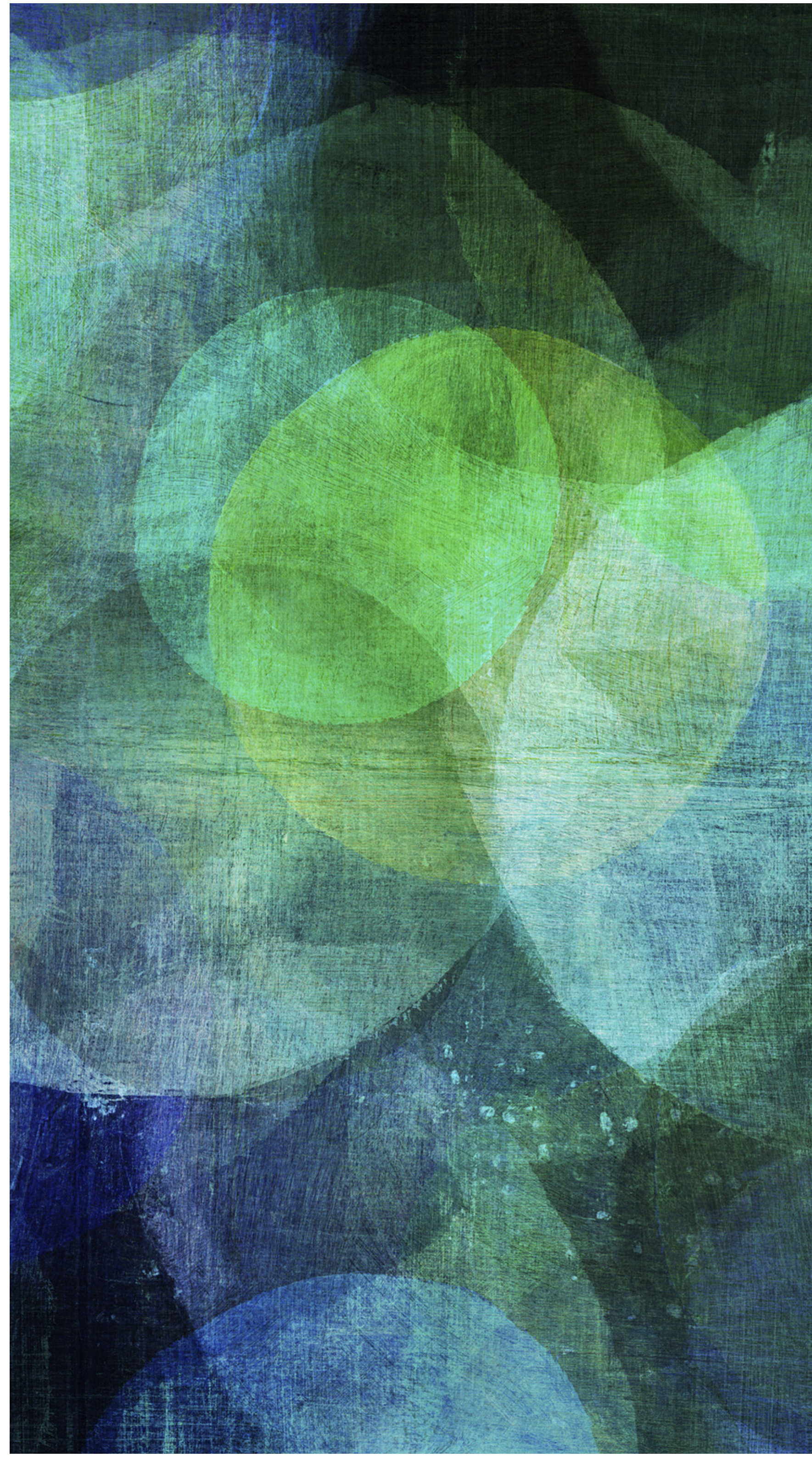
CRASH COURSE OF MACHINE LEARNING WITH EXAMPLES IN R

Session I

Dillon R. Gardner



PREAMBLE

- Examples will be in R using RStudio IDE
  - R:
    - https://www.r-project.org
  - RStudio:
    - https://www.rstudio.com/products/rstudio/download/
- Code on github
  - https://github.com/mohsseha/ArchConfRML
- Install necessary packages by running:

```
Rscript 1-RBasics/loadPackages.R
```

WHAT IS MACHINE LEARNING FROM 10,000 FEET

- Traditional programming's goal is automation
- Machine learning: automating automation
- Getting programs to write themselves
- How? Let DATA do the hard work!

TRADITIONAL PROGRAM

MACHINE LEARNING

TRADITIONAL PROGRAM

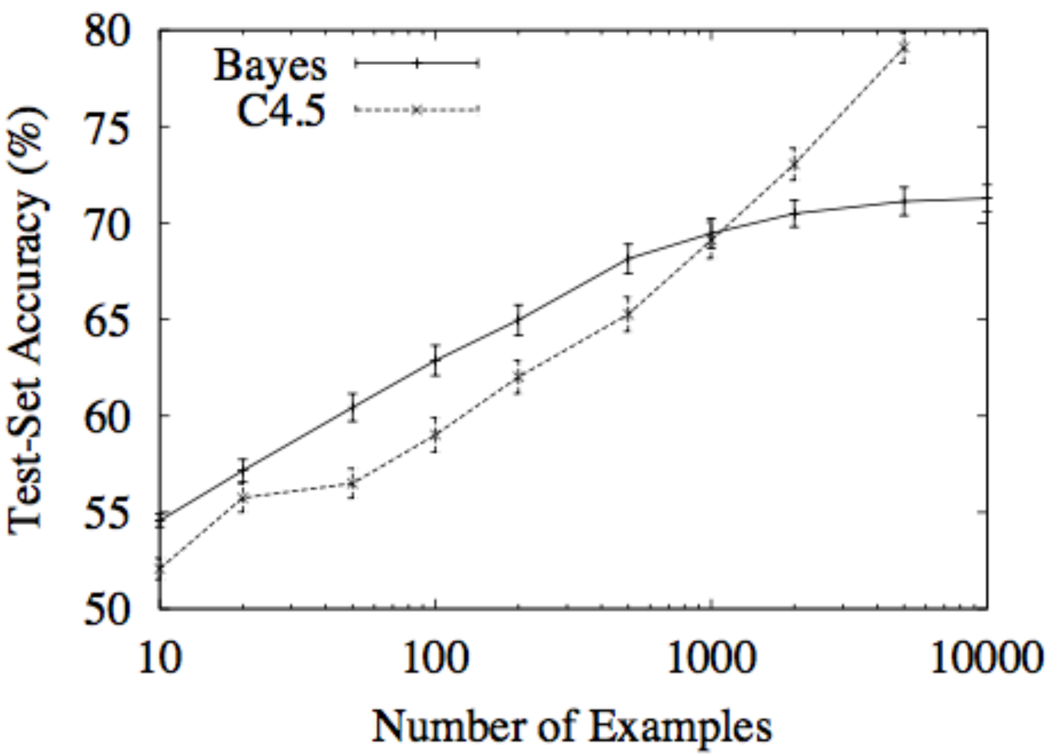
- Knowledge used to design a blueprint for program
- Engineering task of constructing a program that meets specifications

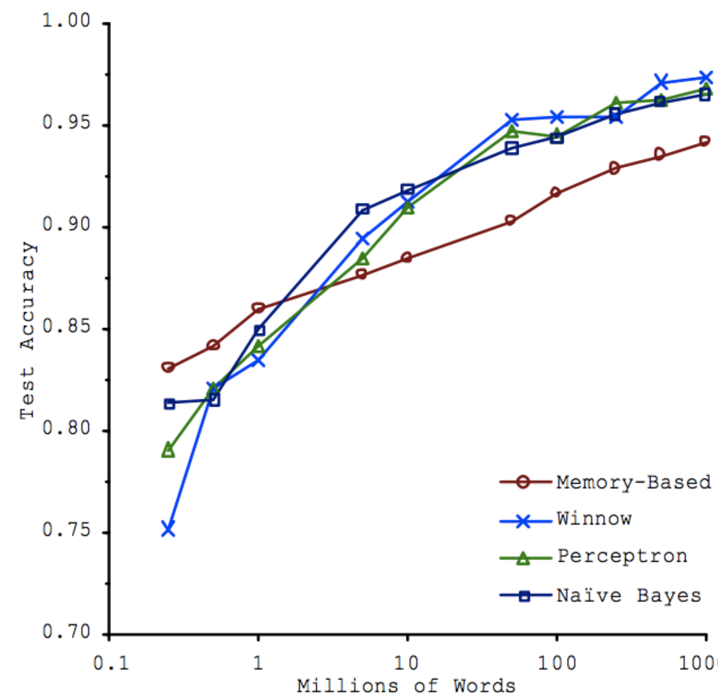
MACHINE LEARNING

- Knowledge is used to decide the final form a program should take
- Engineering task that of a farmer.
  - Plant the seed (algorithm)
  - Feed/water (data)
  - Reap the plants (programs)

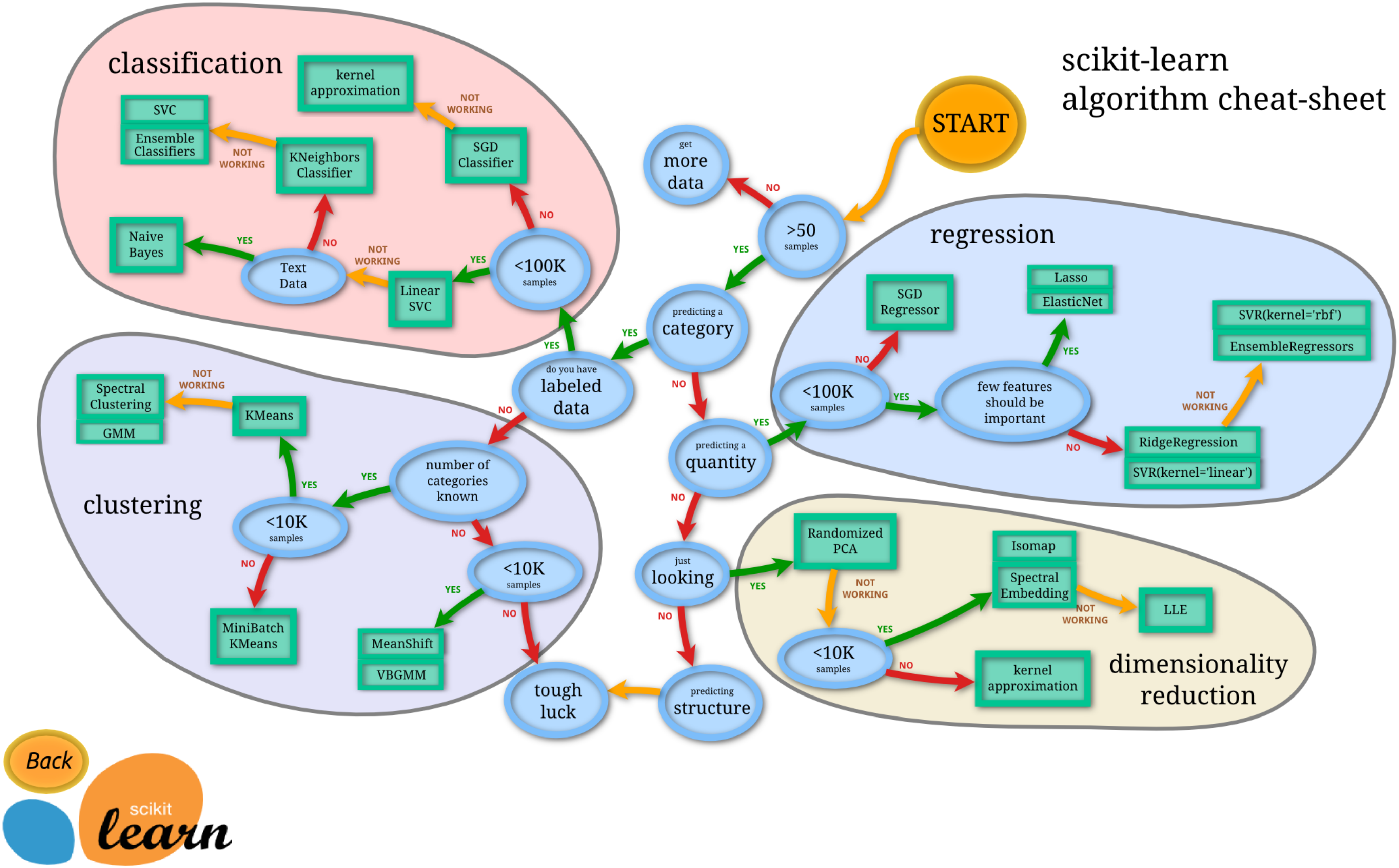
MACHINE LEARNING

- "Learners combine knowledge with data to grow programs" — Pedro Domingos





[Peter Norvig]

