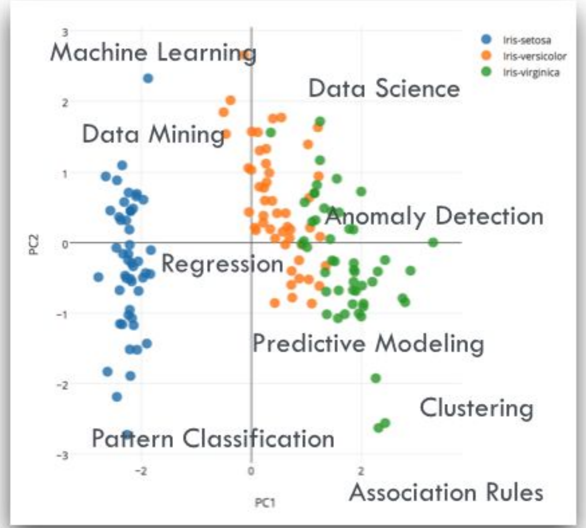
Practical DS in NLP

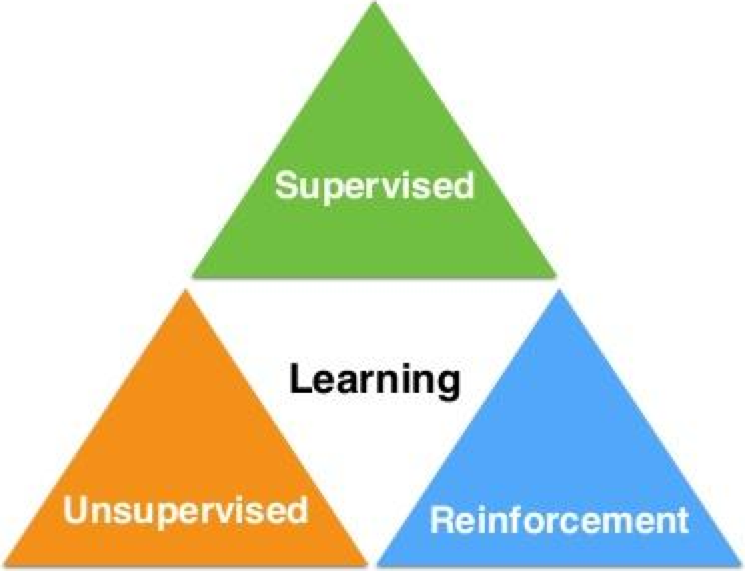
Supervised Learning and Pattern Recognition

laampt@gmail.com

Big picture



- Labeled data
- Direct feedback
- Predict outcome/future

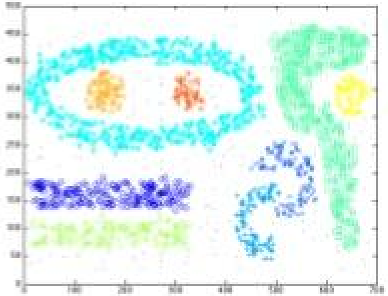


- No labels
- No feedback
- “Find hidden structure”

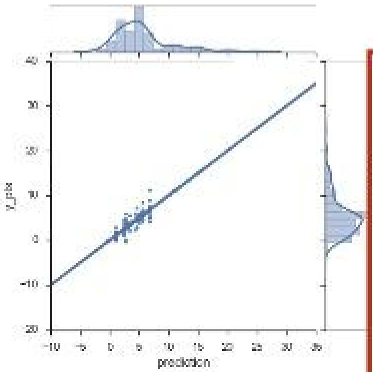
- Decision process
- Reward system
- Learn series of actions

Unsupervised Learning

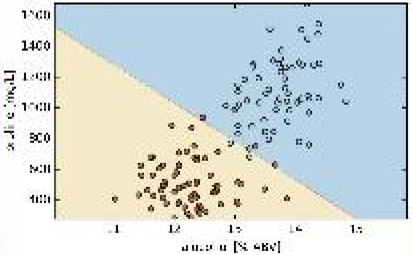
Supervised Learning



Clustering:
[DBSCAN on a toy dataset]



Regression:
[Soccer Fantasy Score prediction]

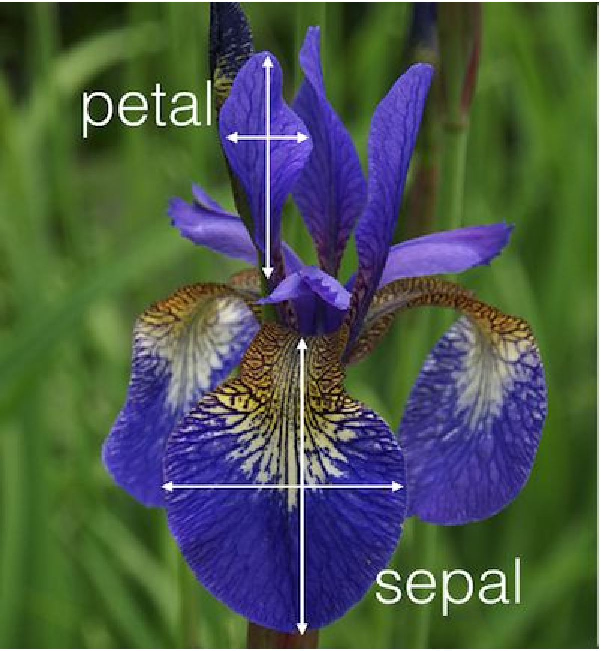


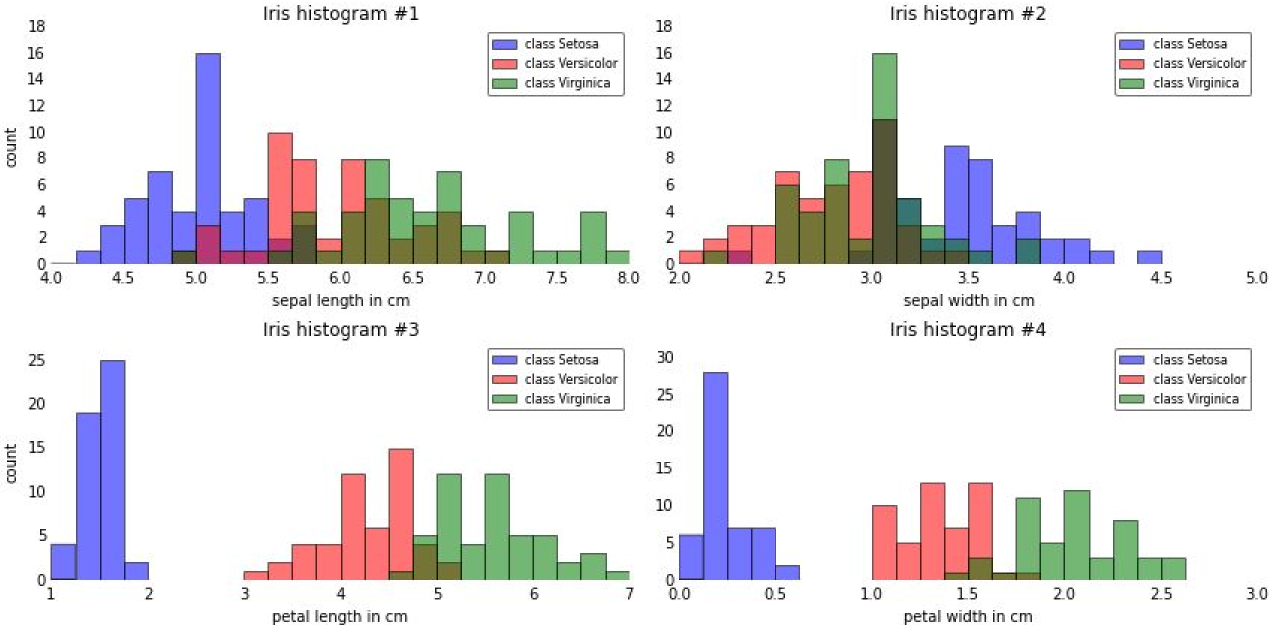
Classification:
[SVM on 2 classes of the Wine dataset]

