What is this thing called Data Science?

Phil Henrickson, PhD
AE Business Solutions
Delivered April 18, 2019

about me

about me

PhD in political science

about me

PhD in political science

Reaction #1)

"why would you do that"

"is political science really a science"

GOP lawmaker doubles down on criticizing Kerry's political science degree as not 'science'

BY CHRIS MILLS RODRIGO - 04/11/19 07:17 PM EDT

591 COMMENTS

Rep. Thomas Massie (R-Ky.), who has two engineering degrees from the Massachusetts Institute of Technology, began his line of questioning attempting to undermine Kerry's authority on the issue by asking him about his "science degree" from Yale.

Kerry explained it was actually a bachelor of arts in political science.

"So, it's not really science," Massie said. "I think it's somewhat appropriate that someone with a pseudoscience degree is here pushing pseudoscience in front of our committee today."

about me

PhD in political science

background in academic research
on international conflict and
political violence

about me

PhD in political science

background in academic research on international conflict and political violence

machine learning, predictive modeling, and causal inference

about me

PhD in political science

background in academic research on international conflict and political violence

machine learning, predictive modeling, and causal inference

about me

PhD in political science

Reaction #1)

"why would you do that"

"is political science really a science"

about me

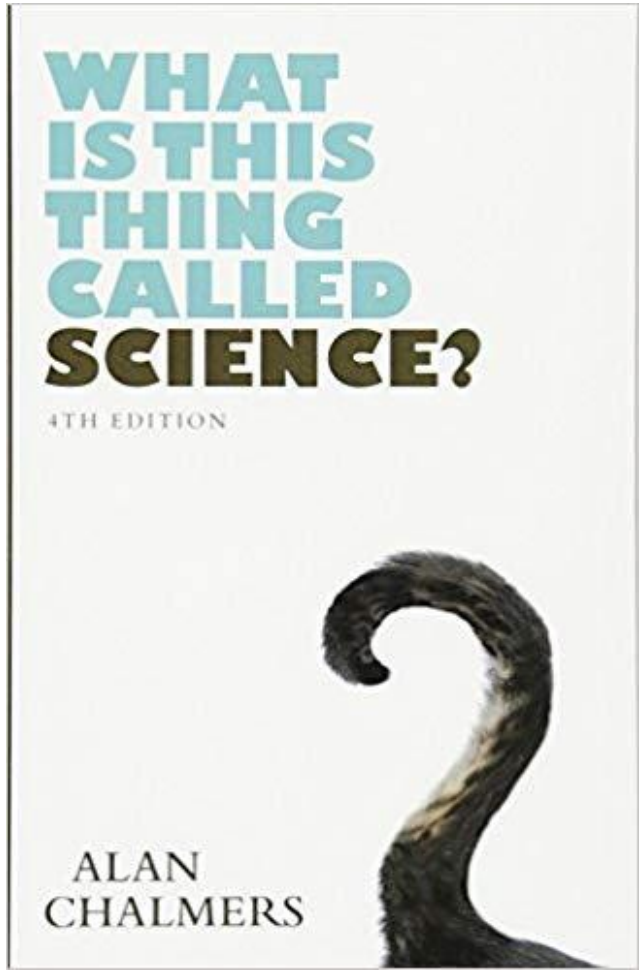
PhD in political science

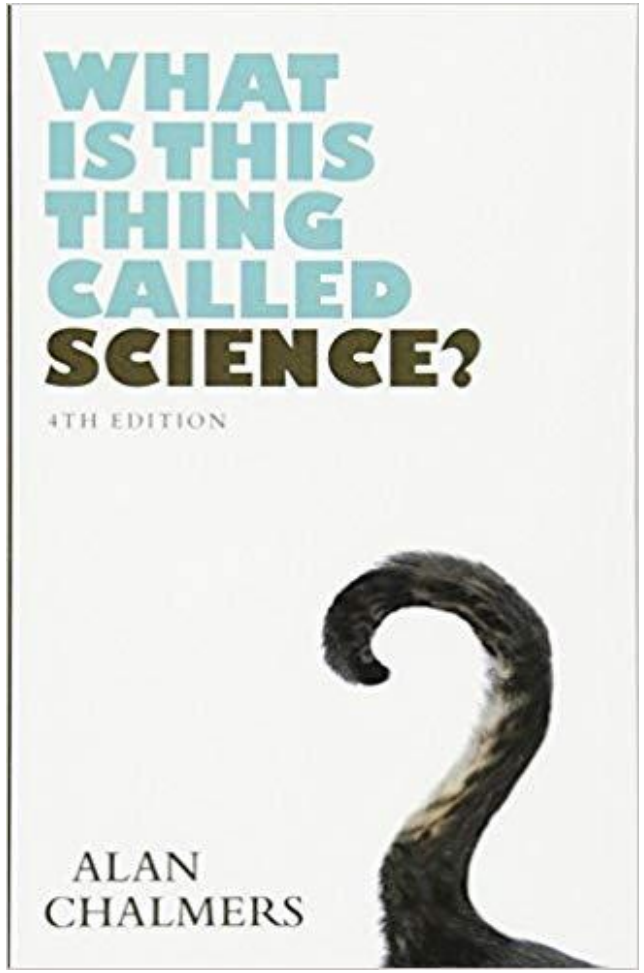
Reaction #1) "why would you do that"

"is political science really a science"

Reaction #2) "you're Dr. Phil!"

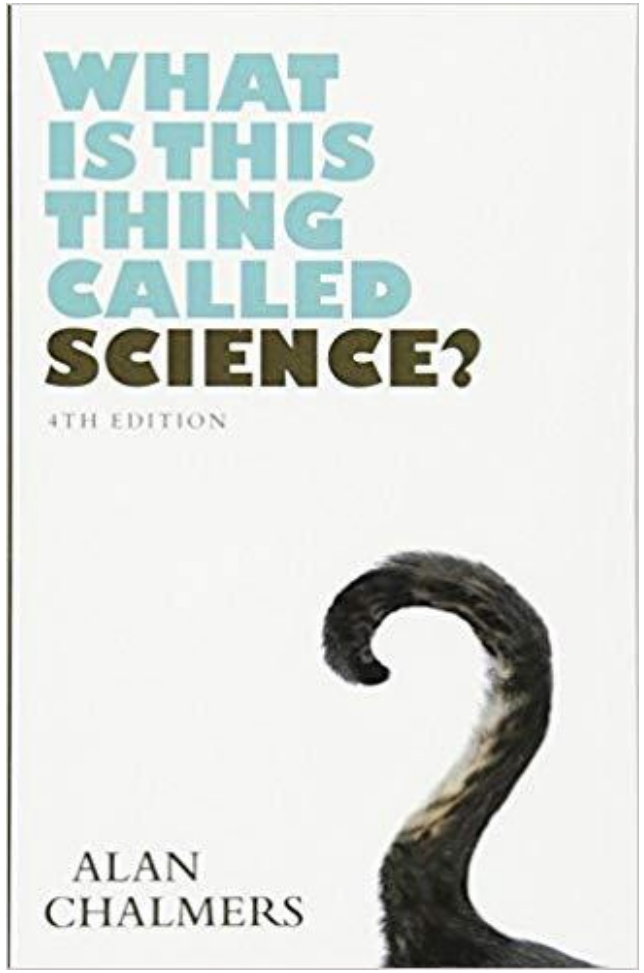
What is this thing called Data Science?


WHAT IS THIS THING CALLED SCIENCE?
4TH EDITION
ALAN CHALMERS



…it is a widely held belief that there is something special about **science** and its methods.

The naming of some claim or line of reasoning or piece of research [as] "**scientific**" is done in a way that is intended to imply some kind of merit or special kind of reliability.



But what, if anything, is so special about science?

Harvard Business Review

ARTWORK: TAMAR COHEN, ANDREW J BUBOLTZ, 2011, SILK SCREEN ON A PAGE FROM A HIGH SCHOOL YEARBOOK, 8.5" X 12"

DATA

Data Scientist: The Sexiest Job of the 21st Century

WHAT TO READ NEXT

ON-DEMAND WEBINAR

BI Trend #1: AI and machine learning

WATCH NOW

SHARE

Keeping up with the rapid pace of change in the BI ecosystem gets more challenging every year. With that in mind, we've put together a series of webinars focused on the Top 10 BI Trends in 2018.

LENGTH:
60 MIN

FORTUNE

HOME SUBSCRIBE

TECH • SCIENCE

'Sexiest Job' Ignites Talent Wars as Demand for Data Scientists Soars

Forbes

Billionaires Innovation Leadership Money Consumer Industry Lifestyle

IBM Predicts Demand For Data Scientists Will Soar 28% By 2020

Louis Columbus Contributor

InfoWorld FROM IDG

INSIDER Sign In | Register

Why data science and machine learning are the fastest growing jobs in the US

The US could have as many as 250,000 open data science jobs by 2024, and the data science skills gap will find companies scrambling to train or hire talent in the coming years

MORE LIKE

Forbes

Billionaires Innovation Leadership Money Consumer Industry Lifestyle

15,685 views | Dec 11, 2017, 08:32pm

LinkedIn's Fastest-Growing Jobs Today Are In Data Science And Machine Learning

Louis Columbus Contributor

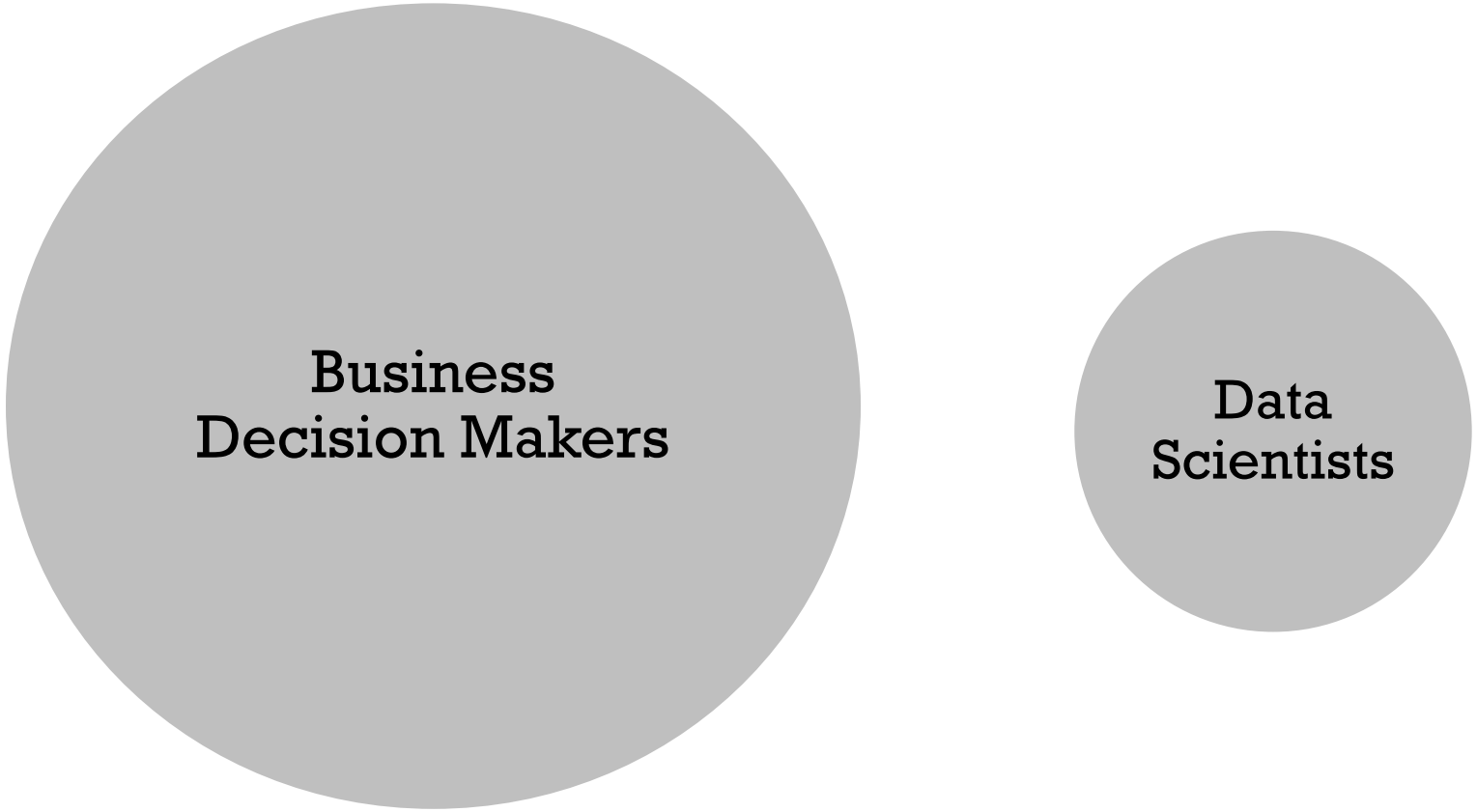
TWEET THIS

Machine Learning Engineers, Data Scientists, and Big Data Engineers rank among the top emerging jobs on LinkedIn.

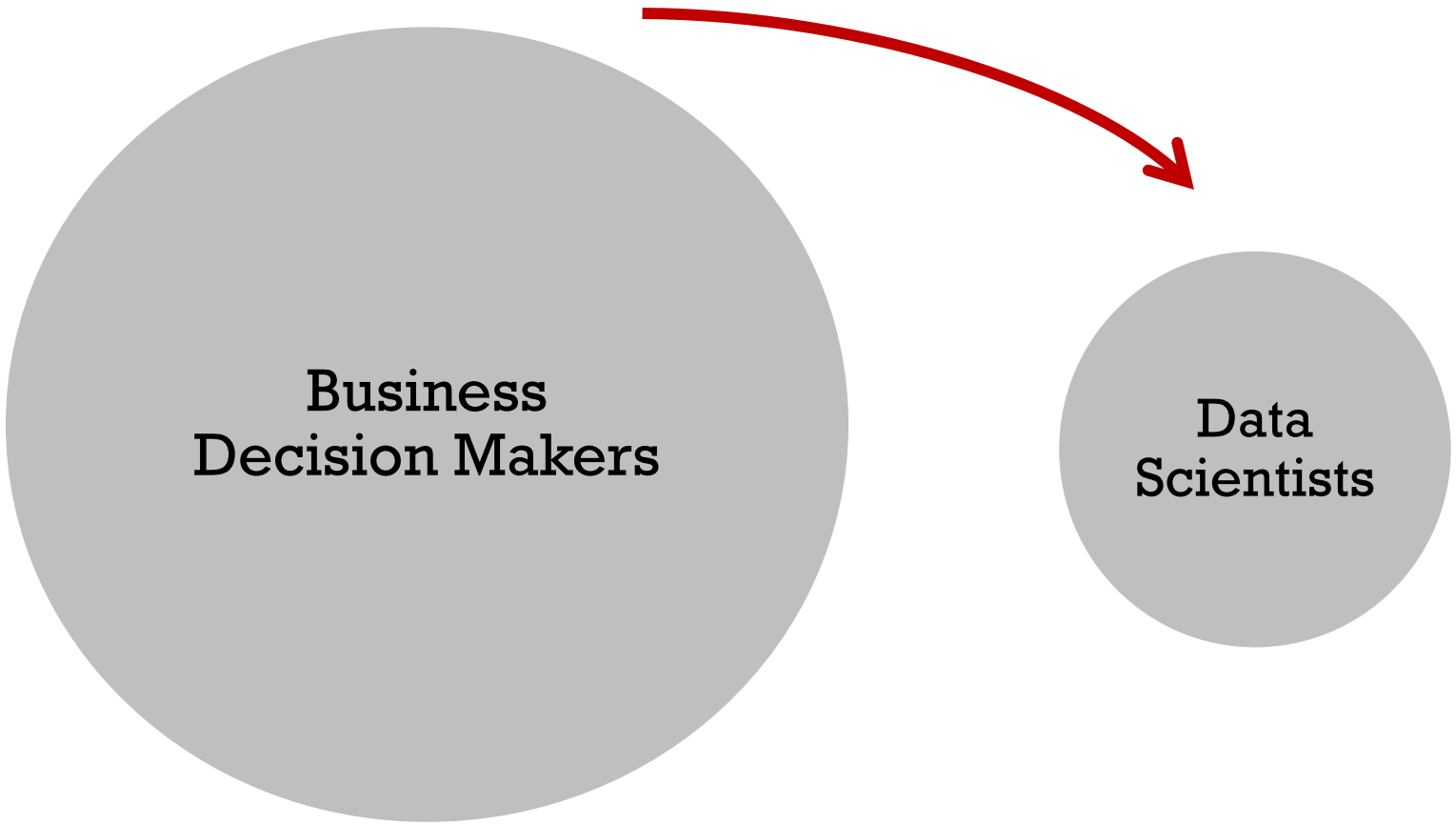
There are currently 1,829 open Machine Learning Engineering positions on LinkedIn.

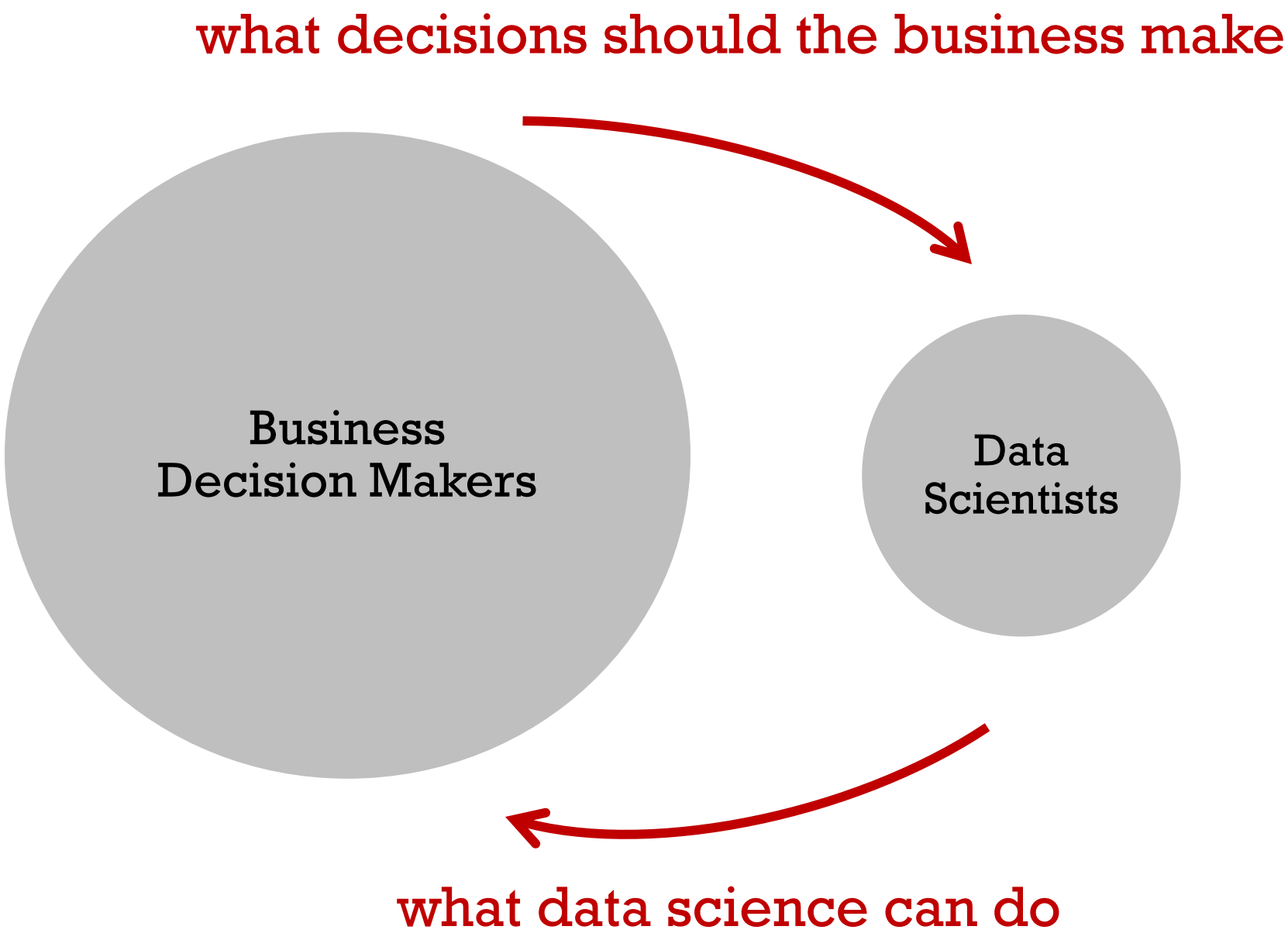
What, if anything, is special about **data science**?

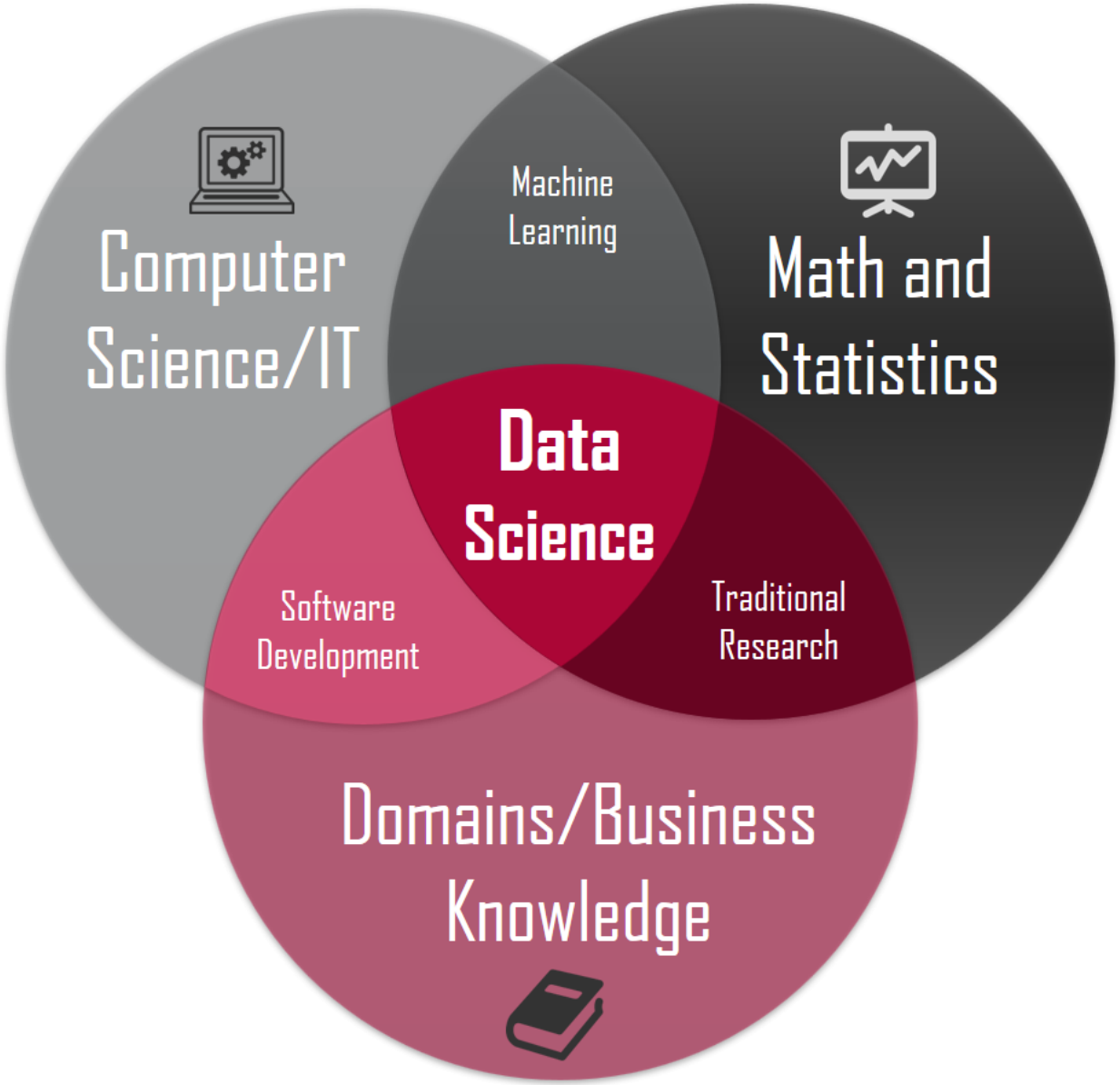
Business
Decision Makers

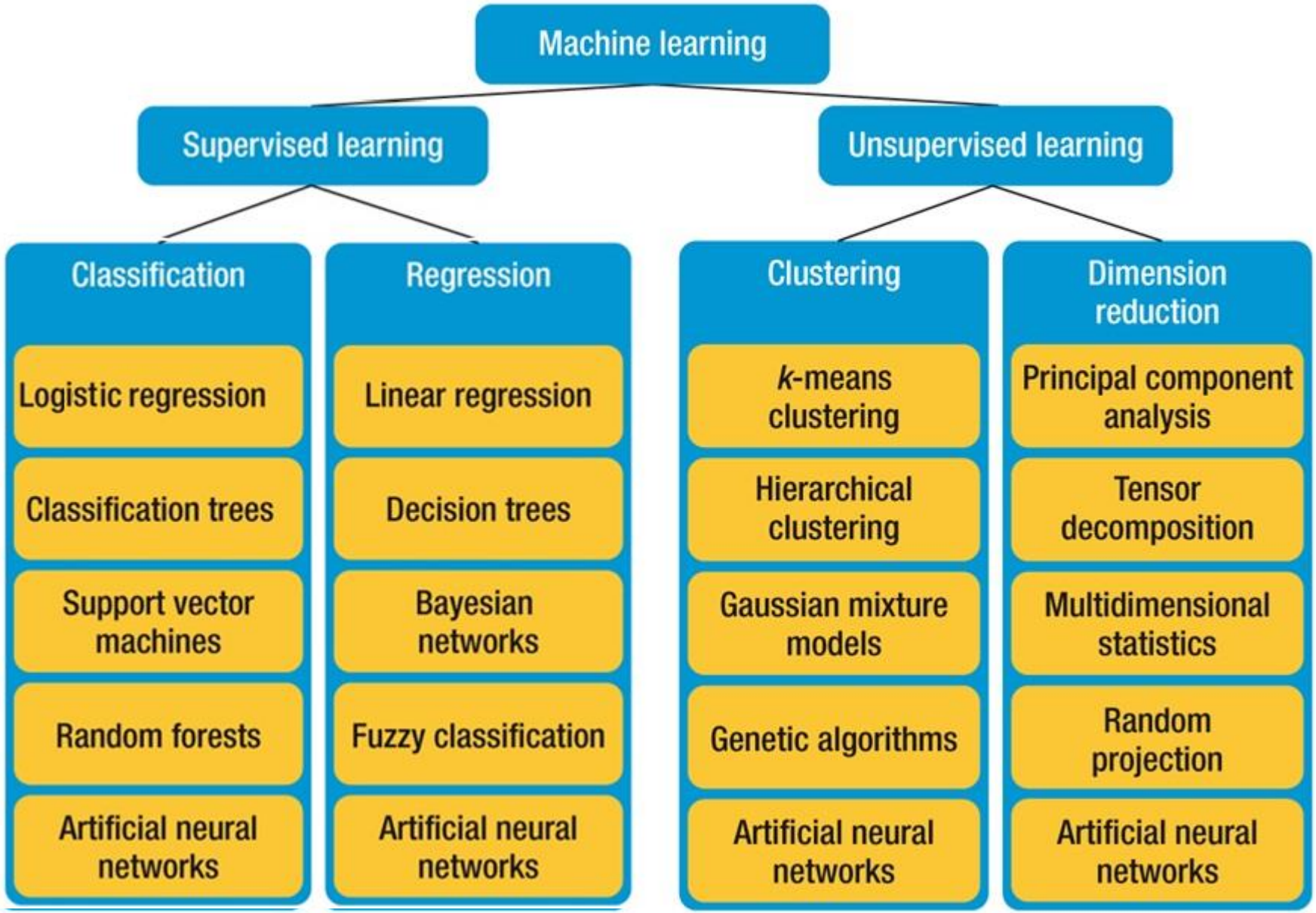

Business
Decision Makers
Data
Scientists

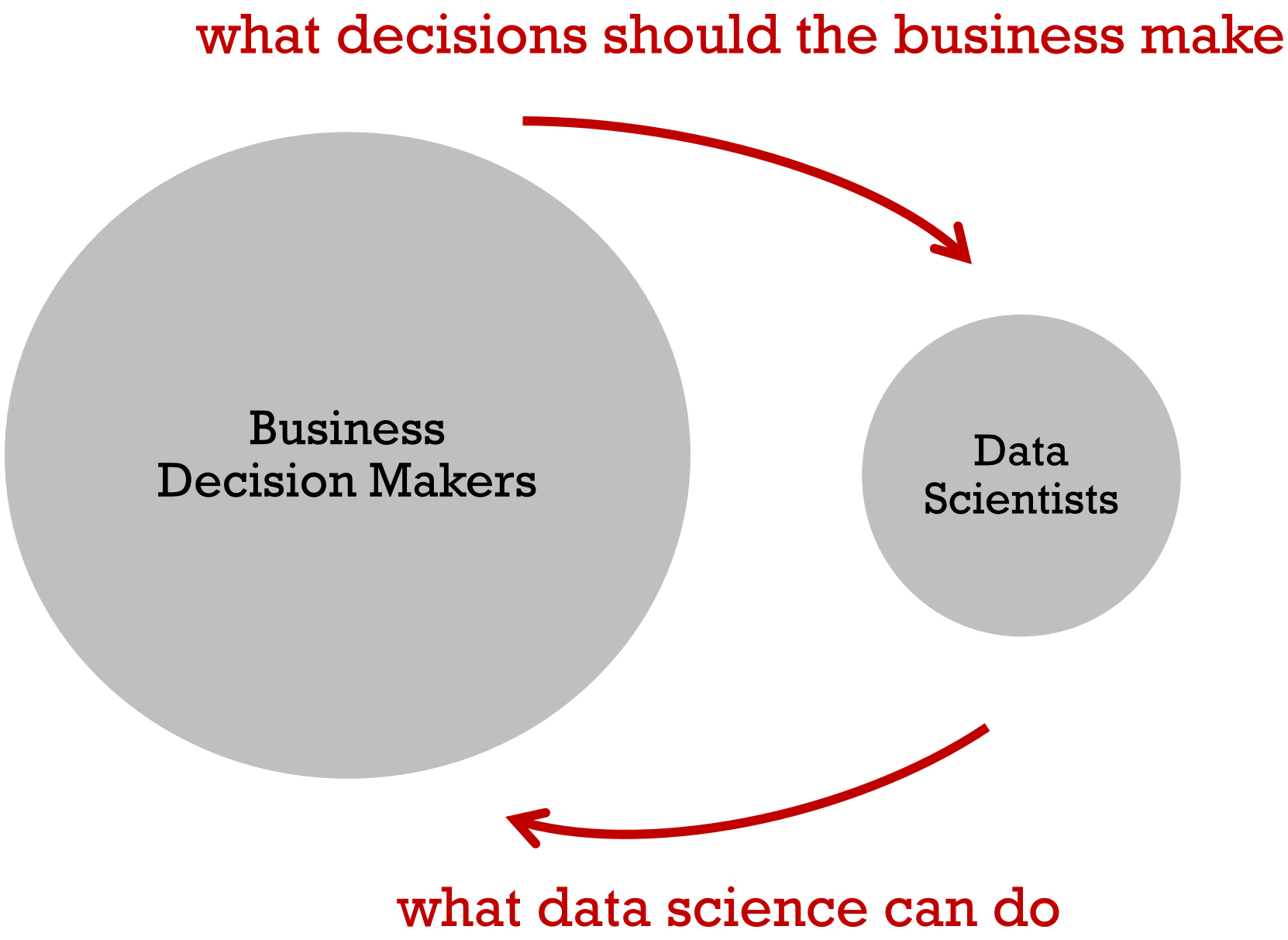
what decisions should the business make




what decisions should the business make
Business
Decision Makers
Data
Scientists
what data science can do


Computer Science/IT
Machine Learning
Math and Statistics
Data Science
Software Development
Traditional Research
Domains/Business Knowledge


Machine learning
Supervised learning
Unsupervised learning
Classification
Logistic regression
Classification trees
Support vector machines
Random forests
Artificial neural networks
Regression
Linear regression
Decision trees
Bayesian networks
Fuzzy classification
Artificial neural networks
Clustering
k-means clustering
Hierarchical clustering
Gaussian mixture models
Genetic algorithms
Artificial neural networks
Dimension reduction
Principal component analysis
Tensor decomposition
Multidimensional statistics
Random projection
Artificial neural networks


what decisions should the business make
Business
Decision Makers
Data
Scientists
what data science can do

what decisions should the business make

Business
Decision Makers

Data
Scientists

what data science can do

what decisions should government make

Political
Decision Makers

Scientific
Researchers

what science can do

what decisions should the business make

Business
Decision Makers

Data
Scientists

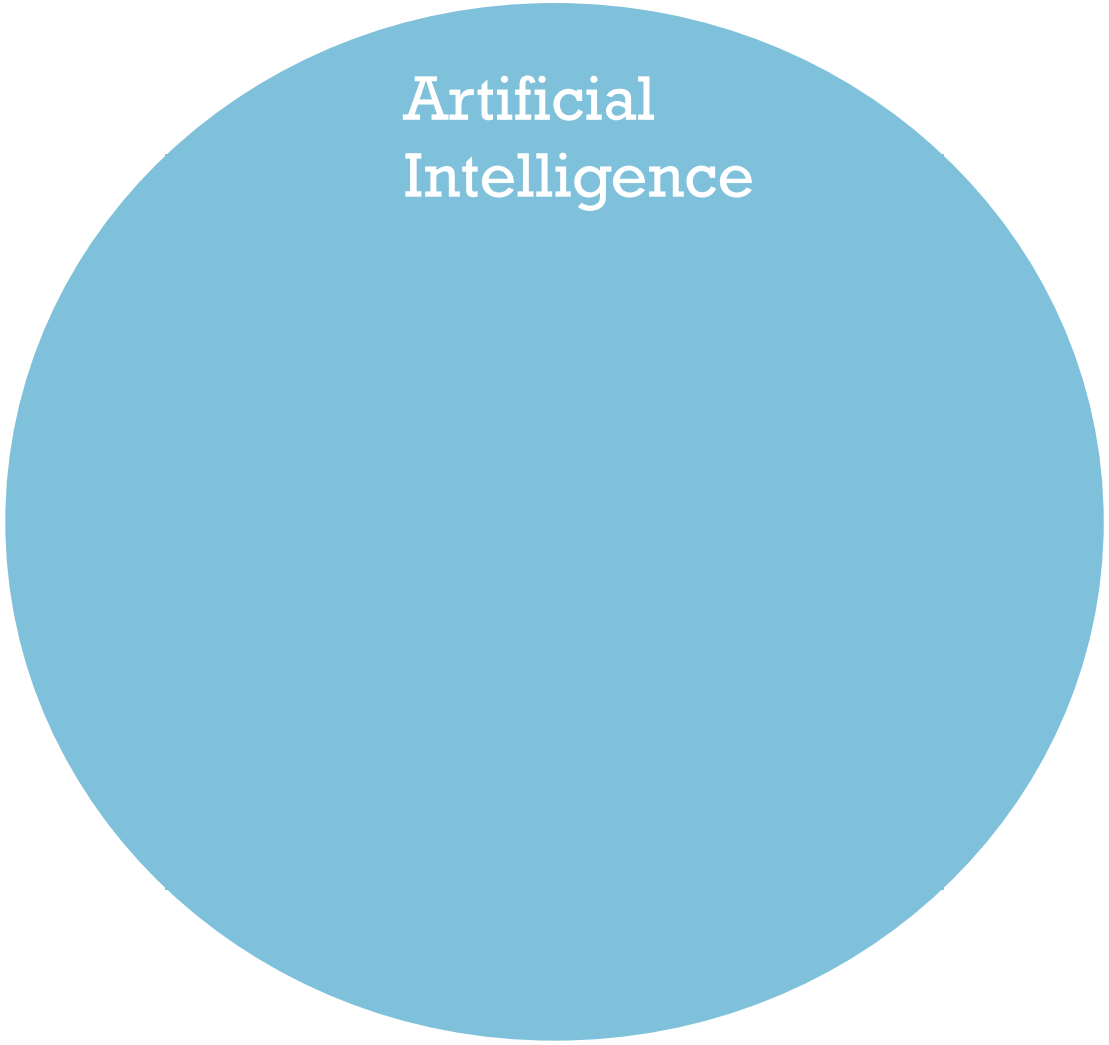
what data science can do

1) What can data science do?