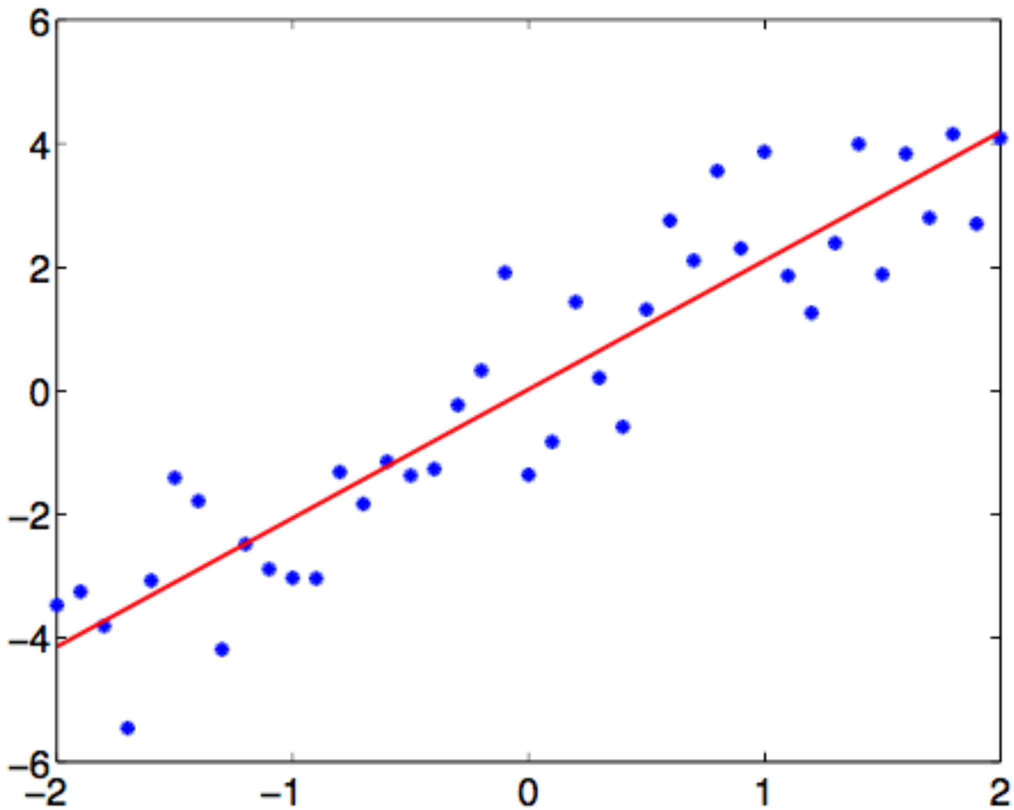


scikit-learn
algorithm cheat-sheet
START
>50 samples
get more data
predicting a category
do you have labeled data
predicting a quantity
just looking
predicting structure
tough luck
classification
<100K samples
SGD Classifier
kernel approximation
Linear SVC
Text Data
Naive Bayes
KNeighbors Classifier
SVC
Ensemble Classifiers
clustering
number of categories known
<10K samples
KMeans
Spectral Clustering
GMM
MiniBatch KMeans
MeanShift
VBGMM
regression
SGD Regressor
few features should be important
Lasso
ElasticNet
RidgeRegression
SVR(kernel='linear')
SVR(kernel='rbf')
EnsembleRegressors
dimensionality reduction
Randomized PCA
Isomap
Spectral Embedding
LLE
kernel approximation
YES
NO
NOT WORKING
Back
scikit learn

COMPONENTS OF ML ALGORITHM

- Representation
  - Language for the output program from the machine learner
  - Decision trees, neural networks, linear regression, etc.
- Evaluation
  - How do we compare candidate programs from the ML algorithm?
- Optimization
  - How can we rapidly find the “best” program

LINEAR MODELING



Representation

$$f(x;w) = w_0 + w_1 x$$

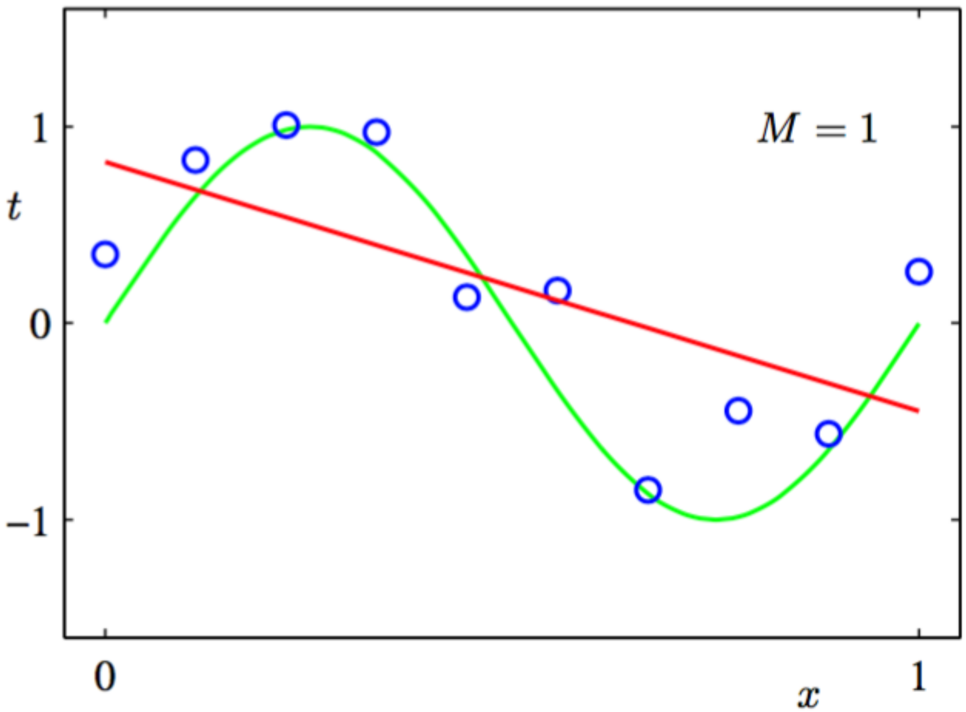
Evaluation

$$Cost(w) = \sum_{j}^{n} ((y_j - f(x_j;w))^2$$

Optimization

Gradient descent to minimize cost

LINEAR MODELING



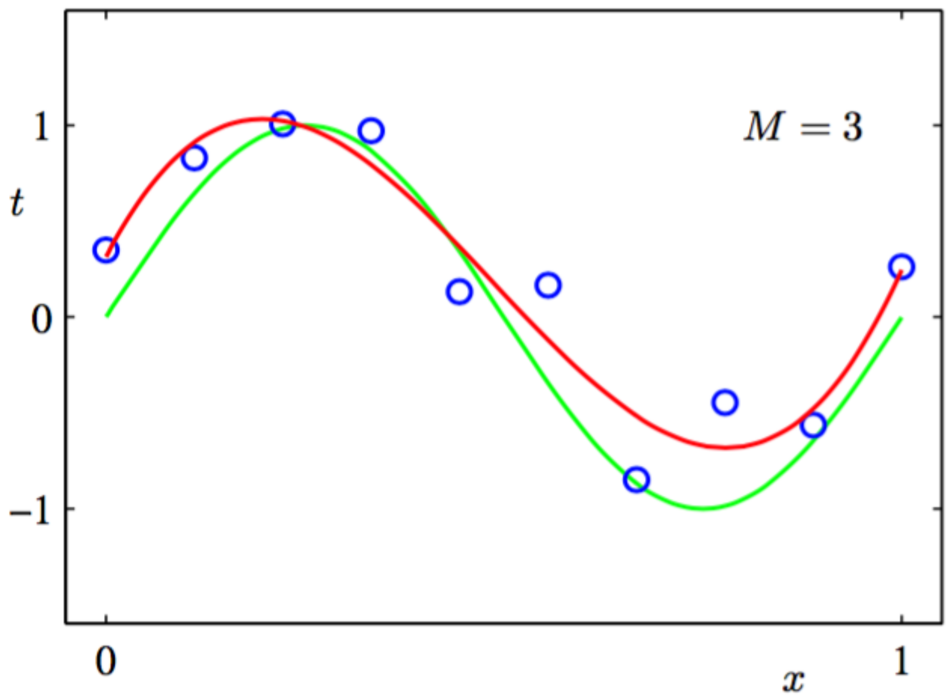
$$f(x;w) = w_0 + w_1 x$$

$$f(x;w) = w_0 + w_1 x + w_2 x^2$$

$$f(x;w) = w_0 + \sum_{i}^{m} w_i x^i$$

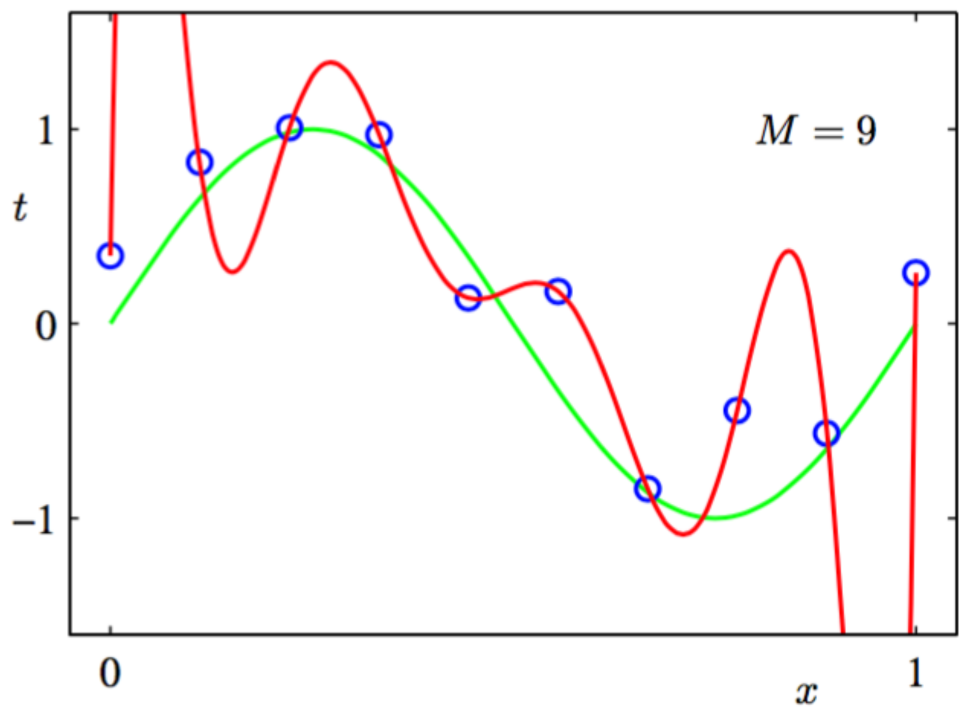
$$f(x;w) = w_0 + \sum_{i}^{m} w_i \phi_i(x)$$

LINEAR MODELING



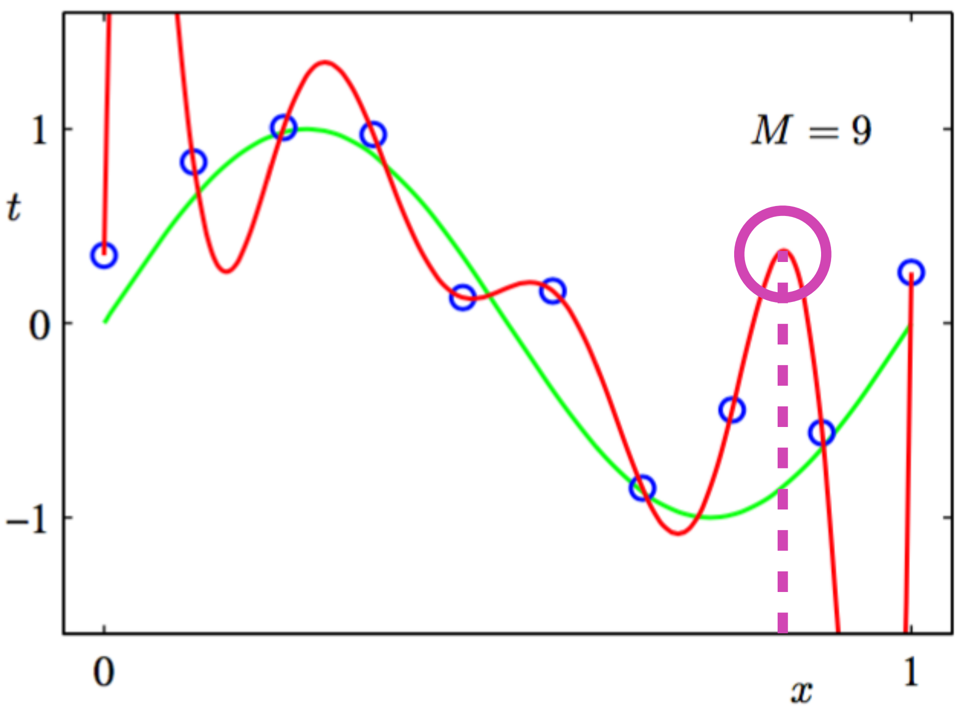
$$f(x;w) = w_0 + \sum_{i}^{m} w_i x^i$$

LINEAR MODELING



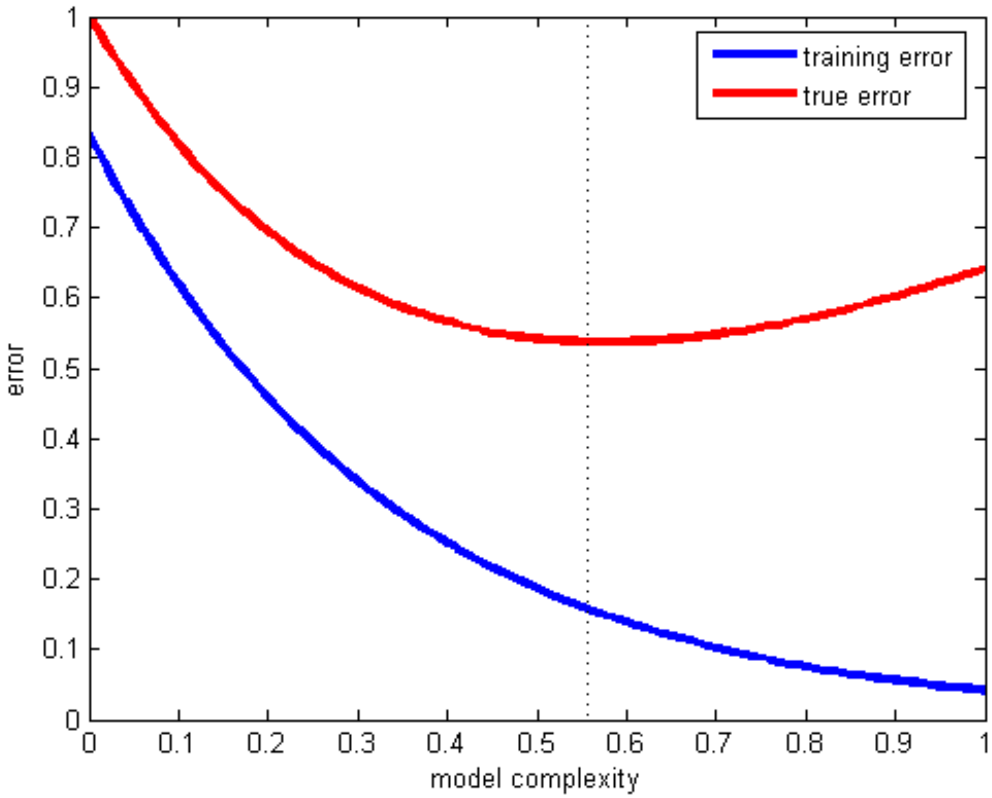
$$f(x;w) = w_0 + \sum_{i}^{m} w_i x^i$$

OVERFITTING



Do we ***REALLY*** think
this is a good estimate?