Today's topic

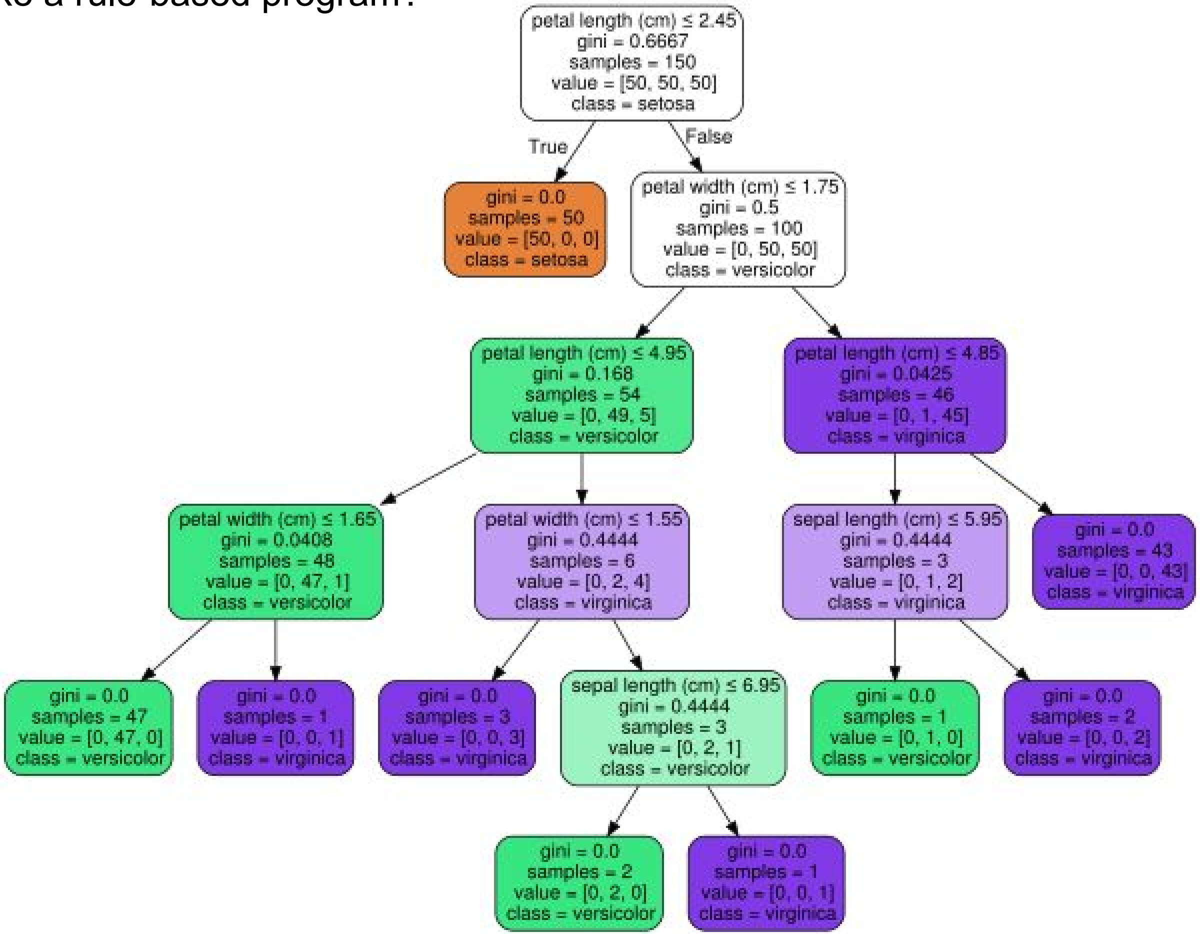
SUPERVISED LEARNING

Super classical example: IRIS

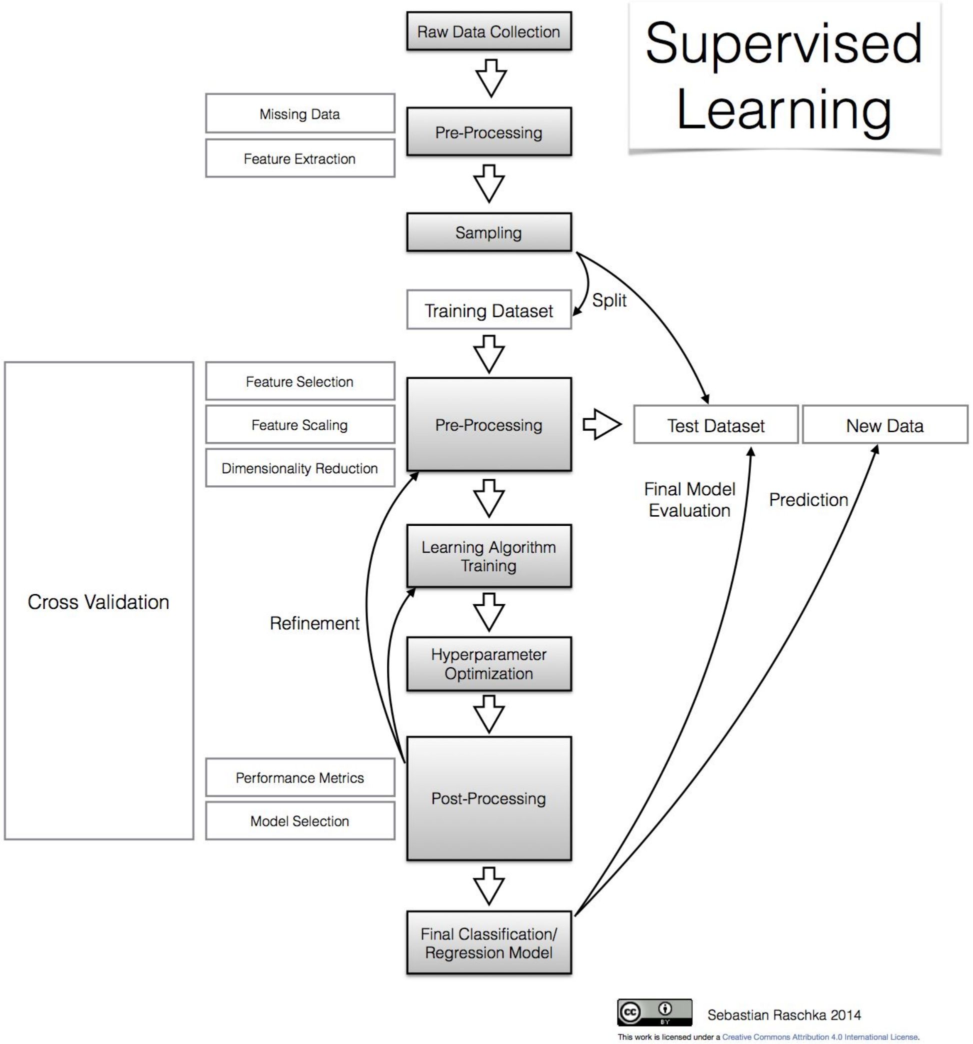




Easily to explain like a rule-based program?

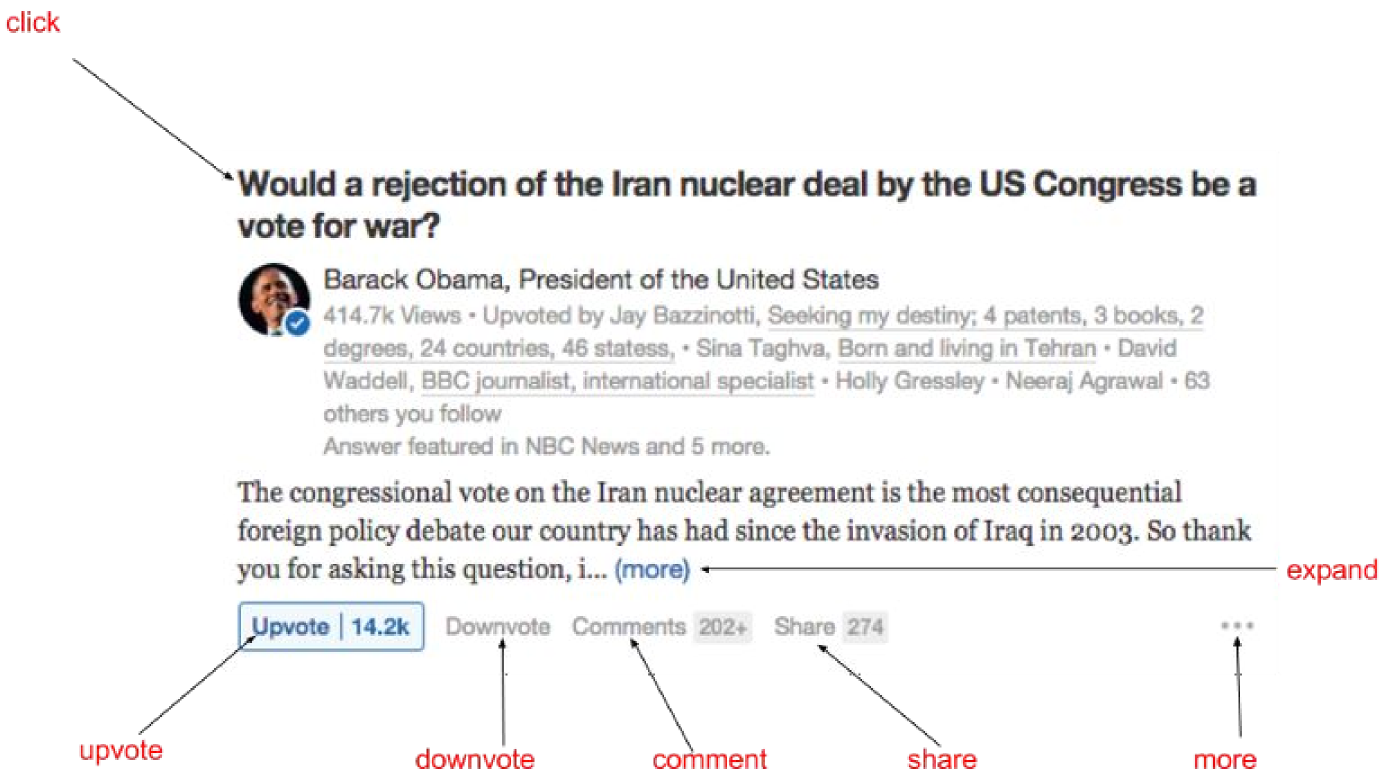


WorkFlow


Supervised Learning
Raw Data Collection
Missing Data
Feature Extraction
Pre-Processing
Sampling
Split
Training Dataset
Cross Validation
Feature Selection
Feature Scaling
Dimensionality Reduction
Pre-Processing
Test Dataset
New Data
Final Model Evaluation
Prediction
Learning Algorithm Training
Refinement
Hyperparameter Optimization
Performance Metrics
Model Selection
Post-Processing
Final Classification/ Regression Model
Sebastian Raschka 2014

FEATURE ENGINEERING

"At the end of the day, some machine learning projects succeed and some fail. What makes the difference? Easily the most important factor is the features used. If you have many independent features that each correlate well with the class, learning is easy. On the other hand, if the class is a very complex function of the features, you may not be able to learn it. Often, the raw data is not in a form that is amenable to learning, but you can construct features from it that are. This is typically where most of the effort in a machine learning project goes." Pedro Domingos



Would a rejection of the Iran nuclear deal by the US Congress be a vote for war?

Barack Obama, President of the United States

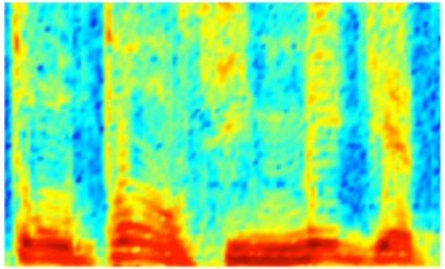
414.7k Views • Upvoted by Jay Bazzinotti, Seeking my destiny; 4 patents, 3 books, 2 degrees, 24 countries, 46 statess, • Sina Taghva, Born and living in Tehran • David Waddell, BBC journalist, international specialist • Holly Gressley • Neeraj Agrawal • 63 others you follow

Answer featured in NBC News and 5 more.

