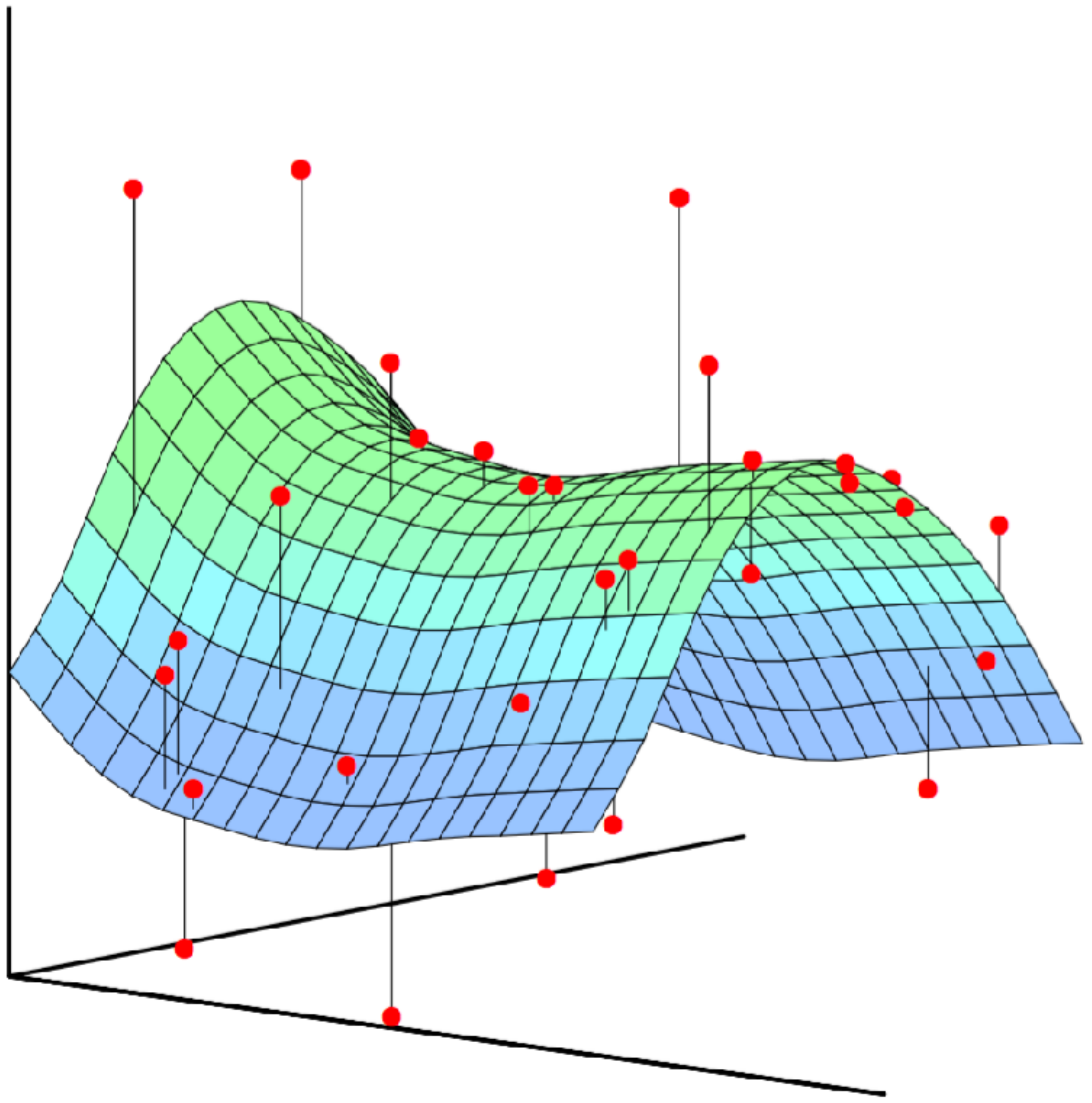
Introduction to Statistical Learning and Machine Learning

Chap 1 - Introduction

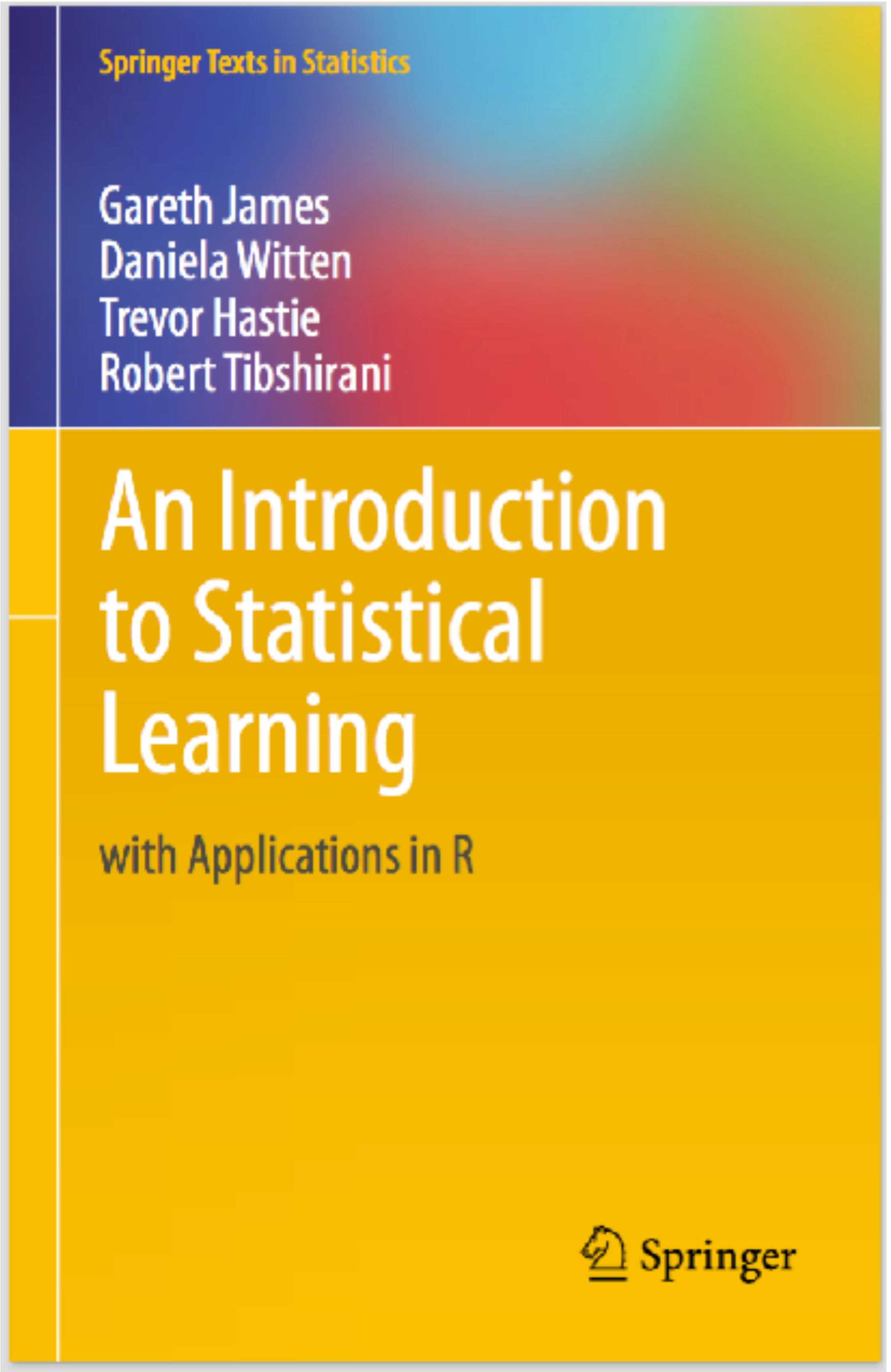
Yanwei Fu
SDS, Fudan University

Course Information

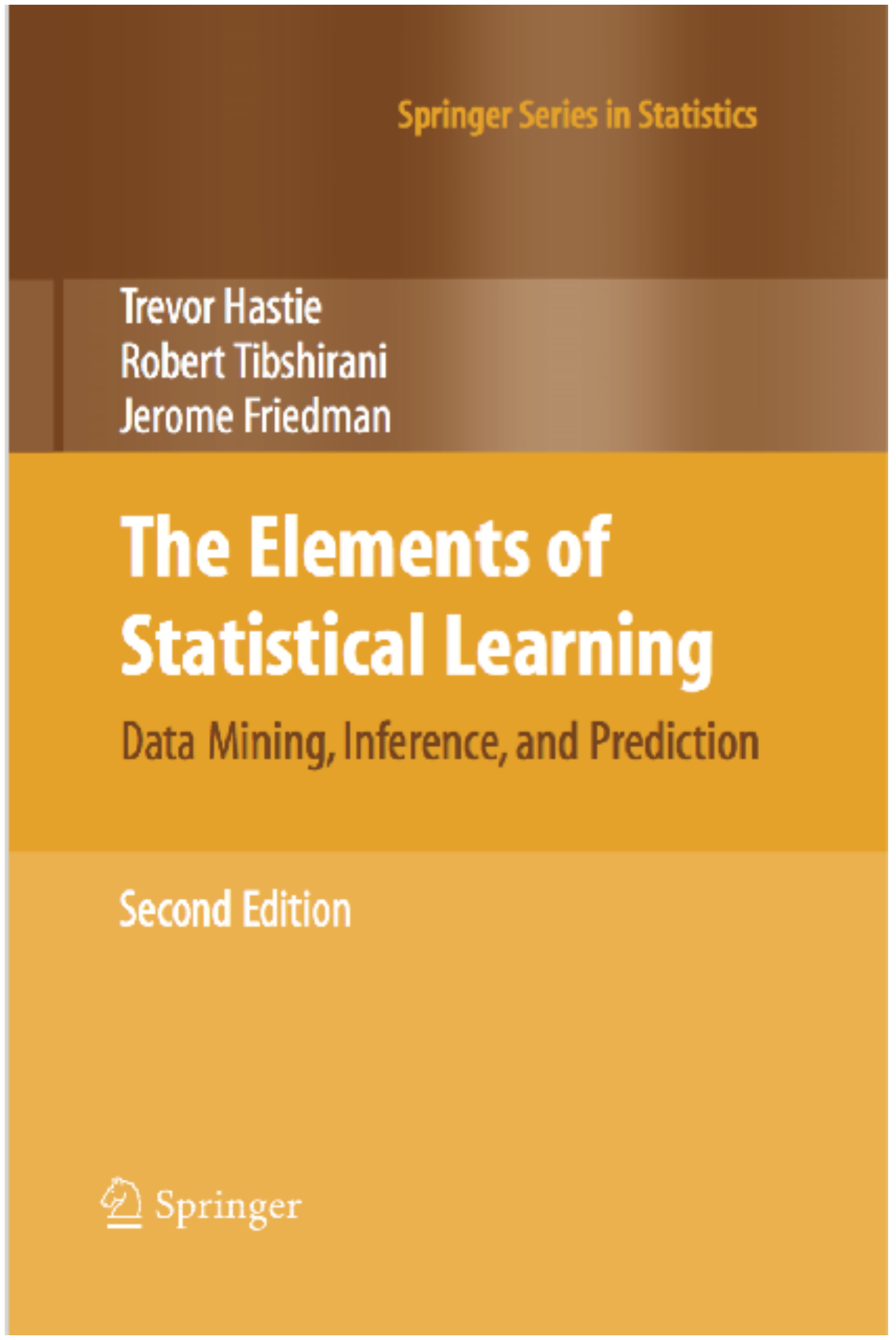
- Instructor: 付彦伟
- Email: yanweifu@fudan.edu.cn
- Course Websites:
  - http://www.sdspeople.fudan.edu.cn/fuyanwei/course/SLML.html
- Times&Venue:
  - Tue (11-13), H4204
- TA：张晟中 582195191@qq.com
- Office Hour: Wed. 4:00-5:30pm,
  - 子彬楼N211



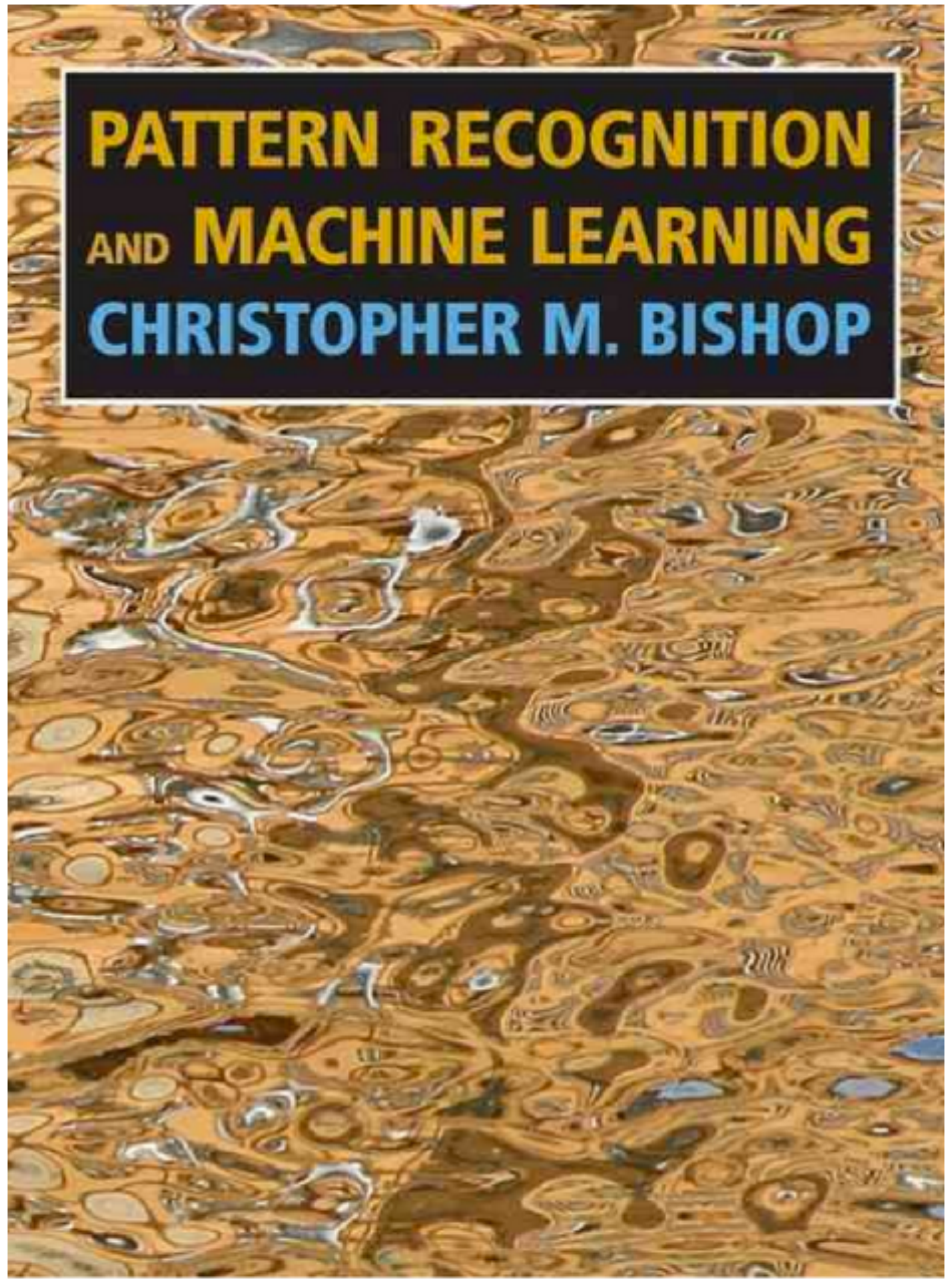
Textbook



James, Witten, Hastie and Tibshirani
An Introduction to Statistical Learning, with applications in R. Springer. 2013.

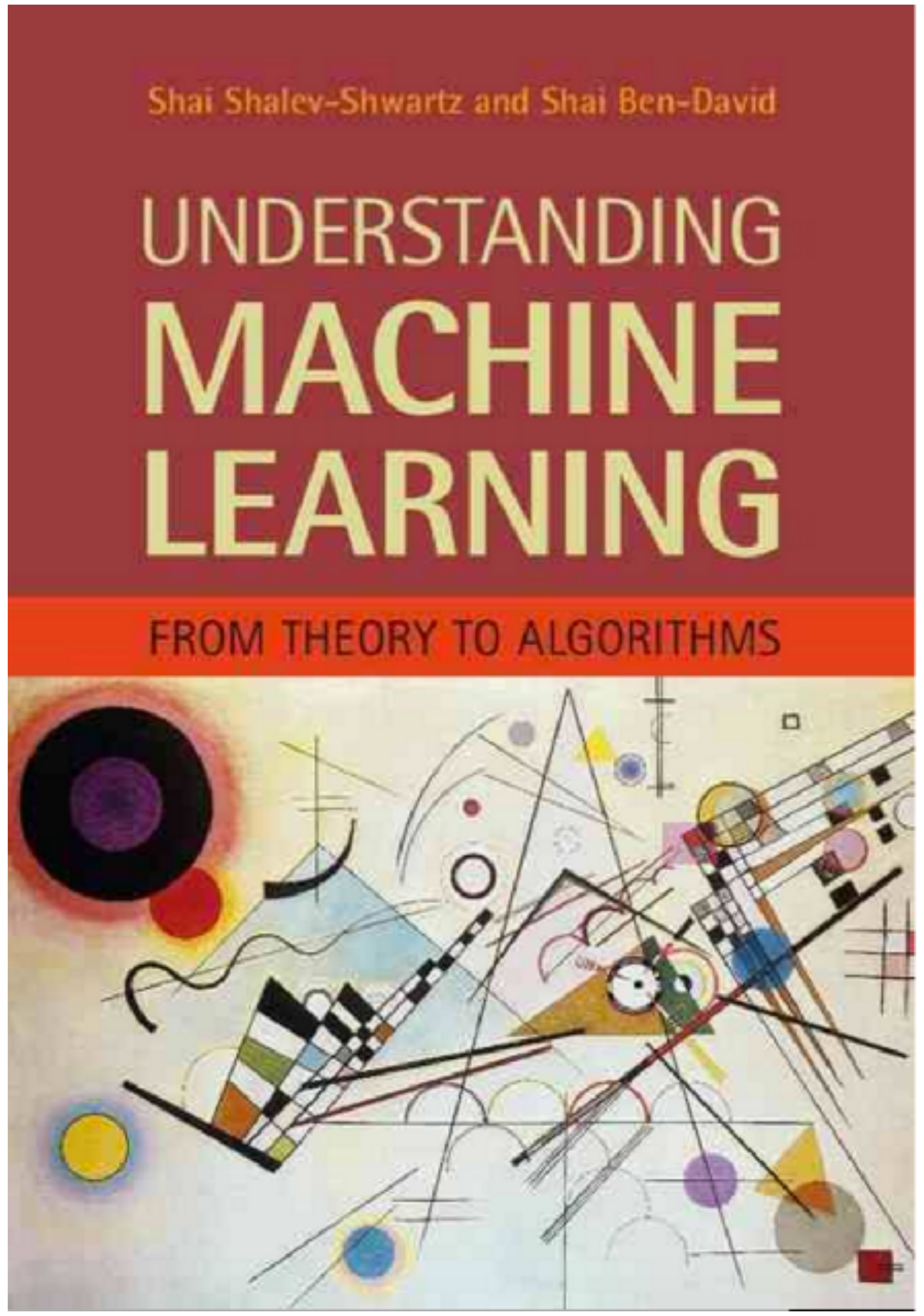


Hastie, Tibshurani, and Friedman
The Elements of Statistical Learning, data mining, inference and Prediction, 2nd Edition. Springer. 2011

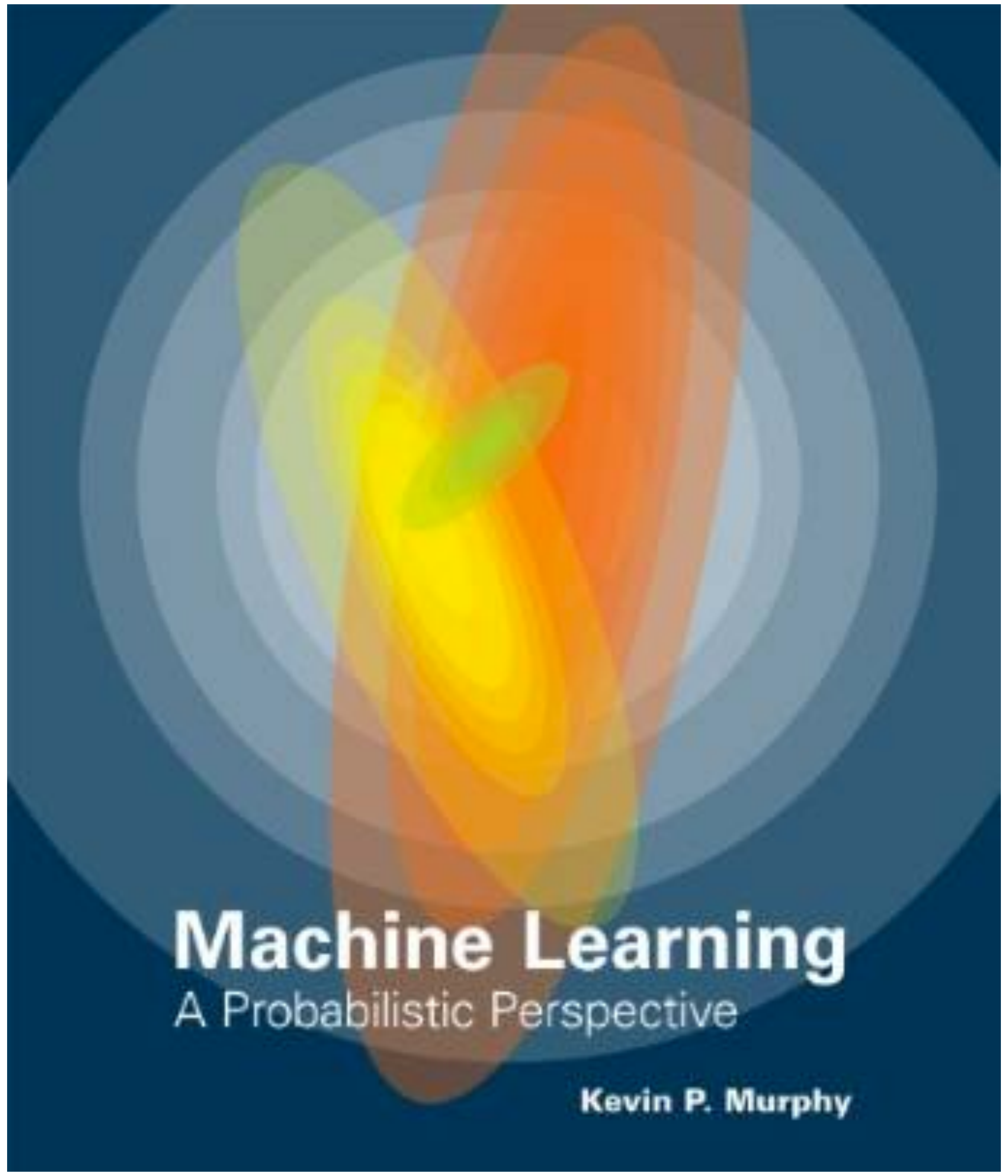


Bishop,
Pattern recognition and Machine Learning,
Springer. 2006

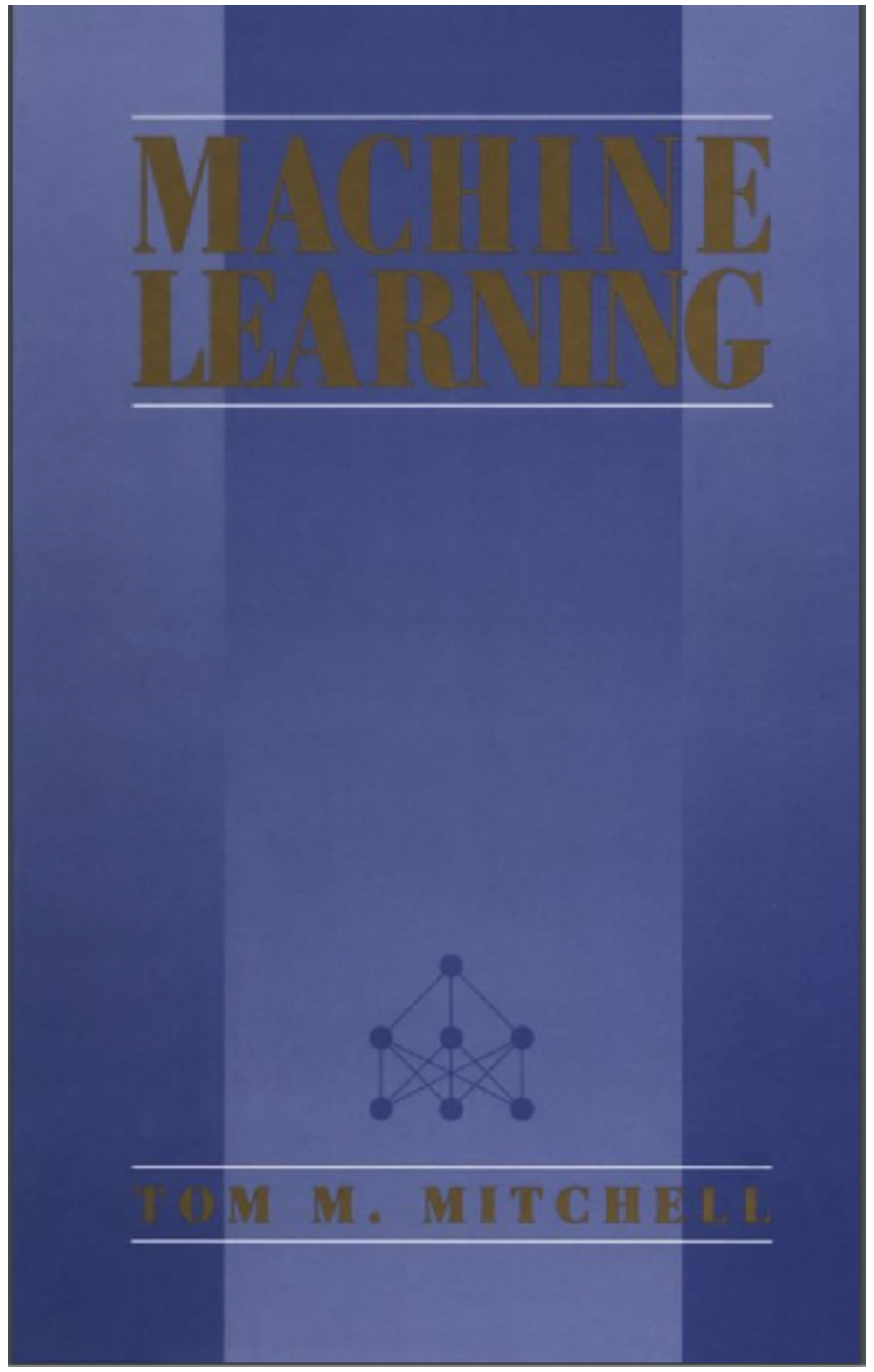
Reference Books



Shalev-Shwartz,and Ben-David, *Understanding Machine Learning: From Theory to Algorithms,* Cambridge University Press 2014.