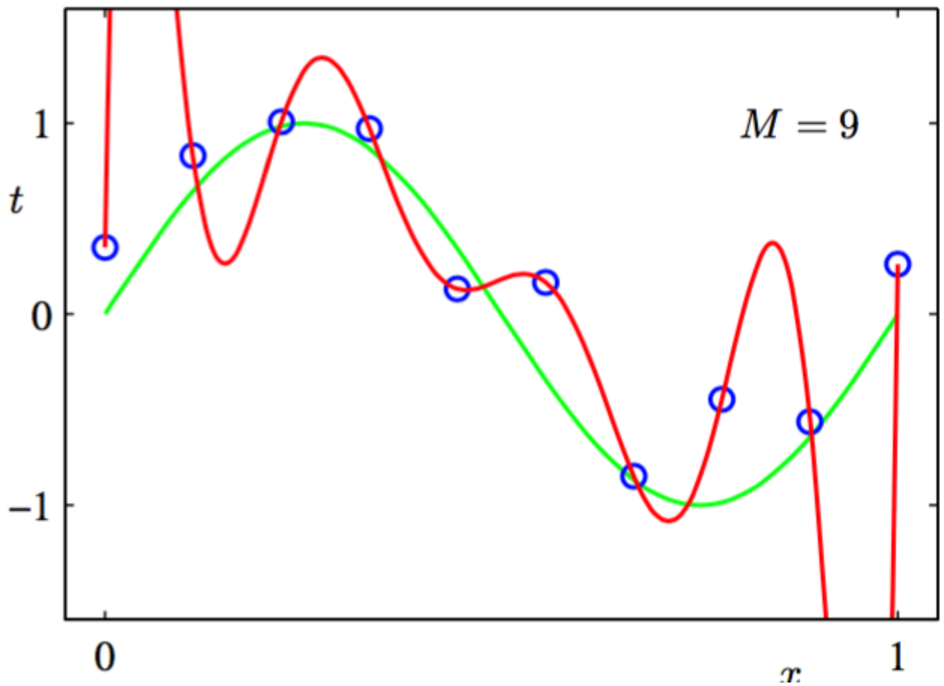
OVERFITTING



Solutions?

- Save some data for validation
- More data needed for more complex model
- Introduce regularization (penalty on large weights)

LINEAR MODELING – RIDGE REGRESSION



Representation

$$f(x; w) = w_0 + \sum_{i}^{m} w_i x^i$$

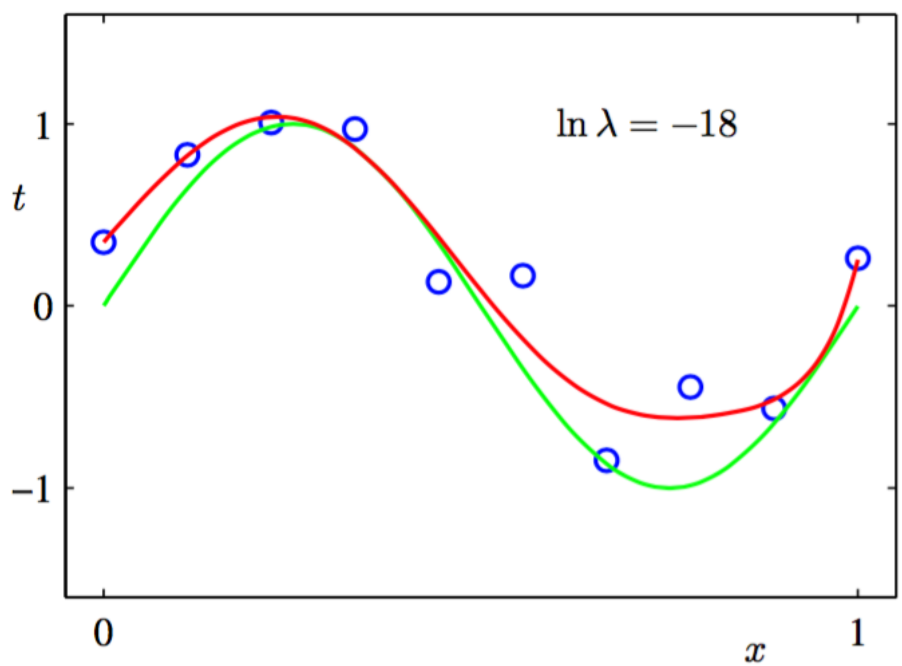
Evaluation

$$Cost(w) = \sum_{j}^{n} ((y_j - f(x_j; w))^2 + \lambda \sum_{i}^{m} w_i^2$$

Optimization

Gradient descent to minimize cost

LINEAR MODELING – RIDGE REGRESSION



Representation

$$f(x; w) = w_0 + \sum_{i}^{m} w_i x^i$$

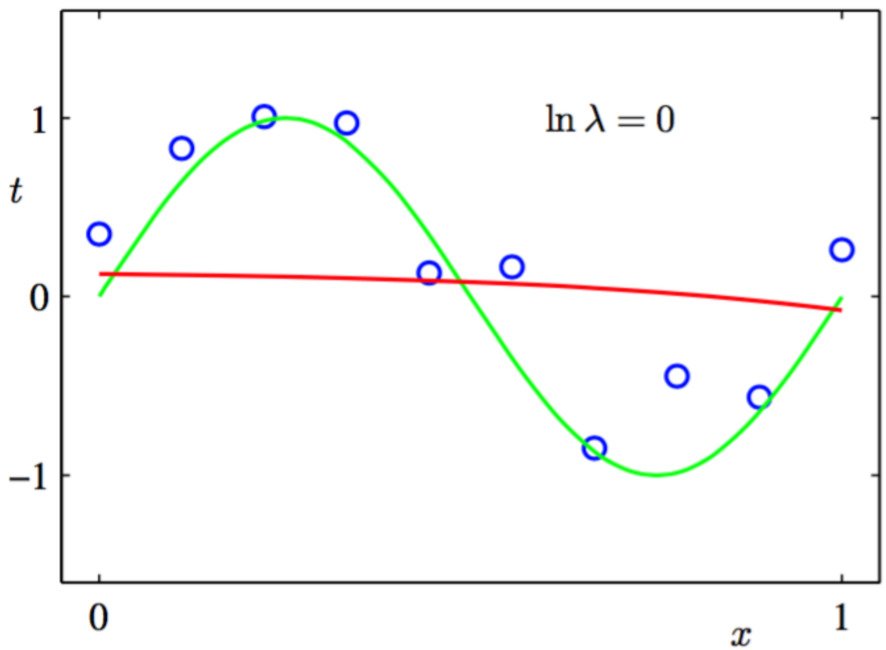
Evaluation

$$Cost(w) = \sum_{j}^{n}((y_j - f(x_j; w))^2 + \lambda \sum_{i}^{m} w_i^2$$

Optimization

Gradient descent to minimize cost

LINEAR MODELING – RIDGE REGRESSION



Representation

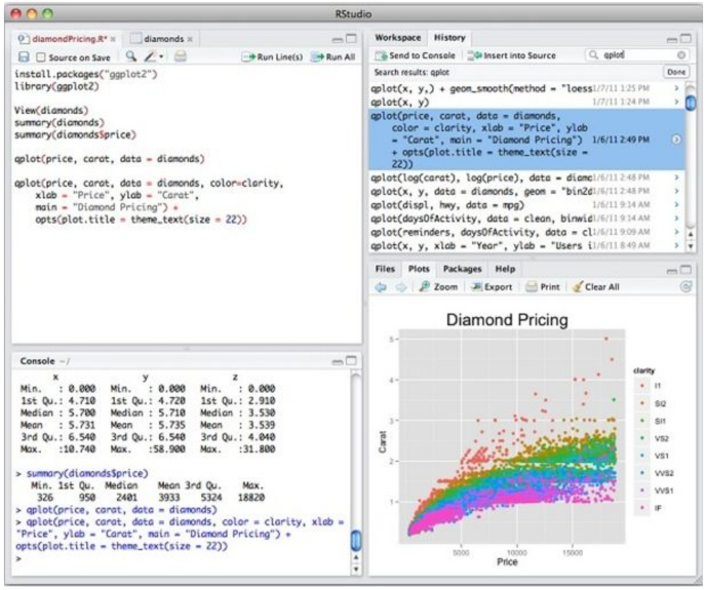
$$f(x; w) = w_0 + \sum_i^m w_i x^i$$

Evaluation

$$Cost(w) = \sum_j^n ((y_j - f(x_j; w))^2 + \lambda \sum_i^m w_i^2$$

Optimization

Gradient descent to minimize cost


RStudio
install.packages("ggplot2")
library(ggplot2)
View(diamonds)
summary(diamonds)
summary(diamonds$price)
qplot(price, carat, data = diamonds)
Console
Diamond Pricing
Price
Carat
clarity

THE GOOD

- Powerful and easy-to-use visualizations
- Extremely flexible
- Vast library of packages for wide range of tasks
- Fast and easy for exploratory data analysis

THE BAD

- Extremely Flexible
- Poor memory management
- Slow and inefficient
- Hard to productionize
  - Poor support for modules, private namespaces etc.
  - Exceptions hard to manage

THE UGLY

- Atypical syntax
- Flexible naming convention (confusing mixture of . and _)
- Multiple OO systems
- Methods typically belong to functions, not classes
- Indexing starts at 1

BASIC R SYNTAX

More Syntax in 1-RBasics in github repo

- Assignment operator is **->**

  `> a <- 5`

  = is used for default parameter values in function definitions

- c is for combine or convert/coerce

  `> c(1,4,5)`

  `> c(1, “Hello”, 2.5, “World”)`

- Ranges can be succinctly created with :

  `> c(1:10, 5,6, 2:5)`

BASIC R DATA STRUCTURES

- data.frames are extremely common objects for handling tabular data

```
>myDF <- data.frame(a=c("Hello", "World", "!"),
                    b=c(1,2,3))
```

- data.frames can be accessed either positionally or by name

```
>myDF[1]        >myDF["a"]       > myDF$a
>myDF[1,1]      >myDF[1, "a"]    > myDF$a[1]
>myDF[2, 1:2]   >myDF[2, c(1,2)]
```

BASIC R SYNTAX

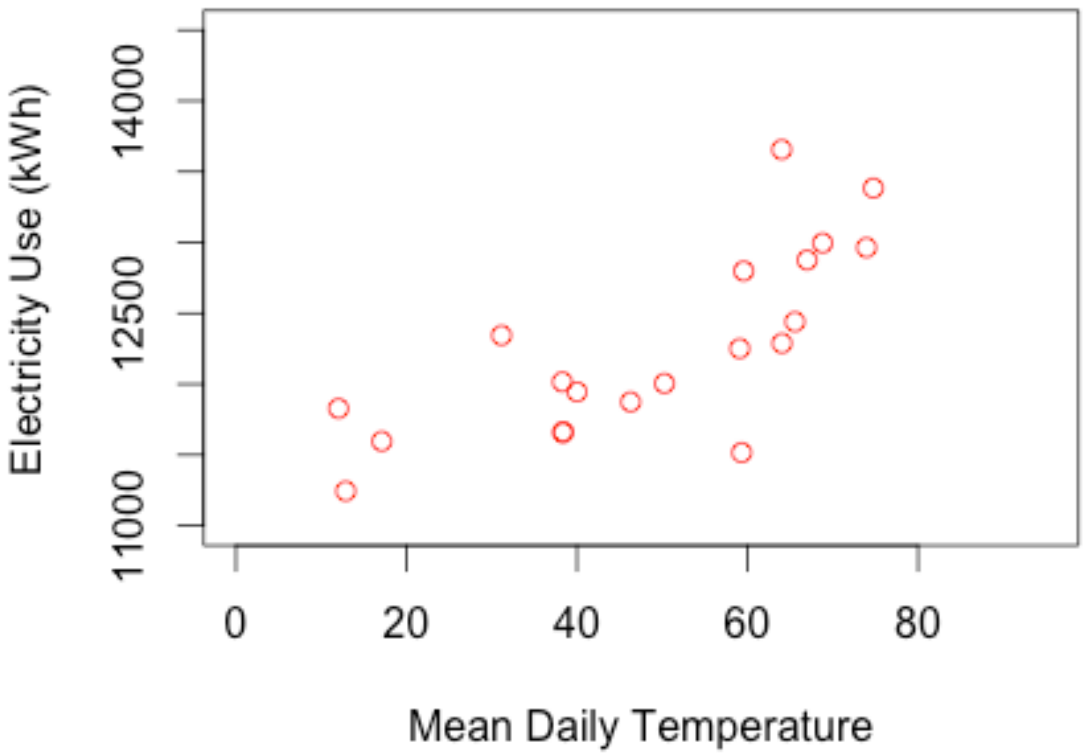
- Functions are objects

```
> hello <- function(){
        print(“Hello World”)
    }
> hello()
```