The congressional vote on the Iran nuclear agreement is the most consequential foreign policy debate our country has had since the invasion of Iraq in 2003. So thank you for asking this question, i... (more)

Data2Vec

AUDIO



Audio Spectrogram

DENSE

IMAGES

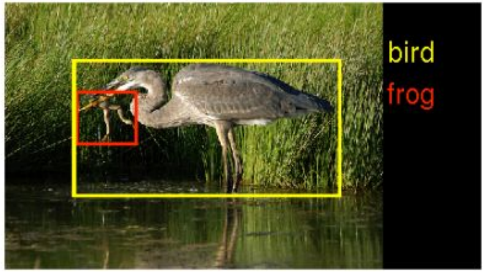


Image pixels

DENSE

TEXT

0	0	0	0.2	0	0.7	0	0	0	...	...

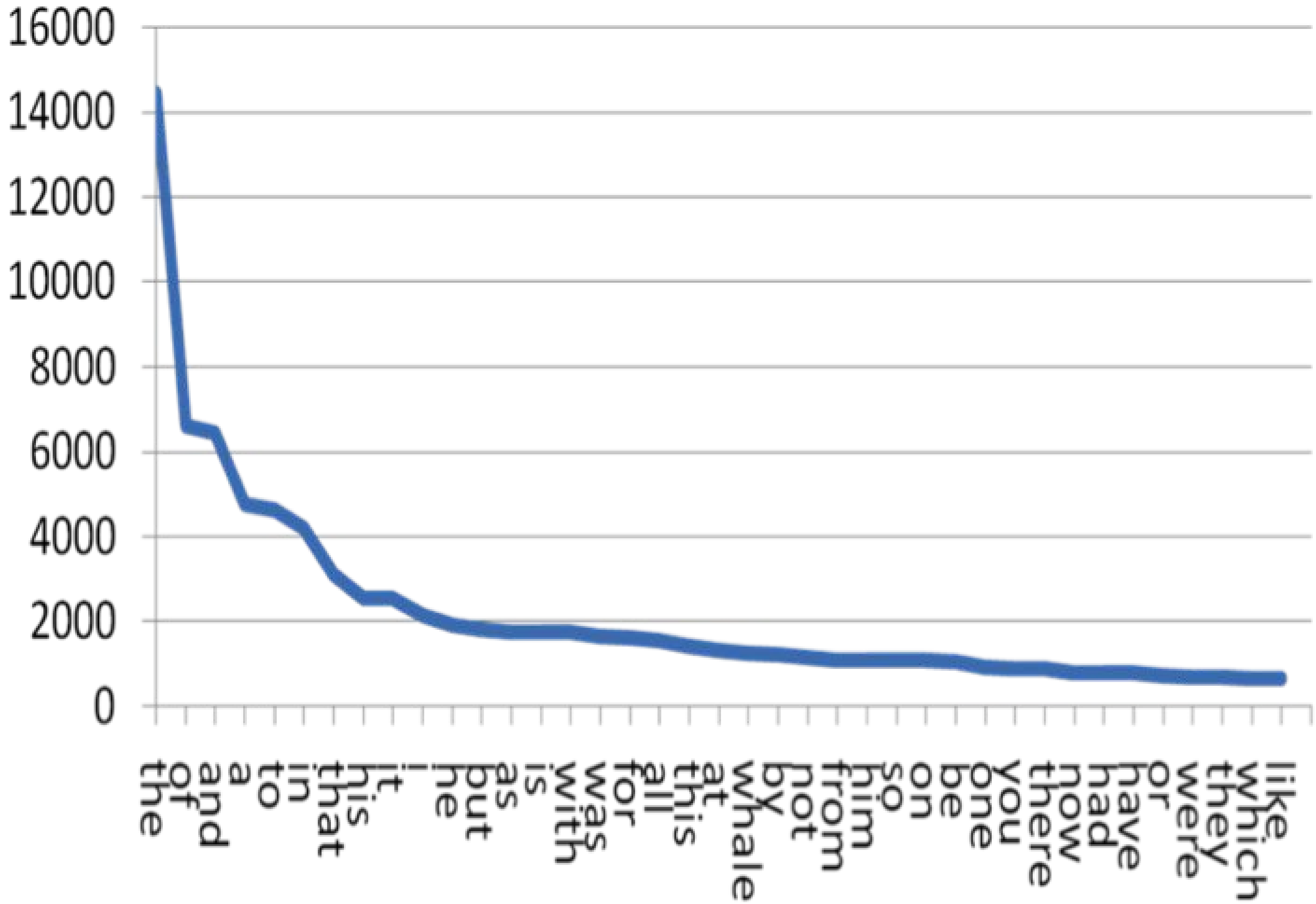
Word, context, or
document vectors

SPARSE

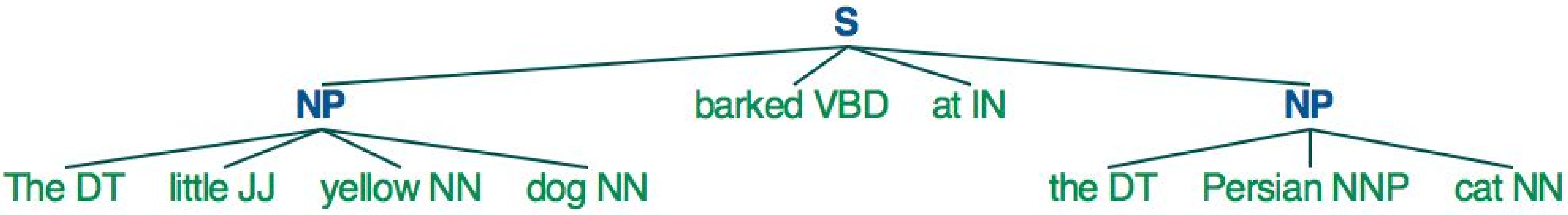
NLP

Understanding and Representing text and meaning

Zipf’s law


16000
14000
12000
10000
8000
6000
4000
2000
0
the
of
and
a
to
in
that
his
it
I
he
but
as
is
with
was
for
all
this
at
whale
by
not
from
him
so
on
be
one
you
there
now
had
have
or
were
they
which
like

Structural/Hierarchical



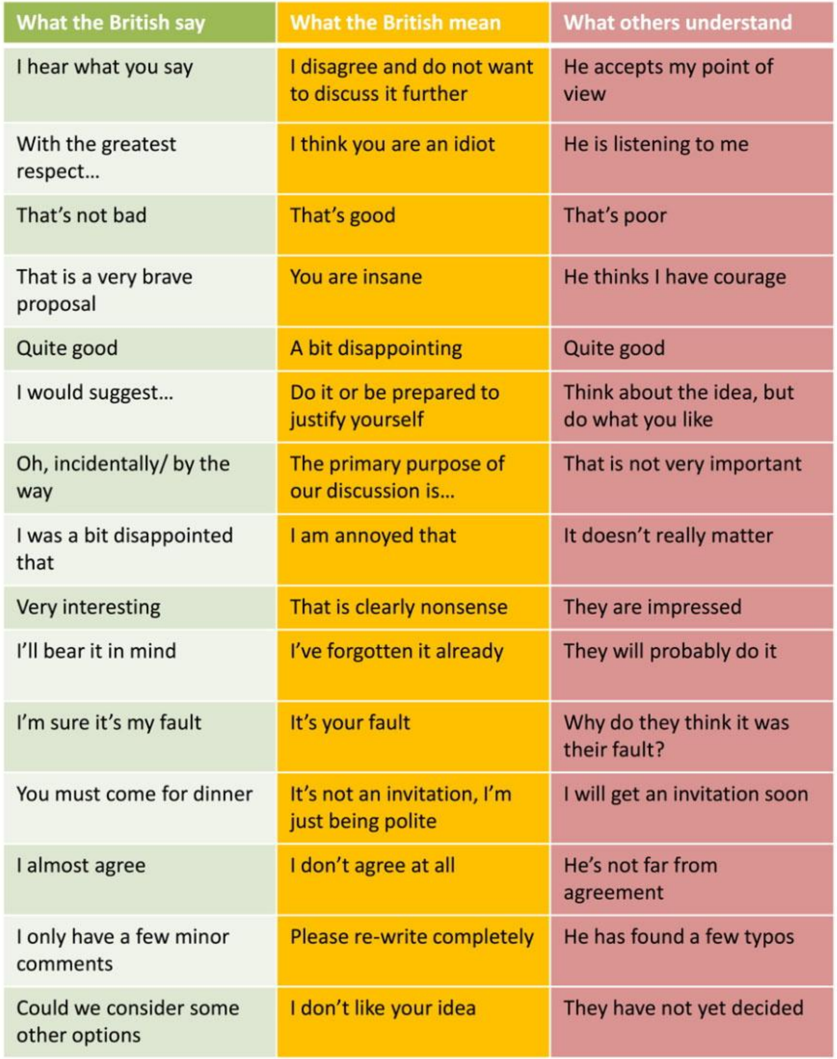
Ambiguity

Anh oi anh dang o dau vay, den day le em dang coi quan ne, anh den nho mua bao nha, o day toan bao cu ko ah, ma thoi anh khoi mua vi em vua mat kinh roi, anh den le di em chiu het noi roi, anh oi

gare (nhà ga)

Hát câu i tờ đón Xuân về ...

