DATA SCIENCE FOR SECURITY

BENNETT KLEINBERG

25 OCT 2019

UCL

TODAY

- The big promise: A primer of data science
- The pitfalls and problems
- Data Science for security

(if we have time: The do-or-die problem of data science)

YOUR THOUGHTS?

1. More data = better problem solving.
2. Every problem will become a “data problem”.
3. The big challenge for data science is a technological one.

AAAH: SO WE’RE TALKING BIG DATA!

PROBLEMS WITH “BIG DATA”

- what is “big”?
- data = data?
- complexity of data?
- sexiness of small data

BEFORE WE START

What are the chances that this man is a terrorist?

Problem 1: A secret government agency has developed a scanner which determines whether a person is a terrorist. The scanner is fairly reliable; 95% of all scanned terrorists are identified as terrorists, and 95% of all upstanding citizens are identified as such. An informant tells the agency that exactly one passenger of 100 aboard an aeroplane in which you are seated is a terrorist. The agency decide to scan each passenger and the shifty looking man sitting next to you is tested as "TERRORIST". What are the chances that this man *is* a terrorist? Show your work!

THE BIG PROMISE: A PRIMER OF DATA SCIENCE

MACHINE LEARNING?

- core idea: a system learns from experience
- no precise instructions

Applications?

WHY DO WE WANT THIS?

Step back…

How did you perform regression analysis in school?

OKAY …

- you've got one outcome variable (e.g. number of shooting victims)
- and two predictors (e.g. gender of shooter, age)
- typical approach $victims = gender + age$
- regression equation with: intercept, beta coefficients and inferred error term

BUT!

Often we have no idea about the relationships.

- too many predictors
- too diverse a problem
- simply unknown

ML IN GENERAL

- concered with patterns in data
- learning from data
- more experience results typically in better models
- data, data, data

TYPES OF MACHINE LEARNING

BROAD CATEGORIES

- **Supervised learning**
- **Unsupervised learning**
- Hybrid models
- Deep learning
- Reinforcement learning

DEEP LEARNING

Inspired by the human brain.

- MIT's course website https://deeplearning.mit.edu/
- Lex Fridman's courses from MIT –> YouTube

REINFORCEMENT LEARNING

- Excellent YouTube examples from code bullet
- e.g. AI Learns to play the Worlds Hardest Game

Demo

SUPERVISED LEARNING

WHAT IS SUPERVISED?

- who is the supervisor?
- supervised = labelled data
- i.e. you know the outcome
- flipped logic

Contrary: unsupervised.

CLASSES OF SUPERVISED LEARNING

- classification (e.g. death/alive, fake/real)
- regression (e.g. income, number of deaths)

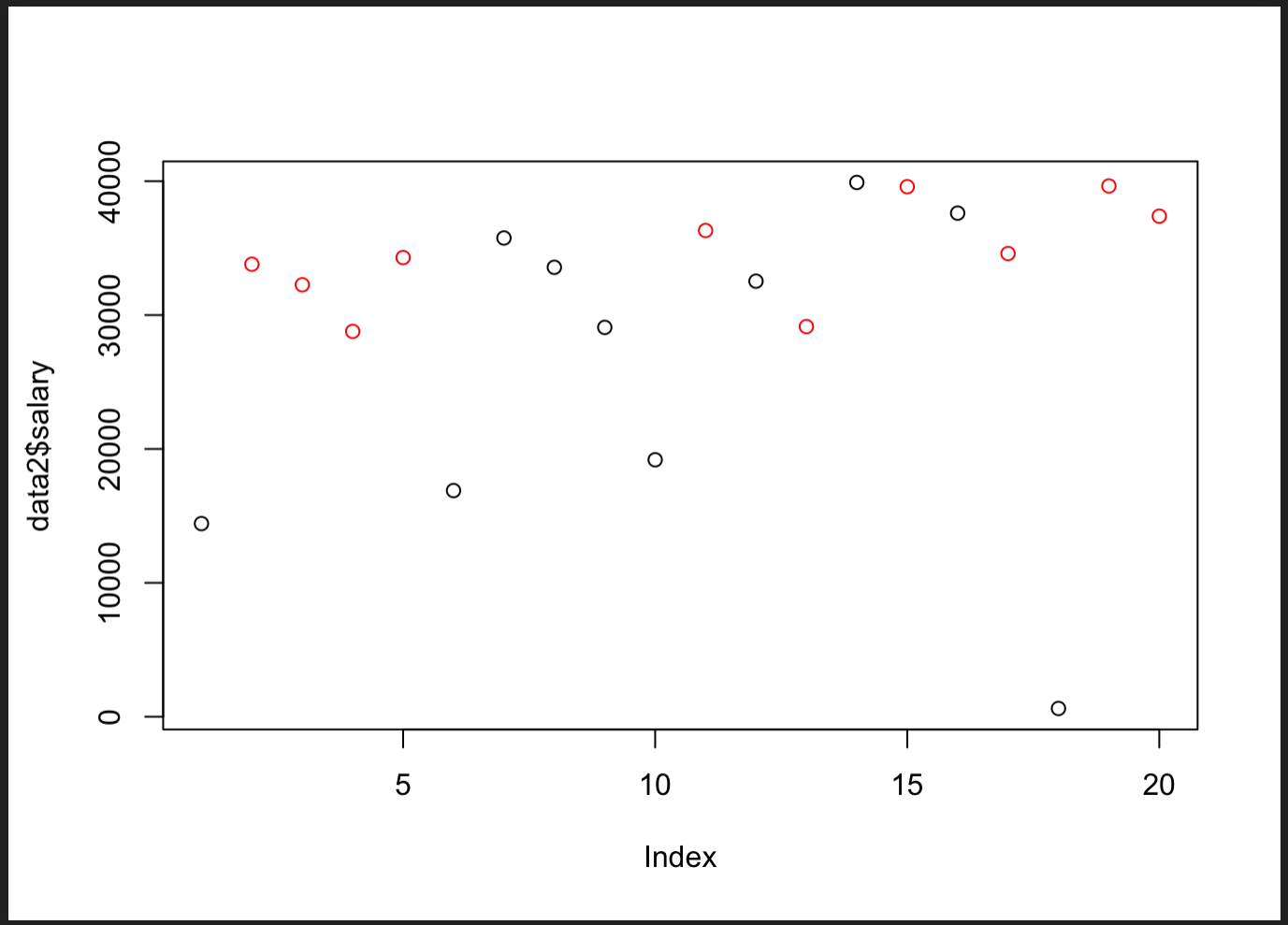
MINI EXAMPLE

Supervised classification

SIMPLE EXAMPLE

- gender prediction
- based on salary

gender	salary
male	33796
male	34597
male	34296
male	32262
female	19190
female	14424
female	37614
female	29079

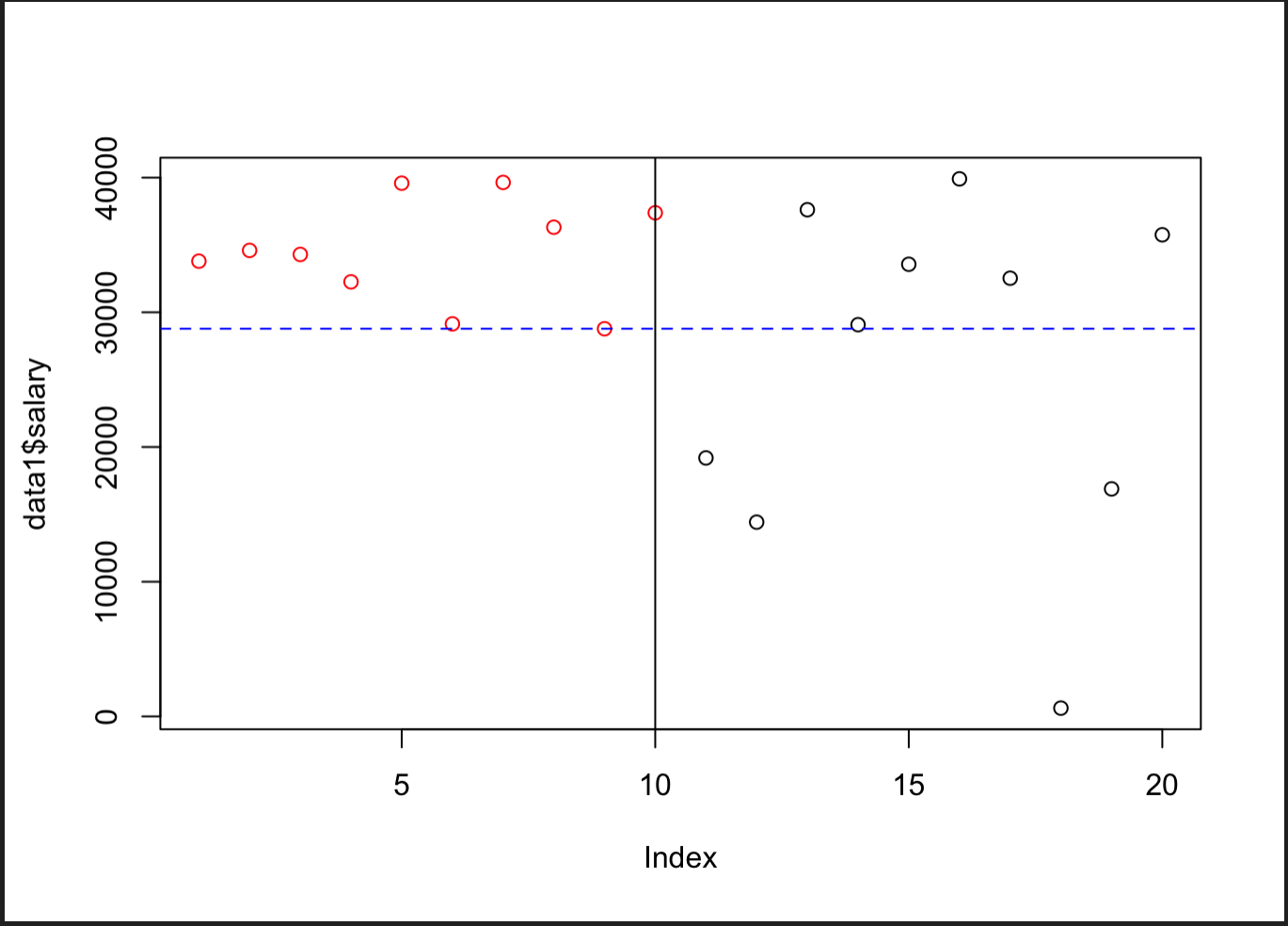

data2$salary
0
10000
20000
30000
40000
5
10
15
20
Index

How to best separate the data into two groups?

