MGMT 655: Machine Learning for Decision Making

Lesson 1: The Fundamentals & Workflow of Data-Driven Decision Making

Prof. SUNGJONG ROH

MBA May Term| 2024-05-27

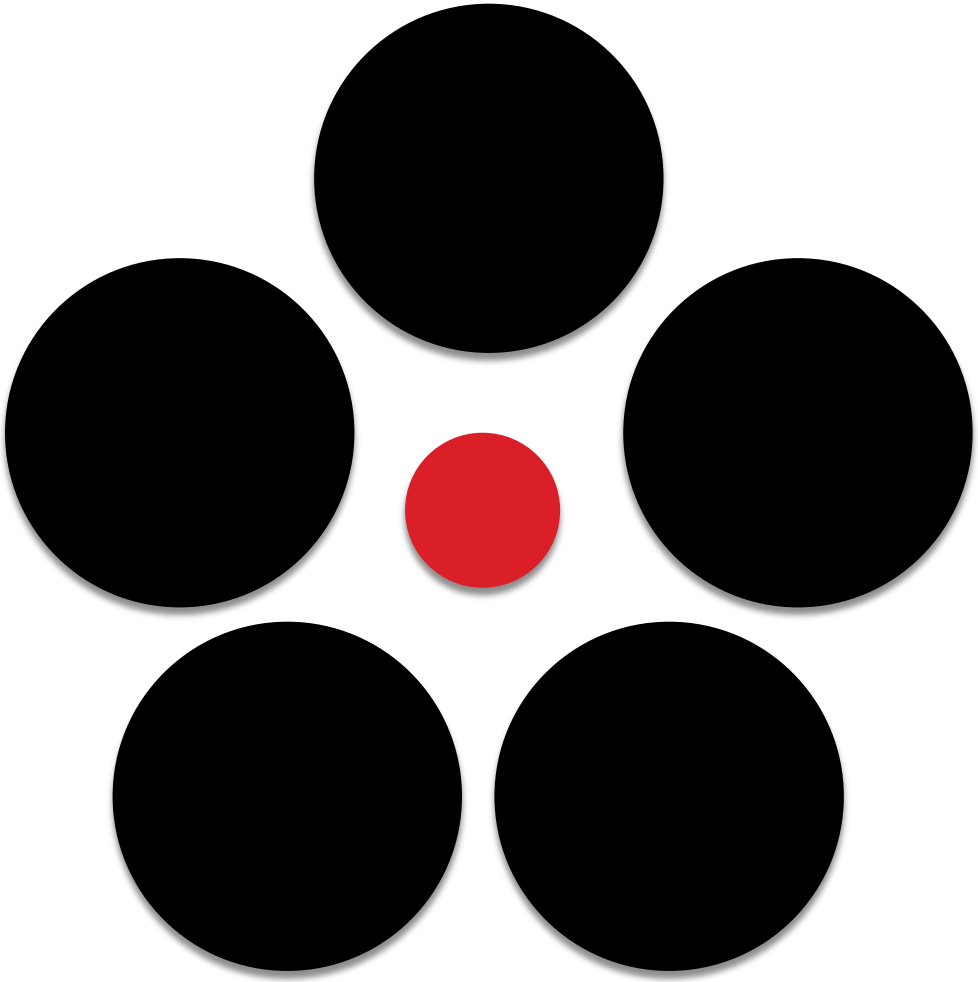
SMU SINGAPORE MANAGEMENT UNIVERSITY
Lee Kong Chian School of Business

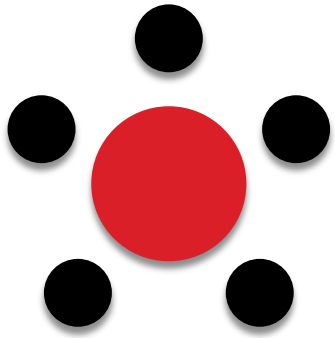
Human intelligence
has its own idiosyncratic nature
that sometimes promotes
but often prevents
optimal decision making.

Two objects cannot occupy the same space.

13

Human intelligence
has its own idiosyncratic nature
that sometimes promotes
but often prevents
optimal decision making.





Welcome to the course, Team ☺

Today, we will learn the processes of data-driven decision making.

Today, we will learn the processes of data-driven decision making.

The goal of today's learning is two-folds.

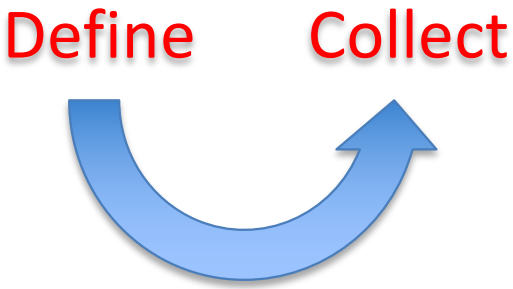
On one hand, it is intended to give you understanding about
what are the steps of work that is required
to make informed strategic decision-making using data.

Importantly, for each step of work,
I'll discuss elaborated learning pointers
(things to keep in mind as strategic thinkers in organizations)
pertaining to each of the step of work.

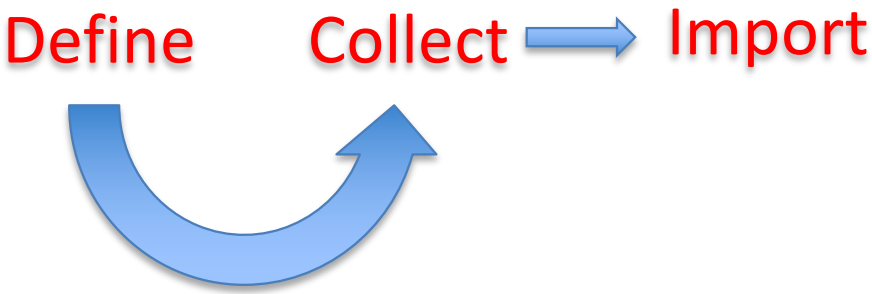
Workflow of Data-Driven Decision Making

Define

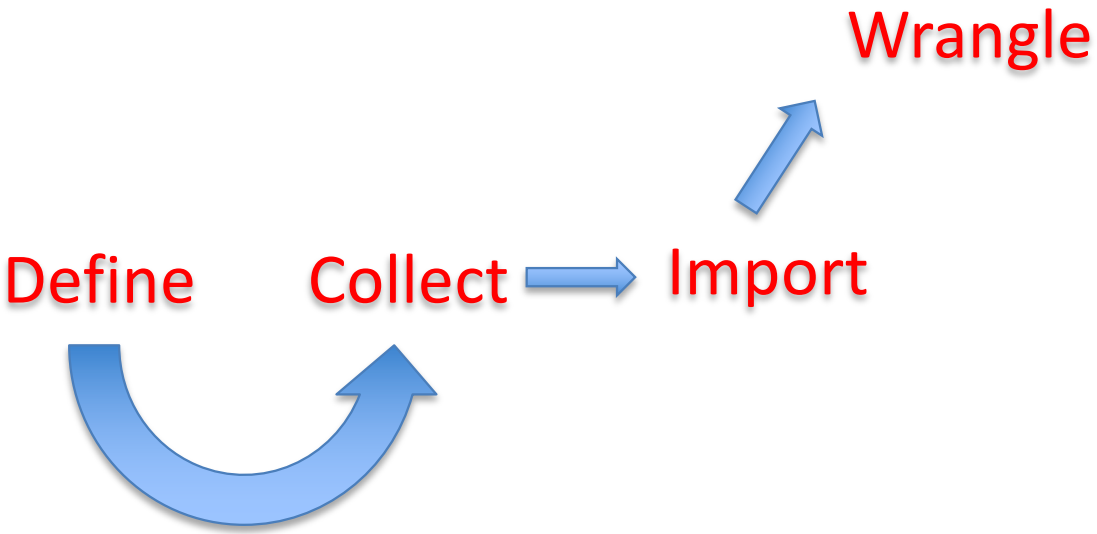
Workflow of Data-Driven Decision Making



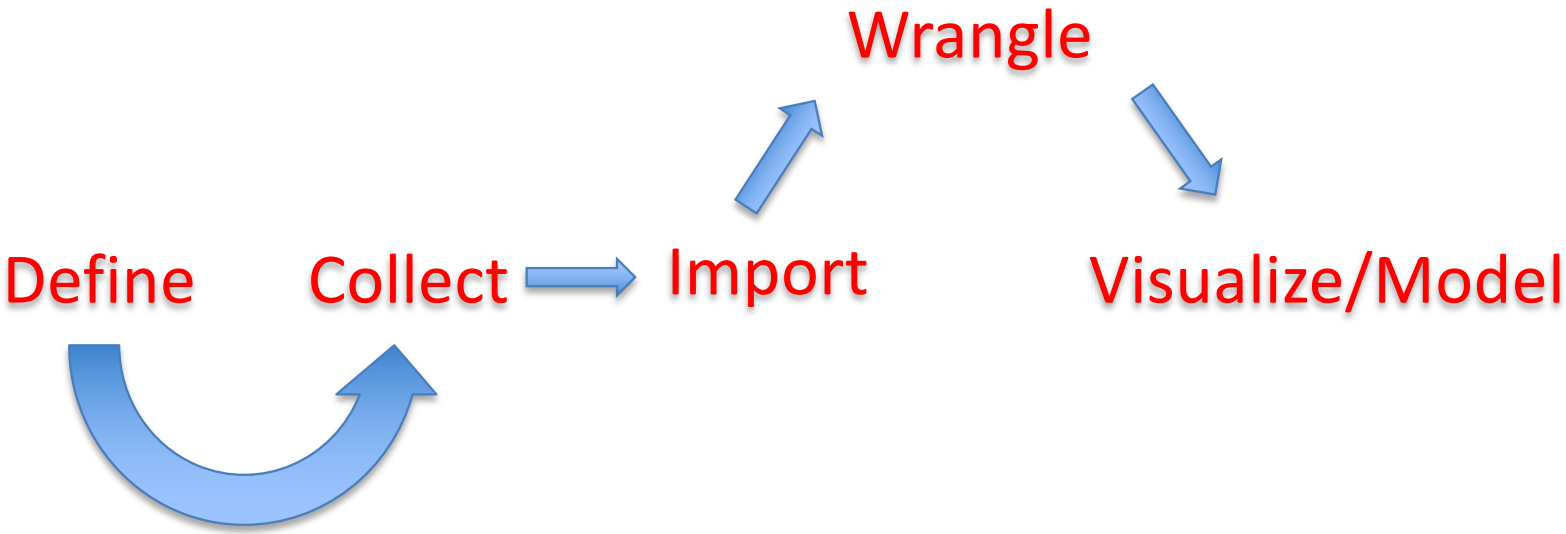
Workflow of Data-Driven Decision Making



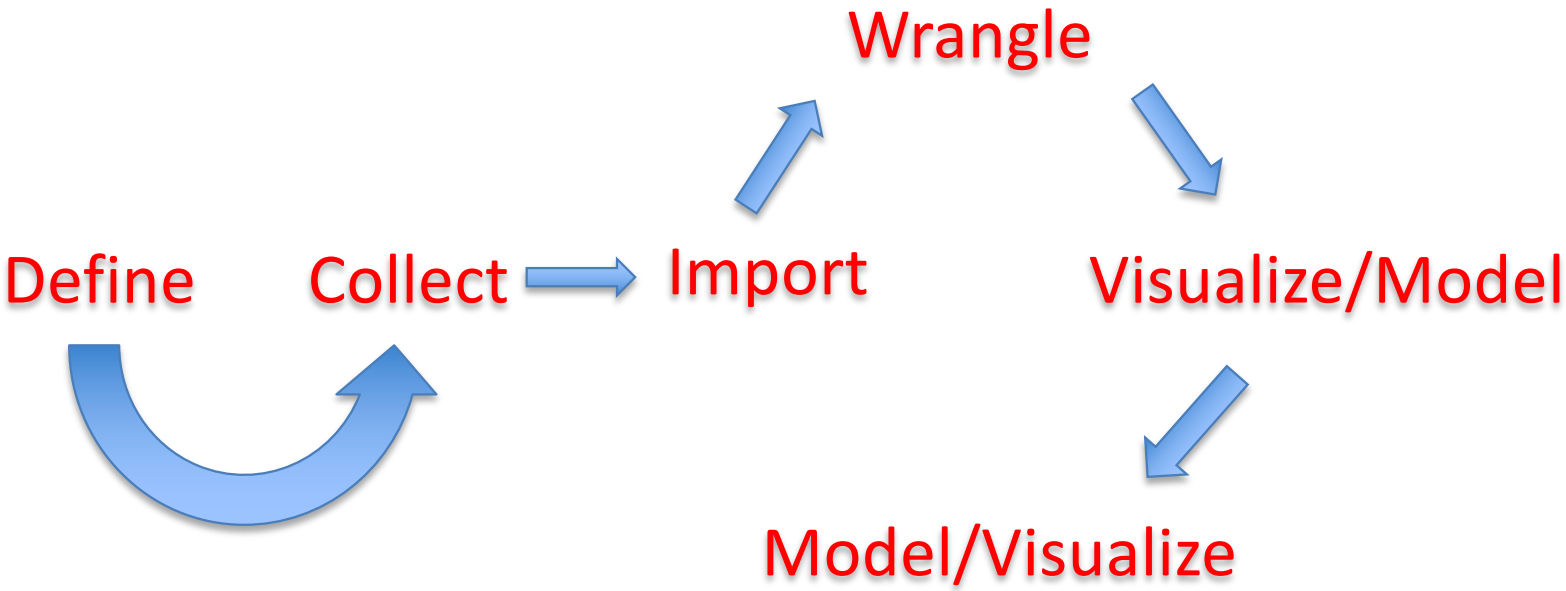
Workflow of Data-Driven Decision Making



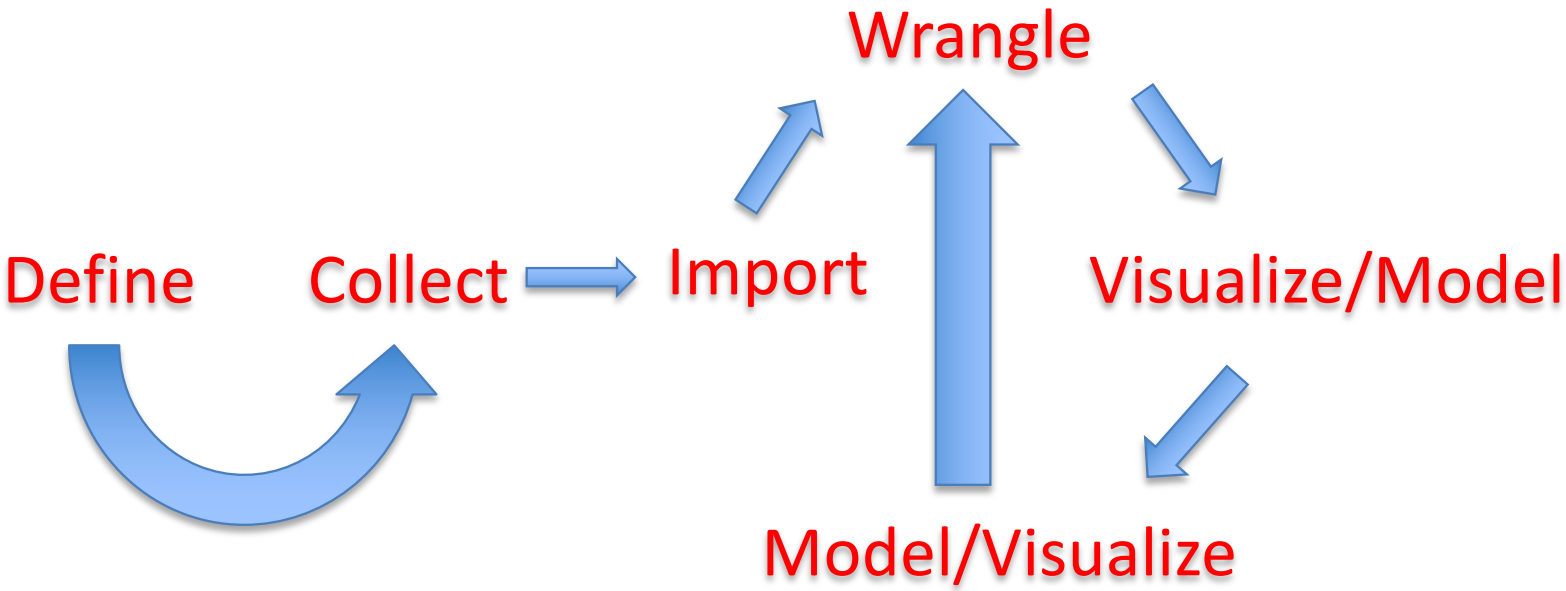
Workflow of Data-Driven Decision Making



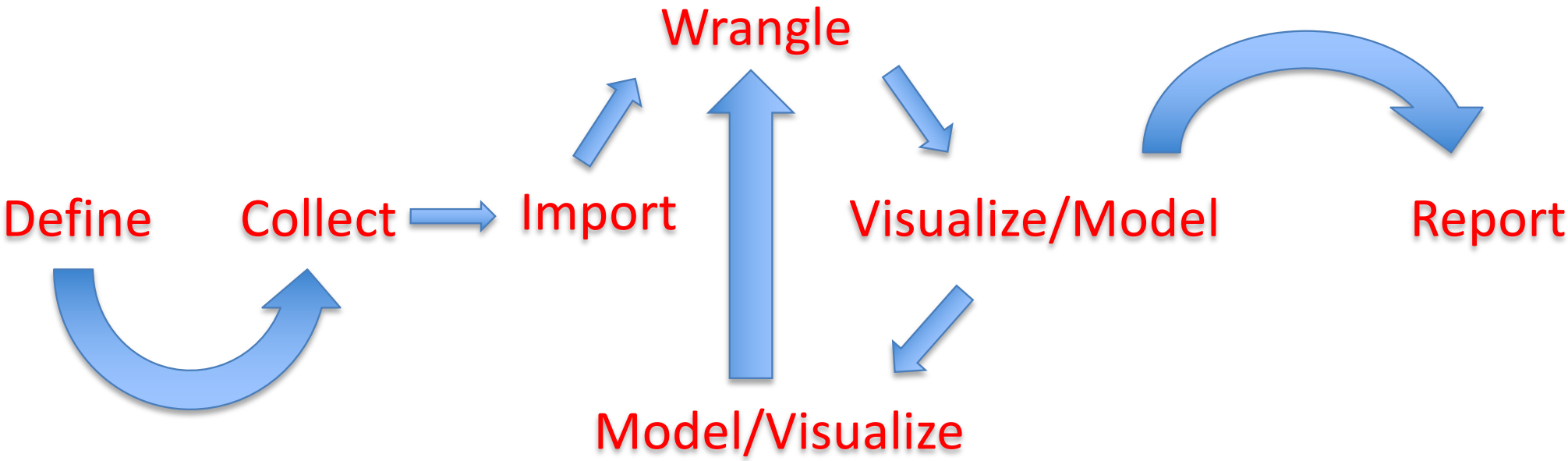
Workflow of Data-Driven Decision Making



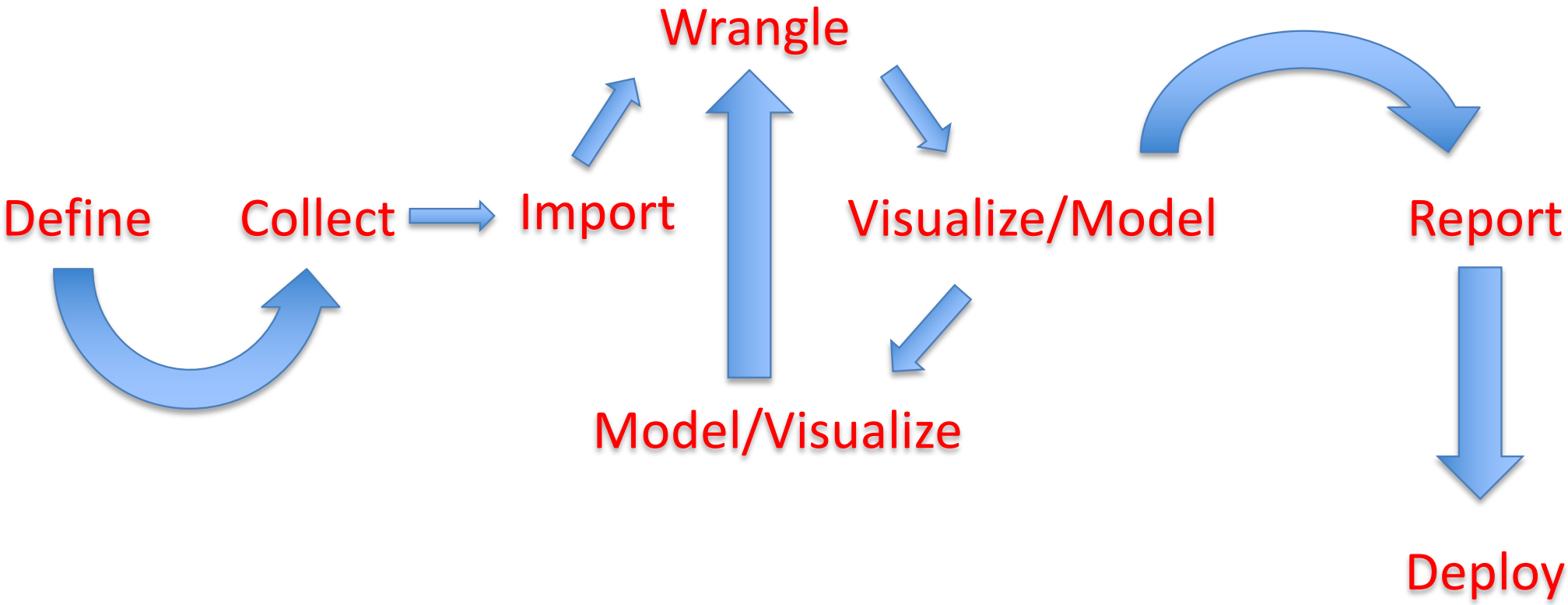
Workflow of Data-Driven Decision Making



Workflow of Data-Driven Decision Making



Workflow of Data-Driven Decision Making



Agenda

Define

Collect

Import

Wrangle

Visualize

Model

Report

Deploy

Agenda (Workflow of Data-Driven Decision Making)

Define

Collect

Import

Wrangle

Visualize

Model

Report

Deploy

Agenda (Workflow of Data-Driven Decision Making)

Define

Collect

Import

Wrangle

Visualize

Model

Report

Deploy

Agenda (Workflow of Data-Driven Decision Making)

Define

Collect

Import

Wrangle

Visualize

Model

Report

Deploy

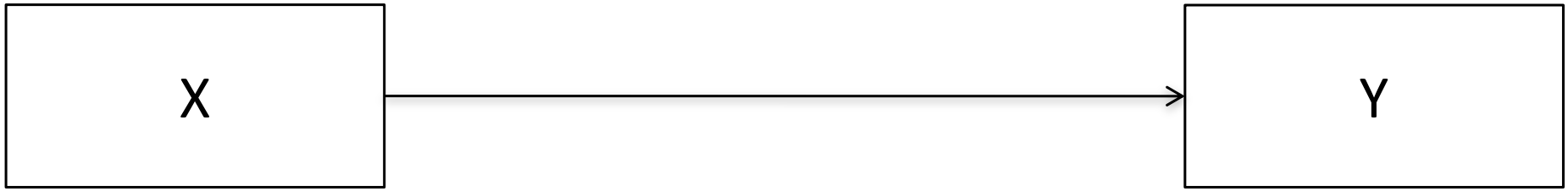
http://a.tgcdn.net/images/products/additional/large/f2cc_laboratory_shot_glasses_inuse.jpg

How to make a confident claim of X leads to Y

(Understanding randomized experiment, a.k.a., A/B Testing)

Elements of Causal Effects

$$y = f\ (x)$$



Three Necessary (NOT sufficient) Conditions of Causality

Three Necessary (NOT sufficient) Conditions of Causality

1. Correlation

Three Necessary (NOT sufficient) Conditions of Causality

1. Correlation

2. Time Lag (Temporal Precedence)

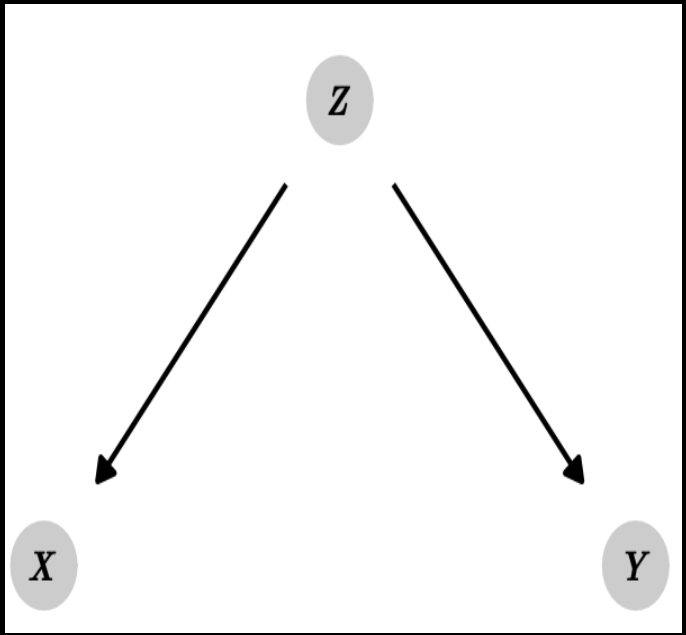
Three Necessary (NOT sufficient) Conditions of Causality

1. Correlation

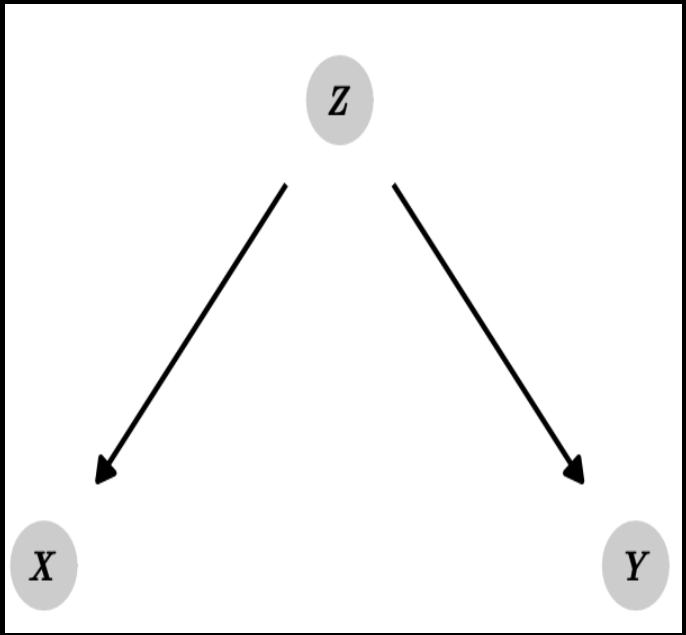
2. Time Lag (Temporal Precedence)

3. Non-Spuriousness
   (Addressing Confounders)

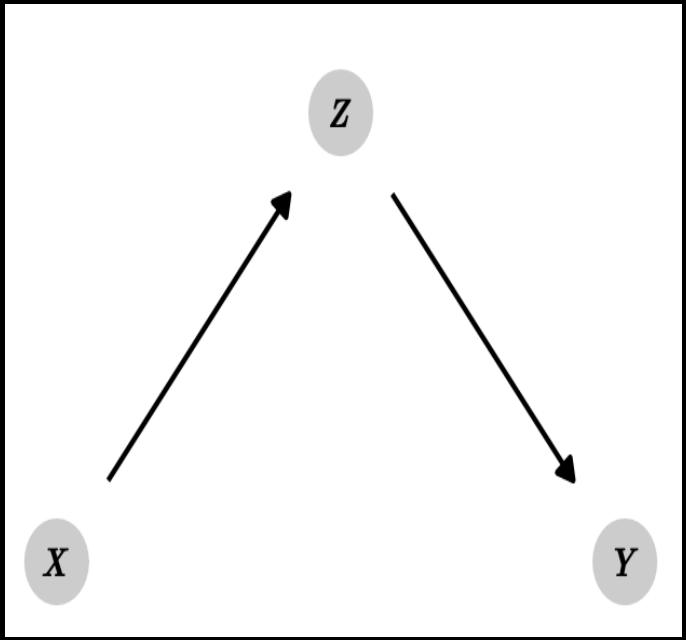
Confounder



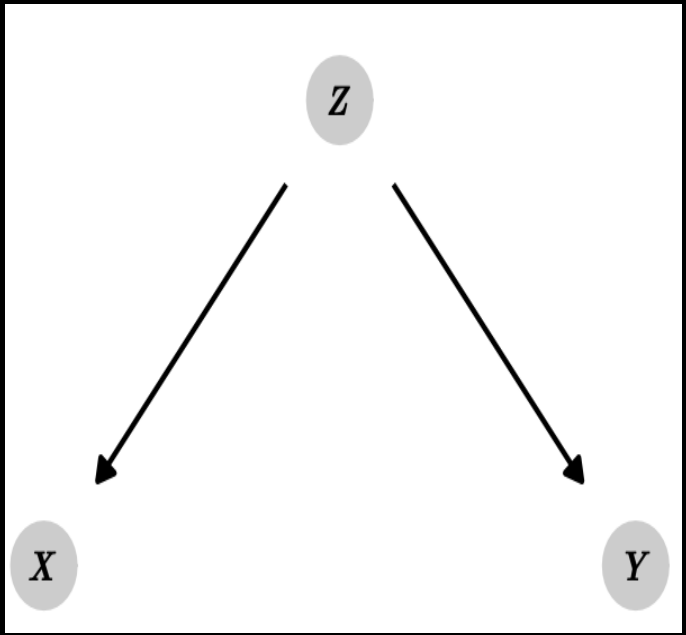
Confounder



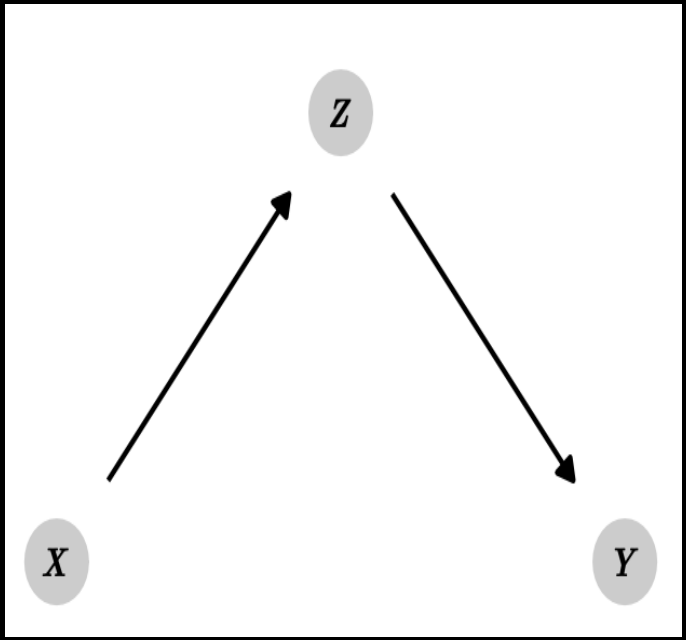
Mediator



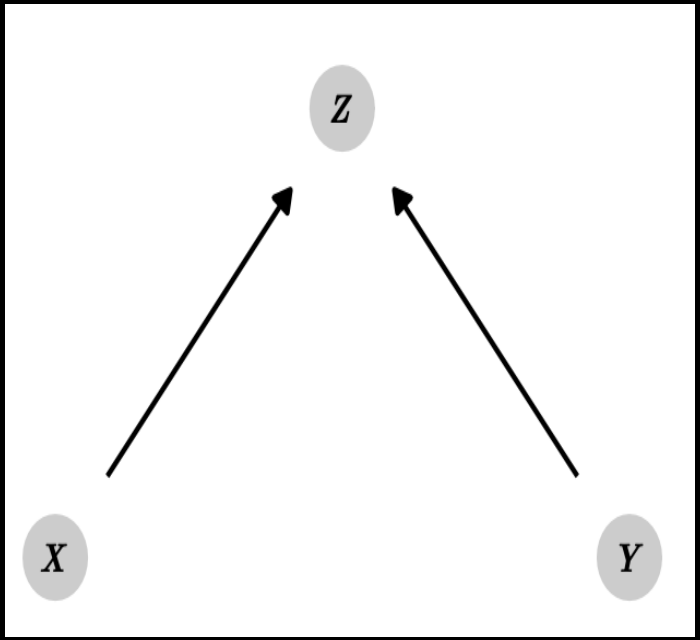
Confounder



Mediator



Collider



Data Collection Methods (Study Design)

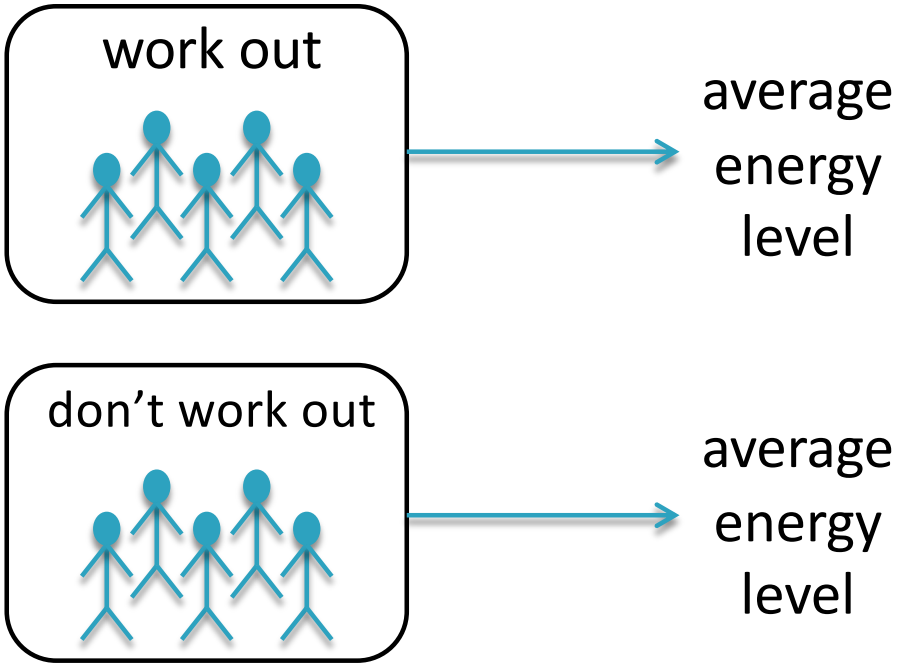
Observational Studies	Experimental Studies
Collect data in a way that does not directly interfere with how the data arise (“observe”)	Randomly assign subjects to treatments
Only establish an association	Establish causal connections

Q: Evaluate the relationship between regularly working out and energy level

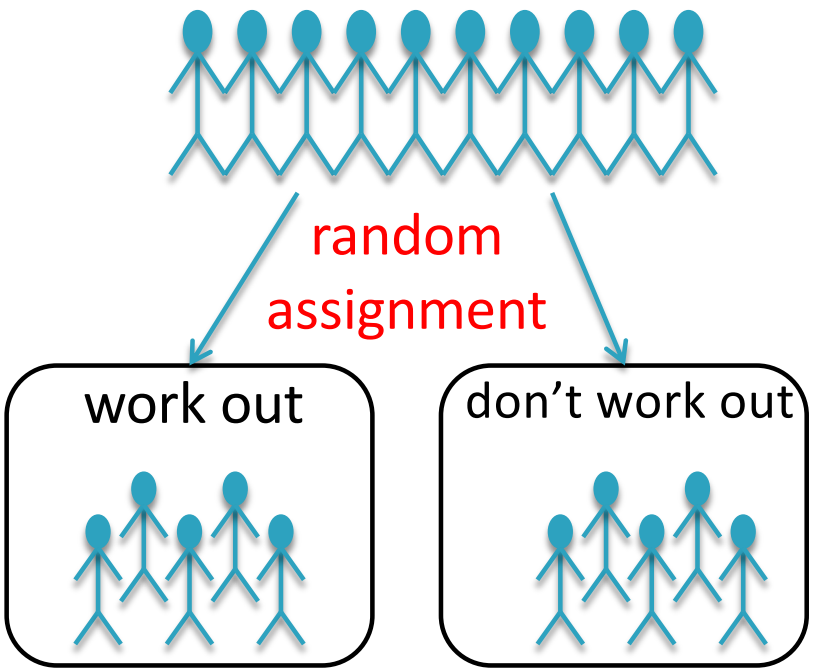
Random Assignment

Q: Evaluate the relationship between regularly working out and energy level

Observational Studies



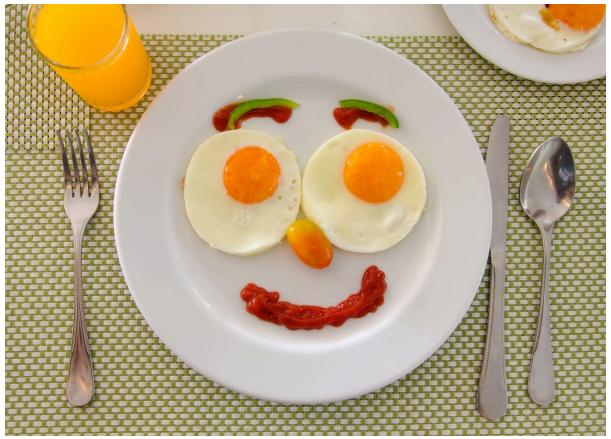
Experimental Studies

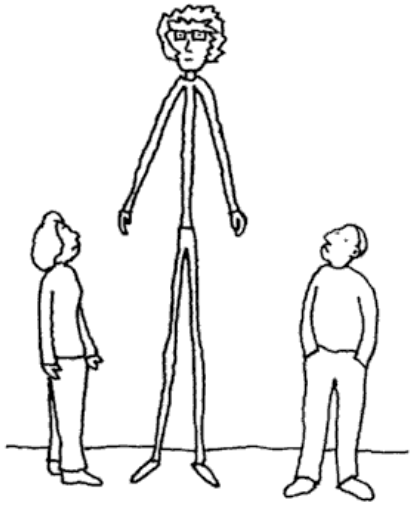


Let’s do some exercise.