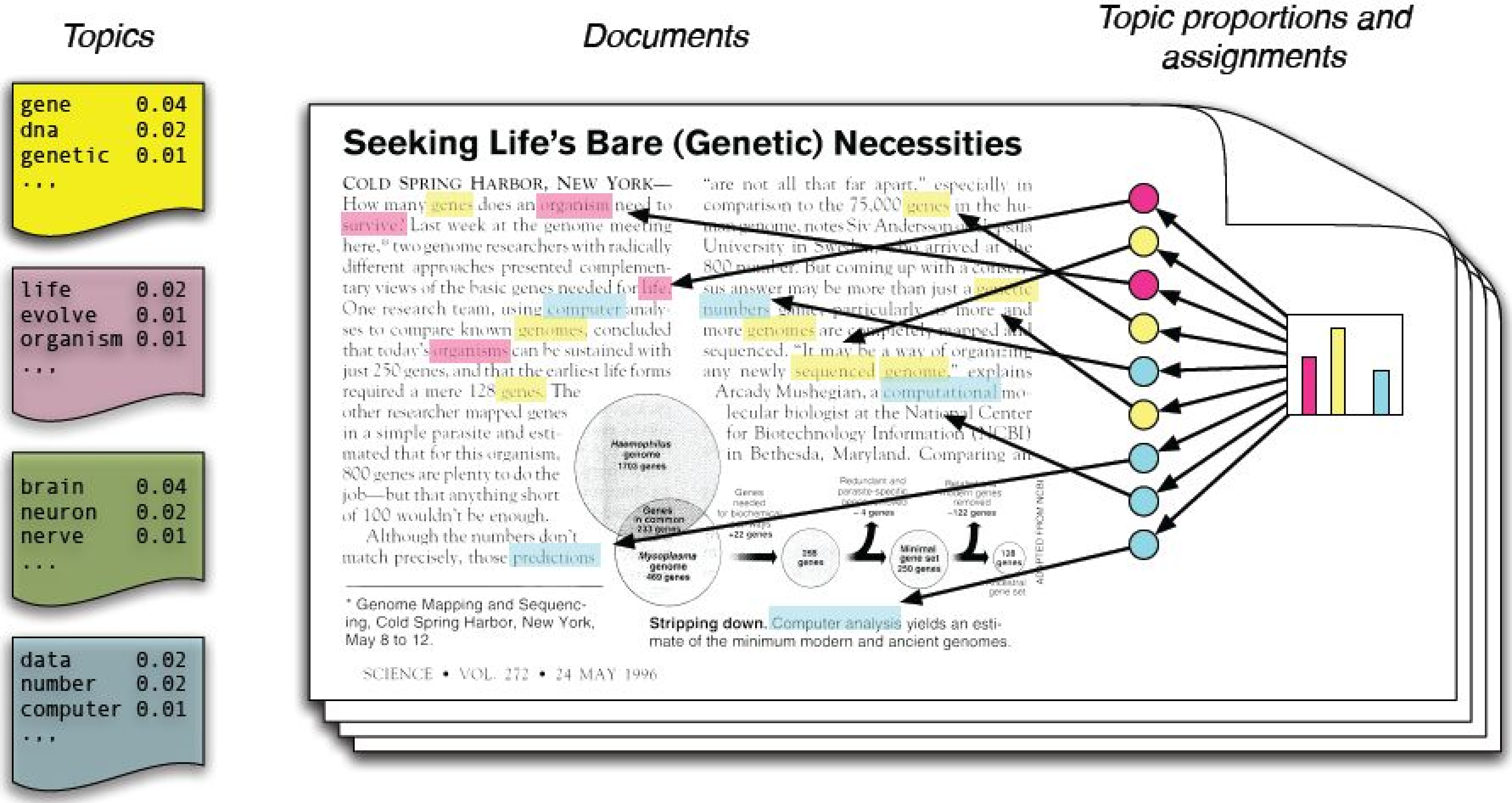
Sarcasm



What the British say	What the British mean	What others understand
I hear what you say	I disagree and do not want to discuss it further	He accepts my point of view
With the greatest respect…	I think you are an idiot	He is listening to me
That's not bad	That's good	That's poor
That is a very brave proposal	You are insane	He thinks I have courage
Quite good	A bit disappointing	Quite good
I would suggest…	Do it or be prepared to justify yourself	Think about the idea, but do what you like
Oh, incidentally/ by the way	The primary purpose of our discussion is…	That is not very important
I was a bit disappointed that	I am annoyed that	It doesn't really matter
Very interesting	That is clearly nonsense	They are impressed
I'll bear it in mind	I've forgotten it already	They will probably do it
I'm sure it's my fault	It's your fault	Why do they think it was their fault?
You must come for dinner	It's not an invitation, I'm just being polite	I will get an invitation soon
I almost agree	I don't agree at all	He's not far from agreement
I only have a few minor comments	Please re-write completely	He has found a few typos
Could we consider some other options	I don't like your idea	They have not yet decided

What they say: Merry Christmas!
What they mean: It’s the middle of summer; let’s have a BBQ and drink outdoors.

Australian guy :)



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Redundant modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

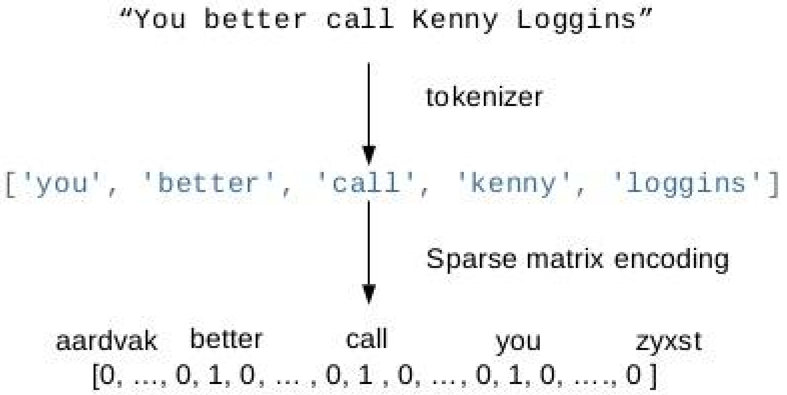
SCIENCE • VOL. 272 • 24 MAY 1996

VSM

Vector Space Model

Bag Of Word Representations

`CountVectorizer / TfidfVectorizer`



“You better call Kenny Loggins”

tokenizer

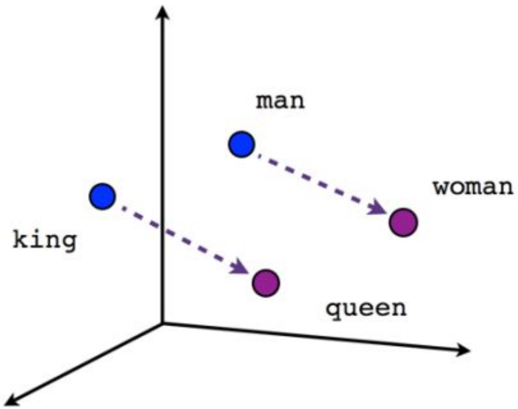
['you', 'better', 'call', 'kenny', 'loggins']

Sparse matrix encoding

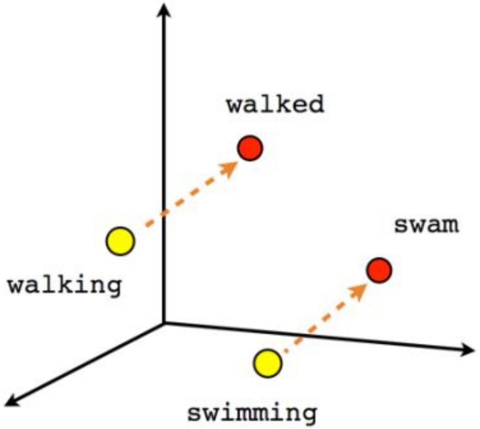
aardvak better call you zyxst

[0, …, 0, 1, 0, … , 0, 1 , 0, …, 0, 1, 0, …., 0]

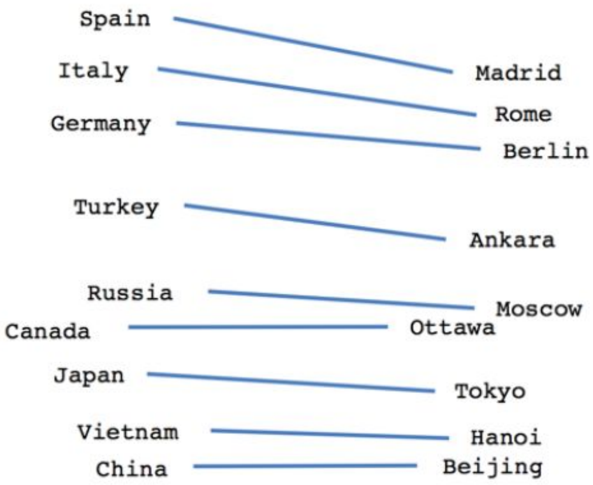
Word2Vec



Male-Female



Verb tense



Country-Capital

Categorical Variables

	color	size	prize	class
0	green	M	10.1	class1
1	red	L	13.5	class2
2	blue	XL	15.3	class1