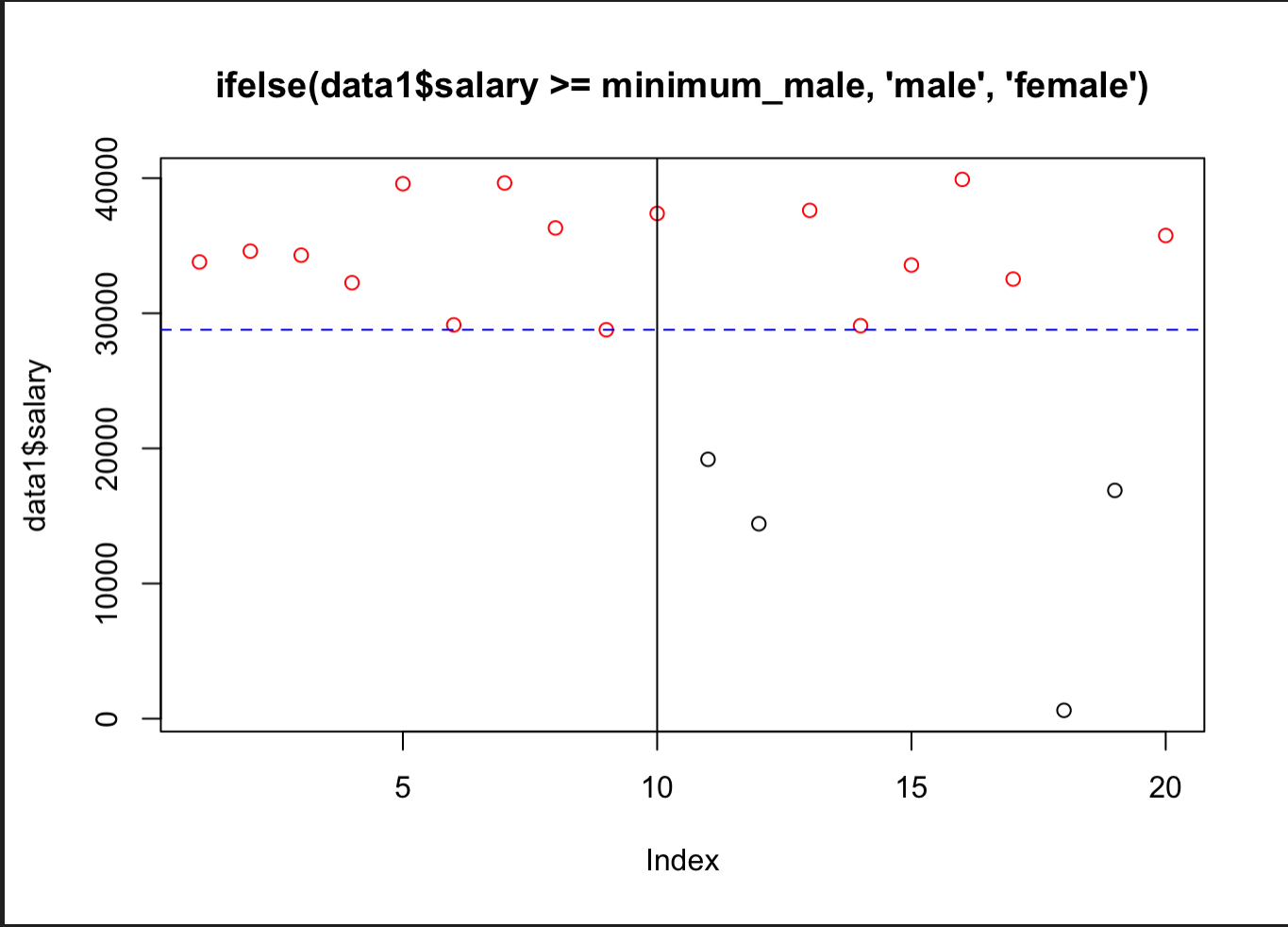
CORE IDEA

- learn relationship between
  - outcome (target) variable
  - features (predictors)
- “learning” is done through an algorithm
  - simplest algorithm: `if A then B`

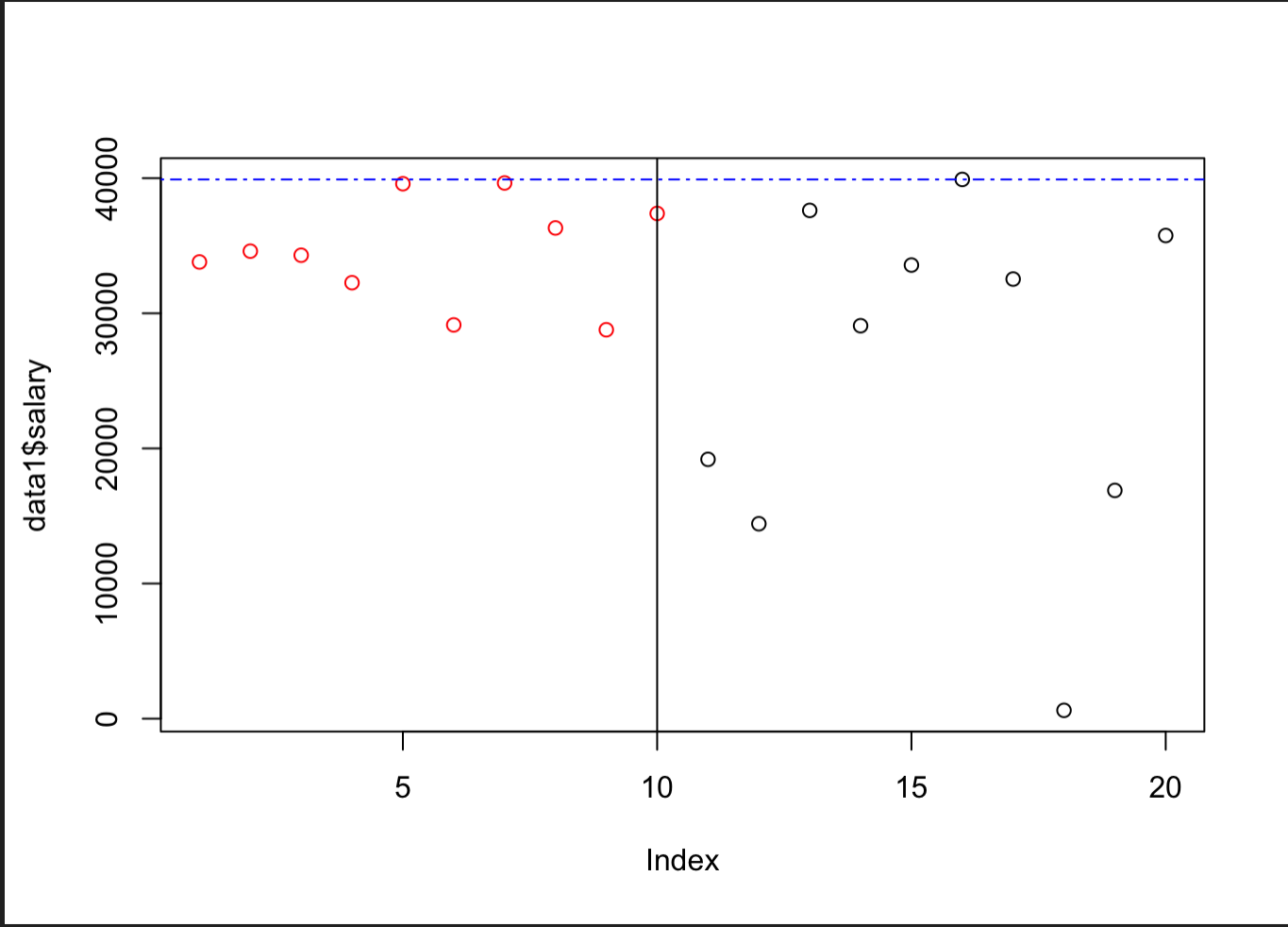
IDEA 1: MALE SALARY THRESHOLD



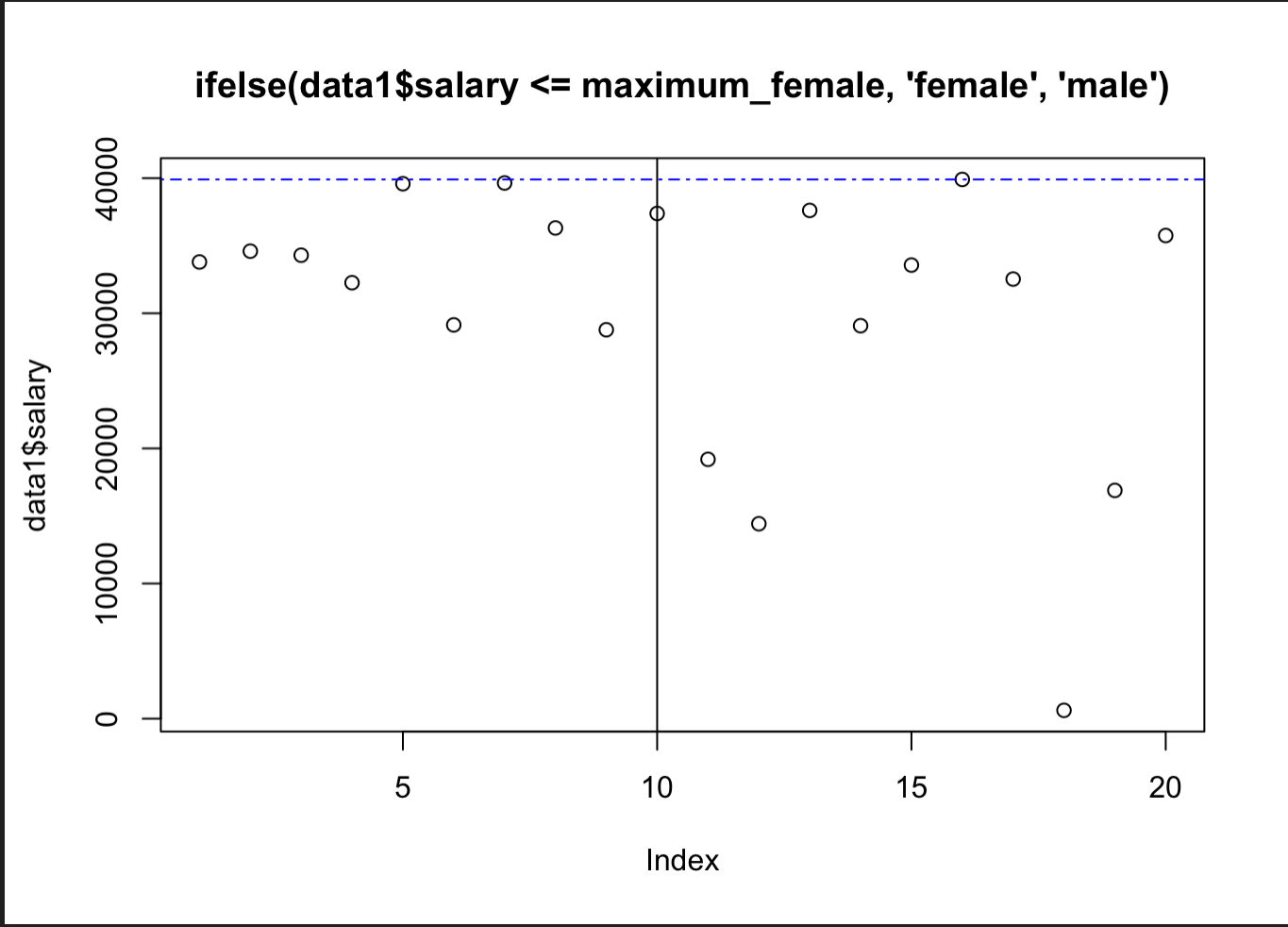
IDEA 1: MALE SALARY THRESHOLD



IDEA 2: FEMALE SALARY THRESHOLD



IDEA 2: FEMALE SALARY THRESHOLD



But this is not learning!