3 Plausible Explanations

Eating breakfast causes girls to be slimmer



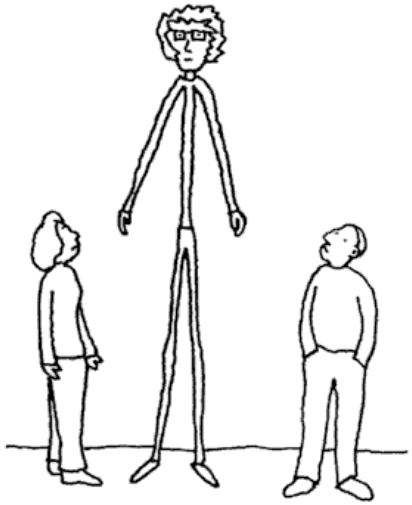


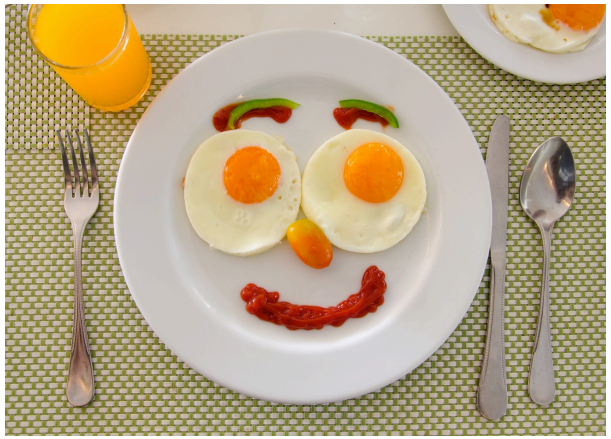
http://pas-wordpress-media.s3.amazonaws.com/wp-content/uploads/2014/03/Culture-Eats-Strategy-For-Breakfast.jpg

http://2.bp.blogspot.com/-b_7DdJVRo_A/UsUGJfKRfBI/AAAAAAAACWg/vzKksyjLvTI/s1600/me-tall.gif

3 Plausible Explanations

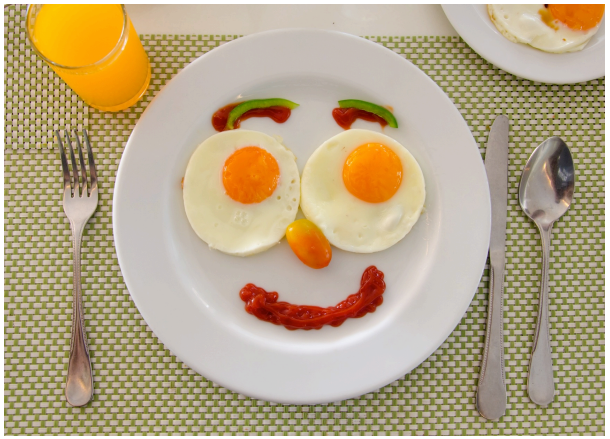
Being slim causes girls to eat breakfast

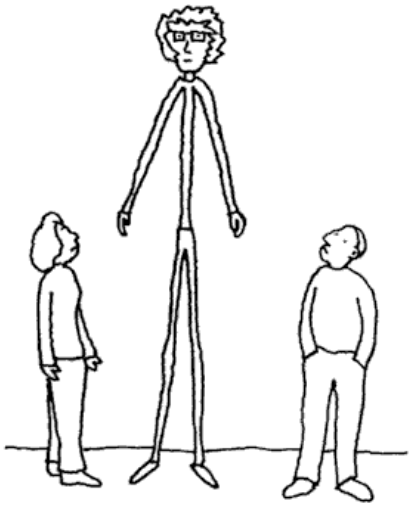




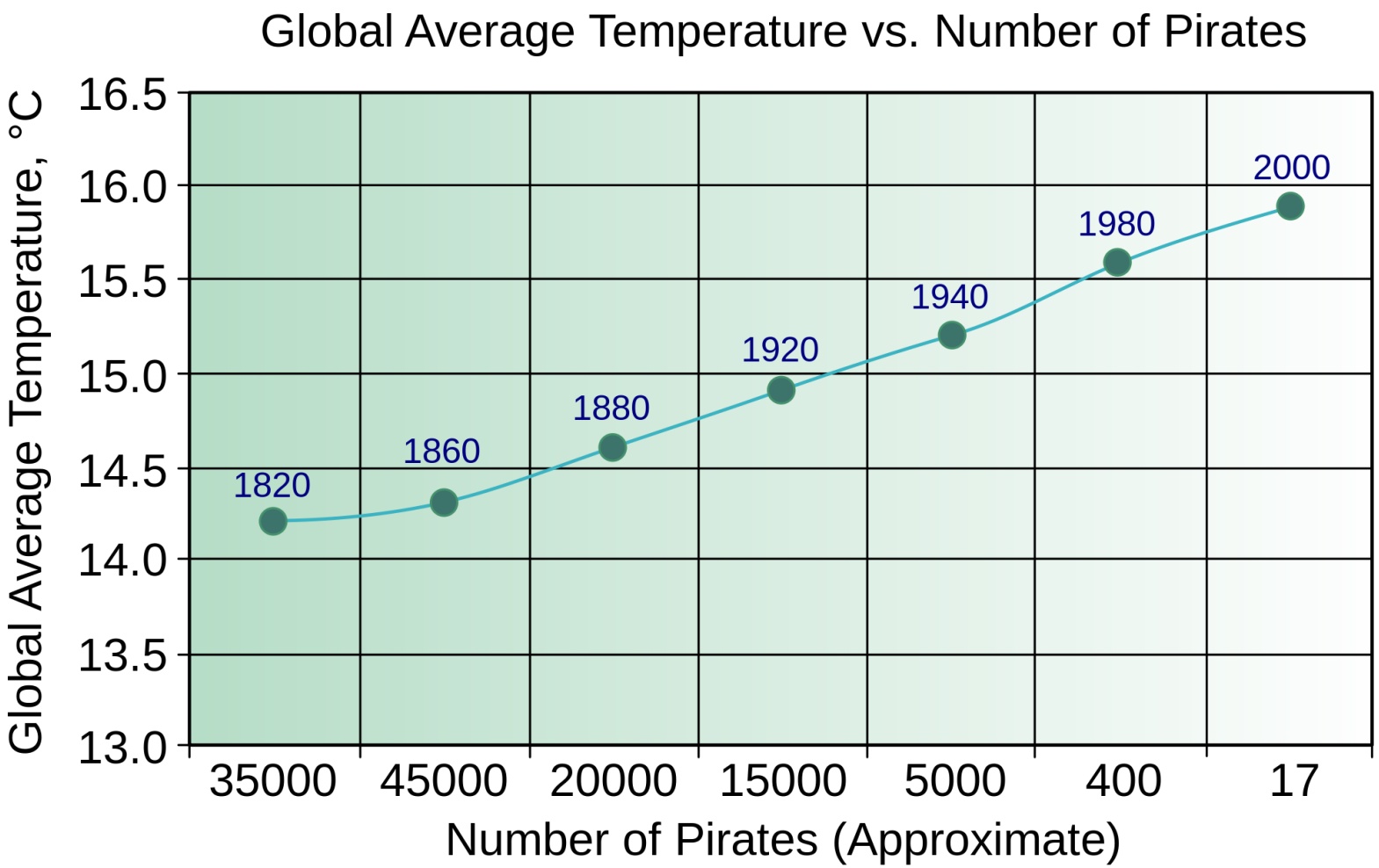
Extraneous variables that affect both the independent and the dependent variable, and that make it seem like there is a relationship between them

Confounding Variables





Correlation Does NOT Imply Causation



https://upload.wikimedia.org/wikipedia/commons/thumb/d/de/PiratesVsTemp(en).svg/2000px-PiratesVsTemp(en).svg.png

Some executives still mistakenly believe that all they need to do is establish correlation, and causality can be inferred.

Oh Dear… That's Wrong T.T

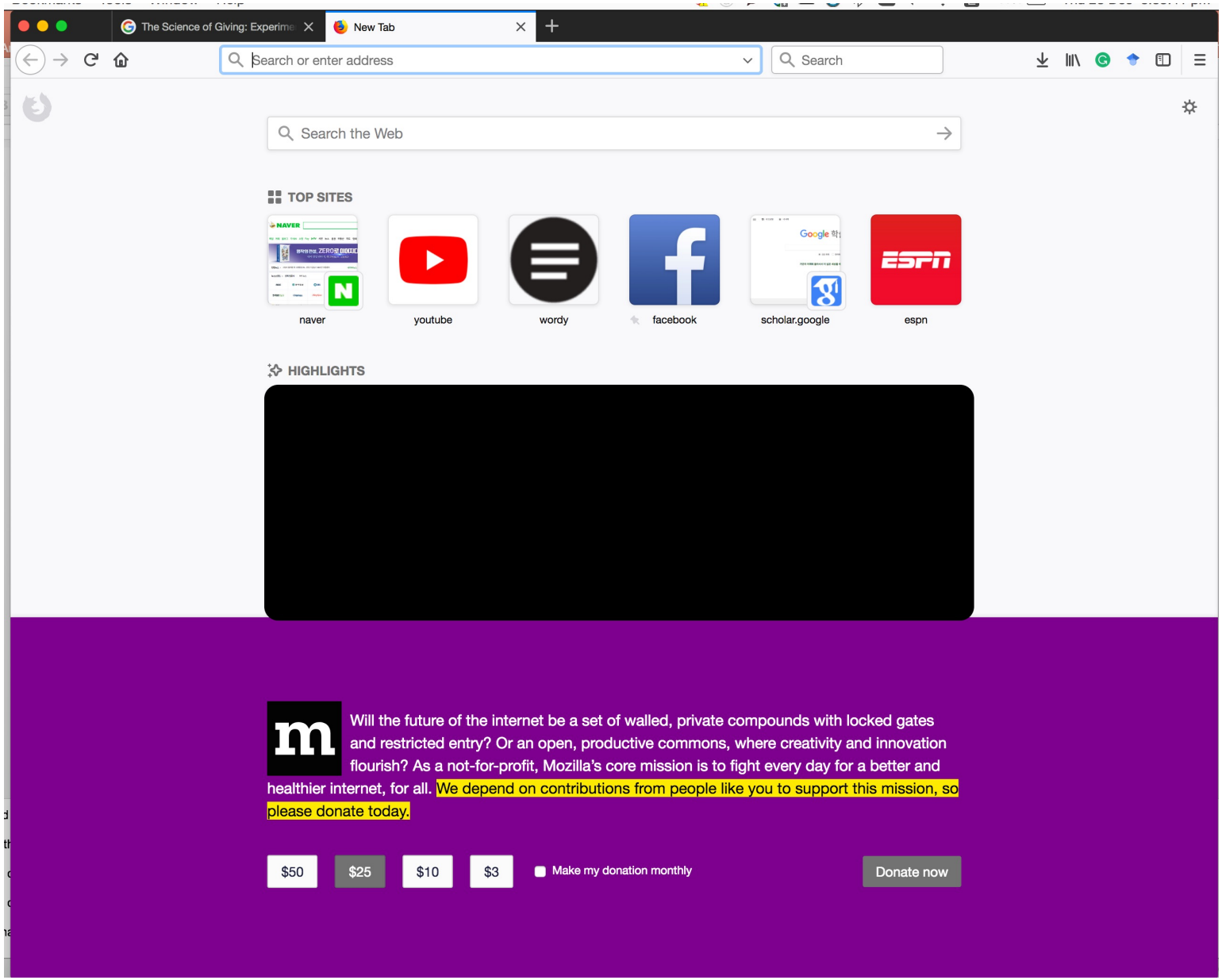
Let’s do some further exercise.

Web Bowser & Job Performance

There is good evidence that
Firefox and Chrome users significantly
outperform Internet Explorer and Safari users.

They also stay in their jobs 15 percent longer.

Why? It's about how you got the browser.


The Science of Giving: Experime
New Tab
Search or enter address
Search
Search the Web
TOP SITES
naver
youtube
wordy
facebook
scholar.google
espn
HIGHLIGHTS
Will the future of the internet be a set of walled, private compounds with locked gates and restricted entry? Or an open, productive commons, where creativity and innovation flourish? As a not-for-profit, Mozilla's core mission is to fight every day for a better and healthier internet, for all. We depend on contributions from people like you to support this mission, so please donate today.
$50
$25
$10
$3
Make my donation monthly
Donate now

Bookmarks Tools Window Help

R Korea - KRUG(Korean R User | New Tab

Search or enter address | Search

Search the Web

TOP SITES

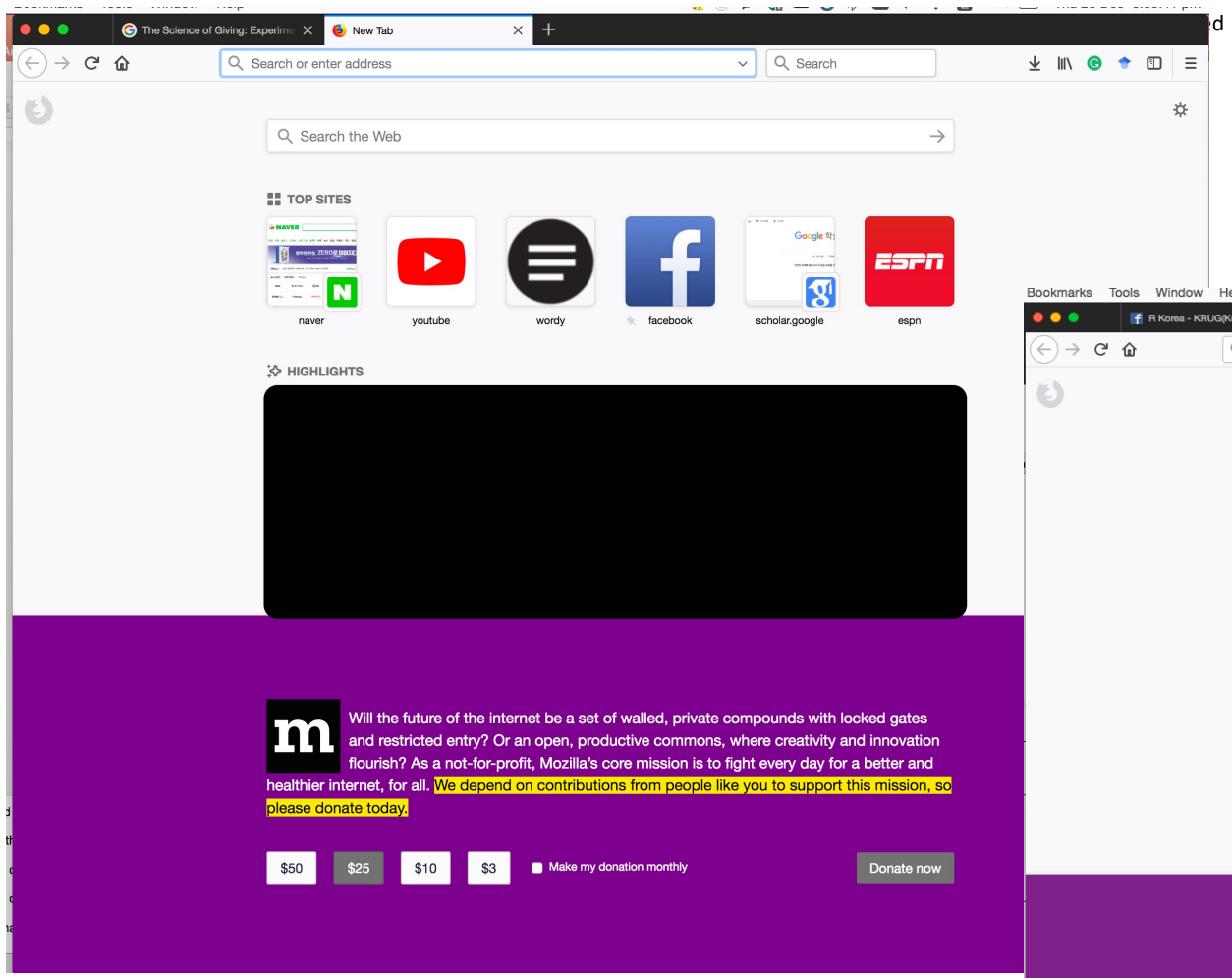
naver | youtube | scholar.google | facebook | wordy | espn

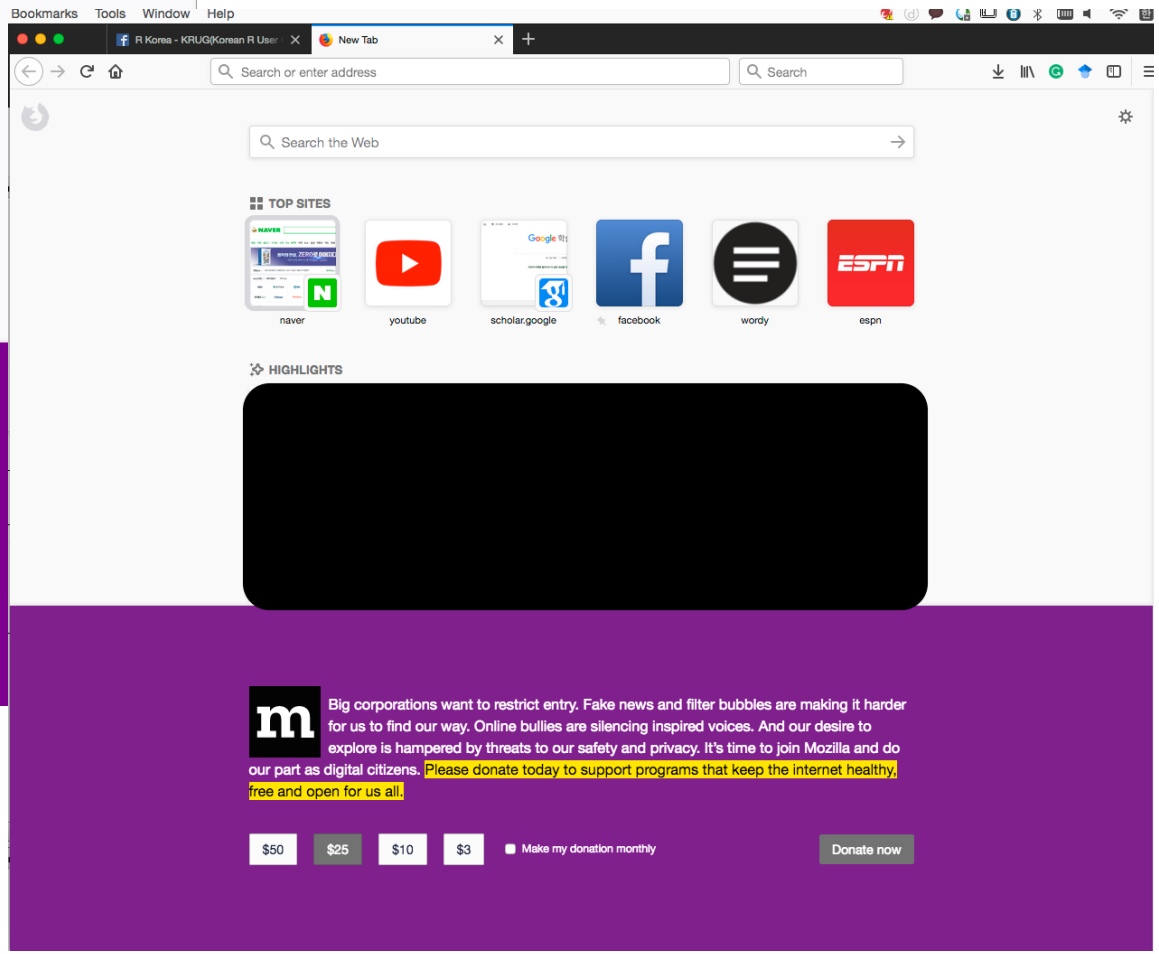
HIGHLIGHTS

[redacted]

Big corporations want to restrict entry. Fake news and filter bubbles are making it harder for us to find our way. Online bullies are silencing inspired voices. And our desire to explore is hampered by threats to our safety and privacy. It's time to join Mozilla and do our part as digital citizens. Please donate today to support programs that keep the internet healthy, free and open for us all.

$50 | $25 | $10 | $3 | Make my donation monthly | Donate now


The Science of Giving: Experime
New Tab
Search or enter address
Search
Search the Web
TOP SITES
naver
youtube
wordy
facebook
scholar.google
espn
HIGHLIGHTS
Will the future of the internet be a set of walled, private compounds with locked gates and restricted entry? Or an open, productive commons, where creativity and innovation flourish? As a not-for-profit, Mozilla's core mission is to fight every day for a better and healthier internet, for all. We depend on contributions from people like you to support this mission, so please donate today.
$50
$25
$10
$3
Make my donation monthly
Donate now


Bookmarks
Tools
Window
Help
R Korea - KRUG(Korean R User
New Tab
Search or enter address
Search
Search the Web
TOP SITES
naver
youtube
scholar.google
facebook
wordy
espn
HIGHLIGHTS
Big corporations want to restrict entry. Fake news and filter bubbles are making it harder for us to find our way. Online bullies are silencing inspired voices. And our desire to explore is hampered by threats to our safety and privacy. It's time to join Mozilla and do our part as digital citizens. Please donate today to support programs that keep the internet healthy, free and open for us all.
$50
$25
$10
$3
Make my donation monthly
Donate now

“Experiment with Everything” Approach

Today, Microsoft and several other leading companies (including Amazon, Booking.com, Facebook, and Google) each conduct more than 10,000 online controlled experiments annually, with many tests engaging millions of users.

A

B

“Experiment with Everything” Approach

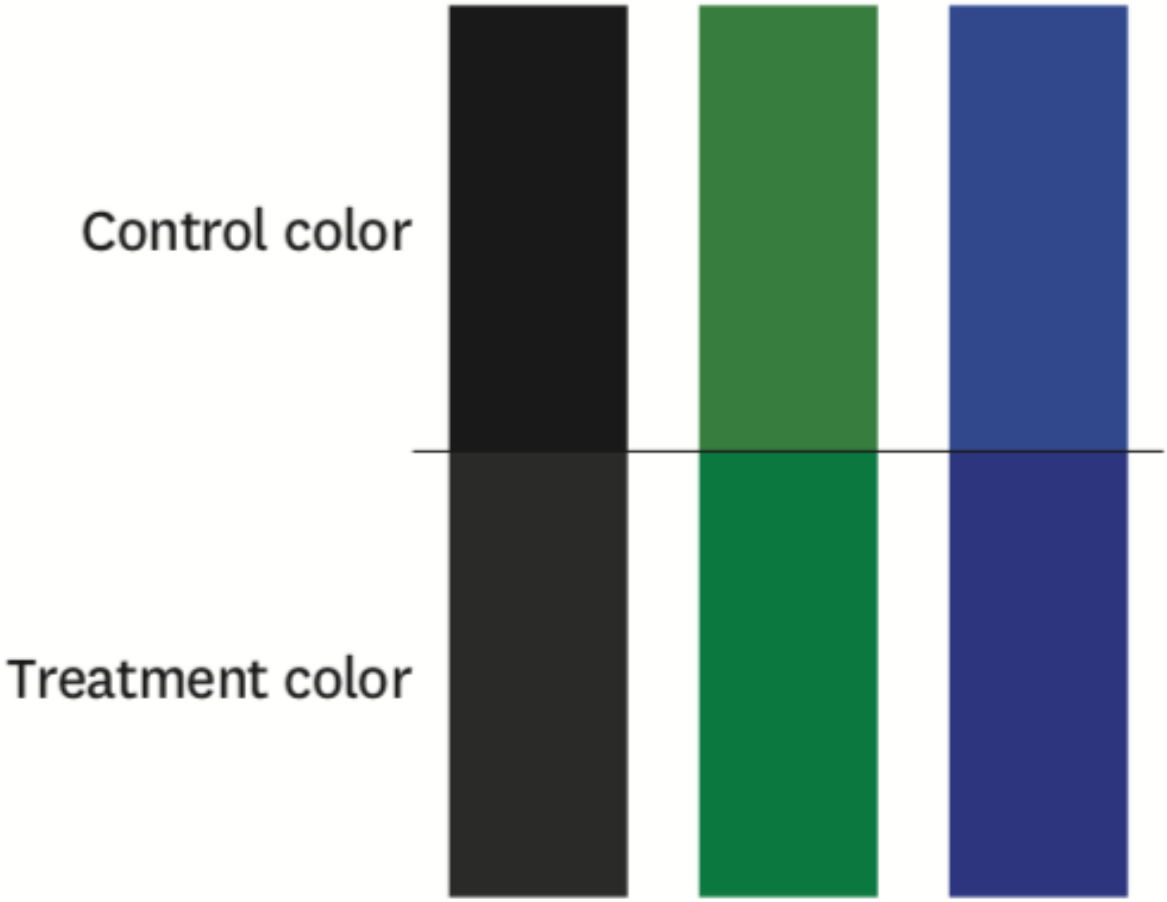
Start-ups and companies without digital roots, such as Walmart, Hertz, and *Singapore Airlines*, also run them regularly, though on a smaller scale.

“Experiment with Everything” Approach

These organizations have discovered that an “experiment with everything” approach has surprisingly large payoffs.

Case Study:
Bing’s small changes with a HUGE impact

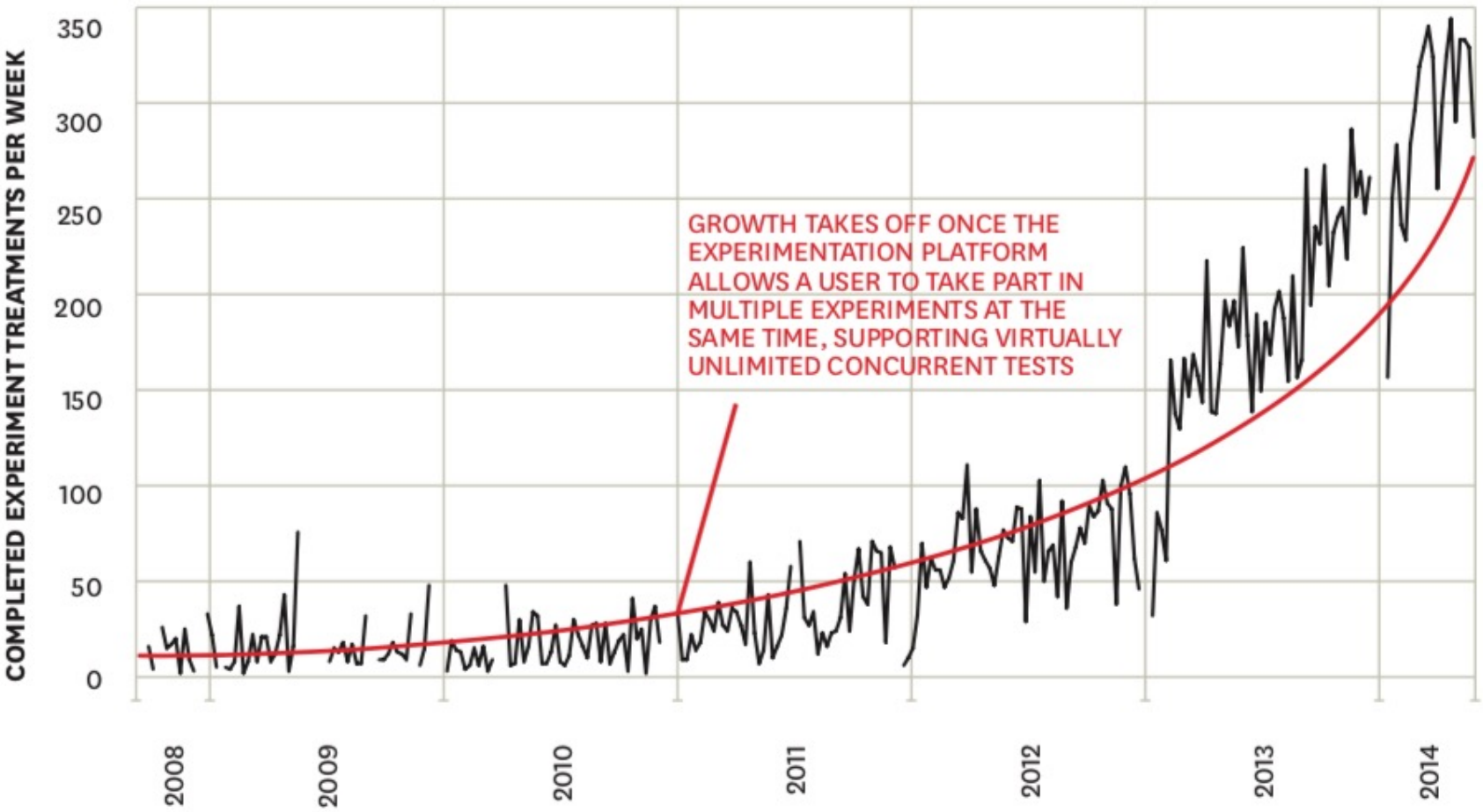
A



B

- Bing's experiments showed that slightly darker blues and greens in titles and a slightly lighter black in captions improved the users' experience.
- When rolled out to all users, the color changes boosted revenue by more than $10 million annually.

The Growth of Experimentation (A/B Testing) at Bing



A Case from *booking.com*

(Thomke, 2020)

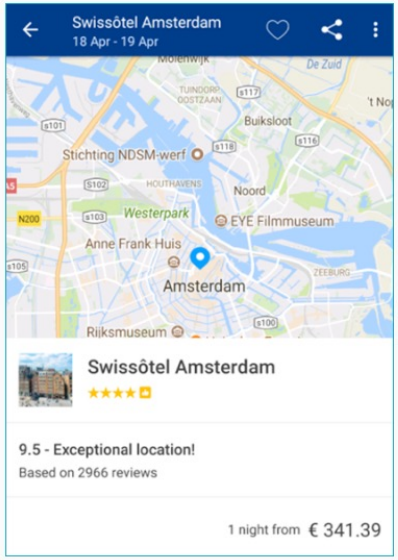
Scenario #1

Hypothesis

Highlighting a neighborhood's walkability helps users make better decisions about property location.

A: The Control

Shows the site's current practice



A Case from *booking.com*

(Thomke, 2020)

Scenario #1

Hypothesis

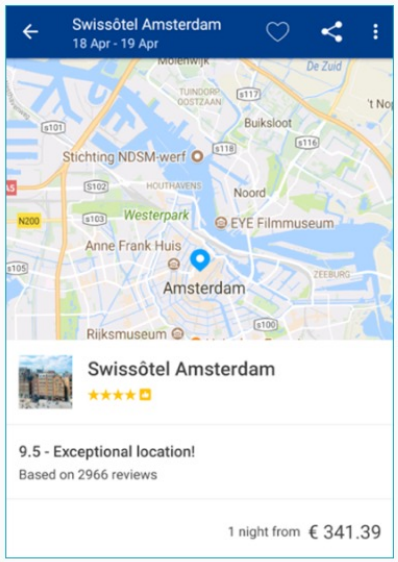
Highlighting a neighborhood's walkability helps users make better decisions about property location.

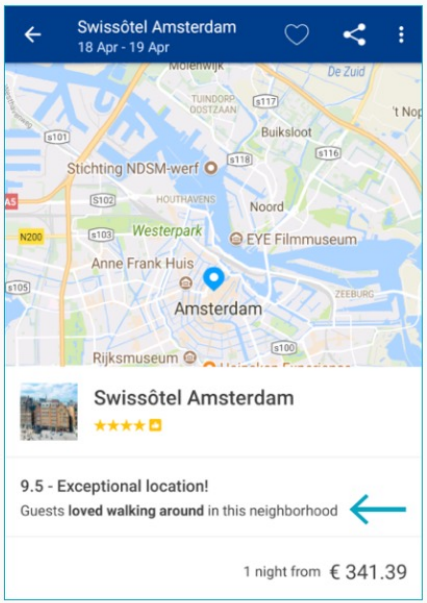
A: The Control

Shows the site's current practice

B: The Treatment

Adds walkability information





A Case from *booking.com*

(Thomke, 2020)

Scenario #1

Hypothesis

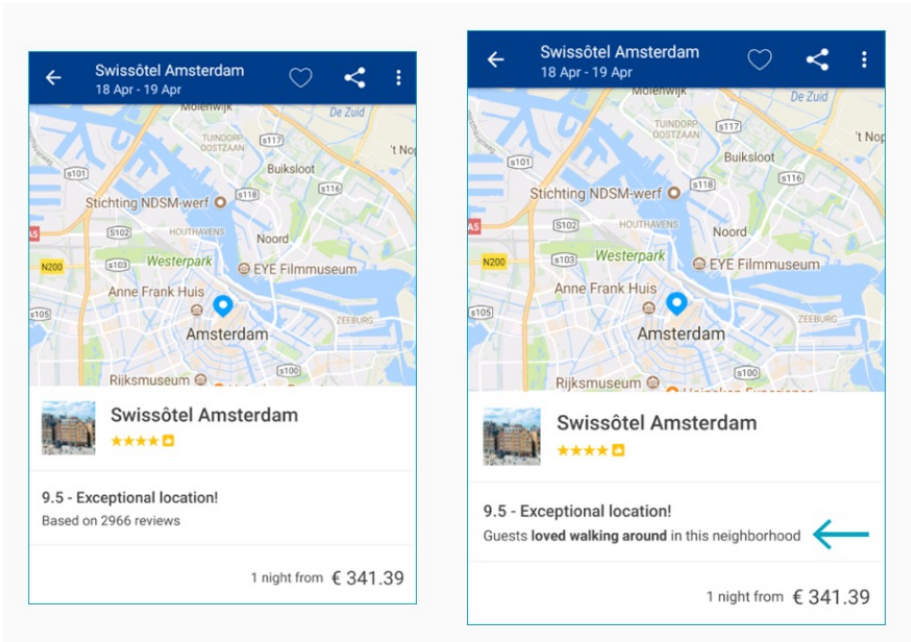
Highlighting a neighborhood's walkability helps users make better decisions about property location.

A: The Control

Shows the site's current practice

B: The Treatment

Adds walkability information



The Result

- The treatment had no significant impact on the key metric.
- The current practice is kept in place.

A Case from *booking.com*

(Thomke, 2020)

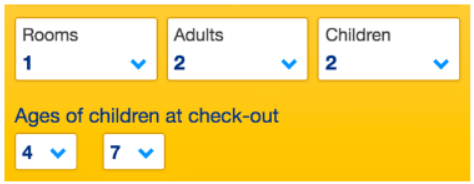
Scenario #2

Hypothesis

Displaying the checkout date when users select the age of children in their party improves their experience.

A: The Control

Shows the site's current practice



A Case from *booking.com*

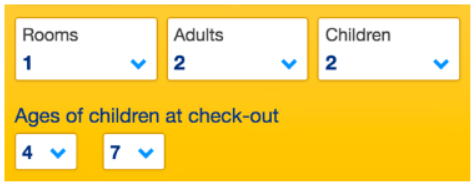
(Thomke, 2020)

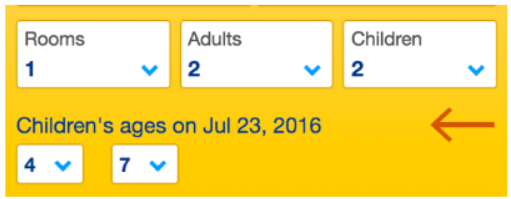
Scenario #2

Hypothesis

Displaying the checkout date when users select the age of children in their party improves their experience.

A: The Control	B: The Treatment
Shows the site's current practice	Adds the checkout date above children's ages



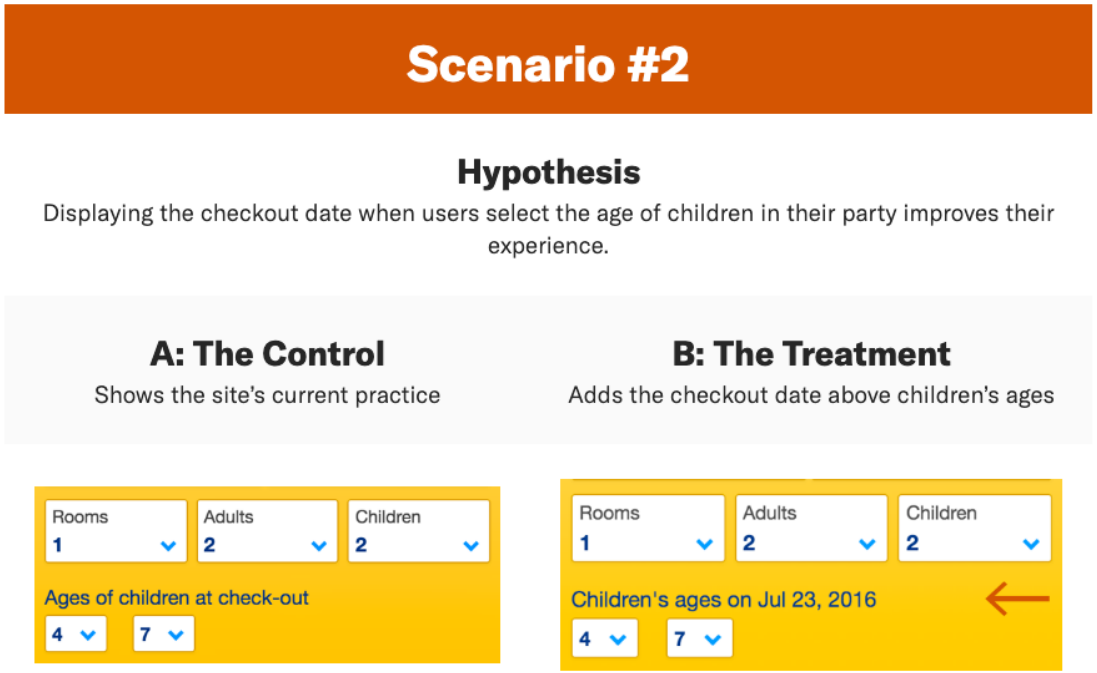


A Case from *booking.com*

(Thomke, 2020)

The Result

- The treatment had a significant positive impact on the key metric, and the change is implemented.



