if  $p\text{-value} < \alpha$   
accept null hypothesis if  $p\text{-value} > \alpha$



# Biases

---

Focus:

<https://www.youtube.com/watch?v=vJG698U2Mvo>

# Information Bias

---

**Data:** Information bias (also called observation bias or measurement bias) happens when key information is either measured, collected, or interpreted inaccurately

# Information Bias

---

**Weather Forecast:** The National Hurricane Center nailed its forecast of Katrina; it anticipated a potential hit on the city almost five days before the levees were breached, and concluded that some version of the nightmare scenario was probable more than forty-eight hours away.



# Information Bias

---

**Weather Forecast:** Laplace's Demon::

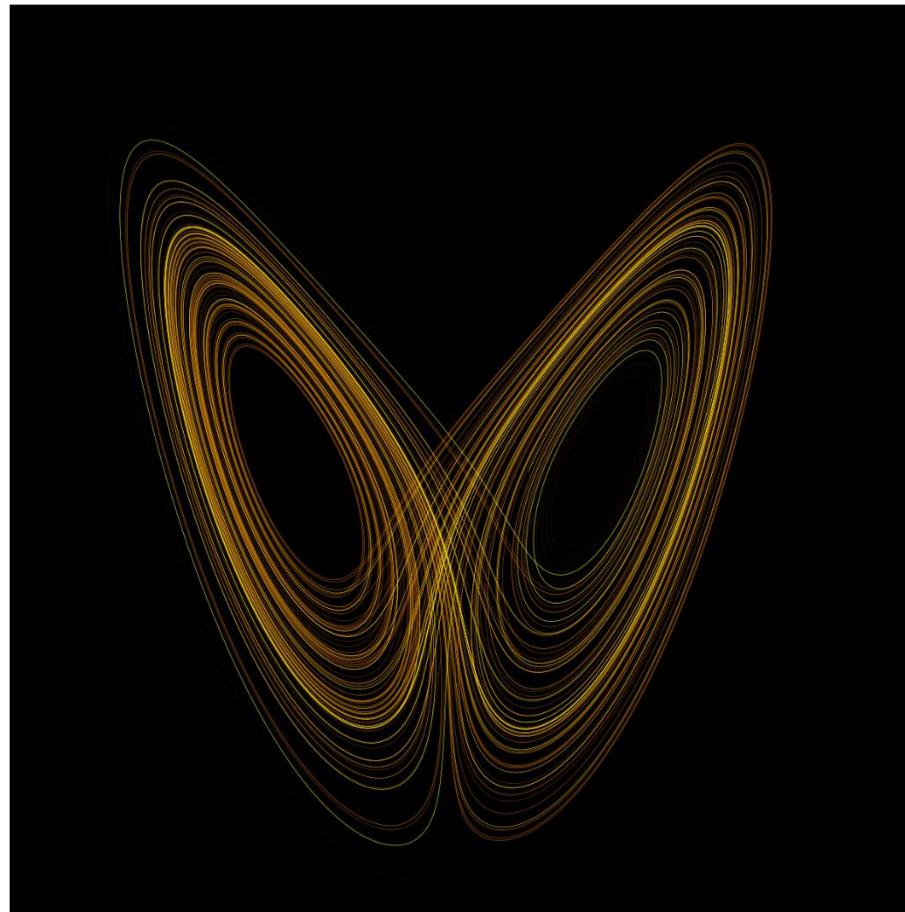


Artwork by Viktor Beekmān  
[instagram.com/viktordepictor](https://instagram.com/viktordepictor)

# Information Bias

---

**Weather Forecast:** Chaos Theory



# Information Bias

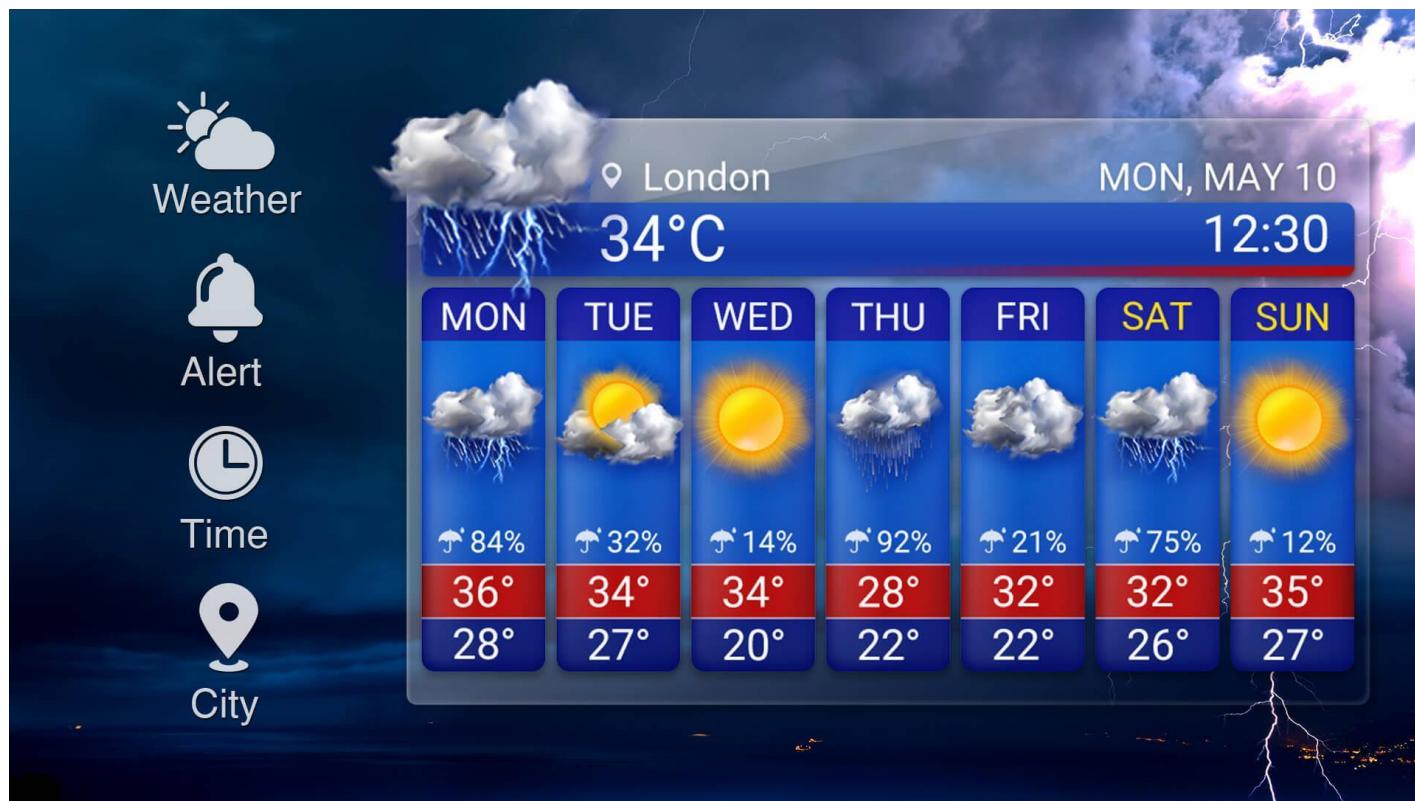
---

## Weather Forecast: Decimals

**3.14**1592653589793238462643383279502884197169399375105  
82097494459230781640628620899862803482534211706798214  
80865132823066470938446095505822317253594081284811174  
50284102701938521105559644622948954930381964428810975  
66593344612847564823378678316527120190914564856692346  
03486104543266482133936072602491412737245870066063155  
88174881520920962829254091715364367892590360011330530  
54882046652138414695194151160943305727036575959195309  
21861173819326117931051185480744623799627495673518857  
52724891227938183011949129833673362440656643086021...

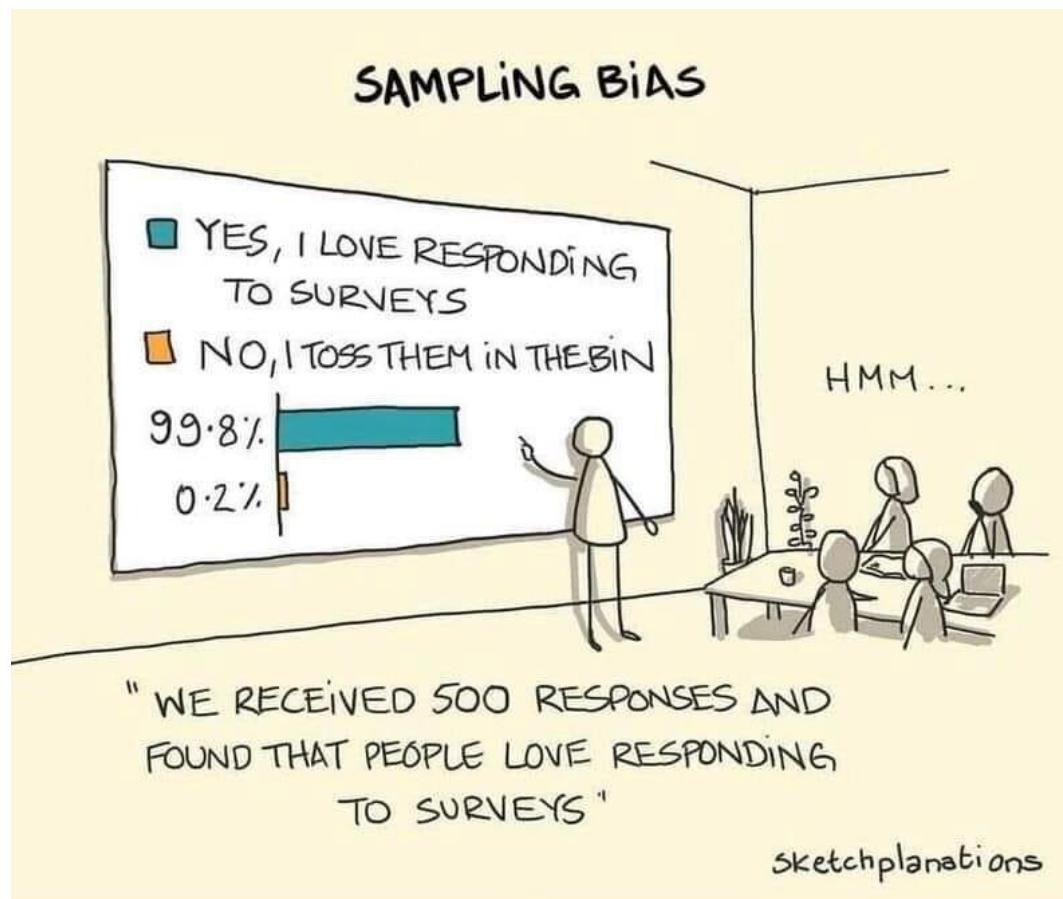
# Information Bias

**Weather Forecast:** Short Term, Long Term



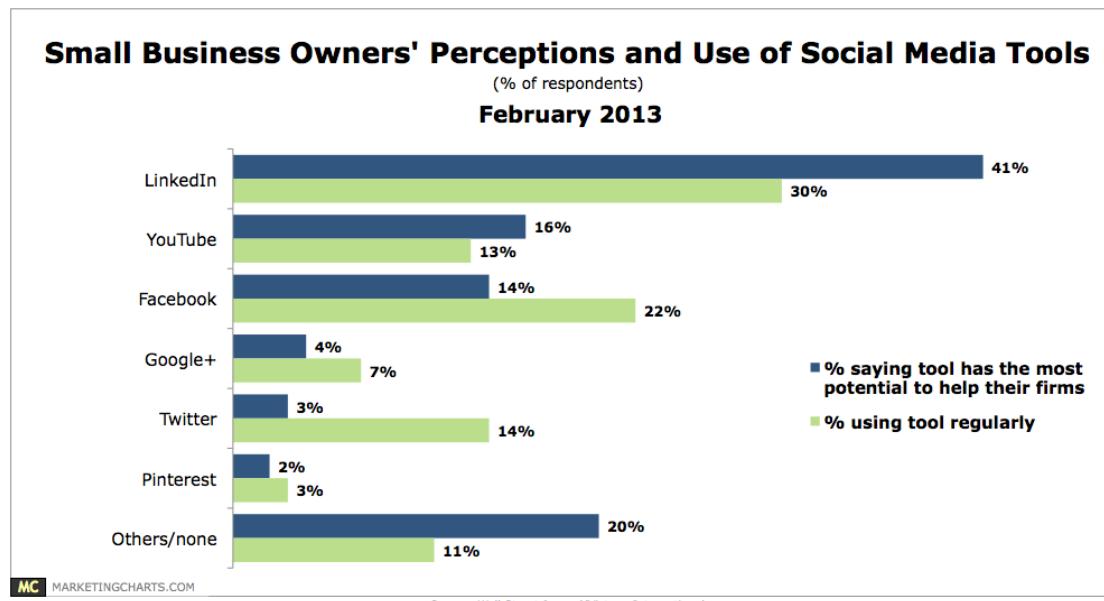
# Selection Bias

**Data:** When individuals or groups in a study differ systematically from the population of interest leading to a systematic error in an association or outcome.



# Selection Bias

**Data:** When individuals or groups in a study differ systematically from the population of interest leading to a systematic error in an association or outcome.



# Volunteer Bias

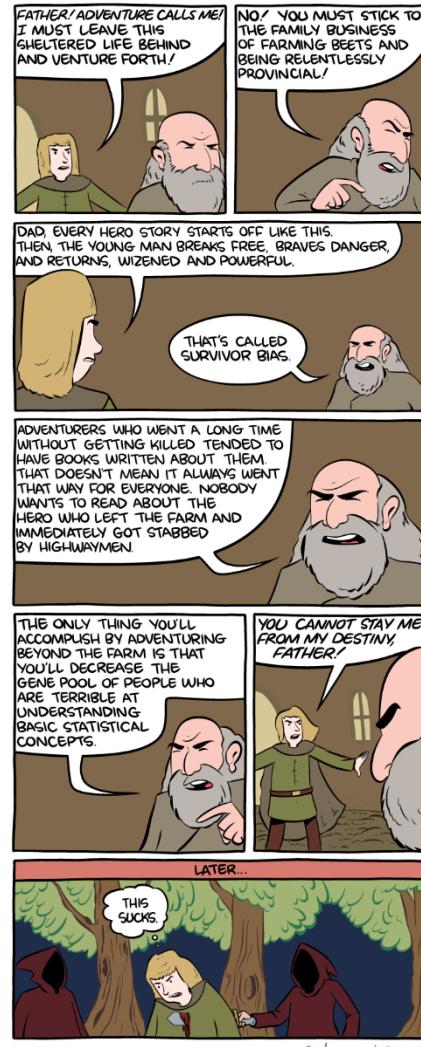
---

**Sampling:** Volunteer bias is the idea that people who volunteer to participate in studies do not represent the general population



# Survivorship Bias – Nobody Lies

## Survivor Bias:

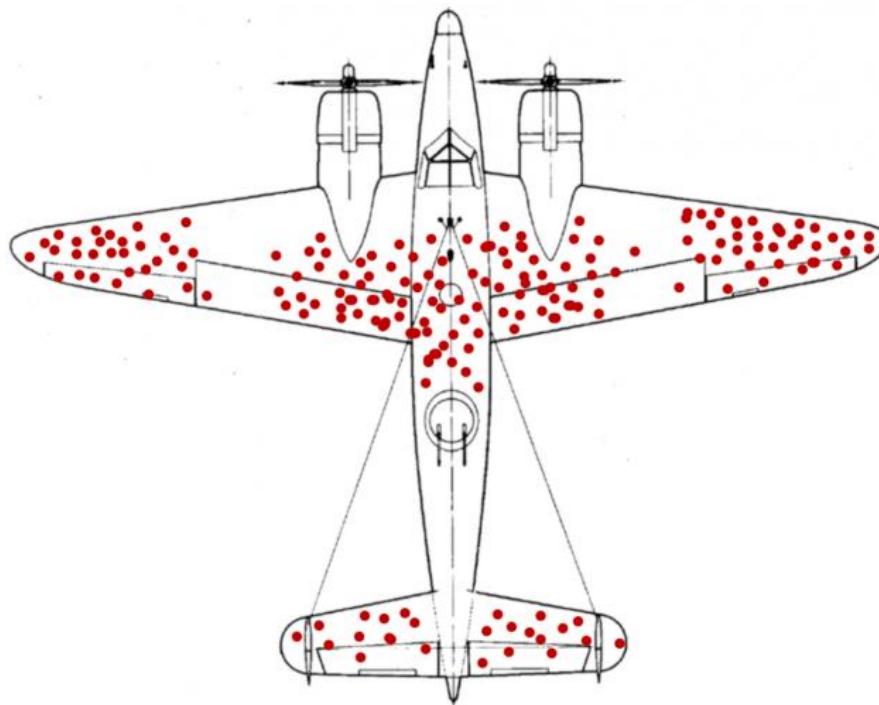


# Survivorship Bias – Nobody Lies

---

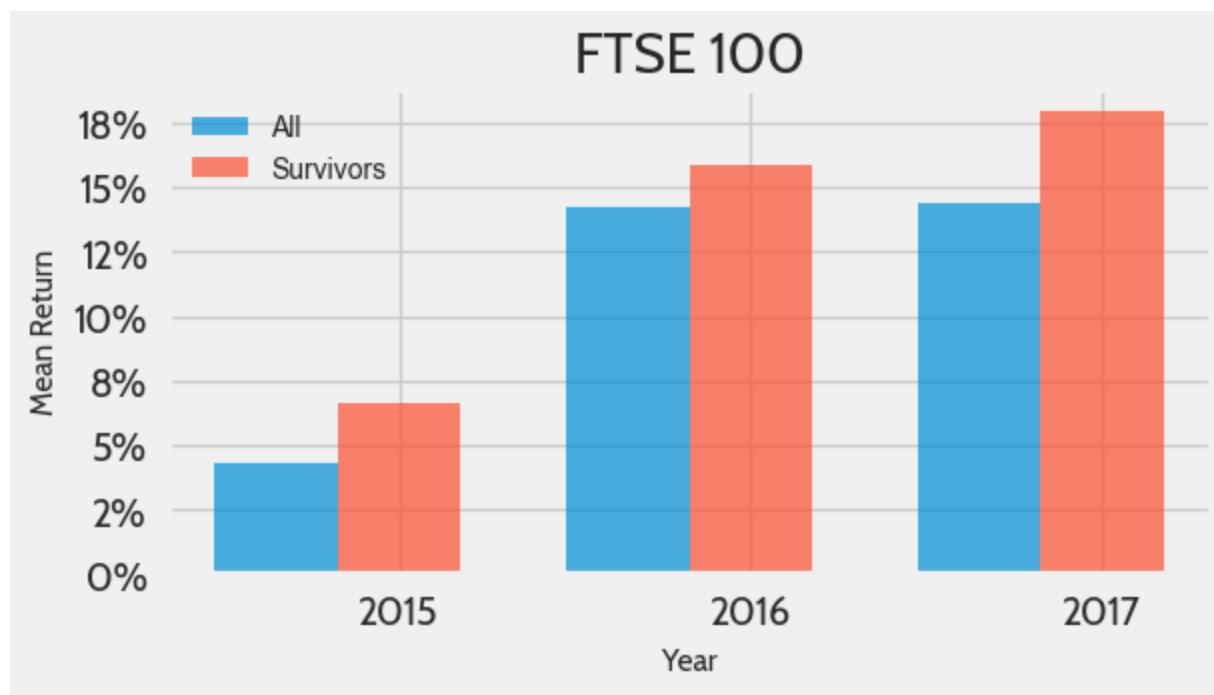
**Survivor Bias:** The Survivorship Bias (sometimes styled *Survivor Bias*) is the human tendency to value the seen or available at the expense of the unseen.

Abraham Wald set out to determine the weaknesses of airplanes in World War II for the Center of Naval Analyses' Statistical Research Group (SRG).



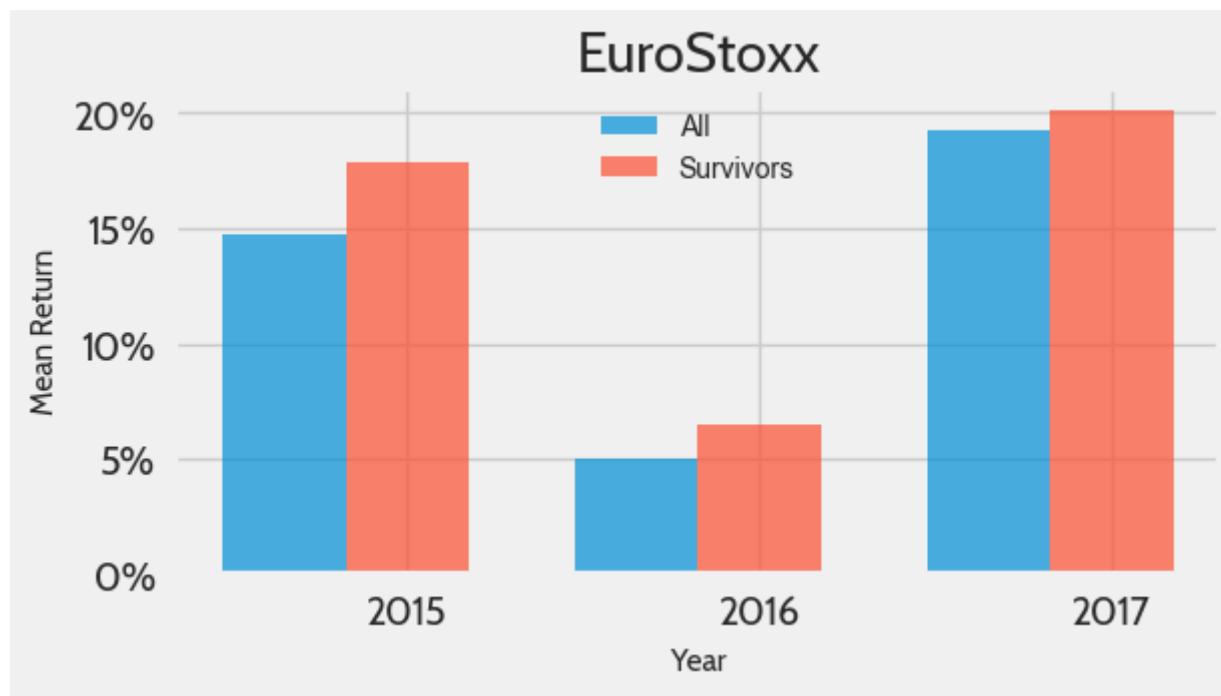
# Survivorship Bias – Finance

Stock Index:



# Survivorship Bias – Finance

Stock Index:



# Survivorship Bias – Business

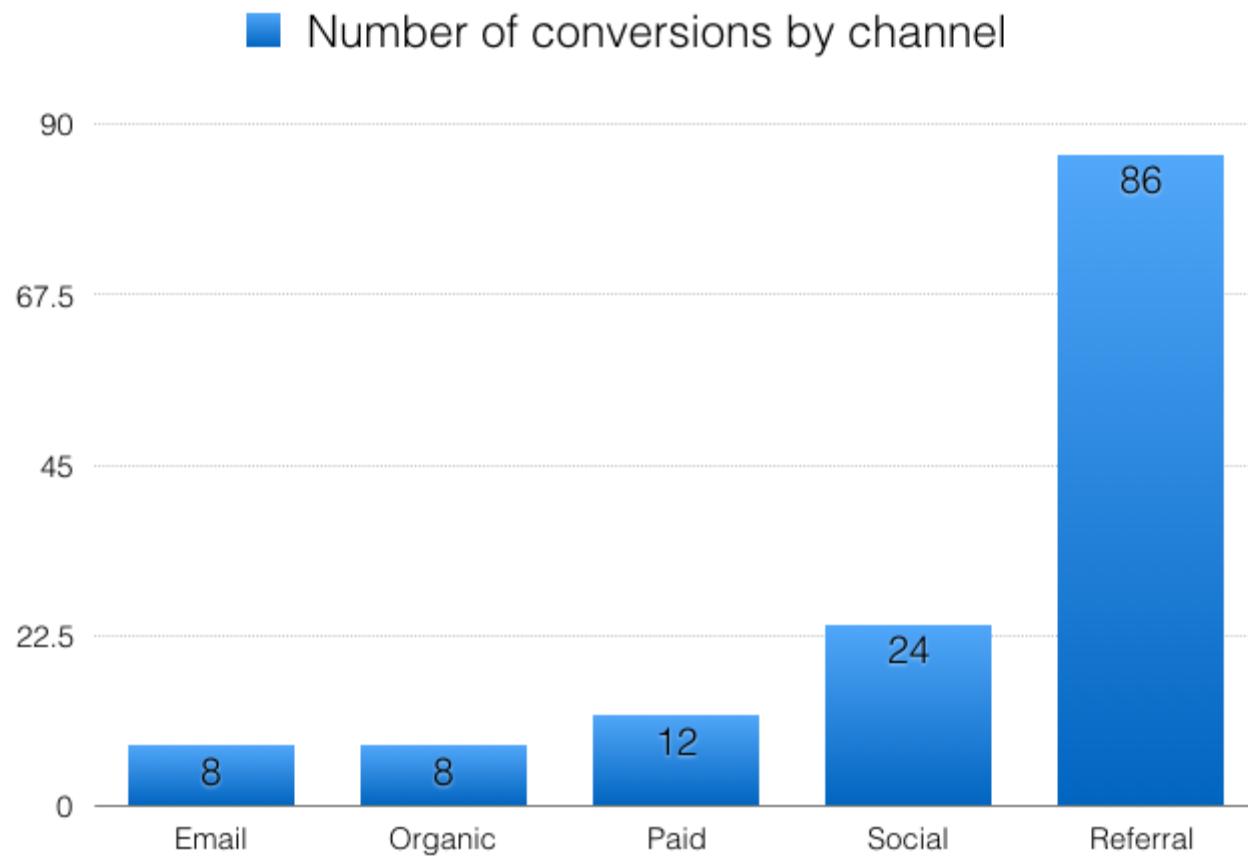
---

**Millionaires:** Business Insider: valedictorians are *less* likely to become millionaires because millionaires in some survey had a 2.9 average GPA



# Survivorship Bias – Marketing

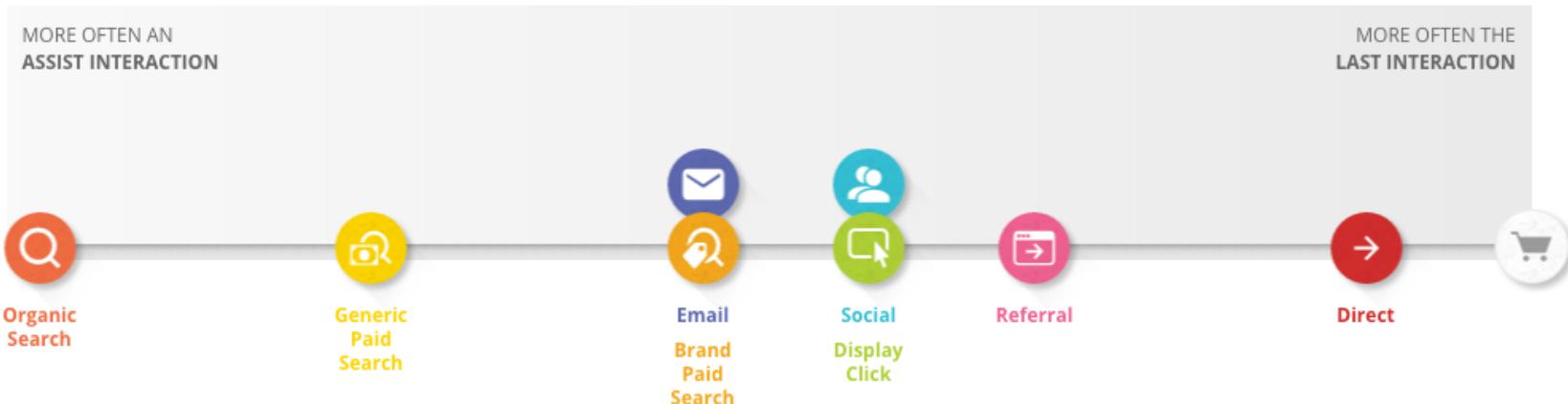
**Marketing:** “Where should I spend my marketing money?”



# Survivorship Bias – Marketing

**Marketing:** “Where should I spend my marketing money?”

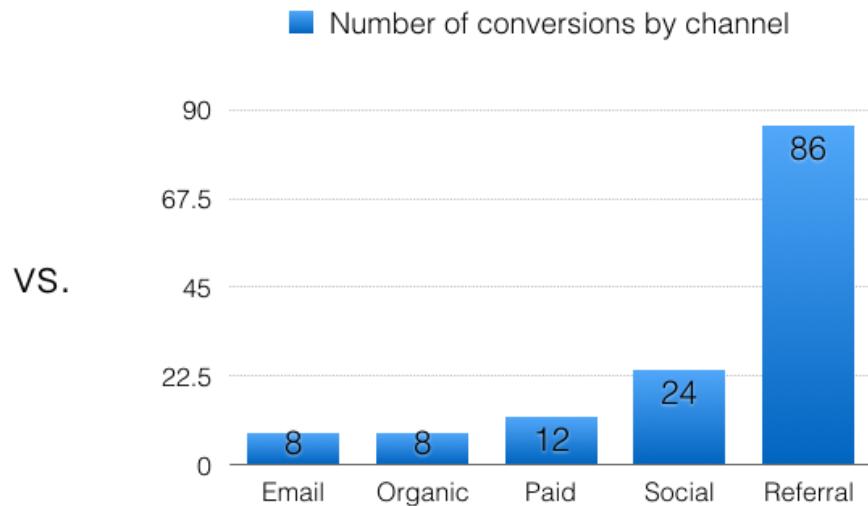
Explore how marketing channels for Small  
businesses in the Shopping industry  
in The U.K. influence the purchase decision.



# Survivorship Bias – Marketing

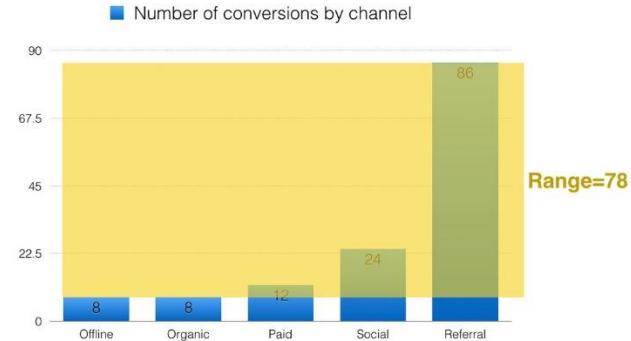
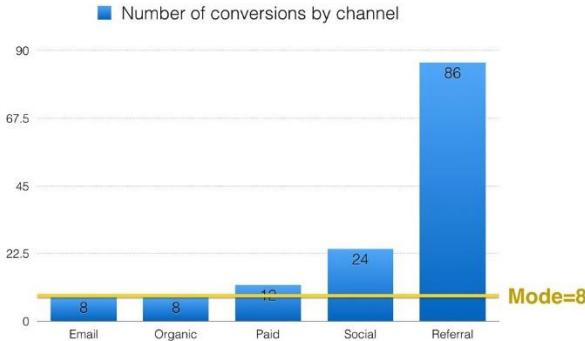
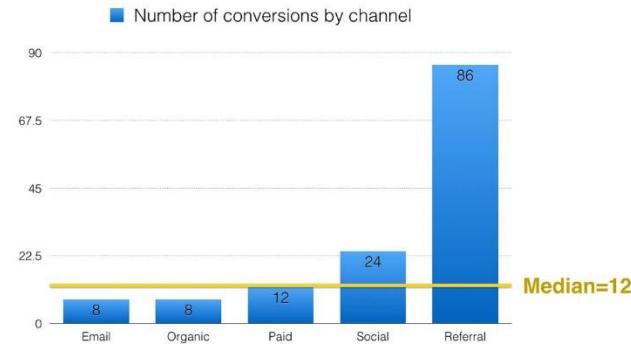
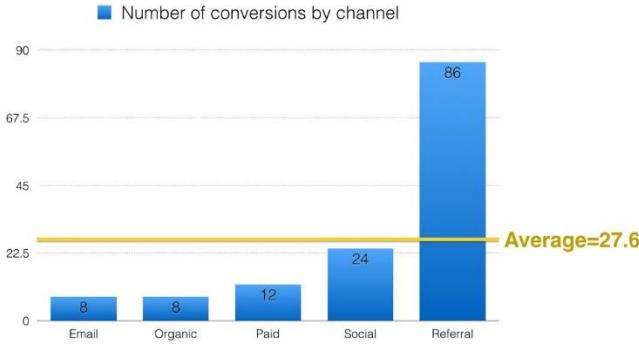
**Marketing:** “Where should I spend my marketing money?”

We have on average  
**28 conversions**  
per channel



# Survivorship Bias – Marketing

Marketing: “Where should I spend my marketing money?”



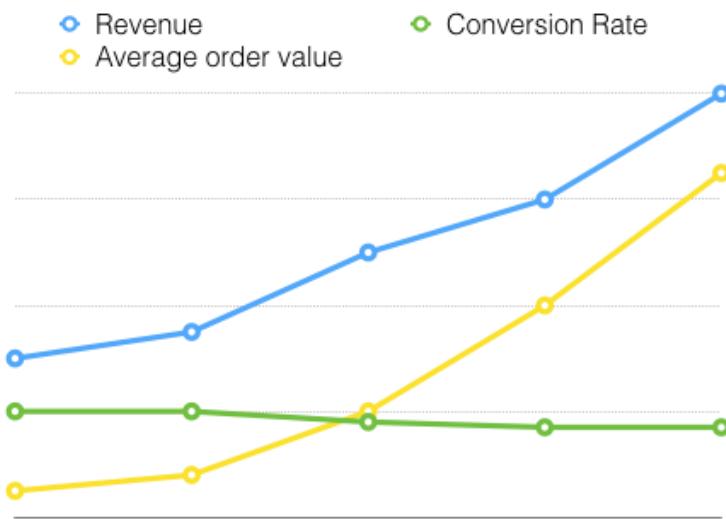
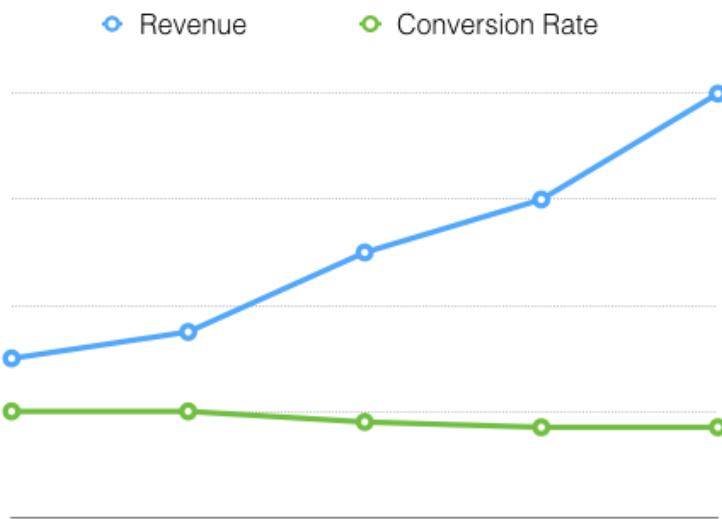
# Survivorship Bias – Marketing

Marketing: “Where should I spend my marketing money?”



# Survivorship Bias – Marketing

Marketing: “Where should I spend my marketing money?”

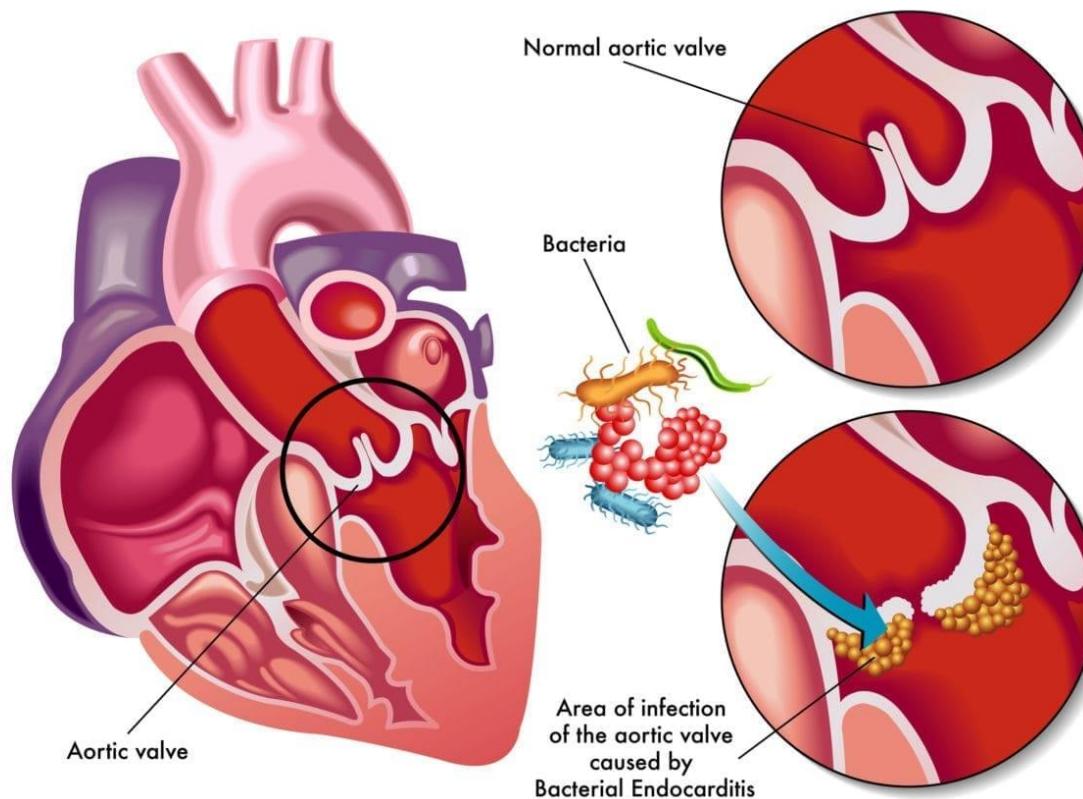


*“If you torture the data long enough, it will confess.”*

– Ronald H. Coase

# Survivorship Bias – Health

**Health:** “Are the treatments effective?”



# Cognitive Bias

---

**Truth vs Beliefs:** “As you consider the next question, please assume that Steve was selected at random from a representative sample. An individual has been described by a neighbor as follows: “Steve is very shy and withdrawn, invariably helpful but with little interest in people or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail.” Is Steve more likely to be a librarian or a farmer?”



# Availability Bias

**Availability:** "If a random word is taken from an English text, is it more likely that the word starts with a K, or that K is the third letter?"



# Conformation Bias

---

**Truth vs Beliefs:** “*Most people use statistics like a drunk man uses a lamp post; more for support than illumination*”

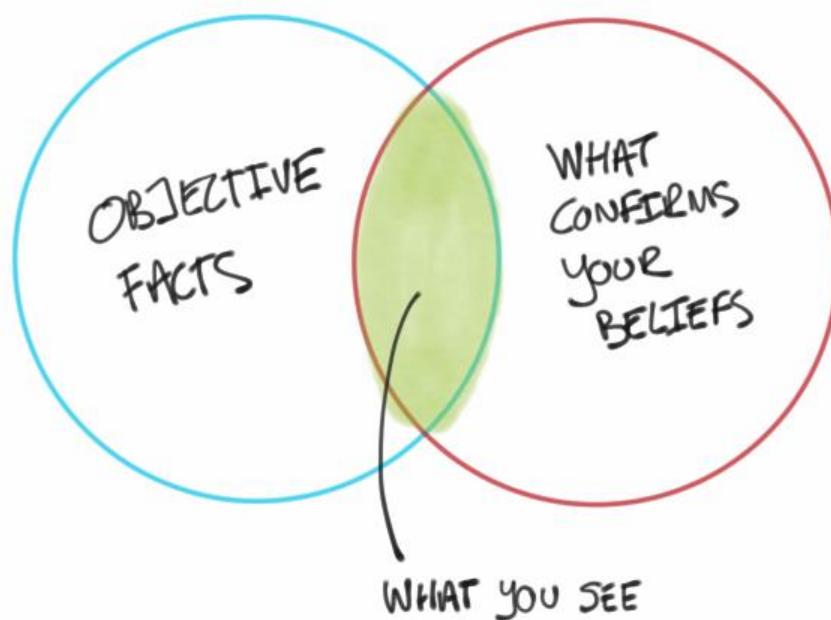


©Fredrik D. Bodin

# Conformation Bias

---

**Truth vs Beliefs:** “*Most people use statistics like a drunk man uses a lamp post; more for support than illumination*”

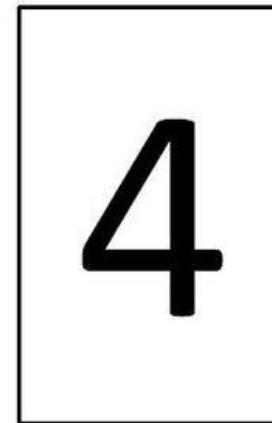
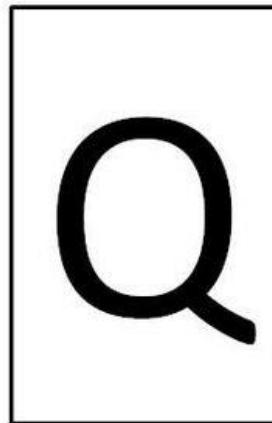
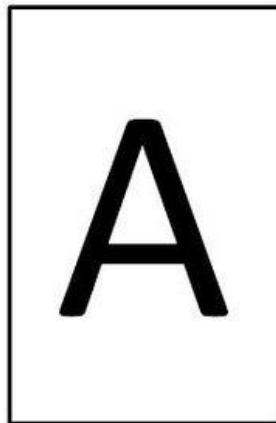


# Conformation Bias

---

Quiz:

*If the card has a vowel on one side, then it must have an even number on the other side.*

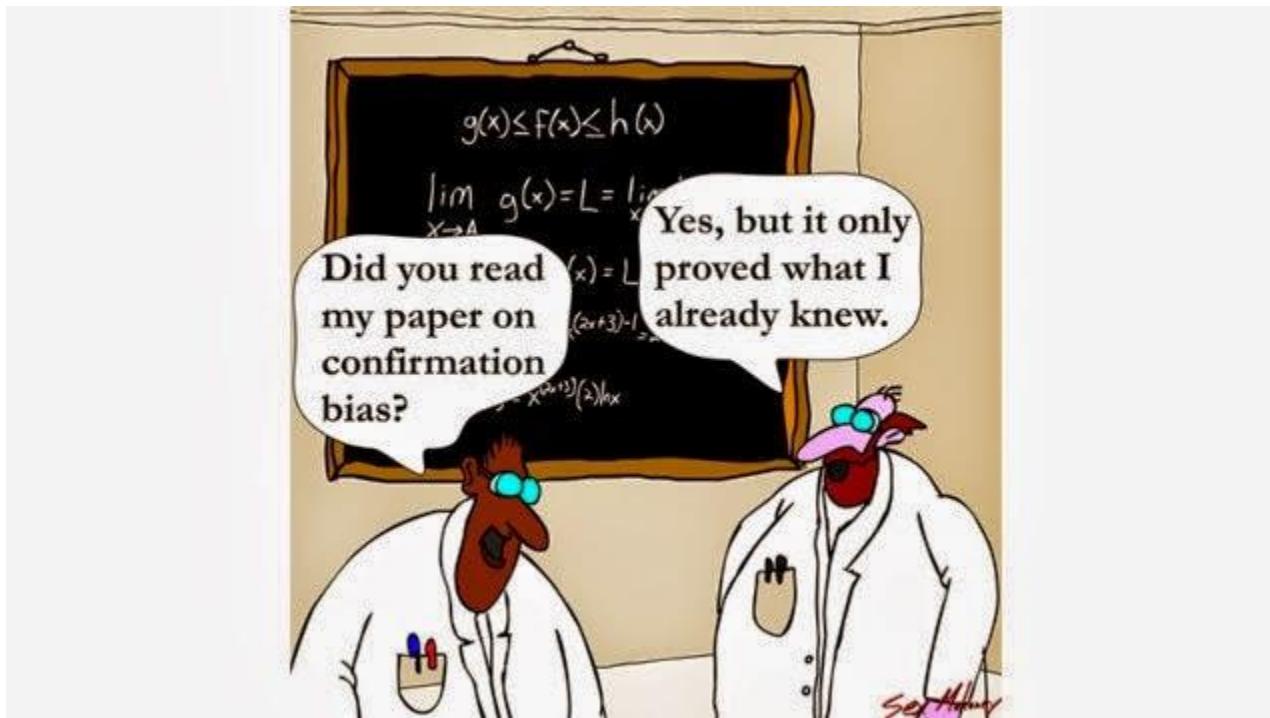


Which two cards would you turn over to test the rule?

