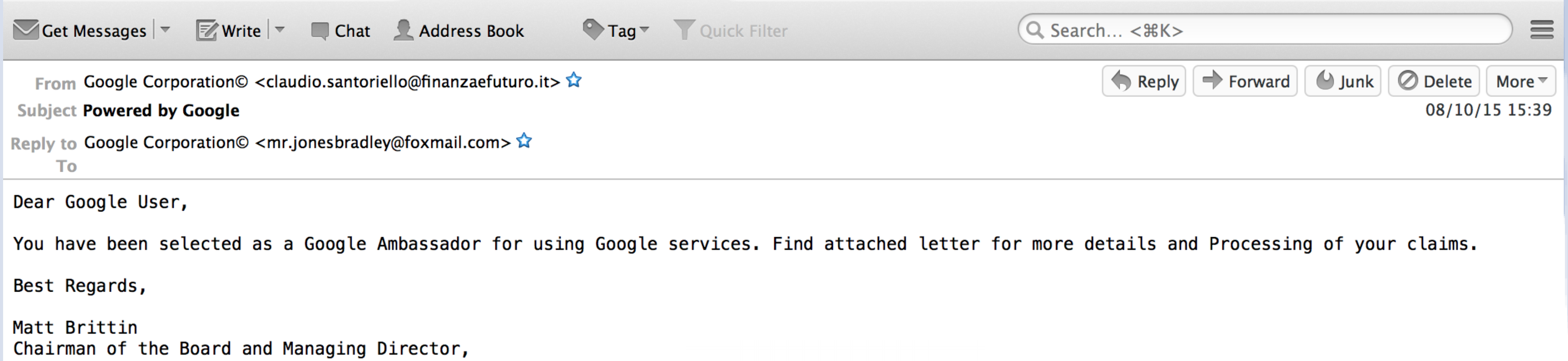
Spam Classification



From Google Corporation© <claudio.santoriello@finanzaefuturo.it>

Subject **Powered by Google**

08/10/15 15:39

Reply to Google Corporation© <mr.jonesbradley@foxmail.com>

To

```
Dear Google User,

You have been selected as a Google Ambassador for using Google services. Find attached letter for more details and Processing of your claims.

Best Regards,

Matt Brittin
Chairman of the Board and Managing Director,
Google United Kingdom
```

1 attachment: Google Notification Letter.pdf 367 KB

Spam image by Qwertyxp2000 from https://commons.wikimedia.org/wiki/File:Spam_can.png

Spam Classification

Y_i : `spam' or `nospam'

X_i: e-mails (pre-processed)

Spam Classification

Y_i : `spam' or `nospam'
X_i: e-mails (pre-processed)

- Trivial $\mathcal{F} = \{\text{always say `spam', always say `nospam'}\}$

Spam Classification

Y_i : `spam' or `nospam'
X_i: e-mails (pre-processed)

- Trivial $\mathcal{F} = \{\text{always say `spam', always say `nospam'}\}$
- $\mathcal{F} =$ all functions of the form

$$\text{IF } \%(\text{word}_1) > \theta_1 \text{ OR } \%(\text{word}_2) < \theta_2 \text{ THEN `spam'; ELSE `nospam'}$$

Spam Classification

Y_i : `spam' or `nospam'
X_i: e-mails (pre-processed)

- Trivial $\mathcal{F} = \{\text{always say `spam', always say `nospam'}\}$
- $\mathcal{F} =$ all functions of the form

$$\text{IF } \%(\text{word}_1) > \theta_1 \text{ OR } \%(\text{word}_2) < \theta_2 \text{ THEN `spam'; ELSE `nospam'}$$

- Empirical Risk Minimization:
  - Pick $\hat{f} \in \mathcal{F}$ with smallest nr. of mistakes on $\mathcal{T}$

