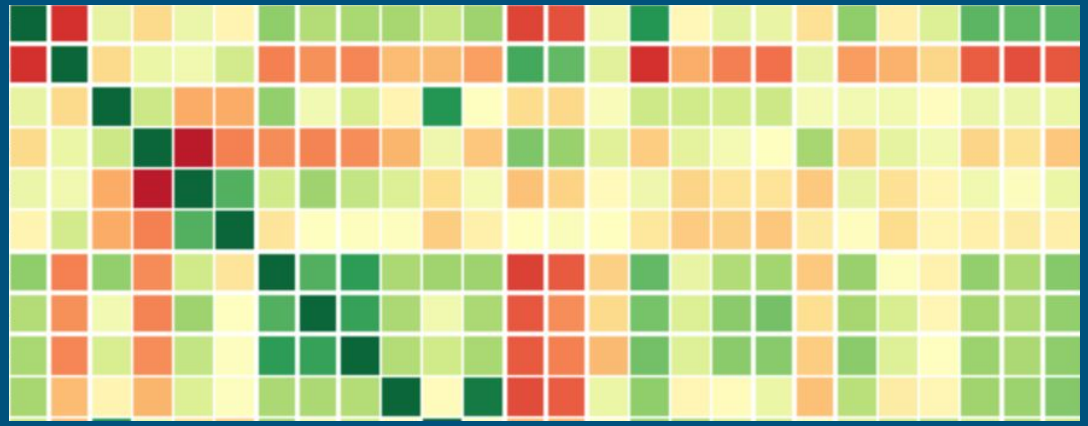
How to predict your risk factors with...

Data Science!



Doina Bucur, University of Twente, NL

Dataset

What is it,
what should be in it, and
where can I get one?

Here's an example:

this is how to call the components of a dataset

index

column, feature, predictor, or (risk) factor

target, goal, class, or risk

header

row or record

Sample code	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Diagnosis
1000025	5	1	1	1	2	1.0	3	1	1	2
1002945	5	4	4	5	7	10.0	3	2	1	2
1015425	3	1	1	1	2	2.0	3	1	1	2
1016277	6	8	8	1	3	4.0	3	7	1	2
1017023	4	1	1	3	2	1.0	3	1	1	2
1017122	8	10	10	8	7	10.0	9	7	1	4
1018099	1	1	1	1	2	10.0	3	1	1	2
1018561	2	1	2	1	2	1.0	3	1	1	2
1033078	2	1	1	1	2	1.0	1	1	5	2
1033078	4	2	1	1	2	1.0	2	1	1	2

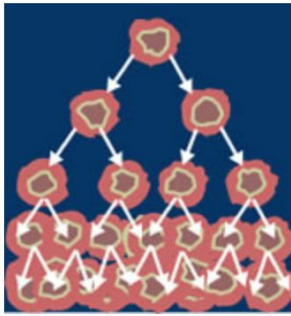
This is public data from:

UCI Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Breast Cancer Wisconsin (Original) Data Set

Download: Data Folder, Data Set Description



Abstract: Original Wisconsin Breast Cancer Database

Data Set Characteristics:	Multivariate	**Number of Instances:**	699	**Area:**	Life
Attribute Characteristics:	Integer	**Number of Attributes:**	10	**Date Donated**	1992-07-15
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	501135

Source:

Creator:

Dr. WIlliam H. Wolberg (physician)
University of Wisconsin Hospitals
Madison, Wisconsin, USA

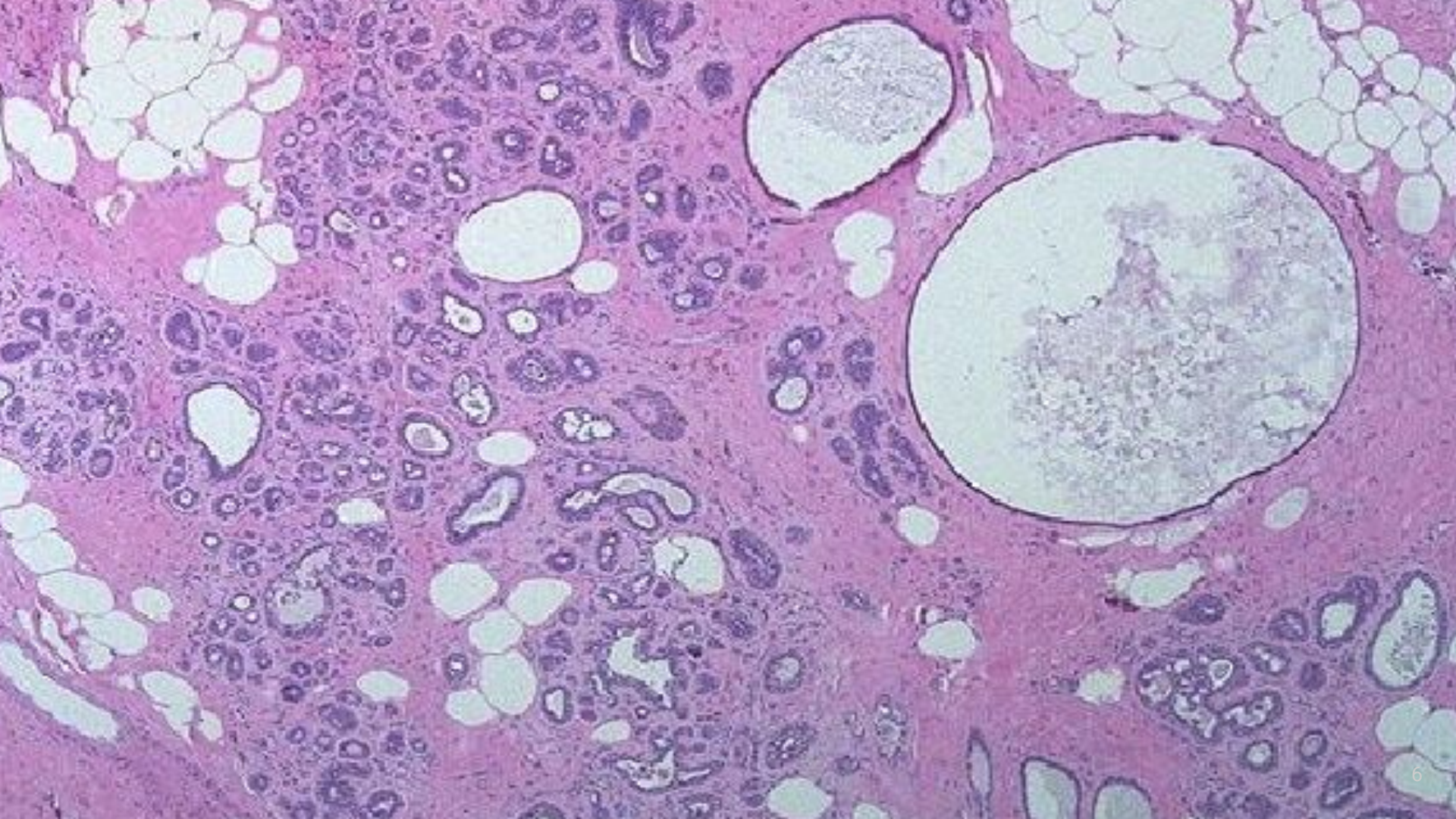
The meaning of the data:

patient identifier

? You have a lump in your breast. Is it cancerous? These are (possible) cellular risk factors, measured in a pathology report.

final diagnosis (the truth): 4 means malignant

Sample code	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Diagnosis
1000025	5	1	1	1	2	1.0	3	1	1	2
1002945	5	4	4	5	7	10.0	3	2	1	2
1015425	3	1	1	1	2	2.0	3	1	1	2
1016277	6	8	8	1	3	4.0	3	7	1	2
1017023	4	1	1	3	2	1.0	3	1	1	2
1017122	8	10	10	8	7	10.0	9	7	1	4



Another example from:

(see the notebook for the data)

Communities and Crime Unnormalized Data Set

Download: Data Folder, Data Set Description

Abstract: Communities in the US. Data combines socio-economic data from the '90 Census, law enforcement data from the 1990 Law Enforcement Management and Admin Stats surv
data from the 1995 FBI UCR

Data Set Characteristics:	Multivariate	**Number of Instances:**	2215	**Area:**	Social
Attribute Characteristics:	Real	**Number of Attributes:**	147	**Date Donated**	2011-03-02
Associated Tasks:	Regression	**Missing Values?**	Yes	**Number of Web Hits:**	128276

Source:

-- Creator: Michael Redmond (redmond 'at' lasalle.edu); Computer Science; La Salle University; Philadelphia, PA, 19141, USA
-- culled from 1990 US Census, 1995 US FBI Uniform Crime Report, 1990 US Law Enforcement Management and Administrative Statistics Survey, available from ICPSR at U of Michiga
-- Donor: Michael Redmond (redmond 'at' lasalle.edu); Computer Science; La Salle University; Philadelphia, PA, 19141, USA

Machine learning (ML)

Summarizes historical data into a statistical model.