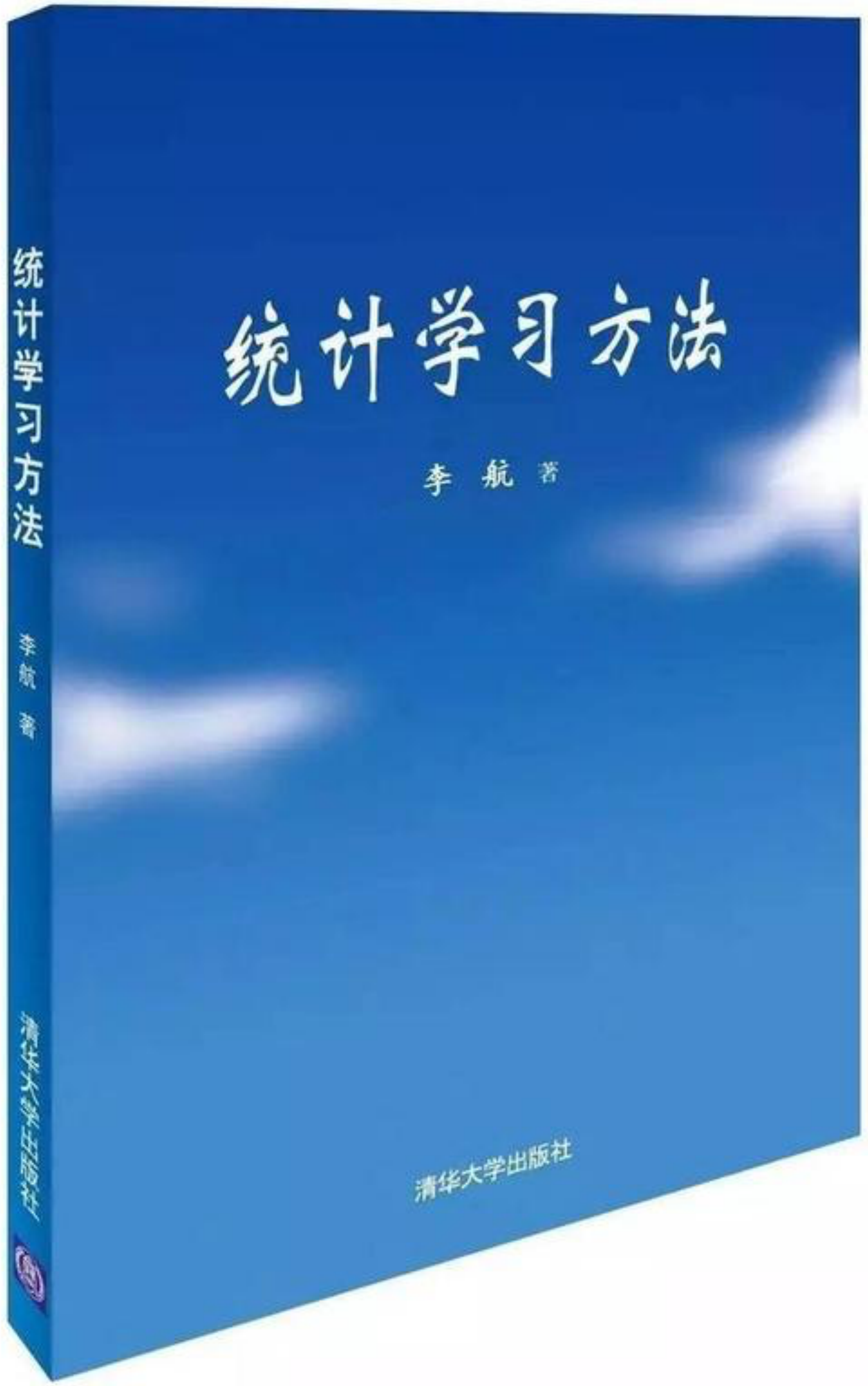


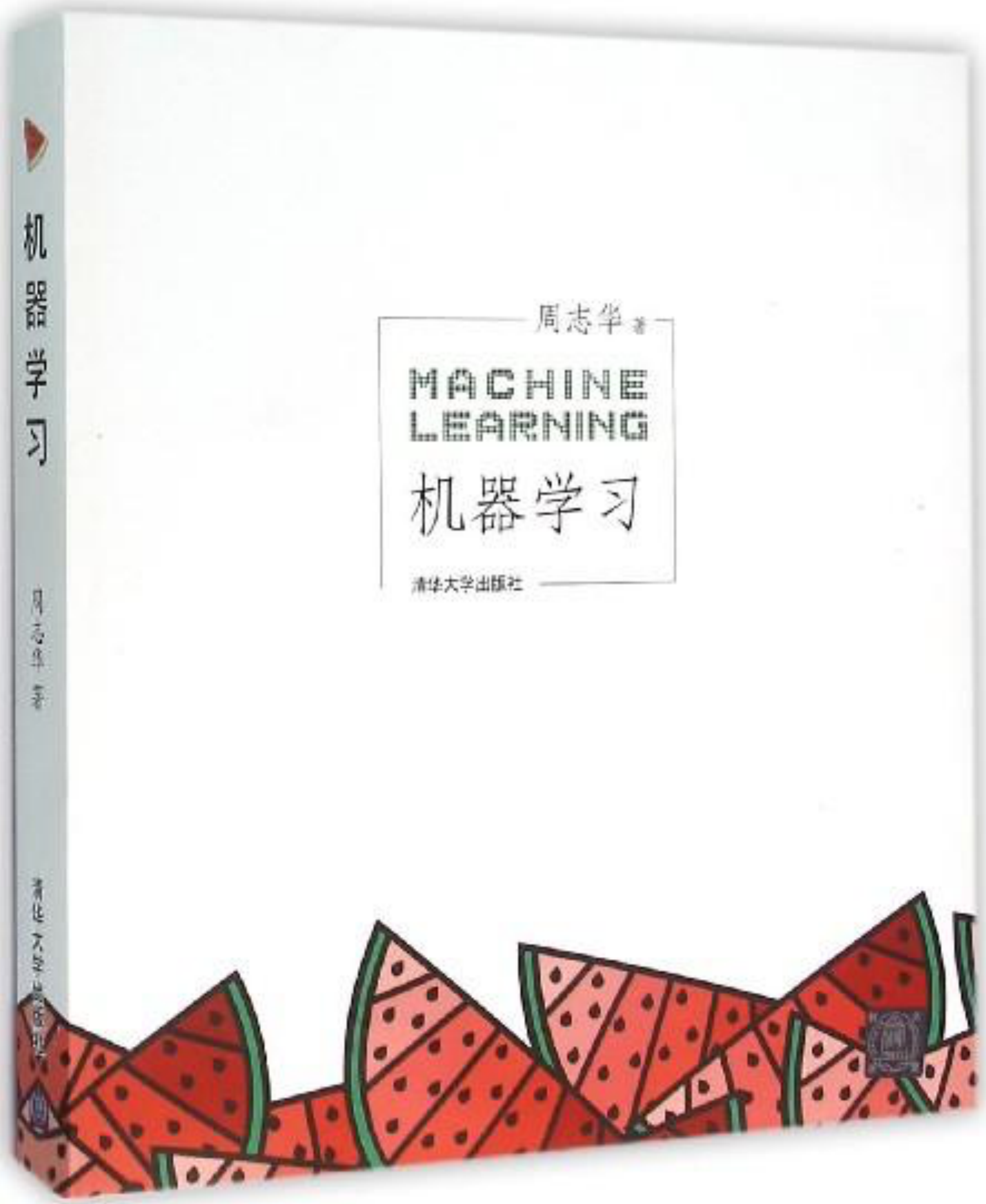
Murphy, *Machine Learning: A Probabilistic Perspective,* The MIT Press 2014.



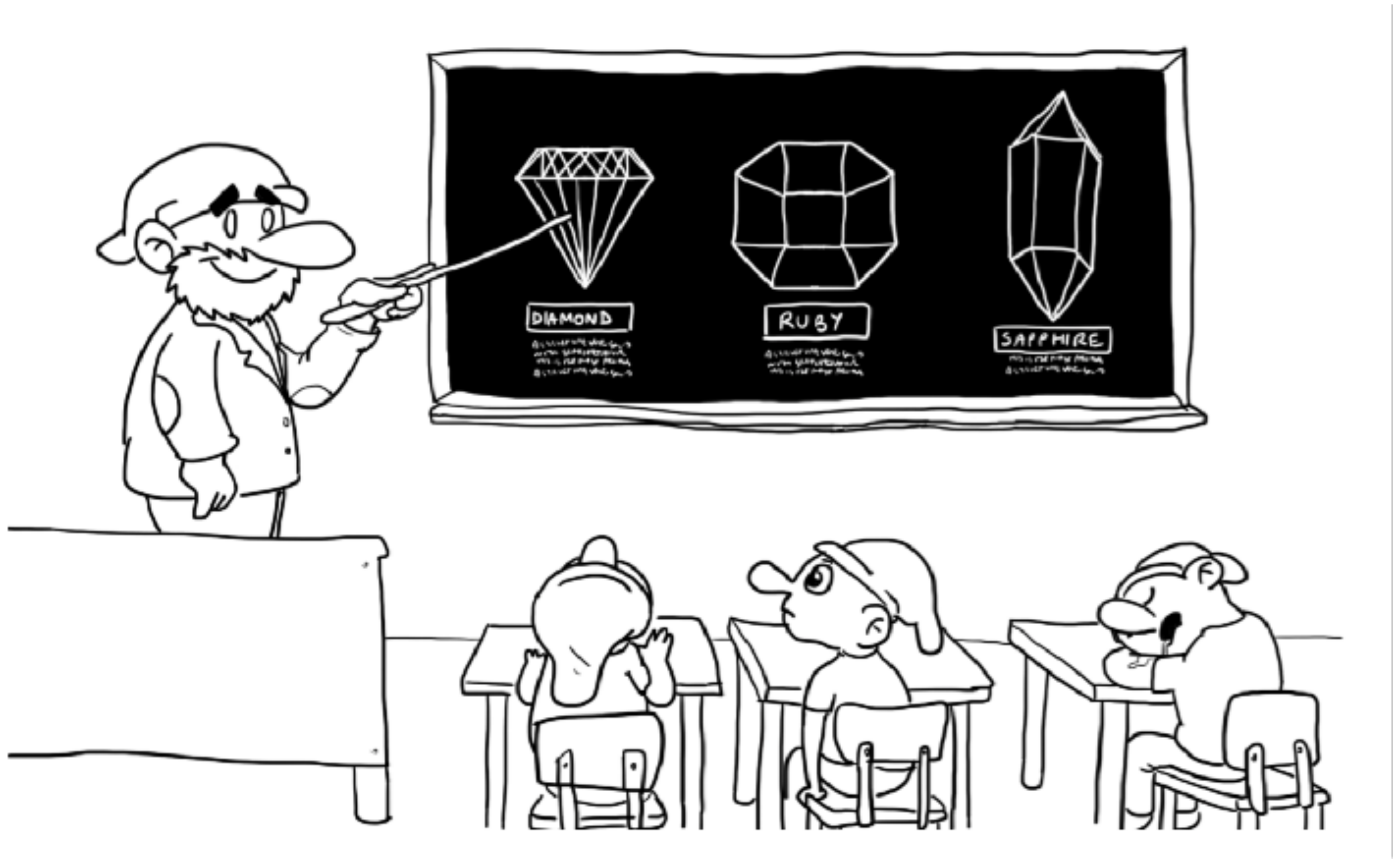
Tom M. Mitchell, *Machine Learning, McGraw Hill,* 1997.

Reference Books in Chinese



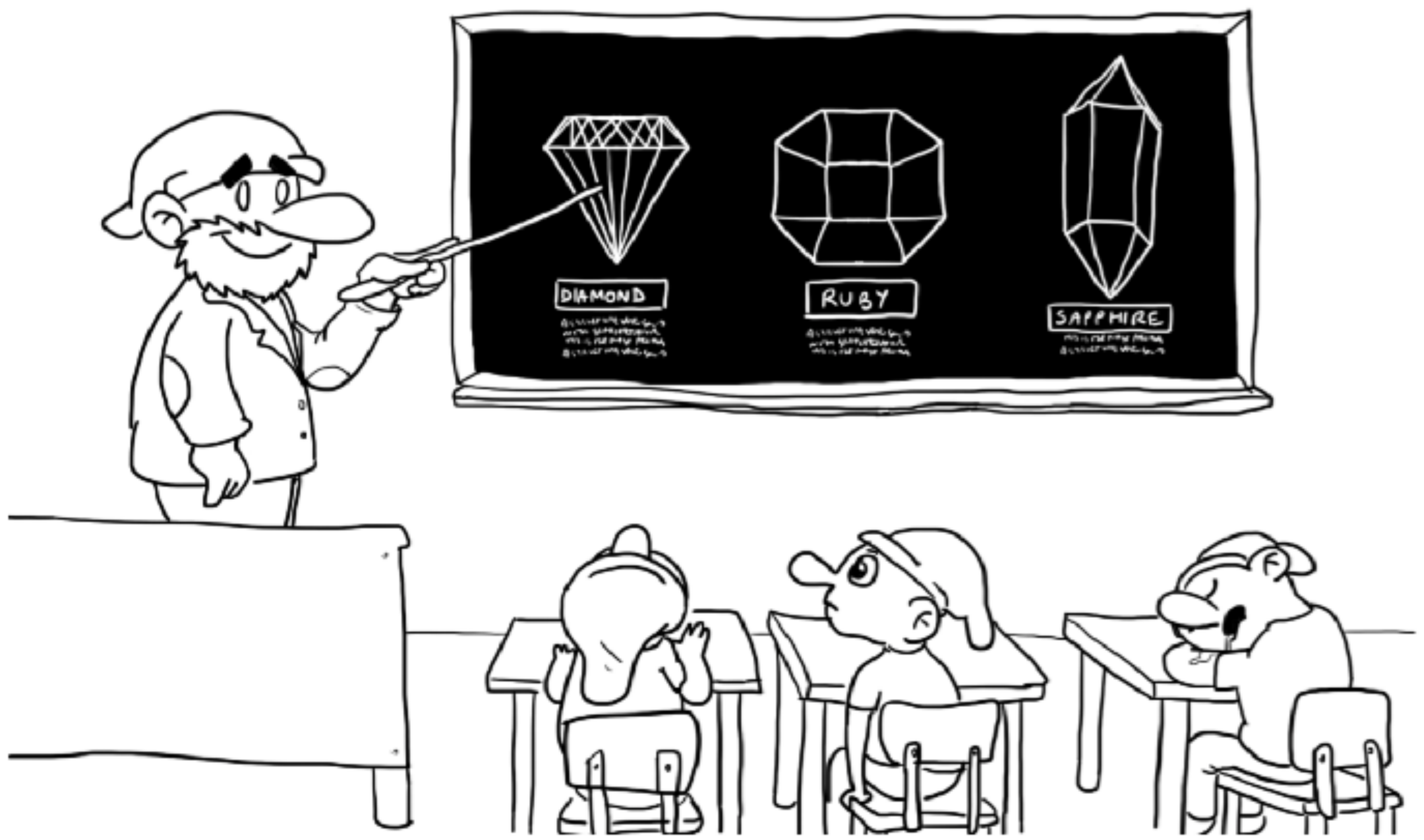


Course Requirement



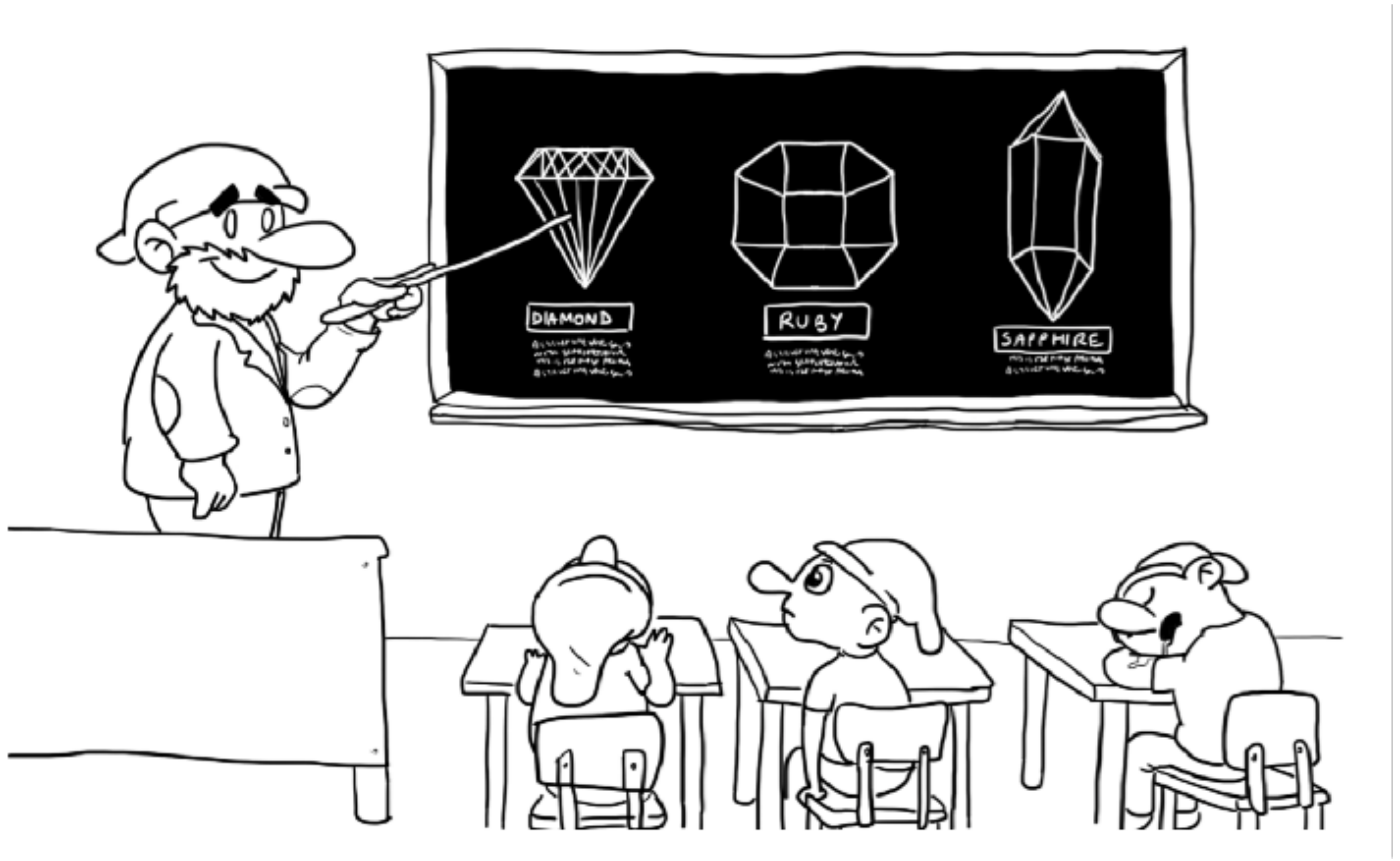
Course Requirement

- Prerequisites on Math:
  - Basic linear Algebra/Calculus: vectors, matrices, eigenvalues;
  - Probability: conditional probability, expectations;
  - Multivariate calculus: gradients, optima;



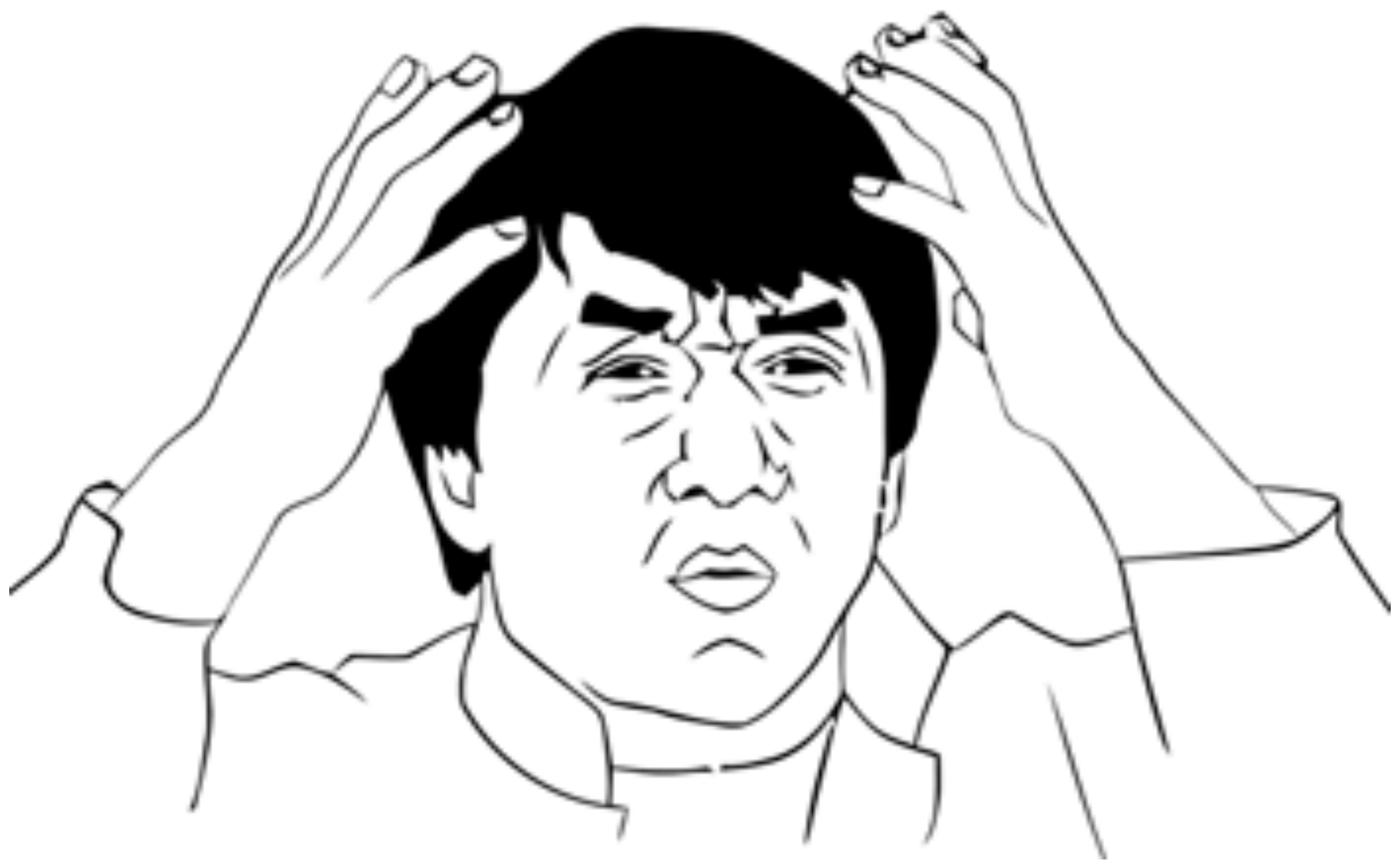
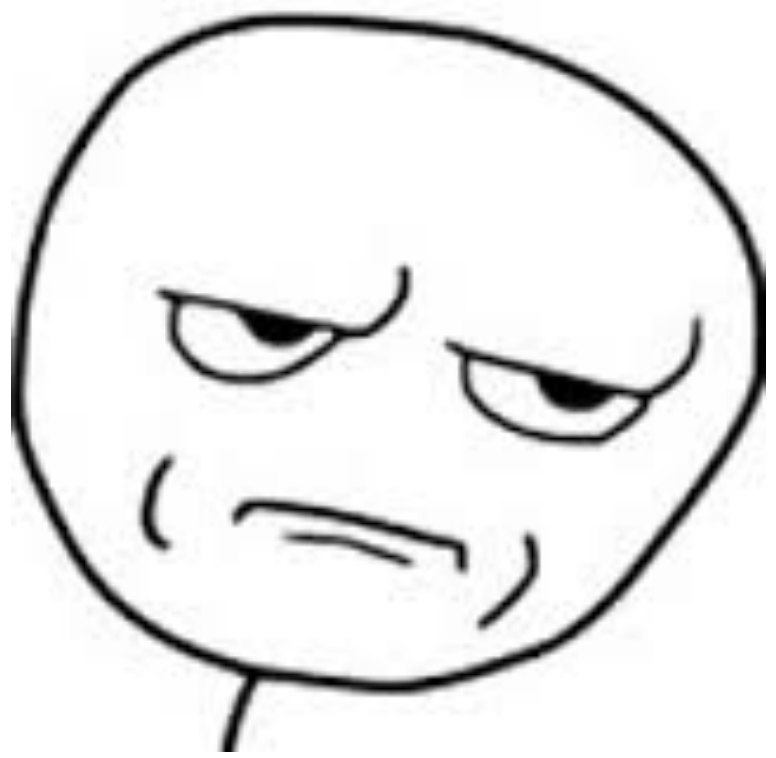
Course Requirement

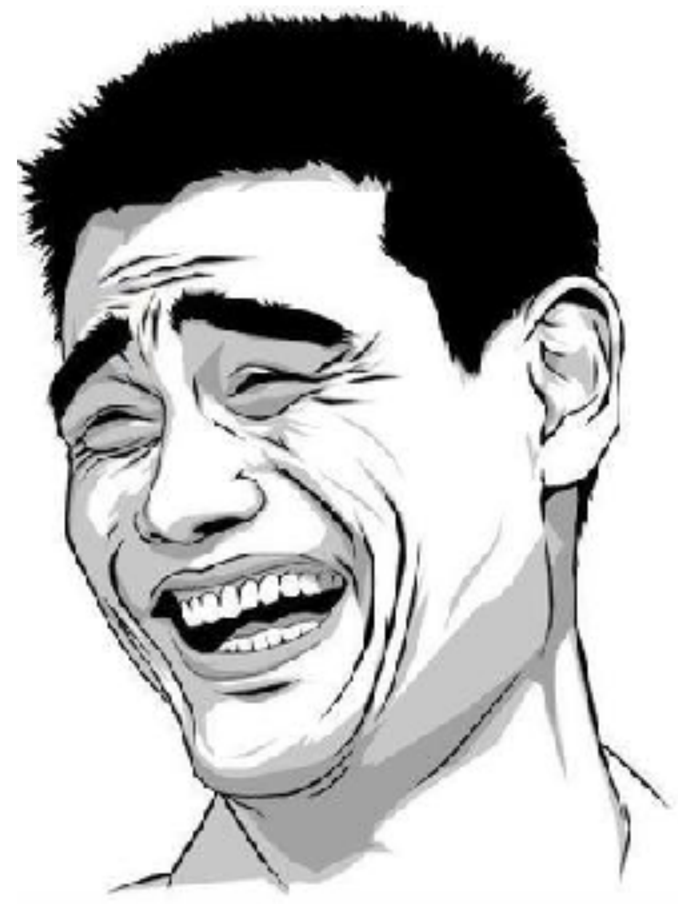
- Prerequisites on Math:
  - Basic linear Algebra/Calculus: vectors, matrices, eigenvalues;
  - Probability: conditional probability, expectations;
  - Multivariate calculus: gradients, optima;
- Prerequisites on programming:
  - Data structures: pointers, trees, heaps, hash maps, graphs;
  - Scientific computing: matrix factorisation .