STEPWISE SUPERVISED ML

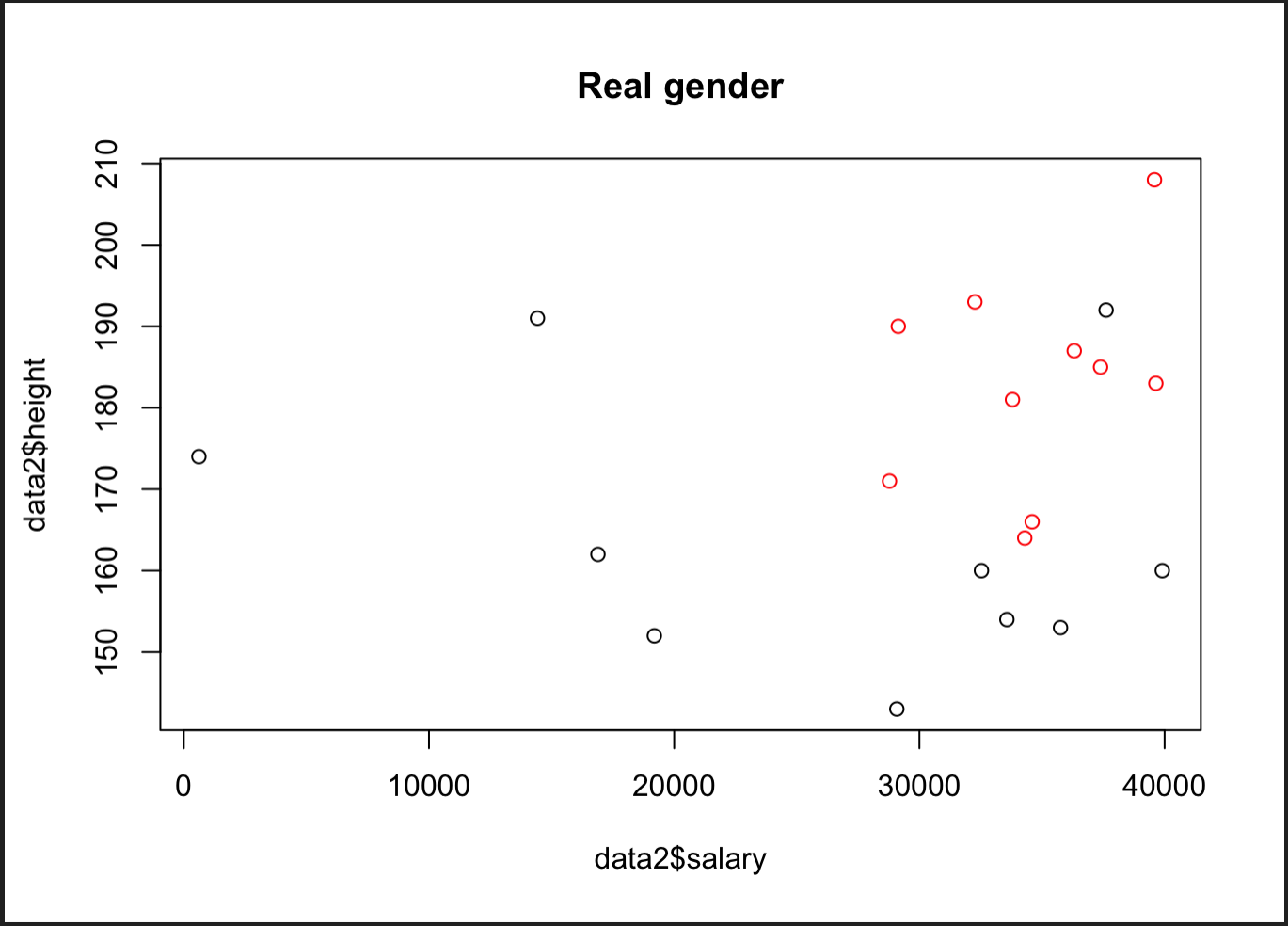
- clarify what `outcome` and `features` are
- determine which classification algorithm to use
- train the model

ENTER: `CARET`

```
library(caret)
```

- excellent package for ML in R
- well-documented website
- common interface for 200+ models

CARET IN PRACTICE



`CARET` IN PRACTICE

```
my_first_model = train(gender ~ .
                       , data = data2
                       , method = "svmLinear"
                       )
```

Now you have trained a model!

= you have taught an algorithm to learn to predict gender from salary & height

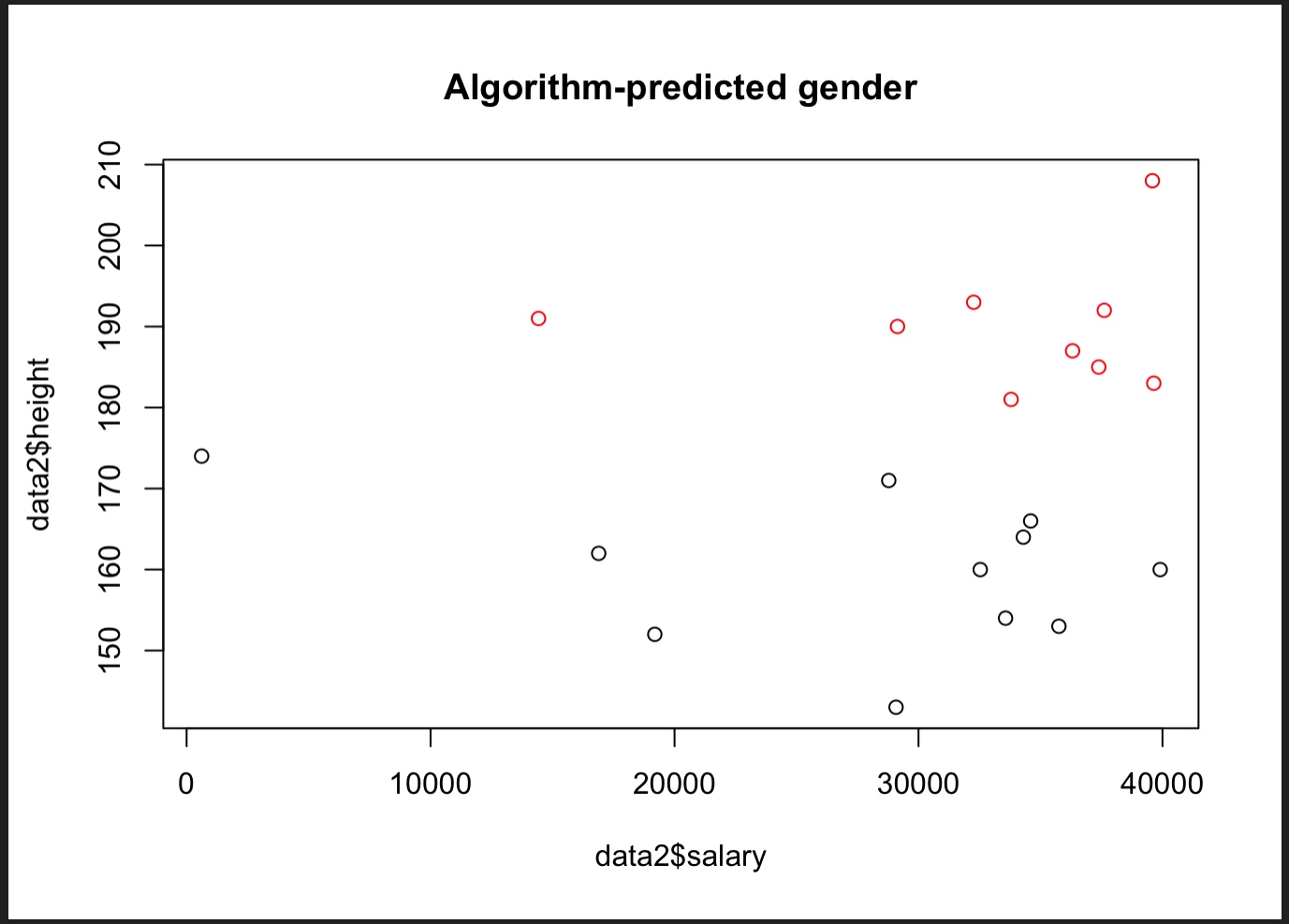
But now what?

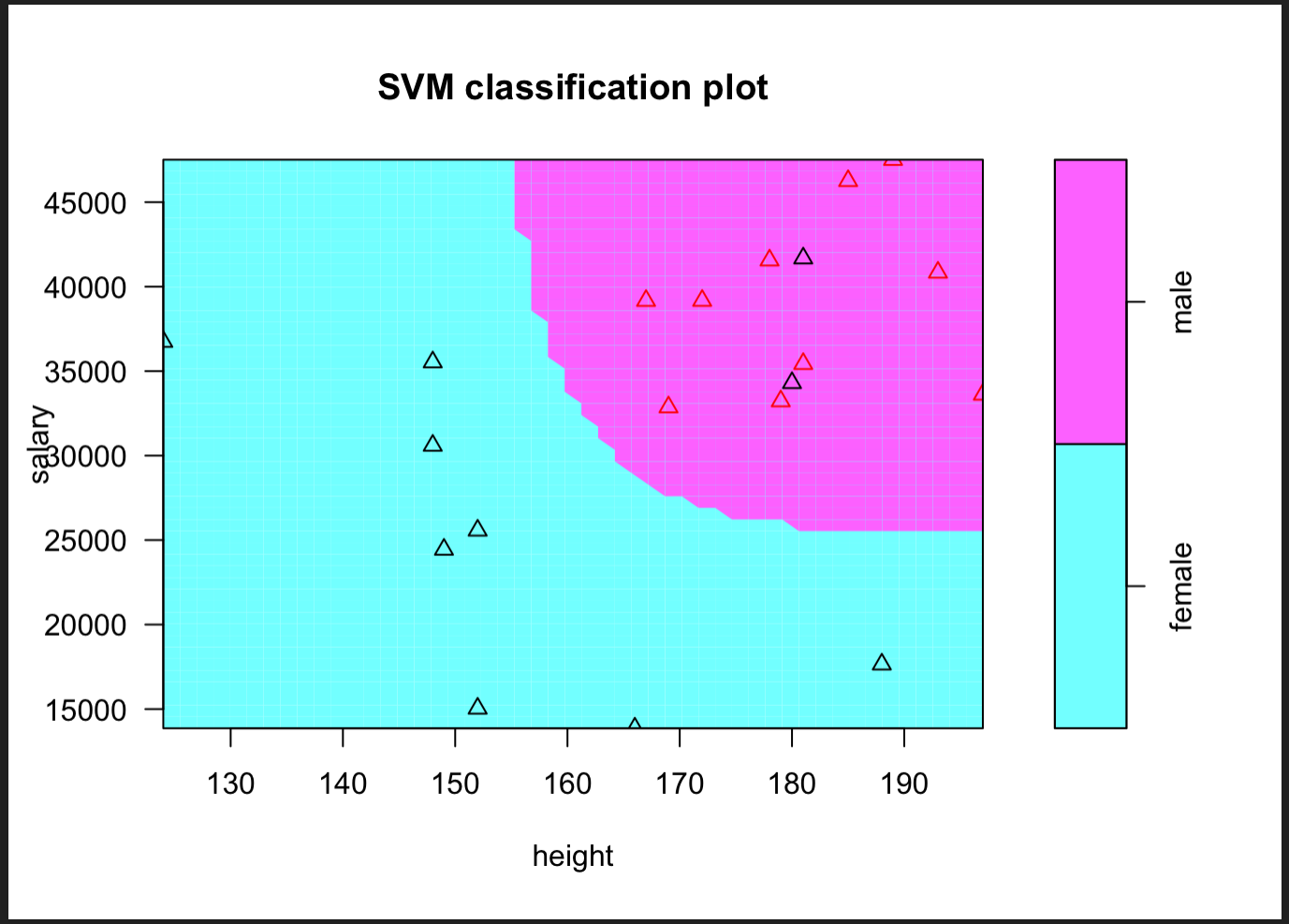
PUT YOUR MODEL TO USE

Make predictions:

```
data2$model_predictions = predict(my_first_model, data2)
```