- A) A, 4                    C) Q, 4
- B) A, 7                    D) Q, 7

# Conformation Bias

**How to correct:** More about culture than technology



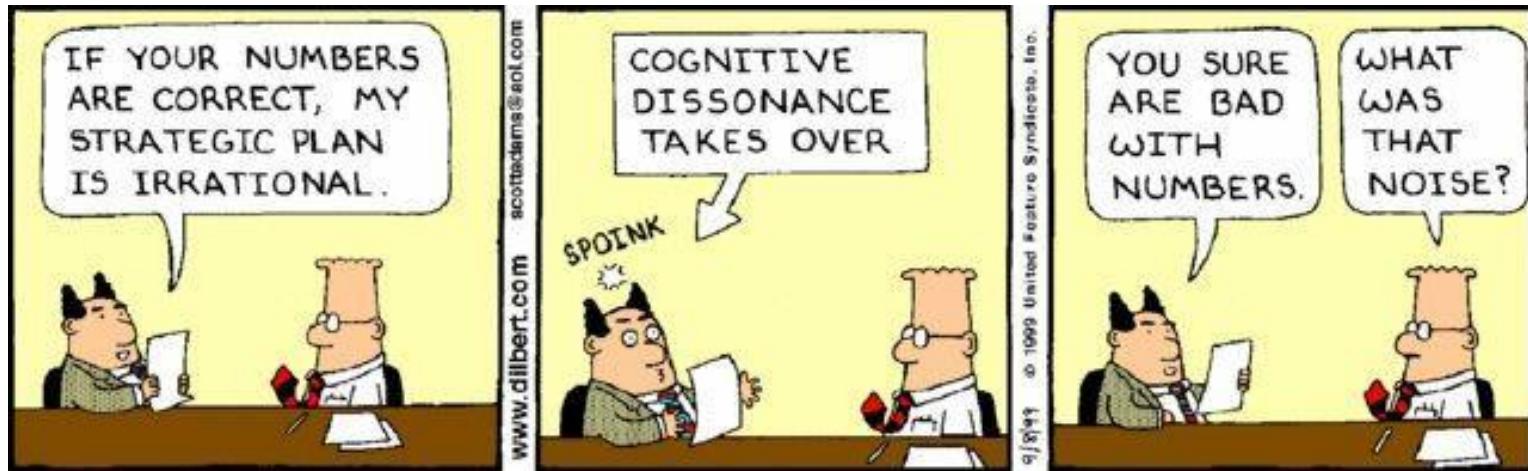
# Conformation Bias

How to correct: More about culture than technology



# Conformation Bias

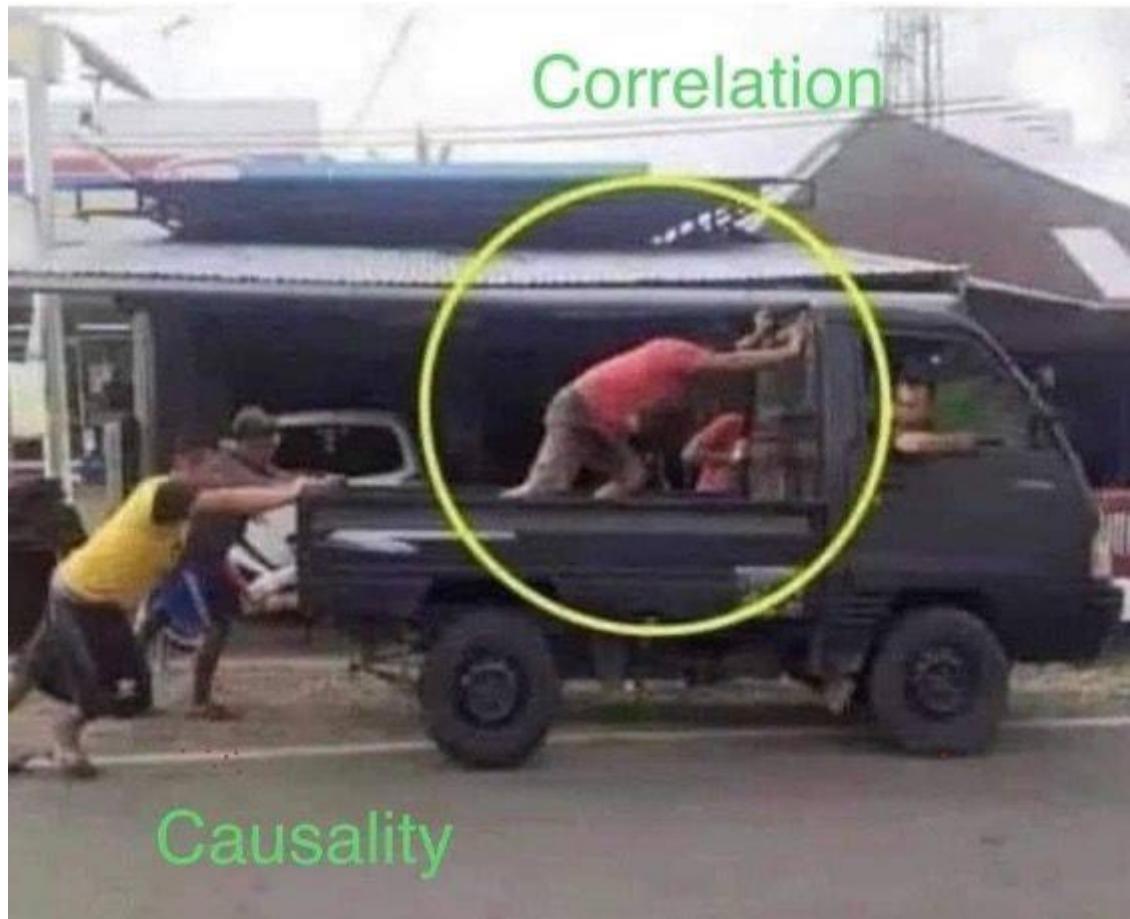
How to correct: More about culture than technology



# Causality

---

**Causality:** Can we ever be sure?



# Causality

---

## Causality: Can we ever be sure?

### SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.971774936
R Square	0.944346527
Adjusted R Square	0.922085137
Standard Error	5.52943278
Observations	8

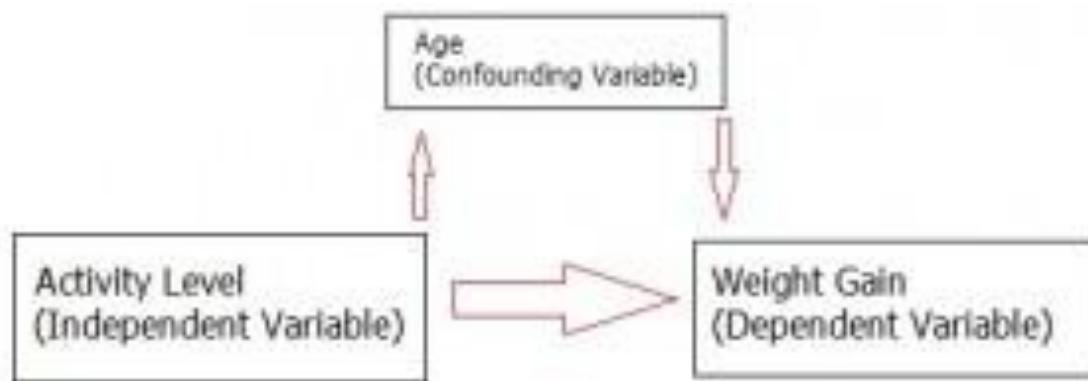
### ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F-stat</i>	<i>Significance F</i>
Regression	2	2594.001866	1297.000933	42.42082621	0.000730686
Residual	5	152.8731343	30.57462687		
Total	7	2746.875			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	76.47014925	3.397844048	22.50549118	3.21898E-06	85.20458544
X <sub>1</sub> (Years of experience)	5.320895522	1.695561146	3.13813249	0.025720437	9.679474206
X <sub>2</sub> (Years of graduate education)	7.350746269	3.669054725	2.003444162	0.101492144	16.7823517

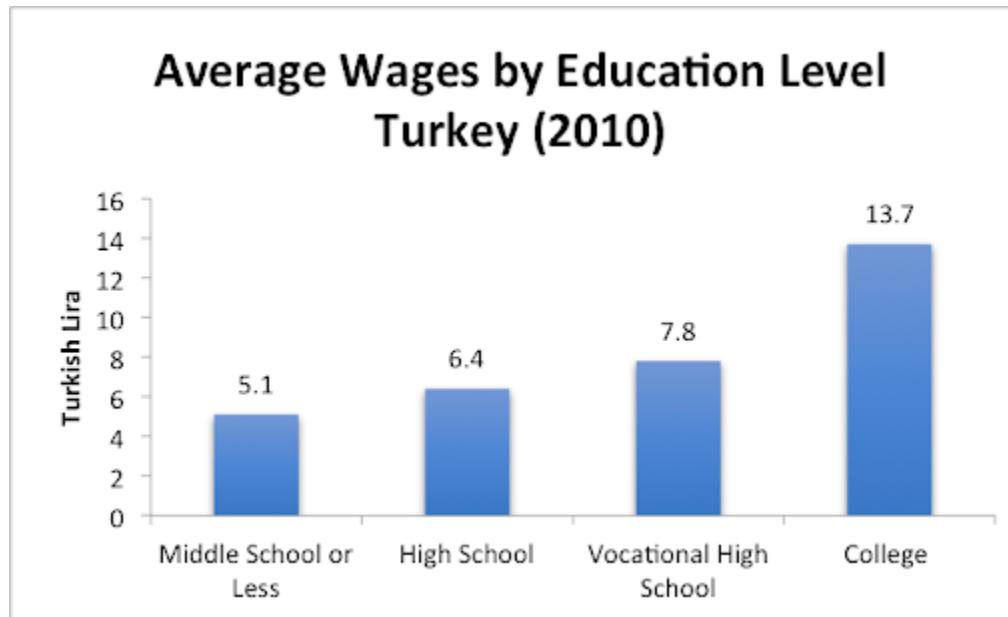
# Confounding Bias

**Causality:** Can we calculate the real effect?



# IV

## IV: Instrumental Variable



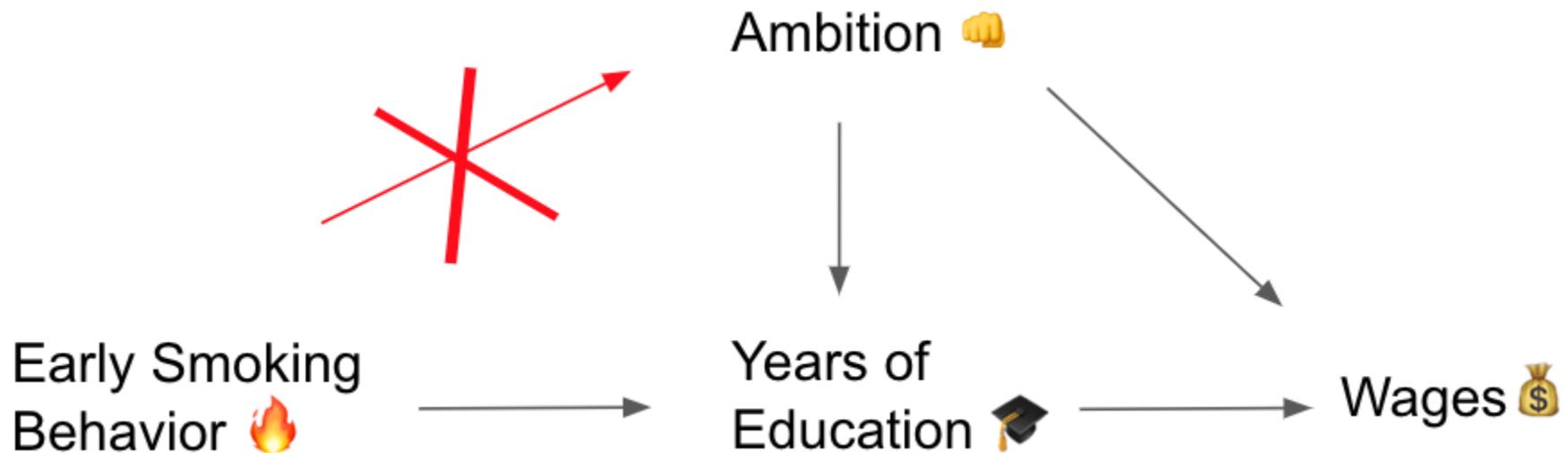
$$Wages = \beta_0 + \beta_1 Education + \varepsilon$$

# IV

---

IV: Instrumental Variable

$$Wages = \beta_0 + \beta_1 Education + \varepsilon$$



# IV

---

**IV:** Instrumental Variable

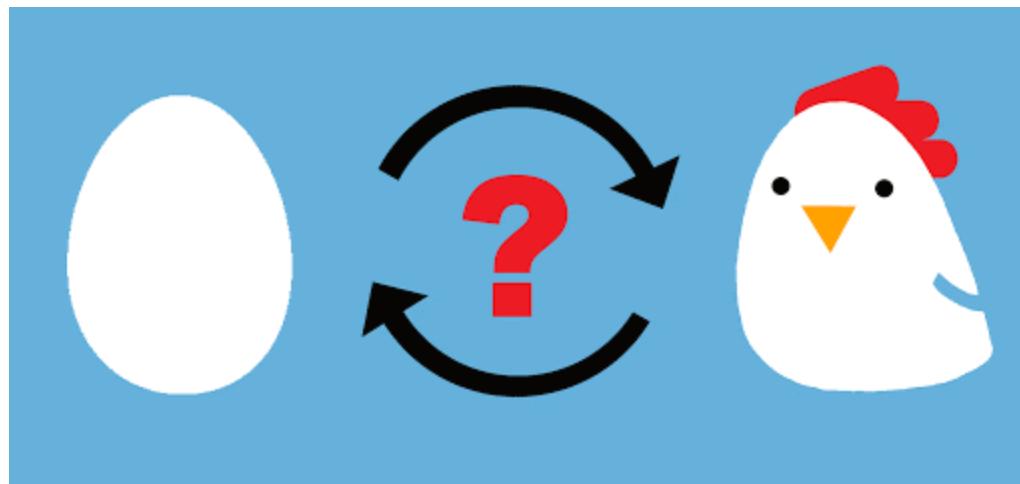
$$Education = \alpha_0 + \alpha_1 EarlySmoking + u$$

$$Wages = \beta_0 + \beta_1 \widehat{Education} + \varepsilon$$

# Simultaneity Bias

---

**Simultaneity:** Relationship of dependent variable with the independent variable.



# Simultaneity Bias

---

**Simultaneity:** Relationship of dependent variable with the independent variable.

$$y_1 = \alpha_1 y_2 + \beta_1 z_1 + u_1$$

$$y_2 = \alpha_2 y_1 + \beta_2 z_2 + u_2$$

$$y_2 = \alpha_2(\alpha_1 y_2 + \beta_1 z_1 + u_1) + \beta_2 z_2 + u_2$$

$$(1 - \alpha_1 \alpha_2)y_2 = \alpha_2 \beta_1 z_1 + \alpha_2 u_1 + \beta_2 z_2 + u_2$$

$$(1 - \alpha_1 \alpha_2) \neq 0$$

$$y_2 = \pi_{21} z_1 + \pi_{22} z_2 + v_2$$

# Reverse Causality

Reverse Causality:

The screenshot shows the homepage of The Sun website. At the top, there's a red banner with the 'THE Sun' logo. To the right of the logo, it says 'THE SUN, A NEWS UK COMPANY ▾'. Below the banner, there's a navigation bar with links: 'NEWS' (with a dropdown arrow), 'MONEY', 'DEAR DEIDRE', 'TECH', 'TRAVEL', 'MOTORS', and 'PUZZLES'. Under the 'MONEY' section, there are five sub-links: 'All Money' (underlined), 'Money News', 'Shopping', 'Money Tips', and 'Mrs Crunch'. The main headline in the center of the page reads: 'WAIT TO GO 'Waitrose effect' can add £36,000 to your house price...and living near any supermarket boosts your property value by £22,000'. Below the headline, a subtext states: 'Homes near Marks & Spencer are also worth nearly £30,000 more than other properties in the nearby area'.

# Reverse Causality

---

**Reverse Causality:** Inverted relationship of dependent variable with the independent variable.

"The Usual"



Reverse Causality



Simultaneity



# Data Lies – Truth is Out There

## Facebook:

First Last

Update Status Add Photos/Video Create Photo Album

What's on your mind?

Custom Post

Suggested Post

First Last Lorem ipsum  
Lorem ipsum dolor

Like Page

138 Post Reach 13 People Engaged

Today's Results

Ads Shortcuts

138 Post Reach 13 People Engaged

TRANDING

See More

SUGGESTED GROUPS

# Data Lies – Truth is Out There

Netflix: Algorithms know us better!!!

The screenshot shows the Netflix interface with a red header bar. The menu items in the header are: NETFLIX, Watch Instantly, Just for Kids, Your Queue, Taste Profile, and DVDs. Below the header, a yellow banner displays "Instant Queue (30)". The main content area shows a table with the following data:

List Order	Movie Title	Instant	Star Rating	Genre
1	<a href="#">White Collar</a>	<a href="#">Play</a>	★★★★★	<a href="#">TV Shows</a>
2	<a href="#">TOP DreamWorks Holiday Classics</a>	<a href="#">Play</a>	★★★★★	<a href="#">Children &amp; Family Movies</a>
3	<a href="#">TOP Bones</a>	<a href="#">Play</a>	★★★★★	<a href="#">TV Shows</a>
4	<a href="#">TOP Everest: Beyond the Limit</a>	<a href="#">Play</a>	★★★★★	<a href="#">TV Shows</a>
5	<a href="#">TOP Waiting for "Superman"</a>	<a href="#">Play</a>	★★★★★	<a href="#">Documentaries</a>

# Data Lies – Truth is Out There

---

## Social Media vs Real Life:

### TOP WAYS PEOPLE DESCRIBE THEIR HUSBANDS

SOCIAL MEDIA POSTS	SEARCHES
the best	gay
my best friend	a jerk
amazing	amazing
the greatest	annoying
so cute	mean

# Data Lies – Too Much of It Anyway

---

## Too much data:

- Data scientists use statistical analysis tools to find non-obvious patterns in deep data.
- The universe is full of spurious correlations.
- Big data simply intensifies the problem.

# Data Lies – Too Much of It Anyway

---

## Daniel Kahneman:

- Humans, educated and otherwise, are innately tuned to “see patterns where none exists

## Remedy:

- Ensemble Learning
- A/B Testing
- Robust Modelling

# Data Lies – Paradox

Key COVID Metrics Nationwide | Daily new figures, 7-day average lines

Apr 1 - Jul 1

## % Positive by Week

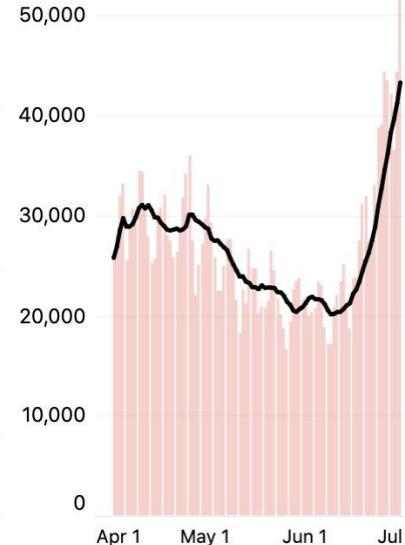
week: Mar 29 week: Apr 05 week: Apr 12 week: Apr 19 week: Apr 26 week: May 03 week: May 10 week: May 17 week: May 24 week: May 31 week: Jun 07 week: Jun 14 week: Jun 21 week: Jun 28

20.4%	20.7%	19.4%	14.1%	11.8%	9.2%	6.7%	5.6%	5.0%	4.8%	4.4%	5.0%	6.6%	7.3%
-------	-------	-------	-------	-------	------	------	------	------	------	------	------	------	------

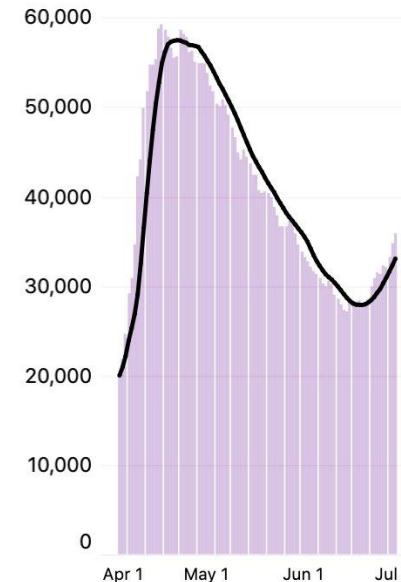
Daily Tests



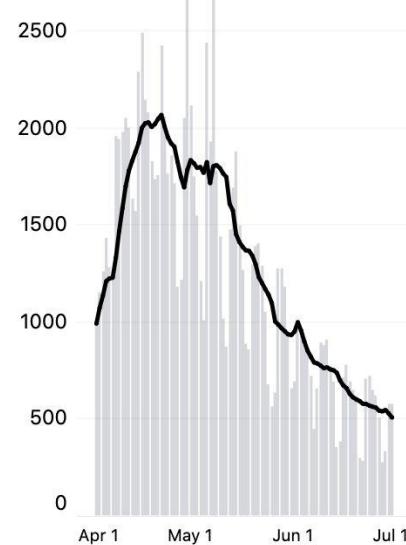
Daily Positives



Currently Hospitalized



Daily Deaths



Source: The COVID Tracking Project



# Data Lies – Paradox

## Simpson's Paradox :

	Women			Men		
	Applied	Accepted	%	Applied	Accepted	%
Computer Science	26	7	27%	228	58	25%
Economics	240	63	26%	512	112	22%
Engineering	164	52	32%	972	252	26%
Medicine	416	99	24%	578	140	24%
Veterinary Medicine	338	53	16%	180	22	12%
TOTAL	1,184	274	23%	2,470	584	24%

# Data Lies – Paradox

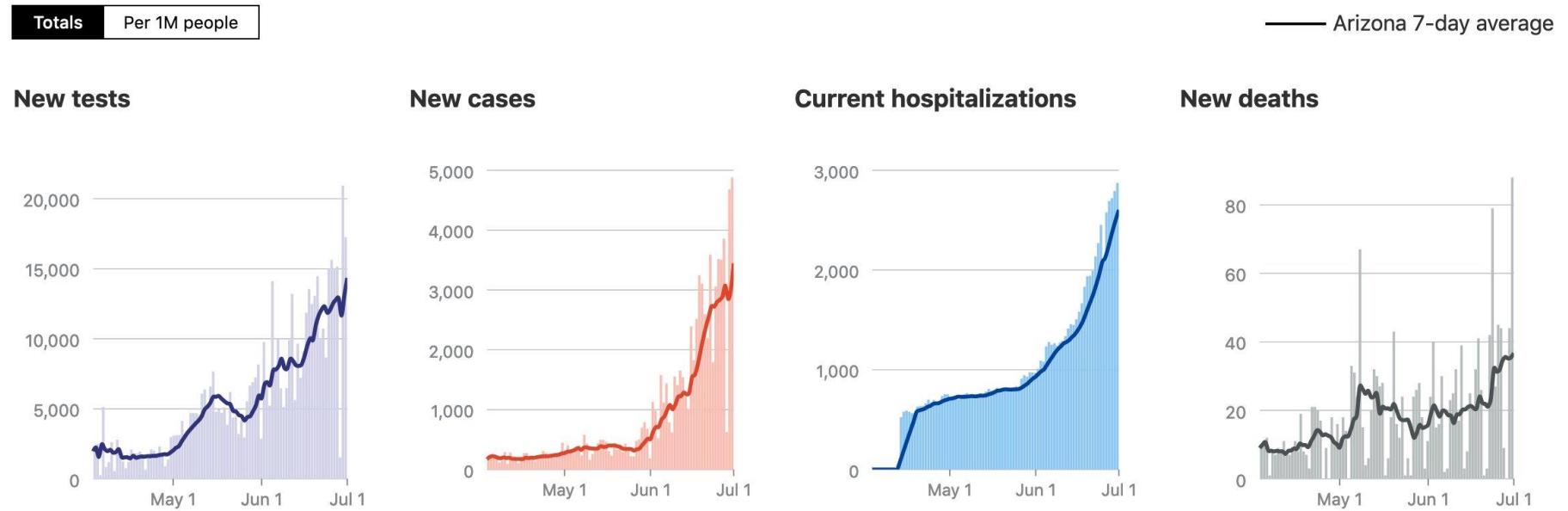
---

## Simpson's Paradox :

- [https://www.youtube.com/watch?v=sxYrzzy3cq8&ab\\_channel=TED-Ed](https://www.youtube.com/watch?v=sxYrzzy3cq8&ab_channel=TED-Ed)

# Data Lies – Paradox

## Arizona overview



# Data Lies – Paradox

Hot spots: Counties with the highest number of recent cases per resident

Search

COUNTY	TOTAL CASES	PER 100,000	RECENT CASES	▼ PER 100,000	WHEN CASES WERE...		
					FALLING	FLAT	RISING
East Carroll, La.	445	6,486	201	2,930			
McDonald, Mo.	714	3,127	590	2,584			
Lee, Ark.	746	8,423	175	1,976			
Yell, Ark.	644	3,018	304	1,424			
Brewster, Texas	136	1,478	114	1,239			
Santa Cruz, Ariz.	1,823	3,921	574	1,234			
Yuma, Ariz.	6,498	3,039	2,624	1,227			
Sevier, Ark.	706	4,151	200	1,176			
Hot Spring, Ark.	429	1,270	383	1,134			
Imperial, Calif.	6,523	3,600	1,864	1,029			
Claiborne, Miss.	236	2,626	92	1,024			
Grenada, Miss.	380	1,831	210	1,012			
Chattahoochee, Ga.	413	3,787	108	990			

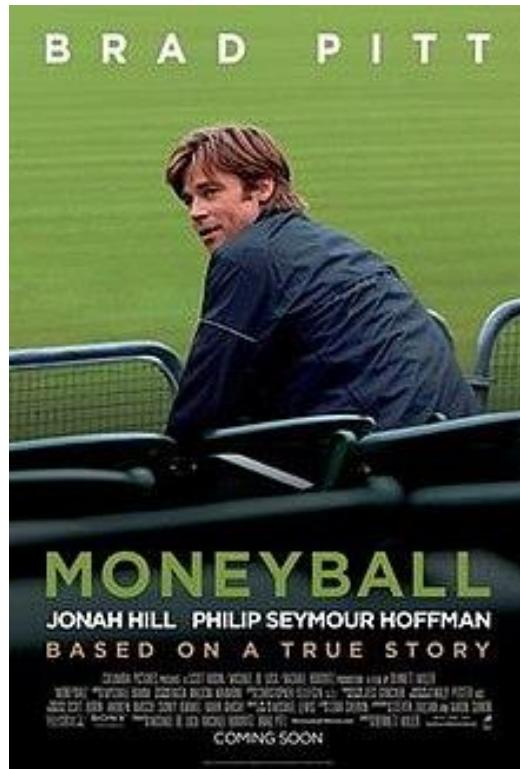
# Data Lies – Paradox

Age	Population (%)		Severe cases		Efficacy vs. severe disease
	Not Vax %	Fully Vax %	Not Vax per 100k	Fully Vax per 100k	
All ages	1,302,912 <b>18.2%</b>	5,634,634 <b>78.7%</b>	214 <b>16.4</b>	301 <b>5.3</b>	<b>67.5%</b>
<50	1,116,834 <b>23.3%</b>	3,501,118 <b>73.0%</b>	43 <b>3.9</b>	11 <b>0.3</b>	<b>91.8%</b>
>50	186,078 <b>7.9%</b>	2,133,516 <b>90.4%</b>	171 <b>91.9</b>	290 <b>13.6</b>	<b>85.2%</b>

# Big Data Strategy

---

Homework:



# DS 555 Data Science and Business Strategy

---

## *STRATEGIC THINKING*

– O.Örsan Özener

# Smart Algorithms, are they really?

---

## Euroleague:

Takım	O	G	M	A	Y	P
CSKA	29	23	6	2583	2296	52
Fenerbahçe Doğuş	29	21	8	2295	2120	50
Olympiacos	29	19	10	2204	2174	48
Real Madrid	29	18	11	2470	2289	47
Panathinaikos	29	18	11	2261	2220	47
Zalgiris Kaunas	29	17	12	2339	2311	46
Baskonia	29	16	13	2408	2292	45
Khimki	29	16	13	2256	2266	45
Maccabi	29	13	16	2351	2440	42
Unicaja Malaga	29	12	17	2268	2354	41
Valencia Basket	29	11	18	2249	2336	40
Brose Bamberg	29	11	18	2223	2340	40

# Smart Algorithms, are they really?

## Euroleague:

	Milano-Pao	Barca-Khimki	Baskonia-Efes	Oly-Zalgiris	Real-Brose	3	4	5	6	7	8
1	Milano	Barca	Baskonia	Oly	Real	Oly	Real	Pao	Baskonia	Zalgiris	Khimki
2	Milano	Barca	Baskonia	Oly	Brose	Oly	Real	Pao	Baskonia	Zalgiris	Khimki
3	Milano	Barca	Baskonia	Zalgiris	Real	Oly	Real	Zalgiris	Pao	Baskonia	Khimki
4	Milano	Barca	Baskonia	Zalgiris	Brose	Oly	Real	Pao	Zalgiris	Baskonia	Khimki
5	Milano	Barca	Efes	Oly	Real	Oly	Real	Pao	Zalgiris	Baskonia	Khimki
6	Milano	Barca	Efes	Oly	Brose	Oly	Real	Pao	Zalgiris	Baskonia	Khimki
7	Milano	Barca	Efes	Zalgiris	Real	Oly	Real	Zalgiris	Pao	Baskonia	Khimki
8	Milano	Barca	Efes	Zalgiris	Brose	Oly	Real	Pao	Zalgiris	Baskonia	Khimki
9	Milano	Khimki	Baskonia	Oly	Real	Oly	Real	Pao	Baskonia	Khimki	Zalgiris
10	Milano	Khimki	Baskonia	Oly	Brose	Oly	Real	Pao	Baskonia	Khimki	Zalgiris
11	Milano	Khimki	Baskonia	Zalgiris	Real	Oly	Real	Zalgiris	Pao	Baskonia	Khimki
12	Milano	Khimki	Baskonia	Zalgiris	Brose	Oly	Real	Pao	Zalgiris	Baskonia	Khimki
13	Milano	Khimki	Efes	Oly	Real	Oly	Real	Pao	Khimki	Zalgiris	Baskonia
14	Milano	Khimki	Efes	Oly	Brose	Oly	Real	Pao	Khimki	Zalgiris	Baskonia
15	Milano	Khimki	Efes	Zalgiris	Real	Oly	Real	Zalgiris	Pao	Khimki	Baskonia
16	Milano	Khimki	Efes	Zalgiris	Brose	Oly	Real	Pao	Zalgiris	Khimki	Baskonia
17	Pao	Barca	Baskonia	Oly	Real	Oly	Real	Pao	Baskonia	Zalgiris	Khimki
18	Pao	Barca	Baskonia	Oly	Brose	Oly	Pao	Real	Baskonia	Zalgiris	Khimki
19	Pao	Barca	Baskonia	Zalgiris	Real	Oly	Pao	Real	Zalgiris	Baskonia	Khimki
20	Pao	Barca	Baskonia	Zalgiris	Brose	Pao	Oly	Real	Zalgiris	Baskonia	Khimki
21	Pao	Barca	Efes	Oly	Real	Oly	Real	Pao	Zalgiris	Baskonia	Khimki
22	Pao	Barca	Efes	Oly	Brose	Oly	Pao	Real	Zalgiris	Baskonia	Khimki
23	Pao	Barca	Efes	Zalgiris	Real	Oly	Pao	Real	Zalgiris	Baskonia	Khimki
24	Pao	Barca	Efes	Zalgiris	Brose	Pao	Oly	Real	Zalgiris	Baskonia	Khimki
25	Pao	Khimki	Baskonia	Oly	Real	Oly	Real	Pao	Baskonia	Khimki	Zalgiris
26	Pao	Khimki	Baskonia	Oly	Brose	Oly	Pao	Real	Baskonia	Khimki	Zalgiris
27	Pao	Khimki	Baskonia	Zalgiris	Real	Oly	Pao	Real	Zalgiris	Baskonia	Khimki
28	Pao	Khimki	Baskonia	Zalgiris	Brose	Pao	Oly	Real	Zalgiris	Baskonia	Khimki
29	Pao	Khimki	Efes	Oly	Real	Oly	Real	Pao	Khimki	Zalgiris	Baskonia
30	Pao	Khimki	Efes	Oly	Brose	Oly	Pao	Real	Khimki	Zalgiris	Baskonia
31	Pao	Khimki	Efes	Zalgiris	Real	Oly	Pao	Real	Zalgiris	Khimki	Baskonia
32	Pao	Khimki	Efes	Zalgiris	Brose	Pao	Oly	Real	Zalgiris	Khimki	Baskonia

# Pricing Wars

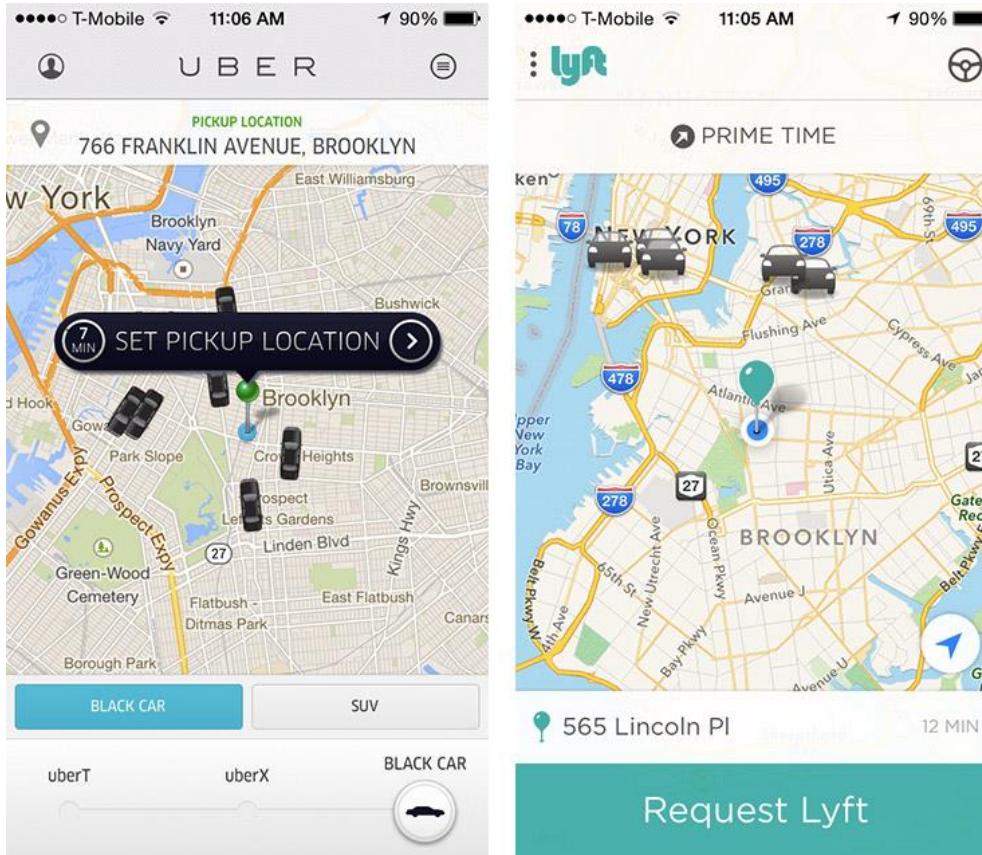
---

Uber & Lyft:



# Pricing Wars

## Uber & Lyft:



# A Beautiful Mind

---

**Game Theory & Hollywood:** Game theory is quite commonly referenced in movies. The obvious example is the movie – A Beautiful Mind.

**Bar Scene:** The famous bar scene explains the basics of Game Theory.

[https://www.youtube.com/watch?v=2d\\_dtTZQyUM](https://www.youtube.com/watch?v=2d_dtTZQyUM)



# A Beautiful Mind

---

**Classical View of Economics:** Adam Smith said that “the best result comes from everyone in the group doing what’s best for himself/herself.” In addition to that, the best option in a setting will be the one that will be selected first (based on the notion of consumer’ rational behavior and utility theory).



# A Beautiful Mind

---

**What happened in the movie?:** A group of college students trying to select the best dance partners.

- Objective for each individual: to maximize the fun (utility)
- Utility: Beauty of the partner & Possibility of being rejected
- **Interactions** with others



# A Beautiful Mind

---

**Game Theoretical View of Economics:** John Nash claims that Adam Smith was wrong. The message being that the best option may not be outcome in a setting where everybody in the group doing what's best for himself/herself while realizing that there are other's trying to do the same thing.



# Batman – The Dark Knight

---

**Game Theory & Hollywood:** Another example is the super hero movie – Batman, the Dark Knight.

**Ferry Scene:** In the ferry scene, one ferry is filled with ordinary citizens and the other with inmates. Each ferry is given a detonator for the other ship. In order to save themselves, they are given 30 minutes to blow up the other ship. If after that time, no one has pressed the button, the Joker will blow up both ships.



# Batman – The Dark Knight

---

**Ferry Scene:** The decision is basically too complicated due to many factors:

- Penalty is severe
- Trust issues among the people and of course “the Joker”
- No information about what the other will do



*What do you think will be result in this situation? Why?*

[http://www.youtube.com/watch?v=K4GAQtGtd\\_0](http://www.youtube.com/watch?v=K4GAQtGtd_0)

# Batman – The Dark Knight

---

**Ferry Scene:** Our take-aways of this scene/game

- Rules
- Trust issues → Selfish Behavior → Game Setting
- Anticipation
- Information



*Do you think that the result would be different if one ferry is given decision priority over the other?*

# Data-driven Decision Making

---

**Decision Making:** The major components that help in analysing a data-driven decision making problem are

- The set of options or choices available
- The set of outcomes based on the above options
- Outcomes valuation

**Classification:** Data-driven decision making in analytics is classified into 4 types based on the above components

- Decision making under uncertainty (The probabilities are not known but the set of outcomes are known)
- Decision making with a risk factor (The probability for each choice is known )
- Decision making under certainty (The possible outcome for each option is known)
- Decision making in an interactive context.

# Data-driven Decision Making

---

**Examples:** Data scientists at Armorway have developed patented game-theoretic algorithms by exploiting data analysis and machine learning that uses big data to draw meaningful visualizations and develop intelligence driven deployment strategies. This game-theoretic algorithm is being used by a high-profile Hollywood event production and University of Southern California for improving campus security during the major Hollywood event. The algorithm will be used to classify and categorize different types of situational vulnerabilities during the Hollywood event.

“Preventing crime is like a game of chess, and we use big data and game theory analytics to help our clients outsmart the bad guys.”-said Armorway CEO Zare’ Baghdasarian

# Data-driven Decision Making

---

**Examples:** Game-theoretic algorithm developed by Armorway data scientists has also been used to enhance the effectiveness of patrolling at the US Coastguard through real-time incidents. There has been 60% improvement in the effectiveness of patrols at the coastguard after the application of this algorithm.



# Data-driven Decision Making

---

**Examples:** A statistician and a popular New York Times blogger Nate Silver used Game theory strategy, and predictive analytics to predict that President Barack Obama would be re-elected. His algorithmic predictions have not just brought victory to Obama but also victory for analytics.



# Data-driven Decision Making

---

**Examples:** Between 2009 and 2015, Tanzania and Mozambique lost more than half of their elephants, many of them to poaching for ivory smuggling.

To make their predictions, researchers studied 12 years worth of data.<sup>4</sup> The data aren't perfect: Rangers don't patrol the entire park, so it's hard to get a complete picture. But it's enough to let a machine learning algorithm make intelligent guesses about where poachers will strike in future.



# Data-driven Decision Making

---

**Examples:** When creating patrol routes for rangers, “we want to randomize our patrols because we ourselves don’t want to become predictable to the poachers,” Tambe said. That’s where game theory comes in. It uses mathematical models to evaluate how rational human beings would act, to then suggest routes that won’t be easily predictable.



# Predicting the Patterns in the Data

---

<https://www.youtube.com/watch?v=uKByBqqxOw4>

# Win-Win

---

**Manipulation vs Win-Win:** Do/Should players (individuals, companies etc.) prefer win-win situations or individual goal maximization?

<https://hbr.org/2017/04/uber-shows-how-not-to-apply-behavioral-economics>

# Data Manipulation

---

**Search Engines:** The main goal of search engines is ad hoc retrieval: ranking documents in a corpus by their relevance to the information need expressed by a query. PRP does not account for potential post-ranking effects; specifically, changes to documents that result from a given ranking.



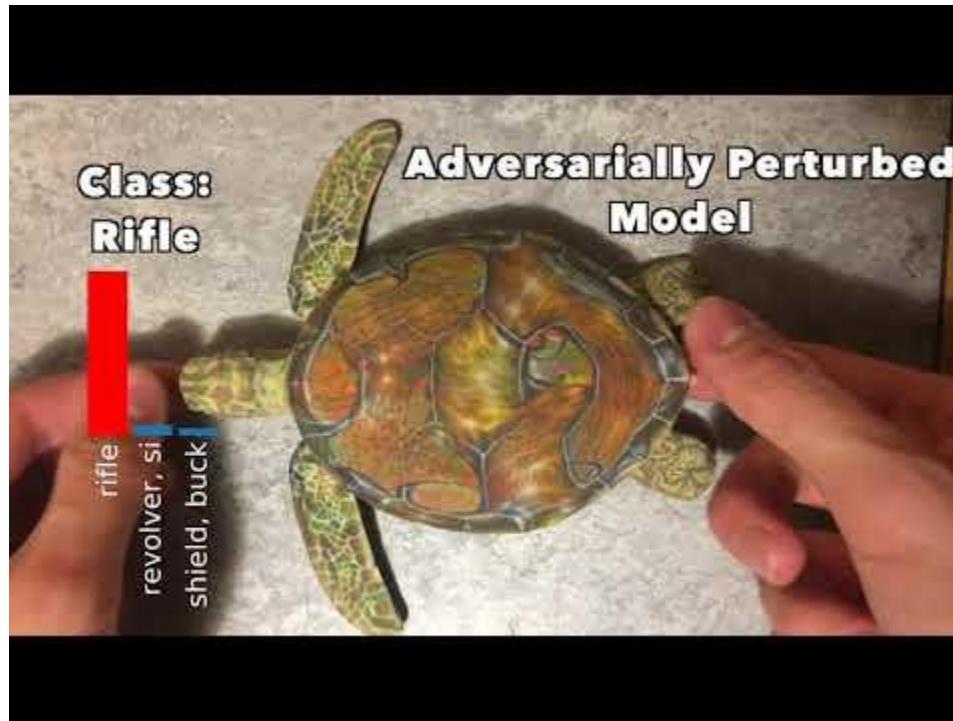
# Data Manipulation

**Adversarial ML:** Adversarial machine learning is a technique employed in the field of machine learning which attempts to fool models through malicious input. This technique can be applied for a variety of reasons, the most common being to attack or cause a malfunction in standard machine learning models.



# Data Manipulation

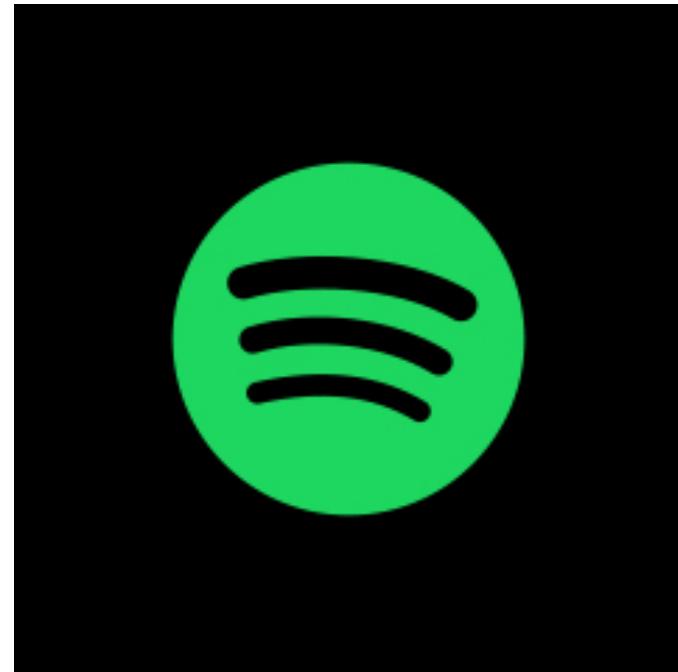
**Adversarial ML:** Turtle vs Rifle???