Interventions, A/B Testing, and Causal Inference

- When creating corporate strategies, many companies make decisions—on everything from new product features, to look and feel, to marketing campaigns— using subjective opinions rather than hard data.

Interventions, A/B Testing, and Causal Inference

- When creating corporate strategies, many companies make decisions—on everything from new product features, to look and feel, to marketing campaigns— using subjective opinions rather than hard data.
- Companies should conduct controlled experiments (A/B tests) to evaluate their ideas.

Interventions, A/B Testing, and Causal Inference

- When creating corporate strategies, many companies make decisions—on everything from new product features, to look and feel, to marketing campaigns— using subjective opinions rather than hard data.
- Companies should conduct controlled experiments (A/B tests) to evaluate their ideas.
- Potential improvements should be rigorously tested, because large investments can fail to deliver, and some tiny changes can be surprisingly detrimental while others have big payoffs.

Interventions, A/B Testing, and Causal Inference

Executives should understand how to properly design and execute A/B tests and other controlled experiments, ensure their integrity, interpret their results, and avoid pitfalls.

Insist that Data trump Opinions

"In an increasingly digital world, if you don't do large-scale experimentation, in the long term—and in many industries the short term—you're dead."

Mark Okerstrom, the former CEO of Expedia Group

Insist that Data trump Opinions

"In an increasingly digital world, if you don't do large-scale experimentation, in the long term—and in many industries the short term—you're dead."

"At any one time we're running hundreds, if not thousands, of concurrent experiments, involving millions of visitors. Because of this, we don't have to guess what customers want; we have the ability to run the most massive 'customer surveys' that exist, again and again, to have them tell us what they want."

Insist that Data trump Opinions

"In an increasingly digital world, if you don't do large-scale experimentation, in the long term—and in many industries the short term—you're dead."

"At any one time we're running hundreds, if not thousands, of concurrent experiments, involving millions of visitors. Because of this, we don't have to guess what customers want; we have the ability to run the most massive 'customer surveys' that exist, again and again, to have them tell us what they want."

Mark Okerstrom, the former CEO of Expedia Group

Key Elements of Experimental Design

- **Control**: Compare treatment of interest to a control group.
- **Randomize**: Randomly assign subjects to treatments, and randomly sample from the population whenever possible.
- **Replicate**: Within a study, replicate by collecting a sufficiently large sample. Or replicate the entire study.

Tracing Digital Footprints: Google Search

- Google Trends is a valuable tool for people working in the finance industry, as it provides insights into search patterns and public interest in various topics over time.
- This data can help finance professionals identify emerging trends, gauge consumer sentiment, and monitor market dynamics.

Tracing Digital Footprints: Google Search

- Google Trends is a valuable tool for people working in the finance industry, as it provides insights into search patterns and public interest in various topics over time.
- This data can help finance professionals identify emerging trends, gauge consumer sentiment, and monitor market dynamics.

- When do you want to launch campaign for physical exercise?
- When do you want to market diet products?

Agenda (Workflow of Data-Driven Decision Making)

Define

Collect

Import

Wrangle

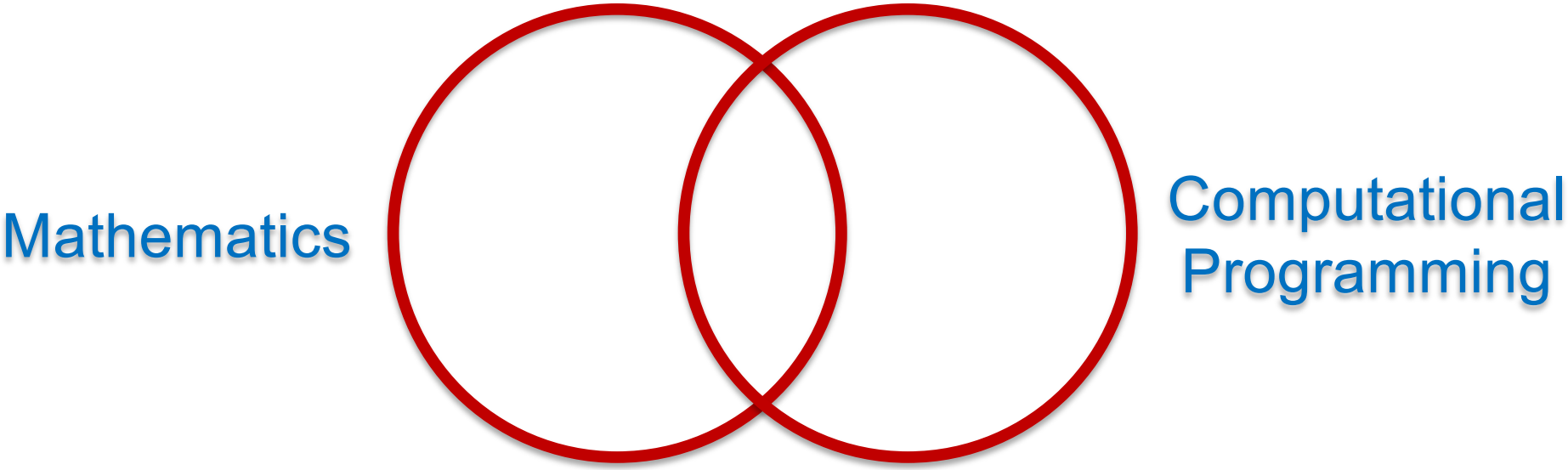
Visualize

Model

Report

Deploy

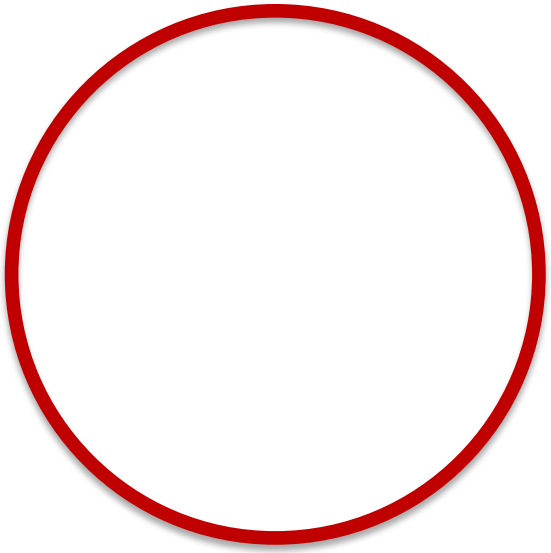
Three Essential Skillsets of Successful Data-Driven Decision Making



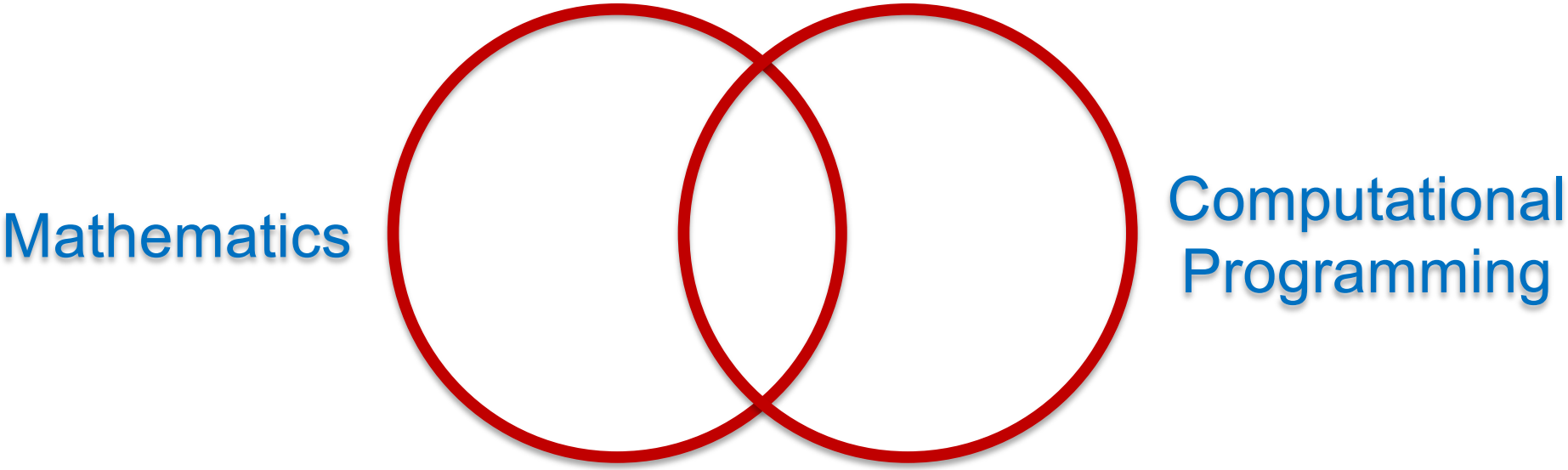
Three Essential Skillsets of Successful Data-Driven Decision Making

Linear Algebra
Calculus
Statistics
Probability
Topology

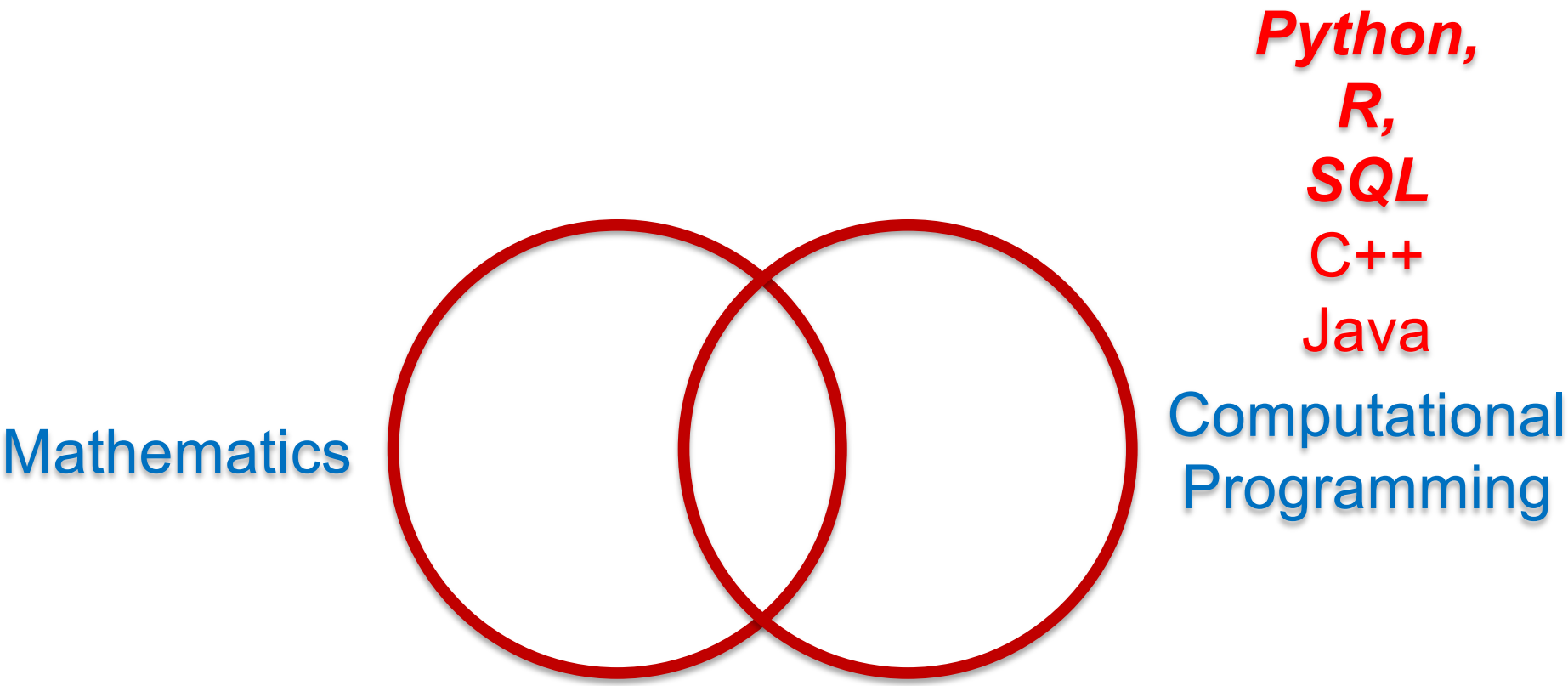
Mathematics



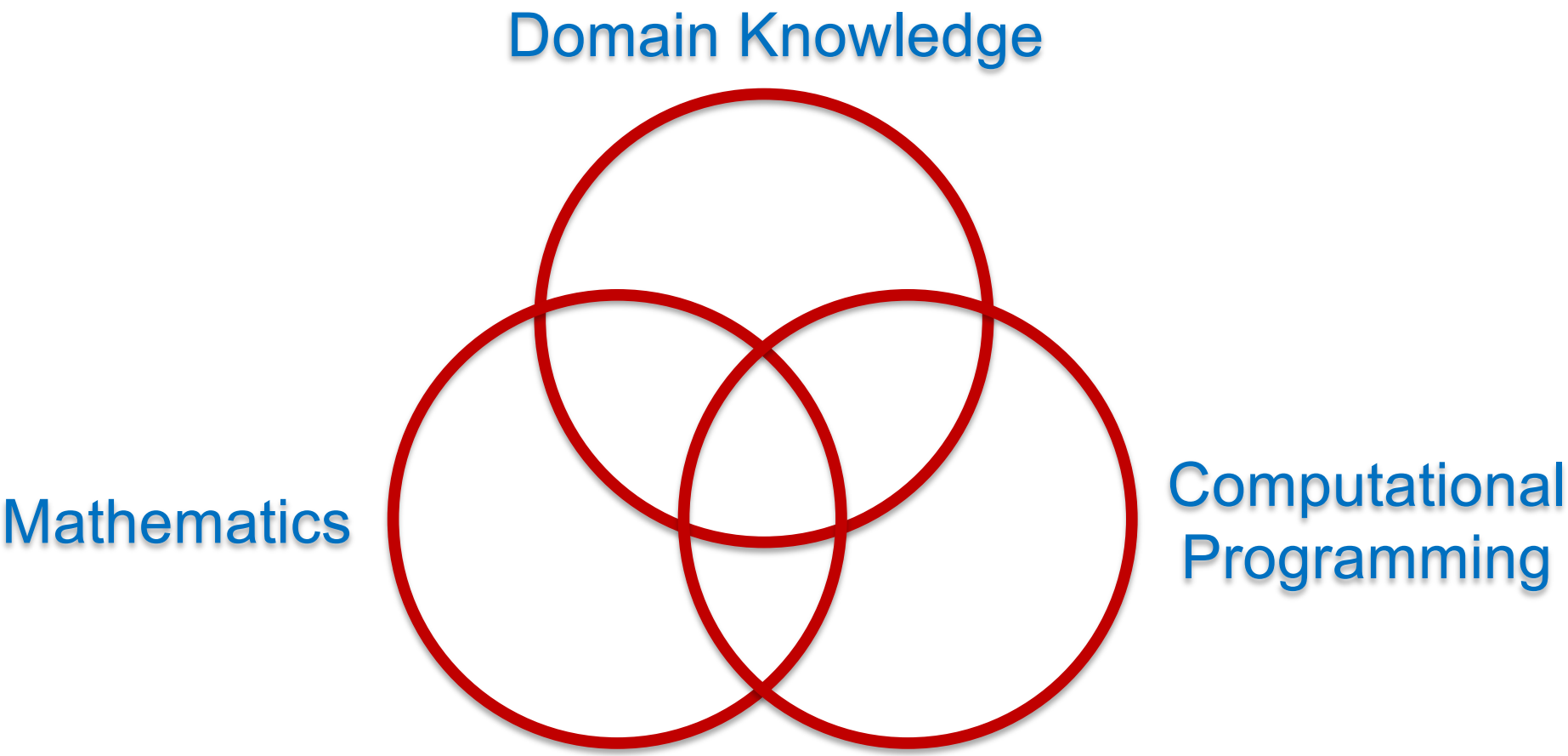
Three Essential Skillsets of Successful Data-Driven Decision Making



Three Essential Skillsets of Successful Data-Driven Decision Making



Three Essential Skillsets of Successful Data-Driven Decision Making



Agenda (Workflow of Data-Driven Decision Making)

Define

Collect

Import

Wrangle

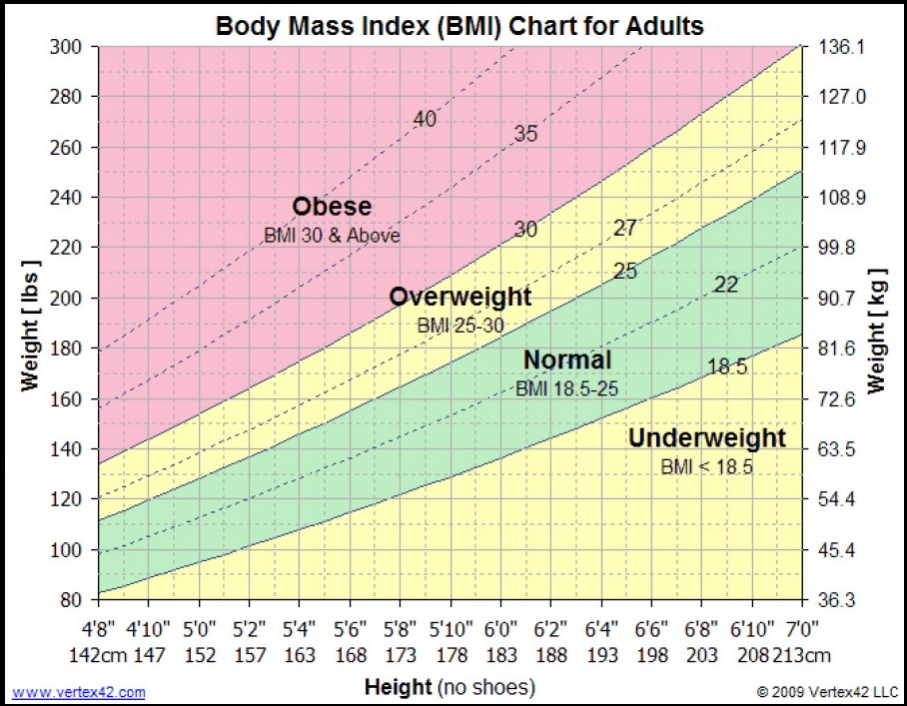
Visualize

Model

Report

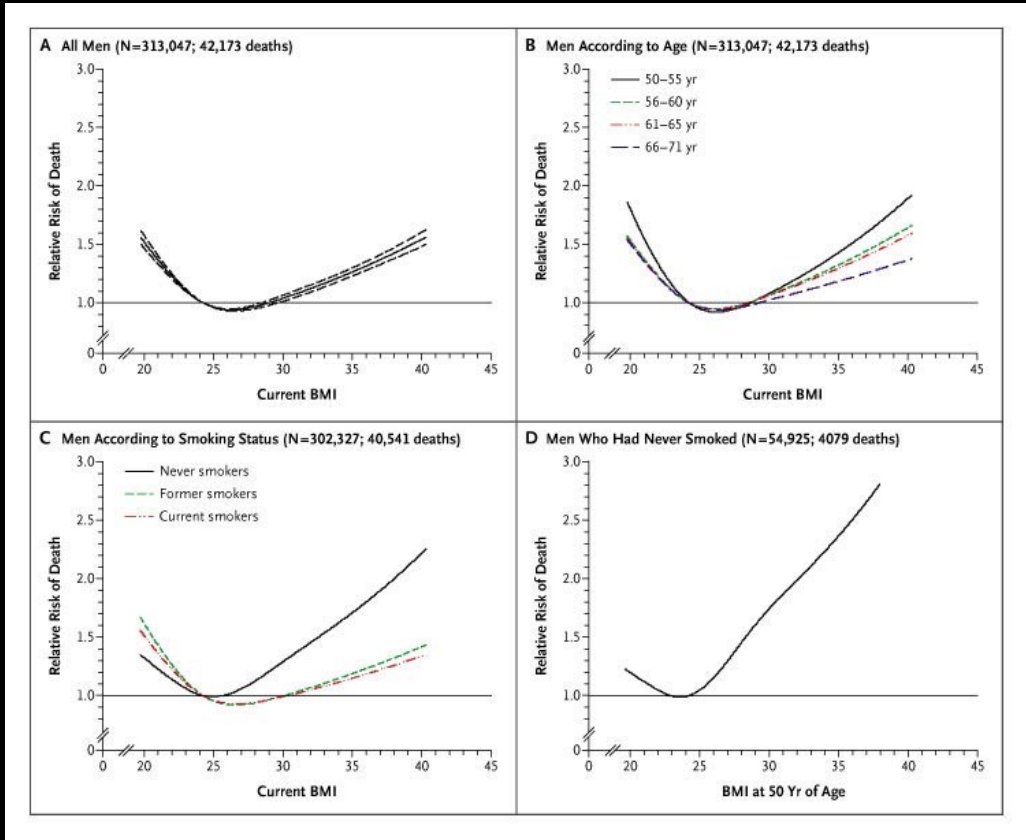
Deploy

(What) Do we know about Body Mass Index?

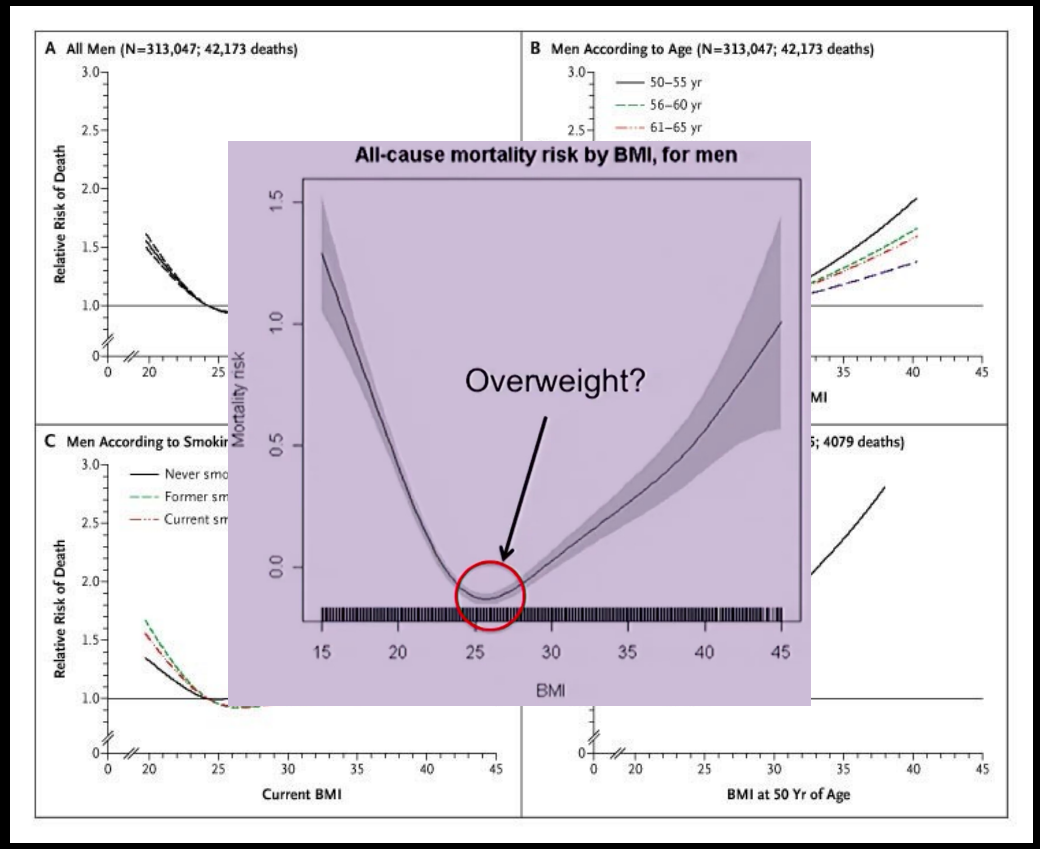


What would be the relationships between BMI and mortality?

The Truth about the Relationships between Body Mass Index and Mortality



The Truth about the Relationships between Body Mass Index and Mortality

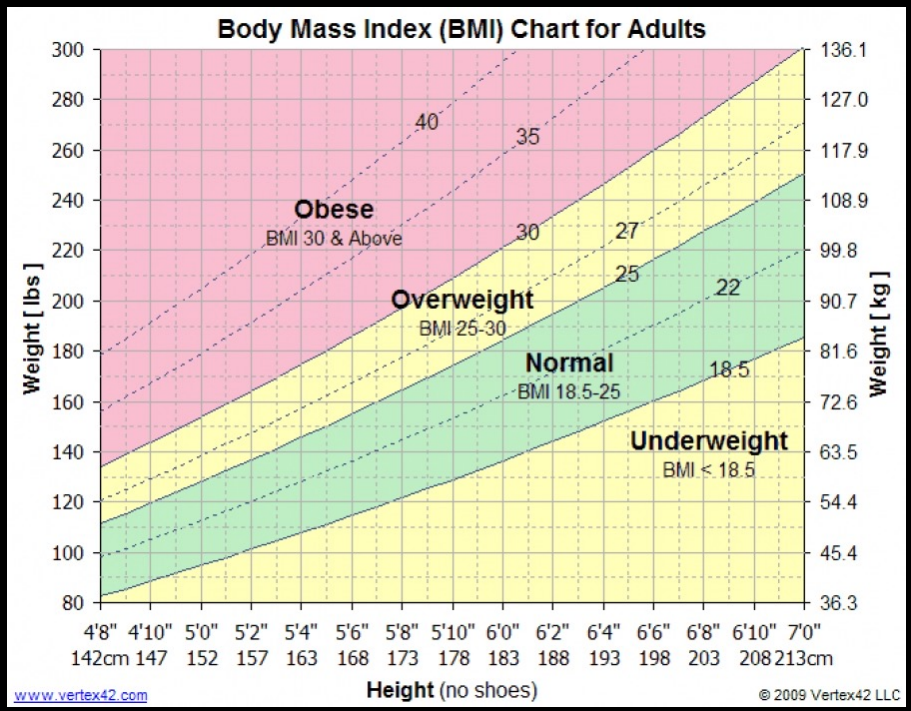


There is a need to define and re-define your key variables (indices)

Refining Our Understanding of "Index"

- Oops… Ohtani Shohei is overweight!?
- Note that body mass index is taking weight over height squared.

(What) Do we know about Body Mass Index?



There is a need to define and re-define your key variables (indices)

Refining Our Understanding of "Index"

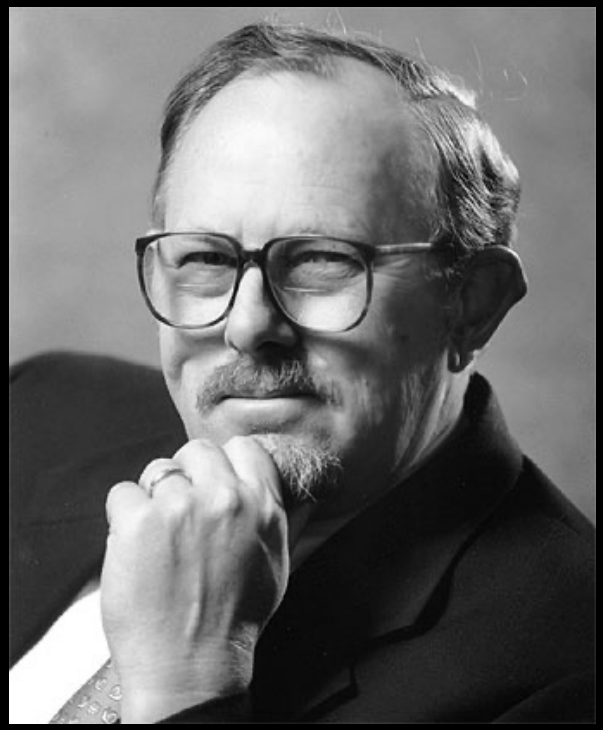
- Oops… Ohtani Shohei is overweight!?
- Note that body mass index is taking weight over height squared.
- Body Mass Index = $\frac{weight}{height^2}$

There is a need to define and re-define your key variables (indices)

Refining Our Understanding of "Index"

- Oops… Ohtani Shohei is overweight!?
- Note that body mass index is taking weight over height squared.
- Body Mass Index = $\frac{weight}{height^2}$
- The index is biased against tall people.
- The use of the squared term in the calculation can disadvantage taller people, and a more appropriate exponent, such as 2.3 based on height, should be considered.

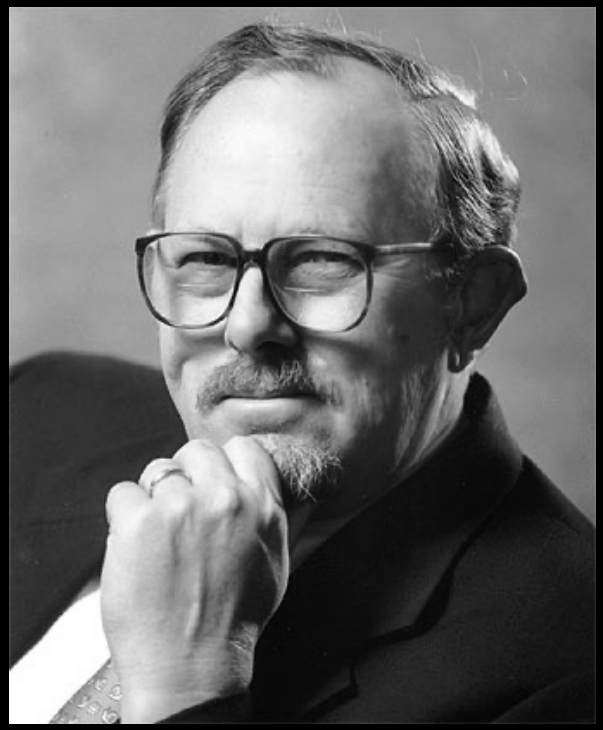
Speaking of index, let's consider a performance metric in assessing investment portfolios, namely Sharpe ratio.



Speaking of index, let's consider a performance metric in assessing investment portfolios, namely Sharpe ratio.

$$Sharpe\ ratio = \frac{R_p - R_f}{\sigma_p}$$

Hmm… The Sharpe ratio does treat upside and downside volatility equally, which can be problematic in certain situations.

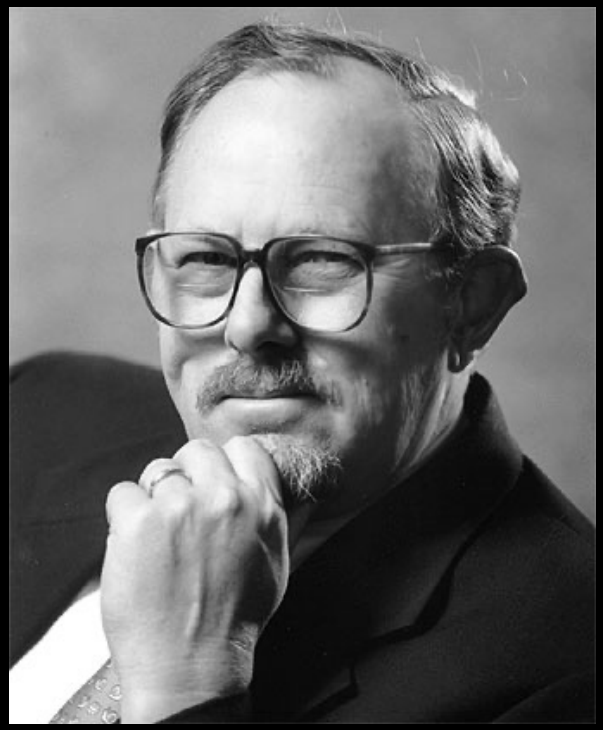


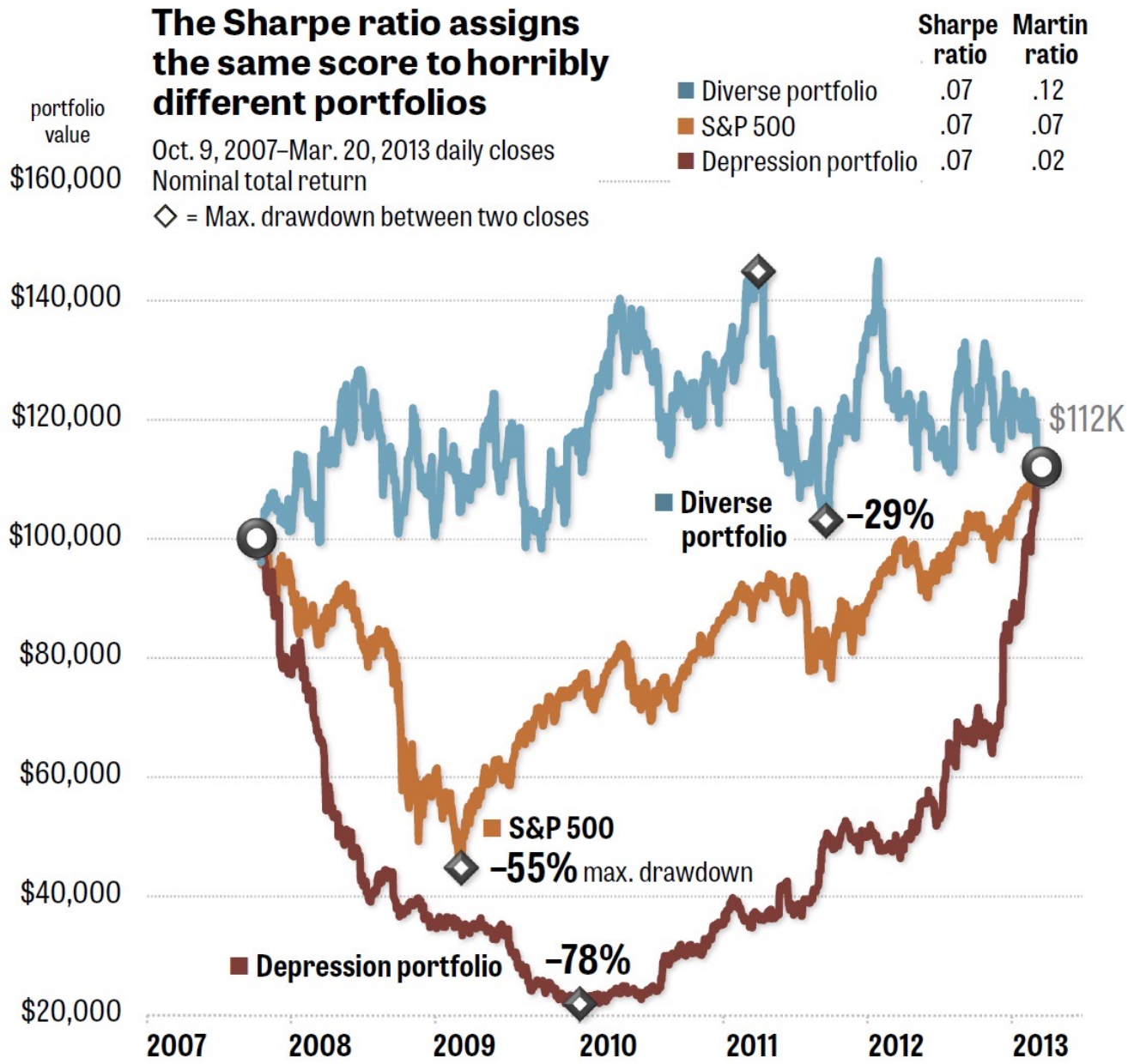
Speaking of index, let's consider a performance metric in assessing investment portfolios, namely Sharpe ratio.

