Machine Learning & Data Mining: Conclusion

J. Savoy
University of Neuchatel

I. H. Witten, E. Frank, M.A. Hall: Data Mining. Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

T. Hastie, R. Tibshirani, J. Friedman: The Elements of Statistical Learning. Springer, New York

C.M. Bishop: Pattern Recognition and Machine Learning. Springer.

References: to go further

- I. H. Witten, E. Frank, M.A. Hall: Data Mining. Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2011.
- A. Rajaraman, D. Ullman: Mining of Massive Datasets. Cambridge University Press, 2012.
- P. Flach: Machine Learning. The Art and Science of Algorithms that Make Sense of Data. Cambridge University Press, 2012.
- T. Mitchell: Machine Learning. McGraw Hill, 1997.
- C.M. Bishop: Pattern Recognition and Machine Learning. Springer, 2006.
- T. Hastie, R. Tibshirani, J. Friedman: The Elements of Statistical Learning. Springer, New York, 2009
- S. Chakrabarti: Mining the Web. Discovering Knowledge from Hypertext Data. Morgan Kaufmann, 2003.
- M. Bramer: Principles of Data Mining. Springer, 2007.

Terminology

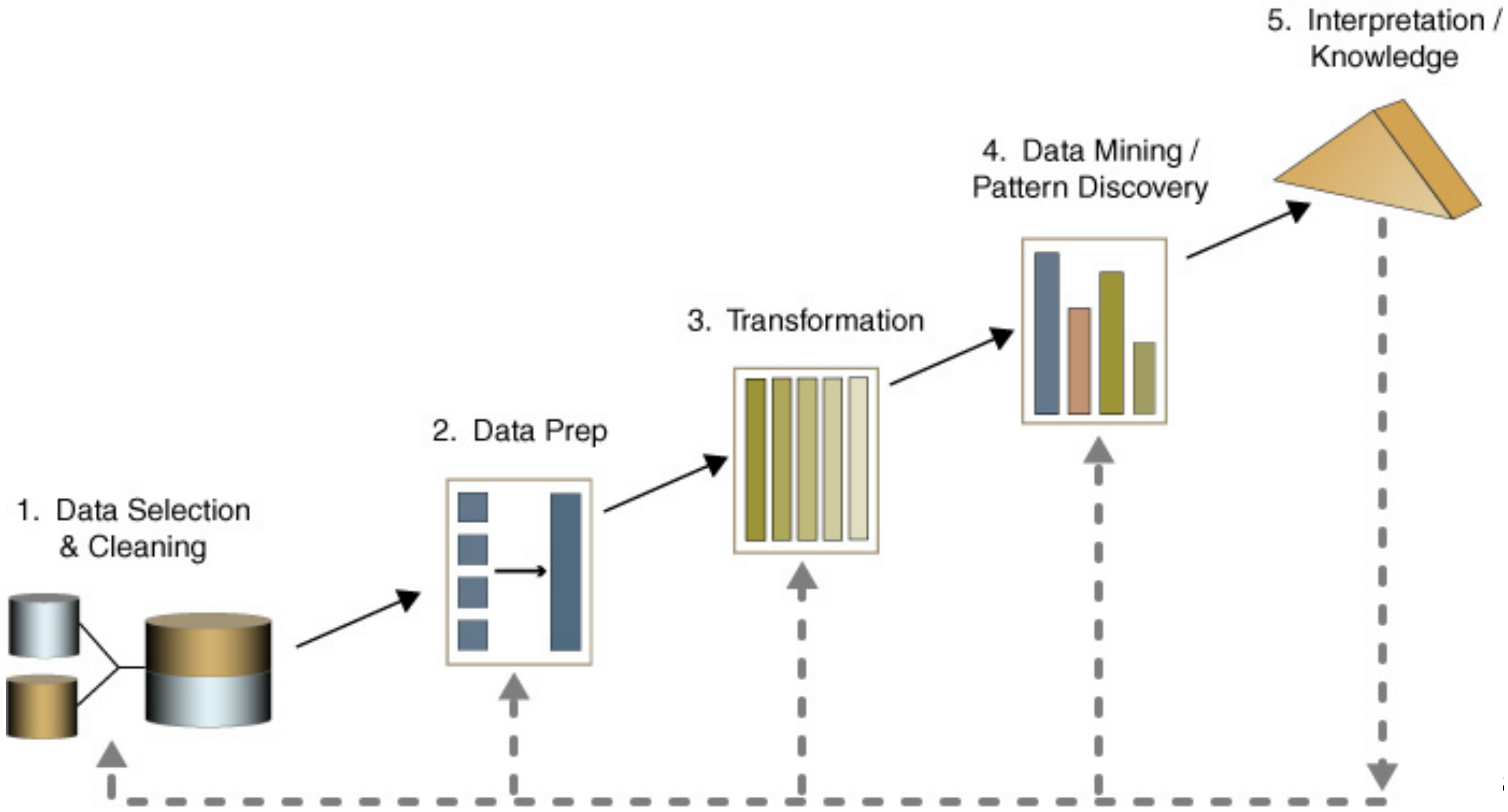
- CS: input $\rightarrow$ model $\rightarrow$ output
- Statistician: independent var. $\rightarrow$ model $\rightarrow$ dependant var.
- ML: features $\rightarrow$ model $\rightarrow$ responses
- ML: predictors $\rightarrow$ function approximation $\rightarrow$ responses
- Nature:
  - nominal (discrete, categorical)
  - ordinal (ordered categorical)
  - quantitative
- Response:
  - classification (qualitative)
    binary (numeric code or target)
  - regression (quantitative)

Main Models / Chapters

- 1R
- Bayes learning
- Decision Tree
- Associations rules
- Attribute Selection
- Nearest Neighbors (*k*-NN)
- Nearest Neighbors Search (minhash)
- Linear Models: Winnow
- Ensemble Learning
- Clustering
- Evaluation

Real Data, Real Life

A whole process is often hidden



An Overview of the Steps That Compose the KDD Process

Attribute Selection

Need to be able to predict the decision class

Cleaning the data

- Errors?
- Outliers?
- Variable types?
- Statistical relationship or causality?

