Statistical methods for Machine Learning

Lecture 1: Introduction
4.2 2014

Aasa Feragen
aasa@diku.dk

Teachers and Instructors

- Christian Igel (igel@diku.dk), **course responsible**
- Aasa Feragen (aasa@diku.dk)

Instructors

- Asja Fischer (asja.fischer@ini.ruhr-uni-bochum.de)
- Oswin Krause (oswin.krause@diku.dk)
- Niklas Kasenburg (niklas.kasenburg@diku.dk)
- Kristoffer Stensbo-Smidt (k.stensbo@diku.dk)
- Pengfei Diao (diao@di.ku.dk)

Learning goals

Knowledge of

- the **general principles** of machine learning
- basic **probability theory** for modeling and analyzing data
- the theoretical concepts underlying **classification, regression, and clustering**
- the **mathematical foundations** of **selected machine learning algorithms**
- **common pitfalls** in machine learning

Skills in

- applying **linear and non-linear techniques for classification and regression**
- performing elementary **dimensionality reduction**
- elementary data **clustering**
- **implementing** selected machine learning algorithms
- **visualizing** and **evaluating** results obtained with machine learning techniques
- using **software libraries** for solving machine learning problems
- identifying and handling common **pitfalls** in machine learning

Competences in

- recognizing and describing possible **applications** of machine learning
- **comparing, appraising and selecting** machine learning methods of for specific tasks
- solving **real-world data mining and pattern recognition problems** by using machine learning techniques

We assume that you know

- Basic mathematical analysis (high school level and DiMS or MatIntro) and linear algebra (vectors and matrices)
- Take home exam 1 has a math brush-up quiz – use it as a guide!
- Probability theory at high school level
- Programming at an introductory level (we will use either Matlab, R, Python, or C/C++ - it is up to you)

Be aware:

- You are a mixed crowd with different backgrounds!
- There might be parts you find trivial and other parts you won't.
- Use the TAs, the lecturers, the forum!

Form

- Lectures:
  - Tuesday 10:15 - 12:00, Room: DIKU Aud. 4.1.22
  - Thursday 13:15 - 15:00, Room: DIKU Aud. 4.1.22
- Exercise classes:
  - Thursday 9:15 - 12:00, Rooms:
    - Class 1: DIKU-NC 1.0.04
    - Class 2: DIKU-NC 3.1.25
    - Class 3: DIKU-NC 1.0.37
    - Class 4: DIKU-NC 1.0.26
    - Class 5: DIKU-NC 1.0.10
- You have been assigned to one of these exercise classes (you can see which in Absalon).

Format of exercise classes

- The teaching assistent will lead a general discussion of the current lectures and assignment as well as provide general feedback on finished assignments (approx. 1 hour)
- You can also get individual help with the assignments while you work on them (approx. 2 hours)
- Bring your laptop!
- The exercise rooms have no computer terminals.

Mandatory assignments

3 mandatory assignments:

- Mix of theoretical and practical problems
- Two weeks to solve each of them
- Solutions can be made individually or in groups of no more than 3 participants
- Help from the TAs at the exercise class
- Feedback at exercise class
- Use the discussion forum!

How do I pass this course?

- Must pass the 3 mandatory assignments to be eligible for participating in the exam
- If you do not pass an assignment the first time you will be given a second chance to submit a new solution (assuming that you have made a SERIOUS attempt at every exercise the first time).
- **Exam assignment:** Larger written assignment similar to the other mandatory assignments
- This assignment must be solved individually, but we encourage you to discuss it with your fellow students.
- Final grading for the course is: 7-point grading based on the exam assignment only.

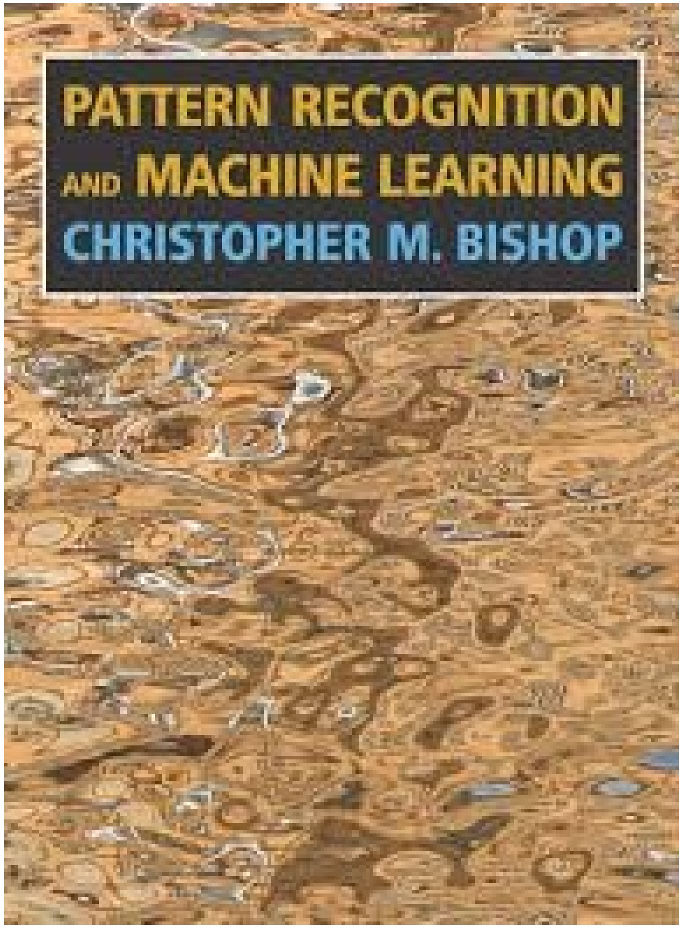
How much time should I spend on this course?

- KU expects 20 hours / week for a 7.5 ECTS course, 40 hours/wk for full time study

  (yes, it is more than the 37.5 hours/wk common out in real life, i.e. according to Danish union agreements)

- **How should I spend my time?**
  - Lectures and exercise class = 2 + 2 + 3 = 7 hours/wk
  - Reading and assignment = 20 – 7 = 13 hours/wk
- **Recommended**:
  - Prepare by reading the current week's material prior to each lecture (approx. 6 hours/wk). Be prepared for the exercise classes – come with questions!
  - Work on the assignment at home (approx. 7 hours/wk - and you will spend 2-3 hours on the assignment in exercise class)

Course material



Challenge:

- If you find the book not sufficiently mathematical, write out the proofs yourself.
- If you find the book too mathematical, draw figures to understand what the math describes.

Course Home Page and Information

- Through Absalon (access via your KUnet account)
  - Updated course information
  - Updated lecture and exercise schedules
  - Links to lecture slides (after the lecture)
  - Exercise material
  - Course material (reading material)
  - Links to additional material (reading, programming, etc.)
  - A discussion forum for course related topics

Tentative lecture schedule

4/2	AF	Introduction; starting probability theory and estimation
6/2	AF	Probability theory and estimation; Bayes theorem
11/2	CI	Ingredients of statistical learning theory (loss, risk minimization, bounds)
13/2	CI	Linear classification (LDA, perceptron, margin bounds)
18/2	AF	Linear models for regression I
20/2	AF	Linear models for regression II
25/2	CI	Neural networks (MLPs)
27/2	CI	Kernel methods I (RKHS, kernel NN, representer theorem, regularization networks)
4/3	CI	Kernel methods II (SVMs)
6/3	CI	Unsupervised Learning and Clustering
11/3	AF	Principal Component Analysis
13/3	AF	Visualization (MDS/PCA, Isomap, etc.)
18/3	CI	Trees and Forests
20/3	CI	Basics of learning theory; Course evaluation and questions about the exam

Exercises

4/2 - 18/2	Foundations of statistical machine learning
8/2 - 4/3	Basic supervised learning algorithms
4/3 - 18/3	Neural networks and support vector machines
18/3 - 3/4	Exam assignment

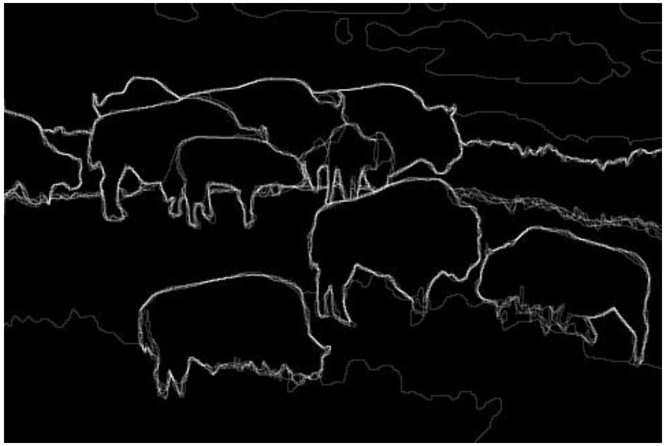
Any questions to the course setup?

Let's get started!

After today's lecture you should

- Be familiar with the types of questions answered and the types of problems solved with Machine Learning and Pattern Recognition techniques
- Recall basic probability theory:
    - discrete and continuous distributions
    - probability mass/density functions
    - Gaussian distributions

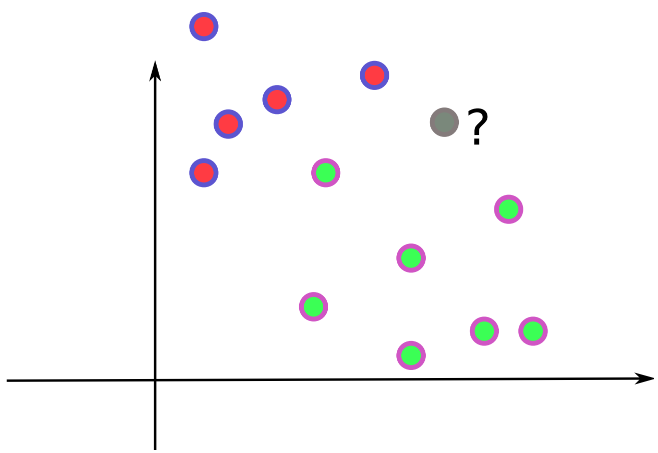
Machine Learning/Data Mining/Pattern Recognition



- **Example 1: Image segmentation**
  - Split the image into “objects” (foreground) and “irrelevant” (background).
- Classification of voxels x into classes:
  - y(x) = 1 (foreground)
  - y(x) = 0 (background)

Images from Berkeley Segmentation Dataset and Benchmark,
https://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/
D. Martin and C. Fowlkes and D. Tal and J. Malik, ICCV 2001

Machine Learning/Data Mining/Pattern Recognition



Classification splits data x into a finite number of classes:

- y(x) = 1 (foreground)
- y(x) = 0 (background)

General goal of ML:

- Model a mapping (rule) between data x and some abstract description y(x) of the data.

Supervised learning

- We know the rule y for a set of data (the training set) and try to learn a general rule y

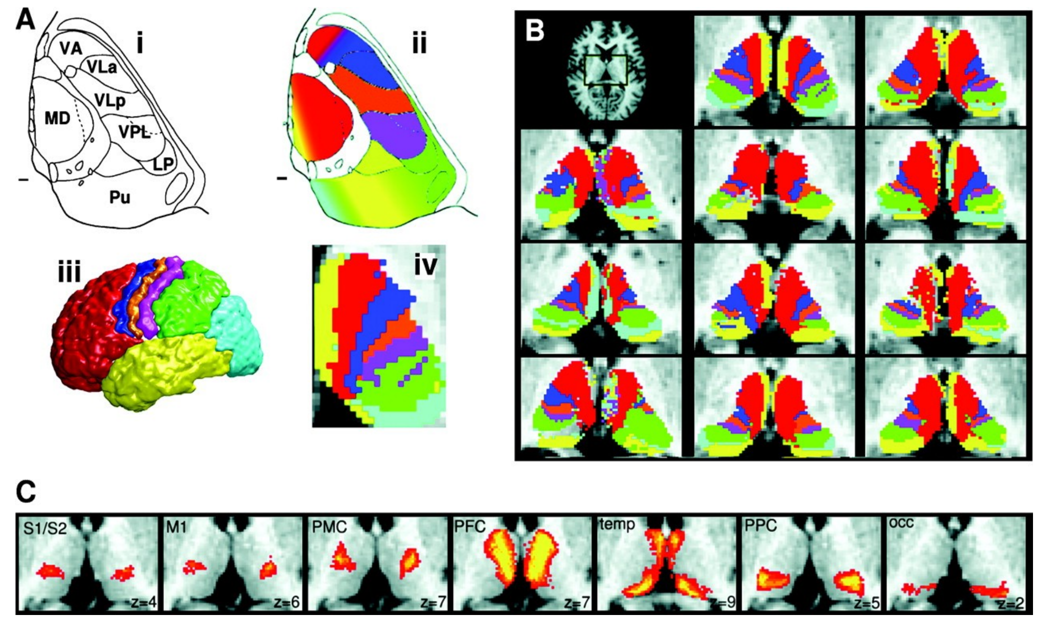
Machine Learning/Data Mining/Pattern Recognition

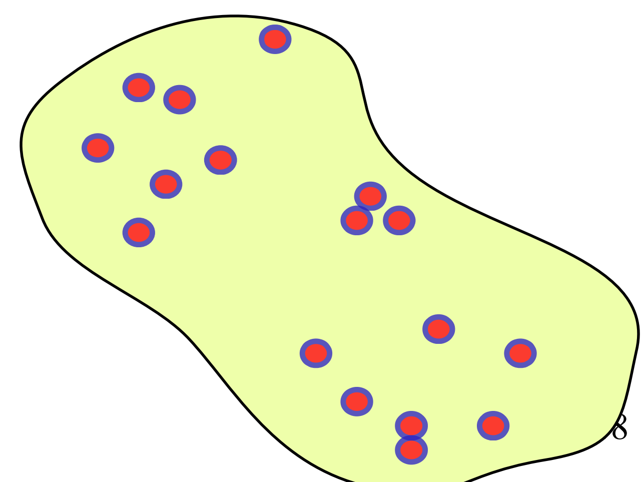
- **Example 2: Stock market prediction**
- Regression
  - $y(x)$ = stock price
  - Predicting a continuous variable
- Also a case of **supervised** learning

Images from Berkeley Segmentation Dataset and Benchmark, https://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/
D. Martin and C. Fowlkes and D. Tal and J. Malik, ICCV 2001

Machine Learning and Pattern Recognition?