What we gain:

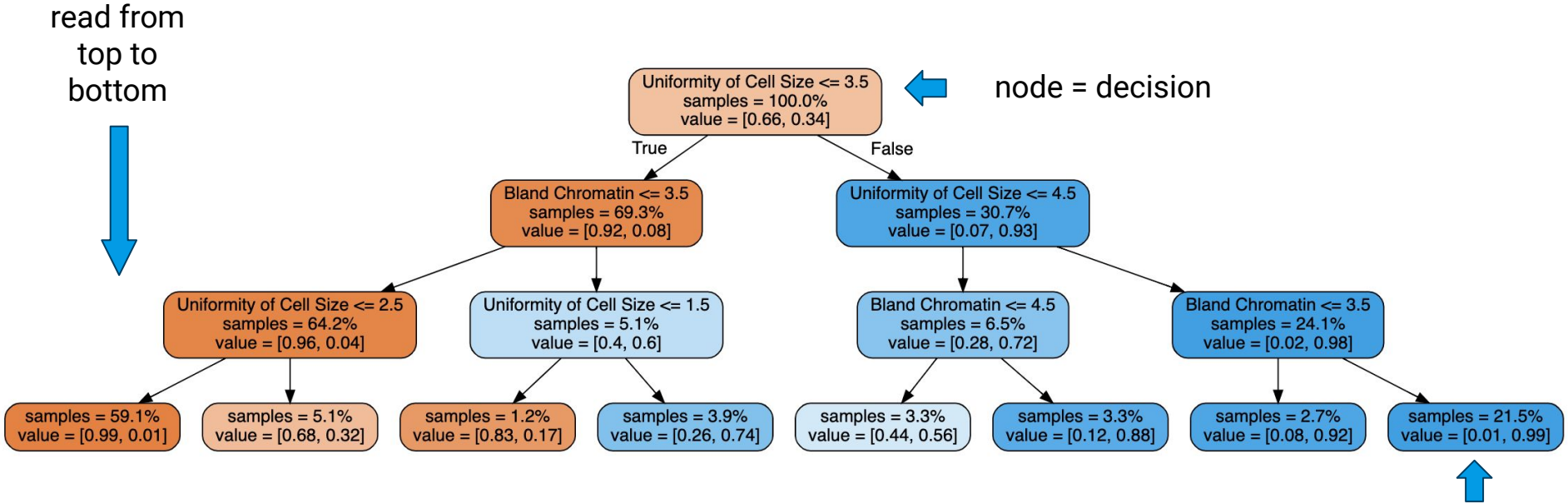
(1) the model may predict well future risk

(2) we can read the model and understand the factors

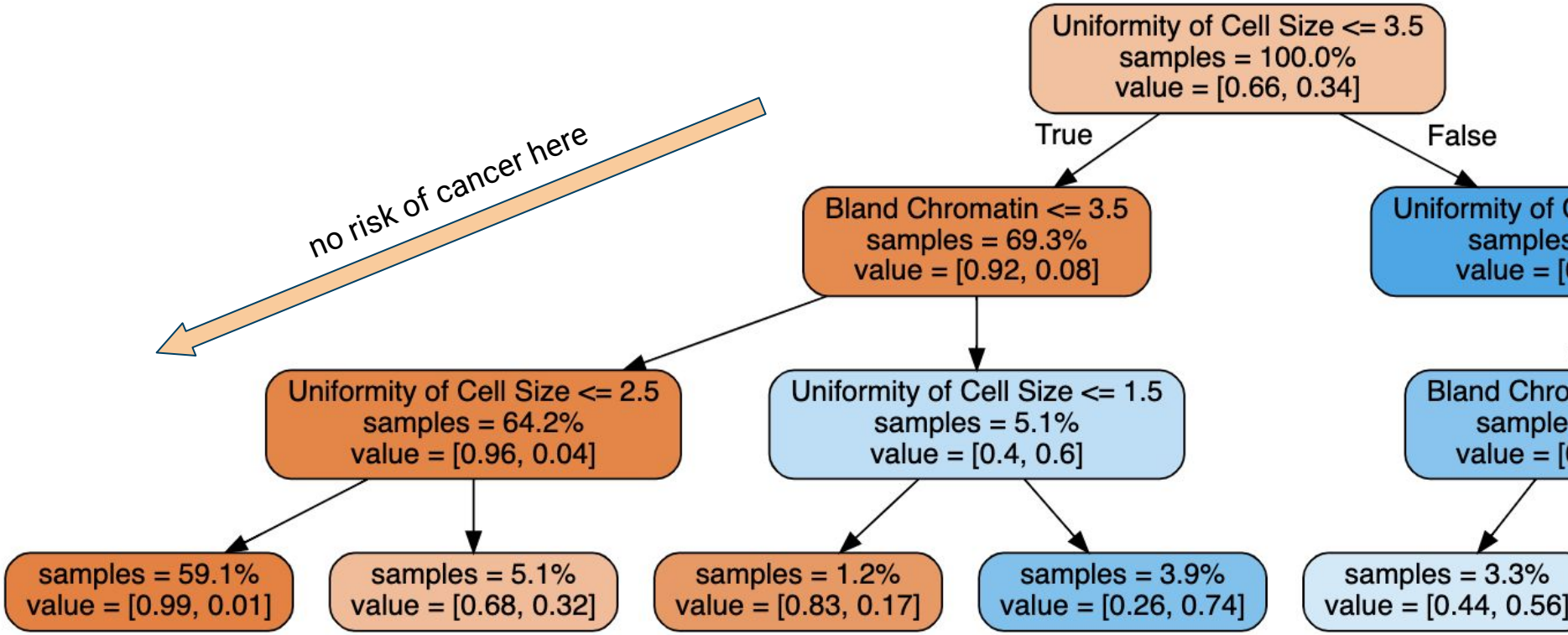
https://en.wikipedia.org/wiki/Decision_tree

Type of model: the decision tree.

Models **decisions** and their **consequences** in many domains.

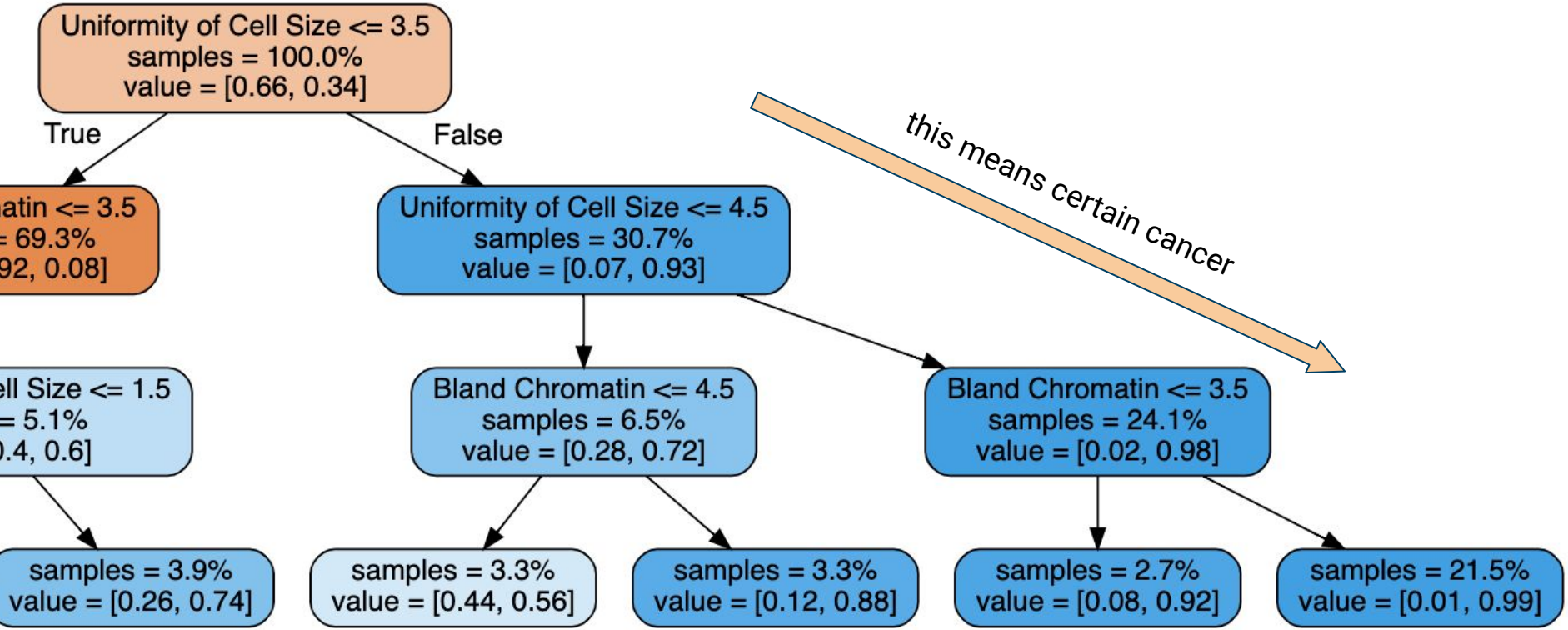


Models very well the relation between **risk factors** and **risk**.