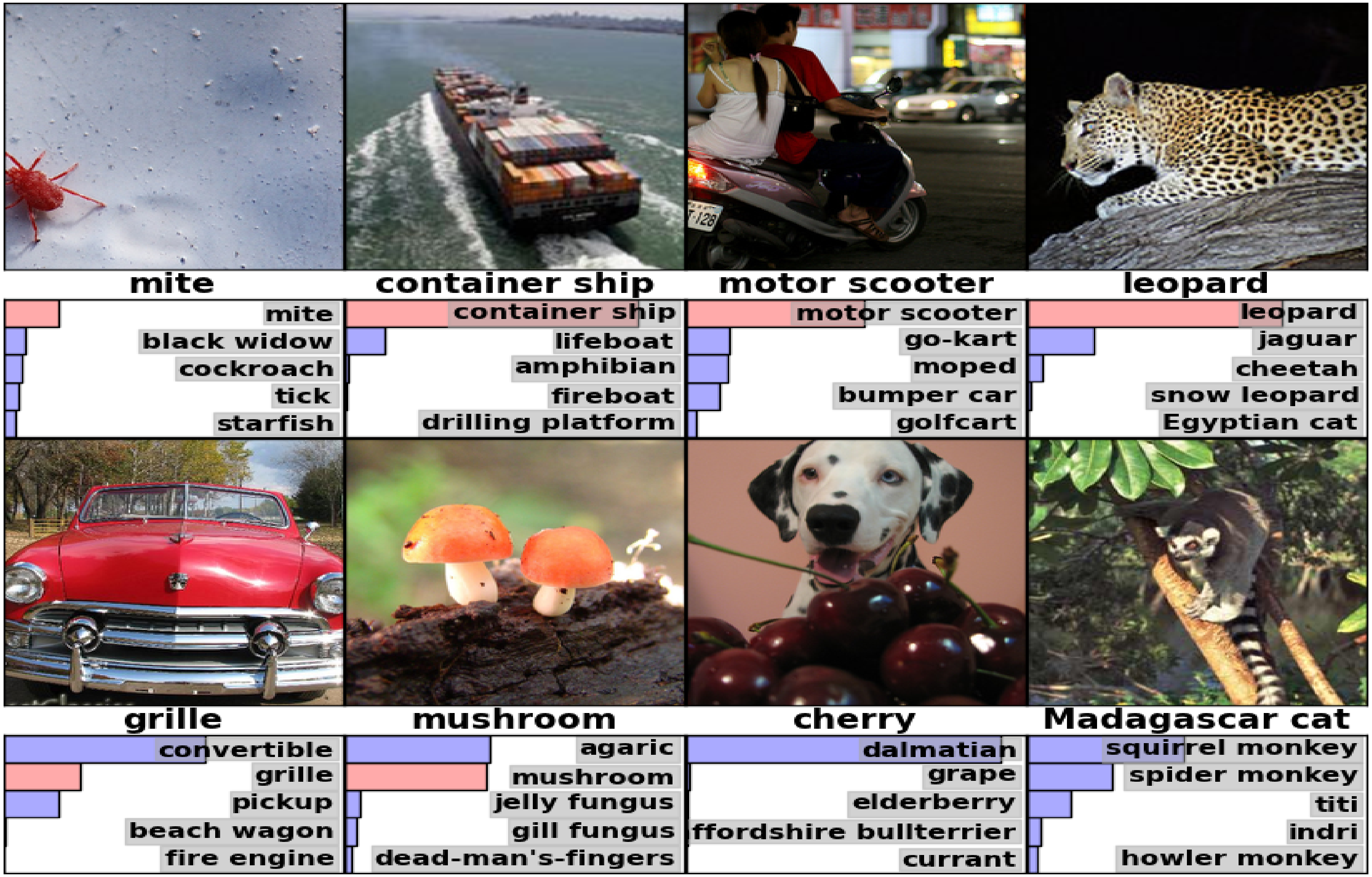
Handwritten Digit Classification

Y_i: 0, 1, 2, ..., 9
X_i: picture of one digit =
$k \times k$ matrix of grey-values

- Nearest neighbour:
  - measure distance of new picture to all pictures in $\mathcal{T}$
  - choose same class as closest picture