- %>% is a commonly defined pipe operator

  > a %>% f(b,c) is equivalent to f(a,b,c)

LINEAR MODELING – RIDGE REGRESSION



Representation

$$f(x; w) = w_0 + \sum_{i}^{m} w_i x^i$$

Evaluation

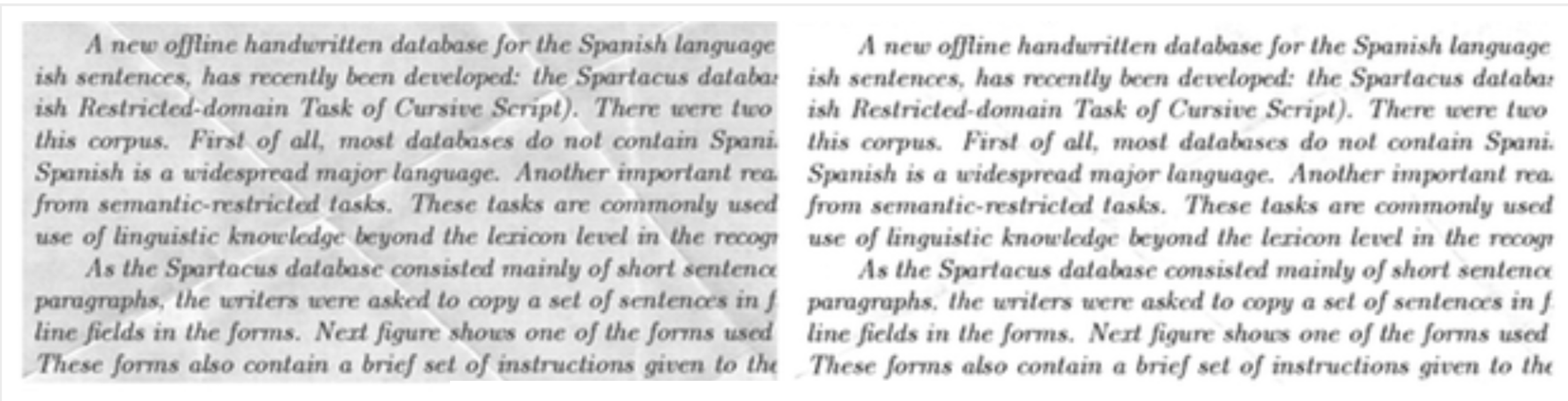
$$Cost(w) = \sum_{j}^{n} ((y_j - f(x_j; w))^2 + \lambda \sum_{i}^{m} w_i^2$$

Optimization

Gradient descent to minimize cost

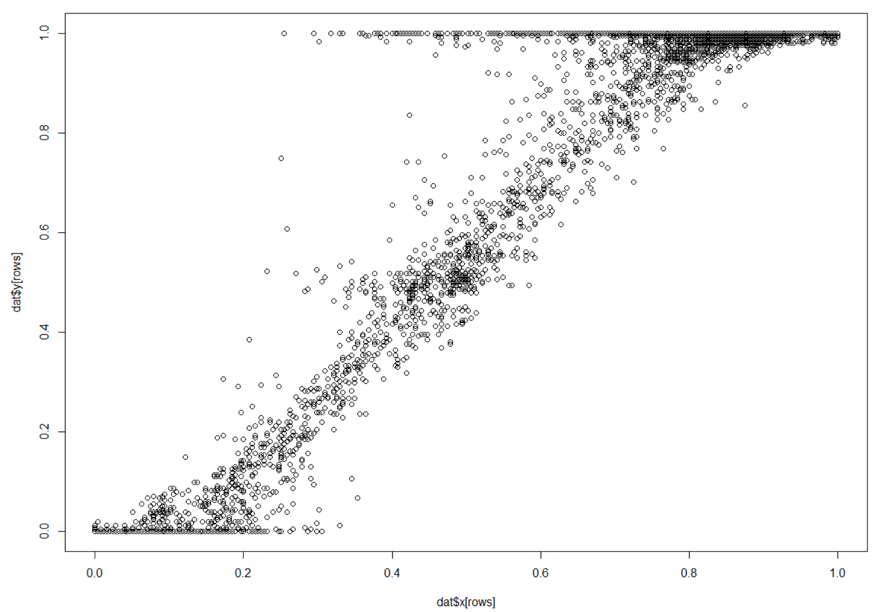
LINEAR MODELING EXAMPLE

LINEAR MODEL EXAMPLES: DENOISING DOCUMENTS

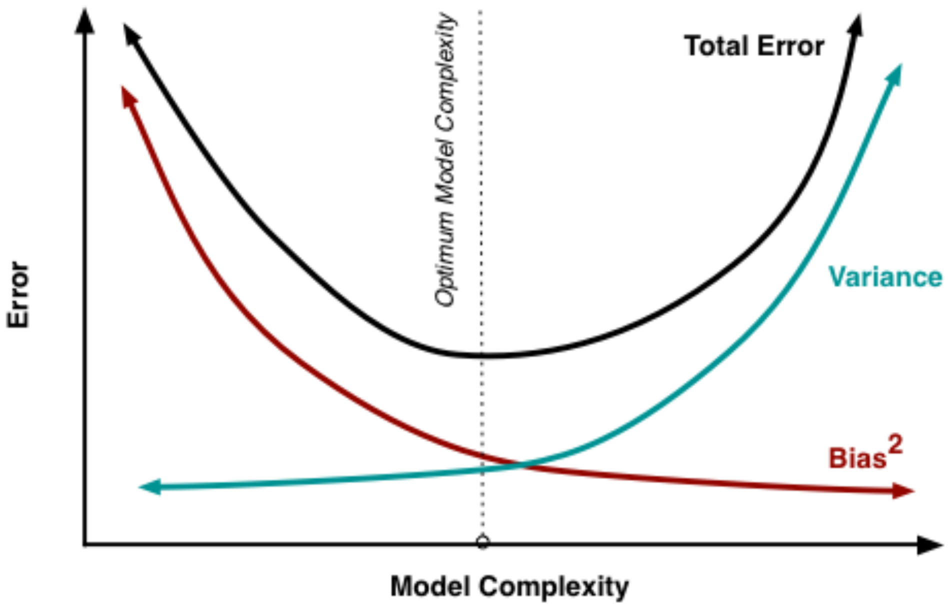


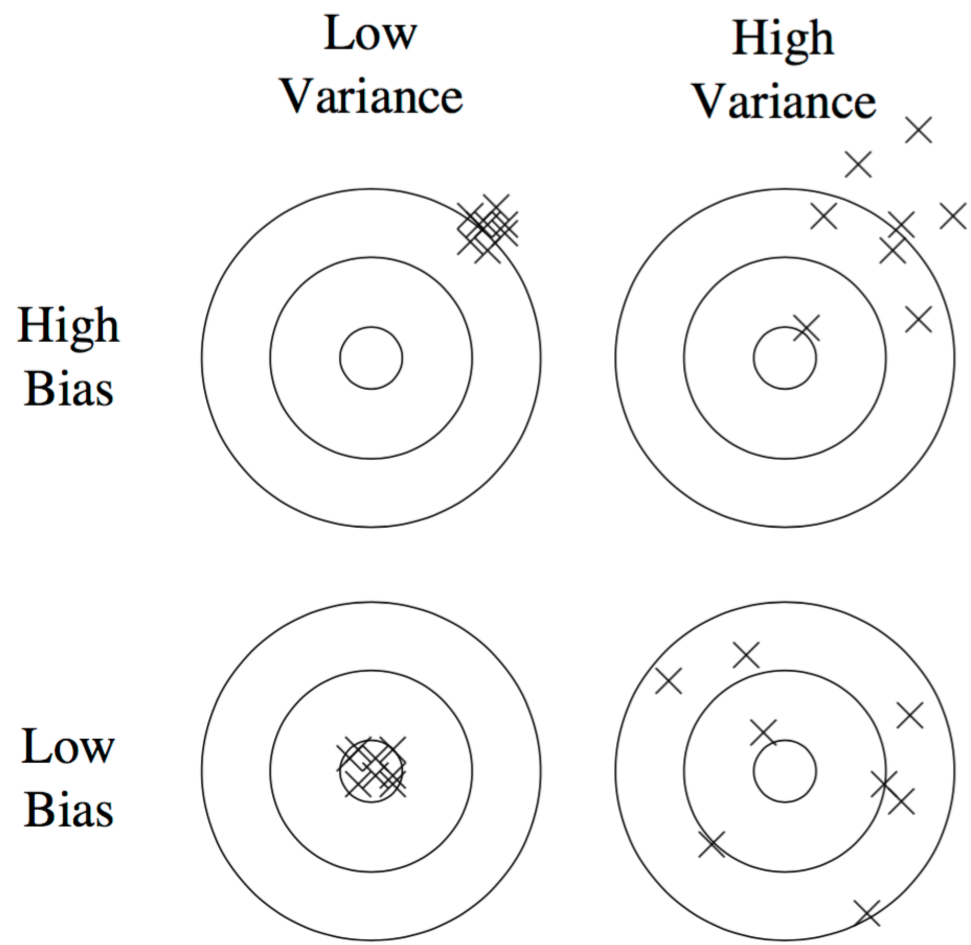
A new offline handwritten database for the Spanish language
ish sentences, has recently been developed: the Spartacus databa
ish Restricted-domain Task of Cursive Script). There were two
this corpus. First of all, most databases do not contain Spani
Spanish is a widespread major language. Another important rea
from semantic-restricted tasks. These tasks are commonly used
use of linguistic knowledge beyond the lexicon level in the recog
As the Spartacus database consisted mainly of short sentenc
paragraphs, the writers were asked to copy a set of sentences in f
line fields in the forms. Next figure shows one of the forms used
These forms also contain a brief set of instructions given to the

A new offline handwritten database for the Spanish language
ish sentences, has recently been developed: the Spartacus databa
ish Restricted-domain Task of Cursive Script). There were two
this corpus. First of all, most databases do not contain Spani
Spanish is a widespread major language. Another important rea
from semantic-restricted tasks. These tasks are commonly used
use of linguistic knowledge beyond the lexicon level in the recog
As the Spartacus database consisted mainly of short sentenc
paragraphs, the writers were asked to copy a set of sentences in f
line fields in the forms. Next figure shows one of the forms used
These forms also contain a brief set of instructions given to the