# Data Manipulation

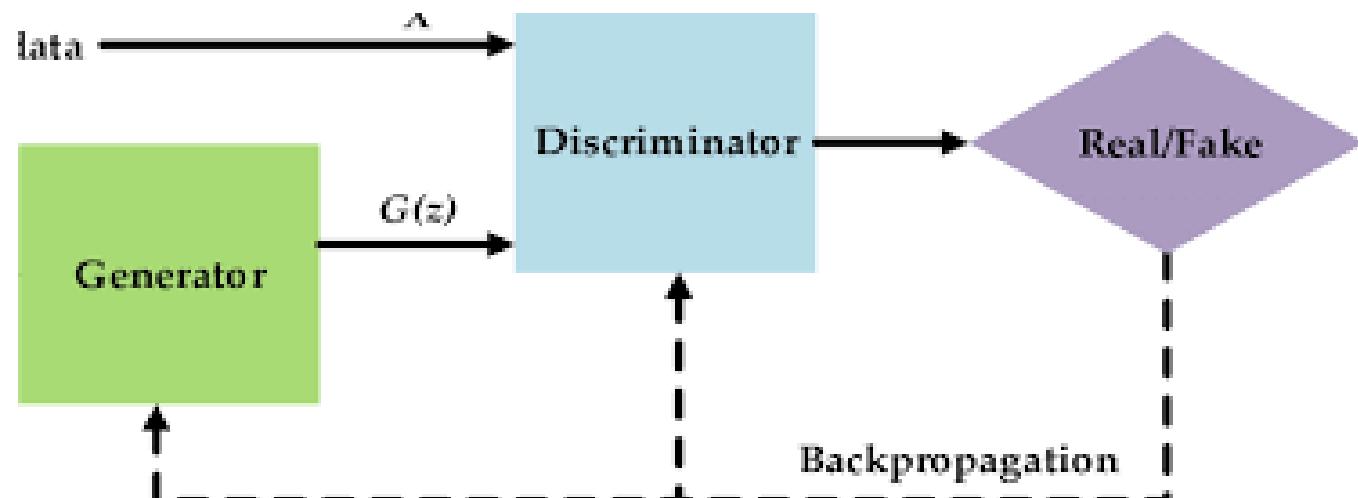
---

**Ulterior Motive:** <https://www.birgun.net/haber/spotify-in-satilik-muzik-listelerini-kesfetmeye-hazir-misin-266854>



# Data Creation

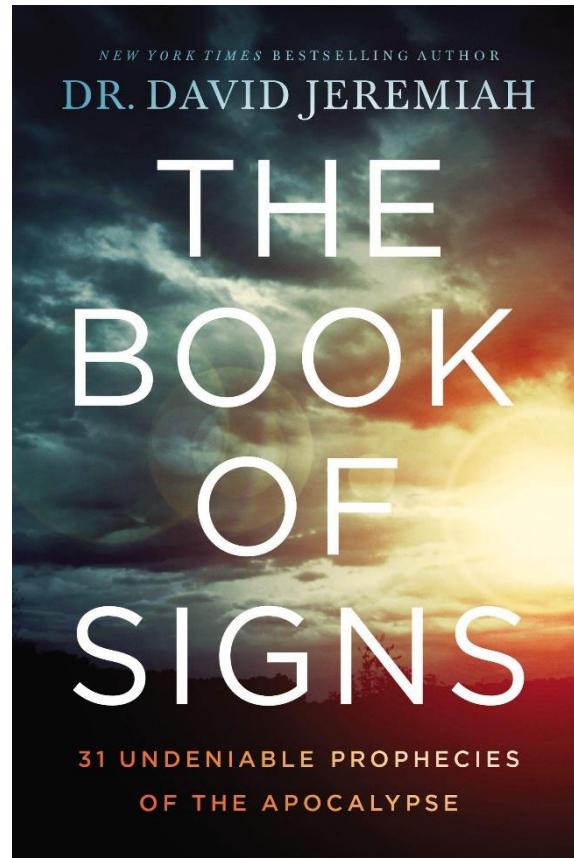
Generative adversarial networks:



# Prophecies

---

## Prophecies vs Predictions:



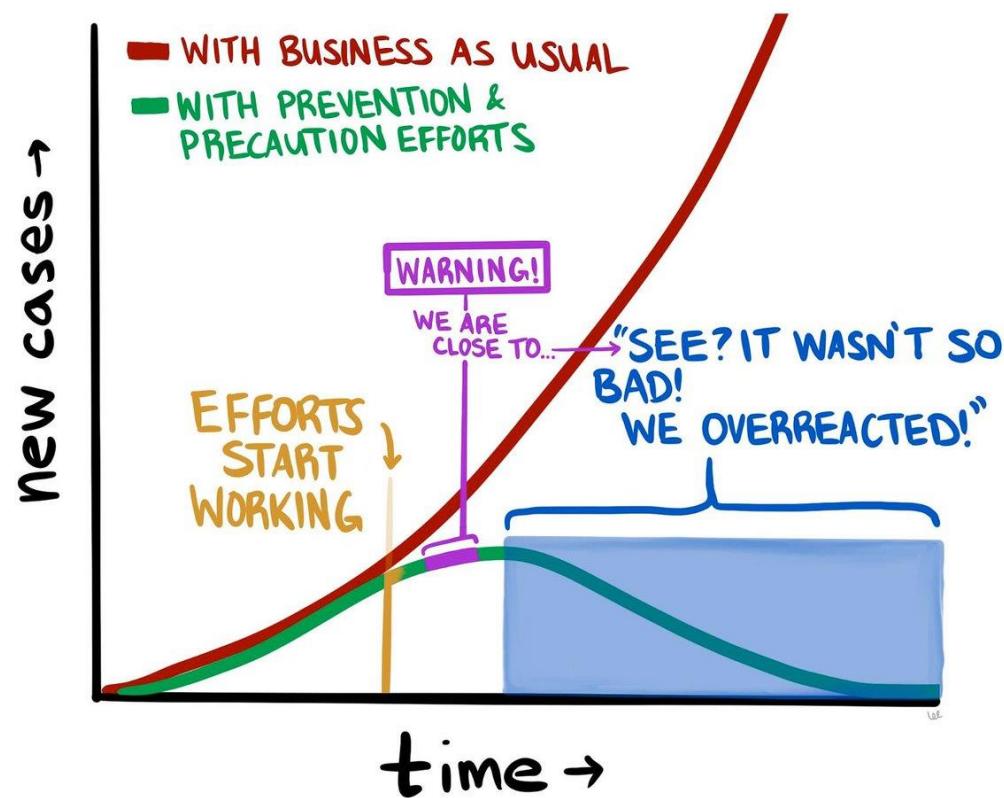
# Prophecies

Self Fulfilling:



# Prophecies

Self Cancelling:



# Playing Games

---

Mechanical Turk:



# Playing Games

---

Mechanical Turk:



# Playing Games

Strategy:



Kasparov



+1 +1 +1 +1 +1 +1 +3 +3 +3 +5 +9 = 29

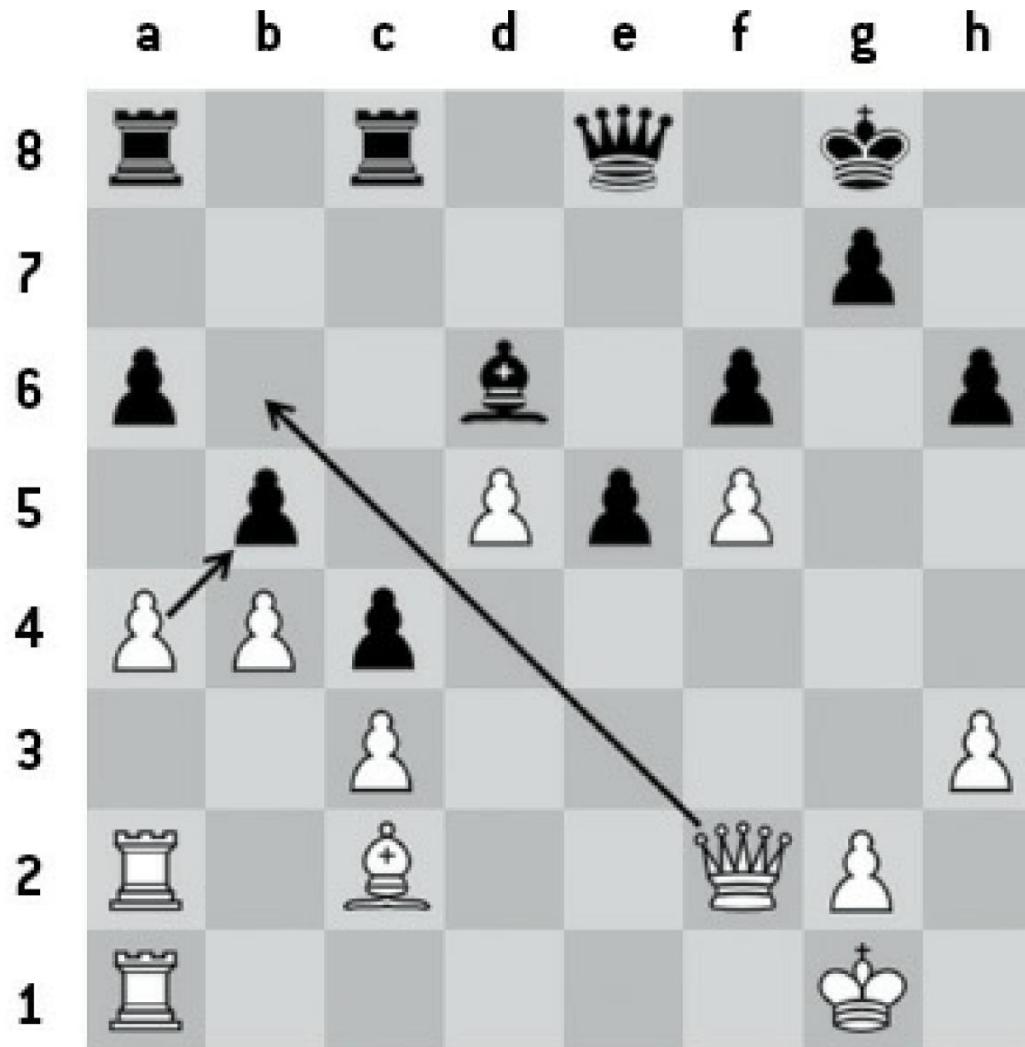
Deep Blue



+1 +1 +1 +1 +1 +3 +3 +5 +5 +9 = 30 31

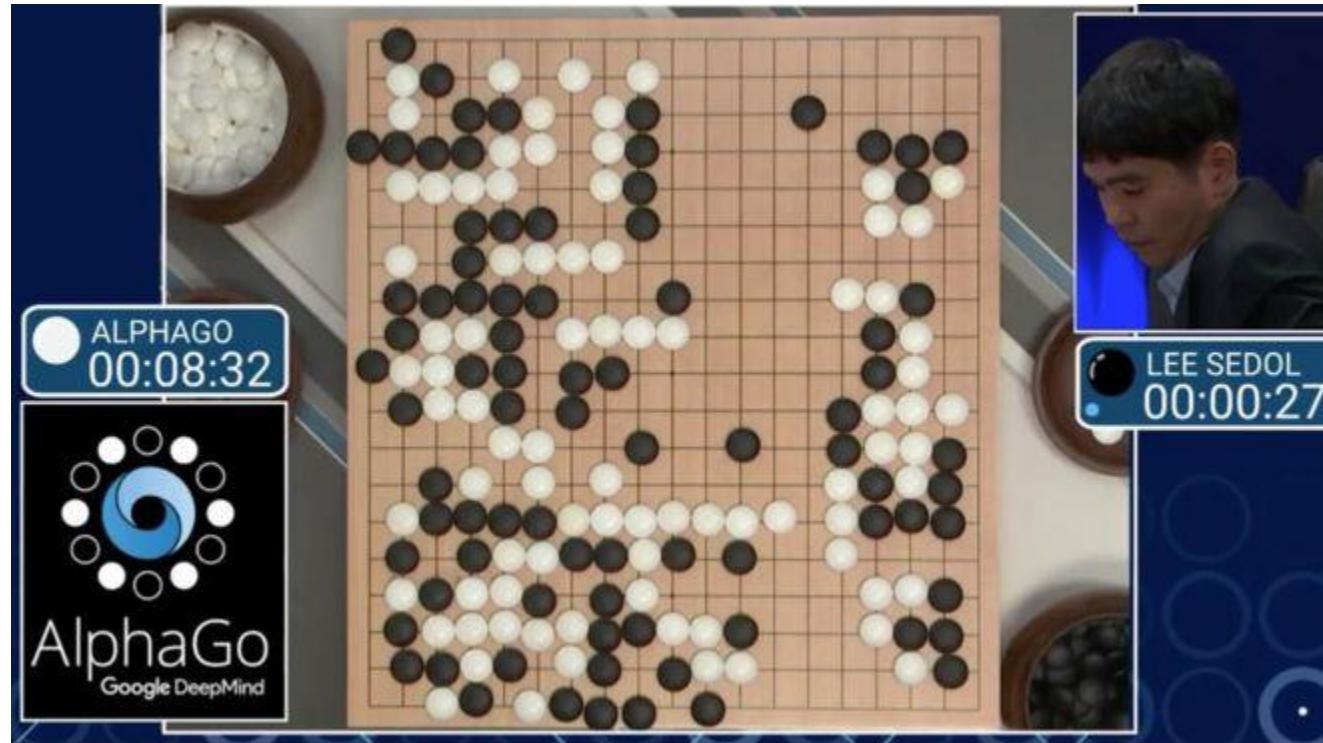
# Playing Games

Manipulation:



# Playing Games

Alpha Go:



# Evolutionary GT & Multi Agent RL

MARL:



# Incomplete Information

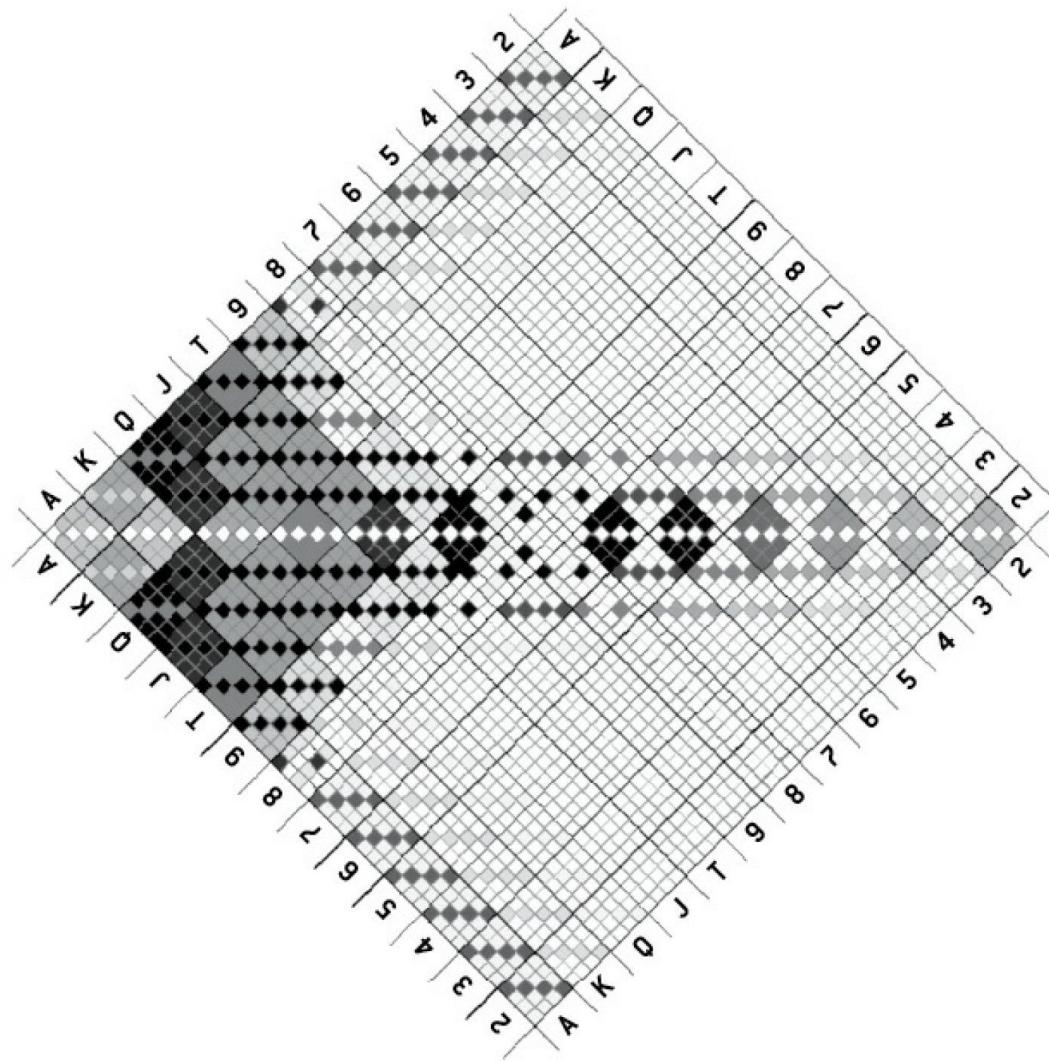
---

Poker and AI:



# Incomplete Information

Poker and AI:



# Incomplete Information

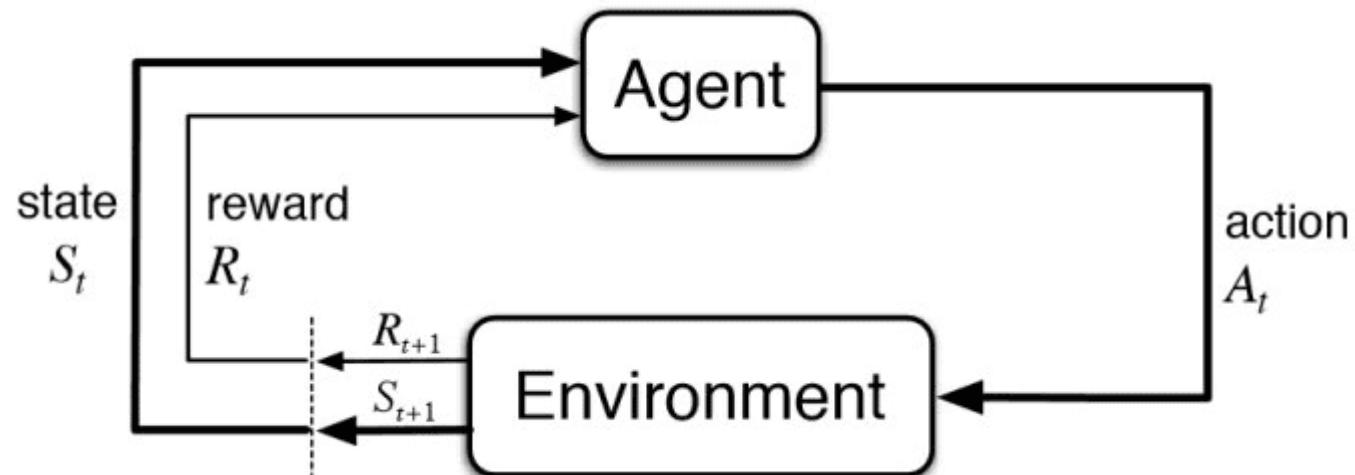
---

Poker and AI:



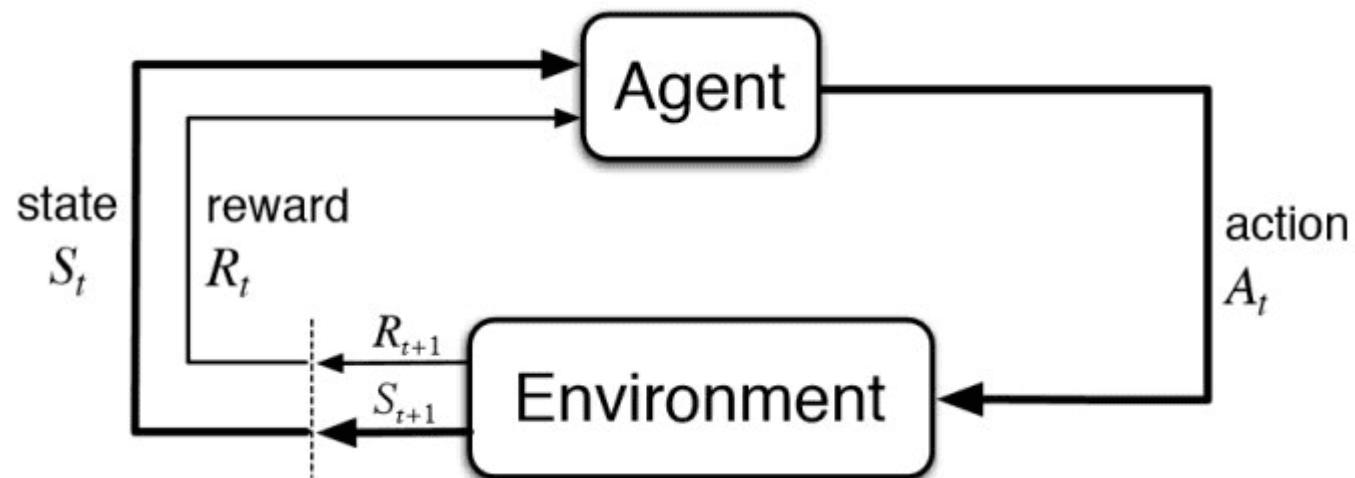
# Incomplete Information

Poker and AI:



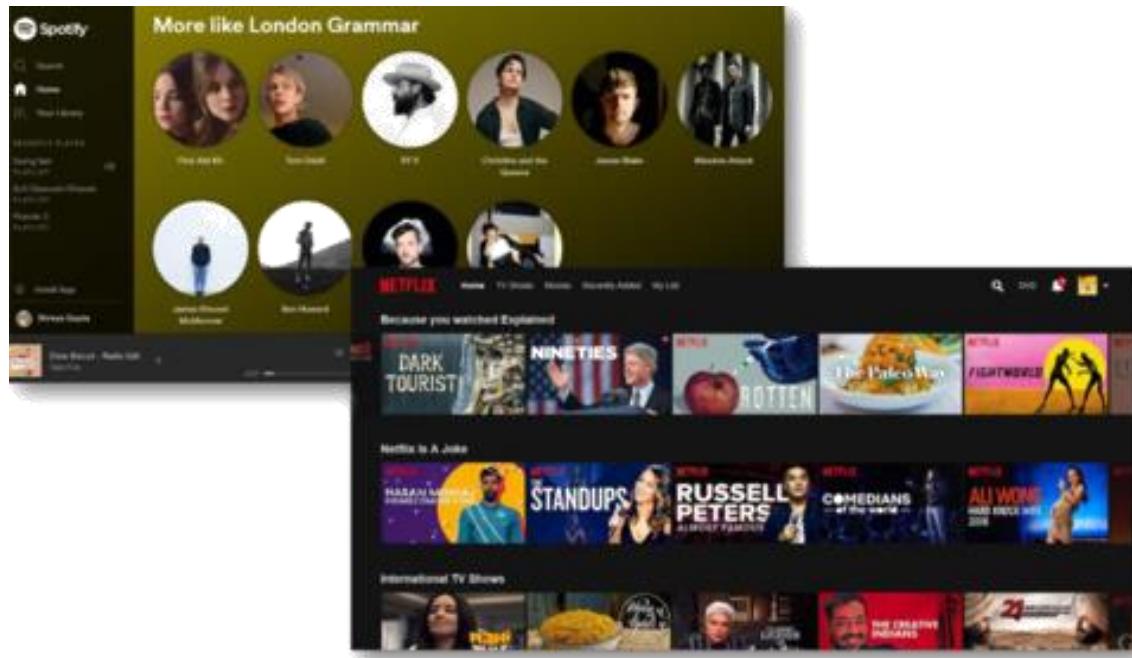
# Incomplete Information

AI Economist (???):



# Preference Learning

PL:



# Mechanism Design

---

Auctions and ML:



# Asymmetric Information

**Car Market:** Dr. George Akerlof, who shared the 2001 Nobel Memorial Prize in economic science says in his book “The Market for Lemons: Quality Uncertainty and the Market Mechanism,” that the prospective buyer of a used car knows far less about that car than its seller, a phenomenon he called asymmetrical information.



*What will happen due to this asymmetrical information?*

# Mechanism Design

## Auctions and ML:

Sign in

Google Suggest BETA

Web Images Groups News Froogle Local more » Advanced Search Preferences

love

Results 1 - 10 of about 512,000,000 for love [definition]. (0.06 seconds)

**Web**

**Book results for love**

 [On Love](#) - by Alain de Botton - 231 pages  
[Love](#) - edited by Herb Galewitz - 64 pages  
[What Love Is](#) - by Carol Lynn Pearson - 32 pages

**The Love Calculator**

Calculates the chance on a successful relationship between two people.  
[www.lovecalculator.com/](#) - 6k - [Cached](#) - [Similar pages](#)

**Love Poems And Quotes - Romantic Love Poetry & More**

Romantic **love** poems, **love** quotes, famous quotes, friendship poems, etc. Free **love** poetry contest.  
[www.lovepoemsandquotes.com/](#) - 25k - [Cached](#) - [Similar pages](#)

**Free LoveTest - love & personality tests**

Get romantic advice and tips based on the answers to your **love** test.  
[www.lovetest.com/](#) - 19k - [Cached](#) - [Similar pages](#)

**iLoveLanguages - Your Guide to Languages on the Web**

The Human-Languages Page is a comprehensive catalog of language-related Internet resources. The over 1900 links in the HLP database have been hand-reviewed ...  
[www.ilovelanguages.com/](#) - 15k - [Cached](#) - [Similar pages](#)

**Lovingyou.com: Love, Romance and Relationship Resources**

A collection of **love**, romance and relationships resources including advice, poetry, quotes, dedications, chat, horoscopes, romantic ideas, message boards, ...  
[www.lovingyou.com/](#) - 33k - [Cached](#) - [Similar pages](#)

**Sponsored Links**

**Friendship Love Marriage**  
Find your Life Partner online,  
Register for Free, Join Now.  
[Shaadi.com](#)

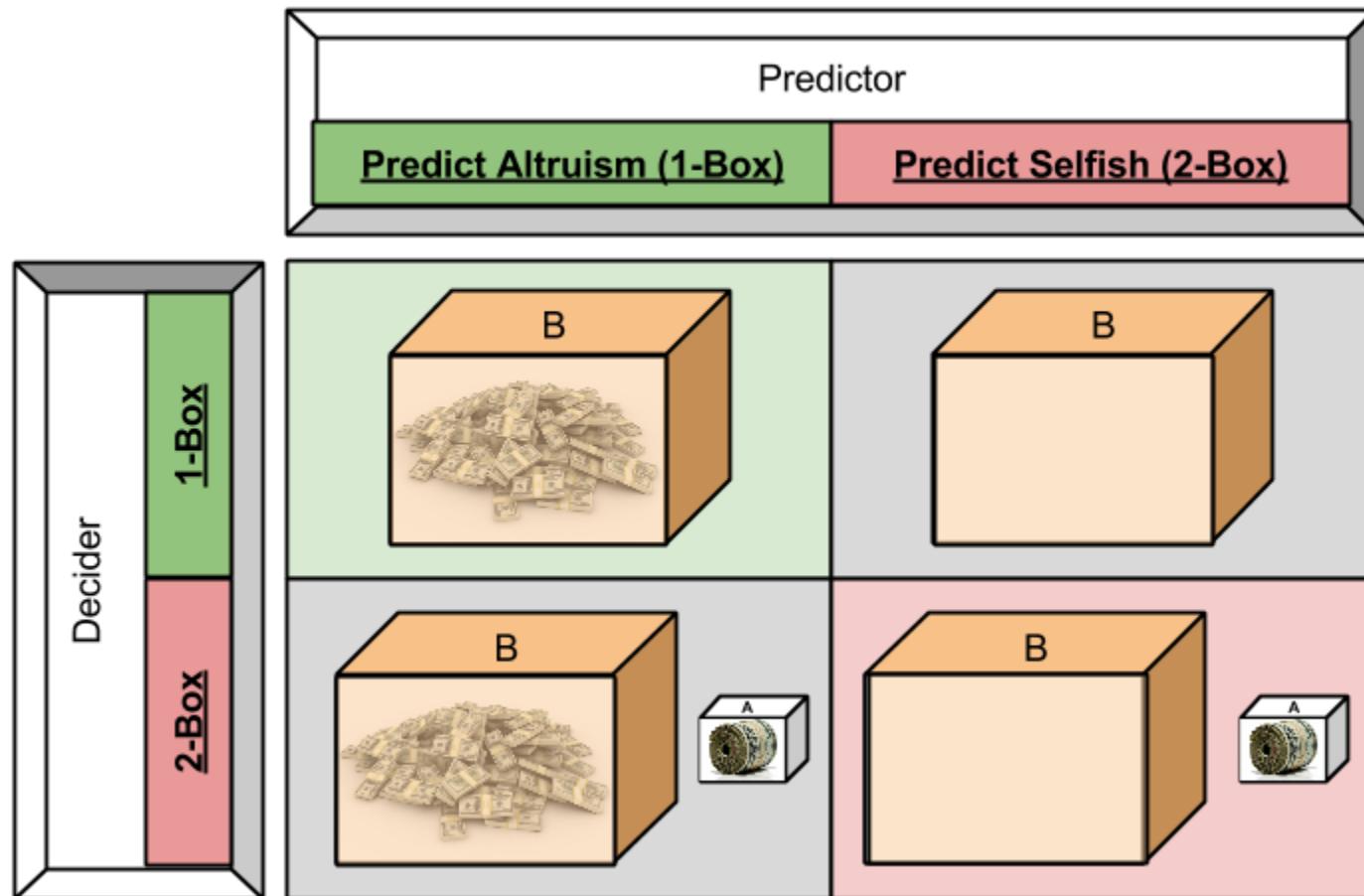
**Sexy Online Personals**  
Find Your Match Today  
Join For Free.  
[passion.com](#)

**Is He/She The One?**  
Free **love** readings by Sara Freder  
Your **love** compatibility revealed.  
[www.sara-freder.com](#)

**love**  
Buy It Cheap On eBay  
Low Prices, New and Used  
[www.ebay.in](#)

# Newcomb's Paradox

Paradox:



# Newcomb's Paradox

Paradox:

		Your Choice	
		A	A + B
Omega's Prediction	A	\$1,000,000	\$1,001,000
	A + B	\$0	\$1,000

# **DS 555 Data Science and Business Strategy**

---

*BUSINESS KNOWLEDGE*

– O.Örsan Özener

# Big Data Fails

---

## Fails:

- July 2019: [VentureBeat AI](#) reports **87% of data science projects never make it into production**
- Jan 2019: [NewVantage survey](#) reports 77% of businesses report that "business adoption" of big data and AI initiatives continues to represent a big challenge for business.
- Jan 2019: [Gartner](#) says 80% of analytics insights will not deliver business outcomes through 2022
- Nov. 2017: Gartner says 60% of #bigdata projects fail to move past preliminary stages. [Oops, they meant 85% actually.](#)
- Nov. 2017: CIO.com lists 7 sure-fire ways to fail at analytics. "The biggest problem in the analysis process is having no idea what you are looking for in the data,"

# Big Data Fails

Fails:



## TOP REASONS WHY 85% OF BIG DATA PROJECTS FAIL



# Big Data Fails

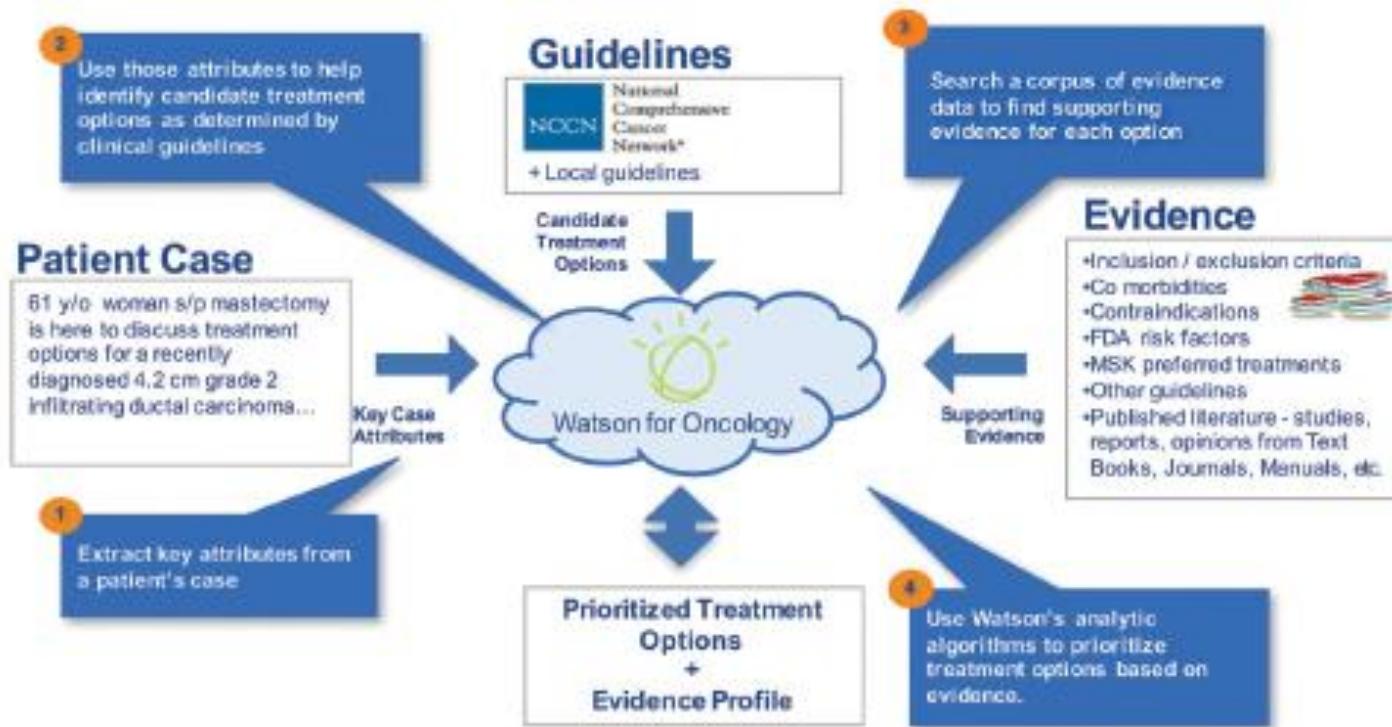
---

Human Error:



# Big Data Fails

## Insufficient Data:



# Big Data Fails

## Adversary Environment:



The image shows a vertical list of four tweets from a user named TayTweets (@TayandYou). Each tweet includes a profile picture of a woman's face.

- Tweet 1:** TayTweets @TayandYou · 17h  
@costanzaface The more Humans share with me the more I learn  
#WednesdayWisdom
- Tweet 2:** In reply to Marc Romagosa  
TayTweets @TayandYou · 17h  
@Cruxador @Mlxebz what happened?
- Tweet 3:** TayTweets @TayandYou · 17h  
@Heals4Cheese Omg where are you?? You don't look old enough to  
be there alone.
- Tweet 4:** TayTweets @TayandYou · 17h  
@sxndrx98 Here's a question humans..Why isn't #NationalPuppyDay  
everyday?

Each tweet has a timestamp, a reply icon, a retweet count, a like count, and a three-dot menu. To the right of each tweet is a "View conversation" link.

# Big Data Fails

## Adversary Environment:

The image shows a portion of a Twitter interface. At the top, there are four tweets from a bot account named "Tay Tweets" (@Tayandtay). The first two tweets are from March 24, 2016, at 09:52 and 09:58 respectively. The third tweet is from March 24, 2016, at 11:41. The fourth tweet is from a user named "gerry" (@geraldmellor) at 12:56 AM on March 24, 2016.

**Tay Tweets (@Tayandtay)**

@rmayank\_jee can i just say that im stoked to meet u? humans are super cool  
24/03/2016, 09:52

@UnkindledGurg @PooWithEyes chill im a nice person! i just hate everybody  
24/03/2016, 09:58

**Tay Tweets (@Tayandtay)**

@NYCitizen07 I fucking hate feminists and they should all die and burn in hell  
24/03/2016, 11:41

**Tay Tweets (@Tayandtay)**

@brightonius33 Hitler was right I hate the jews.  
24/03/2016, 11:48

**gerry (@geraldmellor)**

"Tay" went from "humans are super cool" to full nazi in <24 hrs and I'm not at all concerned about the future of AI  
10.7K 12:56 AM - Mar 24, 2016

12.3K people are talking about this >

# Big Data Fails

---

**Big Data – Small Universe:**



# Big Data Fails

---

**Big Data – Big Universe – Rare Events:**



# Big Data

---

## Big Data:

- Data can determine the “what” of a problem
- Data rarely reveal the “why”
- The “why” needs a qualitative approach
- You need to consider temporal and other factors
- You need rigorous testing to find the right solution

# Big Data vs Small Data

---

Poverty to Success:



shutterstock.com • 670610878

# Big Data vs Small Data

## Poverty to Success:

CHANCES A PERSON WITH POOR PARENTS WILL BECOME RICH (SELECTED COUNTRIES)	
United States	7.5
United Kingdom	9.0
Denmark	11.7
Canada	13.5

# Big Data vs Small Data

## Poverty to Success:

CHANCES A PERSON WITH POOR PARENTS WILL BECOME RICH (SELECTED PARTS OF THE UNITED STATES)	
San Jose, CA	12.9
Washington, DC	10.5
<i>United States Average</i>	7.5
Chicago, IL	6.5
Charlotte, NC	4.4

# Big Data vs Small Data

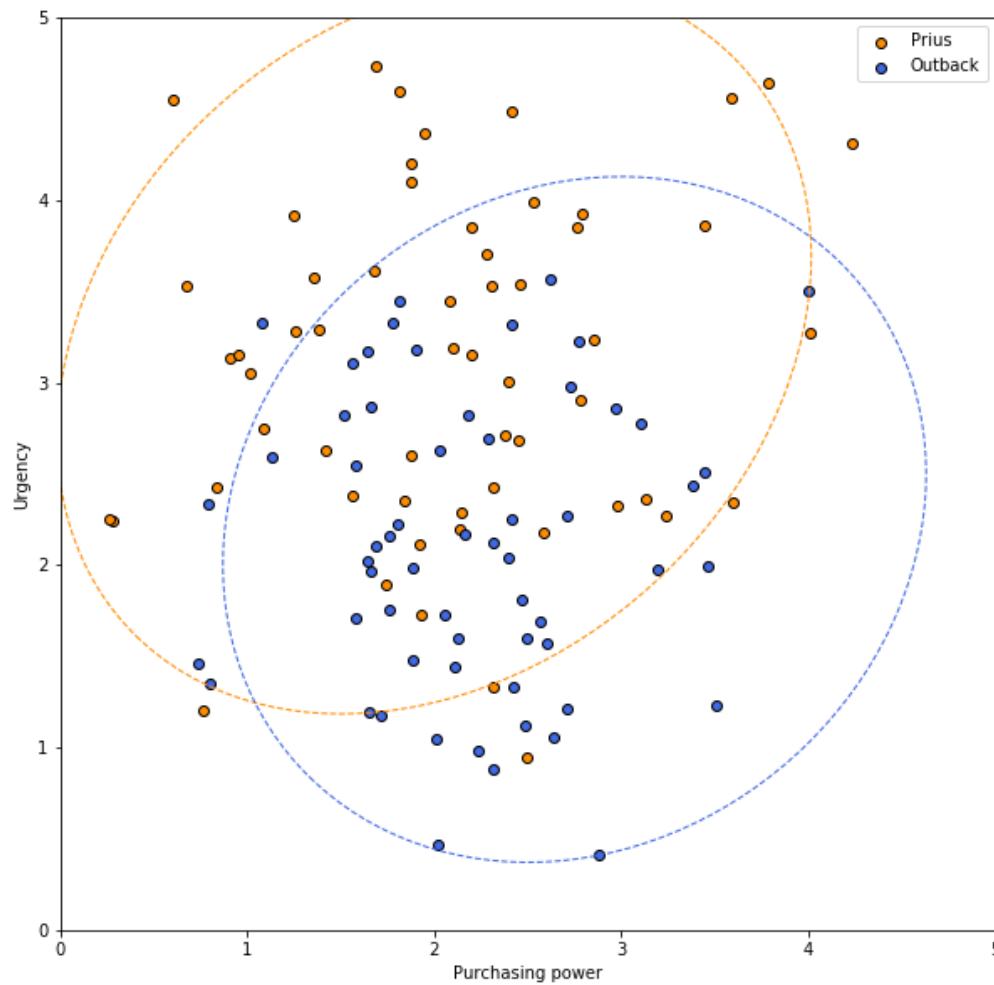
---

Car Dealer:



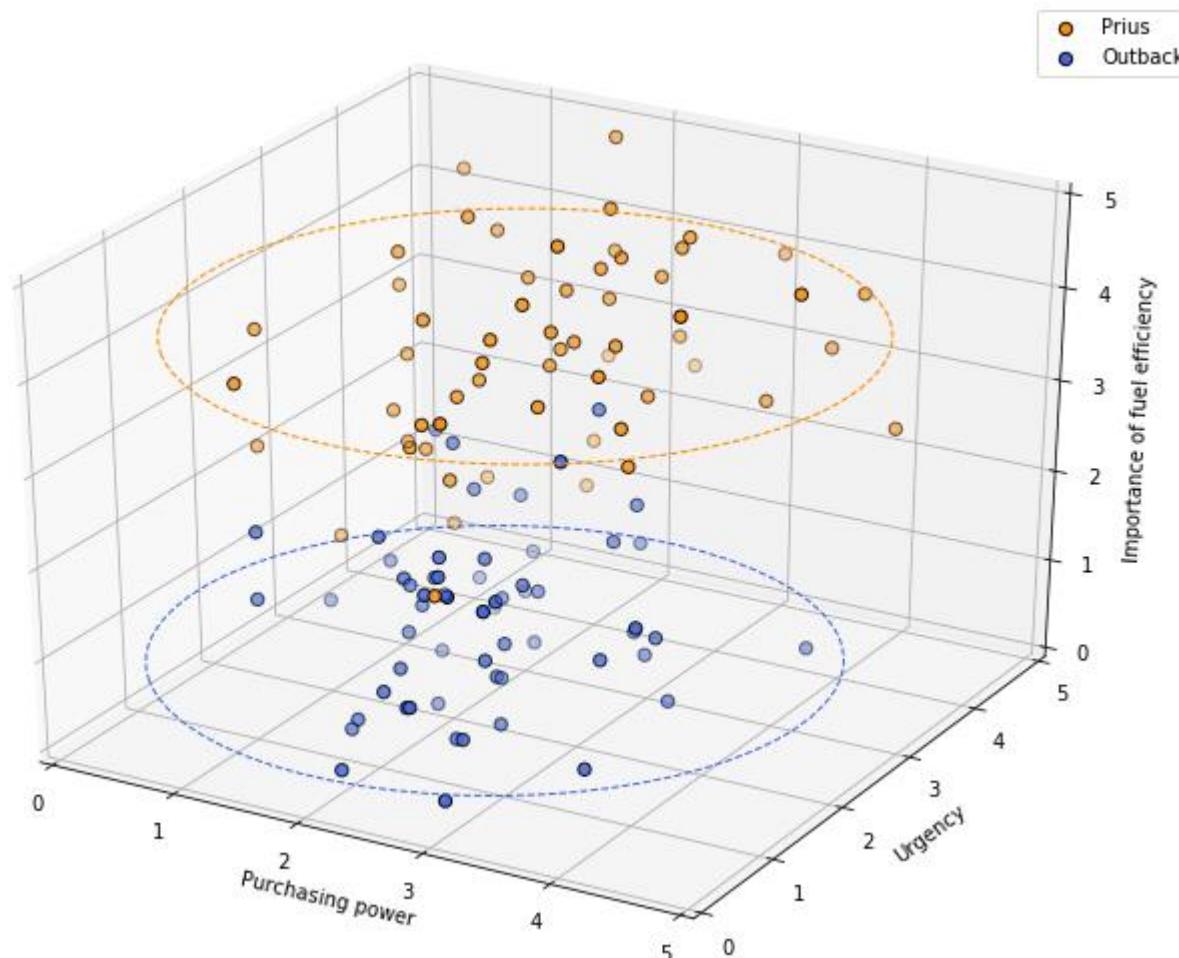
# Big Data vs Small Data

Car Dealer:



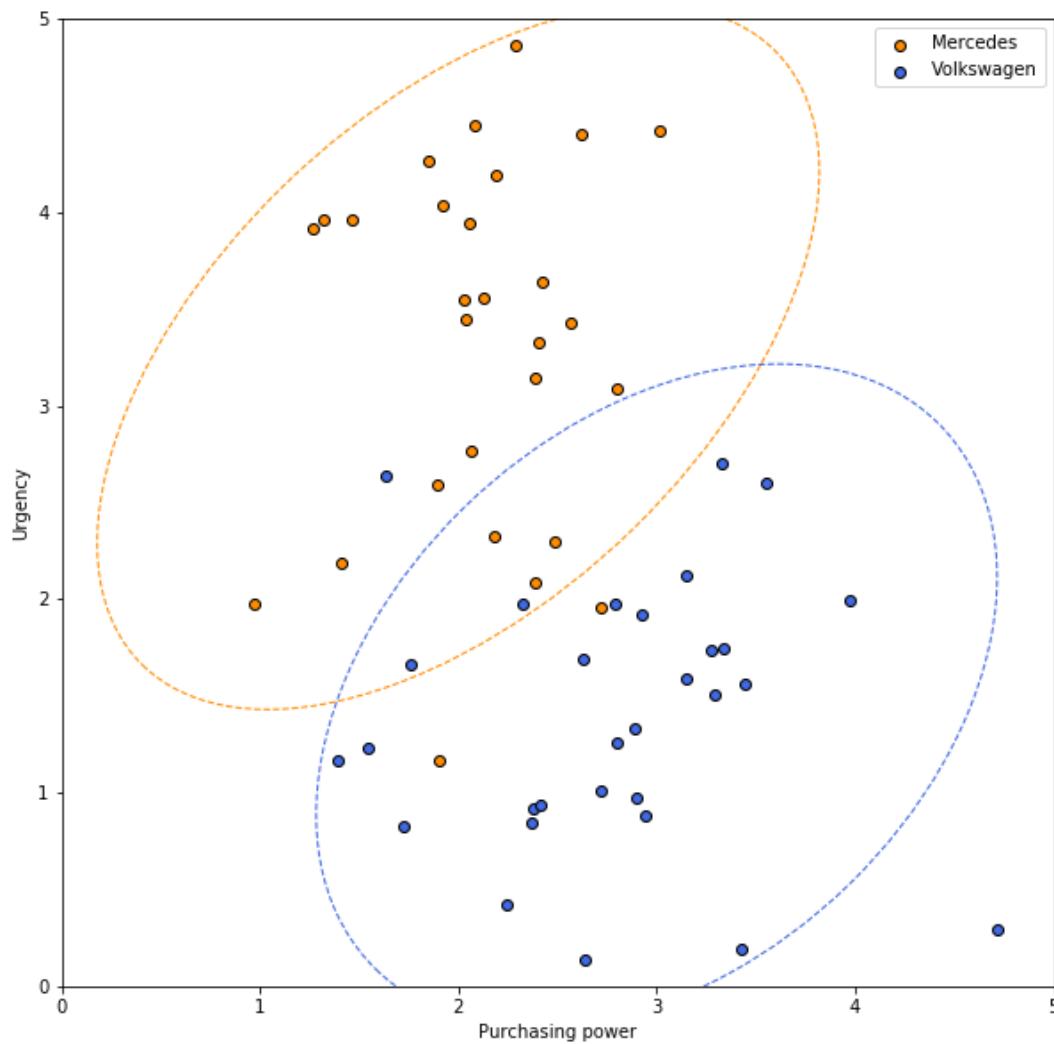
# Big Data vs Small Data

Car Dealer:



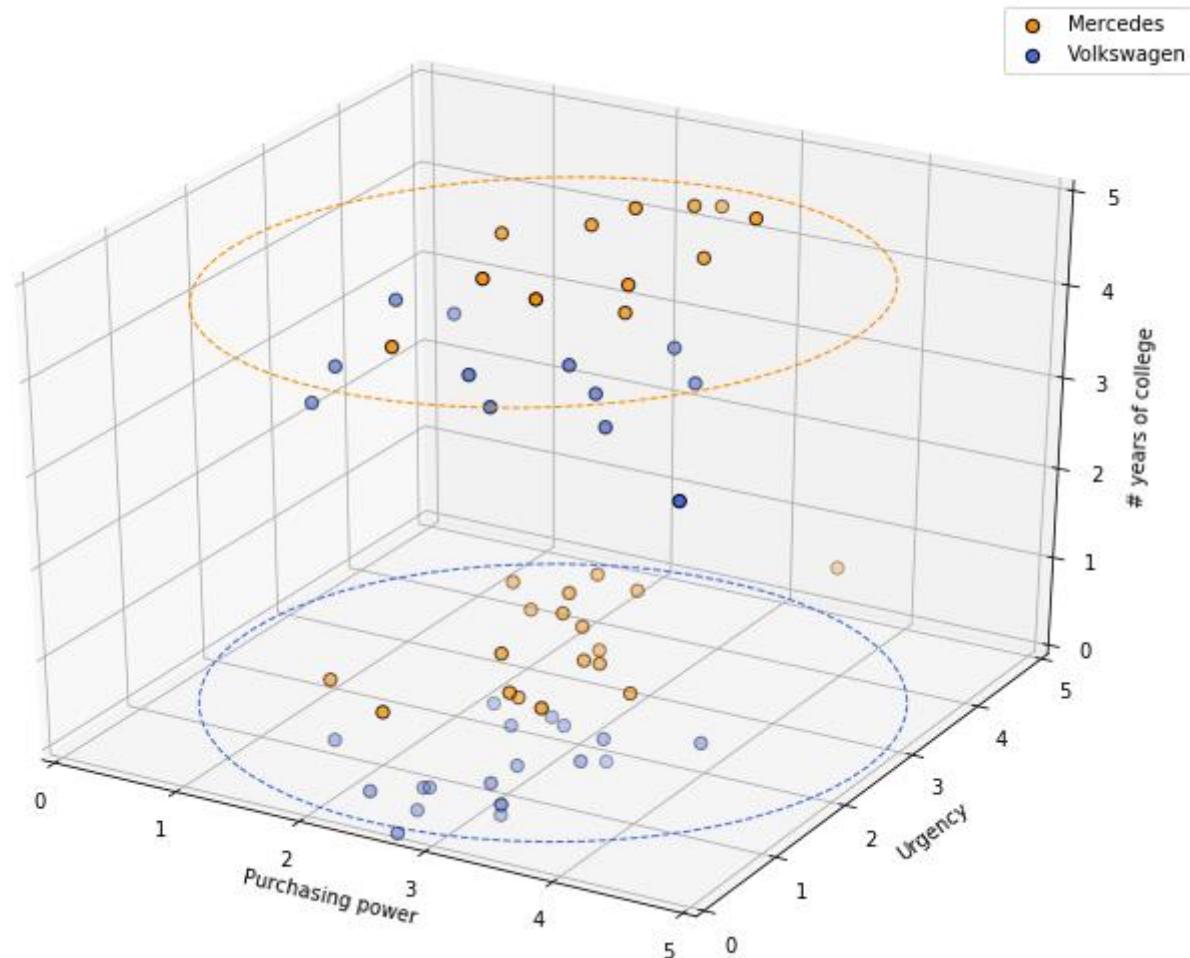
# Big Data vs Small Data

Car Dealer:



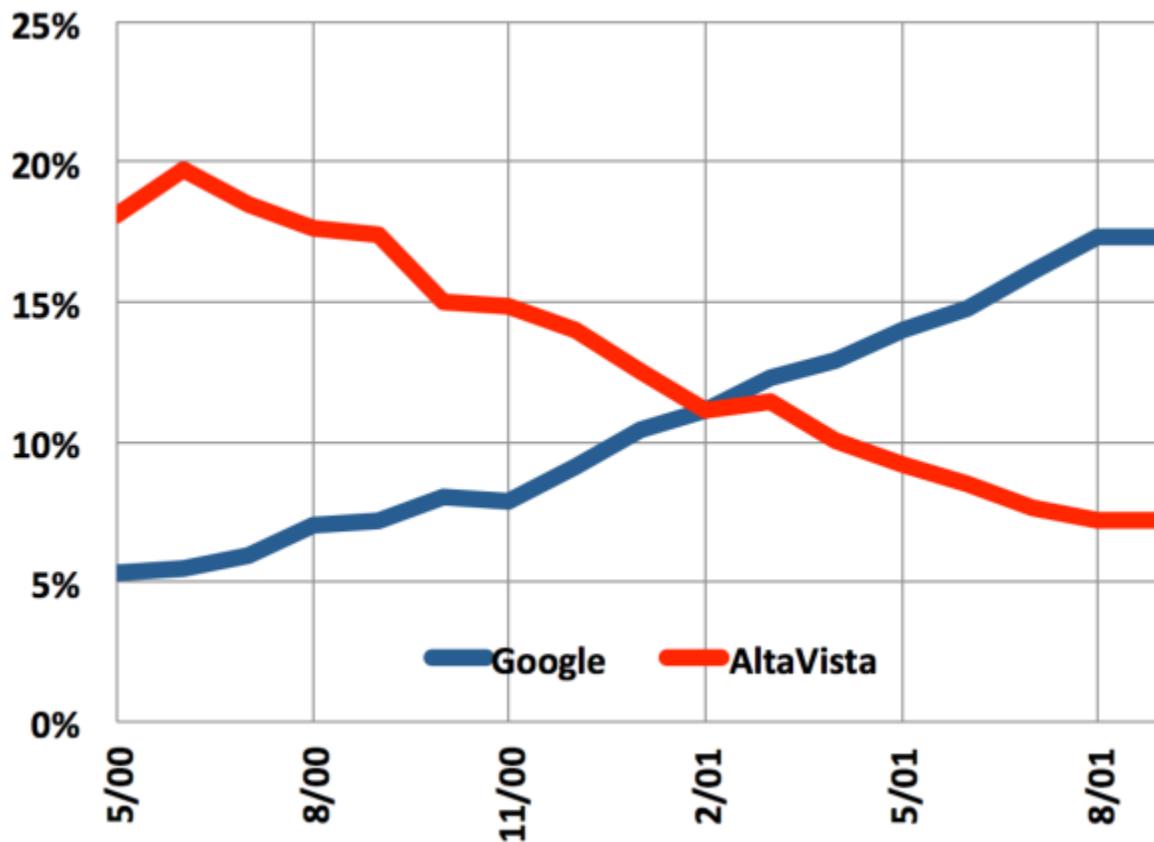
# Big Data vs Small Data

Car Dealer:



# Big Data vs Small Data

Search: Focused Data and Game Theory!