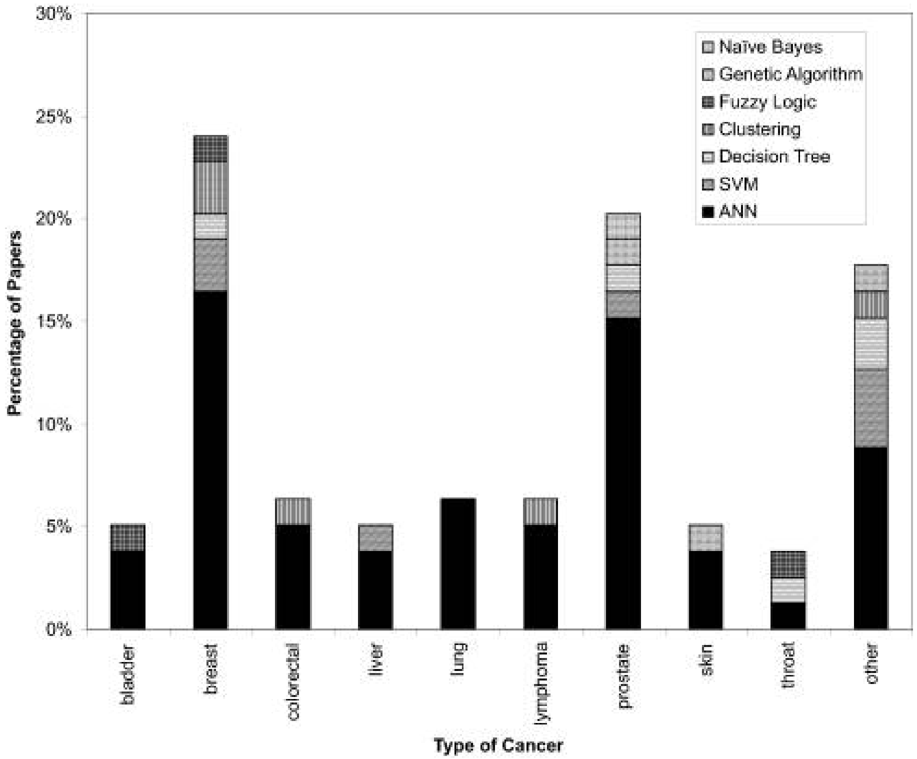
Hubert Eichner's Online Demo

General Image Classification



Krizhevsky, Sutskever, Hinton, ImageNet Classification with Deep Convolutional Neural Networks, NIPS 2012

Applications in Cancer Research



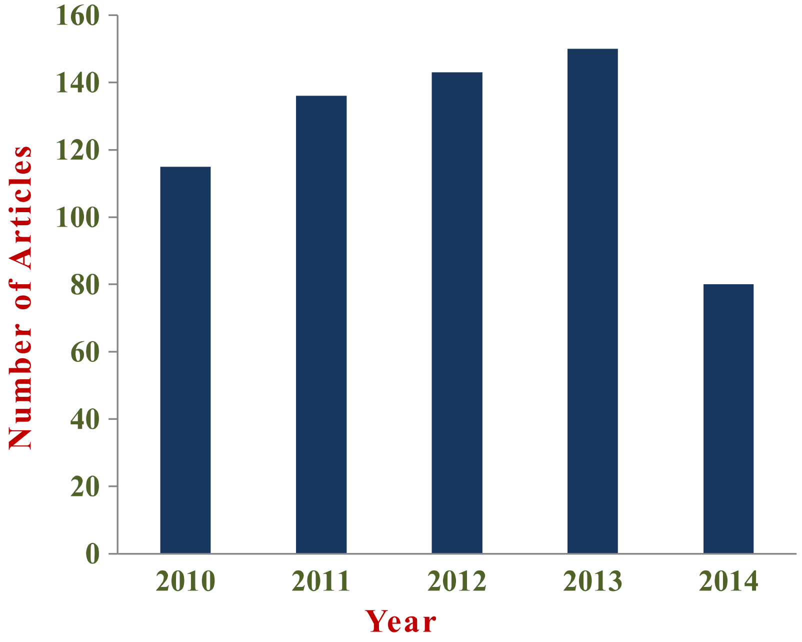


Fig. 7. Distribution of published studies, within the last 5 years, that employ ML techniques for cancer prediction.

Cruz, Wishart, Applications of Machine Learning in Cancer Prediction and Prognosis, Cancer Informatics, 2:59-77, 2006.

Kourou et al, Machine learning applications in cancer prognosis and prediction, Computational and Structural Biotechnology Journal, 13:8-17, 2015.

Regression: Prostate Cancer

- Goal: predict level of PSA (prostate specific antigen) for men with prostate cancer