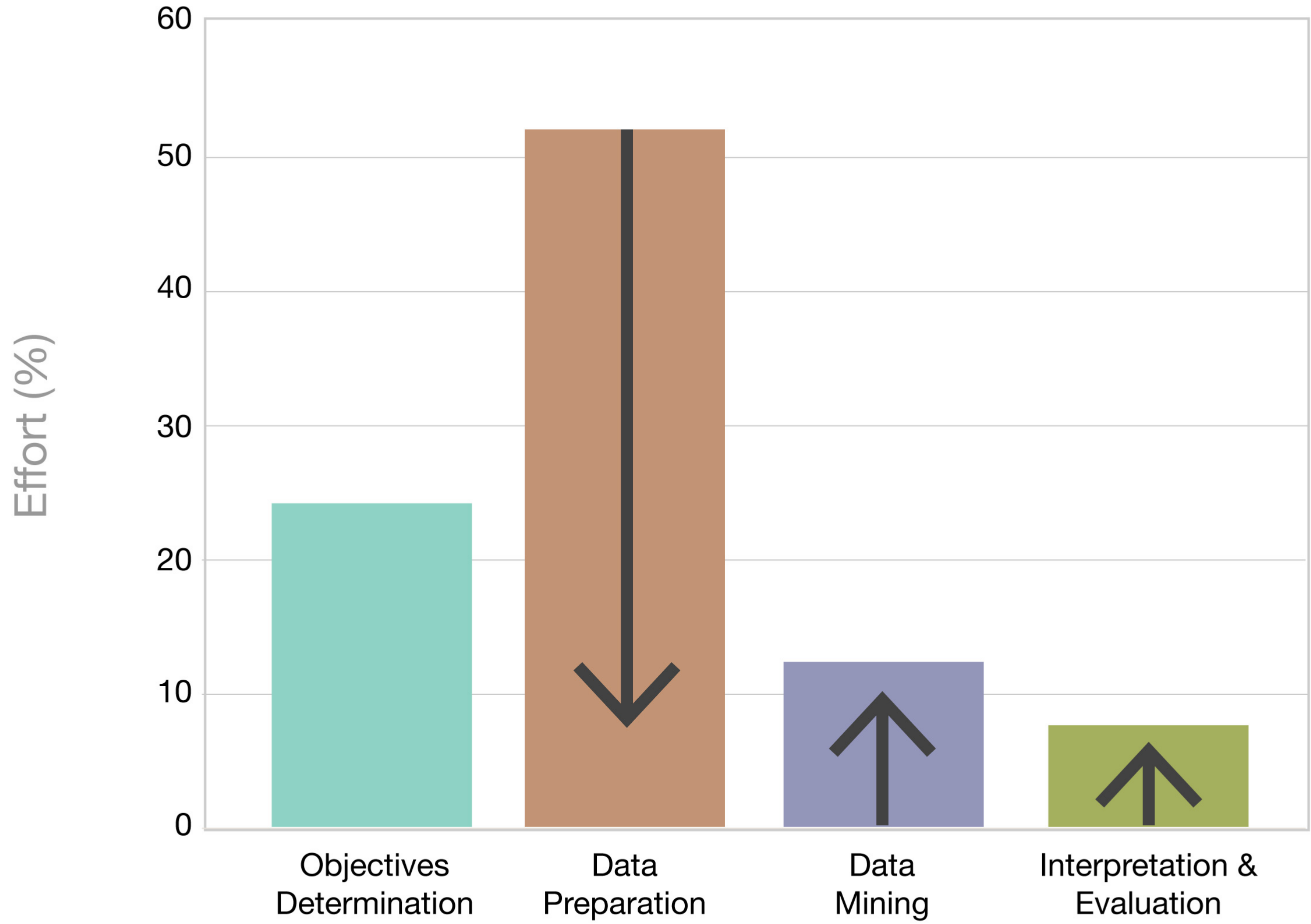
Attribute Selection

Need to be able to predict the decision class
Not always easy to verify the (statistical) relationship between attributes and the decision

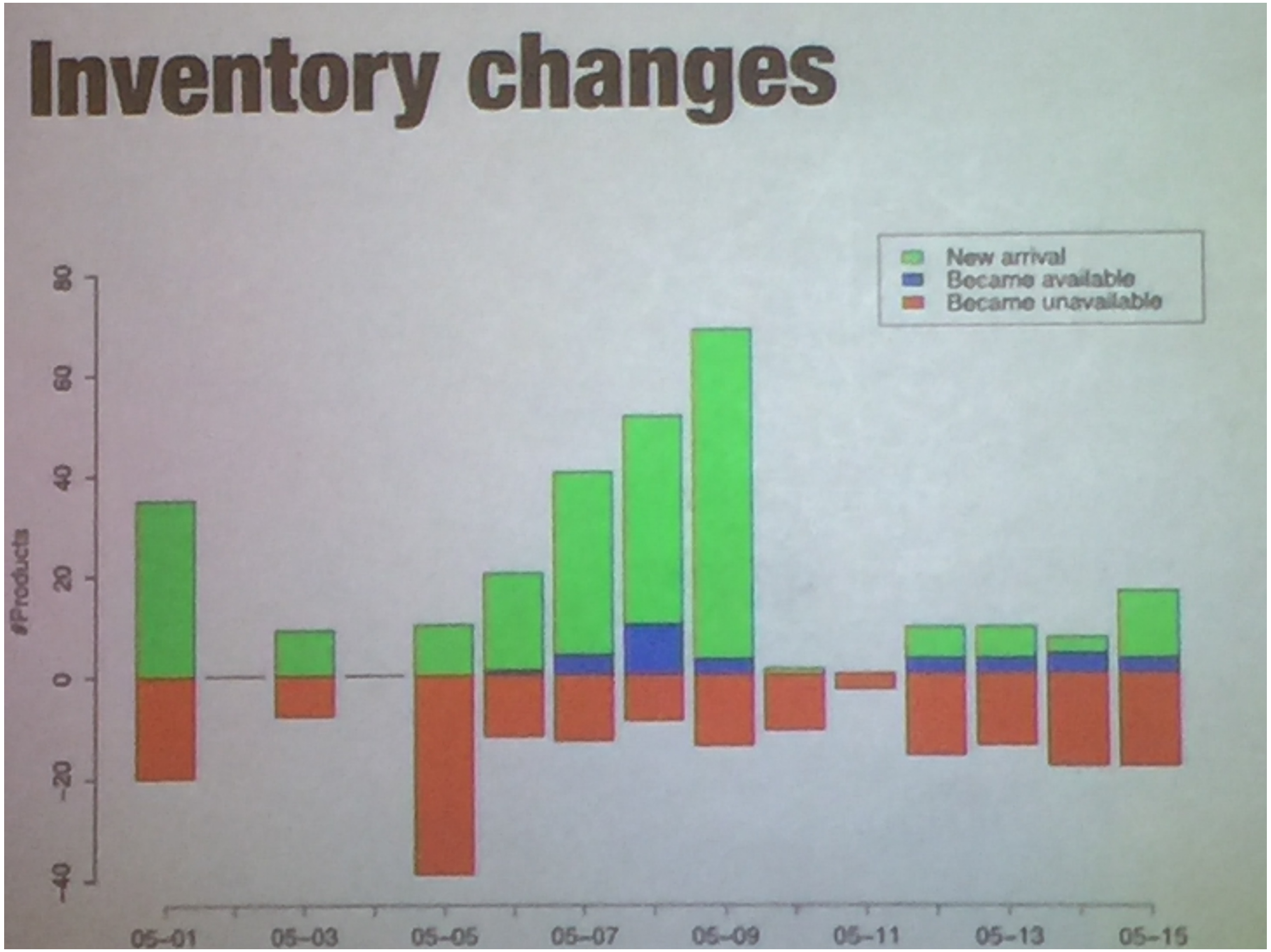
A	B	C	D	E	F	Category
2	blue	0	1	0	1	1
5	red	1	0	1	0	1
6	red	0	1	0	0	0
3	green	1	0	0	1	0

Real Data: What do you Learn

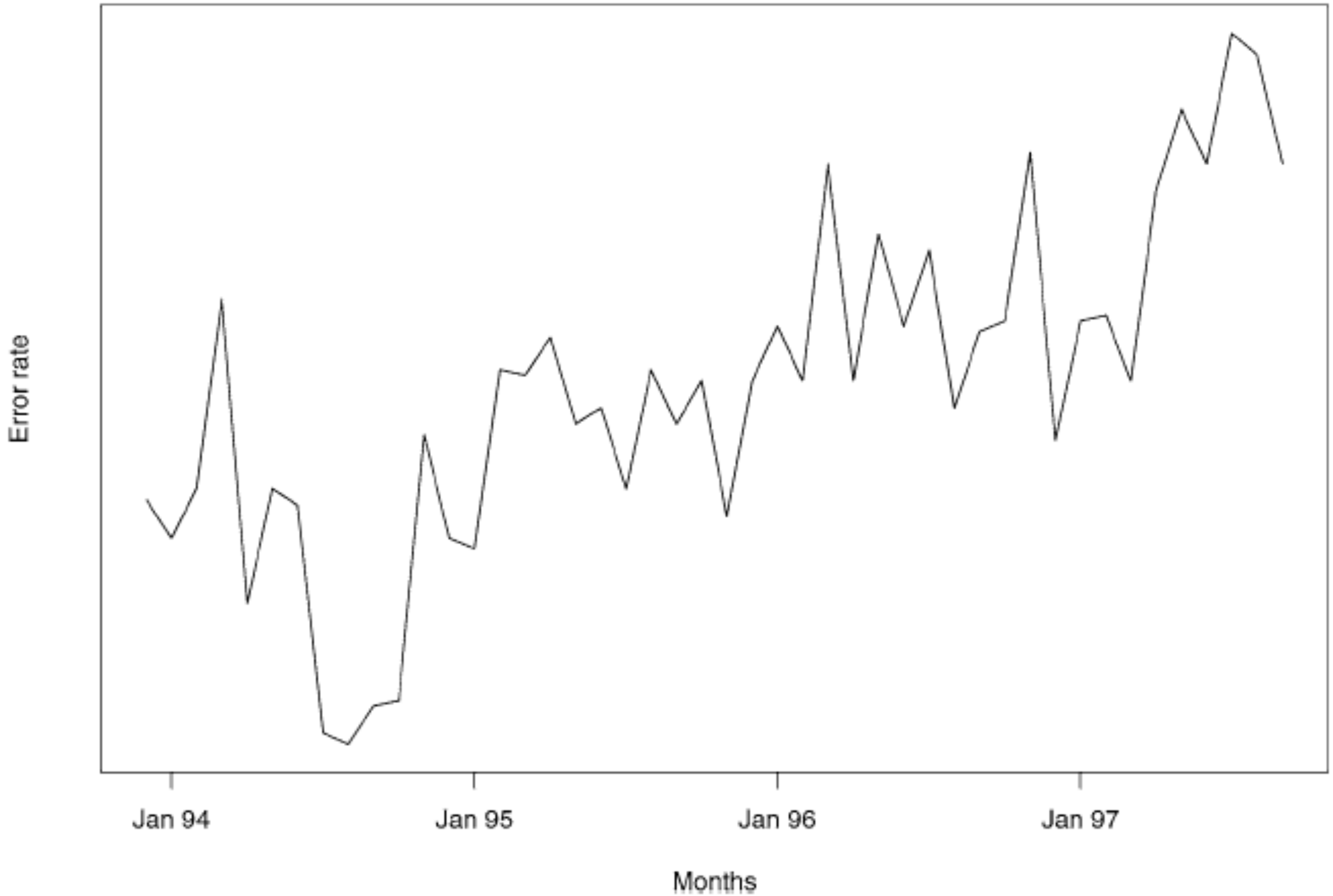
Arrows indicate the direction we hope the effort should go



