Generalized Bayesian Learning

Benjamin Guedj, Ph.D.

https://bguedj.github.io
Inria Lille - Nord Europe & UCL AI

2018–2019

[https://bguedj.github.io]

8h de cours (4 $\times$ 2h, 6 février & 5 mars 2019)

Evaluation : projet à rendre. Deadline : **3 avril 2019 à 23h59**.

Slides : [https://bguedj.github.io/teaching/bguedj.pdf]

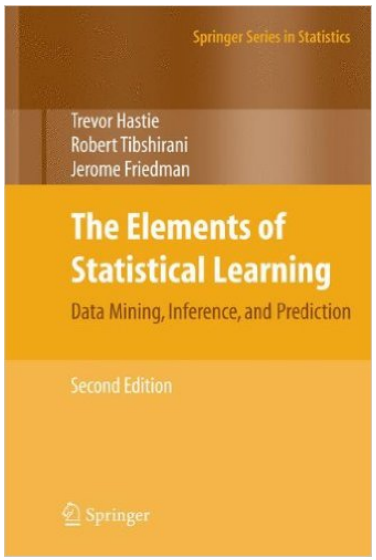
Projet : [https://bguedj.github.io/teaching/projet.pdf]

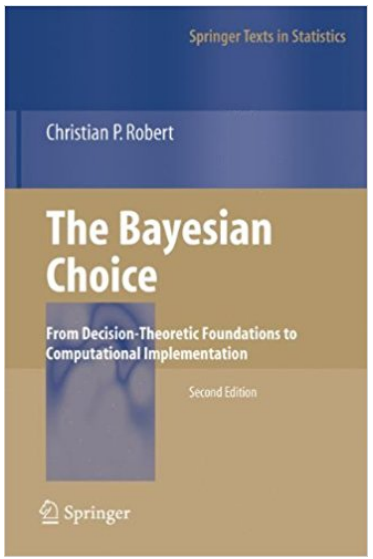
References

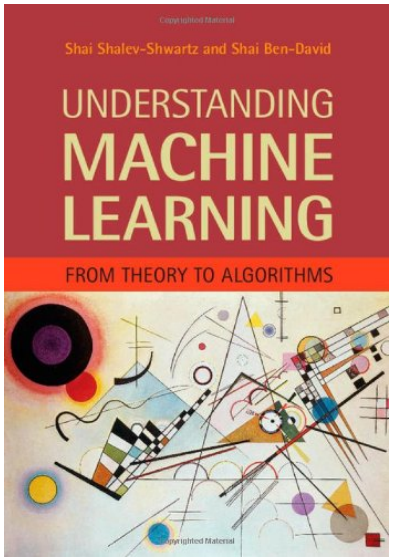
T. Hastie, R. Tibshirani and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, 2009. [Link]

C. P. Robert. *The Bayesian Choice*, Springer, 2007. [Link]

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





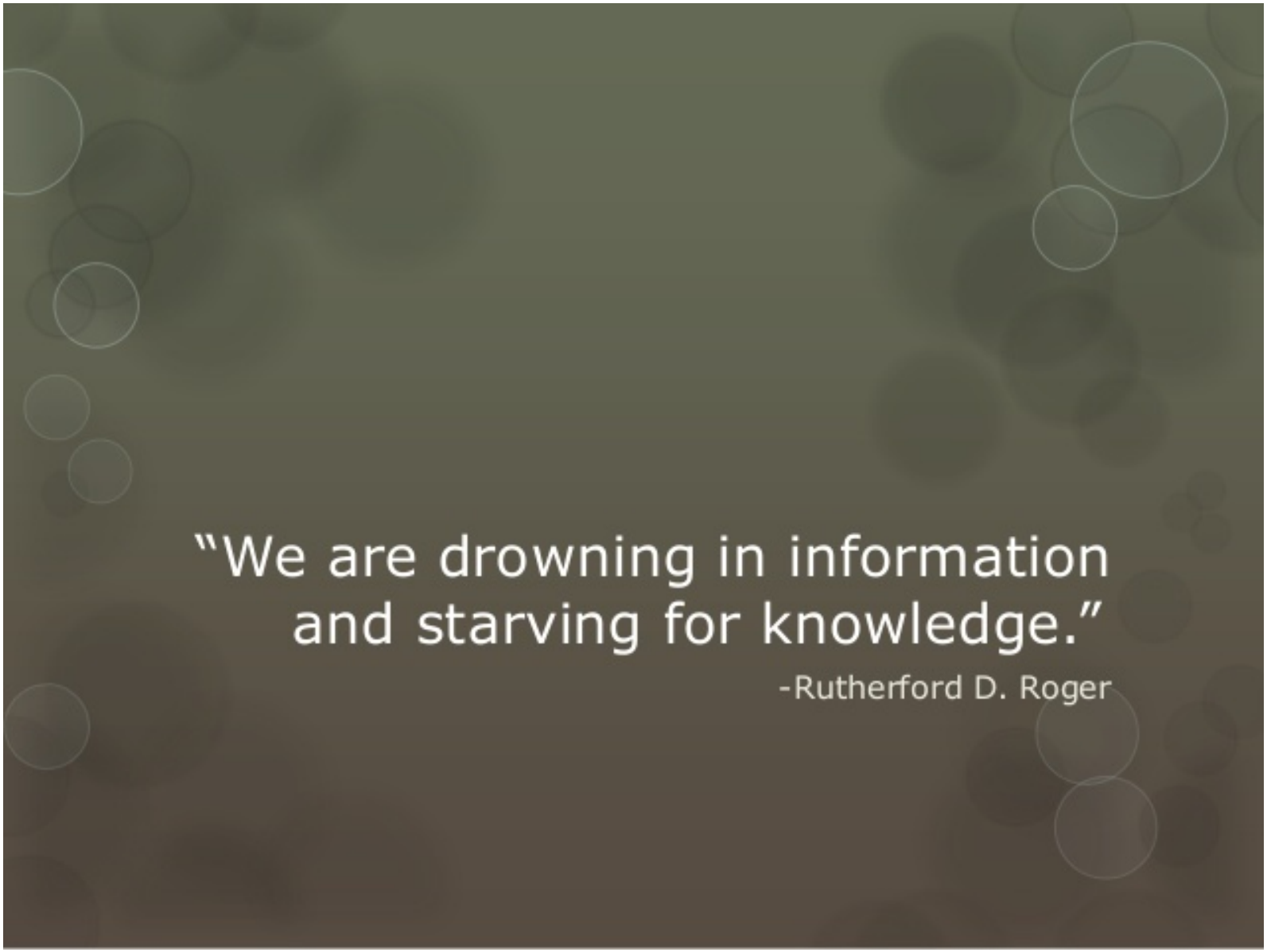


Outline

1. Introduction to statistical and machine learning
2. The Bayesian framework
3. Generalized Bayesian learning
4. Bayesian learning in practice: implementation

The rising of AI

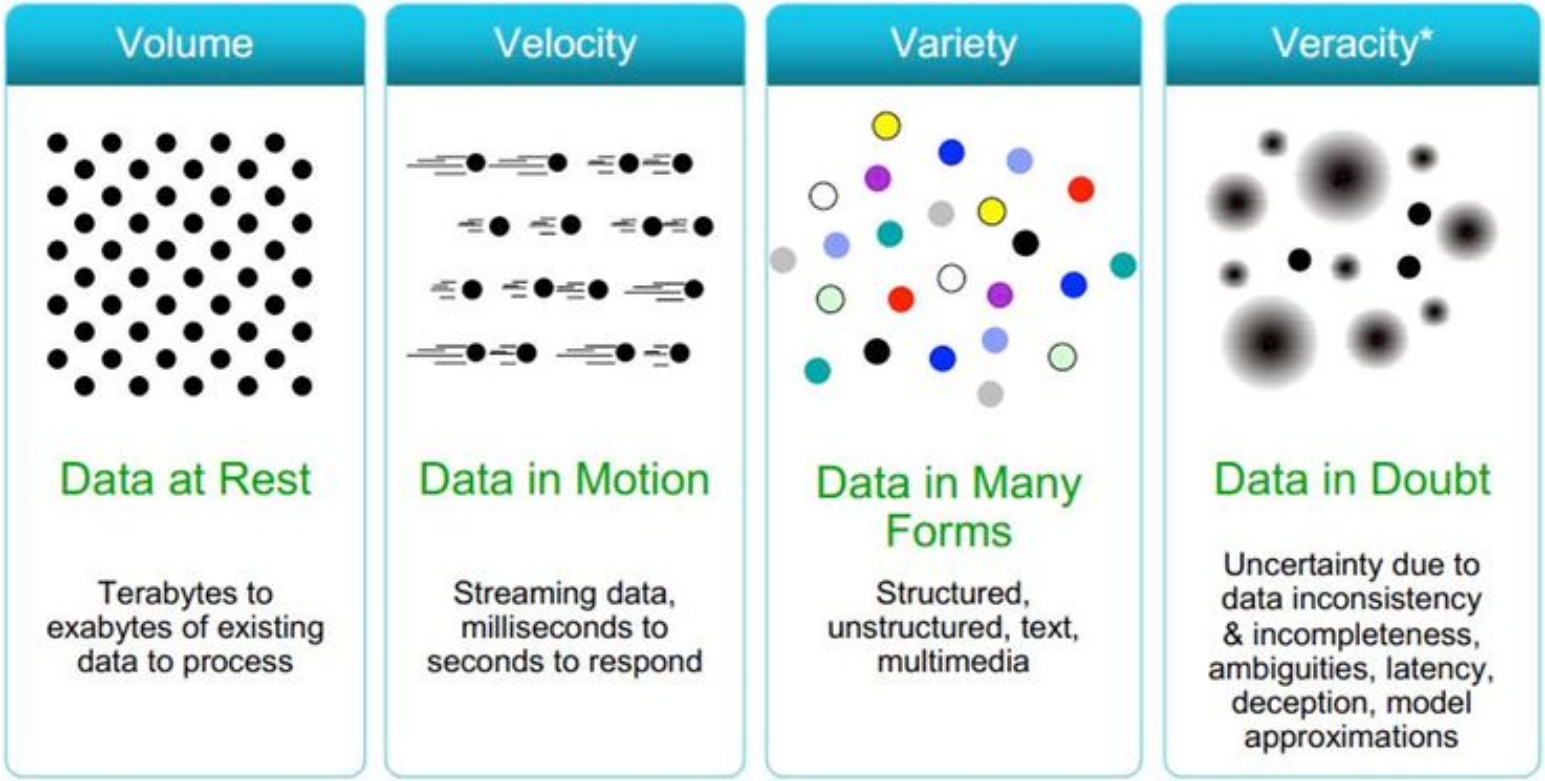
Introduction


"We are drowning in information
and starving for knowledge."
-Rutherford D. Roger