- **Example 3: Clustering**
  - Cluster brain MRI voxels with respect to connectivity
- Example of **unsupervised learning:**
  - No known values y(x)
  - Don't know which clusters we are looking for

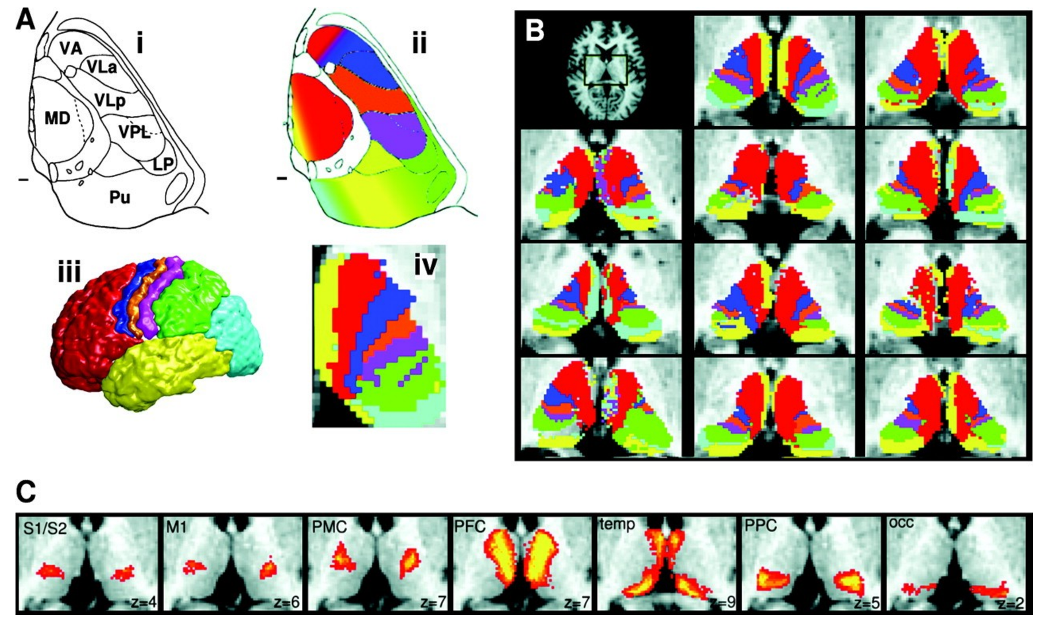


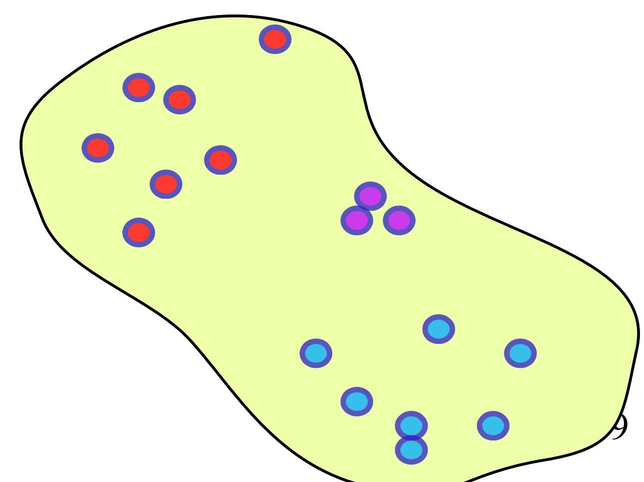


Johansen-Berg et al, Cerebral Cortex 2005

Machine Learning and Pattern Recognition?

- **Example 3: Clustering**
  - Cluster brain MRI voxels with respect to connectivity
- Example of **unsupervised learning:**
  - No known values y(x)
  - Don't know which clusters we are looking for

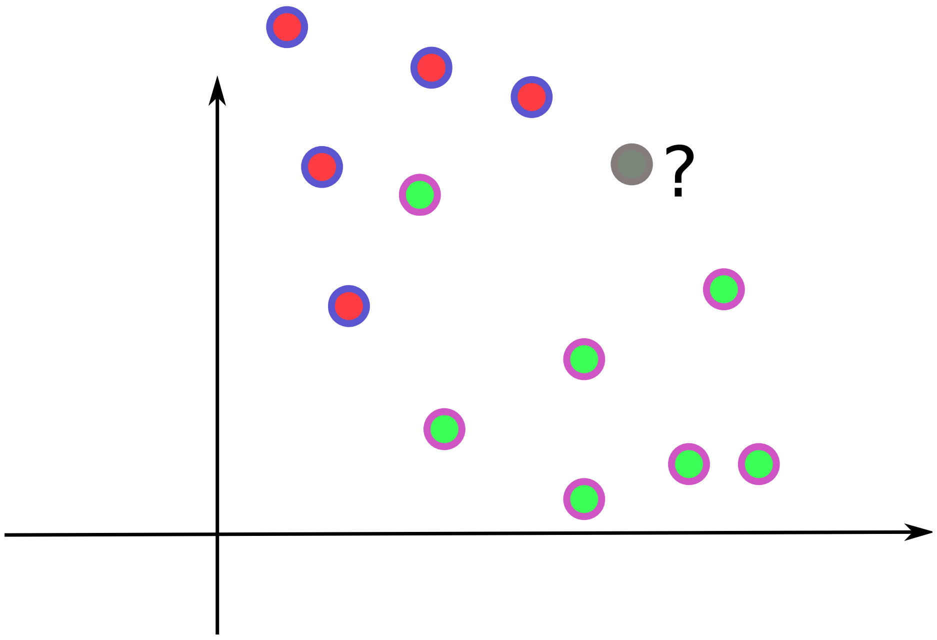




Johansen-Berg et al, Cerebral Cortex 2005

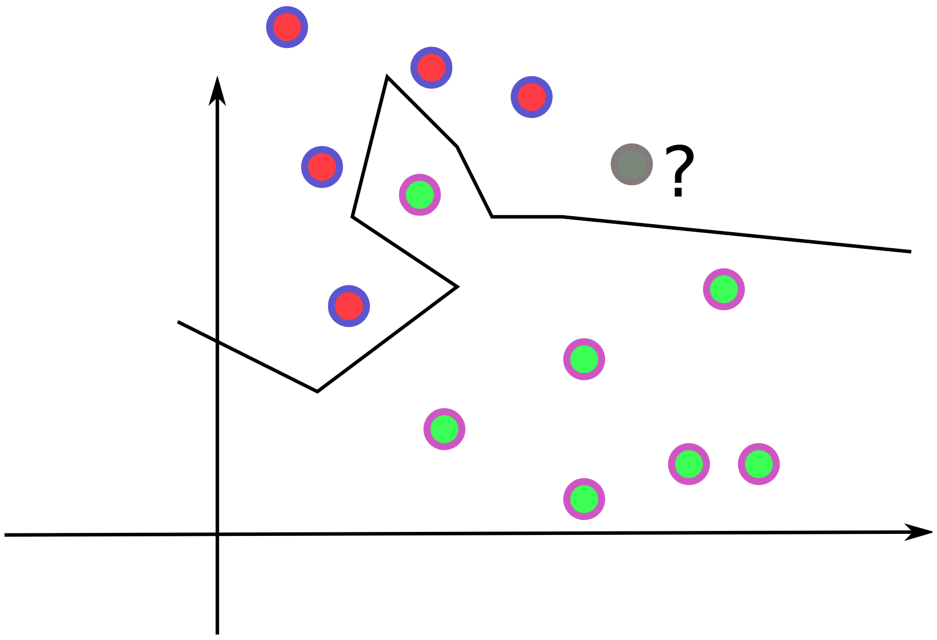
Generalizability

- Make sure the model y(x) generalizes to new unseen data (the test set).



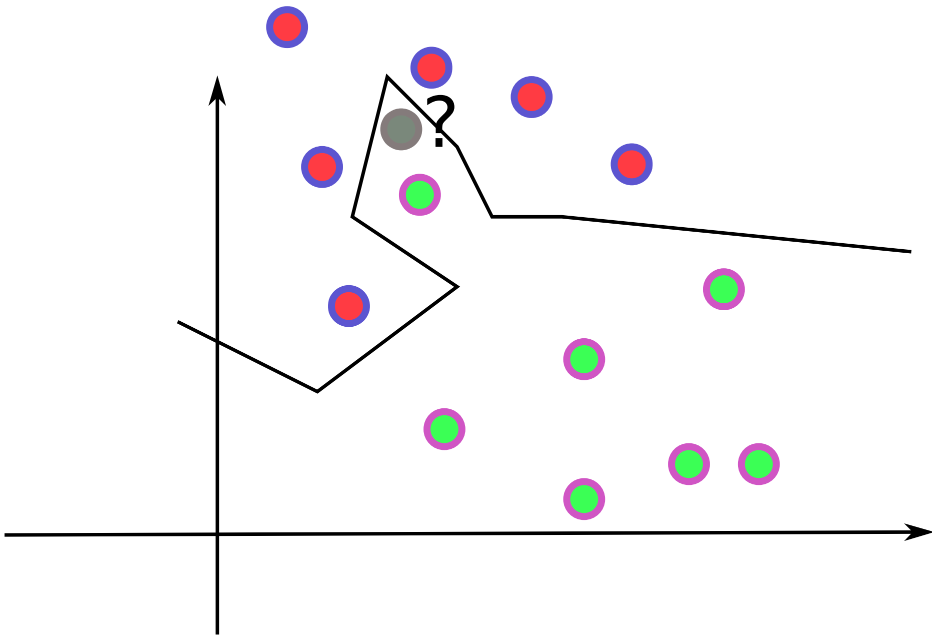
Generalizability

- Make sure the model y(x) generalizes to new unseen data (the test set).



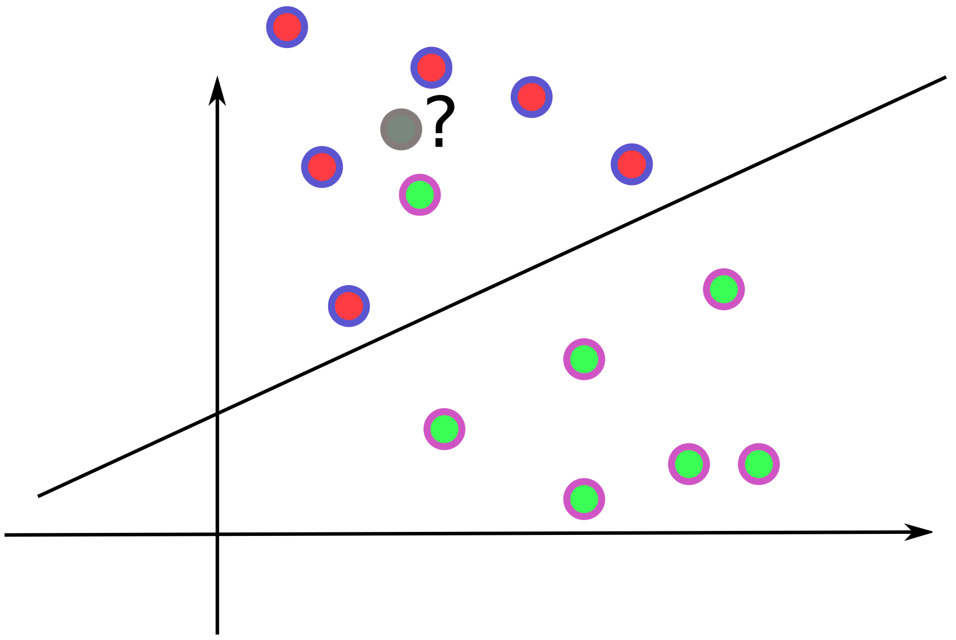
Generalizability

- Make sure the model y(x) generalizes to new unseen data (the test set).