Real Data: Change quickly



Real Data: Background changes



Hand, D.J.. (2006). Classifier Technology and the Illusion of Progress. *Statistical Science, 21(1), 1-14*.

Model Selection

Binary Response (yes or no)

Set of instances with the corresponding (correct) decisions. Each instance owns two attributes (x_1, x_2), real values.

The pertinent questions:

- Which model can you apply? Why?
- Which one is the best? Why?
- Which is the complexity of your representation?
- All attributes are useful?

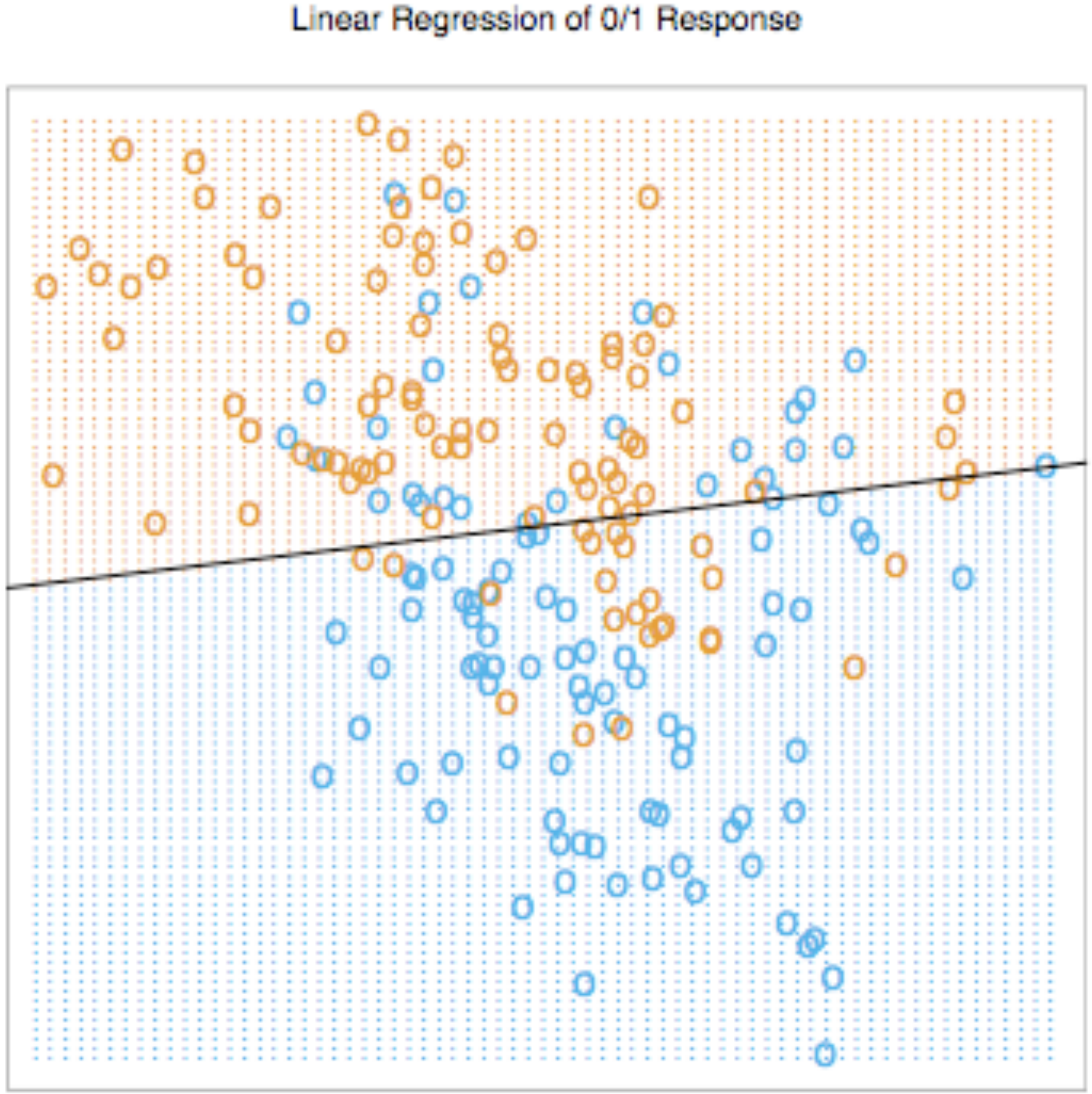
Model Selection

Binary Response
(orange vs. blue)

Linear Regression
(draw a line to split
the two classes)

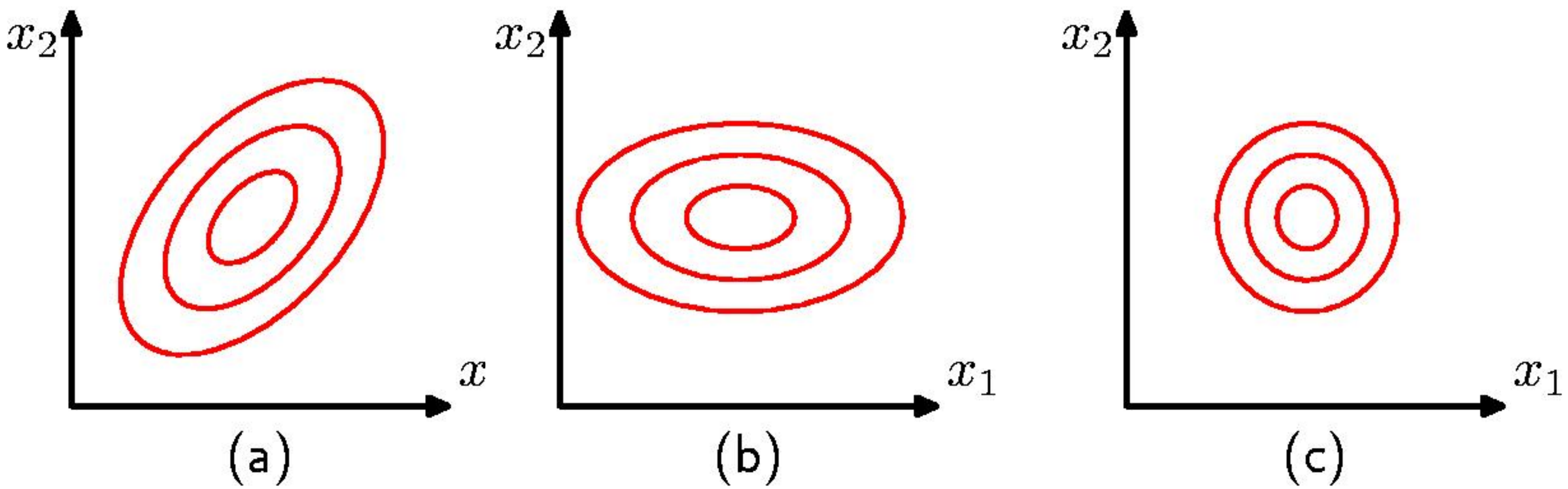
No error is an utopia!