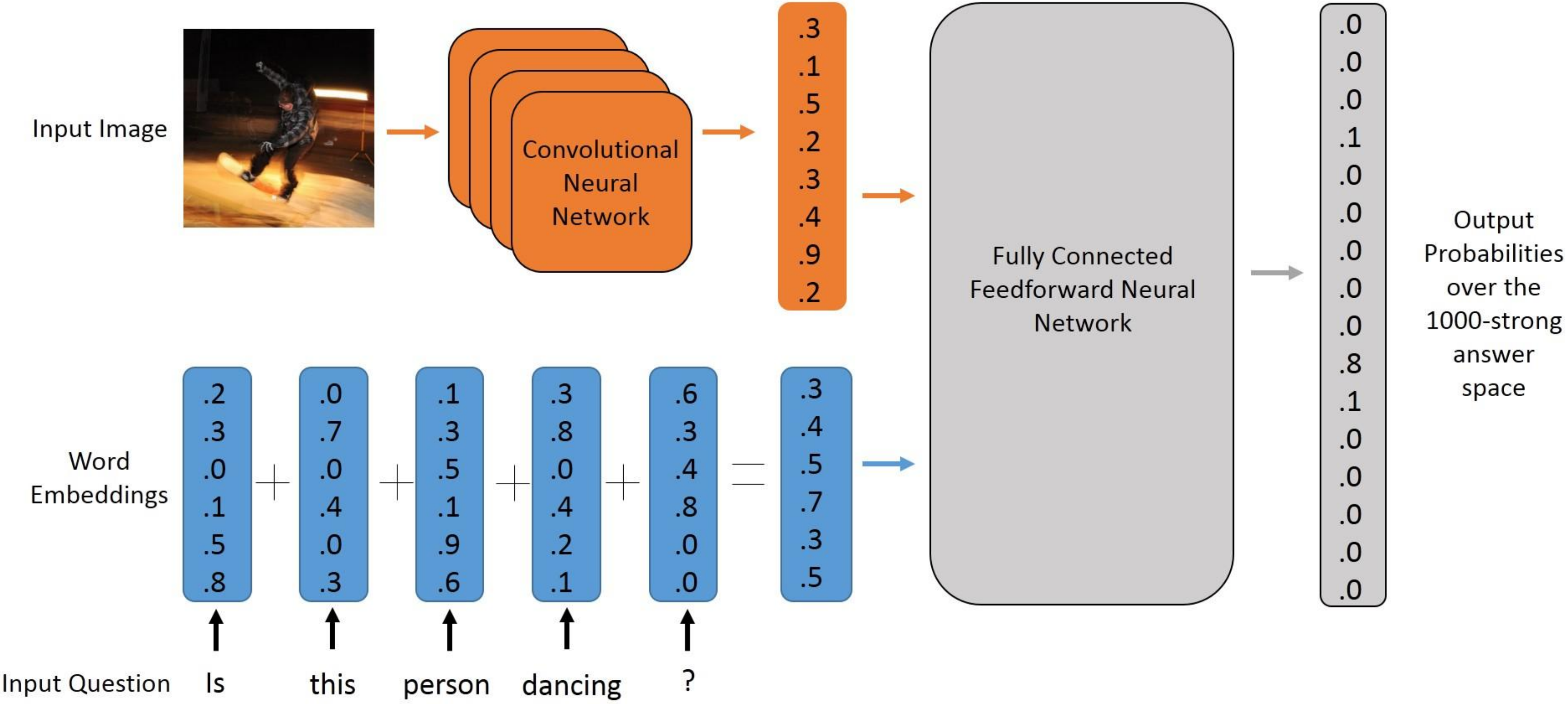
nominal

green → (1,0,0)
red → (0,1,0)
blue → (0,0,1)

ordinal

M → 1
L → 2
XL → 3

	class	color=blue	color=green	color=red	prize	size
0	0	0	1	0	10.1	1
1	1	0	0	1	13.5	2
2	0	1	0	0	15.3	3



Input Image

Convolutional Neural Network

Fully Connected Feedforward Neural Network

Output Probabilities over the 1000-strong answer space

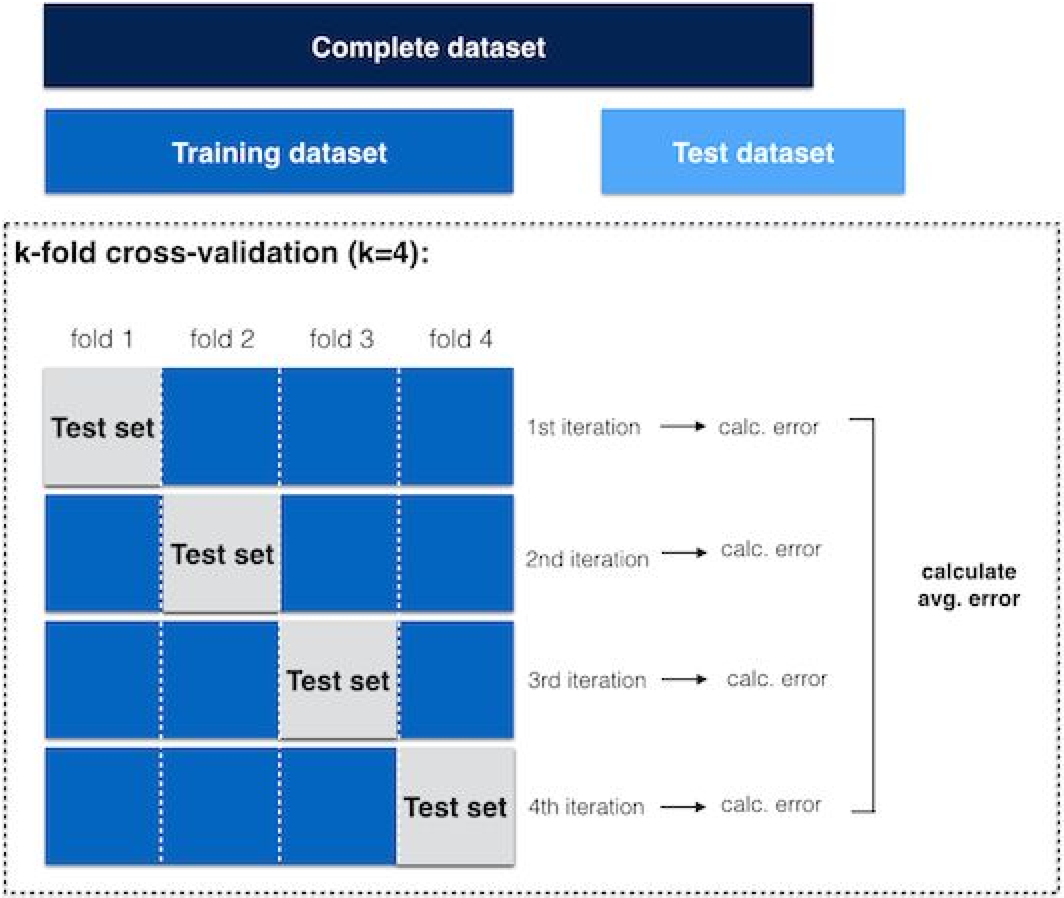
Word Embeddings

Input Question

Is this person dancing ?

Cross Validation

CV



Complete dataset

Training dataset

Test dataset

k-fold cross-validation (k=4):

fold 1 fold 2 fold 3 fold 4

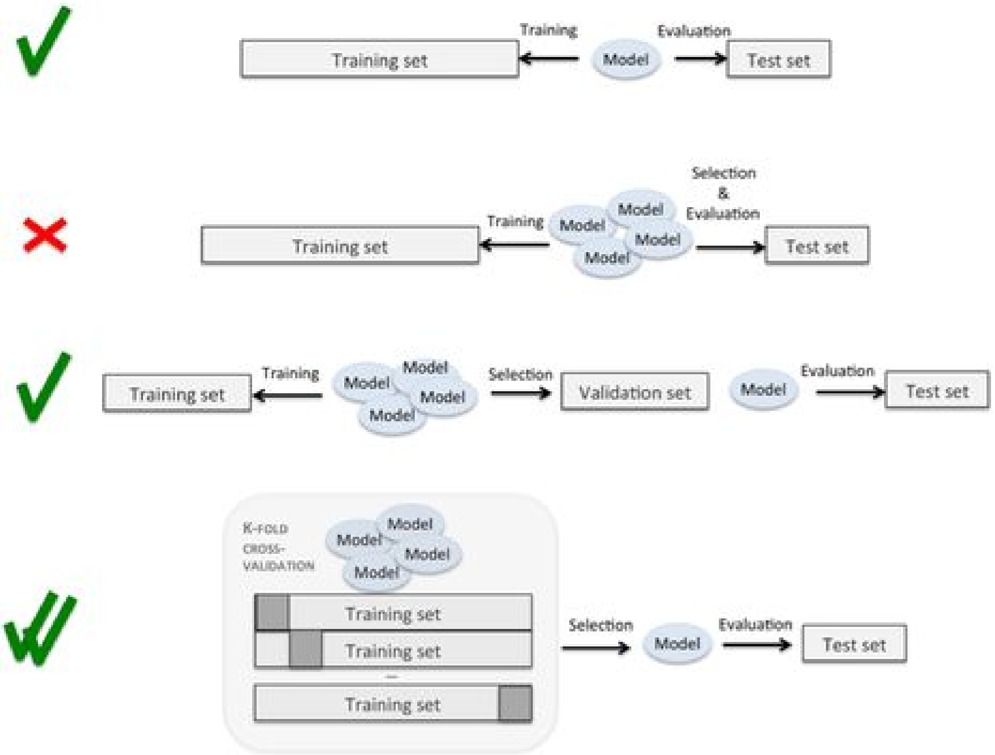
Test set — 1st iteration → calc. error

Test set — 2nd iteration → calc. error

Test set — 3rd iteration → calc. error

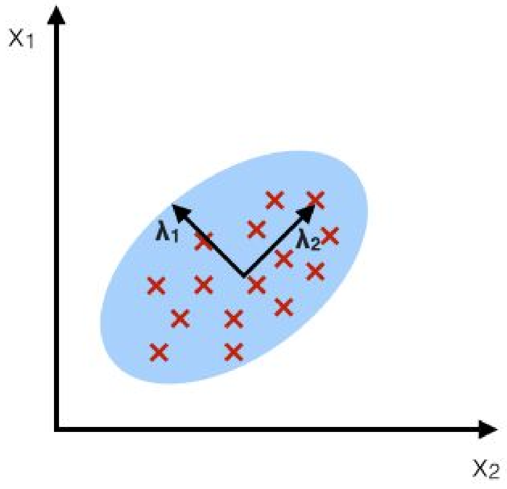
Test set — 4th iteration → calc. error

calculate avg. error

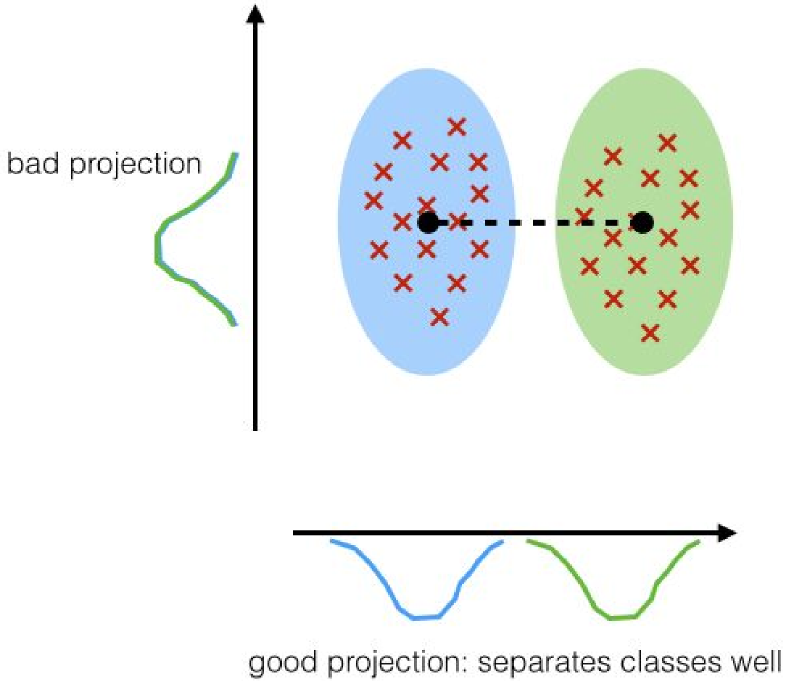

✓
Training set
Training
Model
Evaluation
Test set
✗
Training set
Training
Model
Model
Model
Model
Selection
&
Evaluation
Test set
✓
Training set
Training
Model
Model
Model
Model
Selection
Validation set
Model
Evaluation
Test set
✓✓
K-fold
cross-
validation
Model
Model
Model
Model
Training set
Training set
...
Training set
Selection
Model
Evaluation
Test set