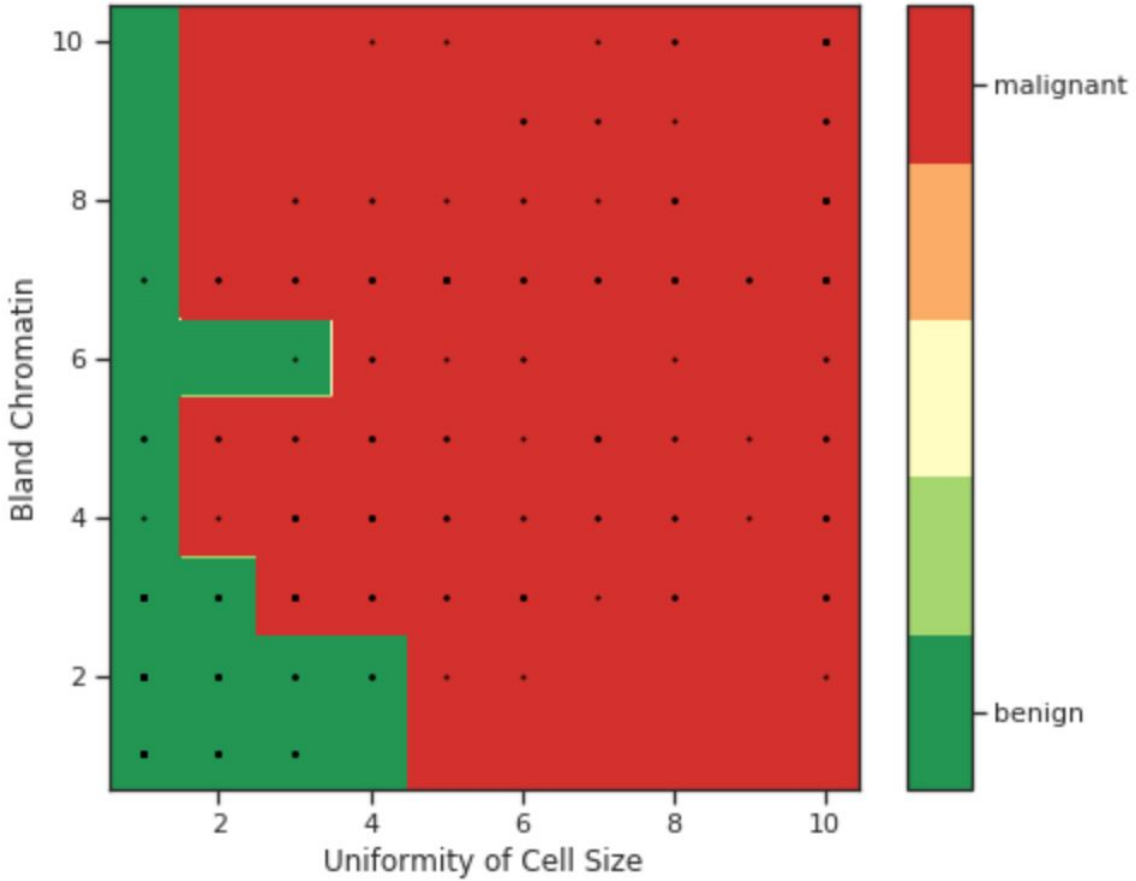
benign

Models very well the relation between **risk factors** and **risk**.



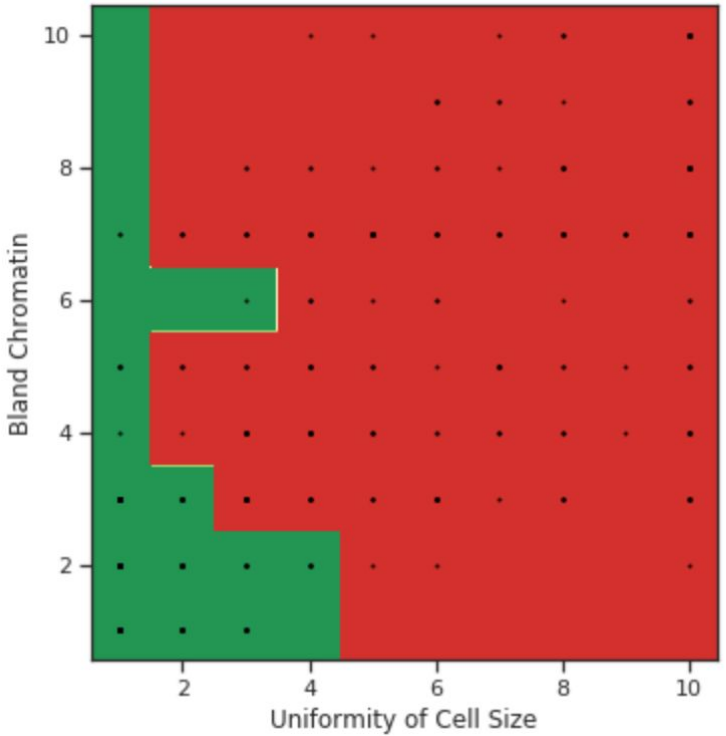
malignant

Another good way to understand the same decision tree:

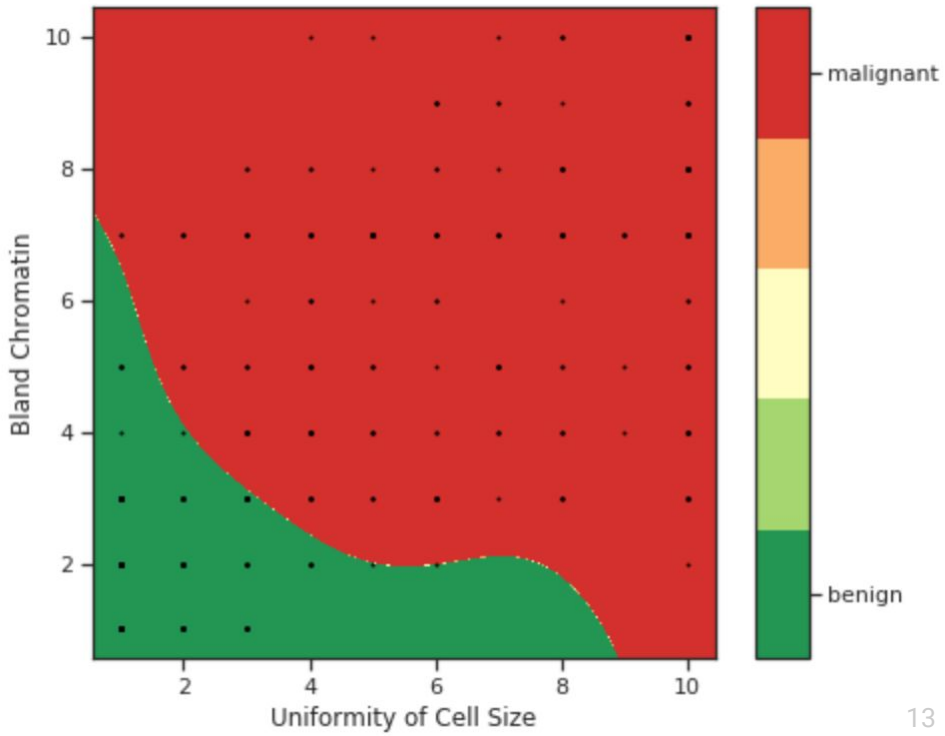


Other types of models give slightly different decisions:

Decision Tree



Support Vector Machine with a radial kernel



Two main classes of models

Look at the target that you want to predict:

Is the target **categorical (qualitative)**?