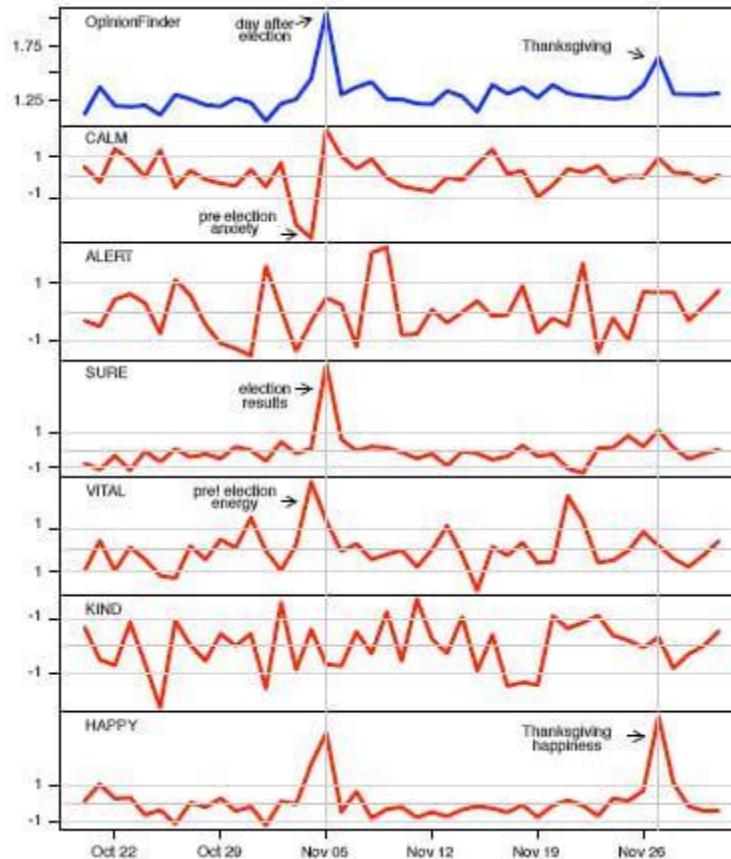
# Big Data vs Big Data

**Billion Dollar Question:** Can you predict Stock Market with Big Data?



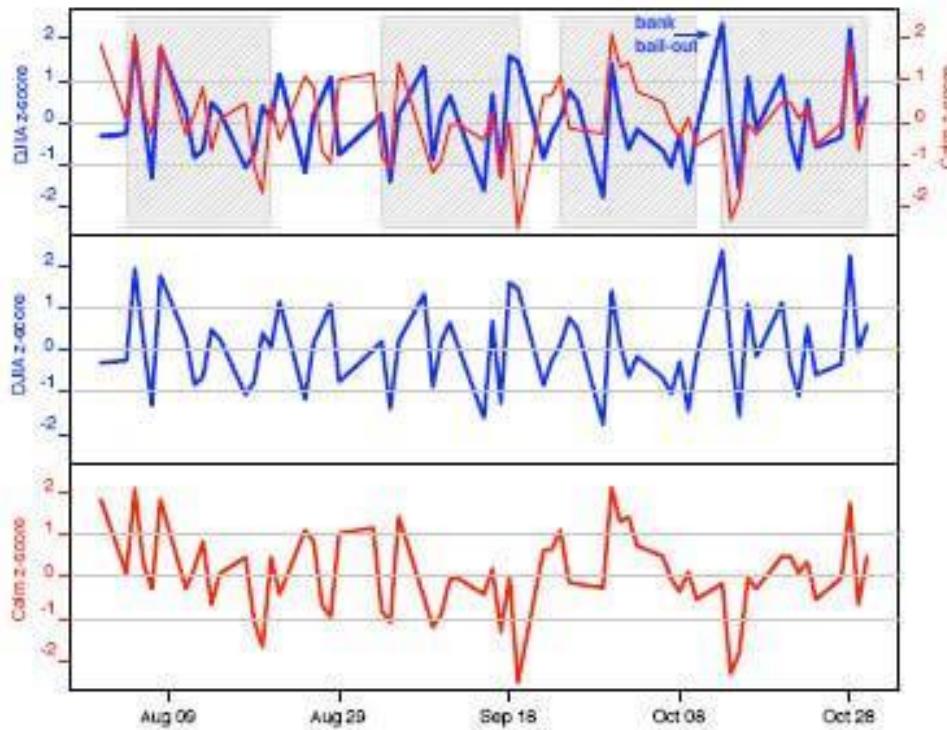
# Stock Market

**Billion Dollar Question:** Can you predict Stock Market with Big Data?



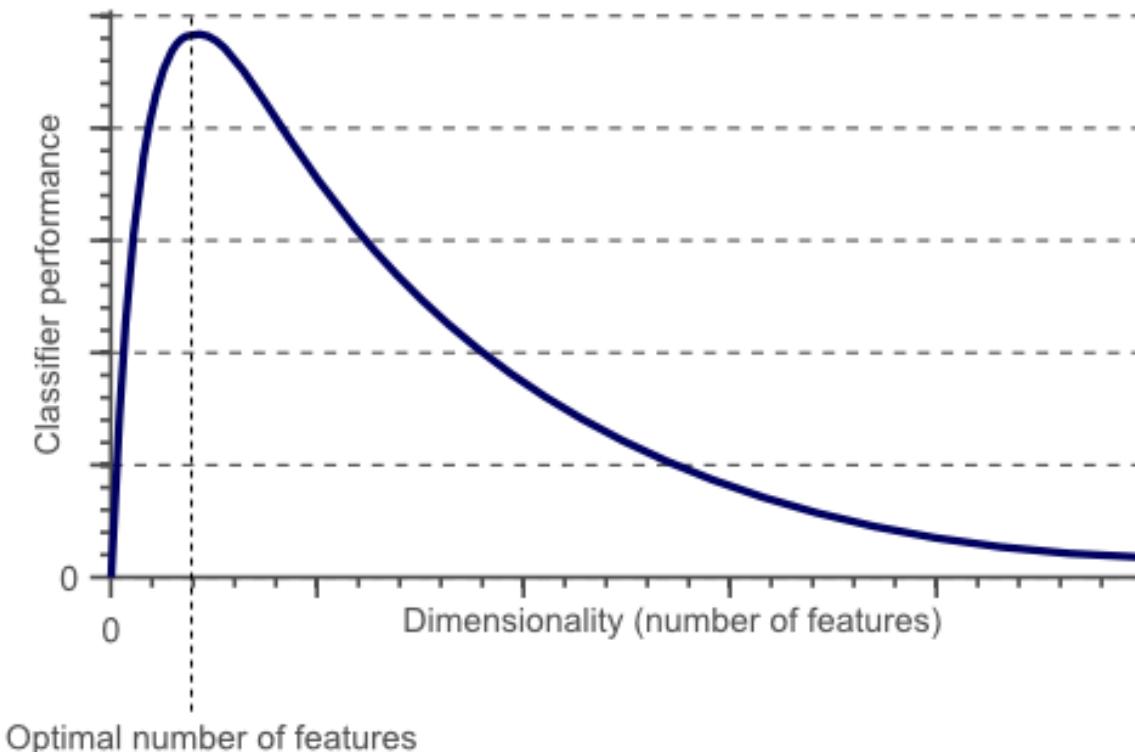
# Stock Market

**Billion Dollar Question:** Can you predict Stock Market with Big Data?



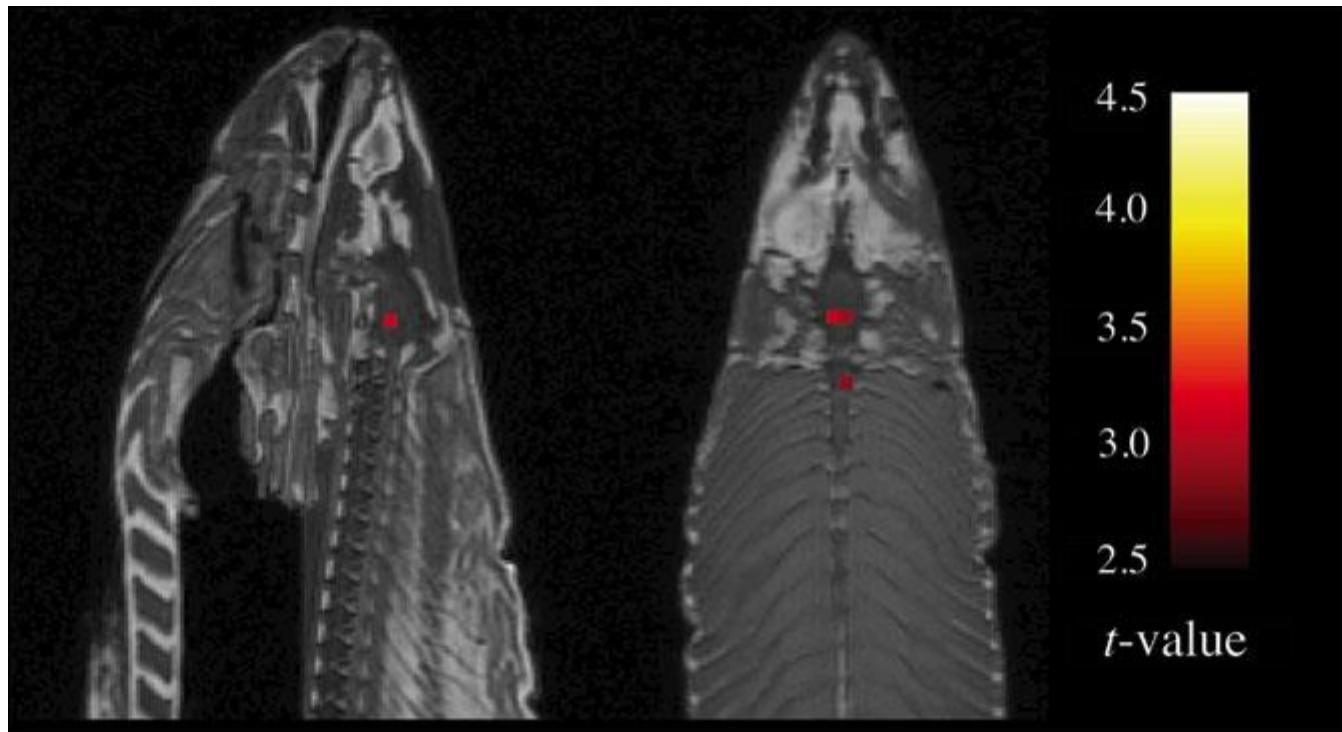
# Stock Market

## Curse of Dimensionality:



# Salmon and MRI

Curse of Dimensionality:



# Salmon and MRI

---

## Curse of Dimensionality:

### METHODS

Subject. One mature Atlantic Salmon (*Salmo salar*) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

Task. The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing.

Design. Stimuli were presented in a block design with each photo presented for 10 seconds followed by 12 seconds of rest. A total of 15 photos were displayed. Total scan time was 5.5 minutes.

# Domain Knowledge vs Data

---

**Hurricanes:** New York Times: Hurricane Frances threatening a direct hit on Florida's Atlantic coast. 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 landfall, 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)



# Domain Knowledge vs Data

**Hurricanes:** It would be valuable to discover patterns due to the hurricane that were not obvious. To do this, analysts might examine the huge volume of Wal-Mart data from prior, similar situations (such as Hurricane Charley) to identify *unusual* local demand for products. From such patterns, the company might be able to anticipate unusual demand for products and rush stock to the stores ahead of the hurricane's landfall.



# Domain Knowledge vs Data

## Image Processing:



# Domain Knowledge vs Data

---

**Suburban Brazil: TV Ownership**



# Domain Knowledge vs Data

## Big Data Fails:

İTÜ EN BÜYÜK ÜNİVERSİTE KÜTÜPHANESİ SİZLERLE KURUYOR DESTEK OL

İTÜ

▼ İTÜ HAKKINDA ▼ AKADEMİK ▼ ARAŞTIRMA ▼ ADAY ÖĞRENCİ ▼ MEZUN

YERLEŞKELER	FAKÜLTELER	ENSTİTÜLER	ÖĞRENİM BİRİMLERİ
Ayazağa	İnşaat	Enerji	Yabancı Diller Yüksekokulu
Taşkışla	Mimarlık	Fen Bilimleri	Türk Musikisi Devlet Konservatuvarı
Maçka	Makina	Sosyal Bilimler	Müzik İleri Araştırmalar Merkezi (MIAM)
Gümüşsuyu	Elektrik - Elektronik	Avrasya Yer Bilimleri	Uluslararası Ortak Lisans Programları
Tuzla	Maden	Bilişim	İnsan ve Toplum Bilimleri Bölümü
KKTC	Kimya - Metalurji	Deprem Mühendisliği ve Afet Yönetimi	Güzel Sanatlar Bölümü
	İşletme		Atatürk İlkeleri ve İnkılap Tarihi Bölümü
	Gemi İnşası ve Deniz Bilimleri		Beden Eğitimi Bölümü
	Fen - Edebiyat		Türk Dili Bölümü
	Uçak ve Uzay Bilimleri		Tüm Bölümler
	Denizcilik		
	Tekstil Teknolojileri ve Tasarımı		
	Bilgisayar ve Bilişim		

Milli Eğitim Bakanlığı ve İTÜ'den Bir İlk  
"İTÜ Mesleki ve Teknik Anadolu Lisesi"

Haberler [Tüm Haberler >>](#)

29.01.2019 Tekstilde Yenilikçi Projeler İTÜ'de Hayat Buluyor

İTÜ Mezun Platformu [Katıl](#) [Birleş](#) [Paylaş](#) [Destek Ol](#)

# Domain Knowledge vs Data

## Quantitative vs Qualitative:

Unique Pageviews

1,008

% of Total: 100.00% (1,008)



Unique Visitors

398

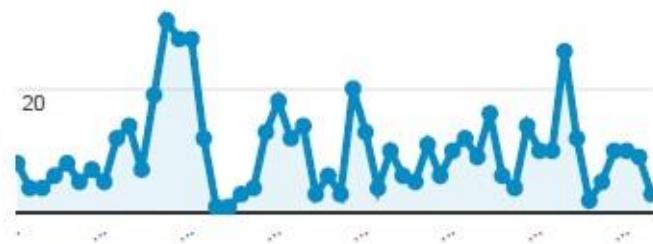
% of Total: 100.00% (398)



Visits

Sessions

40



Visits by Keyword

Keyword

Sessions

(not set)

359

(not provided)

126

ads right column

14

christine brown potential u  
nlocked perth

1

colour selection for brandin

.

# Domain Knowledge vs Data

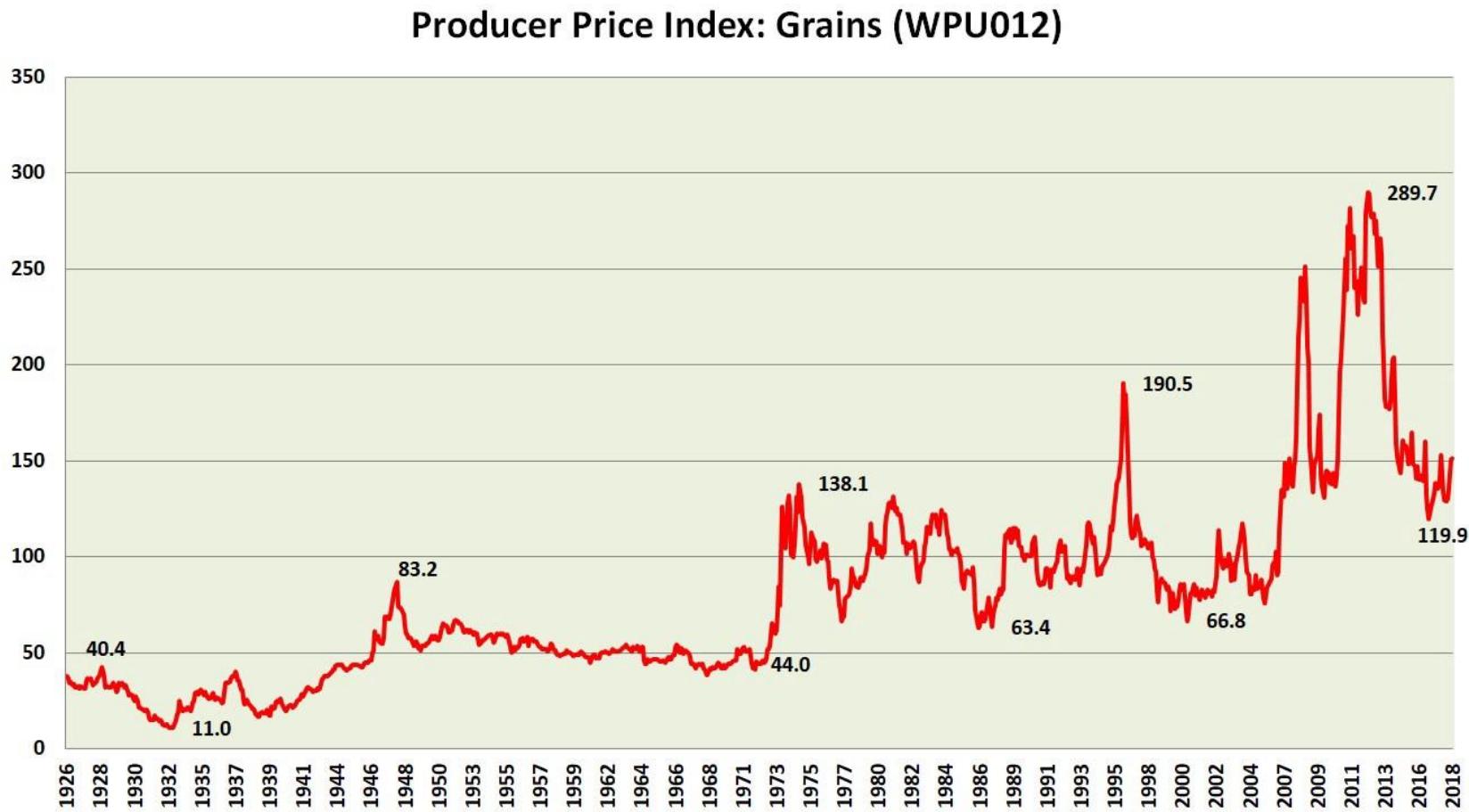
---

Same Problem?:



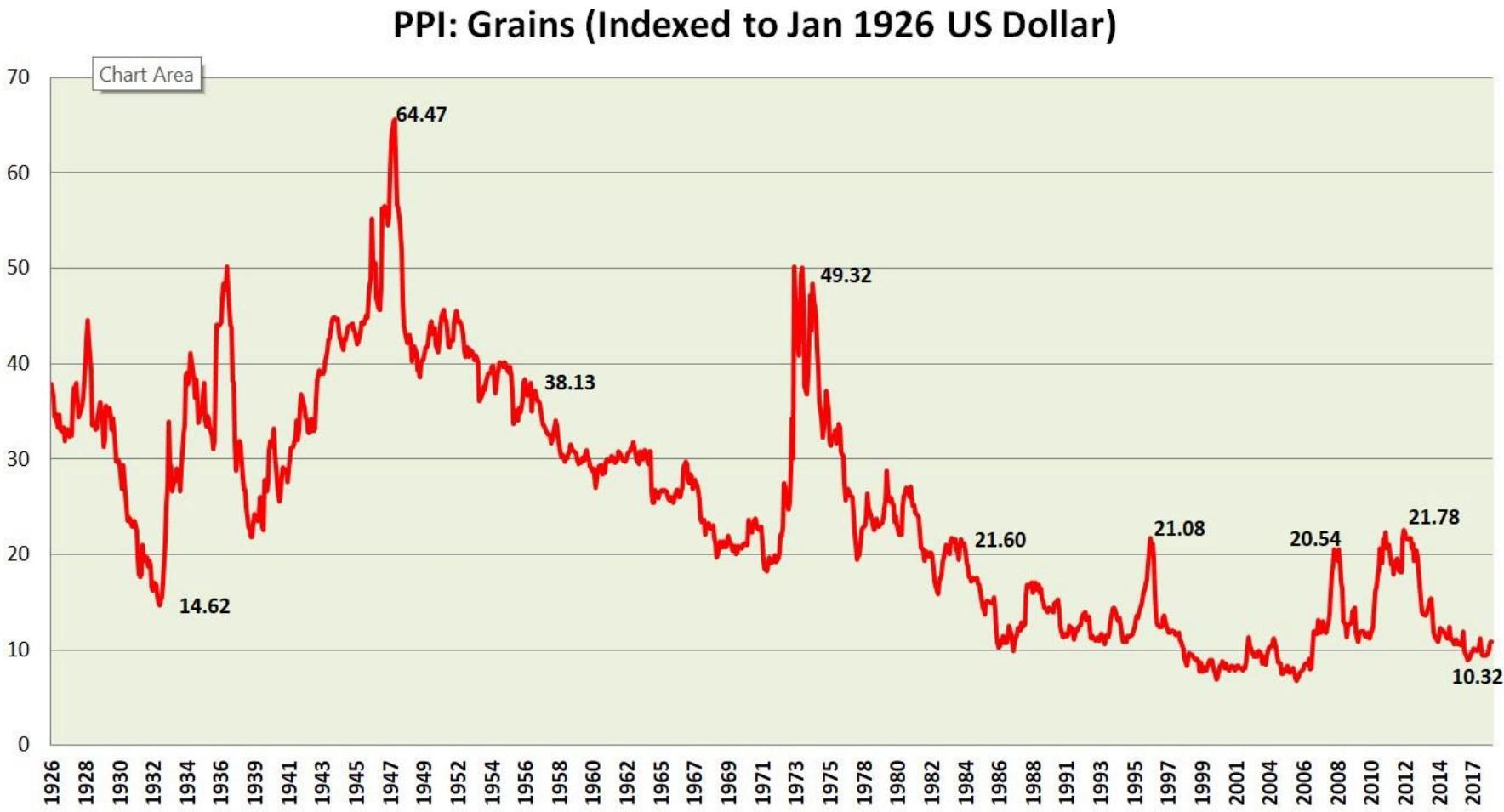
# Domain Knowledge vs Data

**Simple Example:** What happens to prices?



# Domain Knowledge vs Data

**Simple Example:** What happens to prices?



# Sources of Business Data

## In-House vs Outsourced:

Yurt içi üretici fiyat endeksi, bir önceki yılın aynı ayına göre değişim oranı, Aralık 2018  
[2003=100]



# Sources of Business Data

---

**Big Brother:**

<https://www.paraanaliz.com/2017/ekonomi/cinin-sosyal-kredi-sistemi-15531/>

# Sources of Business Data

---

## Data & Talent:

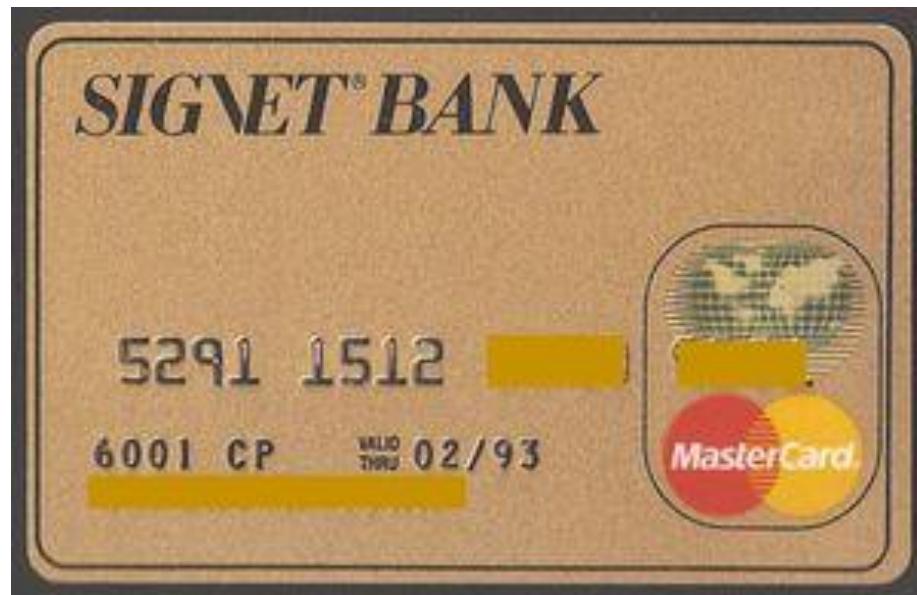
- The right data to best make decisions and/or the right talent to best support making decisions from the data.
- The best data science team can yield little value without the appropriate data.
- The right data often cannot substantially improve decisions without suitable data science talent

# Sources of Business Data

---

## Acquiring Data:

- Predictive modelling for default
- Predicting profitability rather than probability
- No data!!!



# Sources of Business Data

---

## Acquiring Data:

- What will you do if this is a new business model and therefore no data is available?
- Pay for the data!!!

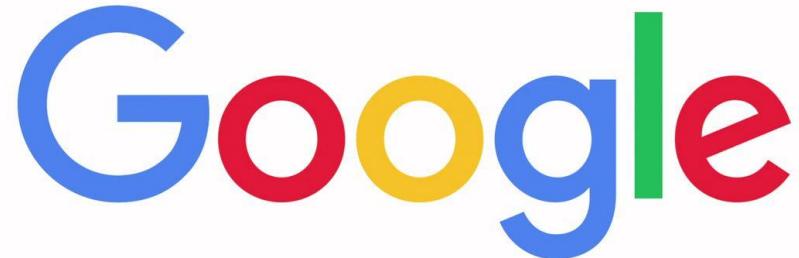


# Sources of Business Data

---

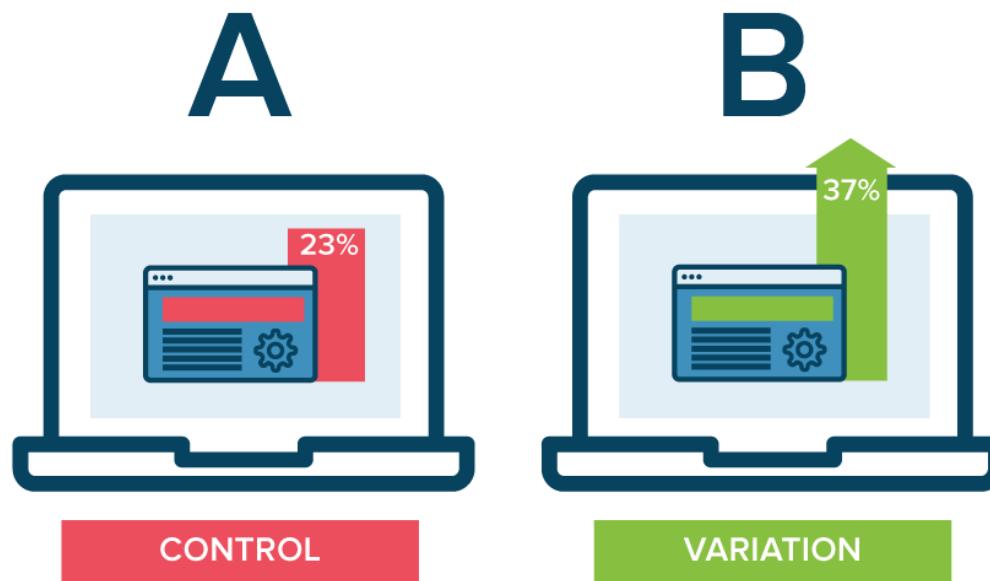
## Create your Data: Experiments!

- A few engineers decided to perform an experiment on Google's site
- The treatment group was shown twenty links on the search results pages.
- The control group was shown the usual ten. The engineers then compared the satisfaction of the two groups based on how frequently they returned to Google.



# Sources of Business Data

**A/B Testing:** A/B testing (also known as split testing or bucket testing) is a method of comparing two versions of a webpage or app against each other to determine which one performs better. AB testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal.



# Sources of Business Data

---

## A/B Testing:

- Discovery A/B tested the components of their video player to engage with their TV show 'super fan.' The result? A 6% increase in video engagement.
- ComScore A/B tested logos and testimonials to increase social proof on a product landing page and increased leads generated by 69%.
- Secret Escapes tested variations of their mobile signup pages, doubling conversion rates and increasing lifetime value.

# Sources of Business Data

A/B Testing:



# Sources of Business Data

## A/B Testing:



# Sources of Business Data

## A/B Testing:

The screenshot shows the homepage of Hurriyet.com.tr. At the top, there is a navigation bar with the logo 'Hürriyet' and links for GÜNDEM, DÜNYA, EKONOMİ, SPOR, YAZARLAR, KELEBEK, and HÜRRIYET DÜNYASI. There is also a search bar and a login area with fields for E-posta and Şifre, and buttons for Giriş Yap and Yeni Üye.

The main banner features a large image of a stadium and the text 'Gören şaştı kaldı' and 'DÜNYADA EŞİ BENZERİ YOK!' with numbered arrows (1, 2, 3, 4, 5) pointing right. Below the banner are four news thumbnails:

- F.Bahçe'nin Rize'de tadi kaçtı!
- O sözleri tekrar etti: Kaosun nedeni...
- PKK'nın bomba deposu bulundu
- Sibel Can'dan Mutlu Kaya'ya cevap

At the bottom, there are two partial images: one of a person's head and another of a stadium scene with the text 'Rize'de yine penaltı krizi!'. The overall layout illustrates a live A/B testing experiment on the website's homepage.

# Sources of Business Data

---

## A/B Testing:

	HEADLINE A	HEADLINE B
1.	Can the SnotBot drone save the whales?	Can this drone help save the whales?
2.	Of course “deflated balls” is a top search term in Massachusetts	This top Mass. Google search term is pretty embarrassing
3.	Hookup contest at heart of St. Paul rape trial	No charges in prep school sex scandal
4.	Woman makes bank off rare baseball card	Woman makes \$179,000 off rare baseball card
5.	MBTA projects annual operating deficit will double by 2020	Get ready: the MBTA's deficit is about to double
6.	How Massachusetts helped win you the right to birth control access	How Boston University helped end “crimes against chastity”
7.	When the first subway opened in Boston	Cartoons from when the first subway opened in Boston
8.	Victim and family in prep-school rape trial blame toxic culture	Victim and family in prep-school rape trial releases statement
9.	Guy in “Free Brady” hat is only one able to foil Miley Cyrus prank	Pats fan gets an eyeful for recognizing an undercover Miley Cyrus

# Sources of Business Data

## A/B Testing:

	HEADLINE A	HEADLINE B	WINNER?
1.	<b>Can the SnotBot drone save the whales?</b>	Can this drone help save the whales?	53% more clicks for A
2.	Of course “deflated balls” is a top search term in Massachusetts	<b>This top Mass. Google search term is pretty embarrassing</b>	986% more clicks for B

	HEADLINE A	HEADLINE B	WINNER?
3.	Hookup contest at heart of St. Paul rape trial	<b>No charges in prep school sex scandal</b>	108% more clicks for B
4.	<b>Woman makes bank off rare baseball card</b>	Woman makes \$179,000 off rare baseball card	38% more clicks for A
5.	MBTA projects annual operating deficit will double by 2020	<b>Get ready: the MBTA's deficit is about to double</b>	62% more clicks for B
6.	How Massachusetts helped win you the right to birth control access	<b>How Boston University helped end “crimes against chastity”</b>	188% more clicks for B
7.	<b>When the first subway opened in Boston</b>	Cartoons from when the first subway opened in Boston	33% more clicks for A
8.	Victim and family in prep-school rape trial blame toxic culture	<b>Victim and family in prep-school rape trial releases statement</b>	76% more clicks for B
9.	Guy in “Free Brady” hat is only one able to foil Miley Cyrus prank	<b>Pats fan gets an eyeful for recognizing an undercover Miley Cyrus</b>	67% more clicks for B

# Sources of Business Data

---

## A/B Testing:

### **Hotels**

www.example.com

Special rates until the end of the month. No booking fees, book your room now!



### **Dublin Hotels**

www.example.com

Browse hundreds of hotels in Dublin, sort by price, location and user reviews.



### **Hotels in Ireland**

www.example.com

Compare prices of 1000s of hotels all over Ireland!



AdChoices

# Sources of Business Data

## Uncommon Places:

- Potential borrowers write a brief description of why they need a loan and about 13 percent of borrowers defaulted on their loan.
- Language that potential borrowers use is a strong predictor of their probability of paying back.



# Sources of Business Data

---

## Uncommon Places:

- Which five are related with the default?

God	will pay
Promise	graduate
debt-free	thank you
lower interest rate	after-tax
minimum payment	hospital

# Sources of Business Data

---

## Uncommon Places:

- Which five are related with the default?
- Is it OK to use this a factor in prediction?
- What about social media information for hiring, etc?

God	will pay
Promise	graduate
debt-free	thank you
lower interest rate	after-tax
minimum payment	hospital

# Sources of Business Data

## No Data:

- When Donald Trump first declared his candidacy, most analysts predicted only a small chance of him becoming the Republican nominee
- Presidential elections are a relatively rare event, the historical data is limited; in other words, the sample size is relatively small and outdated.



# Sources of Business Data

## No Data:

- Transfer Learning identify the areas of knowledge which are “transferable” to the target domain. This broader set of data can then be used to help “train” the model. These algorithms identify the commonalities between the target task, recent tasks, previous tasks, and similar-but-not-the-same tasks. Thus, they help guide the algorithm to learn only from the relevant parts of the data.



# Sources of Business Data

## No Data:

- Consider a company with a successful operation in the U.S. that wants to expand to the German market.

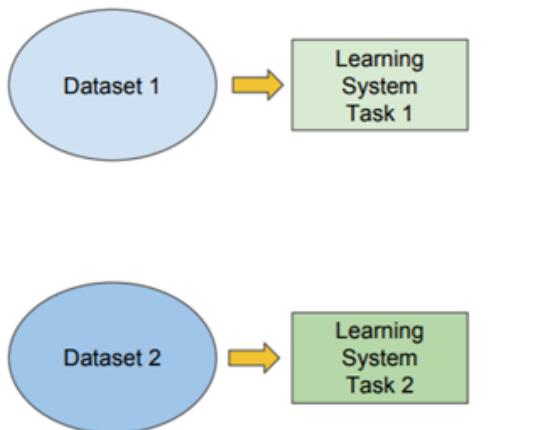


# Sources of Business Data

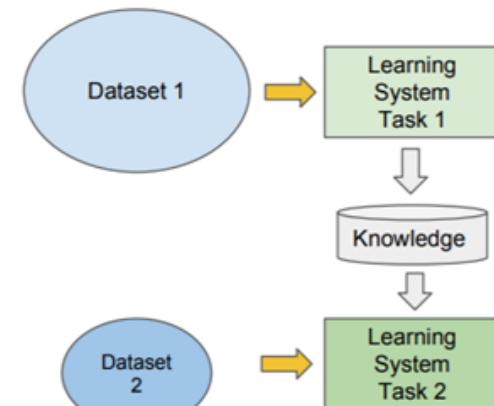
## Transfer Learning:

### Traditional ML      vs      Transfer Learning

- Isolated, single task learning:
  - Knowledge is not retained or accumulated. Learning is performed w.o. considering past learned knowledge in other tasks

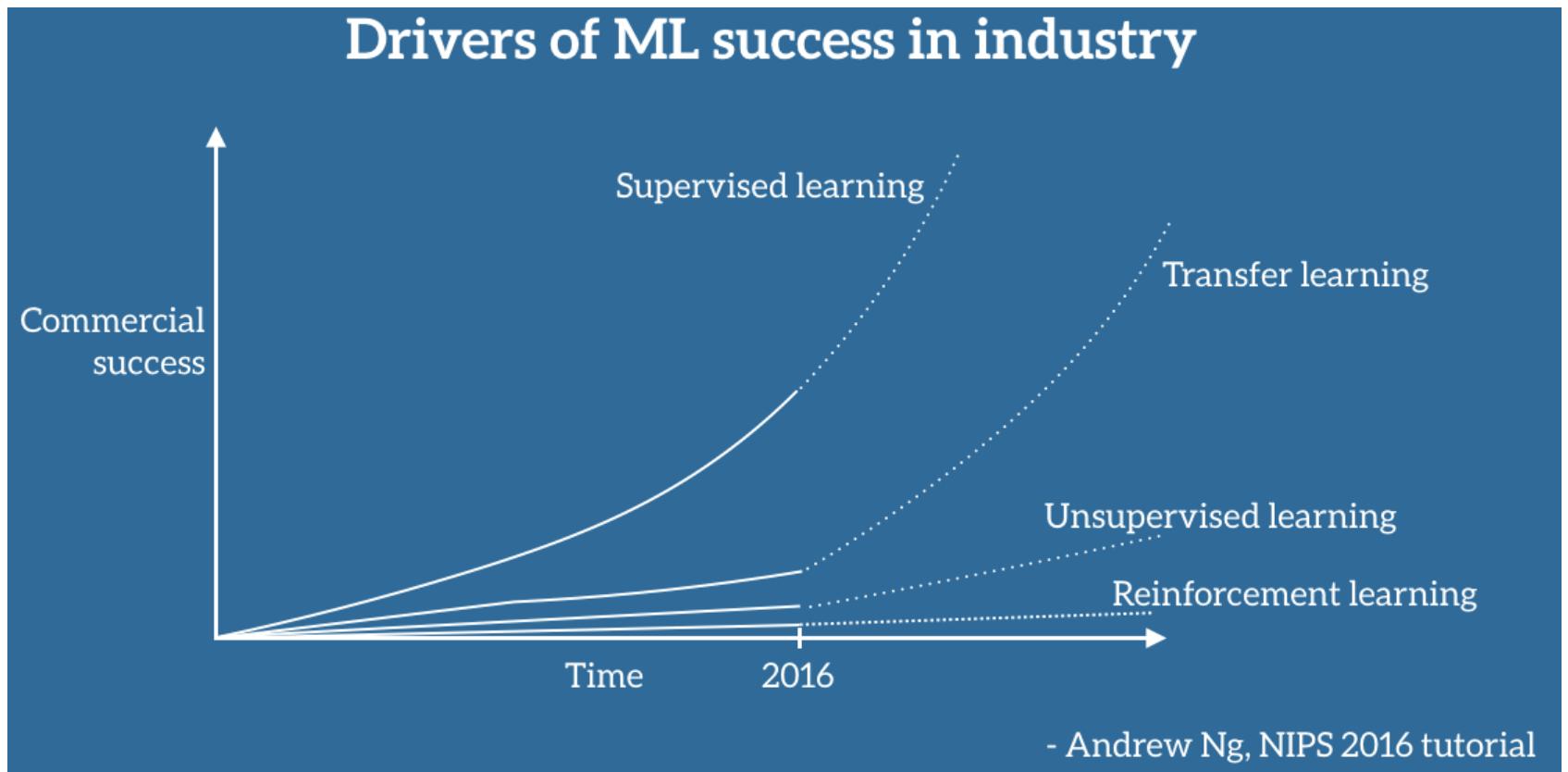


- Learning of a new tasks relies on the previous learned tasks:
  - Learning process can be faster, more accurate and/or need less training data



# Sources of Business Data

Transfer Learning:



# Sources of Business Data

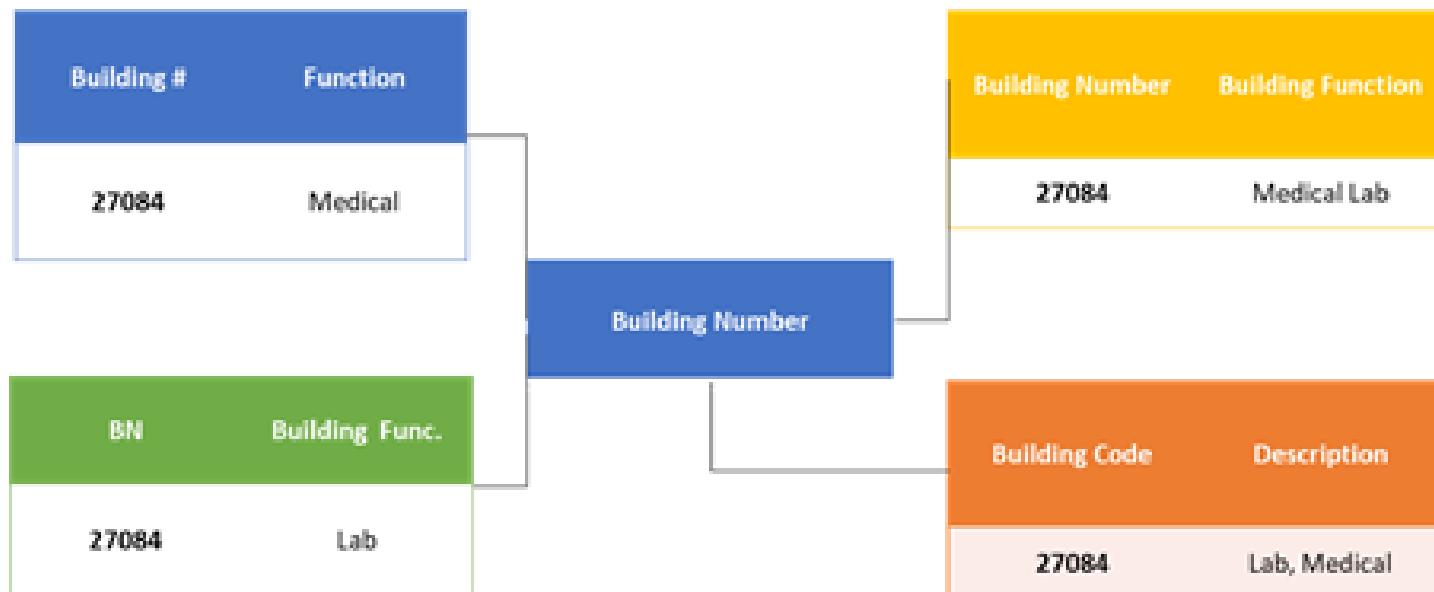
---

## Data Unification:

- The biggest problem data scientist face today is dirty data
- Today data scientists often end up spending 60% of their time cleaning and unifying dirty data before they can apply any analytics or machine learning.
- Data cleaning is essentially the task of removing errors and anomalies or replacing observed values with true values from data to get more value in analytics.
- There are the traditional types of data cleaning like imputing missing data and data transformations and there also more complex data unification problems like deduplication and repairing integrity constraint violations.