T. Hastie, R. Tibshirani, J. Friedman:
The Elements of Statistical Learning.
Springer, New York



Model Selection

We have two variables (x_1, x_2) following a Gaussian distribution. We show the contour of constant probability density and in (a) we have the general case (positive covariance). In (b) the covariance is zero, but $\sigma_{x1} \neq \sigma_{x2}$. In (c), the two variables have the same variance.



Model Selection

Binary Response
(orange vs. blue)

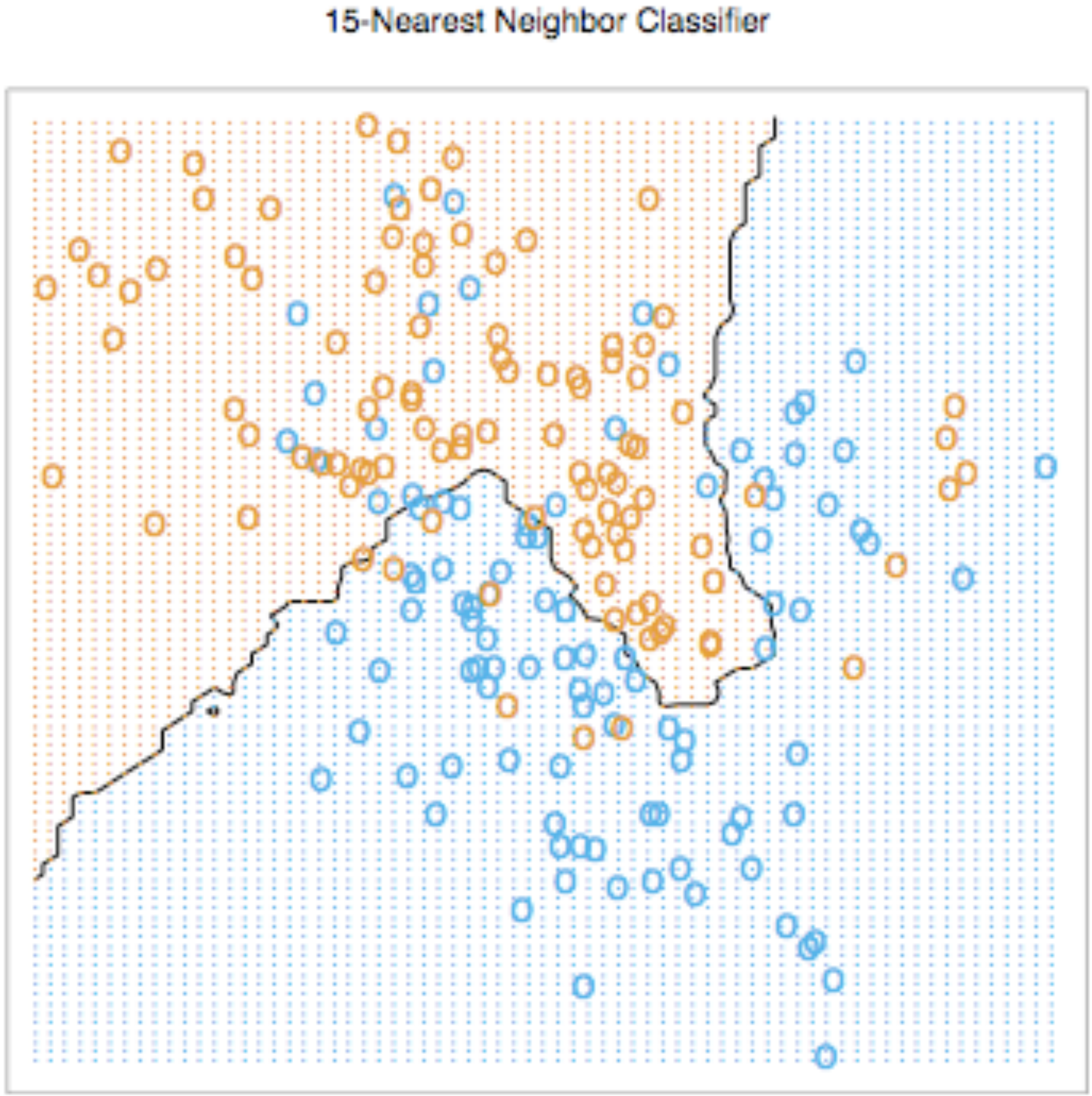
Possible scenario (data model)

1. The data on each class are generated from a bivariate Gaussian distributions with uncorrelated components and different means.
2. The data in each class came from a mixture of 10 low-variance Gaussian distributions, with individual means themselves distributed as Gaussian
3. …