Feature Selection

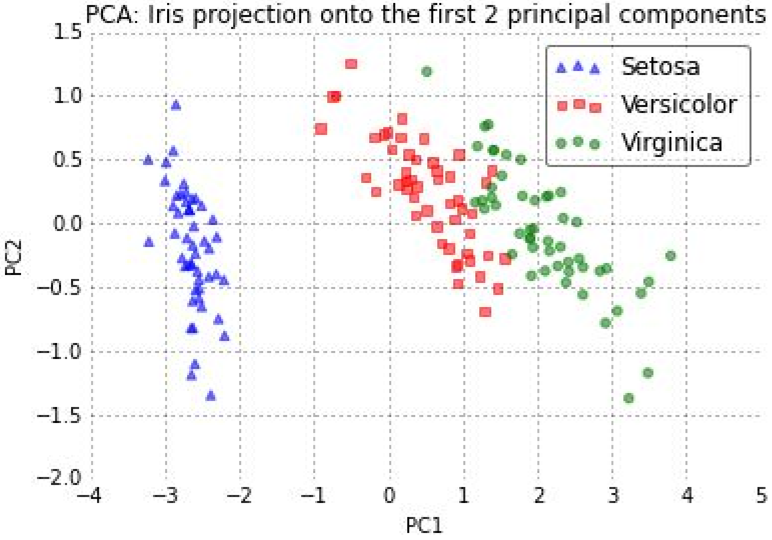
PCA:
component axes that maximize the variance



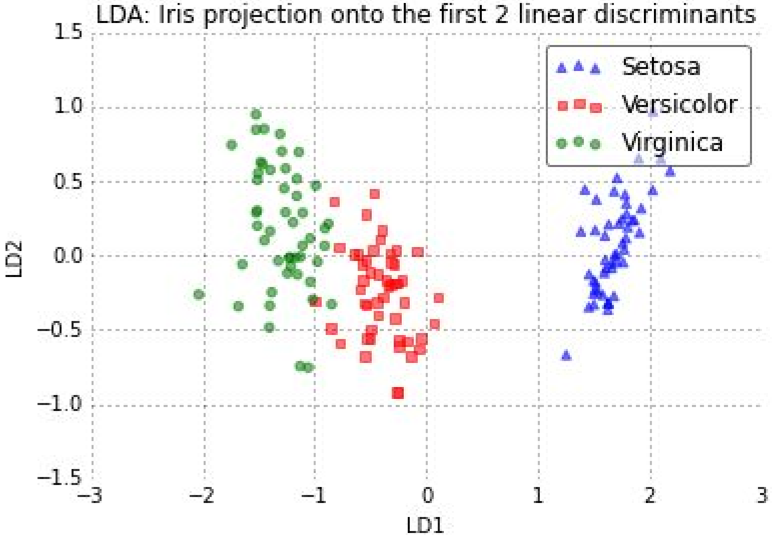
LDA:
maximizing the component axes for class-separation

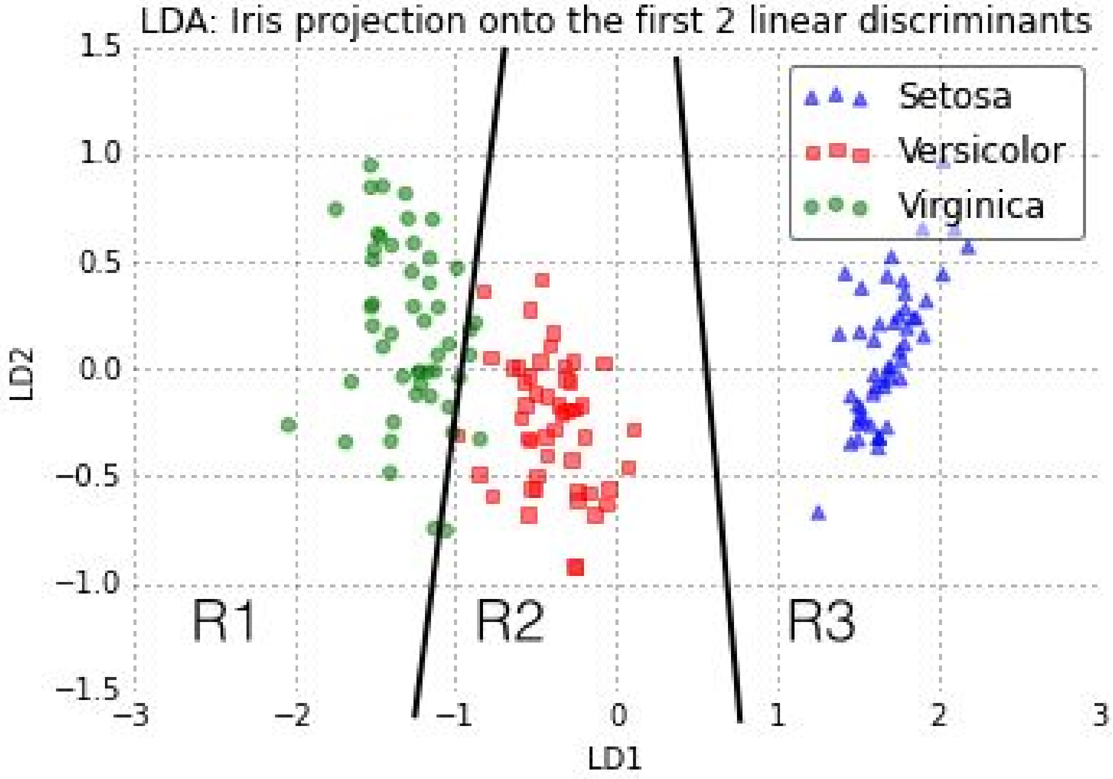


PCA

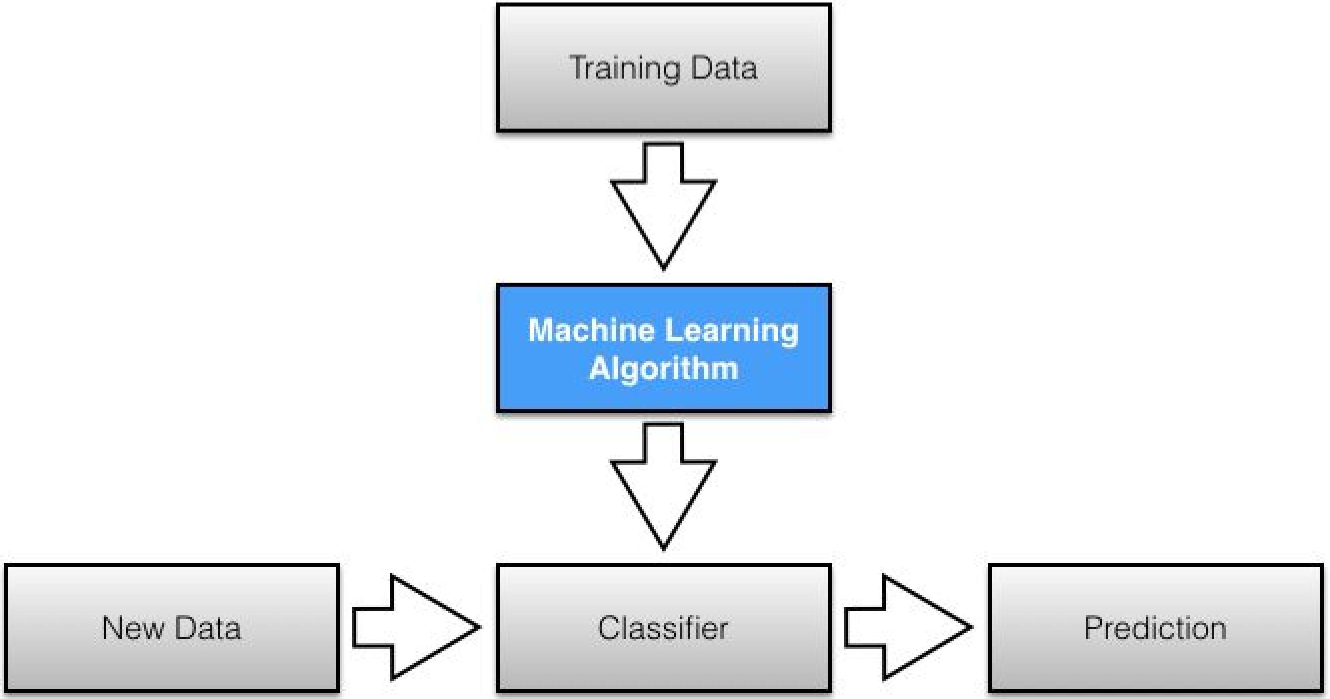

PCA: Iris projection onto the first 2 principal components
▲ Setosa
■ Versicolor
● Virginica
PC1
PC2
−4
−3
−2
−1
0
1
2
3
4
5
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0

LDA

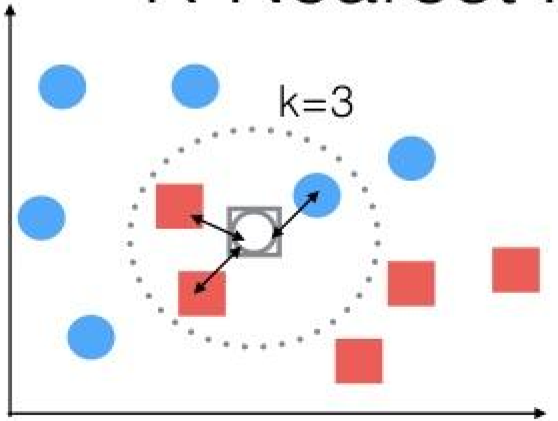

LDA: Iris projection onto the first 2 linear discriminants
▲ Setosa
■ Versicolor
● Virginica
LD1
LD2
−3
−2
−1
0
1
2
3
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