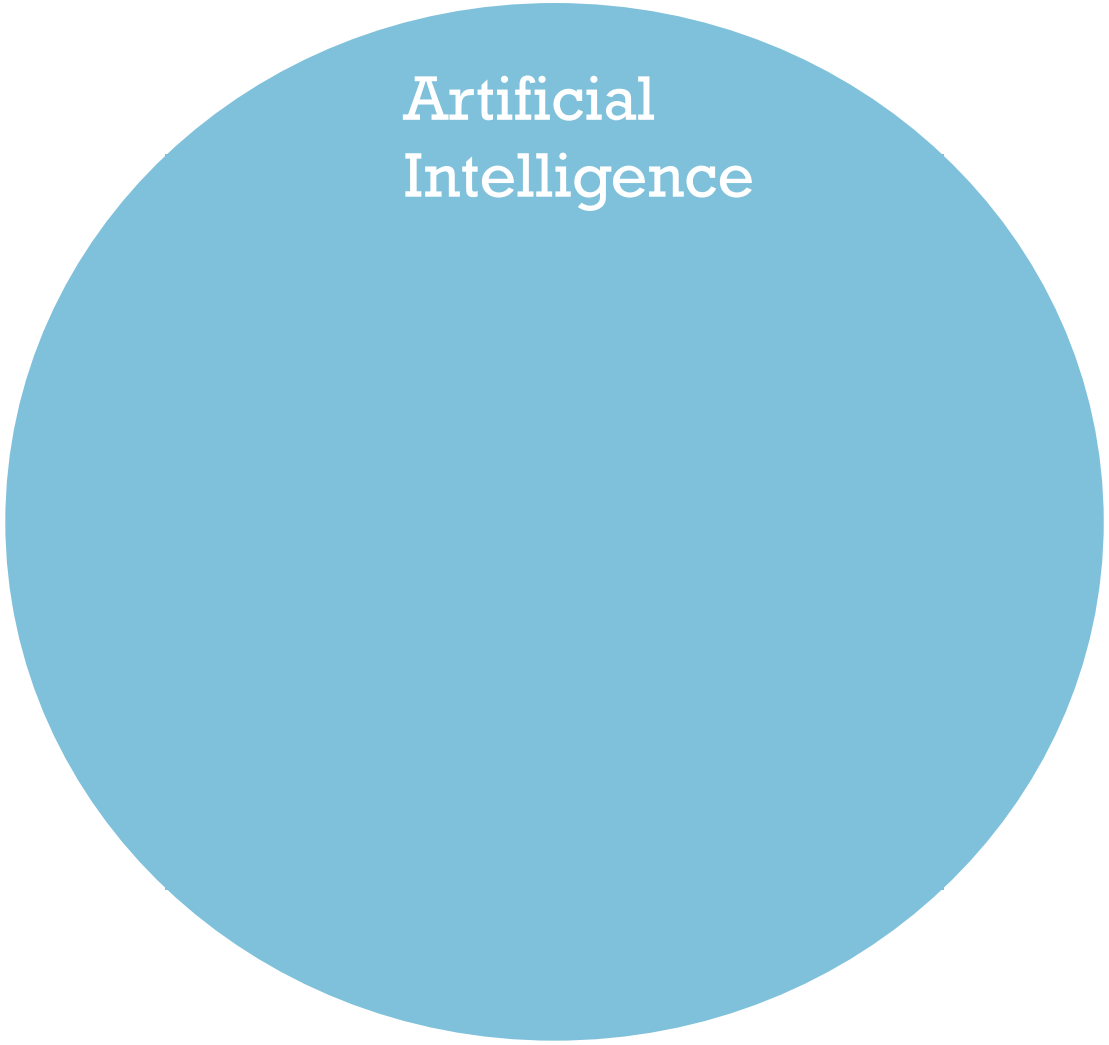
1) What can data science do?

or

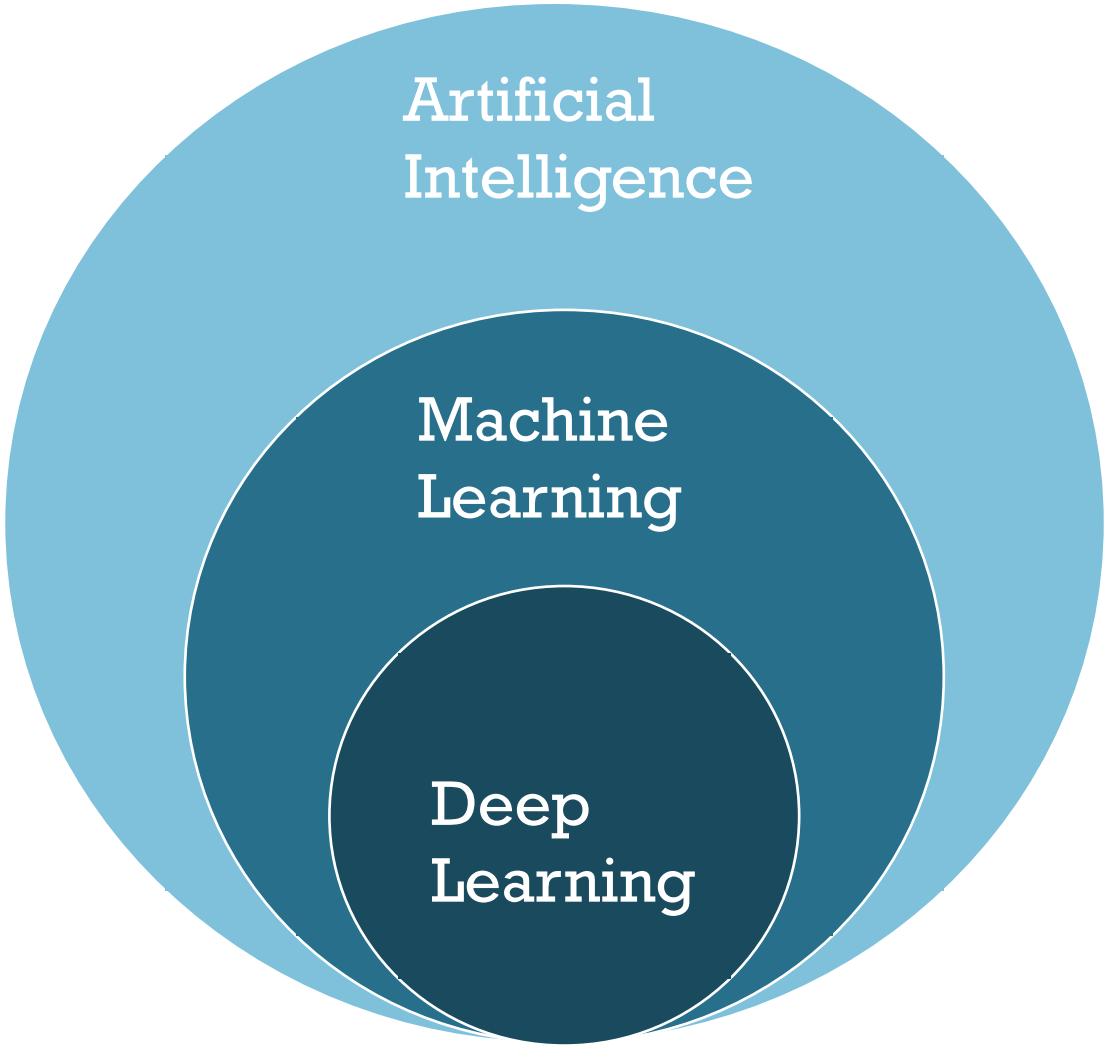
why is data science the new hotness


Artificial
Intelligence

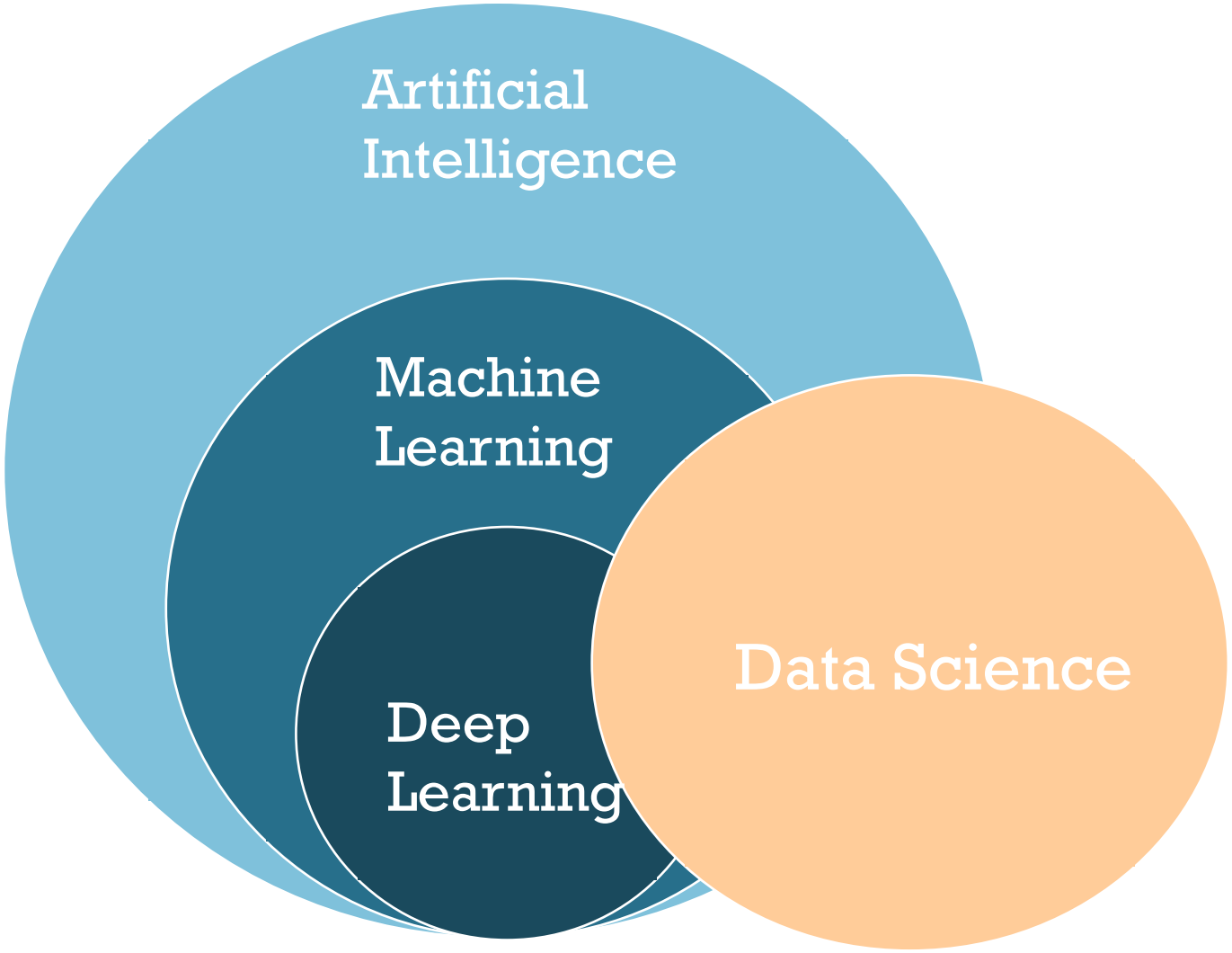
enabling machines to approximate and imitate human intelligence



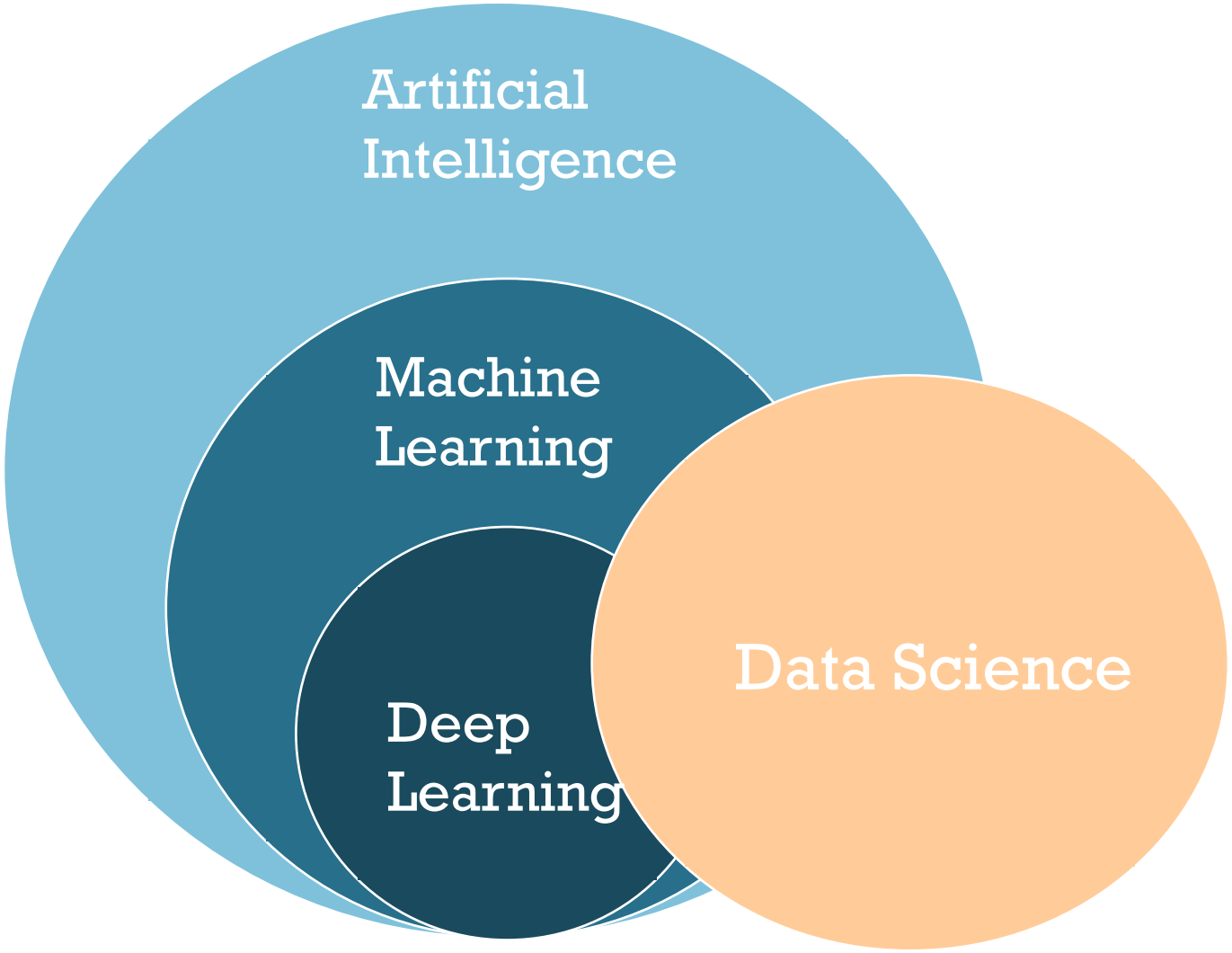
enabling machines to approximate and imitate human intelligence



enabling machines to approximate and imitate human intelligence



This is a bad way to understand data science.



This is a better way to understand data science.

This is a better way to understand data science.

This is a better way to understand data science.

Data science gives us a set of tools to solve problems.

This is a better way to understand data science.

Data science gives us a set of tools to solve problems.

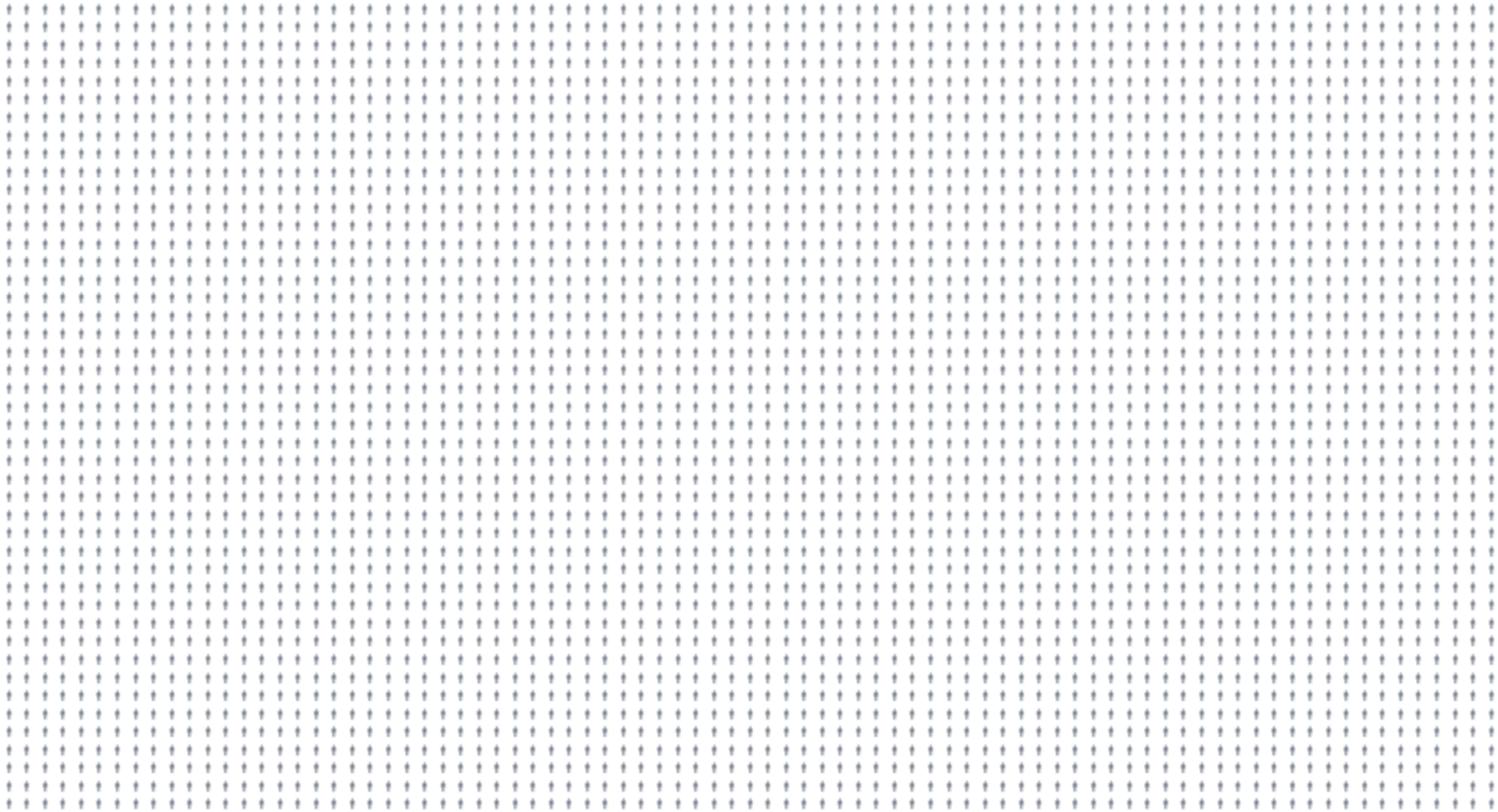
Let's look at a problem.

Suppose you are a business with subscribers.

Every subscriber you lose costs you money (ex, $1000).

You have historical data on customers that you have retained and customers that you have lost.

historical customers



historical customers

gender

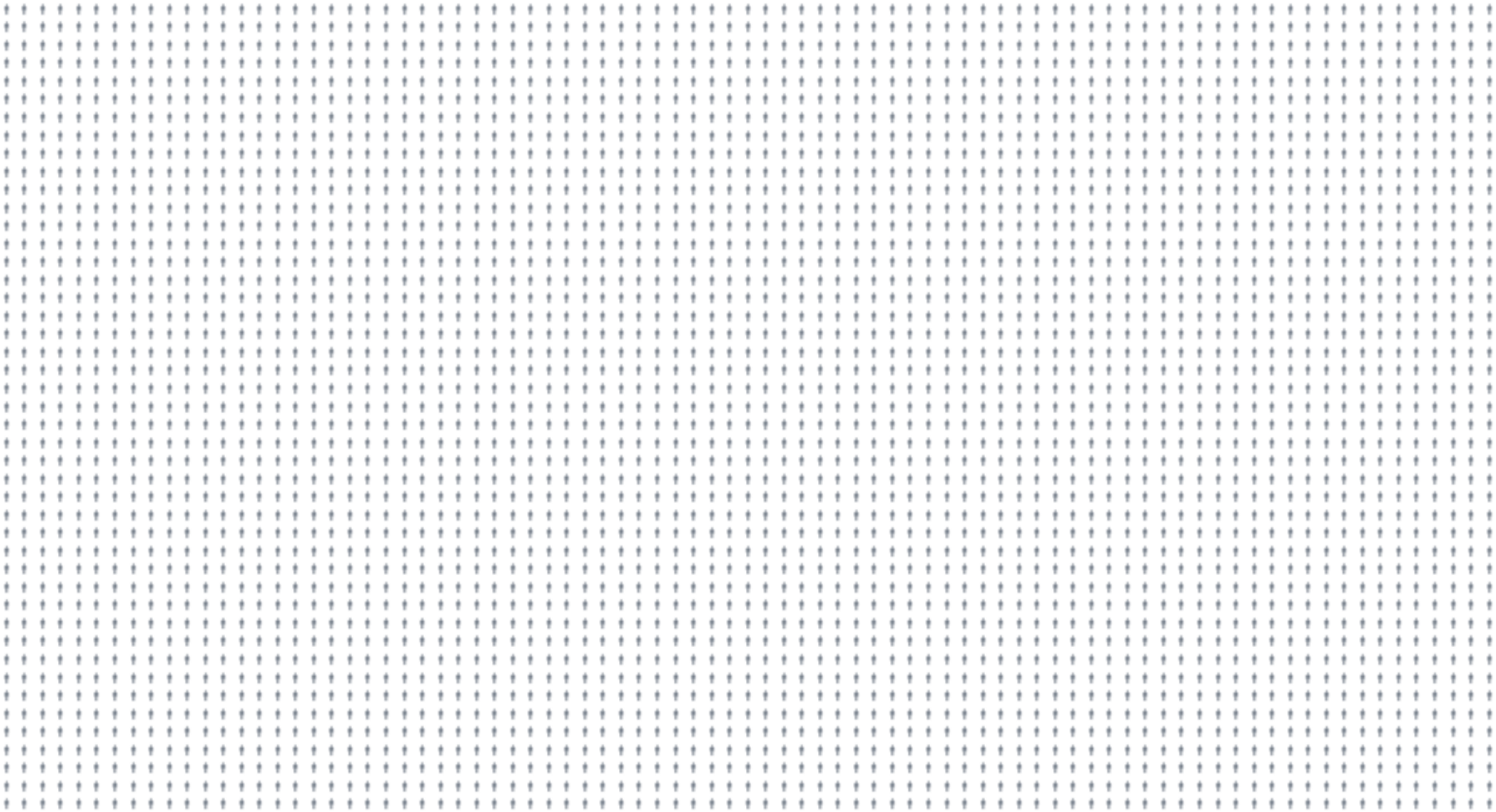
senior citizen

contract

payment method

monthly charges

total charges



historical customers

gender

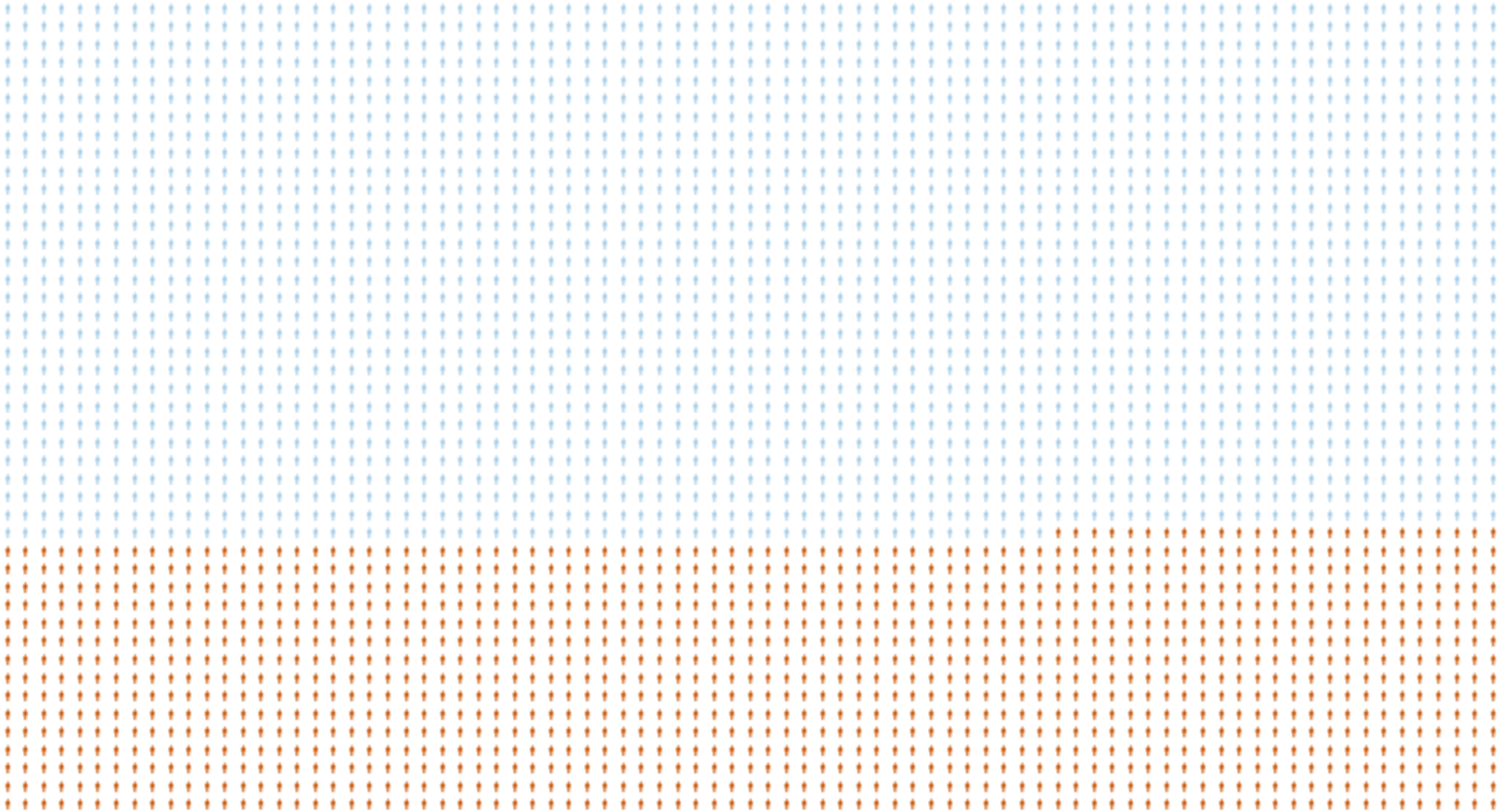
senior citizen

contract

payment method

monthly charges

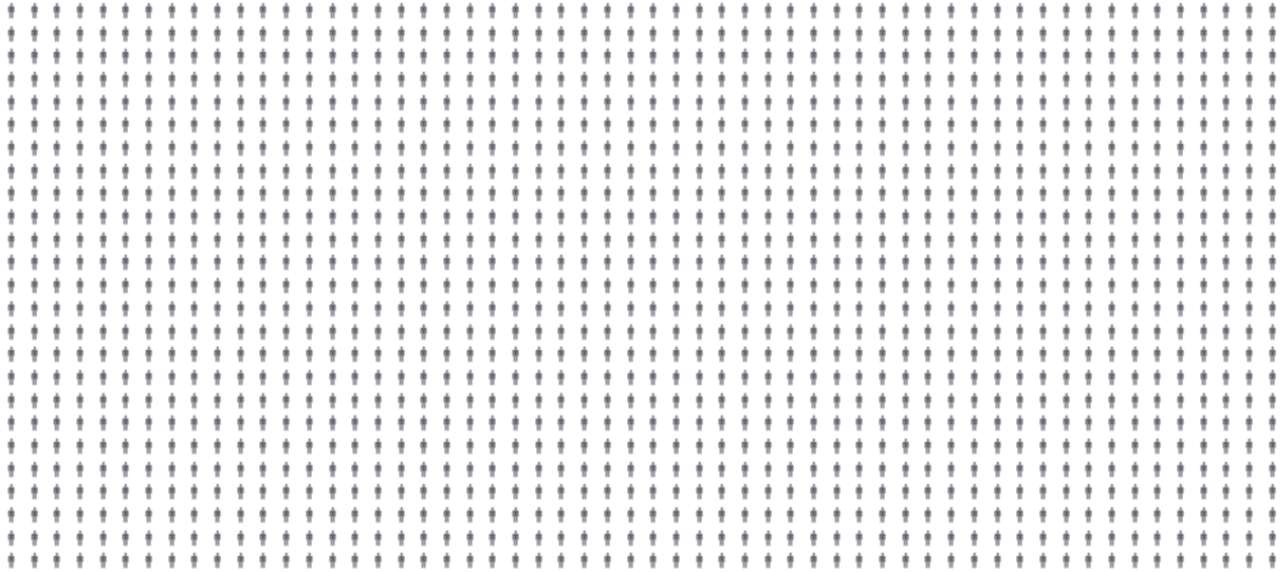
total charges



retained

lost

current customers



current customers

gender

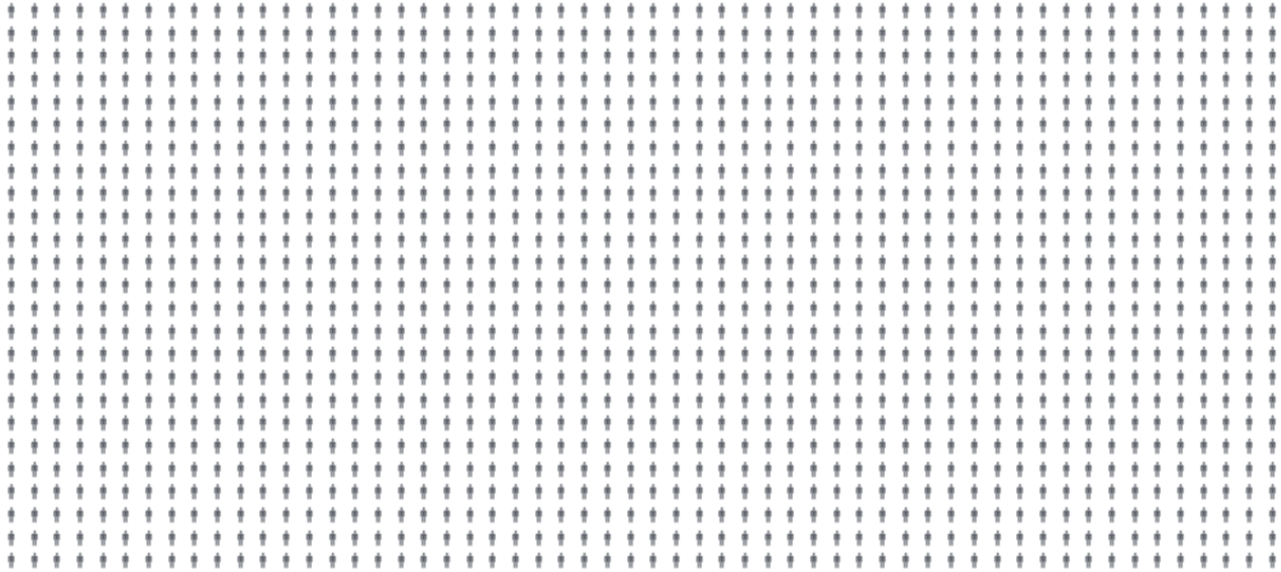
senior citizen

contract

payment method

monthly charges

total charges



current customers

gender

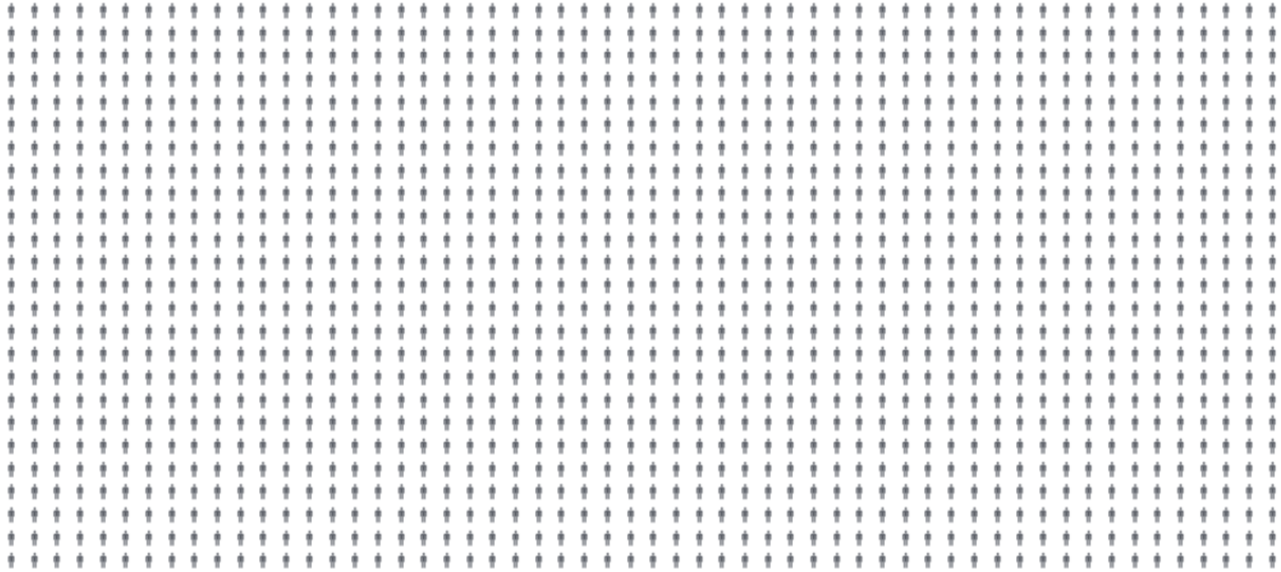
senior citizen

contract

payment method

monthly charges

total charges



don't know

current customers

gender

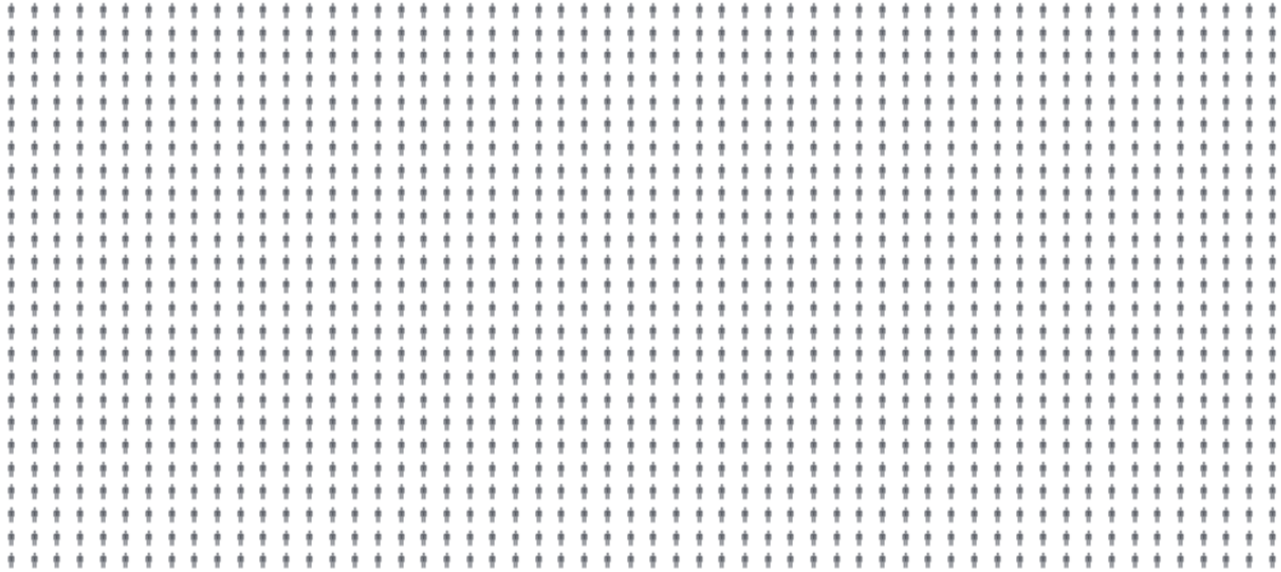
senior citizen

contract

payment method

monthly charges

total charges



don't know

Every customer who leaves will cost you money. To keep customers, you can offer a discount (ex, $300).

current customers

gender

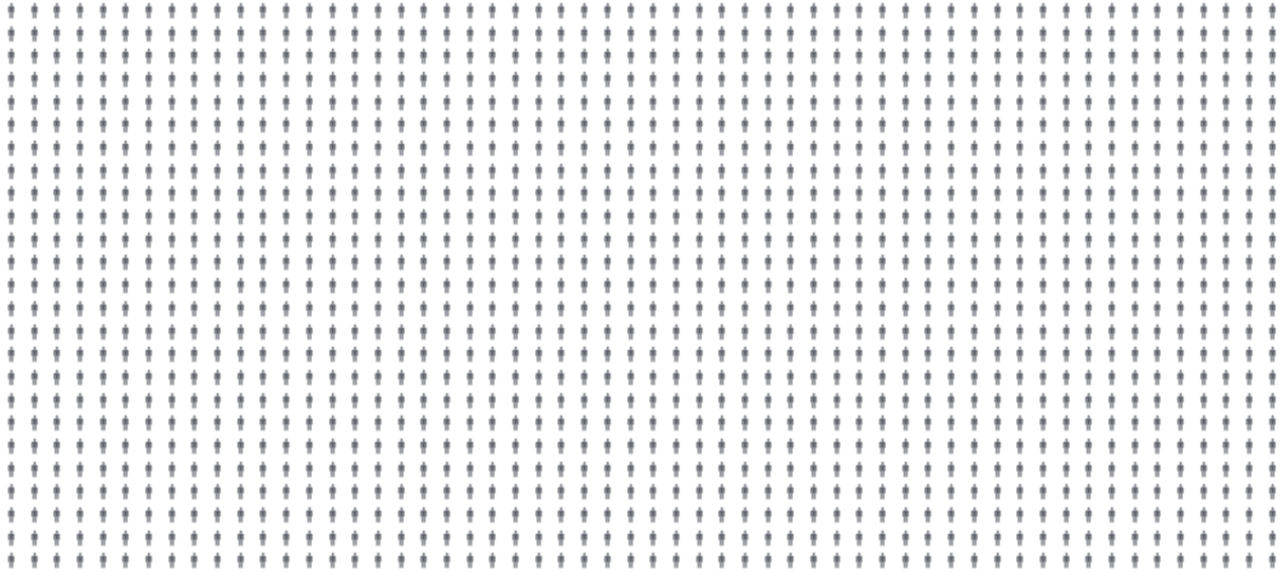
senior citizen

contract

payment method

monthly charges

total charges



don't know

Every customer who leaves will cost you money. To keep customers, you can offer a discount (ex, $300).