Model Selection

Binary Response

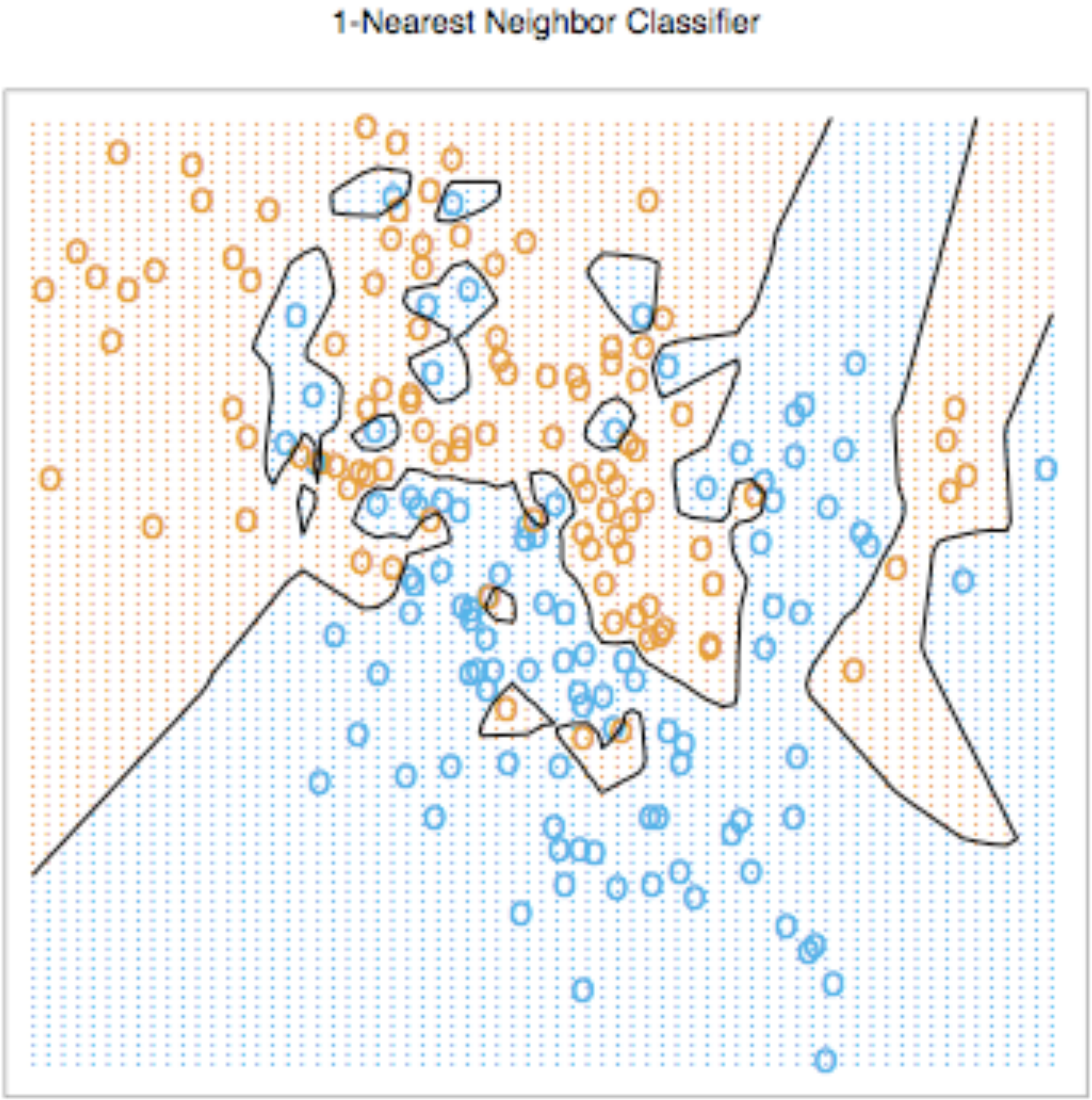
15-Nearest Neighbors



Model Sele

Binary Response

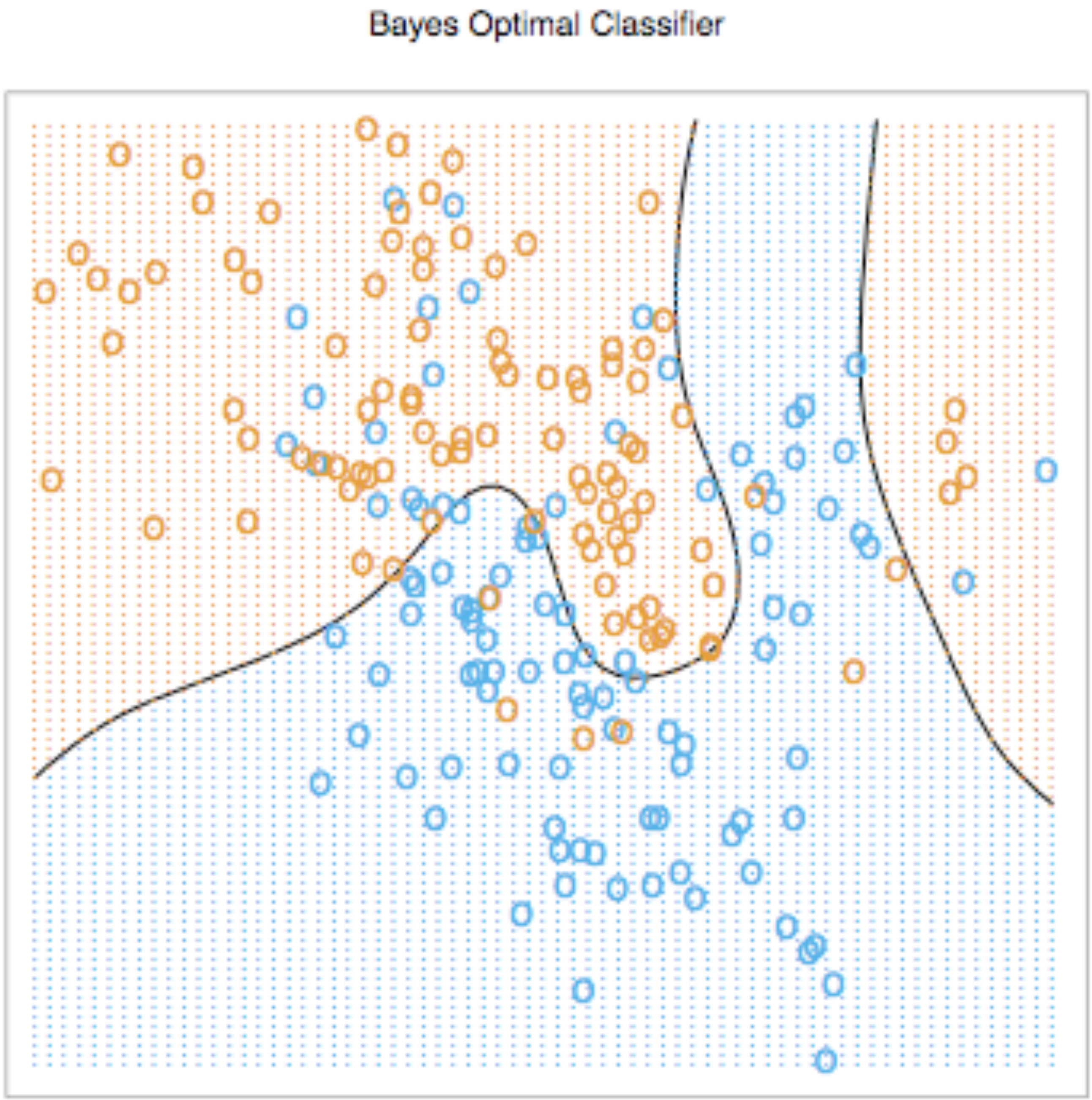
1-Nearest Neighbor



Model Selec

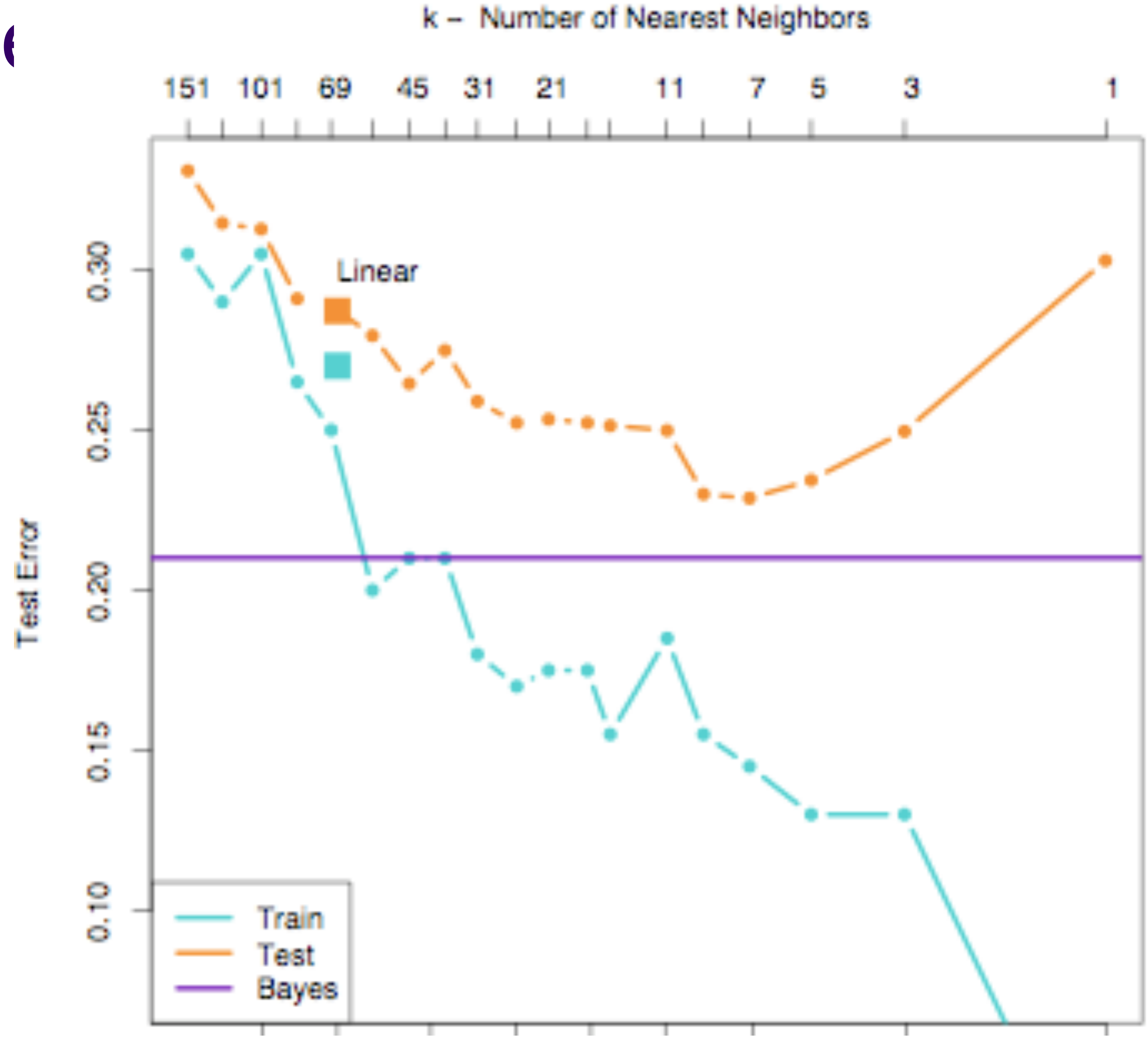
Bayes Classifier