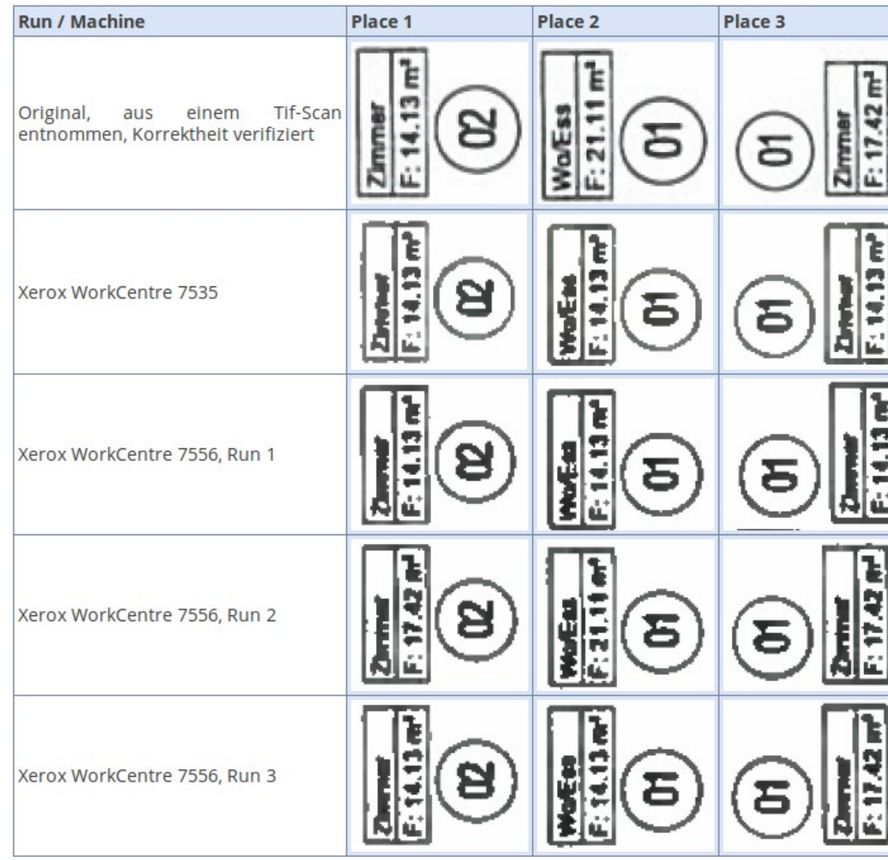
Generalizability

- Make sure the model y(x) generalizes to new unseen data (the test set).



Generalizability

- Make sure the model y(x) generalizes to new unseen data (the test set).
- **Example 3: Xerox**
  - **July 2013:** Xerox scanners were found to mangle numbers in documents
  - **Cause:** JBIG2 compression algorithm replacing image patches by “similar” image patches from a database
  - Model for “similar” did not generalize
  - Unexpected impact of ML: legal documents scanned, etc...



http://www.dkriesel.com/en/blog/2013/0802_xerox-workcentres_are_switching_written_numbers_when_scanning

Summary of ML principles

- General ML task:

  Learn rule y(x) which predicts a target t from measured data x

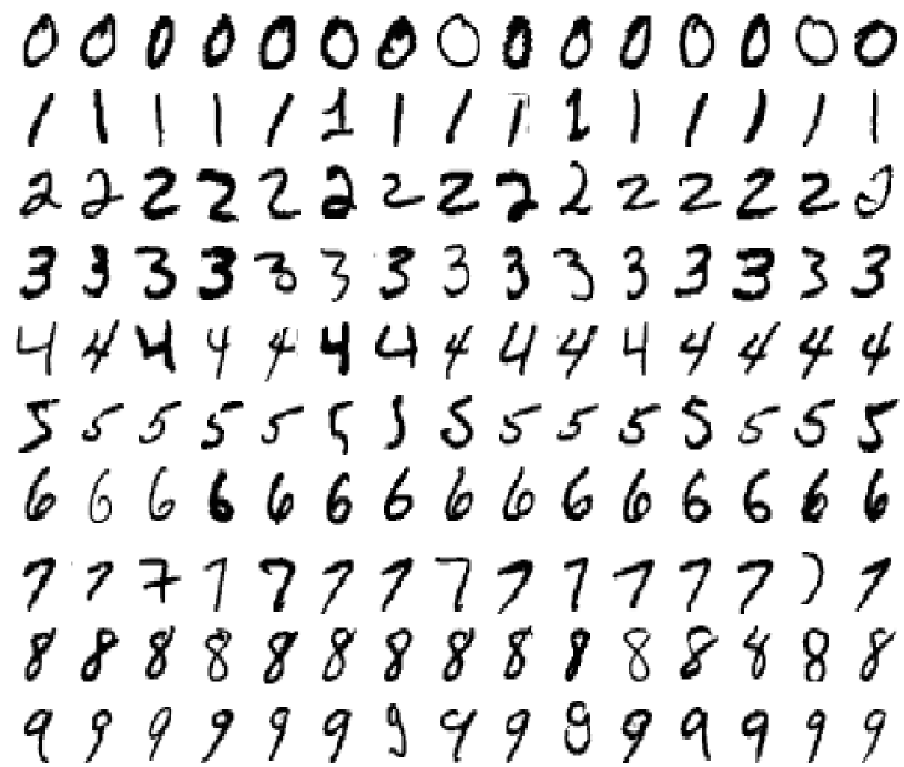
- **Unsupervised learning:**
  - No examples of y(x)
  - Examples:
    - Clustering

- **Supervised learning:**
  - Have a set of examples x for which y(x) is known (training set)
  - Learn a function y from the training set
  - Check generalizability to test set
  - Examples:
    - Classification – discrete target t
    - Regression – continuous target t

Why *Statistical* Machine Learning

- Often the data we are modeling:
  - Have too large variability and/or complexity to be described by deterministic rules

- Example: Variability in handwriting

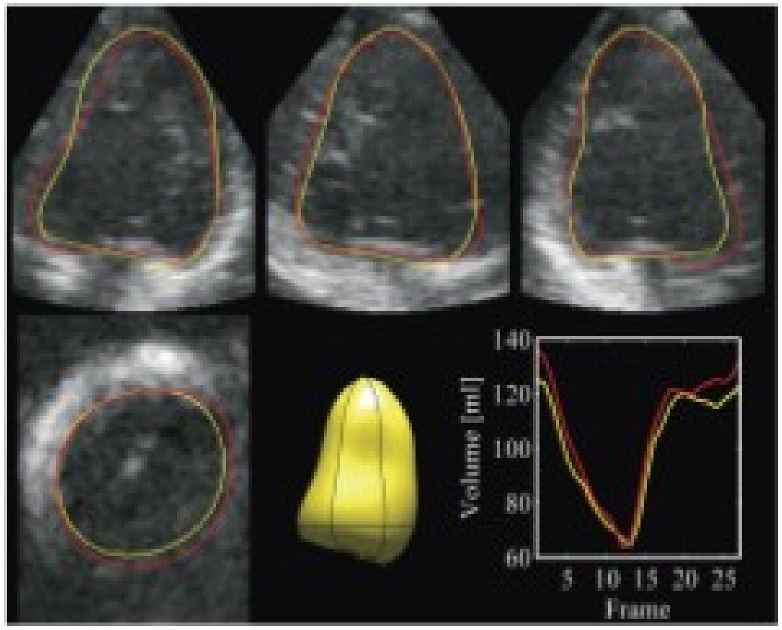


http://yann.lecun.com/exdb/mnist/index.html

Why *Statistical* Machine Learning

- Often the data we are modeling:
  - Have too large variability and/or complexity to be described by deterministic rules
  - Are inherently stochastic

    (e.g. image projection angle)
  - Are noisy (e.g. caused by sensory noise)



http://www-prima.inrialpes.fr/FGnet/data/01-TalkingFace/talking_face.html

Why *Statistical* Machine Learning

- Often the data we are modeling:
  - Have too large variability and/or complexity to be described by deterministic rules
  - Are inherently stochastic

    (e.g. image projection angle)
  - Are noisy (e.g. caused by sensory noise)
- Hence a probabilistic description is most often needed.
- For probabilistic models we need to be able to represent and estimate probability distributions either:
  - Parametric
  - Non-parametric

Probability Theory 101

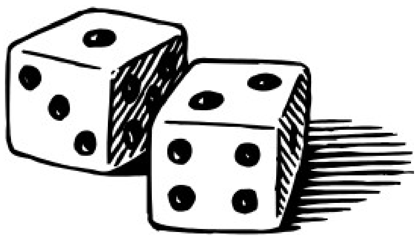
Probability theory and Estimation

Example: Throwing two dice X_1, X_2

Two random variables X, Y

$$X = X_1$$
$$Y = X_1 + X_2$$

Trial with N throws



$$X = x_i \in \{1, 2, \ldots, 6\}$$
$$Y = y_j \in \{2, 3, \ldots, 12\}$$

Probability theory and Estimation

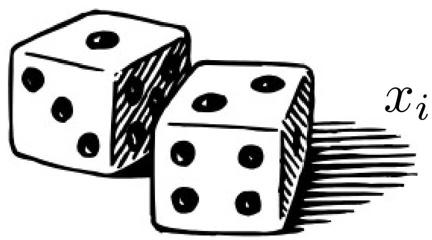
Example: Throwing two dice X_1, X_2

Two random variables X, Y

$$X = X_1$$
$$Y = X_1 + X_2$$

Trial with N throws

$$X = x_i \in \{1, 2, \ldots, 6\}$$
$$Y = y_j \in \{2, 3, \ldots, 12\}$$



$$n_{ij} = \sharp \text{ trials with } (X = x_i, Y = y_j)$$
$$c_i = \textstyle\sum_j n_{ij} = \sharp \text{ trials with } X = x_i$$
$$r_j = \textstyle\sum_i n_{ij} = \sharp \text{ trials with } Y = y_j$$

y_j

x_i	**2**	**3**	**4**	**5**	**...**	**9**	**10**	**11**	**12**
1									
2									
3					n_{ij}				
4									
5									
6									

Probability theory and Estimation

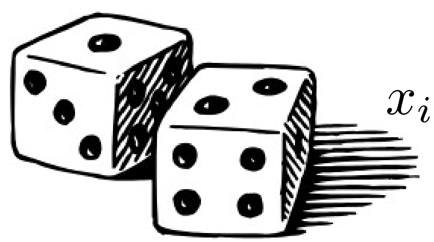
Example: Throwing two dice X_1, X_2

Two random variables X, Y

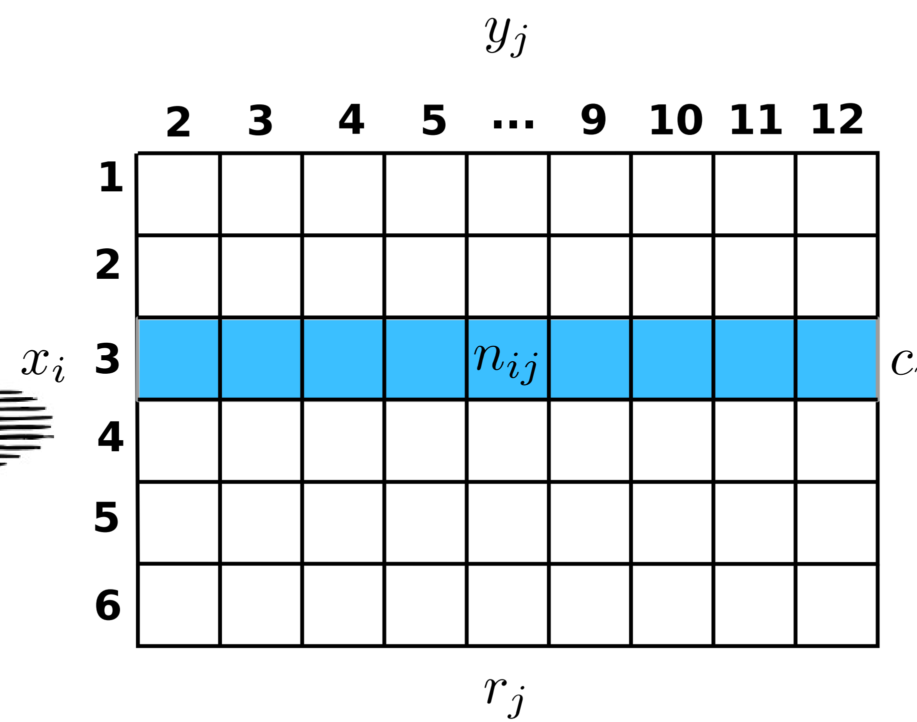
$$X = X_1$$
$$Y = X_1 + X_2$$

Trial with N throws

$$X = x_i \in \{1, 2, \ldots, 6\}$$
$$Y = y_j \in \{2, 3, \ldots, 12\}$$



$$n_{ij} = \sharp \text{ trials with } (X = x_i, Y = y_j)$$
$$c_i = \sum_j n_{ij} = \sharp \text{ trials with } X = x_i$$
$$r_j = \sum_i n_{ij} = \sharp \text{ trials with } Y = y_j$$