Course Work

- Final scores=
  - +Class attendance/discussion (10%);
  - +Weekly homework (20%): 8-10;
  - +Monthly Mini-projects (50%); 3-4;
  - +Final Project (20%): (No final exam).
- Each Course:
  - the first 2/3 time for lectures;
  - the rest 1/3 for lab/discussion.
- Pain and Happiness
  - Huge efforts to code, debug, read and think;
  - Worth doing it!! A fundamental ingredient in the training of a modern data scientist.



Syllabus

	Topic	Slides	Exec &Notes
1	Overview	Introduction	ex1 Notes
2	Linear Regression(1)	Simple Linear Regression	ex2
3	Linear Regression(2)	Multiple&Ridge	
4	Linear Classification	LC	
5	Project-1	Project_1.zip	deadline: Oct-9 (5:00pm), send to cliao15@fudan.edu.cn
6	SVM and Kernel Methods(1)	SVM(1)	ex3: Chap4(Page170) 6, 7; Chap 9 (Page 368) 1, 2
7	SVM and Kernel Methods (2)	SVM(2)	ex4
8	Neural Network	Neural_network	http://www.robots.ox.ac.uk/~vgg/practicals/cnn/ http://cs231n.github.io/convolutional-networks/
9	Project-2	Project2.zip	other mirror: https://pan.baidu.com/s/1cyzfHk
10	PAC	PAC	Andrew Ng's Note
11	Mid-term Review	Midterm	ex5
12	unsupervised learning	unsupervised	
13	tree-based	tree-based	
14	Project-3	Project3.zip	
15	semi-supervised learning	ssl	
16	Final Project	final_project	
17	Graphical model	graphical_model	

http://yanweifu.github.io/courses/SLML/SLML.html

Academic Integrity（学术诚信)

this slide is referring to：Eton: http://114.212.80.3/wiki/index.php/%E9%9A%8F%E6%9C%BA%E7%AE%97%E6%B3%95_(Fall_2015)

Academic Integrity（学术诚信)

- **Academic integrity** is the moral code or ethical policy of academia. This includes values such as avoidance of cheating or plagiarism; maintenance of academic standards; honesty and rigor in research and academic publishing. (https://en.wikipedia.org/wiki/Academic_integrity)

this slide is referring to：Eton: http://114.212.80.3/wiki/index.php/%E9%9A%8F%E6%9C%BA%E7%AE%97%E6%B3%95_(Fall_2015)

Academic Integrity（学术诚信)

- **Academic integrity** is the moral code or ethical policy of academia. This includes values such as avoidance of cheating or plagiarism; maintenance of academic standards; honesty and rigor in research and academic publishing. (https://en.wikipedia.org/wiki/Academic_integrity)
- No cheating and plagiarism,
  - How to define *Plagiarism*? We follow ACM Policy on Plagiarism.
  - 抄袭和被抄袭双方的成绩都将被取消.
  - 作业、报告、期末论文的署名原则：署你名字的工作必须由自己完成；允许讨论，但作业必须独立完成，并在作业中列出所有参与讨论的人。不允许其他任何形式的合作——尤其是与已经完成作业的同学“讨论”。
  - 这是学术底线。

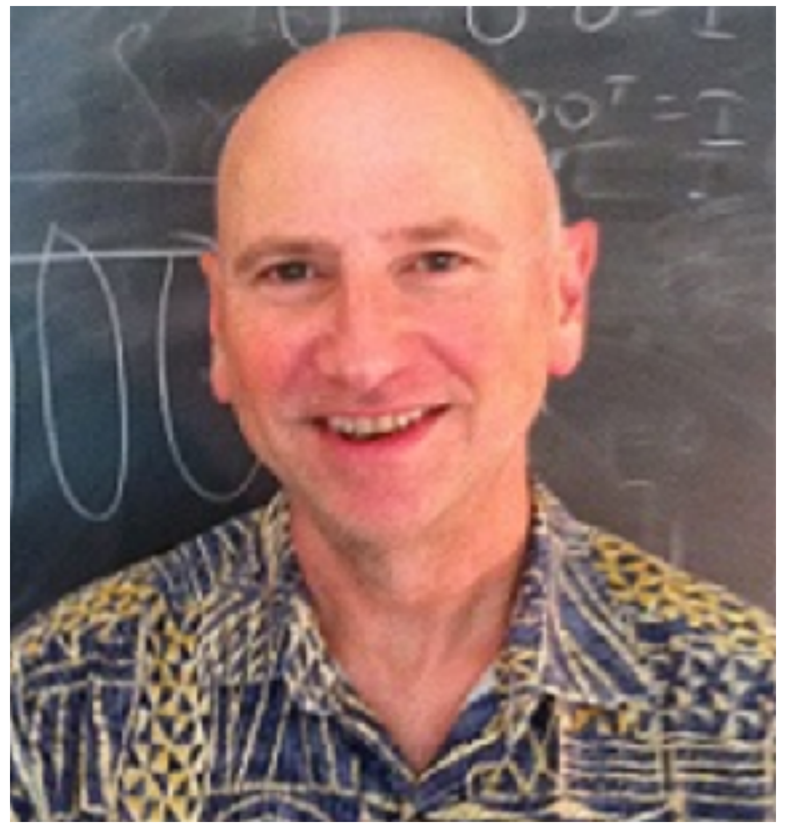
this slide is referring to：Eton: http://114.212.80.3/wiki/index.php/%E9%9A%8F%E6%9C%BA%E7%AE%97%E6%B3%95_(Fall_2015)

Chap1-Introduction

- Overview of statistical learning&machine learning;
- Various applications.
- Tutorial: Recap of R,vector calculus/algebra;

Statistics Vs. Machine Learning

Some ideas from Larry A. Wasserman (Statistician&Machine learning, Prof. in CMU)



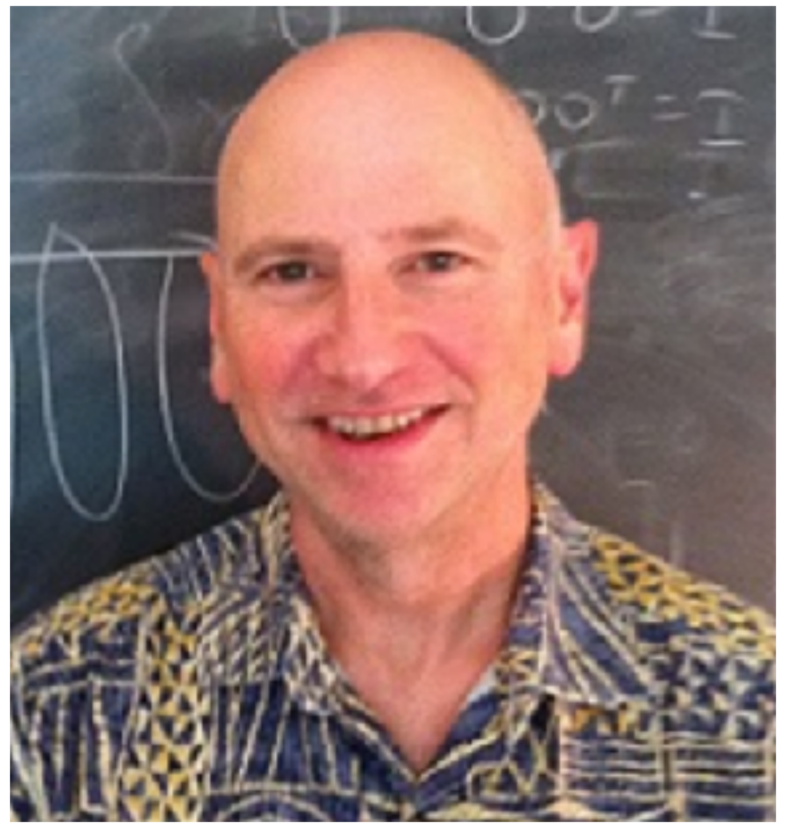
	Statistics	Machine Learning
Age	very old	young
flagship journal	The Annals of Statistics	The Journal of Machine Learning Research
interested topics	survival analysis, spatial analysis, multiple testing, minimax theory, deconvolution, semiparametric inference, bootstrapping, time series.	online learning, semisupervised learning, manifold learning, active learning, boosting

Credits: https://normaldeviate.wordpress.com/2012/06/12/statistics-versus-machine-learning-5-2/

Statistics Vs. Machine Learning

Some ideas from Larry A. Wasserman (Statistician&Machine learning, Prof. in CMU)

1. What is the difference between these two fields?
    - short answer: No!
    - both concerned the same question: how do we learn from data?



	Statistics	Machine Learning
Age	very old	young
flagship journal	The Annals of Statistics	The Journal of Machine Learning Research
interested topics	survival analysis, spatial analysis, multiple testing, minimax theory, deconvolution, semiparametric inference, bootstrapping, time series.	online learning, semisupervised learning, manifold learning, active learning, boosting

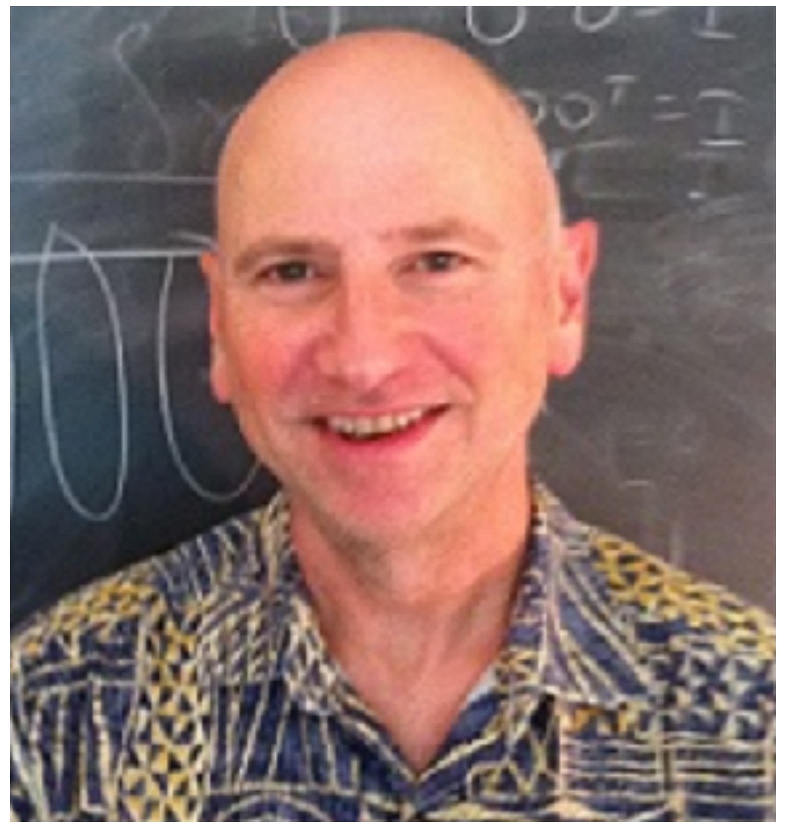
Credits: https://normaldeviate.wordpress.com/2012/06/12/statistics-versus-machine-learning-5-2/

Statistics Vs. Machine Learning

Some ideas from Larry A. Wasserman (Statistician&Machine learning, Prof. in CMU)

1. What is the difference between these two fields?
   - short answer: No!
   - both concerned the same question: how do we learn from data?
2. “Overall, the two fields are blending together more and more and I think this is a good thing.”

   — Larry Wasserman



	Statistics	Machine Learning
Age	very old	young
flagship journal	The Annals of Statistics	The Journal of Machine Learning Research
interested topics	survival analysis, spatial analysis, multiple testing, minimax theory, deconvolution, semiparametric inference, bootstrapping, time series.	online learning, semisupervised learning, manifold learning, active learning, boosting

Credits: https://normaldeviate.wordpress.com/2012/06/12/statistics-versus-machine-learning-5-2/

Statistics&Machine Learning

Glossary

Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant = $1,000,000	large grant = $50,000
nice place to have a meeting: Snowbird, Utah, French Alps	nice place to have a meeting: Las Vegas in August

A "joke" by Prof. Robert Tibshiriani. He is both statistician and machine learning expert.

credit: Mark Schmidt

Statistics&Machine Learning

- **Machine learning** is defined as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty. (Murphy, 2014);

Glossary

Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant = $1,000,000	large grant = $50,000
nice place to have a meeting: Snowbird, Utah, French Alps	nice place to have a meeting: Las Vegas in August

A “joke” by Prof. Robert Tibshiriani. He is both statistician and machine learning expert.

credit: Mark Schmidt

Statistics&Machine Learning

- **Machine learning** is defined as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty. (Murphy, 2014);
- Machine learning (ML) is very similar to statistics. But ML places more emphasis on:
    1. Computation and large datasets.
    2. Predictions rather than descriptions.
    3. Non-asymptotic performance.
    4. Models that work across domains.

Glossary

Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant = $1,000,000	large grant = $50,000
nice place to have a meeting: Snowbird, Utah, French Alps	nice place to have a meeting: Las Vegas in August

A "joke" by Prof. Robert Tibshiriani. He is both statistician and machine learning expert.

credit: Mark Schmidt

Machine Learning Vs. Data Mining

Boundary lines are blurred: many ML problems involve tons of data — Big Data.

But in general,

Data-mining: Typically using very simple machine learning techniques on very large databases because computers are too slow to do anything more interesting with ten billion examples.

But problems with AI favor (e.g., recognition, robot navigation) still domain of ML.

Statistical Learning Vs. Machine Learning

Statistical Learning Vs. Machine Learning

- **Statistical learning** refers to a set of tools for modeling and understanding complex datasets (Hastie, 2011);

Statistical Learning Vs. Machine Learning

- **Statistical learning** refers to a set of tools for modeling and understanding complex datasets (Hastie, 2011);
- **Machine learning** is defined as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty. (Murphy, 2014);
    - *Statistical learning theory* is a framework for machine learning drawing from the fields of statistics and functional analysis. Statistical learning theory deals with the problem of finding a predictive function based on data. Statistical learning theory has led to successful applications in fields such as computer vision, speech recognition, bioinformatics and baseball. (https://en.wikipedia.org/wiki/Statistical_learning_theory)

Statistical Learning Vs. Machine Learning

- **Statistical learning** refers to a set of tools for modeling and understanding complex datasets (Hastie, 2011);
- **Machine learning** is defined as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty. (Murphy, 2014);
  - *Statistical learning theory* is a framework for machine learning drawing from the fields of statistics and functional analysis. Statistical learning theory deals with the problem of finding a predictive function based on data. Statistical learning theory has led to successful applications in fields such as computer vision, speech recognition, bioinformatics and baseball. (https://en.wikipedia.org/wiki/Statistical_learning_theory)
- Machine learning arose as a subfield of *Artificial Intelligence*.

Statistical Learning Vs. Machine Learning

- **Statistical learning** refers to a set of tools for modeling and understanding complex datasets (Hastie, 2011);
- **Machine learning** is defined as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty. (Murphy, 2014);
  - *Statistical learning theory* is a framework for machine learning drawing from the fields of statistics and functional analysis. Statistical learning theory deals with the problem of finding a predictive function based on data. Statistical learning theory has led to successful applications in fields such as computer vision, speech recognition, bioinformatics and baseball. (https://en.wikipedia.org/wiki/Statistical_learning_theory)
- Machine learning arose as a subfield of *Artificial Intelligence*.
- Statistical learning arose as a subfield of *Statistics*.

Statistical Learning Vs. Machine Learning

- **Statistical learning** refers to a set of tools for modeling and understanding complex datasets (Hastie, 2011);
- **Machine learning** is defined as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty. (Murphy, 2014);
  - *Statistical learning theory* is a framework for machine learning drawing from the fields of statistics and functional analysis. Statistical learning theory deals with the problem of finding a predictive function based on data. Statistical learning theory has led to successful applications in fields such as computer vision, speech recognition, bioinformatics and baseball. (https://en.wikipedia.org/wiki/Statistical_learning_theory)
- Machine learning arose as a subfield of *Artificial Intelligence*.
- Statistical learning arose as a subfield of *Statistics*.
- There is much overlap | both fields focus on supervised and unsupervised problems:
  - Machine learning has a greater emphasis on large scale applications and prediction accuracy.
  - Statistical learning emphasizes models and their interpretability, and precision and uncertainty.

Machine Learning & AI

人工智能研究的主要方法：

1，**符号主义方法**：认知是一种符号处理过程，人类思维过程可用符号来描述，思维就是计算，这种思想一度构成了人工智能的基础理论。

代表人物：司马贺（西蒙，Herbert Alexander Simon）和纽厄尔（Allen Newell），物理符号系统，1975 年图灵奖获得者。

2，**联结主义方法**：模拟人的智能要依靠仿生学，特别是需要模拟人脑，建立脑模型。人类思维的基本单元是神经元，而不是符号，智能是相互联结的神经元竞争与协作结果。

代表人物：麦卡洛克（Warren McCulloch), 皮茨（Walter Pitts)提出的神经元的数理模型。

3，**行为主义方法**：模拟人在控制过程中的智能行为和作用，研制所谓的控制论动物。

代表人物：博德（H.W.Bode) 和埃文斯（W.R.Evans)等。

Resources — Conferences

Resources — Conferences

Machine Learning:

- Neural Information Processing Systems (NIPS)
- International Conference on Machine Learning (ICML)
- European Conference on Machine Learning (ECML)
- Uncertainty in Artificial Intelligence (UAI)
- Computational Learning Theory (COLT)
- International Conference on AI & Statistics (AISTATS)

Resources — Conferences

Machine Learning:

- Neural Information Processing Systems (NIPS)
- International Conference on Machine Learning (ICML)
- European Conference on Machine Learning (ECML)
- Uncertainty in Artificial Intelligence (UAI)
- Computational Learning Theory (COLT)
- International Conference on AI & Statistics (AISTATS)

AI In general

- AAAI Conference on Artificial Intelligence (AAAI) (AAAI: Association for the Advancement of Artificial Intelligence);
- International Joint Conference on Artificial Intelligence (IJCAI);

Resources—Conferences

Machine Learning:

- Neural Information Processing Systems (NIPS)
- International Conference on Machine Learning (ICML)
- European Conference on Machine Learning (ECML)
- Uncertainty in Artificial Intelligence (UAI)
- Computational Learning Theory (COLT)
- International Conference on AI & Statistics (AISTATS)

AI In general

- AAAI Conference on Artificial Intelligence (AAAI) (AAAI: Association for the Advancement of Artificial Intelligence);
- International Joint Conference on Artificial Intelligence (IJCAI);

Pattern Recognition&Computer Vision:

- European Conference on Computer Vision (ECCV)
- IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
- IEEE International Conference on Computer Vision (ICCV)
- ICPR/ACCV/BMVC;

Resources — Conferences

Machine Learning:

- Neural Information Processing Systems (NIPS)
- International Conference on Machine Learning (ICML)
- European Conference on Machine Learning (ECML)
- Uncertainty in Artificial Intelligence (UAI)
- Computational Learning Theory (COLT)
- International Conference on AI & Statistics (AISTATS)

AI In general

- AAAI Conference on Artificial Intelligence (AAAI) (AAAI: Association for the Advancement of Artificial Intelligence);
- International Joint Conference on Artificial Intelligence (IJCAI);

Pattern Recognition&Computer Vision:

- European Conference on Computer Vision (ECCV)
- IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
- IEEE International Conference on Computer Vision (ICCV)
- ICPR/ACCV/BMVC;

- Data Mining:
  - ACM SIGKDD (Knowledge discovery and Data Mining);
  - ACM SIGIR/ICDM;

Resources—Conferences

Machine Learning:

- Neural Information Processing Systems (NIPS)
- International Conference on Machine Learning (ICML)
- European Conference on Machine Learning (ECML)
- Uncertainty in Artificial Intelligence (UAI)
- Computational Learning Theory (COLT)
- International Conference on AI & Statistics (AISTATS)

AI In general

- AAAI Conference on Artificial Intelligence (AAAI) (AAAI: Association for the Advancement of Artificial Intelligence);
- International Joint Conference on Artificial Intelligence (IJCAI);

Pattern Recognition&Computer Vision:

- European Conference on Computer Vision (ECCV)
- IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
- IEEE International Conference on Computer Vision (ICCV)
- ICPR/ACCV/BMVC;

- Data Mining:
  - ACM SIGKDD (Knowledge discovery and Data Mining);
  - ACM SIGIR/ICDM;

Natural Language Processing:

- ACL/EMNLP/COLING

Resources—Conferences

Machine Learning:

- Neural Information Processing Systems (NIPS)
- International Conference on Machine Learning (ICML)
- European Conference on Machine Learning (ECML)
- Uncertainty in Artificial Intelligence (UAI)
- Computational Learning Theory (COLT)
- International Conference on AI & Statistics (AISTATS)

AI In general

- AAAI Conference on Artificial Intelligence (AAAI) (AAAI: Association for the Advancement of Artificial Intelligence);
- International Joint Conference on Artificial Intelligence (IJCAI);

Pattern Recognition&Computer Vision:

- European Conference on Computer Vision (ECCV)
- IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
- IEEE International Conference on Computer Vision (ICCV)
- ICPR/ACCV/BMVC;

- Data Mining:
  - ACM SIGKDD (Knowledge discovery and Data Mining);
  - ACM SIGIR/ICDM;

Natural Language Processing:

- ACL/EMNLP/COLING

Resources — Conferences

Machine Learning:

- Neural Information Processing Systems (NIPS)
- International Conference on Machine Learning (ICML)
- European Conference on Machine Learning (ECML)
- Uncertainty in Artificial Intelligence (UAI)
- Computational Learning Theory (COLT)
- International Conference on AI & Statistics (AISTATS)

AI In general

- AAAI Conference on Artificial Intelligence (AAAI) (AAAI: Association for the Advancement of Artificial Intelligence);
- International Joint Conference on Artificial Intelligence (IJCAI);

Pattern Recognition&Computer Vision:

- European Conference on Computer Vision (ECCV)
- IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
- IEEE International Conference on Computer Vision (ICCV)
- ICPR/ACCV/BMVC;

- Data Mining:
  - ACM SIGKDD (Knowledge discovery and Data Mining);
  - ACM SIGIR/ICDM;

Natural Language Processing:

- ACL/EMNLP/COLING

Resources—Journals

Resources—Journals

Machine Learning&AI:

- Journal of Machine Learning Research (JMLR)
- IEEE Trans on Pattern Analysis and Machine Intelligence (TPAMI);
- Artificial Intelligence;
- International Journal of Computer Vision (IJCV);

Statistics:

- The Annals of Statistics;

Resources—Journals

Machine Learning&AI:

- Journal of Machine Learning Research (JMLR)
- IEEE Trans on Pattern Analysis and Machine Intelligence (TPAMI);
- Artificial Intelligence;
- International Journal of Computer Vision (IJCV);

Statistics:

- The Annals of Statistics;

Pattern Recognition:

- Neural Computation
- Neural Networks;

Resources—Journals

Machine Learning&AI:

- Journal of Machine Learning Research (JMLR)
- IEEE Trans on Pattern Analysis and Machine Intelligence (TPAMI);
- Artificial Intelligence;
- International Journal of Computer Vision (IJCV);

Statistics:

- The Annals of Statistics;

Pattern Recognition:

- Neural Computation
- Neural Networks;

Data Mining:

- IEEE Transactions on Knowledge and Data Engineering (TKDE)

What is Machine Learning?

What is Machine Learning?

- **Definition of ML (Mitchell, 1997): WELL-POSED LEARNING PROBLEMS.**
  - A computer program is said to learn from experience ***E*** with respect to some class of tasks ***T*** and performance measure ***P***, if its performance at tasks in ***T***, as measured by ***P***, improves with experience ***E***.

What is Machine Learning?

- **Definition of ML (Mitchell, 1997): WELL-POSED LEARNING PROBLEMS.**
  - A computer program is said to learn from experience ***E*** with respect to some class of tasks ***T*** and performance measure ***P***, if its performance at tasks in ***T***, as measured by ***P***, improves with experience ***E***.
- **Example: A computer program that learns to play checkers**
  - **T**ask: playing checkers games;
  - **E**xperience: obtained by playing games against itself;
  - **P**erformance Measure: percent of games won against opponents

A handwriting recognition learning problem:

- Task ***T***: recognizing and classifying handwritten words within images
- Performance measure ***P***: percent of words correctly classified
- Training experience ***E***: a database of handwritten words with given classifications

A handwriting recognition learning problem:

- Task ***T***: recognizing and classifying handwritten words within images
- Performance measure ***P***: percent of words correctly classified
- Training experience ***E***: a database of handwritten words with given classifications

A robot driving learning problem: an example from (Mitchell, 1997)

- Task ***T:*** driving on public four-lane highways using vision sensors;
- Performance measure ***P***: average distance traveled before an error (as judged by human overseer)
- Training experience ***E***: a sequence of images and steering commands recorded while observing a human driver;

A handwriting recognition learning problem:

- Task ***T***: recognizing and classifying handwritten words within images
- Performance measure ***P***: percent of words correctly classified
- Training experience ***E***: a database of handwritten words with given classifications

A robot driving learning problem: an example from (Mitchell, 1997)

- Task ***T:*** driving on public four-lane highways using vision sensors;
- Performance measure ***P***: average distance traveled before an error (as judged by human overseer)
- Training experience ***E***: a sequence of images and steering commands recorded while observing a human driver;

Example: Spam classification

- Task ***T***: determine if emails are Spam or non-Spam.
- Experience ***E:*** Incoming emails with human classification
- Performance Measure ***P***: percentage of correct decisions

Notations, formally

Task:

$\mathcal{X}$ input variables (from input set), a.k.a., features, predictors, independent variables.

$\mathcal{Y}$ output variables (from output set), a.k.a., response or dependent variable.

$f : \mathcal{X} \to \mathcal{Y}$ Prediction function,

Performance:

$l : \mathcal{X} \to \mathcal{Y}$ Loss function,

$l\left(y, y^{'}\right)$ is the cost of predicting $y^{'}$ if y is correct.

Experience: task-dependent, many different scenarios

- Supervised Learning, Unsupervised Learning, Reinforcement Learning,
- Semi$-$supervised Learning, Multiple Instance Learning, Active Learning.

Supervised Learning

Supervised Learning

- A labeled training set examples with outputs provided by an expert,

- Regression Vs. Classification problems,
  - ***Regression***: *Y* is is **quantitative** (e.g price, blood pressure);
  - ***Classification***: *Y* takes values in a finite, unordered set (survived/died, digit 0-9, cancer class of tissue sample), **qualitative**.

Supervised Learning

- A labeled training set examples with outputs provided by an expert,

$$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\} \subset \mathcal{X} \times \mathcal{Y}$$

- Regression Vs. Classification problems,
  - ***Regression***: *Y* is is **quantitative** (e.g price, blood pressure);
  - ***Classification***: *Y* takes values in a finite, unordered set (survived/died, digit 0-9, cancer class of tissue sample), **qualitative**.

Other problems such as ranking is often formulated as either problem.

Supervised Learning

- A labeled training set examples with outputs provided by an expert,

$$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\} \subset \mathcal{X} \times \mathcal{Y}$$

- Regression Vs. Classification problems,
  - ***Regression***: *Y* is is **quantitative** (e.g price, blood pressure);
  - ***Classification***: *Y* takes values in a finite, unordered set (survived/died, digit 0-9, cancer class of tissue sample), **qualitative**.

Other problems such as ranking is often formulated as either problem.