Borders more smooth



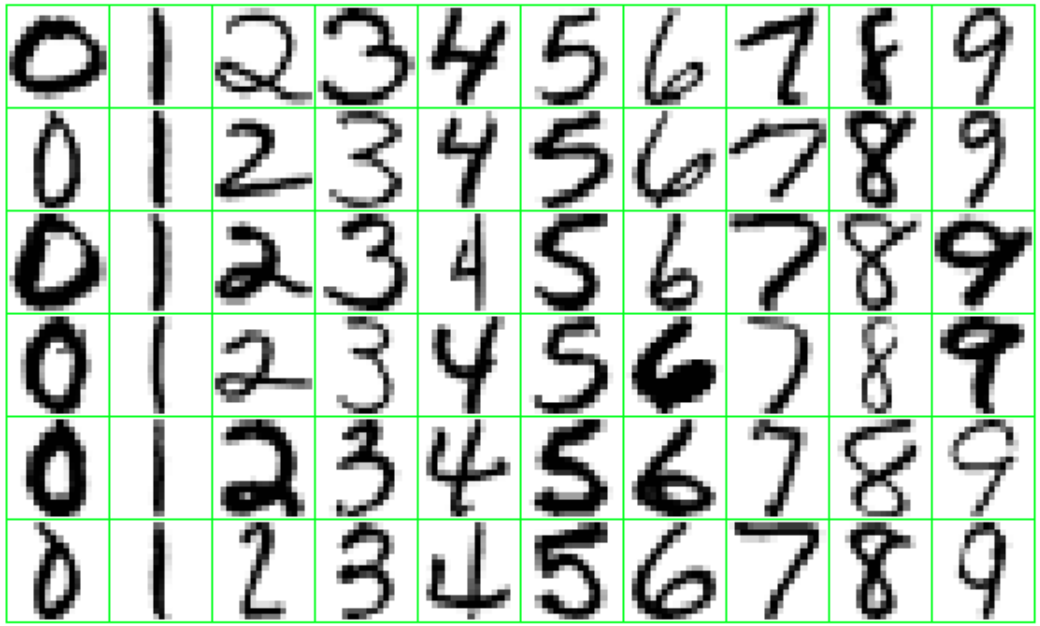
Model Se

Test Error
10,000 test
200 training



What does that mean ?

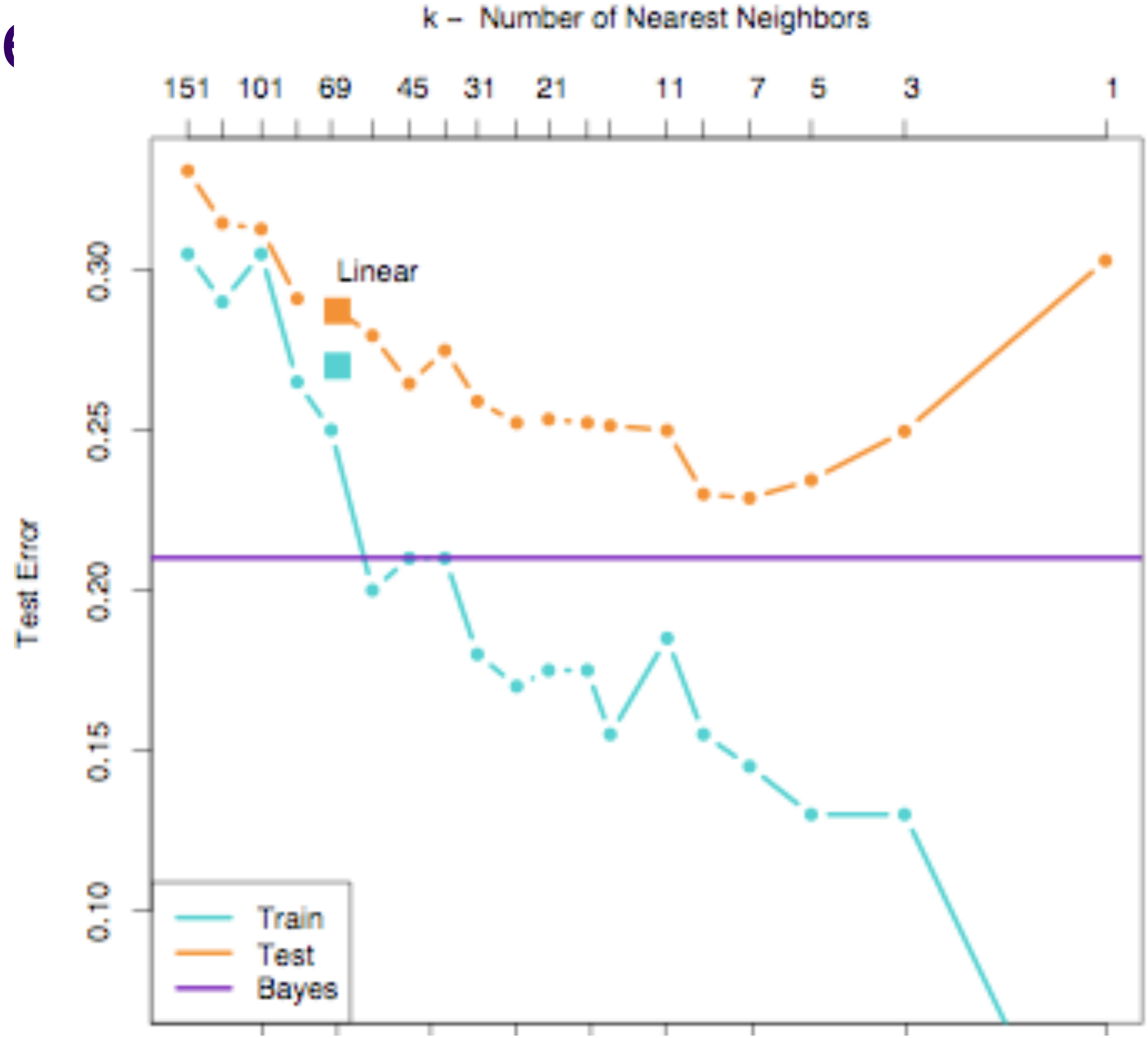
An error rate of 0.2?



For the Swiss postal code (4 digits)?
For a credit card (20 digits)?