Probability theory and Estimation

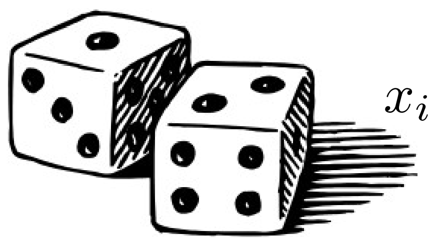
Example: Throwing two dice X_1, X_2

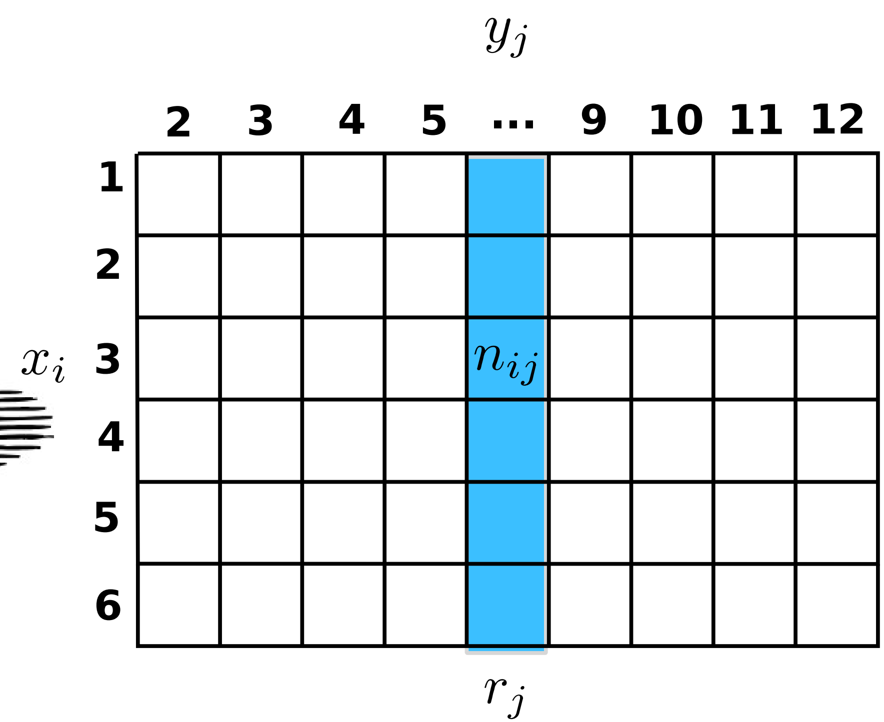
Two random variables X, Y

$$X = X_1$$
$$Y = X_1 + X_2$$

Trial with N throws

$$X = x_i \in \{1, 2, \ldots, 6\}$$
$$Y = y_j \in \{2, 3, \ldots, 12\}$$





$$n_{ij} = \sharp \text{ trials with } (X = x_i, Y = y_j)$$
$$c_i = \sum_j n_{ij} = \sharp \text{ trials with } X = x_i$$
$$r_j = \sum_i n_{ij} = \sharp \text{ trials with } Y = y_j$$

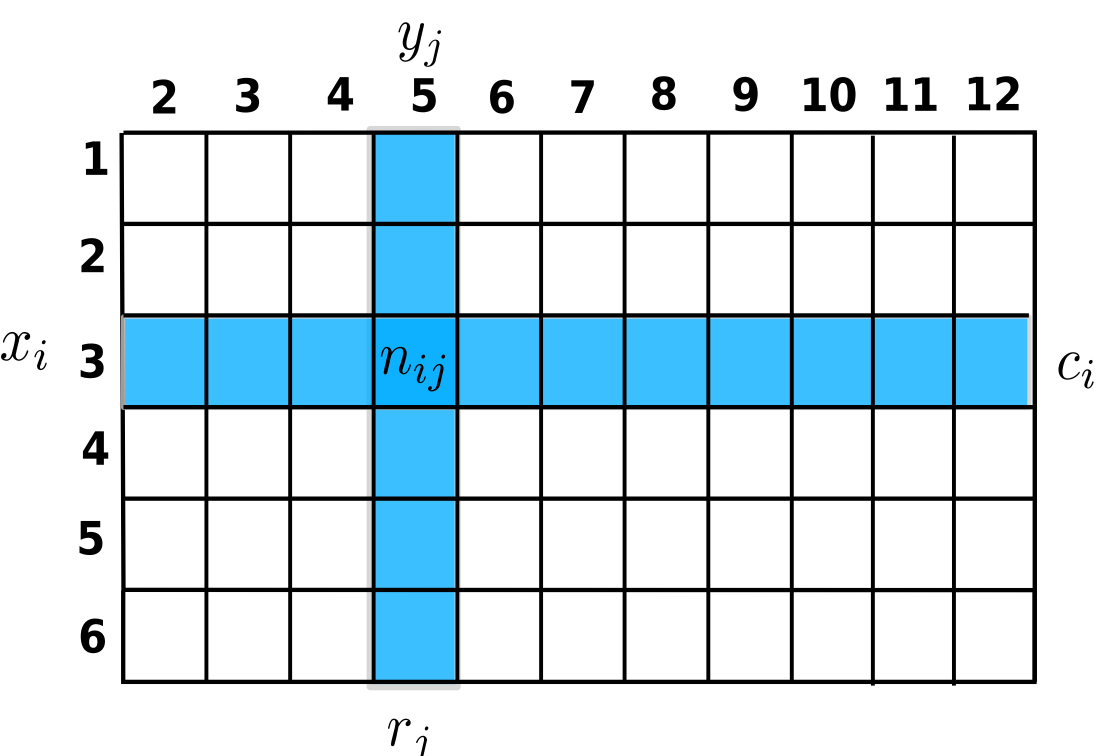
Probability theory and Estimation

Example: Throwing two dice X_1, X_2

Two random variables X, Y

$$X = X_1$$
$$Y = X_1 + X_2$$

Trial with N throws



Joint probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Marginal probability

$$p(X = x_i) = \frac{c_i}{N} \qquad p(Y = y_j) = \frac{r_j}{N}$$

Conditional probability

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$
$$p(X = x_i | Y = y_j) = \frac{n_{ij}}{r_j}$$

In the limit $N \to \infty$!

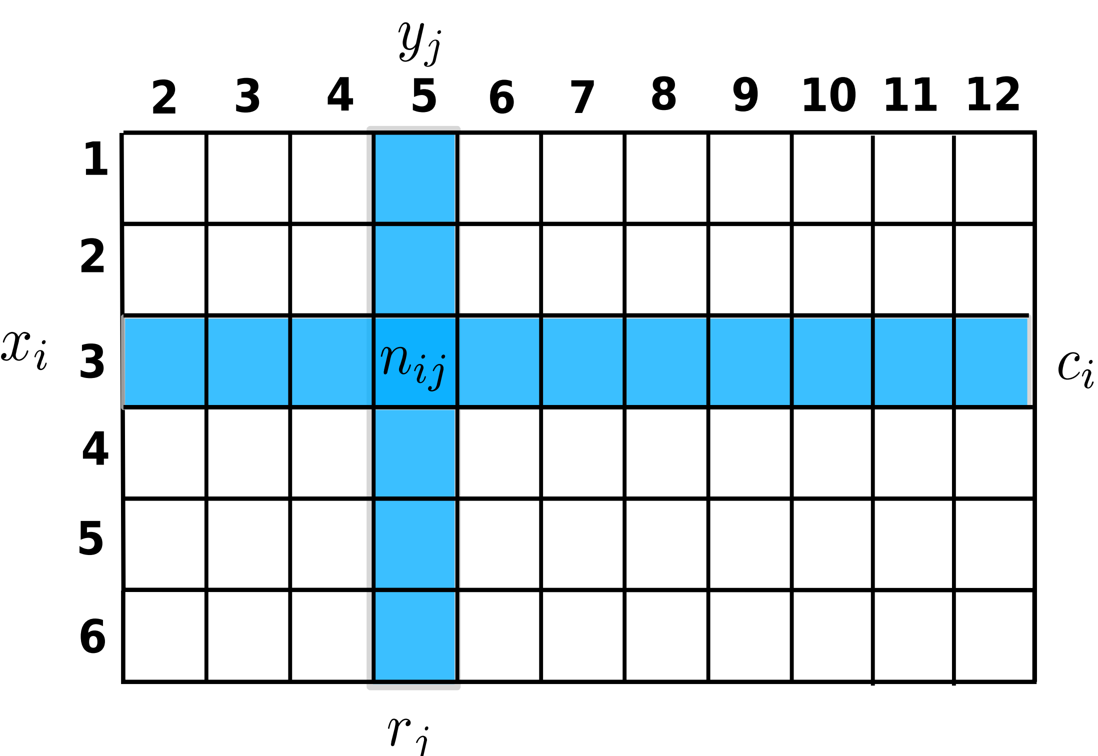
Probability theory and Estimation

Example: Throwing two dice X_1, X_2

Two random variables X, Y

$$X = X_1$$
$$Y = X_1 + X_2$$

Trial with N throws



Joint probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Marginal probability

$$p(X = x_i) = \frac{c_i}{N} \qquad p(Y = y_j) = \frac{r_j}{N}$$

Conditional probability

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$
$$p(X = x_i | Y = y_j) = \frac{n_{ij}}{r_j}$$

In the limit $N \to \infty$!

$\mathrm{p}(\mathrm{X} = \mathrm{i},\ \mathrm{Y} = \mathrm{j}) = ?$

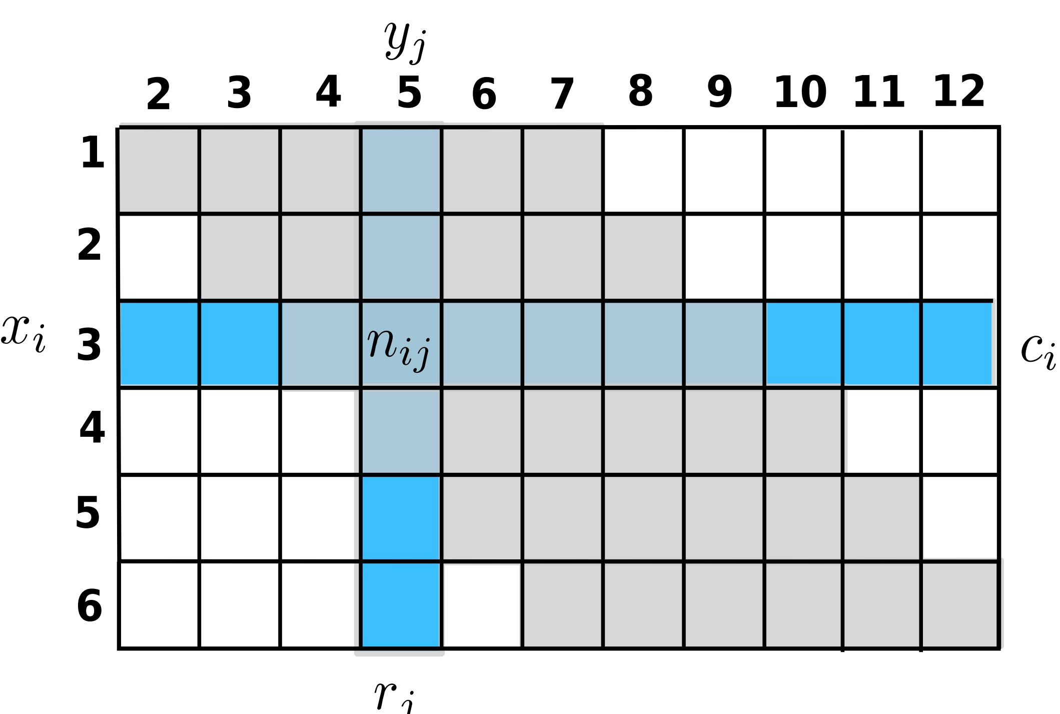
Probability theory and Estimation

Example: Throwing two dice X_1, X_2

Two random variables X, Y

$$X = X_1$$
$$Y = X_1 + X_2$$

Trial with N throws



Joint probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Marginal probability

$$p(X = x_i) = \frac{c_i}{N} \qquad p(Y = y_j) = \frac{r_j}{N}$$

Conditional probability

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$
$$p(X = x_i | Y = y_j) = \frac{n_{ij}}{r_j}$$

In the limit $N \to \infty$!