	female	male
female	8	2
male	3	7


Algorithm-predicted gender
data2$height
150
160
170
180
190
200
210
0
10000
20000
30000
40000
data2$salary


SVM classification plot
45000
40000
35000
30000
25000
20000
15000
salary
130
140
150
160
170
180
190
height
male
female

THE KEY CHALLENGE?

Think about what we did…

PROBLEM OF INDUCTIVE BIAS

- remember: we learn from the data
- but what we really want to know is: how does it work on "**unseen**" data

How to solve this?

KEEP SOME DATA FOR YOURSELF

Train/test split

- split the data (e.g. 80%/20%, 60%/40%)
- use one part as TRAINING SET
- use the other as TEST SET

```
training_data = data2[ in_training,]
test_data = data2[-in_training,]
```

	gender	salary	height
3	male	33225	179
9	male	40841	193
11	female	15039	152
20	female	30597	148

PIPELINE AGAIN

- define outcome (DONE)
- define features (DONE)
- build model (DONE)
  - but this time: on the TRAINING SET
- evaluate model
  - this time: on the TEST SET

Teach the SVM:

```
my_second_model = train(gender ~ .
                        , data = training_data
                        , method = "svmLinear"
                        )
```

Fit/test the SVM:

```
model_predictions = predict(my_second_model, test_data)
```

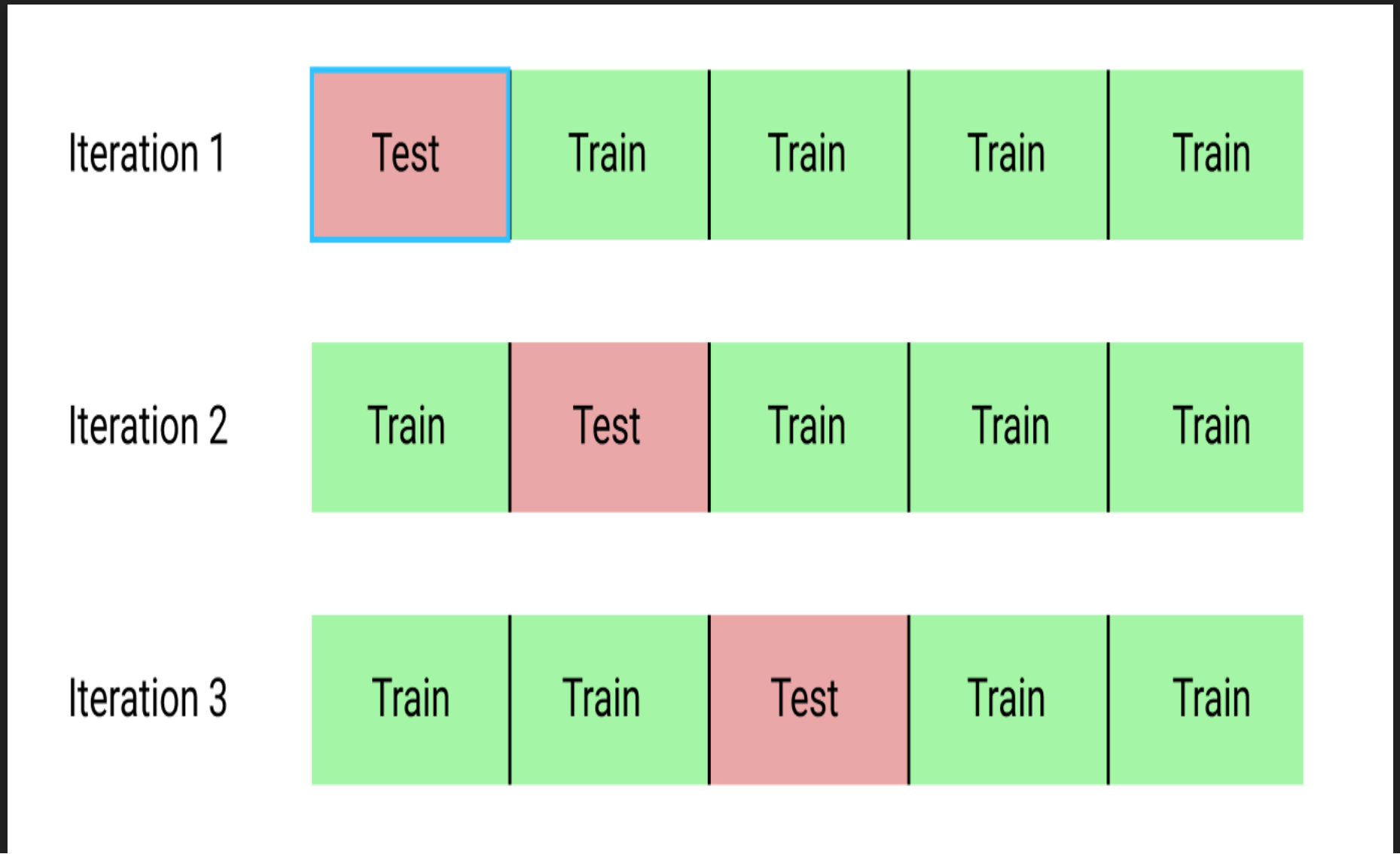
	female	male
female	2	0
male	0	2

BUT!

- our model might be really dependent on the training data
- we want to be more careful
- Can we do some kind of safeguarding in the training data?

CROSS-VALIDATION

K-fold cross-validation




Iteration 4
Train
Train
Train
Test
Train
Iteration 5
Train
Train
Train
Train
Test

Img source

HOW DO WE KNOW WHETHER A MODEL IS GOOD?

MODEL PERFORMANCE METRICS

Example: 400 fake/real news articles:

		Prediction	
		Fake	Real
Reality	Fake	159	41
	Real	42	158

How do you evaluate that model?

(159+158)/400 = 0.73

INTERMEZZO

THE CONFUSION MATRIX

CONFUSION MATRIX

	Fake	**Real**
Fake	True positives	False negatives
Real	False positives	True negatives

CONFUSION MATRIX

- true positives (TP): correctly identified fake ones
- true negatives (TN): correctly identified real ones
- false positives (FP): false accusations
- false negatives (FN): missed fakes

OKAY: LET'S USE ACCURACIES

$$acc = \frac{(TP+TN)}{N}$$

Any problems with that?

ACCURACY

Model 1

	Fake	**Real**
Fake	252	48
Real	80	220

Model 2

	Fake	**Real**
Fake	290	10
Real	118	182

PROBLEM WITH ACCURACY

- same accuracy, different confusion matrix
- relies on thresholding idea
- not suitable for comparing models (don’t be fooled by the literature!!)

Needed: more nuanced metrics

THE PROBLEM FROM THE BEGINNING:

What are the chances that this man is a terrorist?

Problem 1: A secret government agency has developed a scanner which determines whether a person is a terrorist. The scanner is fairly reliable; 95% of all scanned terrorists are identified as terrorists, and 95% of all upstanding citizens are identified as such. An informant tells the agency that exactly one passenger of 100 aboard an aeroplane in which you are seated is a terrorist. The agency decide to scan each passenger and the shifty looking man sitting next to you is tested as "TERRORIST". What are the chances that this man *is* a terrorist? Show your work!

FORMALISING THE PROBLEM

Probability of TERRORIST **given** that there is an ALARM

Looking for: `P(terrorist GIVEN alarm)`

Formal: `P(terrorist|alarm)`

SOLVING THE PROBLEM

	Terrorist	Passenger	
Terrorist	950	50	1,000
Passenger	4,950	94,050	99,000
	5,900	94,100	100,000

```
P(terrorist|alarm) = 950/5900 = 16.10%
```

BEYOND ACCURACY

```
##         prediction
## reality Fake Real Sum
##    Fake  252   48 300
##    Real   80  220 300
##    Sum   332  268 600
```

```
##         prediction
## reality Fake Real Sum
##    Fake  290   10 300
##    Real  118  182 300
##    Sum   408  192 600
```

PRECISION

How often is the prediction correct when predicting class *X*?

Note: we have two classes, so we get *two* precision values

Formally:

- $Pr_{fake} = \frac{TP}{(TP+FP)}$
- $Pr_{real} = \frac{TN}{(TN+FN)}$

PRECISION

```
##         prediction
## reality Fake Real Sum
##    Fake  252   48 300
##    Real   80  220 300
##    Sum   332  268 600
```

- $Pr_{fake} = \frac{252}{332} = 0.76$
- $Pr_{real} = \frac{220}{268} = 0.82$

COMPARING THE MODELS

	Model 1	Model 2
acc	0.79	0.79
Pr_{fake}	0.76	0.71
Pr_{real}	0.82	0.95

RECALL

How many cases of class *X* are detected?

Note: we have two classes, so we get *two* recall values

Also called sensitivity and specificity!

Formally:

- $R_{fake} = \frac{TP}{(TP+FN)}$
- $R_{real} = \frac{TN}{(TN+FP)}$

RECALL

```
##          prediction
## reality Fake Real Sum
##    Fake  252   48 300
##    Real   80  220 300
##    Sum   332  268 600
```

- $R_{fake} = \frac{252}{300} = 0.84$
- $R_{real} = \frac{220}{300} = 0.73$

COMPARING THE MODELS

	Model 1	Model 2
acc	0.79	0.79
Pr_{fake}	0.76	0.71
Pr_{real}	0.82	0.95
R_{fake}	0.84	0.97
R_{real}	0.73	0.61

COMBINING PR AND R

The *F1* measure.