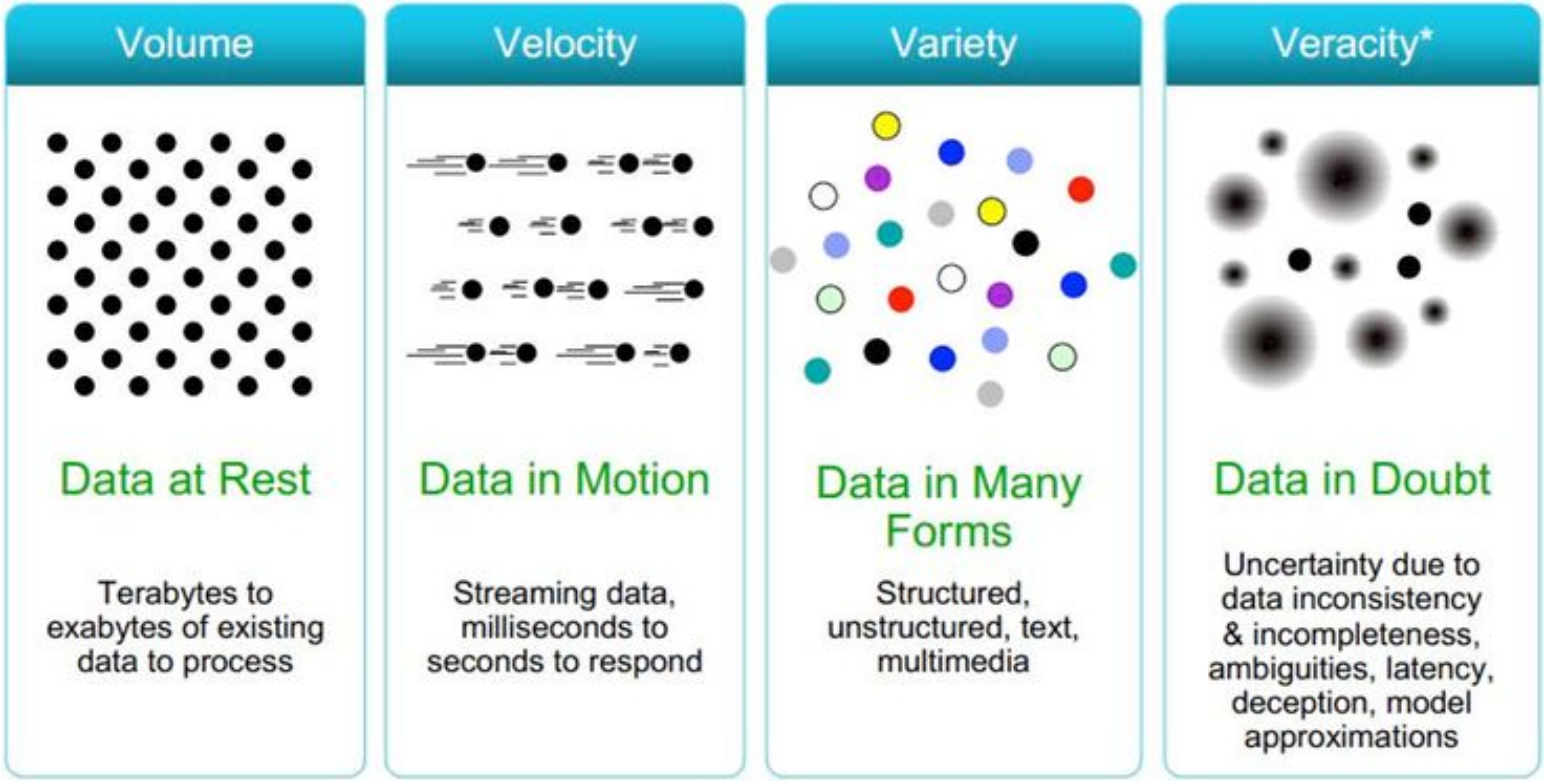
It is vital to remember that information - in the sense of raw data - is not knowledge, that knowledge is not wisdom, and that wisdom is not foresight. But information is the first essential step to all of these.

Arthur C Clarke

Big Data 4 V's


Volume
Data at Rest
Terabytes to exabytes of existing data to process
Velocity
Data in Motion
Streaming data, milliseconds to seconds to respond
Variety
Data in Many Forms
Structured, unstructured, text, multimedia
Veracity*
Data in Doubt
Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Big Data 4 V's



$\longrightarrow$ Value ($)

Data Scientists: voted 'most sexiest job' of the 21st century. Demand is expected to exceed supply by 50 to 60% (McKinsey, 2015)

Volume / Velocity

i	3	6	9	12	15	18	21	24	
10^i	kilo	mega	giga	tera	peta	exa	zeta	yotta	bytes

Volume / Velocity

i	3	6	9	12	15	18	21	24	
10^i	kilo	mega	giga	tera	peta	exa	zeta	yotta	bytes

- Google: 24 petabytes/day

Volume / Velocity

i	3	6	9	12	15	18	21	24	
10^i	kilo	mega	giga	tera	peta	exa	zeta	yotta	bytes

- Google: 24 petabytes/day
- Facebook: 10 terabytes/day

Volume / Velocity

i	3	6	9	12	15	18	21	24	
10^i	kilo	mega	giga	tera	peta	exa	zeta	yotta	bytes

- Google: 24 petabytes/day
- Facebook: 10 terabytes/day
- Twitter: 7 terabytes/day

Volume / Velocity

i	3	6	9	12	15	18	21	24	
10^i	kilo	mega	giga	tera	peta	exa	zeta	yotta	bytes

- Google: 24 petabytes/day
- Facebook: 10 terabytes/day
- Twitter: 7 terabytes/day
- Large Hadron Collider: 40 terabytes/second

Volume / Velocity

i	3	6	9	12	15	18	21	24	
10^i	kilo	mega	giga	tera	peta	exa	zeta	yotta	bytes

- Google: 24 petabytes/day
- Facebook: 10 terabytes/day
- Twitter: 7 terabytes/day
- Large Hadron Collider: 40 terabytes/second
- Google Street View: 20 petabytes

Volume / Velocity

i	3	6	9	12	15	18	21	24	
10^i	kilo	mega	giga	tera	peta	exa	zeta	yotta	bytes

- Google: 24 petabytes/day
- Facebook: 10 terabytes/day
- Twitter: 7 terabytes/day
- Large Hadron Collider: 40 terabytes/second
- Google Street View: 20 petabytes
- AT&T network: 30 petabytes/day

Volume / Velocity

i	3	6	9	12	15	18	21	24	
10^i	kilo	mega	giga	tera	peta	exa	zeta	yotta	bytes

- Google: 24 petabytes/day
- Facebook: 10 terabytes/day
- Twitter: 7 terabytes/day
- Large Hadron Collider: 40 terabytes/second
- Google Street View: 20 petabytes
- AT&T network: 30 petabytes/day
- Human brain may store about 2.5 petabytes of binary data

Volume / Velocity

i	3	6	9	12	15	18	21	24	
10^i	kilo	mega	giga	tera	peta	exa	zeta	yotta	bytes

- Google: 24 petabytes/day
- Facebook: 10 terabytes/day
- Twitter: 7 terabytes/day
- Large Hadron Collider: 40 terabytes/second
- Google Street View: 20 petabytes
- AT&T network: 30 petabytes/day
- Human brain may store about 2.5 petabytes of binary data
- ...