$\mathrm{p}(\mathrm{X} = \mathrm{i}, \mathrm{Y} = \mathrm{j}) = 1/36$ or 0

$\mathrm{p}(\mathrm{Y} = 3) = ?$

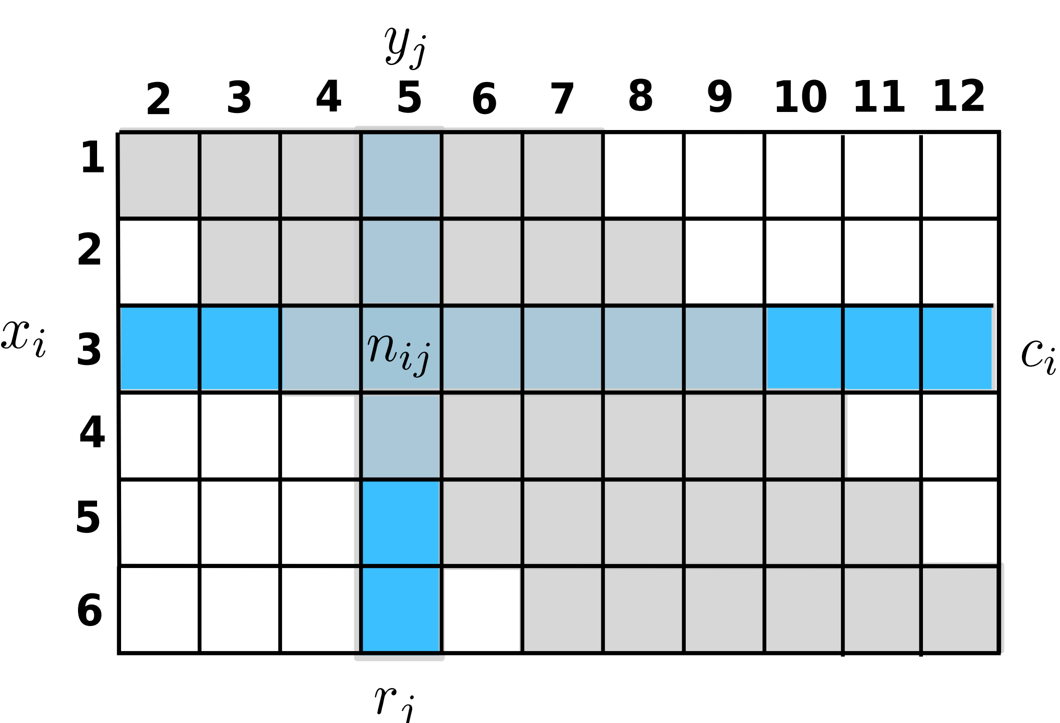
Probability theory and Estimation

Example: Throwing two dice X_1, X_2

Two random variables X, Y

$$X = X_1$$
$$Y = X_1 + X_2$$

Trial with N throws



Joint probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Marginal probability

$$p(X = x_i) = \frac{c_i}{N} \qquad p(Y = y_j) = \frac{r_j}{N}$$

Conditional probability

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$
$$p(X = x_i | Y = y_j) = \frac{n_{ij}}{r_j}$$

In the limit $N \to \infty$!

$$\mathrm{p(X = i,\ Y = j) = 1/36 \text{ or } 0}$$
$$\mathrm{p(Y = 3) = 2/36}$$
$$\mathrm{p(Y = 3\ |X = 2) = ?}$$

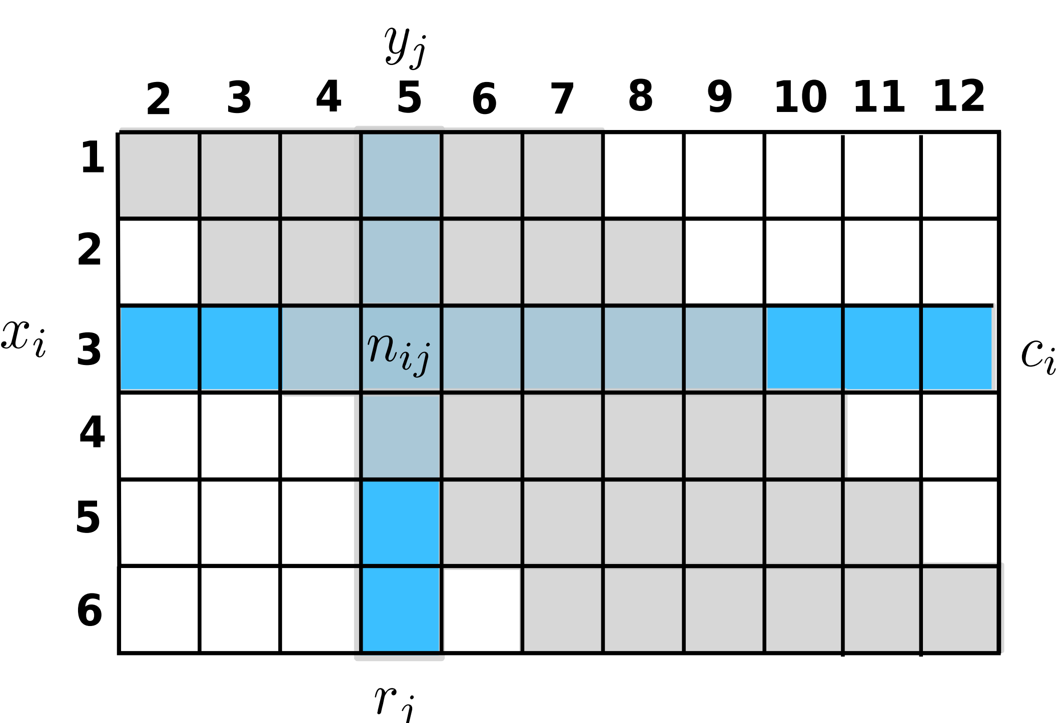
Probability theory and Estimation

Example: Throwing two dice X_1, X_2

Two random variables X, Y

$$X = X_1$$
$$Y = X_1 + X_2$$

Trial with N throws



Joint probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Marginal probability

$$p(X = x_i) = \frac{c_i}{N} \qquad p(Y = y_j) = \frac{r_j}{N}$$

Conditional probability

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$
$$p(X = x_i | Y = y_j) = \frac{n_{ij}}{r_j}$$

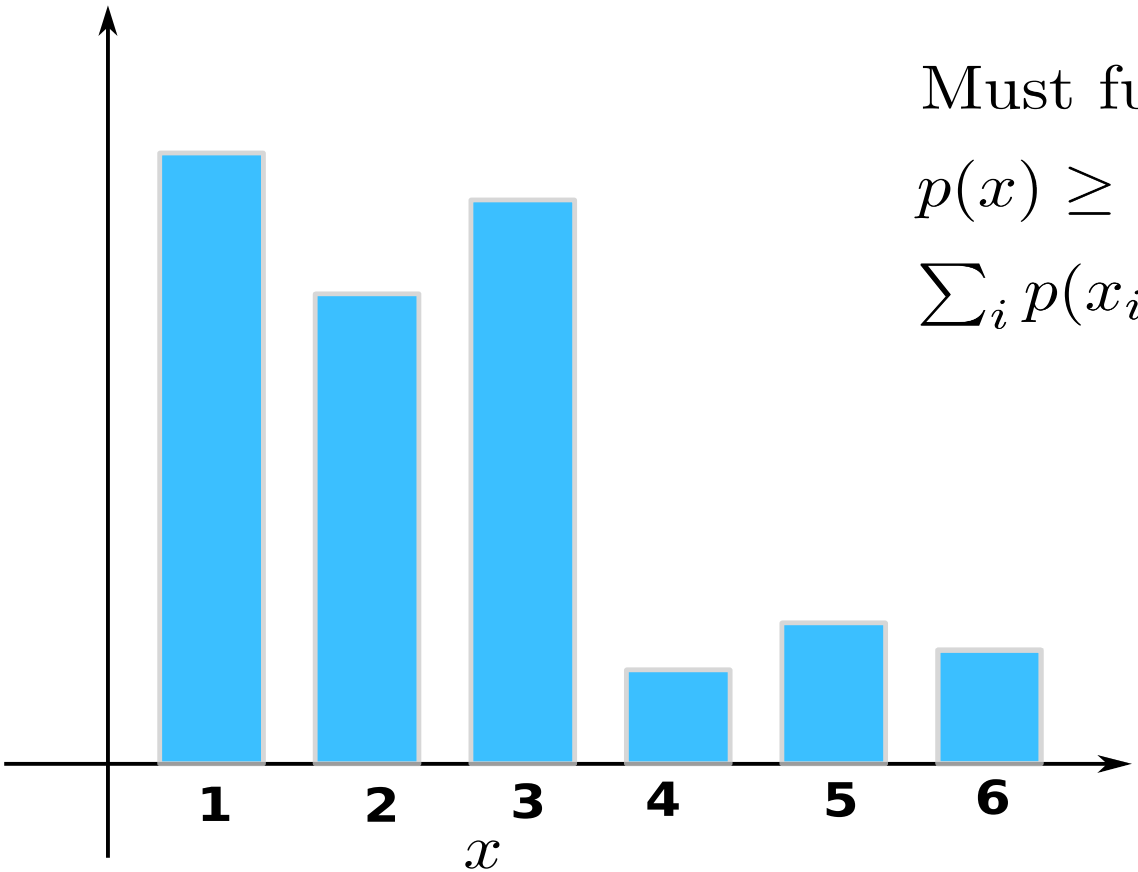
In the limit $N \to \infty$!

$$\mathrm{p(X = i,\ Y = j) = 1/36 \text{ or } 0}$$
$$\mathrm{p(Y = 3) = 2/36}$$
$$\mathrm{p(Y = 3\ |X = 2) = 1/6}$$

Probability mass function

Discrete random variables:

$p(x) = p(X = x)$ is called a *probability mass function*



Must fulfill

$p(x) \geq 0$ for all x

$\sum_i p(x_i) = 1$

Continuous random variables

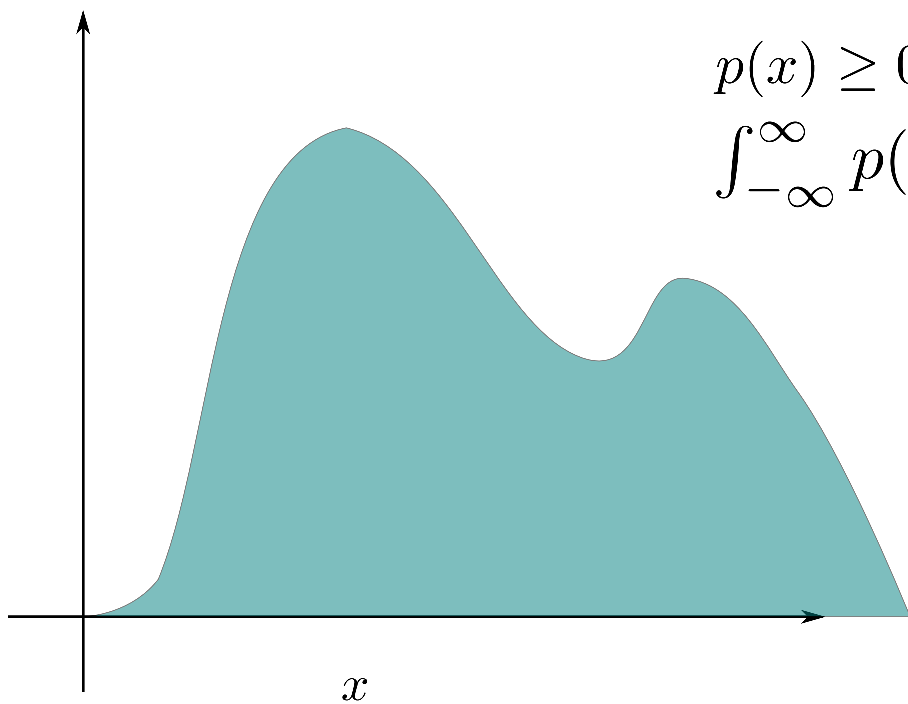
$x \in \mathbb{R}$ real random variable

$p \colon \mathbb{R} \to \mathbb{R}$

Must fulfill

$p(x) \geq 0$ for all x

$\int_{-\infty}^{\infty} p(x)dx = 1$



Continuous random variables

$x \in \mathbb{R}$ real random variable

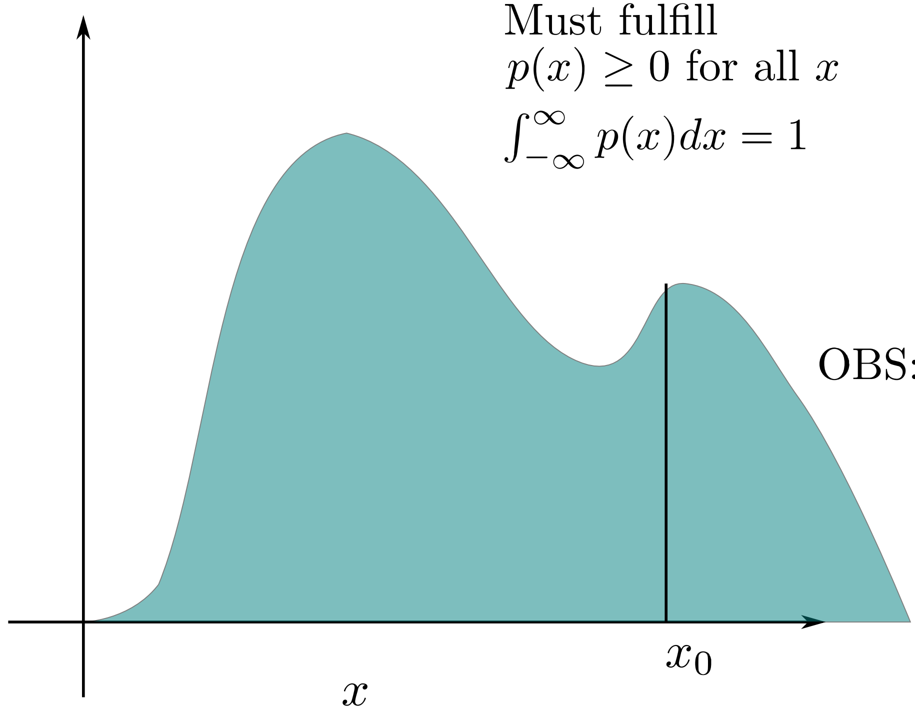
$p \colon \mathbb{R} \to \mathbb{R}$

$p(x)$ is the *probability density function* of X

Must fulfill

$p(x) \geq 0$ for all x

$\int_{-\infty}^{\infty} p(x)dx = 1$



OBS: $p(X = x_0) = 0$

Continuous random variables

$x \in \mathbb{R}$ real random variable

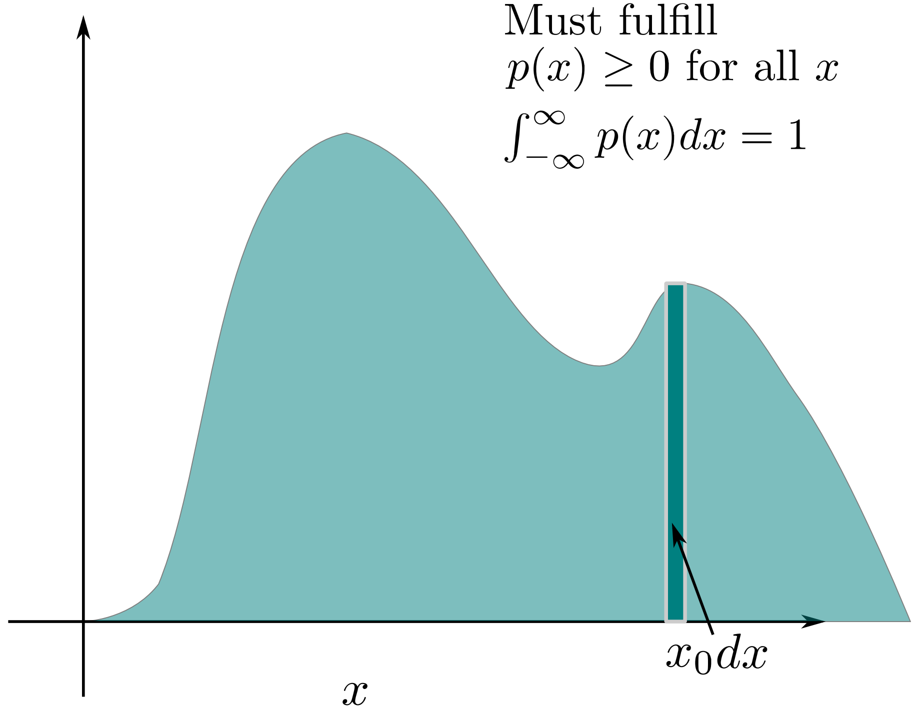
$p\colon \mathbb{R} \to \mathbb{R}$

$p(x)$ is the *probability density function* of X

Must fulfill

$p(x) \geq 0$ for all x

$\int_{-\infty}^{\infty} p(x)dx = 1$



Continuous random variables

$x \in \mathbb{R}$ real random variable

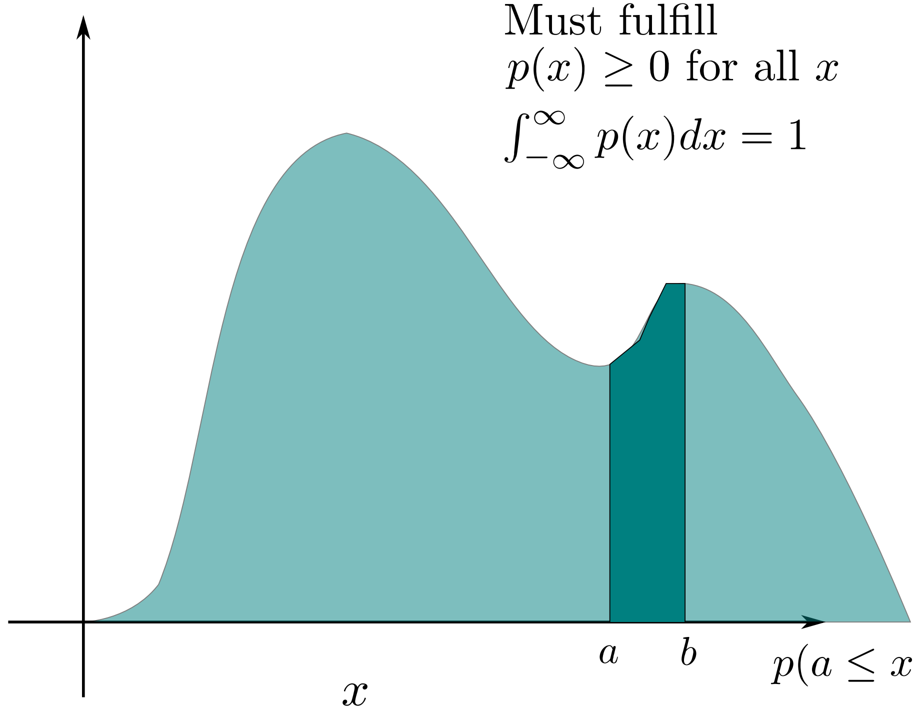
$p \colon \mathbb{R} \to \mathbb{R}$

$p(x)$ is the *probability density function* of X

Must fulfill

$p(x) \geq 0$ for all x

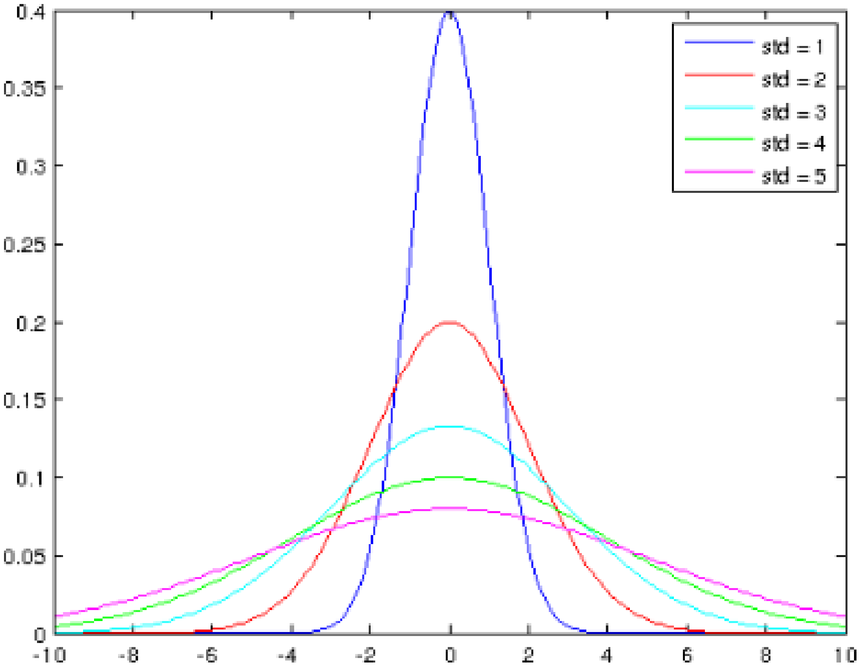
$\int_{-\infty}^{\infty} p(x)dx = 1$



$p(a \leq x \leq b) = \int_a^b p(x)dx$

Example: The Gaussian distribution (normal distribution)

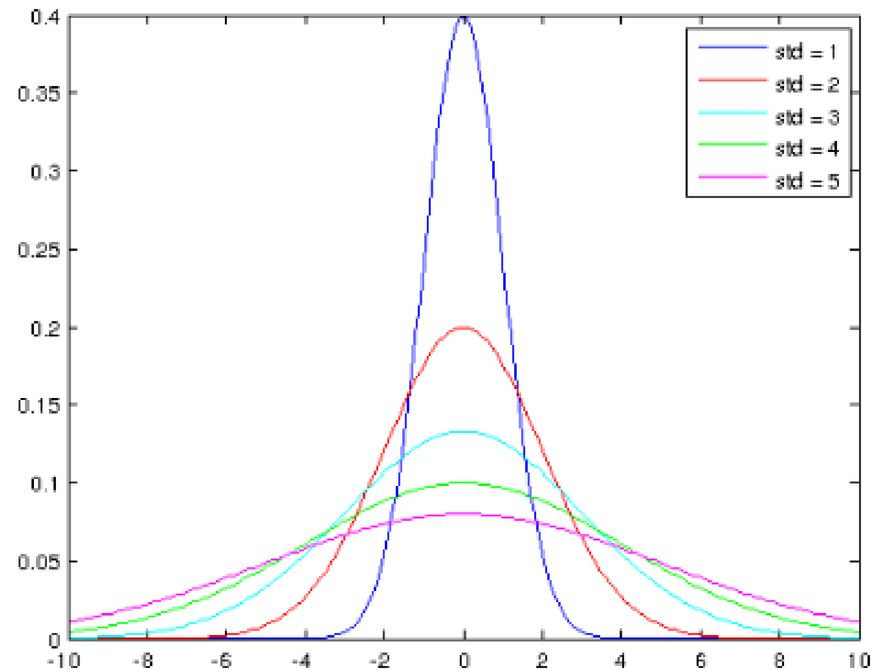
$$p(x) = \mathcal{N}(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)}$$