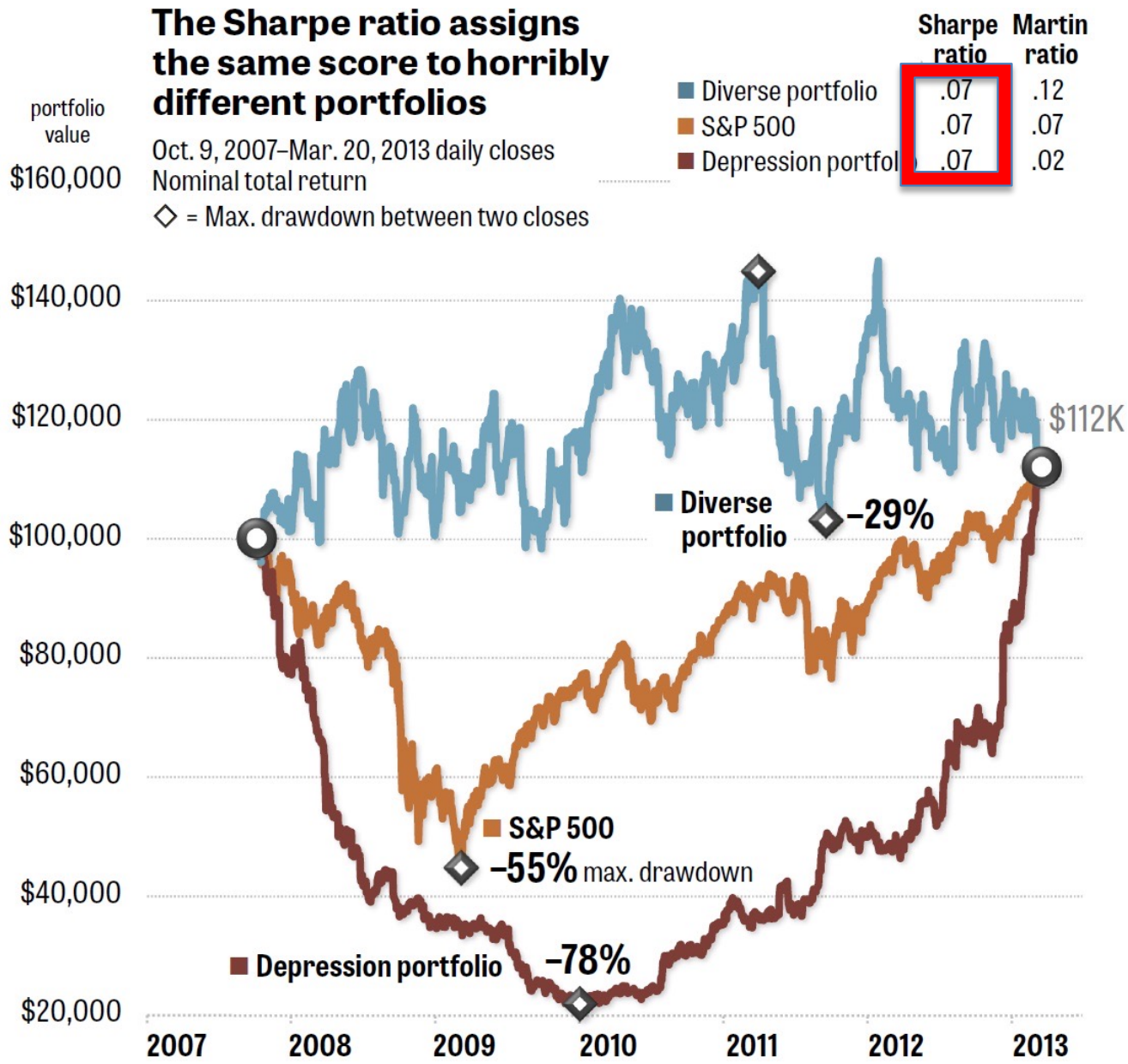
$$Sharpe\ ratio = \frac{R_p - R_f}{\sigma_p}$$

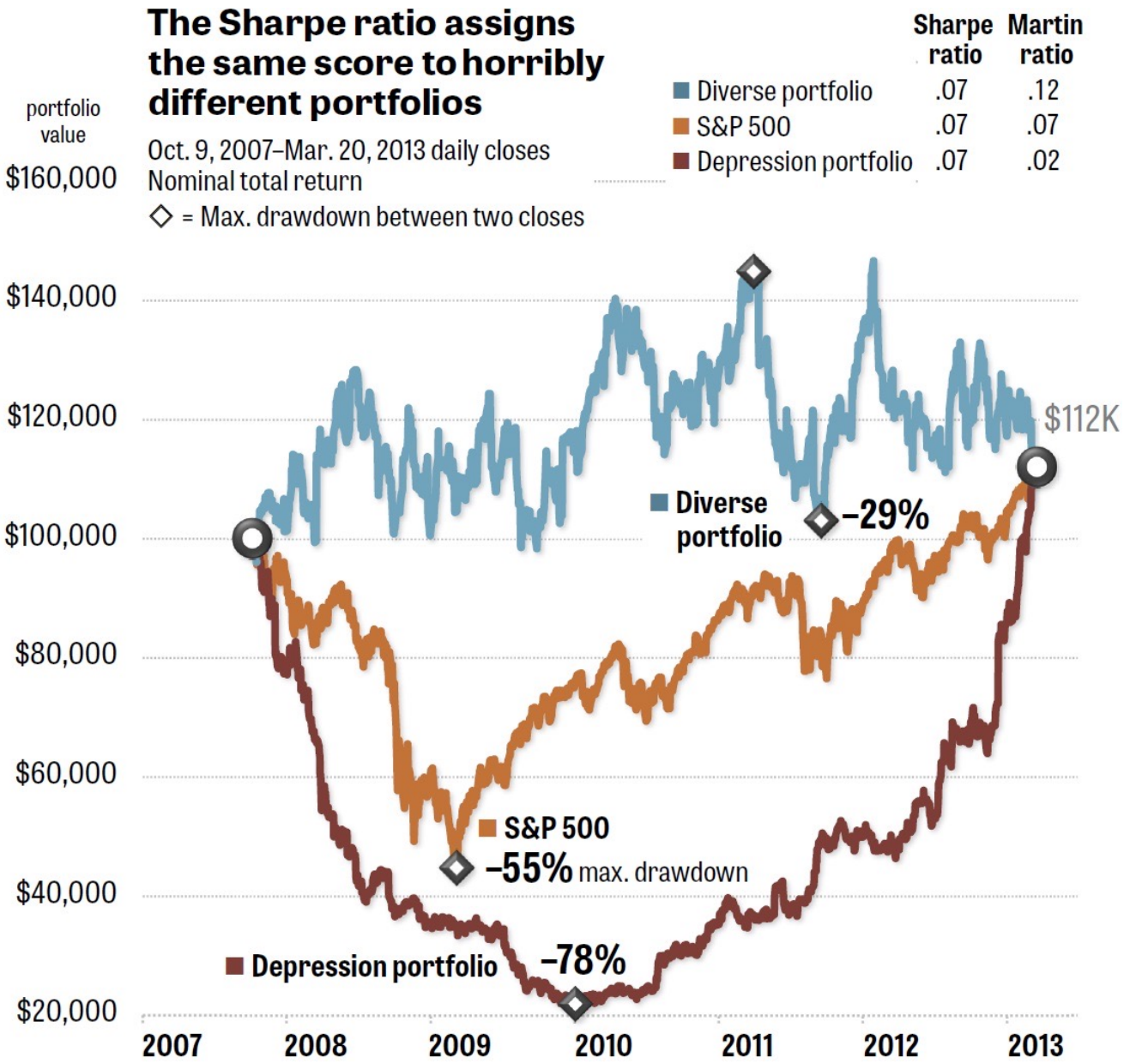
Hmm… The Sharpe ratio does treat upside and downside volatility equally, which can be problematic in certain situations.

Investors may be willing to tolerate more upside volatility (i.e., volatility that leads to higher returns) while being more concerned about downside volatility (i.e., volatility that leads to lower returns).




The Sharpe ratio assigns the same score to horribly different portfolios
Oct. 9, 2007–Mar. 20, 2013 daily closes
Nominal total return
◇ = Max. drawdown between two closes
Sharpe ratio
Martin ratio
Diverse portfolio .07 .12
S&P 500 .07 .07
Depression portfolio .07 .02
portfolio value
$160,000
$140,000
$120,000
$100,000
$80,000
$60,000
$40,000
$20,000
2007
2008
2009
2010
2011
2012
2013
$112K
Diverse portfolio
–29%
S&P 500
–55% max. drawdown
Depression portfolio
–78%

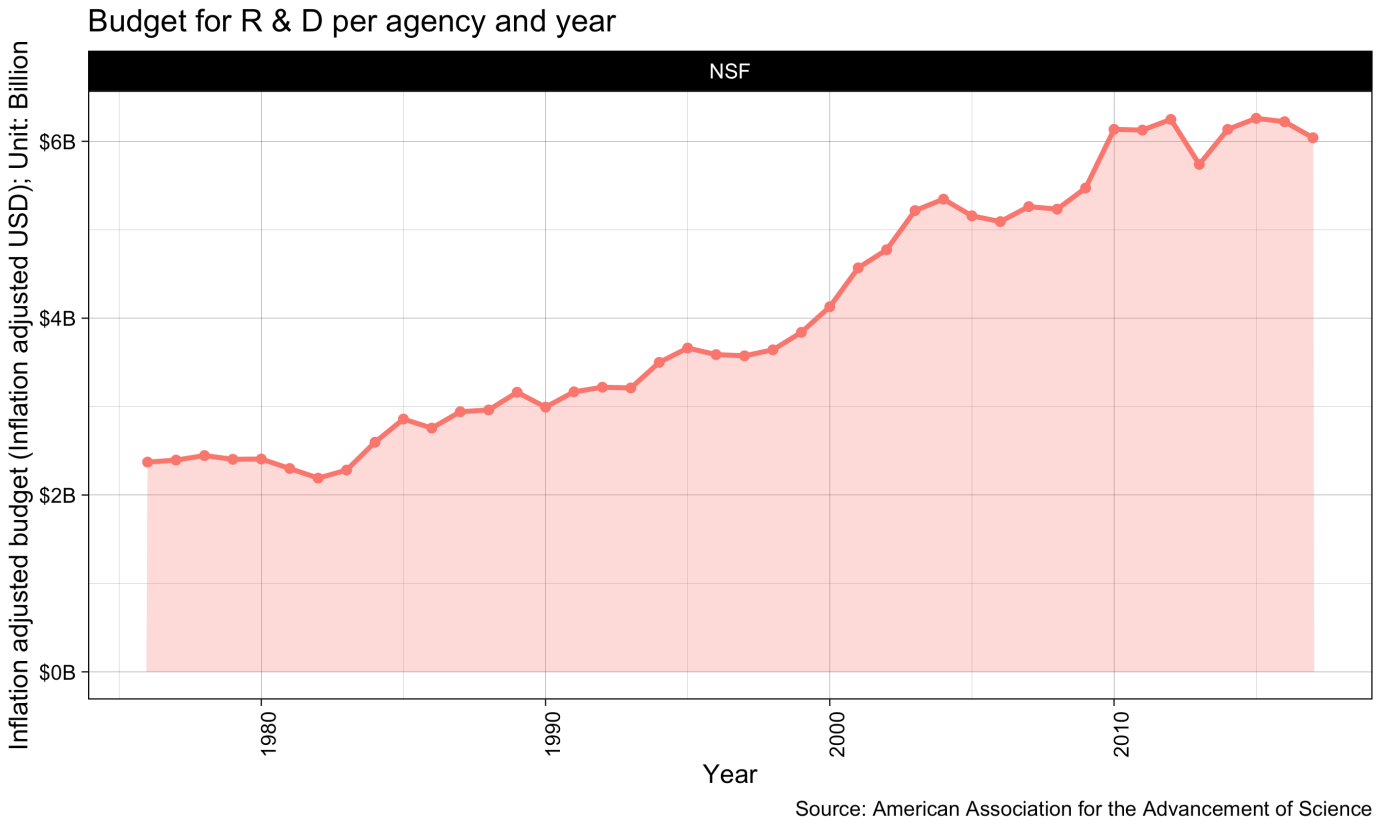

The Sharpe ratio assigns the same score to horribly different portfolios
Oct. 9, 2007–Mar. 20, 2013 daily closes
Nominal total return
◇ = Max. drawdown between two closes
portfolio value
Sharpe ratio
Martin ratio
Diverse portfolio .07 .12
S&P 500 .07 .07
Depression portfolio .07 .02
$160,000
$140,000
$120,000
$100,000
$80,000
$60,000
$40,000
$20,000
$112K
Diverse portfolio
–29%
S&P 500
–55% max. drawdown
Depression portfolio
–78%
2007
2008
2009
2010
2011
2012
2013



$$Sortino\ ratio = \frac{R_p - R_f}{\sigma_d}$$

- In situations like this, alternative risk-adjusted performance measures may be more appropriate.
- For example, the *Sortino ratio* is similar to the Sharpe ratio, but only considers downside volatility when calculating risk-adjusted performance.

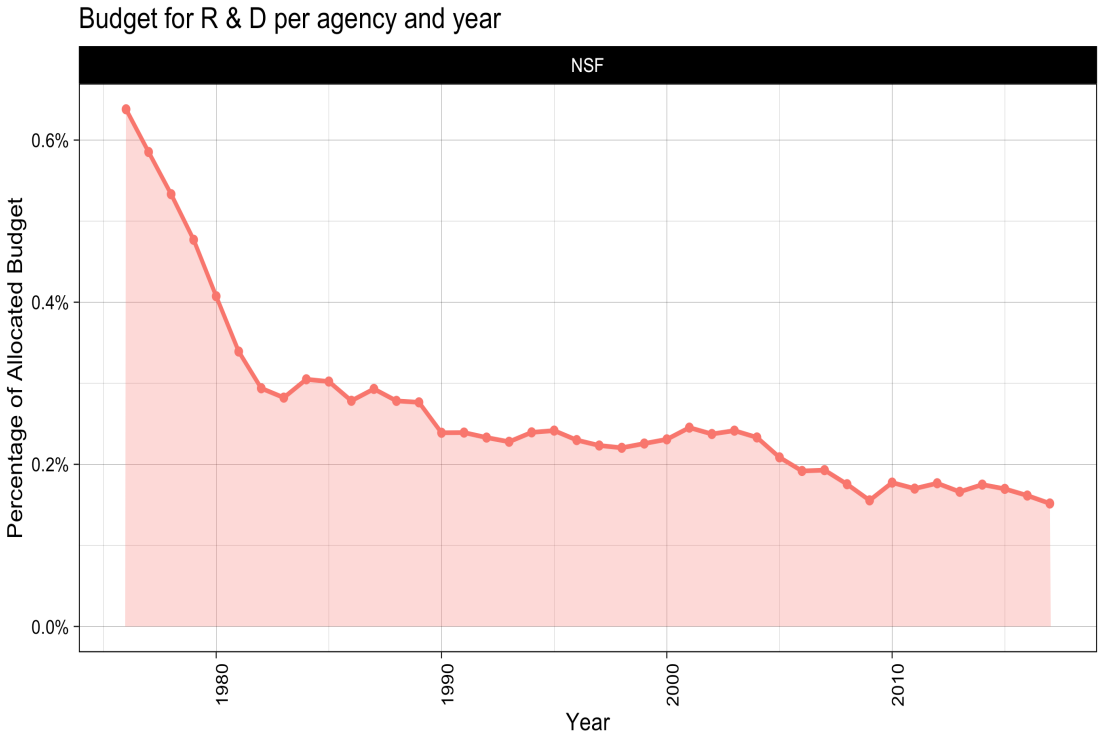
Standardization of the Scales



Standardization of the Scales

Standardization & Normalization

- To conduct a comparable analysis on time-series data, it is essential to take into account the total budget outlays when allocating the year-to-year R&D budget.

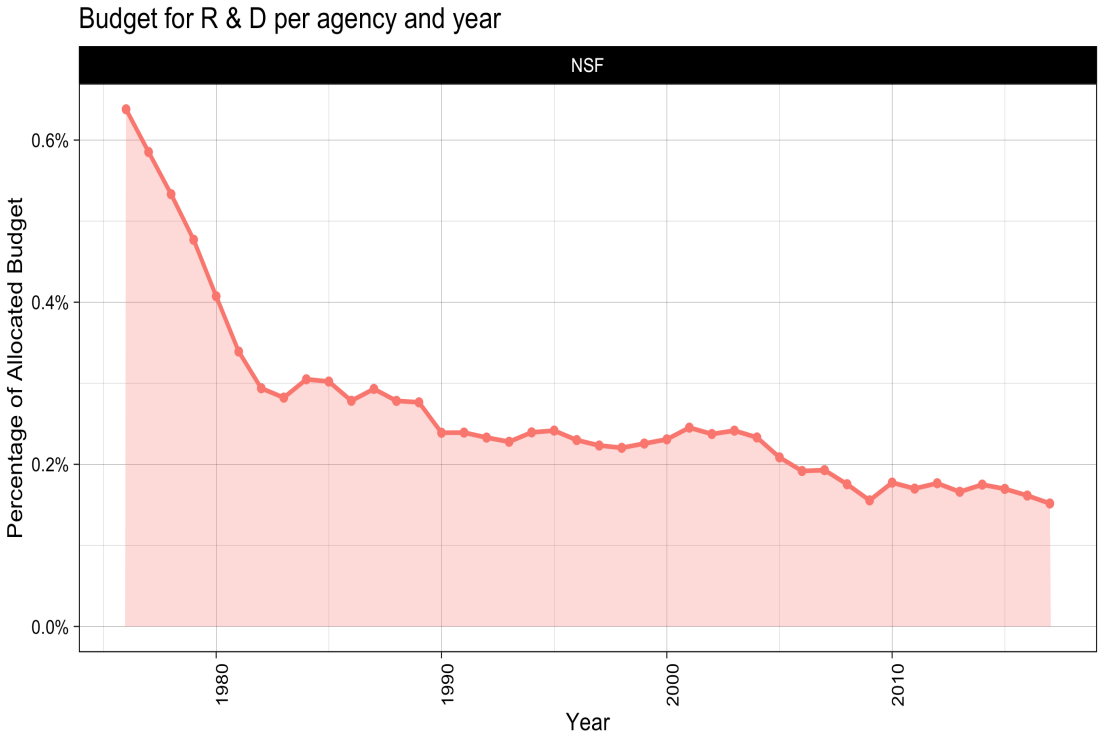


Source: American Association for the Advancement of Science

Standardization of the Scales

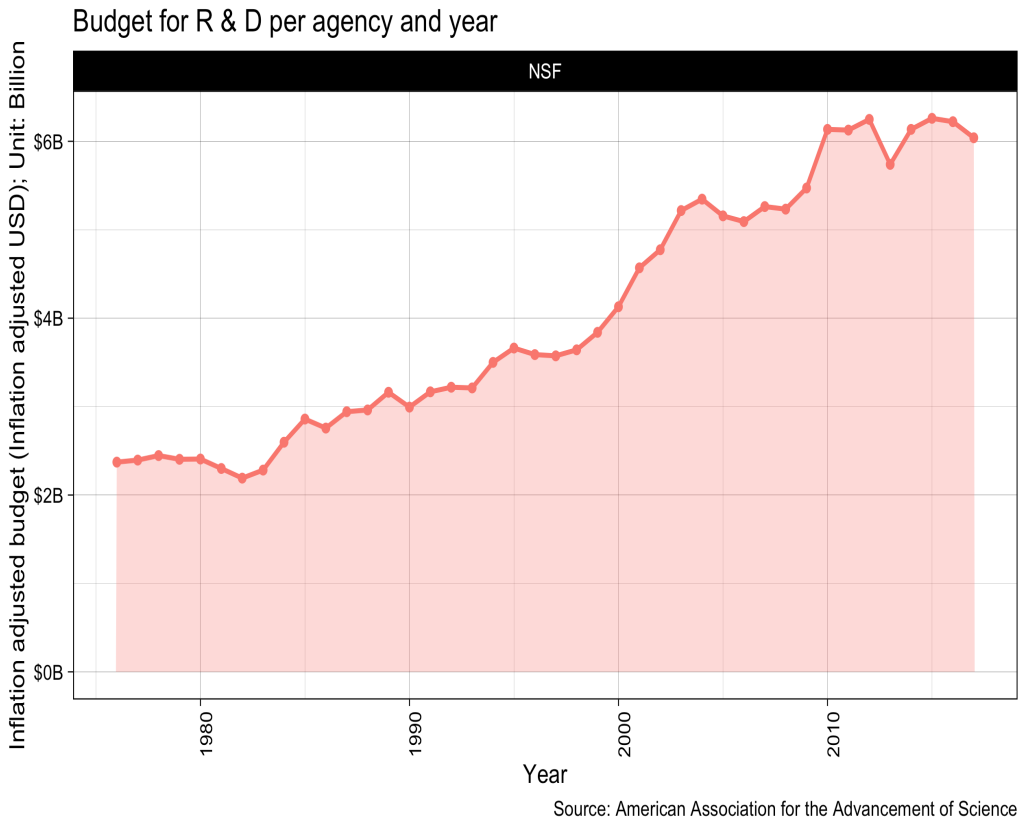
Standardization & Normalization

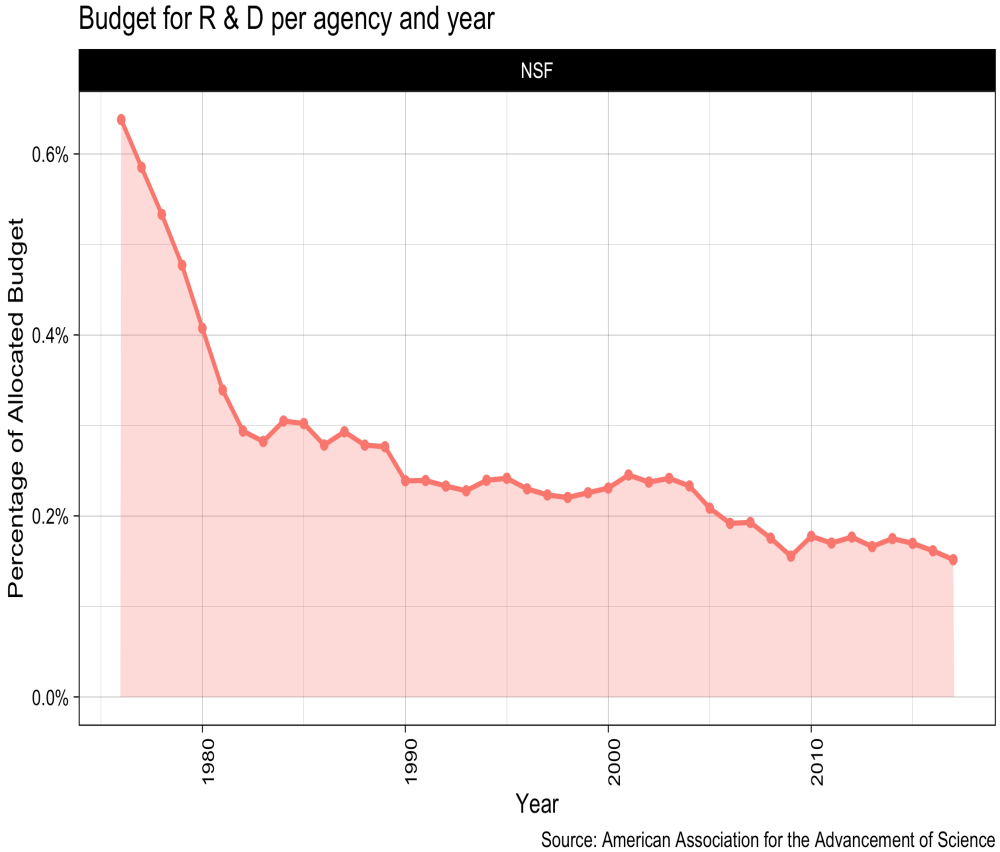
- Failure to do so is akin to not factoring in inflation rates, which makes it difficult to determine the actual value of your money year-to-year.
- In real-world data analysis, it is crucial to go beyond inflation and time value of money considerations and consider additional hidden factors for proper normalization and standardization of your target scale.



Source: American Association for the Advancement of Science

The Same (Data) But Different (Conclusion)





Even with the same factual information, misinformation can still arise.

Agenda (Workflow of Analytics)

Define

Collect

Import

Wrangle

Visualize

Model

Report

Deploy

Communicating Quantitative Information

Search

Visualizing Friendships

By Paul Butler on Tuesday, December 14, 2010 at 9:16am

Visualizing data is like photography. Instead of starting with a blank canvas, you manipulate the lens used to present the data from a certain angle.

When the data is the social graph of 500 million people, there are a lot of lenses through which you can view it. One that piqued my curiosity was the locality of friendship. I was interested in seeing how geography and political borders affected where people lived relative to their friends. I wanted a visualization that would show which cities had a lot of friendships between them.

I began by taking a sample of about ten million pairs of friends from Apache Hive, our data warehouse. I combined that data with each user's current city and summed the number of friends between each pair of cities. Then I merged the data with the longitude and latitude of each city.

At that point, I began exploring it in R, an open-source statistics environment. As a sanity check, I plotted points at some of the latitude and longitude coordinates. To my relief, what I saw was roughly an outline of the world. Next I erased the dots and plotted lines between the points. After a few minutes of rendering, a big white blob appeared in the center of the map. Some of the outer edges of the blob vaguely resembled the continents, but it was clear that I had too much data to get interesting results just by drawing lines. I thought that making the lines semi-transparent would do the trick, but I quickly realized that my graphing environment couldn't handle enough shades of color for it to work the way I wanted.

Instead I found a way to simulate the effect I wanted. I defined weights for each pair of cities as a function of the Euclidean distance between them and the number of friends between them. Then I plotted lines between the pairs by weight, so that pairs of cities with the most friendships between them were drawn on top of the others. I used a color ramp from black to blue to white, with each line's color depending on its weight. I also transformed some of the lines to wrap around the image, rather than spanning more than halfway around the world.

After a few minutes of rendering, the new plot appeared, and I was a bit taken aback by what I saw. The blob had turned into a surprisingly detailed map of the world. Not only were continents visible, certain international borders were apparent as well. What really struck me, though, was knowing that the lines didn't represent coasts or rivers or political borders, but real human relationships. Each line might represent a friendship made while travelling, a family member abroad, or an old college friend pulled away by the various forces of life.

Later I replaced the lines with great circle arcs, which are the shortest routes between two points on the Earth. Because the Earth is a sphere, these are often not straight lines on the projection.

When I shared the image with others within Facebook, it resonated with many people. It's not just a pretty picture, it's a reaffirmation of the impact we have in connecting people, even across oceans and borders.

Paul is an intern on Facebook's data infrastructure engineering team.

Visualizing Facebook Friends

Paul Butler – December 18, 2010

Earlier this week I published a data visualization on the Facebook Engineering blog which, to my surprise, has received a lot of media coverage

I've received a lot comments about the image, many asking for more details on how I created it. When I tell people I used R, the reaction I get is roughly what I would expect if I told them I made it with Microsoft Paint and a bottle of Jägermeister. Some people even questioned whether it was actually done in R. The truth is, aside from the addition of the logo and date text, the image was produced entirely with about 150 lines of R code with no external dependencies. In the process I learned a few things about creating nice-looking graphs in R.

Search

Visualizing Friendships

By Paul Butler on Tuesday, December 14, 2010 at 9:16am

Visualizing data is like photography. Instead of starting with a blank canvas, you manipulate the lens used to present the data from a certain angle.

When the data is the social graph of 500 million people, there are a lot of lenses through which you can view it. One that piqued my curiosity was the locality of friendship. I was interested in seeing how geography and political borders affected where people lived relative to their friends. I wanted a visualization that would show which cities had a lot of friendships between them.

I began by taking a sample of about ten million pairs of friends from Apache Hive, our data warehouse. I combined that data with each user's current city and summed the number of friends between each pair of cities. Then I merged the data with the longitude and latitude of each city.

At that point, I began exploring it in R, an open-source statistics environment. As a sanity check, I plotted points at some of the latitude and longitude coordinates. To my relief, what I saw was roughly an outline of the world. Next I erased the dots and plotted lines between the points. After a few minutes of rendering, a big white blob appeared in the center of the map. Some of the outer edges of the blob vaguely resembled the continents, but it was clear that I had too much data to get interesting results just by drawing lines. I thought that making the lines semi-transparent would do the trick, but I quickly realized that my graphing environment couldn't handle enough shades of color for it to work the way I wanted.

Instead I found a way to simulate the effect I wanted. I defined weights for each pair of cities as a function of the Euclidean distance between them and the number of friends between them. Then I plotted lines between the pairs by weight, so that pairs of cities with the most friendships between them were drawn on top of the others. I used a color ramp from black to blue to white, with each line's color depending on its weight. I also transformed some of the lines to wrap around the image, rather than spanning more than halfway around the world.

After a few minutes of rendering, the new plot appeared, and I was a bit taken aback by what I saw. The blob had turned into a surprisingly detailed map of the world. Not only were continents visible, certain international borders were apparent as well. What really struck me, though, was knowing that the lines didn't represent coasts or rivers or political borders, but real human relationships. Each line might represent a friendship made while travelling, a family member abroad, or an old college friend pulled away by the various forces of life.

Later I replaced the lines with great circle arcs, which are the shortest routes between two points on the Earth. Because the Earth is a sphere, these are often not straight lines on the projection.

When I shared the image with others within Facebook, it resonated with many people. It's not just a pretty picture, it's a reaffirmation of the impact we have in connecting people, even across oceans and borders.

Paul is an intern on Facebook's data infrastructure engineering team.

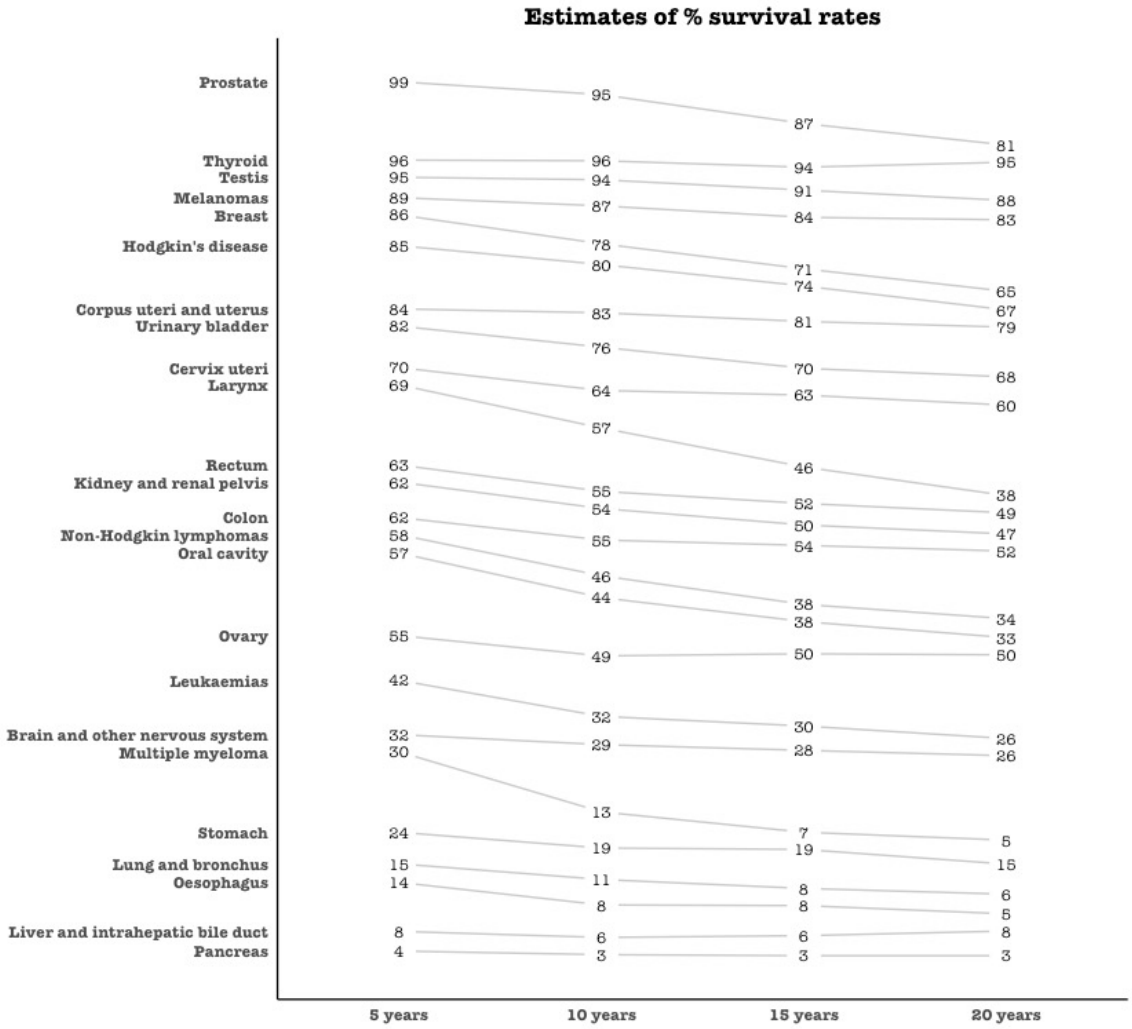
Visualizing Facebook Friends

Paul Butler – December 18, 2010

Earlier this week I published a data visualization on the Facebook Engineering blog which, to my surprise, has received a lot of media coverage

I've received a lot comments about the image, many asking for more details on how I created it. When I tell people I used R, the reaction I get is roughly what I would expect if I told them I made it with Microsoft Paint and a bottle of Jägermeister. Some people even questioned whether it was actually done in R. The truth is, aside from the addition of the logo and date text, the image was produced entirely with about 150 lines of R code with no external dependencies. In the process I learned a few things about creating nice-looking graphs in R.

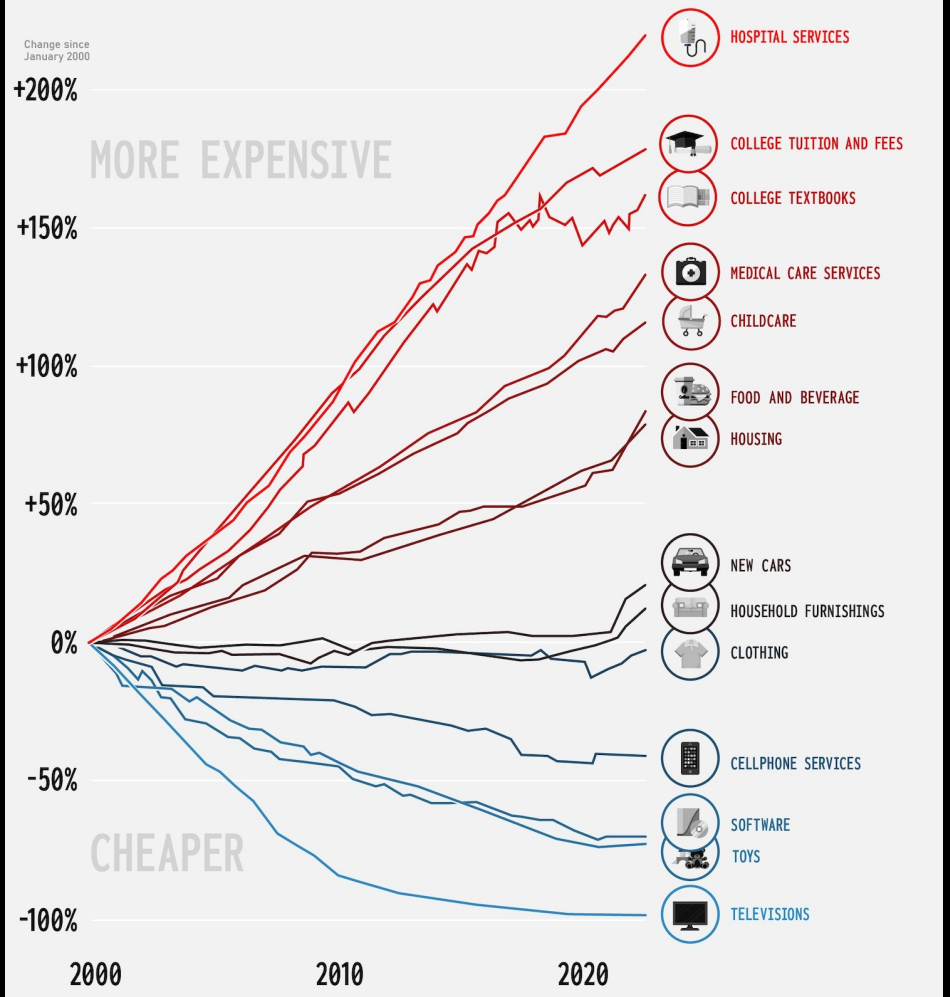
Below image was produced entirely about 50 lines of Python code.



Price Changes

OF CONSUMER GOODS AND SERVICES

Broadly speaking, price levels have increased by 74% since 2000. That said, the actual numbers vary wildly depending on the type of good or service. Many consumer goods like toys and TVs have gotten cheaper, while critical categories like healthcare and education have skyrocketed.