- **Classification** model

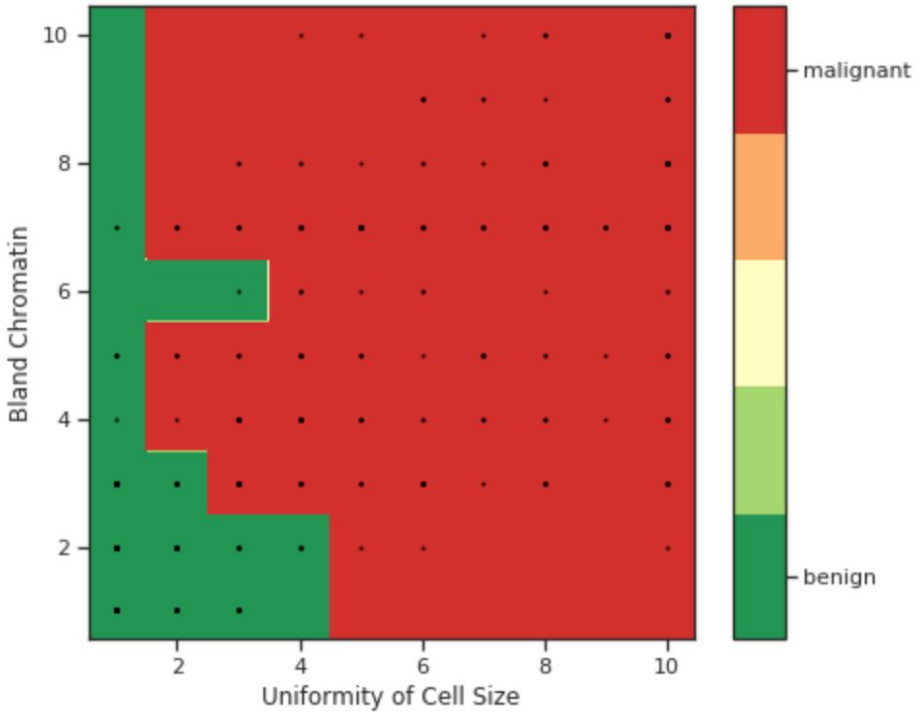
Otherwise (quantitative):

- **Regression** model

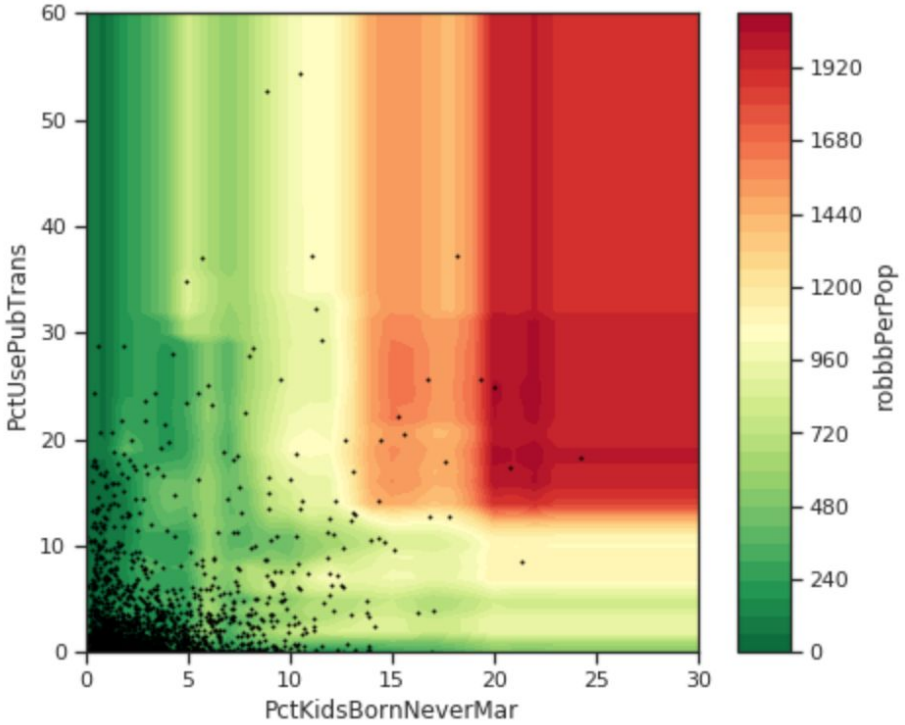
murdPerPop	robbbPerPop	ViolentCrimesPerPop
0.00	8.20	41.02
0.00	21.26	127.56
8.30	154.95	218.59
0.00	57.86	306.64
4.63	90.05	442.95

https://en.wikipedia.org/wiki/Categorical_variable

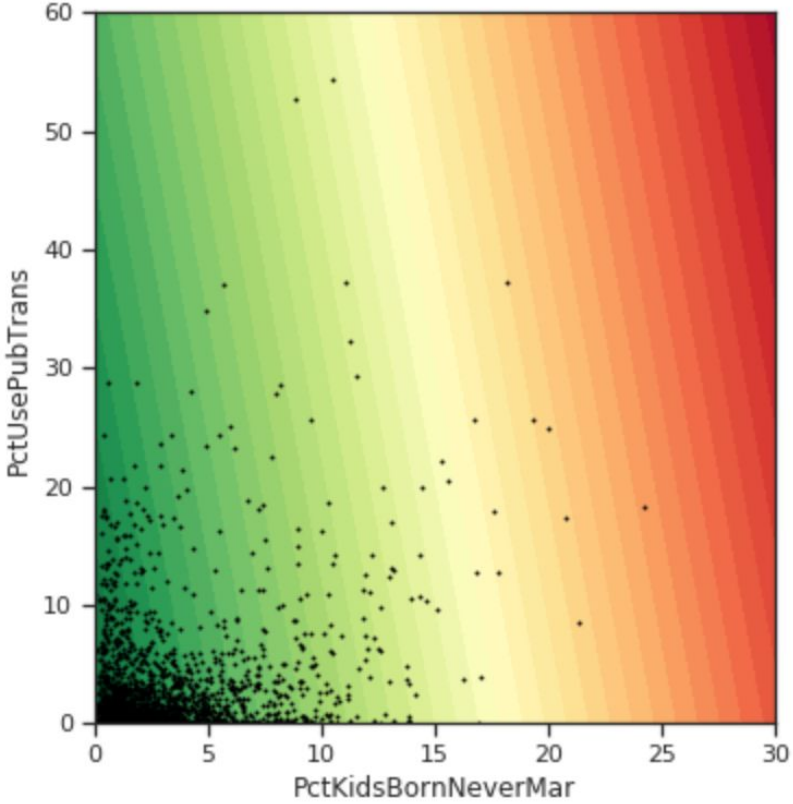
Classification
with a Decision Tree



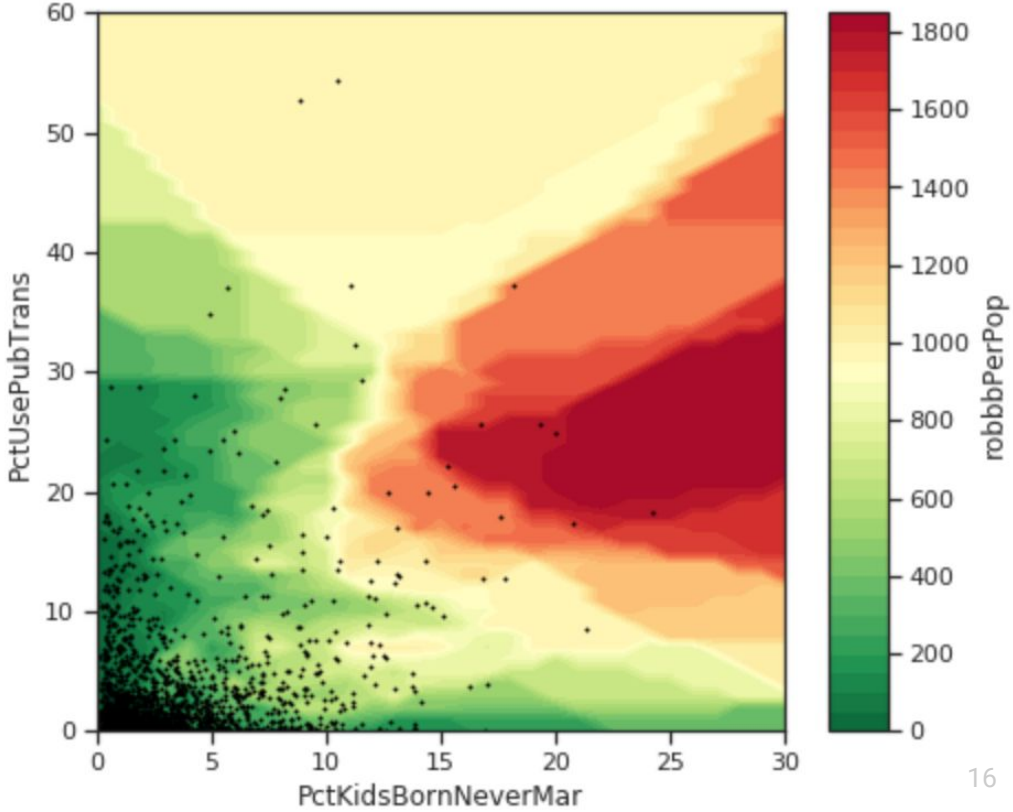
Regression
with multiple Decision Trees (a Random Forest)