Variety / Veracity

categorical
Marital status
maried

integer
Children
3

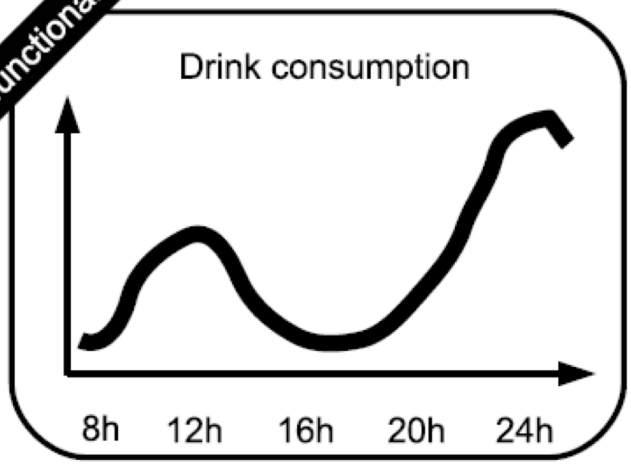
missing
Size (m)
?

rank
Drink preference
beer > soda > water

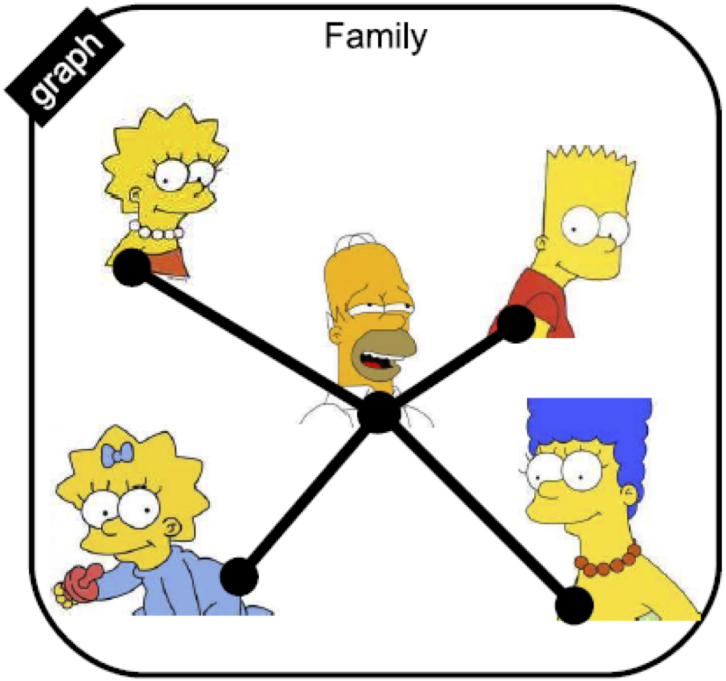
ordinal
Intelligence
low

continuous
Weight (kg)
119.5

functional

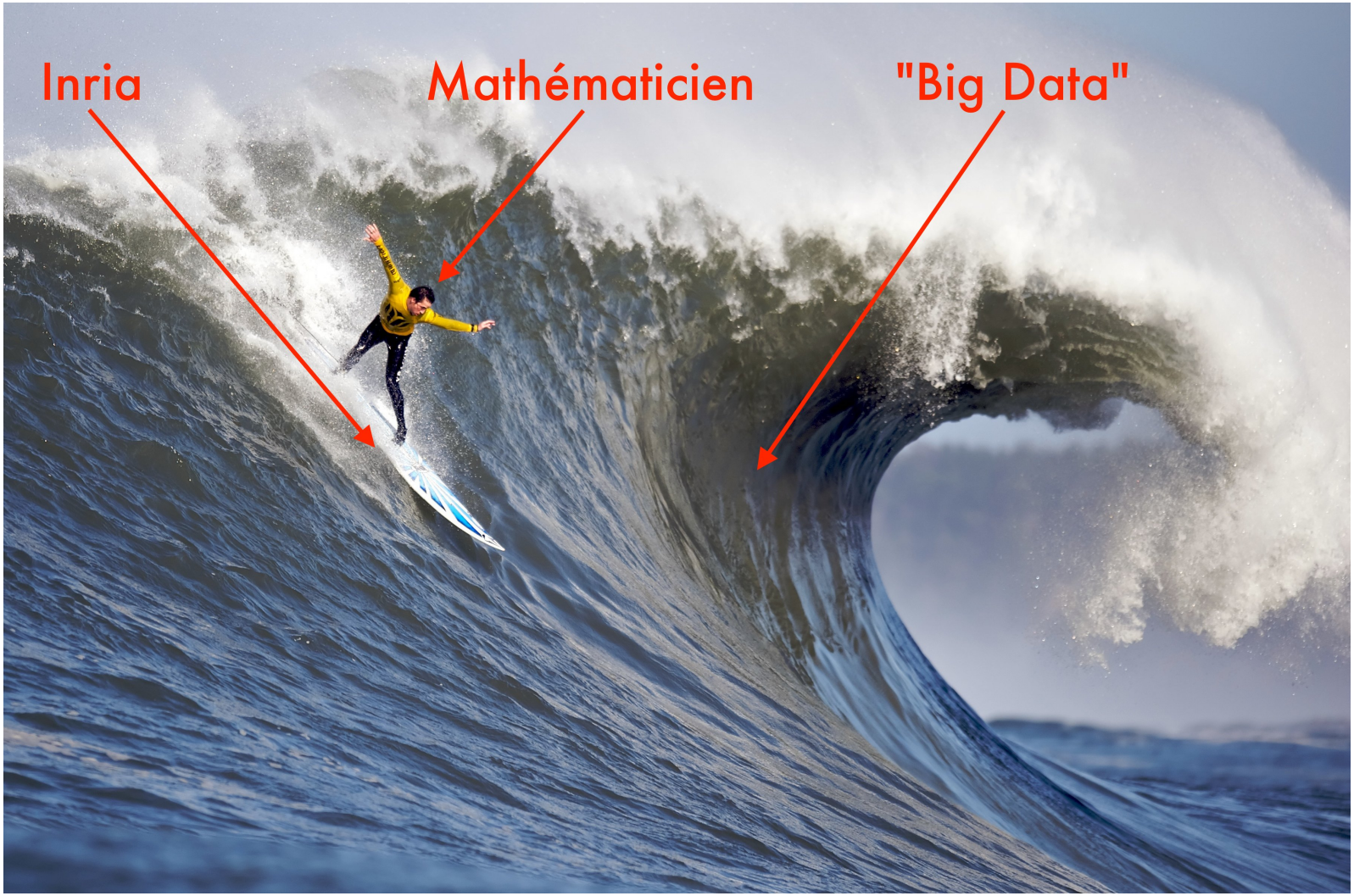


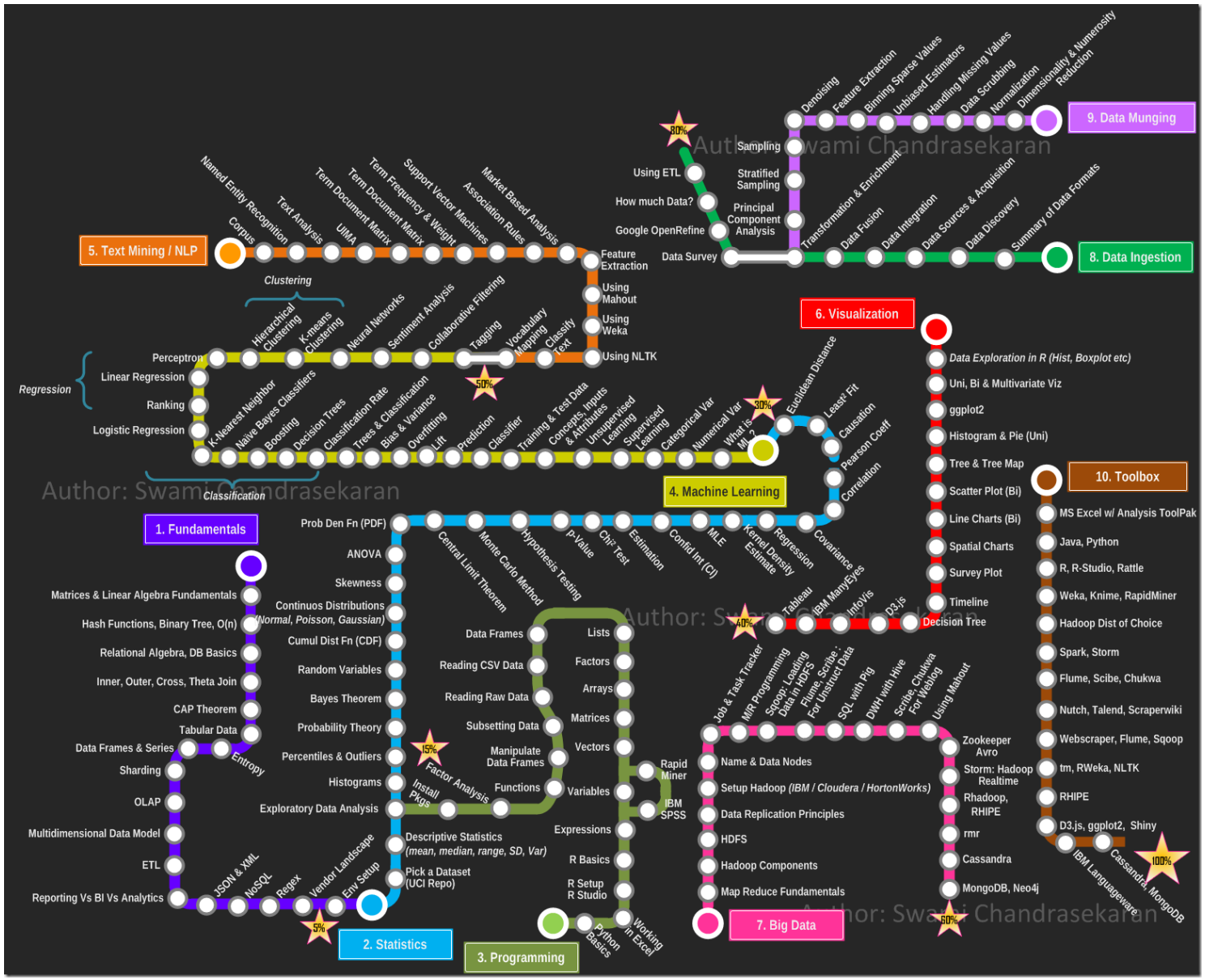
graph

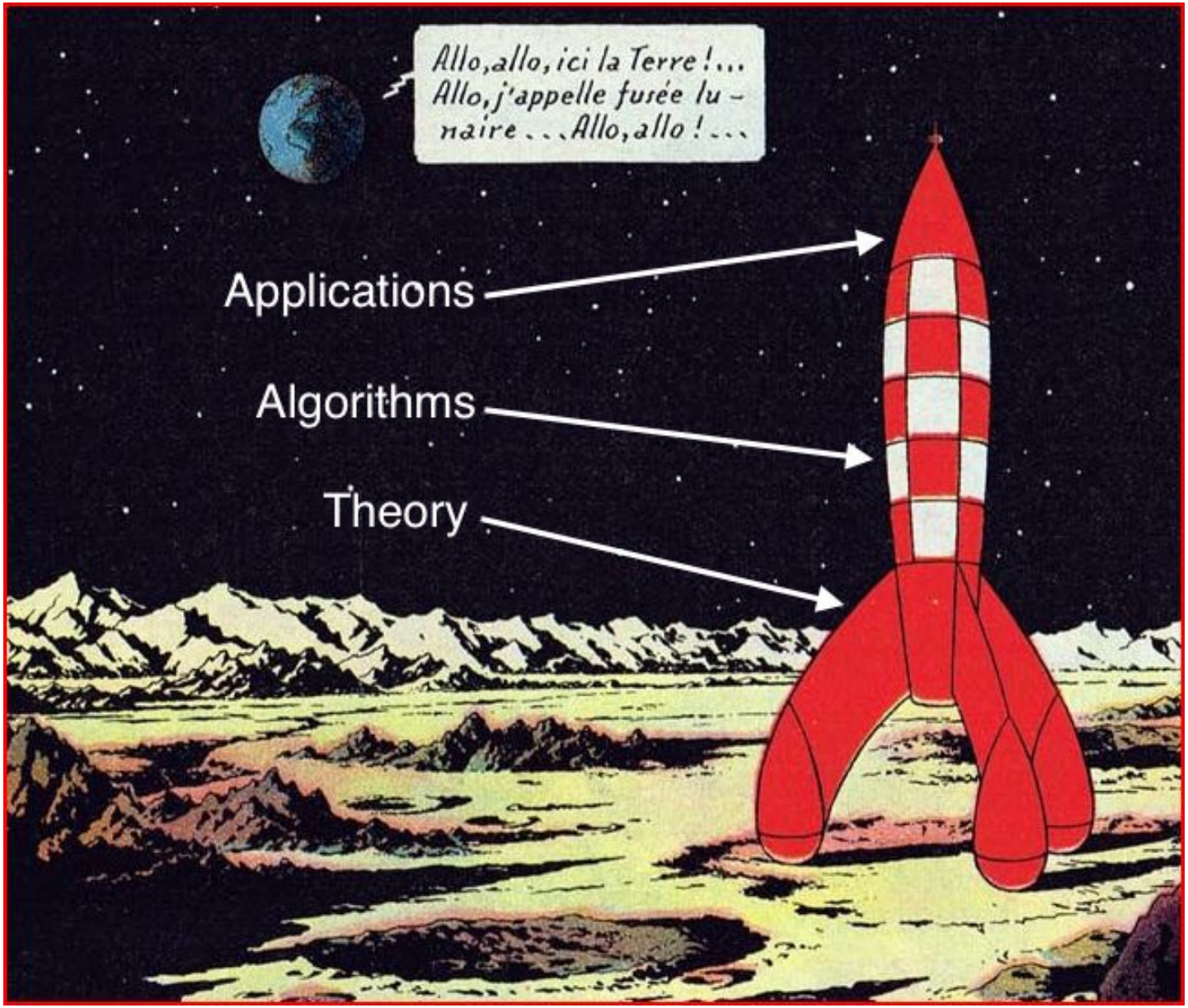


And so on...

My job (allegory)




1. Fundamentals
Matrices & Linear Algebra Fundamentals
Hash Functions, Binary Tree, O(n)
Relational Algebra, DB Basics
Inner, Outer, Cross, Theta Join
CAP Theorem
Tabular Data
Data Frames & Series
Entropy
Sharding
OLAP
Multidimensional Data Model
ETL
Reporting Vs BI Vs Analytics
JSON & XML
NoSQL
Regex
Vendor Landscape
Env Setup
5%
2. Statistics
Prob Den Fn (PDF)
ANOVA
Skewness
Continuos Distributions (Normal, Poisson, Gaussian)
Cumul Dist Fn (CDF)
Random Variables
Bayes Theorem
Probability Theory
Percentiles & Outliers
Histograms
Exploratory Data Analysis
Descriptive Statistics (mean, median, range, SD, Var)
Pick a Dataset (UCI Repo)
Central Limit Theorem
Monte Carlo Method
Hypothesis Testing
p-Value
Chi² Test
Estimation
Confid Int (CI)
MLE
Kernel Density Estimate
Regression
Covariance
15%
Factor Analysis
Install Pkgs
3. Programming
Data Frames
Reading CSV Data
Reading Raw Data
Subsetting Data
Manipulate Data Frames
Functions
Lists
Factors
Arrays
Matrices
Vectors
Variables
Expressions
R Basics
R Setup R Studio
Python Basics
Working in Excel
Rapid Miner
IBM SPSS
4. Machine Learning
Regression
Perceptron
Linear Regression
Ranking
Logistic Regression
Classification
K-Nearest Neighbor
Naive Bayes Classifiers
Boosting
Decision Trees
Classification Rate
Trees & Classification
Bias & Variance
Overfitting
Lift
Prediction
Classifier
Training & Test Data
Concepts, Inputs & Attributes
Unsupervised Learning
Supervised Learning
Categorical Var
Numerical Var
What is ML?
50%
Clustering
Hierarchical Clustering
K-means Clustering
Neural Networks
Sentiment Analysis
Collaborative Filtering
Tagging
Vocabulary Mapping
Classify Text
Using NLTK
Using Mahout
Using Weka
Named Entity Recognition
Text Analysis
Corpus
UIMA
Term Document Matrix
Term Document Matrix
Term Frequency & Weight
Support Vector Machines
Association Rules
Market Based Analysis
Feature Extraction
5. Text Mining / NLP
30%
Euclidean Distance
Least² Fit
Causation
Pearson Coeff
Correlation
6. Visualization
Data Exploration in R (Hist, Boxplot etc)
Uni, Bi & Multivariate Viz
ggplot2
Histogram & Pie (Uni)
Tree & Tree Map
Scatter Plot (Bi)
Line Charts (Bi)
Spatial Charts
Survey Plot
Timeline
Decision Tree
40%
Tableau
IBM ManyEyes