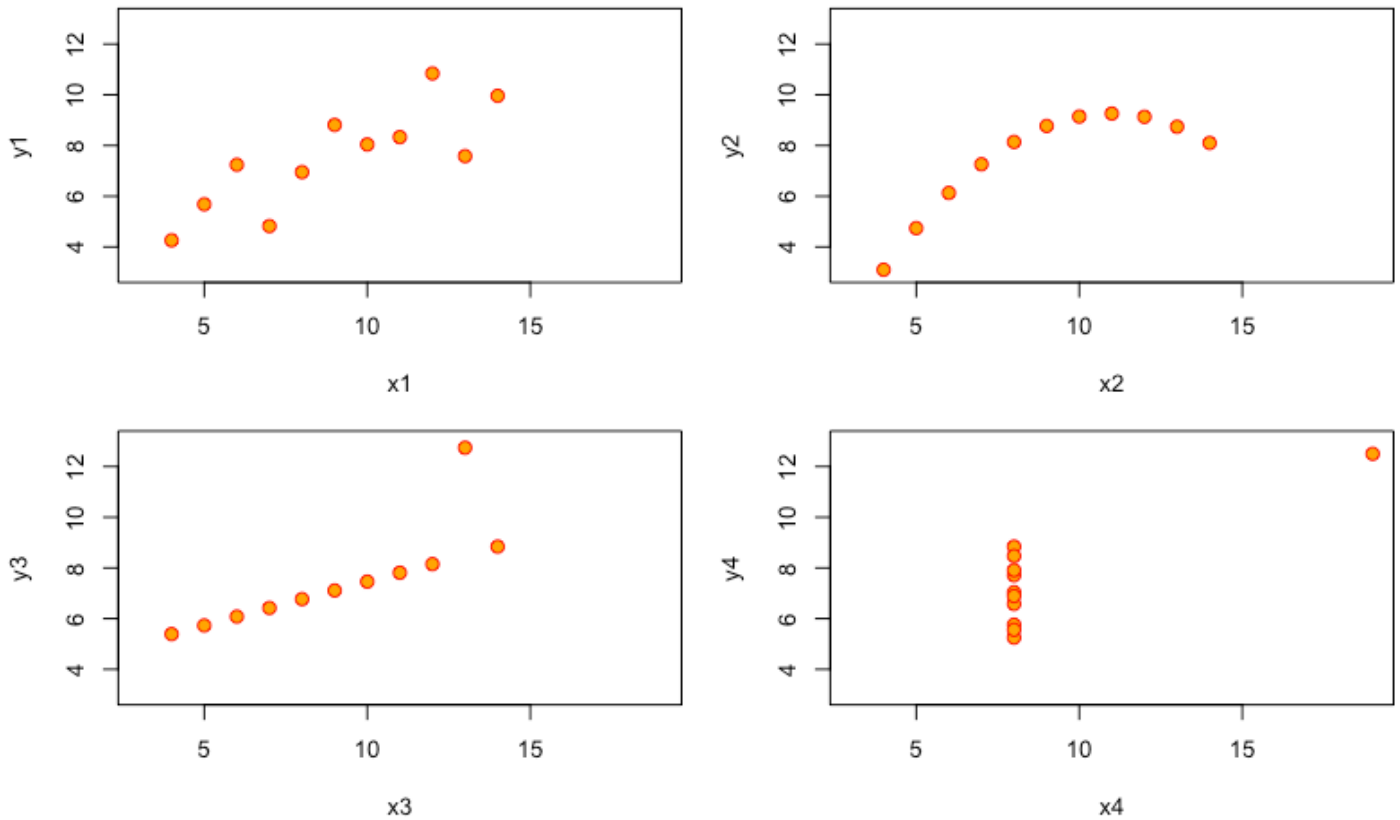
Source: Bureau of Labor Statistics Original design and concept by Mark J. Perry, Senior Fellow Emeritus, AEI

Turning Data Into Insights

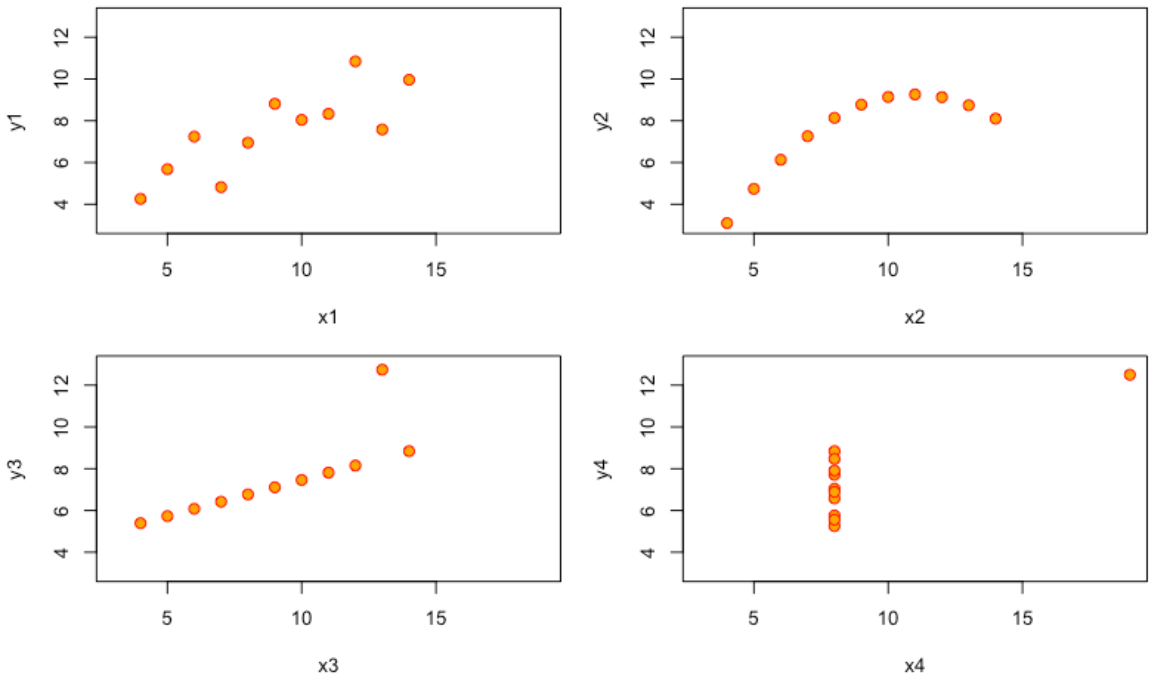
- These datasets were created and published by a statistician named Francis Anscombe back in the 1970s
- They've come to be known as Anscombe's quartet.

	x1	x2	x3	x4	y1	y2	y3	y4
1	10	10	10	8	8.04	9.14	7.46	6.58
2	8	8	8	8	6.95	8.14	6.77	5.76
3	13	13	13	8	7.58	8.74	12.74	7.71
4	9	9	9	8	8.81	8.77	7.11	8.84
5	11	11	11	8	8.33	9.26	7.81	8.47
6	14	14	14	8	9.96	8.10	8.84	7.04
7	6	6	6	8	7.24	6.13	6.08	5.25
8	4	4	4	19	4.26	3.10	5.39	12.50
9	12	12	12	8	10.84	9.13	8.15	5.56
10	7	7	7	8	4.82	7.26	6.42	7.91
11	5	5	5	8	5.68	4.74	5.73	6.89

Turning Data Into Insights



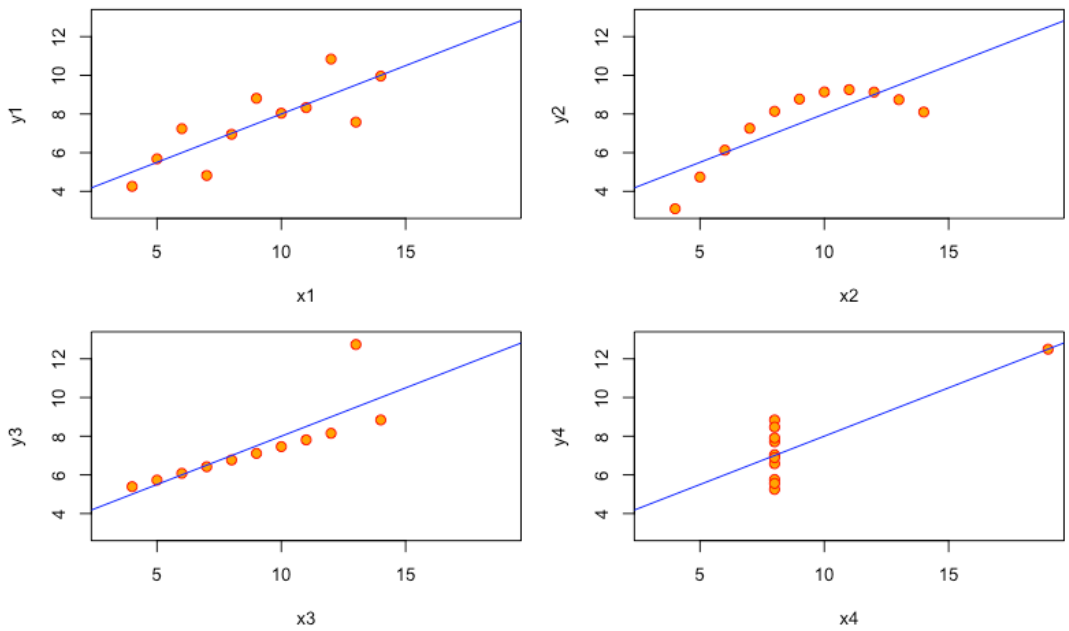
Turning Data Into Insights



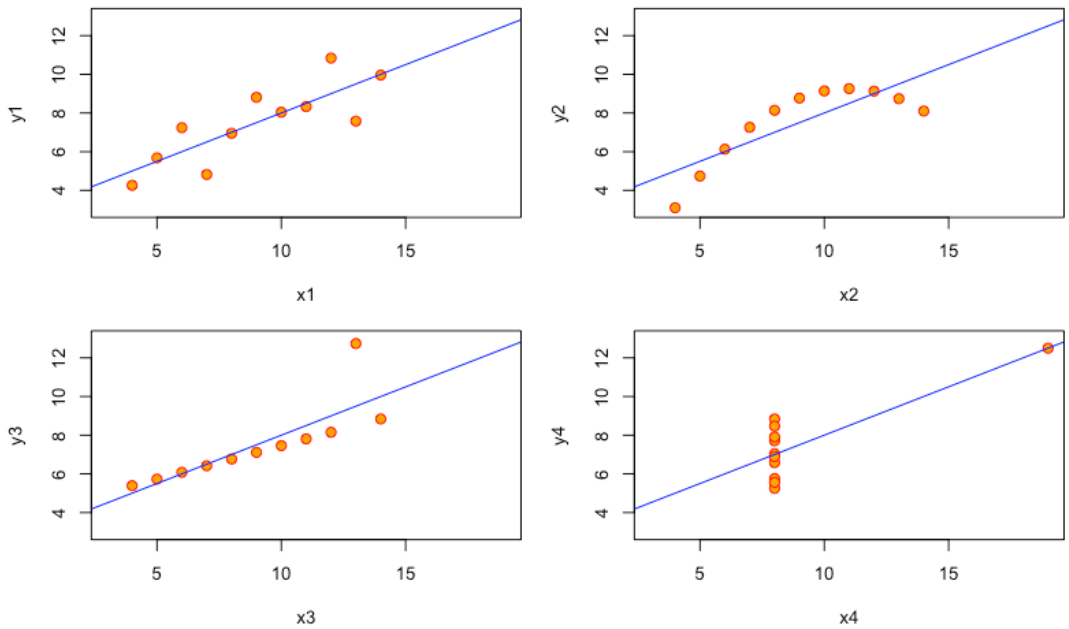
Turning Data Into Insights

- It might be a little bit more obvious if I fit in the least squares.

- The regression line in each case is the interesting thing about these four datasets as all four have the same summary statistics.

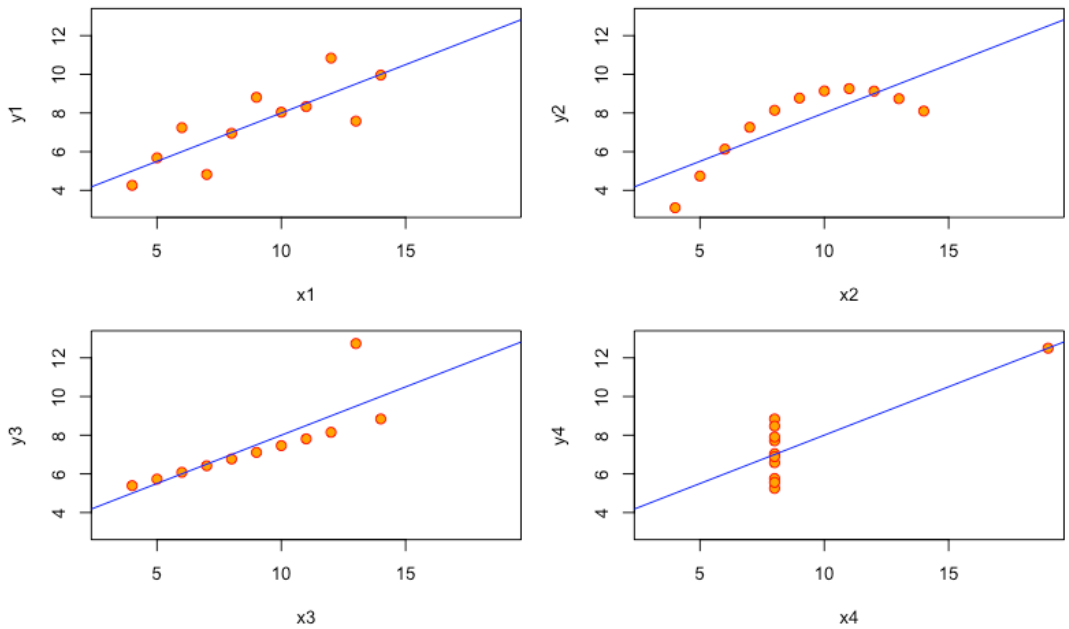


Turning Data Into Insights



- They have the same mean of X, the same variance of X, the same mean of Y, the same variance of Y, the same correlation, the same slope of the line, and the same intercept of the line, so the summary statistics for all four datasets are the same.

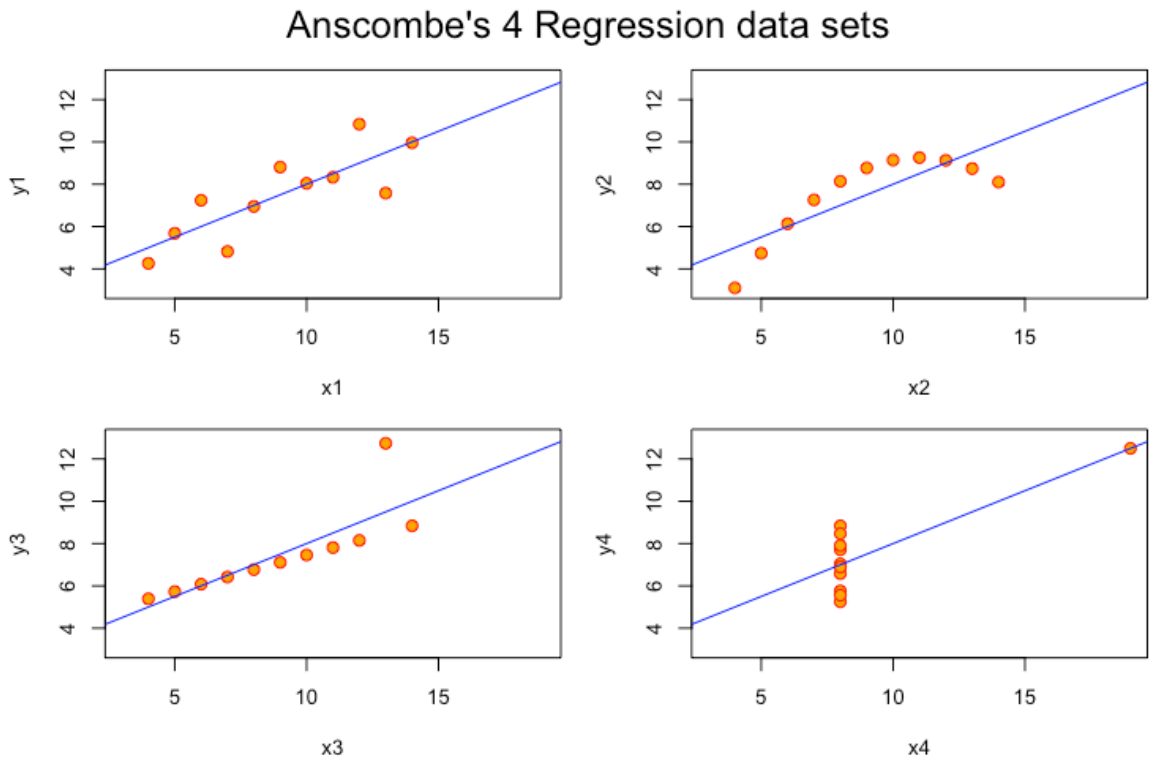
Turning Data Into Insights



- They have the same mean of X, the same variance of X, the same mean of Y, the same variance of Y, the same correlation, the same slope of the line, and the same intercept of the line, so the summary statistics for all four datasets are the same.
- But the overall picture is quite different.

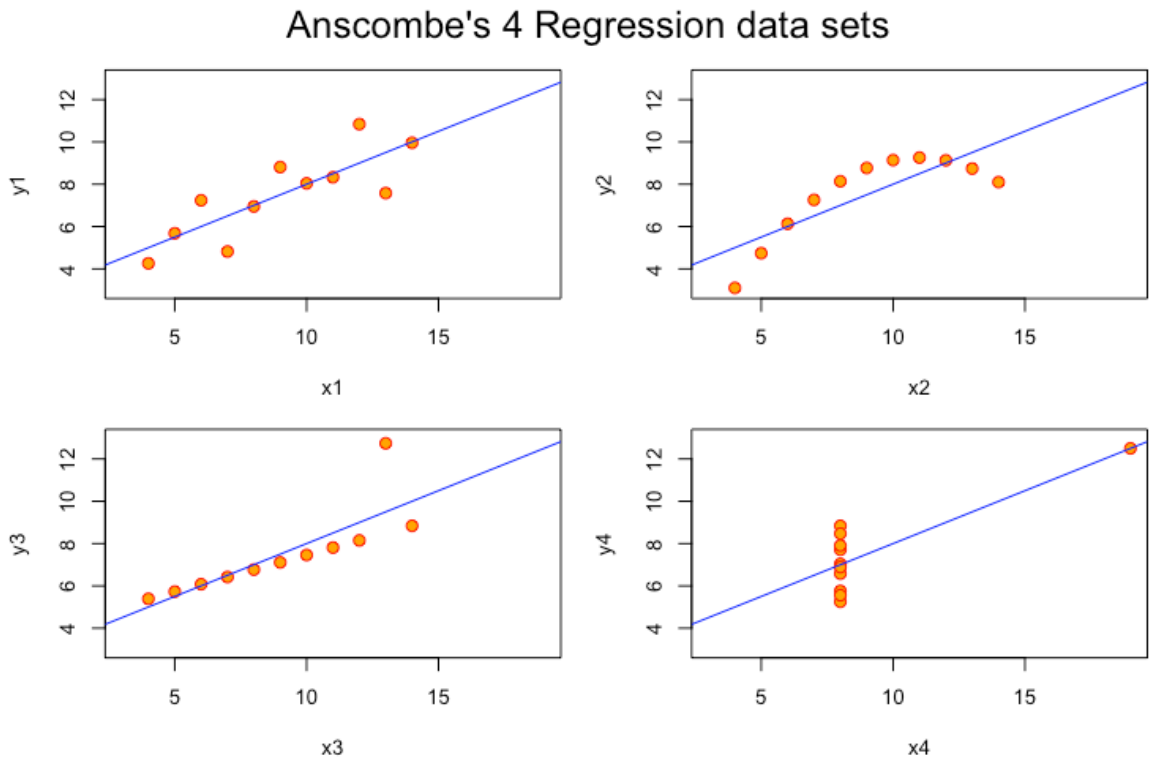
The Medium Could be The Message

- The lesson here is simply that summary statistics don't tell the whole story.



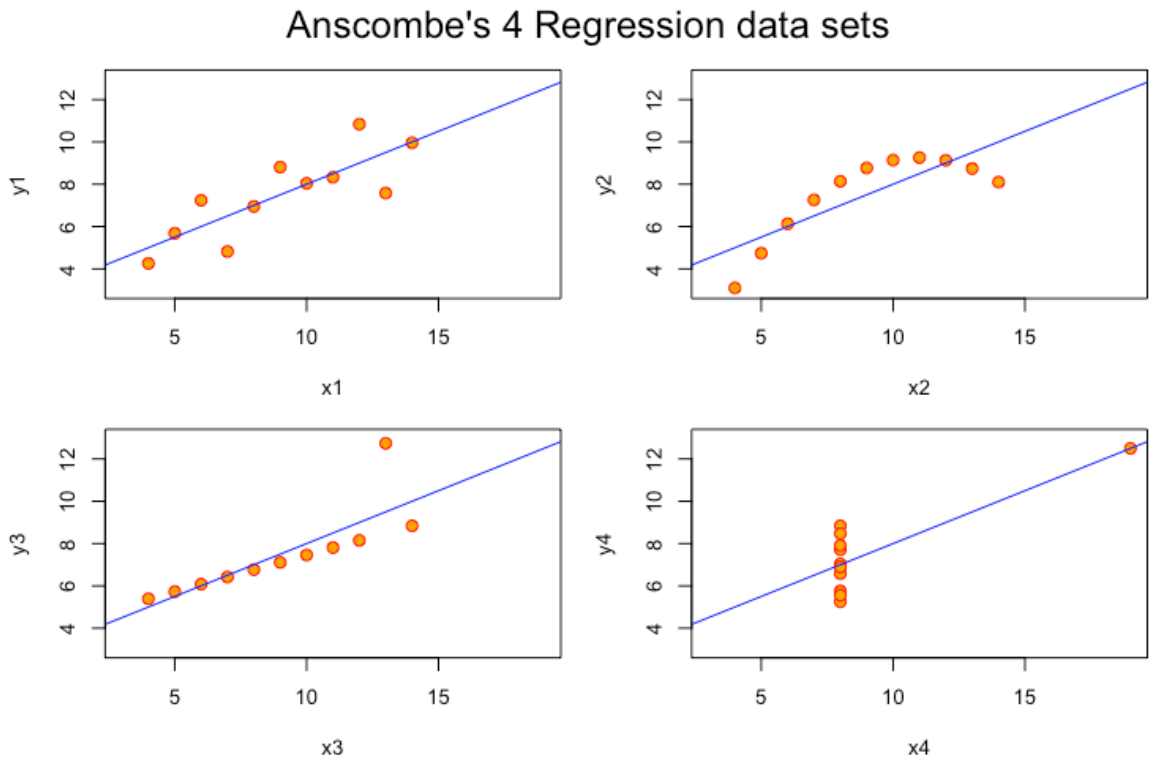
The Medium Could be The Message

- The lesson here is simply that summary statistics don't tell the whole story.
- Imagine if we were to look only at the output of these four datasets for a simple linear regression.

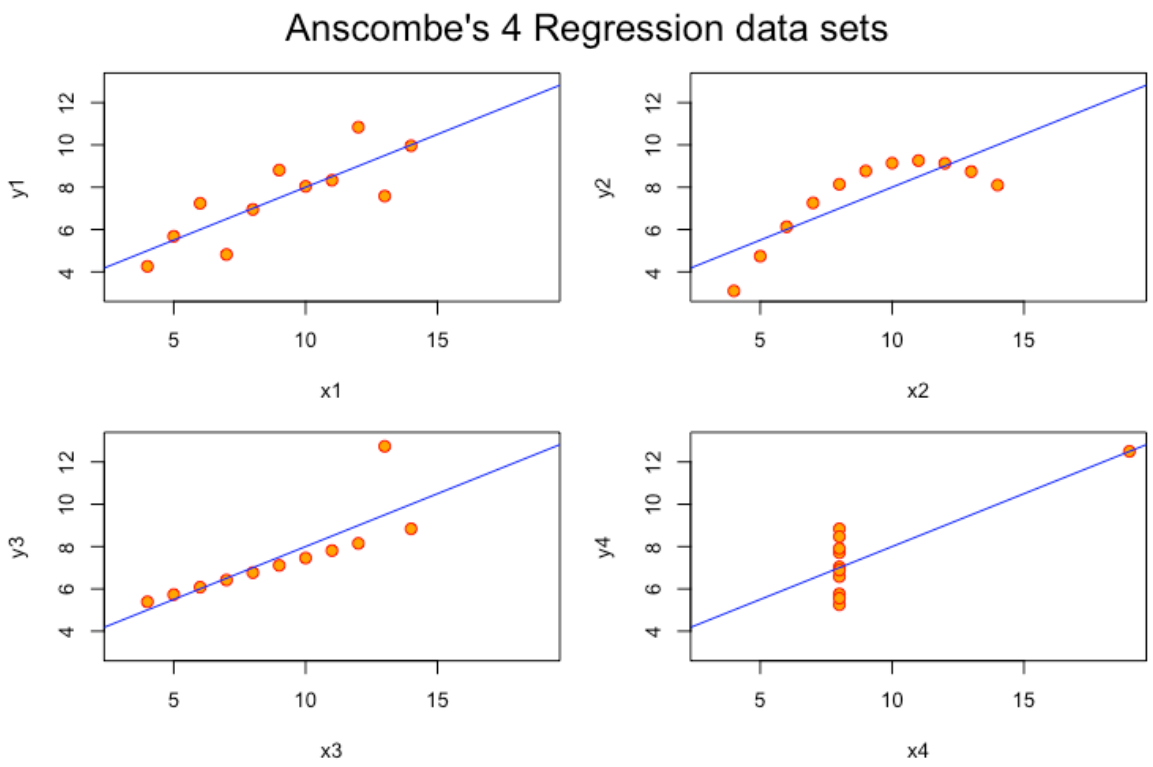


The Medium Could be The Message

- The lesson here is simply that summary statistics don't tell the whole story.
- Imagine if we were to look only at the output of these four datasets for a simple linear regression.
- Everything would look exactly the same, but the reality of the situation is that things are very different.

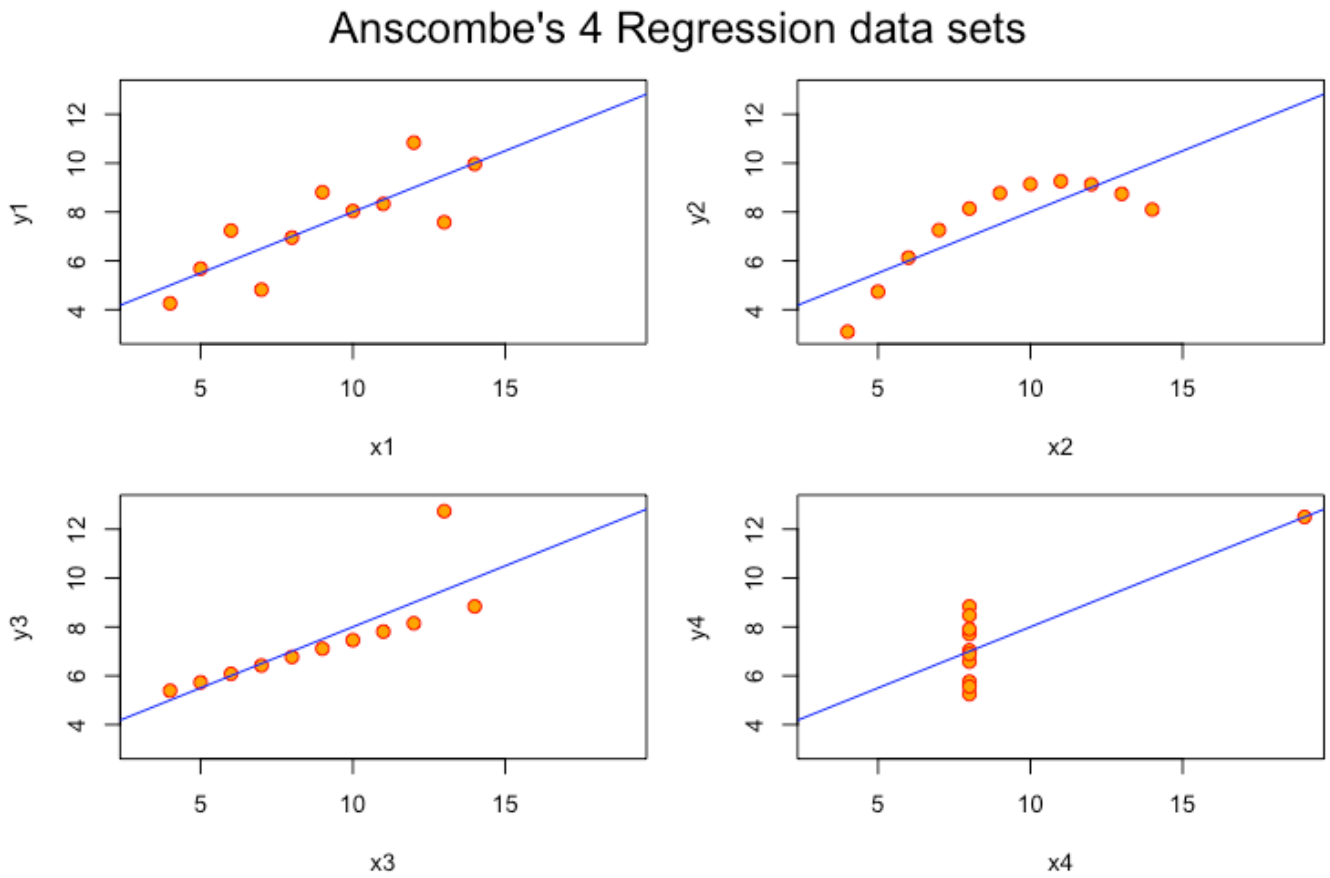


The Medium Could be The Message



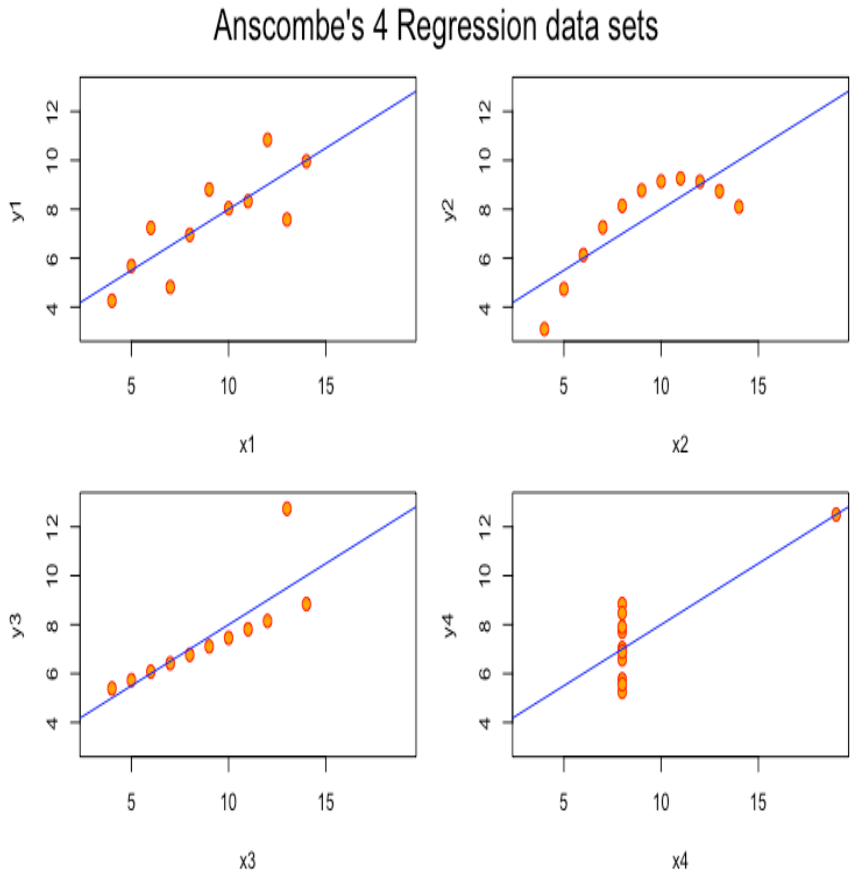
- Visualizing your data will give you a deeper understanding about quantitative information that you might think you know but you were not aware of: namely, data visualization helps transform data into insights.

Turning Data Into Insights



Turning Data Into Insights

- In this section of the lesson, I have prepared an additional bonus lecture to enhance your understanding of why data visualization is a crucial component of decision-making using data.
- You will find this bonus lecture along with my "Life After the Webinar" note. Enjoy ☺



Agenda (Workflow of Data-Driven Decision Making)

Define

Collect

Import

Wrangle

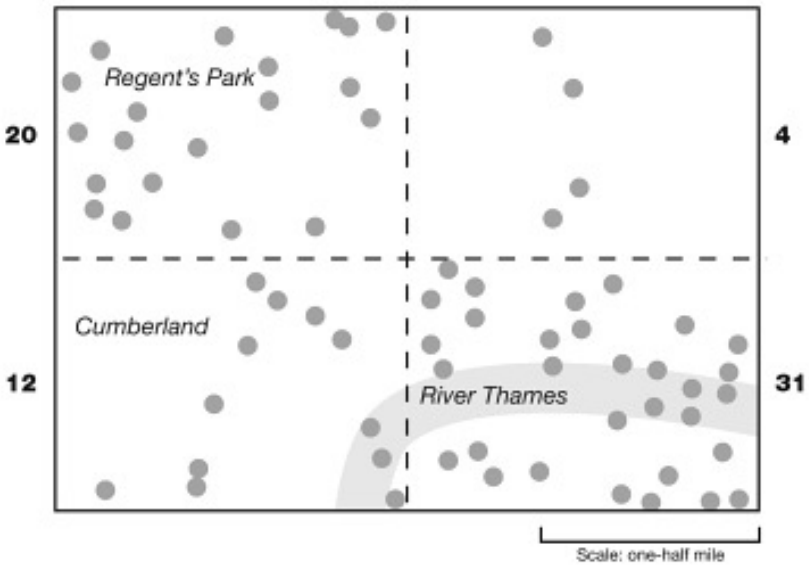
Visualize

Model

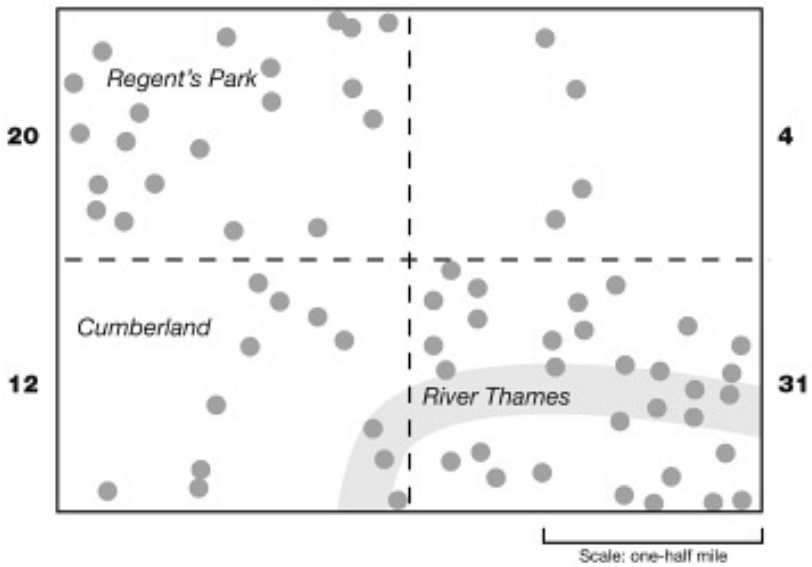
Report

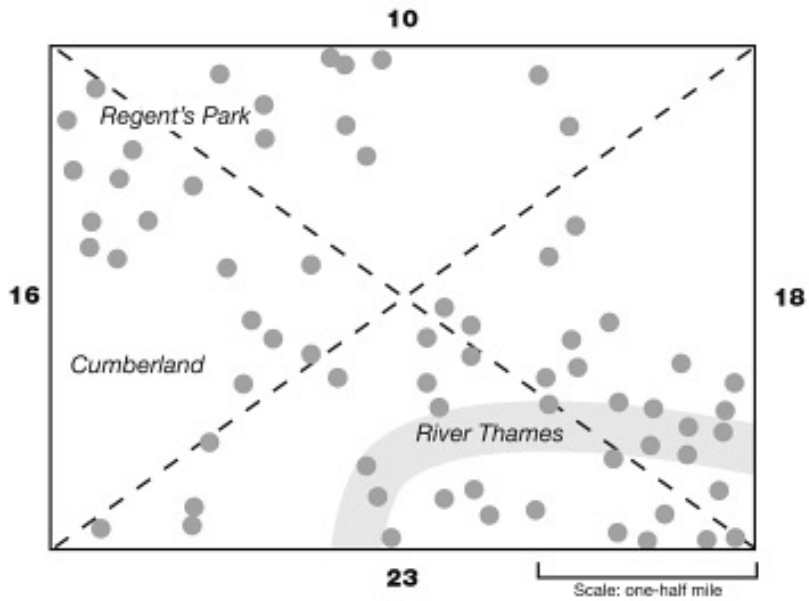
Deploy

Data Visualization & Clustering Bias



Data Visualization & Clustering Bias

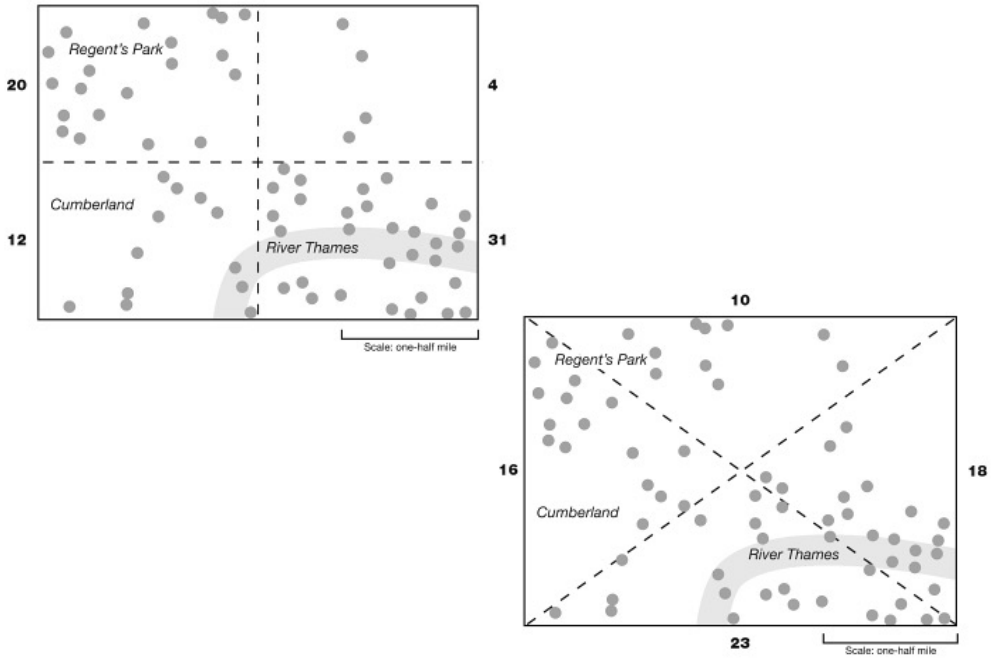




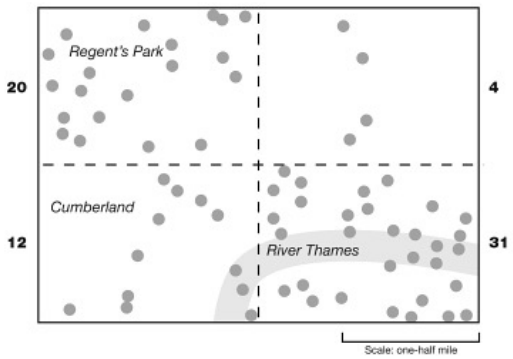
Data Visualization & Clustering Bias

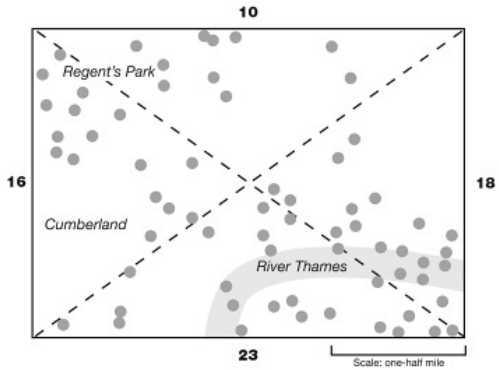
Clustering Bias?