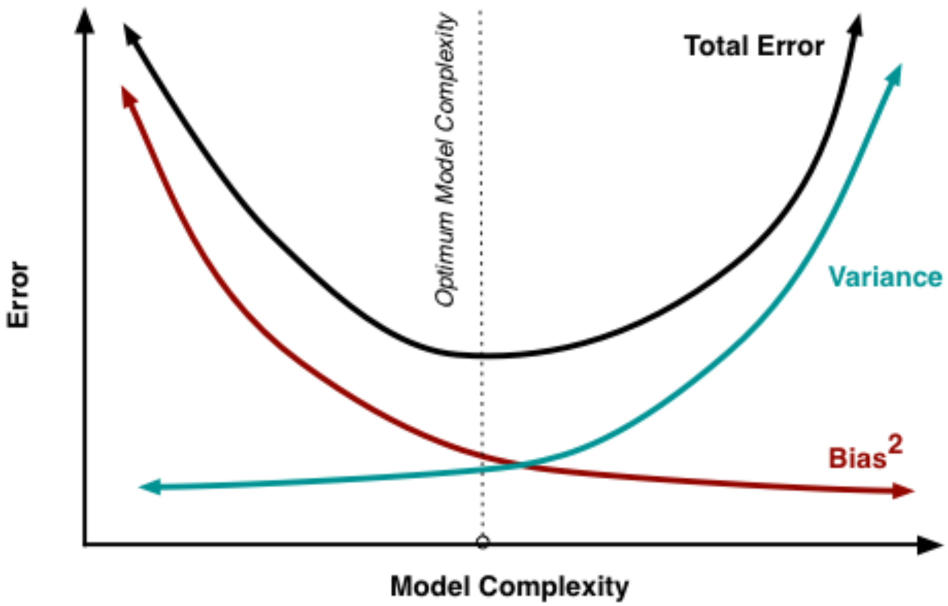


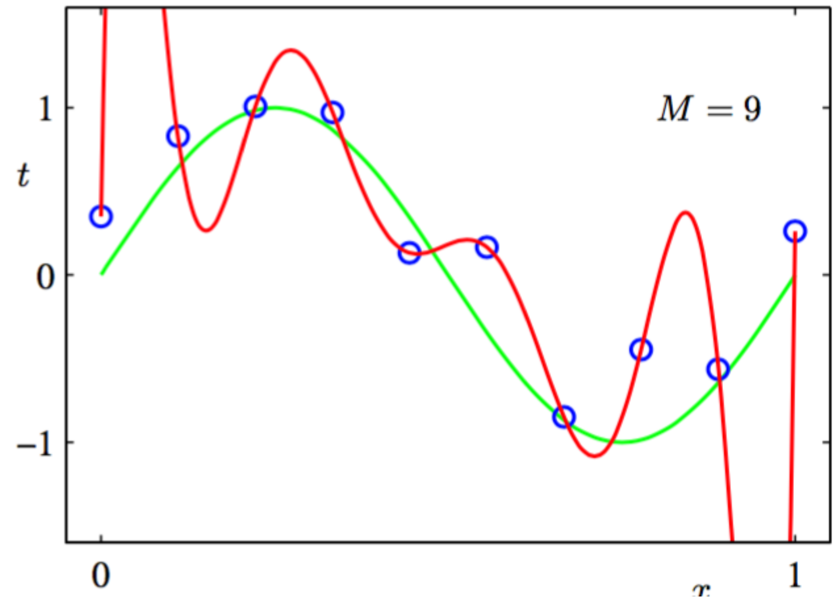
VARIANCE-BIAS TRADE-OFF

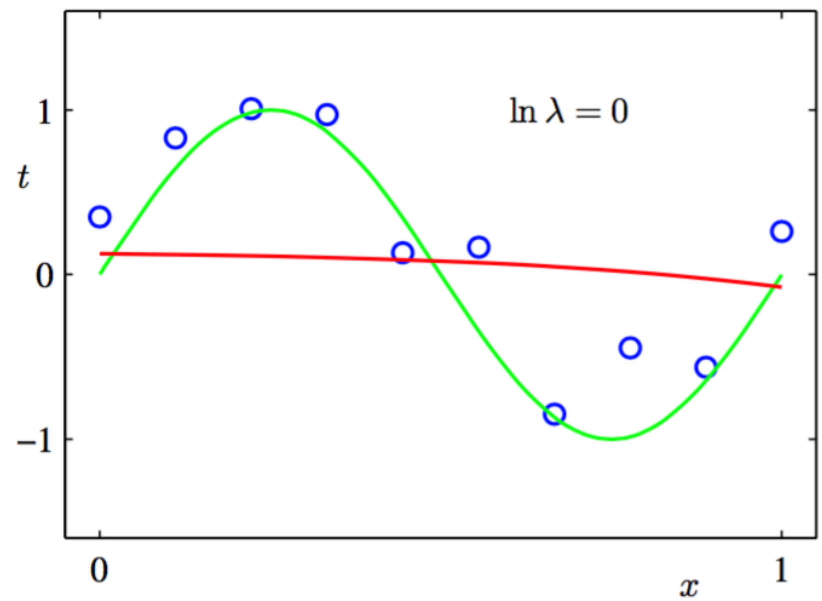




VARIANCE-BIAS TRADE-OFF





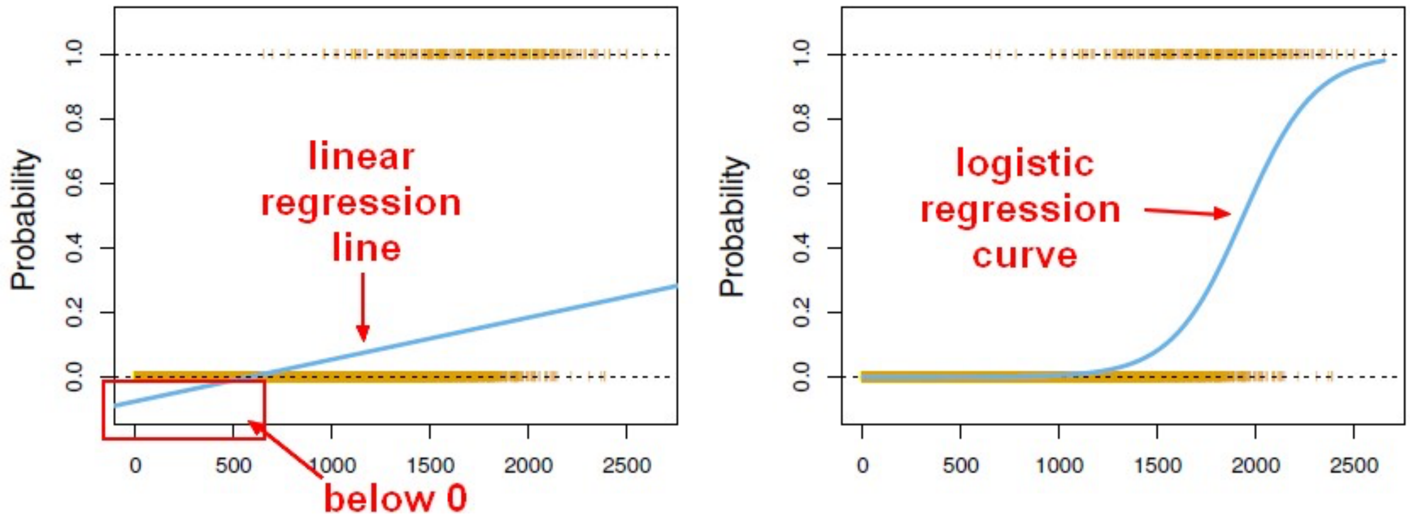


GENERALIZE LINEAR MODEL

- Linear regression output ranges from -∞ to ∞
- How about situation in which the output is a binary variable?
- Generalize linear models:

$$f(x;w) = g(w_0 + \sum_{i}^{m} w_i x^i)$$

LOGISTIC REGRESSION



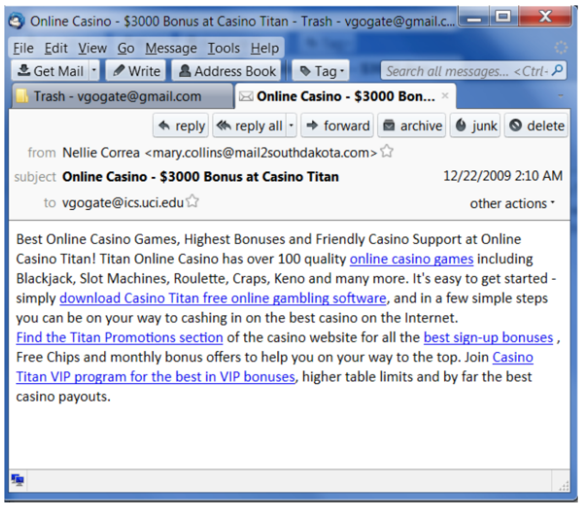
$$P(y=1|\mathbf{x},\mathbf{w}) = g(w_0 + \sum_i w_i x_i)$$

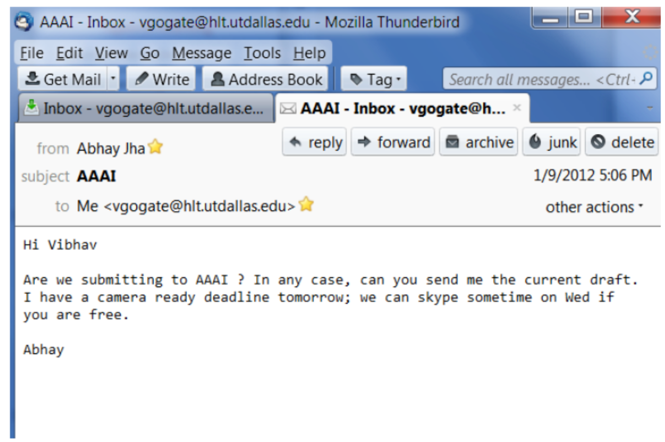
$$g(z) = (1 + exp(-z))^{-1}$$

$$L = \sum_j P(y_j|\mathbf{x_j},\mathbf{w})$$

LOGISTIC REGRESSION EXAMPLES

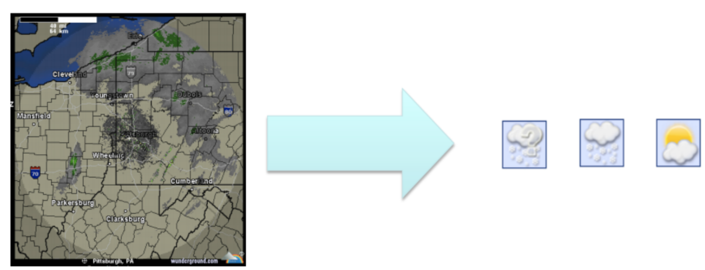
Spam Filters



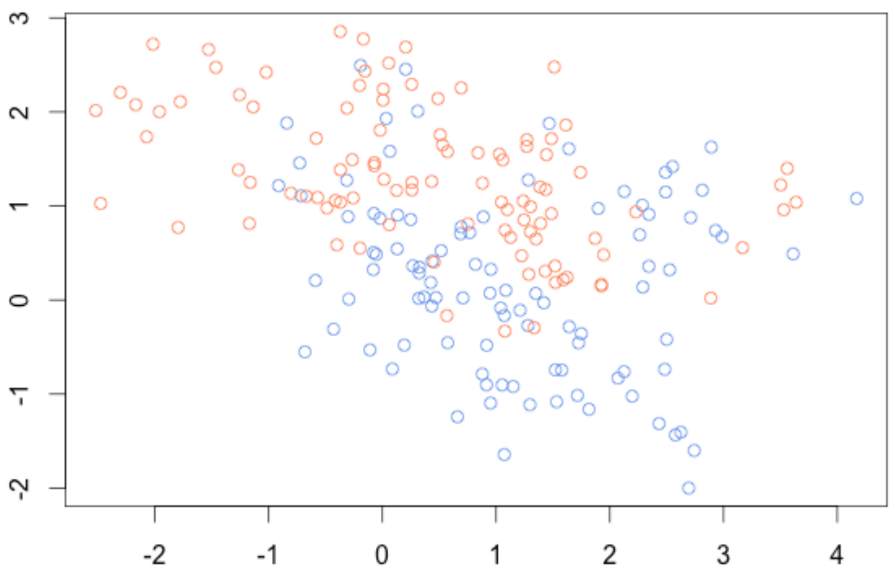


Precipitation Prediction

Classification Example: Weather Prediction

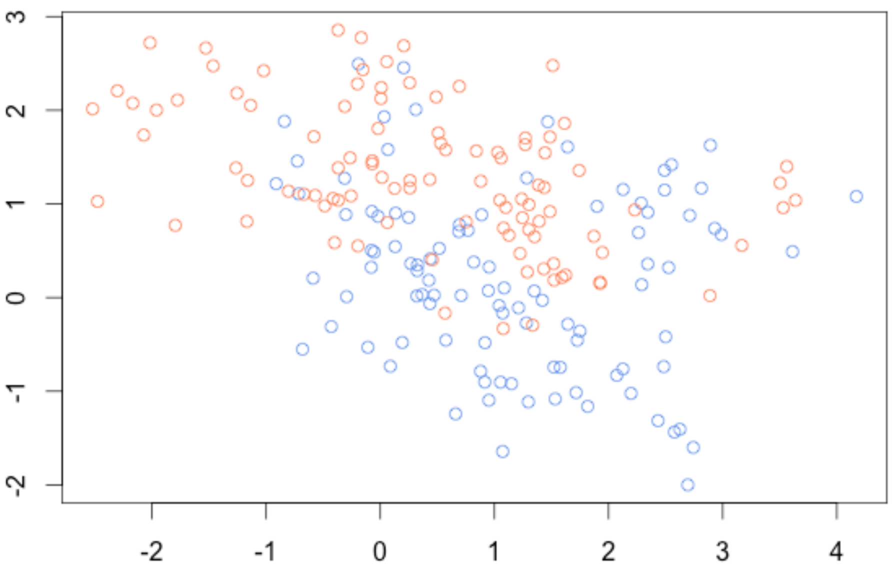


NEAREST NEIGHBORS

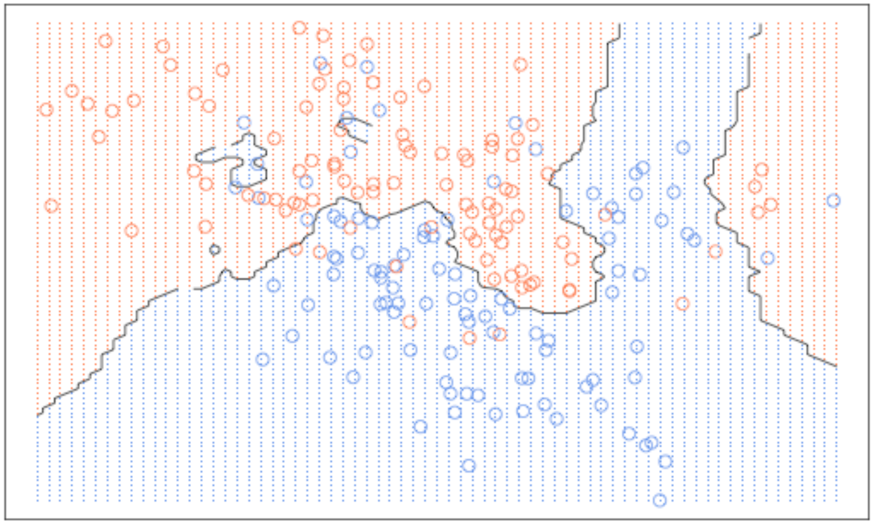


- How can we classify?
- Learn by analogy: I am likely to be similar to what's near me
  - Open question of how to determine distance?
  - How many neighbors do we consider?