Definition,

- A supervised learning system (or learner), *L* is a (computable) function from the set of (finite) training sets to the set of prediction functions:

$$L : \mathbb{P}^{<\infty}(\mathcal{X} \times \mathcal{Y}) \to \mathcal{Y}^{\mathcal{X}}$$
$$L : \mathcal{D} \mapsto f$$

So if presented with a training set $\mathcal{D}$, it provides a decision rule/function $f : \mathcal{X} \to \mathcal{Y}$

Let *L* be a learning system.

- Process of computing is $f = L(\mathcal{D})$ called training (phase).
- Applying f to new data is called prediction, or testing. (phase).

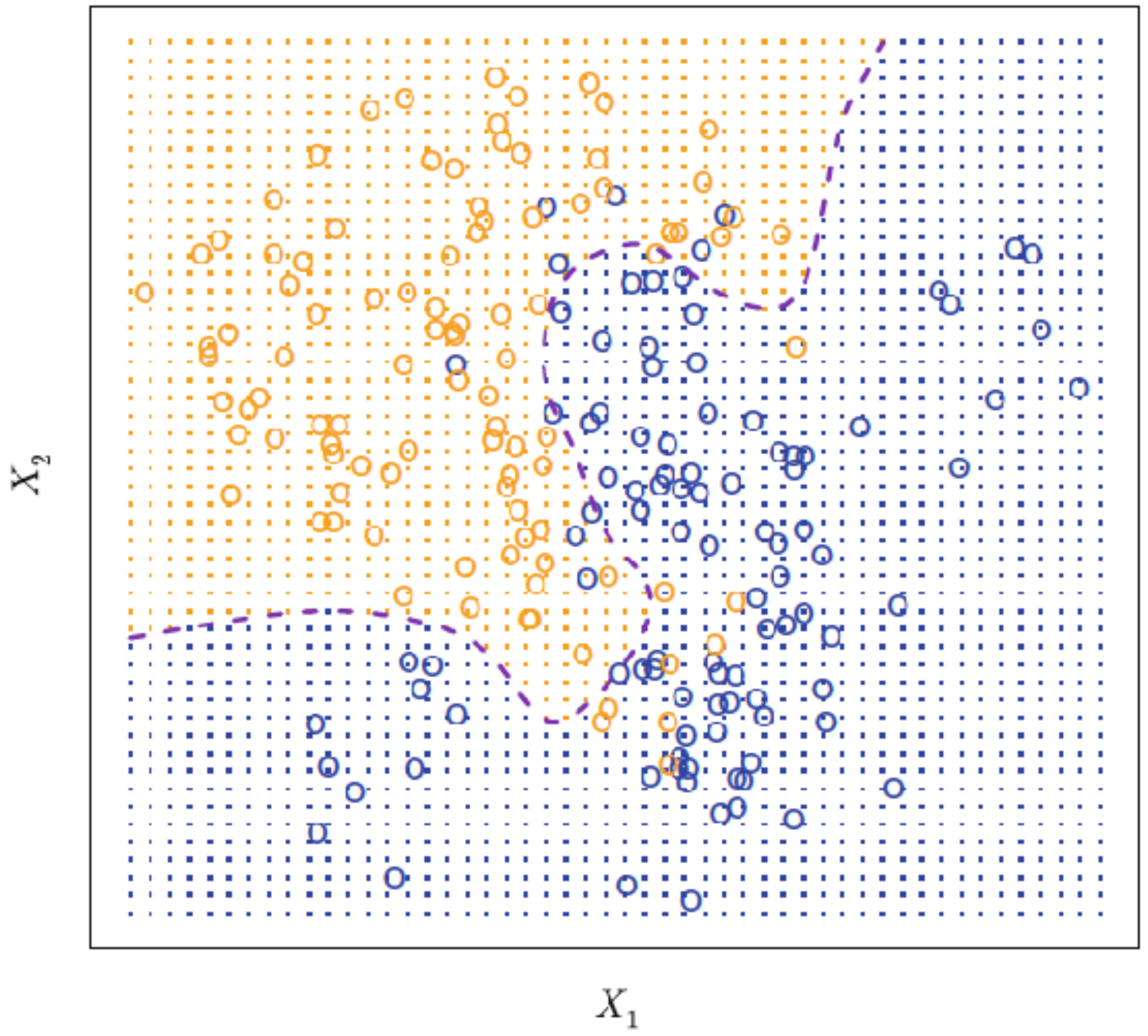
The Classification Setting

Error rate $$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i).$$

Indicator Variable; training errors; testing errors;

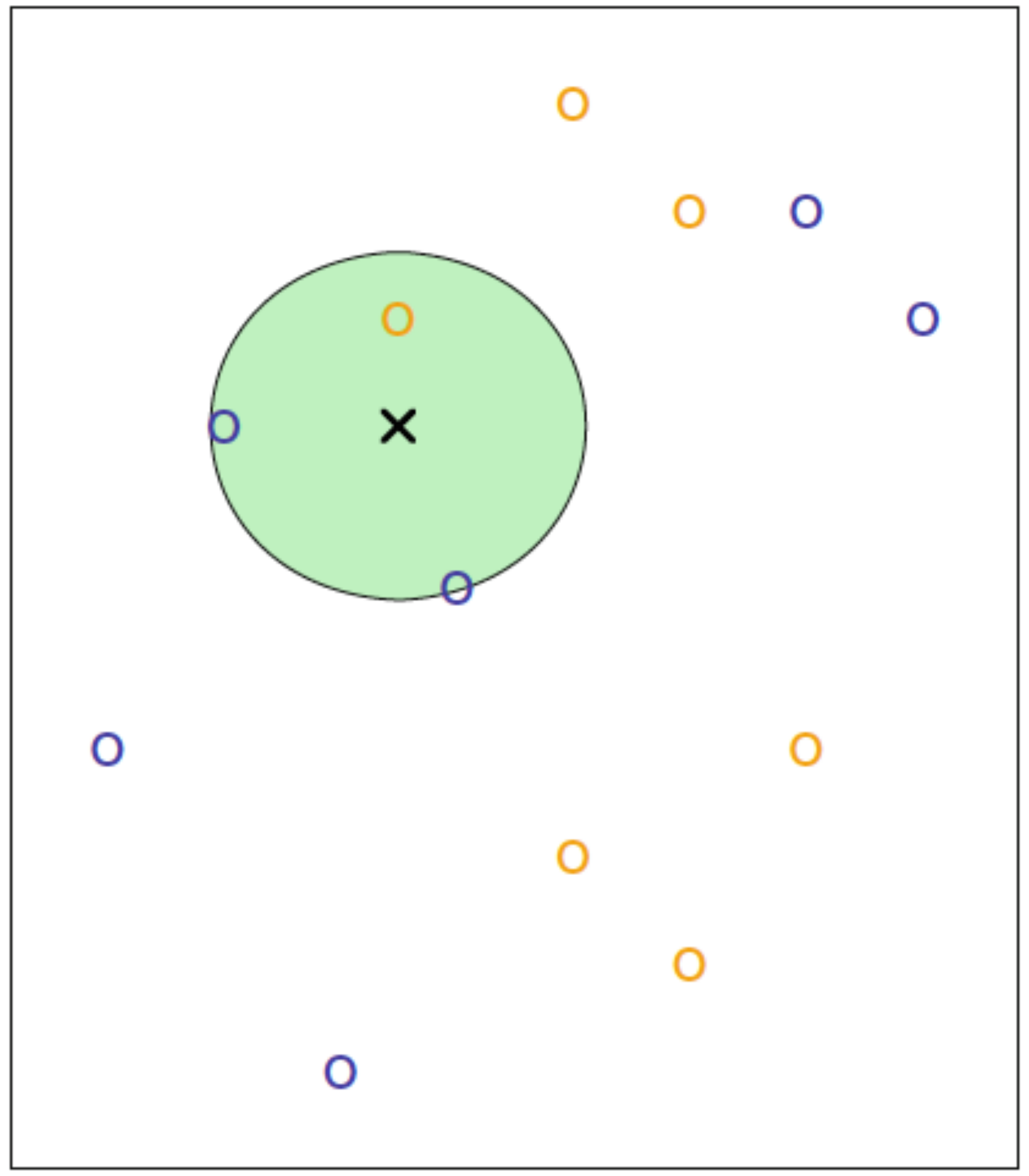
The Bayes Classifier: *assigns each observation to the most likely class, given its predictor values.*

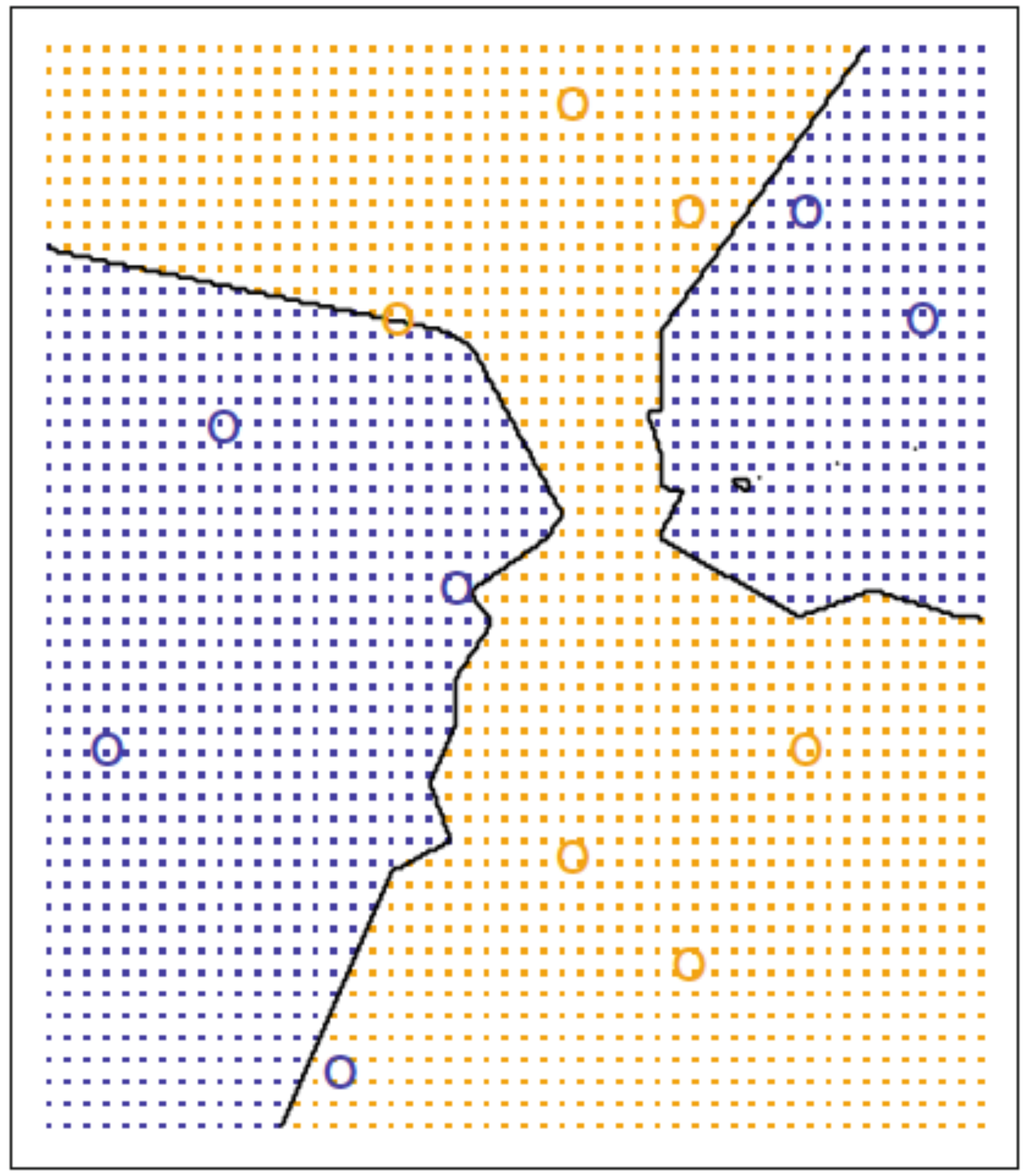
$\Pr(Y = j | X = x_0)$ *Bayes decision boundary*



K-Nearest Neighbors

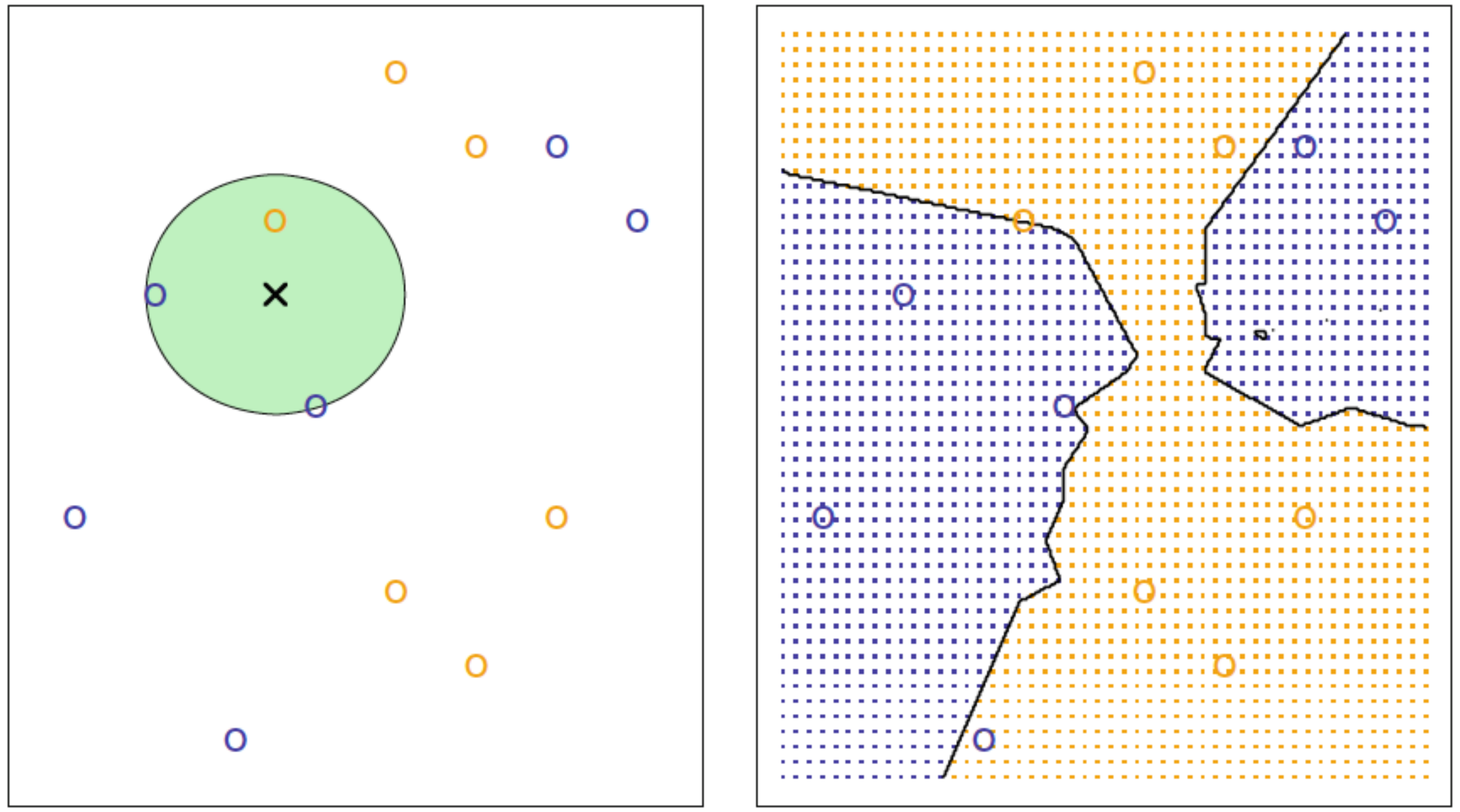
$$\Pr(Y = j | X = x_0) = \frac{1}{K}\sum_{i \in \mathcal{N}_0} I(y_i = j).$$





K-Nearest Neighbors (Non-parametric Method)

$$\Pr(Y = j | X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j).$$



k-Nearest Neighbor – Training

input dataset $\mathcal{D} = \{(x^1, y^1), \ldots, (x^n, y^n)\} \subset \mathbb{R}^d \times \mathcal{Y}$
store all examples $(x^1, y^1), \ldots, (x^n, y^n)$.

k-Nearest Neighbor – Classification

input new example x
for each training example (x^i, y^i) compute $d_i(x) = \|x - x^i\|$ (Euclidean distance)
sort d_i in increasing order
output majority vote among y^is within the k smallest d^i

"nearest neighbor" is 1-nearest neighbour.

Toy example: ***How grade will I get in this course?***

General workflow of SL.

- **Data**: entry survey and marks from previous years
- **Process the data:**
  - Split into **training set; test set;**
  - Representation of **input features**; output
- Choose form of model: **linear regression**
- Decide how to evaluate the system's performance: **objective function**
- Set model parameters to optimize performance
- Evaluate on test set: **generalization**

CSC411/CSC2515: Entry Survey

Which course are you taking?
- CSC411
- CSC2515

Name

Student Number

Major

Years Until Graduation

1 2 3 4 5

Status

Email

Familiarity with Bayes Rule
- Proficient
- Confortable
- Rusty
- Hunh?

Familiarity with Maximum A Posteriori
- Proficient
- Confortable
- Rusty
- Hunh?

Familiarity with Logistic Regression
- Proficient
- Confortable
- Rusty
- Hunh?

Familiarity with Gradient Descent
- Proficient
- Confortable
- Rusty
- Hunh?

Familiarity with Chain Rule
- Proficient
- Confortable
- Rusty
- Hunh?

Familiarity with Matlab
- Proficient
- Confortable
- Rusty
- Hunh?

Familiarity with Python
- Proficient
- Confortable
- Rusty
- Hunh?

Familiarity with Belief Networks
- Proficient
- Confortable
- Rusty
- Hunh?

Familiarity with EigenVectors
- Proficient
- Confortable
- Rusty
- Hunh?

What related courses have you taken?

e.g., CSC321, CSC384

credit: Urtasun & Zemel (University of Toronto)

Toy example: ***How grade will I get in this course?***

Settings:

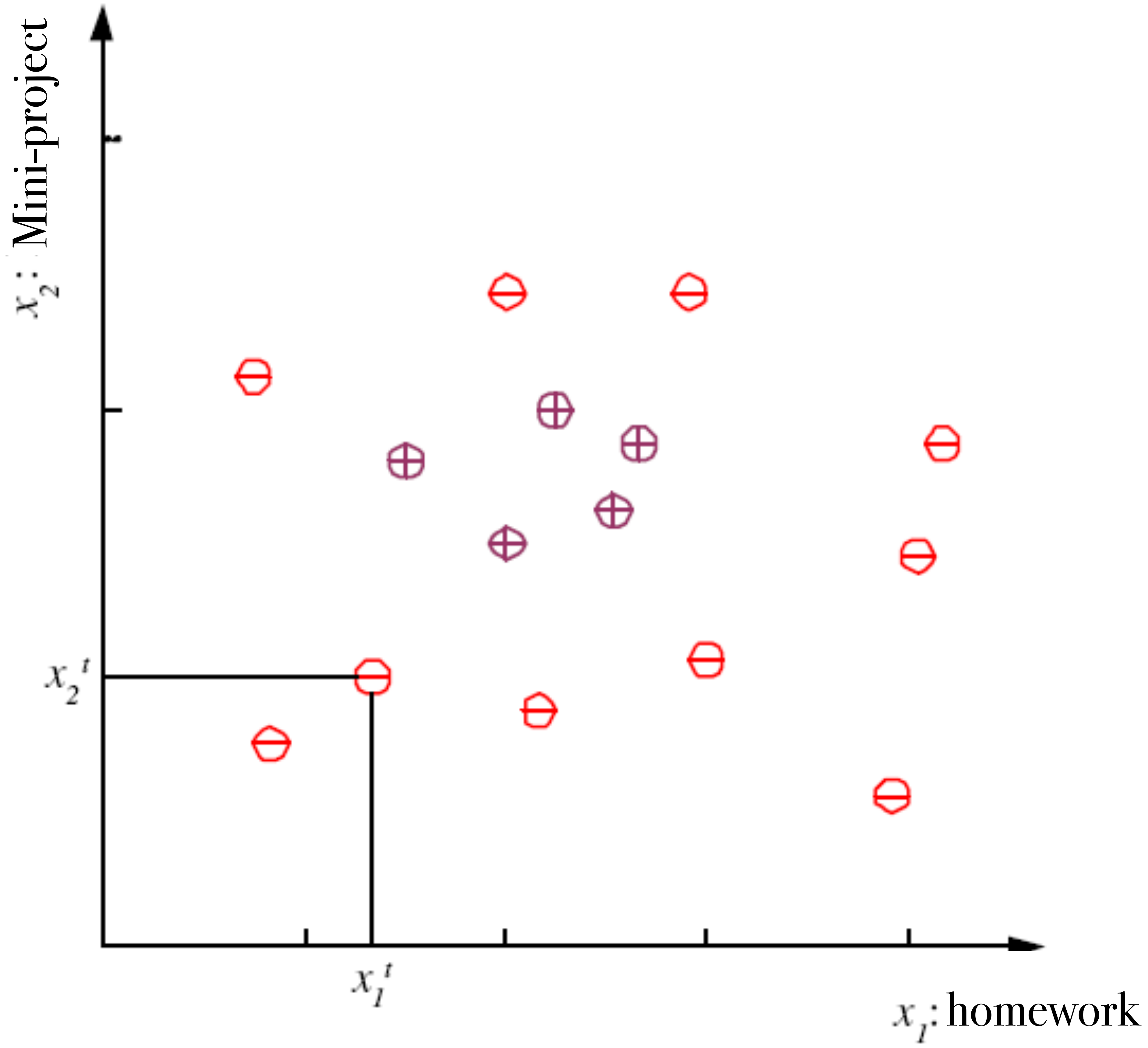
- Class C of a "good score"
  - Knowledge extraction: What do people expect from a good score?
- Output:

  Positive (+) and negative (–) examples
- Input representation:

  x1: homework, x2 : Mini-projects

Training Set

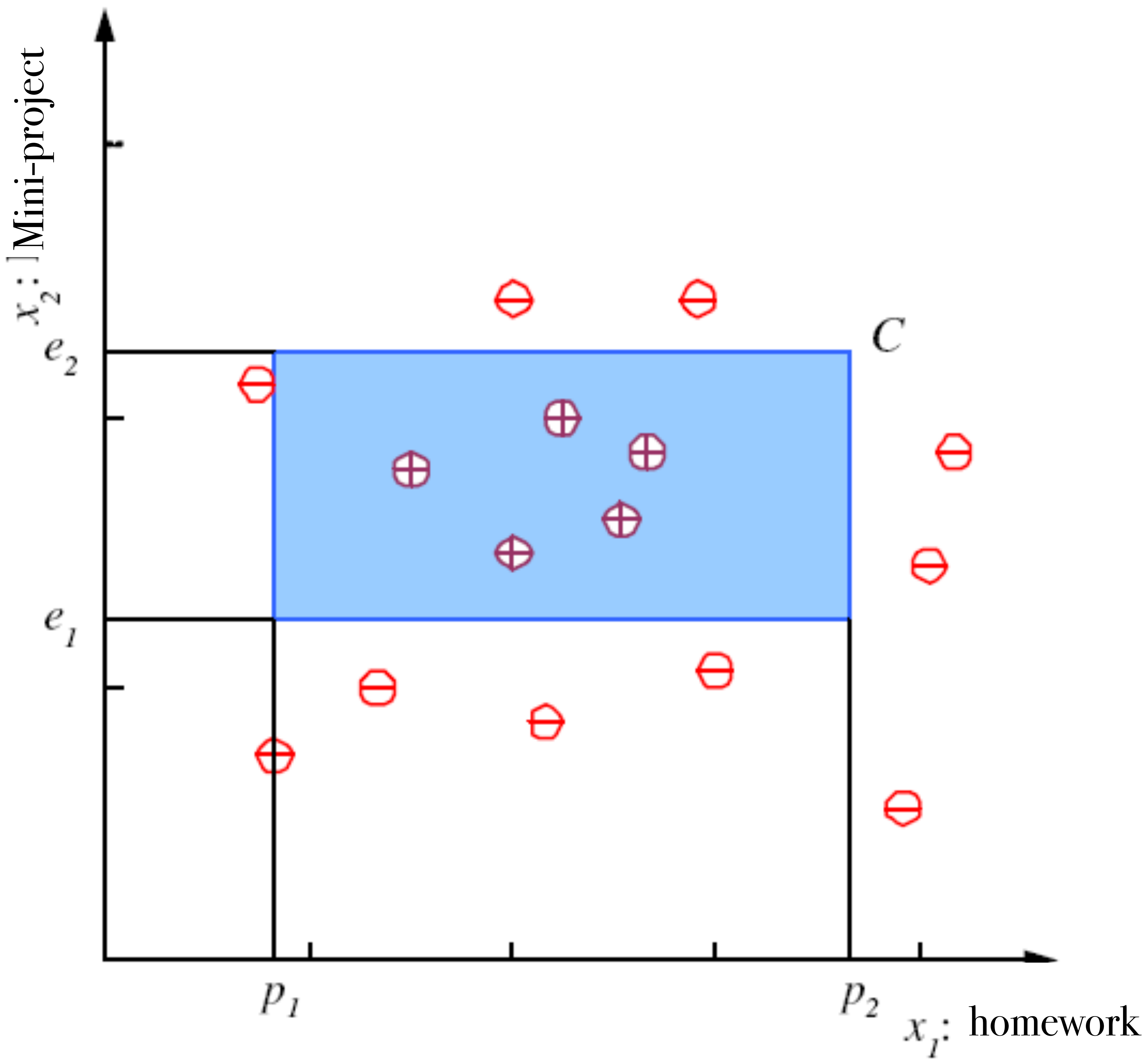
$$\mathcal{D} = \{(x^1, y^1), (x^2, y^2), \cdots, (x^N, y^N)\} \subset \mathcal{X} \times \mathcal{Y}$$



$$y = \begin{cases} 1 & if\, x\, is\, positive \\ 0 & if\, x\, is\, negative \end{cases}$$

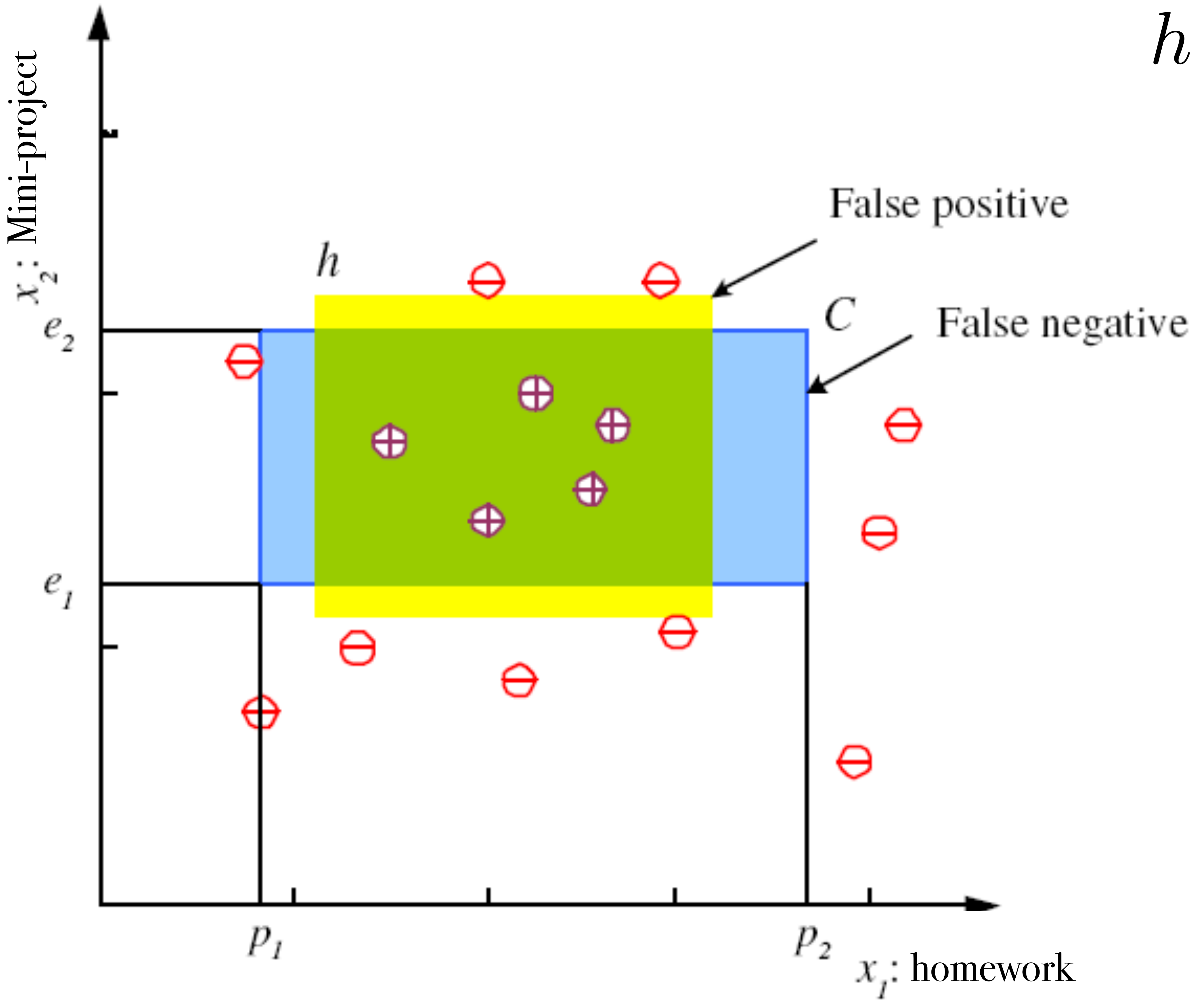
$$x^t = \begin{bmatrix} x_1^t \\ x_2^t \end{bmatrix}$$

Class C



Tom M. Mitchell is an American computer scientist and E. Fredkin University Professor at the Carnegie Mellon University.