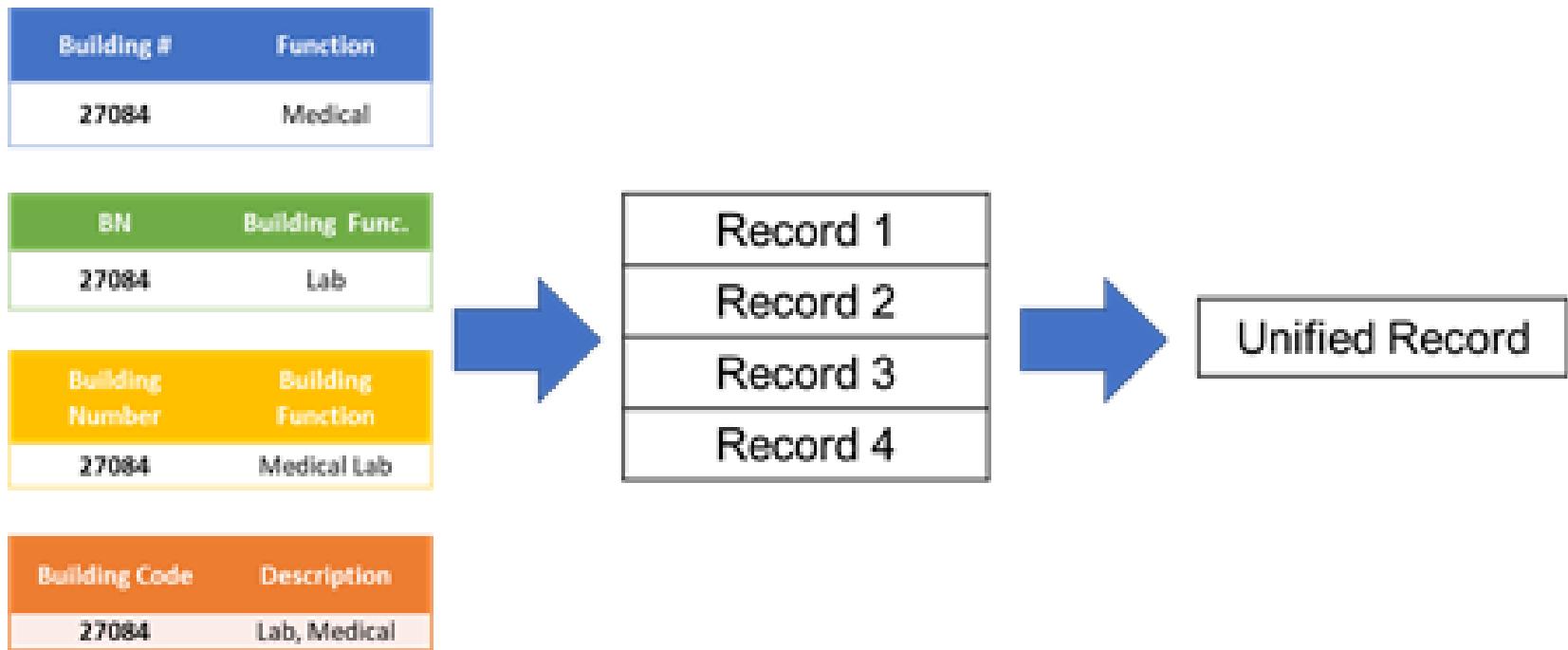
# Sources of Business Data

## Data Unification:



# Sources of Business Data

## Data Unification:



# Sources of Business Data

---

## Data Unification:

Building #	Function	Building Manager	Location
27084	Medical Lab	Jane Smith	123 front str
27093	Administration	John Doe	NULL
NA	Management	William A.	Don't know

# Sources of Business Data

## Data Unification:

Building #	Function	Building Manager	Location	Postal Code
27084	Medical Lab	Jane Smith	123 front str	M3V2V1
27093	Administration	John Doe	212 front str	M3V2V1
6543	Management	William A.	533 front str	M2X1Z1
4432	Management	Jane Smith	114 front street	M4C1Z1

**Business rule:**  
A building manager cannot manage more than one building with different functions.

**Denial Constraint:**  
Two buildings with different locations cannot have the same postal code.

# Sources of Business Data

## Data Unification:

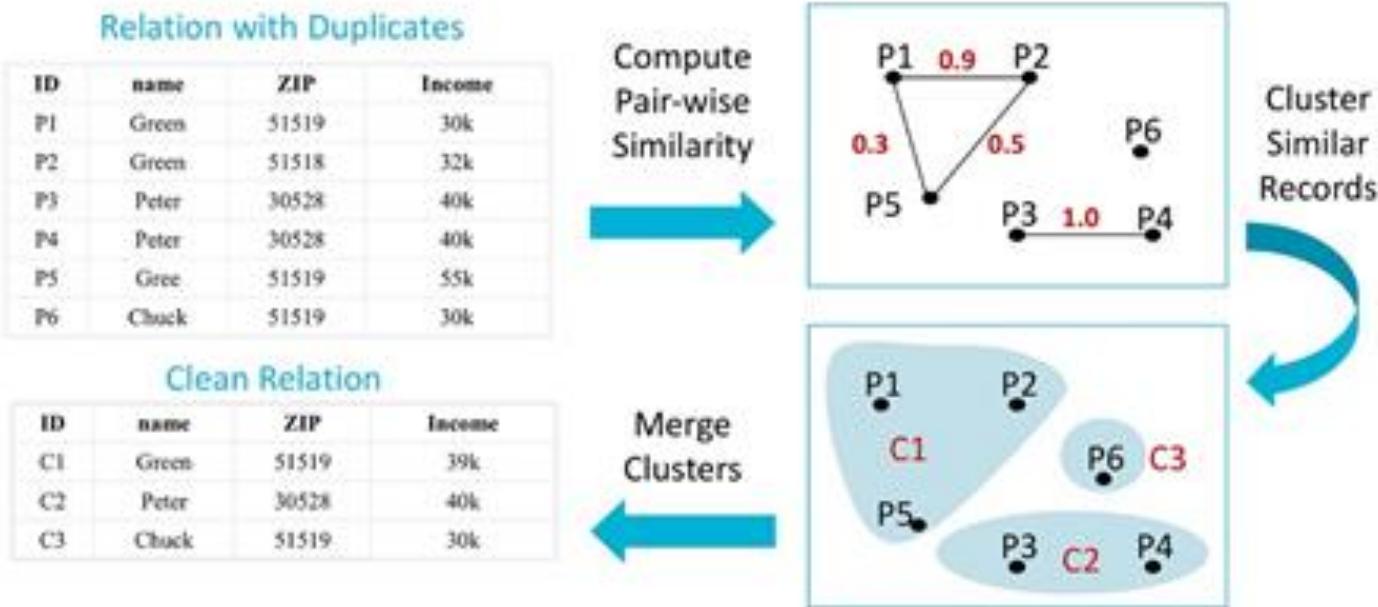
ID	Name	ZIP	City	State	Income
1	Green	60610	Chicago	IL	30k
2	Green	60611	Chicago	IL	32k
3	Peter		New Yrk	NY	40k
4	John	11507	New York	NY	40k
5	Gree	90057	Los Angeles	CA	55k
6	Chuck	90057	San Francisco	CA	30k

Diagram illustrating data unification issues:

- Missing Value:** An arrow points from the empty ZIP value for Peter (row 3) to the "Missing Value" label.
- Duplicates:** An arrow points from the two rows with ZIP 90057 (rows 5 and 6) to the "Duplicates" label.
- Value/Syntactic Error:** An arrow points from the ZIP value 11507 (row 4) to the "Value/Syntactic Error" label.
- Integrity Constraint Violation:** Two arrows point from the ZIP value 11507 (row 4) and the City value "New Yrk" (row 3) to the "Integrity Constraint Violation" label.

# Sources of Business Data

**Data Unification:** The three primary tasks involved in entity resolution are deduplication, record linkage, and canonicalization



# Feature Engineering

---

Data from Unexpected Sources:



# Feature Engineering

---

Data from Unexpected Sources:



# Feature Engineering

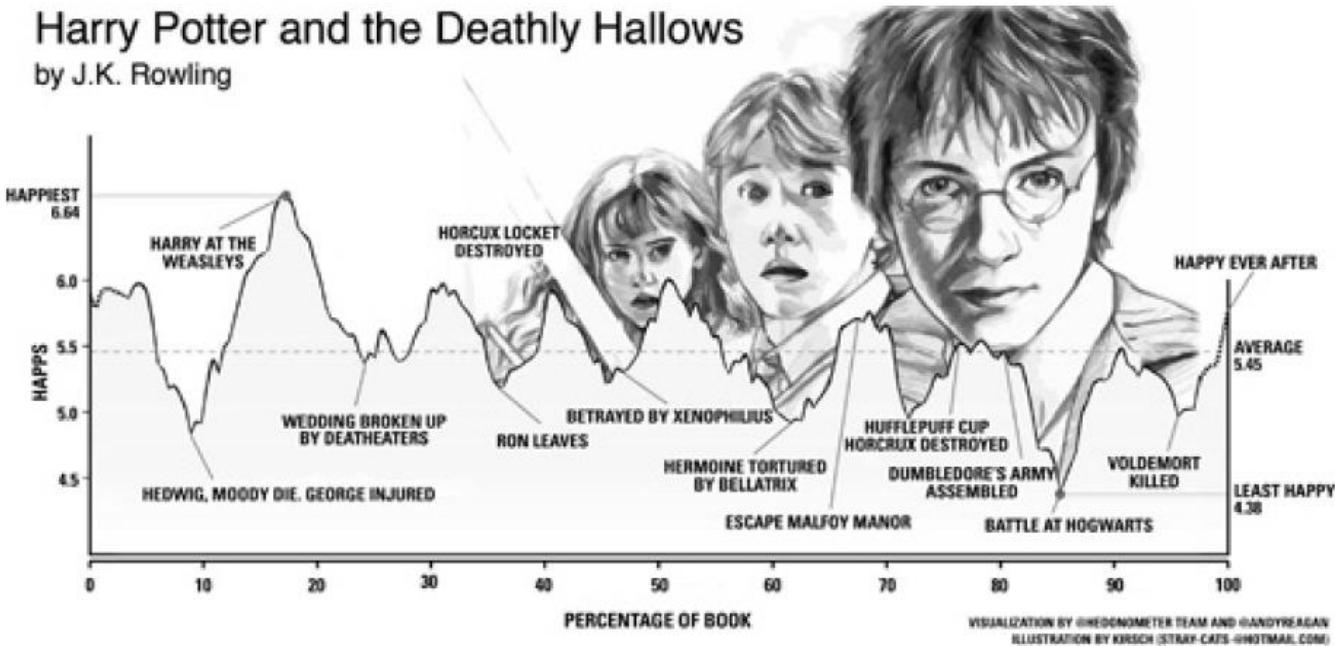
---

## Data from Unexpected Sources:

- If you use new data to revolutionize a field, it is best to go into a field where old methods are lousy
- The second lesson is that, when trying to make predictions, you needn't worry too much about why your models work
- Does that contradict with curse of dimensionality?

# Feature Engineering

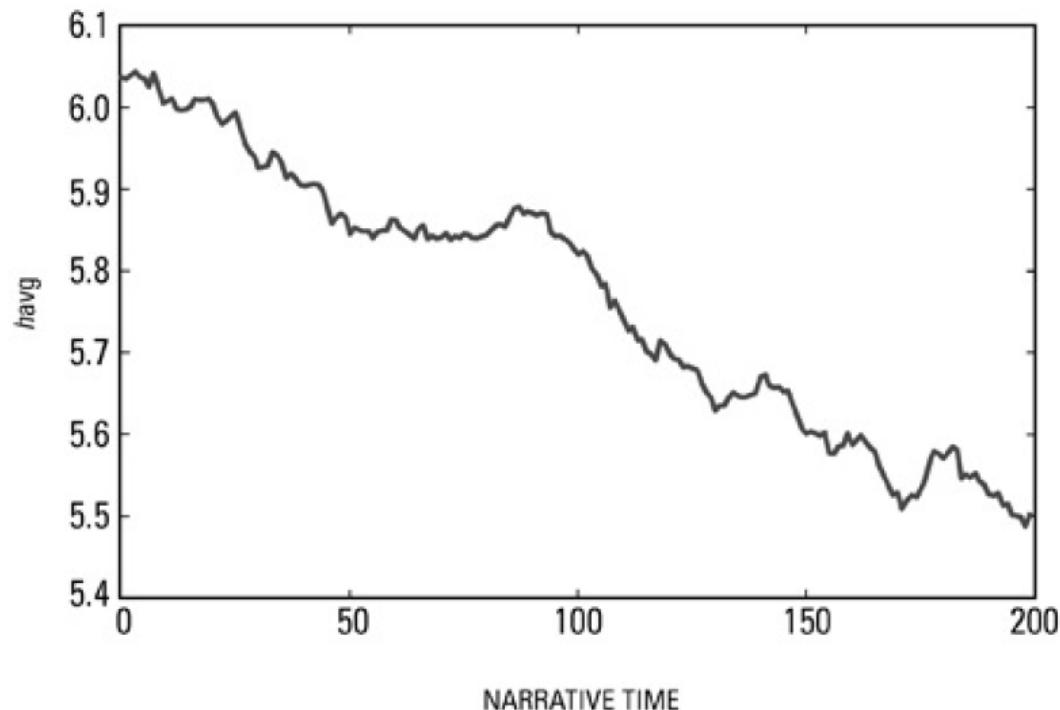
Data from Unexpected Sources:



# Feature Engineering

---

Data from Unexpected Sources:



# Feature Engineering

---

**Data from Unexpected Sources:** Experiments suggest violent movies can incite violent behavior



# Feature Engineering

---

**Data from Unexpected Sources:** On weekends with a popular violent movie, the economists found, crime dropped.



# Feature Engineering

---

**Data from Unexpected Sources:** Utilize the Big Data to zoom in closer.



# Feature Engineering

---

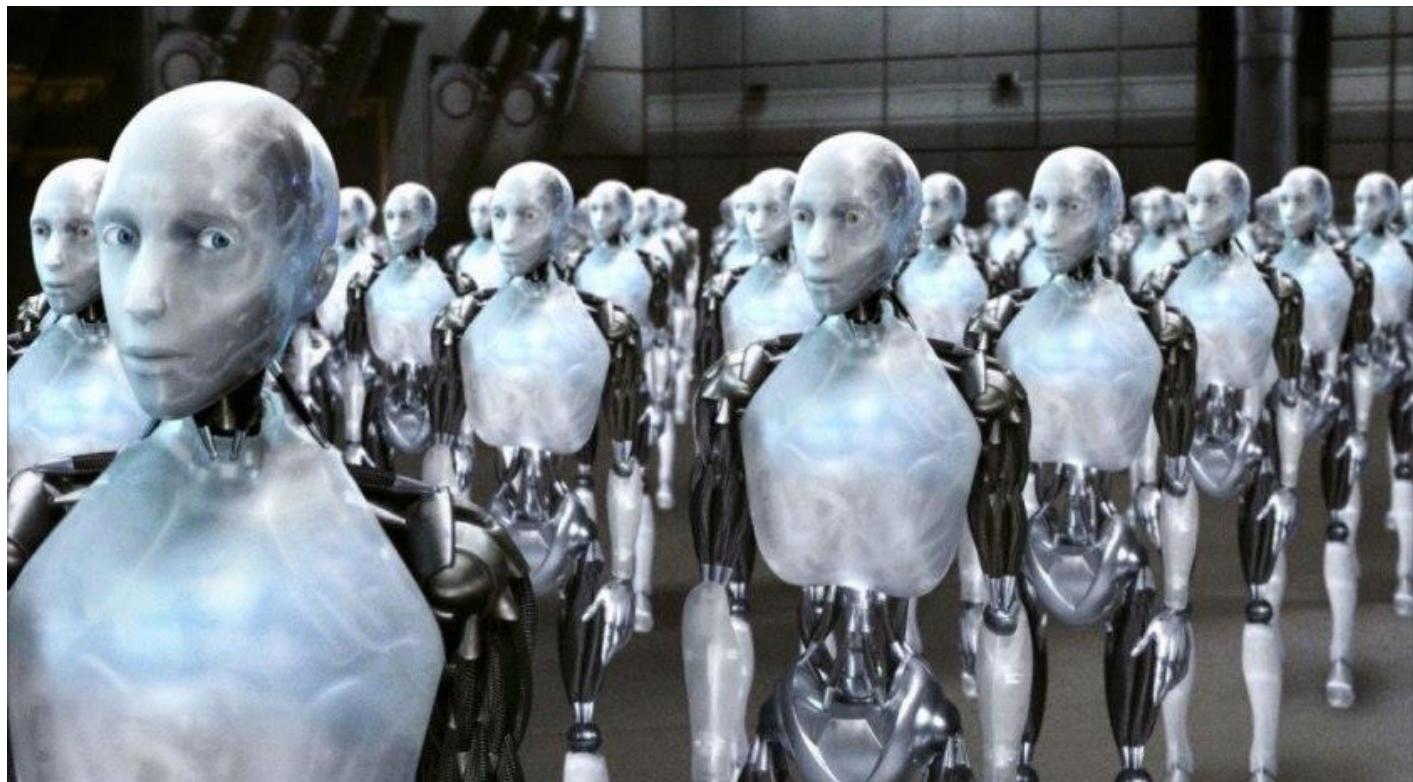
**Data from Unexpected Sources:** Unforeseen causality effect



# Feature Engineering

---

**Last Stronghold:** One of the main reasons Automated ML cannot replace humans. Yet 😊



# DS 555 Data Science and Business Strategy

---

*CUSTOMER ORIENTED THINKING*

– O.Örsan Özener

# Customer Oriented Thinking

## Recommendation Engines:



Jac the Pumpkin Queen

@GirlFromBlupo

Follow



Dear Amazon, I bought a toilet seat because I needed one. Necessity, not desire. I do not collect them. I am not a toilet seat addict. No matter how temptingly you email me, I'm not going to think, oh go on then, just one more toilet seat, I'll treat myself.

12:22 AM - 6 Apr 2018

68,292 Retweets 386,693 Likes



# Customer Oriented Thinking

---

## Human Focus:

- World economy in the 20th century was driven by big brick-and-mortar corporations such as GM, US steel, and Exxon
- In the 21st century, knowledge-based service companies such as Facebook, Amazon, and Google dominate the global market.



# Customer Oriented Thinking

---

## Tesco:

- Tesco, the largest retailer in the UK which grew to become a global retailer
- Its success is mostly due to massive promotion of Tesco Clubcard. Getting on the cycle of customer revisit and loyalty based on the card, Tesco topped the retail market and was able to accumulate sufficient data for understanding types of customers.



# Customer Oriented Thinking

---

## Tesco:

- Accomplishing a remarkable growth by meeting customer needs based on sophisticated data analytics
- However, recently announcing enormous drop in revenue, Tesco admitted a failure in data-based management. Tesco announced an annual pre-tax loss of £6.4 billion in 2015.



# Customer Oriented Thinking

---

## Tesco:

- Fundamental purpose of Tesco Clubcard was not simply loyalty management
- Data-based quantitative marketing ~ individual level analysis of customers
- Why did this strategy fail?



# Customer Oriented Thinking

---

## Tesco:

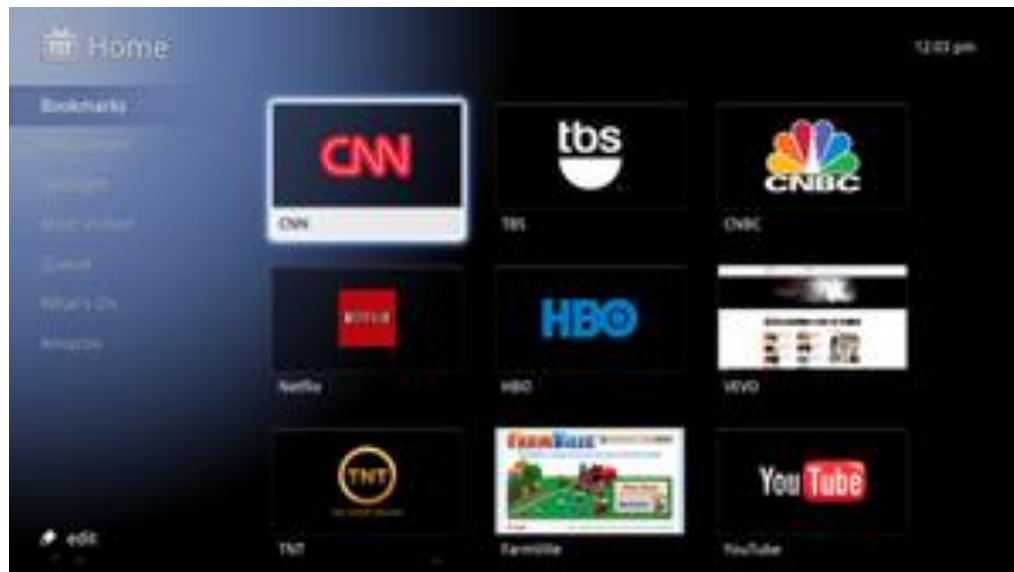
- First, the company put data over business.
- Second, the analysis focused on products instead of customers.
- Third, the rigidity of organization shown in large corporations.



# Customer Oriented Thinking

## Google TV:

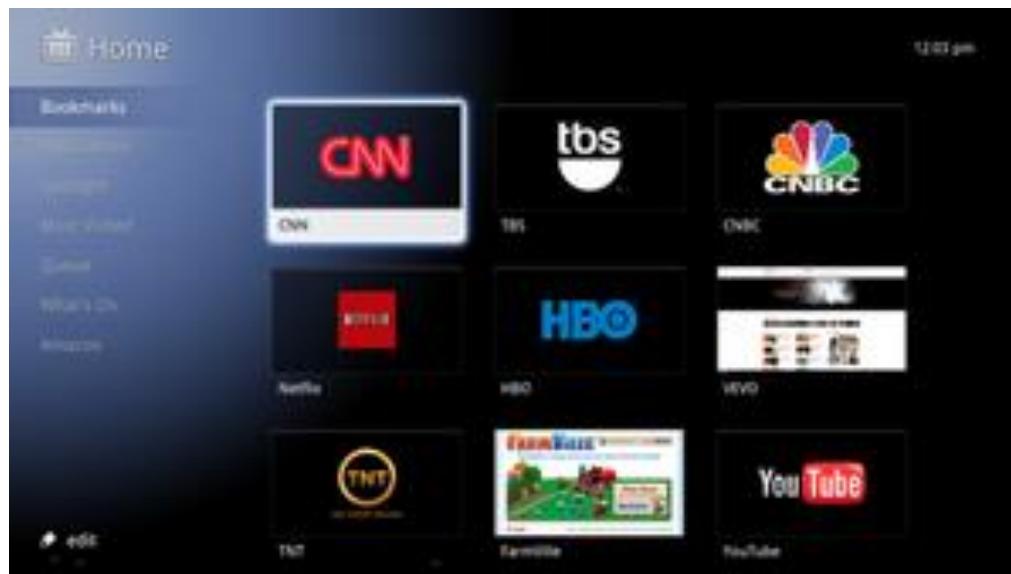
- “TV meets Web, Web meets TV.” In November 2010, Google released Google TV, which uses an Android and Chrome browser.
- Getting results by providing personalized advertisements based on search data on Google.com, Google was aiming to provide personalized advertisements by analysing TV viewing data with the same technology.



# Customer Oriented Thinking

## Google TV:

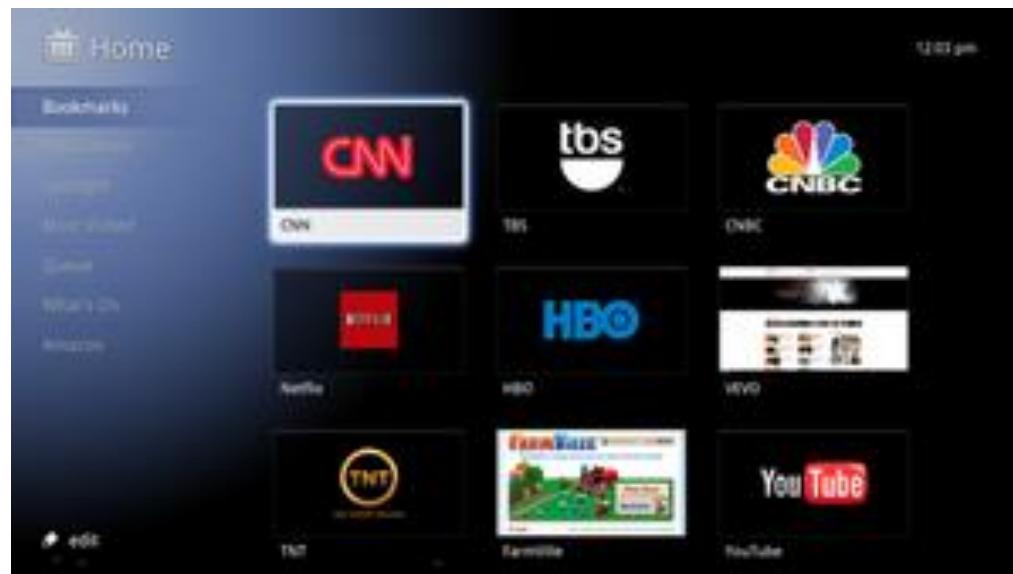
- However, the ambitiously launched Google TV turned out to be an apparent failure.
- Experts suggest that the primary reason for Google TV's failure was Google's inability in securing premium content



# Customer Oriented Thinking

## Google TV:

- The root cause of failure is applying data analytics without understanding consumer behaviour
- Product was too complex for the common usage



# Customer Oriented Thinking

---

## Business Cards:

- When smartphone was booming, business card apps became popular
- These apps scan and recognize characters on business cards, automatically inputting information such as a name, phone number, and e-mail address.



# Customer Oriented Thinking

---

## Business Cards:

- However, by the mid-2010s, the numerous business card apps all disappeared
- The primary target of the business card apps is salespeople. From the salespeople's point of view, card management is a life-support system.
- OCR had limitations. No matter how high accuracy becomes, it was impossible to keep errors below a certain level



# Customer Oriented Thinking

---

## Business Cards:

- *Remember*, a business card app, provides the service of entering information automatically if a picture of a card is taken with a smartphone, like other business card apps. However, unlike other apps that have been struggling, this app has been a great success
- Card recognition accuracy of Remember is nearly 100 percent. How?



# Customer Oriented Thinking

---

## Amazon:

- Online commerce companies, recently have a deadly shortcoming, they cannot meet with customers.
- Therefore, they can do is make inferences about customers. Most online companies still understand customers in numbers from the recency, frequency, and monetary (RFM) perspectives.



# Customer Oriented Thinking

---

## Amazon:

- “What’s the best way to sell refrigerators in the North Pole?”
- Relying on data without consideration of customers’ needs is just to scratch the surface
- The reason why the success of data analytics is difficult is because data analytics experts lack the ability to understand customers.



# User Experience

---

## Telco:

- Monday, August 13: An e-mail from the cell phone provider warning the customer that Amelia was about to exceed her monthly data usage limit of 2GB.
- She was very upset that she was about to go over her limit, and it would start costing her an additional \$10.00 per GB over the limit.



# User Experience

---

## Telco:

- Understand the user experience

Question	Answer
How much of my data plan do I have left?	Current usage as of August 13 is 65 percent
When does my new month start?	On August 14, which is 1 day from today
When am I likely to run over my data plan limit?	The probability of you overrunning your data plan is 0.00001 percent...or <b>NEVER!!</b>

# User Experience

---

**Telco:** Cellular provider could have provided a user experience that highlighted the information and insights necessary to help the customer make a decision about data usage.

- Actual usage to date (65 percent)
- A forecast of usage by the end of the period (67 percent)
- The date when the data plan will reset (in 1 day on August 14)

# User Experience

---

**Telco:** What if 82 percent of data usage had been consumed with 50 percent of usage period remaining? How do we make the user experience and the customer engagement useful, relevant, and actionable?

For example, Telco could offer prescriptive advice about how to reduce data consumption such as:

- Transitioning to apps that are more data usage efficient
- Turning off apps in the background that are unnecessarily consuming data such as mapping apps or apps that are using GPS tracking

Alternatively, Telco could even offer Amelia options to avoid paying an overage penalty such as:

- Purchase a 1-month data usage upgrade for \$2.00 (which is cheaper than the \$10 overage penalty)
- Upgrade existing contract (covering 6 months) for \$10.00

# User Experience

---

**Telco:** This level of customer intimacy can open up all sorts of new monetization opportunities such as:

- Leverage your customer's usage patterns and behaviors to recommend apps that move the user into a more profitable, high-retention user category
- Help app developers to be more successful while collecting referral fees, comarketing fees, and other monetization ideas that align with the app developers' business objectives

# DS 555 Data Science and Business Strategy

---

## *OPTIMIZATION*

– O.Örsan Özener

# What Now?

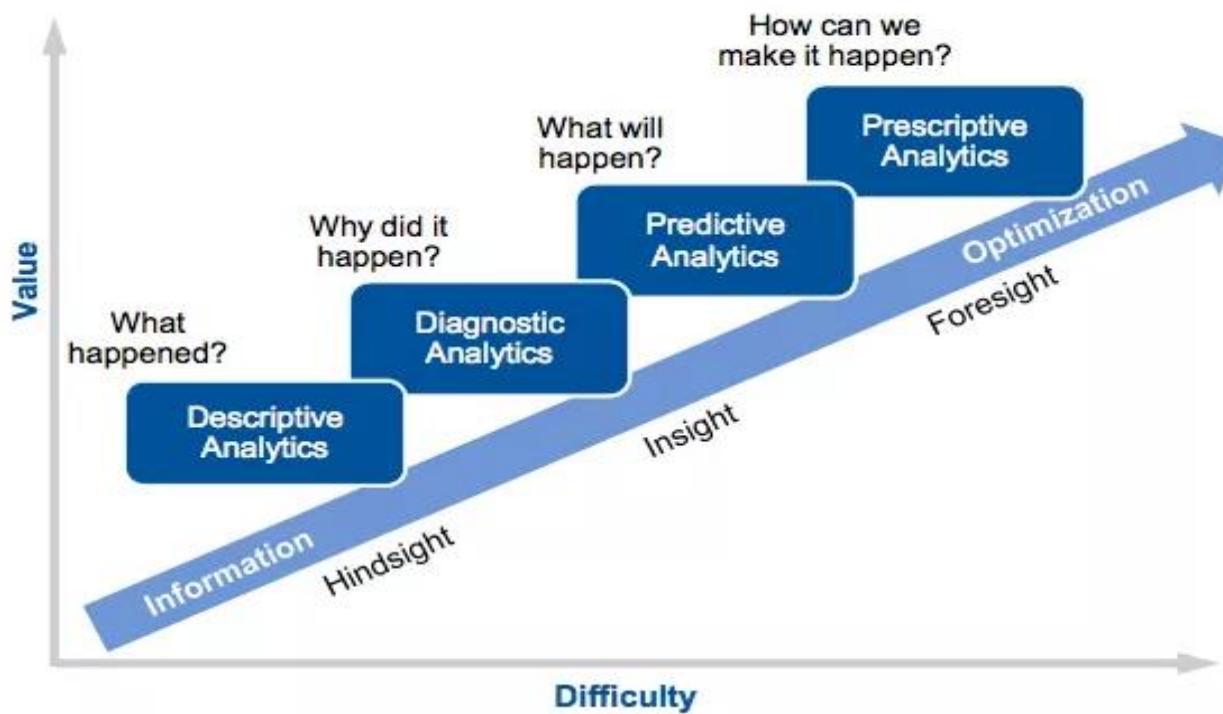
---

**After the Apocalypse:** Suppose that we clean the data, filter out the anomalies, model the consumer patterns, validate the model predict the outcomes, so basically see the end of the tunnel. What is next?



# Data Analytics

## Stages of Data Analytics:



# Roller Coaster Ride

---

**Disney:**

- Forecasting for workforce management
- Forecasting flow in the parks for transportation needs
- Optimizing fast passes and ride wait times
- Optimizing for hotel reservations
- Most importantly revenue management



# Flying High

---

## Airlines:

- Predicting flight delays
- Predicting breakdowns
- Aircraft scheduling
- Crew Scheduling
- Gate Assignment



# Others?

---

**Telecommunication, Banking:**



# Objective

---

**Ultimate Goal:** ML algorithms vs Business Objectives



# Stakes are High

---

**Gambling:**



# More You Buy, the Cheaper It Gets

**Costco:** Costco implements the strategy of reducing price for the products that sell more. Does it make sense?



# More You Buy, the Cheaper It Gets

---

**Costco:** Costco made a steady growth and recorded \$112.6 billion sales and \$20.6 billion net profit in 2014, continuing a winning streak. What is the secret of their success?



# More You Buy, the Cheaper It Gets

---

**Costco:** What would you do?



# More You Buy, the Cheaper It Gets

---

**Costco:** What is the “positioning” of Costco in the grocery market?



# More You Buy, the Cheaper It Gets

---

**Costco:** Solidifying customer base!



# More You Buy, the Cheaper It Gets

## Costco: Solidifying customer base!

Search

Morris/Bloomberg

SHARE THIS ARTICLE

 Share

 Tweet

 Post

 Email

### In this article

COST

**COSTCO WHOLESALE**

297.39 USD

▼ -2.42 -0.81%

AMZN

**AMAZON.COM INC**

1,781.60 USD

▼ -19.20 -1.07%

WMT

**WALMART INC**

119.28 USD

▲ +0.19 +0.16%

BJ

**BJ'S WHOLESALE C**

23.66 USD

▼ -0.04 -0.17%

Bloomberg

Costco Wholesale Corp. rose to a record high Friday after the membership warehouse retailer's August same-store sales topped analysts' estimates, marking an "impressive" end to its fiscal year, according to analysts at both Telsey Advisory Group and Cowen.

Shares rose as much as 2.6% and are now up 49% so far this year, more than double the percentage gain of its closest pure-play rival, BJ's Wholesale Club Holdings Inc.

### Costco Outperforms

**Costco's impressive sales results drives outperformance versus BJ's**



# Donations

---

**Charity:** Consider a real example of targeted marketing: targeting the best prospects for a charity mailing. Fundraising organizations need to manage their budgets and the patience of their potential donors. In any given campaign segment, they would like to solicit from a “good” subset of the donors. This could be a very large subset for an inexpensive, infrequent campaign, or a smaller subset for a focused campaign that includes a not-so-inexpensive incentive package.



# Donations

---

## Charity: What is our objective?

- We would like to maximize our donation *profit*
- Expected benefit of targeting =  $p(R | x) v_R + [1 - p(R | x)]v_{NR}$
- We do not expect consumers to donate spontaneously without a solicitation
- What if we do?



# Donations

---

**Charity:** Data related part?

- We estimate  $p(R | x)$  from the data
- we may be able to estimate  $v_R(x)$  and/or  $v_{NR}(x)$  from the data as well



# Donations

---

**Charity:** Let  $d_R(x)$  be the estimated donation if consumer  $x$  were to respond, and let  $c$  be the mailing cost. Then:

- $p(R | x) d_R(x) > c$



# Donations

---

## Charity: Other Issues

- Selection bias
- Long term donations and Nudging/Nagging
- Total Solicitation Budget



# Churn

**Churn:** Targeting the customers that are likely to churn

- Let's call  $u_S(x)$  the profit from customer  $x$  if she stays; and  $u_{NS}(x)$  the profit from customer  $x$  if she leaves, both not including the incentive cost.
- Furthermore, for simplicity, let's assume that we incur the incentive cost  $c$  no matter whether the customer stays or leaves.



# Churn

**Churn:** The expected benefit of targeting is:

$$EB_T(x) = p(S | x, T) (u_s(x) - c) + [1 - p(S | x, T)] (u_{NS}(x) - c)$$

The expected benefit of not targeting is:

$$EB_{notT}(x) = p(S | x, notT) (u_s(x) - c) + [1 - p(S | x, notT)] (u_{NS}(x) - c)$$



# Churn

---

**Churn:** Value of targeting is:

$$VT = EB_T(\mathbf{x}) - EB_{notT}(\mathbf{x}),$$

In other words:

$$\Delta(p) uS(x) - c$$



# LINEAR REGRESSION

---

## Real Life Data

- **Empirical problem:** Class size and educational output
  - Policy question: What is the effect on test scores (or some other outcome measure) of reducing class size by one student per class? by 8 students/class?
  - We must use data to find out (is there any way to answer this *without* data?)

# The California Test Score Data Set

All K-6 and K-8 California school districts ( $n = 420$ )

## Variables:

- 5<sup>th</sup> grade test scores (Stanford-9 achievement test, combined math and reading), district average
- Student-teacher ratio (STR) = no. of students in the district divided by no. full-time equivalent teachers

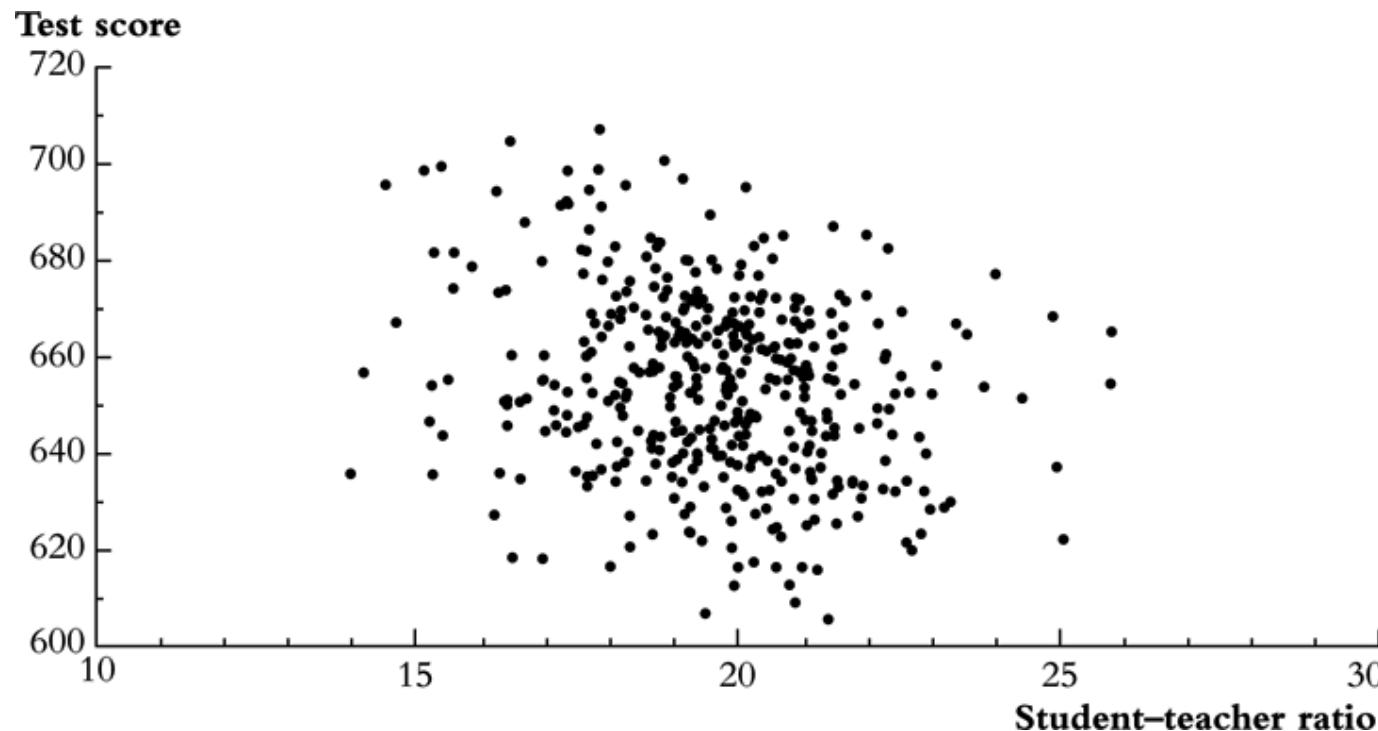
# Initial look at the data:

		Percentile							
	Average	Standard Deviation	10%	25%	40%	50% (median)	60%	75%	90%
Student–teacher ratio	19.6	1.9	17.3	18.6	19.3	19.7	20.1	20.9	21.9
Test score	654.2	19.1	630.4	640.0	649.1	654.5	659.4	666.7	679.1

This table doesn't tell us anything about the relationship between test scores and the *STR*.

# Do districts with smaller classes have higher test scores?

**Scatterplot** of test score v. student-teacher ratio



*What does this figure show?*

We need to get some numerical evidence on whether districts with low STRs have higher test scores – but how?

1. Compare average test scores in districts with low STRs to those with high STRs (“***estimation***”)
2. Test the “null” hypothesis that the mean test scores in the two types of districts are the same, against the “alternative” hypothesis that they differ (“***hypothesis testing***”)
3. Estimate an interval for the difference in the mean test scores, high v. low STR districts (“***confidence interval***”)

Initial data analysis: Compare districts with “small” ( $\text{STR} < 20$ ) and “large” ( $\text{STR} \geq 20$ ) class sizes:

Class Size	Average score ( $\bar{Y}$ )	Standard deviation ( $s_{B_{YB}}$ )	$n$
Small	657.4	19.4	238
Large	650.0	17.9	182

1. ***Estimation*** of  $\Delta$  = difference between group means
2. ***Test the hypothesis*** that  $\Delta = 0$
3. Construct a ***confidence interval*** for  $\Delta$

*This framework allows rigorous statistical inferences about moments of population distributions using a sample of data from that population ...*

1. **Estimation**
2. Testing
3. Confidence Intervals

## **Estimation**

$\bar{Y}$  is the natural estimator of the mean. But:

- a) What are the properties of  $\bar{Y}$ ?
- b) Why should we use  $\bar{Y}$  rather than some other estimator?
  - $YB_{1B}$  (the first observation)
  - maybe unequal weights – not simple average
  - $\text{median}(YB_{1B}, \dots, YB_{nB})$

The starting point is the sampling distribution of ...  $\bar{Y}$

# 1. Estimation

$$\bar{Y}_{\text{small}} - \bar{Y}_{\text{large}}$$

$$= 657.4 - 650.0$$

$$= 7.4$$

Is this a large difference in a real-world sense?

- Standard deviation across districts = 19.1
- Difference between 60<sup>th</sup> and 75<sup>th</sup> percentiles of test score distribution is 667.6 – 659.4 = 8.2
- This is a big enough difference to be important for school reform discussions, for parents, or for a school committee?

# Hypothesis Testing

The ***hypothesis testing*** problem (for the mean): make a provisional decision based on the evidence at hand whether a null hypothesis is true, or instead that some alternative hypothesis is true. That is, test

- $H_0: E(Y) = \mu_{Y,0}$  vs.  $H_1: E(Y) > \mu_{Y,0}$  (1-sided,  $>$ )
- $H_0: E(Y) = \mu_{Y,0}$  vs.  $H_1: E(Y) < \mu_{Y,0}$  (1-sided,  $<$ )
- $H_0: E(Y) = \mu_{Y,0}$  vs.  $H_1: E(Y) \neq \mu_{Y,0}$  (2-sided)

# *Some terminology for testing statistical hypotheses:*

**p-value** = probability of drawing a statistic (e.g.  $\bar{Y}$ ) at least as adverse to the null as the value actually computed with your data, assuming that the null hypothesis is true.

The **significance level** of a test is a pre-specified probability of incorrectly rejecting the null, when the null is true.

**Calculating the p-value** based on :  $\bar{Y}$

p-value =

$$\Pr_{H_0} [| \bar{Y} - m_{Y,0} | > | \bar{Y}^{act} - m_{Y,0} | ]$$

Where  $\bar{Y}^{act}$  is the value of  $\bar{Y}$  actually observed (nonrandom)

# What is the link between the $p$ -value and the significance level?

- The significance level is prespecified. For example, if the prespecified significance level is 5%,
  - you reject the null hypothesis if  $|t| \geq 1.96$ .
  - Equivalently, you reject if  $p \leq 0.05$ .
  - The  $p$ -value is sometimes called the ***marginal significance level***.
  - Often, it is better to communicate the  $p$ -value than simply whether a test rejects or not – the  $p$ -value contains more information than the “yes/no” statement about whether the test rejects.

# Hypothesis Testing:

Size	$\bar{Y}$	$sB_{\gamma_B}$	$n$
small	657.4	19.4	238
large	650.0	17.9	182

$$t = \frac{\bar{Y}_s - \bar{Y}_l}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}} = \frac{657.4 - 650.0}{\sqrt{\frac{19.4^2}{238} + \frac{17.9^2}{182}}} = \frac{7.4}{1.83} = 4.05$$

$|t| > 1.96$ , so reject (at the 5% significance level) the null hypothesis that the two means are the same.

# The Linear Regression Model

The *population regression line*:

$$\text{Test Score} = \beta_0 + \beta_1 \text{STR}$$

$\beta_1$  = slope of population regression line

$$= \frac{\Delta \text{Test score}}{\Delta \text{STR}}$$

= change in test score for a unit change in STR

- Why are  $\beta_0$  and  $\beta_1$  “population” parameters?
- We would like to know the population value of  $\beta_1$ .
- We don’t know  $\beta_1$ , so must estimate it using data.

# The Population Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \quad i = 1, \dots, n$$

- We have  $n$  observations,  $(X_i, Y_i)$ ,  $i = 1, \dots, n$ .
- $X$  is the ***independent variable*** or ***regressor***
- $Y$  is the ***dependent variable***
- $\beta_0$  = ***intercept***
- $\beta_1$  = ***slope***
- $u_i$  = the regression ***error***
- The regression error consists of omitted factors. In general, these omitted factors are other factors that influence  $Y$ , other than the variable  $X$ . The regression error also includes error in the measurement of  $Y$ .