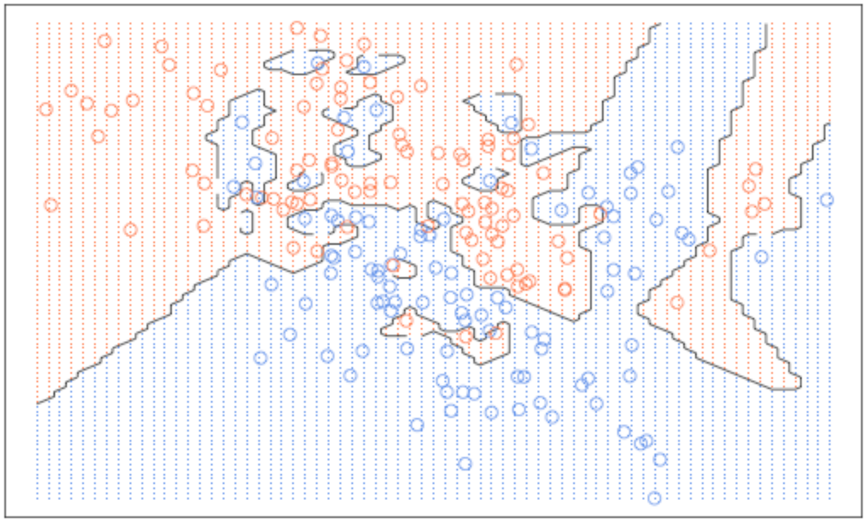
NEAREST NEIGHBORS



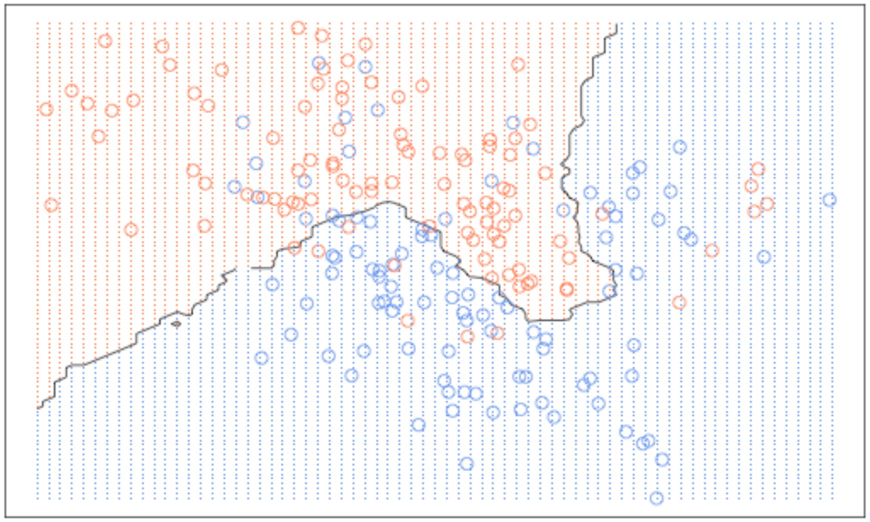
5-nearest neighbour



1-nearest neighbour



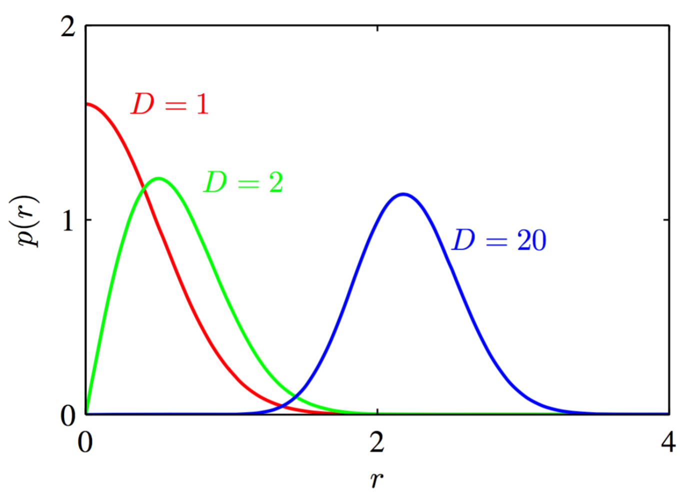
15-nearest neighbour

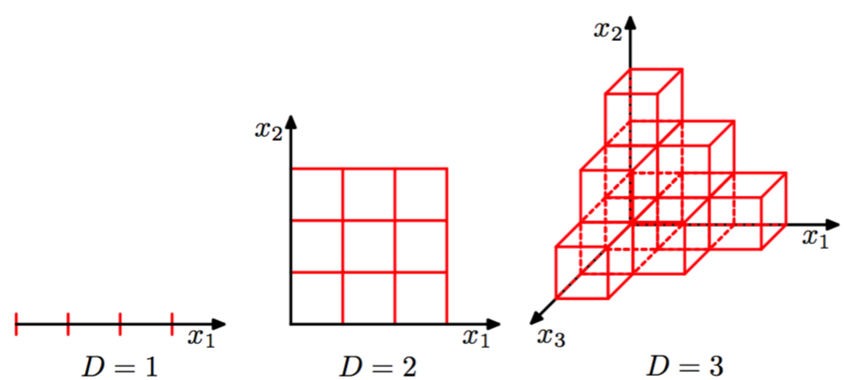


CURSE OF DIMENSIONALITY

Why can't we just blindly apply these tools to massive sets of data with a large number of features?

- Specious connections if we have too much unrelated data (Washington Redskins Rule)
- Challenges due to exponential increases
- Intuition starts to fail



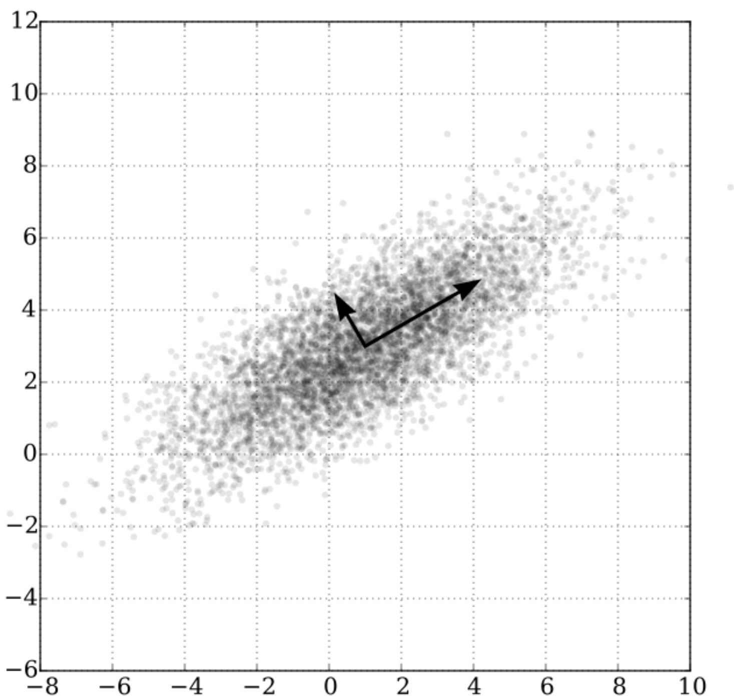


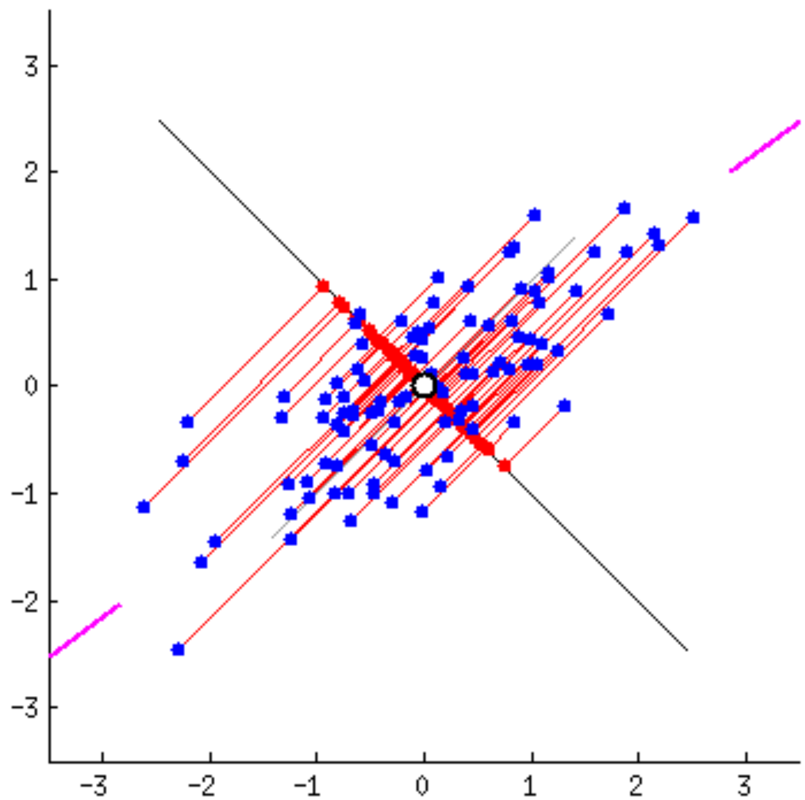
Better data is always better. There is no arguing against that. So any effort you can direct towards "improving" your data is always well invested. The issue is that better data does not mean **more** data. As a matter of fact, sometimes it might mean **less**!

-Xavier Amatrian, Quora

DIMENSIONALITY REDUCTION: PRINCIPLE COMPONENT ANALYSIS (PCA)

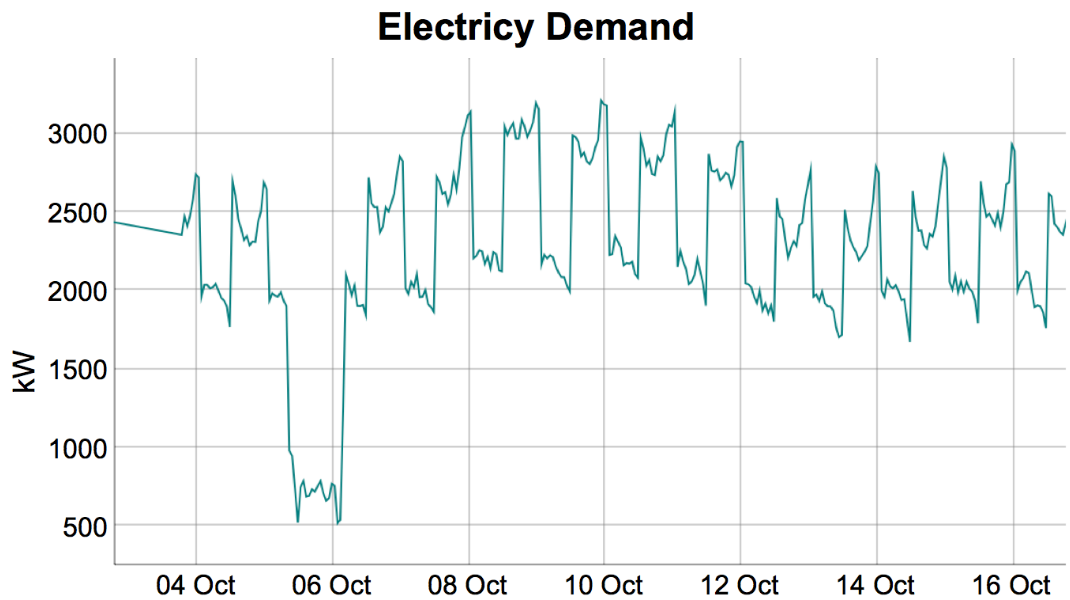
- Reduce feature dimensionality by finding directions of largest variation in the data





PCA EXAMPLE

10/3/2014 12:00	230.5
10/3/2014 1:00	240.2
10/3/2014 2:00	242.4
10/3/2014 3:00	259.3
10/3/2014 4:00	247.3



PCA EXAMPLE: TRANSFORM DATA

10/3/2014 12:00	230.5
10/3/2014 1:00	240.2
10/3/2014 2:00	242.4
10/3/2014 3:00	259.3
10/3/2014 4:00	247.3

Transform Data so each row is a day

HOUR

Date	12	1	2
10/3/2014	230.5	240.2	242.4
10/4/2014	225.8	232.1	238.7
10/5/2014	232.1	248	233.6
10/6/2014	240.1	219.4	215.7
10/7/2014	222.8	230.3	240.5

…

PCA EXAMPLE: DETERMINE PRINCIPLE COMPONENTS

HOUR

Date	12	1	2
10/3/2014	230.5	240.2	242.4
10/4/2014	225.8	232.1	238.7
10/5/2014	232.1	248	233.6
10/6/2014	240.1	219.4	215.7
10/7/2014	222.8	230.3	240.5

. . .

Learn Principle Components

PC

1	2	3
0.07	-0.02	0.0
0.07	-0.02	0.0
0.08	-0.02	0.0
0.08	0.0	0.7
0.08	0.01	1.2

. . .

PCA EXAMPLE: "ROTATE" DATA TO PRINCIPLE COMPONENTS

HOUR

DATE	12	1	2
10/3/2014	230.5	240.2	242.4
10/4/2014	225.8	232.1	238.7
10/5/2014	232.1	248	233.6
10/6/2014	240.1	219.4	215.7
10/7/2014	222.8	230.3	240.5

. . .