InfoVis
D3.js
80%
Sampling
Stratified Sampling
Principal Component Analysis
Transformation & Enrichment
Using ETL
How much Data?
Google OpenRefine
Data Survey
9. Data Munging
Denoising
Feature Extraction
Binning Sparse Values
Unbiased Estimators
Handling Missing Values
Data Scrubbing
Normalization
Dimensionality & Numerosity Reduction
8. Data Ingestion
Data Fusion
Data Integration
Data Sources & Acquisition
Data Discovery
Summary of Data Formats
7. Big Data
Job & Task Tracker
MR Programming
Sqoop: Loading Data in HDFS
Flume, Scribe : For Unstruct Data
SQL with Pig
DWH with Hive
Scribe, Chukwa For Weblog
Using Mahout
Name & Data Nodes
Setup Hadoop (IBM / Cloudera / HortonWorks)
Data Replication Principles
HDFS
Hadoop Components
Map Reduce Fundamentals
Zookeeper Avro
Storm: Hadoop Realtime
Rhadoop, RHIPE
rmr
Cassandra
MongoDB, Neo4j
60%
10. Toolbox
MS Excel w/ Analysis ToolPak
Java, Python
R, R-Studio, Rattle
Weka, Knime, RapidMiner
Hadoop Dist of Choice
Spark, Storm
Flume, Scibe, Chukwa
Nutch, Talend, Scraperwiki
Webscraper, Flume, Sqoop
tm, RWeka, NLTK
RHIPE
D3.js, ggplot2, Shiny
IBM Languageware
Cassandra, MongoDB
100%


Allo, allo, ici la Terre !...
Allo, j'appelle fusée lu-
naire ...Allo, allo !...
Applications
Algorithms
Theory

A foretaste of Learning Theory

{Statistical,Machine} Learning: building automatic procedures to infer general rules from examples.

{Statistical,Machine} Learning: building automatic procedures to infer general rules from examples.

In the (~~rather~~ not so?) long term: mimic the inductive functioning of the humain brain to develop an artificial intelligence.

{Statistical,Machine} Learning: building automatic procedures to infer general rules from examples.

In the (~~rather~~ not so?) long term: mimic the inductive functioning of the humain brain to develop an artificial intelligence.

In the Big Data Era, very dynamic field at the crossroads of Computer Science, Optimization and Statistics.

Probabilistic framework: n-sample $\mathcal{D}_n = (\mathbf{X}_i, \mathbf{Y}_i)_{i=1}^n$ of i.i.d. replications of some random variable

$$(\mathbf{X}, \mathbf{Y}) \in \mathcal{X} \times \mathcal{Y}, \quad \dim(\mathcal{X}) = d.$$

Probabilistic framework: n-sample $\mathcal{D}_n = (\mathbf{X}_i, \mathbf{Y}_i)_{i=1}^n$ of i.i.d. replications of some random variable

$$(\mathbf{X}, \mathbf{Y}) \in \mathcal{X} \times \mathcal{Y}, \quad \dim(\mathcal{X}) = d.$$

We want to infer the link between the explanatory variable $\mathbf{X}$ and the response variable $\mathbf{Y}$, *i.e.*, use $\mathcal{D}_n$ to build up $\hat{\phi}$ such that $\hat{\phi}(\mathbf{X})$ is a "good" approximation of $\mathbf{Y}$.