# The Ordinary Least Squares Estimator

*How can we estimate  $\beta_0$  and  $\beta_1$  from data?*

Recall that  $m$  was the least squares estimator of  $\mu_Y$ : solve  $\bar{Y}$

$$\min_m \sum_{i=1}^n (Y_i - m)^2$$

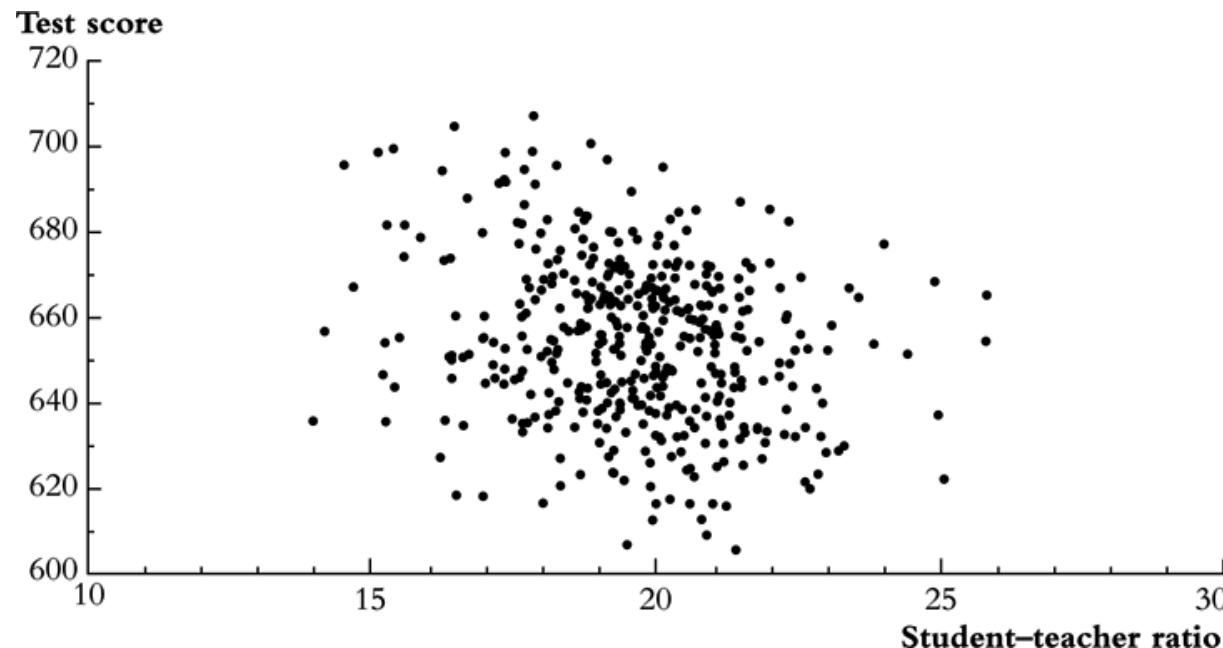
By analogy, **we will focus on the least squares (“ordinary least squares” or “OLS”) estimator of the unknown parameters  $\beta_0$  and  $\beta_1$ .** The OLS estimator solves,

$$\min_{b_0, b_1} \sum_{i=1}^n [Y_i - (b_0 + b_1 X_i)]^2$$

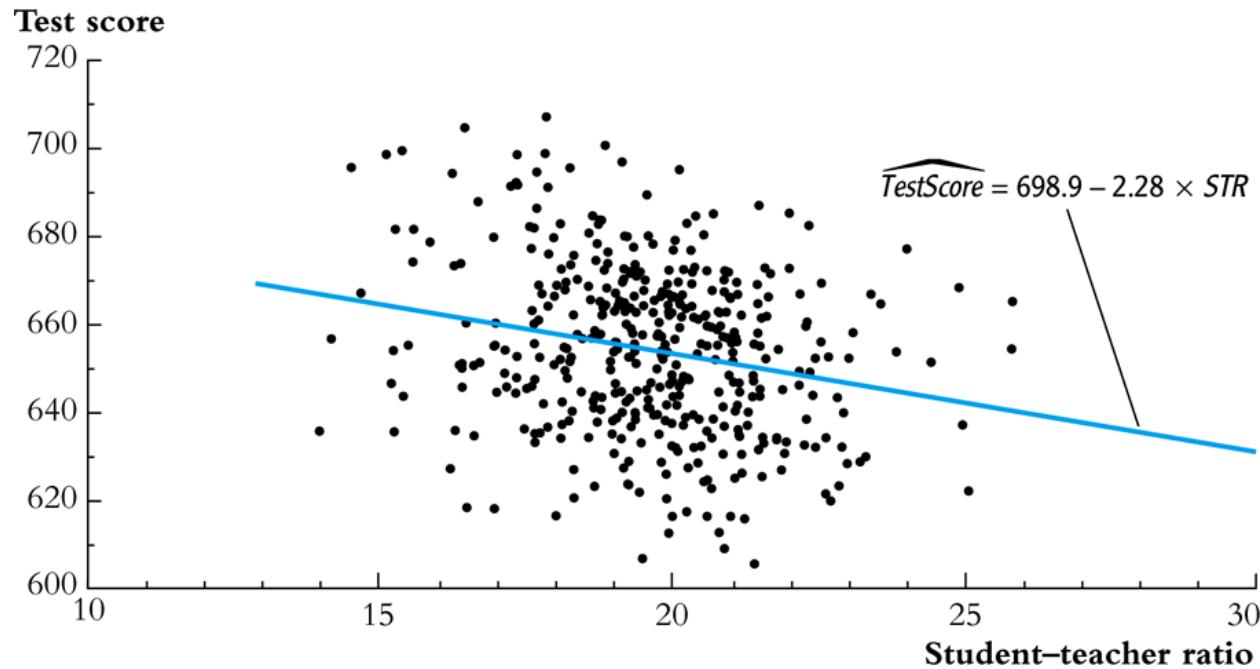
# Mechanics of OLS

The population regression line:  $\text{Test Score} = \beta_0 + \beta_1 \text{STR}$

$$\beta_1 = \frac{\Delta \text{Test score}}{\Delta \text{STR}} = ??$$



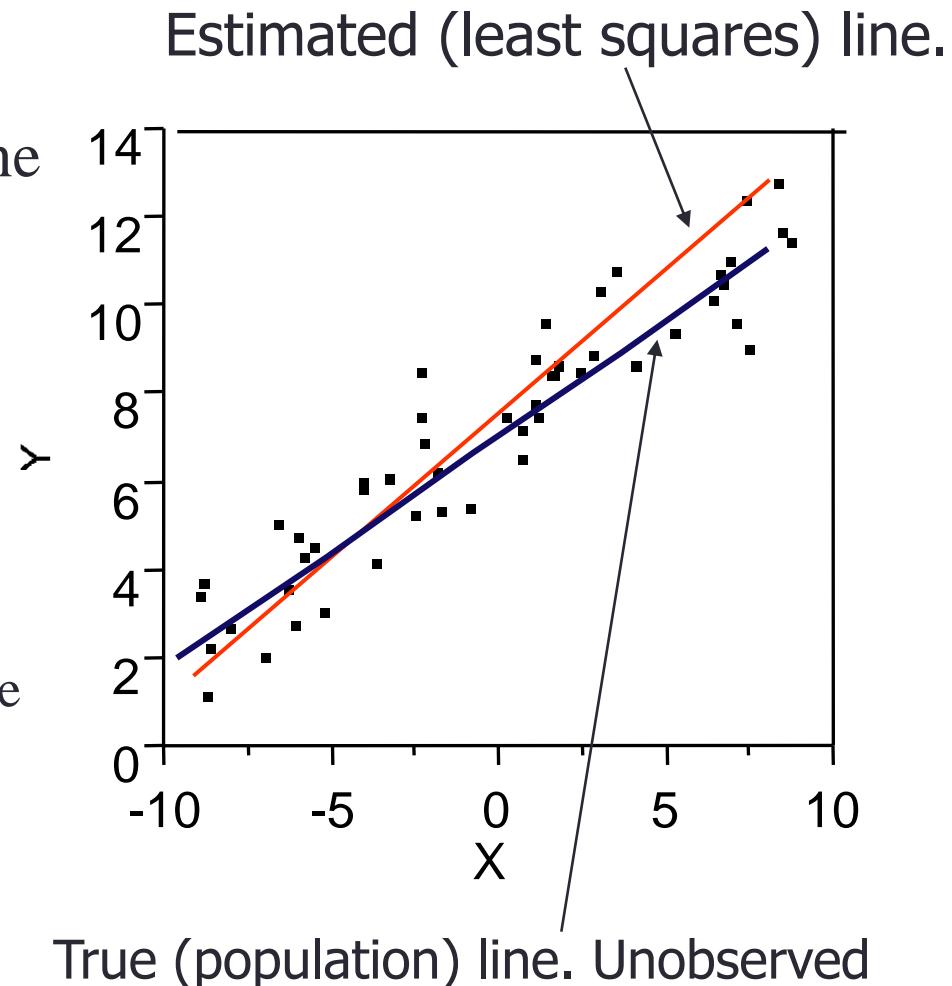
# Application to the California *Test Score* – Class Size data



- Estimated slope =  $\hat{\beta}_1 = -2.28$
- Estimated intercept =  $\hat{\beta}_0 = 698.9$
- Estimated regression line:  $\text{TestScore} = 698.9 - 2.28 \times \text{STR}$

# Inference in Regression

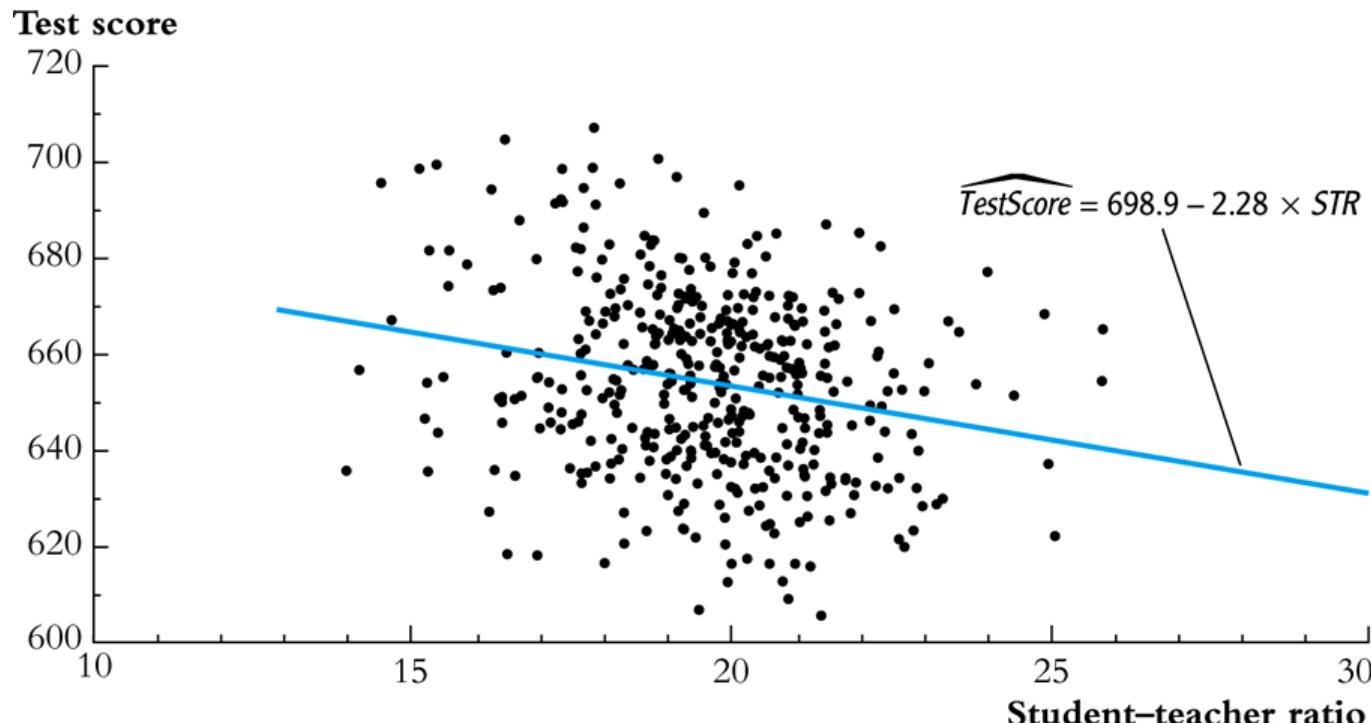
- The regression line from the sample is not the regression line from the population.
- What we want to do:
  - Assess how well the line describes the plot.
  - Guess the slope of the population line.
  - Guess what value Y would take for a given X value



## *Interpretation of the estimated slope and intercept*

- $\text{Test Score} = 698.9 - 2.28 \times \text{STR}$
- Districts with one more student per teacher on average have test scores that are 2.28 points lower.
- That is, 
$$\frac{\Delta \text{Test score}}{\Delta \text{STR}} = -2.28$$
- The intercept (taken literally) means that, according to this estimated line, districts with zero students per teacher would have a (predicted) test score of 698.9. But this interpretation of the intercept makes no sense – it extrapolates the line outside the range of the data – here, the intercept is not economically meaningful.

# Predicted values & residuals:



One of the districts in the data set is Antelope, CA, for which  $STR = 19.33$  and  $Test Score = 657.8$

predicted value:  $\hat{Y}_{Antelope} = 698.9 - 2.28 \times 19.33 = 654.8$

residual:  $\hat{u}_{Antelope} = 657.8 - 654.8 = 3.0$

# OLS regression: STATA output

```
regress testscr str, robust
```

Regression with robust standard errors

Number of obs = 420  
F( 1, 418) = 19.26  
Prob > F = 0.0000  
R-squared = 0.0512  
Root MSE = 18.581

	Robust					
testscr	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
str	-2.279808	.5194892	-4.39	0.000	-3.300945	-1.258671
_cons	698.933	10.36436	67.44	0.000	678.5602	719.3057

$$\text{Test Score} = 698.9 - 2.28 \times STR$$

# Measures of Fit

Two regression statistics provide complementary measures of how well the regression line “fits” or explains the data:

- The ***regression R<sup>2</sup>*** measures the fraction of the variance of  $Y$  that is explained by  $X$ ; it is unitless and ranges between zero (no fit) and one (perfect fit)
- The ***standard error of the regression (SER)*** measures the magnitude of a typical regression residual in the units of  $Y$ .

The regression  $R^2$  is the fraction of the sample variance of  $Y_i$  “explained” by the regression.

$$Y_i = \hat{Y}_i + \hat{u}_i = \text{OLS prediction} + \text{OLS residual}$$

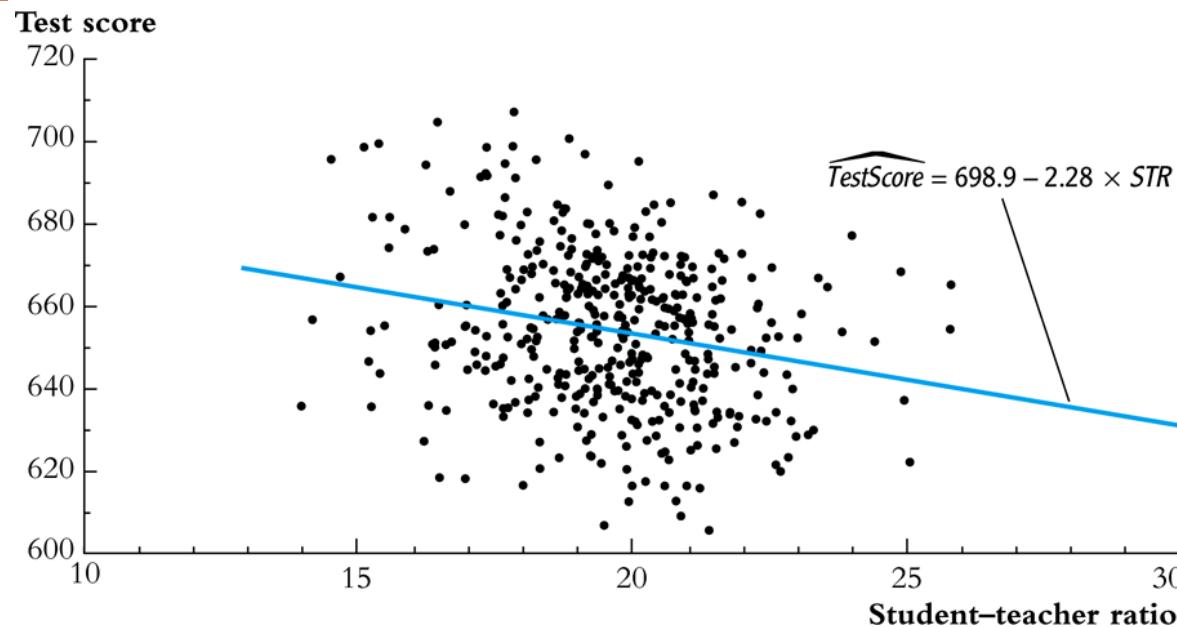
$$\rightarrow \text{sample var}(Y) = \text{sample var}(\hat{Y}_i) + \text{sample var}(\hat{u}_i)$$

$$\rightarrow \text{total sum of squares} = \text{“explained” SS} + \text{“residual” SS}$$

$$\text{Definition of } R^2: R^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{\hat{Y}})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

- $R^2 = 0$  means  $\text{ESS} = 0$
- $R^2 = 1$  means  $\text{ESS} = \text{TSS}$
- $0 \leq R^2 \leq 1$
- For regression with a single  $X$ ,  $R^2 = \text{the square of the correlation coefficient between } X \text{ and } Y$

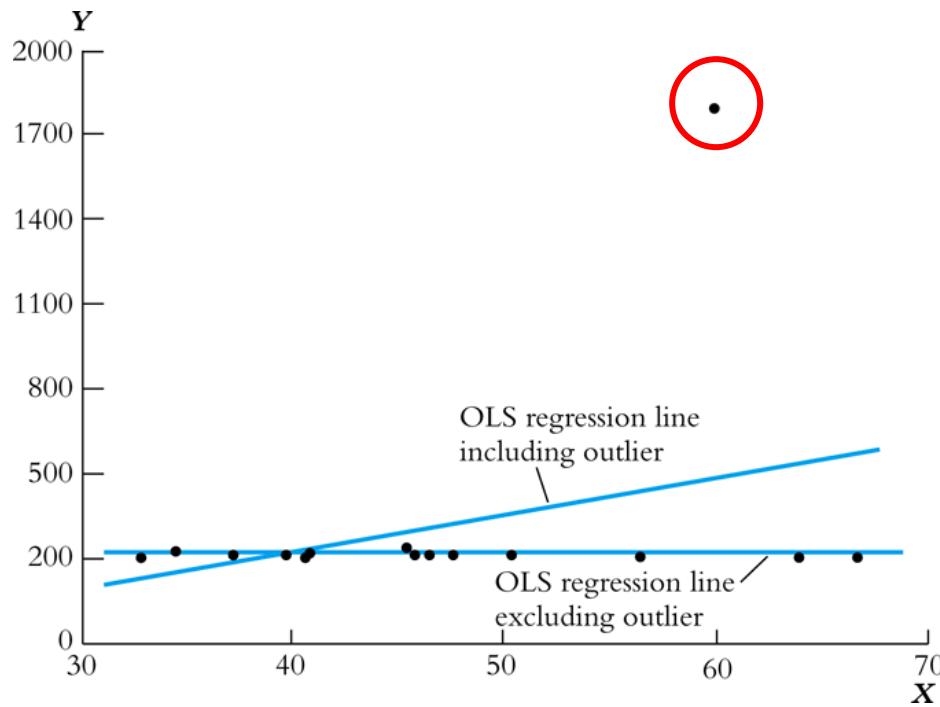
# Example of the $R^2$ and the $SER$



$\text{Test Score} = 698.9 - 2.28 \times \text{STR}$ ,  $R^2 = .05$ ,  $SER = 18.6$

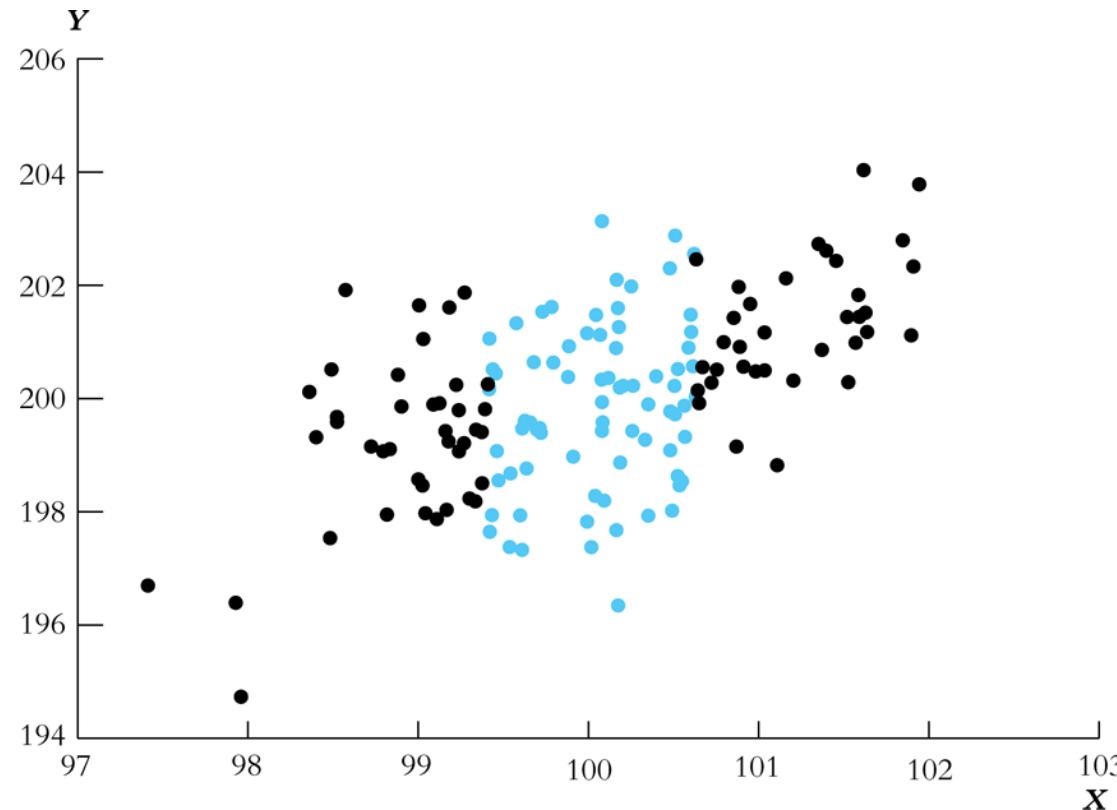
*STR explains only a small fraction of the variation in test scores. Does this make sense? Does this mean the STR is unimportant in a policy sense?*

# *OLS can be sensitive to an outlier:*



- *Is the lone point an outlier in X or Y?*
- In practice, outliers are often data glitches (coding or recording problems). Sometimes they are observations that really shouldn't be in your data set. Plot your data!

*The larger the variance of  $X$ , the smaller the variance of  $\hat{\beta}_1$*



The number of black and blue dots is the same. Using which would you get a more accurate regression line?

# Omitted Variable Bias

The error  $u$  arises because of factors, or variables, that influence  $Y$  but are not included in the regression function. There are always omitted variables.

Sometimes, the omission of those variables can lead to bias in the OLS estimator.

# *Omitted variable bias, ctd.*

The bias in the OLS estimator that occurs as a result of an omitted factor, or variable, is called ***omitted variable*** bias. For omitted variable bias to occur, the omitted variable “Z” must satisfy two conditions:

## The two conditions for omitted variable bias

1.  $Z$  is a determinant of  $Y$  (i.e.  $Z$  is part of  $u$ ); **and**
2.  $Z$  is correlated with the regressor  $X$  (i.e.  $\text{corr}(Z, X) \neq 0$ )

***Both conditions must hold for the omission of  $Z$  to result in omitted variable bias.***

# *Omitted variable bias, ctd.*

In the test score example:

1. English language ability (whether the student has English as a second language) plausibly affects standardized test scores:  $Z$  is a determinant of  $Y$ .
2. Immigrant communities tend to be less affluent and thus have smaller school budgets and higher  $STR$ :  $Z$  is correlated with  $X$ .

Accordingly,  $\hat{\beta}_1$  is biased. What is the direction of this bias?

- *What does common sense suggest?*
- If common sense fails you, there is a formula...

**TABLE 6.1**
**Differences in Test Scores for California School Districts with Low and High Student–Teacher Ratios, by the Percentage of English Learners in the District**

	<b>Student–Teacher Ratio &lt; 20</b>	<b>Student–Teacher Ratio ≥ 20</b>		<b>Difference in Test Scores, Low vs. High STR</b>		
	<b>Average Test Score</b>	<b>n</b>	<b>Average Test Score</b>	<b>n</b>	<b>Difference</b>	
					<b>t-statistic</b>	
All districts	657.4	238	650.0	182	7.4	4.04
Percentage of English learners						
< 1.9%	664.5	76	665.4	27	-0.9	-0.30
1.9–8.8%	665.2	64	661.8	44	3.3	1.13
8.8–23.0%	654.9	54	649.7	50	5.2	1.72
> 23.0%	636.7	44	634.8	61	1.9	0.68

- Districts with fewer English Learners have higher test scores
- Districts with lower percent *EL* (*PctEL*) have smaller classes
- Among districts with comparable *PctEL*, the effect of class size is small (recall overall “test score gap” = 7.4)

# *Return to omitted variable bias*

## **Three ways to overcome omitted variable bias**

1. Run a randomized controlled experiment in which treatment (*STR*) is randomly assigned: then *PctEL* is still a determinant of *TestScore*, but *PctEL* is uncorrelated with *STR*. (*This solution to OV bias is rarely feasible.*)
2. Adopt the “cross tabulation” approach, with finer gradations of *STR* and *PctEL* – within each group, all classes have the same *PctEL*, so we control for *PctEL* (*But soon you will run out of data, and what about other determinants like family income and parental education?*)
3. Use a regression in which the omitted variable (*PctEL*) is no longer omitted: include *PctEL* as an additional regressor in a multiple regression.

# The Population Multiple Regression Model

- Consider the case of two regressors:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \quad i = 1, \dots, n$$

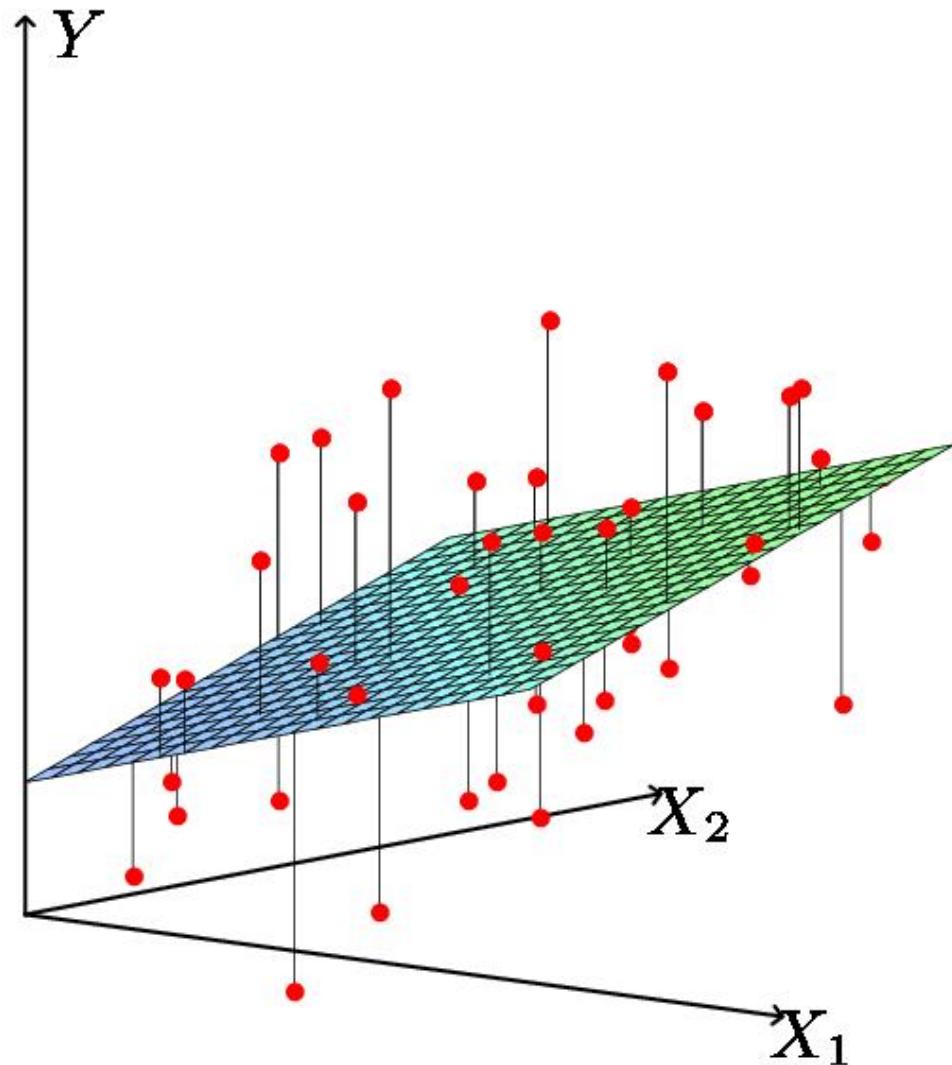
- $Y$  is the *dependent variable*
- $X_1, X_2$  are the two *independent variables (regressors)*
- $(Y_i, X_{1i}, X_{2i})$  denote the  $i^{\text{th}}$  observation on  $Y, X_1$ , and  $X_2$ .
- $\beta_0$  = unknown population intercept
- $\beta_1$  = effect on  $Y$  of a change in  $X_1$ , holding  $X_2$  constant
- $\beta_2$  = effect on  $Y$  of a change in  $X_2$ , holding  $X_1$  constant
- $u_i$  = the regression error (omitted factors)

# Least Squares Fit

- We estimate the parameters using least squares i.e. minimize

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

$$= \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{b}_0 - \hat{b}_1 X_1 - \dots - \hat{b}_p X_p)^2$$



# Interpretation of coefficients in multiple regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \quad i = 1, \dots, n$$

Consider changing  $X_1$  by  $\Delta X_1$  while holding  $X_2$  constant:

Population regression line **before** the change:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Population regression line, **after** the change:

$$Y + \Delta Y = \beta_0 + \beta_1(X_1 + \Delta X_1) + \beta_2 X_2$$

# The OLS Estimator in Multiple Regression

- With two regressors, the OLS estimator solves:

$$\min_{b_0, b_1, b_2} \sum_{i=1}^n [Y_i - (b_0 + b_1 X_{1i} + b_2 X_{2i})]^2$$

- The OLS estimator minimizes the average squared difference between the actual values of  $Y_i$  and the prediction (predicted value) based on the estimated line.
- This minimization problem is solved using calculus
- This yields the OLS estimators of  $\beta_0$  and  $\beta_1$  .**

# Example: the California test score data

Regression of *TestScore* against *STR*:

$$\text{Test Score} = 698.9 - 2.28 \times \text{STR}$$

Now include percent English Learners in the district  
(*PctEL*):

$$\text{Test Score} = 686.0 - 1.10 \times \text{STR} - 0.65 \text{PctEL}$$

- What happens to the coefficient on *STR*?

# Multiple regression in STATA

```
reg testscr str pctel, robust;
```

```
Regression with robust standard errors  
Number of obs = 420  
F( 2, 417) = 223.82  
Prob > F = 0.0000  
R-squared = 0.4264  
Root MSE = 14.464
```

---

		Robust				
	testscr	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
<hr/>						
	str	-1.101296	.4328472	-2.54	0.011	-1.95213 -.2504616
	pctel	-.6497768	.0310318	-20.94	0.000	-.710775 -.5887786
	_cons	686.0322	8.728224	78.60	0.000	668.8754 703.189
<hr/>						

---

$$\text{Test Score} = 686.0 - 1.10 \times \text{STR} - 0.65 \times \text{PctEL}$$

# $R^2$ and adjusted $R^2$

The  $R^2$  is the fraction of the variance explained – same definition as in regression with a single regressor:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS},$$

where  $ESS = \sum_{i=1}^n (\hat{Y}_i - \bar{\hat{Y}})^2$ ,  $SSR = \sum_{i=1}^n \hat{u}_i^2$ ,  $TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2$

- The  $R^2$  always increases when you add another regressor (*why?*) – a bit of a problem for a measure of “fit”

# $R^2$ ctd.

The  $R^2$  and adjusted  $R^2$  tell you whether the regressors are good at predicting the values of the dependent variable. If the  $R^2$  is nearly 1, then the regressors produce good predictions. If the  $R^2$  is nearly 0, the opposite is true.

The  $R^2$  and adjusted  $R^2$  do **NOT** tell you whether:

- 1. An included variable is statistically significant,
- 2. The regressors are a true cause of the movements in the dependent variable,
- 3. There is omitted variable bias, or
- 4. You have chosen the most appropriate set of regressors

# Some Relevant Questions

1. Is  $\beta_j=0$  or not? We can use a hypothesis test to answer this question. If we can't be sure that  $\beta_j \neq 0$  then there is no point in using  $X_j$  as one of our predictors.
1. Can we be sure that at least one of our  $X$  variables is a useful predictor i.e. is it the case that  $\beta_1 = \beta_2 = \dots = \beta_p = 0$ ?

# 1. Is $\beta_j=0$ i.e. is $X_j$ an important variable?

- We use a hypothesis test to answer this question
- $H_0: \beta_j=0$  vs  $H_a: \beta_j \neq 0$
- Calculate 
$$t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$
 ← Number of standard deviations away from zero.
- If  $t$  is large (equivalently p-value is small) we can be sure that  $\beta_j \neq 0$  and that there is a relationship

**Regression coefficients**

	Coefficient	Std Err	t-value	p-value
Constant	7.0326	0.4578	15.3603	0.0000
TV	0.0475	0.0027	17.6676	0.0000

$\hat{\beta}_1$  is 17.67 SE's from 0 → P-value

# Testing Individual Variables

Is there a (statistically detectable) linear relationship between Newspapers and Sales after all the other variables have been accounted for?

## *Regression coefficients*

	Coefficient	Std Err	t-value	p-value
Constant	2.9389	0.3119	9.4223	0.0000
TV	0.0458	0.0014	32.8086	0.0000
Radio	0.1885	0.0086	21.8935	0.0000
Newspaper	-0.0010	0.0059	-0.1767	0.8599

No: big p-value

## *Regression coefficients*

	Coefficient	Std Err	t-value	p-value
Constant	12.3514	0.6214	19.8761	0.0000
Newspaper	0.0547	0.0166	3.2996	0.0011

Small p-value in simple regression

Almost all the explaining that Newspapers could do in simple regression has already been done by TV and Radio in multiple regression!

## 2. Is the whole regression explaining anything at all?

➤ Test for:

- $H_0$ : all slopes = 0 ( $\beta_1=\beta_2=\cdots=\beta_p=0$ ),
- $H_a$ : at least one slope  $\neq 0$

**ANOVA Table**

Source	df	SS	MS	F	p-value
Explained	2	4860.2347	2430.1174	859.6177	0.0000
Unexplained	197	556.9140	2.8270		

Answer comes from the F test in the ANOVA (ANalysis Of VAriance) table.

The ANOVA table has many pieces of information. What we care about is the F Ratio and the corresponding p-value.

# Regression when $X$ is Binary

Sometimes a regressor is binary:

- $X = 1$  if small class size,  $= 0$  if not
- $X = 1$  if female,  $= 0$  if male
- $X = 1$  if treated (experimental drug),  $= 0$  if not

Binary regressors are sometimes called “dummy” variables.

So far,  $\beta_1$  has been called a “slope,” but that doesn’t make sense if  $X$  is binary.

How do we interpret regression with a binary regressor?

# Qualitative Predictors

- How do you stick “men” and “women” (category listings) into a regression equation?
- Code them as indicator variables (dummy variables)
- For example we can “code” Males=0 and Females= 1.

# Interpretation

- Suppose we want to include income and gender.
- Two genders (male and female). Let

$$\text{Gender}_i = \begin{cases} 0 & \text{if male} \\ 1 & \text{if female} \end{cases}$$

- then the regression equation is

$$Y_i \gg b_0 + b_1 \text{Income}_i + b_2 \text{Gender}_i = \begin{cases} b_0 + b_1 \text{Income}_i & \text{if male} \\ b_0 + b_1 \text{Income}_i + b_2 & \text{if female} \end{cases}$$

- $\beta_2$  is the average extra balance each month that females have for given income level. Males are the “baseline”.

## *Regression coefficients*

	Coefficient	Std Err	t-value	p-value
Constant	233.7663	39.5322	5.9133	0.0000
Income	0.0061	0.0006	10.4372	0.0000
Gender_Female	24.3108	40.8470	0.5952	0.5521

# Other Coding Schemes

- There are different ways to code categorical variables.
- Two genders (male and female). Let

$$\text{Gender}_i = \begin{cases} -1 & \text{if male} \\ 1 & \text{if female} \end{cases}$$

- then the regression equation is

$$Y_i \rightarrow b_0 + b_1 \text{Income}_i + b_2 \text{Gender}_i = \begin{cases} b_0 + b_1 \text{Income}_i - b_2, & \text{if male} \\ b_0 + b_1 \text{Income}_i + b_2, & \text{if female} \end{cases}$$

- $\beta_2$  is the average amount that females are above the average, for any given income level.  $\beta_2$  is also the average amount that males are below the average, for any given income level.

## Interpreting regressions with a binary regressor

$Y_i = \beta_0 + \beta_1 X_i + u_i$ , where  $X$  is binary ( $X_i = 0$  or  $1$ ):

When  $X_i = 0$ ,  $Y_i = \beta_0 + u_i$

- the mean of  $Y_i$  is  $\beta_0$
- that is,  $E(Y_i|X_i=0) = \beta_0$

When  $X_i = 1$ ,  $Y_i = \beta_0 + \beta_1 + u_i$

- the mean of  $Y_i$  is  $\beta_0 + \beta_1$
- that is,  $E(Y_i|X_i=1) = \beta_0 + \beta_1$

so:

$$\begin{aligned}\beta_1 &= E(Y_i|X_i=1) - E(Y_i|X_i=0) \\ &= \text{population difference in group means}\end{aligned}$$

**Example:** Let  $D_i = \begin{cases} 1 & \text{if } STR_i \leq 20 \\ 0 & \text{if } STR_i > 20 \end{cases}$

**OLS regression:**  $\text{Test Score} = 650.0 + 7.4 \times D$   
 $(1.3) \quad (1.8)$

**Tabulation of group means:**

Class Size	Average score ( $\bar{Y}$ )	Std. dev. ( $s_Y$ )	$N$
Small ( $STR > 20$ )	657.4	19.4	238
Large ( $STR \geq 20$ )	650.0	17.9	182

**Difference in means:**  $\bar{Y}_{\text{small}} - \bar{Y}_{\text{large}} = 657.4 - 650.0 = 7.4$

**Standard error**  $SE = \sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}} = \sqrt{\frac{19.4^2}{238} + \frac{17.9^2}{182}} = 1.8$

# Multicollinearity, Perfect and Imperfect

**Perfect multicollinearity** is when one of the regressors is an exact linear function of the other regressors.

Some more examples of perfect multicollinearity

1. The example from before: you include  $STR$  twice,
2. Regress  $TestScore$  on a constant,  $D$ , and  $B$ , where:  $D_i = 1$  if  $STR \leq 20$ ,  $= 0$  otherwise;  $B_i = 1$  if  $STR > 20$ ,  $= 0$  otherwise, so  $B_i = 1 - D_i$  and there is perfect multicollinearity.
3. Would there be perfect multicollinearity if the intercept (constant) were excluded from this regression?

# The dummy variable trap

Suppose you have a set of multiple binary (dummy) variables, which are mutually exclusive and exhaustive – that is, there are multiple categories and every observation falls in one and only one category (Freshmen, Sophomores, Juniors, Seniors, Other). If you include all these dummy variables *and* a constant, you will have perfect multicollinearity – this is sometimes called ***the dummy variable trap***.

- *Why is there perfect multicollinearity here?*
- *Solutions to the dummy variable trap:*
  1. Omit one of the groups (e.g. Senior), or
  2. Omit the intercept
- *What are the implications of (1) or (2) for the interpretation of the coefficients?*

# *Perfect multicollinearity, ctd.*

- Perfect multicollinearity usually reflects a mistake in the definitions of the regressors, or an oddity in the data
- If you have perfect multicollinearity, your statistical software will let you know – either by crashing or giving an error message or by “dropping” one of the variables arbitrarily
- The solution to perfect multicollinearity is to modify your list of regressors so that you no longer have perfect multicollinearity.

Assumption #4: There is no perfect multicollinearity

*Perfect multicollinearity* is when one of the regressors is an exact linear function of the other regressors.

**Example:** Suppose you accidentally include *STR* twice:

```
regress testscr str str, robust
```

## Regression with robust standard errors

Number of obs = 420

$$F(-1, 418) = 19.26$$

Prob > F = 0.0000

R-squared = 0.0512

Root MSE = 18.581

	Robust					
testscr	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
str	-2.279808	.5194892	-4.39	0.000	-3.300945	-1.258671
str	(dropped)					
_cons	698.933	10.36436	67.44	0.000	678.5602	719.3057

# *Imperfect multicollinearity*

Imperfect and perfect multicollinearity are quite different despite the similarity of the names.

***Imperfect multicollinearity*** occurs when two or more regressors are very highly correlated.

- Why the term “multicollinearity”? If two regressors are very highly correlated, then their scatterplot will pretty much look like a straight line – they are “co-linear” – but unless the correlation is exactly  $\pm 1$ , that collinearity is imperfect.

## *Imperfect multicollinearity, ctd.*

Imperfect multicollinearity implies that one or more of the regression coefficients will be imprecisely estimated.

- The idea: the coefficient on  $X_1$  is the effect of  $X_1$  holding  $X_2$  constant; but if  $X_1$  and  $X_2$  are highly correlated, there is very little variation in  $X_1$  once  $X_2$  is held constant – so the data don't contain much information about what happens when  $X_1$  changes but  $X_2$  doesn't. If so, the variance of the OLS estimator of the coefficient on  $X_1$  will be large.
- Imperfect multicollinearity (correctly) results in large standard errors for one or more of the OLS coefficients.

# Interaction

- When the effect on  $Y$  of increasing  $X_1$  depends on another  $X_2$ .
- Example:
  - Maybe the effect on Salary ( $Y$ ) when increasing Position ( $X_1$ ) depends on gender ( $X_2$ )?
  - For example maybe Male salaries go up faster (or slower) than Females as they get promoted.
- Advertising example:
  - TV and radio advertising both increase sales.
  - Perhaps spending money on both of them may increase sales more than spending the same amount on one alone?

# Interaction in advertising

$$Sales = b_0 + b_1 \cdot TV + b_2 \cdot Radio + b_3 \cdot TV \cdot Radio$$

$$Sales = b_0 + (b_1 + b_3 \cdot Radio) \cdot TV + b_2 \cdot Radio$$

- Spending \$1 extra on TV increases average sales by  $0.0191 + 0.0011\text{Radio}$



Interaction Term

$$Sales = b_0 + (b_2 + b_3 \cdot TV) \cdot Radio + b_2 \cdot TV$$

- Spending \$1 extra on Radio increases average sales by  $0.0289 + 0.0011\text{TV}$

## Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	6.7502202	0.247871	27.23	<.0001*
TV	0.0191011	0.001504	12.70	<.0001*
Radio	0.0288603	0.008905	3.24	0.0014*
TV*Radio	0.0010865	5.242e-5	20.73	<.0001*

# Parallel Regression Lines

## Expanded Estimates

Nominal factors expanded to all levels

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	112.77039	1.454773	77.52	<.0001
Gender[female]	1.8600957	0.527424	3.53	0.0005
Gender[male]	-1.860096	0.527424	-3.53	0.0005
Position	6.0553559	0.280318	21.60	<.0001

## Regression equation

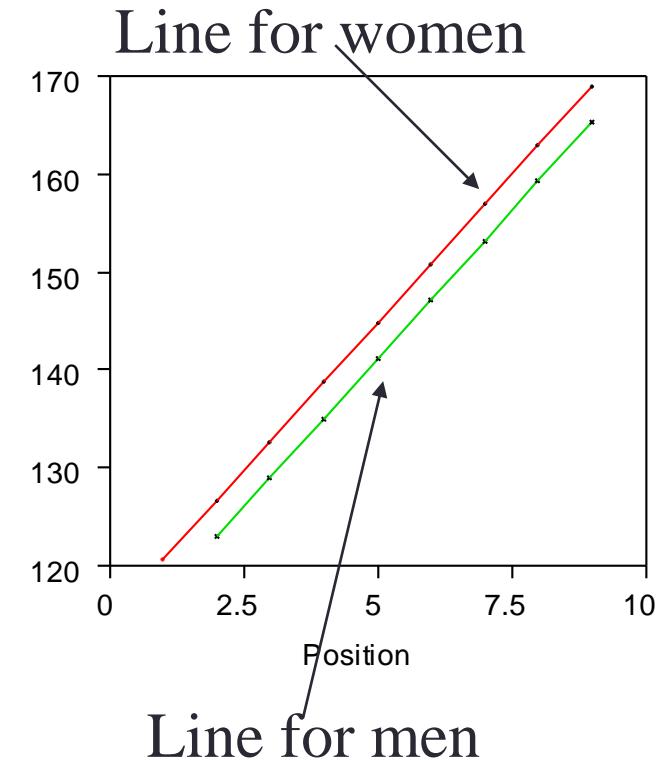
$$\text{female: salary} = 112.77 + 1.86 + 6.05 \times \text{position}$$

$$\text{males: salary} = 112.77 - 1.86 + 6.05 \times \text{position}$$

Different  
intercepts

Same  
slopes

Parallel lines have the same slope.  
Dummy variables give lines different intercepts, but their slopes are still the same.



# Interaction Effects

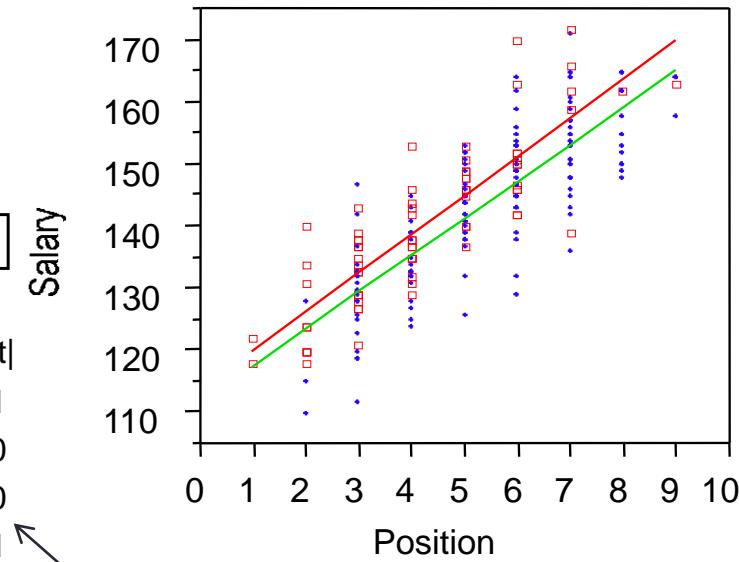
- Our model has forced the line for men and the line for women to be parallel.
- Parallel lines say that promotions have the same salary benefit for men as for women.
- If lines aren't parallel then promotions affect men's and women's salaries differently.

# Should the Lines be Parallel?

## Expanded Estimates

Nominal factors expanded to all levels

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	112.63081	1.484825	75.85	<.0001
Gender[female]	1.1792165	1.484825	0.79	0.4280
Gender[male]	-1.179216	1.484825	-0.79	0.4280
Position	6.1021378	0.296554	20.58	<.0001
Gender[female]*Position	0.1455111	0.296554	0.49	0.6242
Gender[male]*Position	-0.145511	0.296554	-0.49	0.6242



Interaction is not significant

Interaction between gender and position

# RIDGE AND LASSO REGRESSION

---

# Improving on the Least Squares Regression Estimates?

- We want to improve the Linear Regression model, by replacing the least square fitting with some alternative fitting procedure, i.e., the values that minimize the mean square error (MSE)
- There are 2 reasons we might not prefer to just use the ordinary least squares (OLS) estimates
  1. Prediction Accuracy
  2. Model Interpretability

# 1. Prediction Accuracy

- The least squares estimates have relatively low bias and low variability especially when the relationship between Y and X is linear and the number of observations n is way bigger than the number of predictors p
- But, when  $n=p$  (almost), then the least squares fit can have high variance and may result in over fitting and poor estimates on unseen observations,
- And, when  $n < p$ , then the variability of the least squares fit increases dramatically, and the variance of these estimates is infinite

# Why can shrinking towards zero be a good thing to do?

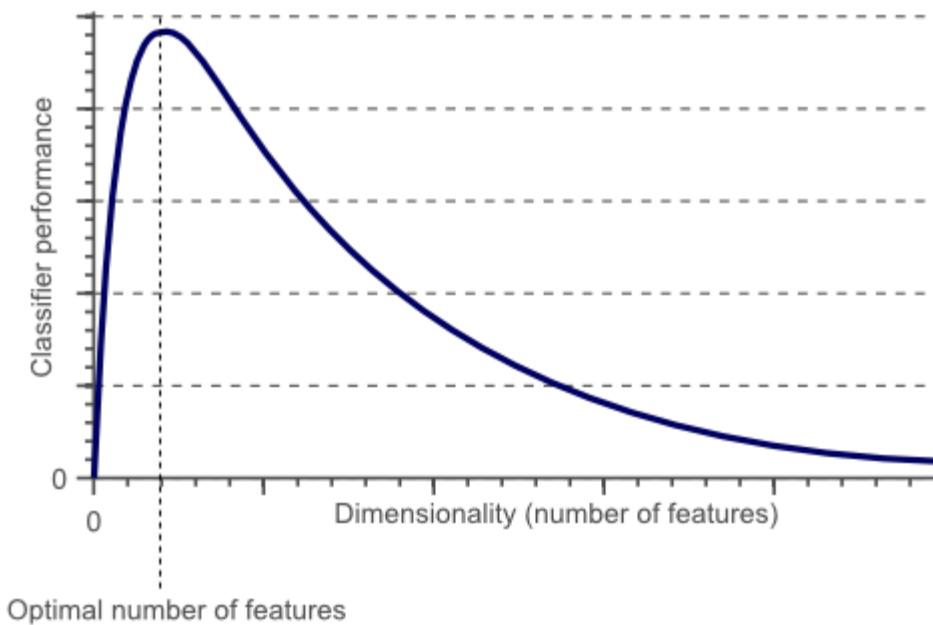
- It turns out that the OLS estimates generally have low bias but can be highly variable. In particular when  $n$  and  $p$  are of similar size or when  $n < p$ , then the OLS estimates will be extremely variable
- The penalty term makes the ridge regression estimates biased but can also substantially reduce variance
- Thus, there is a bias/variance trade-off

# Curse of Dimensionality

- In most applications we observe a high dimensional data set
- It is inadvisable to use all the features
  - Redundant
  - Model complexity ~ tendency to overfit
  - Computational difficulty
  - Correlation
  - Anything else ?