- Human intelligence has a natural inclination to create patterns out of random outcomes.
- In the finance industry, clustering bias can lead to incorrect assumptions, misinterpretation of data, and flawed decision-making in areas such as risk assessment, pricing, and underwriting.



Data Visualization & Clustering Bias





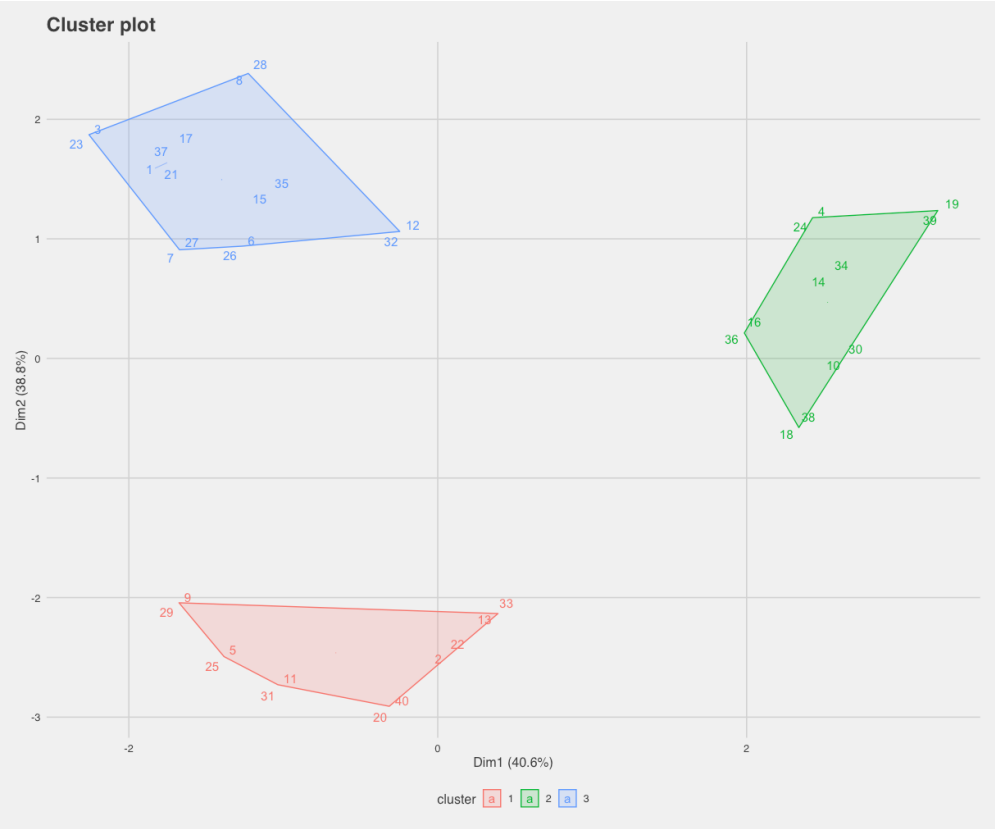
Demographic Clustering Bias

- For example, an insurer finds that young drivers have higher accident rates than older drivers. While age may be a contributing factor, the insurer may overlook other factors that could also influence accident rates, such as driving experience, vehicle type, or driving habits.
- By focusing solely on age, the insurer may charge unfairly high premiums for young drivers, failing to account for the full range of risk factors.

Solving Clustering Bias with Modeling (Unsupervised Machine Learning)

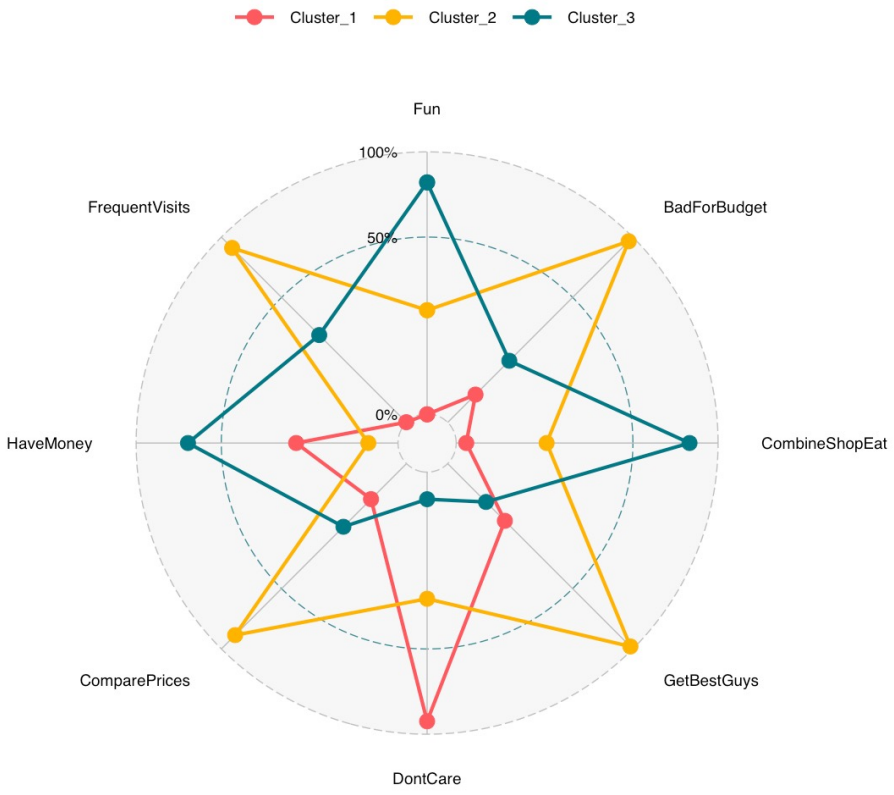
"Unsupervised" Learning

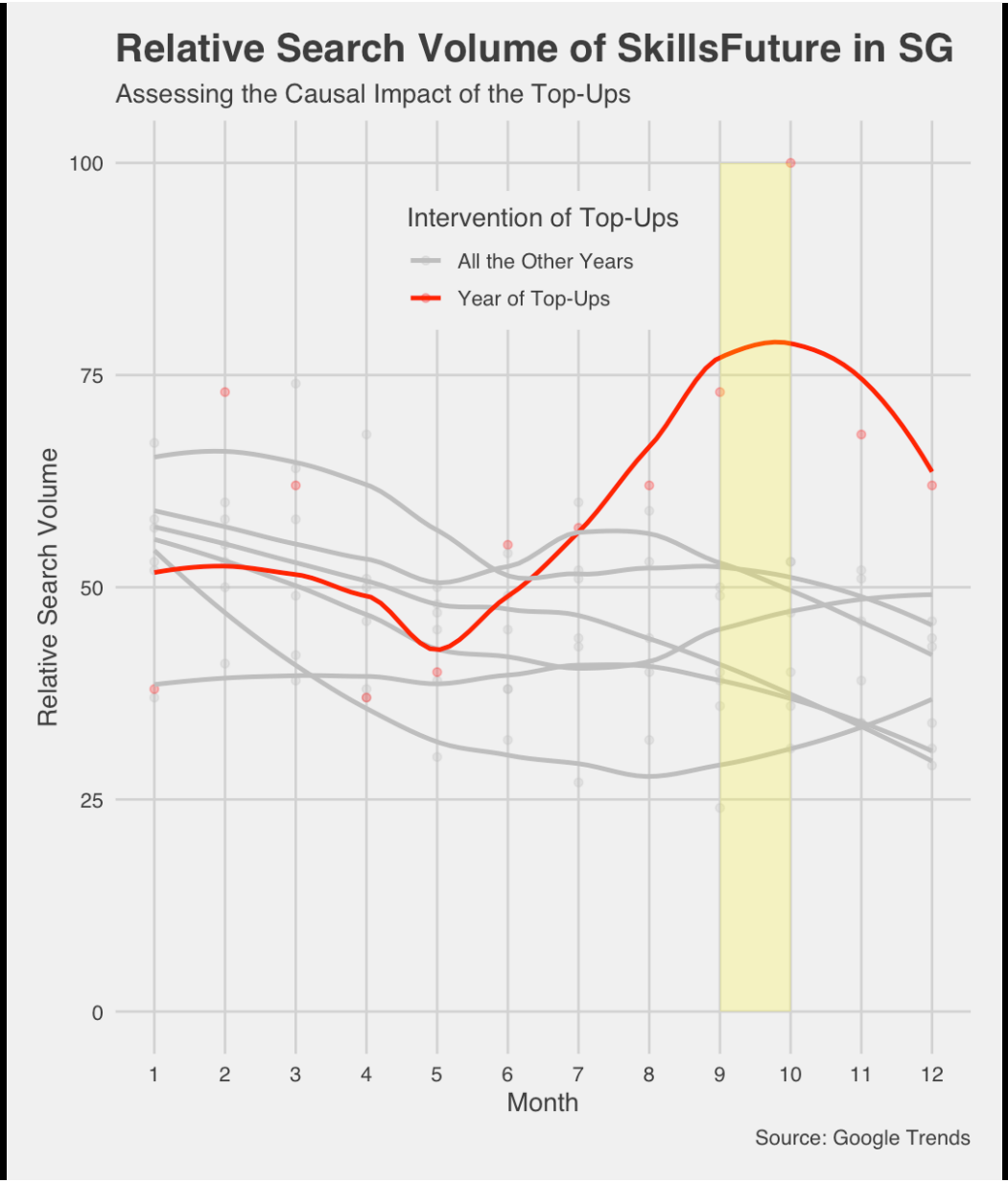
- The cluster analysis sample on the right suggests utilizing unsupervised machine learning algorithms to detect hidden patterns of shoppers in a Singapore mall.
- Using eight attributes, a total of 40 customers were clustered into three different groups.
- On the following page, I will demonstrate how you can use data visualization to inform your strategic decisions by understanding the profiles of these three customer clusters.



Aha!

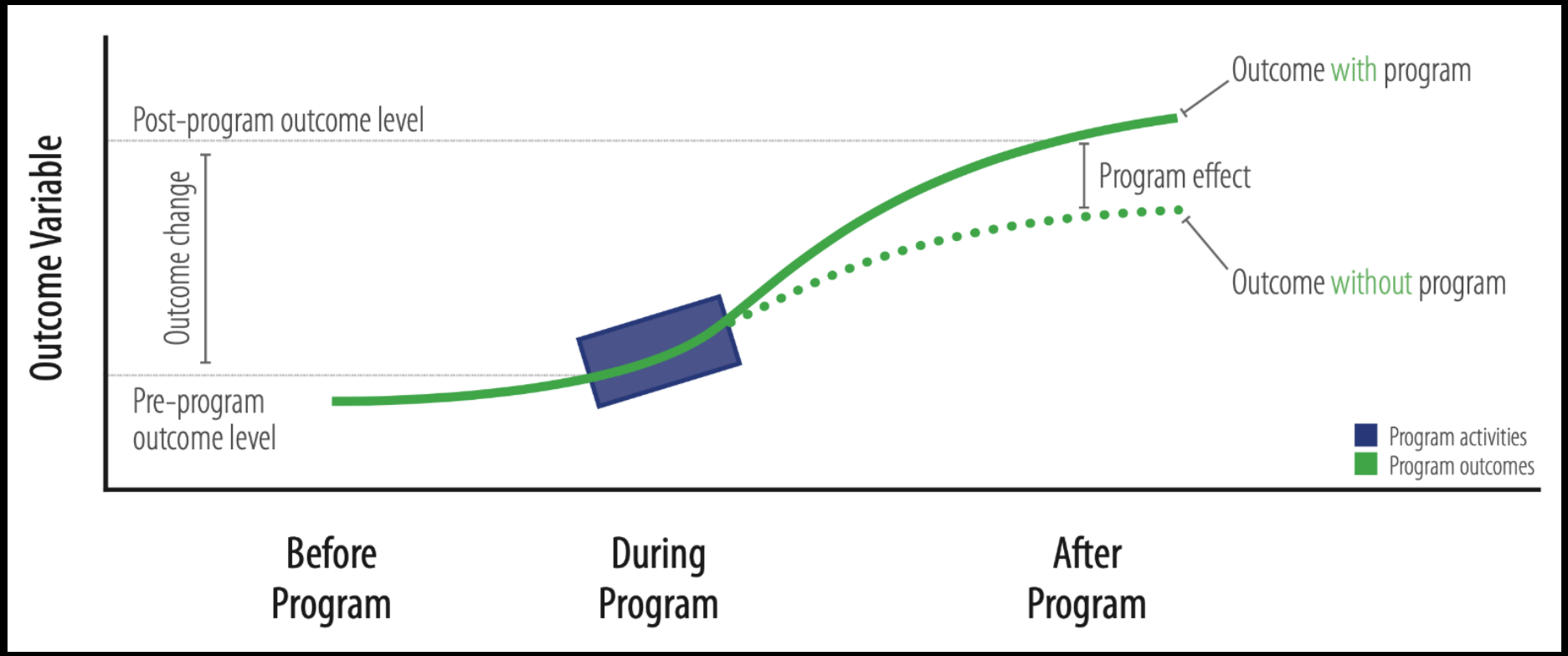
Clustering analysis that reveals the profiles of customer segments can drive data-driven strategies




Relative Search Volume of SkillsFuture in SG
Assessing the Causal Impact of the Top-Ups
Intervention of Top-Ups
All the Other Years
Year of Top-Ups
Relative Search Volume
100
75
50
25
0
1
2
3
4
5
6
7
8
9
10
11
12
Month
Source: Google Trends

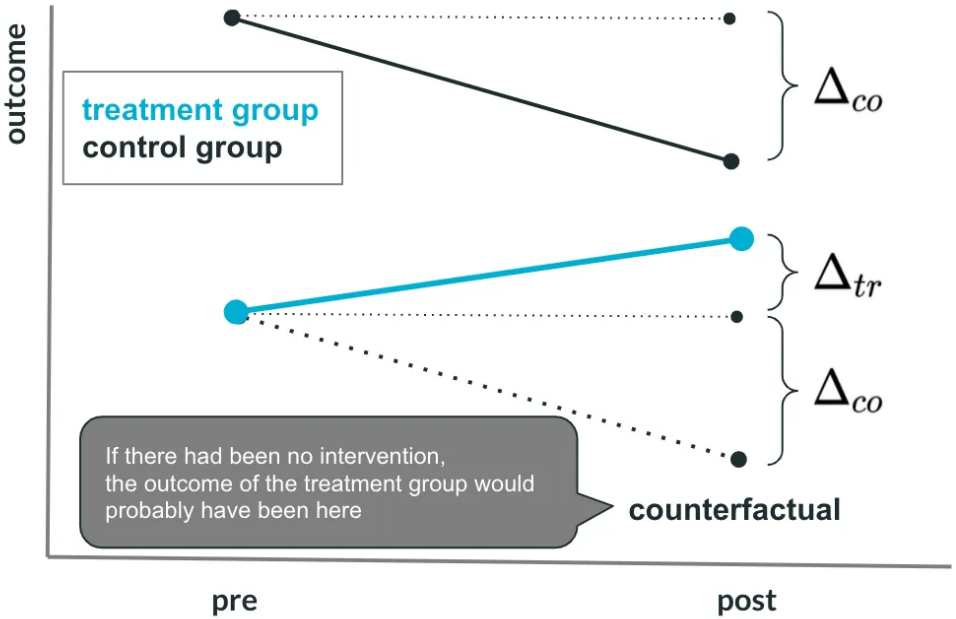
Causal Inference with Econometric Design

Natural (Quasi) Experiment

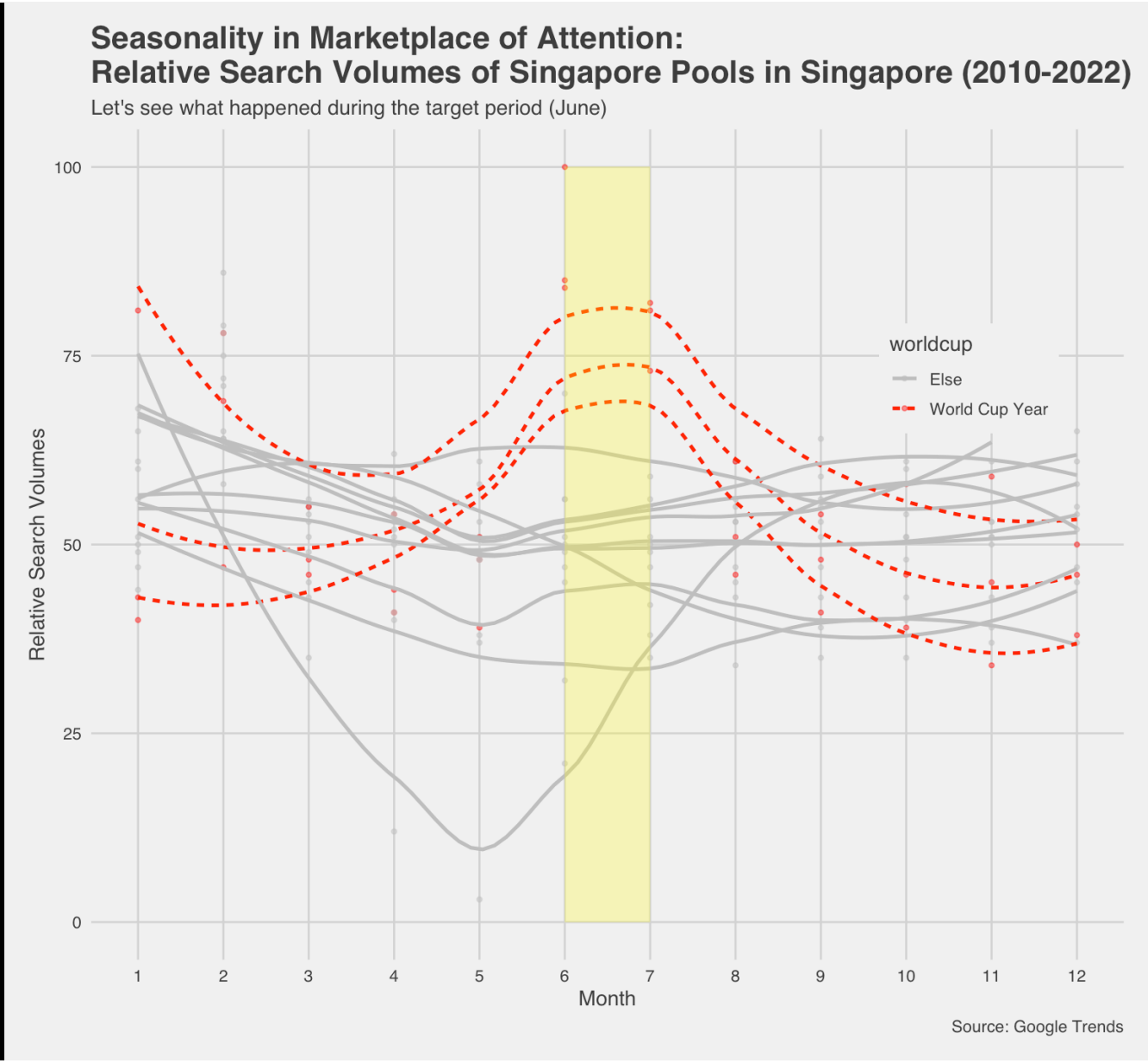


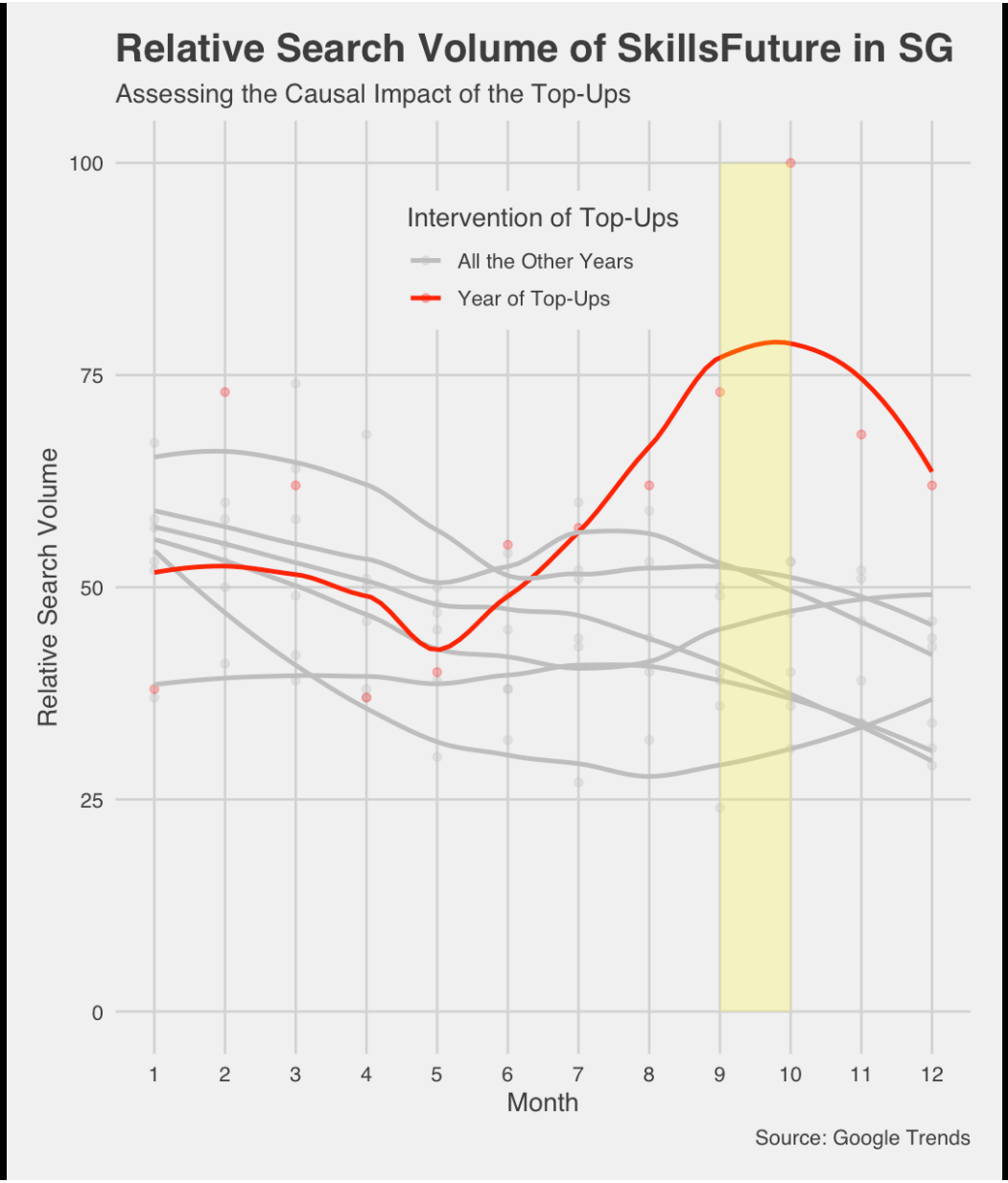
Difference in Differences (DID)

$$\mathrm{ATT}^{\mathrm{did}} = \Delta_{\mathrm{tr}} - \Delta_{\mathrm{co}}$$



pros	Excellent interpretability
cons	Parallel trend assumption • DID requires the selection of a control group that has parallel trends in the treatment group and the pre-intervention period. • This selection may be arbitrary in some cases. e.t.c.


Seasonality in Marketplace of Attention:
Relative Search Volumes of Singapore Pools in Singapore (2010-2022)
Let's see what happened during the target period (June)
Relative Search Volumes
100
75
50
25
0
1
2
3
4
5
6
7
8
9
10
11
12
Month
worldcup
Else
World Cup Year
Source: Google Trends


Relative Search Volume of SkillsFuture in SG
Assessing the Causal Impact of the Top-Ups
Intervention of Top-Ups
All the Other Years
Year of Top-Ups
Relative Search Volume
100
75
50
25
0
1
2
3
4
5
6
7
8
9
10
11
12
Month
Source: Google Trends

Agenda (Workflow of Analytics)

Define

Collect

Import

Wrangle

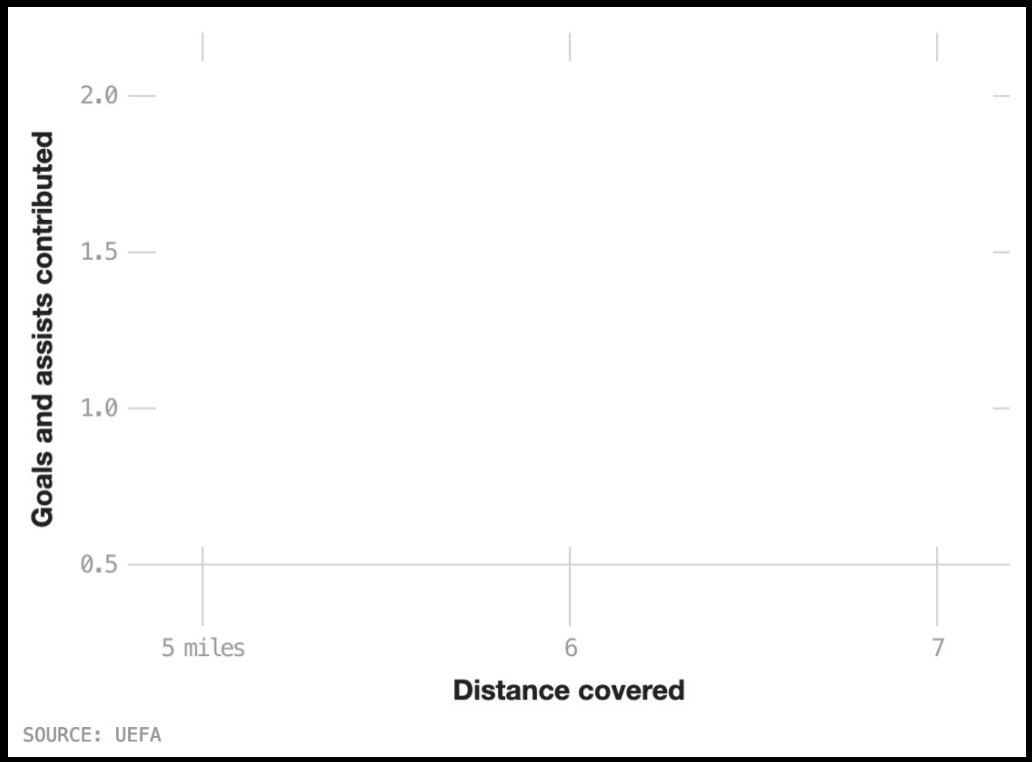
Visualize

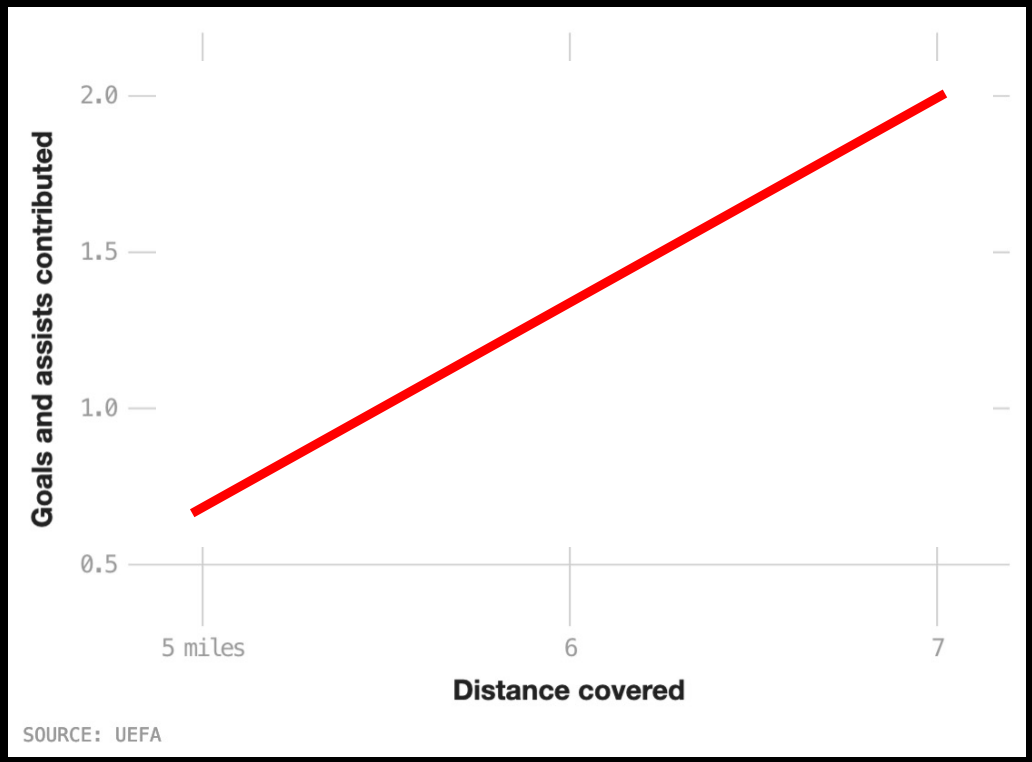
Model

Report

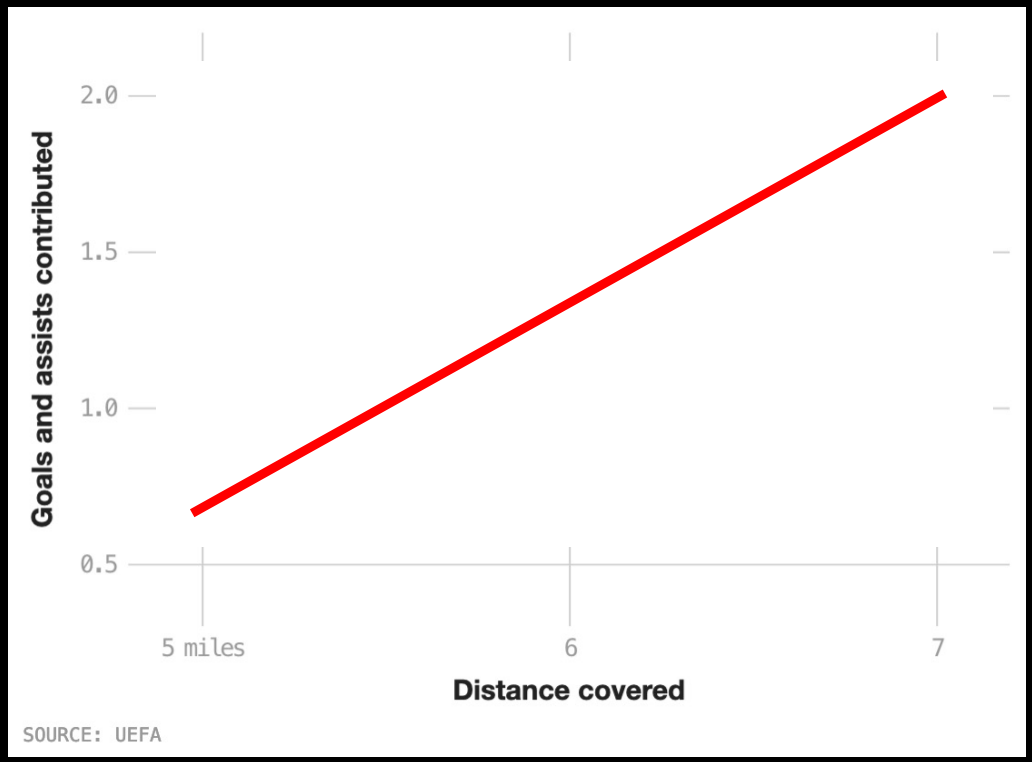
Deploy

What do you think, about the relationship between the two variables?

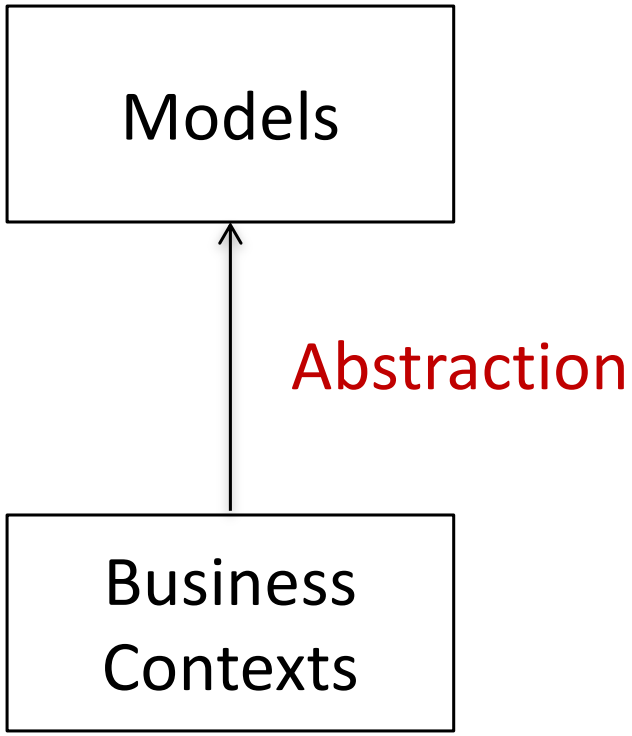



2.0
1.5
1.0
0.5
Goals and assists contributed
5 miles
6
7
Distance covered
SOURCE: UEFA

“Working *harder*, contributing *more*” Hypothesis



Analytics for Decision Making

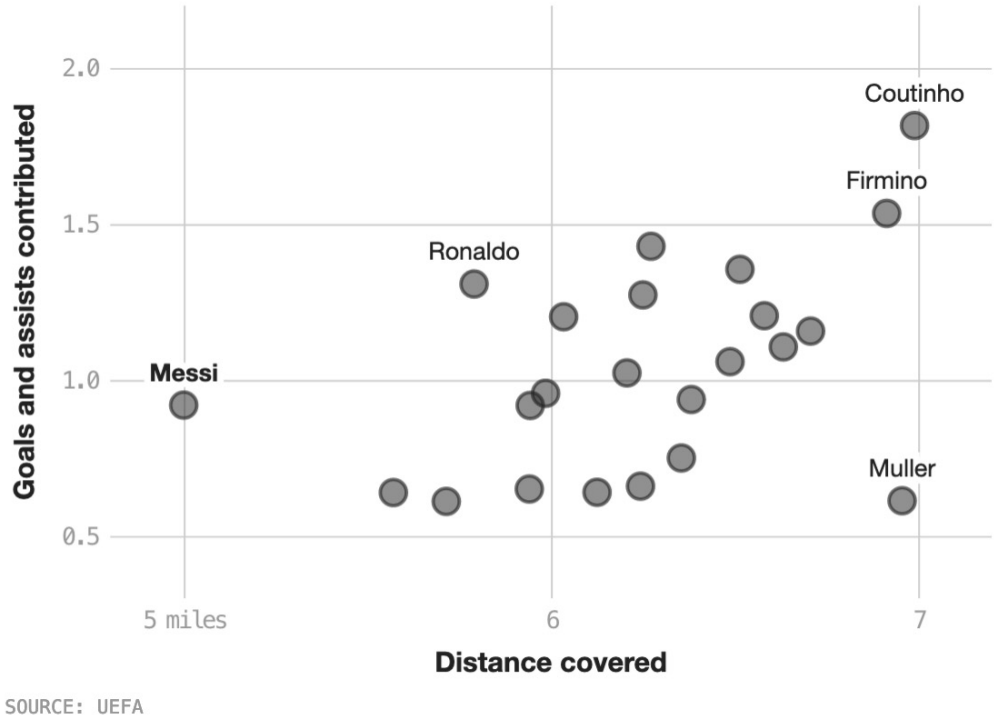


Model and Business Contexts

Refining Our Understanding of "Models"

- Data scientists should provide decision-makers with the information they need to make informed decisions quickly and efficiently.

- By presenting data in a clear and concise way, decision-makers can quickly identify trends, patterns, and outliers, and make decisions based on the insights gained from the data.

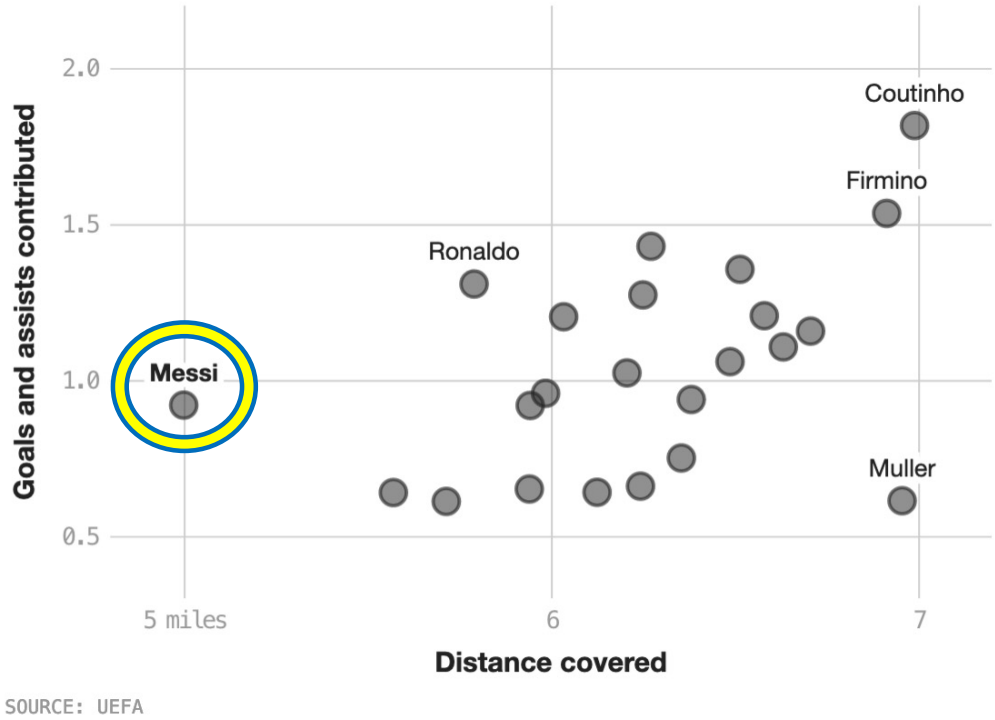


Model and Business Contexts

Refining Our Understanding of "Models"

- Data scientists should provide decision-makers with the information they need to make informed decisions quickly and efficiently.