Hypothesis class $\mathcal{H}$



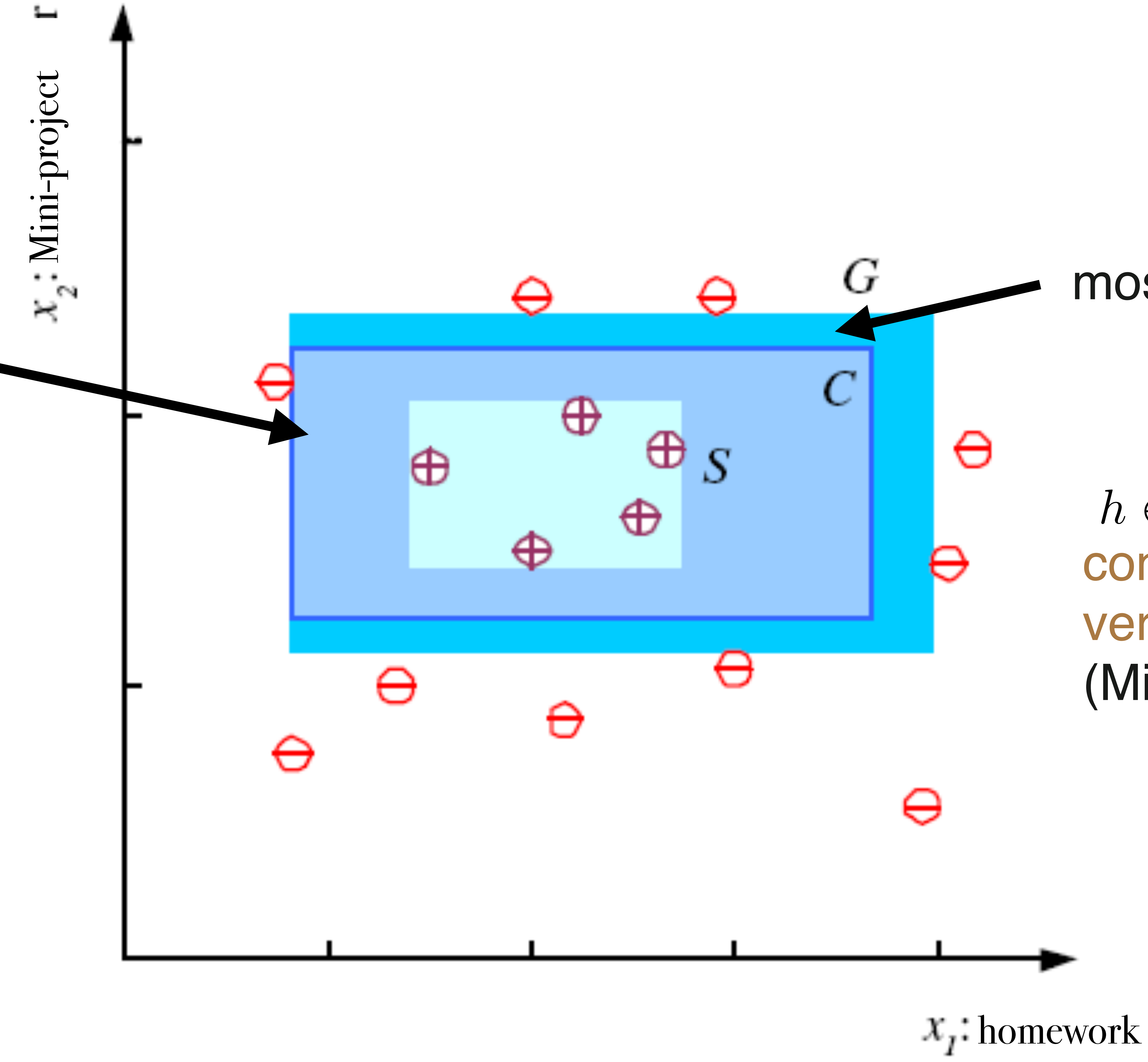
$$h(x) = \begin{cases} 1 & if\, x\, is\, positive \\ 0 & if\, x\, is\, negative \end{cases}$$

Error of *h* on $\mathcal{H}$

$$E\left(h|\mathcal{X}\right) = \sum_{t=1}^{N} \mathbf{1}\left(h\left(x^t\right) \neq y^t\right)$$

Version Space (Mitchell, 1997)

most specific hypothesis, S



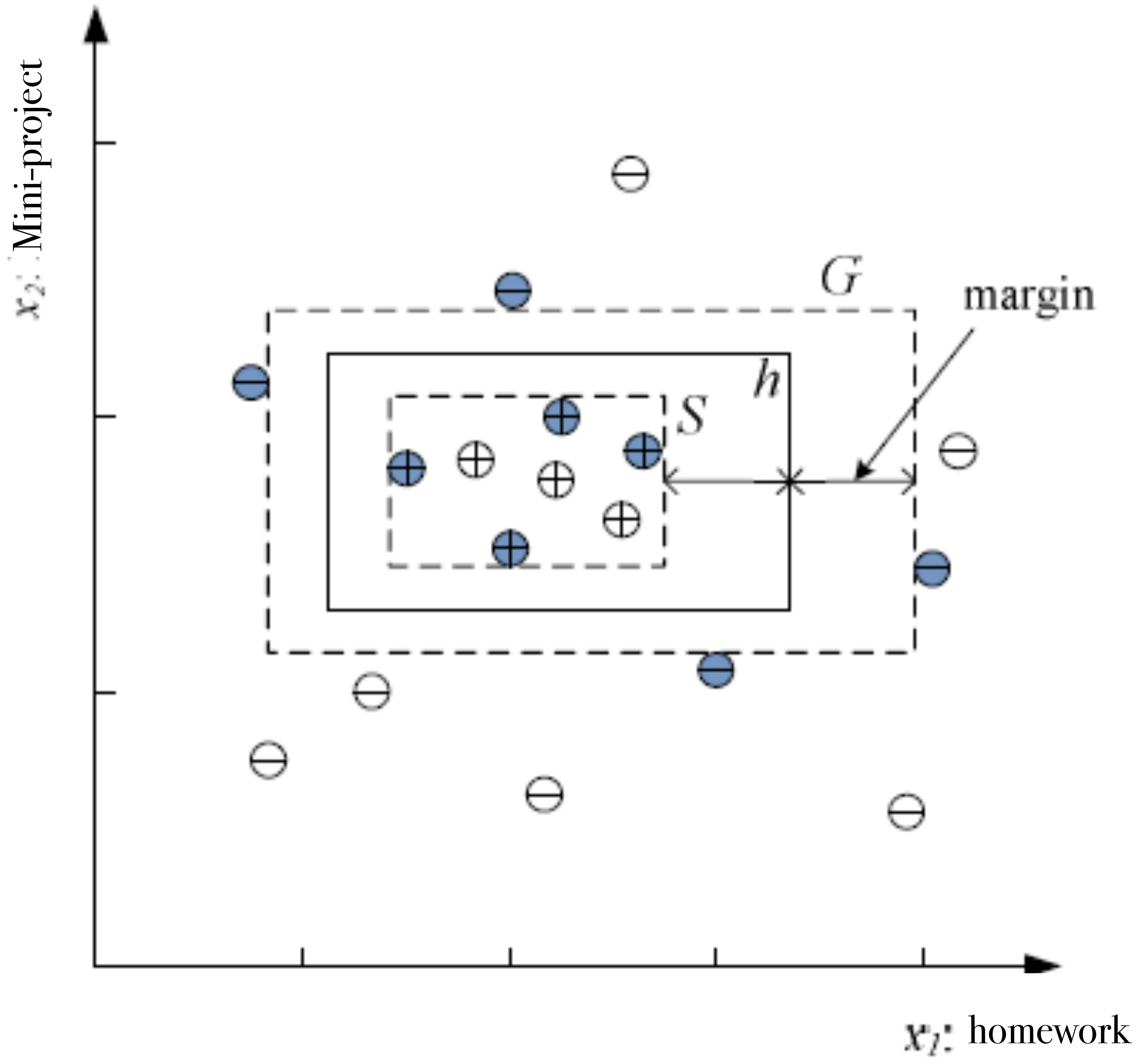
most general hypothesis, G

$h \in \mathcal{H}$ between S and G is consistent and make up the version space (Mitchell, 1997)

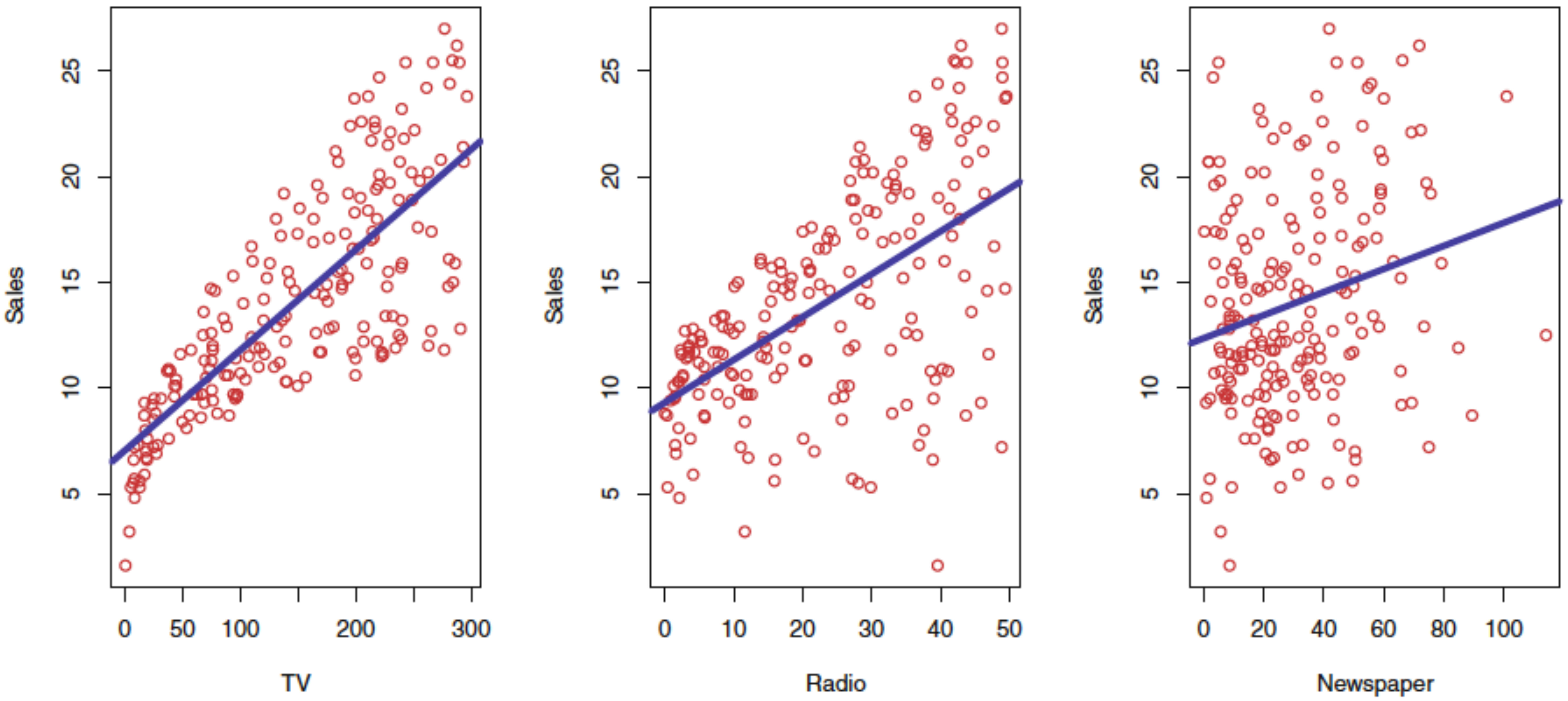
Margin

Choose *h* with largest margin

Occam's Razor

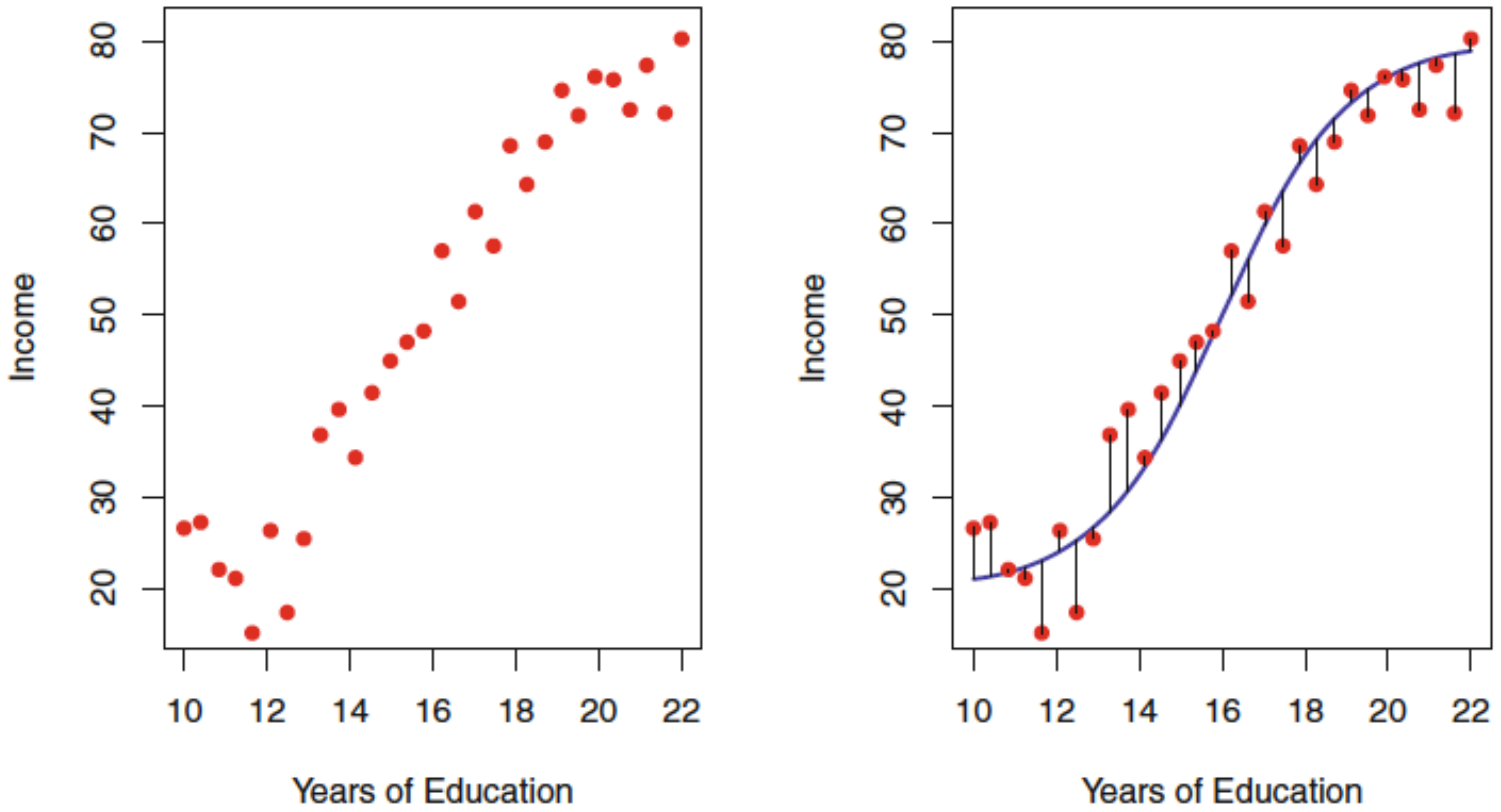


The Advertising data set



The Advertising data set. The plot displays sales, in thousands of units, as a function of TV, radio, and newspaper budgets, in thousands of dollars, for 200 different markets. In each plot we show the simple least squares fit of sales to that variable, as described in Chapter 3. In other words, each blue line represents a simple model that can be used to predict sales using TV, radio, and newspaper, respectively.

The Income data set



Left: The red dots are the observed values of income (in tens of thousands of dollars) and years of education for 30 individuals. Right: The blue curve represents the true underlying relationship between income and years of education, which is generally unknown (but is known in this case because the data were simulated). The black lines represent the error associated with each observation. Note that some errors are positive (if an observation lies above the blue curve) and some are negative (if an observation lies below the curve). Overall, these errors have approximately mean zero.

More examples

1. Classification: Determine which discrete category the example is

2. Recognizing patterns: Speech Recognition, facial identity, etc

3. Recommender Systems: Noisy data, commercial pay-off (e.g., Amazon, Netflix).

4. Information retrieval: Find documents or images with similar content

5. Computer vision: detection, segmentation, depth estimation, optical flow,etc

6. Robotics: perception, planning, etc

7. Learning to play games;

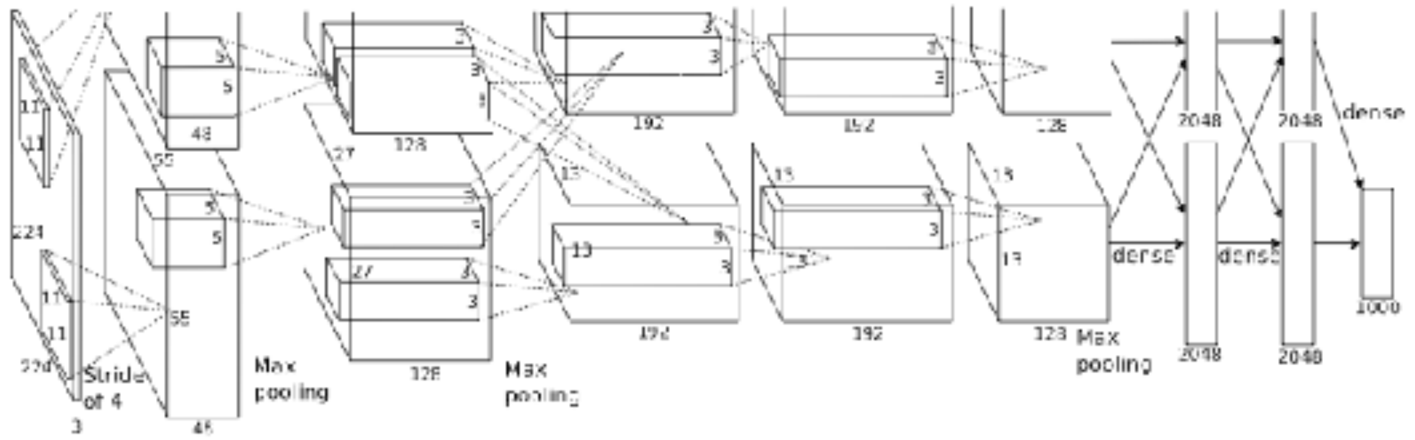
8. Recognizing anomalies: Unusual sequences of credit card transactions, panic situation at an airport