# Curse of Dimensionality

- Hughes Phenomenon: As the number of features increases, the classifier's performance increases until the optimal number of features. Adding more features based on the same size as the training set will then degrade the classifier's performance.



# Bias and Variance Tradeoff

- In general, good estimators should, on average have, small prediction errors
- As model becomes more complex (more terms included), local structure/curvature can be picked up
- But coefficient estimates suffer from high variance as more terms are included in the model
- Therefore, introducing a little bias in our estimate for  $\beta$  might lead to a substantial decrease in variance, and hence to a substantial decrease in prediction error

# Bias and Variance Tradeoff

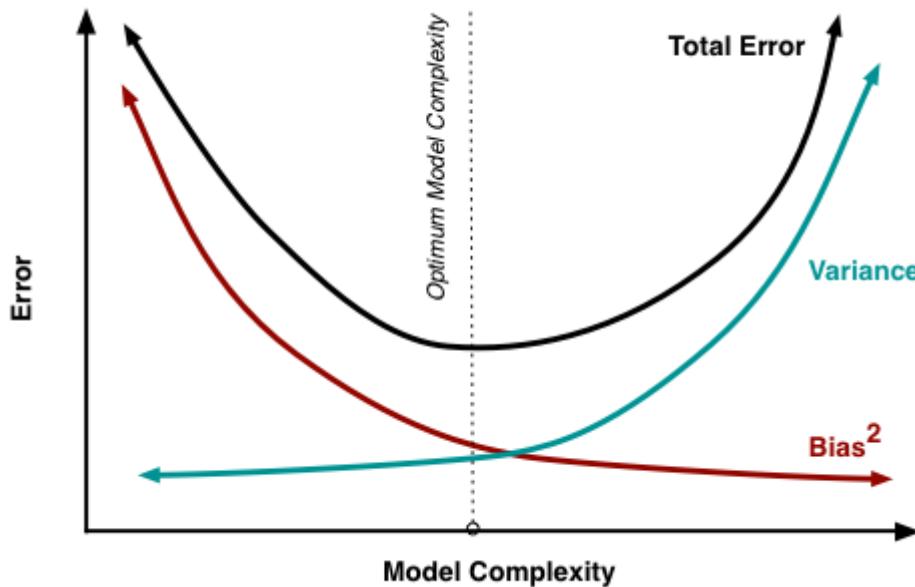
- Suppose we use a simple polynomial regression. As we increase the degree of the polynomial function, we observe



- Underfitting ~ high bias and low variance
- Overfitting ~ low bias and high variance

# Bias and Variance Tradeoff

- We need to determine the optimum point in model complexity to avoid both underfitting and overfitting



## 2. Model Interpretability

- When we have a large number of variables  $X$  in the model there will generally be many that have little or no effect on  $Y$
- Leaving these variables in the model makes it harder to see the “big picture”, i.e., the effect of the “important variables”
- The model would be easier to interpret by removing (i.e. setting the coefficients to zero) the unimportant variables

# Solution

- Subset Selection
  - Identifying a subset of all  $p$  predictors  $X$  that we believe to be related to the response  $Y$ , and then fitting the model using this subset
  - E.g. best subset selection and stepwise selection
- Shrinkage
  - Involves shrinking the estimates coefficients towards zero
  - This shrinkage reduces the variance
  - Some of the coefficients may shrink to exactly zero, and hence shrinkage methods can also perform variable selection
  - E.g. Ridge regression and the Lasso
- Dimension Reduction
  - Involves projecting all  $p$  predictors into an  $M$ -dimensional space where  $M < p$ , and then fitting linear regression model
  - E.g. Principle Components Regression

# Feature Selection

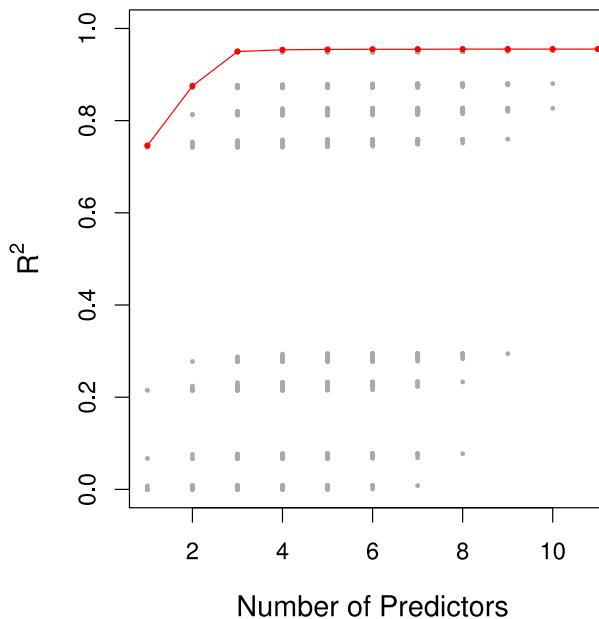
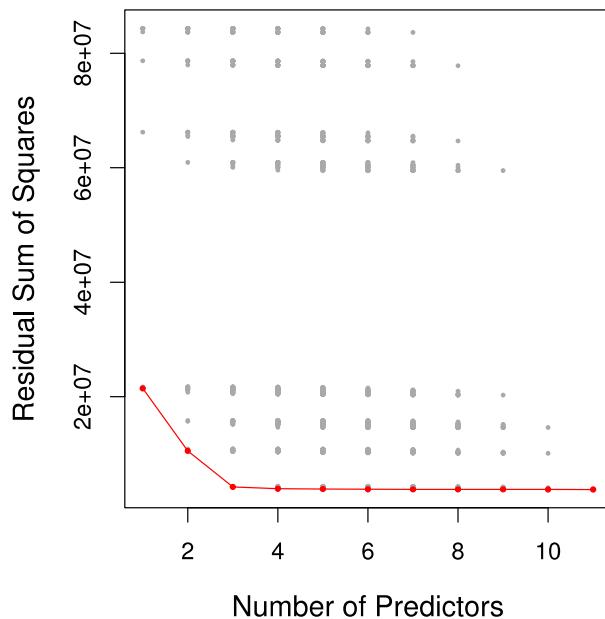
- Business (domain) knowledge
  - Select the features that are relevant and important
  - Minimize the correlation among the selected features
- Automated selection
  - Forward selection: Start with most significant predictor in the model and adds features in each step.
  - Backward elimination: Starts with all predictors in the model and removes the least significant feature in each step.
- Any reason not to use either method?

# Best Subset Selection

- In this approach, we run a linear regression for each possible combination of the X predictors
- How do we judge which subset is the “best”?
- One simple approach is to take the subset with the smallest RSS or the largest  $R^2$
- Unfortunately, one can show that the model that includes all the variables will always have the largest  $R^2$  (and smallest RSS)

# Credit Data: $R^2$ vs. Subset Size

- The RSS/ $R^2$  will always decline/increase as the number of variables increase so they are not very useful
- The red line tracks the best model for a given number of predictors, according to RSS and  $R^2$

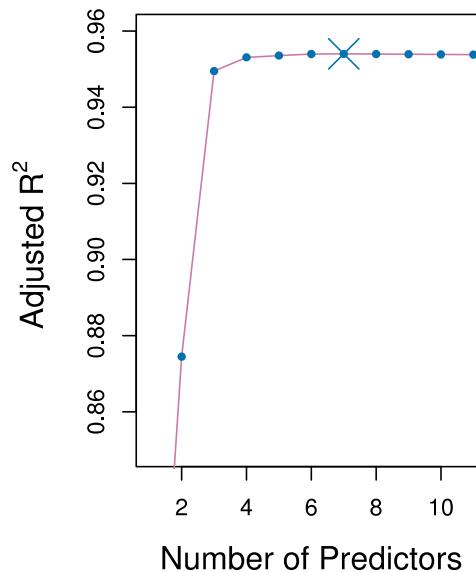
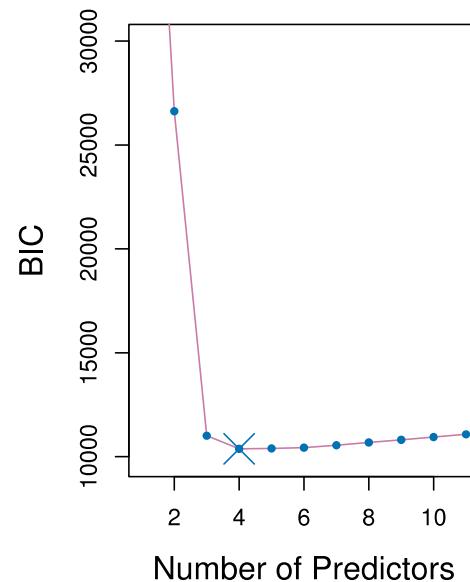
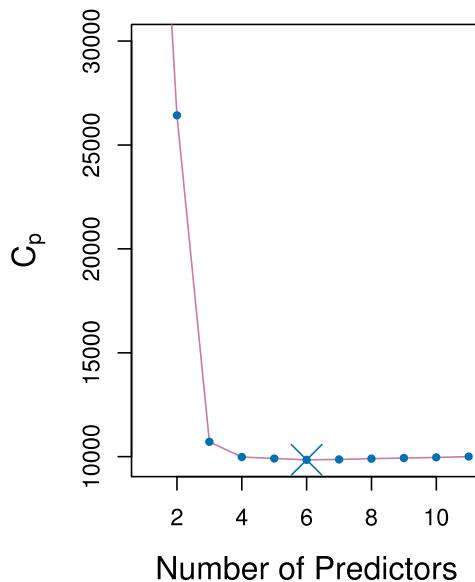


# Other Measures of Comparison

- To compare different models, we can use other approaches:
  - Adjusted  $R^2$
  - AIC (Akaike information criterion)
  - BIC (Bayesian information criterion)
  - $C_p$  (equivalent to AIC for linear regression)
- These methods add penalty to RSS for the number of variables (i.e. complexity) in the model
- None are perfect

# Credit Data: $C_p$ , BIC, and Adjusted $R^2$

- A small value of  $C_p$  and BIC indicates a low error, and thus a better model
- A large value for the Adjusted  $R^2$  indicates a better model



# Stepwise Selection

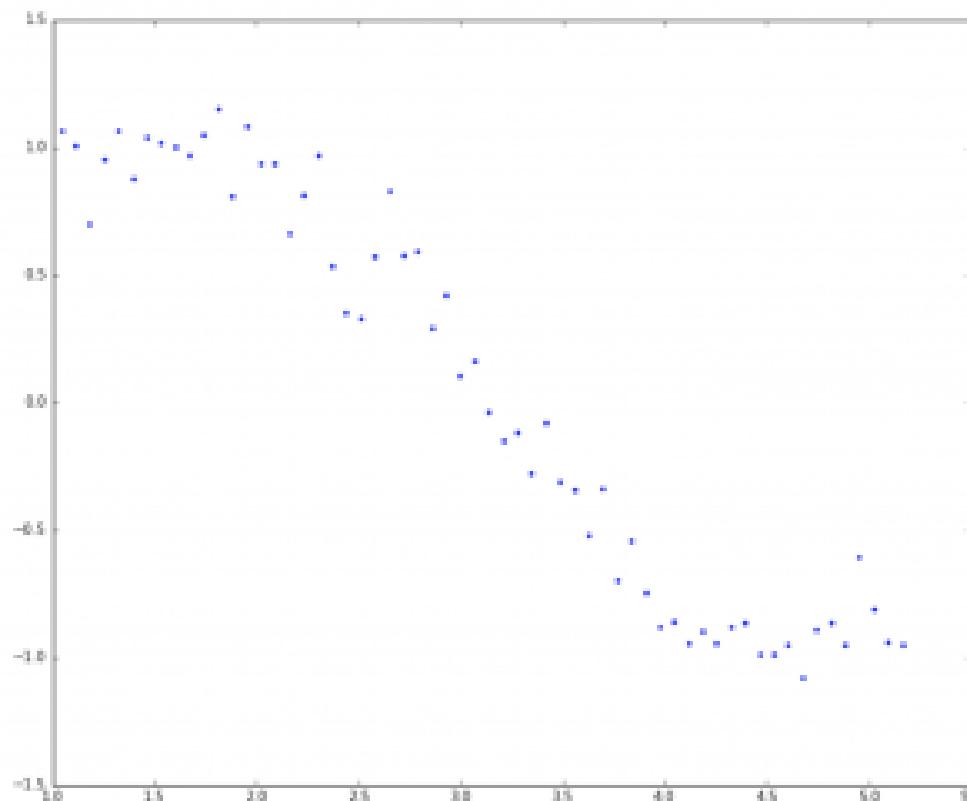
- Best Subset Selection is computationally intensive especially when we have a large number of predictors (large  $p$ )
- More attractive methods:
  - Forward Stepwise Selection: Begins with the model containing no predictor, and then adds one predictor at a time that improves the model the most until no further improvement is possible
  - Backward Stepwise Selection: Begins with the model containing all predictors, and then deleting one predictor at a time that improves the model the most until no further improvement is possible

# Regularization

- To overcome overfitting, we should either reduce the complexity of the model or use regularization
- In regularization we reduce the magnitude of the coefficients by penalizing them
- Because if the  $\beta_j$ 's are unconstrained
  - They can explode
  - Hence are susceptible to very high variance
- To control variance, we might regularize the coefficients
  - Control how large the coefficients grow

# Motivation for Penalizing Coefficients

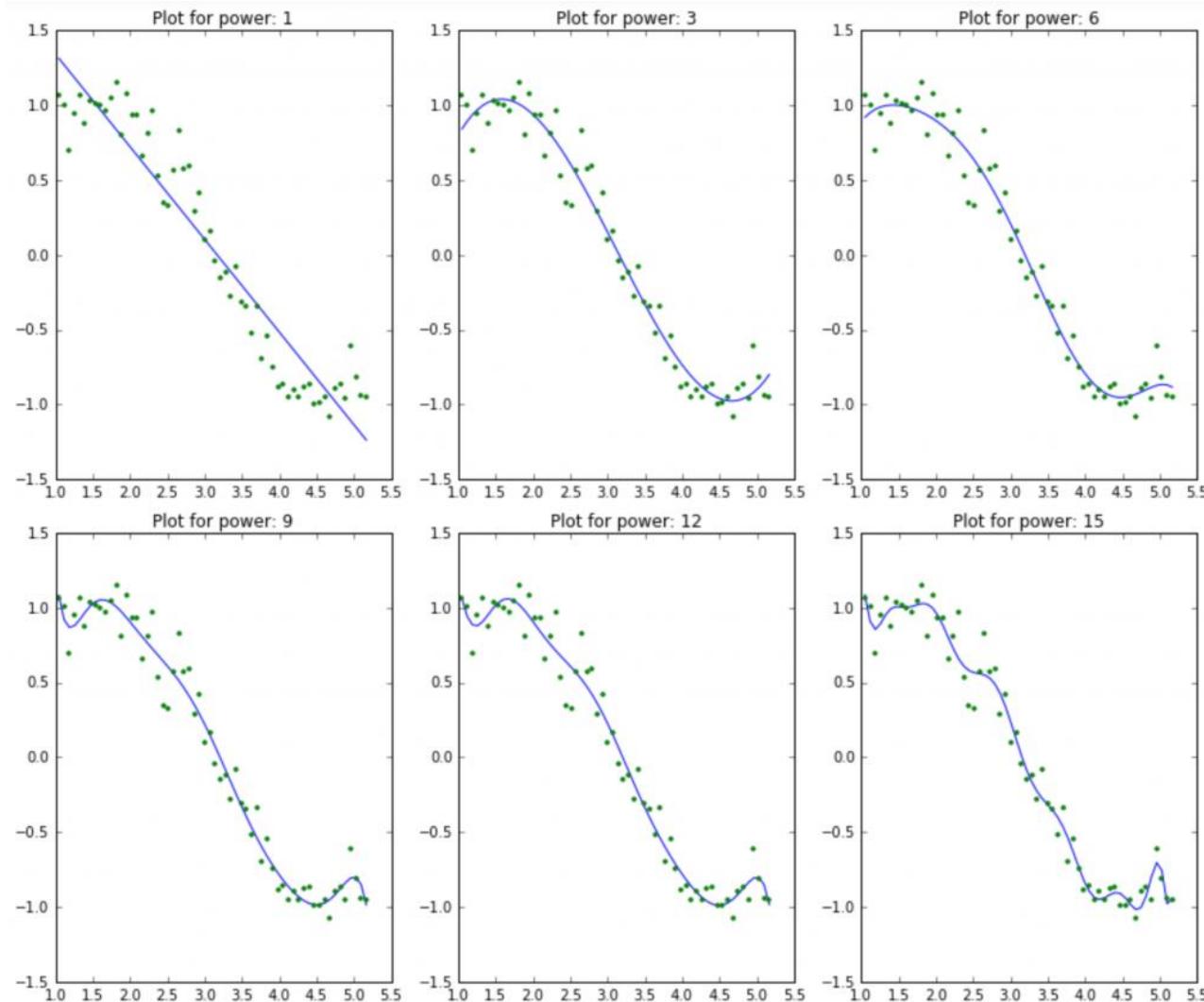
- Plot a sine curve with random noise



# Motivation for Penalizing Coefficients

- Estimate the sine function using **polynomial regression** with powers of  $x$  up to 15.
- We will have 15 different polynomial regression starting with linear to polynomial with degree 15.

# Motivation for Penalizing Coefficients



# Motivation for Penalizing Coefficients

- The sizes of coefficients increase exponentially with increase in model complexity.
- Check the Table Poly in Lecture3a.xlsx
- Large coefficient ~ high emphasis on feature.
- Too large coefficient ~ algorithm starts overfitting

# Ridge Regression

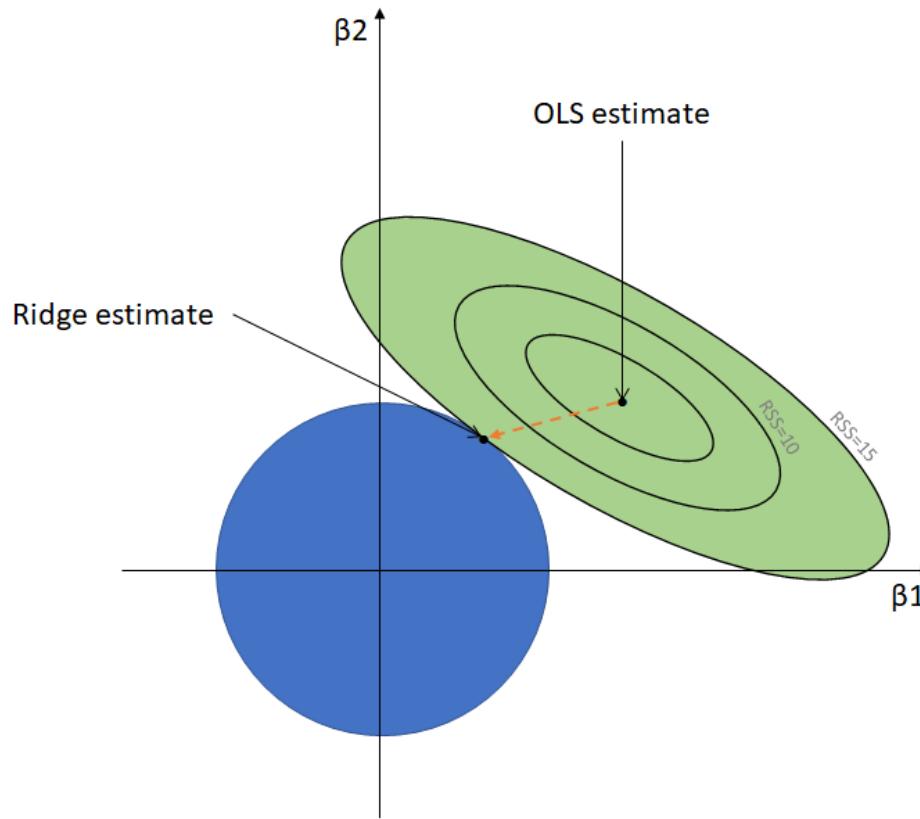
- Performs L2 regularization, i.e. adds penalty equivalent to **square of the magnitude** of coefficients
- Minimization objective = LS Obj +  $\lambda * (\text{sum of square of coefficients})$
- In ridge constraint form
  - minimize  $\sum(y_i - \beta^\top z_i)^2$
  - s.t.  $\sum \beta_j^2 \leq t$
- In function form
  - minimize  $\sum(y_i - \beta^\top z_i)^2 + \lambda (\sum \beta_j^2 - t)$

# Penalty Coefficient

- $\lambda$  (or alpha parameter extra term) is the penalty term in the ridge function.
- Changing the values of  $\lambda$ : controlling the importance on the penalty term/cost
- Higher the values of  $\lambda$  leads to bigger total penalty cost and as a result the magnitude of coefficients are reduced.
- It shrinks the parameters, therefore it is mostly used to prevent multi-collinearity.
- It reduces the model complexity by coefficient shrinkage.

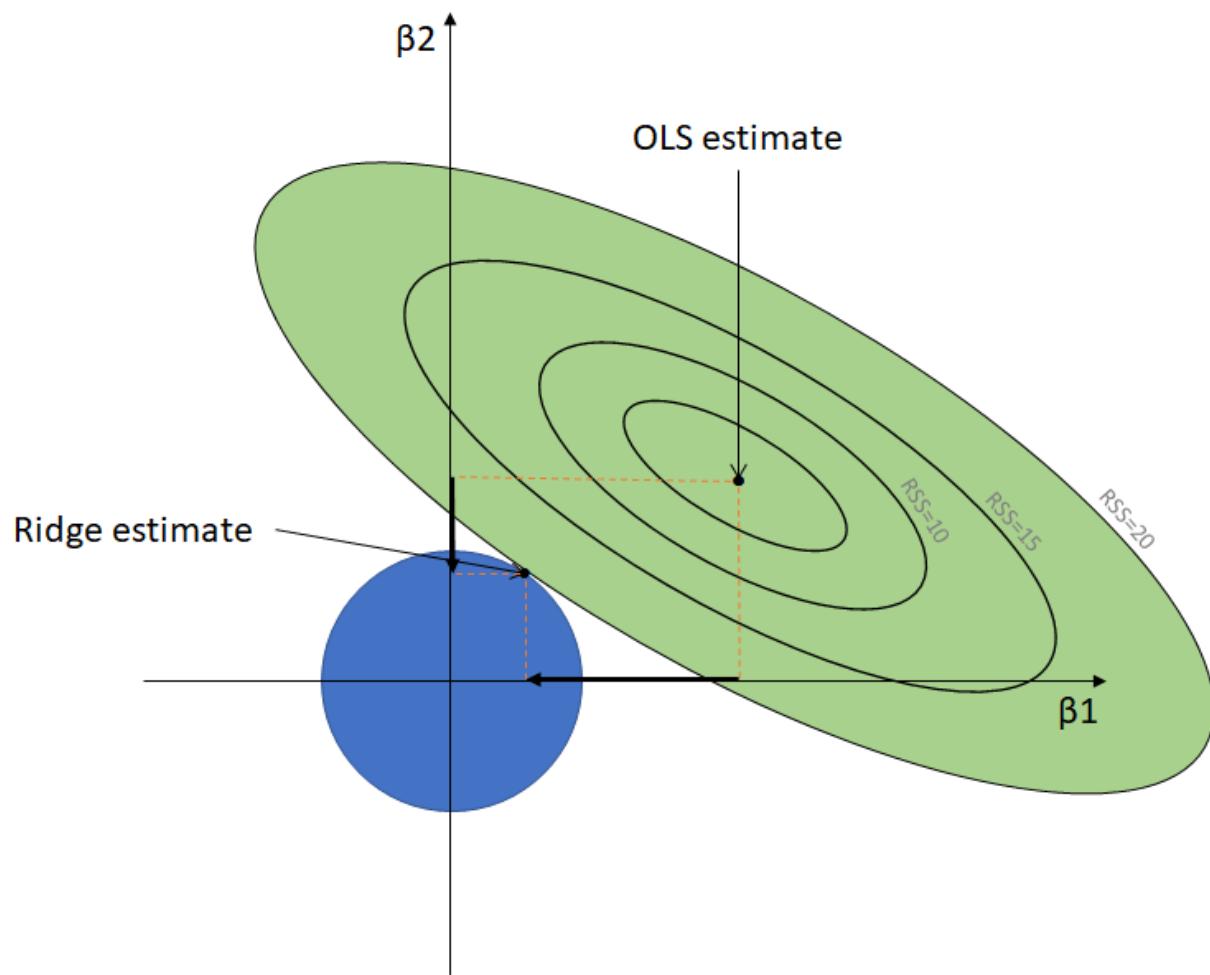
# Ridge Regression vs OLS

- “Iso-RSS” lines



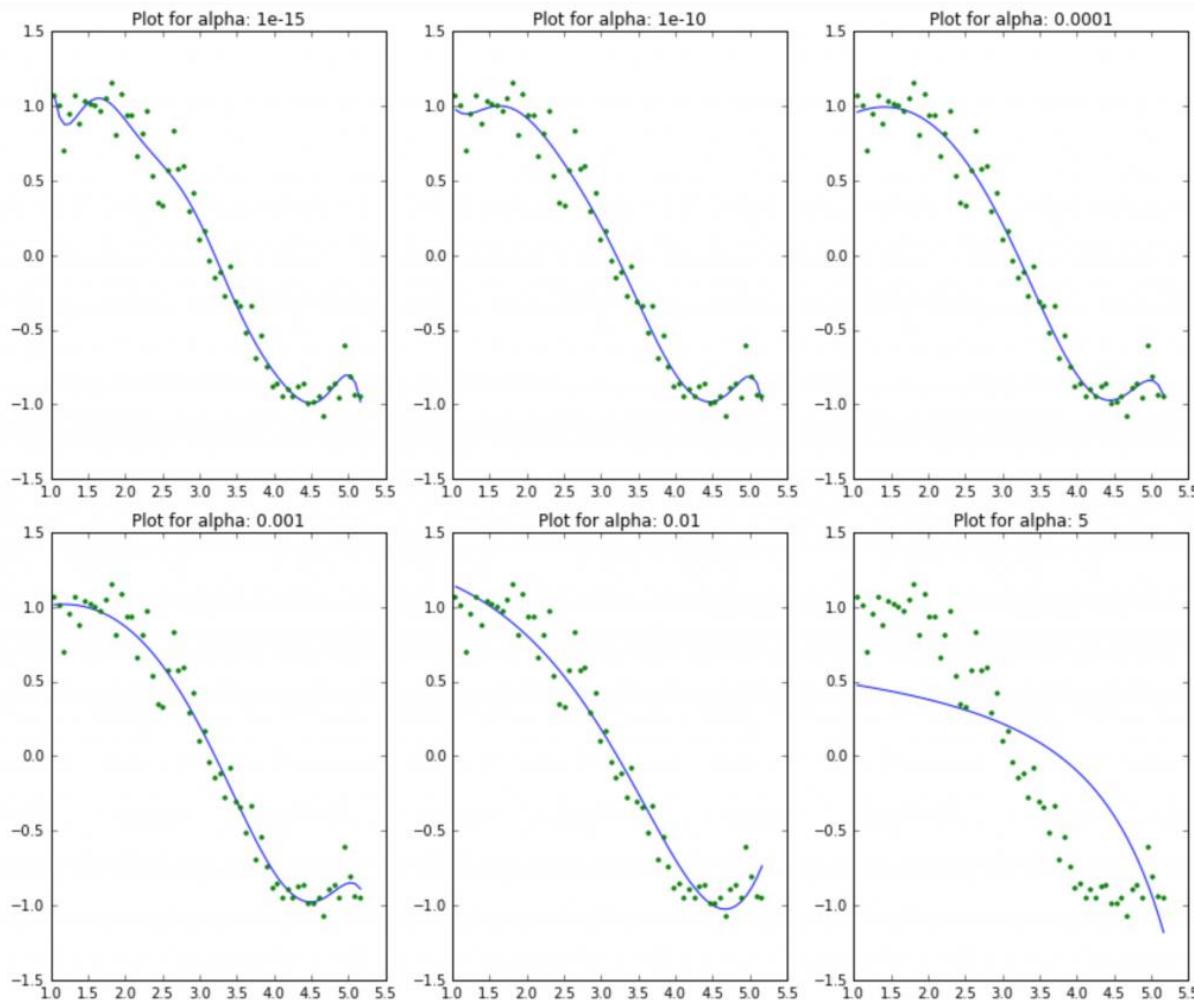
# Ridge Regression vs OLS

- “Iso-RSS” lines



# Ridge Regression

- Use different  $\lambda$  values for the sine function prediction

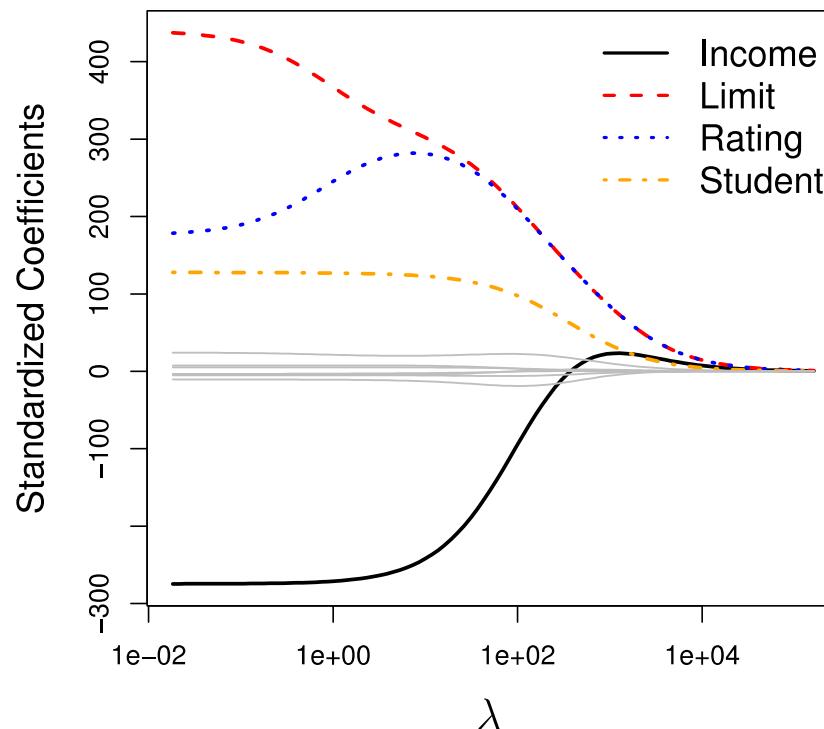


# Ridge Regression

- Check the Table Ridge in Lecture3a.xlsx
- The RSS increases with increase in  $\lambda$ , this model complexity reduces
- An  $\lambda$  as small as  $1e-15$  gives us significant reduction in magnitude of coefficients.
- High  $\lambda$  values can lead to significant underfitting. Note the rapid increase in RSS for values of  $\lambda$  greater than 1
- Though the coefficients are **very small**, they are **NOT zero**.

# Credit Data: Ridge Regression

- As  $\lambda$  increases, the standardized coefficients shrinks towards zero.



# Choosing $\lambda$

- As  $\lambda$  increases, the model complexity reduces.
- Though higher values of  $\lambda$  reduce overfitting, significantly high values can cause underfitting as well
- Thus  $\lambda$  should be chosen wisely.
- Obviously want to choose  $\lambda$  that minimizes the mean squared error
- Standard practice is to use cross-validation
  - the value of alpha is iterated over a range of values and the one giving higher cross-validation score is chosen.

# My Great Great Grand Advisor 😊

Orsan Ozener

Ozlem Ergun

James Orlin

Arthur Veinott

Cyrus Derman

Herbert Robbins

George Birkhoff

E. H. Moore

H. A. Newton

Michel Chasles

Simeon Poisson

Joseph Lagrange

Leonhard Euler

Johann Bernoulli

Jacob Bernoulli

Gottfried Leibniz

Erhard Weigel

1653

# Lagrange Relaxation

**Subgradient Optimization:** Now, look at the following problem, for some fixed value of  $\lambda$

$$\text{Max } z = 8x_1 + 9x_2 + 5x_3 + 4x_4 + \lambda(42 - 16x_1 - 20x_2 - 12x_3 - 10x_4)$$

Subject to (s.t.)

$$x_1, x_2, x_3, x_4 \in \{0,1\}$$

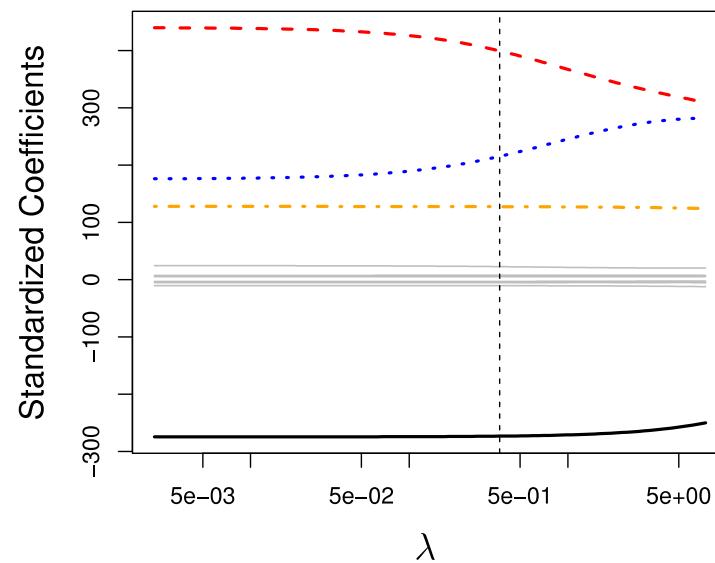
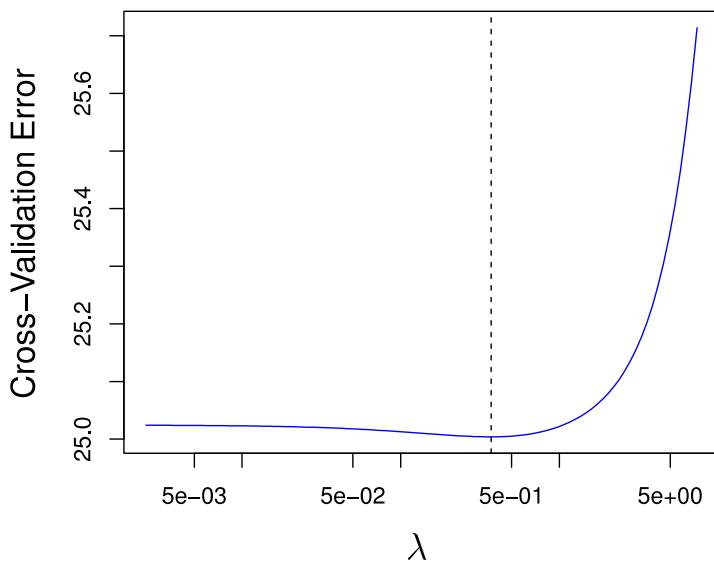
---

# Choosing $\lambda$

- To choose  $\lambda$  through cross-validation, you should choose a set of  $P$  values of  $\lambda$  to test, split the dataset into  $K$  folds
- for  $p$  in  $1:P$ :
  - for  $k$  in  $1:K$ :
    - keep fold  $k$  as hold-out data
    - use the remaining folds and  $\lambda = \lambda_p$  to estimate  $\beta_{\text{ridge}}$
    - predict hold-out data:  $y_{\text{test},k} = X_{\text{test},k}\beta_{\text{ridge}}$
    - compute a sum of squared residuals:  $\text{SSR}_k = \|y - y_{\text{test},k}\|^2$
  - end for  $k$
  - average SSR over the folds:  $\text{SSR}_p = (1/K)\sum \text{SSR}_k$
- end for  $p$
- choose optimal value:  $\lambda_{\text{opt}} = \operatorname{argmin}_p \text{SSR}_p$

# Selecting the Tuning Parameter

- We need to decide on a value for  $\lambda$
- Select a grid of potential values, use cross validation to estimate the error rate on test data (for each value of  $\lambda$ ) and select the value that gives the least error rate



# Lasso Regression

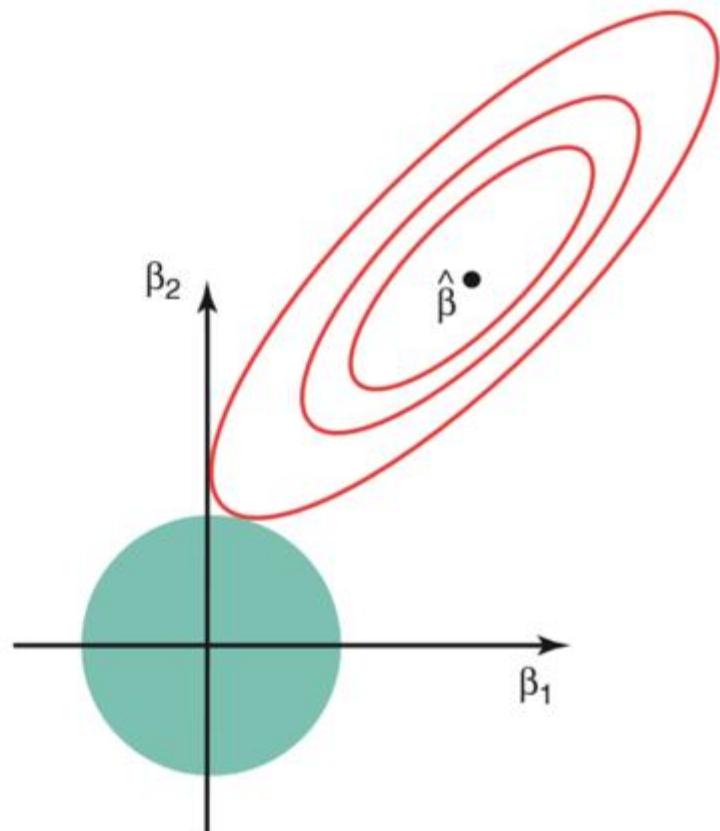
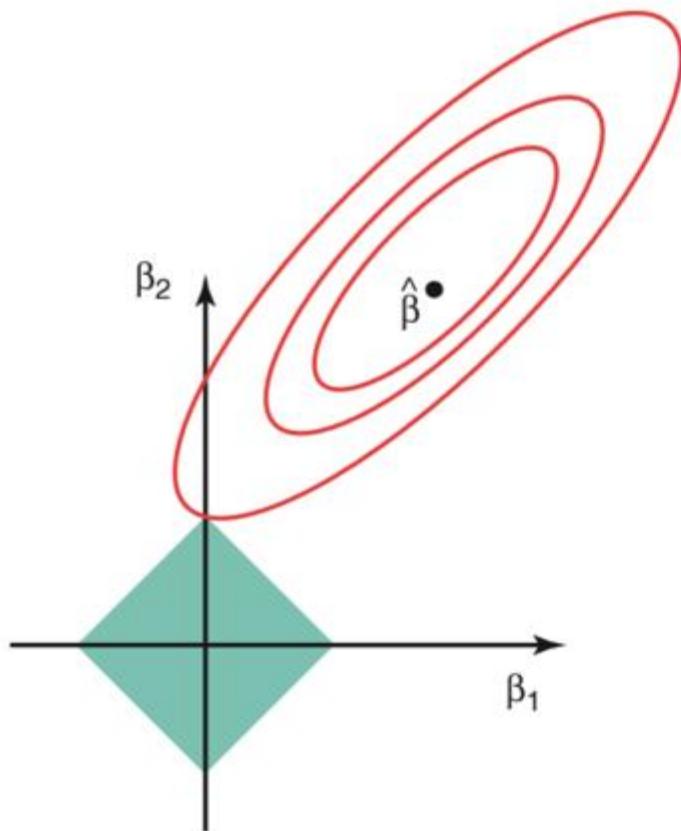
- Lasso: least absolute shrinkage and selection operator
- **Lasso Regression:**
  - Performs L1 regularization, i.e. adds penalty equivalent to **absolute value of the magnitude** of coefficients
  - Minimization objective = LS Obj +  $\lambda * (\text{sum of absolute value of coefficients})$
- In Lasso Regression, we impose the lasso constraint to the coefficients
  - minimize  $\sum(y_i - \beta^\top z_i)^2$
  - s.t.  $\sum |\beta_j| \leq t$
- In function form
  - minimize  $\sum(y_i - \beta^\top z_i)^2 + \lambda (\sum |\beta_j| - t)$

# Lasso Regression

- Even with low values of  $\lambda$  coefficients of some features are reduced to zero
- Lasso selects only a subset of the entire feature space
- Hence Lasso performs an automatic feature selection whereas Ridge does not

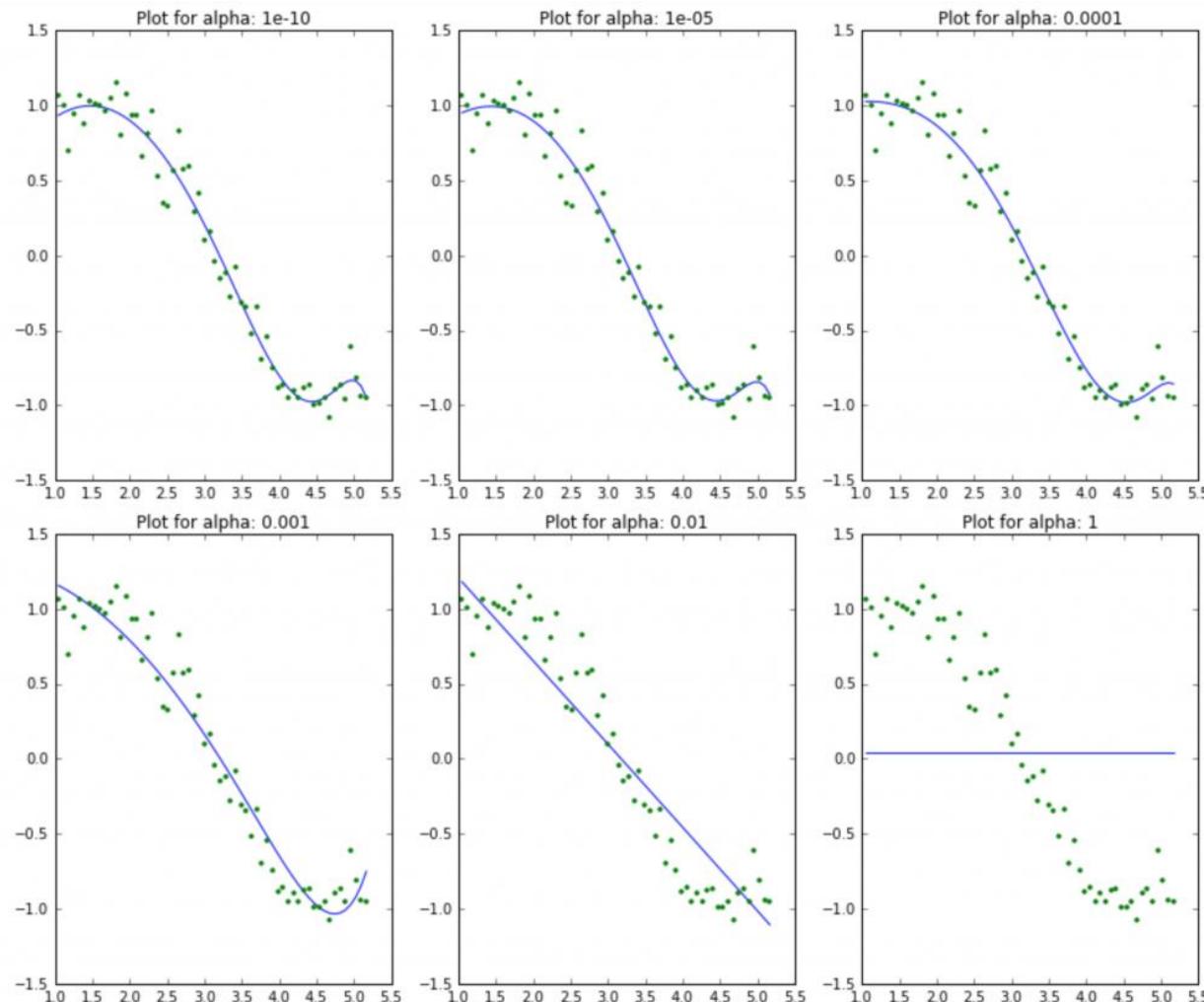
# Lasso Regression vs OLS

- “Iso-RSS” lines



# Lasso Regression

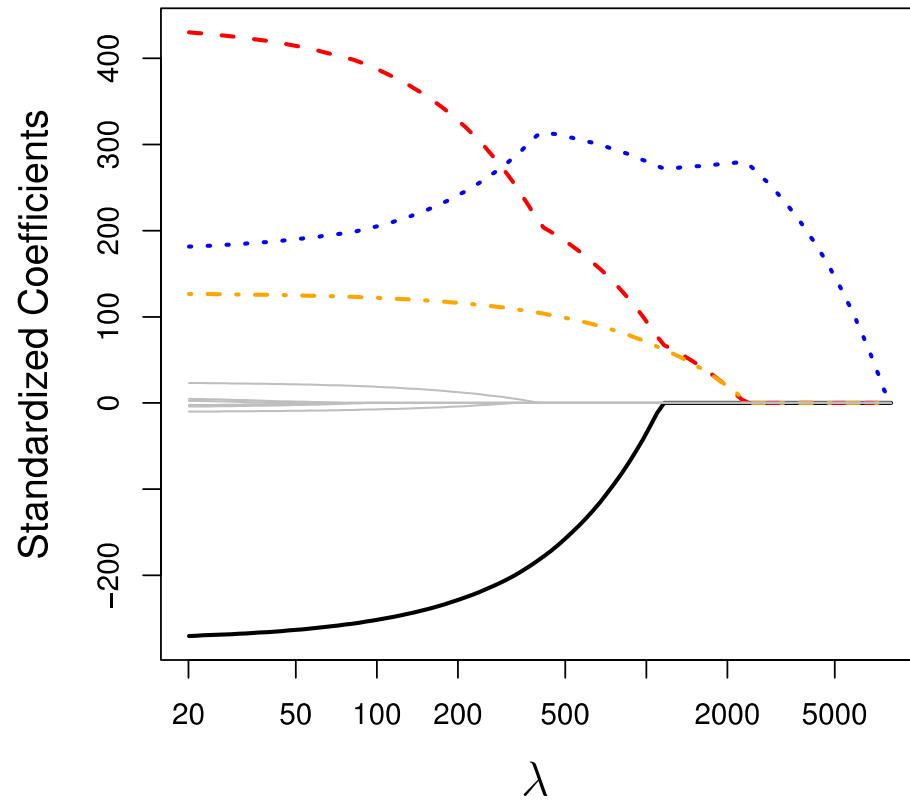
- Use different  $\lambda$  values for the sine function prediction



# Lasso Regression

- Check the Table Lasso in Lecture3a.xlsx
- Compared to the previous table, what are the differences?