- By presenting data in a clear and concise way, decision-makers can quickly identify trends, patterns, and outliers, and make decisions based on the insights gained from the data.

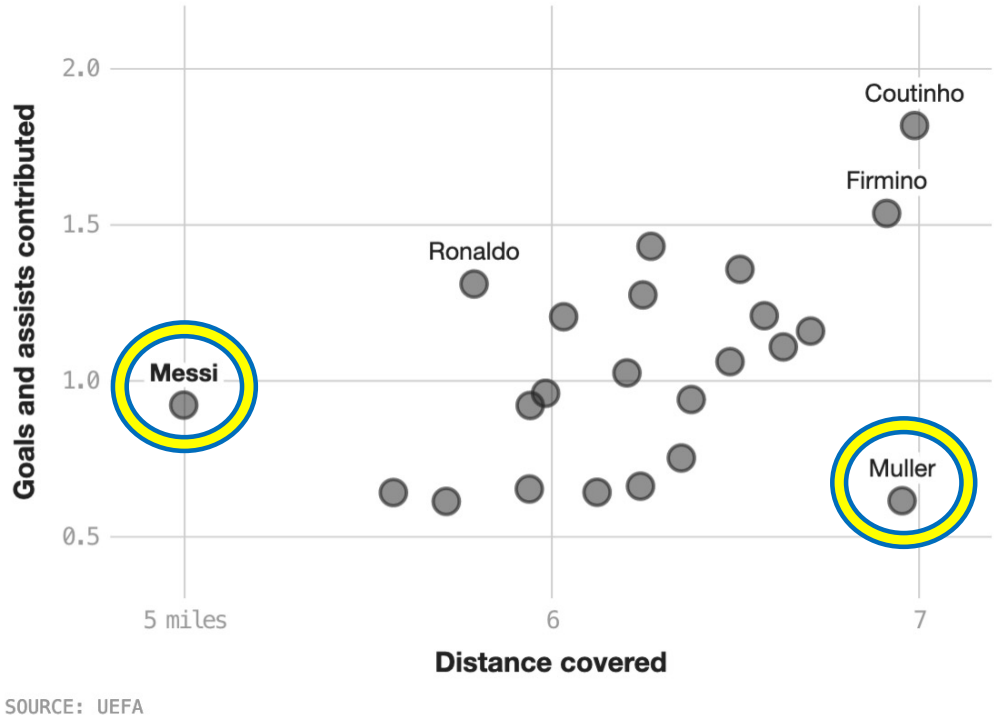


Model and Business Contexts

Refining Our Understanding of "Models"

- Data scientists should provide decision-makers with the information they need to make informed decisions quickly and efficiently.

- By presenting data in a clear and concise way, decision-makers can quickly identify trends, patterns, and outliers, and make decisions based on the insights gained from the data.

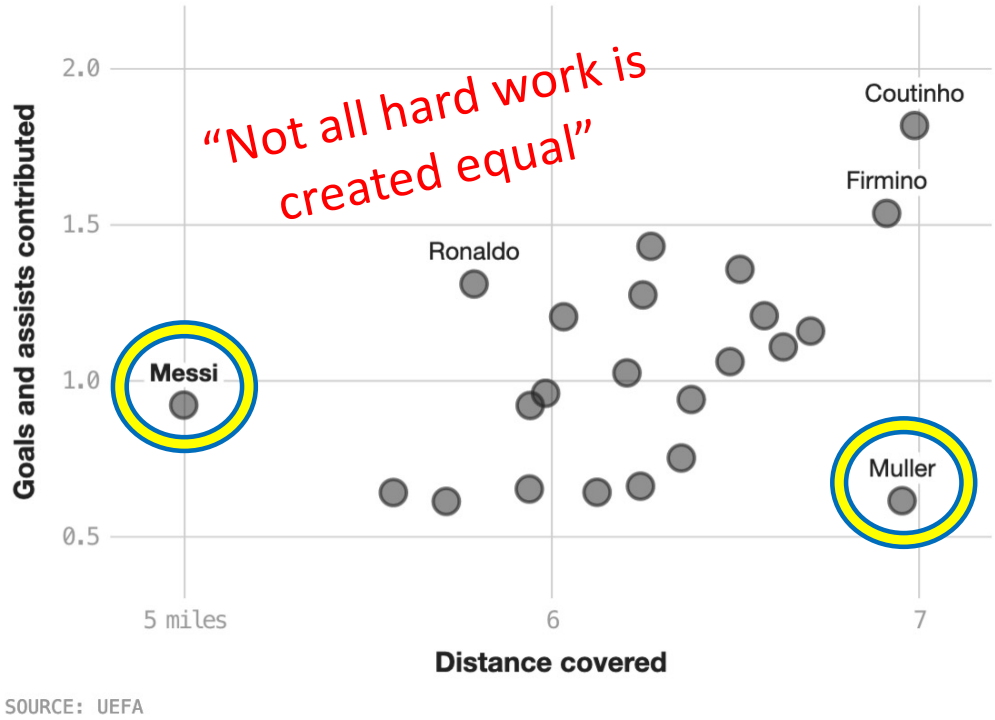


Model and Business Contexts

Refining Our Understanding of "Models"

- ☐ Data scientists should provide decision-makers with the information they need to make informed decisions quickly and efficiently.

- ☐ By presenting data in a clear and concise way, decision-makers can quickly identify trends, patterns, and outliers, and make decisions based on the insights gained from the data.

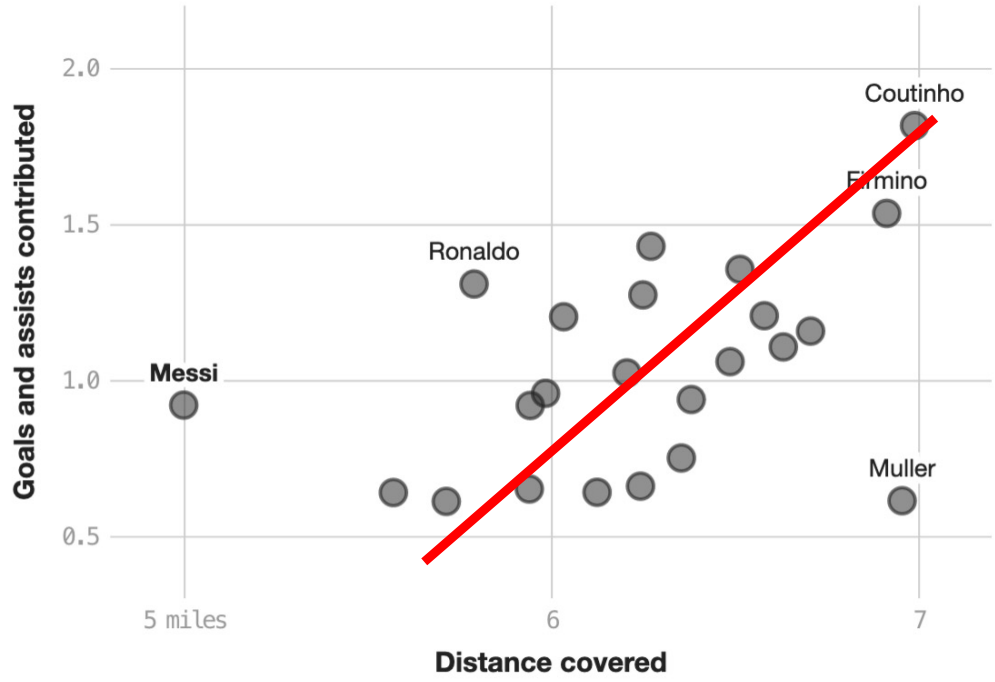


Model and Business Contexts

Refining Our Understanding of "Models"

- ☐ Data scientists should provide decision-makers with the information they need to make informed decisions quickly and efficiently.

- ☐ By presenting data in a clear and concise way, decision-makers can quickly identify trends, patterns, and outliers, and make decisions based on the insights gained from the data.



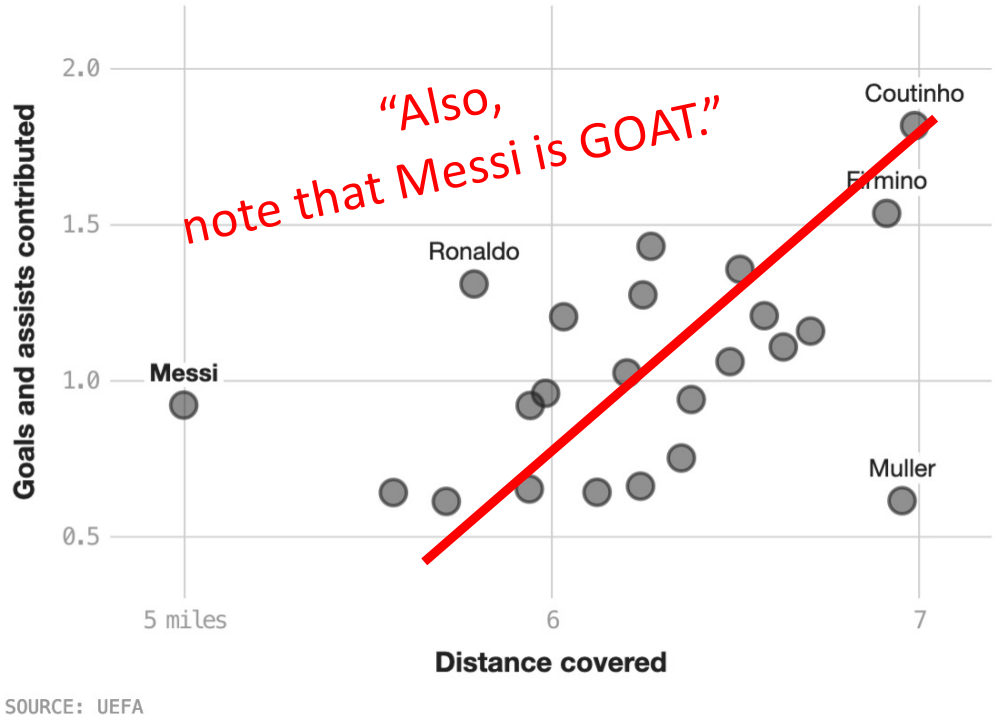
SOURCE: UEFA

Model and Business Contexts

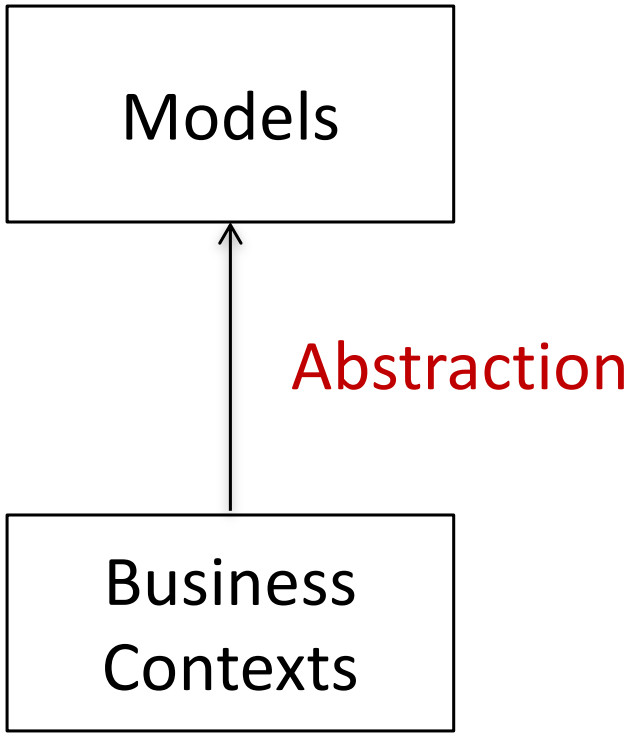
Refining Our Understanding of "Models"

- ☐ Data scientists should provide decision-makers with the information they need to make informed decisions quickly and efficiently.

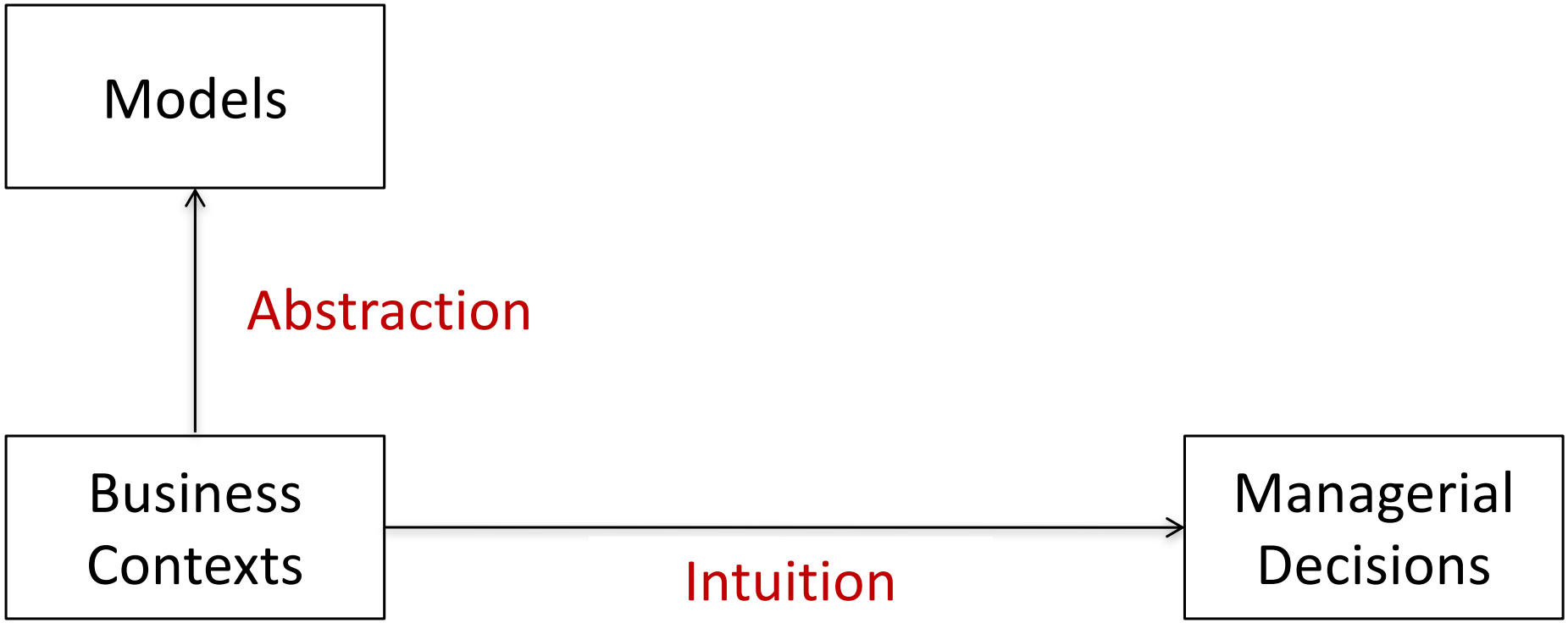
- ☐ By presenting data in a clear and concise way, decision-makers can quickly identify trends, patterns, and outliers, and make decisions based on the insights gained from the data.



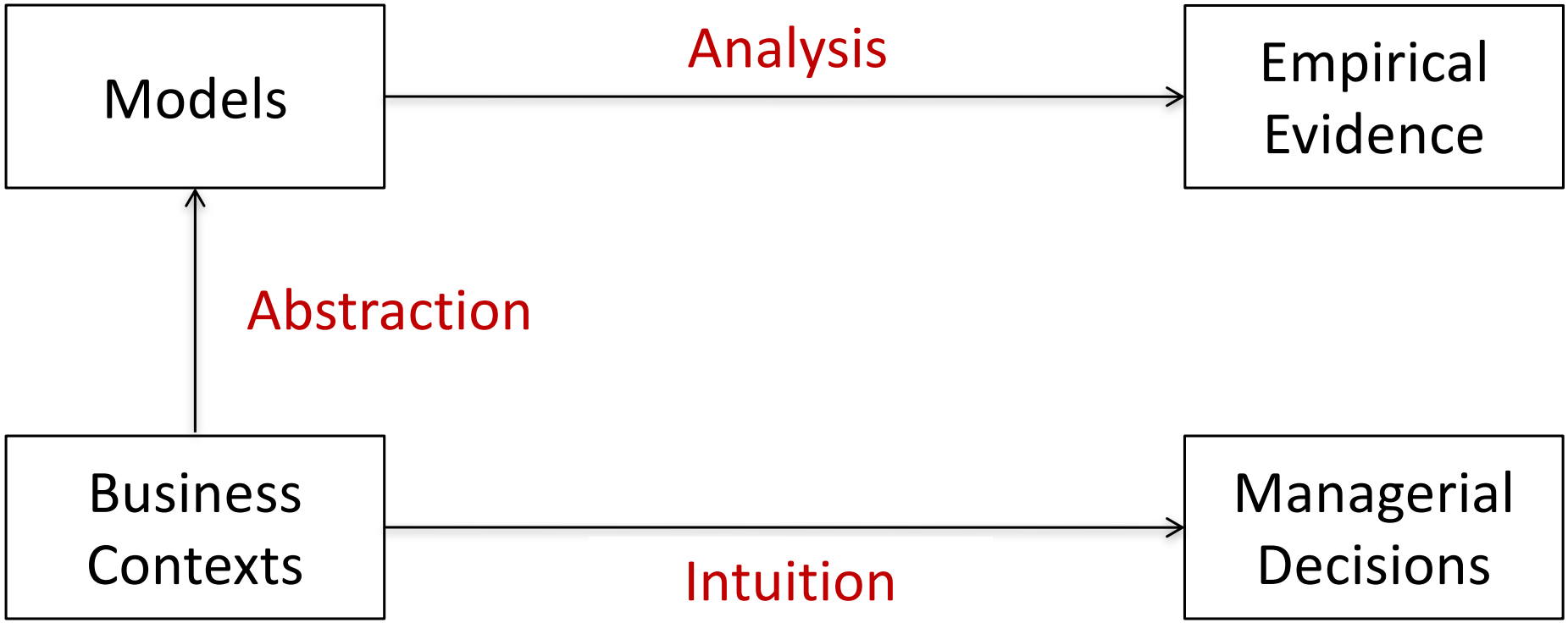
Analytics for Decision Making



Analytics for Decision Making



Analytics for Decision Making



Three Elements of Effects

Baseline Causal Mechanism: X leads to Y

Easy-to-pronounce
Difficult-to-pronounce
Medications

Sales

X: Strategies → Y: Outcomes

Three Elements of Effects

Baseline Causal Mechanism: X leads to Y

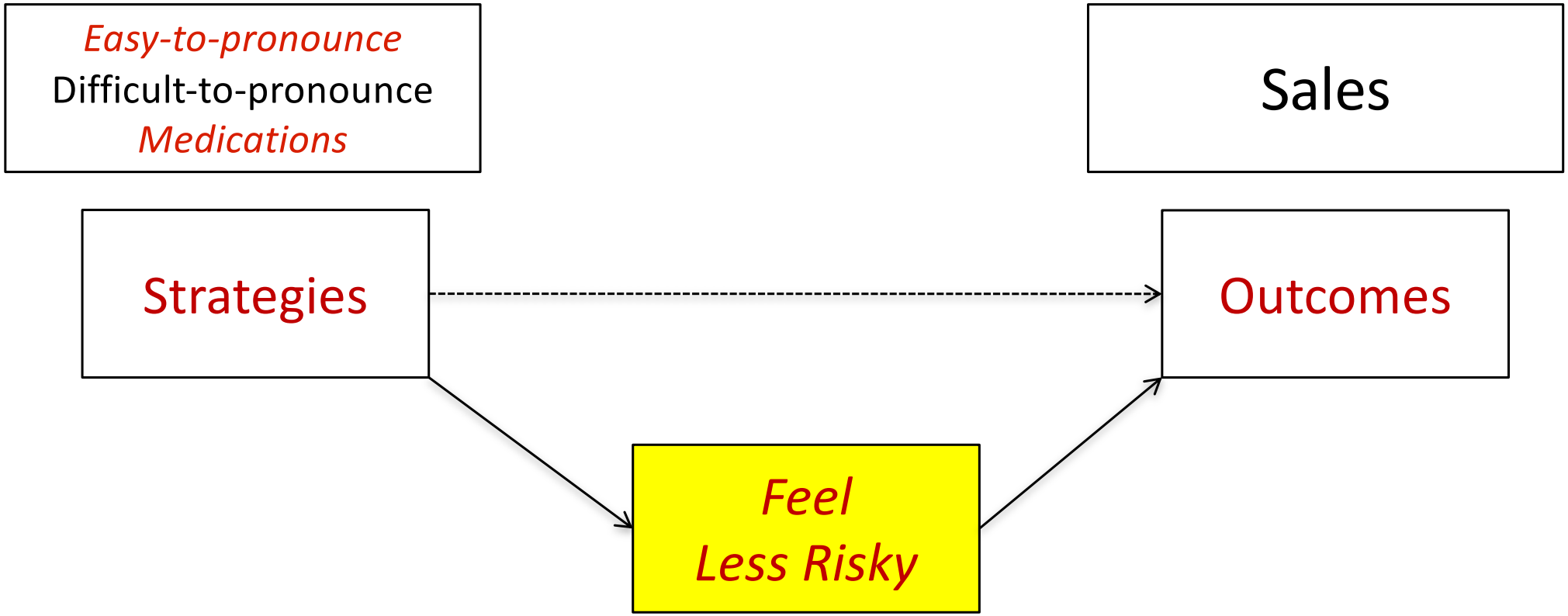
Easy-to-pronounce
Difficult-to-pronounce
Medications

Sales

X: Strategies → Y: Outcomes

Three Elements of Effects

Underlying Mechanism: X leads to Y because of Z



Three Elements of Effects

Baseline Causal Mechanism: X leads to Y

Easy-to-pronounce ?
Difficult-to-pronounce
Roller Coaster

Sales

X: Strategies → Y: Outcomes

Three Elements of Effects

Baseline Causal Mechanism: X leads to Y

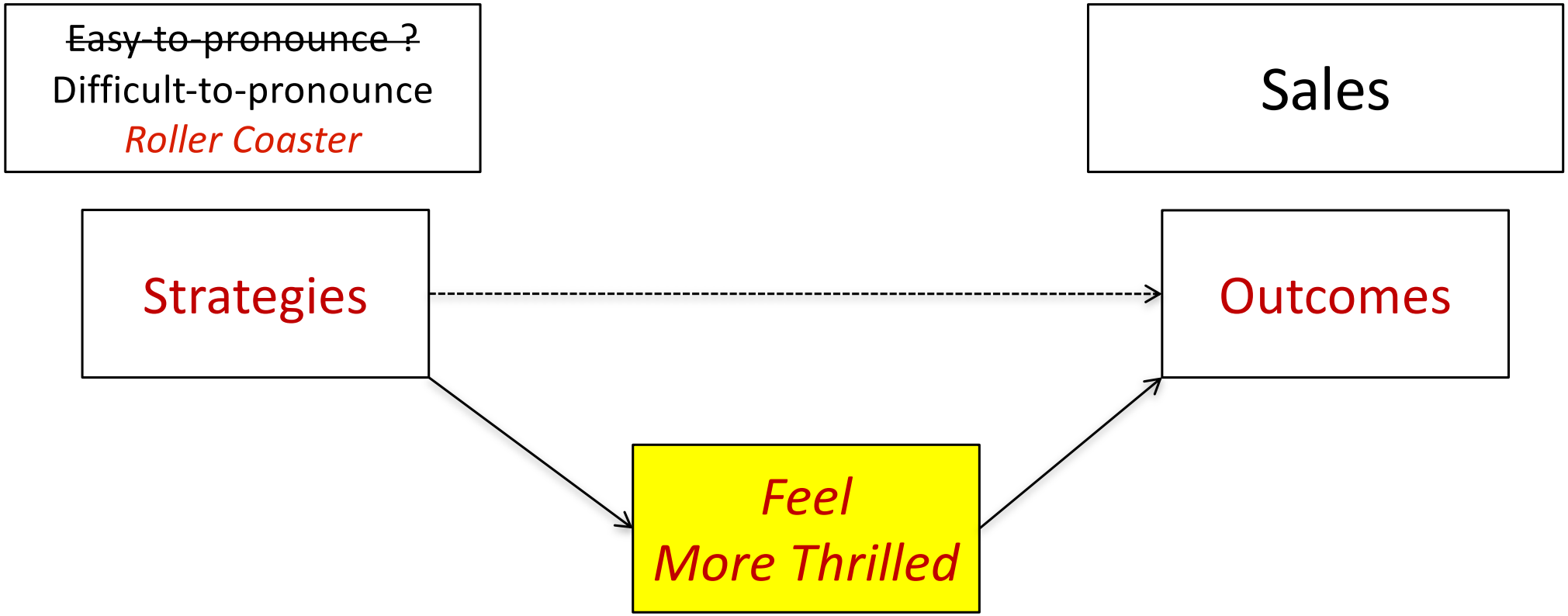
~~Easy-to-pronounce ?~~
Difficult-to-pronounce
Roller Coaster

Sales

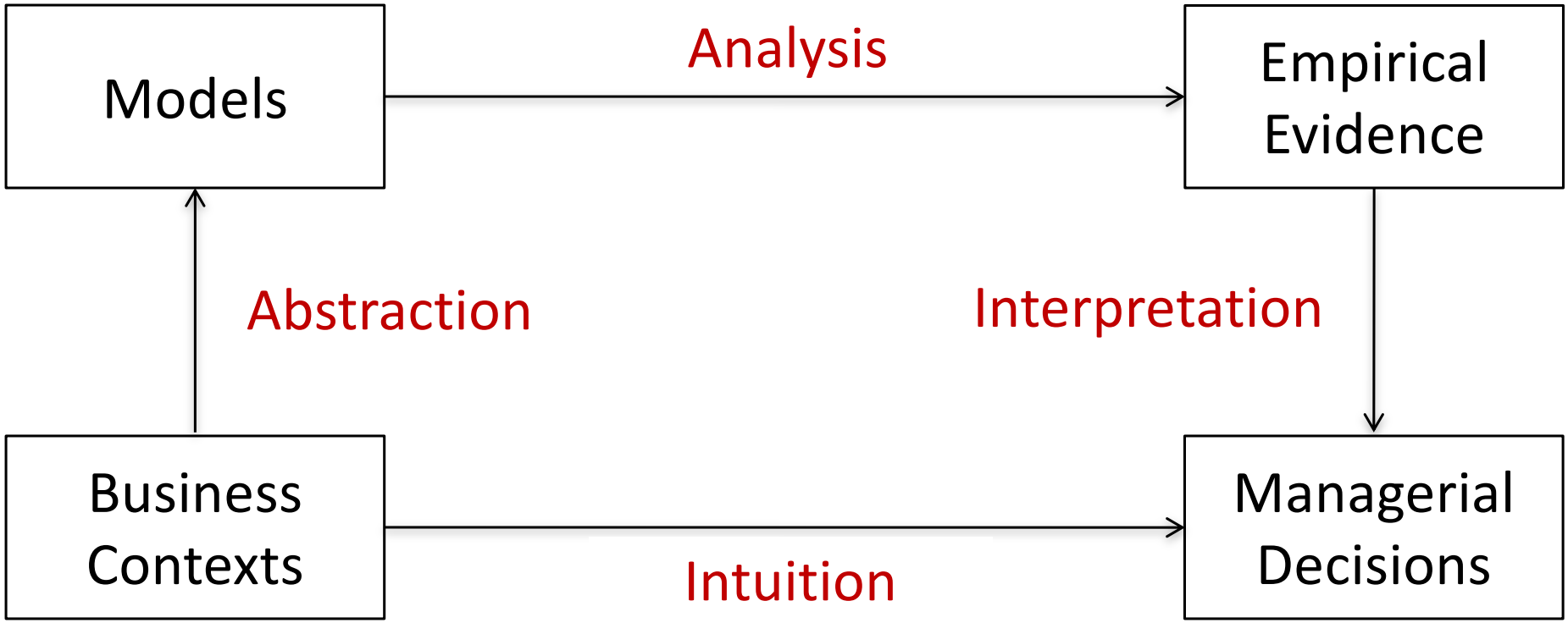
X: Strategies → Y: Outcomes

Three Elements of Effects

Underlying Mechanism: X leads to Y because of Z



Analytics for Decision Making



Empirical Evidence vs. Interpretation?

		Predicted Outcome (y-hat)	
		Positive (presence)	Negative (absence)
Actual Outcome (y)	Positive (presence)		
	Negative (absence)		

Empirical Evidence vs. Interpretation?

		Predicted Outcome (y-hat)	
		Positive (presence)	Negative (absence)
Actual Outcome (y)	Positive (presence)	True Positive	
	Negative (absence)		True Negative

Empirical Evidence vs. Interpretation?

Actual Outcome (y)	Predicted Outcome (y-hat): Positive (presence)	Predicted Outcome (y-hat): Negative (absence)
Positive (presence)	True Positive	False Negative
Negative (absence)	False Positive	True Negative

Data literacy is essential for drawing valid recommendation

Caution!

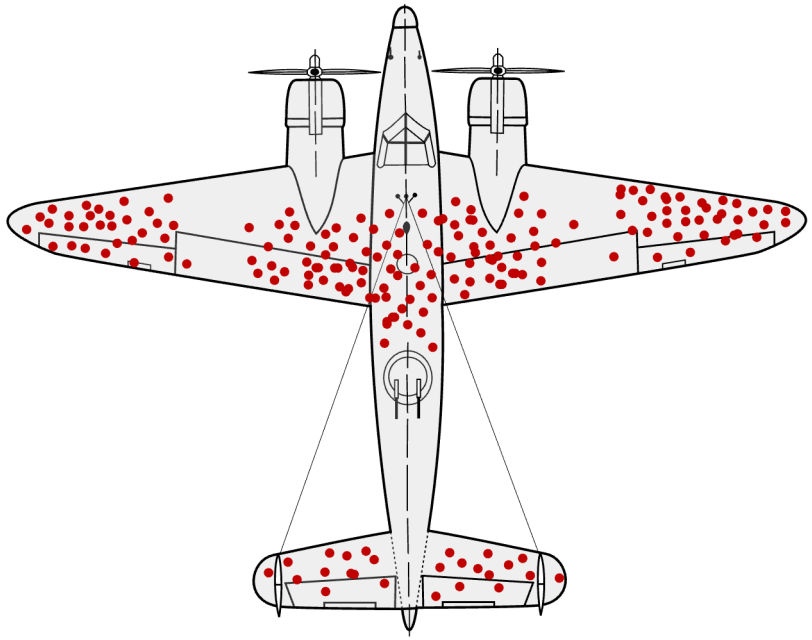
- As we have learned in previous sessions, effective data visualization can assist decision-makers in communicating their findings to others in a clear and concise manner.
- However, it is important to note that data visualization is not a substitute for critical thinking and careful analysis.

Data literacy is essential for drawing valid recommendation

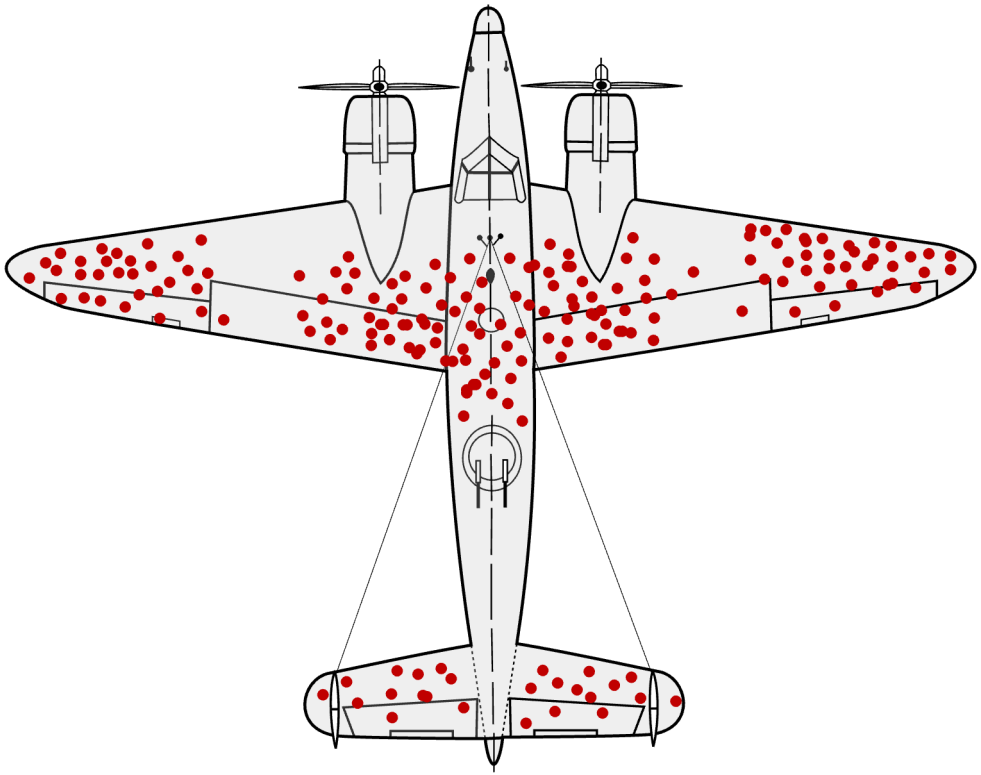
Caution!

- While data visualization can be a useful tool in decision-making, they should be used in conjunction with other data analysis methods to ensure that decisions are based on accurate and complete information.

Case: Survivorship Bias



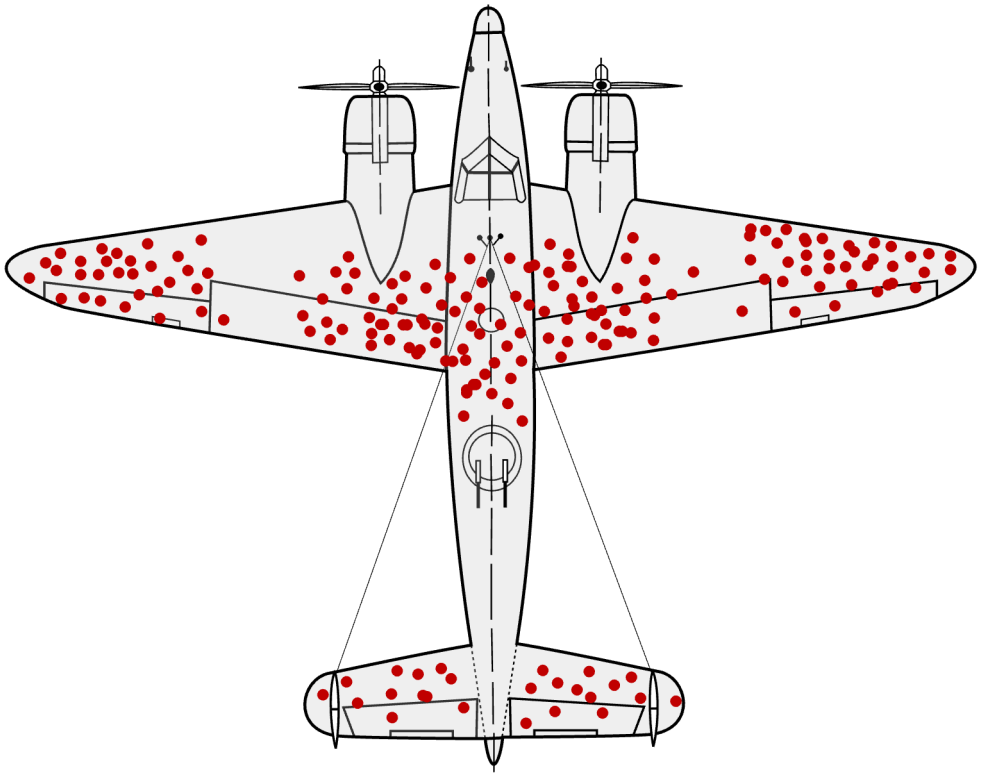
Data literacy is essential for drawing valid recommendation



Survivorship Bias?

- Survivorship bias refers to the tendency to focus on successful outcomes and ignore the unsuccessful outcomes when analyzing a particular set of data.