Who should get the discount?

Who should get the discount?

We need to know which customers are likely to leave.

Who should get the discount?

This is a problem a data scientist can solve.

Who should get the discount?

Inputs → **Output**

Gender
Senior Citizen
Contract
Payment Method
Total Charges
…

Customer Churn

This is a problem a data scientist can solve.

Who should get the discount?

Inputs → **Output**

Gender
Senior Citizen
Contract
Payment Method
Total Charges
…

Customer Churn

We could take a guess at the relationship between these inputs and the output.

Inputs → **Output**

Gender
Senior Citizen
Contract
Payment Method
Total Charges
…

Customer Churn

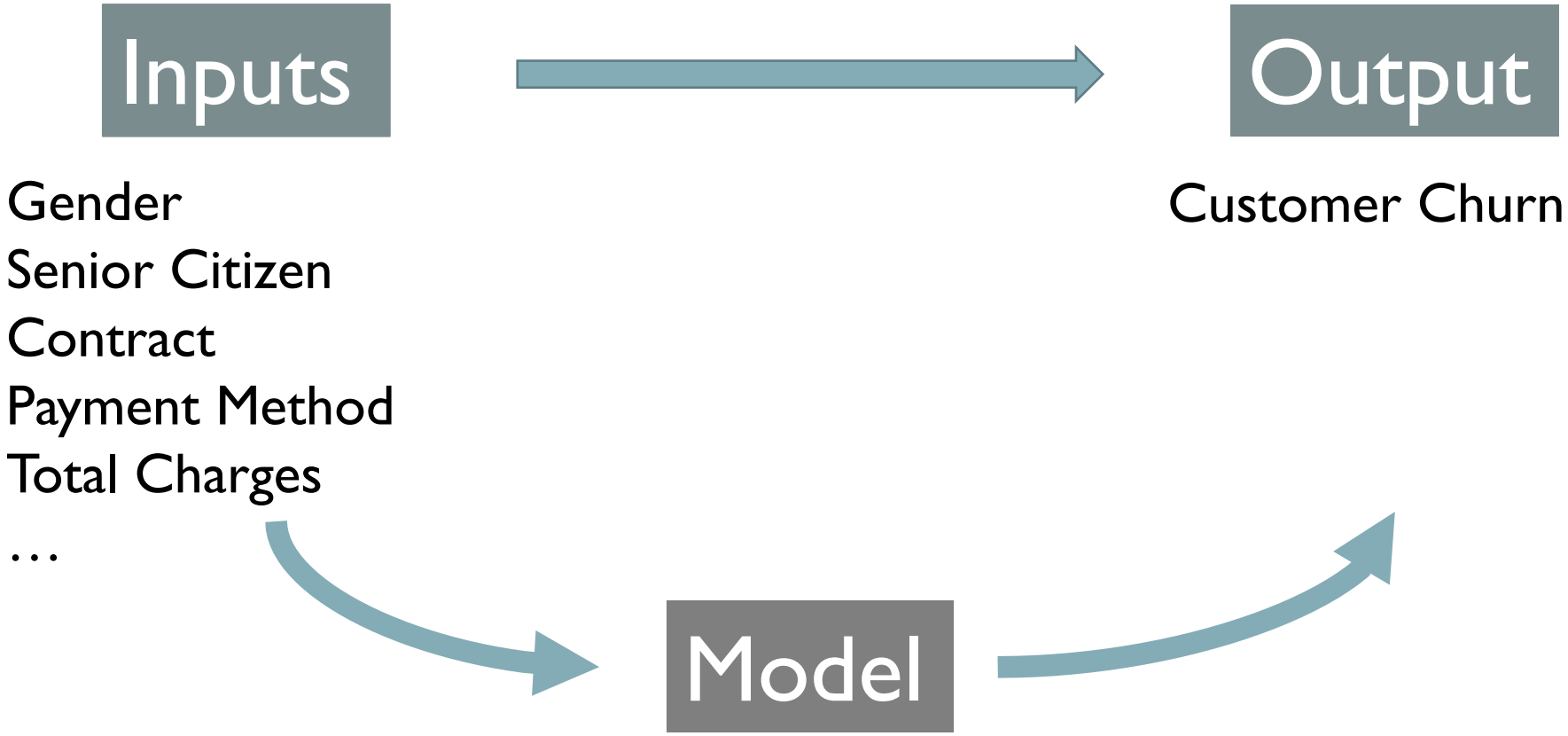
But our guess could be wrong.

Inputs → **Output**

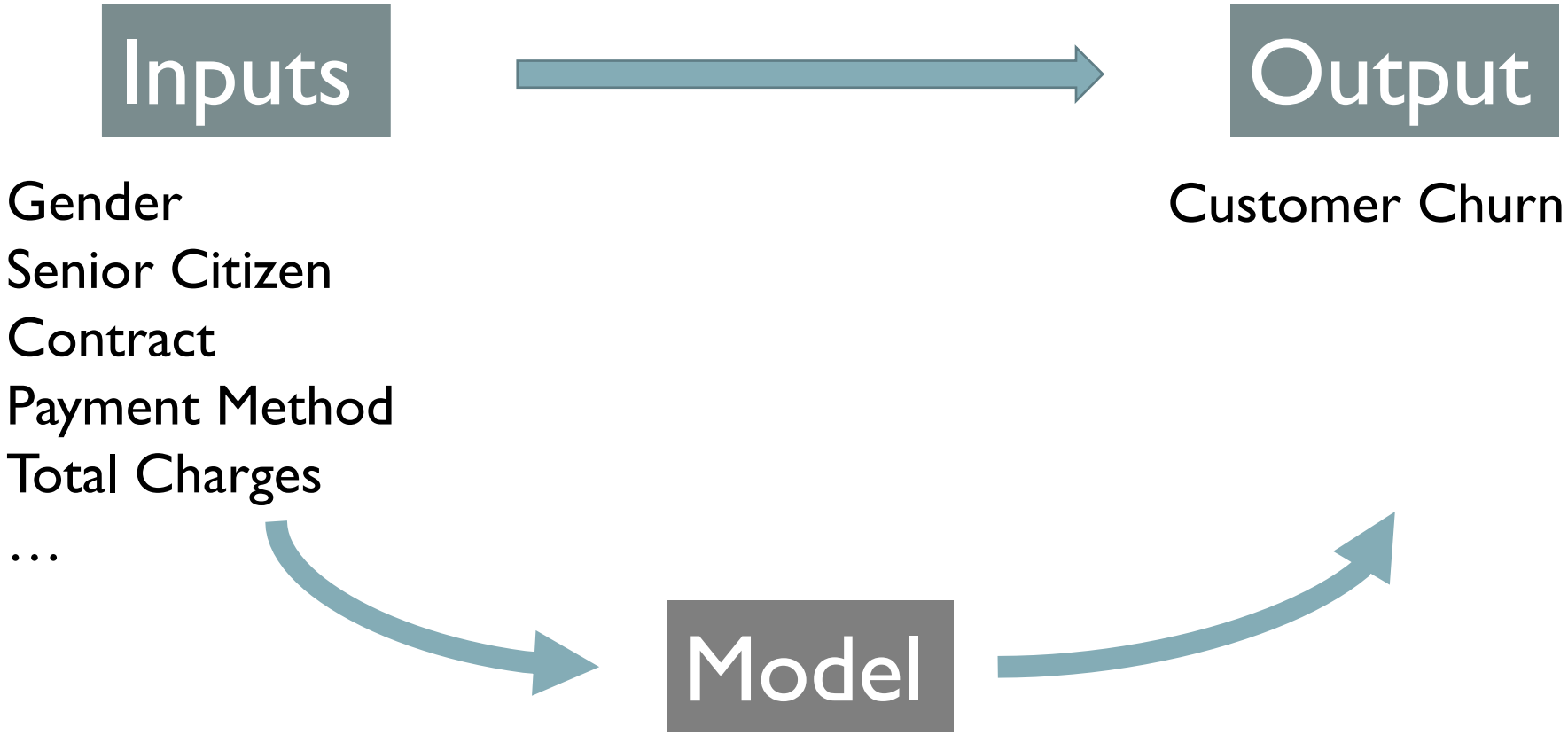
Gender
Senior Citizen
Contract
Payment Method
Total Charges
…

Customer Churn

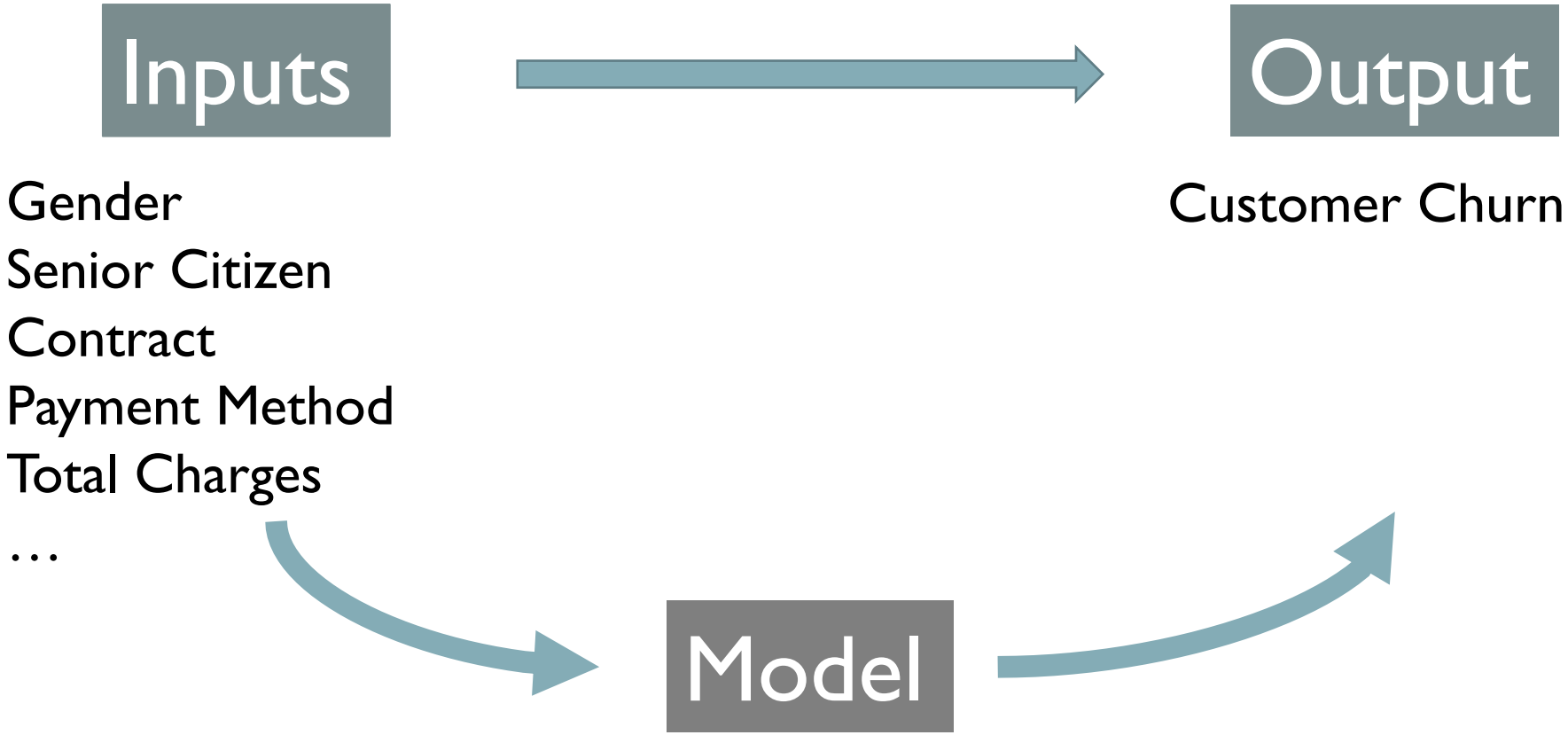
But our guess could be wrong.
And we don't have time to keep guessing.



But our guess could be wrong.
And we don't have time to keep guessing.

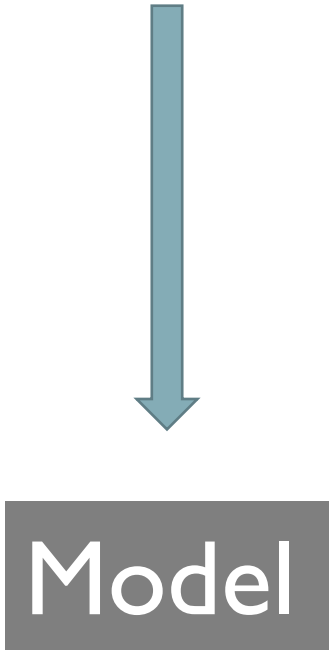


A data scientist can train a model that will be faster and more accurate than a human.



A data scientist can use a **machine** to **learn** the relationship between the inputs and the output.*

Most **data science** problems reduce to trying to figure out this box.

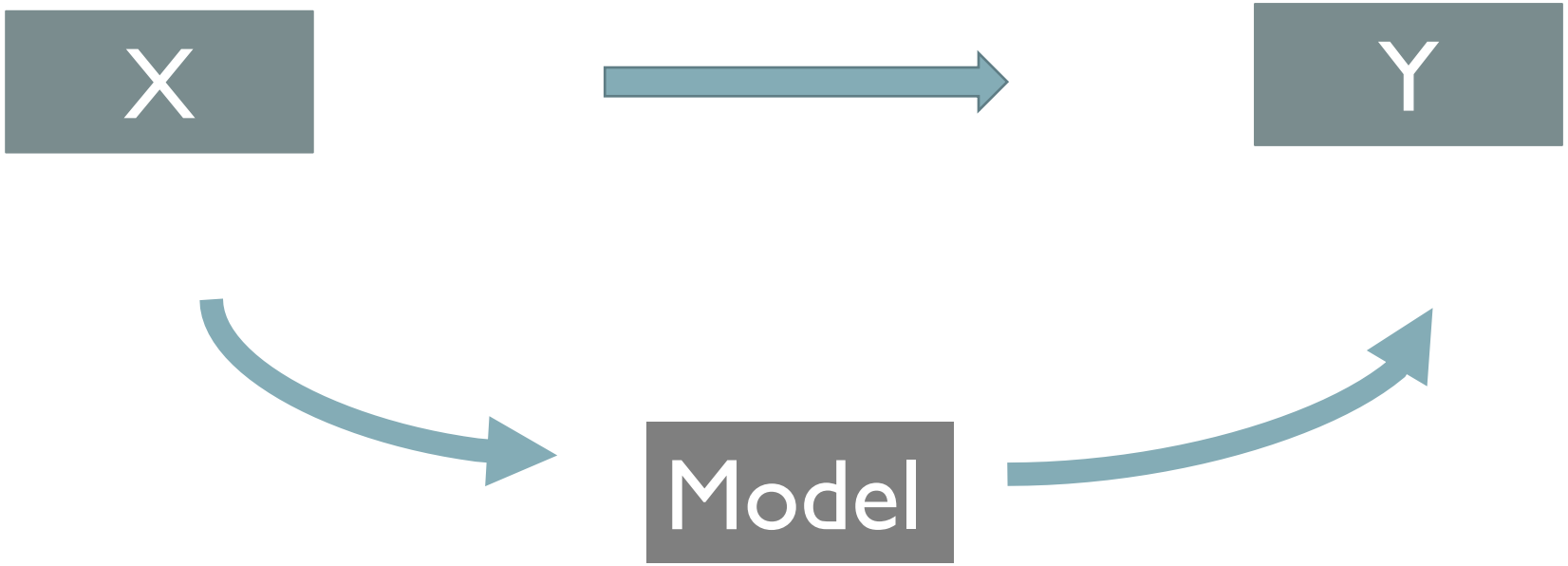


Most **data science** problems reduce to trying to figure out this box.

Model

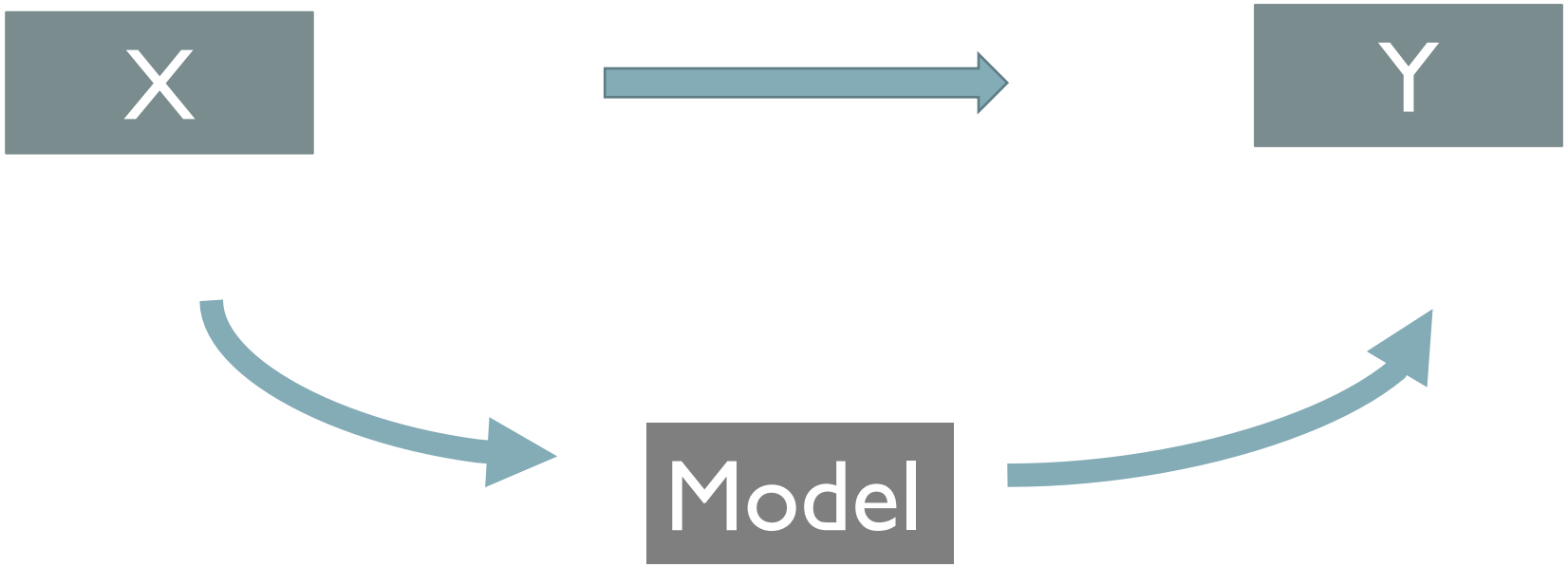
the data-generating process

Most **data science** problems reduce to trying to figure out this box.



the data-generating process

How do we figure out the data generating process?



the data-generating process

SOMM

© 2012 FORGOTTEN MAN FILMS

FORGOTTEN MAN FILMS PRESENTS

EXECUTIVE PRODUCER WILLIAM FOWLER PRODUCED BY CHRISTINA WISE JASON WISE AND JACKSON MYERS

STARRING IAN CAUBLE, BRIAN MCCLINTIC, DLYNN PROCTOR AND DUSTIN WILSON

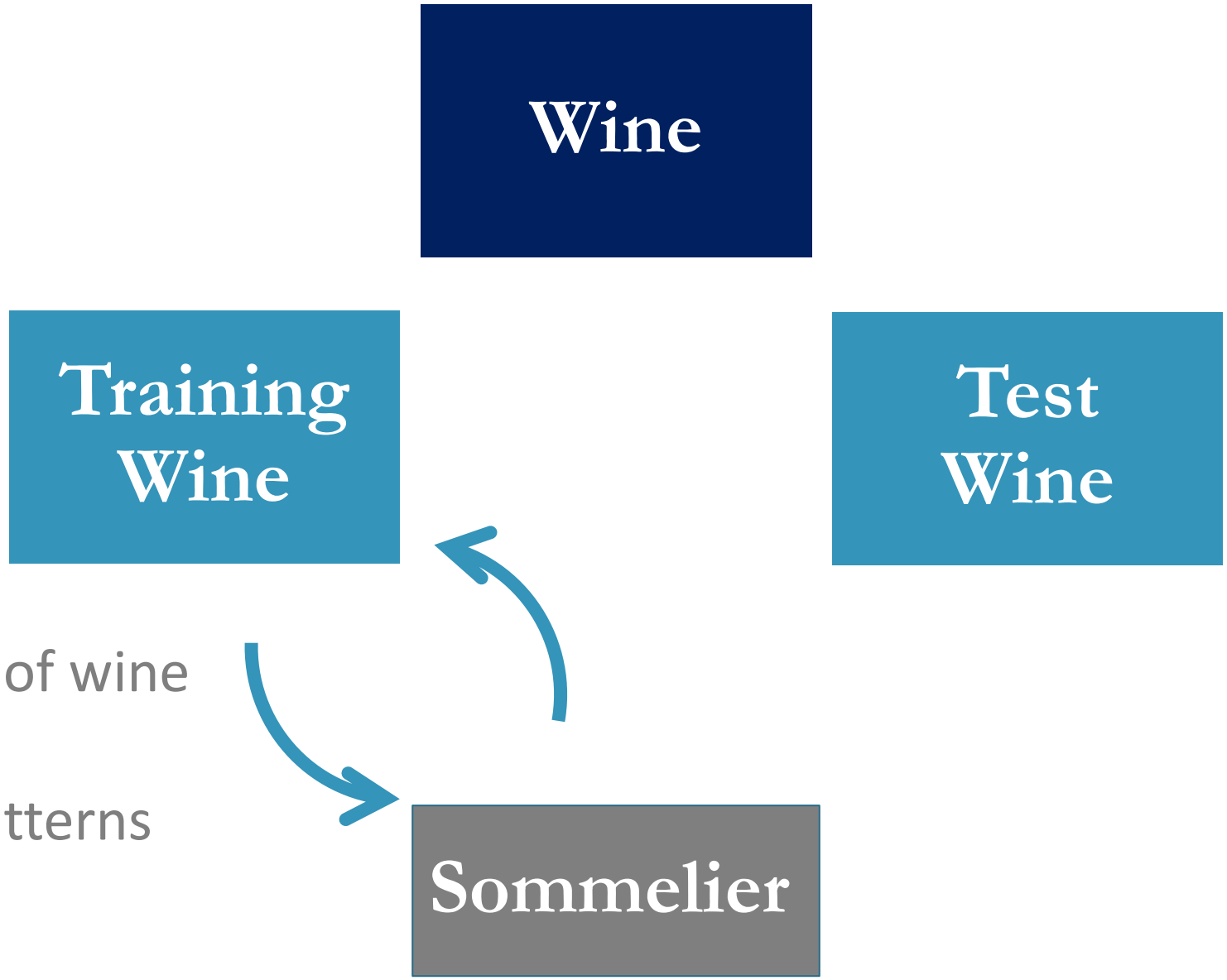
EDITED BY JASON WISE JACKSON MYERS AND BRYAN CARR GRAPHICS BY FRANCIS KMIECIK

CINEMATOGRAPHER JACKSON MYERS DIRECTED BY JASON WISE

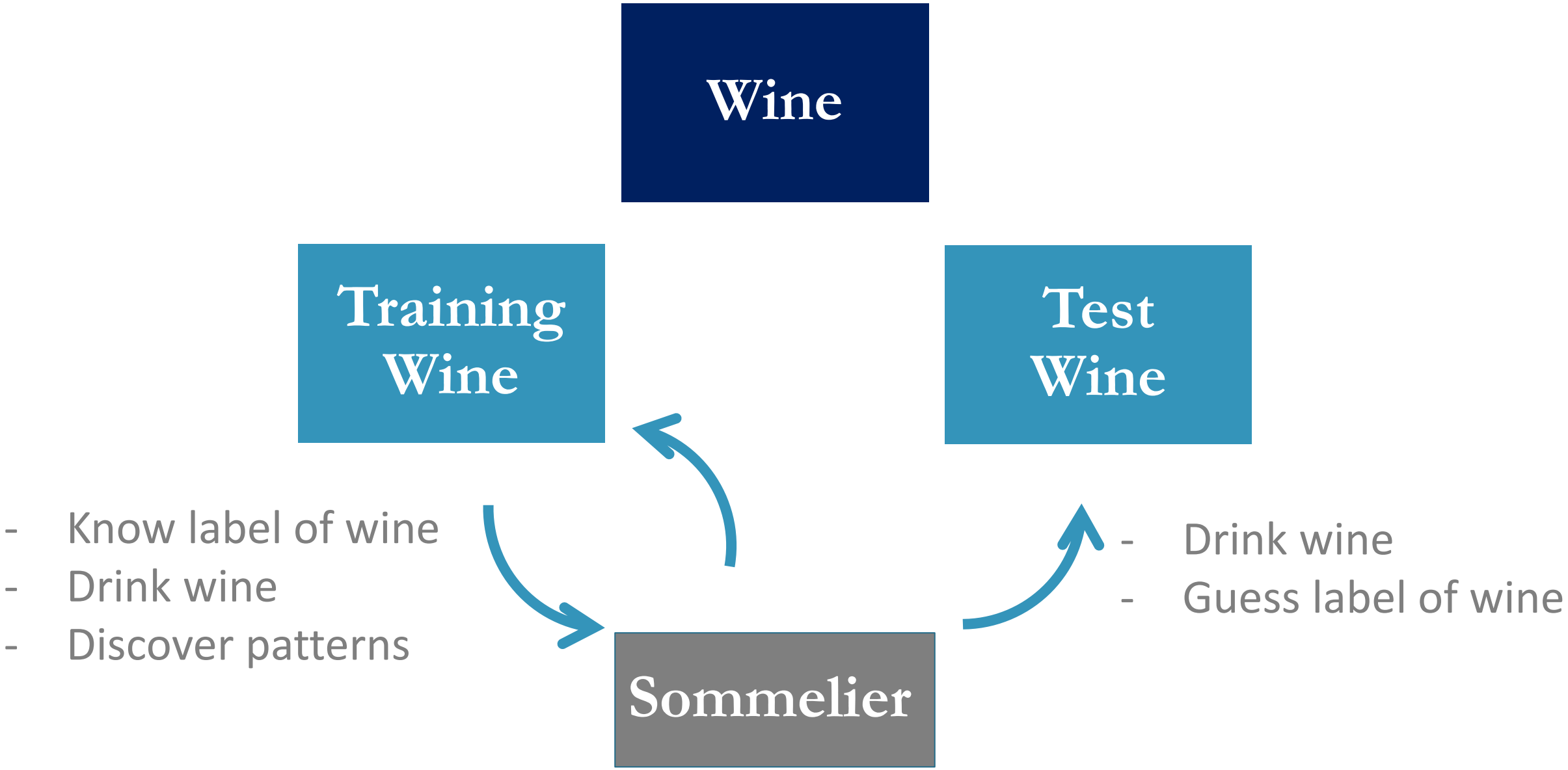
SUBSCRIBE

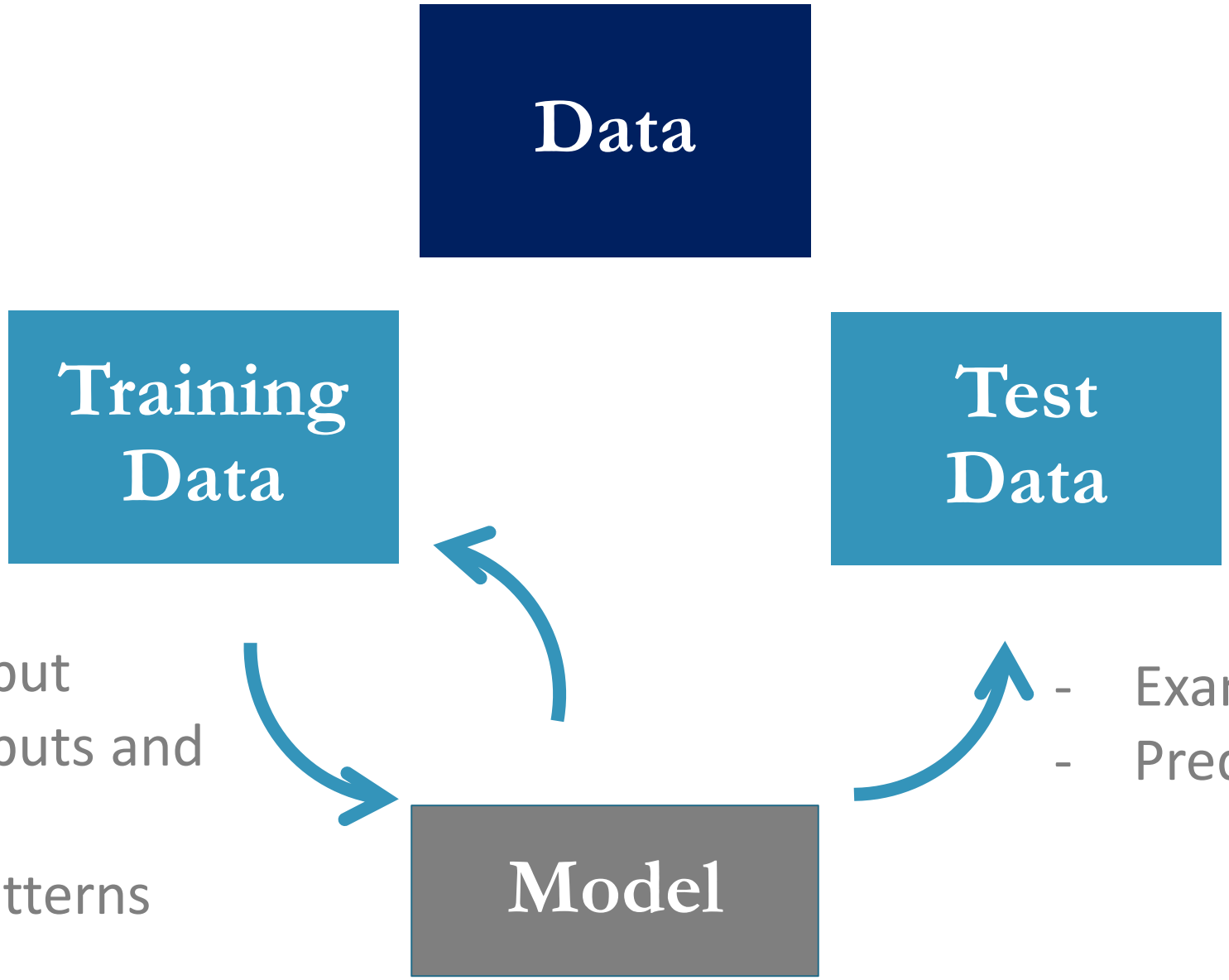
Wine

Wine
Training Wine
Test Wine



- Know label of wine
- Drink wine
- Discover patterns


Wine
Training Wine
Test Wine
- Know label of wine
- Drink wine
- Discover patterns
Sommelier
- Drink wine
- Guess label of wine