Example: The Gaussian distribution (normal distribution)

$$p(x) = \mathcal{N}(x|\mu,\sigma) = \underbrace{\frac{1}{\sqrt{2\pi}\sigma}}_{\text{normalize}} \underbrace{e^{-\frac{1}{2\sigma^2}(x-\mu)}}_{\text{PDF shape}}$$

$$\int_{-\infty}^{\infty} p(x)dx = 1$$



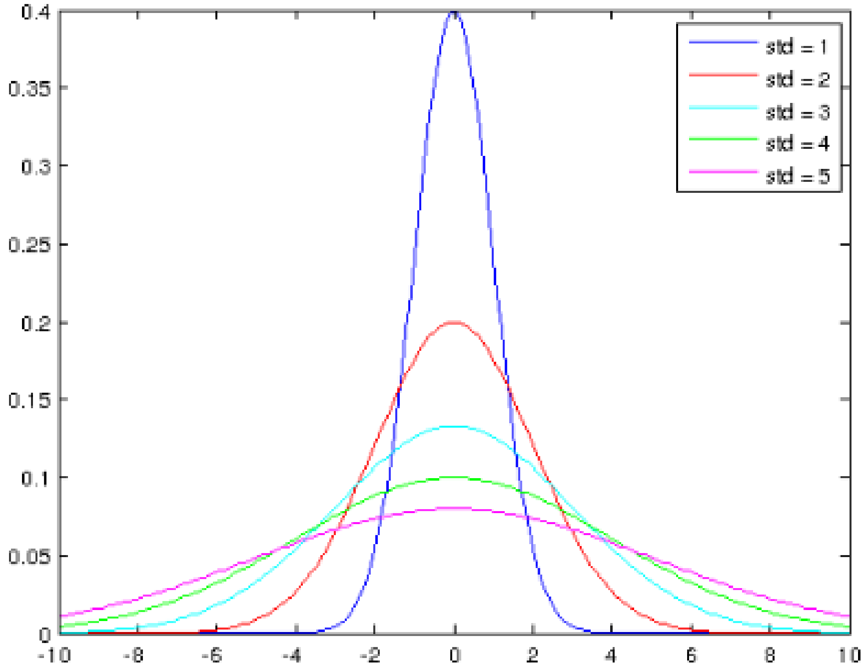
Example: The Gaussian distribution (normal distribution)

$$p(x) = \mathcal{N}(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

$$= Ce^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

bandwidth

maximum



Example: The Gaussian distribution (normal distribution)

$$p(x) = \mathcal{N}(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

$$= Ce^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

bandwidth

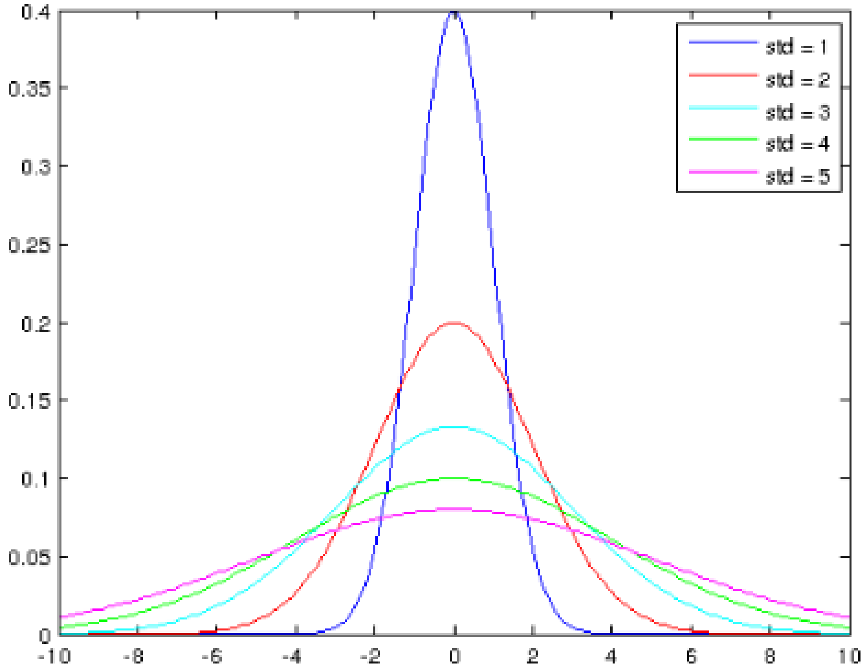
maximum

μ is mean

σ^2 is variance

σ is standard deviation

$\beta = \frac{1}{\sigma^2}$ is precision



Multivariate Gaussian distribution

$$\mathcal{N}(\mathbf{x}|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)}$$

covariance matrix

mean

Multivariate Gaussian distribution

$$\mathcal{N}(\mathbf{x}|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)}$$

The interesting part!

normalization
(boring)

Multivariate Gaussian distribution

$$\mathcal{N}(\mathbf{x}|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)}$$

The interesting part!

normalization
(boring)

Simplified:

$$\mathcal{N}(\mathbf{x}|\mu,\Sigma) = c_1 e^{-c_2\mathbf{x}^T\Sigma^{-1}\mathbf{x}}$$

Translated so that $\mu = \mathbf{0}$

Multivariate Gaussian distribution

$$\mathcal{N}(\mathbf{x}|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)}$$

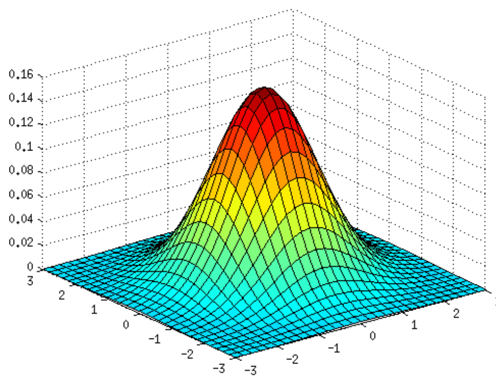
The interesting part!

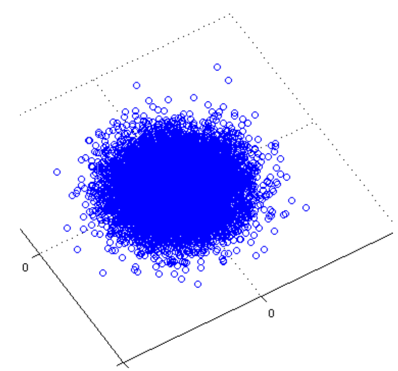
normalization
(boring)

Simplified:

$$\mathcal{N}(\mathbf{x}|\mu,\Sigma) = c_1 e^{-c_2\mathbf{x}^T\Sigma^{-1}\mathbf{x}}$$

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$





Multivariate Gaussian distribution

$$\mathcal{N}(\mathbf{x}|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)}$$

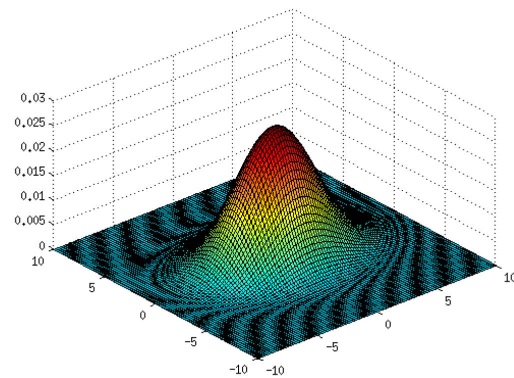
normalization
(boring)

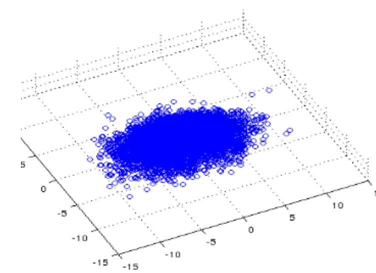
The interesting part!

Simplified:

$$\mathcal{N}(\mathbf{x}|\mu,\Sigma) = c_1 e^{-c_2\mathbf{x}^T\Sigma^{-1}\mathbf{x}}$$

$$\Sigma = \begin{pmatrix} 9 & 0 \\ 0 & 4 \end{pmatrix}$$





Multivariate Gaussian distribution

$$\mathcal{N}(\mathbf{x}|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)}$$

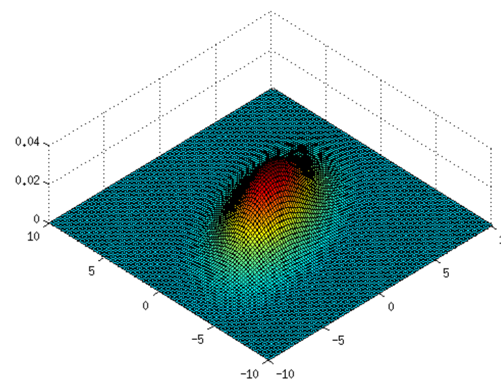
The interesting part!

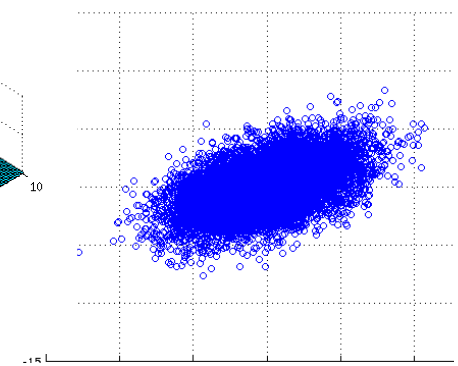
normalization
(boring)

Simplified:

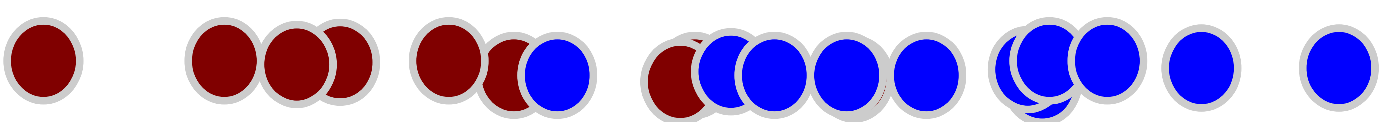
$$\mathcal{N}(\mathbf{x}|\mu,\Sigma) = c_1 e^{-c_2\mathbf{x}^T\Sigma^{-1}\mathbf{x}}$$