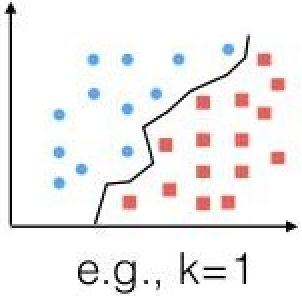

LDA: Iris projection onto the first 2 linear discriminants
▲ Setosa
■ Versicolor
● Virginica
R1
R2
R3
LD1
LD2
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−3
−2
−1
0
1
2
3

Learning Algos and hyperparameters tuning



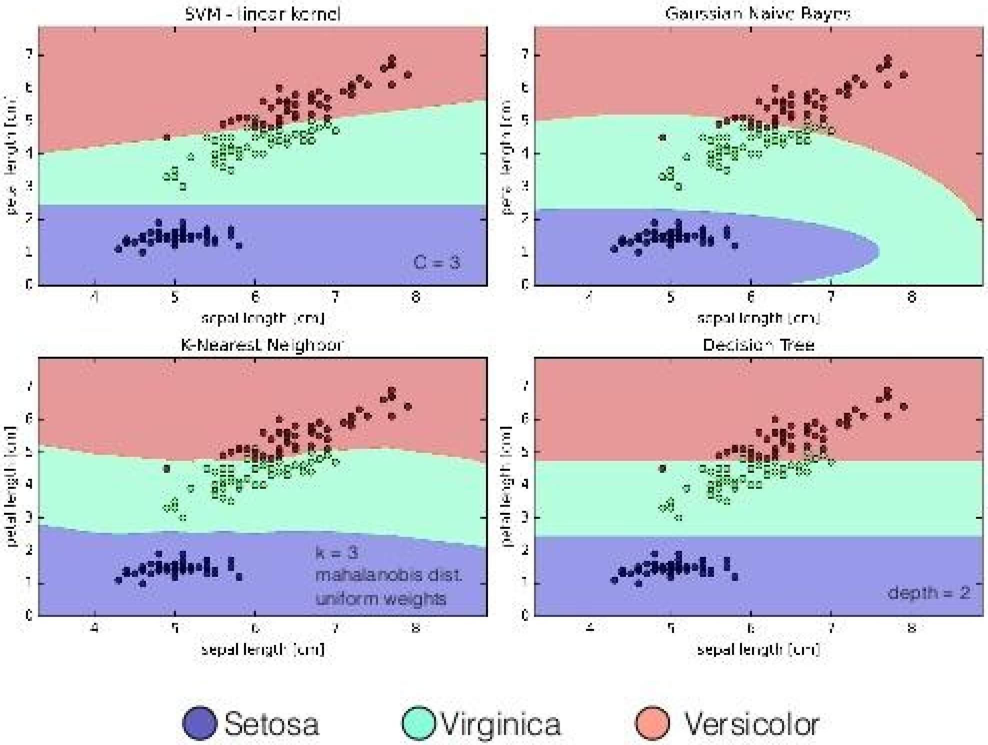
Non-Parametric Classifiers: K-Nearest Neighbor



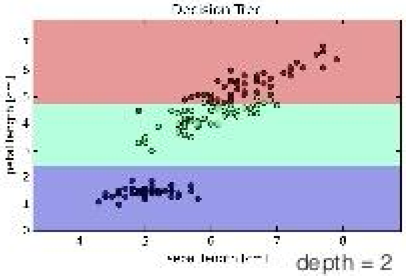


- Simple!
- Lazy learner
- Very susceptible to curse of dimensionality

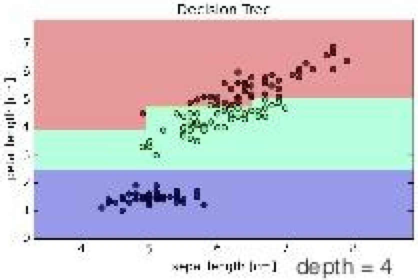
Iris Example



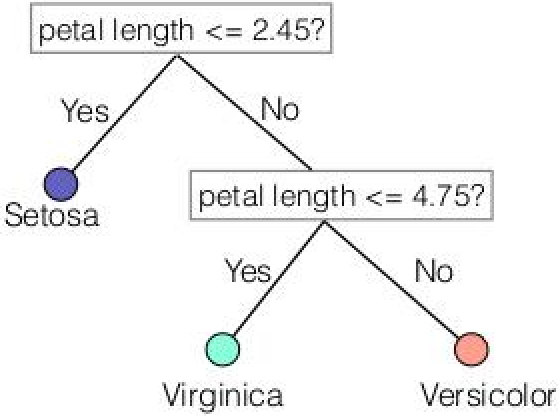
Decision Tree



depth = 2



depth = 4



$$\text{Entropy} = \sum_i -p_i \log_k p_i$$

e.g., $2\,(-\,0.5 \log_2(0.5)) = 1$

Information Gain =
entropy(parent) – [avg entropy(children)]

"No Free Lunch" :(

D. H. Wolpert. The supervised learning no-free-lunch theorems. In Soft Computing and Industry, pages 25–42. Springer, 2002.

Our model is a simplification of reality

⇩

Simplification is based on assumptions (model bias)

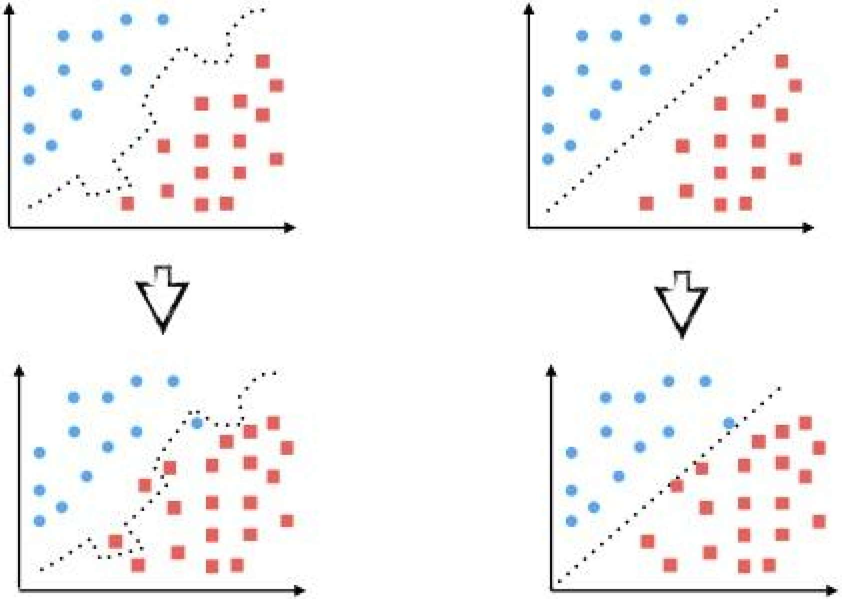
⇩

Assumptions fail in certain situations

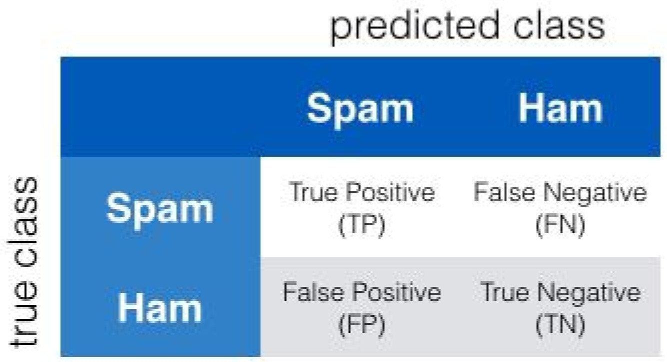
Roughly speaking:

"No one model works best for all possible situations."

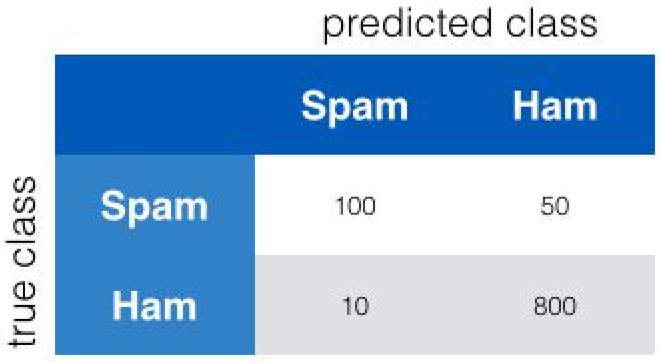
Generalization Error and Overfitting



Evaluation metrics



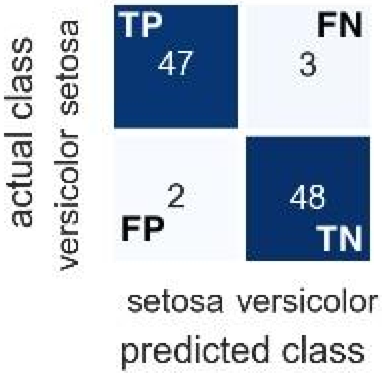
		predicted class	
		Spam	Ham
true class	Spam	True Positive (TP)	False Negative (FN)
	Ham	False Positive (FP)	True Negative (TN)



		predicted class	
		Spam	Ham
true class	Spam	100	50
	Ham	10	800

Error Metrics

here: “setosa” = “positive”



		predicted class	
		setosa	versicolor
actual class	setosa	TP 47	FN 3
	versicolor	2 FP	48 TN

[Linear SVM on sepal/petal lengths]