Data literacy is essential for drawing valid recommendation



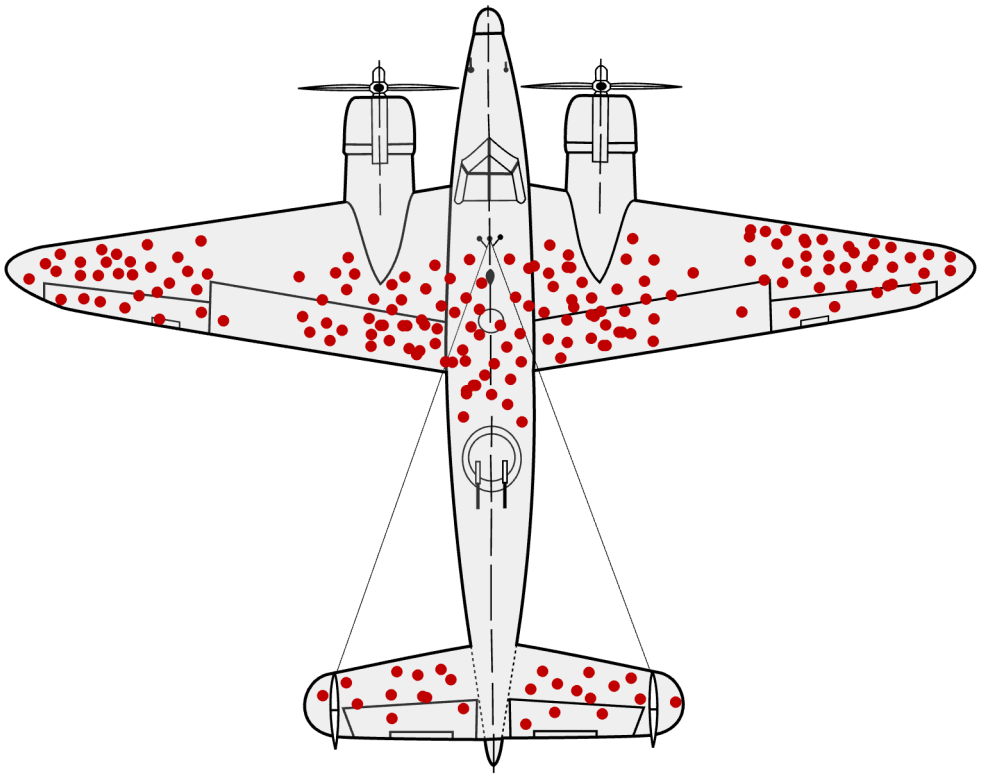
Survivorship Bias?

- This bias occurs when only the data from surviving or successful individuals or entities are analyzed, while the data from those that failed or were unsuccessful are not included in the analysis.
- This can lead to inaccurate conclusions and an incomplete understanding of the data.

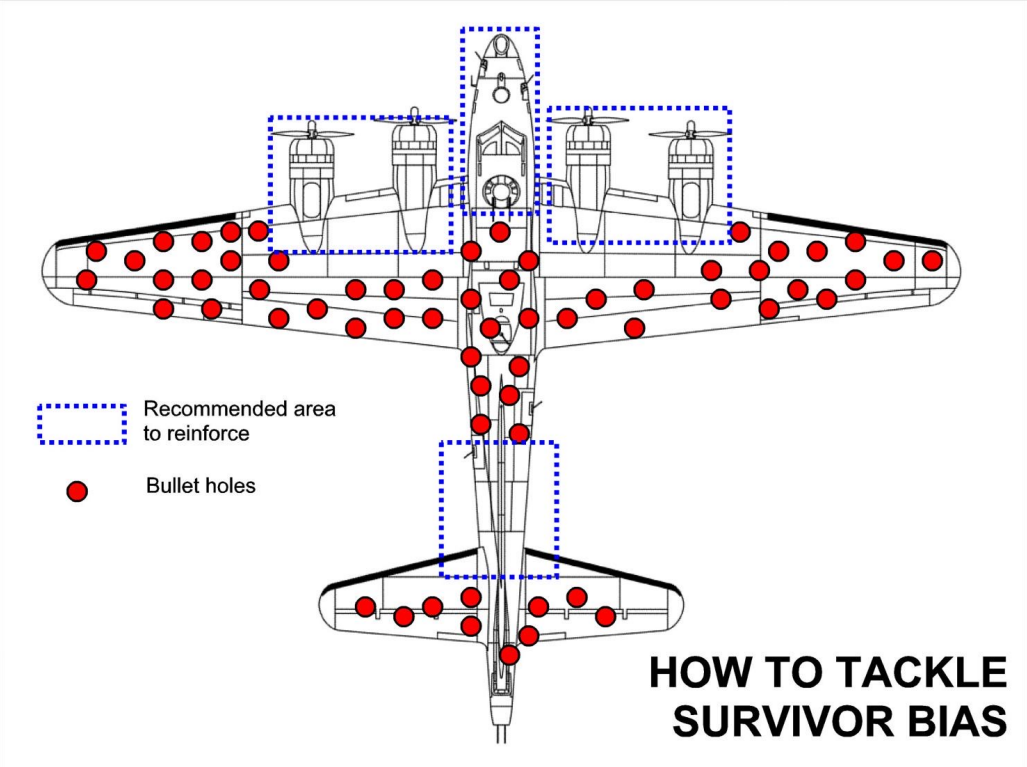
Data literacy is essential for drawing valid recommendation

Why It Matters?

- In the finance industry, survivorship bias can lead to incorrect assumptions and faulty decision-making in areas such as risk assessment, product development, and pricing.



Data literacy is essential for drawing valid recommendation



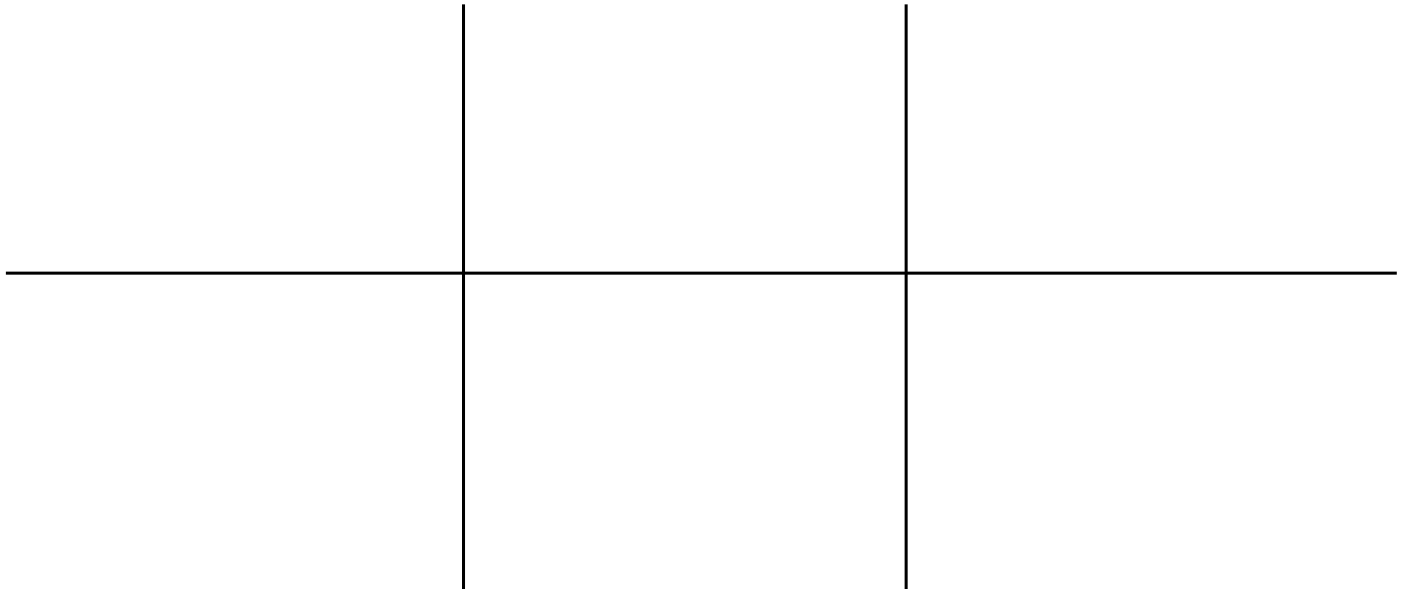
How to Tackle Survivorship Bias?

- To address clustering bias in the finance industry, it's essential to analyze data from various sources, consider multiple factors that could contribute to observed patterns, and use advanced analytics to account for potential bias.
- By doing so, insurers can make more informed decisions about risk assessment, pricing, and underwriting, leading to fairer premiums and more effective risk management.

"Compared with What?"
Difference-in-Differences Revisited

Implementation of Your Strategy

Before After



"Compared with What?"
Difference-in-Differences Revisited

Implementation of Your Strategy

Before	After
+19%	+38%

"Compared with What?"
Difference-in-Differences Revisited

Implementation of Your Strategy

	Before	After	
Year with strategy	+19%	+38%	+19%-p

"Compared with What?"
Difference-in-Differences Revisited

Implementation of Your Strategy

Years with vs. w/o strategy	Before	After	
Year with strategy	+19%	+38%	+19%-p
Years without strategy	+18%	+54%	+36%-p

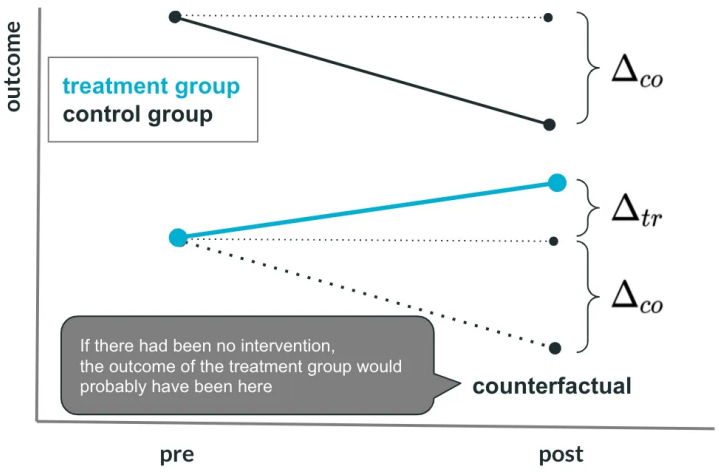
"Compared with What?"
Difference-in-Differences Revisited

Implementation of Your Strategy

Years with vs. w/o strategy	Before	After	Impact (Invalid)
Year with strategy	+19%	+38%	+19%-p
Years without strategy	+18%	+54%	+36%-p

Difference in Differences (DID)

$$\mathrm{ATT}^{\mathrm{did}} = \Delta_{\mathrm{tr}} - \Delta_{\mathrm{co}}$$



pros	Excellent interpretability
cons	Parallel trend assumption • DID requires the selection of a control group that has parallel trends in the treatment group and the pre-intervention period. • This selection may be arbitrary in some cases. e.t.c.

	After	Before
Treatment	A	B
Control	C	D

Difference-in-Differences

(A — B) — (C — D) > 0

(A — B) — (C — D) = 0

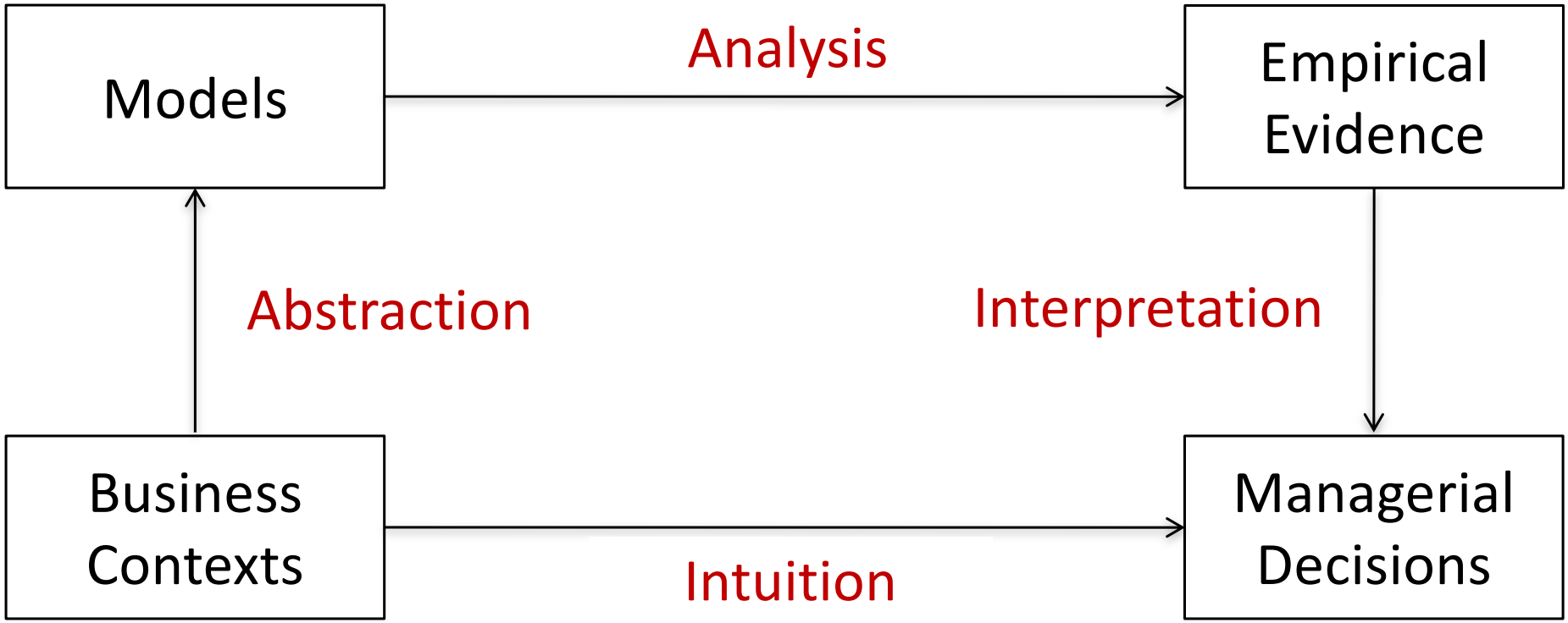
(A — B) — (C — D) < 0

"Compared with What?"
Difference-in-Differences Revisited

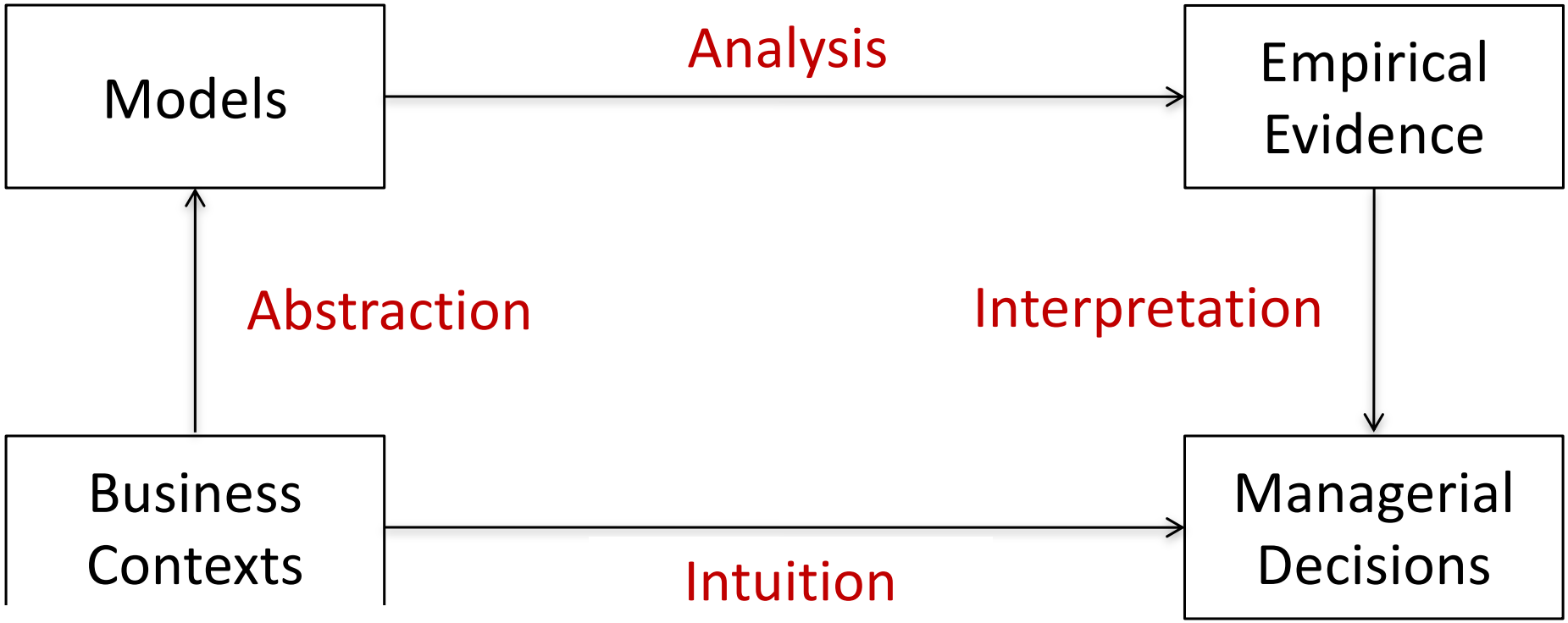
Implementation of Your Strategy

Years with vs. w/o strategy	Before	After	*True Impact = — 17%-p*
Year with strategy	+19%	+38%	+19%-p
Years without strategy	+18%	+54%	+36%-p

Analytics for Decision Making



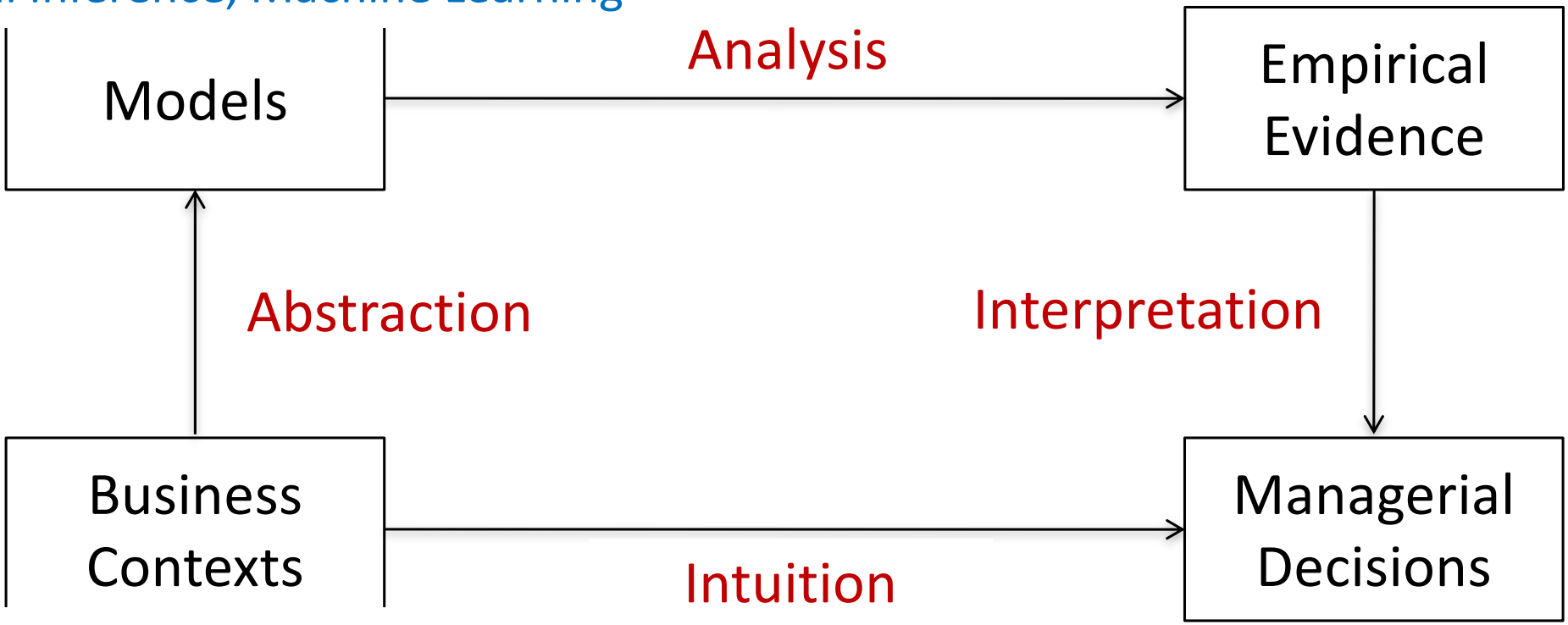
Analytics for Decision Making



How much to charge for the policy?

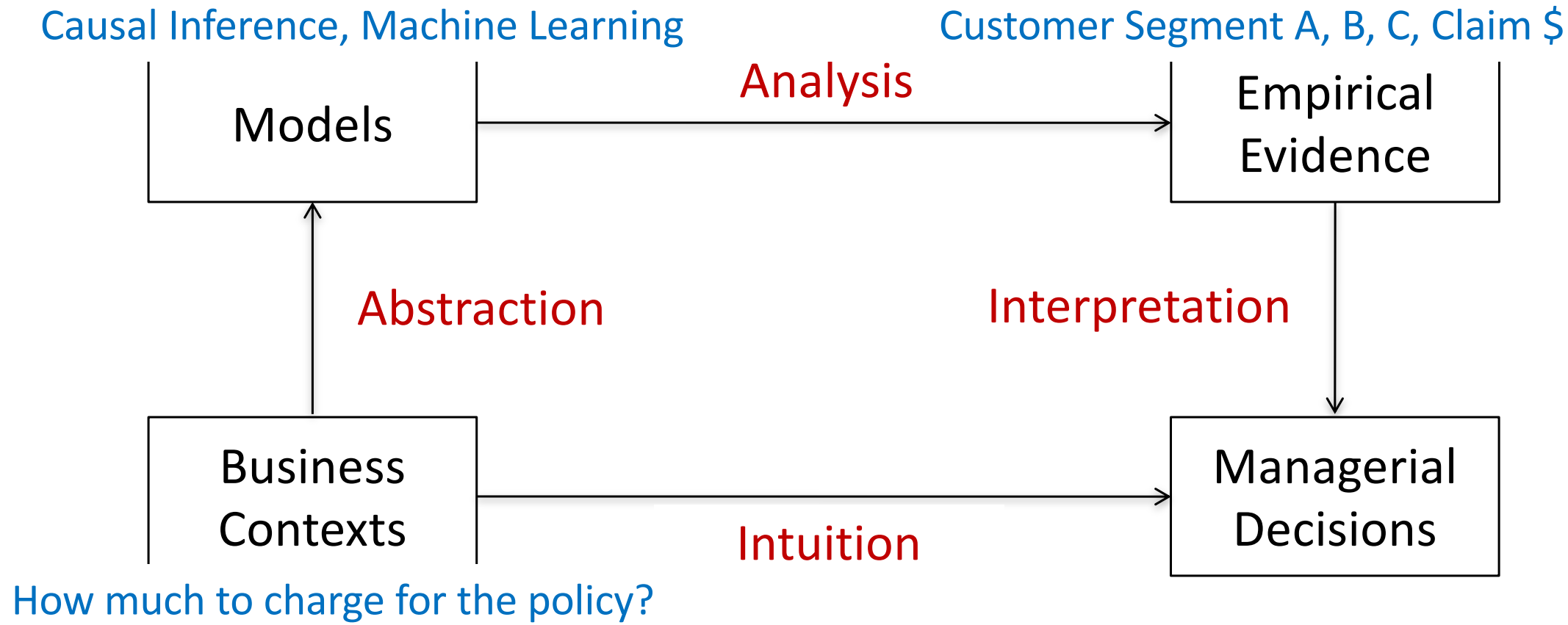
Analytics for Decision Making

Causal Inference, Machine Learning

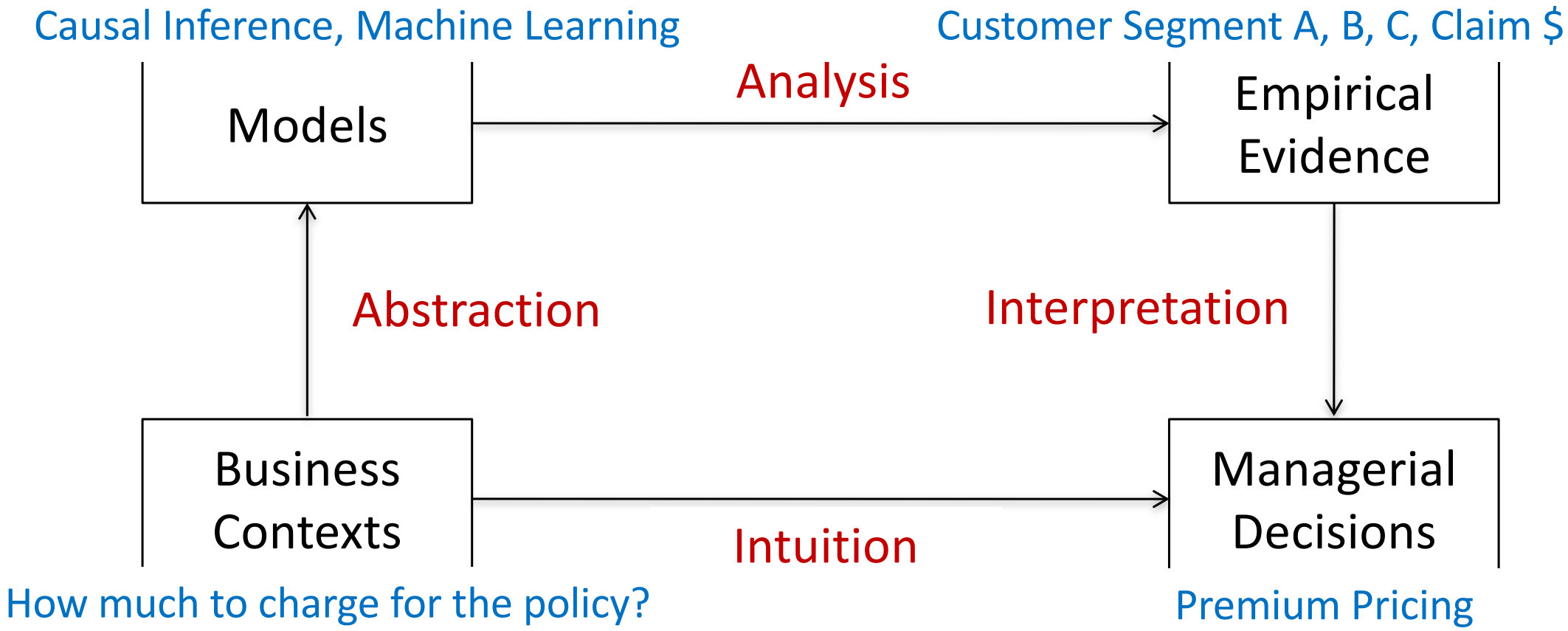


How much to charge for the policy?

Analytics for Decision Making



Analytics for Decision Making



Agenda (Workflow of Data-Driven Decision Making)

Define

Collect

Import

Wrangle

Visualize

Model

Report

Deploy

Remember this? (circa 2002)

"…because as we know,
there are known knowns;
there are things we know we know.

We also know there are known unknowns;
that is to say
we know there are some things we do not know.

But there are also unknown unknowns…"

Quadrants of Knowledge

	known	unknown
known		
unknown		

Quadrants of Knowledge

Knowledge

known unknown

Awareness

known

unknown

Quadrants of Knowledge

Knowledge

Awareness	known	unknown
known	Known-Knowns Things we are aware of and understand	
unknown		

Quadrants of Knowledge

Knowledge

Awareness	known	unknown
known	Known-Knowns Things we are aware of and understand	
unknown		

- These are the factors we understand and have data on.
- In the finance industry, this may include historical claims data, demographics, and other relevant information.

Quadrants of Knowledge

Knowledge

Awareness	known	unknown
known	Known-Knowns Things we are aware of and understand	
unknown		

- DDDM enables insurers to analyze these known-knowns to make informed decisions, such as setting appropriate premium rates based on risk profiles, identifying patterns in fraudulent claims, and optimizing underwriting processes.

Quadrants of Knowledge

Knowledge

Awareness	known	unknown
known		Known-Unknowns Things we are aware of but don't understand
unknown		

Quadrants of Knowledge

Knowledge

known | unknown

Awareness

known

unknown

- These are factors that we have data on, but their implications may not be immediately apparent.
- In the finance industry, this may include emerging trends, such as new technologies and social changes, that could impact the risk landscape.

Known-Unknowns

Things we are aware of but don't understand

Quadrants of Knowledge

Knowledge

known unknown

Awareness

known

unknown

Unknowns-Known
Things we understand
but are not aware of

Quadrants of Knowledge

Knowledge

Awareness

	known	unknown
known		
unknown	Unknowns-Known Things we understand but are not aware of	

- These are factors that we know exist but don't have sufficient data on.
- DDDM can help insurers identify gaps in their data and prioritize efforts to gather more information on these known-unknowns.
- Additionally, advanced analytics techniques like predictive modeling can be used to make educated guesses based on the available data.

Quadrants of Knowledge

Knowledge

known | unknown

Awareness

known

unknown

Unknown-Unknowns

Things we are neither aware of nor understand

Quadrants of Knowledge

Knowledge

Awareness

- These are factors that we are not aware of and don't have data on.
- While DDDM cannot directly address unknown-unknowns, it can help insurers build a more resilient and adaptable decision-making framework.
- By continually monitoring and analyzing data, insurers can become more agile in identifying and responding to emerging risks and trends.

unknown

Unknown-Unknowns
Things we are neither aware of nor understand

Quadrants of Knowledge

Awareness	Knowledge: known	Knowledge: unknown
known	Known-Knowns	Known-Unknowns
unknown	Unknowns-Known	Unknown-Unknowns

Data-Driven Decision Making

Decision-Making with Deployment

- Business analytics should provide decision-makers with the information they need to make informed decisions quickly and efficiently.

- By presenting data in a clear and concise way, decision-makers can quickly identify trends, patterns, and outliers, and make decisions based on the insights gained from the data.

Data-Driven Decision Making

A Case: New York Fed (circa 2008)

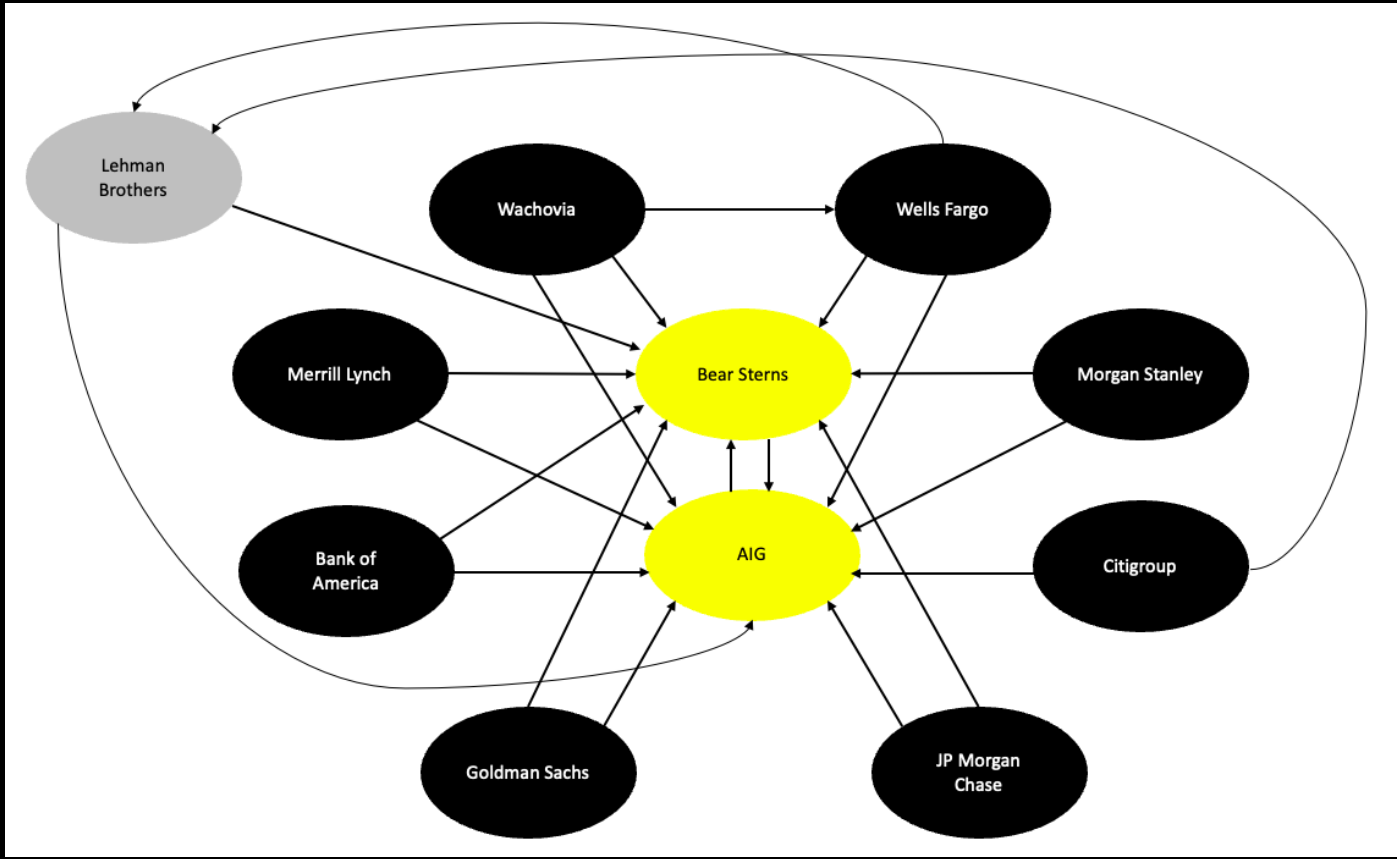
Data-Driven Decision Making

What Happened?

- In 2008 during the financial crisis, Timothy Geithner, who was the head of the New York Fed at the time, had to decide which banks to bail out.
- Instead of just looking at the size or resources of the banks, he used models that visualized the interconnections between financial institutions.

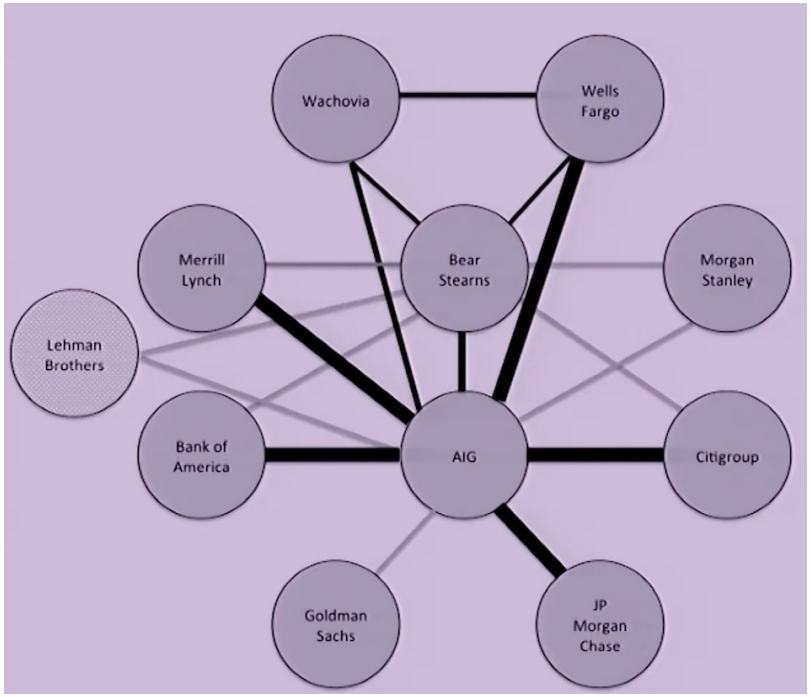
A Case: New York Fed (circa 2008)

Deployment to Decision Making:
A Network Model of Risk Feedbacks
among major financial institutions



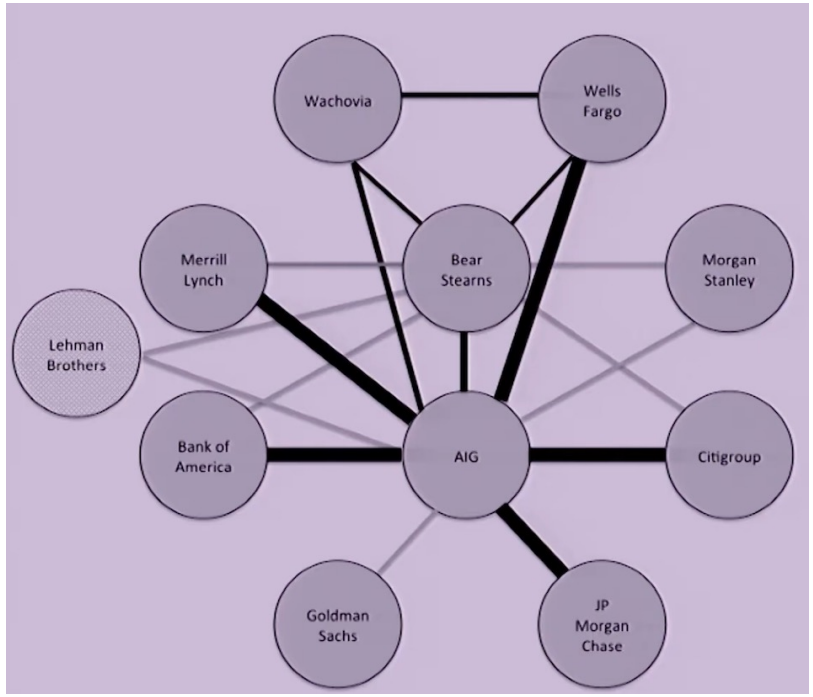
Data-Driven Decision Making

Deployment of Analytic Models



Data-Driven Decision Making

Deployment of Analytic Models



Aid Decision Making

- As you could see on the left (a simplified network plot), AIG had thick lines connecting it to other institutions, meaning that if AIG were to fail, it would have a significant impact on the others.
- On the other hand, Lehman Brothers had only thin lines connecting it to other institutions, indicating that if it were to fail, it would have a limited impact.
- By using these models, Geithner was able to make more informed decisions.

Thank you for being curious ☺ See you in June!