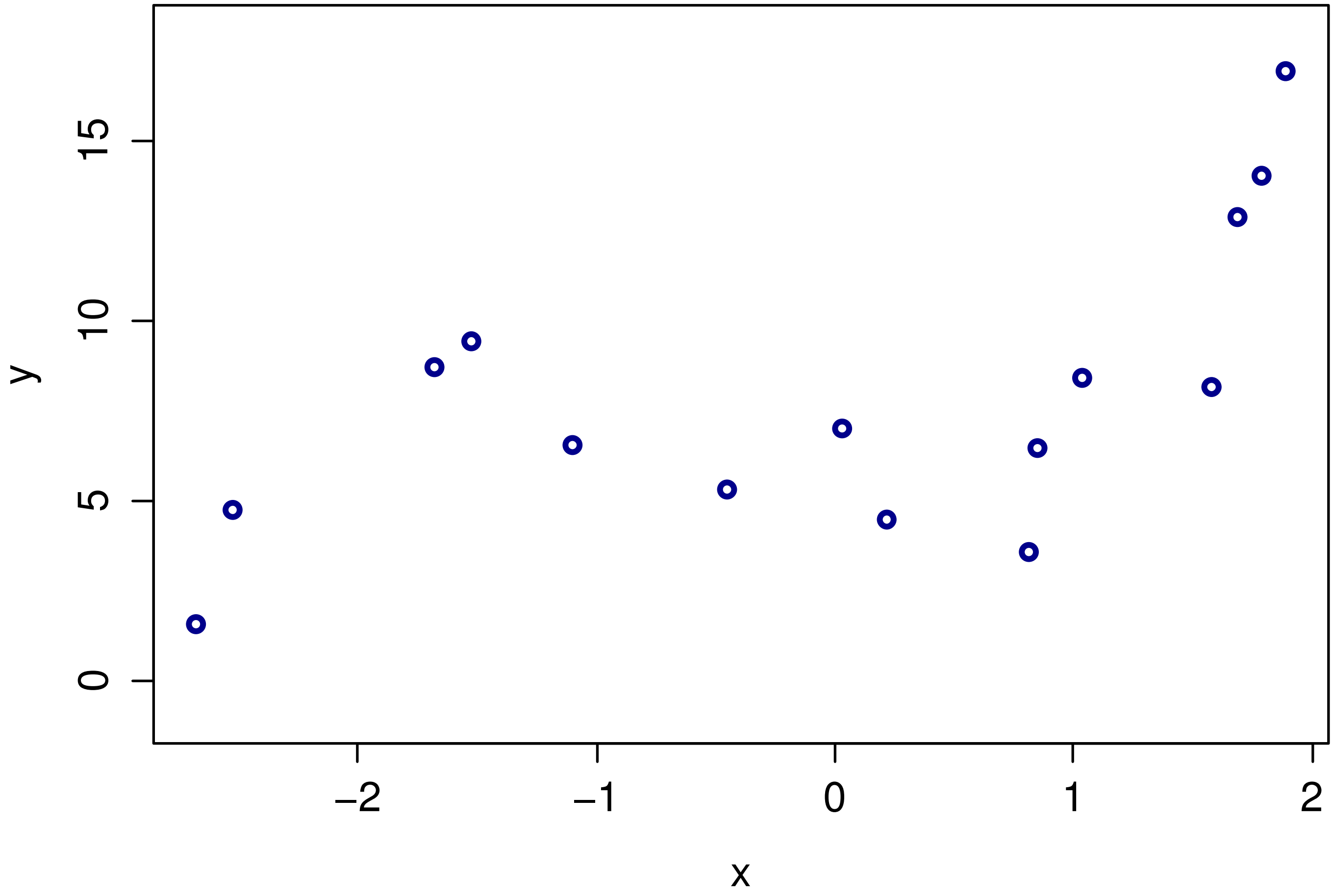
Y_i : **log psa**

X_i : **97 clinical measures**, including:

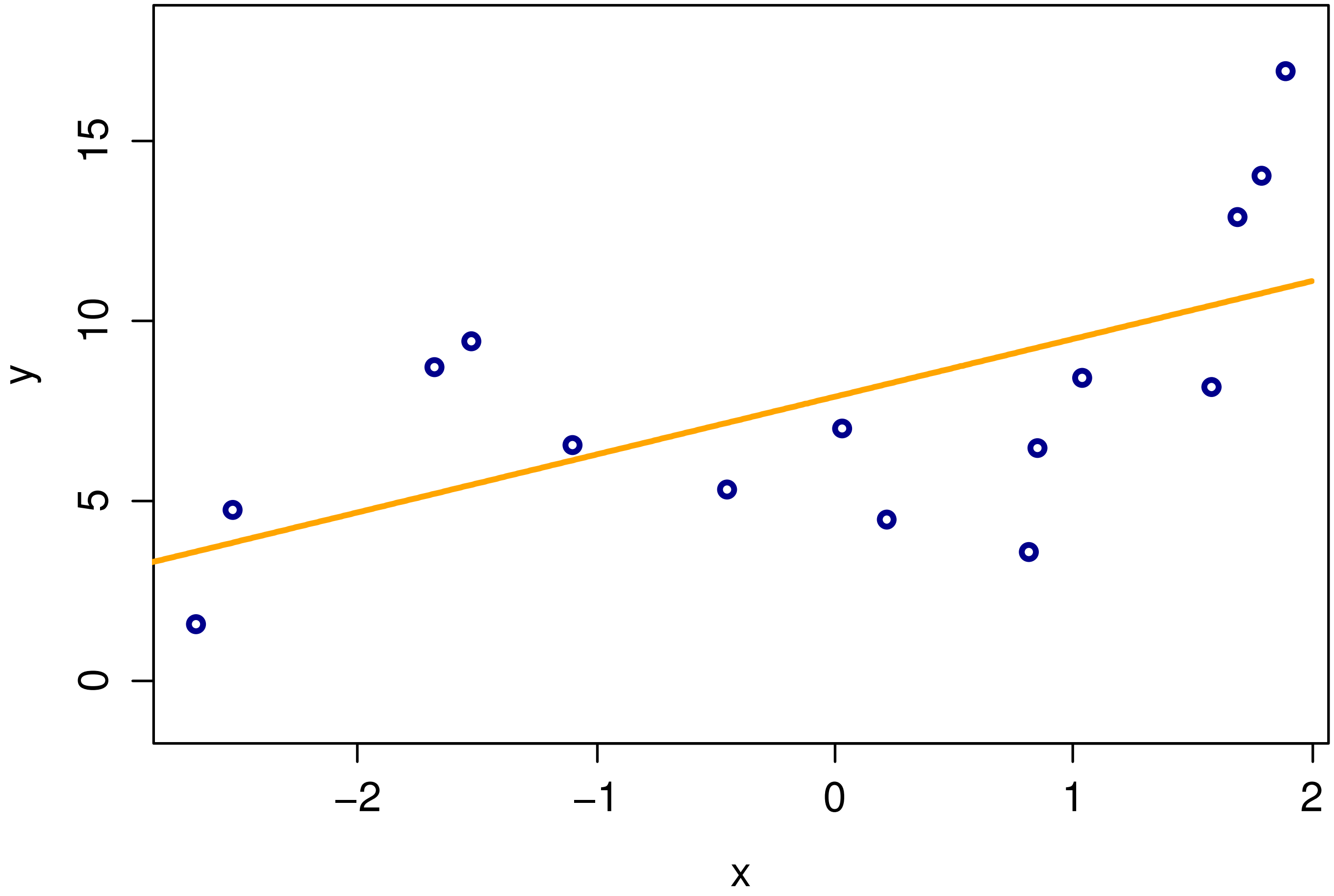
- log cancer volume
- log prostate weight
- Gleason score
- ...