Probabilistic framework: n-sample $\mathcal{D}_n = (\mathbf{X}_i, \mathbf{Y}_i)_{i=1}^n$ of i.i.d. replications of some random variable

$$(\mathbf{X}, \mathbf{Y}) \in \mathcal{X} \times \mathcal{Y}, \quad \dim(\mathcal{X}) = d.$$

We want to infer the link between the explanatory variable $\mathbf{X}$ and the response variable $\mathbf{Y}$, *i.e.*, use $\mathcal{D}_n$ to build up $\hat{\phi}$ such that $\hat{\phi}(\mathbf{X})$ is a "good" approximation of $\mathbf{Y}$.

- Classification: $\mathcal{Y}$ is discrete.
- Regression: $\mathcal{Y}$ is a continuum.

- Online learning: observations are revealed over time.
- Batch learning: all observations are revealed at once.

- Online learning: observations are revealed over time.
- Batch learning: all observations are revealed at once.

Big Data Era: easy/cheap to collect massive amounts of data, hence typically $\mathbf{X}_1, \ldots, \mathbf{X}_n, \ldots \in \mathbb{R}^{n \times d}$ where d and n may be (extremely) large.

- Online learning: observations are revealed over time.
- Batch learning: all observations are revealed at once.

Big Data Era: easy/cheap to collect massive amounts of data, hence typically $\mathbf{X}_1, \ldots, \mathbf{X}_n, \ldots \in \mathbb{R}^{n \times d}$ where d and n may be (extremely) large.

- Good ol' statistics: n and d small

- Online learning: observations are revealed over time.
- Batch learning: all observations are revealed at once.

Big Data Era: easy/cheap to collect massive amounts of data, hence typically $\mathbf{X}_1, \ldots, \mathbf{X}_n, \ldots \in \mathbb{R}^{n \times d}$ where d and n may be (extremely) large.

- Good ol' statistics: n and d small
- Tall data: $n \gg d$

- Online learning: observations are revealed over time.
- Batch learning: all observations are revealed at once.

Big Data Era: easy/cheap to collect massive amounts of data, hence typically $\mathbf{X}_1, \ldots, \mathbf{X}_n, \ldots \in \mathbb{R}^{n \times d}$ where d and n may be (extremely) large.

- Good ol' statistics: n and d small
- Tall data: $n \gg d$
- Fat data: $d \gg n$

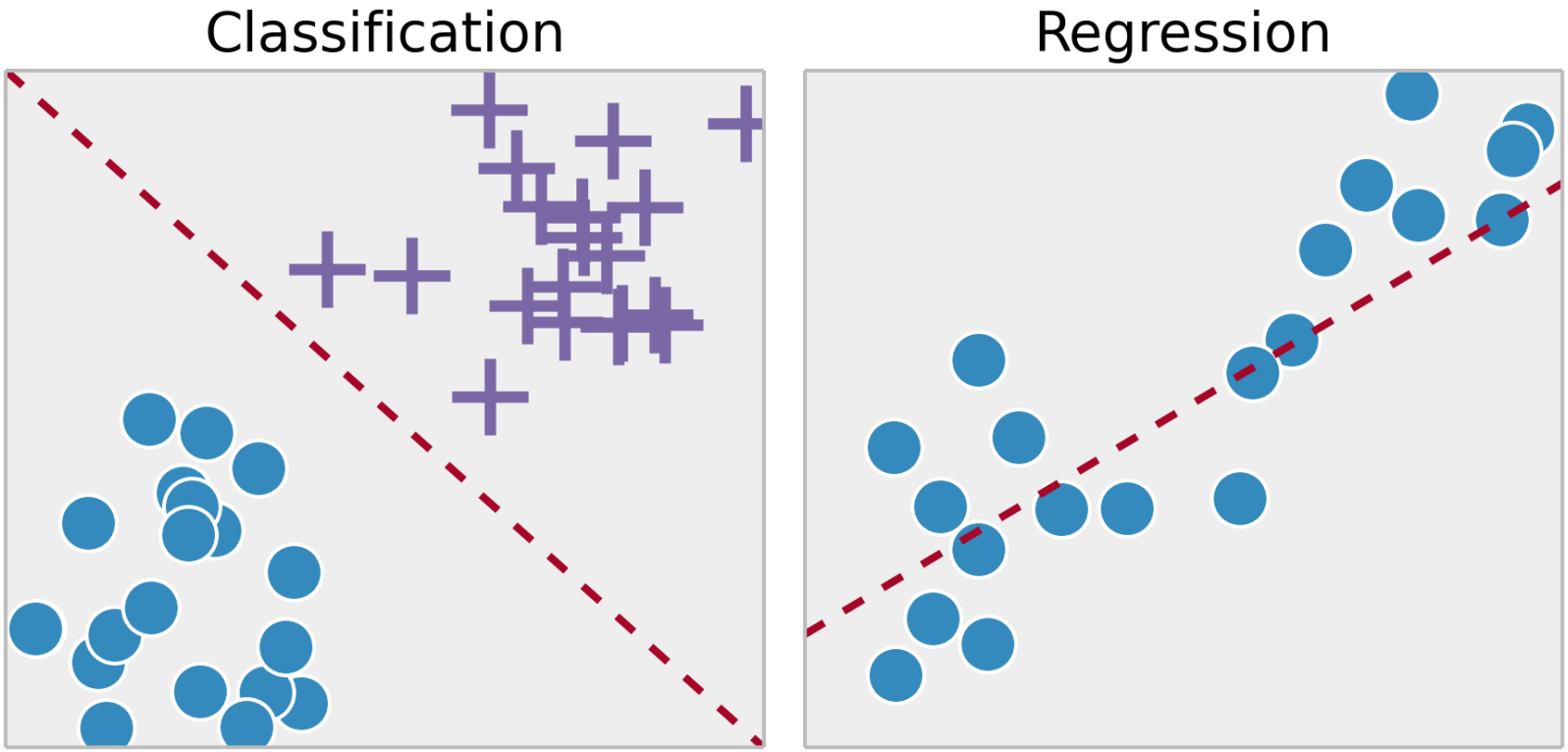
- Online learning: observations are revealed over time.
- Batch learning: all observations are revealed at once.

Big Data Era: easy/cheap to collect massive amounts of data, hence typically $\mathbf{X}_1, \ldots, \mathbf{X}_n, \ldots \in \mathbb{R}^{n \times d}$ where d and n may be (extremely) large.

- Good ol' statistics: n and d small
- Tall data: $n \gg d$
- Fat data: $d \gg n$
- Big/massive data: n and d huge

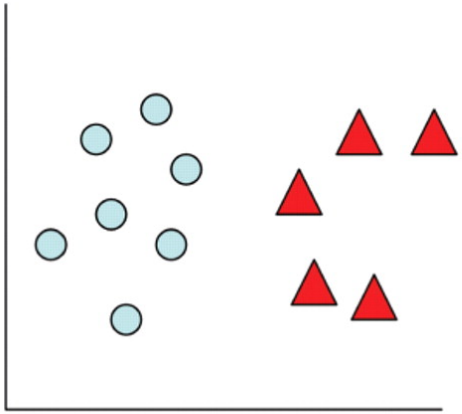
- Supervised learning: all of the $\mathbf{Y}_i$s are observed.

- Supervised learning: all of the $\mathbf{Y}_i$s are observed.

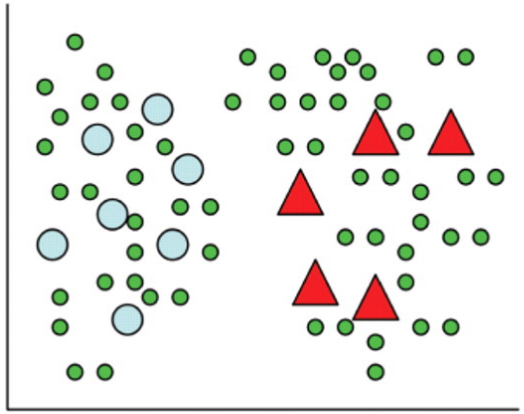


- Semi-supervised learning: some of the $\mathbf{Y}_i$s are observed (labeling is expensive or difficult).

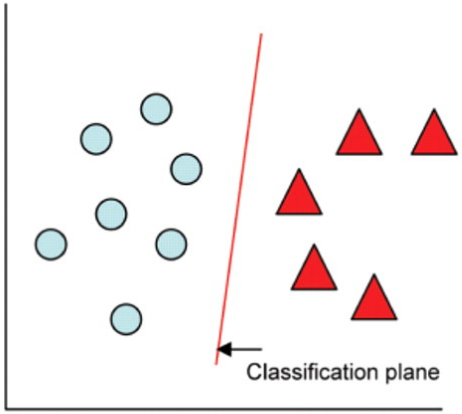
- Semi-supervised learning: some of the $\mathbf{Y}_i$s are observed (labeling is expensive or difficult).