- Known output
- Examine inputs and output
- Discover patterns

- Examine inputs
- Predict output

Back to our customer churn example

historical customers
aka
training data

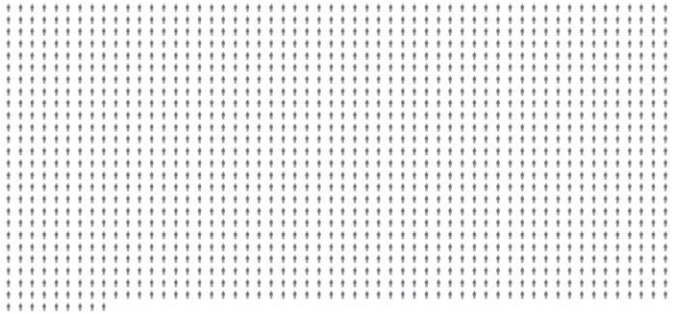
gender
senior citizen
contract type
monthly charges
total charges
payment method

retained

lost

current customers
aka
test data

gender
senior citizen
contract type
monthly charges
total charges
payment method



don't know

historical customers
aka
training data

gender
senior citizen
contract type
monthly charges
total charges
payment method

retained

lost

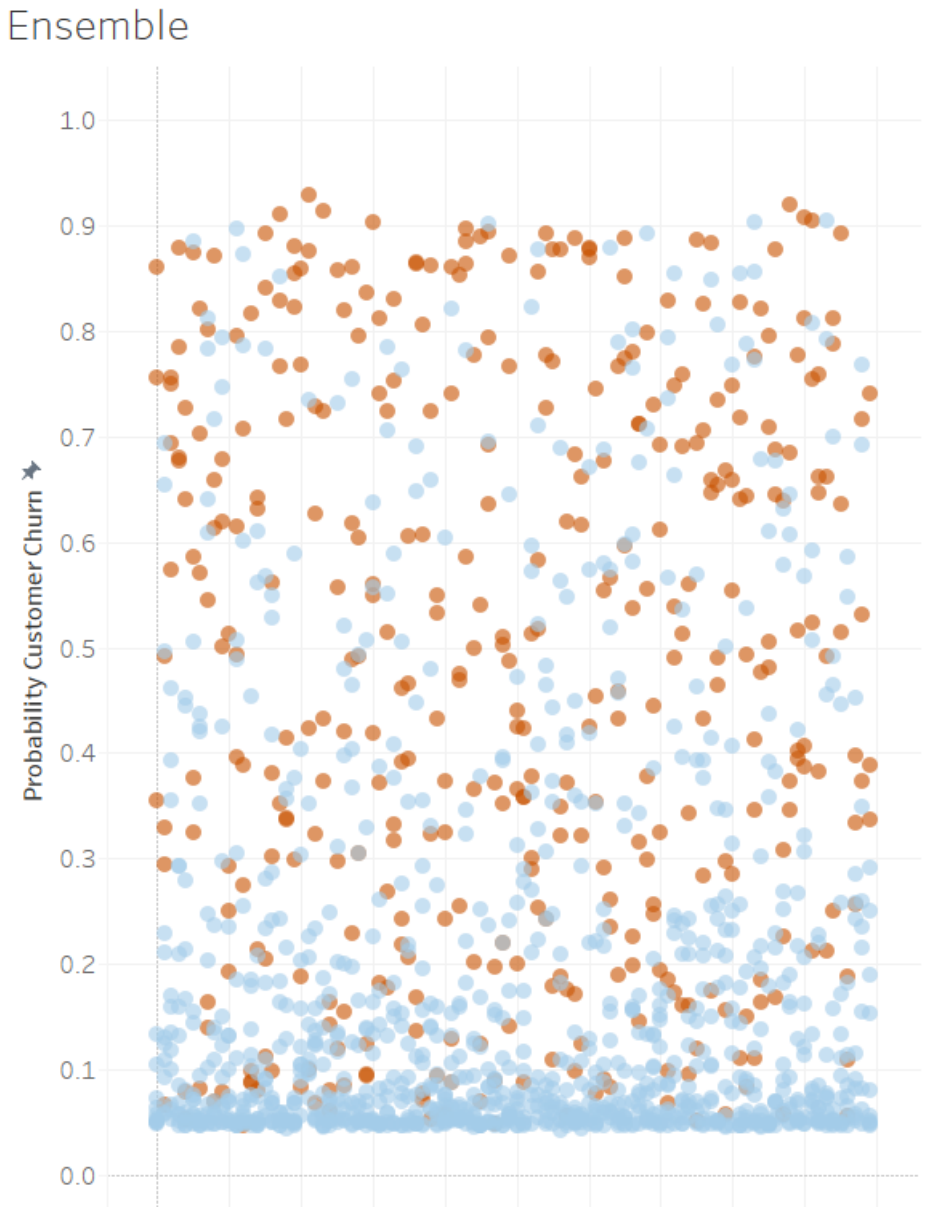
1) Train models to predict probability that a customer will churn

2) Identify optimal probability for giving discounts

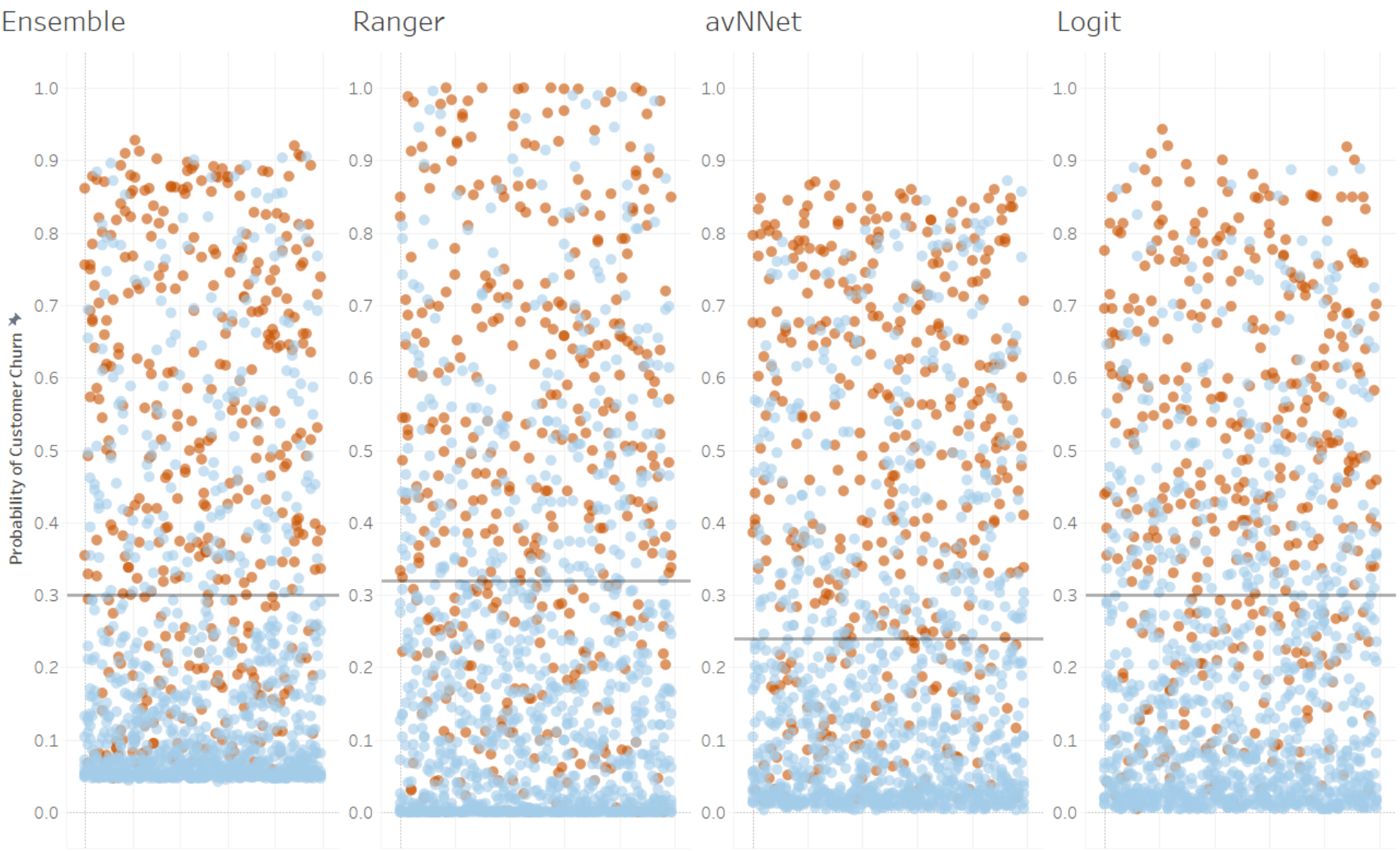
3) Predict test set and see how well we did at saving $$$

1) Train models to predict probability that a customer will churn

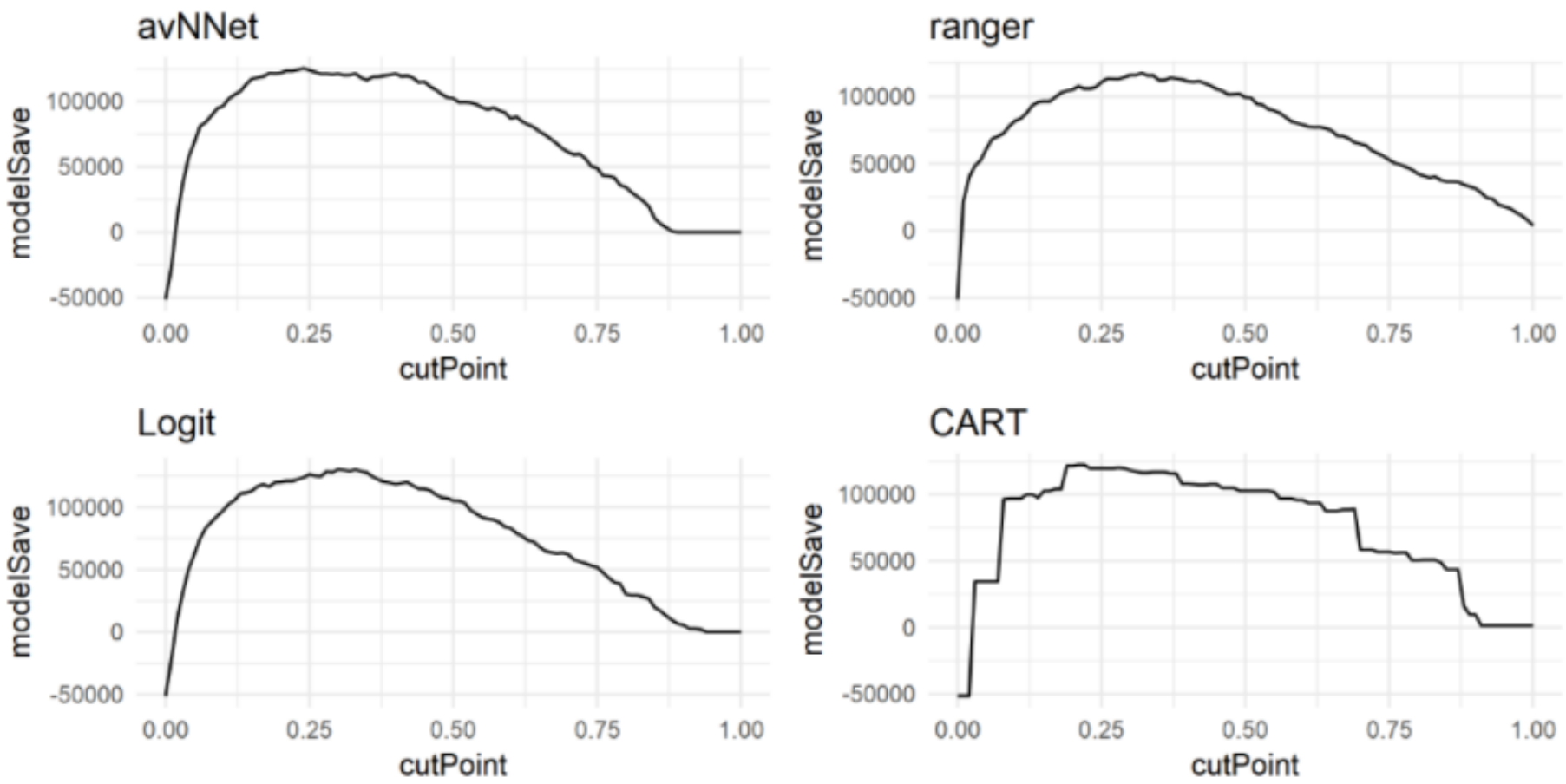
```r
# MARS
set.seed(1999)
baseline_earth<-train(Churn ~ .,
                      data = dplyr::select(trainDat, -one_of(omitVars),-customerID),
                      method = "earth",
                      metric="Spec",
                      tuneLength=5,
                      trControl = ctrlParallel,
                      preProcess=c("center", "scale", "pca"))

# ranger
set.seed(1999)
baseline_ranger<-train(Churn ~ .,
                       data = dplyr::select(trainDat, -one_of(omitVars), -customerID),
                        method = "ranger",
                        importance="permutation",
                        tuneLength=5,
                        trControl = ctrlParallel)

# parRF
# register parallel backend
numCore <- detectCores() - 2
registerDoParallel(cores = numCore)
set.seed(1999)
baseline_parRF<-train(Churn ~ .,
                       data = dplyr::select(trainDat, -one_of(omitVars), -customerID),
                      method = "parRF",
                      metric="Spec",
                      tuneLength=5,
                      trControl = ctrlParallel)
registerDoSEQ()

# avNNet
set.seed(1999)
baseline_avNNet<-train(Churn ~ .,
                        data = dplyr::select(trainDat, -one_of(omitVars), -customerID),
                       method = "avNNet",
                        trControl = ctrlParallel,
                        preProcess=c("center", "scale", "pca"))
```



2) Identify optimal probability for giving discounts



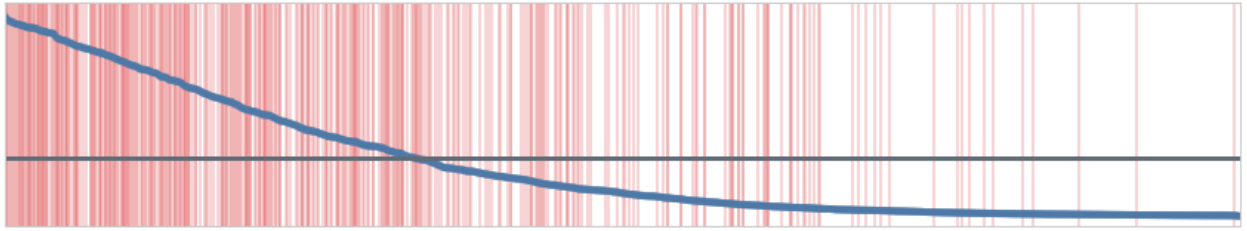
2) Identify optimal probability for giving discounts



3) Predict test set and see how well we did at saving $$$

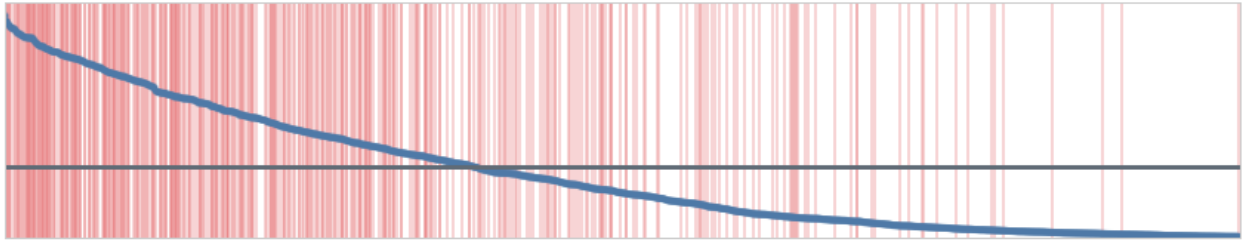
3) Predict test set and see how well we did at saving $$$

Ensemble



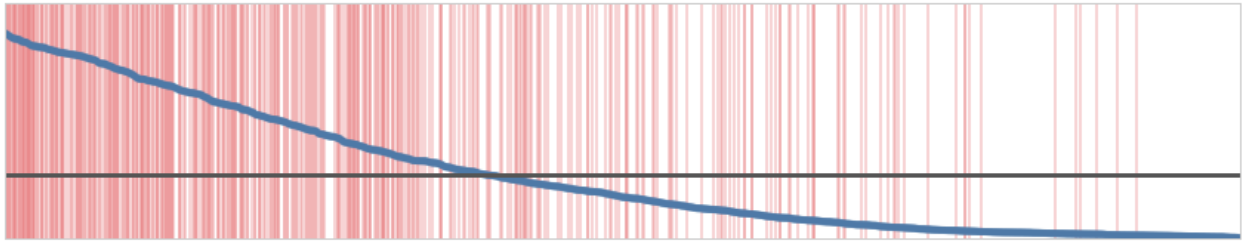
Savings	TP	FP	FN	TN
$159,600	269	198	123	819

Logit



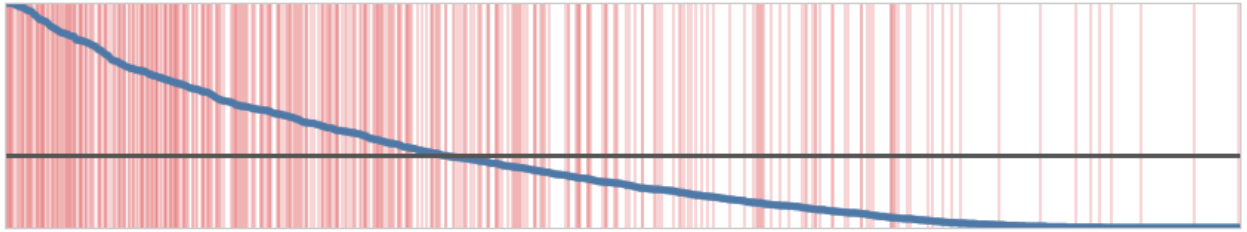
Savings	TP	FP	FN	TN
$158,200	288	247	104	770

avNNet



Savings	TP	FP	FN	TNs
$157,500	305	289	87	728

Ranger



Savings	TP	FP	FN	TNs
$145,200	266	239	126	778

Data science gives us the tools to learn about and recreate data generating processes.

Data science gives us the tools to learn about and recreate data generating processes.

The data may be different, but the methodology for other problems is the same.

UNIVERSITY OF WISCONSIN | RETENTION OF STUDENTS

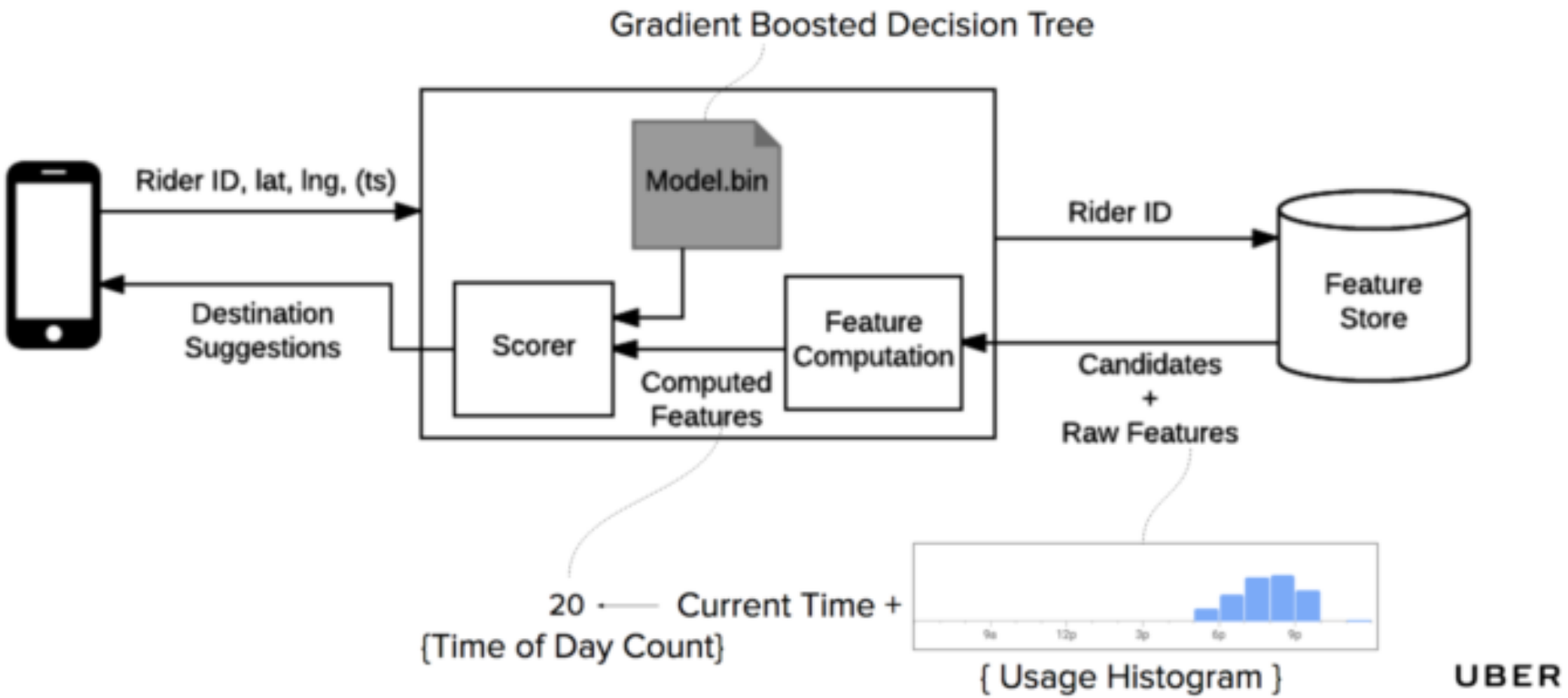
UW campuses using big data to predict students at risk of dropping out

KELLY MEYERHOFER kmeyerhofer@madison.com Mar 13, 2019

Students pass by UW-Madison's education building on Tuesday. About 95 percent of new freshmen enrolling full-time in the fall semester at the university returned the following fall, according to the most recent UW System data. At other UW campuses, the freshman retention rate was 77 percent. Officials say a big-data consulting firm will help faculty and staff at other UW campuses identify students on the cusp of dropping out.

AMBER ARNOLD, STATE JOURNAL

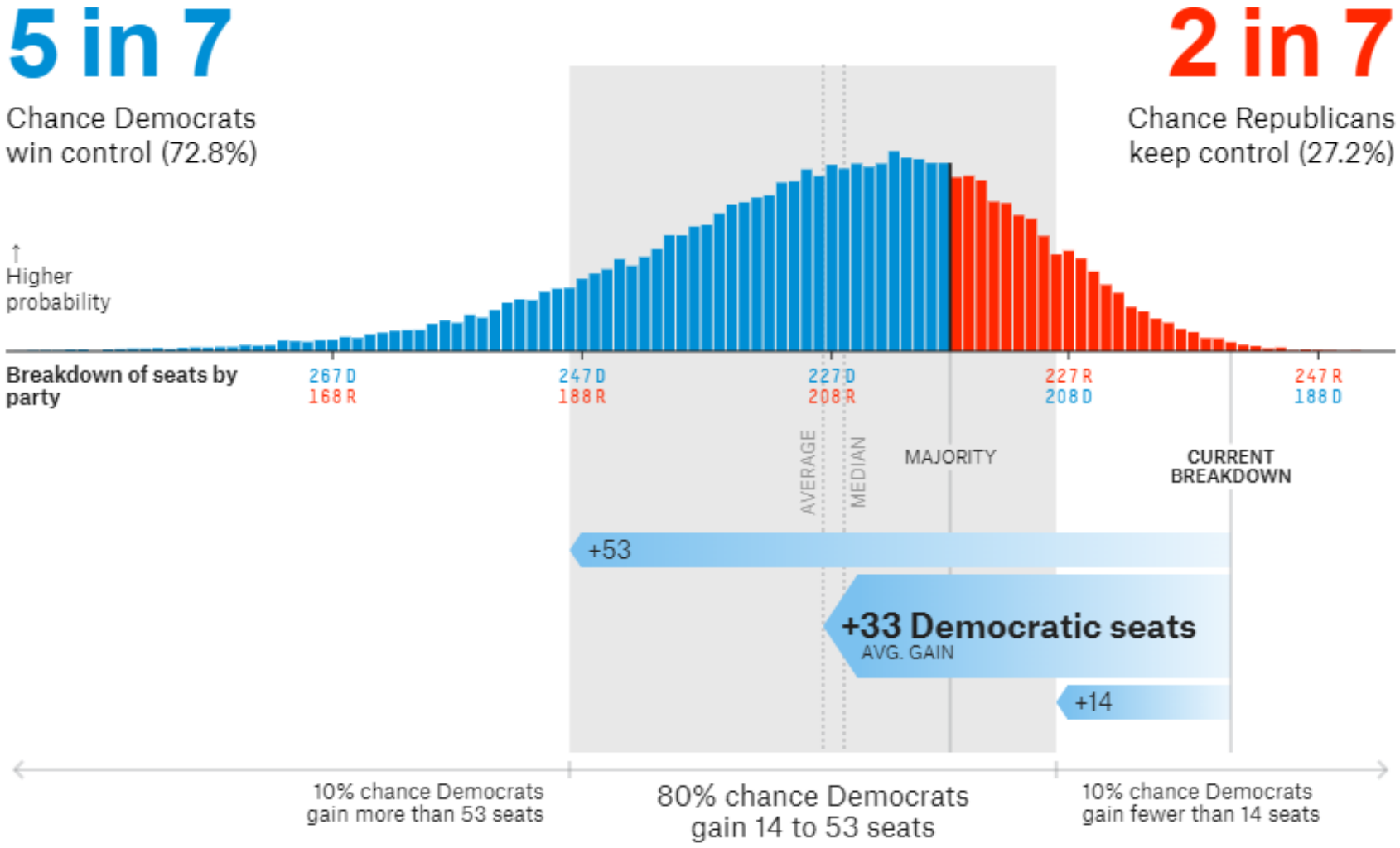
Destination Prediction Ranking



UBER

Forecasting the race for the House

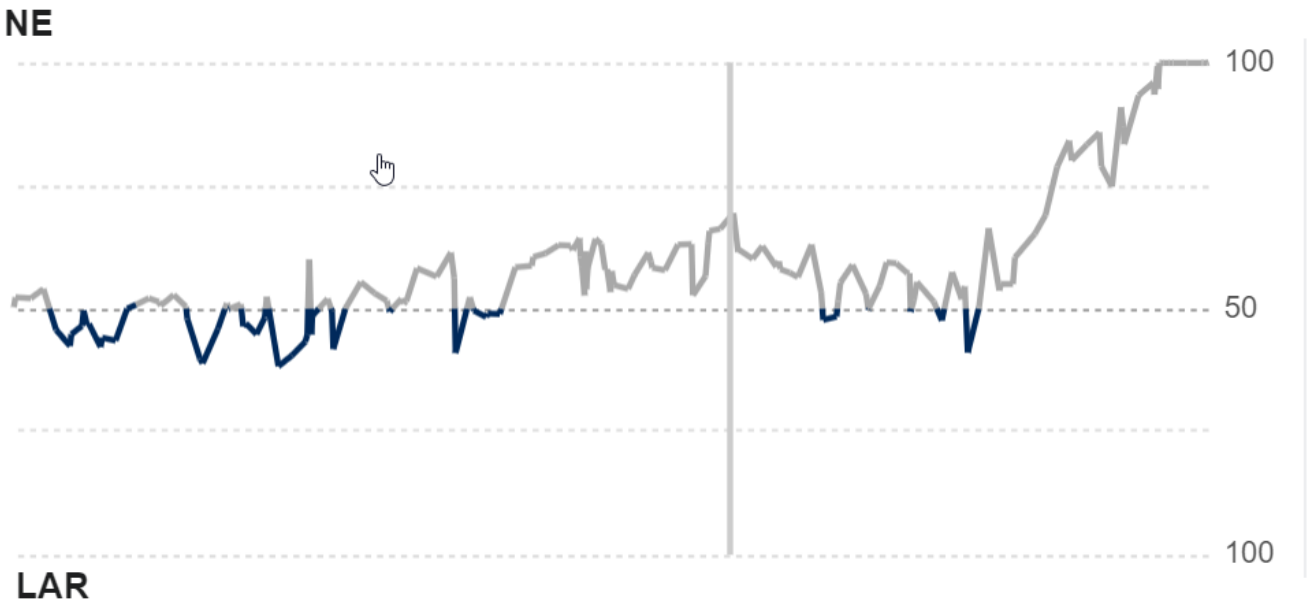
Updated Oct. 5, 2018, at 2:07 PM



ESPN

Win Probability

100%



NE 100%
NE 13 - LAR 3

(0:00 - 4th) END GAME

Playoff Pool Standings

Games Playoffs Ranks Teams Players Goalies About

Previous Day **Wednesday April 17, 2019** Next Day

Chance of Winning: 47.2%		7:00 PM ET		Chance of Winning: 52.8%
Chance of Winning: 44.9%		8:00 PM ET		Chance of Winning: 55.1%
Chance of Winning: 46.5%		10:00 PM ET		Chance of Winning: 53.5%

We've Suggested Tags for Your Photos

We've automatically grouped together similar pictures and suggested the names of friends who might appear in them. This lets you quickly label your photos and notify friends who are in this album.

Tag Your Friends

This will quickly label your photos and notify the friends you tag. Learn more

Who is this?

Who is this?

Who is this?

Who is this?

Who is this?

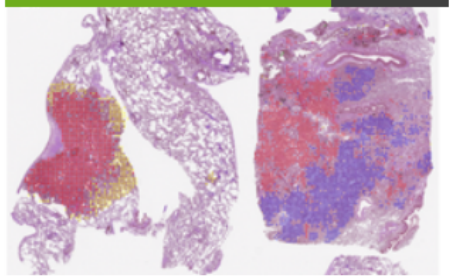
Who is this?

Francis Luu ×

Skip Tagging Friends

Save Tags

Comments 53 Shares



AI Helps Classify Lung Cancer at the Pathologist Level

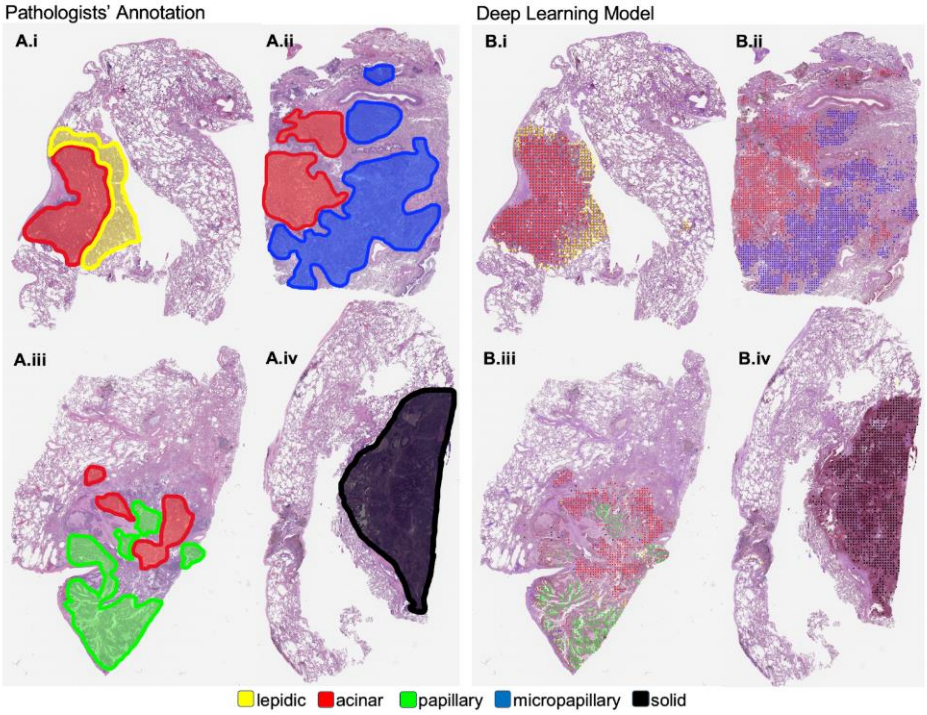
April 11, 2019

According to the American Cancer Society, more than 229,000 people will be diagnosed with lung cancer in the United States this year, with adenocarcinoma being the most common type. To help with diagnosis, researchers from Dartmouth's Norris Cotton Cancer Center and the Hassanpour Lab at Dartmouth University developed a deep learning-based system for automated classification of histologic subtypes on lung adenocarcinoma surgical resection slides on par with pathologists.

"Classification of histologic patterns in lung adenocarcinoma is critical for determining tumor grade and treatment for patients," the researchers stated in their paper. "Our model can potentially be used to aid pathologists in the classification of these histologic patterns and ultimately contribute to more accurate grading of lung adenocarcinoma," the team said.

Using NVIDIA Tesla GPUs with the cuDNN-accelerated PyTorch deep learning framework, the team trained a convolutional neural network on hundreds of whole-slide images from patients with a diagnosis of lung adenocarcinoma who underwent lobectomies at the Dartmouth-Hitchcock Medical Center (DHMC).

"Our model uses a convolutional neural network to identify regions of neoplastic cells, then aggregates those classifications to infer predominant and minor histologic patterns for any given whole-slide image," the researchers said.



AI in the Sky Aids Feet on the Ground Spotting Human Rights Violations

April 4, 2019 by ISHA SALIAN

How Predictive Modeling Has Revolutionized Insurance

July 2, 2012

Email This | Subscribe to Newsletter

Predictive modeling has forever changed the way insurance policies are priced. The tool allows insurers to design more sophisticated models that tap more detailed data sets to refine precisely how much each customer should be charged.

Casualty actuaries got an overview of how far the revolution will change insurance pricing at the session,"The Revolution and Evolution of Predictive Modeling," presented at the recent Casualty Actuarial Society (CAS) Spring Meeting in Phoenix.

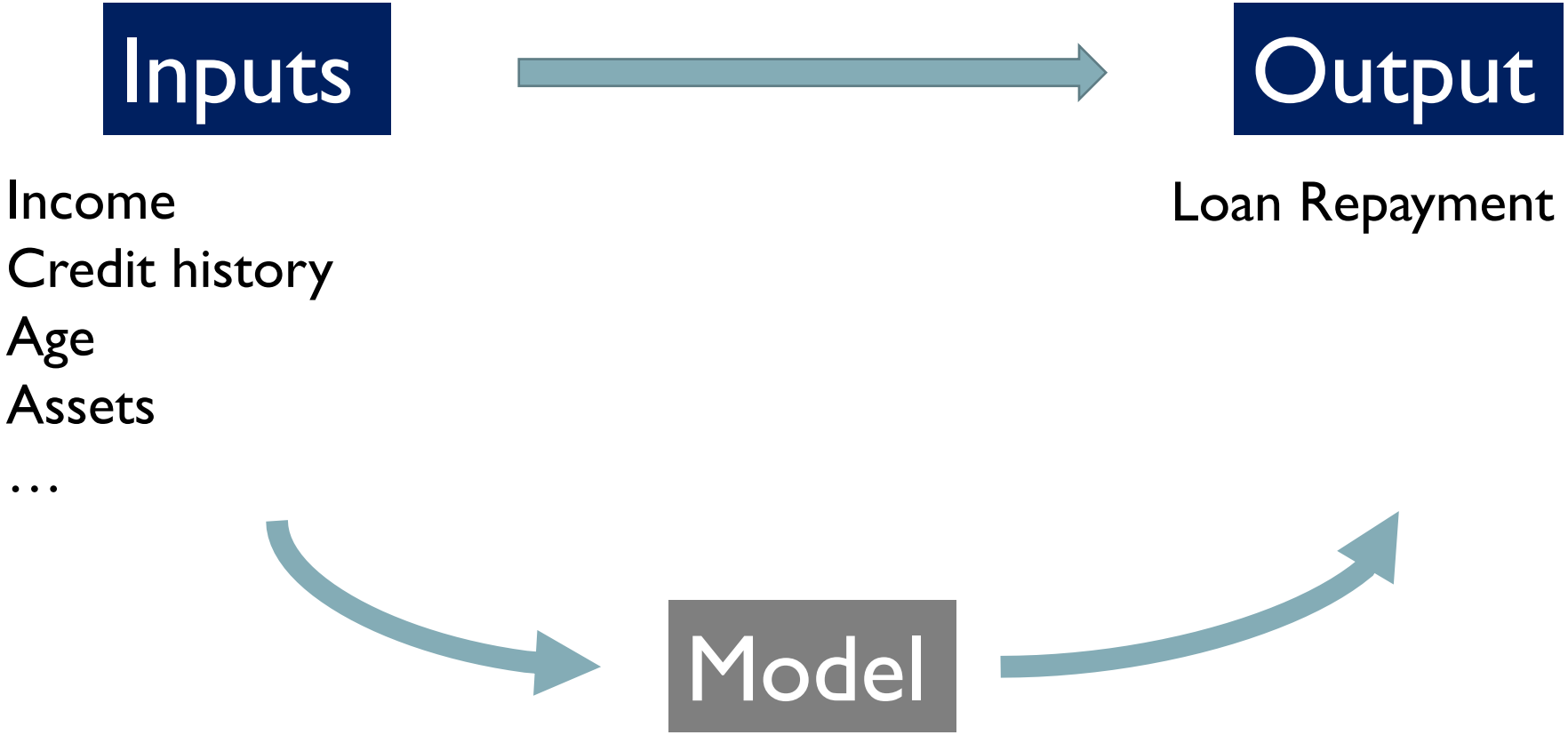
How Southwest Airlines Chooses Big Impact Analytics Projects

Southwest Airlines operates a fleet of 700 aircraft flying 4,000 flights a day. Here's how it uses analytics to manage all the complexity that comes with that.