“micro” and “macro”
averaging for multi-class

$$\textit{Accuracy} = \frac{TP + TN}{FP + FN + TP + TN}$$

$$= 1 - \textit{Error}$$

$$\textit{False Positive Rate} = \frac{FP}{N}$$

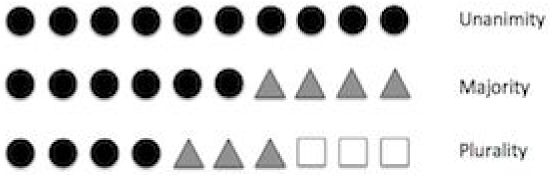
$$\begin{array}{c}\textit{True Positive Rate} \\ \textit{(Recall)}\end{array} = \frac{TP}{P}$$

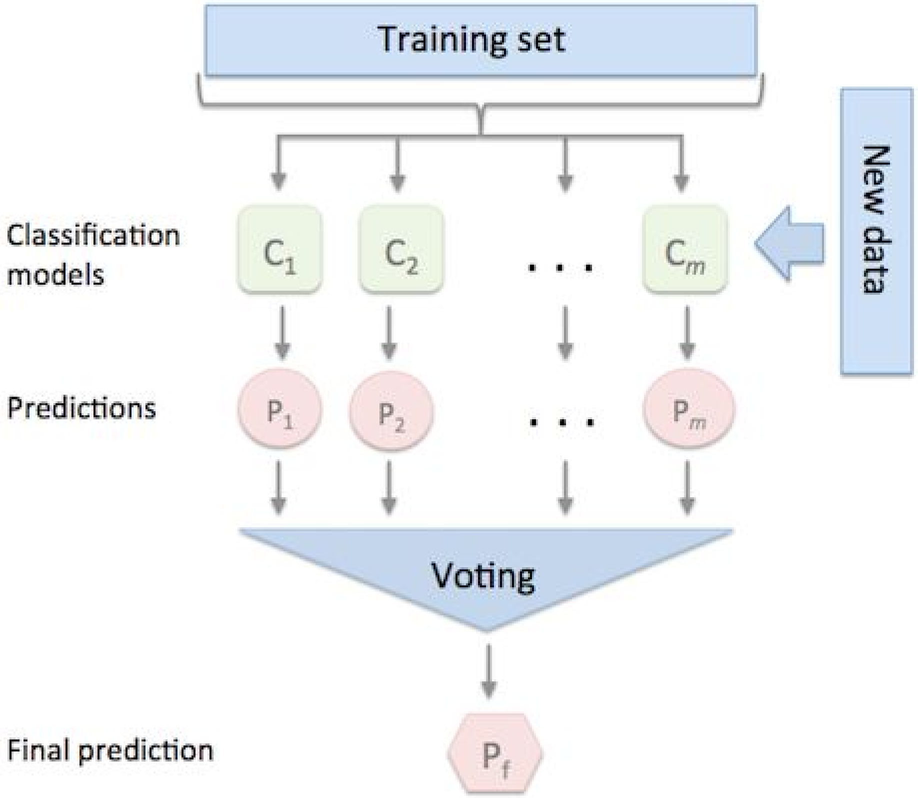
$$\textit{Precision} = \frac{TP}{TP + FP}$$

Part 2: Black art in ML

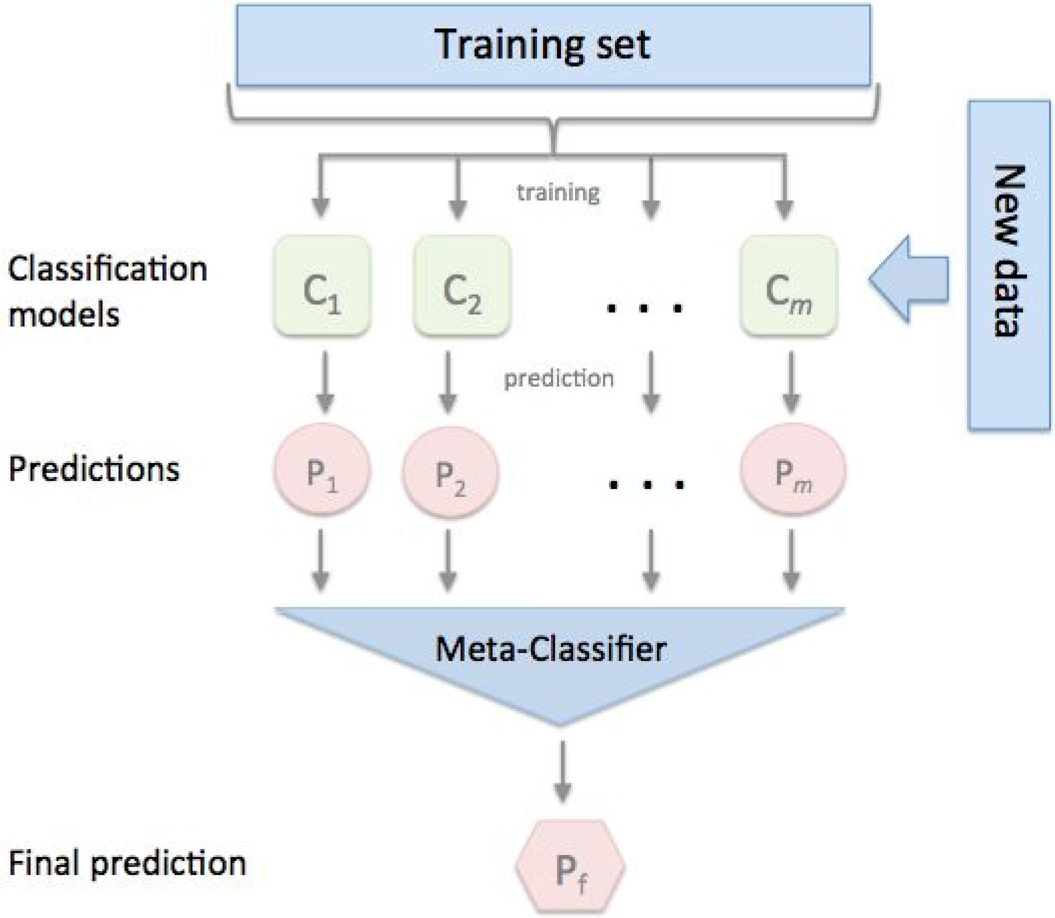
laampt@gmail.com

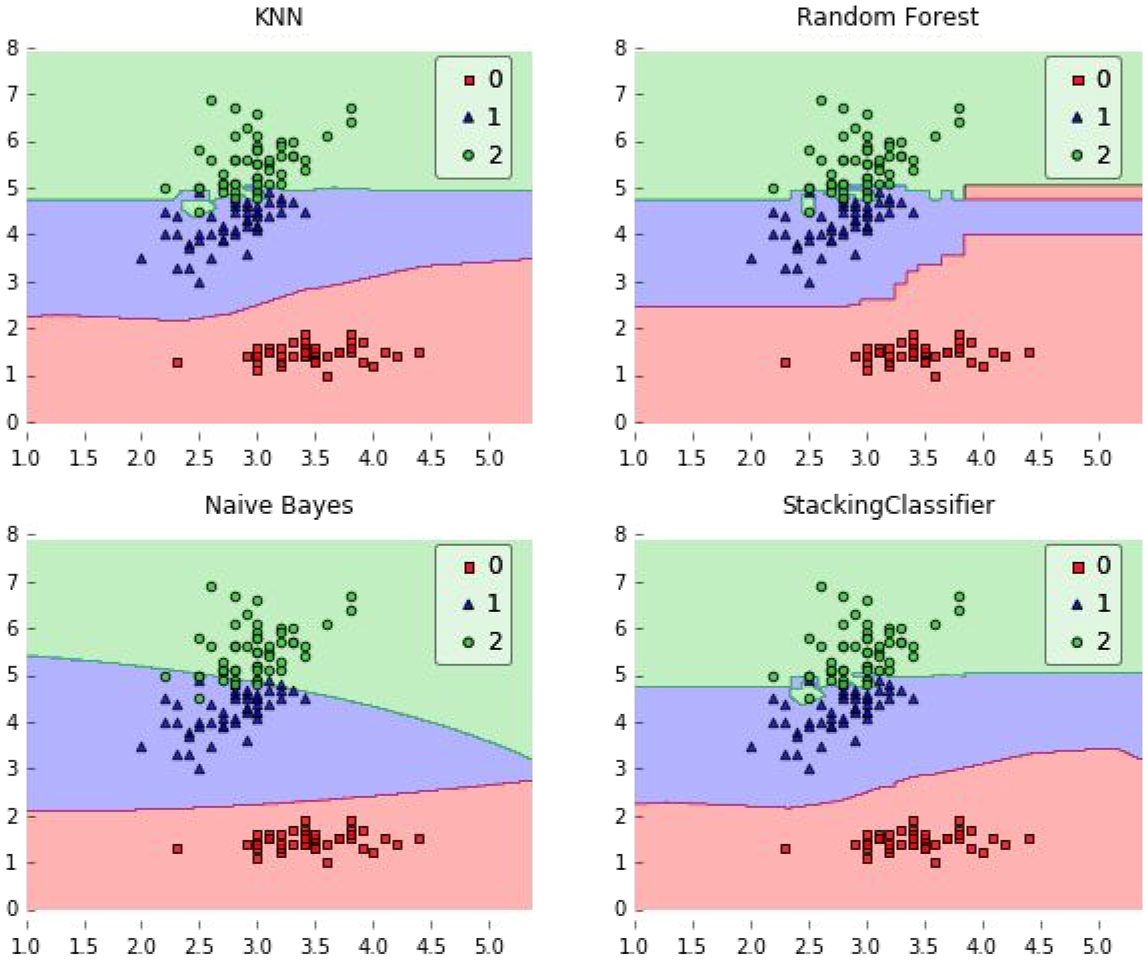
Ensemble: Voting

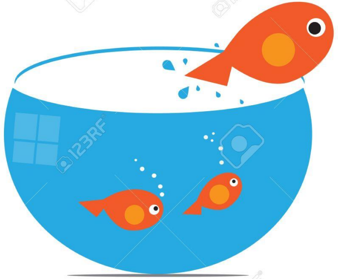




Ensemble: Stacking




KNN
Random Forest
Naive Bayes
StackingClassifier
0
1
2
0
1
2
3
4
5
6
7
8
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0



Q&A

Sometimes the questions are complicated but the answers are simple. Dr. Seuss