See book for more details

Some Data

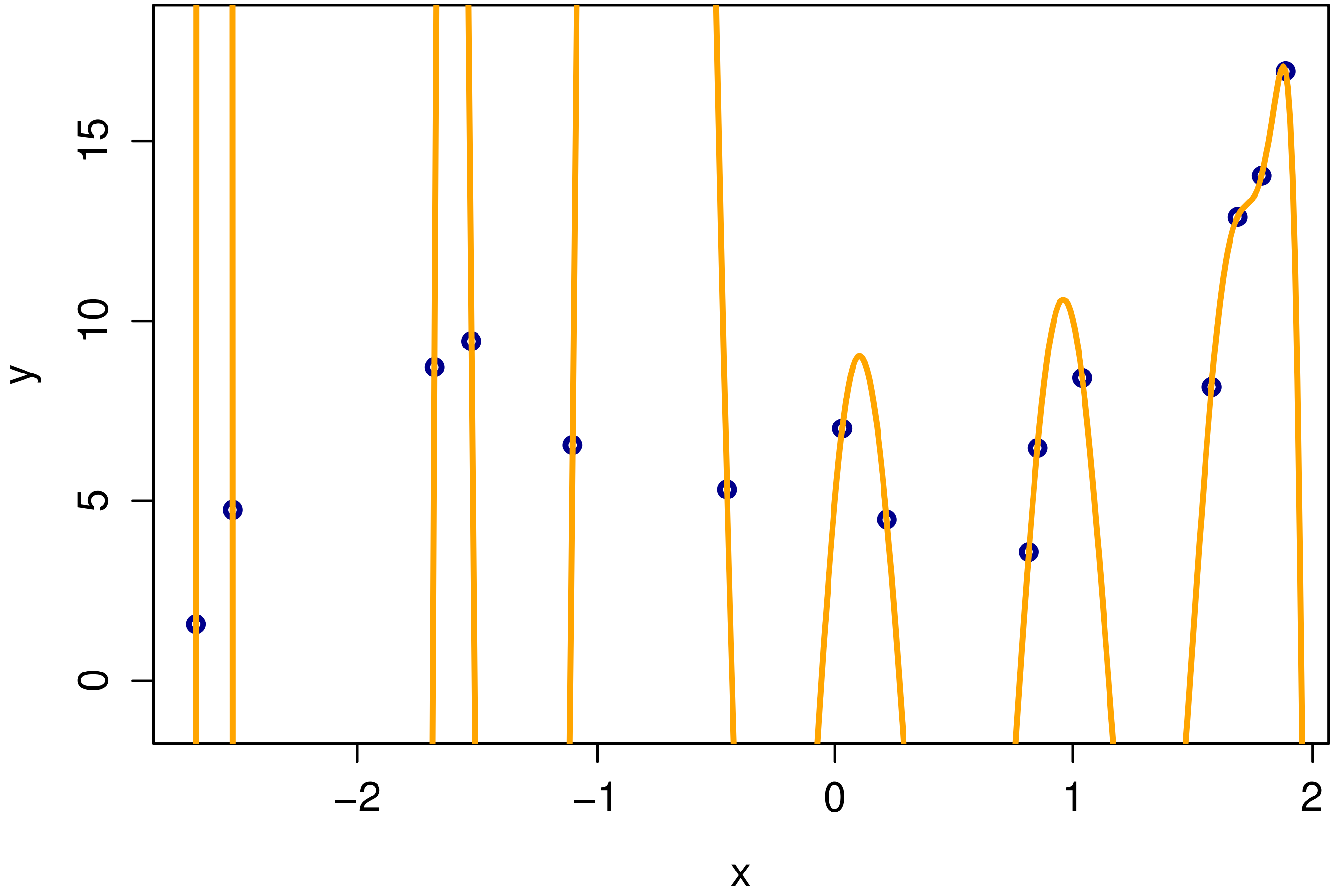


1-st Degree Polynomial



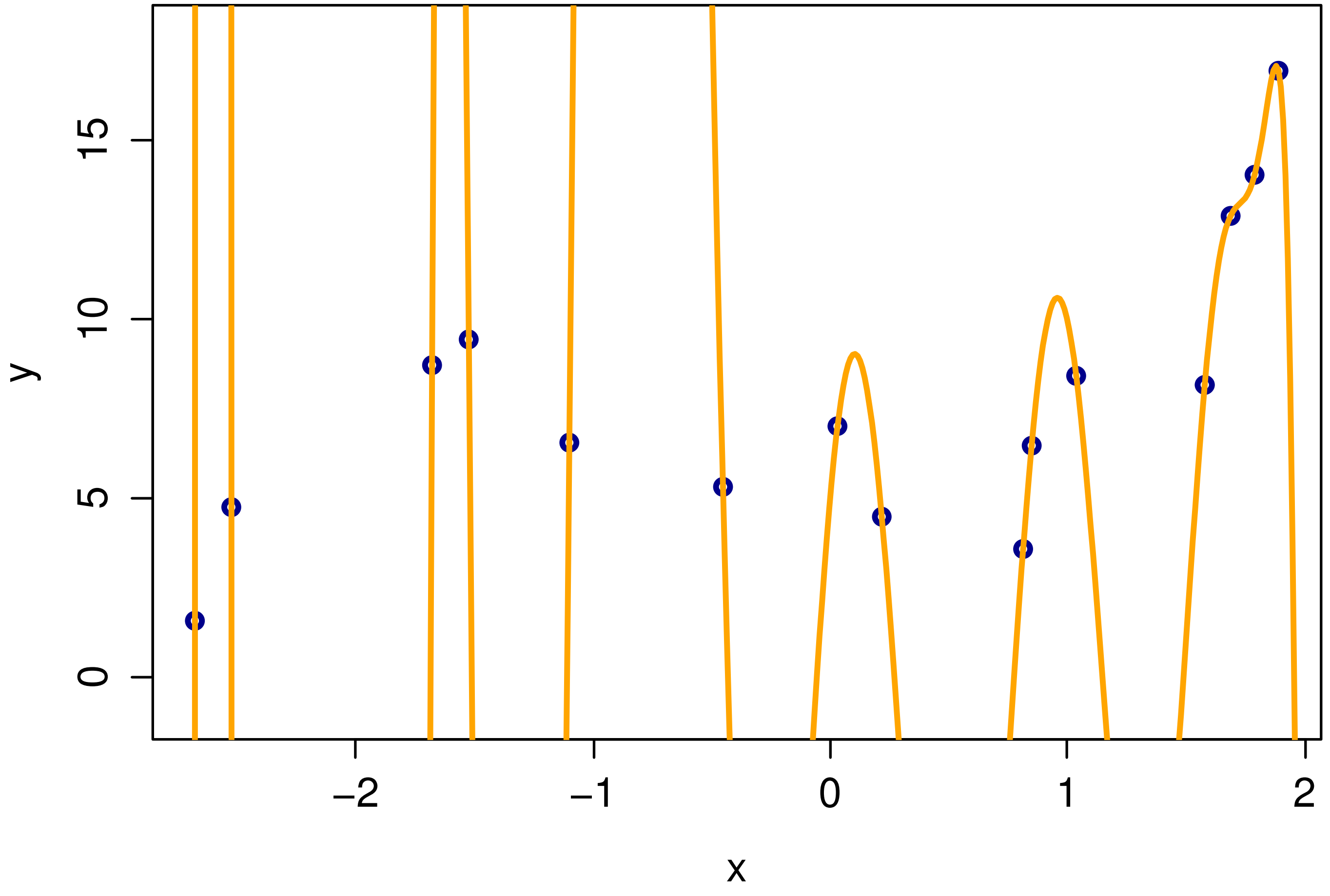