Everyone wants their analytics investment to yield successful business results, but not everyone can show the kind of serious return on investment that captures the attention of C-suite executives. So when Southwest Airlines did a pilot program to look at the potential impact of investing in an analytics package, Doug Gray, director of analytical data services at the company, knew he needed to choose something with a big impact.

Of course, analytics work was nothing new to Southwest Airlines. The company had been working with analytics projects for about 20 years already. Such projects are absolutely necessary to manage the complexity of airline operations efficiently and cost effectively.

Supervised Learning: Mortgage Default



Supervised Learning: Spam Detection

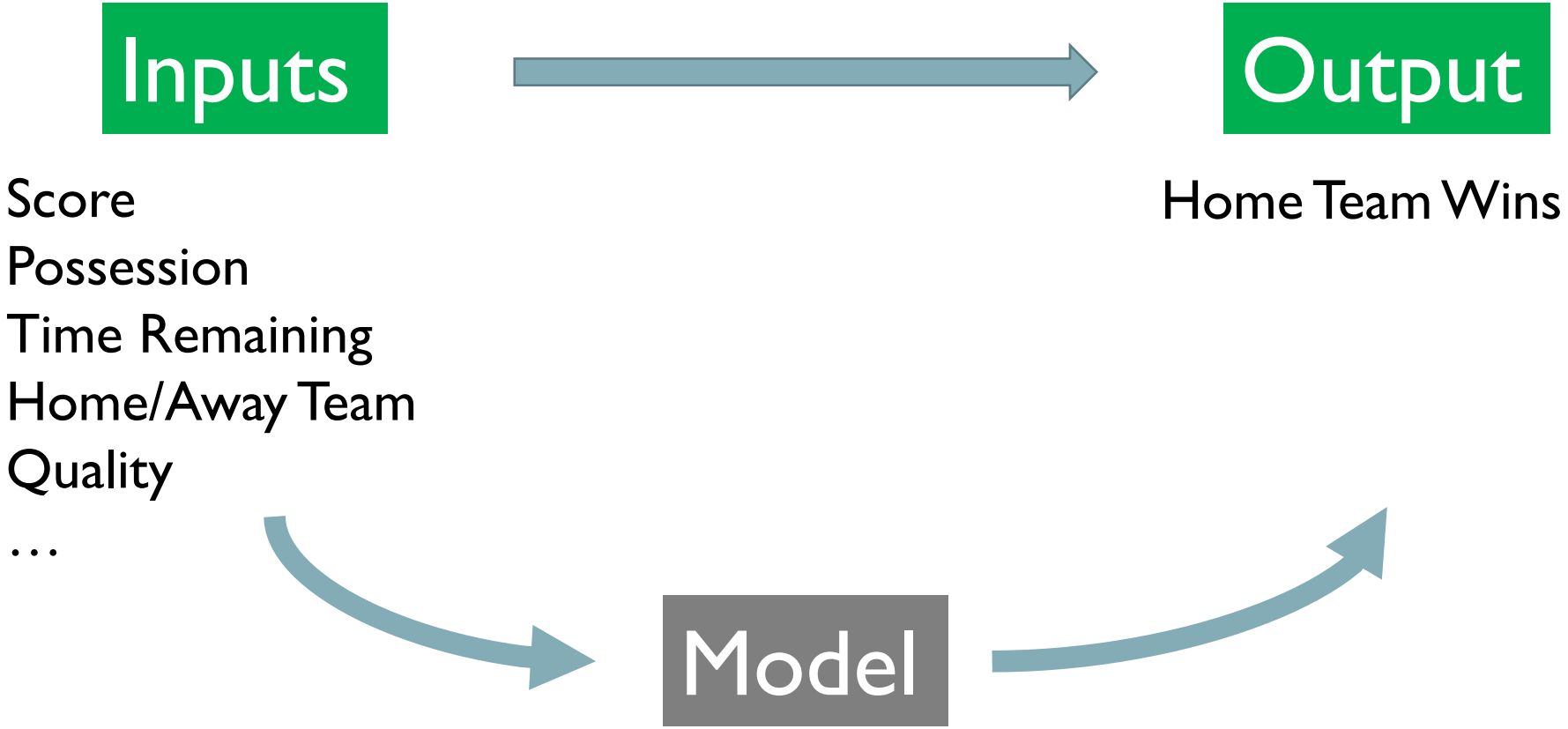
Inputs → **Output**

Words in email
Time received
Sender
Sender location
…

Email is Spam

Model

Supervised Learning: NFL In-Game Probs.



Supervised Learning: Netflix Recommender

Inputs

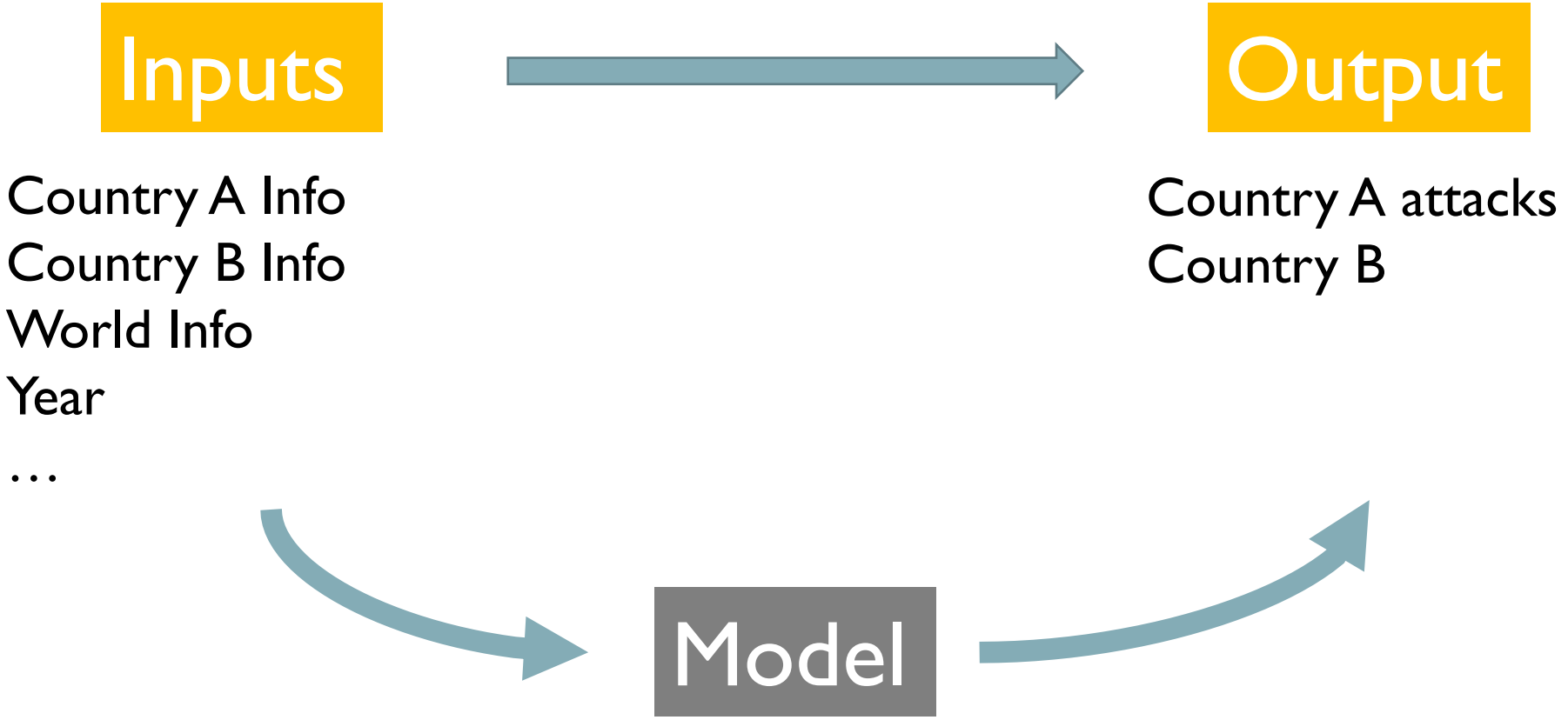
User Characteristics
Movie Genre
Actors
Year of Release
Time of Viewing
…

Output

User's Movie Rating

Model

Supervised Learning: Phil's Dissertation



Data science allows us to learn from data in order to recover the data generating process and use it to make better decisions.

Data science allows us to learn from data in order to recover the data generating process and use it to make better decisions.

But,

Data science allows us to learn from data in order to recover the data generating process and use it **to make better decisions**.

But,

2) What can data science not do?

2) What can data science not do?

or
why artificial intelligence requires human intelligence

Big data strategies disappoint with 85 percent failure rate

BY JAMES WALKER NOV 23, 2017 IN TECHNOLOGY LISTEN | PRINT

Big data projects aren't delivering the transformations companies had anticipated. 85% of all big data strategies are failing due to a combination of challenges, including the problems presented by legacy technologies and pre-existing corporate biases.

Big data strategies disappoint with 85 percent failure rate

BY JAMES WALKER NOV 23, 2017 IN TECHNOLOGY LISTEN | PRINT

Big data projects aren't delivering the transformations companies had anticipated. 85% of all big data strategies are failing due to a combination of challenges, including the problems presented by legacy technologies and pre-existing corporate biases.

Analytics
DRIVING BETTER BUSINESS DECISIONS
Brought to you by informs.

The Data Economy: Why do so many analytics projects fail?

July/August 2014

Key considerations for deep analytics on big data, learning and insights.

By (l-r) Haluk Demirkan and Bulent Dal

What is big data? Big data, which means many things to many people, is not a new technological fad. In addition to providing innovative solutions and operational insights to enduring challenges and opportunities, big data with deep analytics instigate new ways to transform processes, organizations, entire industries and even society. Pushing the boundaries of deep data analytics uncovers new insights and opportunities, and "big" depends on where you start and how you proceed.

Big data is not just "big." The exponentially growing volume of data is only one of many characteristics that are often associated with big data, such as variety, velocity, veracity and others (the six Vs; see box).

According to Gartner Research, the worldwide market for analytics will remain the top focus for CIOs through 2017 [1]. According to Gartner, more than half of all analytics projects fail because they aren't completed within budget or on schedule, or because they fail to deliver the features and benefits that are optimistically agreed on at their outset.

Big data strategies disappoint with 85 percent failure rate

BY JAMES WALKER NOV 23, 2017 IN TECHNOLOGY LISTEN | PRINT

Big data projects aren't delivering the transformations companies had anticipated. 85% of all big data strategies are failing due to a combination of challenges, including the problems presented by legacy technologies and pre-existing corporate biases.

We need to spend more time talking about data science failures

Passiona Cottee Follow

Dec 29, 2017 · 6 min read

In *Star Wars: The Last Jedi*, Yoda reminds us, "the greatest teacher, failure is". As a student of data science, sporting a 'data science' google alert, my news feed brims with the boasts of data science research applications. There's very

The Data Economy: Why do so many analytics projects fail?

July/August 2014

Key considerations for deep analytics on big data, learning and insights.

By (l-r) Haluk Demirkan and Bulent Dal

What is big data? Big data, which means many things to many people, is not a new technological fad. In addition to providing innovative solutions and operational insights to enduring challenges and opportunities, big data with deep analytics instigate new ways to transform processes, organizations, entire industries and even society. Pushing the boundaries of deep data analytics uncovers new insights and opportunities, and "big" depends on where you start and how you proceed.

Big data is not just "big." The exponentially growing volume of data is only one of many characteristics that are often associated with big data, such as variety, velocity, veracity and others (the six Vs; see box).

According to Gartner Research, the worldwide market for analytics will remain the top focus for CIOs through 2017 [1]. According to Gartner, more than half of all analytics projects fail because they aren't completed within budget or on schedule, or because they fail to deliver the features and benefits that are optimistically agreed on at their outset.

Big data strategies disappoint with 85 percent failure rate

BY JAMES WALKER NOV 23, 2017 IN TECHNOLOGY LISTEN | PRINT

Big data projects aren't delivering the transformations companies had anticipated. 85% of all big data strategies are failing due to a combination of challenges, including the problems presented by legacy technologies and pre-existing corporate biases.

We need to spend more time talking about data science failures

Passiona Cottee Follow

Dec 29, 2017 · 6 min read

In *Star Wars: The Last Jedi*, Yoda reminds us, "the greatest teacher, failure is". As a student of data science, sporting a 'data science' google alert, my news feed brims with the boasts of data science research applications. There's very

The Data Economy: Why do so many analytics projects fail?

July/August 2014

Key considerations for deep analytics on big data, learning and insights.

By (l-r) Haluk Demirkan and Bulent Dal

What is big data? Big data, which means many things to many people, is not a new technological fad. In addition to providing innovative solutions and operational insights to enduring challenges and opportunities, big data with deep analytics instigate new ways to transform processes, organizations, entire industries and even society. Pushing the boundaries of deep data analytics uncovers new insights and opportunities, and "big" depends on where you start and how you proceed.

Big data is not just "big." The exponentially growing volume of data is only one of many characteristics that are often associated with big data, such as variety, velocity, veracity and others (the six Vs; see box).

According to Gartner Research, the worldwide market for analytics will remain the top focus for CIOs through 2017 [1]. According to Gartner, more than half of all analytics projects fail because they aren't completed within budget or on schedule, or because they fail to deliver the features and benefits that are optimistically agreed on at their outset.

MANAGING PEOPLE

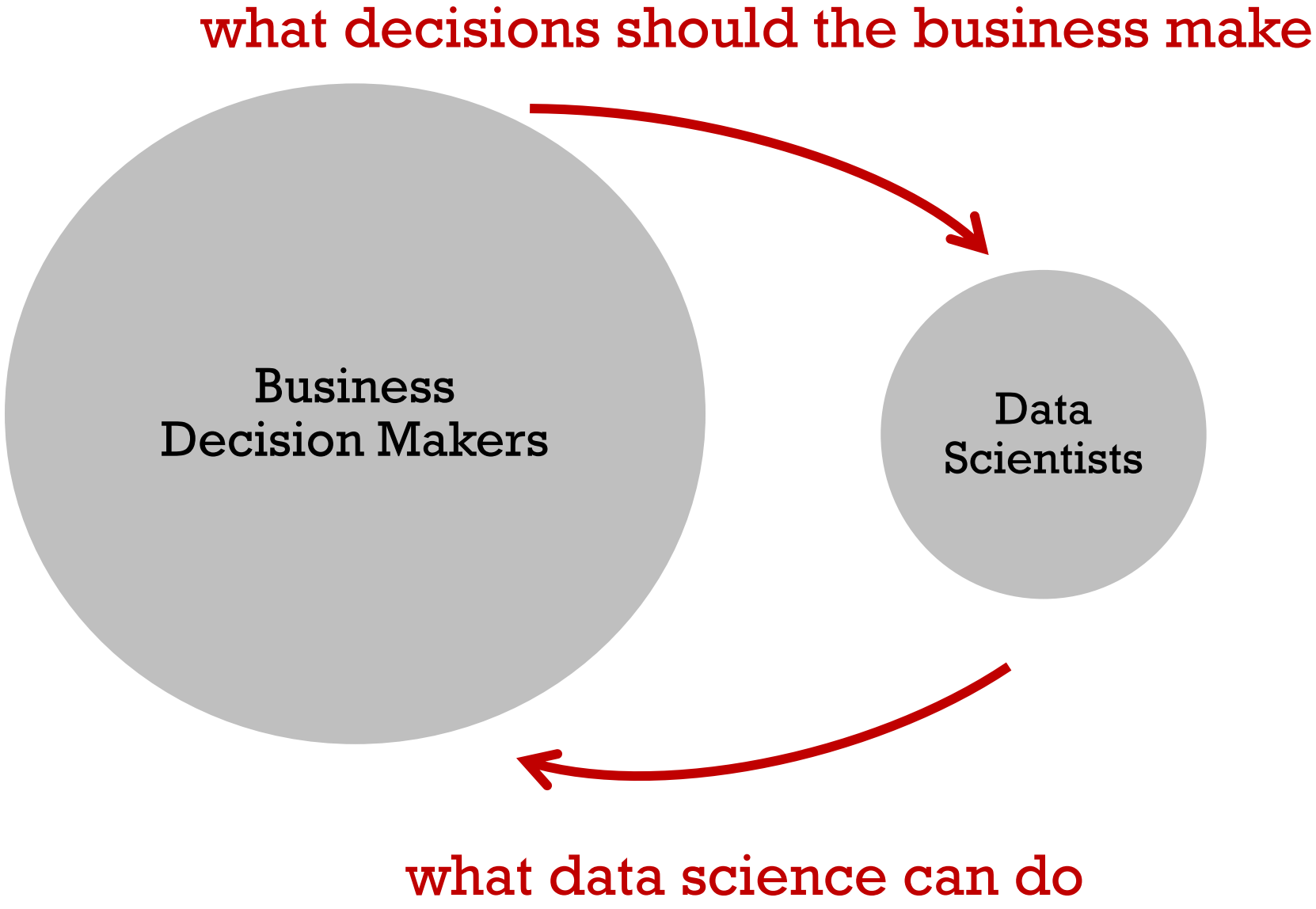
Are You Setting Your Data Scientists Up to Fail?

by Thomas C. Redman

JANUARY 25, 2018

Why does data science fail?

Why does data science fail?

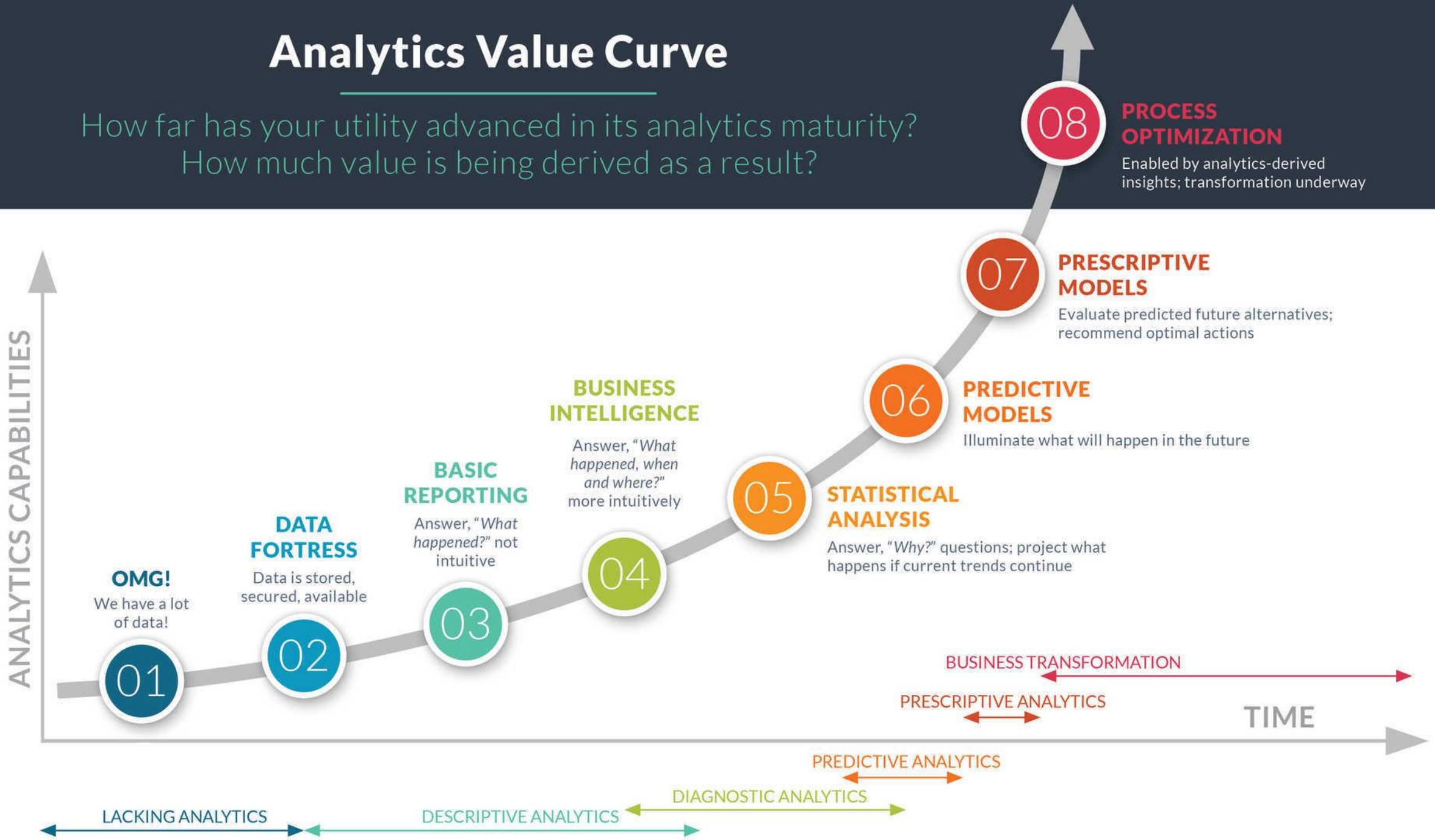


Why does data science fail?

Putting the cart before the horse

Analytics Value Curve

How far has your utility advanced in its analytics maturity?
How much value is being derived as a result?



Why does data science fail?

Putting the cart before the horse

Data scientists not meeting organizational needs

***NEW YORK TIMES* BESTSELLER**

GHETTOSIDE

A TRUE STORY OF MURDER IN AMERICA

JILL LEOVY

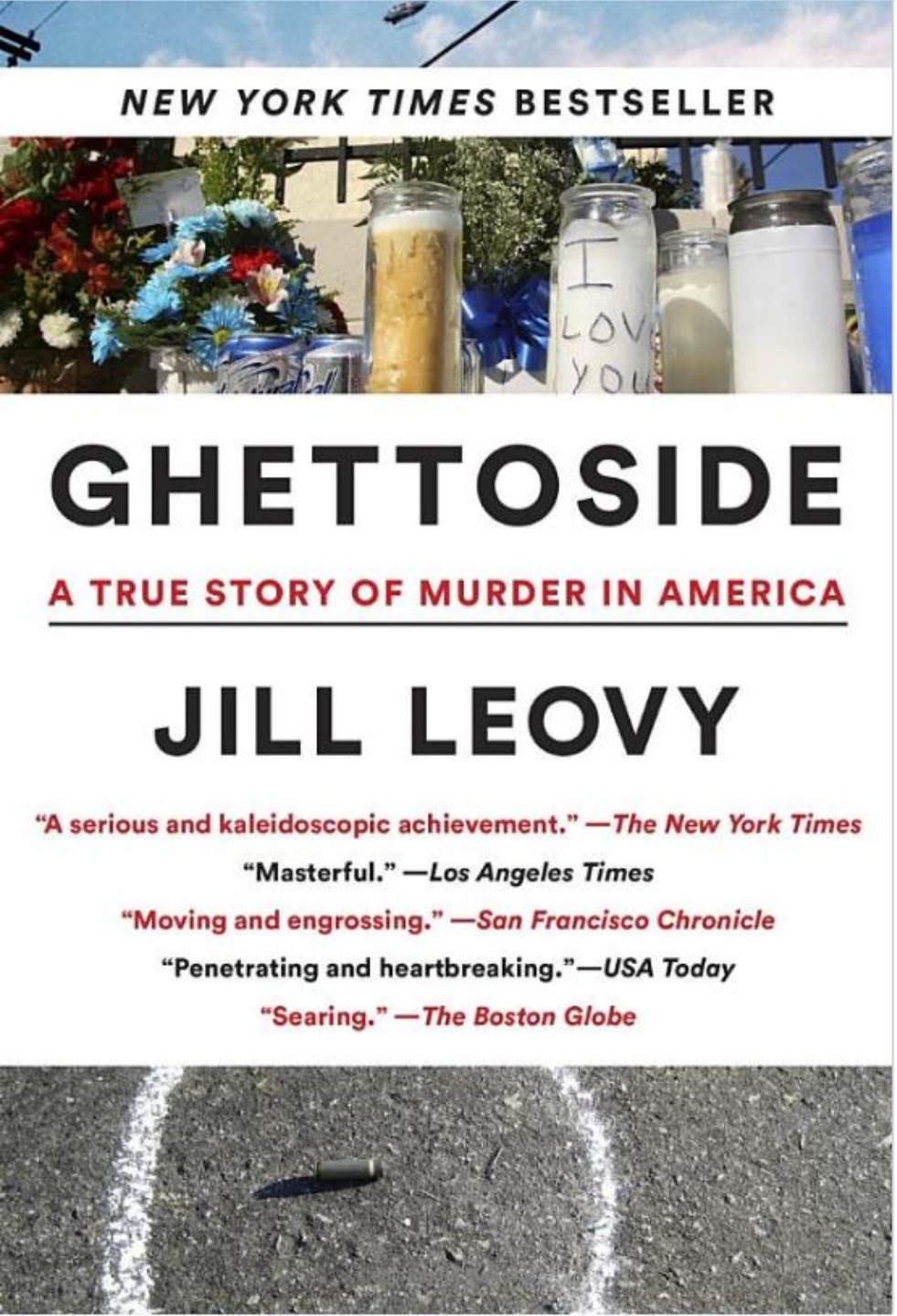
"A serious and kaleidoscopic achievement." —*The New York Times*

"Masterful." —*Los Angeles Times*

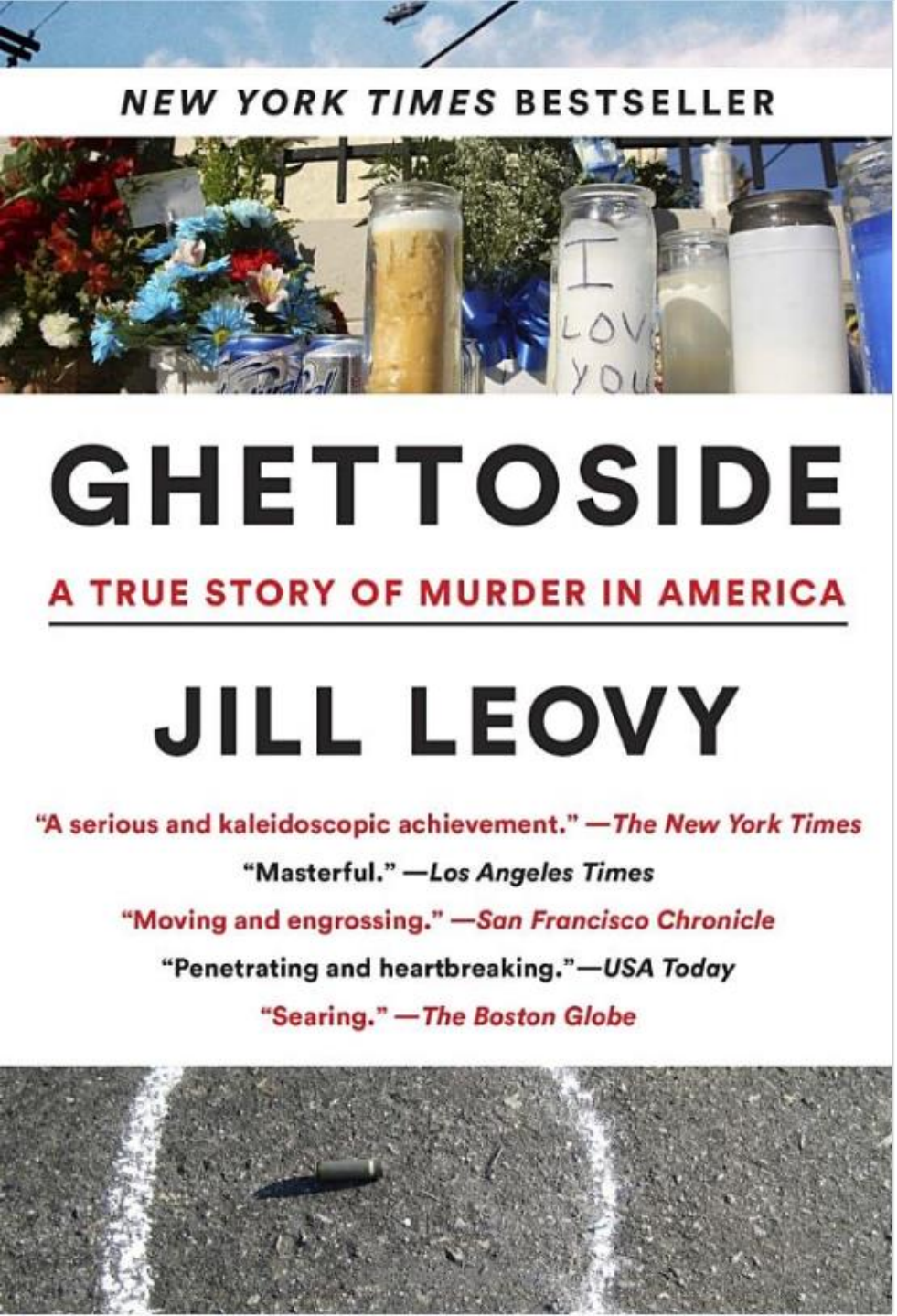
"Moving and engrossing." —*San Francisco Chronicle*

"Penetrating and heartbreaking."—*USA Today*

"Searing." —*The Boston Globe*

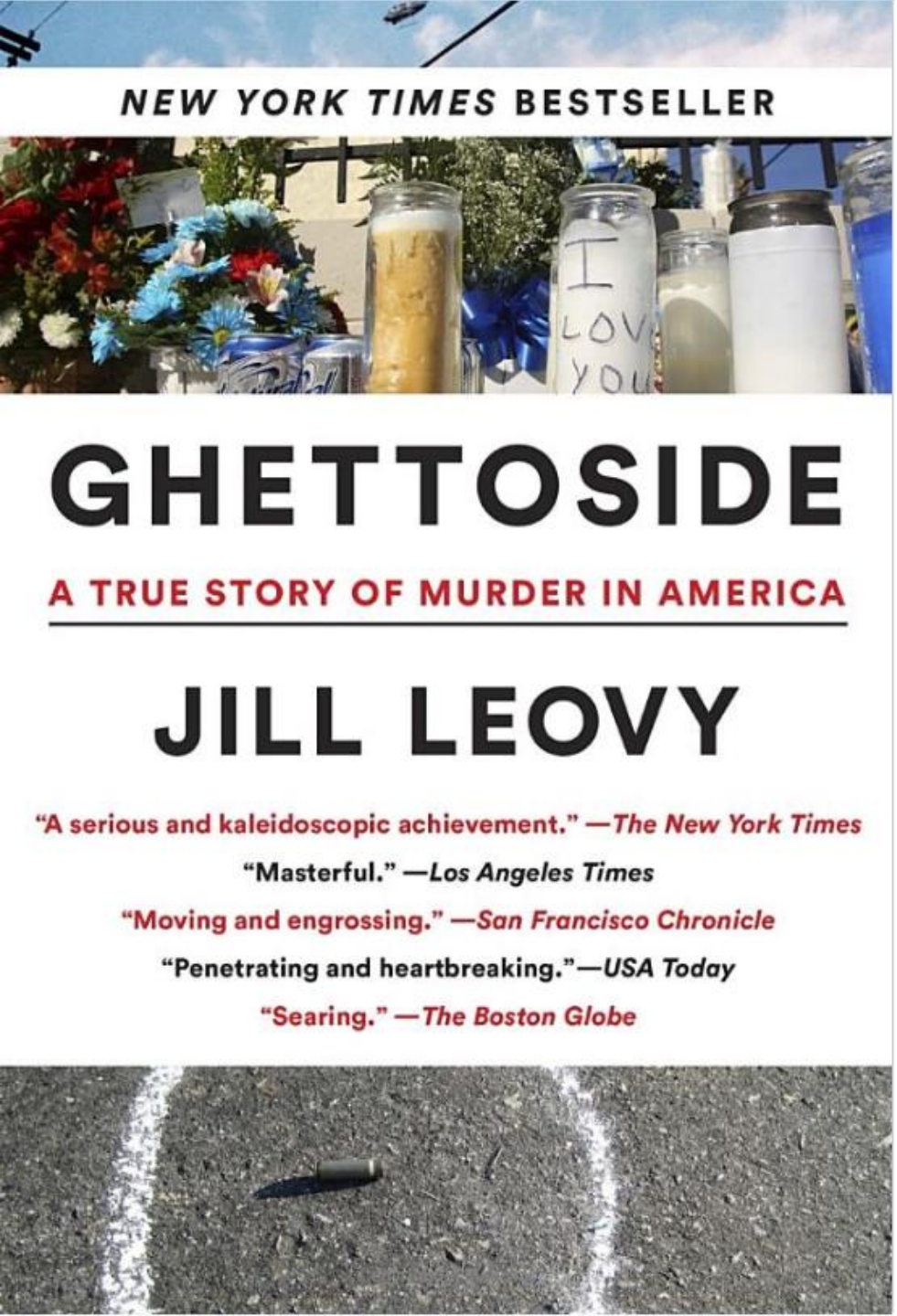


The National Integrated Ballistic Information database catalogued digitized images of bullets and cartridge cases from crime scenes and seized guns.



The National Integrated Ballistic Information database catalogued **digitized images of bullets and cartridge cases** from crime scenes and seized guns.