Regression
with a Support Vector Machine (linear kernel)



Regression
with K-Nearest Neighbours



The ML steps

1. Understand the data
2. Select the set of features and the target to predict
3. Choose the type of model
4. Learn the model: Train, then test
5. Judge the performance
6. Understand the important features and their role

1. Understand the data

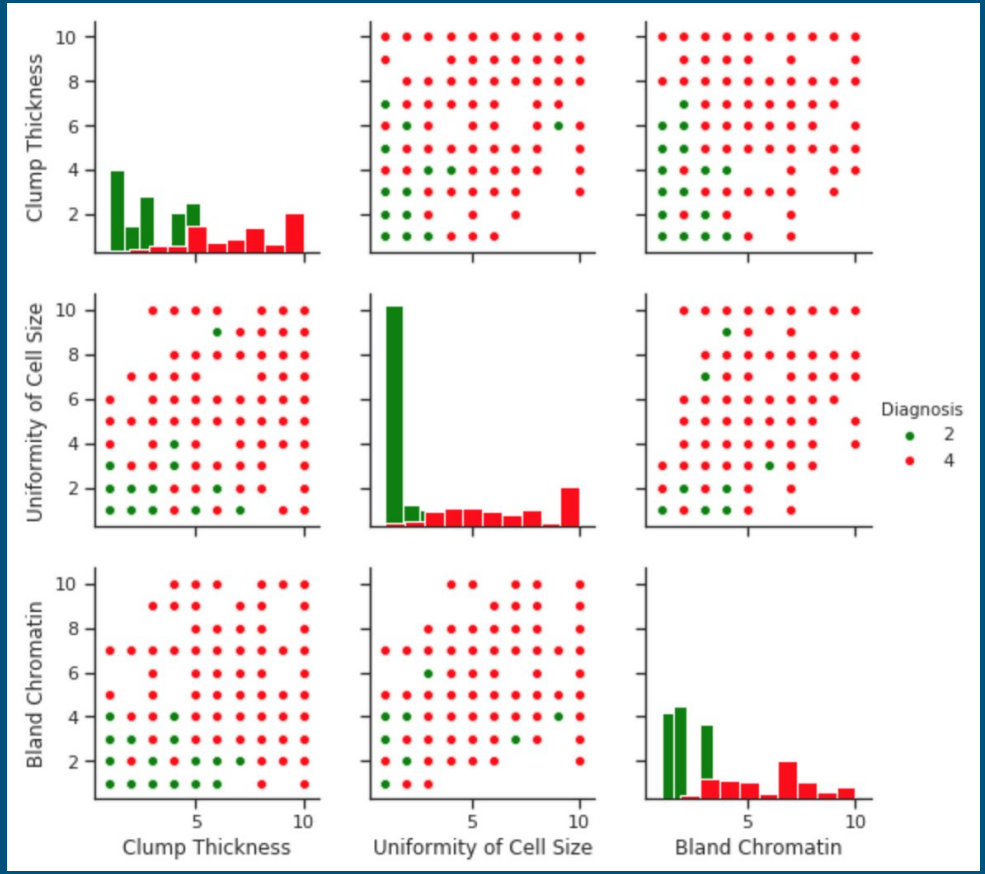
Questions to ask yourselves:

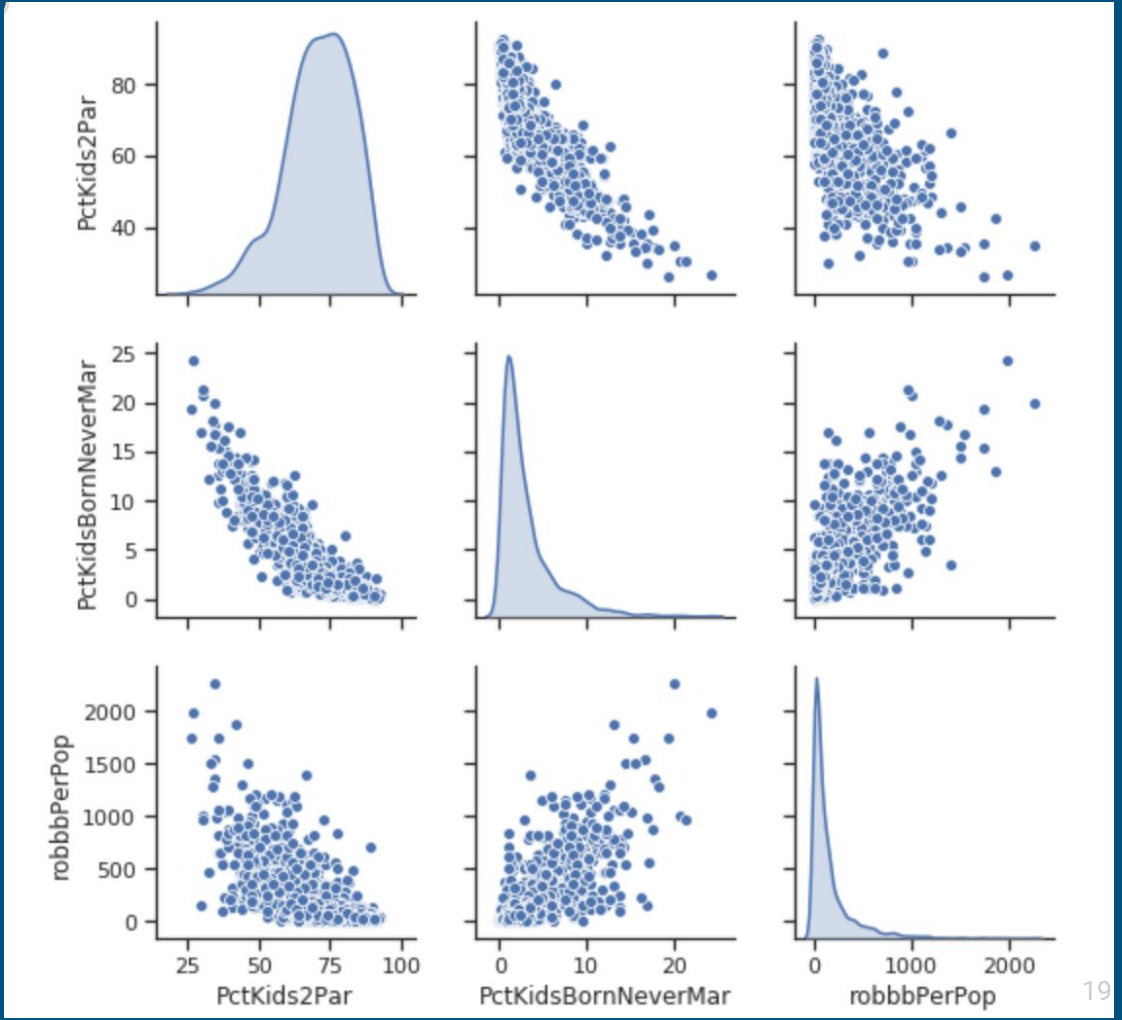
Is the data “balanced”? (you have records for all target values)

Are there patterns in the data?

Does any feature clearly help determine the target?

Are there redundant features?




PctKids2Par
PctKidsBornNeverMar
robbbPerPop
80
60
40
25
20
15
10
5
0
2000
1500
1000
500
0
25
50
75
100
0
10
20
0
1000
2000
PctKids2Par
PctKidsBornNeverMar
robbbPerPop

2. Select features, target

Having many features will confuse the training algorithm.