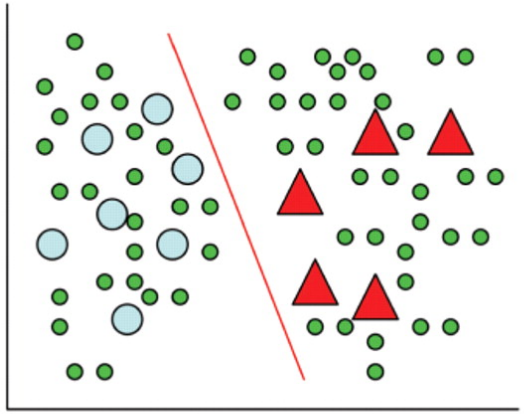
Labeled Data
(a)



Labeled and Unlabeled Data
(b)



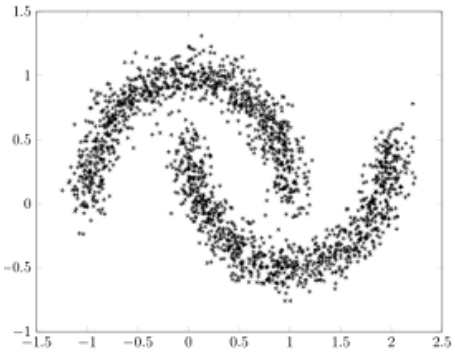
Supervised Learning
(c)

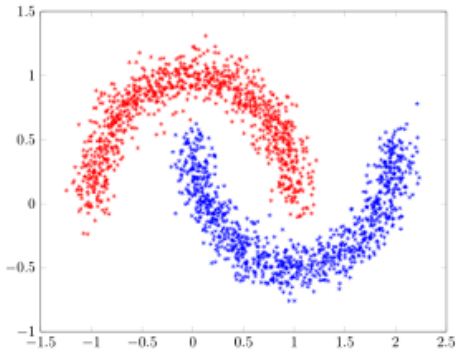


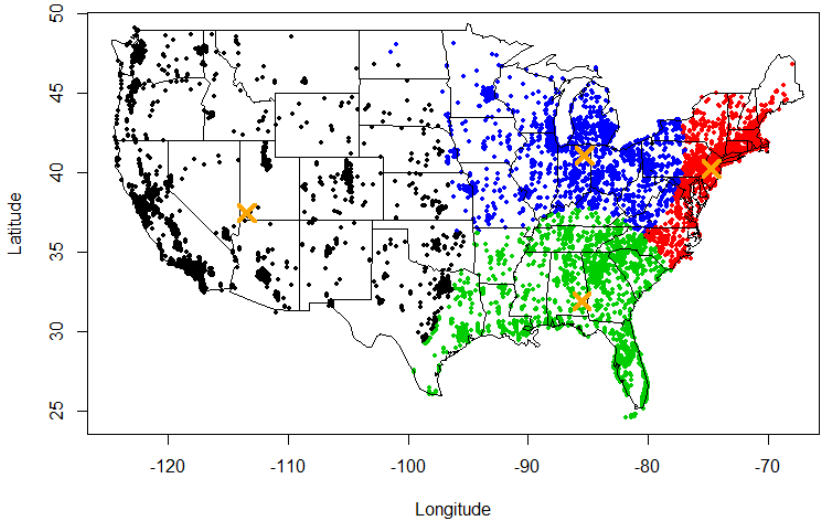
Semi-Supervised Learning
(d)

- Unsupervised learning: none of the $\mathbf{Y}_i$s are observed (detect patterns).

- Unsupervised learning: none of the $\mathbf{Y}_i$s are observed (detect patterns).





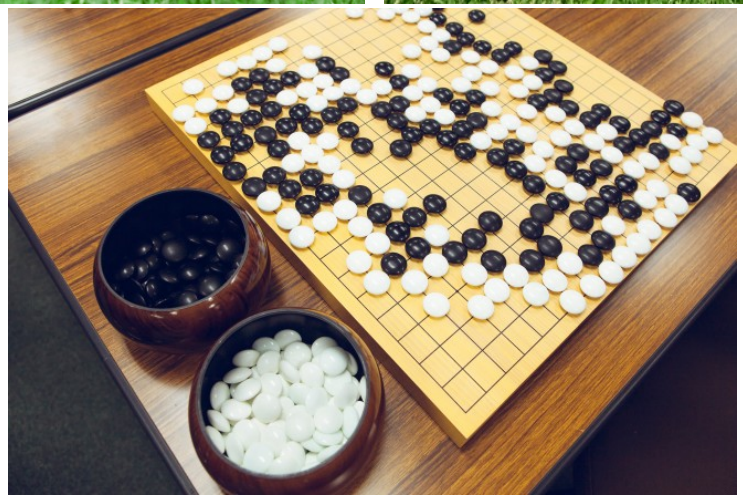


- Unsupervised learning: none of the $\mathbf{Y}_i$s are observed (detect patterns).

- Unsupervised learning: none of the $\mathbf{Y}_i$s are observed (detect patterns).

- Reinforcement learning: feedback (possibly adversarial) from the environment (robotics, adversarial environments, training...).

- Reinforcement learning: feedback (possibly adversarial) from the environment (robotics, adversarial environments, training...).



Loss function:

$$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+.$$

Loss function:

$$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+.$$

$\ell(\hat{\phi}(\mathbf{X}), Y)$ (random) quantifies how a predictor $\hat{\phi}(\mathbf{X})$ is a "good" approximation of $\mathbf{Y}$.

Loss function:

$$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+.$$

$\ell(\hat{\phi}(\mathbf{X}), Y)$ (random) quantifies how a predictor $\hat{\phi}(\mathbf{X})$ is a "good" approximation of $\mathbf{Y}$.

A predictor is any mapping

$$\hat{\phi} : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{Y}^{\mathcal{X}}.$$

Loss function:

$$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+.$$

$\ell(\hat{\phi}(\mathbf{X}), Y)$ (random) quantifies how a predictor $\hat{\phi}(\mathbf{X})$ is a "good" approximation of $\mathbf{Y}$.

A predictor is any mapping

$$\hat{\phi} : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{Y}^{\mathcal{X}}.$$

Risk

$$R(\hat{\phi}) = \mathbb{E}\left[\ell\left(\hat{\phi}(\mathbf{X}), Y\right)\right],$$

Loss function:

$$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+.$$

$\ell(\hat{\phi}(\mathbf{X}), Y)$ (random) quantifies how a predictor $\hat{\phi}(\mathbf{X})$ is a "good" approximation of $\mathbf{Y}$.

A predictor is any mapping

$$\hat{\phi} : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{Y}^{\mathcal{X}}.$$

Risk

$$R(\hat{\phi}) = \mathbb{E}\left[\ell\left(\hat{\phi}(\mathbf{X}), Y\right)\right],$$

Empirical risk

$$R_n(\hat{\phi}) = \frac{1}{n}\sum_{i=1}^{n}\left[\ell\left(\hat{\phi}(\mathbf{X}_i), Y_i\right)\right].$$

- Quadratic loss (regression): $\ell(a, b) = (a - b)^2$.

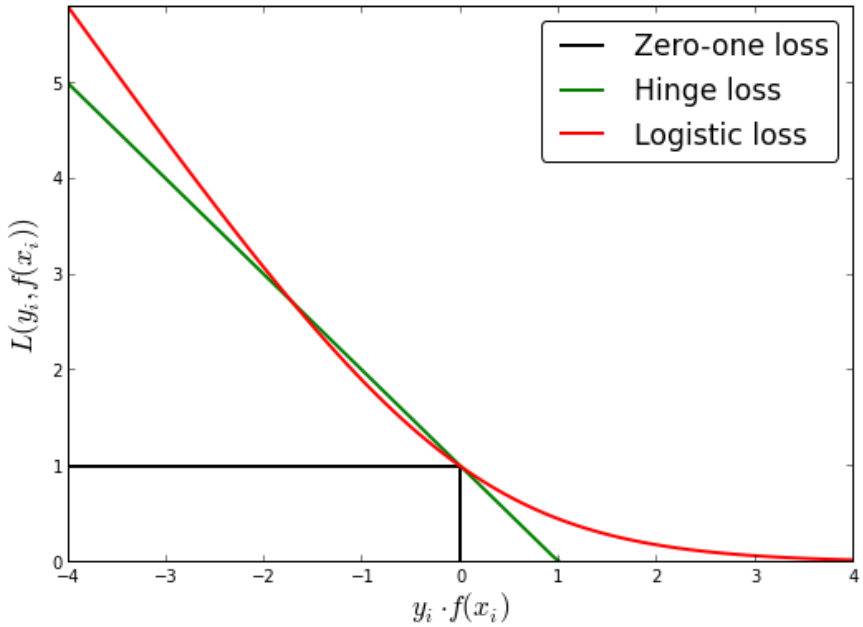
- Quadratic loss (regression): $\ell(a, b) = (a - b)^2$.
- Absolute loss (regression): $\ell(a, b) = |a - b|$.

- Quadratic loss (regression): $\ell(a, b) = (a - b)^2$.
- Absolute loss (regression): $\ell(a, b) = |a - b|$.
- 0-1 loss (classification): $\ell(a, b) = \mathbb{1}_{\{a \neq b\}}$.

- Quadratic loss (regression): $\ell(a, b) = (a - b)^2$.
- Absolute loss (regression): $\ell(a, b) = |a - b|$.
- 0-1 loss (classification): $\ell(a, b) = \mathbb{1}_{\{a \neq b\}}$.
- Hinge loss (classification): $\ell(a, b) = \max(0, 1 - ab)$.