Underfitting: too few parameters

14-th Degree Polynomial: Perfect Fit on this Data



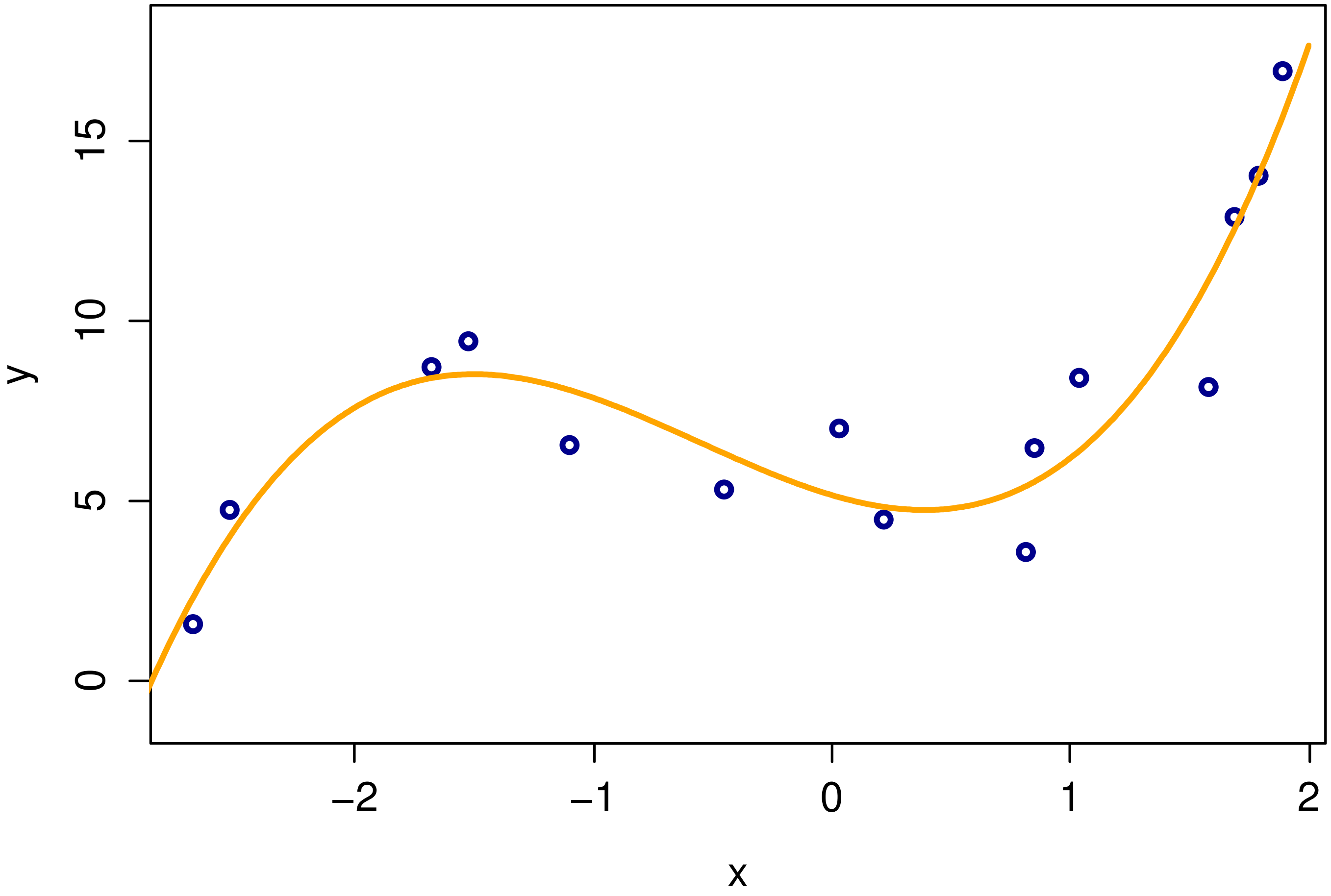
Smaller error on the data, so must be better, right?