9. Spam filtering, fraud detection: The enemy adapts so we must adapt too.

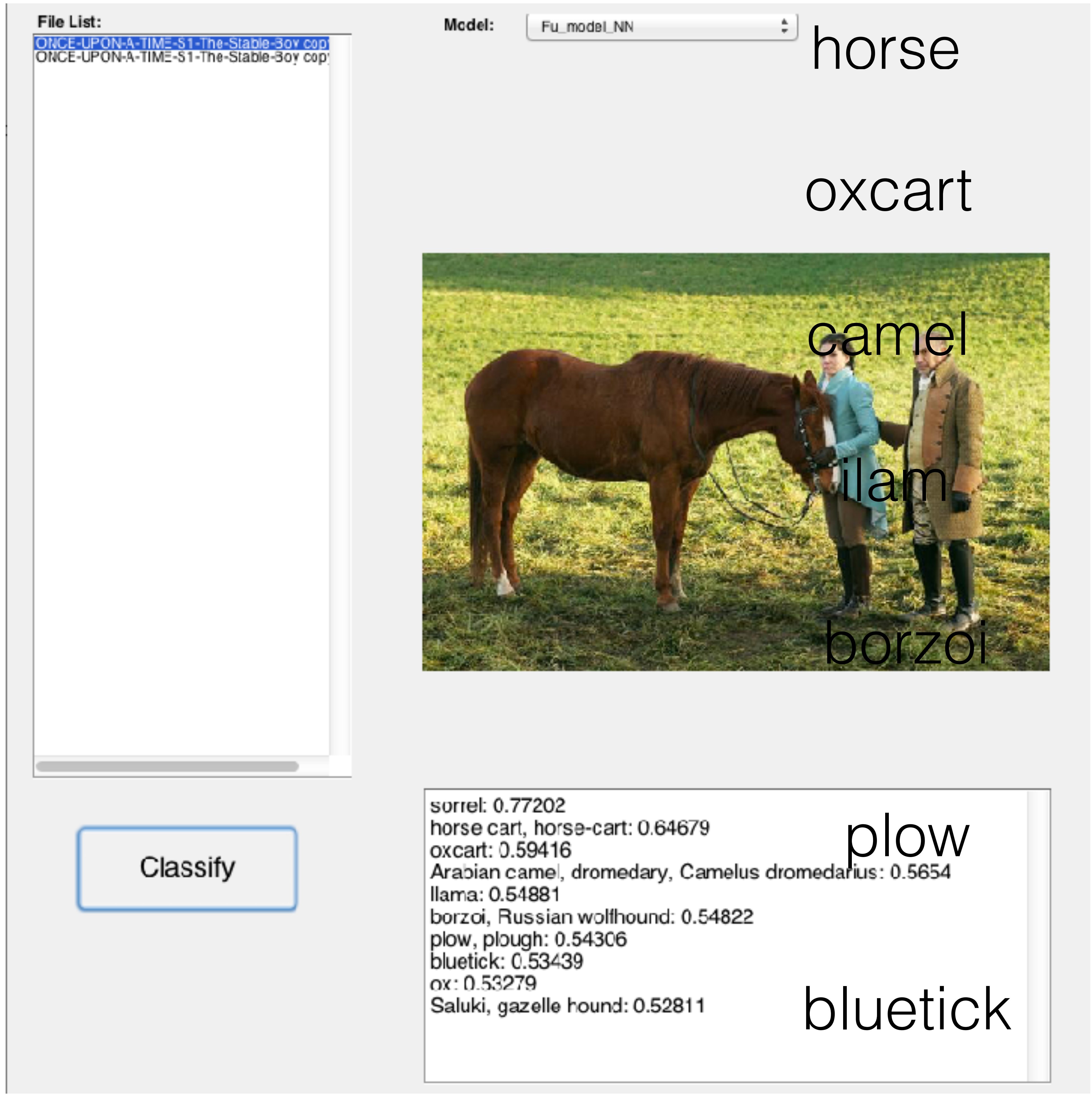
Object Recognition

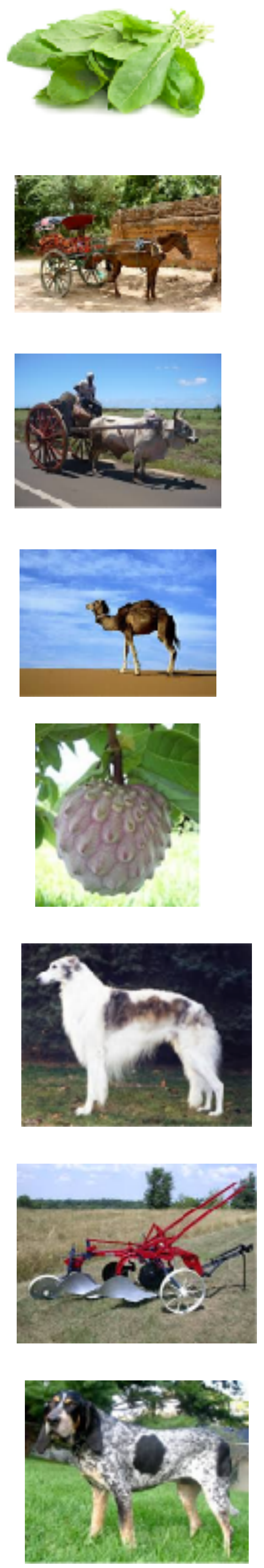
Computer Vision

Deep Convolutional Neural Network



horse
person
boots

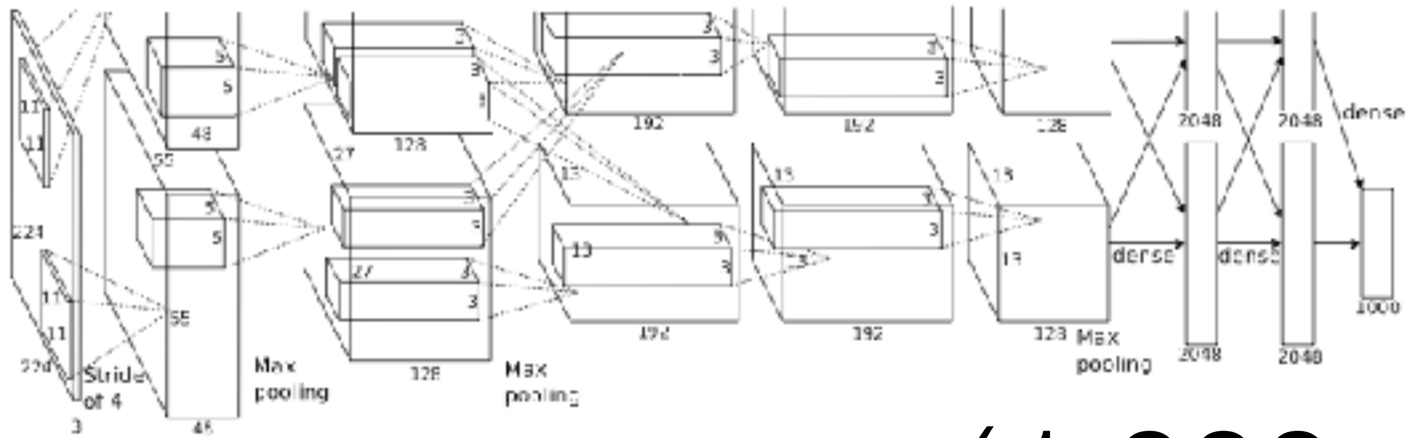




Object Recognition

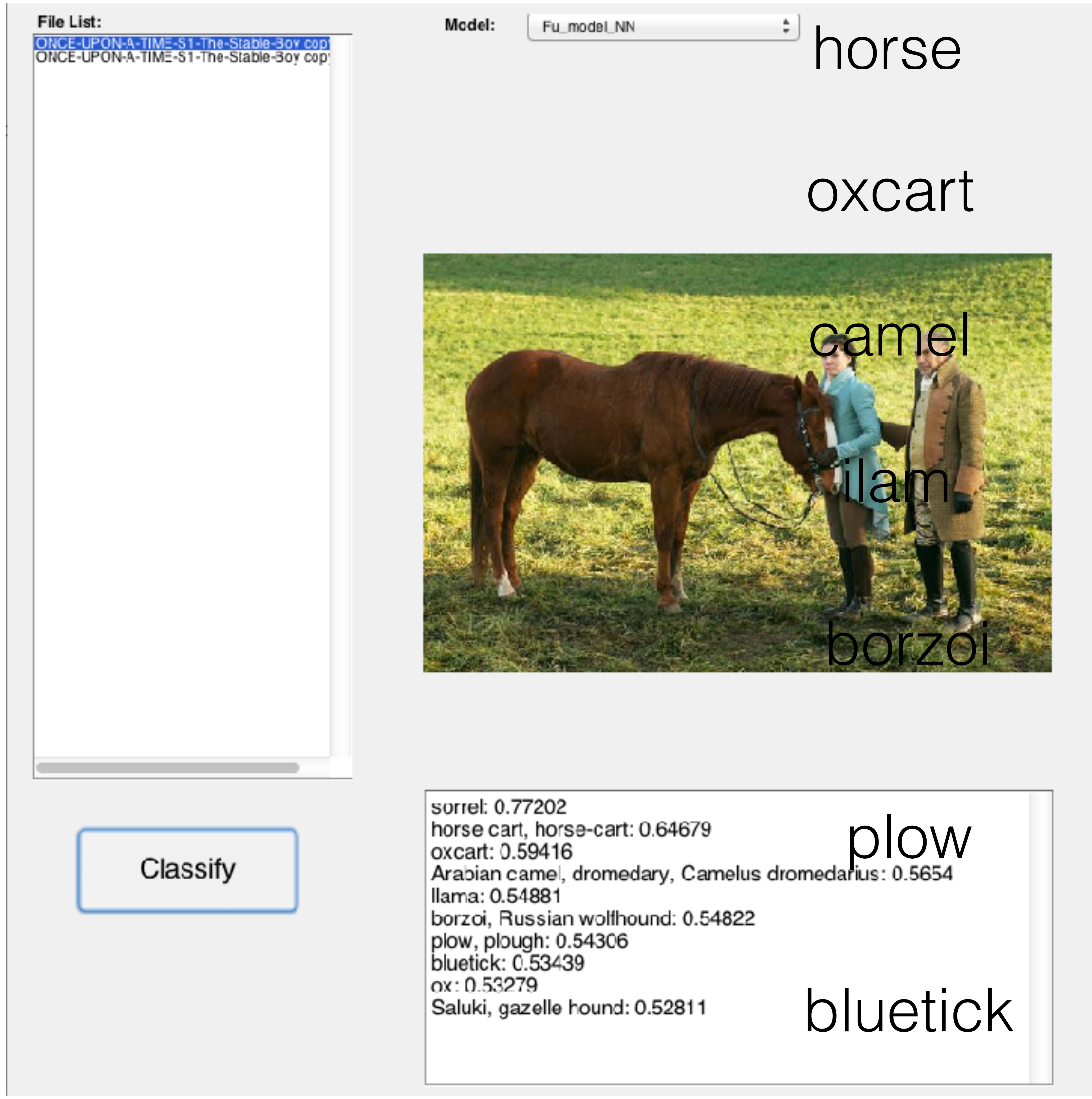
Computer Vision

Deep Convolutional Neural Network



horse
person
boots

(1,000 classes)



sorrel 

horse 

oxcart 

camel 

llam 

borzoi 

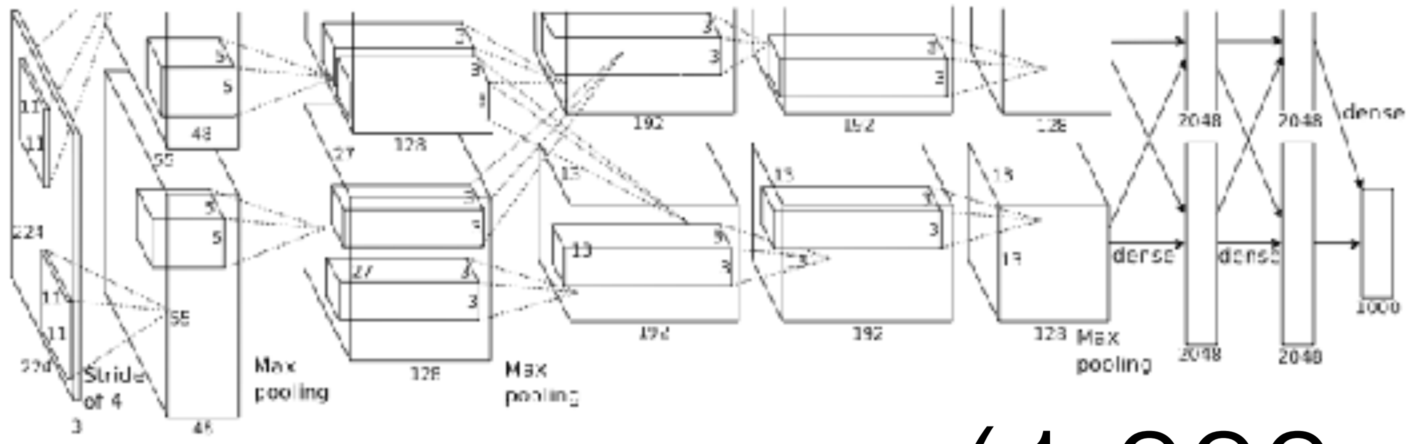
plow 

bluetick 

Object Recognition

Computer Vision

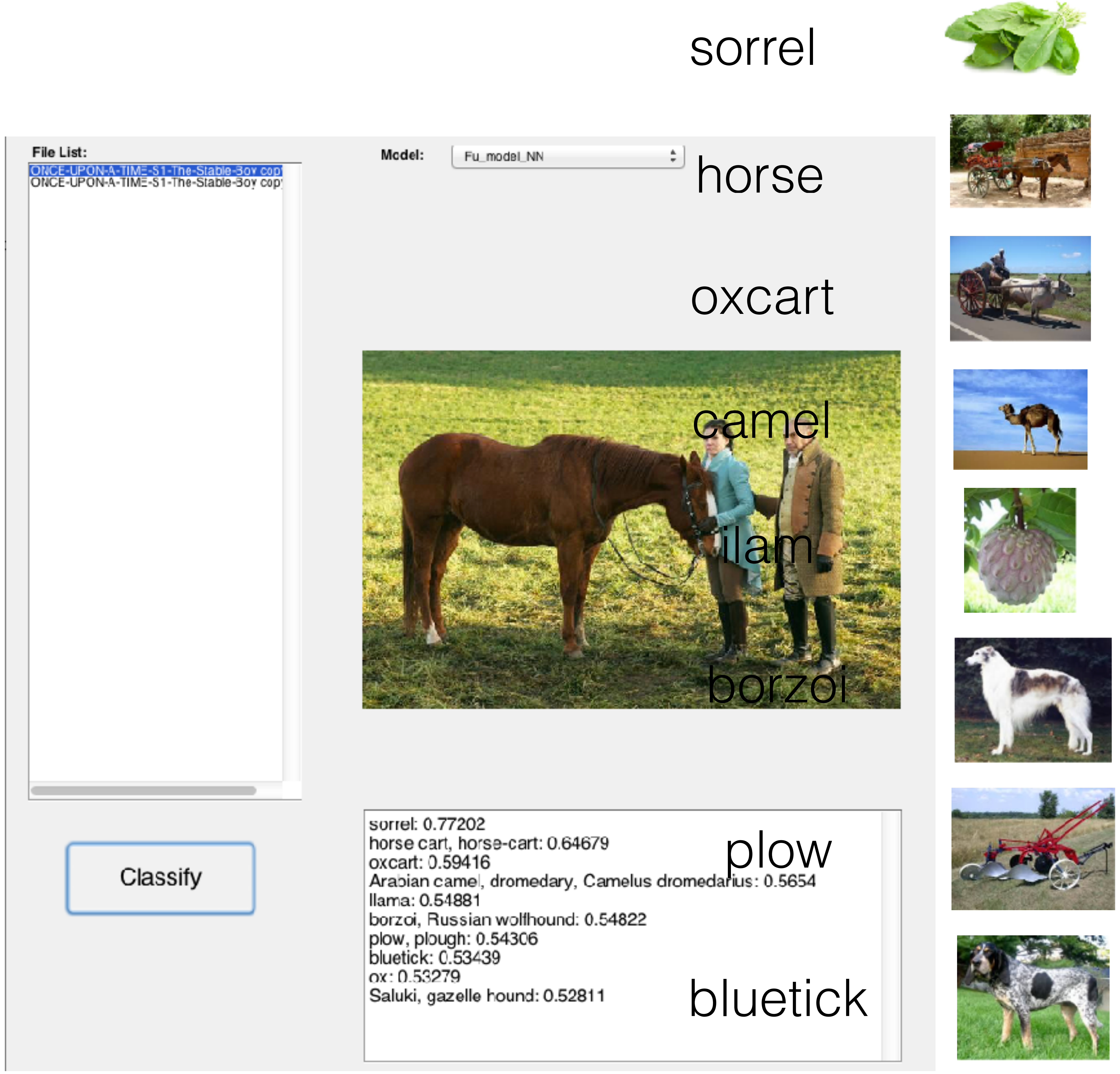
Deep Convolutional Neural Network



horse
person
boots

(1,000 classes)

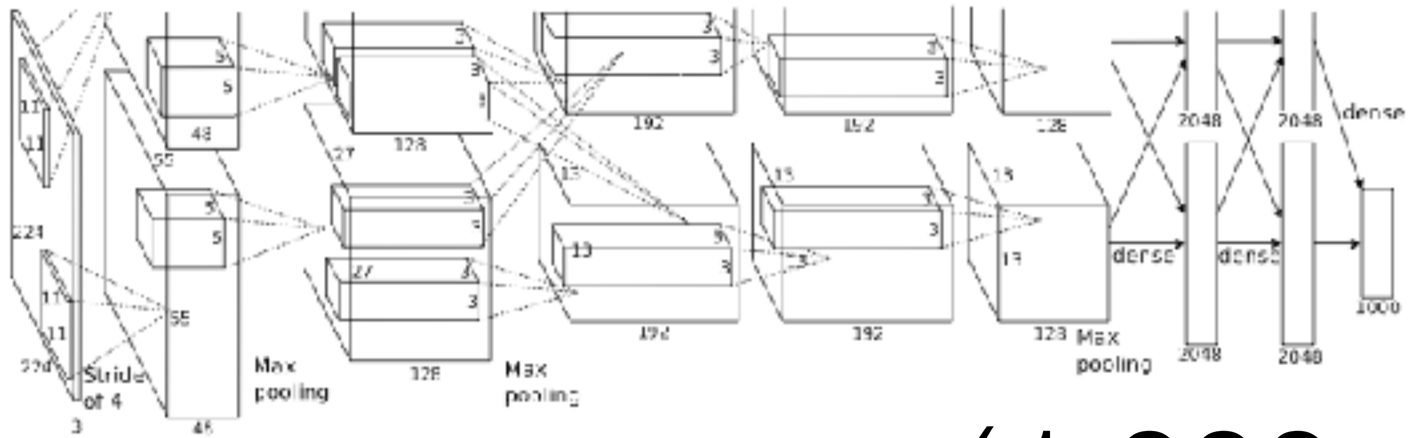
ImageNet: ~1M labeled images



Object Recognition

Computer Vision

Deep Convolutional Neural Network

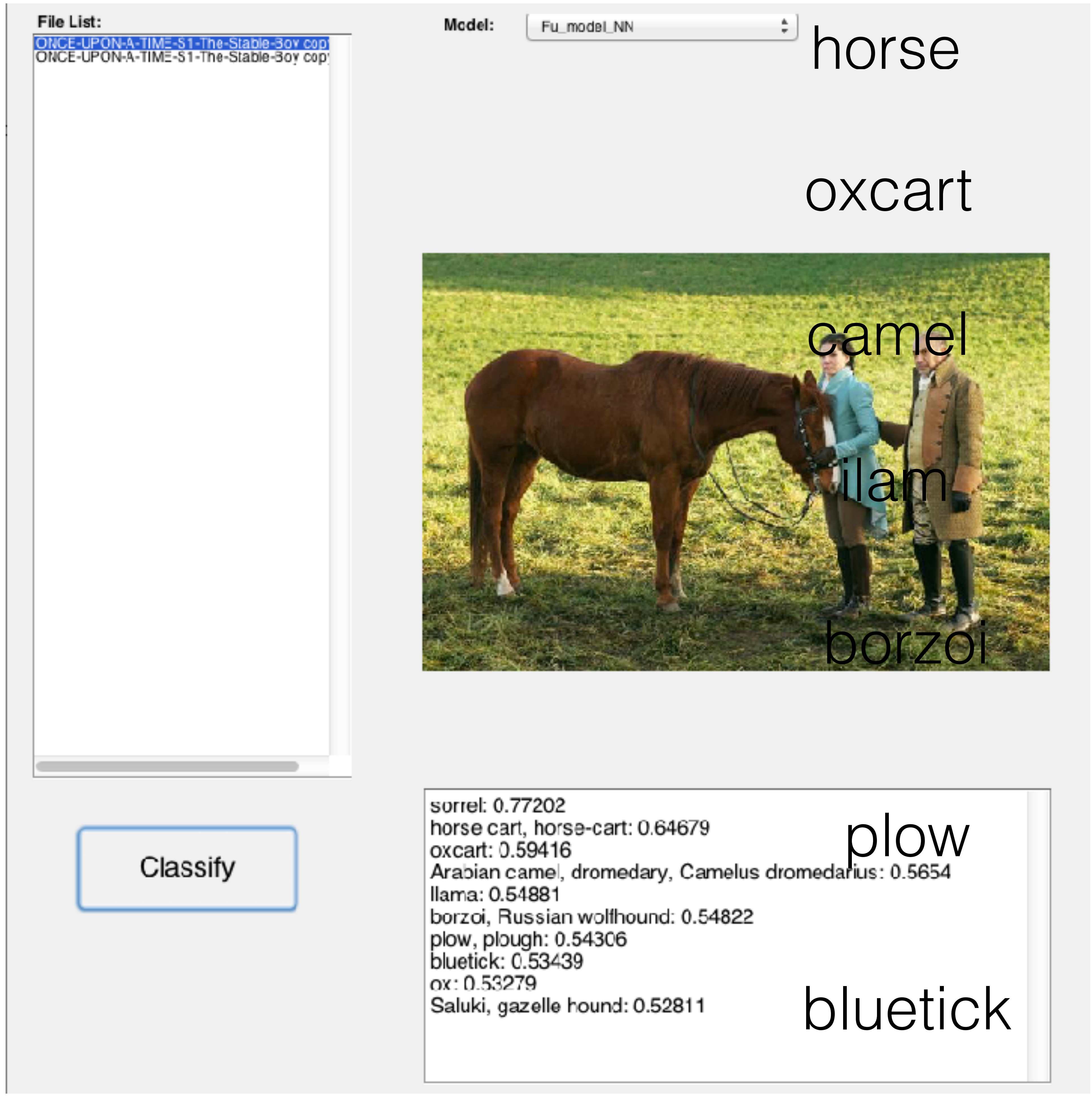


horse
person
boots

(1,000 classes)

ImageNet: ~1M labeled images

~1,000 images per class



sorrel 

horse 

oxcart 

camel 

llam 

borzoi 

plow 

bluetick 

Video Emotion Recognition

Affective Computing

300 hours of video uploaded to YouTube every minute

?