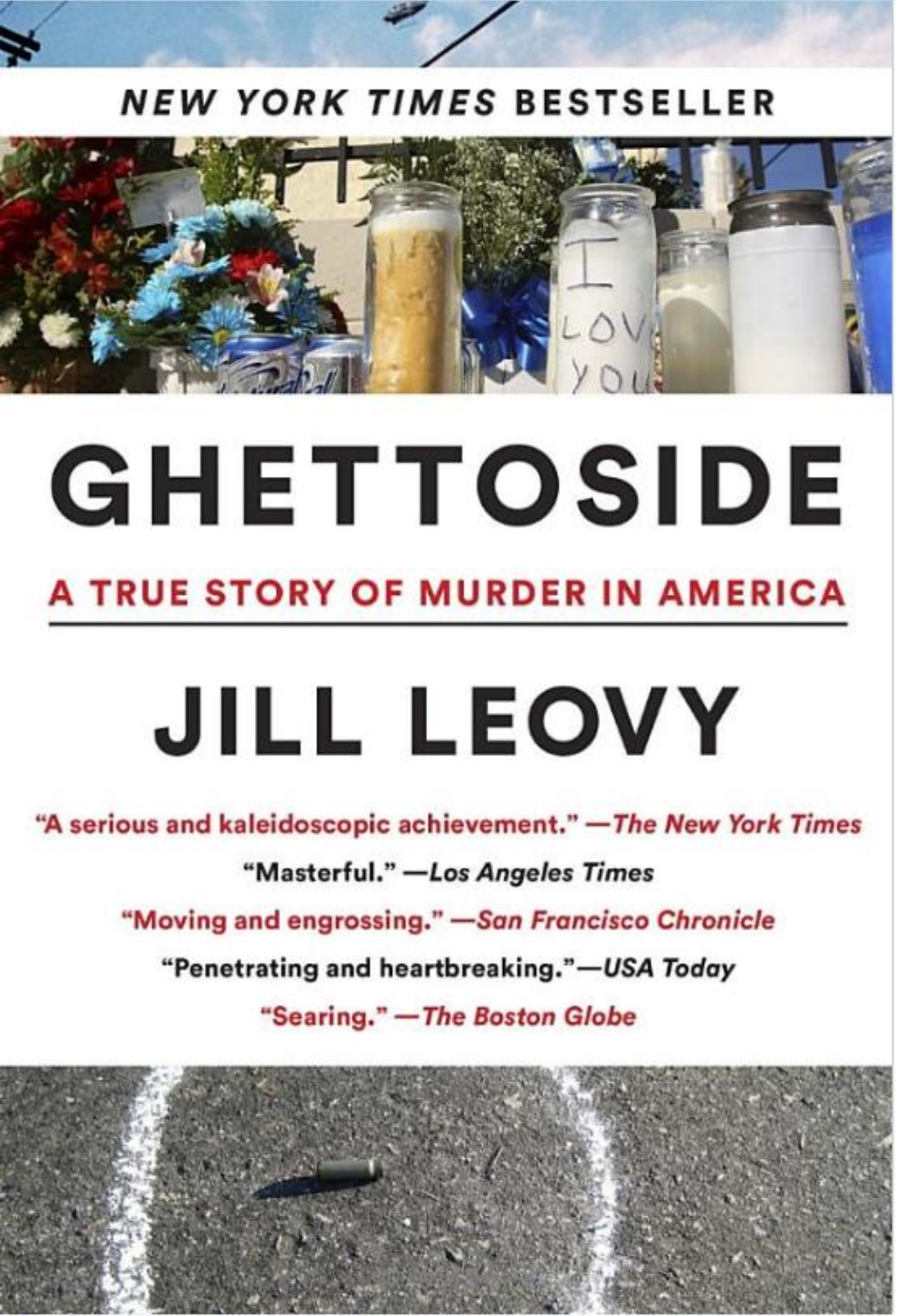
The database could be searched by an **algorithm**. This allowed **fast, cheap searches**, matching ammunition used in crimes to individual weapons.



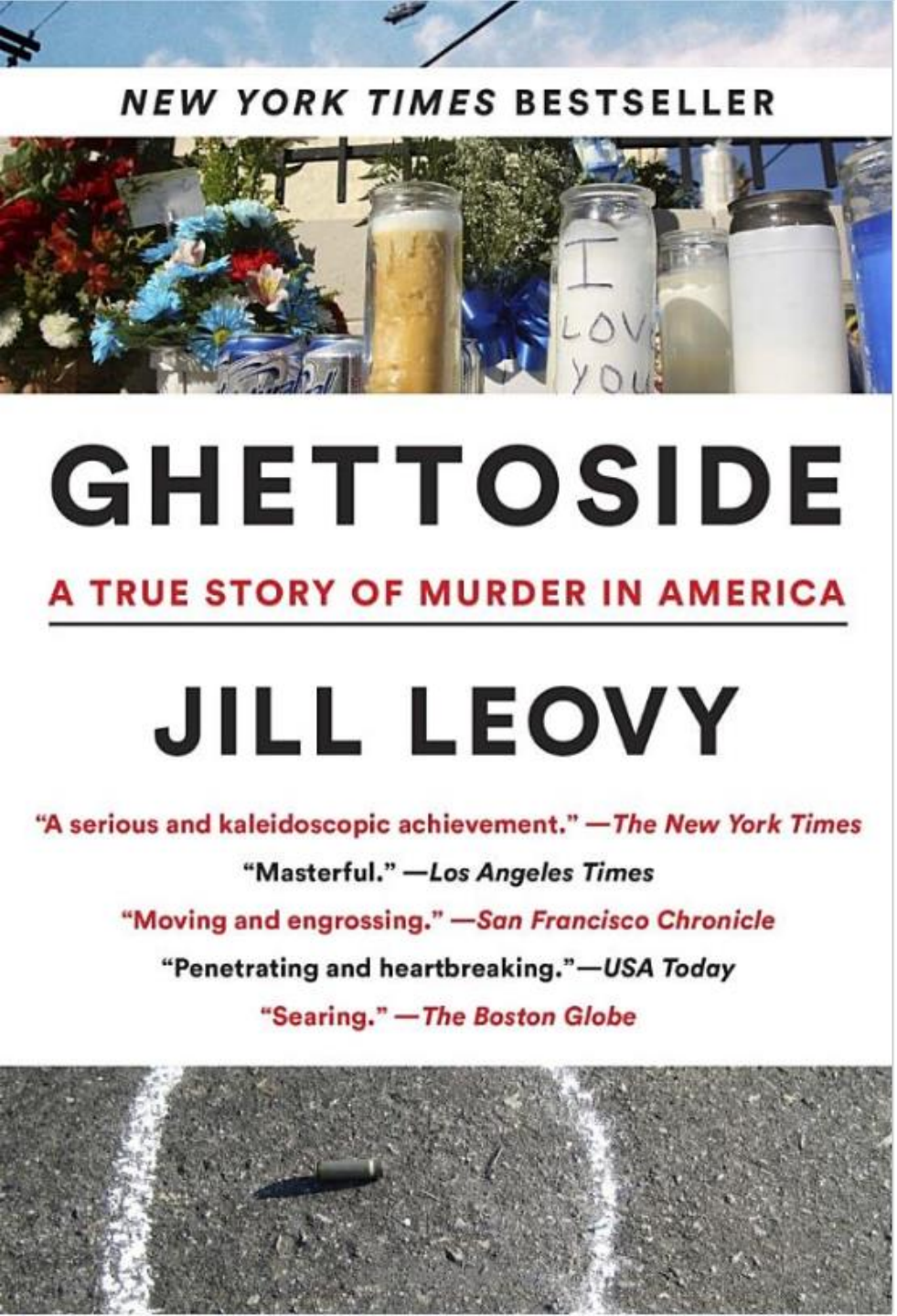
The National Integrated Ballistic Information database catalogued **digitized images of bullets and cartridge cases** from crime scenes and seized guns.

The database could be searched by an **algorithm**. This allowed **fast, cheap searches**, matching ammunition used in crimes to individual weapons.

But,



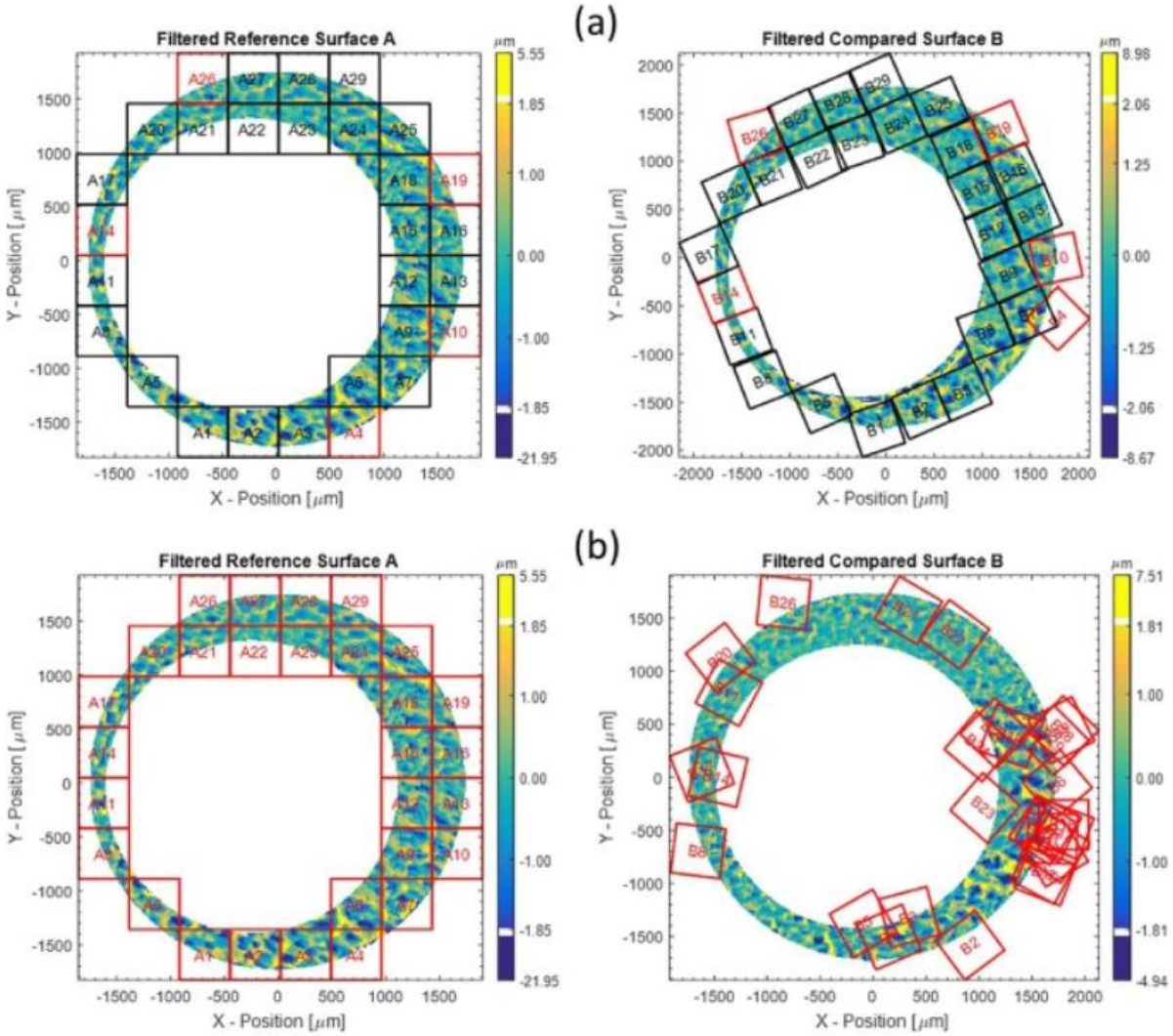
Revolver matches are more difficult than other types of firearm analysis.

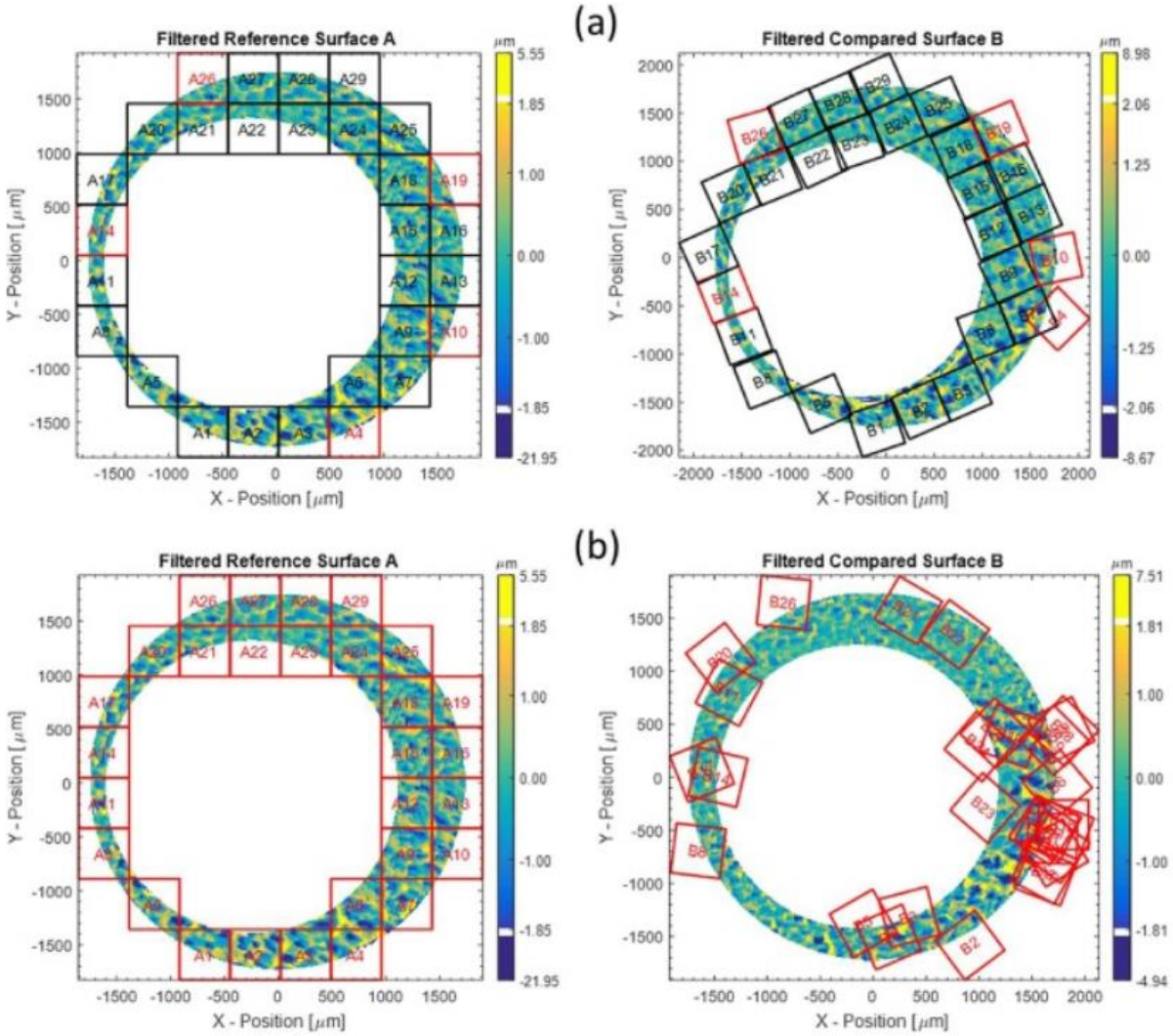


Revolver matches are more difficult than other types of firearm analysis.

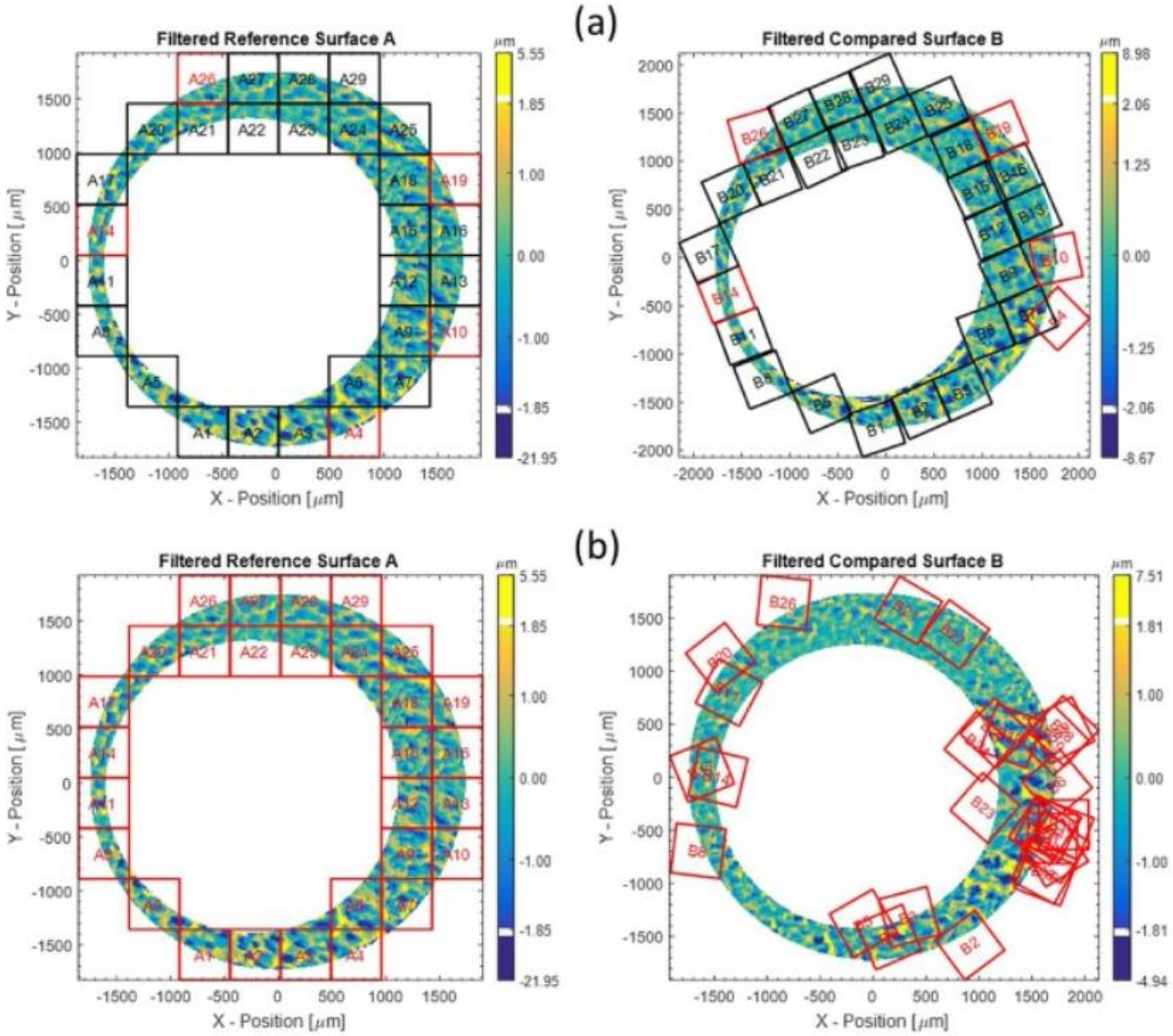
Although the LAPD and other agencies had dutifully entered test-fired bullets from revolvers into the database for years, by the summer of 2007 the system had never successfully matched a bullet used in an LA crime to a revolver.

Not one.


(a)
Filtered Reference Surface A
Filtered Compared Surface B
(b)
Filtered Reference Surface A
Filtered Compared Surface B
Y - Position [μm]
X - Position [μm]
μm



No matter how sophisticated, algorithms have blind spots.



No matter how sophisticated, algorithms have blind spots.

Whether those blind spots matter depends on what you need.

Why does data science fail?

Putting the cart before the horse

Data scientists not meeting organizational needs

Lack of subject matter expertise aka
the mountains cause civil war problem

Ethnicity, Insurgency, and Civil War

JAMES D. FEARON and DAVID D. LAITIN *Stanford University*

An influential conventional wisdom holds that civil wars proliferated rapidly with the end of the Cold War and that the root cause of many or most of these has been ethnic and religious antagonisms. We show that the current prevalence of internal war is mainly the result of a steady accumulation of protracted conflicts since the 1950s and 1960s rather than a sudden change associated with a new, post-Cold War international system. We also find that after controlling for per capita income, more ethnically or religiously diverse countries have been no more likely to experience significant civil violence in this period. We argue for understanding civil war in this period in terms of insurgency or rural guerrilla warfare, a particular form of military practice that can be harnessed to diverse political agendas. The factors that explain which countries have been at risk for civil war are not their ethnic or religious characteristics but rather the conditions that favor insurgency. These include poverty—which marks financially and bureaucratically weak states and also favors rebel recruitment—political instability, rough terrain, and large populations.

Between 1945 and 1999, about 3.33 million battle deaths occurred in the 25 interstate wars that killed at least 1,000 and had at least 100 dead on each side. These wars involved just 25 states that suffered casualties of at least 1,000 and had a median duration of not quite 3 months. In contrast, in the same period there were roughly 127 civil wars that killed at least 1,000, 25 of which were ongoing in 1999. A conservative estimate of the total dead as a direct result of these conflicts is 16.2 million, five times the interstate toll. These civil wars occurred in 73 states—more than a third of the United Nations system—and had a median duration of roughly six years.[1] The civil conflicts in this period surely produced refugee flows far greater than their death toll and far greater than the refugee flows associated with interstate wars since 1945. Cases such as Afghanistan, Somalia, and Lebanon testify to the economic devastation that civil wars can produce. By these crude measures, civil war has been a far greater scourge than interstate war in this period, though it has been studied far less.

What explains the recent prevalence of violent civil conflict around the world? Is it due to the end of the Cold War and associated changes in the international system, or is it the result of longer-term trends? Why have some countries had civil wars while others have not? and Why did the wars break out when they did? We address these questions using data for the period 1945 to 1999 on the 161 countries that had a population of at least half a million in 1990.

The data cast doubt on three influential conventional wisdoms concerning political conflict before and after the Cold War. First, contrary to common opinion, the prevalence of civil war in the 1990s was *not* due to the end of the Cold War and associated changes in the international system. The current level of about one in six countries had already been reached prior to the breakup of the Soviet Union and resulted from a steady, gradual accumulation of civil conflicts that began immediately after World War II.

Second, it appears *not* to be true that a greater degree of ethnic or religious diversity—or indeed any particular cultural demography—by itself makes a country more prone to civil war. This finding runs contrary to a common view among journalists, policy makers, and academics, which holds "plural" societies to be especially conflict-prone due to ethnic or religious tensions and antagonisms.

Third, we find little evidence that one can predict where a civil war will break out by looking for where ethnic or other broad political grievances are strongest. Were this so, one would expect political democracies and states that observe civil liberties to be less civil war-prone than dictatorships. One would further anticipate that state discrimination against minority religions or languages would imply higher risks of civil war. We show that when comparing states at similar levels of per capita income, these expectations are not borne out.

The main factors determining both the secular trend and the cross-sectional variation in civil violence in this period are not ethnic or religious differences or broadly

TABLE 1. Logit Analyses of Determinants of Civil War Onset, 1945–99

	(1) Civil War
Prior war	−0.954**
	(0.314)
Per capita income[a,b]	−0.344***
	(0.072)
log(population)[a,b]	0.263***
	(0.073)
log(% mountainous)	0.219**
	(0.085)
Noncontiguous state	0.443
	(0.274)
Oil exporter	0.858**
	(0.279)
New state	1.709***
	(0.339)
Instability[a]	0.618**
	(0.235)
Democracy[a,c]	0.021
	(0.017)
Ethnic fractionalization	0.166
	(0.373)
Religious fractionalization	0.285
	(0.509)
Anocracy[a]	
Democracy[a,d]	
Constant	−6.731***
	(0.736)
N	6327

TABLE 1. Logit Analyses of Determinants of Civil War Onset, 1945–99

	(1) Civil War
Prior war	−0.954**
	(0.314)
Per capita income[a,b]	−0.344***
	(0.072)
log(population)[a,b]	0.263***
	(0.073)
log(% mountainous)	0.219**
	(0.085)
Noncontiguous state	0.443
	(0.274)
Oil exporter	0.858**
	(0.279)
New state	1.709***
	(0.339)
Instability[a]	0.618**
	(0.235)
Democracy[a,c]	0.021
	(0.017)
Ethnic fractionalization	0.166
	(0.373)
Religious fractionalization	0.285
	(0.509)
Anocracy[a]	
Democracy[a,d]	
Constant	−6.731***
	(0.736)
N	6327

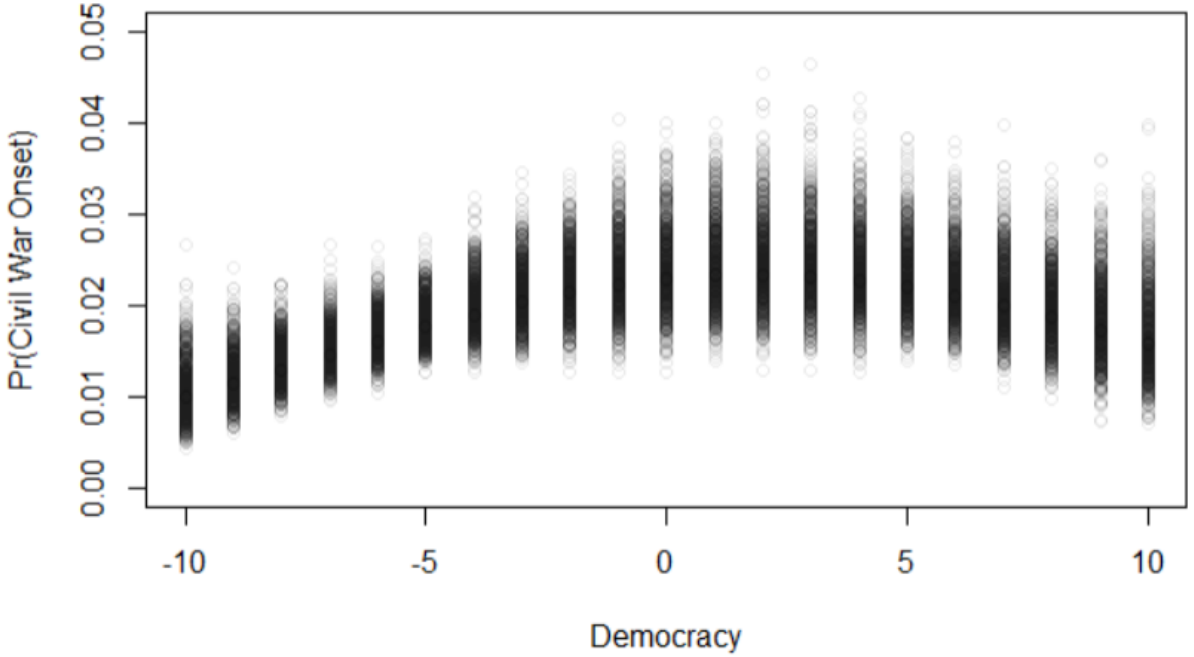
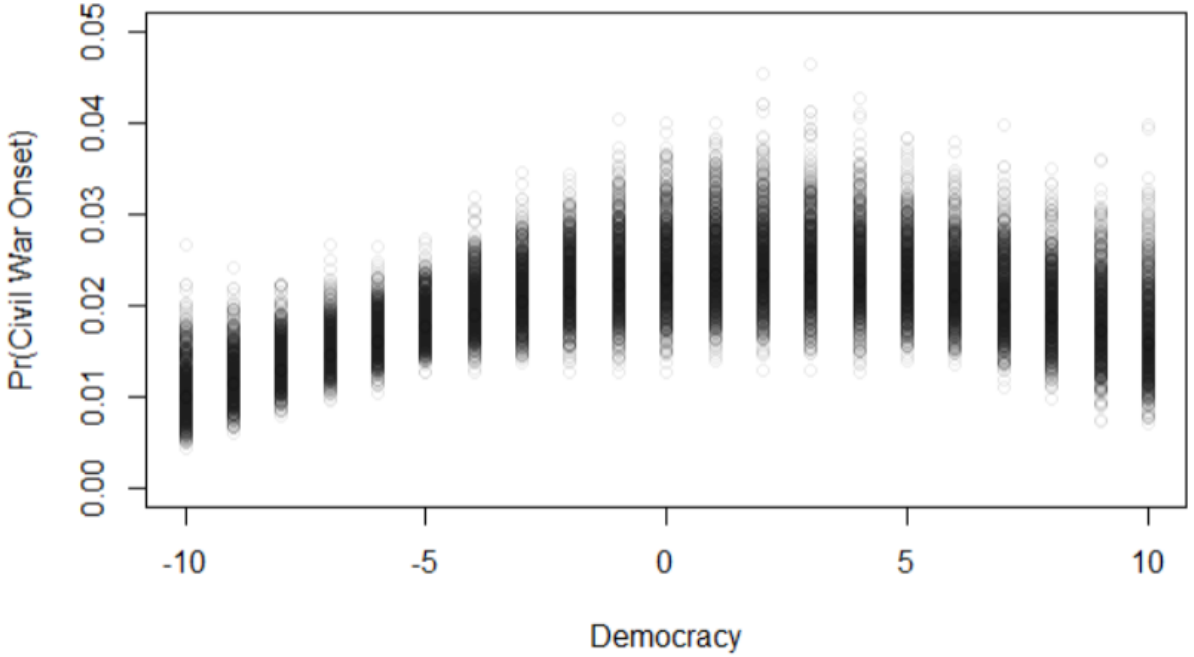


TABLE 1. Logit Analyses of Determinants of Civil War Onset, 1945–99

	(1) Civil War
Prior war	−0.954**
	(0.314)
Per capita income[a,b]	−0.344***
	(0.072)
log(population)[a,b]	0.263***
	(0.073)
log(% mountainous)	0.219**
	(0.085)
Noncontiguous state	0.443
	(0.274)
Oil exporter	0.858**
	(0.279)
New state	1.709***
	(0.339)
Instability[a]	0.618**
	(0.235)
Democracy[a,c]	0.021
	(0.017)
Ethnic fractionalization	0.166
	(0.373)
Religious fractionalization	0.285
	(0.509)
Anocracy[a]	
Democracy[a,d]	
Constant	−6.731***
	(0.736)
N	6327



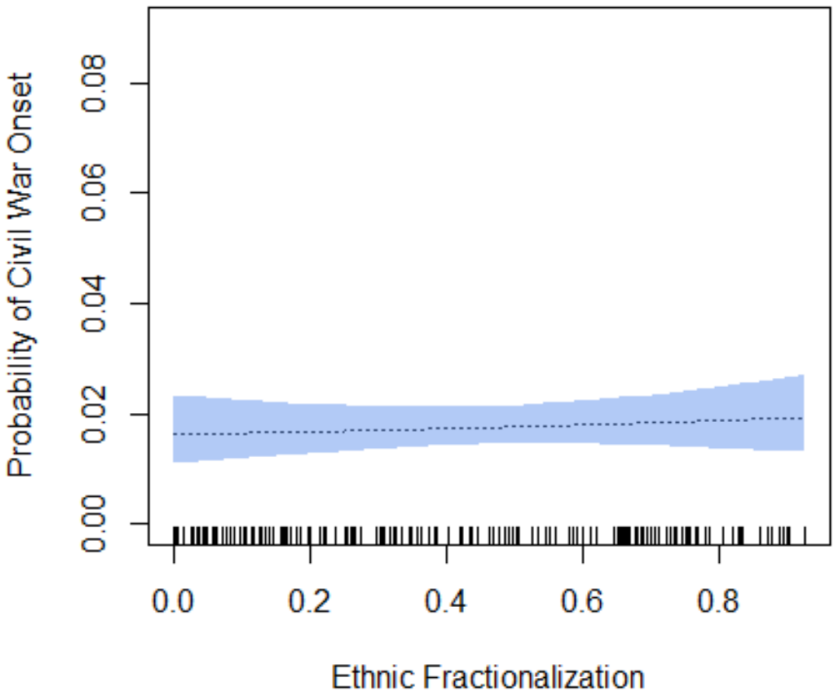
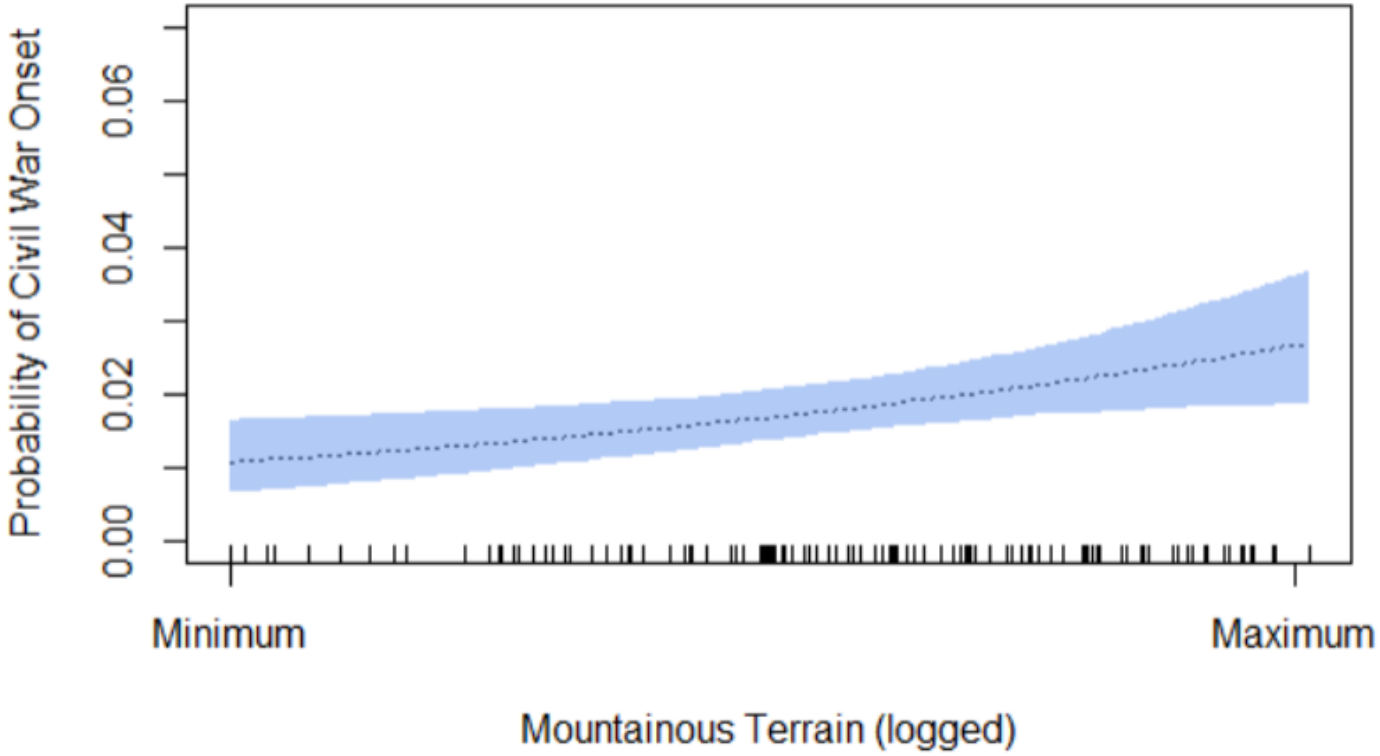
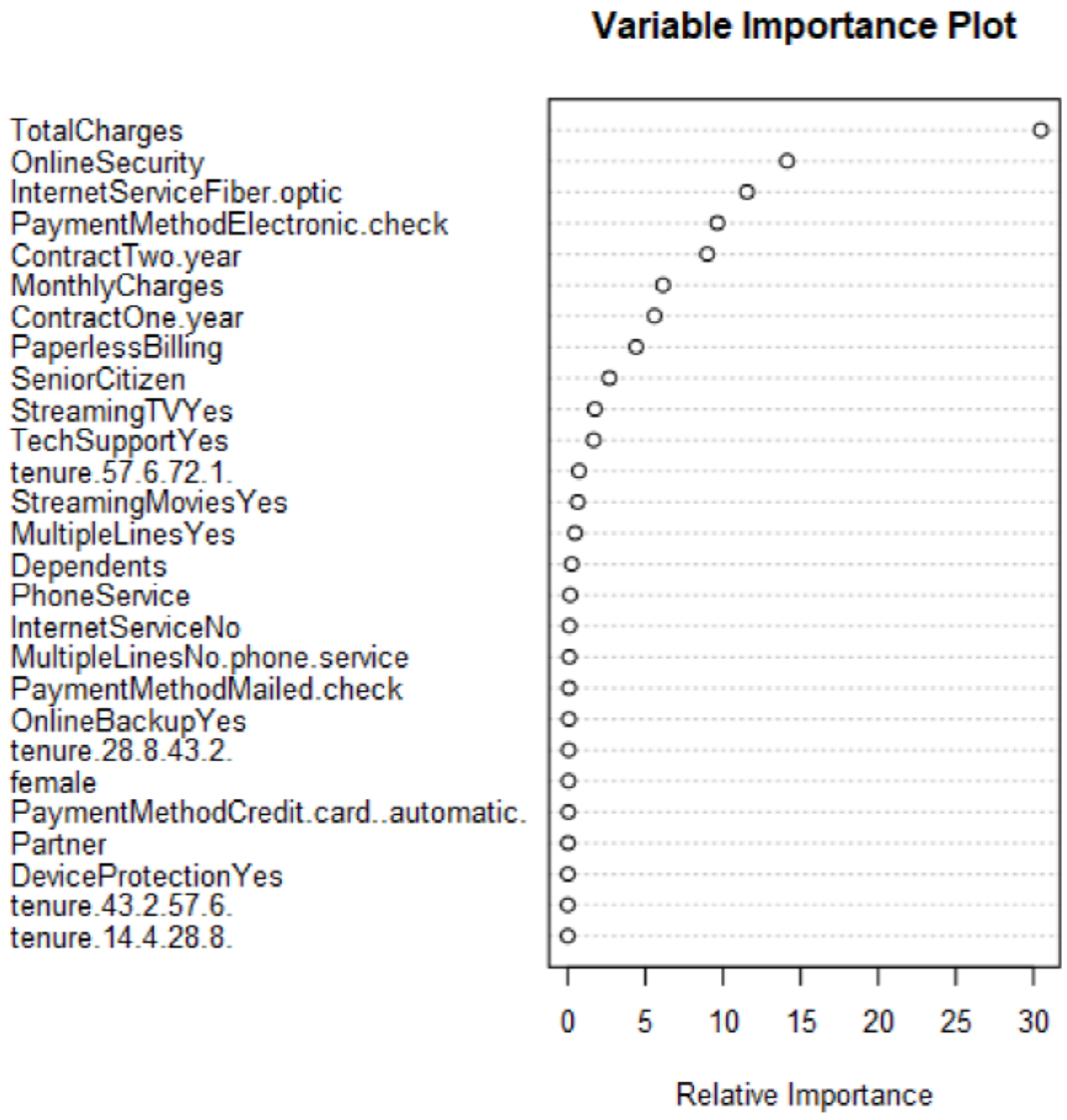