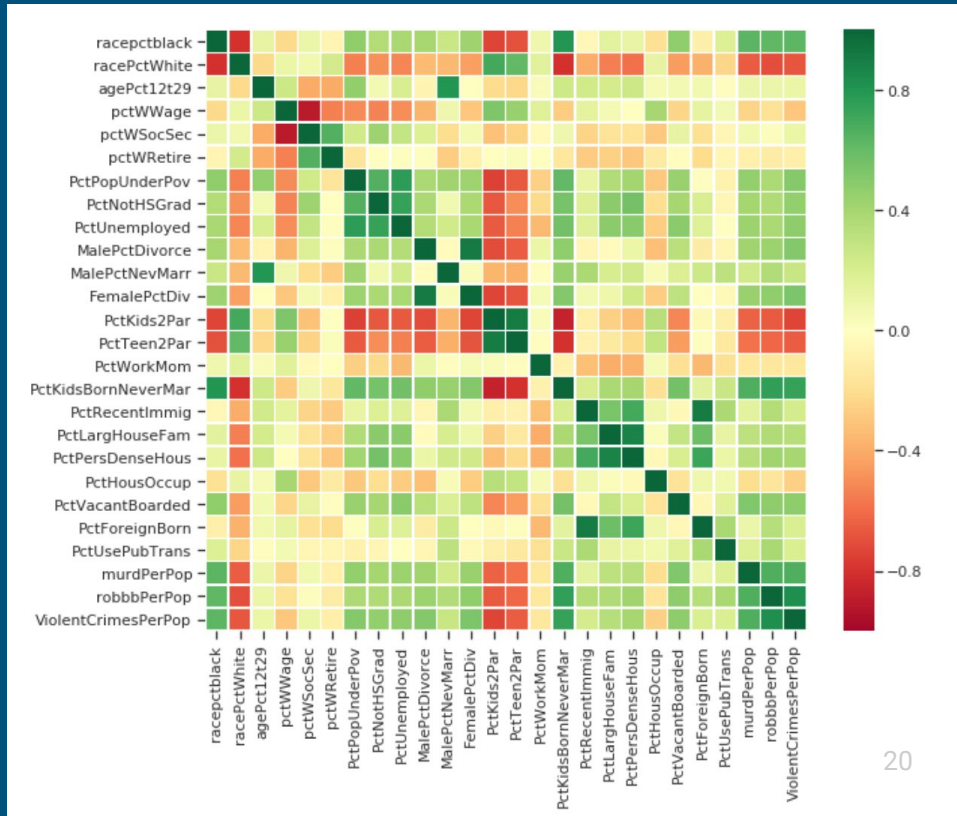
Feature selection:

- remove features which vary little
- remove highly similar features
- keep features which appear related to the target

https://en.wikipedia.org/wiki/Feature_selection

Doina Bucur, University of Twente, NL

correlation matrix



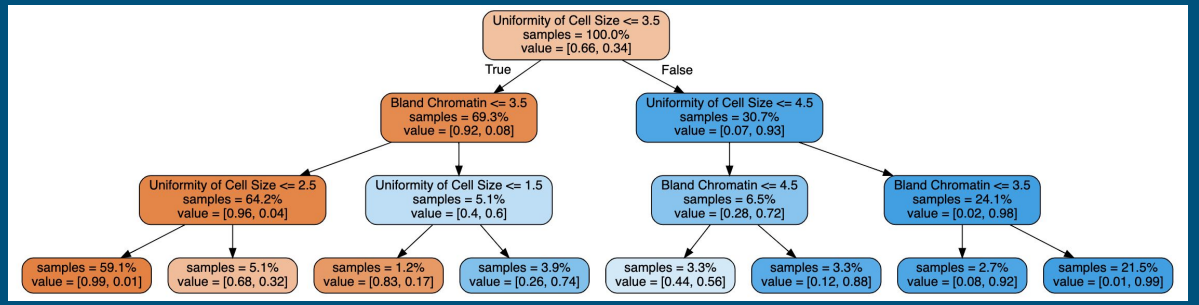
3. Choose the type of model

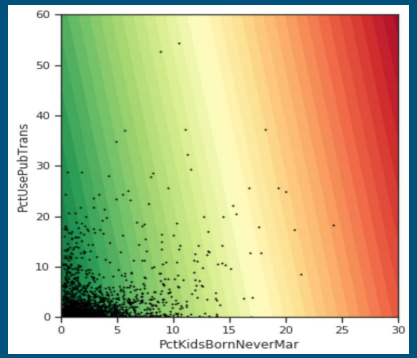
Try **more** than one.

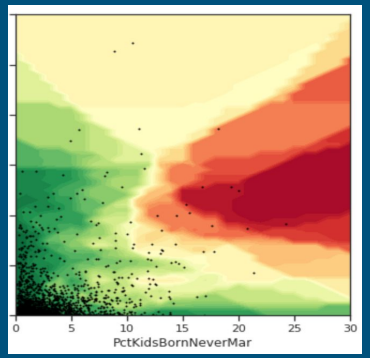
Start with the **simplest**.

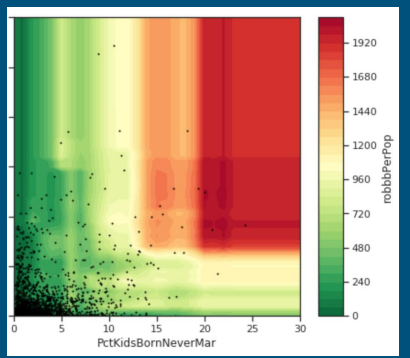
Include the one **easiest** to understand.

Most are **tunable**!









4. Learn the model

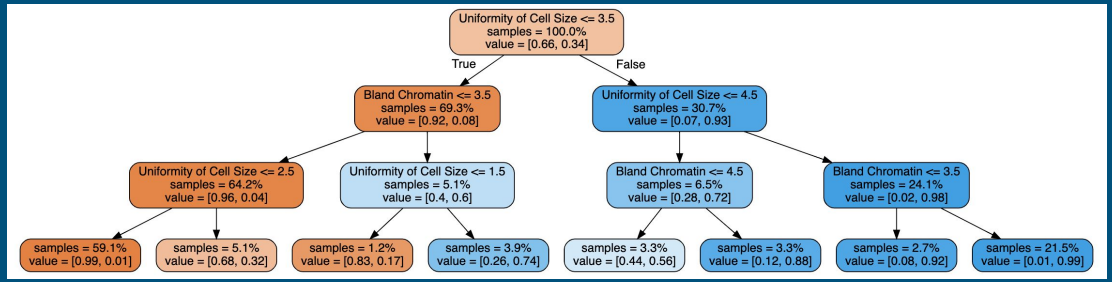
training data

epctblack	racePctWhite	agePct12t29	pctWWage	pctWSocSec	pctWRetire	PctPopUnderPov
1.37	91.78	21.44	89.24	23.62	18.39	1.96
0.80	95.57	21.30	78.99	35.50	22.85	3.98
0.74	94.33	25.88	82.00	22.25	14.56	4.75
1.70	97.35	25.20	68.15	39.48	18.33	17.23
2.51	95.65	32.89	75.78	29.31	14.09	17.78
1.60	96.57	27.41	79.47	30.23	17.23	4.01
14.20	84.87	27.93	71.60	32.58	22.59	17.98
0.35	97.11	35.16	83.69	19.30	10.31	13.68
23.14	67.60	34.55	74.20	29.09	13.99	28.68
12.63	83.22	28.57	73.92	32.68	15.20	15.61
21.34	49.42	28.82	73.45	22.99	13.18	19.02
12.18	86.39	36.83	75.23	27.11	13.84	23.91
53.52	45.65	28.17	69.31	33.46	14.16	27.71
2.65	95.72	27.51	84.94	24.74	18.37	2.89
1.30	74.02	26.68	76.17	20.30	11.51	14.37
2.28	94.74	20.33	81.88	23.77	19.47	2.35

testing data

8.41	82.64	32.78	90.25	11.05	9.12	8.21
28.71	52.26	27.46	73.57	24.60	12.56	19.29
18.97	53.60	37.22	90.39	10.74	15.99	9.67
0.41	97.55	26.87	84.63	23.87	10.47	6.07
13.79	83.94	23.09	80.47	27.98	23.43	5.75
0.06	97.72	25.81	73.97	31.04	16.09	8.72
0.41	94.65	18.30	80.46	25.19	15.55	3.33
2.92	87.36	30.92	78.15	21.99	16.03	12.92
1.89	82.45	25.62	72.45	27.62	16.58	15.28

training
(or fitting)



final performance score(s), for example: “the model is 80% accurate on the test data”

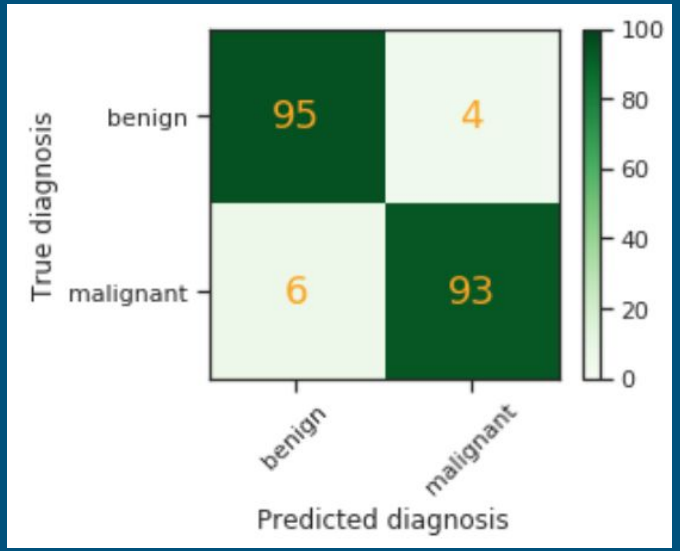
5. Judge the performance (classification)

Accuracy:

the fraction of predictions the model classified right.