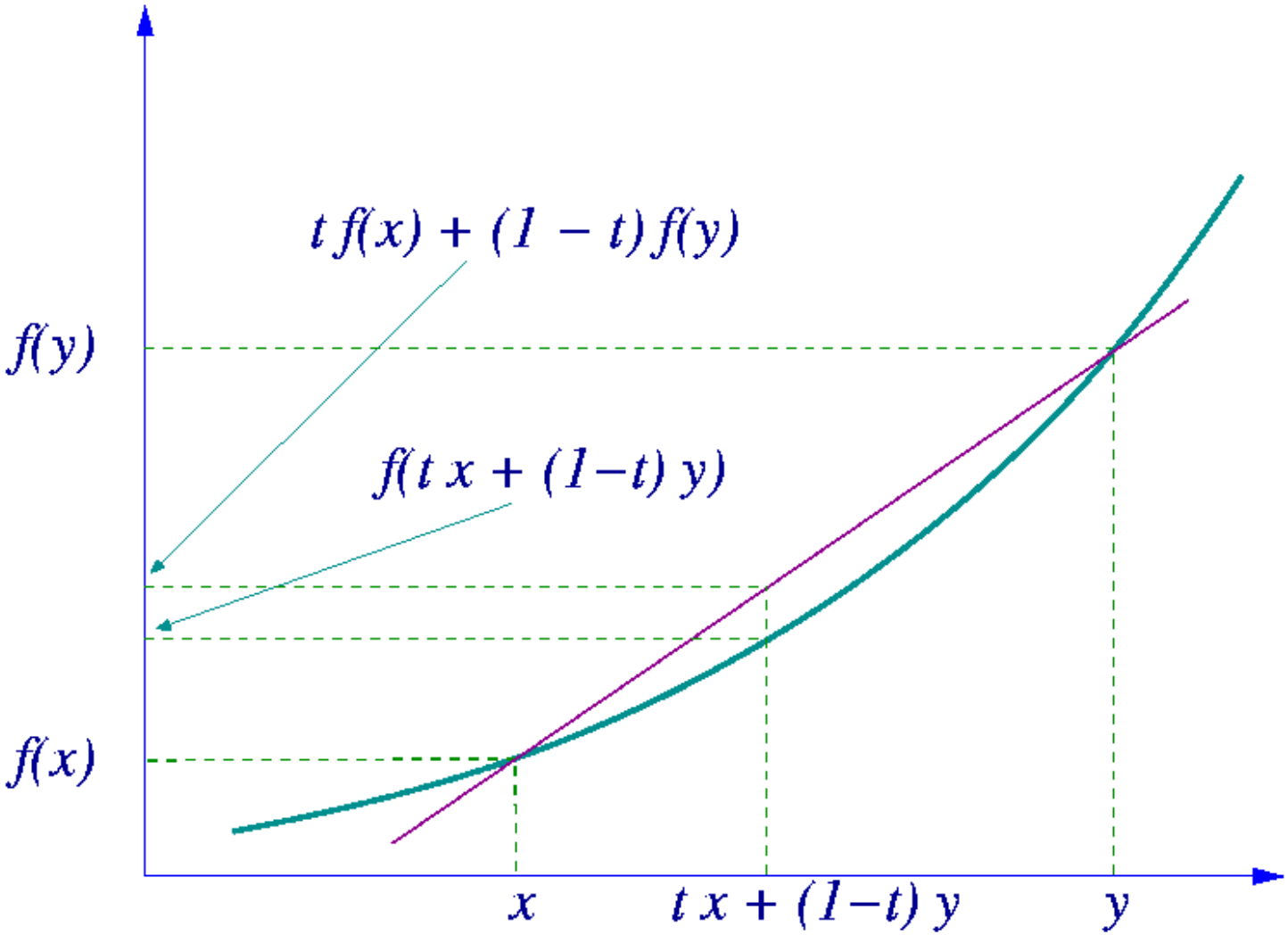
- Quadratic loss (regression): $\ell(a, b) = (a - b)^2$.
- Absolute loss (regression): $\ell(a, b) = |a - b|$.
- 0-1 loss (classification): $\ell(a, b) = \mathbb{1}_{\{a \neq b\}}$.
- Hinge loss (classification): $\ell(a, b) = \max(0, 1 - ab)$.
- Logistic loss (classification): $\ell(a, b) = \log[1 + \exp(-ab)]$.

- Quadratic loss (regression): $\ell(a, b) = (a - b)^2$.
- Absolute loss (regression): $\ell(a, b) = |a - b|$.
- 0-1 loss (classification): $\ell(a, b) = \mathbb{1}_{\{a \neq b\}}$.
- Hinge loss (classification): $\ell(a, b) = \max(0, 1 - ab)$.
- Logistic loss (classification): $\ell(a, b) = \log[1 + \exp(-ab)]$.



Convexity is (often) crucial

Convexity is (often) crucial



Statistical Learning vs. Machine Learning

Same task, different approaches:

Statistical Learning vs. Machine Learning

Same task, different approaches:

- In machine learning, given some deterministic sequence $(\mathbf{x}_i, y_i)$, solve

$$\widehat{\phi}(\cdot) = \arg\min_{m} \left\{ \sum_{i=1}^{n} \ell(y_i, m(\mathbf{x}_i) \right\}.$$

Statistical Learning vs. Machine Learning

Same task, different approaches:

- In machine learning, given some deterministic sequence $(\mathbf{x}_i, y_i)$, solve

$$\widehat{\phi}(\cdot) = \arg\min_m \left\{ \sum_{i=1}^{n} \ell(y_i, m(\mathbf{x}_i) \right\}.$$

- In statistical learning, assume that the Y_is are realisations of some random variable Y (given $\mathbf{X}$) with distribution P. Solve

$$\widehat{\phi}(\cdot) = \arg\max_m \left\{ \sum_{i=1}^{n} \log \mathrm{d}P(Y_i; m(\mathbf{X}_i)) \right\}.$$

SL vs. ML in the simple parametric case

SL vs. ML in the simple parametric case

- In machine learning, given some deterministic sequence $(\mathbf{x}_i, y_i)$, solve

$$\widehat{\phi}(\cdot) = \arg\min_{\theta} \left\{ \sum_{i=1}^{n} \ell(y_i, \langle \theta, \mathbf{x}_i \rangle) \right\}.$$

SL vs. ML in the simple parametric case

- In machine learning, given some deterministic sequence $(\mathbf{x}_i, y_i)$, solve

$$\widehat{\phi}(\cdot) = \arg\min_{\theta} \left\{ \sum_{i=1}^{n} \ell(y_i, \langle \theta, \mathbf{x}_i \rangle) \right\}.$$

- In statistical learning, assume that the Y_is are realisations of some random variable Y (given $\mathbf{X}$) with distribution P. Solve

$$\widehat{\phi}(\cdot) = \arg\max_{\theta} \left\{ \sum_{i=1}^{n} \log \mathrm{d}P(Y_i|\mathbf{X}_i, \theta) \right\}.$$

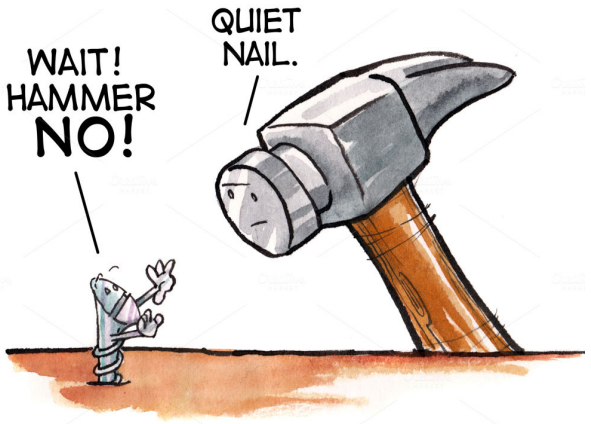
All models are wrong

but some are useful

George E.P. Box

If the only tool you have is a hammer, you tend to see every problem as a nail.

(Abraham Maslow)



A primer on probability distributions

All words are hyperlinks.

A primer on probability distributions

All words are hyperlinks.

- [Normal]

A primer on probability distributions

All words are hyperlinks.

- [Normal]
- [Inverse Gaussian (a.k.a Wald)]

A primer on probability distributions

All words are hyperlinks.

- [Normal]
- [Inverse Gaussian (a.k.a Wald)]
- [Beta]

A primer on probability distributions

All words are hyperlinks.

- [Normal]
- [Inverse Gaussian (a.k.a Wald)]
- [Beta]
- [Poisson]

A primer on probability distributions

All words are hyperlinks.

- [Normal]
- [Inverse Gaussian (a.k.a Wald)]
- [Beta]
- [Poisson]
- [Binomial]

A primer on probability distributions

All words are hyperlinks.

- [Normal]
- [Inverse Gaussian (a.k.a Wald)]
- [Beta]
- [Poisson]
- [Binomial]
- [Bernoulli]

A primer on probability distributions

All words are hyperlinks.

- [Normal]
- [Inverse Gaussian (a.k.a Wald)]
- [Beta]
- [Poisson]
- [Binomial]
- [Bernoulli]
- [Gamma]

A primer on probability distributions

All words are hyperlinks.

- [Normal]
- [Inverse Gaussian (a.k.a Wald)]
- [Beta]
- [Poisson]
- [Binomial]
- [Bernoulli]
- [Gamma]
- [Student]
- ...

The Bayesian paradigm

Introductory example

Consider observations $\mathbf{x} = (x_1, \ldots, x_n)$ generated from a probability distribution with density $f(\cdot|\theta)$.

Introductory example

Consider observations $\mathbf{x} = (x_1, \ldots, x_n)$ generated from a probability distribution with density $f(\cdot|\theta)$.

The associated likelihood is the inverted density:

$$\mathcal{L}(\theta|x) = f(x|\theta).$$

Example $f(\cdot|\theta) = \mathcal{N}(\theta, 1)$.

Bayes' Theorem

Bayes' Theorem

Inversion of probabilities a.k.a actualisation principle.

Bayes' Theorem

Inversion of probabilities a.k.a actualisation principle.

If A and B are events such that $\mathbb{P}(B) \neq 0$,

$$\begin{aligned}\mathbb{P}(A|B) &= \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c)} \\ &= \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}.\end{aligned}$$

(due to Thomas Bayes, published in 1764)

Who was Thomas Bayes?

Who was Thomas Bayes?