Note: we combine Pr and R for each class, so we get *two* F1 measures.

Formally:

- $F1_{fake} = 2 * \frac{Pr_{fake} * R_{fake}}{Pr_{fake} + R_{fake}}$
- $F1_{real} = 2 * \frac{Pr_{real} * R_{real}}{Pr_{real} + R_{real}}$

F1 MEASURE

```
##         prediction
## reality Fake Real Sum
##    Fake  252   48 300
##    Real   80  220 300
##    Sum   332  268 600
```

- $F1_{fake} = 2 * \frac{0.76 * 0.84}{0.76 + 0.84} = 2 * \frac{0.64}{1.60} = 0.80$
- $F1_{real} = 2 * \frac{0.82 * 0.73}{0.82 + 0.73} = 0.78$

COMPARING THE MODELS

	Model 1	Model 2
acc	0.79	0.79
Pr_{fake}	0.76	0.71
Pr_{real}	0.82	0.95
R_{fake}	0.84	0.97
R_{real}	0.73	0.61
$F1_{fake}$	0.80	0.82
$F1_{real}$	0.78	0.74

UNSUPERVISED LEARNING

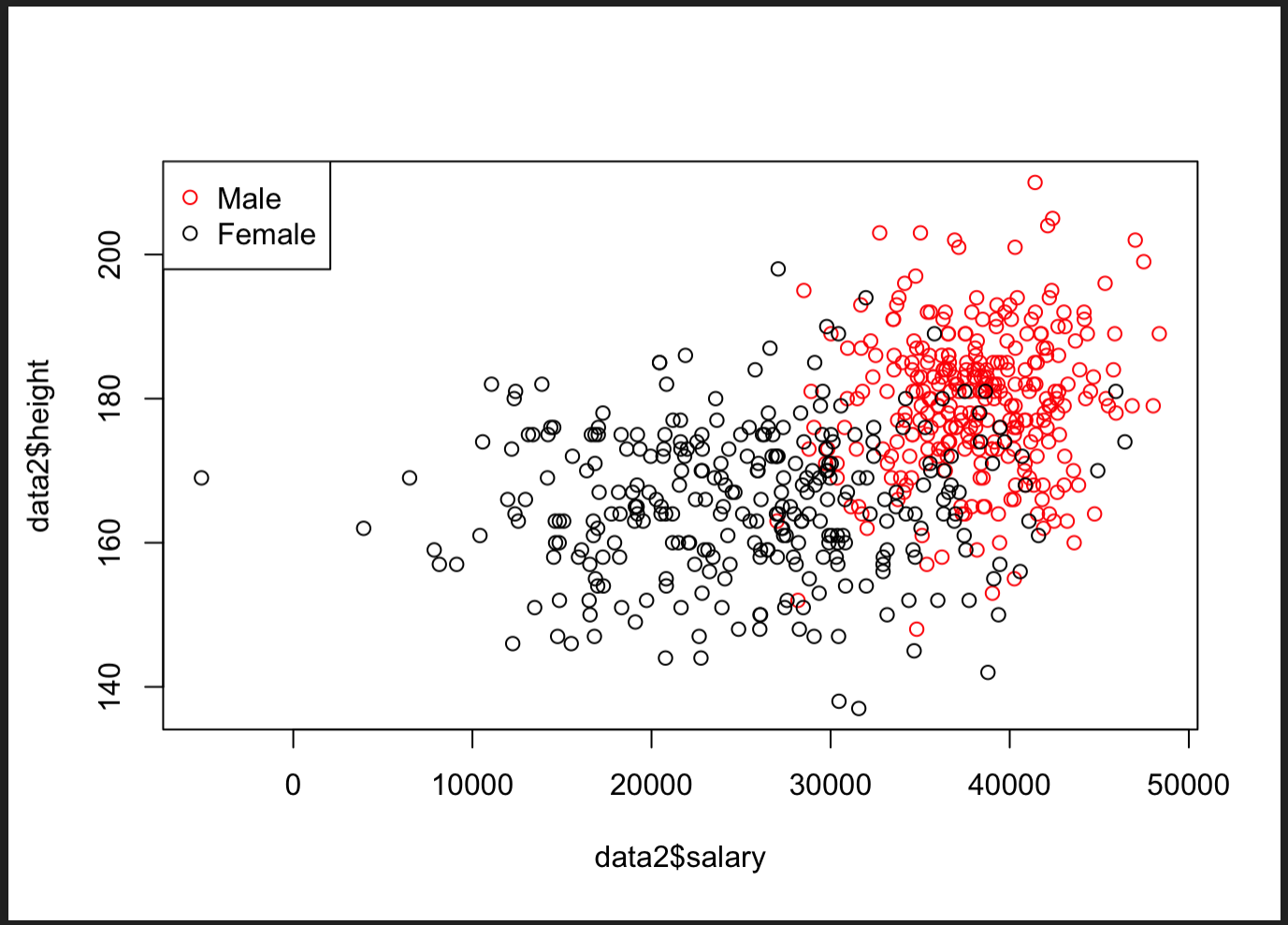
- often we don’t have labelled data
- sometimes there are no labels at all
- core idea: finding clusters in the data

EXAMPLES

- grouping of online ads
- clusters in crime descriptions
- …

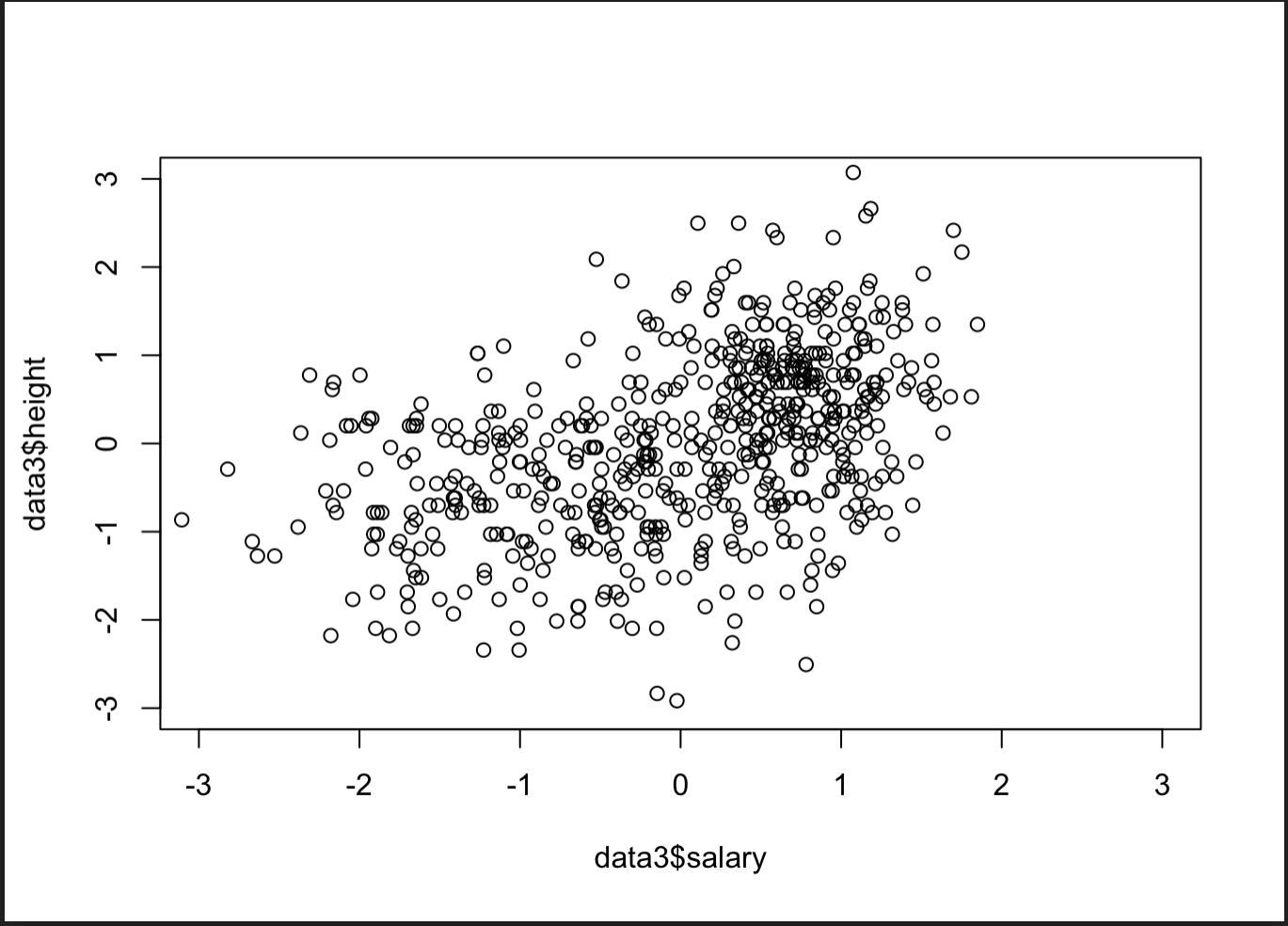
Practically everywhere.

Clustering reduces your data!


Male
Female
data2$height
140
160
180
200
0
10000
20000
30000
40000
50000
data2$salary

THE UNSUPERVISED CASE

You know nothing about groups inherent to the data.



THE K-MEANS IDEA

- separate data in set number of clusters
- find best cluster assignment of observations