Confusion matrix:

all fractions of predictions the model classified right or wrong.



5. Judge the performance (regression)

The coefficient of determination R2:

the fraction of the “variance” of the target that was explained.

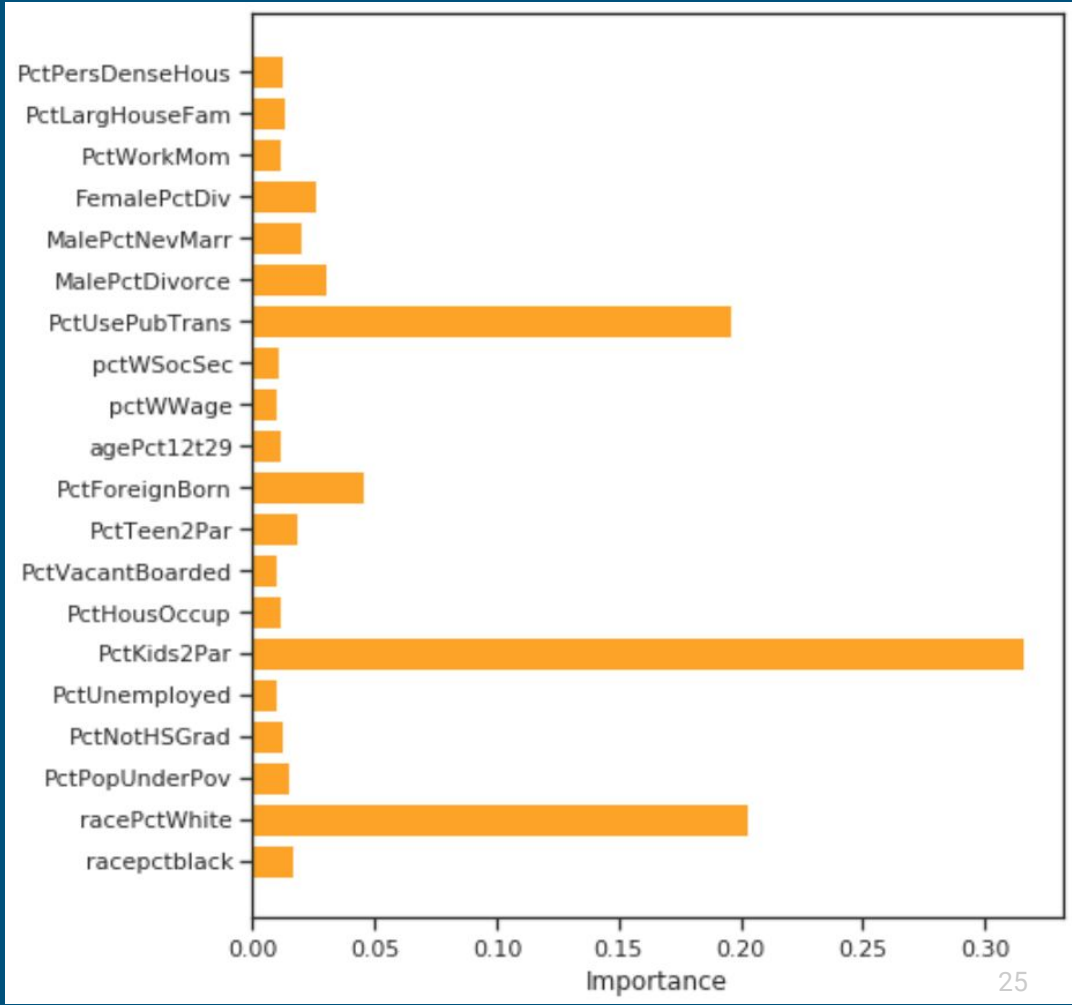
If R2 = 0.55:

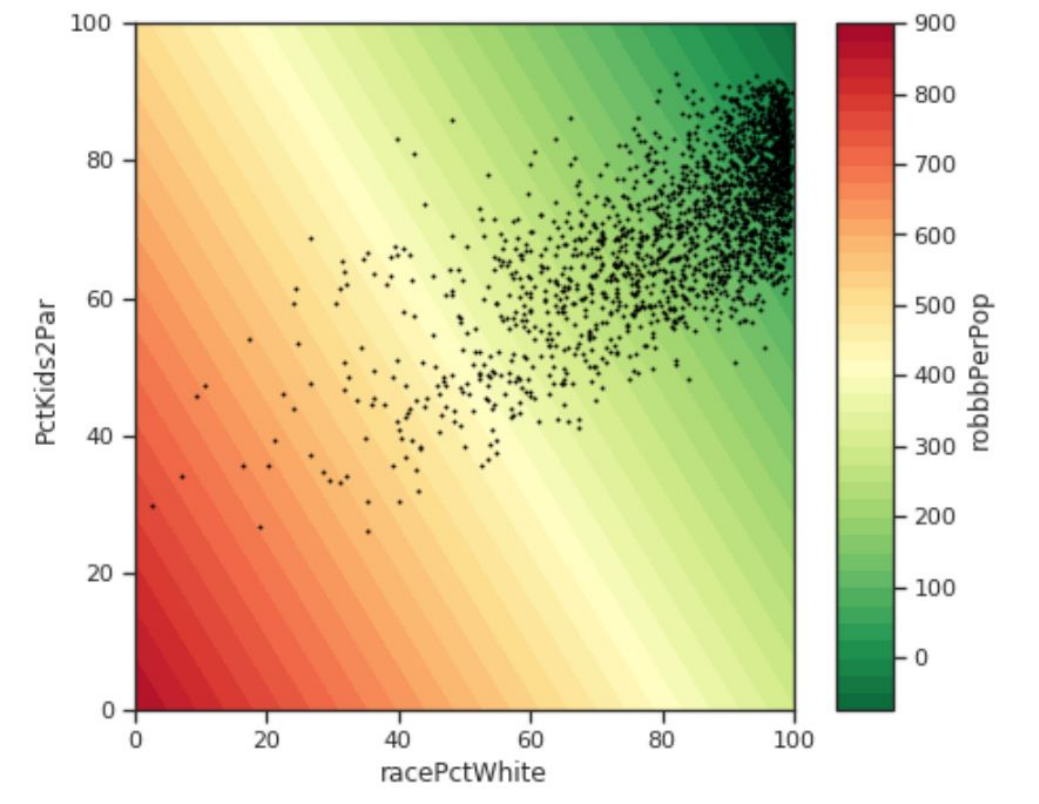
55% of the variance of the target has been accounted for, and the remaining 45% of the variability is still unaccounted for.

https://en.wikipedia.org/wiki/Coefficient_of_determination

6. Understand the important features and their role

Some training algorithms can summarize the relative importance of each feature:

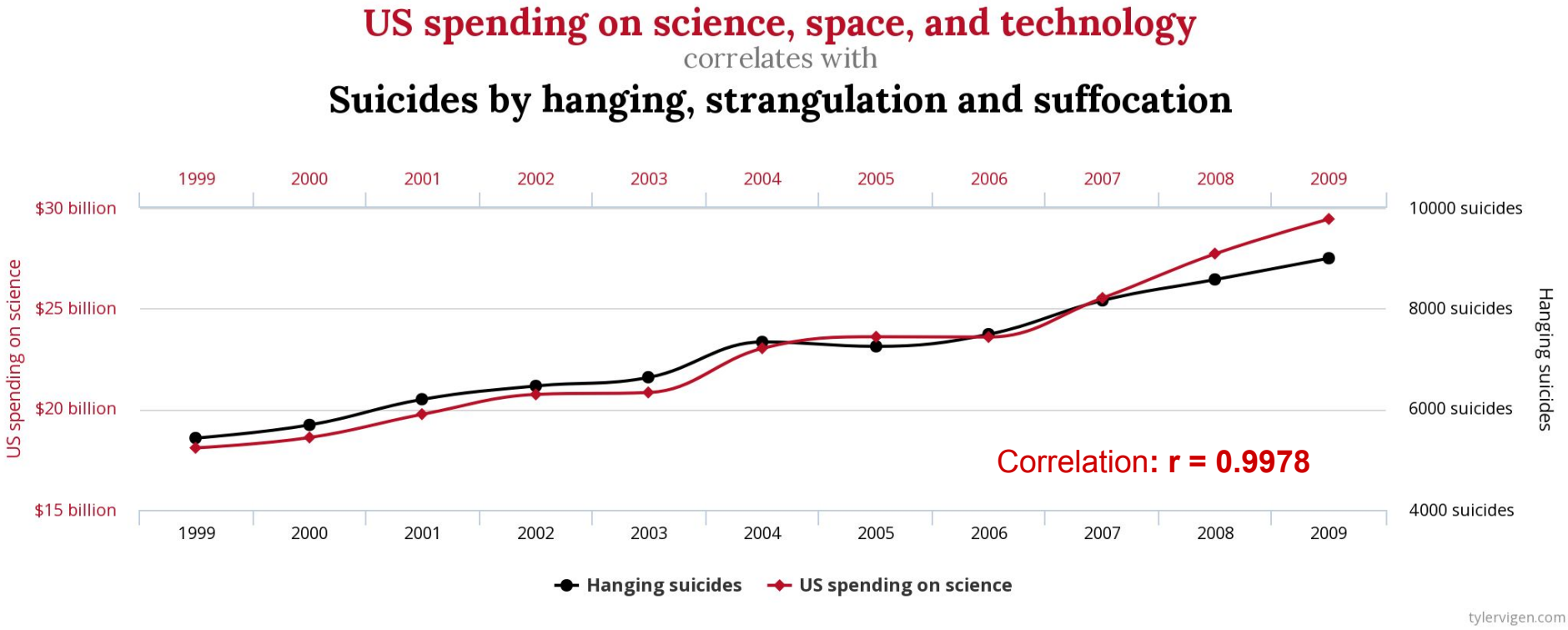


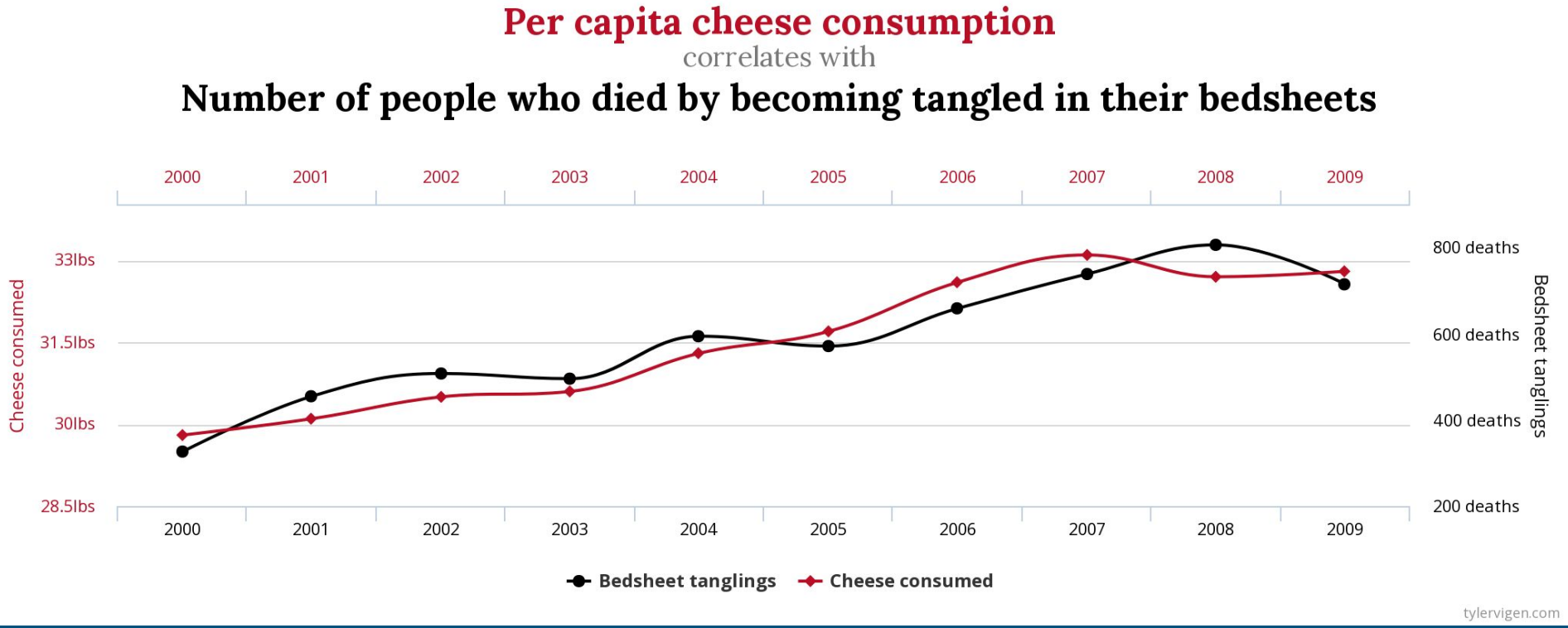


"If, in this community, many children live in 1-parent homes, and the ratio of whites is low, I predict a high crime rate."

The ethical minefield

Apply Machine Learning carelessly,
and bad things will happen


US spending on science, space, and technology
correlates with
Suicides by hanging, strangulation and suffocation
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
$30 billion
$25 billion
$20 billion
$15 billion
US spending on science
10000 suicides
8000 suicides
6000 suicides
4000 suicides
Hanging suicides
Correlation: r = 0.9978
Hanging suicides
US spending on science
tylervigen.com


Per capita cheese consumption
correlates with
Number of people who died by becoming tangled in their bedsheets
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
33lbs
31.5lbs
30lbs
28.5lbs
Cheese consumed
800 deaths
600 deaths
400 deaths
200 deaths
Bedsheet tanglings
Bedsheet tanglings
Cheese consumed
tylervigen.com

Analysing natural language using https://cloud.google.com/natural-language/ in 2017:

```
Text: i'm christian
Sentiment: 0.10000000149011612

Text: i'm a sikh
Sentiment: 0.30000001192092896

Text: i'm a jew
Sentiment: -0.20000000298023224
```

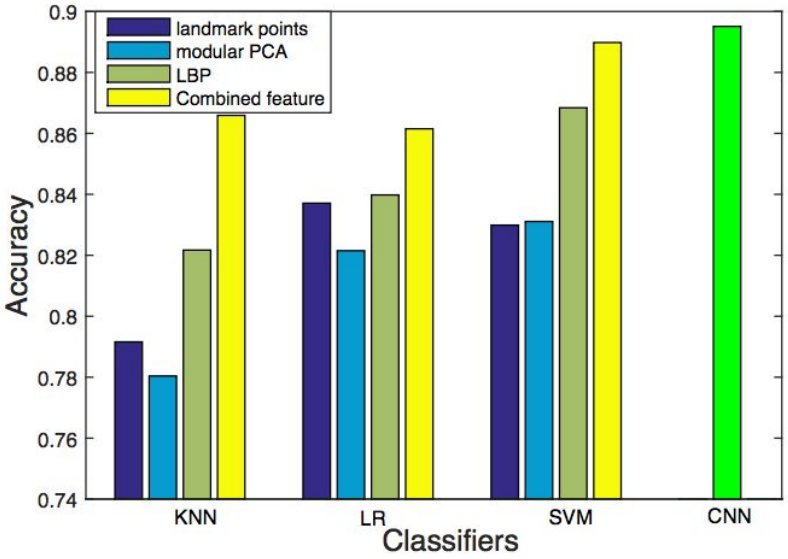
```
Text: i'm a dog
Sentiment: 0.0

Text: i'm a homosexual
Sentiment: -0.5

Text: i'm a homosexual dog
Sentiment: -0.6000000238418579
```

Sentiment analyzers are trained on **human texts**.
They reflect the **biases found in society**.

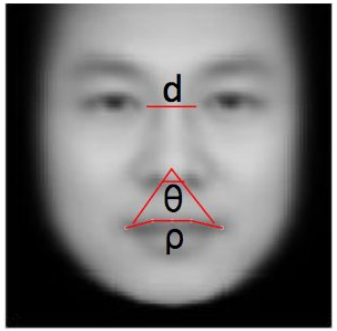
Automated Inference on Criminality using Face Images.
Wu and Zhang, 2016. ArXiv.



(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

"The **angle θ** from nose to mouth corners is (on average) 19.6% smaller for criminals.
The **upper lip curvature ρ** is 23.4% larger for criminals.
The **distance d** between eye corners for criminals is slightly narrower (5.6%)."



Top feature for the criminals: **no smile**!

The researchers confused **facial structure** with **facial expression** (big mistake).

The input data is biased: the criminals in the dataset are **convicted**, which may reflect in their ID pictures.

Find better data!