14-th Degree Polynomial: Perfect Fit on this Data



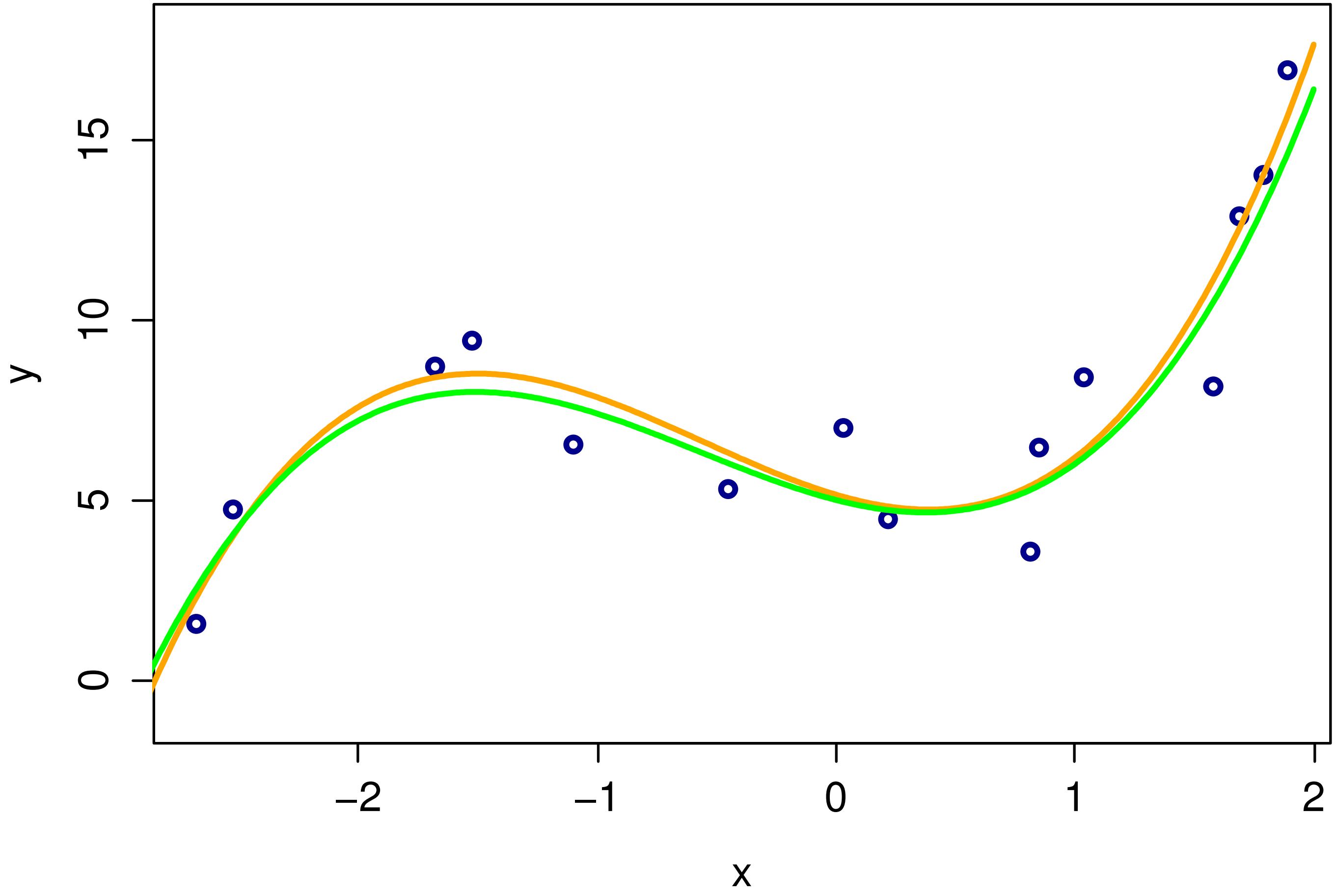
Overfitting: too many parameters!

3-rd Degree Polynomial



Intermediate nr. of parameters

3-rd Degree Polynomial



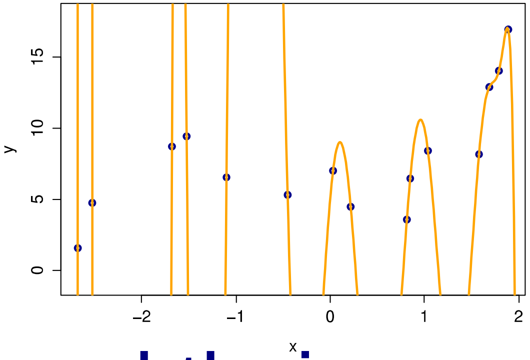
Intermediate nr. of parameters

Why Take This Course?

- Many machine learning methods available as software packages
- Try many different packages with many different parameter settings, and select the one that gets the smallest error on your data

Why Take This Course?

- Many machine learning methods available as software packages
- Try many different packages with many different parameter settings, and select the one that gets the smallest error on your data
- Congratulations, you have **overfit** your data and your method predicts poorly on new data[1]



- This course: understand methods and their parameters, learn proper techniques to select parameters

[1] You may also have **underfit** your data, e.g. because you did not construct more features