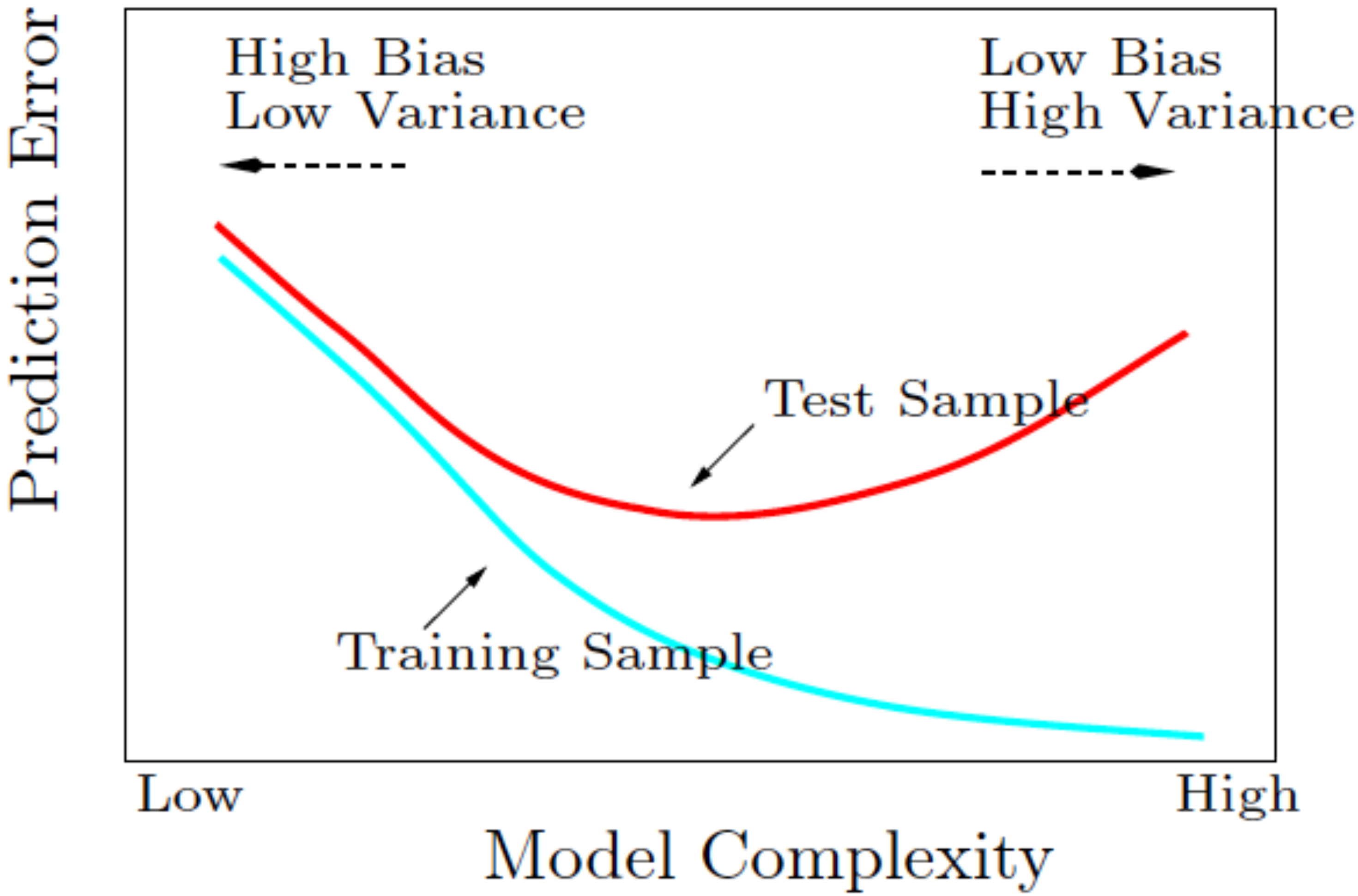
Model Se

Test Error
10,000 test
200 training



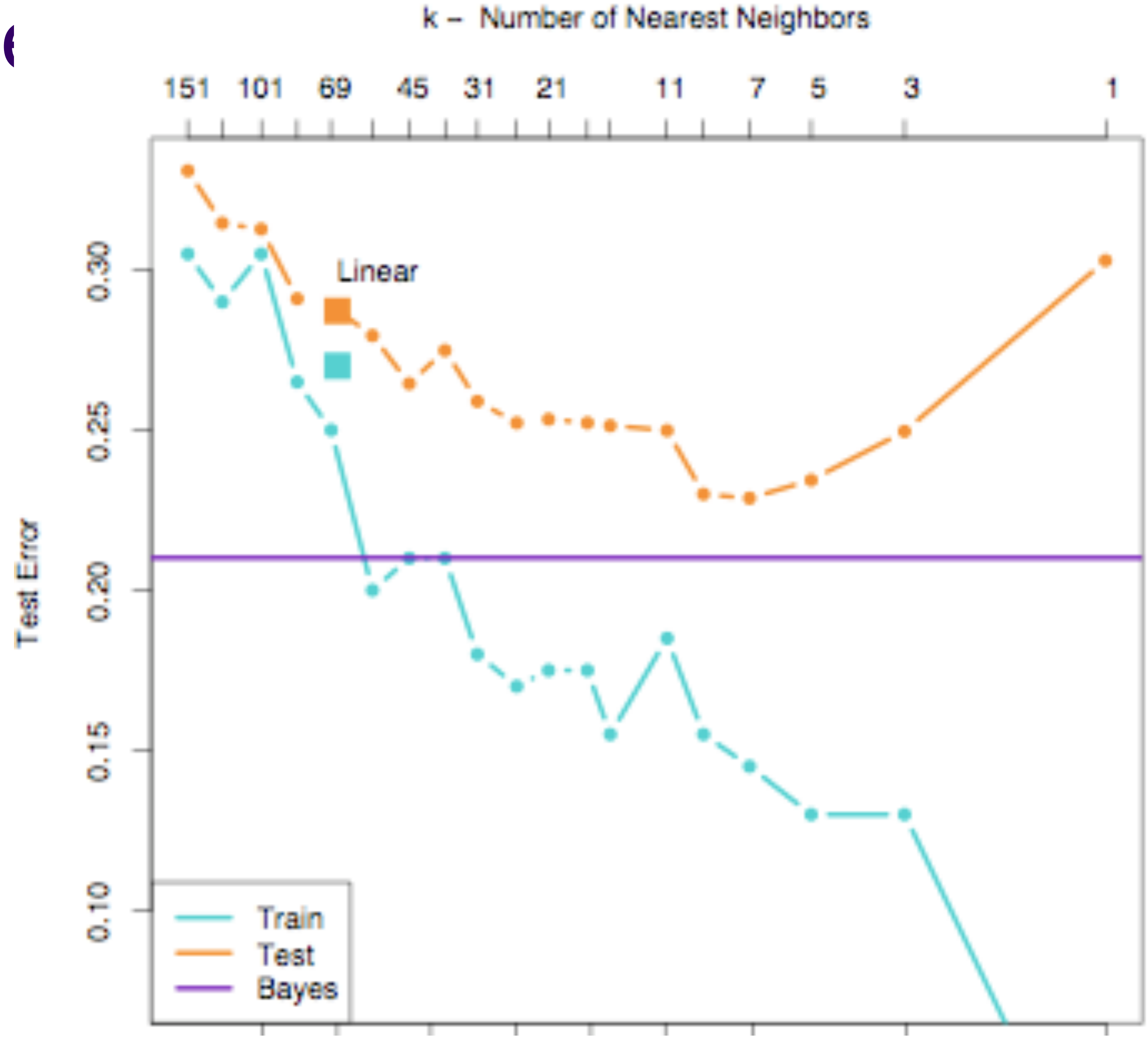
Model Complexity

Error rate and Complexity (over-fitting)
Linear model has high bias
Decision trees suffer from high variance



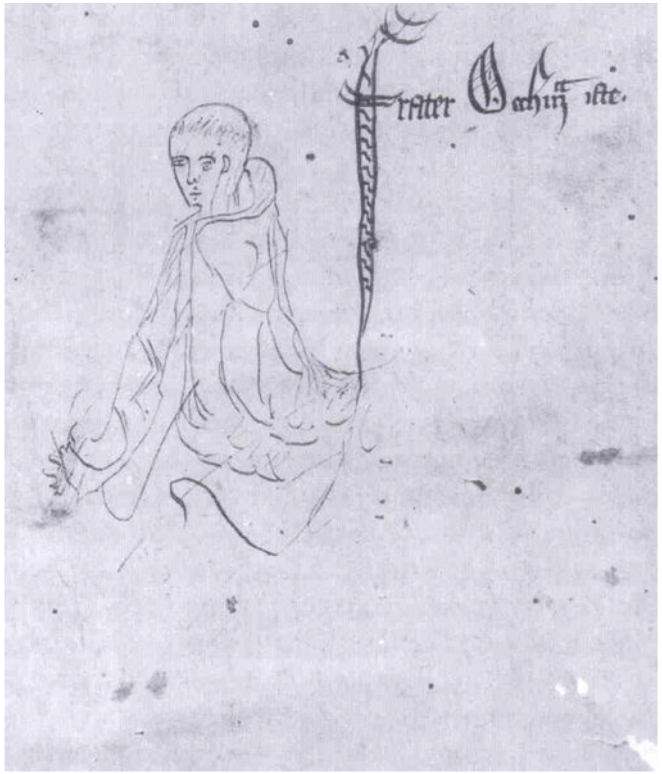
Model Se

Test Error
10,000 test
200 training



Example

- Ockham’s razor: prefer the simplest hypothesis consistent with data. This principle proposed by William of Ockham in the fourteenth century: "Pluralitas non est ponenda sine neccesitate", which translates as "entities should not be multiplied unnecessarily"



Conclusion

- More models than presented in the course
- Many variants of the models
- You have the needed background
- Over-fitting
- Cold start problem (we assume that a training set exists)
- Do not consider that the training sample is error-free (error in the data)
- Many applications
- More than one model can be applied in a given case, some returning similar performance

Conclusion

1. Learning = Representation + Evaluation + Optimization
2. It's generalization that counts
3. Data alone is not enough (prior knowledge)
4. Over-fitting has many faces (bias, variance, regularization term)
5. Intuition fails in high dimensions (curse of dimensionality)
6. Theoretical guarantees are not what they seem (not fully helpful for practical considerations)