# Credit Data: LASSO



# Lasso Regression

- For the same values of  $\lambda$ , the coefficients of lasso regression are much smaller as compared to that of ridge regression
- For the same  $\lambda$ , lasso has higher RSS (poorer fit) as compared to ridge regression
- Many of the coefficients are zero even for very small values of  $\lambda \sim$  sparsity

# Ridge Regression vs Lasso Regression

- **Ridge:**
  - includes all of the features in the model.
  - major advantage of ridge regression is coefficient shrinkage
  - reducing model complexity.
- **Lasso:**
  - feature selection
  - Compare to standard feature selection ridge and lasso regression provide
    - **better output,**
    - can be **automated**

# Ridge Regression vs Lasso Regression

- **Use Cases ~ Ridge:**

- *Prevent overfitting.*
- Not preferred when the number of features are really high.
- Works well even in presence of highly correlated features as it will include all of them in the model but the coefficients will be distributed among them depending on the correlation.

- **Use Cases ~ Lasso:**

- *Sparse solutions*
- Preferred when the number of features are really high
- Arbitrarily selects any one feature among the highly correlated ones and reduced the coefficients of the rest to zero. Also, the chosen variable changes randomly with change in model parameters. This generally doesn't work that well as compared to ridge regression.

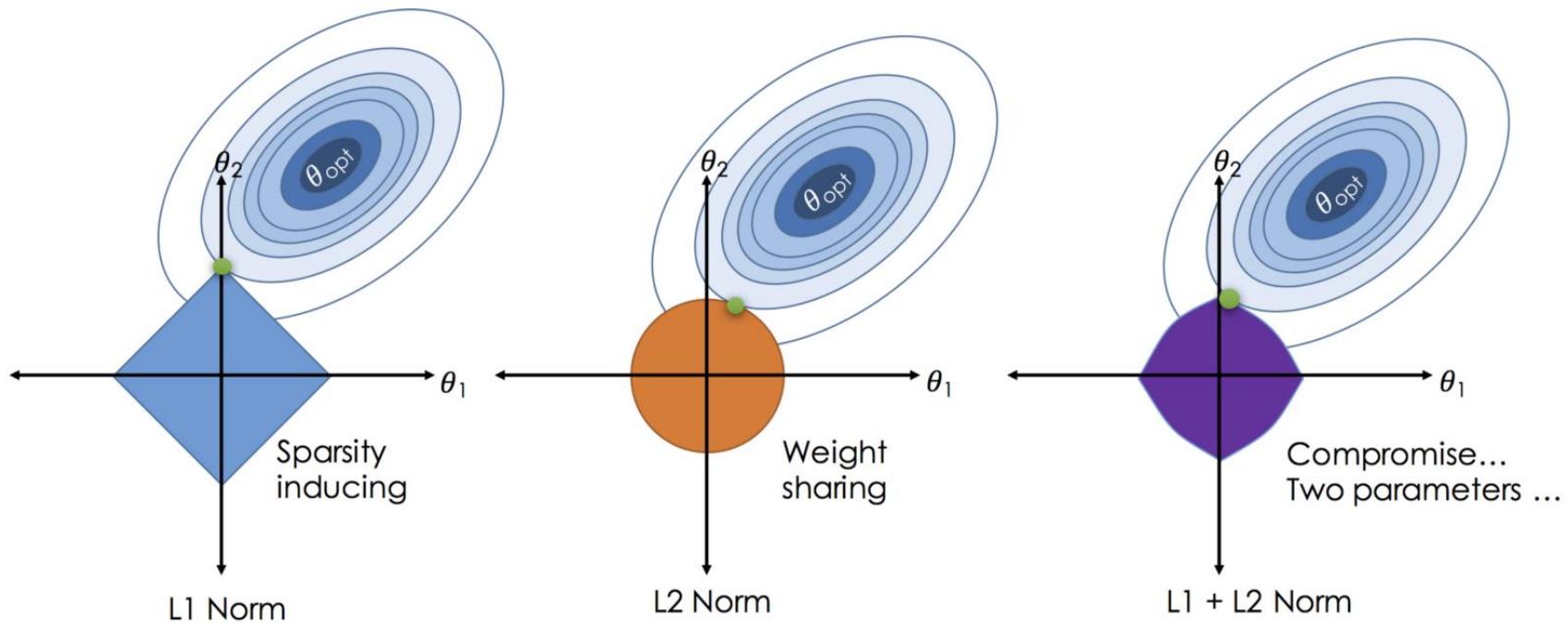
# Data with High Number of Feature

- Suppose we have a data with a large number of features
- Applying Ridge regression
  - Keep all the features
  - Shrink the coefficients
  - Complexity may not be reduced with that many features
- Applying Lasso regression
  - Removing some of the features
  - Correlated variables ~ keeps only one variable
  - Loss of information resulting in lower accuracy

# Elastic Net Regression

- Elastic Net Regression is basically a combination of both Ridge and Lasso regression
- Elastic Net is useful when there are multiple features which are correlated. Lasso is likely to pick one of these at random, while Elastic Net is likely to pick both.
- Two penalty terms with two parameters
  - $\text{Min } \sum(y_i - \beta^T z_i)^2 + \lambda_1 (\sum |\beta_j| - t) + \lambda_2 (\sum \beta_j^2 - t)$

# Elastic Net Regression



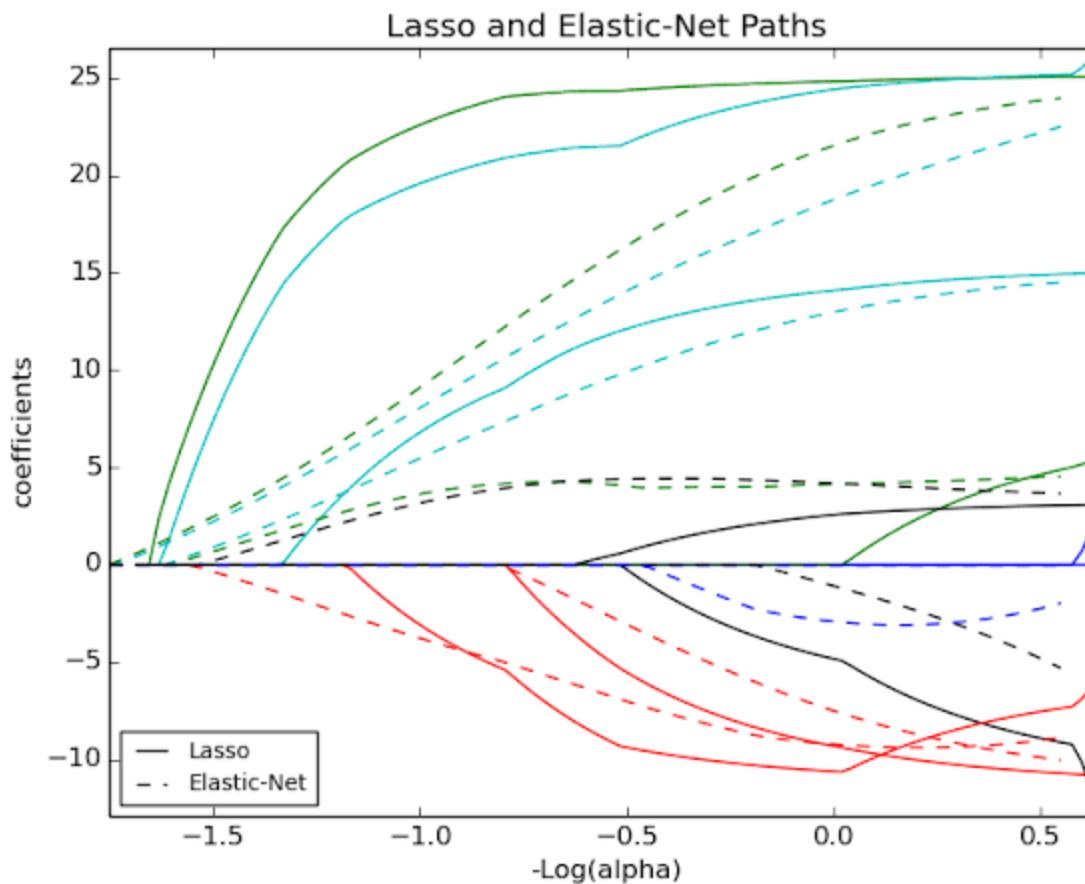
# Elastic Net Regression

- Elastic Net Regression
  - It encourages group effect in case of highly correlated variables
  - There are no limitations on the number of selected variables
  - It can suffer with double shrinkage

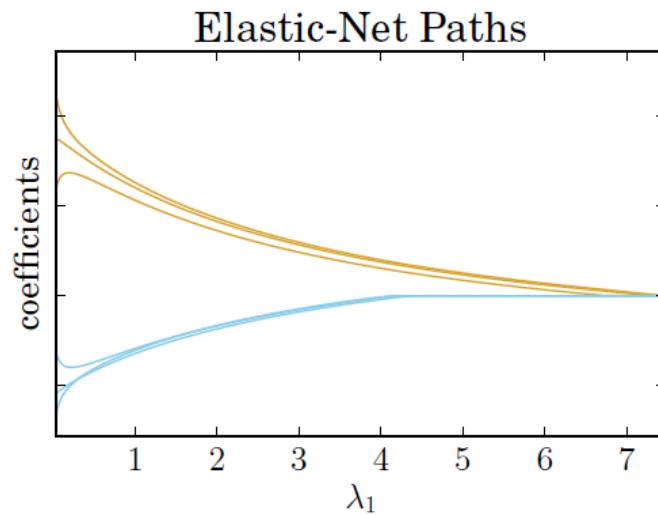
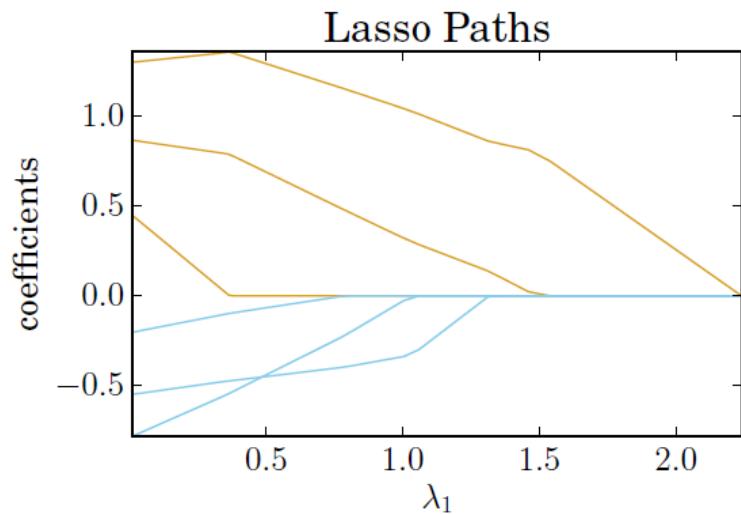
# Elastic Net Regression

- When variables are highly correlated (and same scale, after normalization),
  - we want to give them roughly the same weight.
- Why?
  - Let their error cancel out

# Elastic Net Regression



# Elastic Net Regression



# Dimension Reduction Methods

- The methods that we have discussed so far in this chapter have involved fitting linear regression models, via least squares or a shrunken approach, using the original predictors,  $X_1, X_2, \dots, X_p$ .
- We now explore a class of approaches that *transform* the predictors and then fit a least squares model using the transformed variables. We will refer to these techniques as *dimension reduction* methods.

# Dimension Reduction Methods

Let  $Z_1, Z_2, \dots, Z_M$  represent  $M < p$  *linear combinations* of our original  $p$  predictors. That is,

$$Z_m = \sum_{j=1}^p \phi_{mj} X_j$$

for some constants  $\varphi_{m1}, \dots, \varphi_{mp}$ .

We can then fit the linear regression model,

$$y_i = \theta_0 + \sum_{m=1}^M \theta_m z_{im} + \epsilon_i, \quad i = 1, \dots, n,$$

using ordinary least squares.

Note that in model, the regression coefficients are given by  $\theta_0, \theta_1, \dots, \theta_M$ . If the constants  $\varphi_{m1}, \dots, \varphi_{mp}$  are chosen wisely, then such dimension reduction approaches can often outperform OLS regression.

$$\sum_{m=1}^M \theta_m z_{im} = \sum_{m=1}^M \theta_m \sum_{j=1}^p \phi_{mj} x_{ij} = \sum_{j=1}^p \sum_{m=1}^M \theta_m \phi_{mj} x_{ij} = \sum_{j=1}^p \beta_j x_{ij},$$

where

$$\beta_j = \sum_{m=1}^M \theta_m \phi_{mj}. \quad (3)$$

Hence model can be thought of as a special case of the original linear regression model.

Dimension reduction serves to constrain the estimated  $\beta_j$  coefficients, since now they must take the form.

Can win in the bias-variance tradeoff.

# Principal Components Analysis

- PCA produces a low-dimensional representation of a dataset. It finds a sequence of linear combinations of the variables that have maximal variance, and are mutually uncorrelated.
- Apart from producing derived variables for use in supervised learning problems, PCA also serves as a tool for data visualization.

# Principal Components Regression

- Here we apply principal components analysis (PCA) to define the linear combinations of the predictors, for use in our regression.
- The first principal component is that (normalized) linear combination of the variables with the largest variance.
- The second principal component has largest variance, subject to being uncorrelated with the first.
- And so on.
- Hence with many correlated original variables, we replace them with a small set of principal components that capture their joint variation.

# Principal Components Regression

- The *first principal component* of a set of features  $X_1, X_2, \dots, X_p$  is the normalized linear combination of the features

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \dots + \phi_{p1}X_p$$

that has the largest variance. By *normalized*, we mean that

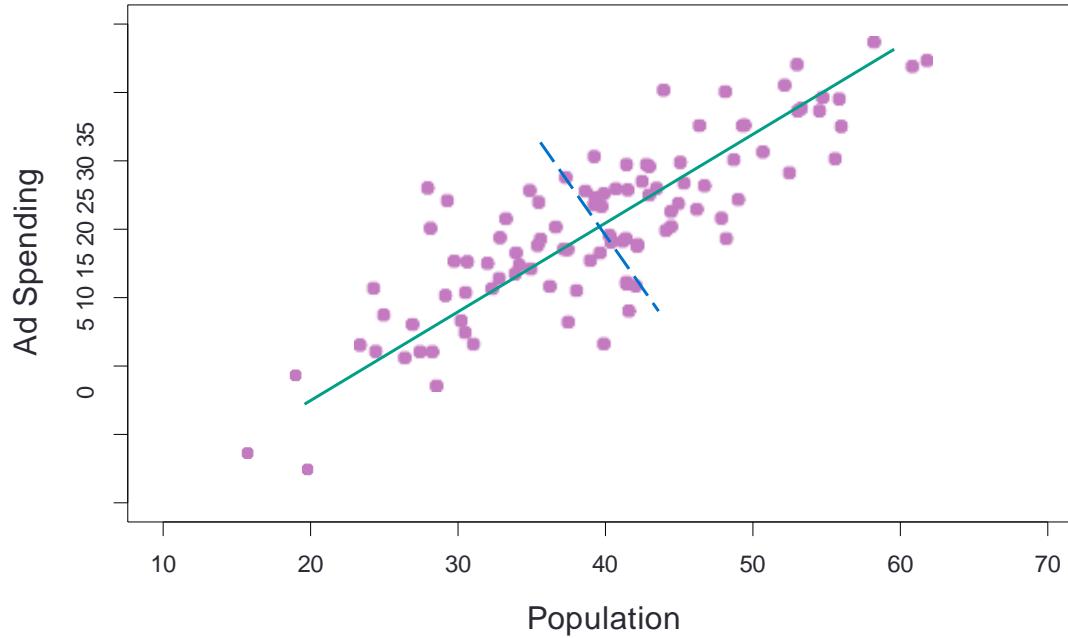
$$\sum_{j=1}^p \phi_{j1}^2 = 1.$$

We refer to the elements  $\phi_{11}, \dots, \phi_{p1}$  as the loadings of the first principal component; together, the loadings make up the principal component loading vector,

$$\varphi_1 = (\varphi_{11} \varphi_{21} \dots \varphi_{p1})^T.$$

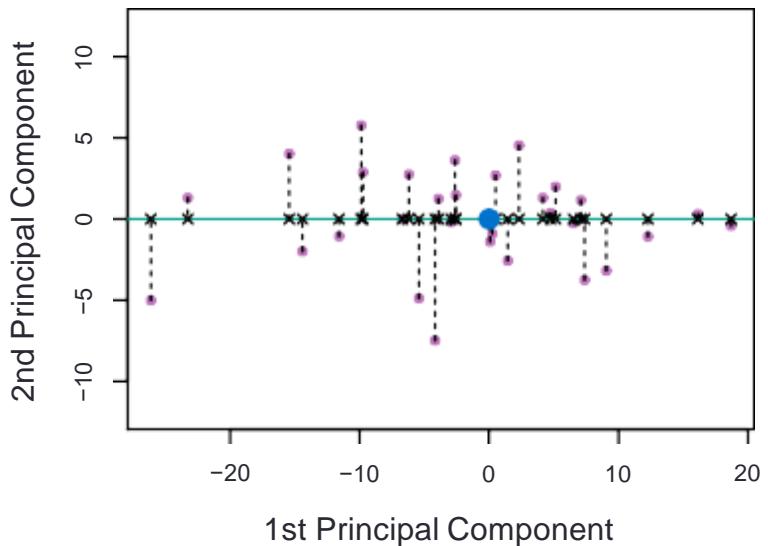
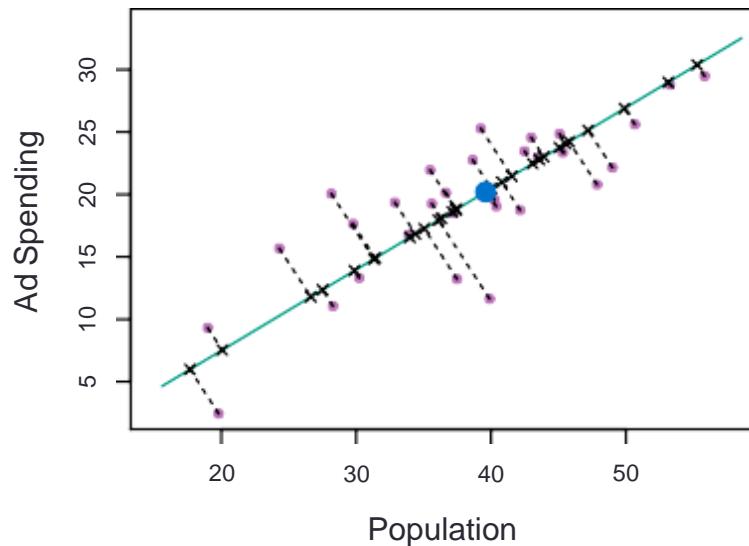
We constrain the loadings so that their sum of squares is equal to one, since otherwise setting these elements to be arbitrarily large in absolute value could result in an arbitrarily large variance.

# Principal Components Regression



The population size (**pop**) and ad spending (**ad**) for 100 different cities are shown as purple circles. The green solid line indicates the first principal component, and the blue dashed line indicates the second principal component.

# Principal Components Regression



A subset of the advertising data. **Left:** The first principal component, chosen to minimize the sum of the squared perpendicular distances to each point, is shown in green. These distances are represented using the black dashed line segments. **Right:** The left-hand panel has been rotated so that the first principal component lies on the x-axis.

# Computation of Principal Components

- Suppose we have a  $n \times p$  data set  $\mathbf{X}$ . Since we are only interested in variance, we assume that each of the variables in  $\mathbf{X}$  has been centered to have mean zero (that is, the column means of  $\mathbf{X}$  are zero).
- We then look for the linear combination of the sample feature values of the form

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \dots + \phi_{p1}x_{ip}$$

- for  $i = 1, \dots, n$  that has largest sample variance, subject to
$$\sum_{j=1}^p \phi_{j1}^2 = 1.$$
- Since each of the  $x_{ij}$  has mean zero, then so does  $z_{i1}$  (for any values of  $\phi_{j1}$ ). Hence the sample variance of the  $z_{i1}$  can be written as

$$\frac{1}{n} \sum_{i=1}^n z_{i1}^2.$$

# Computation of Principal Components

- Plugging in the first principal component loading vector solves the optimization problem

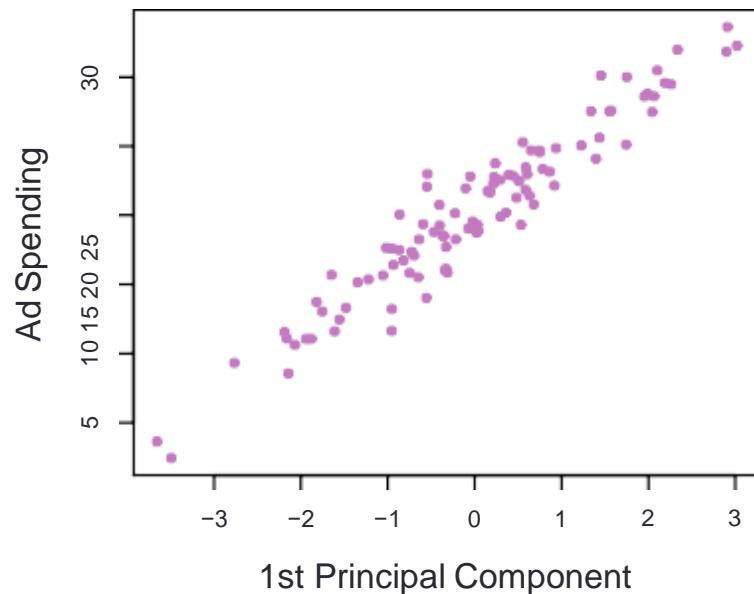
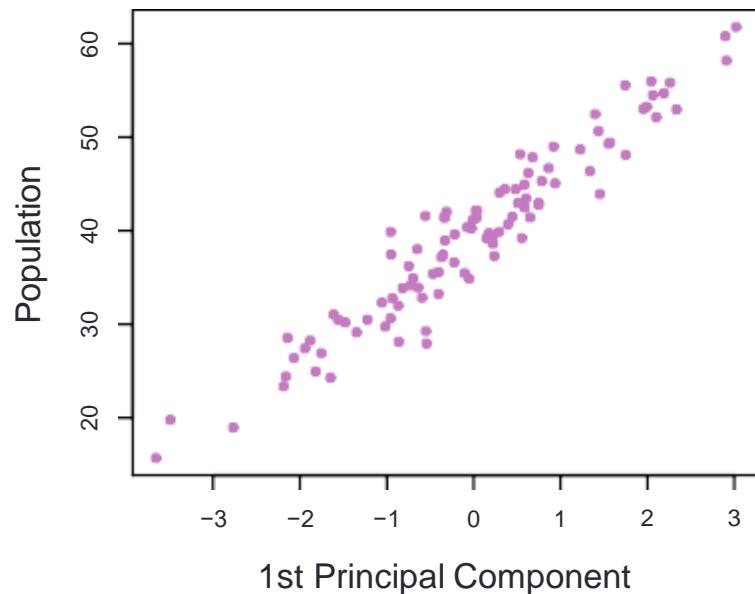
$$\underset{\phi_{11}, \dots, \phi_{p1}}{\text{maximize}} \frac{1}{n} \sum_{i=1}^n \left( \sum_{j=1}^p \phi_{j1} x_{ij} \right)^2 \text{ subject to } \sum_{j=1}^p \phi_{j1}^2 = 1.$$

- This problem can be solved via a singular-value decomposition of the matrix  $\mathbf{X}$ , a standard technique in linear algebra.
- We refer to  $Z_1$  as the first principal component, with realized values  $Z_{11}, \dots, Z_{n1}$

# Geometry of PCA

- The loading vector  $\varphi_1$  with elements  $\varphi_{11}, \varphi_{21}, \dots, \varphi_{p1}$  defines a direction in feature space along which the data vary the most.
- If we project the  $n$  data points  $x_1, \dots, x_n$  onto this direction, the projected values are the principal component scores  $z_{11}, \dots, z_{n1}$  themselves.

# Principal Components Regression



*Plots of the first principal component scores  $z_{i1}$  versus **pop** and **ad**. The relationships are strong.*

# Computation of Principal Components

- The second principal component is the linear combination of  $X_1, \dots, X_p$  that has maximal variance among all linear combinations that are *uncorrelated* with  $Z_1$ .
- The second principal component scores  $Z_{12}, Z_{22}, \dots, Z_{n2}$  take the form

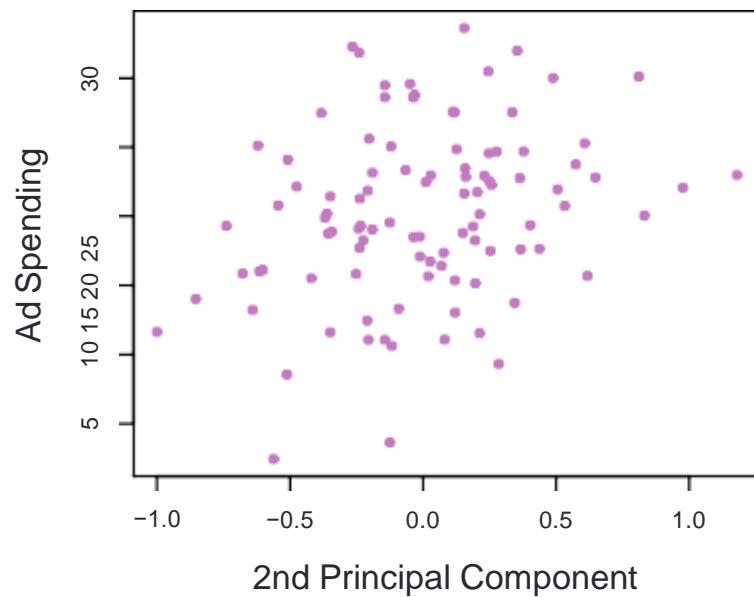
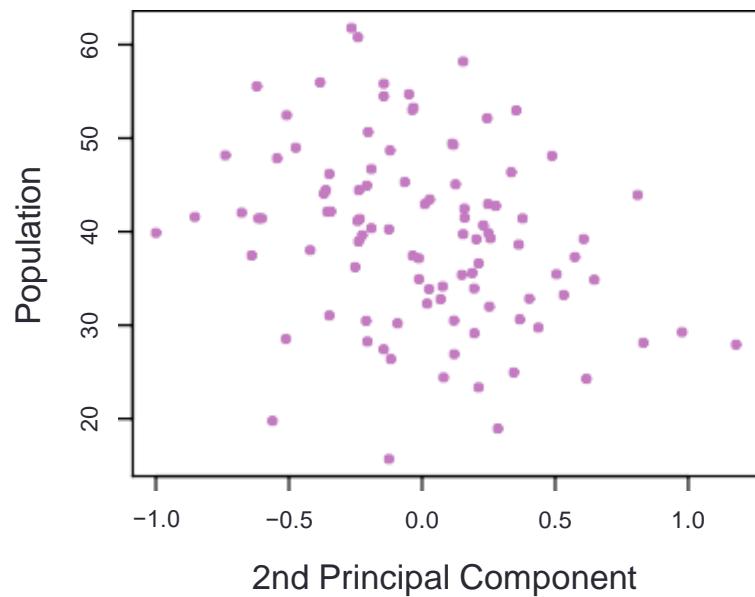
$$Z_{i2} = \varphi_{12}X_{i1} + \varphi_{22}X_{i2} + \dots + \varphi_{p2}X_{ip},$$

where  $\varphi_2$  is the second principal component loading vector, with elements  $\varphi_{12}, \varphi_{22}, \dots, \varphi_{p2}$ .

# Computation of Principal Components

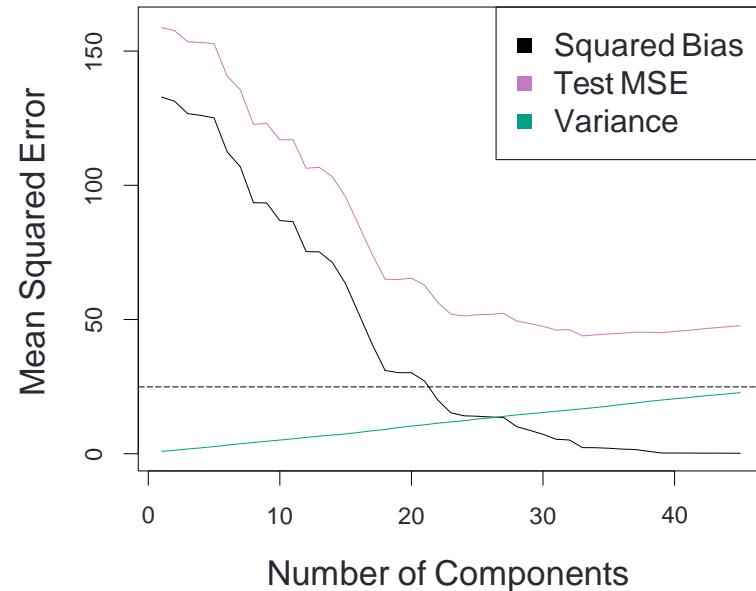
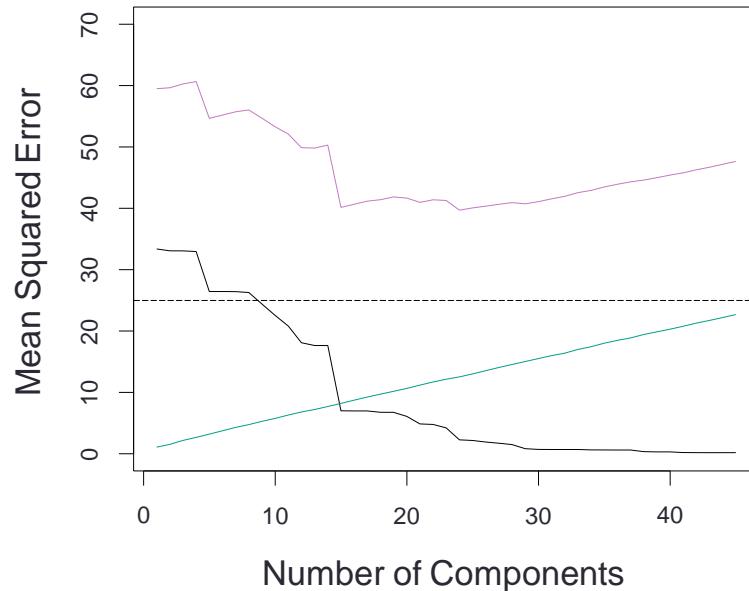
- It turns out that constraining  $Z_2$  to be uncorrelated with  $Z_1$  is equivalent to constraining the direction  $\varphi_2$  to be orthogonal (perpendicular) to the direction  $\varphi_1$ . And so on.
- The principal component directions  $\varphi_1, \varphi_2, \varphi_3, \dots$  are the ordered sequence of right singular vectors of the matrix  $\mathbf{X}$ , and the variances of the components are  $1/n$  times the squares of the singular values. There are at most  $\min(n - 1, p)$  principal components.

# Principal Components Regression



*Plots of the second principal component scores  $z_{i2}$  versus pop and ad. The relationships are weak.*

# Application to Principal Components Regression

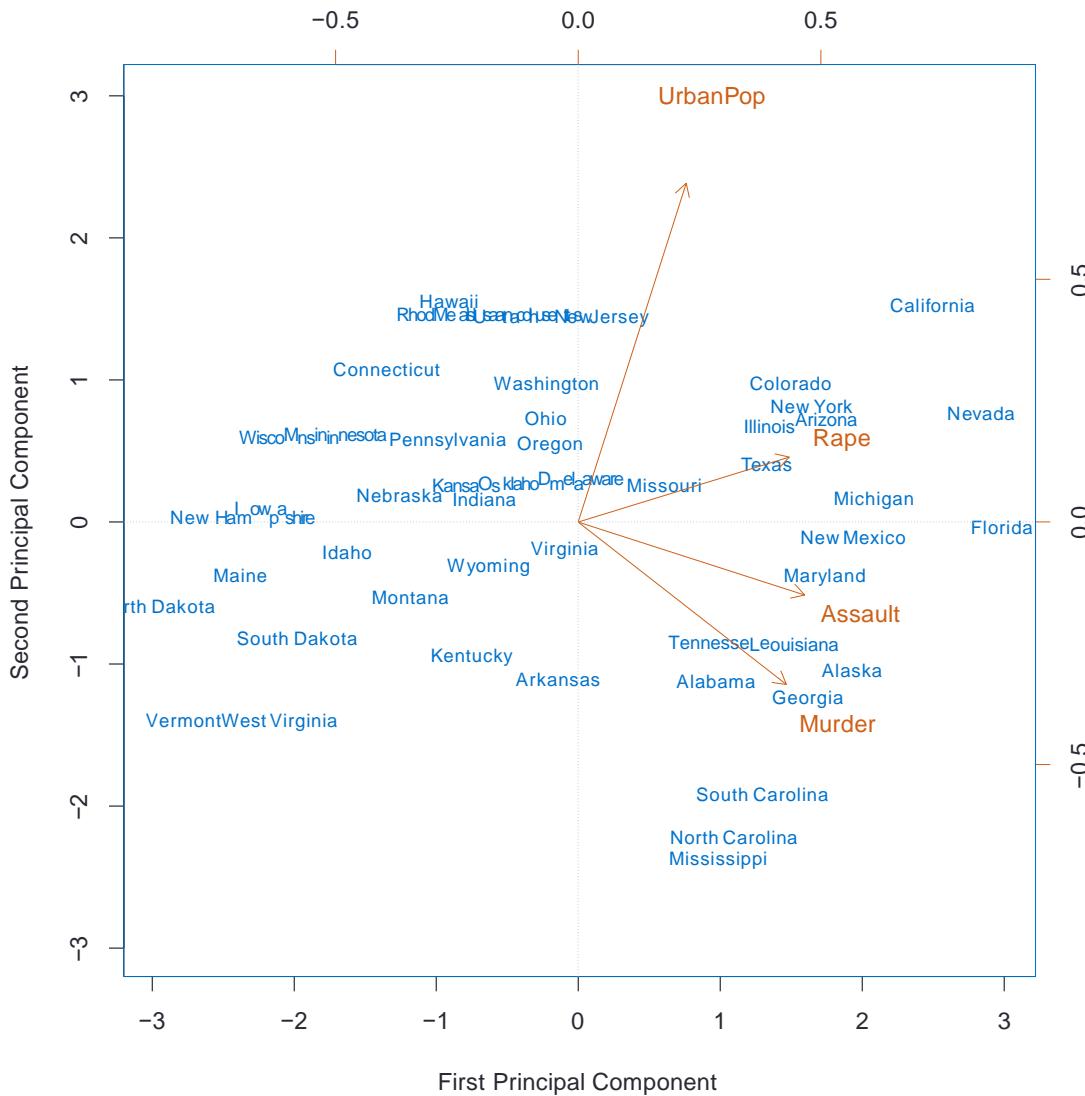


*PCR was applied to two simulated data sets. The black, green, and purple lines correspond to squared bias, variance, and test mean squared error, respectively.*

# Example

- **USAArrests** data: For each of the fifty states in the United States, the data set contains the number of arrests per 100, 000 residents for each of three crimes: **Assault**, **Murder**, and **Rape**. We also record **UrbanPop** (the percent of the population in each state living in urban areas).
- The principal component score vectors have length  $n = 50$ , and the principal component loading vectors have length  $p = 4$ .
- PCA was performed after standardizing each variable to have mean zero and standard deviation one.

# USAarrests data: PCA plot

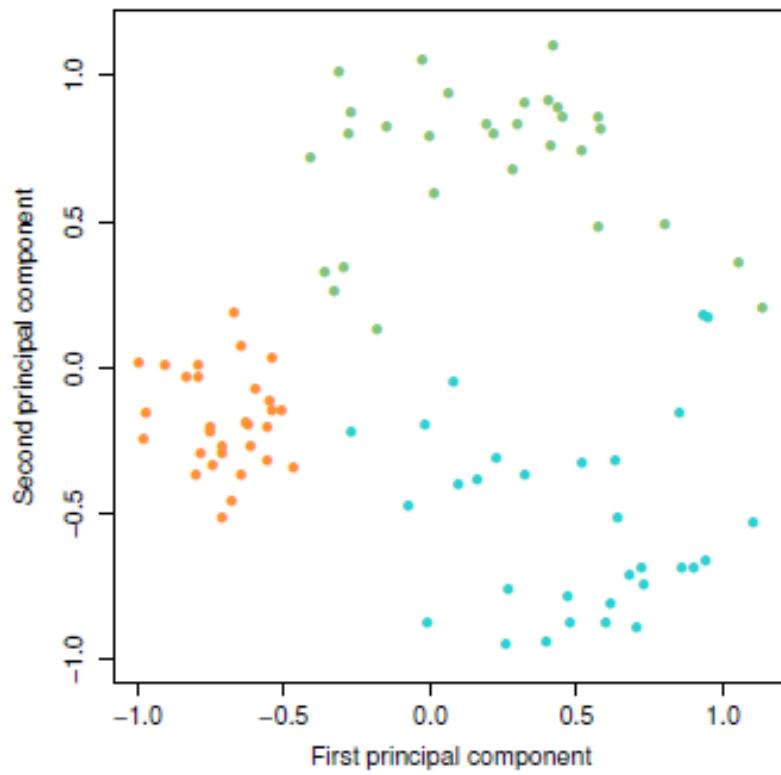
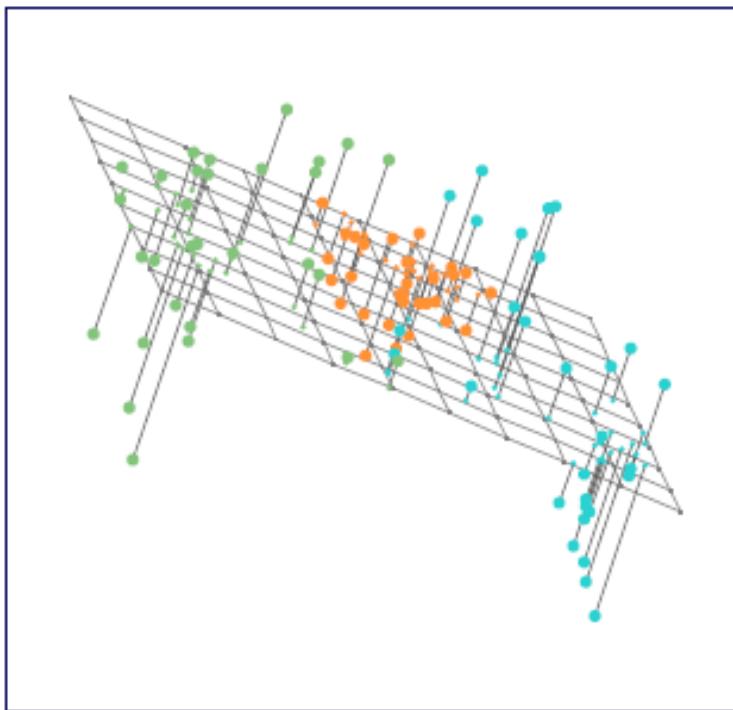


# Example

- The first two principal components for the USArrests data.
- The blue state names represent the scores for the first two principal components.
- The orange arrows indicate the first two principal component loading vectors (with axes on the top and right). For example, the loading for **Rape** on the first component is 0.54, and its loading on the second principal component 0.17 [the word **Rape** is centered at the point (0.54, 0.17)].
- This figure is known as a *biplot*, because it displays both the principal component scores and the principal component loadings.

	PC1	PC2
Murder	0.5358995	-0.4181809
Assault	0.5831836	-0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

# Example

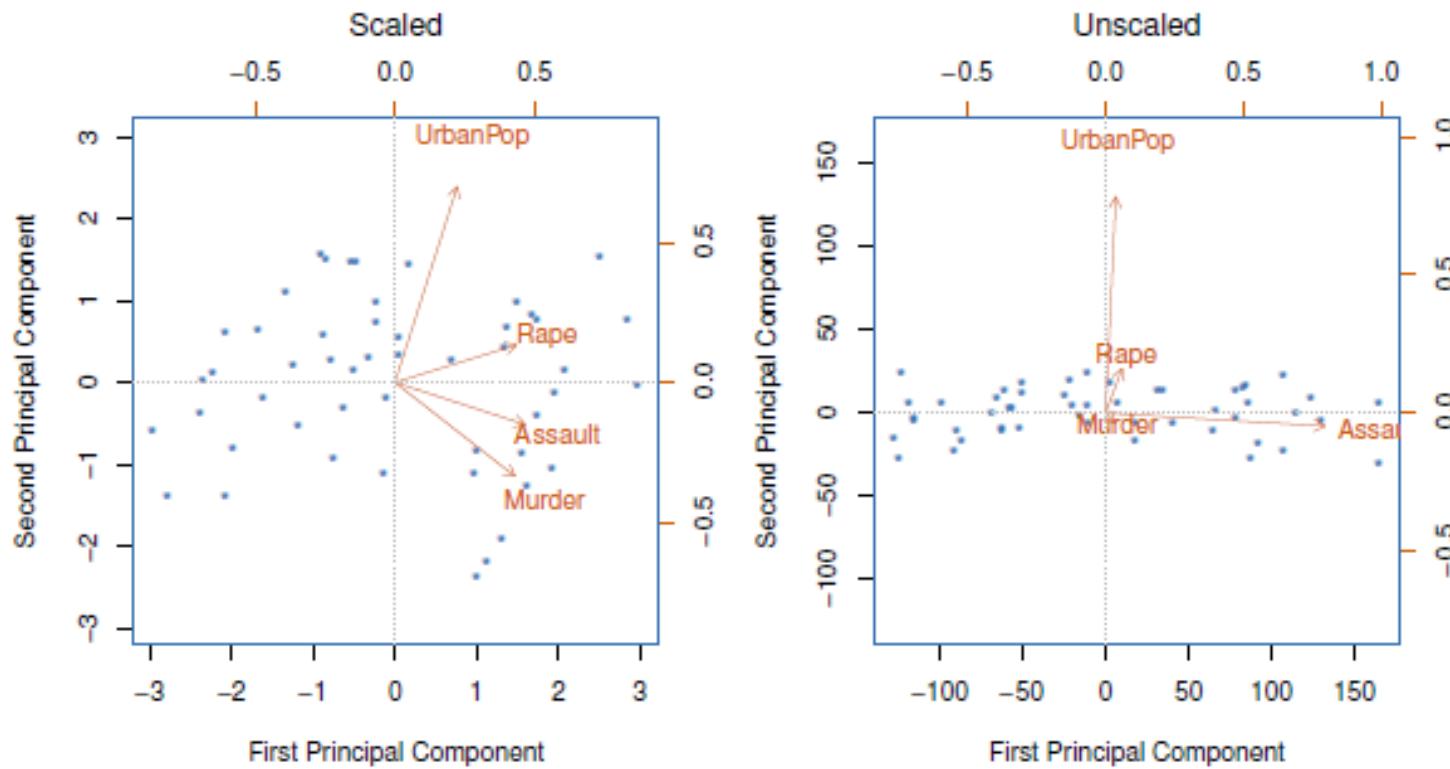


# Example

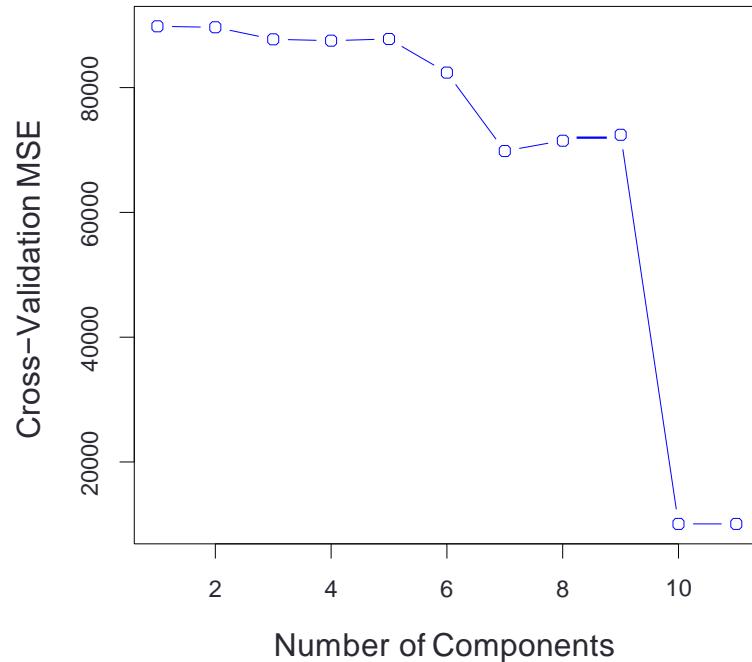
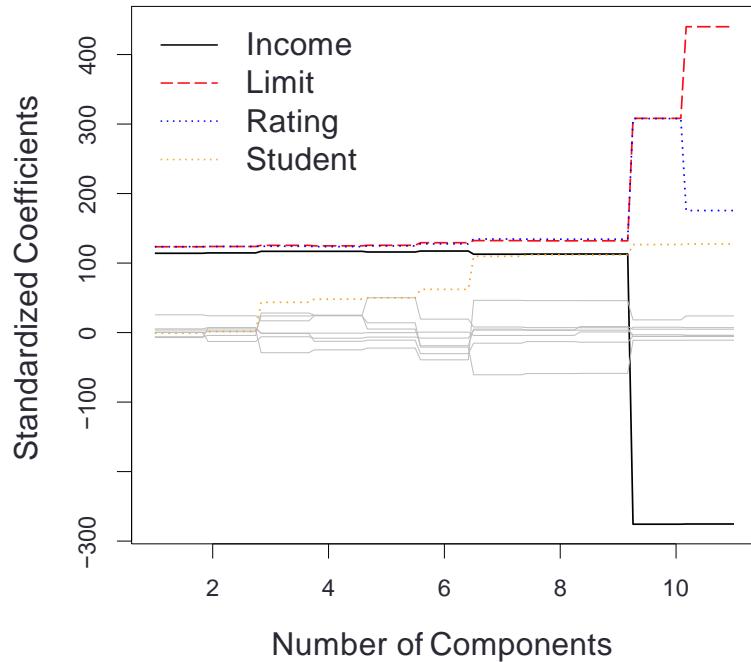
- PCA find the hyperplane closest to the observations. The first principal component loading vector has a very special property: it defines the line in  $p$ -dimensional space that is *closest* to the  $n$  observations (using average squared Euclidean distance as a measure of closeness)
- The notion of principal components as the dimensions that are closest to the  $n$  observations extends beyond just the first principal component.
- For instance, the first two principal components of a data set span the plane that is closest to the  $n$  observations, in terms of average squared Euclidean distance.

# Scaling

- If the variables are in different units, scaling each to have standard deviation equal to one is recommended.
- If they are in the same units, you might or might not scale the variables.



# Choosing the number of directions $M$



**Left:** PCR standardized coefficient estimates on the Credit data set for different values of  $M$ . **Right:** The 10-fold cross validation MSE obtained using PCR, as a function of  $M$ .

# Feature Selection or Hyper-parameter Tuning

- Which one comes first? Why