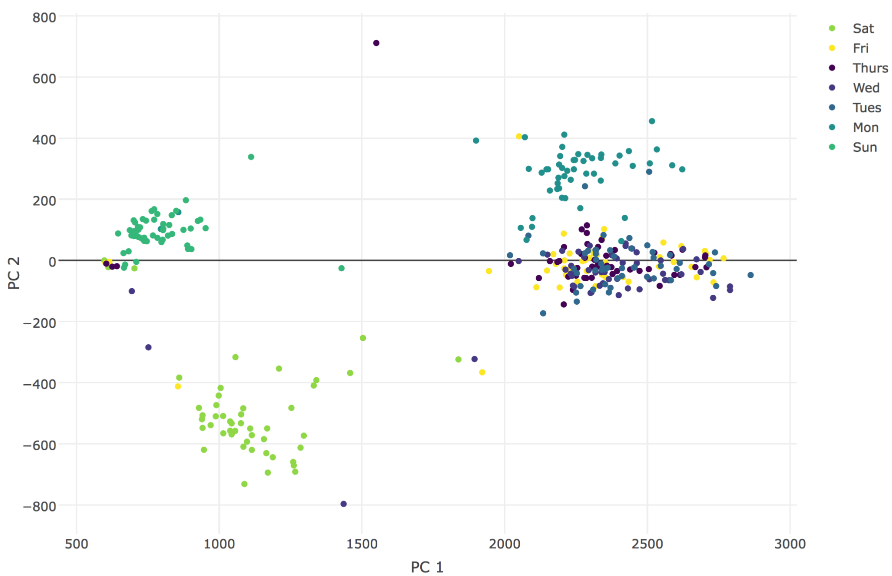
Rotate to Principle Components

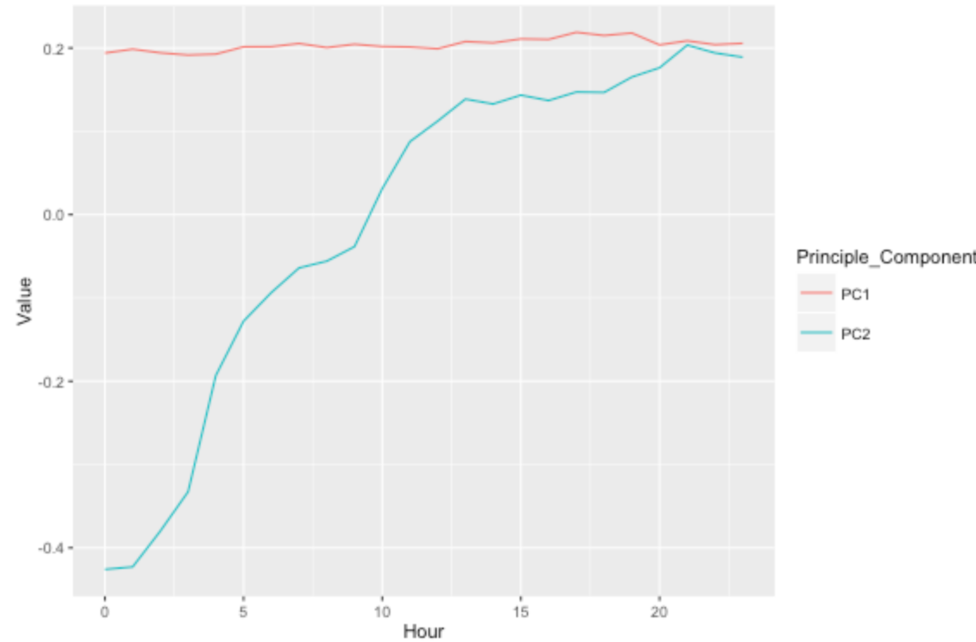
PC

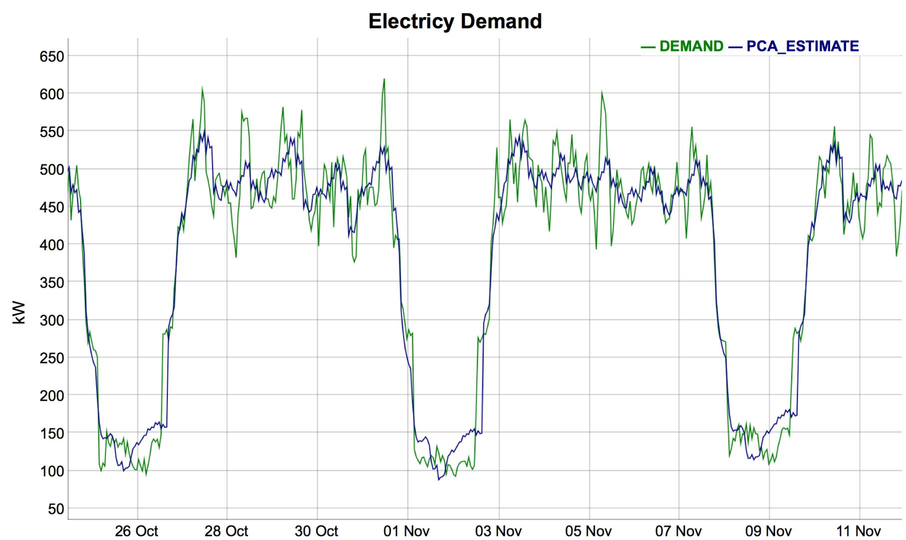
DATE	1	2
10/3/2014	200	5
10/4/2014	251	3
10/5/2014	242	15
10/6/2014	232	9
10/7/2014	210	10

. . .

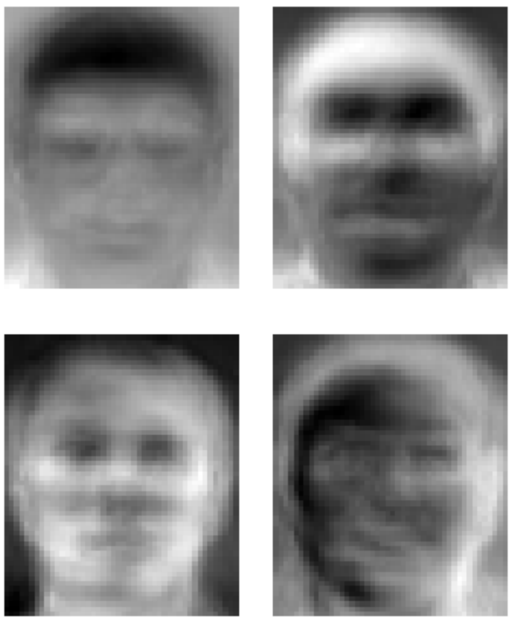
PCA RESULTS

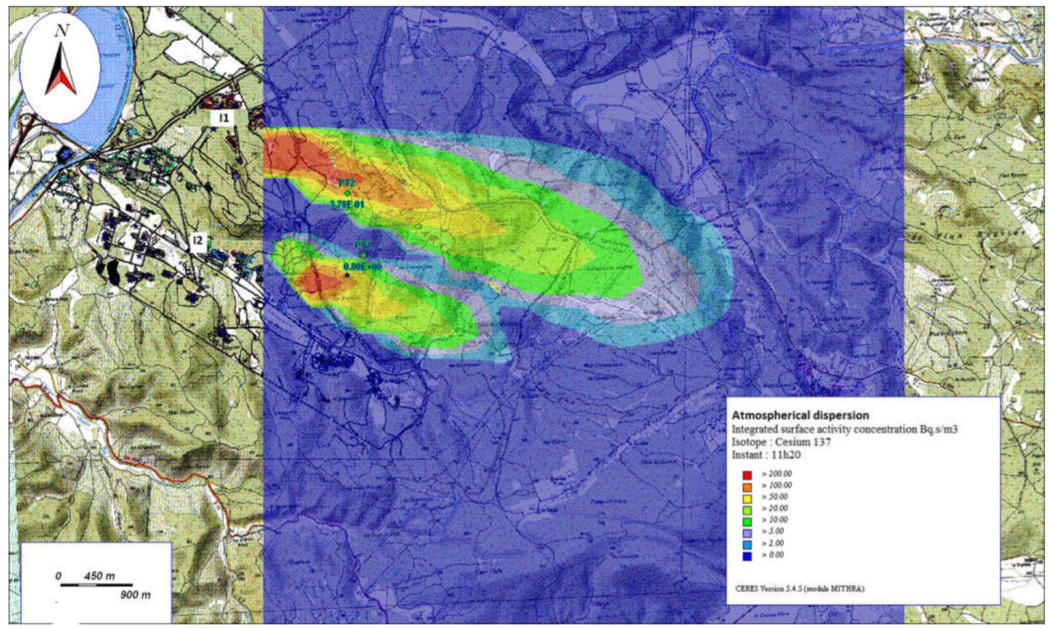

800
600
400
200
0
−200
−400
−600
−800
PC 2
500
1000
1500
2000
2500
3000
PC 1
Sat
Fri
Thurs
Wed
Tues
Mon
Sun


0.2
0.0
-0.2
-0.4
Value
0
5
10
15
20
Hour
Principle_Component
PC1
PC2


Electricy Demand
— DEMAND — PCA_ESTIMATE
650
600
550
500
450
400
350
300
250
200
150
100
50
kW
26 Oct
28 Oct
30 Oct
01 Nov
03 Nov
05 Nov
07 Nov
09 Nov
11 Nov

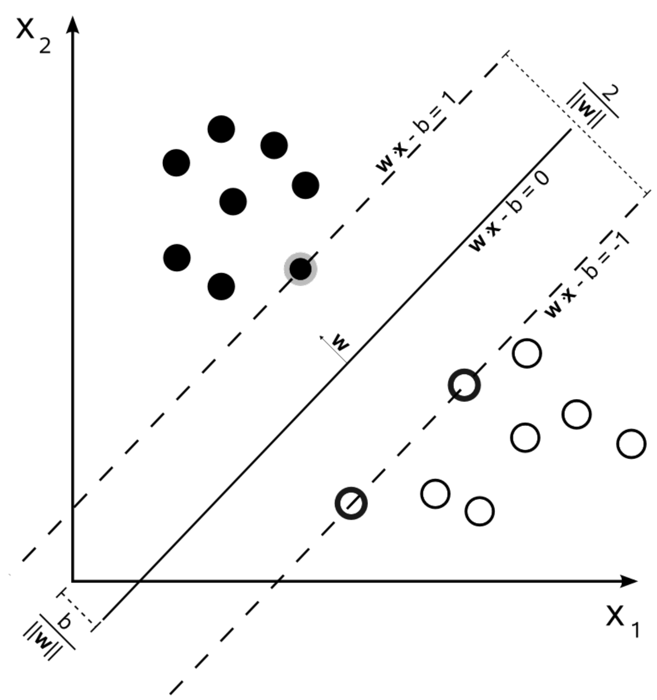
MORE PCA EXAMPLES

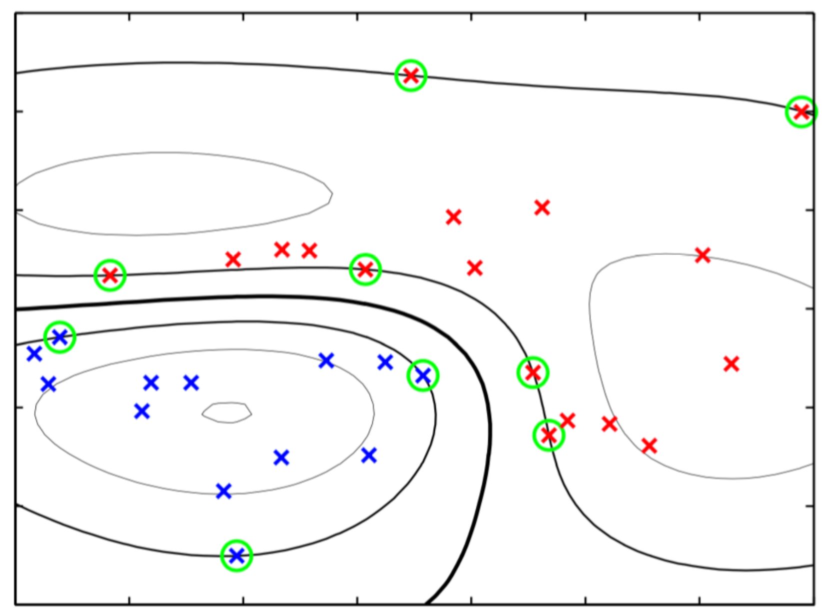




SUPPORT VECTOR MACHINES

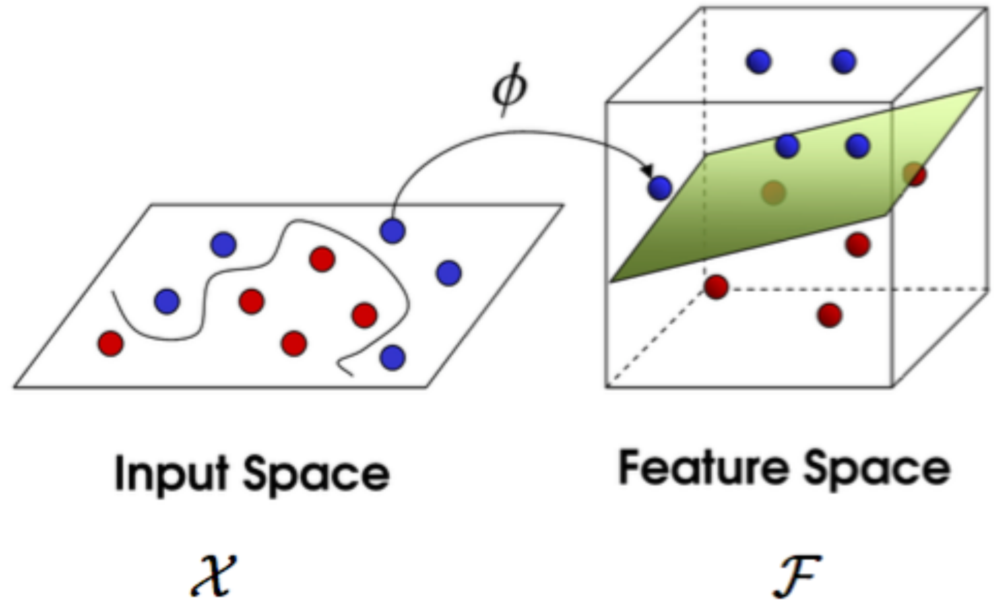
- Classification: like nearest neighbor or logistic regression
- Representation: A plane dividing classes
- Key idea: Maximize Margins
- Keep only the “Support Vectors”