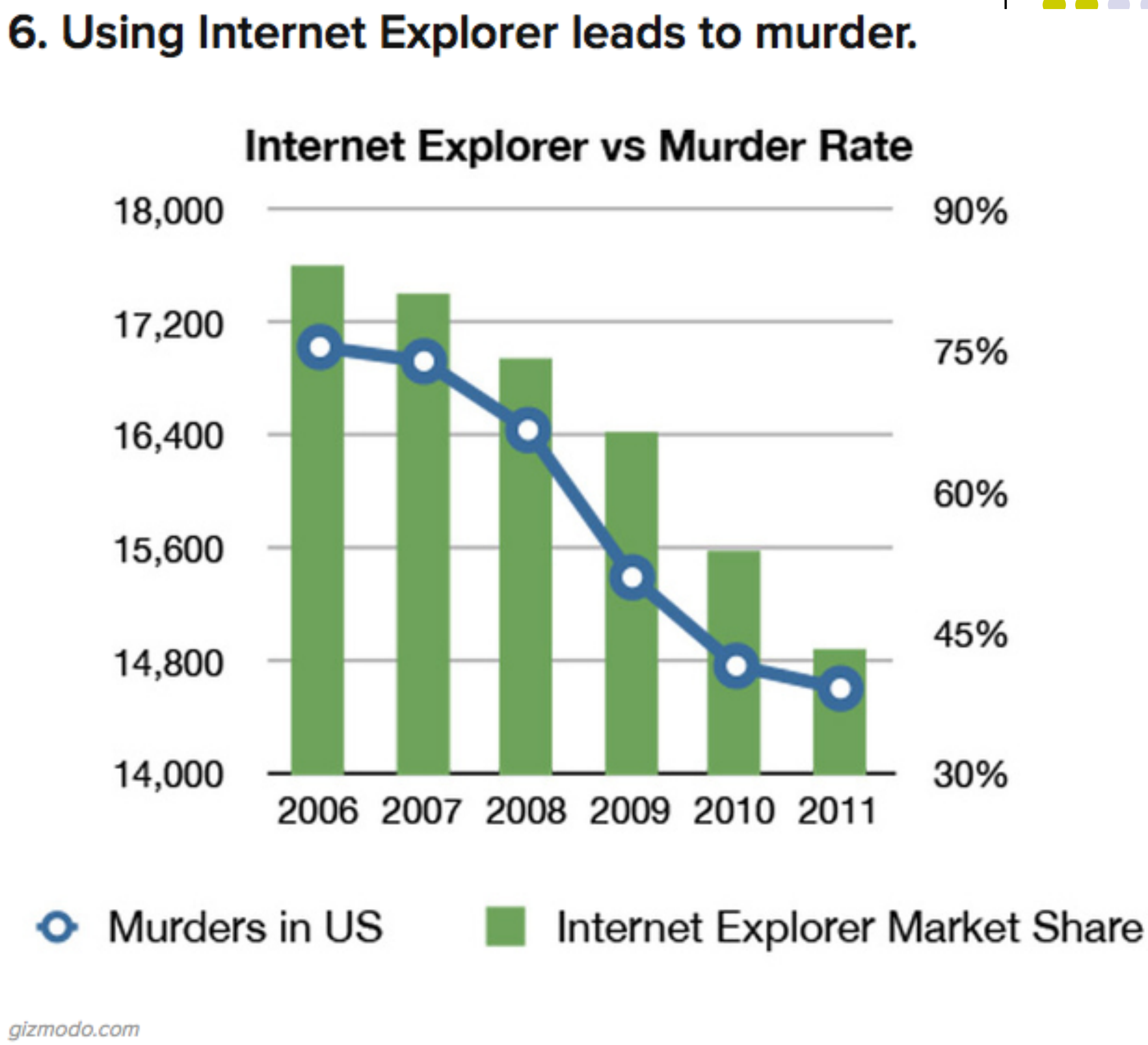
P. Domingos (2012). A few useful things to know about machine learning. CACM, 55(10), 78-87

Conclusion

8. Feature engineering is the key
(which features, creativity, black art in ML)
9. More data beats a cleverer algorithm
(complex models need too many data, thus simplify!)
10. Learn many models, not just one
(best classifier is application dependent)
11. Representation does not imply learnable
(can be learned is not a guarantee for a representation)
12. Correlation does not imply causation

P. Domingos (2012). A few useful things to know about machine learning. CACM, 55(10), 78-87

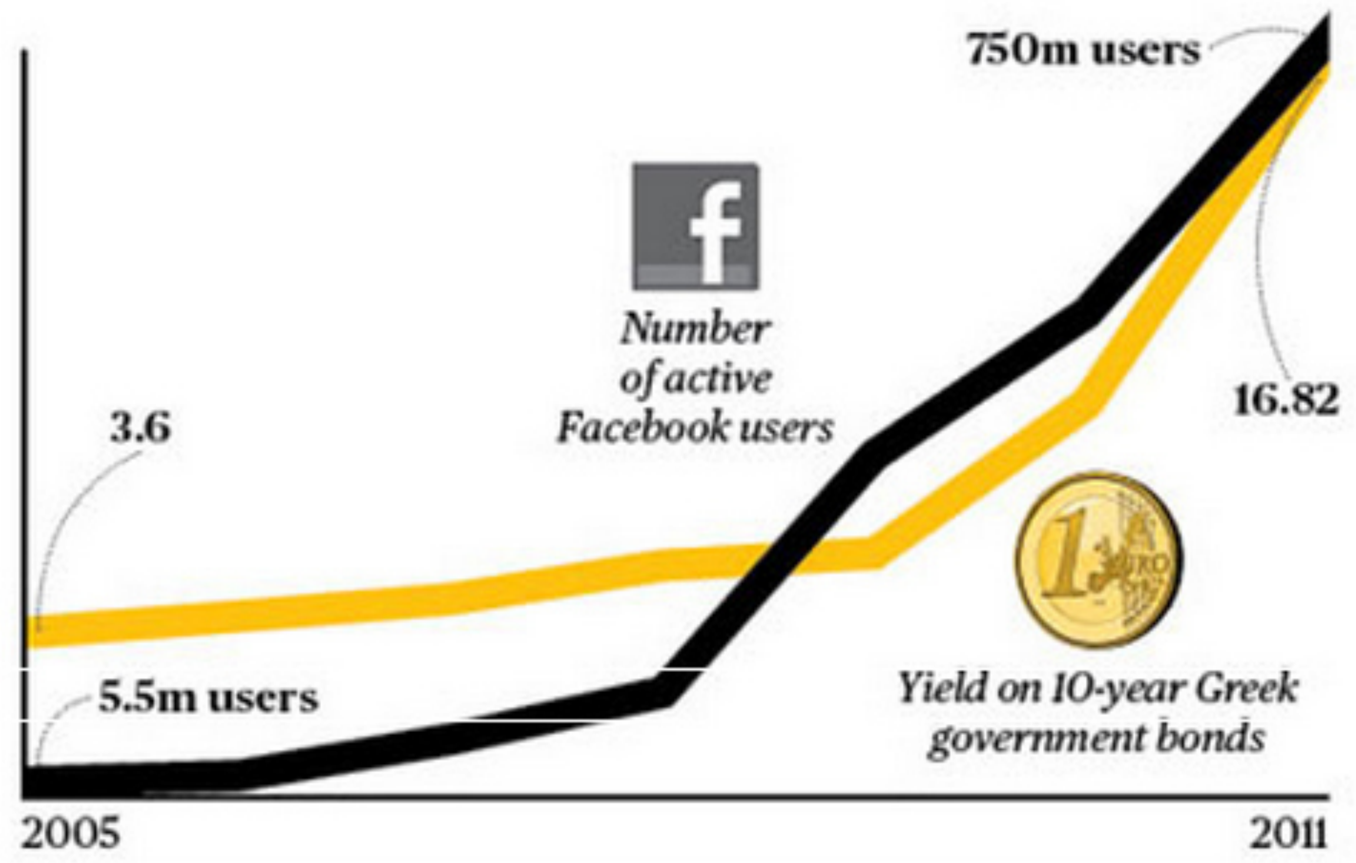
Example



6. Using Internet Explorer leads to murder.
Internet Explorer vs Murder Rate
18,000
17,200
16,400
15,600
14,800
14,000
90%
75%
60%
45%
30%
2006
2007
2008
2009
2010
2011
Murders in US
Internet Explorer Market Share
gizmodo.com

Example

9. Facebook caused the Greek debt crisis.



businessweek.com