STEPWISE

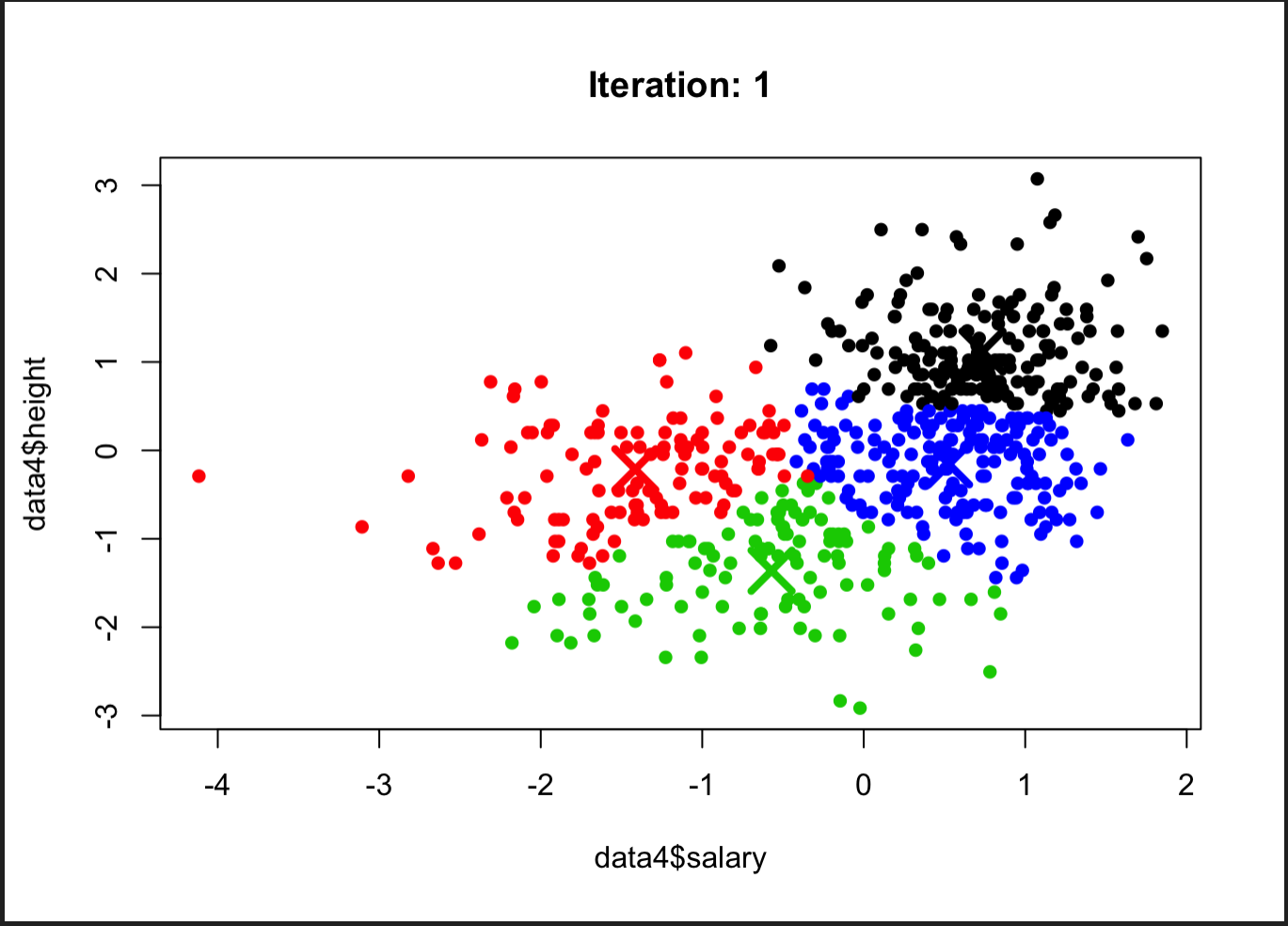
1. set the number of clusters
2. find best cluster assignment

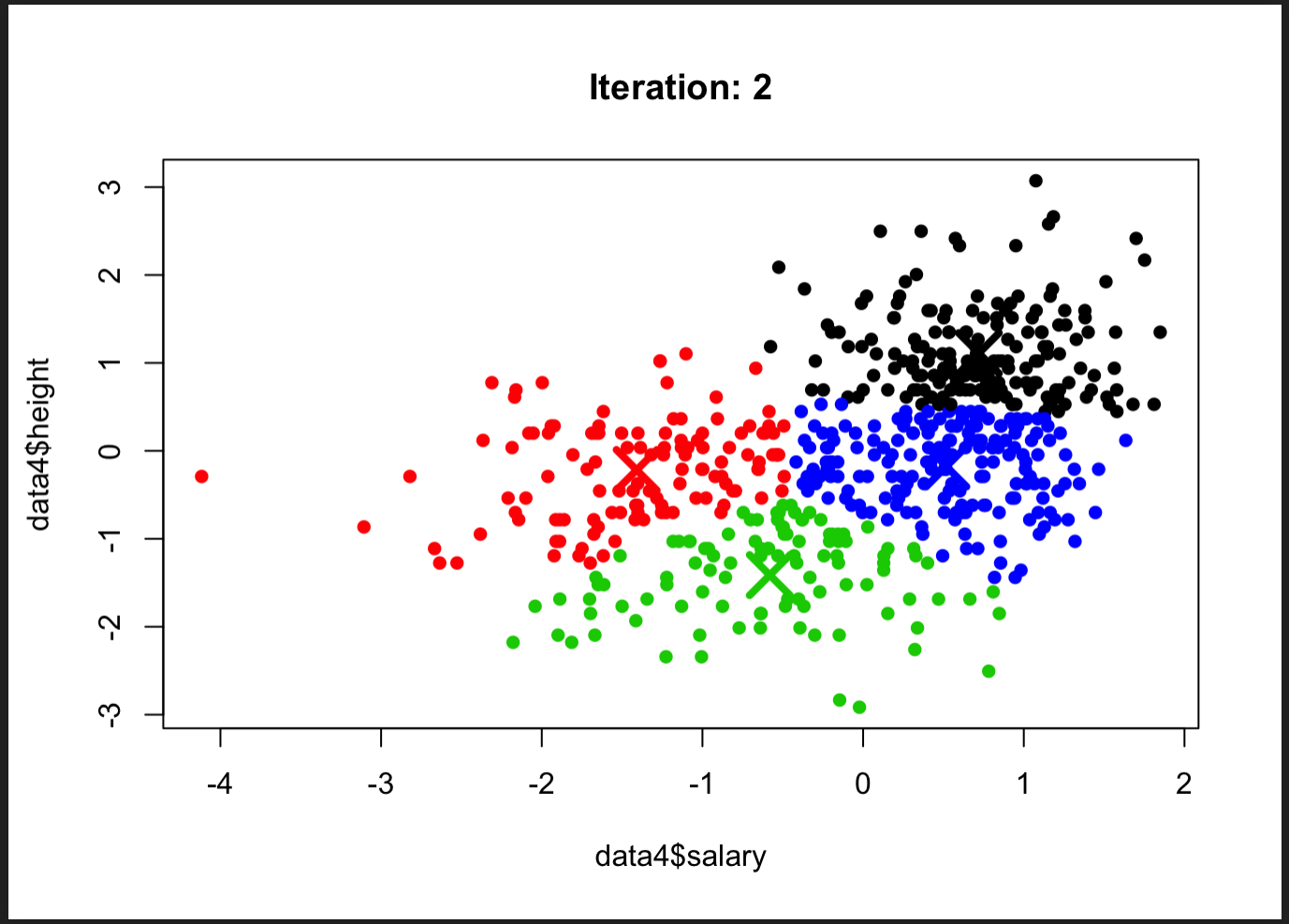
1. NO. OF CLUSTERS

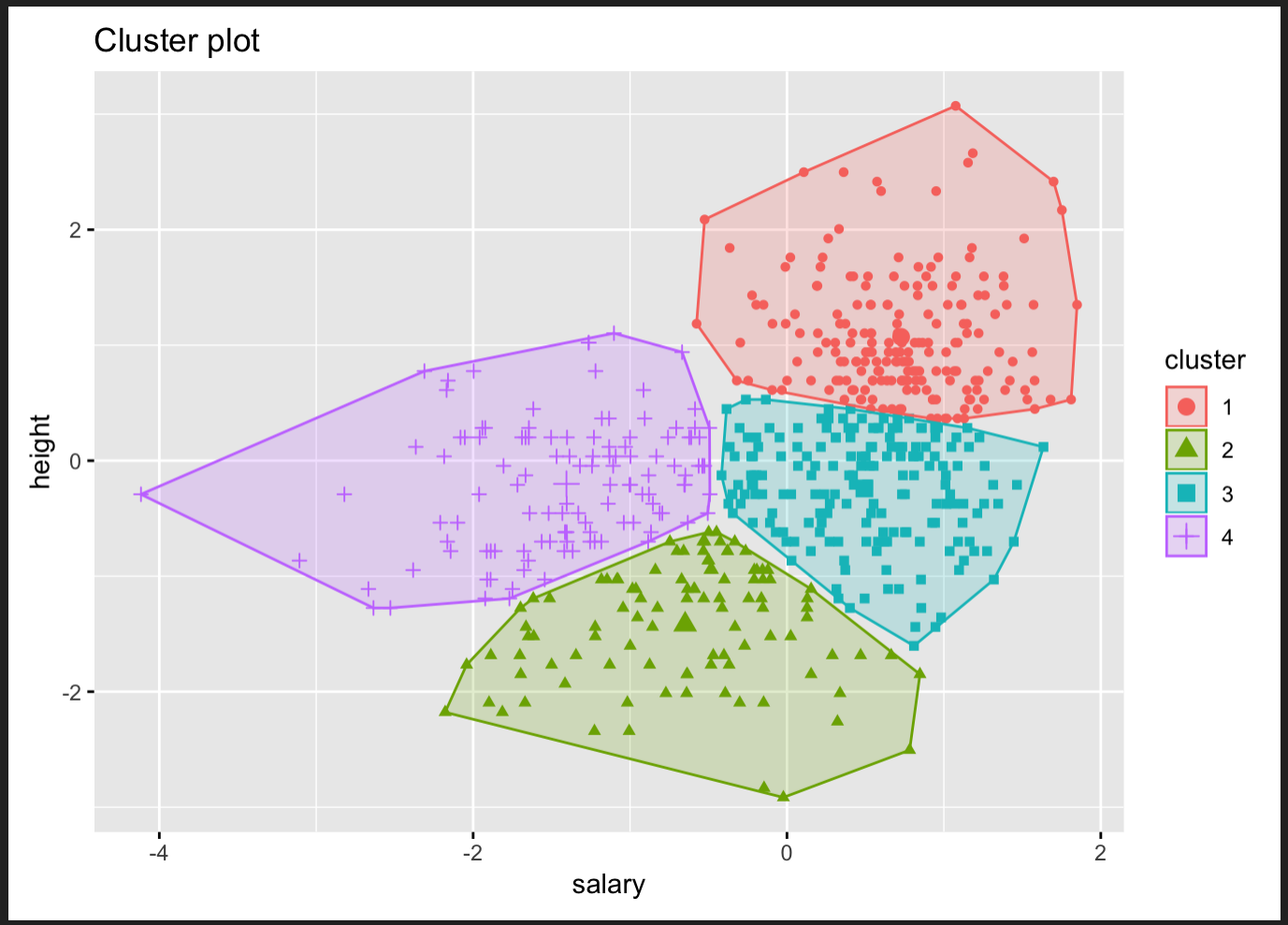
Let’s take 4.

```
unsup_model_1 = kmeans(data4
                       , centers = 4
                       , nstart = 10
                       , iter.max = 10)
```

WHAT'S INSIDE?




Iteration: 2
data4$height
data4$salary
3
2
1
0
-1
-2
-3
-4
-3
-2
-1
0
1
2


Cluster plot
height
salary
2
0
-2
-4
-2
0
2
cluster
1
2
3
4

THE K-MEANS ALGORITHM

- find random centers
- assign each observation to its closest center
- optimise for the WSS

WHAT’S PROBLEMATIC HERE?

BUT HOW DO WE KNOW HOW MANY CENTERS?