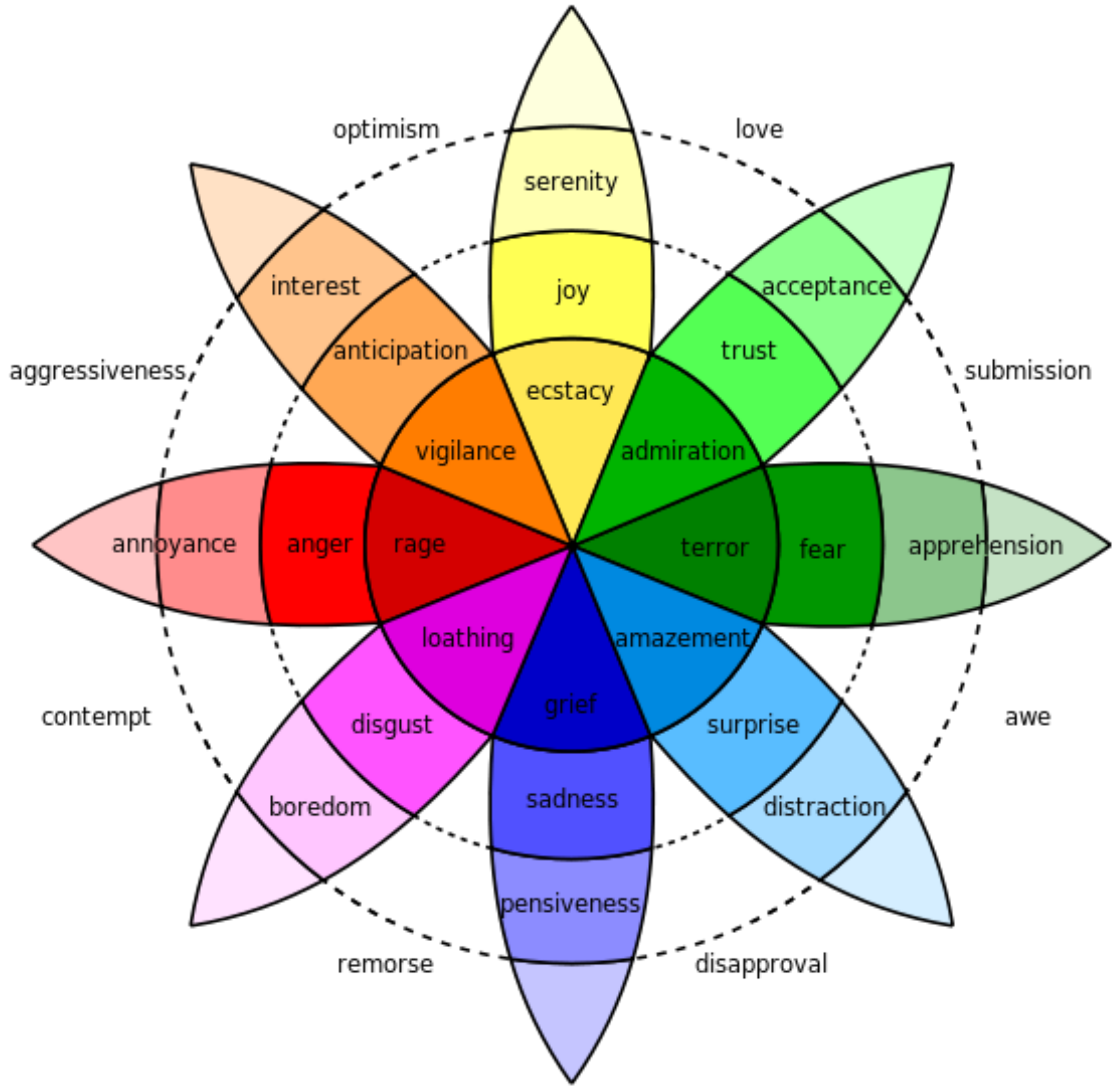
Fear

Sadness

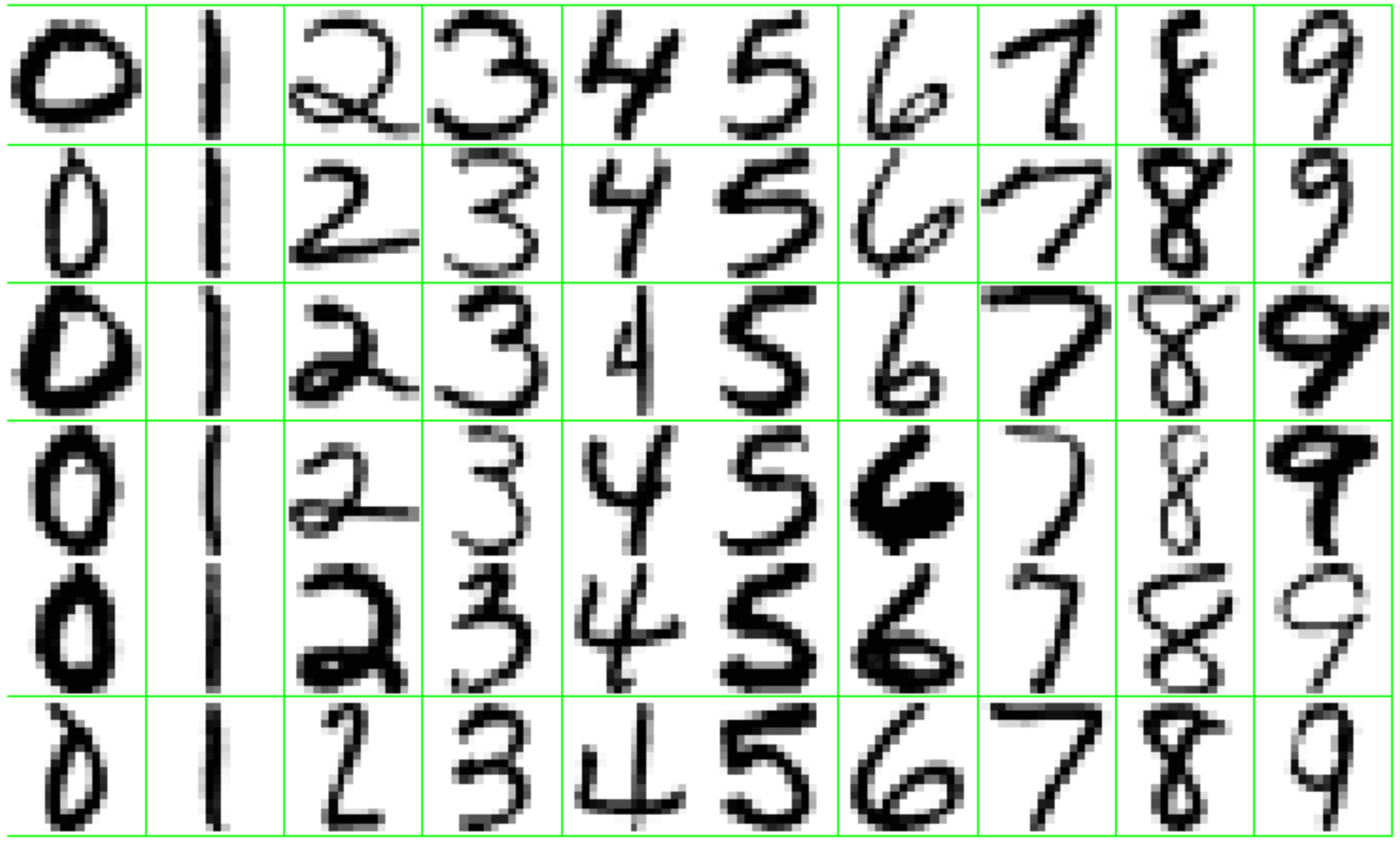
Joy

Anger

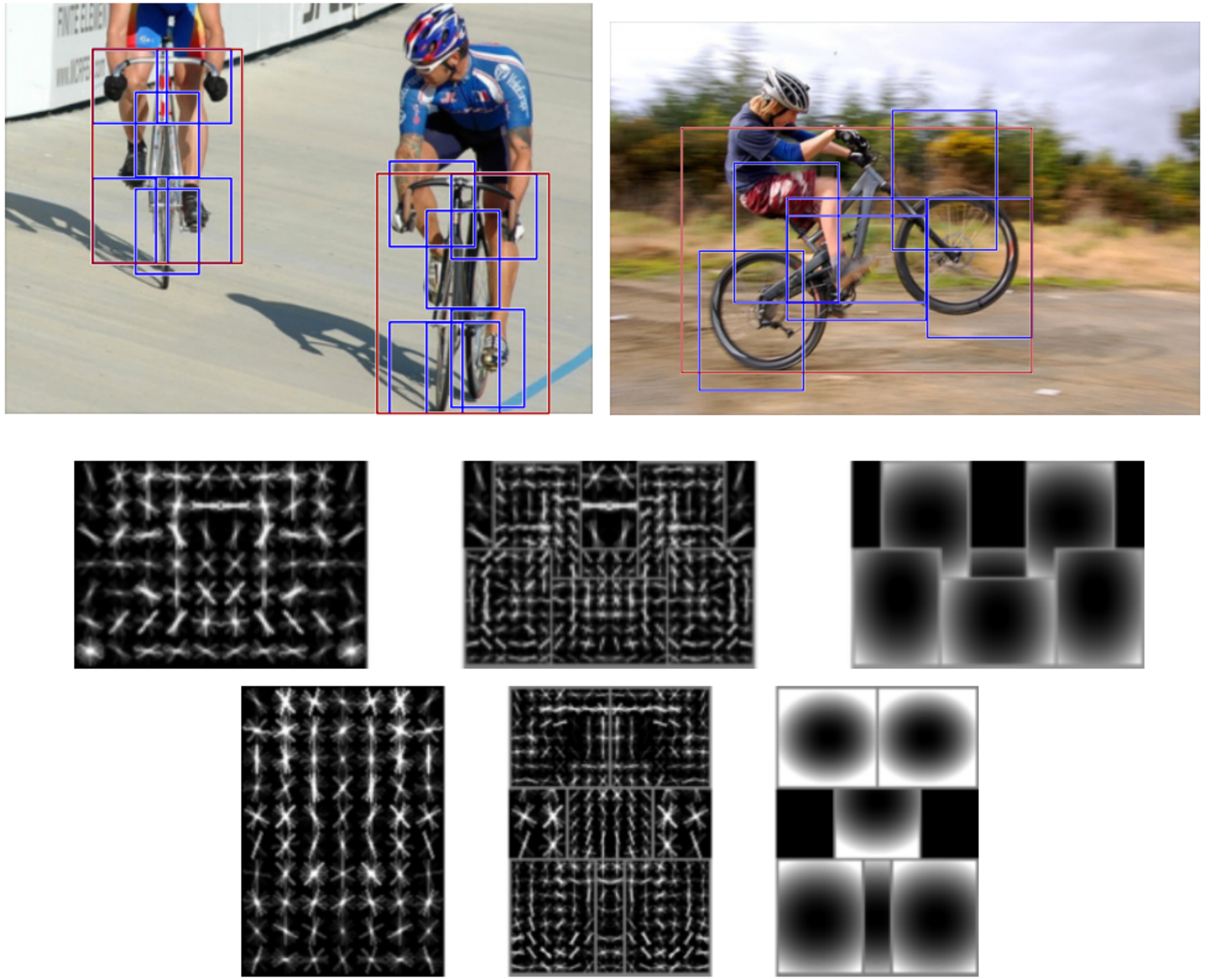
Disgust



plutchik wheel of emotions

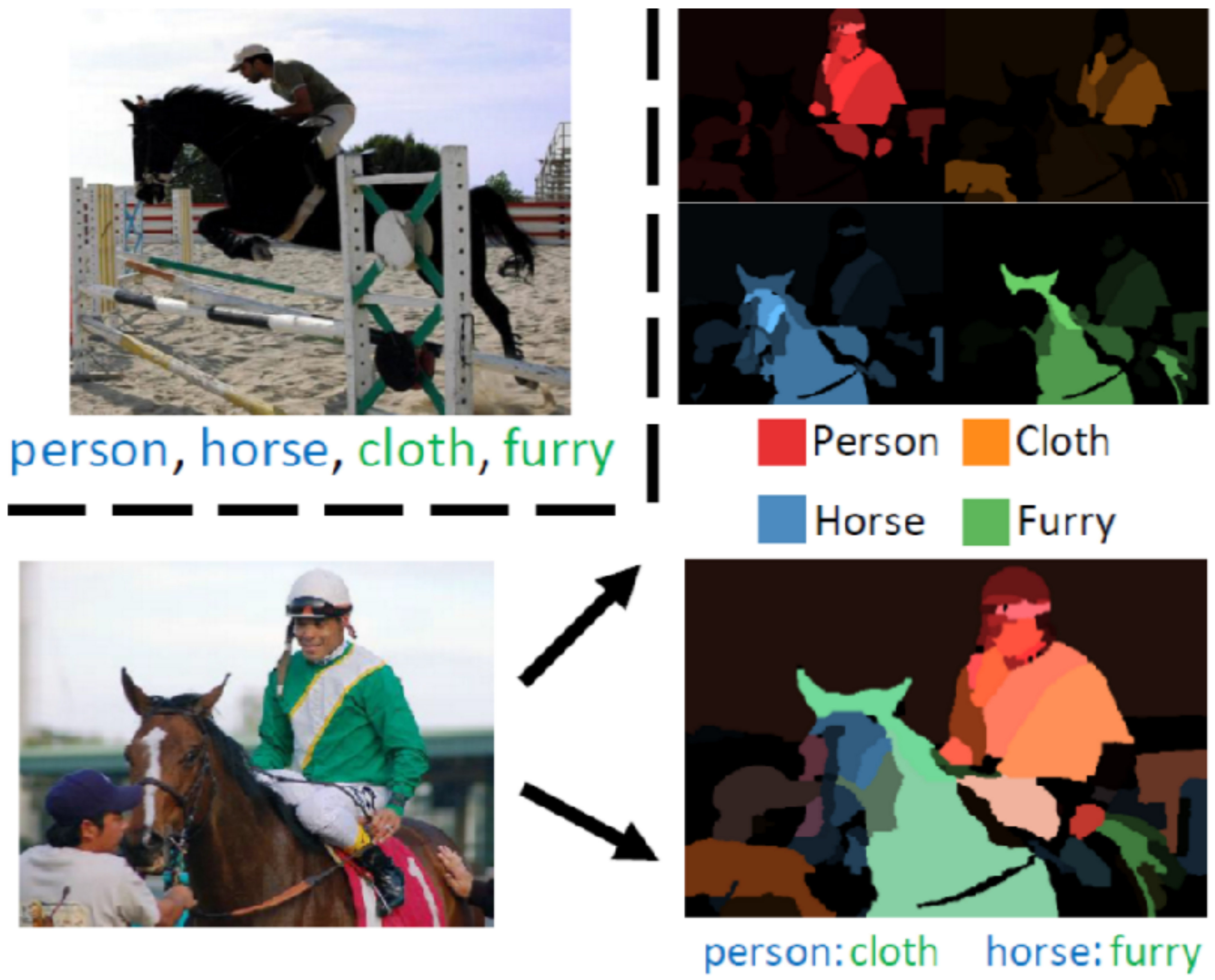


Identify the numbers in a handwritten zip code.



Weakly supervised learning

Weakly supervised learning is a machine learning framework where the model is trained using examples that are only partially annotated or labeled.



Autonomous Driving

Uber@Pittsburgh

Google

Robotics

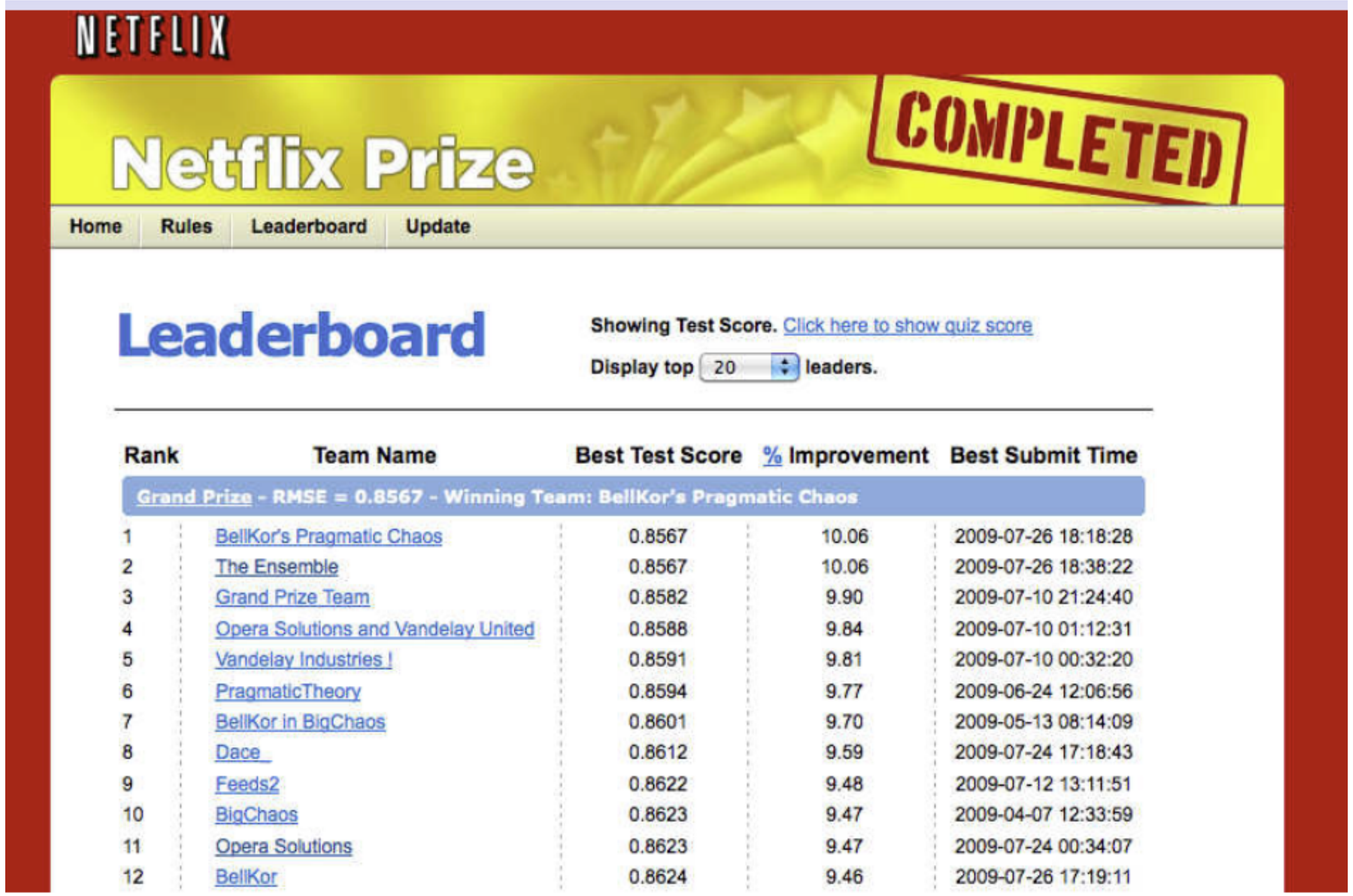
Surgical Robotics

Verb Surgical: http://www.verbsurgical.com/

Playing Catch and Juggling
with a Humanoid Robot

http://forums.wdwmagic.com/threads/disney-research-pittsburgh.866204/

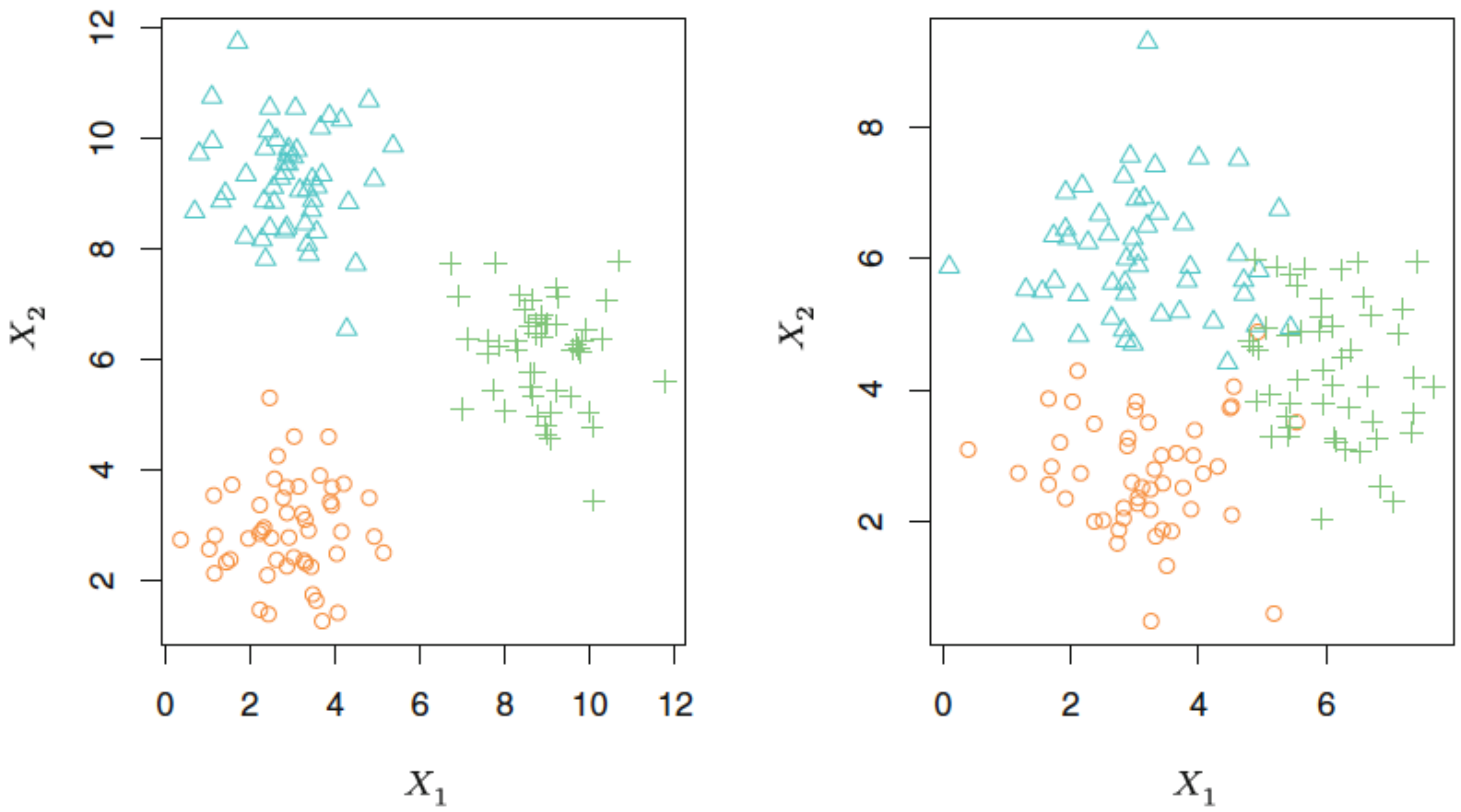
The Netflix Prize



Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace_	0.8612	9.59	2009-07-24 17:18:43
9	Feeds2	0.8622	9.48	2009-07-12 13:11:51
10	BigChaos	0.8623	9.47	2009-04-07 12:33:59
11	Opera Solutions	0.8623	9.47	2009-07-24 00:34:07
12	BellKor	0.8624	9.46	2009-07-26 17:19:11

The competition started in October 2006. Training data is ratings for 18000 movies by 400000 Netflix customers, each rating between 1 and 5. The training data is very sparse about 98% missing. The objective is to predict the rating for a set of 1 million customer-movie pairs that are missing in the training data. Netflix's original algorithm achieved a root MSE of 0.953. The first team to achieve a 10% improvement wins one million dollars.

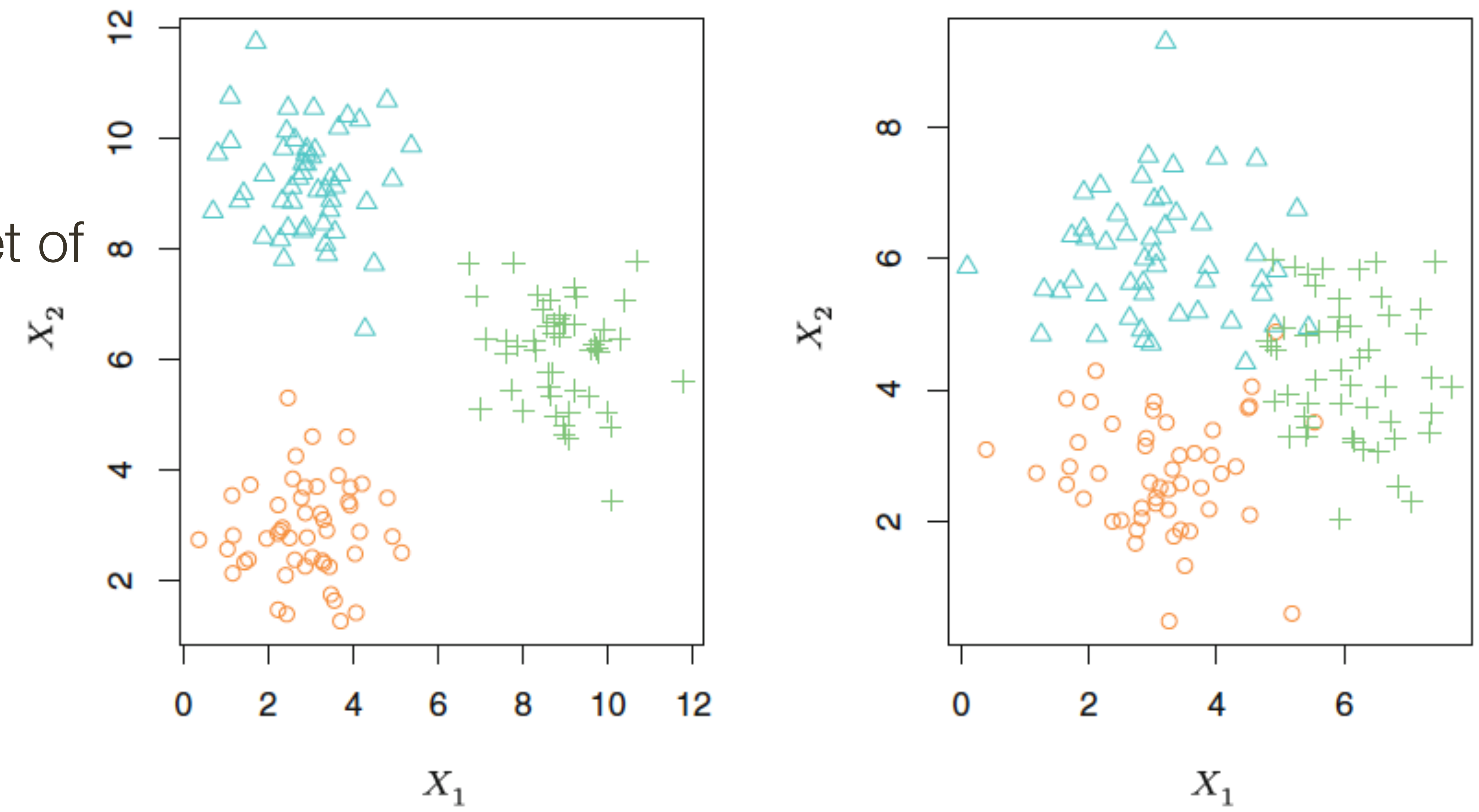
Unsupervised Learning



A clustering data set involving three groups. Each group is shown using a different colored symbol. Left: The three groups are well-separated. In this setting, a clustering approach should successfully identify the three groups. Right: There is some overlap among the groups. Now the clustering task is more challenging.

Unsupervised Learning

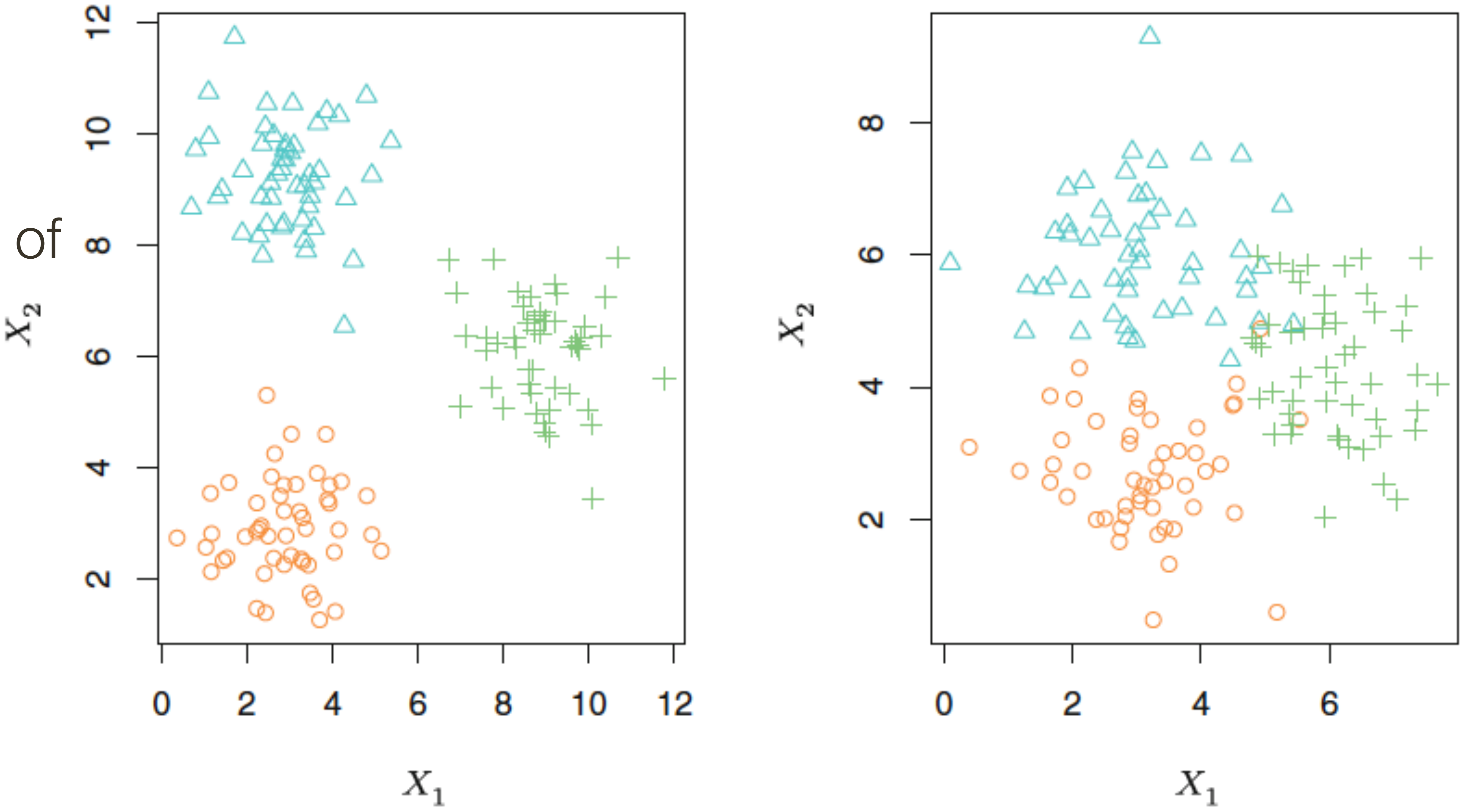
- No outcome variable, just a set of predictors (features) measured on a set of samples.



A clustering data set involving three groups. Each group is shown using a different colored symbol. Left: The three groups are well-separated. In this setting, a clustering approach should successfully identify the three groups. Right: There is some overlap among the groups. Now the clustering task is more challenging.

Unsupervised Learning

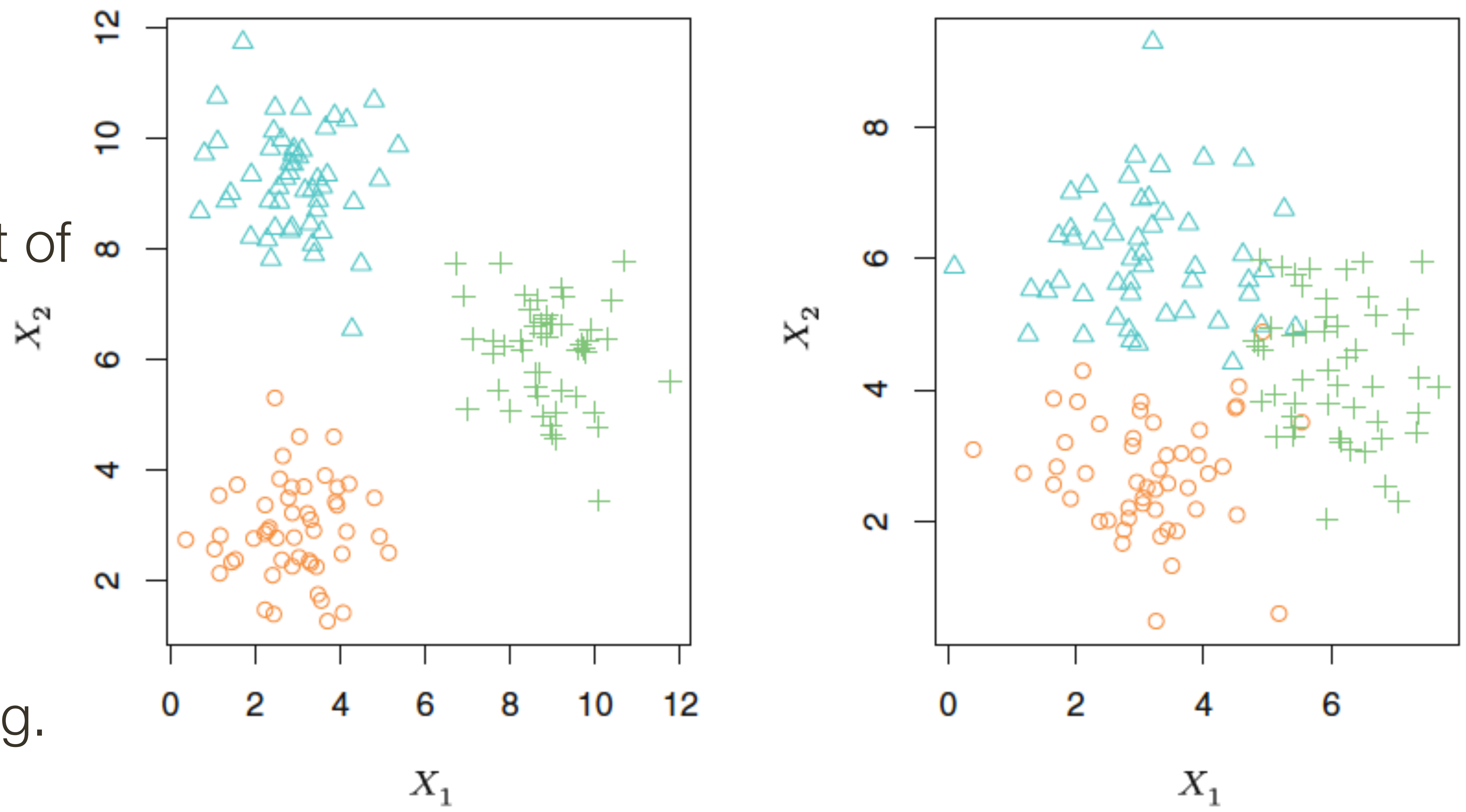
- No outcome variable, just a set of predictors (features) measured on a set of samples.
- objective is more fuzzy
    - find groups of samples that behave similarly



A clustering data set involving three groups. Each group is shown using a different colored symbol. Left: The three groups are well-separated. In this setting, a clustering approach should successfully identify the three groups. Right: There is some overlap among the groups. Now the clustering task is more challenging.