$$\Sigma = \begin{pmatrix} 9 & 3 \\ 3 & 4 \end{pmatrix}$$

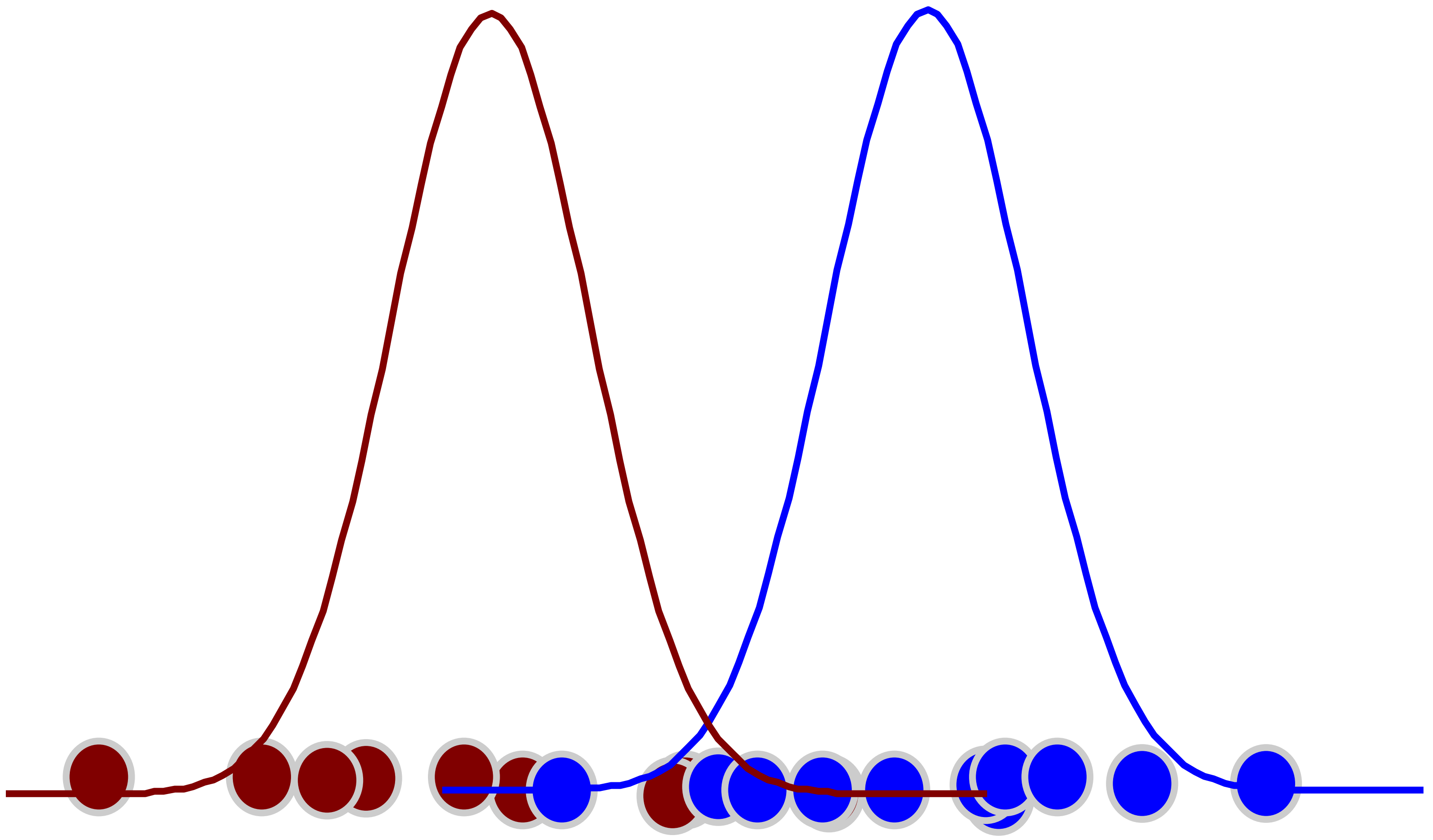




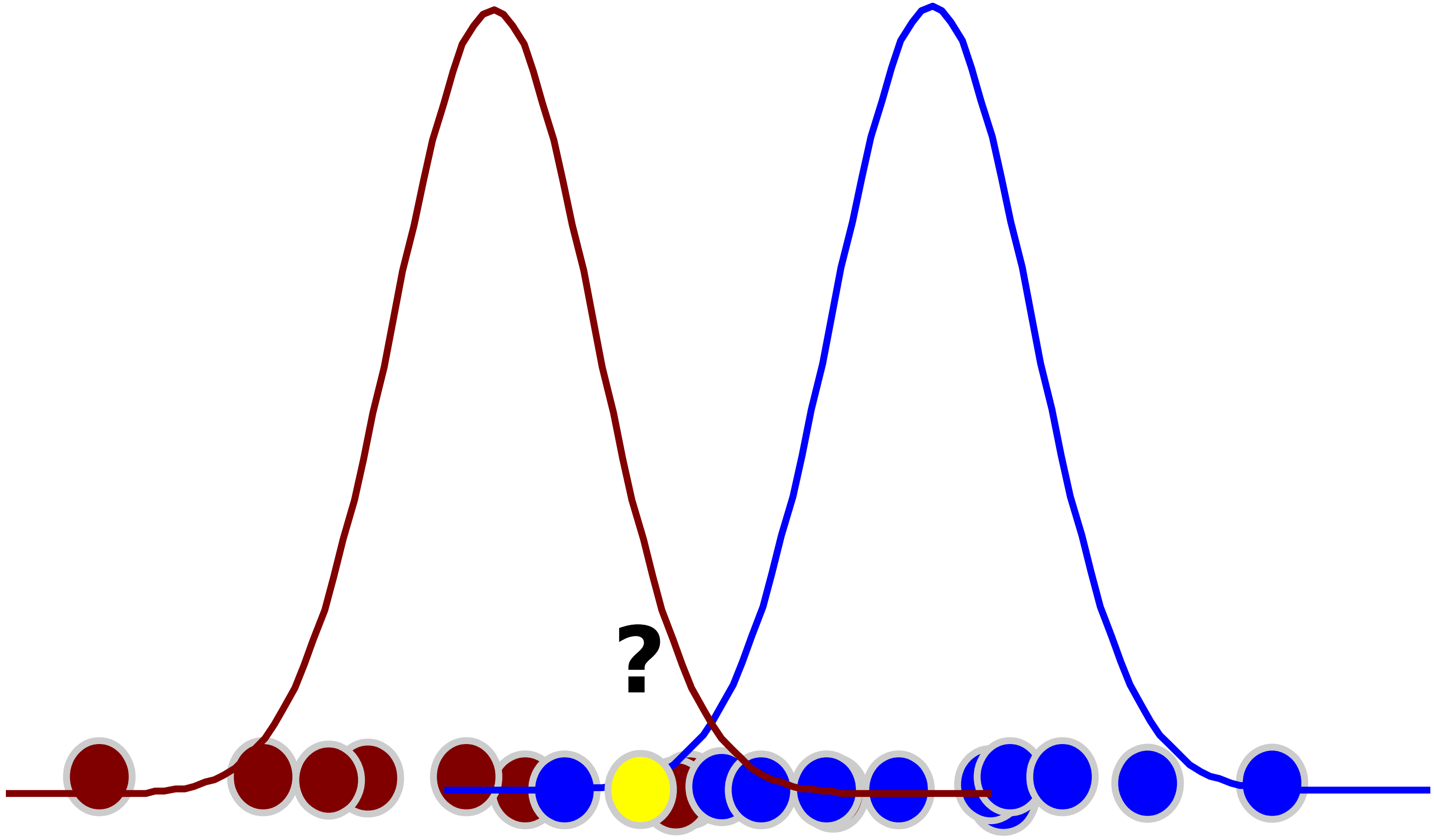
Distributions are useful how? Classification



Distributions are useful how? Classification

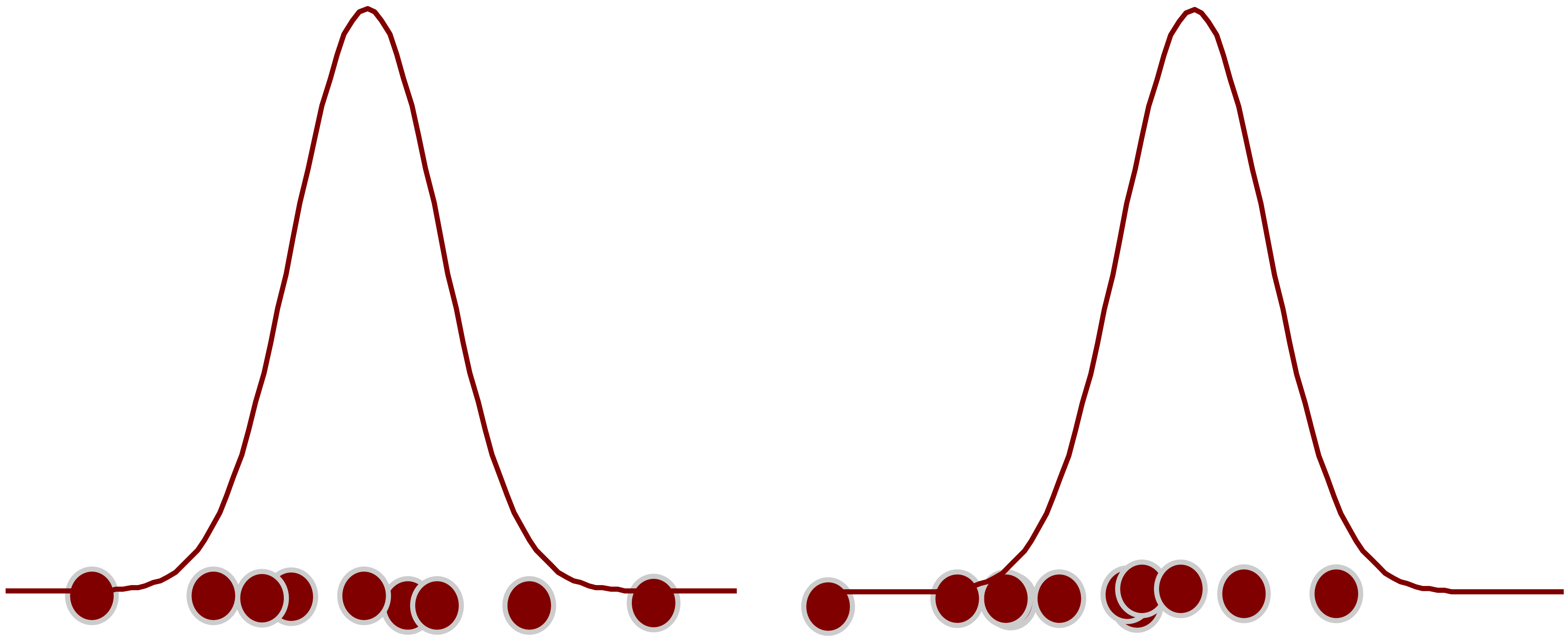


Distributions are useful how? Classification



Distributions are useful how? Clustering

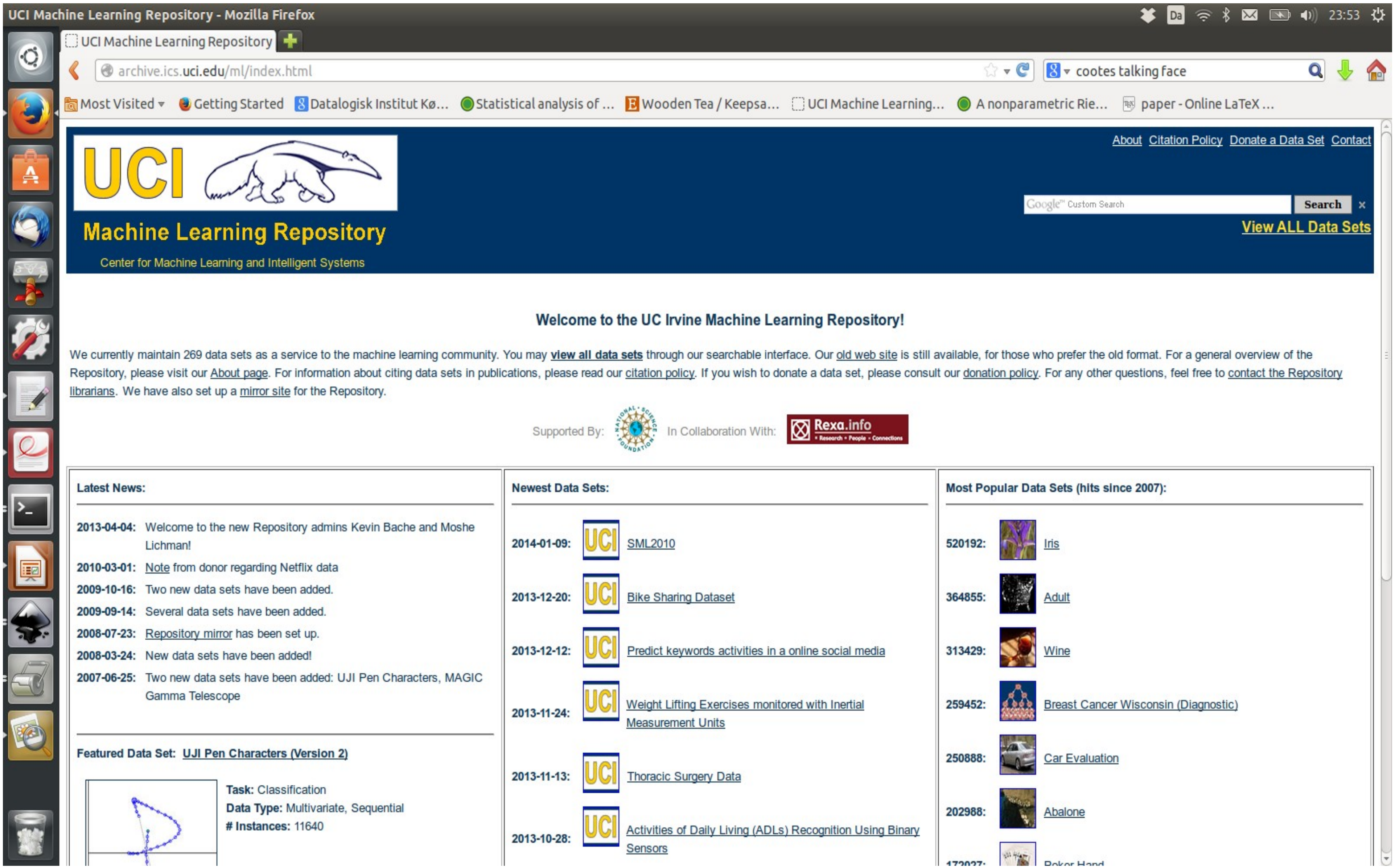
Distributions are useful how? Clustering



Summary

- Today we have learned
  - What are the main types of ML problems
  - Definition and meaning of probability density/mass functions for continuous and discrete random variables
  - Definition of Gaussian probability density distributions and their parameters
- Reading material: CB section 1.1-1.2.4 (p 1-28)
- Any questions?

If you want to play!



Next time!

- Bayes' rule
- Parametric estimation
- Multivariate Gaussian distributions
- Nonparametric estimation
- Reading material: CB sections 1.1-1.2.4 (p 1-28), 2.3 + 2.5 (p 78-113, 120-127)