Possible approach:

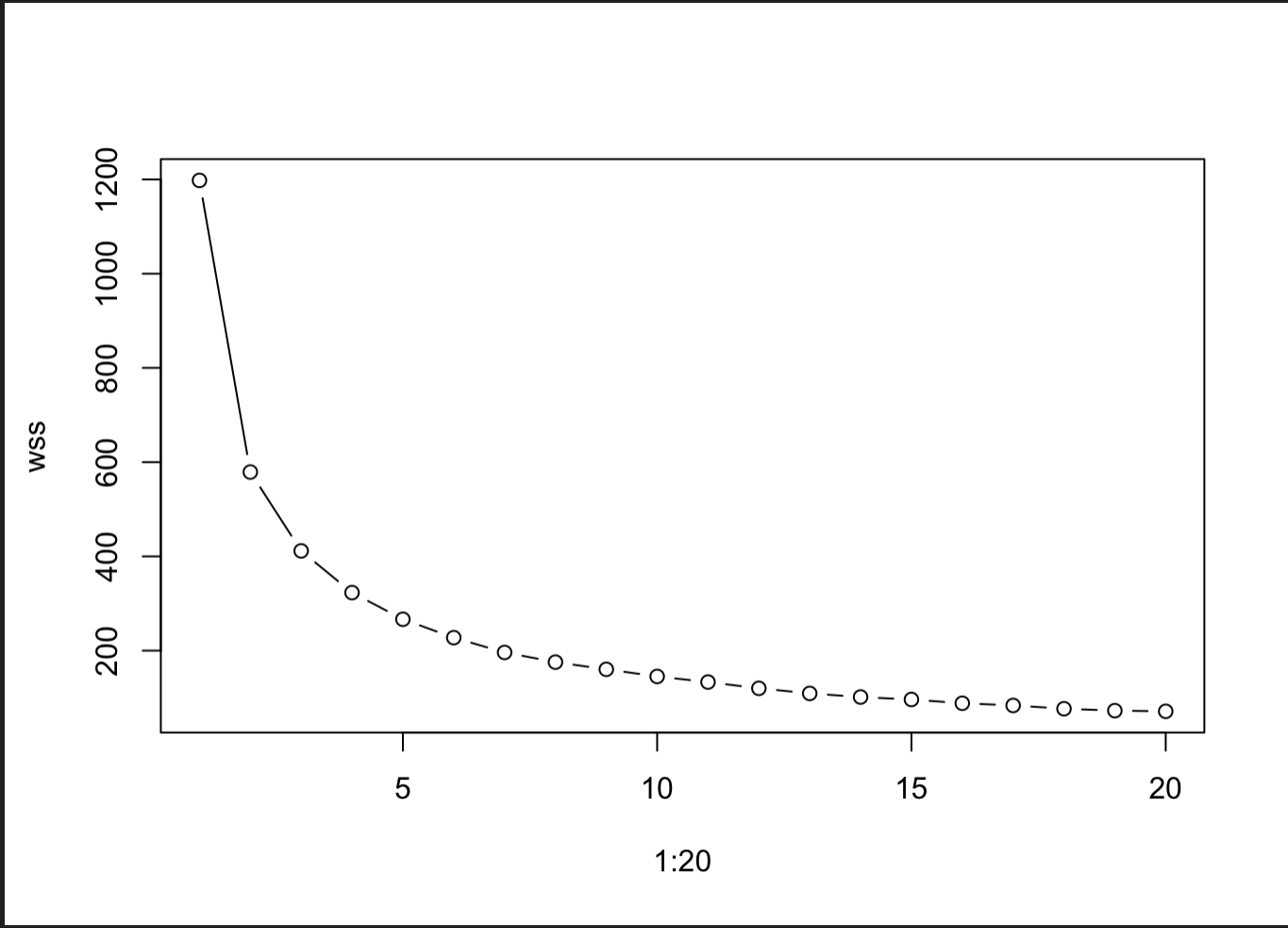
- run it for several combinations
- assess the WSS
- determine based on scree-plot

CLUSTER DETERMINATION

```
wss = numeric()
for(i in 1:20){
  kmeans_model = kmeans(data4, centers = i, iter.max = 20, nstart = 1
  wss[i] = kmeans_model$tot.withinss
}
```

SCREE PLOT (ELBOW METHOD)

Look for the inflexion point at center size *i*.



OTHER METHODS TO ESTABLISH *K*

- Silhoutte method (cluster fit)
- Gap statistic

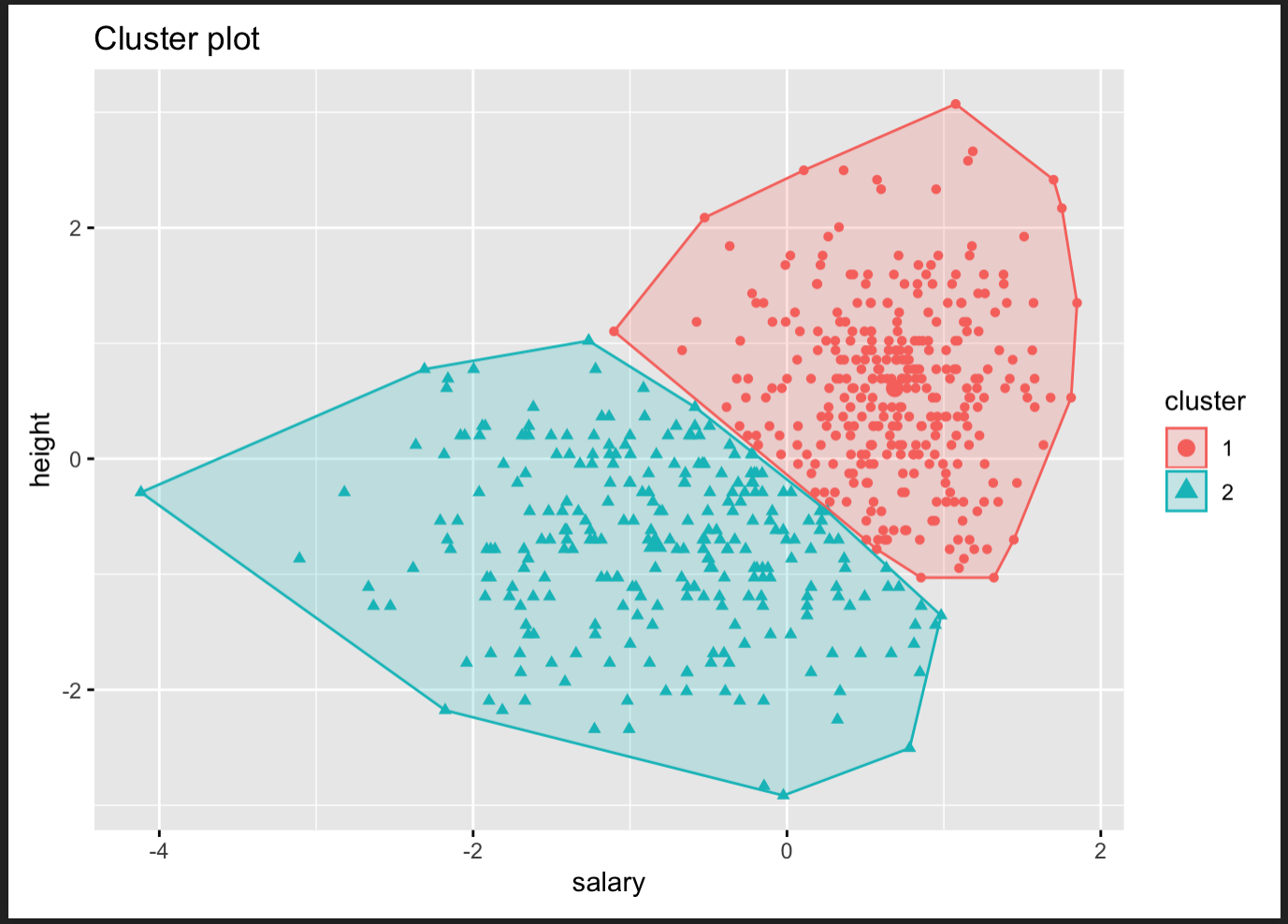
See also this tutorial.

CHOOSING *K*

We settle for $k = 2$

```
unsup_model_final = kmeans(data4
                        , centers = 2
                        , nstart = 10
                        , iter.max = 10)
```

PLOT THE CLUSTER ASSIGNMENT



OTHER UNSUPERVISED METHODS

- k-means (today)
- hierarchical clustering
- density clustering

ISSUES WITH UNSUPERVISED LEARNING

What’s lacking?

What can you (not) say?

CAVEATS OF UNSUP. ML

- there is no “ground truth”
- interpretation/subjectivity
- cluster choice

INTERPRETATION OF FINDINGS

INTERPRETATION OF FINDINGS

```
unsup_model_final$centers
```

```
##       salary     height
## 1  0.6869085  0.6101199
## 2 -0.8395549 -0.7457021
```

- Cluster 1: low salary, small
- Cluster 2: high salary, tall

Note: we cannot say anything about accuracy.

See the k-NN model.

INTERPRETATION OF FINDINGS

- subjective
- labelling tricky
- researchers choice!
- be open about this

PITFALLS AND PROBLEMS

BIAS

Remember supervised learning?

What is the essential characteristic of it?

SUPPOSE ...

... you have to predict the quality of song lyrics.

How would you do it?

EXAMPLES

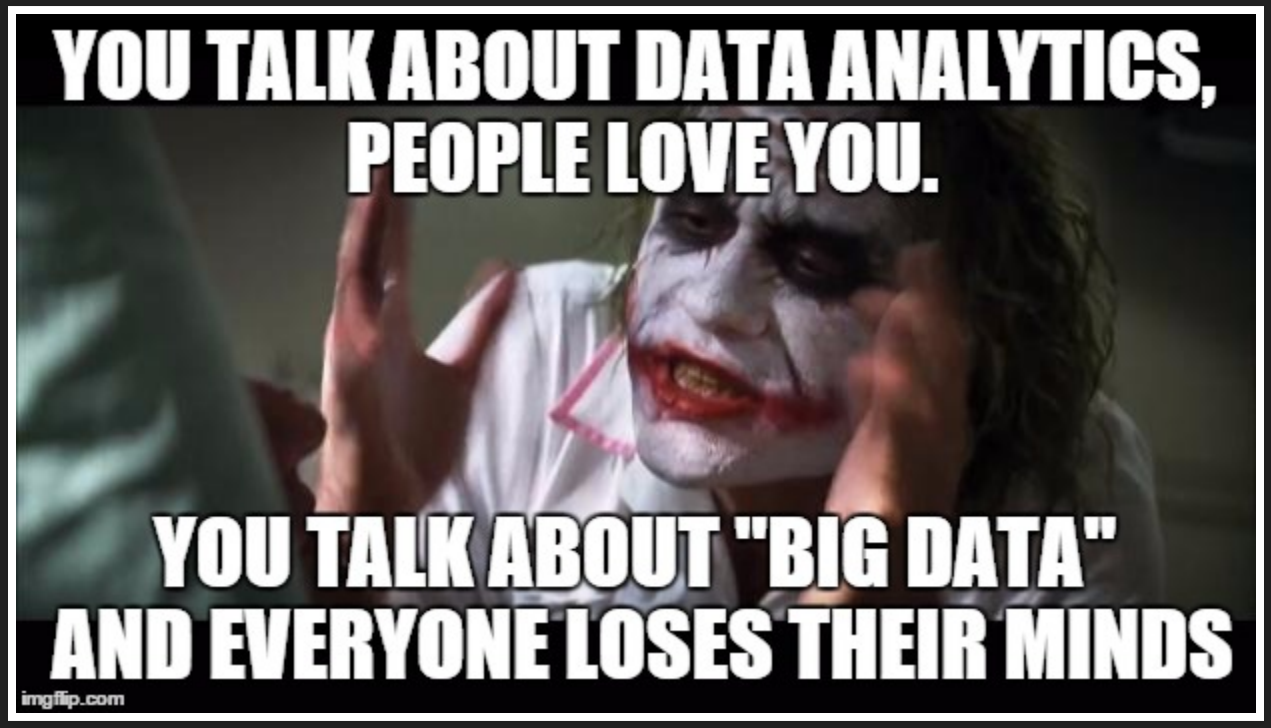
- quality of a football match
- attractiveness of an area
- quality of your degree