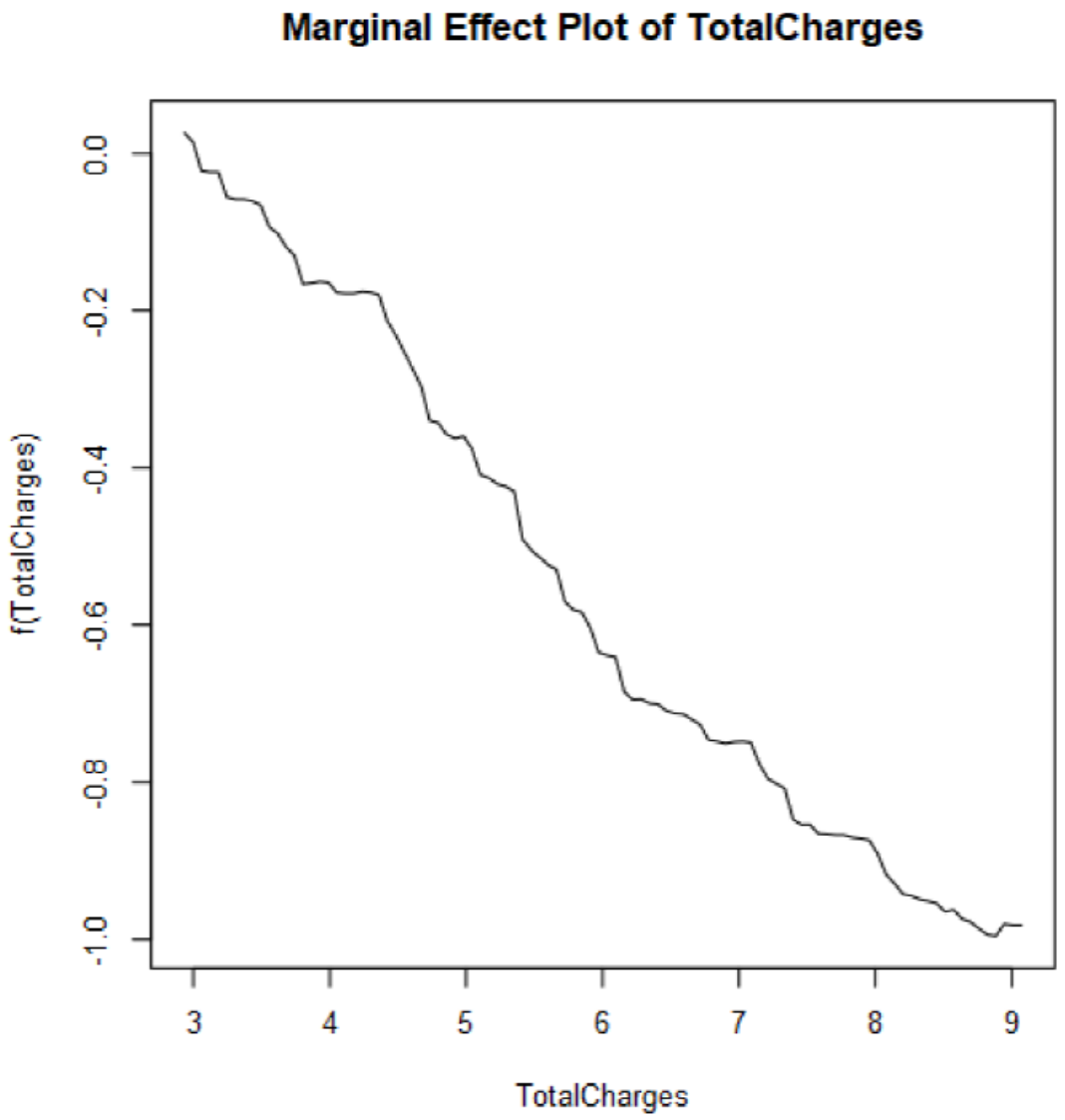
TABLE 1. Logit Analyses of Determinants of Civil War Onset, 1945–99

	(1) Civil War
Prior war	−0.954**
	(0.314)
Per capita income[a,b]	−0.344***
	(0.072)
log(population)[a,b]	0.263***
	(0.073)
log(% mountainous)	0.219**
	(0.085)
Noncontiguous state	0.443
	(0.274)
Oil exporter	0.858**
	(0.279)
New state	1.709***
	(0.339)
Instability[a]	0.618**
	(0.235)
Democracy[a,c]	0.021
	(0.017)
Ethnic fractionalization	0.166
	(0.373)
Religious fractionalization	0.285
	(0.509)
Anocracy[a]	
Democracy[a,d]	
Constant	−6.731***
	(0.736)
N	6327

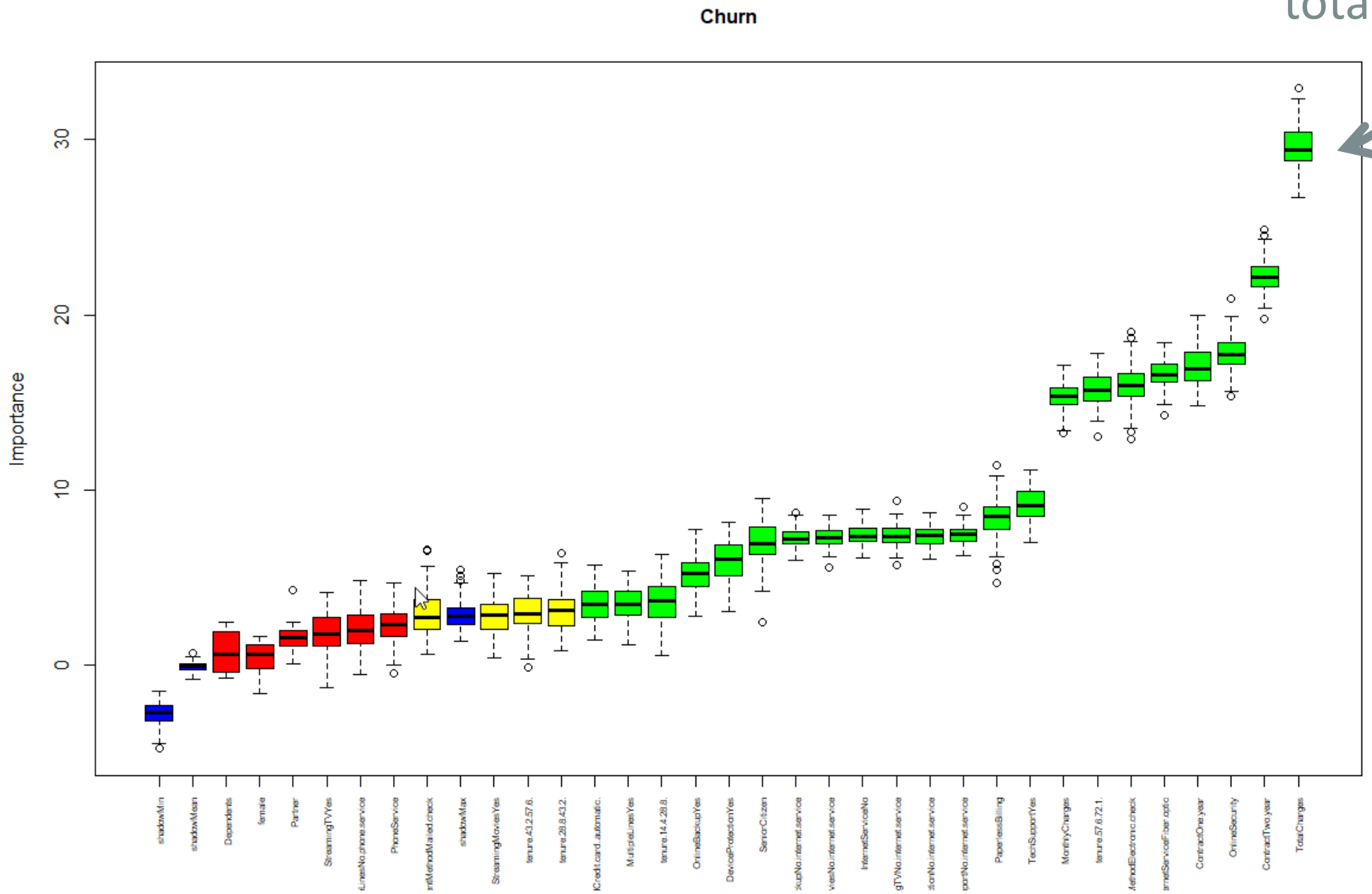


Back to our customer churn example


Variable Importance Plot
TotalCharges
OnlineSecurity
InternetServiceFiber.optic
PaymentMethodElectronic.check
ContractTwo.year
MonthlyCharges
ContractOne.year
PaperlessBilling
SeniorCitizen
StreamingTVYes
TechSupportYes
tenure.57.6.72.1.
StreamingMoviesYes
MultipleLinesYes
Dependents
PhoneService
InternetServiceNo
MultipleLinesNo.phone.service
PaymentMethodMailed.check
OnlineBackupYes
tenure.28.8.43.2.
female
PaymentMethodCredit.card..automatic.
Partner
DeviceProtectionYes
tenure.43.2.57.6.
tenure.14.4.28.8.
0
5
10
15
20
25
30
Relative Importance


Marginal Effect Plot of TotalCharges
f(TotalCharges)
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
3
4
5
6
7
8
9
TotalCharges

total charges



↑ Total Charges → Pr(Customer Churn) ↓

↑ Total Charges → Pr(Customer Churn) ↓

Model: If we **increase total charges**, our customers will stay.

↑ Total Charges → Pr(Customer Churn) ↓

Model: If we **increase total charges**, our customers will stay.

Human: If we **increase customer loyalty**, our customers will stay.

Models require context and interpretation.

Models require context and interpretation.

It is very easy to draw the wrong conclusions if you don’t understand the subject matter.

Models require context and interpretation.

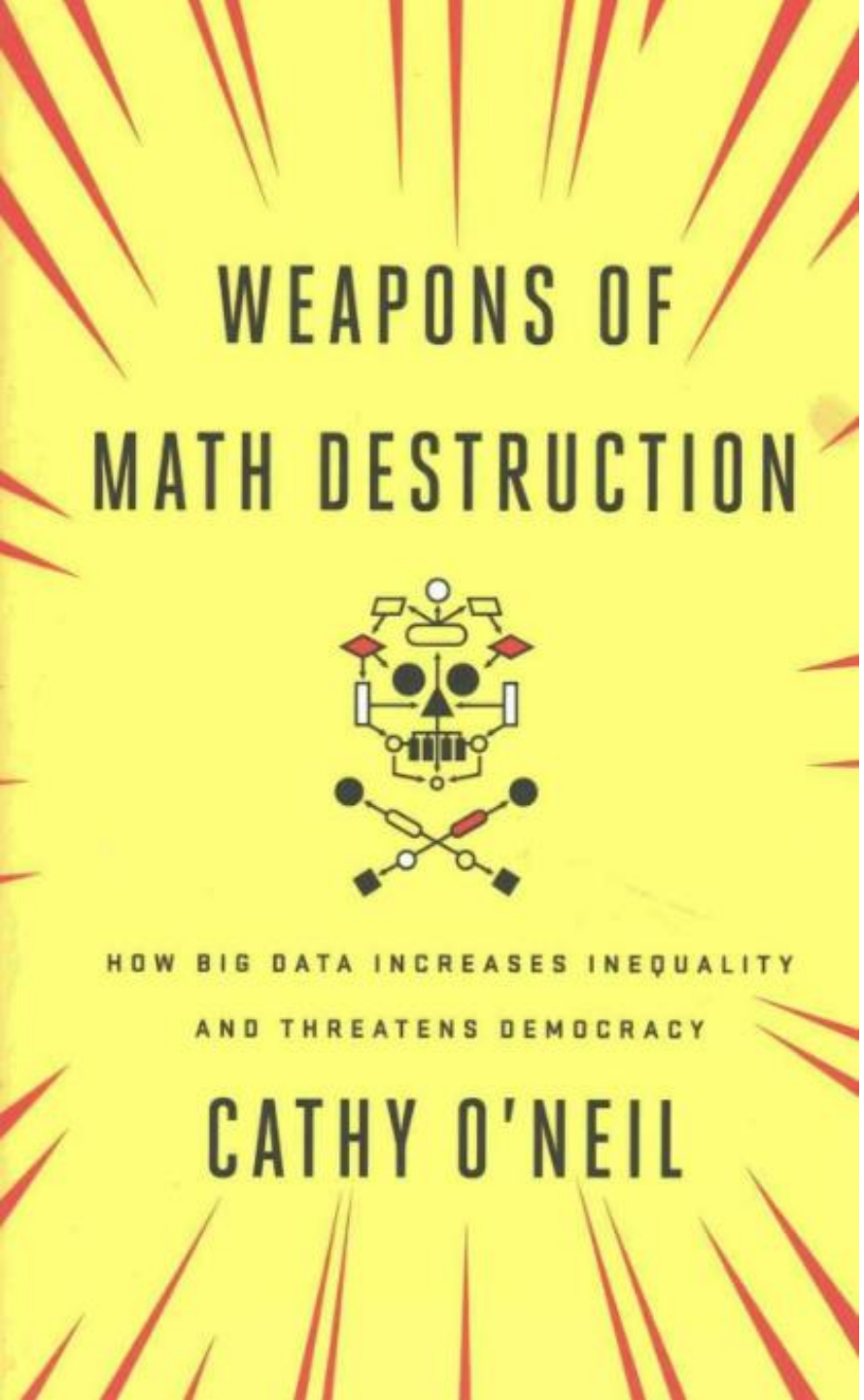
It is very easy to draw the wrong conclusions if you don't understand the subject matter.

Human intelligence is needed to make the appropriate use of algorithms.

WEAPONS OF MATH DESTRUCTION

HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY

CATHY O'NEIL



To create a model, **we make choices** about what's important enough to include, simplifying the world into a toy version that can be easily understood and from which we can infer important facts and actions.

We expect it to handle one job, and accept that it will occasionally act like a clueless machine, **one with enormous blind spots**.

The Washington Post
Democracy Dies in Darkness

Local

Firing of D.C. teacher reveals flaws in value-added evaluation

By **Valerie Strauss**
March 7, 2012

The firing of a D.C. teacher called "creative," "visionary" and "motivating" is the latest example of the many things wrong with value-added methods to evaluate teachers, the newest trend in school reform that is sweeping states with a push from the Obama administration.

My colleague Bill Turque tells the story of teacher Sarah Wysocki, who was let go by D.C. public schools because her students got low standardized test scores, even though she received stellar personal evaluations as a teacher.

The going theory was that the students weren't learning enough because their teachers weren't doing a good job. So in 2009, Rhee implemented a plan to weed out the low-performing teachers. This is the trend in troubled school districts around the country, and from a systems engineering perspective the thinking makes perfect sense: Evaluate the teachers. Get rid of the worst ones, and place the best ones where they can do the most good. In the language of data scientists, this "optimizes" the school system, presumably ensuring better results for the kids.

The going theory was that the students weren't learning enough because their teachers weren't doing a good job. So in 2009, Rhee implemented a plan to weed out the low-performing teachers. This is the trend in troubled school districts around the country, and from a systems engineering perspective the thinking makes perfect sense: Evaluate the teachers. Get rid of the worst ones, and place the best ones where they can do the most good. In the language of data scientists, this "optimizes" the school system, presumably ensuring better results for the kids.

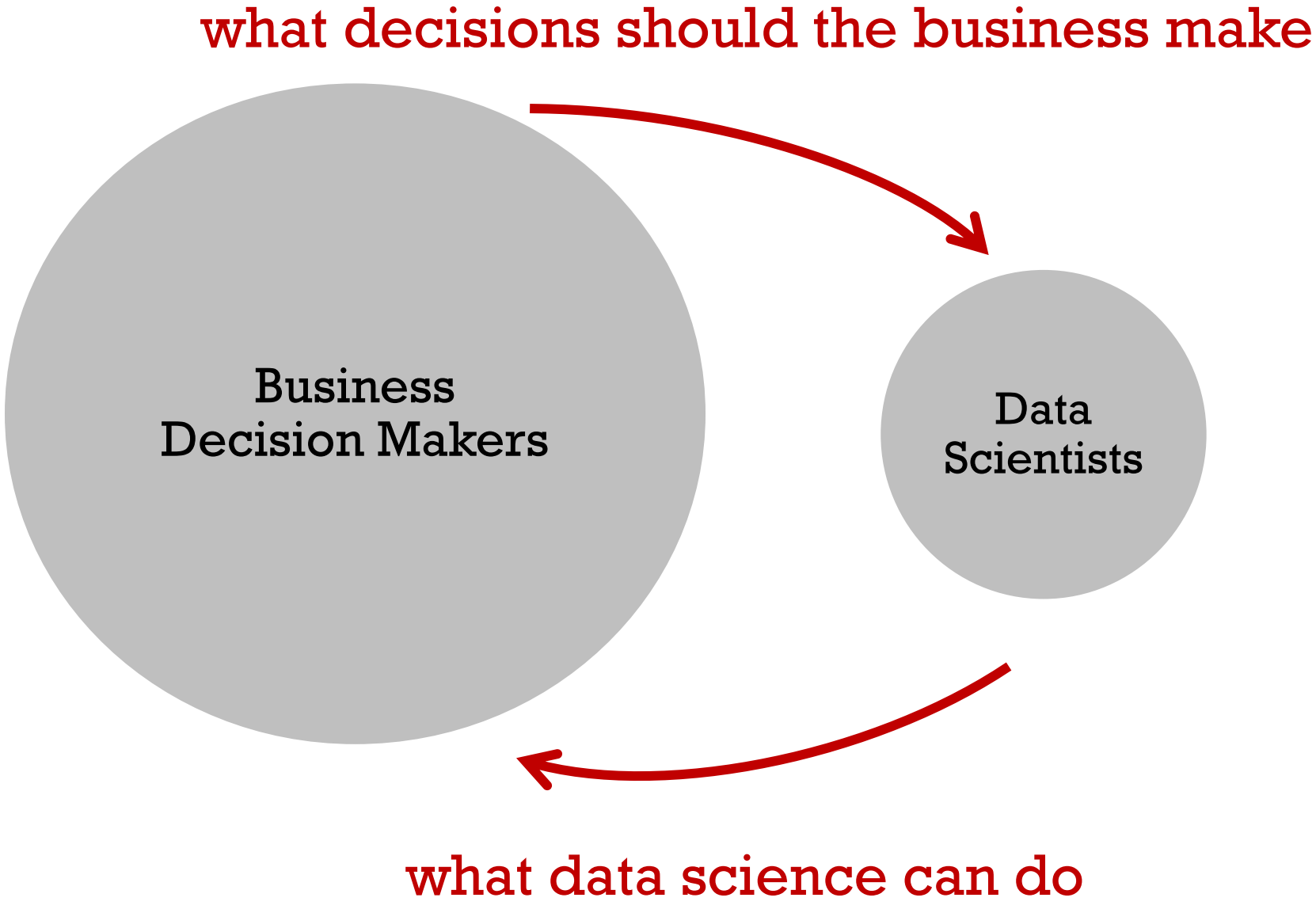
Sarah Wysocki, a fifth-grade teacher, didn't seem to have any reason to worry. She had been at MacFarland Middle School for only two years but was already getting excellent reviews from her principal and her students' parents. One evaluation praised her attentiveness to the children; another called her "one of the best teachers I've ever come into contact with."

Yet at the end of the 2010–11 school year, Wysocki received a miserable score on her IMPACT evaluation. Her problem was a new scoring system known as value-added modeling which purported to measure her effectiveness in teaching math and language skills. That score, generated by an algorithm, represented half of her overall evaluation, and it outweighed the positive reviews from school administrators and the community. This left the district with no choice but to fire her, along with 205 other teachers who had IMPACT scores below the minimal threshold.

Yet at the end of the 2010–11 school year, Wysocki received a miserable score on her IMPACT evaluation. Her problem was a new scoring system known as value-added modeling which purported to measure her effectiveness in teaching math and language skills. That score, generated by an algorithm, represented half of her overall evaluation, and it outweighed the positive reviews from school administrators and the community. This left the district with no choice but to fire her, along with 205 other teachers who had IMPACT scores below the minimal threshold.

Well, she learned, it was complicated. The district had hired a consultancy, Princeton-based Mathematica Policy Research, to come up with the evaluation system. Mathematica's challenge was to measure the educational progress of the students in the district and then to calculate how much of their advance or decline could be attributed to their teachers.

Why does data science fail?



Your scientists were so preoccupied with whether they **could**, they didn't stop to think if they **should**.

The New York Times

TheUpshot

HIDDEN BIAS

When Algorithms Discriminate

By Claire Cain Miller

July 9, 2015

The online world is shaped by forces beyond our control, determining the stories we read on Facebook, the people we meet on OkCupid and the search results we see on Google. Big data is used to make decisions about health care, employment, housing, education and policing.

But can computer programs be discriminatory?

There is a widespread belief that software and algorithms that rely on data are objective. But software is not free of human influence. Algorithms are written and maintained by people, and machine learning algorithms adjust what they do based on people's behavior. As a result, say researchers in computer science, ethics and law, algorithms can reinforce human prejudices.

The New York Times

Knight Capital Says Trading Glitch Cost It $440 Million

BY NATHANIEL POPPER AUGUST 2, 2012 9:07 AM 356

Runaway Trades Spread Turmoil Across Wall St.

Errant trades from the Knight Capital Group began hitting the New York Stock Exchange almost as soon as the opening bell rang on Wednesday. Brendan McDermid/Reuters

Connectivity

Neural Network Learns to Identify Criminals by Their Faces

The effort aimed at identifying criminals from their mugshots raises serious ethical issues about how we should use artificial intelligence.

by Emerging Technology from the arXiv November 22, 2016

Soon after the invention of photography, a few criminologists began to notice patterns in mugshots they took of criminals. Offenders, they said, had particular facial features that allowed them to be identified as law breakers.

That is one big pile of ----.

wrapping up

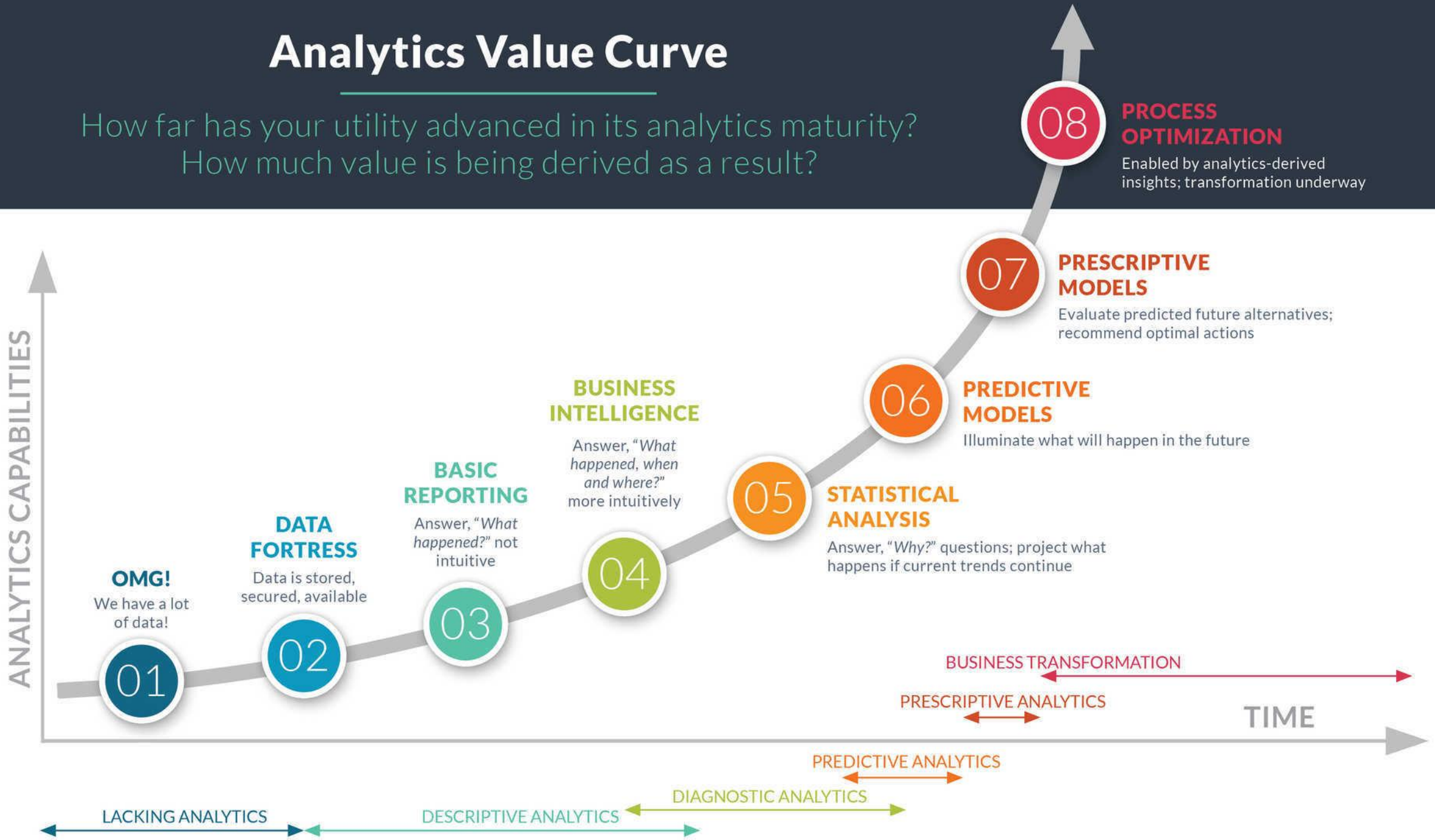
Should I go out and hire a bunch of **data scientists**?

Should I go out and hire a bunch of **data scientists**?

Probably not to start.

Analytics Value Curve

How far has your utility advanced in its analytics maturity?
How much value is being derived as a result?



Should I go out and hire a bunch of data scientists?

Probably not to start.

But you can start looking for opportunities for data science.

Two necessary components to make use of **data science**.

Two necessary components to make use of data science.

- There is an unknown pattern that would be useful to know.

Two necessary components to make use of data science.

- There is an unknown pattern that would be useful to know.
- There is data to learn the pattern.

Two necessary components to make use of **data science**.

- There is an unknown pattern that would be useful to know.
- There is data to learn the pattern.

Two necessary components to make use of **data science**.

- There is an **unknown pattern** that would be **useful to know.**
- **There is data** to learn the pattern.

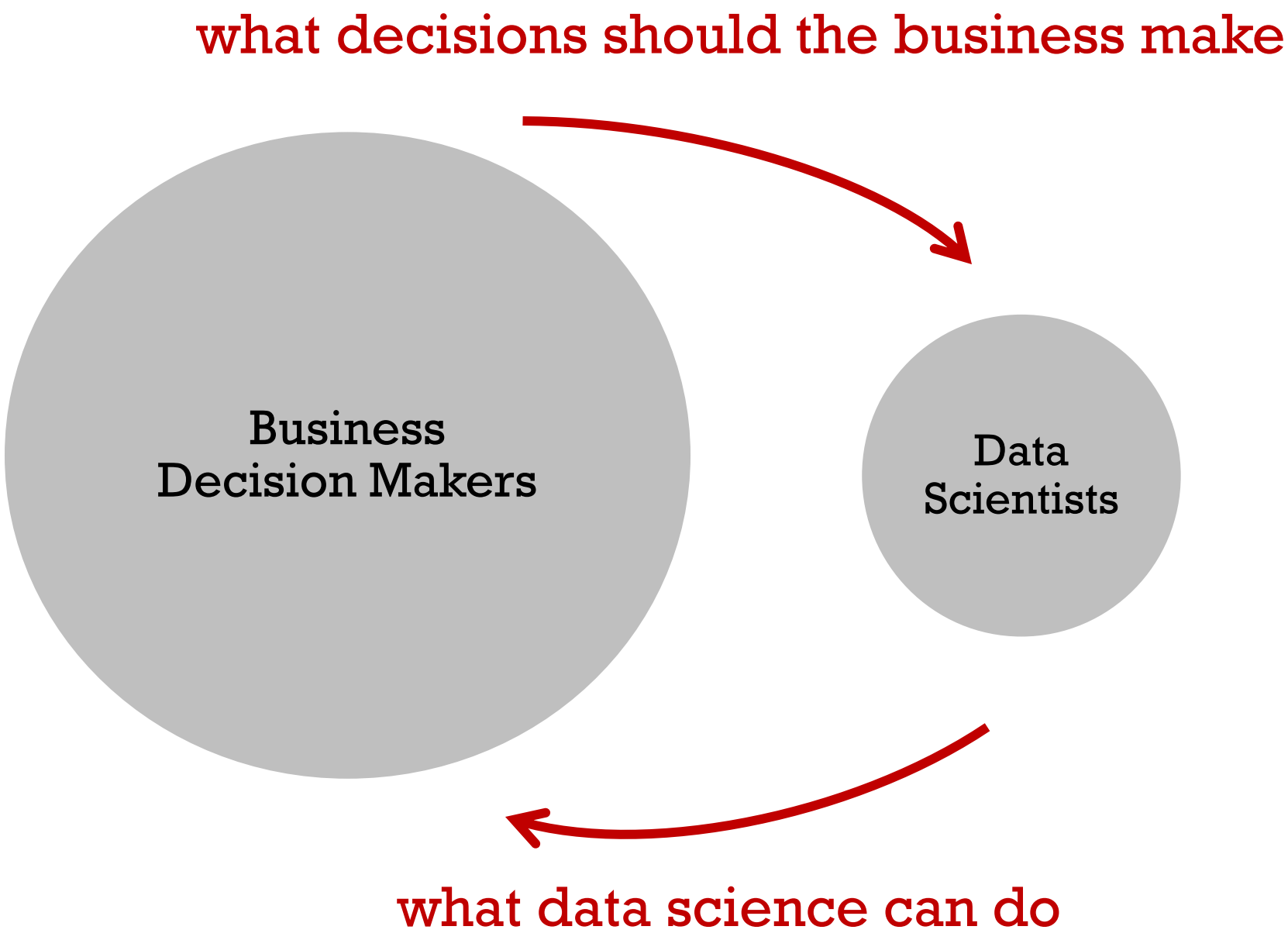
Business Knowledge

Two necessary components to make use of **data science**.

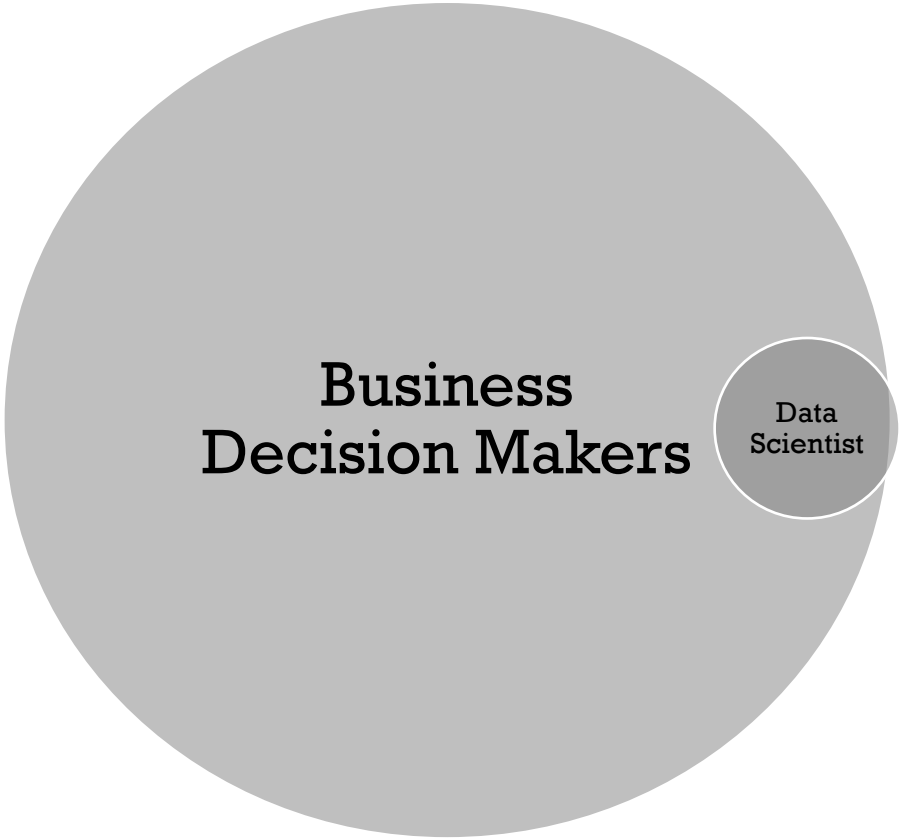
- There is an unknown pattern that would be useful to know.
- There is data to learn the pattern.