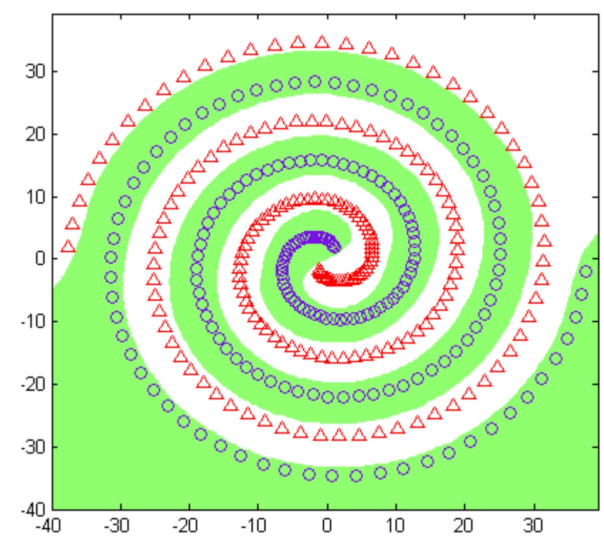




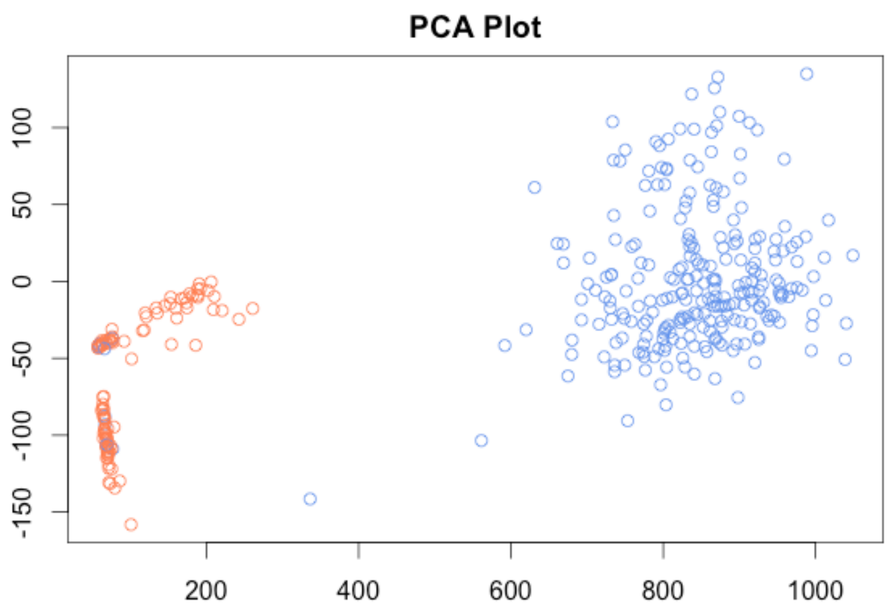
SVM (KERNEL TRICK)

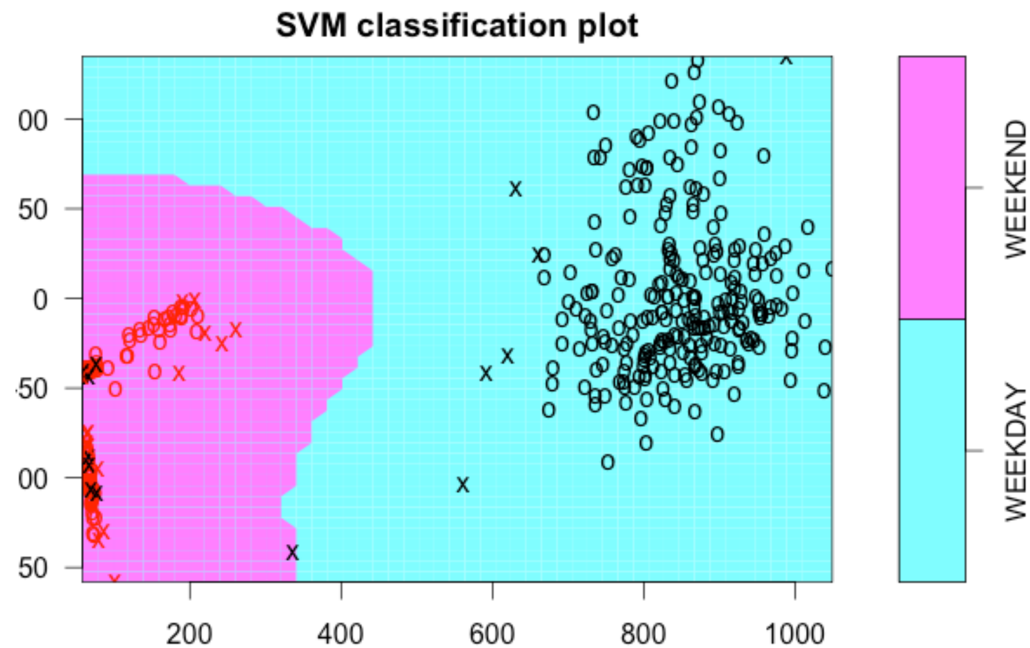
- SVM can only learn a decision plane
- A kernel is a function that maps the input space into a higher-dimension feature space
- Decision plane in feature space can be extremely complex in original space





SVM EXAMPLE



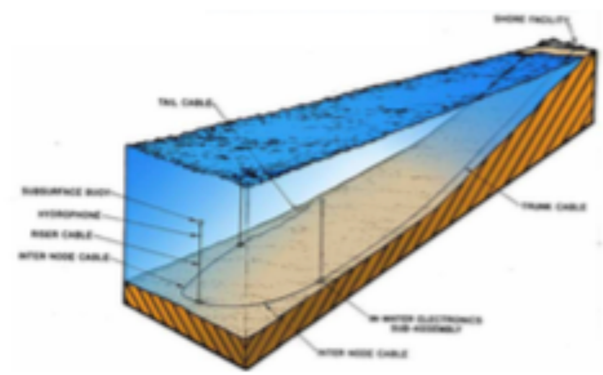


DATE	DAY_TYPE	PREDICTION
2014-09-01	WEEKDAY	WEEKEND
2014-11-27	WEEKDAY	WEEKEND
2014-11-28	WEEKDAY	WEEKEND
2014-12-24	WEEKDAY	WEEKEND
2014-12-25	WEEKDAY	WEEKEND
2015-01-01	WEEKDAY	WEEKEND
2015-01-27	WEEKDAY	WEEKEND
2015-05-25	WEEKDAY	WEEKEND
2015-07-03	WEEKDAY	WEEKEND

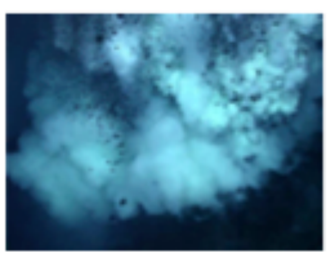
SVM PRACTICAL EXAMPLE

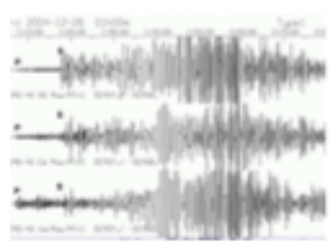
Example: Hydroacoustic signal classification

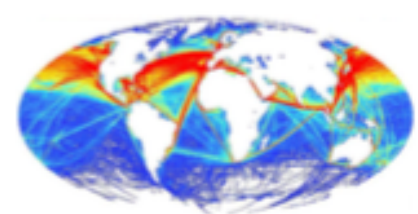
- Verification of the comprehensive nuclear-test-ban treaty
- Data from hydroacoustic network



- SVMs distinguishes explosive events from earthquakes and noise (4.3 % error)







Tuma, Igel, Prior: Hydroacoustic Signal Classification Using Kernel Functions for Variable Feature Sets. *ICPR*, 2010

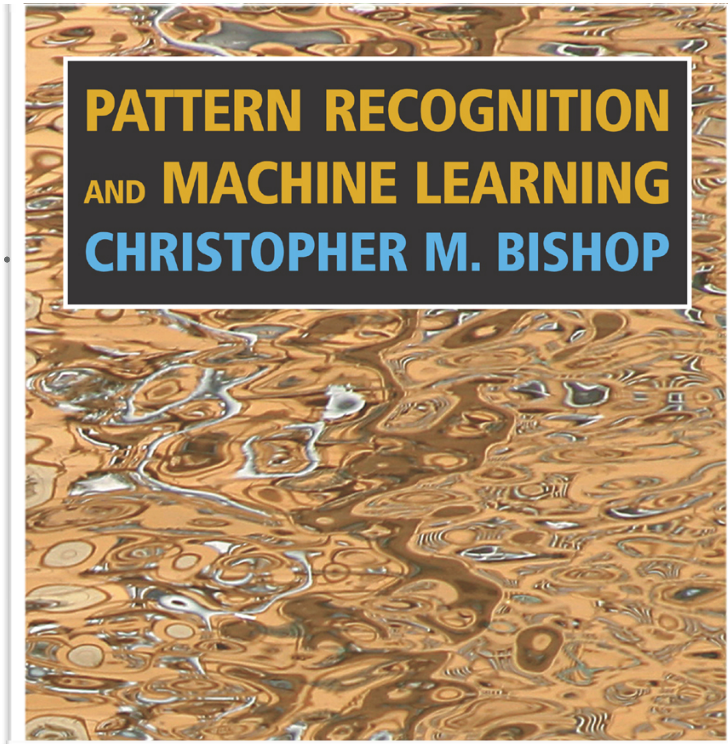
KEY TAKE AWAYS

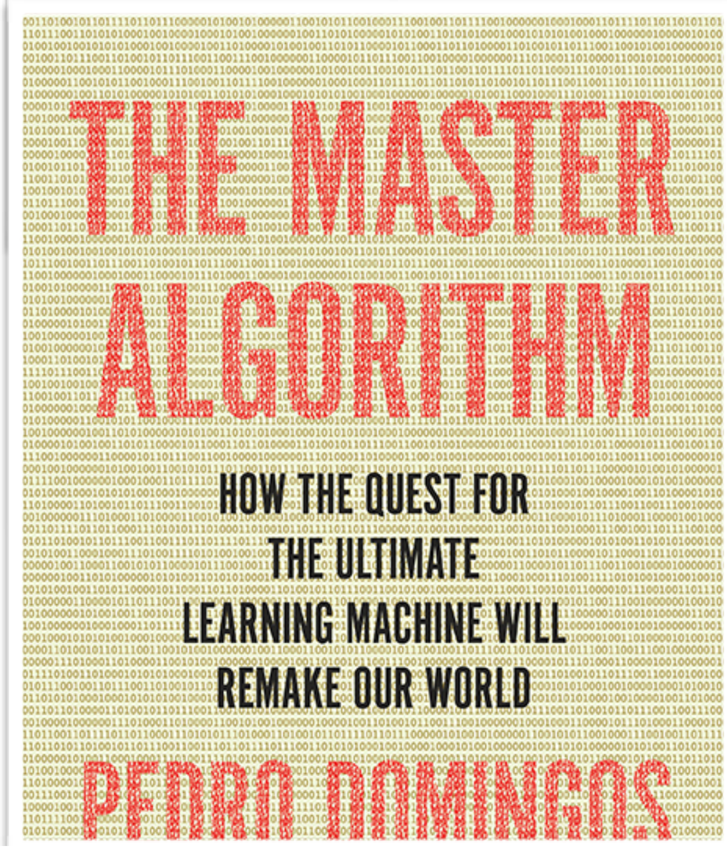
- The path to success in machine learning is quality data
- Simple ideas can lead to powerful results
- Common Pitfalls
  - Overfitting
  - Curse of Dimensionality

THE END
APPENDIX

REFERENCES:

- Google's Jeff Dean About NNs Arch: slides
- Quota's Xavier Amatriain: slides
- This is a good tech debt paper http://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf
- Nando Defreitas @ Oxford Excellent Lectures link
- Video: The Unreasonable Effectiveness of Data
- http://mlss.tuebingen.mpg.de/2015/speakers.html





REFERENCES:

- Tensor Flow: https://news.ycombinator.com/item?id=10532957 Google Deep learning overview
- Speech recognition history: Siri
- DSP vs. ML: https://www.quora.com/What-are-the-connections-between-machine-learning-and-signal-processing
- What's the hype about Deep learning? look at min 5 of this figure: https://www.youtube.com/watch?v=UREUlUDo4Kk
- MS Algo Guide
- Great 20 min overview The data science revolution