Unsupervised Learning

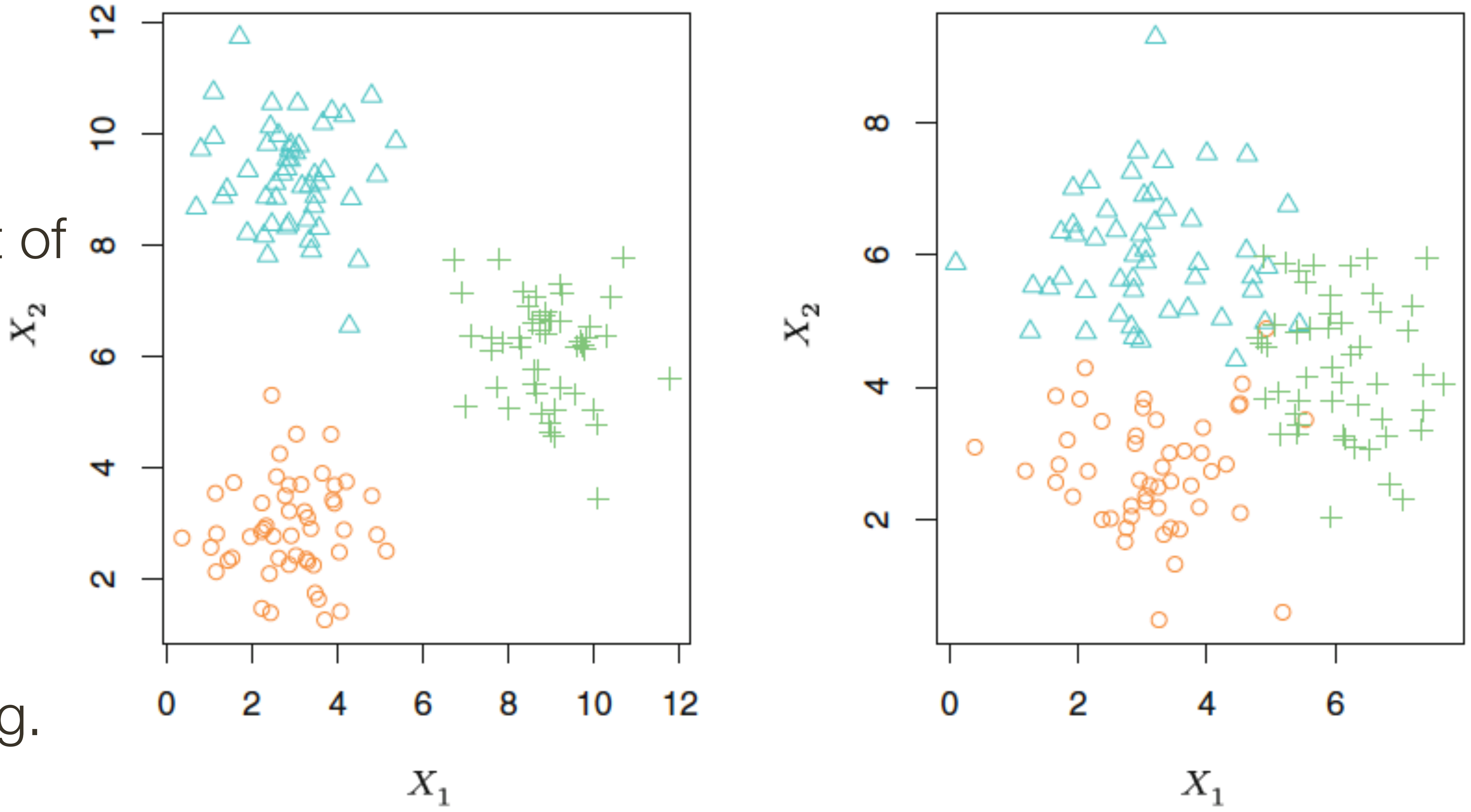
- No outcome variable, just a set of predictors (features) measured on a set of samples.
- objective is more fuzzy
  - find groups of samples that behave similarly
- difficult to know how well your are doing.



A clustering data set involving three groups. Each group is shown using a different colored symbol. Left: The three groups are well-separated. In this setting, a clustering approach should successfully identify the three groups. Right: There is some overlap among the groups. Now the clustering task is more challenging.

Unsupervised Learning

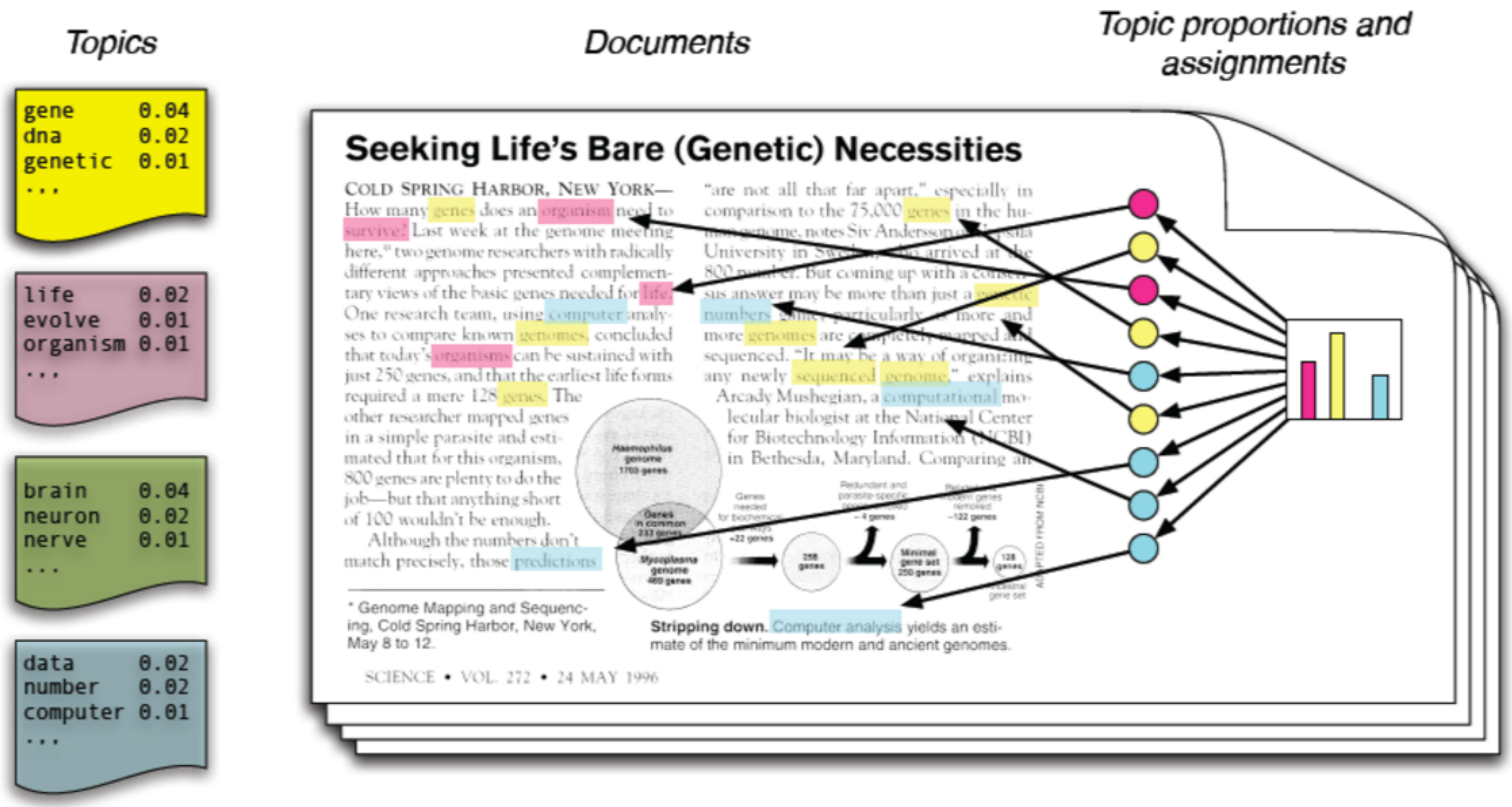
- No outcome variable, just a set of predictors (features) measured on a set of samples.
- objective is more fuzzy
    - find groups of samples that behave similarly
- difficult to know how well your are doing.
- different from supervised learning, but can be useful as a pre-processing step for supervised learning.



A clustering data set involving three groups. Each group is shown using a different colored symbol. Left: The three groups are well-separated. In this setting, a clustering approach should successfully identify the three groups. Right: There is some overlap among the groups. Now the clustering task is more challenging.

Applications of Unsupervised Clustering

Topic Model (Latent Dirichlet Allocation)



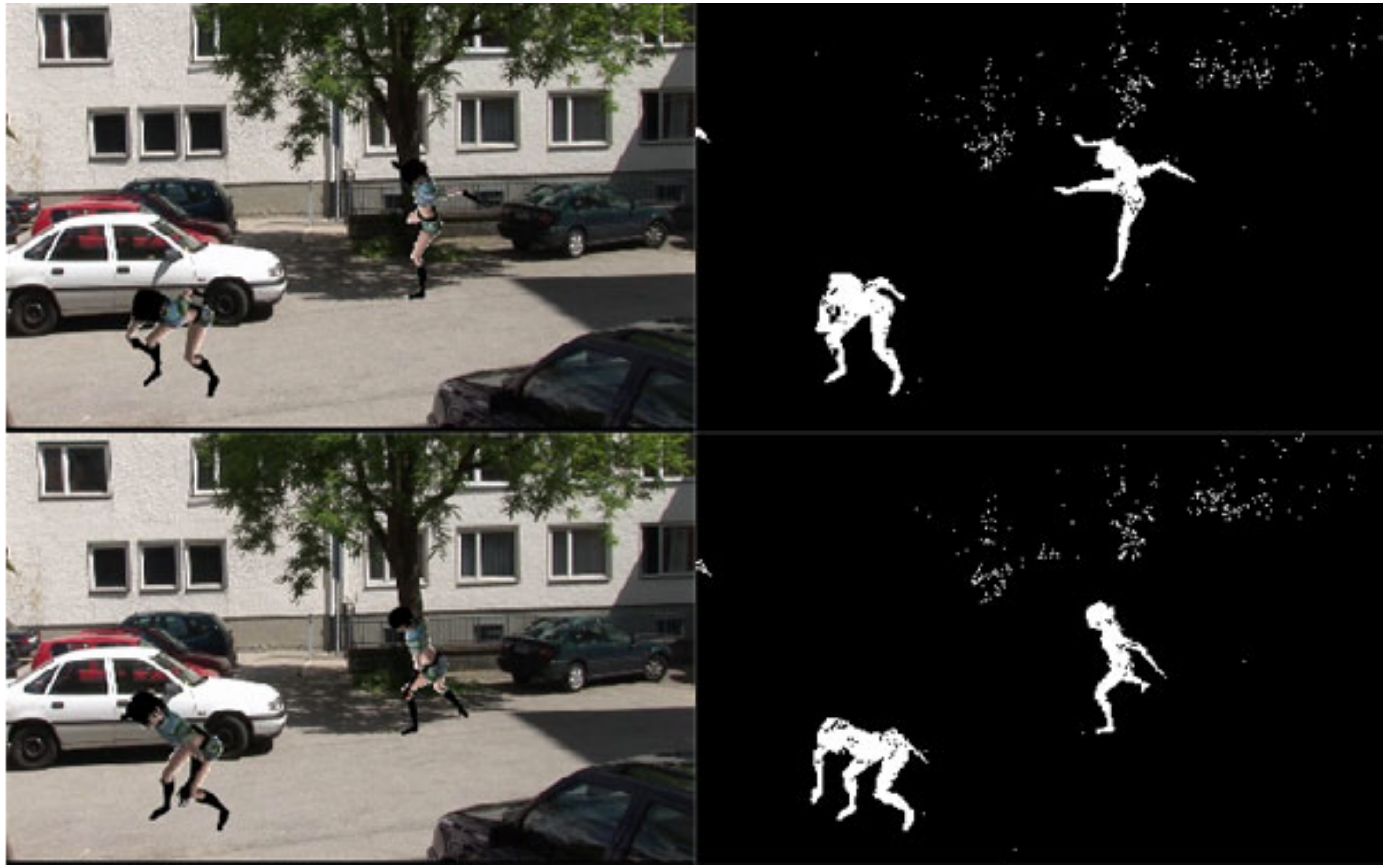
Andrew Y. Ng

Michael Jordan

David M. Blei, Andrew Y. Ng, Michael I. Jordan, Latent Dirichlet Allocation, JMLR 2003

Applications of Unsupervised Clustering

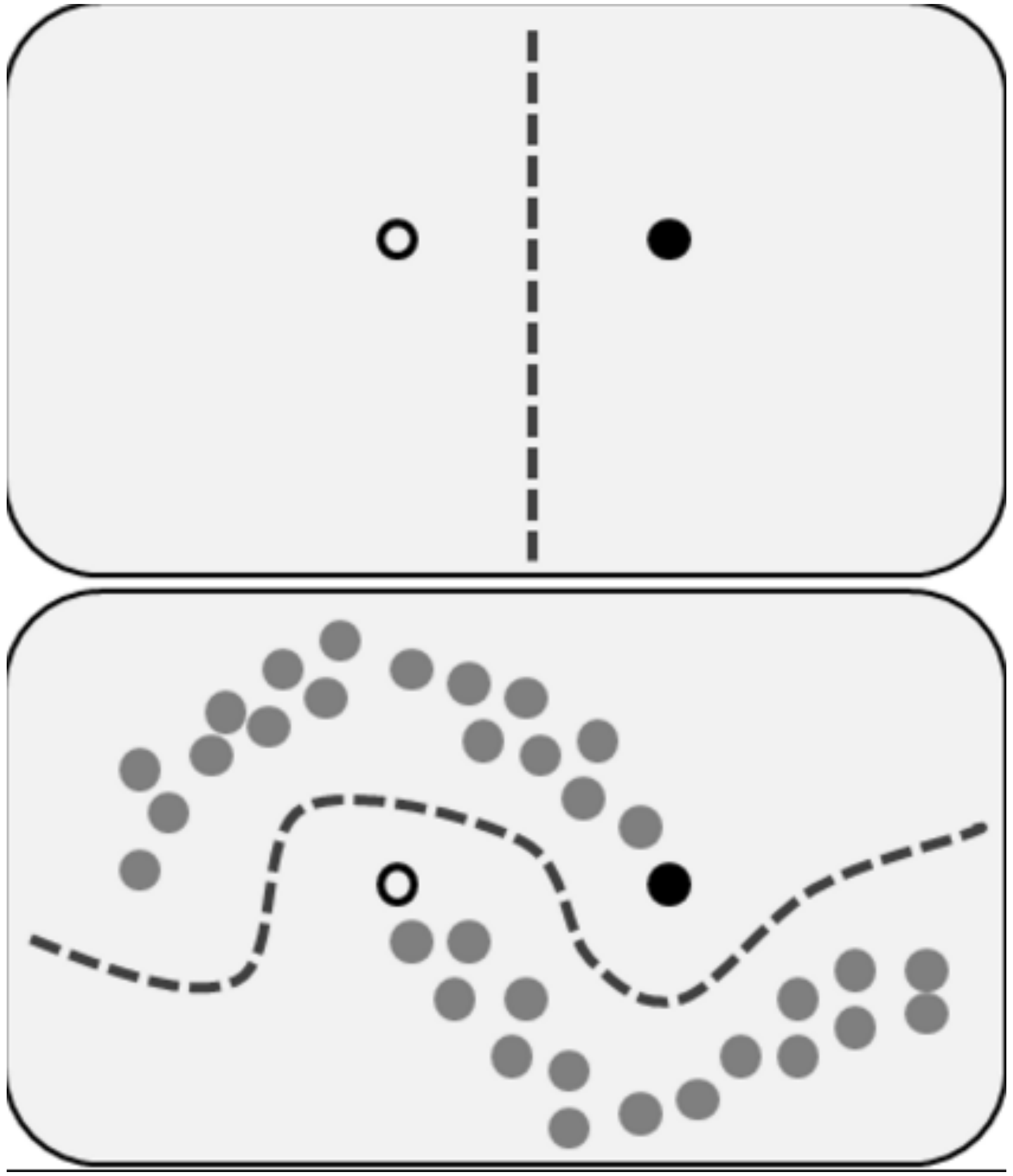
Gaussian Mixture Model (GMM) for Background Subtraction



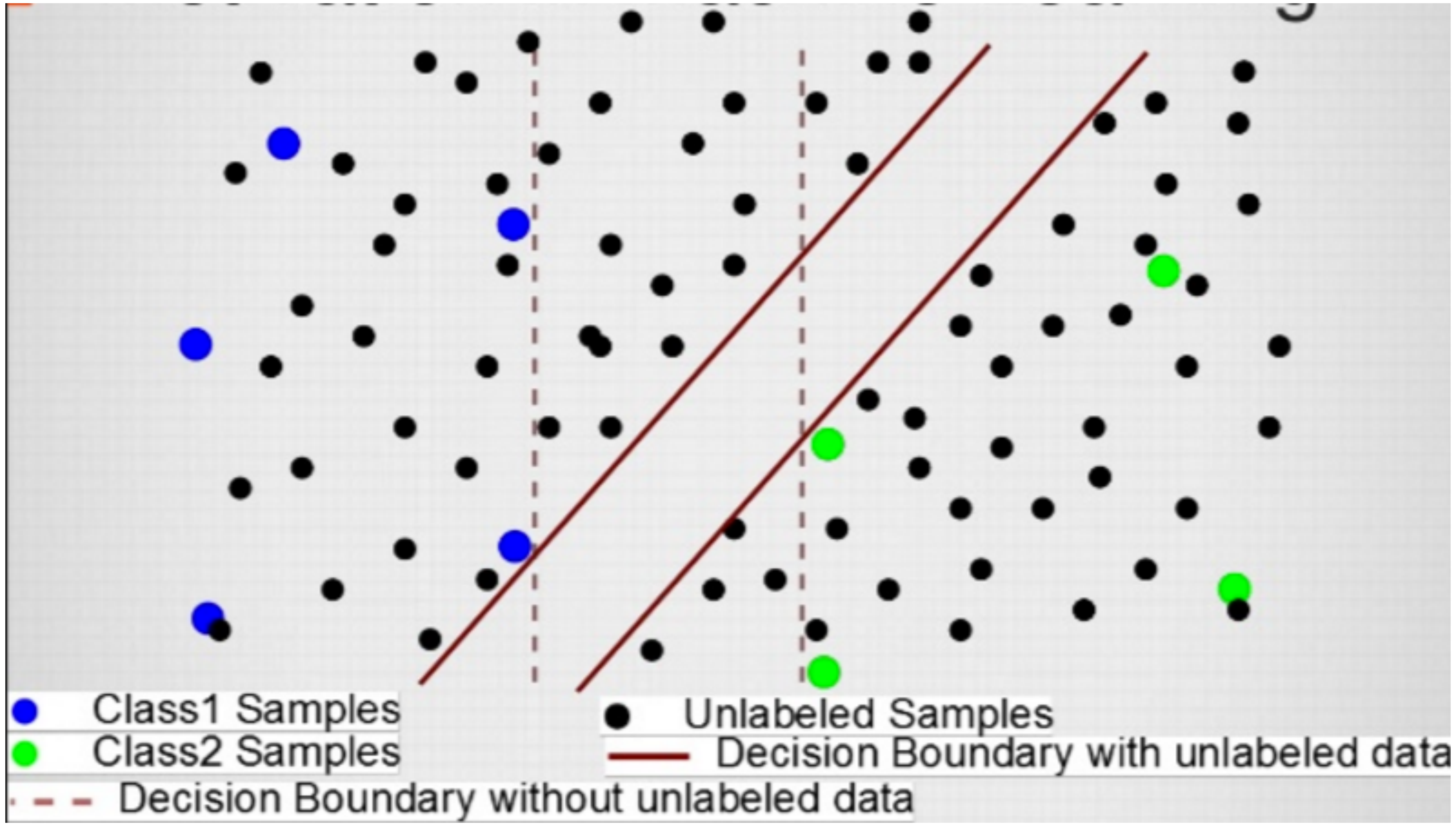
Gaussian Mixture Model (GMM) for Background Subtraction

http://www.codeproject.com/Articles/142859/Extended-GMM-for-Background-Subtraction-on-GPU

Semi-supervised Learning

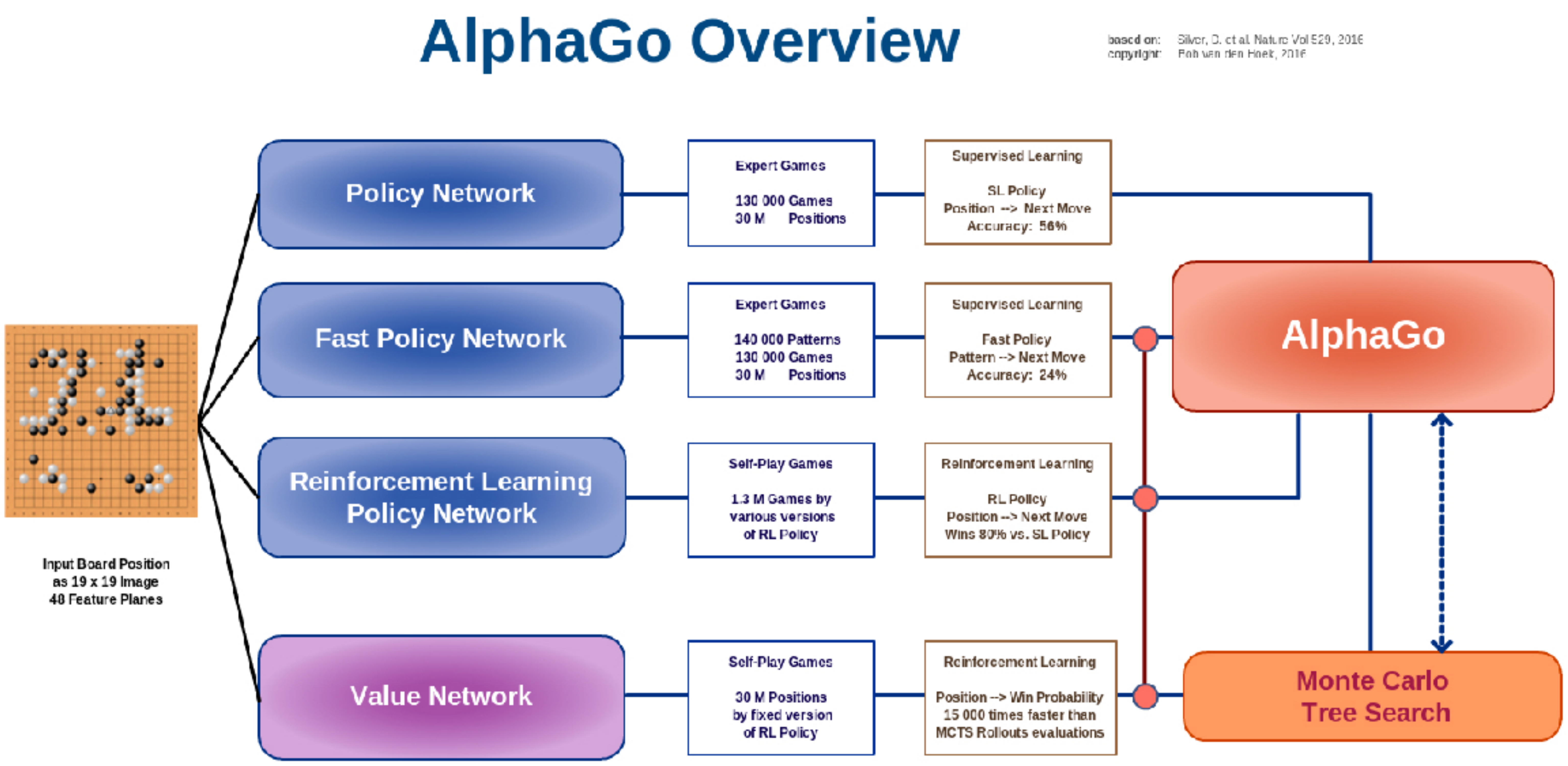


https://en.wikipedia.org/wiki/Semi-supervised_learning

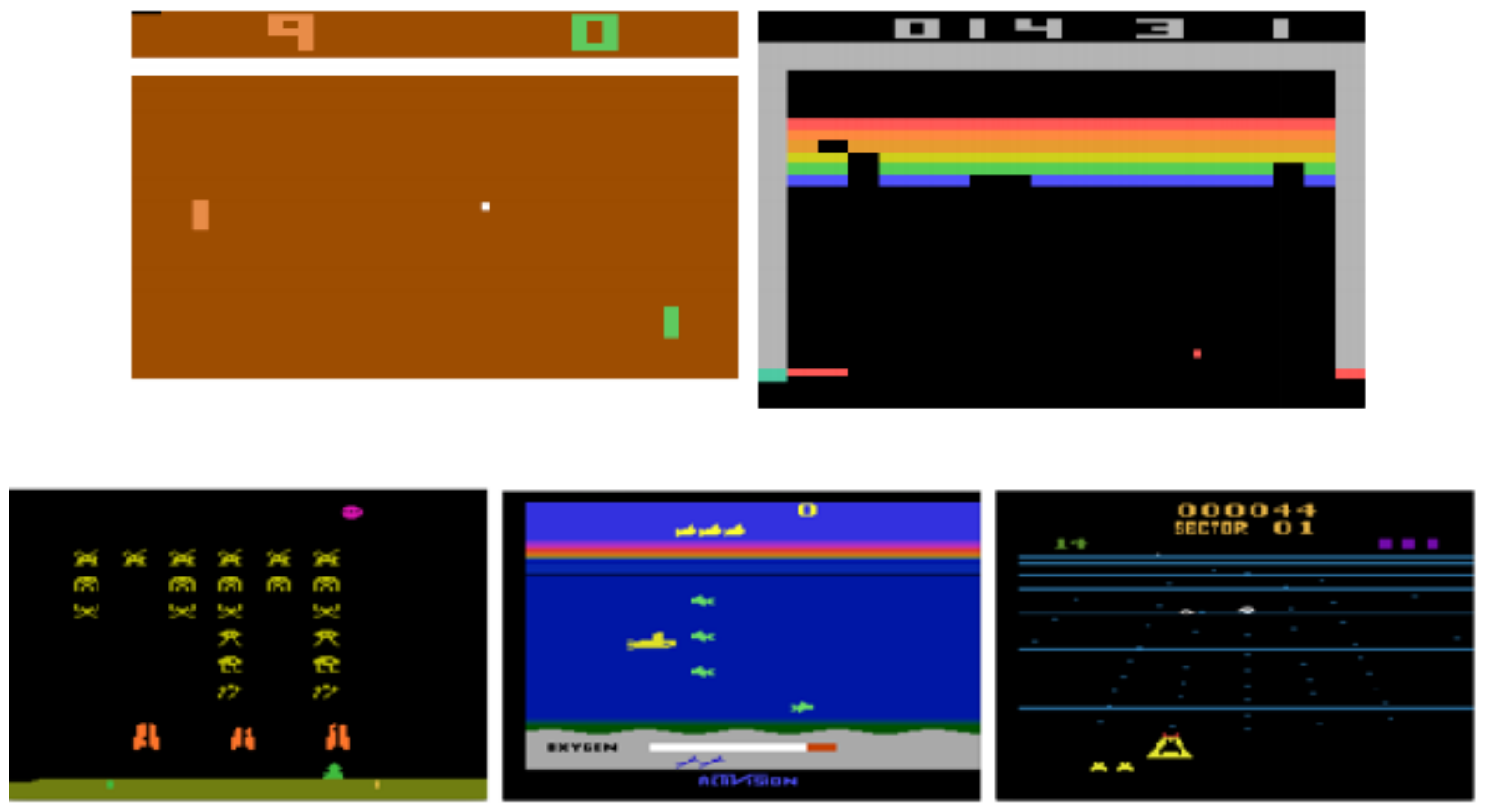


Lukas Tencer: http://www.slideshare.net/lukastencer/semisupervised-learning-42075774

Reinforcement Learning



http://deeplearningskysthelimit.blogspot.com/2016/04/part-2-alphago-under-magnifying-glass.html



Playing Atari with Deep Reinforcement Learning

A toolkit of RL: coding to play games like Pong.
https://gym.openai.com/

http://karpathy.github.io/2016/05/31/rl/

Lab-R
Lab-Matlab

Lab of R

https://github.com/ujjwalkarn/DataScienceR