BIAS THROUGH LABELLED DATA

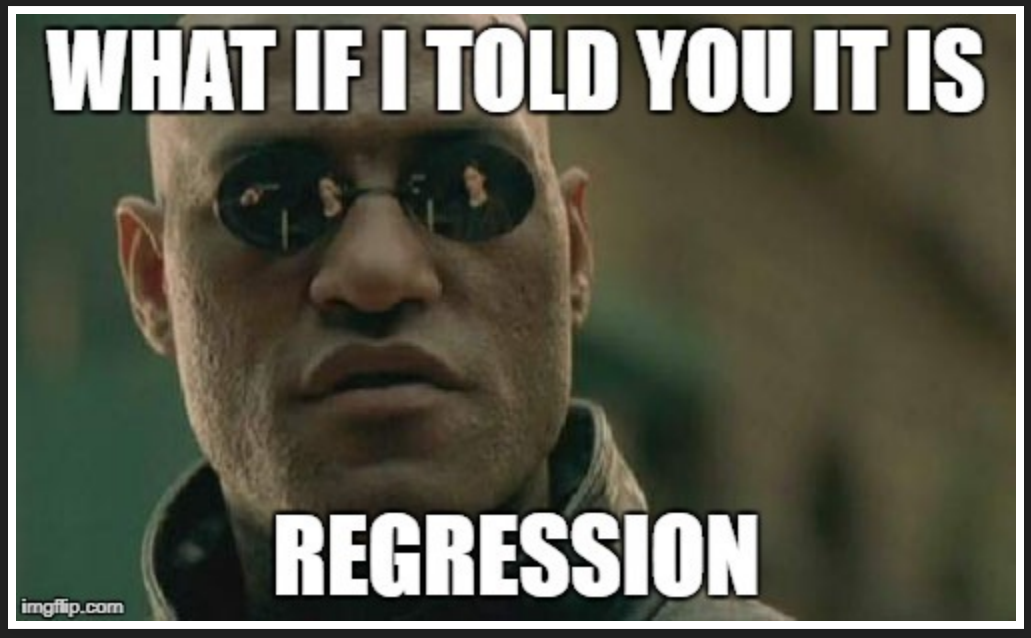
- machine learning is only the tool!
- supervised learning will always predict something
- you need the researcher’s/analyst’s mindset to interpret it

Basic principle: BS in = BS out.

PROBLEMATIC TRENDS



PROBLEMATIC TRENDS



PROBLEMATIC TRENDS

PROBLEMATIC TRENDS

Assumptions, assumptions. Everywhere assumptions.

THE NAIVITÉ FALLACY

UK government reveals new AI tool for flagging extremist content

THE NAIVITÉ FALLACY

The UK Home Office on Monday unveiled a £600,000 artificial intelligence (AI) tool to automatically detect terrorist content.

The Home Office cited tests that show the new tool can automatically detect 94% of Daesh propaganda with 99.995% accuracy. That accuracy rate translates into only 50 out of one million randomly selected videos that would require human review. The tool can run on any platform and can integrate into the video upload process to stop most extremist content before it ever reaches the internet.

source

THE NAIVITÉ FALLACY

THE NAIVITÉ FALLACY

Put simply: you can sell anything.

HERE'S AN IDEA

```
ai_terrorism_detection = function(person){
  person_classification = 'no terrorist'
  return(person_classification)
}
```

"UCL RESEARCHERS USE AI TO FIGHT TERRORISM!"

"AI 99.9999% ACCURATE IN SPOTTING TERRORISTS!"

THE CATEGORY MISTAKE OF DATA SCIENCE

https://www.youtube.com/watch?v=fCLI6kxFFTE

CATEGORY MISTAKE

- So we are getting there with self-driving cars.
- Hence: we can also address the other challenges.

!!!!

CATEGORY MISTAKE

Geller, 1999, 538 article

“I would not be at all surprised if earthquakes are just practically, inherently unpredictable.”

(Ned Field)

CATEGORY MISTAKE

- Building a sophisticated visual recogntion system != predicting everything
- Static phenomena vs. complex systems

Human behaviour might be the ultimate frontier in prediction.

ETHICAL ISSUES

- data sources
- (machine) learning systems
- reinforcing systems
- responsible practices

ETHICS & DATA SCIENCE

Your turn: do you see problems for these aspects?

- data sources
- (machine) learning systems

ETHICS & DATA SCIENCE

What about “reinforcing systems”?

ETHICS & DATA SCIENCE

Choose 1:

1. FP/FN issue in the hand of practitioners
2. academics' responsibility

AN OUTLOOK

What would an ideal Data Science look like?

BE SPECIFIC…

Academic data science

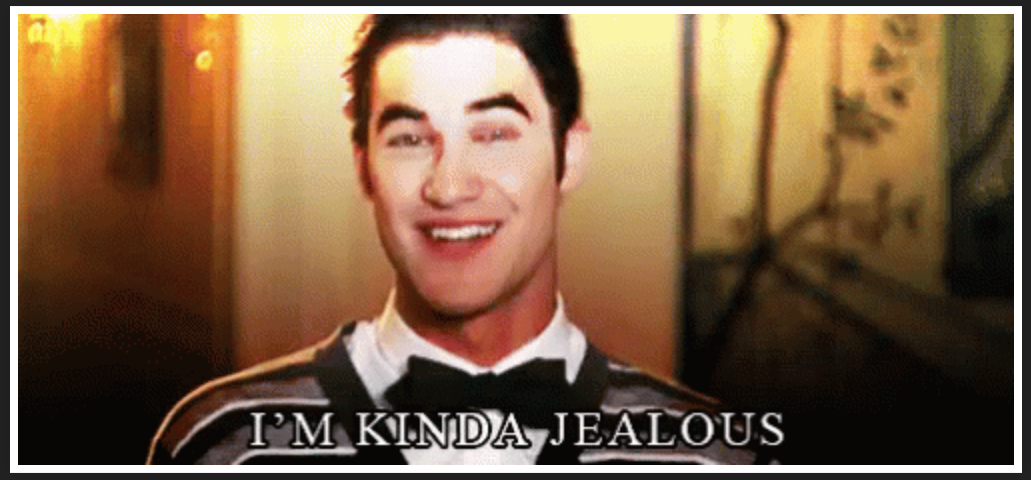
vs

“Industry” data science

Question: which one is leading?

Extreme view:

current academic data science is catering hype to compensate the Google envy.


I'M KINDA JEALOUS

ACADEMIC DATA SCIENCE

What it is doing	What it should be doing
creating “cool” studies	testing assumptions
pumping out non-reproducible papers	investing in fundamental data science research
hiring people to do cool things with our data	starting with the problem
getting on the data science train	focus on methods of data science

OUTLOOK

- we need *boring* studies!
  - longitudinal studies
  - assumption checks
  - replications
- we need to accept that Google & Co. are a different league in applying things
- we need to focus on the “ACADEMIC” part
- we need unis as control mechanism, not as a player

DATA SCIENCE FOR SECURITY

Where can data science help?

USING DATA SCIENCE

1. Automating human work
2. Exceeding human capacity
3. Augmenting human decision-making

AUTOMATING HUMAN WORK

Examples:

- scanning images for guns
- moderating content on social media
- access control to buildings

AUTOMATING HUMAN WORK

Why?

- reliability
- costs per unit
- scalability

EXCEEDING HUMAN CAPACITY

Examples:

- remote sensing applications
- deception detection
- tumor detection

EXCEEDING HUMAN CAPACITY

Why?

- processing capacity problem
- complex relationships
- limited attention of humans

AUGMENTING HUMAN DECISION-MAKING

Human-in-the-loop systems:

1. ML system makes a decision
2. Human revises the decision
3. Final decision reached

AUGMENTING HUMAN DECISION-MAKING

Why?

- uses best of both worlds
- context (human) + scale (machine)
- allows your system to gain traction

THE BIGGEST PROBLEM FOR DATA SCIENCE

EVERYTHING IS MATTER

EVERYTHING CAN BE MEASURED

EVERYTHING CAN BE REPRESENTED IN DATA

PEP THE SUPER DATA SCIENTIST

PEP THE SUPER DATA SCIENTIST

- knows everything about football
  - knows what happens if you hit a ball from angle X from distance Y at speed Z, etc.
  - knows everything about the physiology of the players, about the physical properties of the ball, about the rules
- has got access to all the data that you can possibly collect from a football game

But:

- Pep experiences the world from an isolated room…
- … through his python editor…
- … and only has access to the data

… and never saw a football match.

- Put differently: Pep knows everything about football but has never experienced it
- Pep is thrilled by Cristiano Ronaldo and Lionel Messi
- And has all their data

(Adaptation from The Knowledge Argument / Mary's Room by Frank Jackson, 1982)

ONE DAY ...

One day, Pep goes out to the ”real” world and watches a match between Juventus (C. Ronaldo) and Barca (Messi).

Will Pep learn anything?

WHAT DOES THIS MEAN?

- Qualia problem
- Originates from the philosophy of mind (consciousness problem)
- But reaches far beyond that

PEP’S PROBLEM & SECURITY

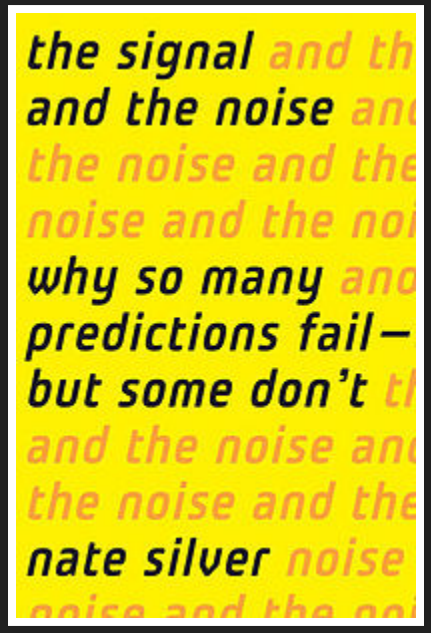
- perception of security
- experience of security
- perception of fairness

RECAP

1. Fastest intro to ML
2. Some problems and pitfalls of data science
3. Possible applications in security problems
4. The hard problem of data science

If you only read one book in 2019…

Read: “The Signal and the noise”, Nate Silver



QUESTIONS?