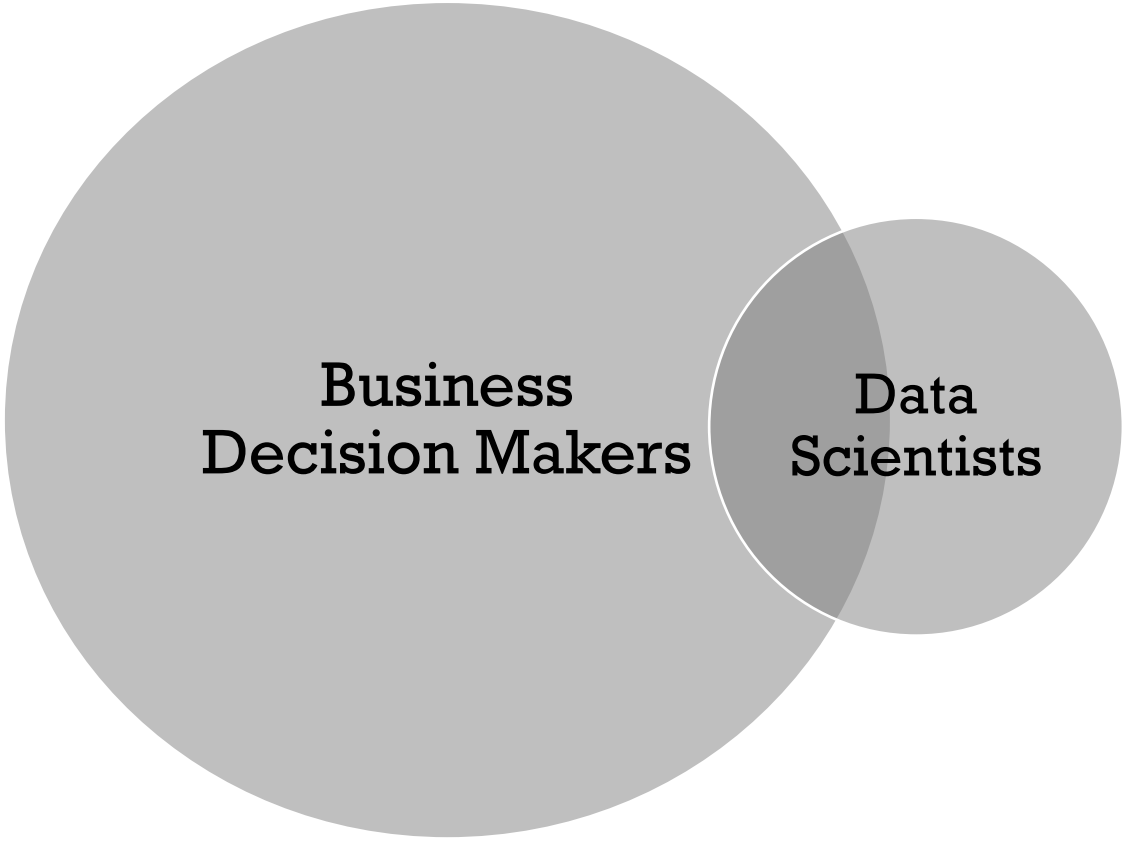
Business Knowledge

Data Science


what decisions should the business make
Business
Decision Makers
Data
Scientists
what data science can do

Business
Decision Makers

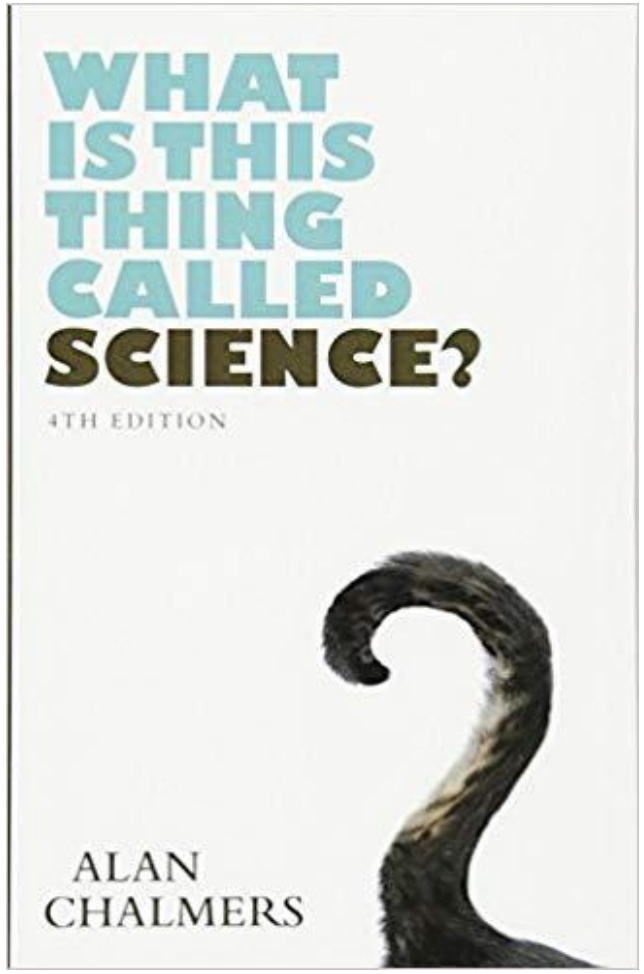

Business
Decision Makers
Data
Scientist

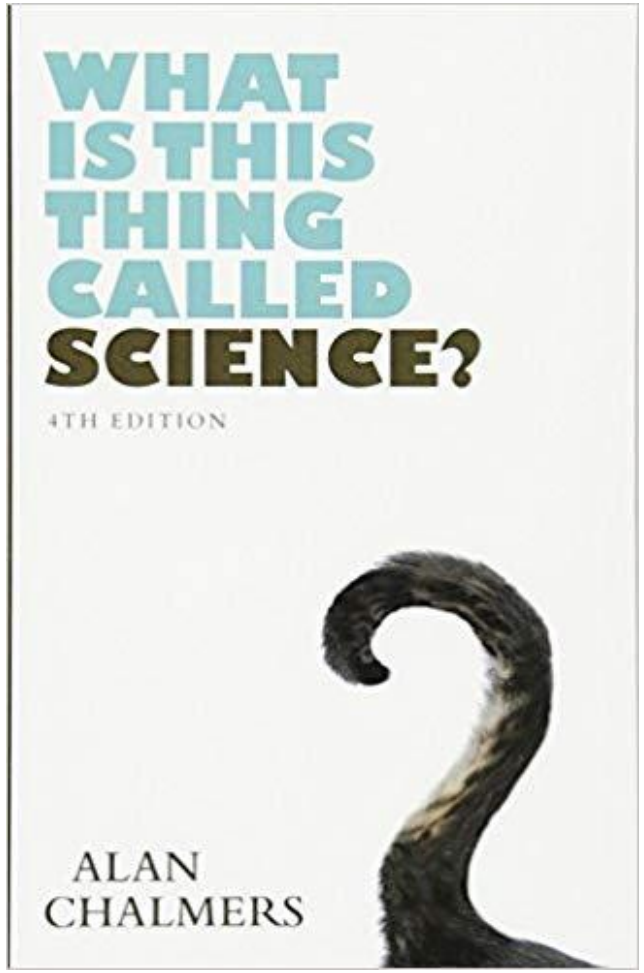

Business
Decision Makers
Data
Scientists

What, if anything, is special about **data science**?

Data science enables us to learn about and recreate a data generating process.

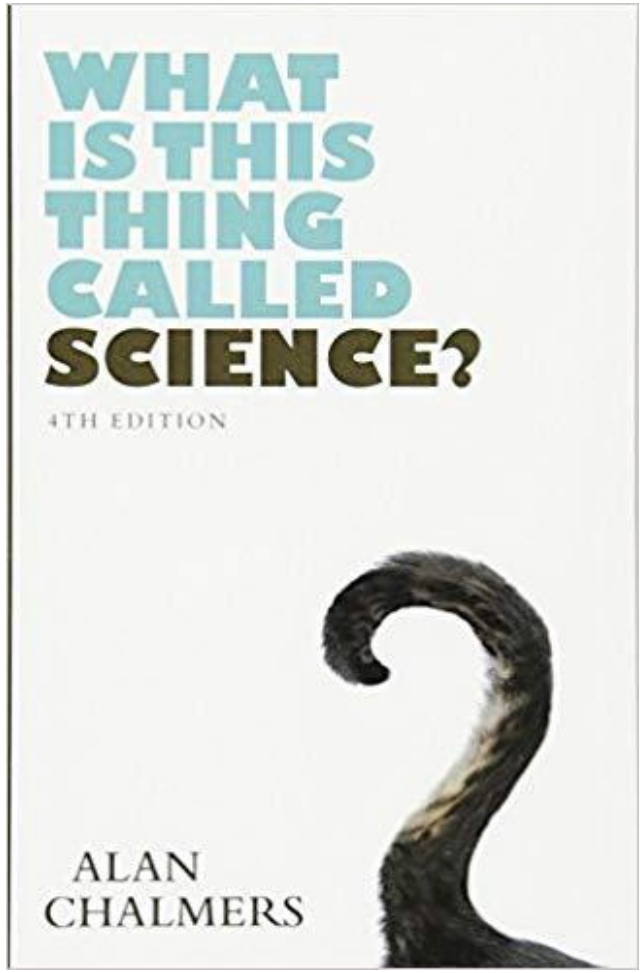
Doing so **can** enable us to make faster, more accurate decisions by discovering where we are wrong.


WHAT IS THIS THING CALLED SCIENCE?
4TH EDITION
ALAN CHALMERS



Many processes are at work in the world around us, and they are all superimposed on, and interact with, each other in complicated ways.

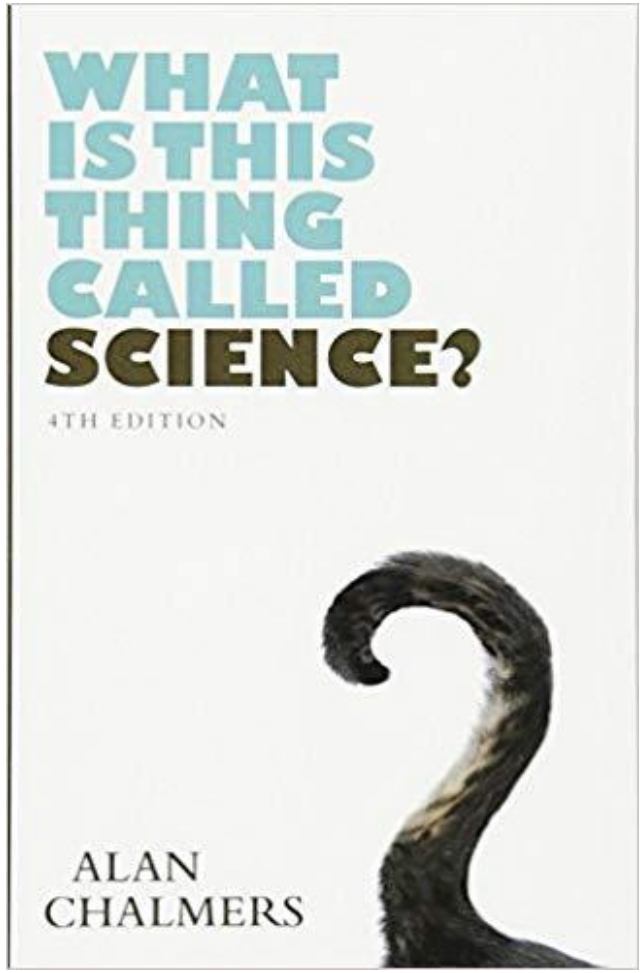
A falling leaf is subject to gravity, air resistance, the force of winds, and will also rot to some degree as it falls.



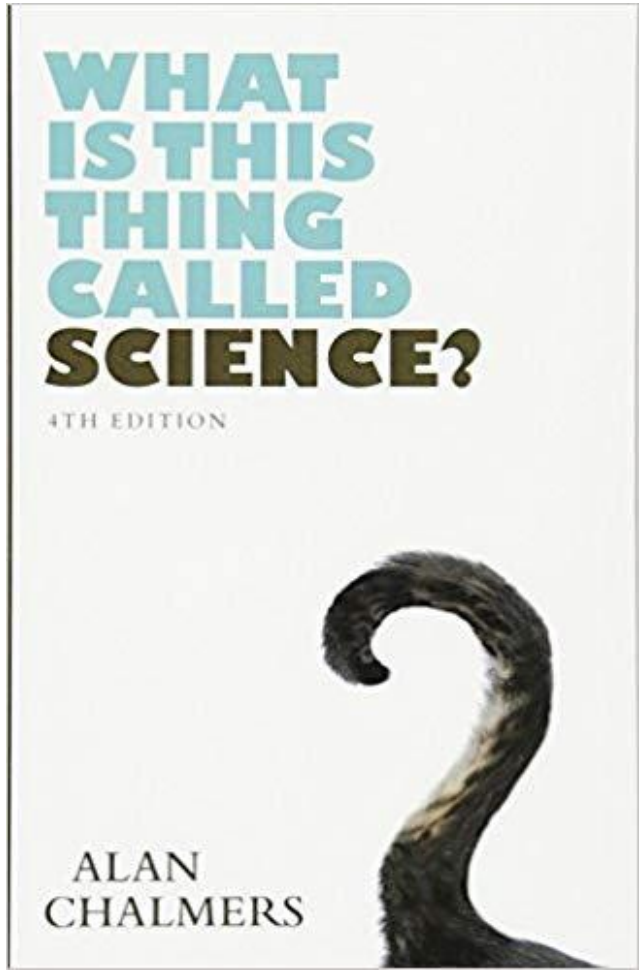
Many processes are at work in the world around us, and they are all superimposed on, and interact with, each other in complicated ways.

A falling leaf is subject to gravity, air resistance, the force of winds, and will also rot to some degree as it falls.

Observation of falling leaves will not yield Galileo's law of fall.



Understandings of nature have to be worked for, and their establishment involves **considerable know-how and practical trial and error**.



Understandings of nature have to be worked for, and their establishment involves **considerable know-how and practical trial and error**.

We learn from our mistakes.

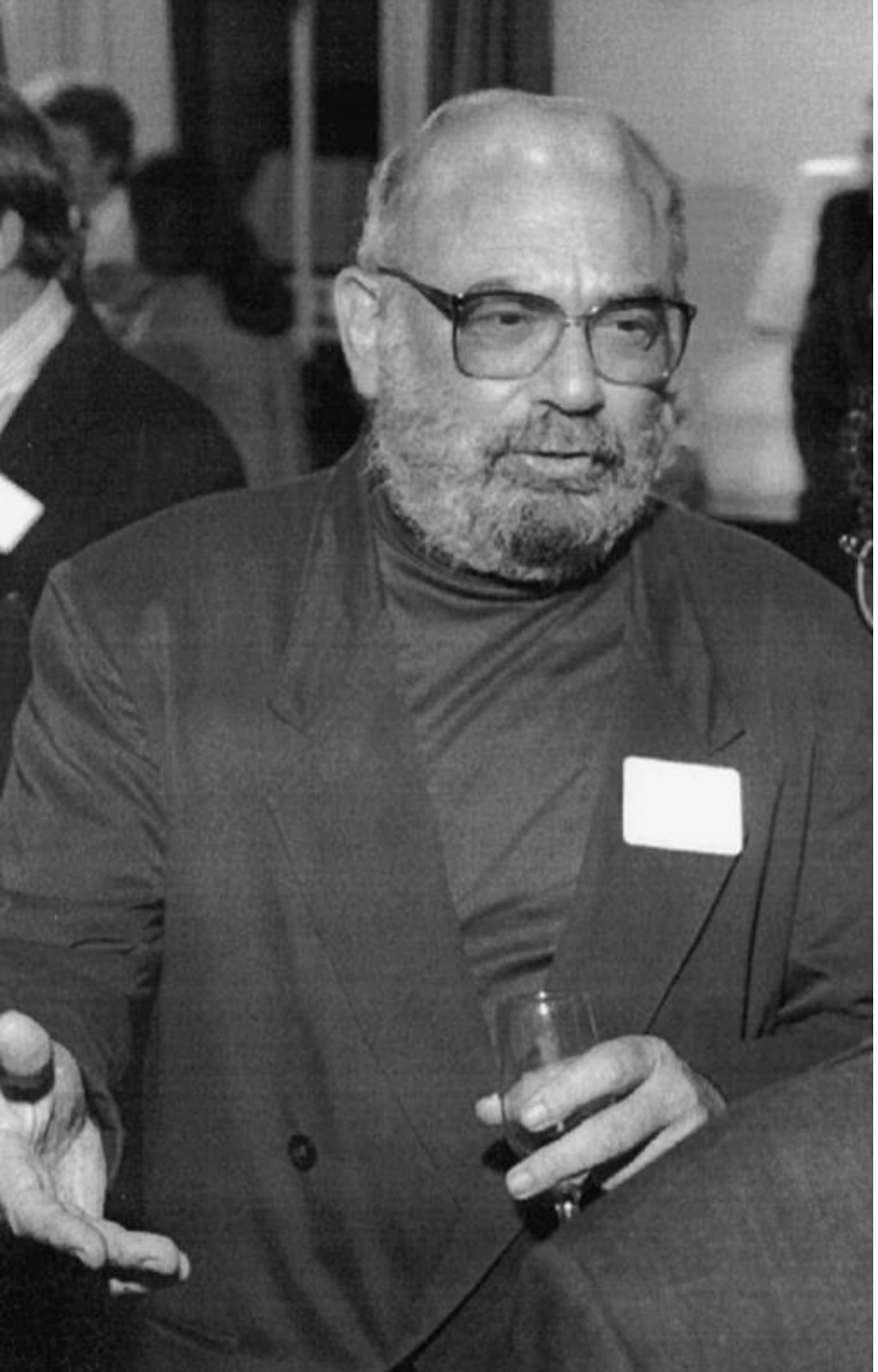
More than anything else,
science progresses by trial and error.

thanks for listening

for data science questions

phil.henrickson@aebs.com

on data science and statistics



Statistical Science
2001, Vol. 16, No. 3, 199–231

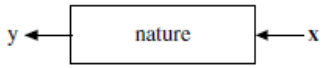
Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

1. INTRODUCTION

Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables **x** (independent variables) go in one side, and on the other side the response variables **y** come out. Inside the black box, nature functions to associate the predictor variables with the response variables, so the picture is like this:



There are two goals in analyzing the data:

Prediction. To be able to predict what the responses are going to be to future input variables;
Information. To extract some information about how nature is associating the response variables to the input variables.

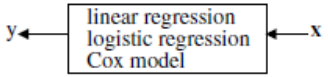
There are two different approaches toward these goals:

The Data Modeling Culture

The analysis in this culture starts with assuming a stochastic data model for the inside of the black box. For example, a common data model is that data are generated by independent draws from

response variables = f(predictor variables, random noise, parameters)

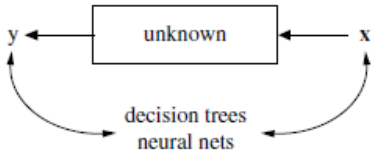
The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:



Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.

The Algorithmic Modeling Culture

The analysis in this culture considers the inside of the box complex and unknown. Their approach is to find a function $f(\mathbf{x})$—an algorithm that operates on **x** to predict the responses **y**. Their black box looks like this:



Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

statistical modeling vs algorithmic modeling

goal: inference

assumed

goal: prediction

learned

neural nets
decision trees
support vector machines

inference: what is the effect of x on y?

x: television

y: protests

Figure 2. East German counties with and without WGTV based on Longley-Rice model and Dresden cutoff level

inference: what is the effect of x on y?

x: television

Figure 2. East German counties with and without WGTV based on Longley-Rice model and Dresden cutoff level

y: protests

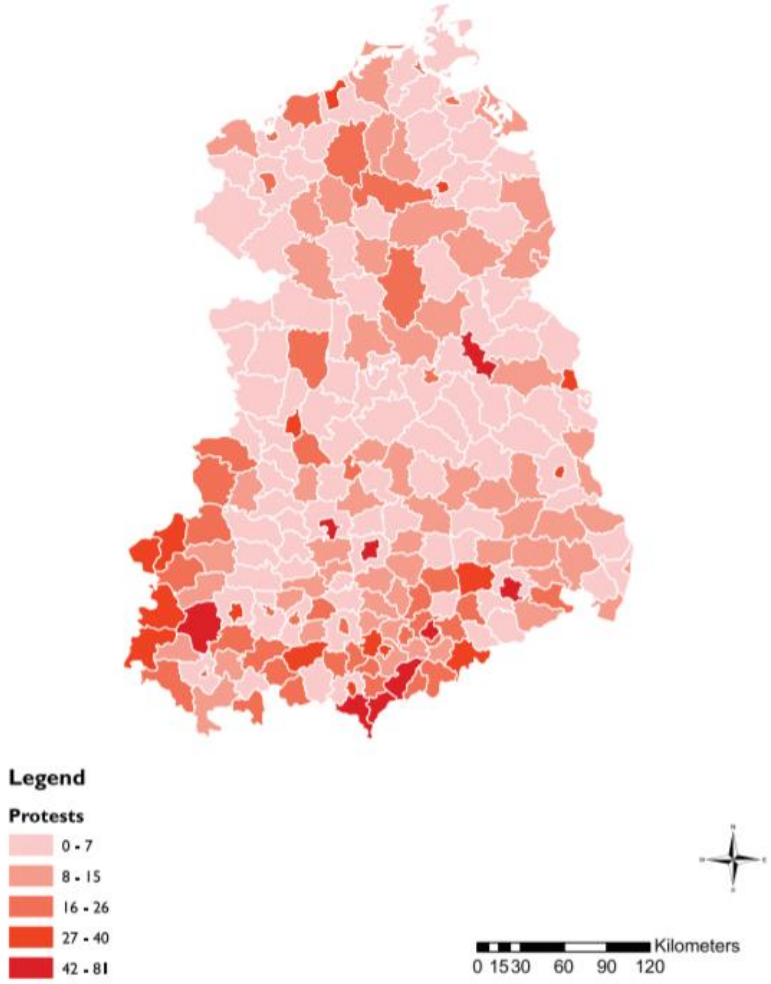
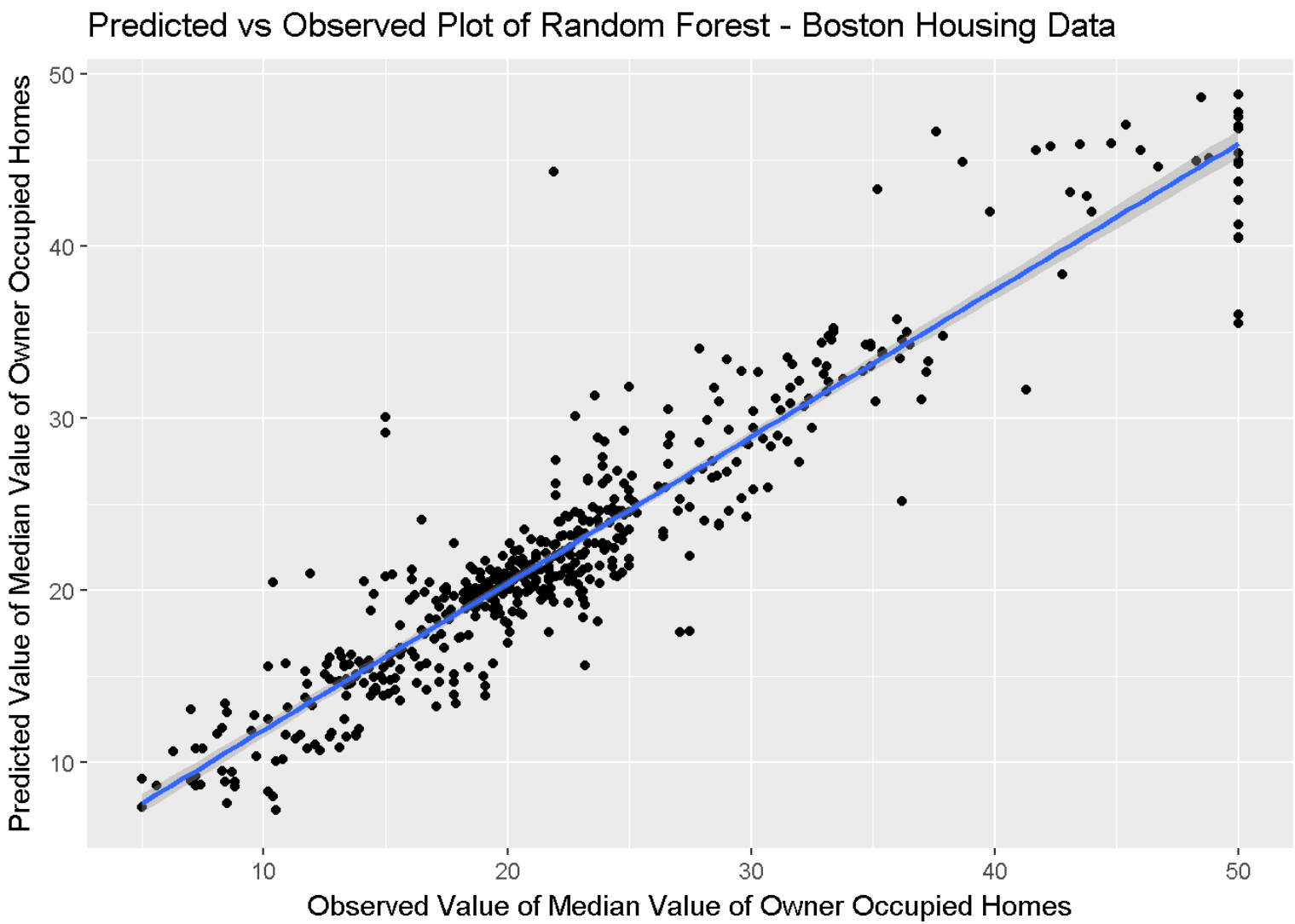
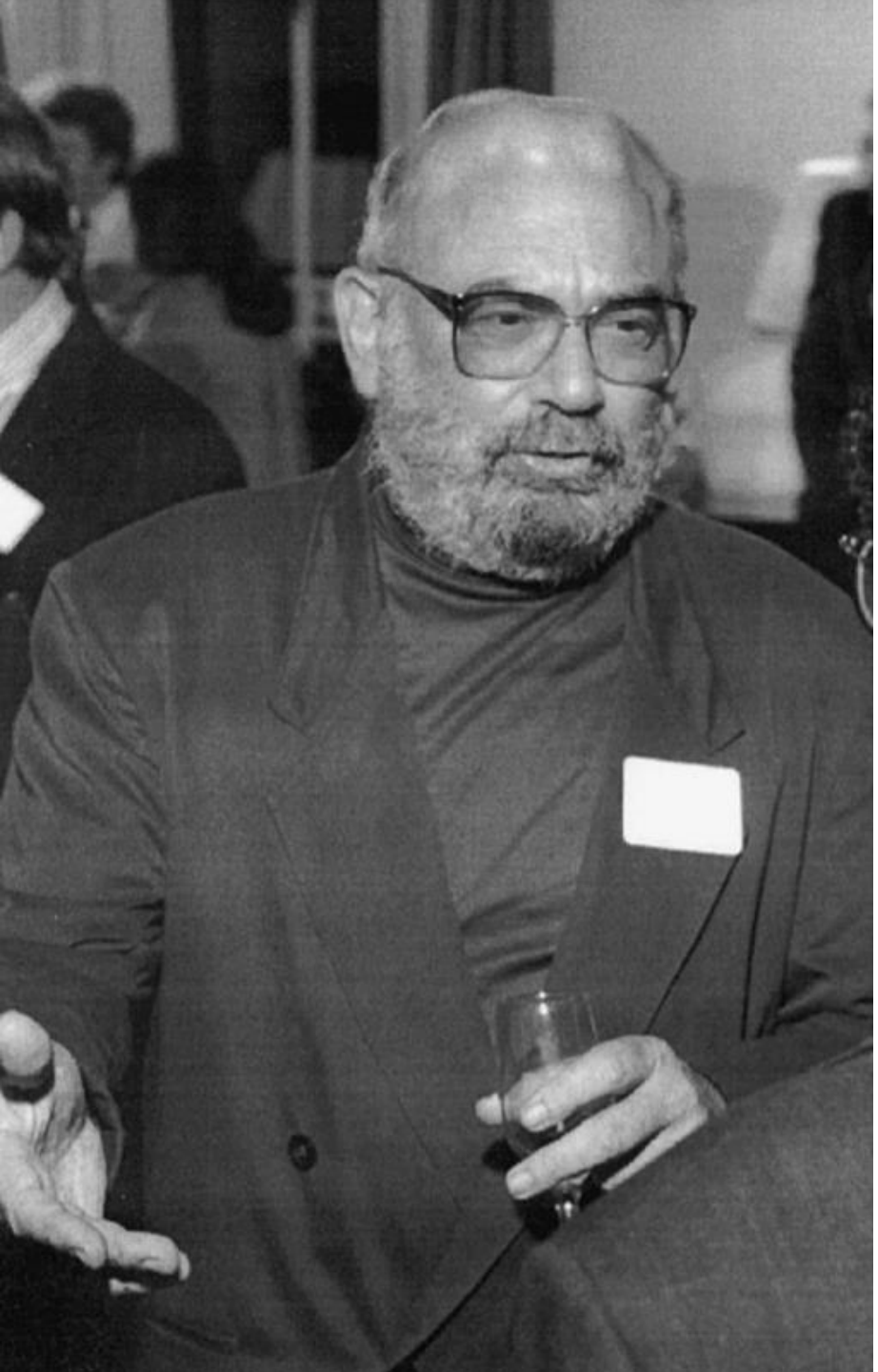


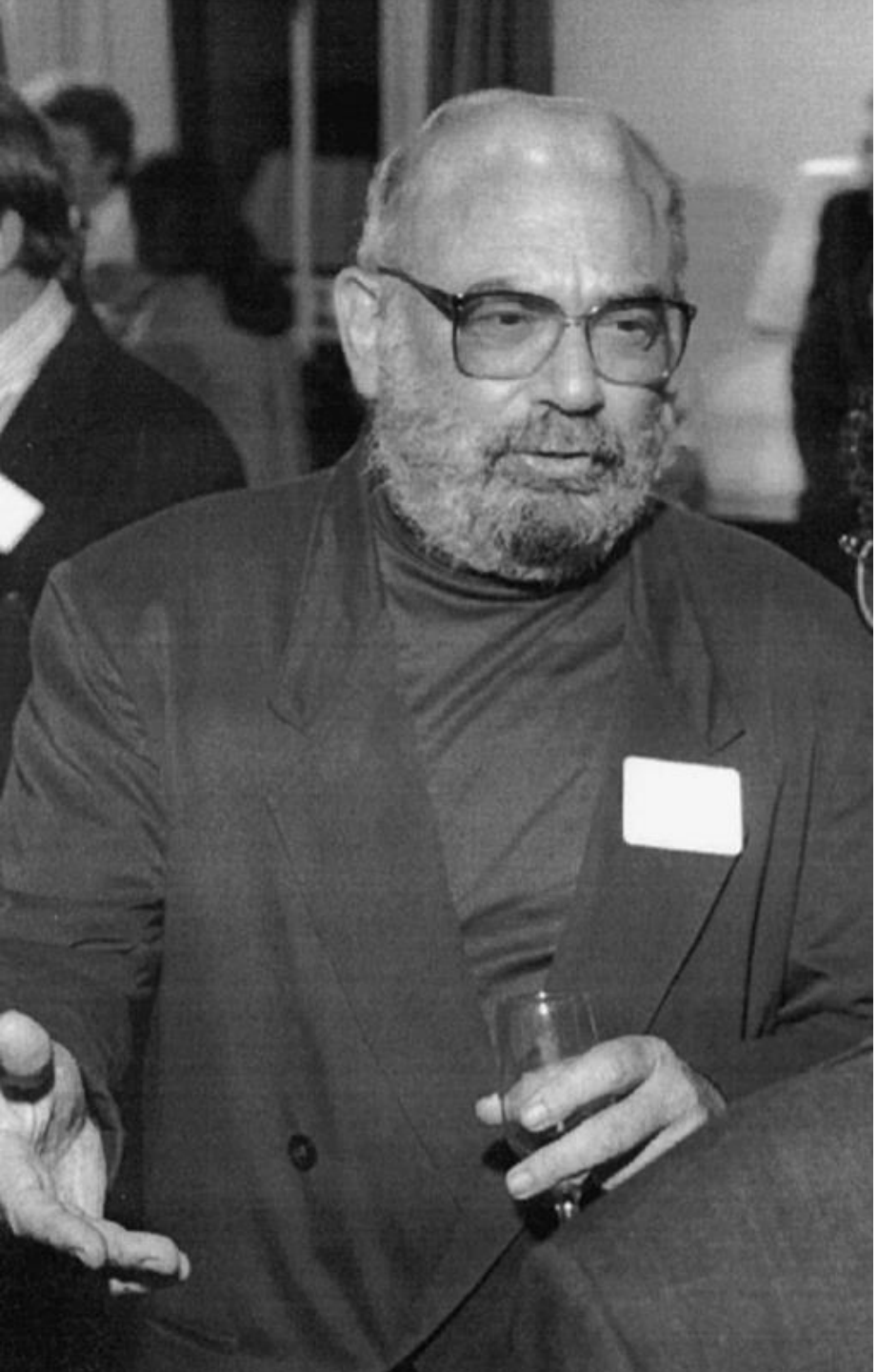
Figure 3. Number of protests per county, East Germany, 4 September 1989–18 March 1990

prediction: how well can the model predict y?





The goals in statistics are to use data to **predict** and to get information about the **underlying data mechanism**. Nowhere is it written on a stone tablet what kind of model should be used to solve problems involving data.



The goals in statistics are to use data to predict and to get information about the underlying data mechanism. Nowhere is it written on a stone tablet what kind of model should be used to solve problems involving data.

The best solution could be an algorithmic model, or maybe a [statistical] model, or maybe a combination. But the trick to being a scientist is to be open to using a wide variety of tools.

all of machine learning in one expression

all of machine learning in one expression

$$\hat{\theta} = argmin_{\theta} \Sigma_i L(y_i - f(x_i, \theta)) + \lambda\gamma(\theta)$$

all of machine learning in one expression

$$\hat{\theta} = argmin_{\theta} \Sigma_i \underline{L(y_i - f(x_i, \theta))} + \underline{\lambda\gamma(\theta)}$$

parameter vector; how X relates to y

predictive loss; actual vs predicted

modeling strategy; trees, nearest-neighbor

strength and type of regularization

Some Machine Learning Algorithms

Function class $\mathcal{F}$ (and its parametrization)	*Regularizer $R(f)$*
Global/parametric predictors	
Linear $\beta' x$ (and generalizations)	Subset selection $\|\|\beta\|\|_0 = \sum_{j=1}^{k} 1_{\beta_j \neq 0}$
	LASSO $\|\|\beta\|\|_1 = \sum_{j=1}^{k} \|\beta_j\|$
	Ridge $\|\|\beta\|\|_2^2 = \sum_{j=1}^{k} \beta_j^2$
	Elastic net $\alpha \|\|\beta\|\|_1 + (1-\alpha) \|\|\beta\|\|_2^2$
Local/nonparametric predictors	
Decision/regression trees	Depth, number of nodes/leaves, minimal leaf size, information gain at splits
Random forest (linear combination of trees)	Number of trees, number of variables used in each tree, size of bootstrap sample, complexity of trees (see above)
Nearest neighbors	Number of neighbors
Kernel regression	Kernel bandwidth
Mixed predictors	
Deep learning, neural nets, convolutional neural networks	Number of levels, number of neurons per level, connectivity between neurons
Splines	Number of knots, order
Combined predictors	
Bagging: unweighted average of predictors from bootstrap draws	Number of draws, size of bootstrap samples (and individual regularization parameters)
Boosting: linear combination of predictions of residual	Learning rate, number of iterations (and individual regularization parameters)
Ensemble: weighted combination of different predictors	Ensemble weights (and individual regularization parameters)