Reverend Thomas Bayes (ca. 1702–1761) Presbyterian minister in Kent from 1731. Election to the Royal Society based on a tract of 1736 where he defended the views and philosophy of Newton. Sole probability paper, "Essay Towards Solving a Problem in the Doctrine of Chances", published posthumously in 1763 and containing the seeds of Bayes' Theorem.

A new paradigm

Bayes introduces a whole new perspective.

A new paradigm

Bayes introduces a whole new perspective.

- Uncertainty on the parameter θ, modeled through a probability distribution π, called *prior distribution*.

A new paradigm

Bayes introduces a whole new perspective.

- Uncertainty on the parameter θ, modeled through a probability distribution π, called *prior distribution*.

- Inference based on the distribution of θ conditional on $\mathbf{X}$ $\pi(\theta|\mathbf{x})$, called *posterior distribution*

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{\int f(\mathbf{x}|\theta)\pi(\theta)\mathrm{d}\theta}.$$

A Bayesian model

. . . is made of a parametric (in this course) statistical model defined through its likelihood $f(\mathbf{x}|\theta)$ and a prior distribution on the parameter $\pi(\theta)$.

Consequences

- Semantic drift from unknown to random

Consequences

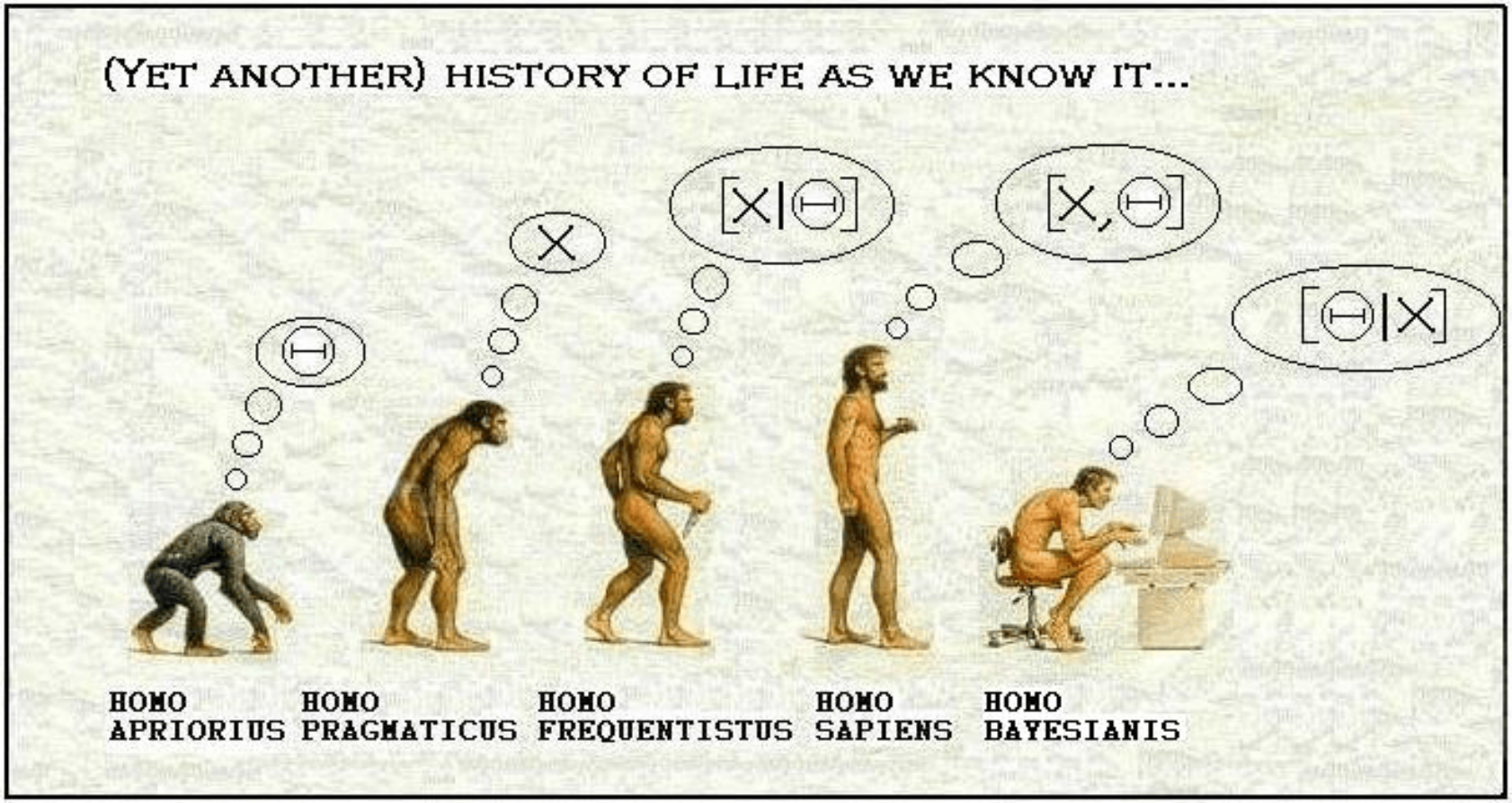
- Semantic drift from unknown to random

- Actualization of θ by extracting the information contained in the observation $\mathbf{x}$

Consequences

- Semantic drift from unknown to random
- Actualization of θ by extracting the information contained in the observation $\mathbf{x}$
- Allows incorporation of imperfect information in the decision process

The advantages of being a Bayesian



Distributions (1/2)

Given the likelihood $f(\mathbf{x}|\theta)$ and the prior $\pi(\theta)$, several distributions of interest:

Distributions (1/2)

Given the likelihood $f(\mathbf{x}|\theta)$ and the prior $\pi(\theta)$, several distributions of interest:

- The *joint distribution* of $(\theta, \mathbf{x})$

$$\varphi(\theta, \mathbf{x}) = f(\mathbf{x}|\theta)\pi(\theta).$$

Distributions (1/2)

Given the likelihood $f(\mathbf{x}|\theta)$ and the prior $\pi(\theta)$, several distributions of interest:

- The *joint distribution* of $(\theta, \mathbf{x})$

$$\varphi(\theta, \mathbf{x}) = f(\mathbf{x}|\theta)\pi(\theta).$$

- The *marginal distribution* of $\mathbf{x}$

$$m(\mathbf{x}) = \int \varphi(\theta, \mathbf{x}) \mathrm{d}\theta = \int f(\mathbf{x}|\theta)\pi(\theta)\mathrm{d}\theta.$$

Distributions (2/2)

Distributions (2/2)

- The *posterior distribution* of θ

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{\int f(\mathbf{x}|\theta)\pi(\theta)\mathrm{d}\theta} = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})}.$$

Distributions (2/2)

- The *posterior distribution* of θ

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{\int f(\mathbf{x}|\theta)\pi(\theta)\mathrm{d}\theta} = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})}.$$

- The *predictive distribution* of y when $y \sim g(\cdot|\theta, \mathbf{x})$

$$g(y|\mathbf{x}) = \int g(y|\theta, \mathbf{x})\pi(\theta|\mathbf{x})\mathrm{d}\theta.$$

A comprehensive example normal-normal

Assume that we model $\mathbf{x} \sim \mathcal{N}(\theta, 1)$ and use the prior $\theta \sim \mathcal{N}(a, 10)$.

A comprehensive example normal-normal

Assume that we model $\mathbf{x} \sim \mathcal{N}(\theta, 1)$ and use the prior $\theta \sim \mathcal{N}(a, 10)$.

$$\begin{aligned}
\pi(\theta|\mathbf{x}) &\propto f(\mathbf{x}|\theta)\pi(\theta) \\
&\propto \exp\left(-\frac{(\mathbf{x}-\theta)^2}{2} - \frac{(\theta-a)^2}{20}\right) \\
&\propto \exp\left(-\frac{11}{20}\theta^2 + \theta(\mathbf{x} + a/10)\right) \\
&\propto \exp\left(-\frac{11}{20}\left(\theta - \frac{10\mathbf{x}+a}{11}\right)^2\right)
\end{aligned}$$

A comprehensive example normal-normal

Assume that we model $\mathbf{x} \sim \mathcal{N}(\theta, 1)$ and use the prior $\theta \sim \mathcal{N}(a, 10)$.

$$\begin{aligned}
\pi(\theta|\mathbf{x}) &\propto f(\mathbf{x}|\theta)\pi(\theta) \\
&\propto \exp\left(-\frac{(\mathbf{x}-\theta)^2}{2} - \frac{(\theta-a)^2}{20}\right) \\
&\propto \exp\left(-\frac{11}{20}\theta^2 + \theta(\mathbf{x} + a/10)\right) \\
&\propto \exp\left(-\frac{11}{20}\left(\theta - \frac{10\mathbf{x}+a}{11}\right)^2\right)
\end{aligned}$$

which means $\theta|\mathbf{x} \sim \mathcal{N}\left(\frac{10x+a}{11}, \frac{10}{11}\right)$.

A comprehensive example uniform-binomial

Bayes' very own example: a billiard ball W rolls on a line of length one, with a uniform probability of stopping anywhere: W stops at p.

A comprehensive example uniform-binomial

Bayes' very own example: a billiard ball W rolls on a line of length one, with a uniform probability of stopping anywhere: W stops at p.

A second ball O then rolls n times under the same assumptions. X denotes the number of times the ball O stopped on the left of W.

A comprehensive example uniform-binomial

Bayes' very own example: a billiard ball W rolls on a line of length one, with a uniform probability of stopping anywhere: W stops at p.

A second ball O then rolls n times under the same assumptions. X denotes the number of times the ball O stopped on the left of W.

Bayes' question: given X, what inference can we make on p?

Mathematical translation

Derive the posterior distribution of p given X, when $p \sim \mathcal{U}(0,1)$ and $X \sim \mathcal{B}(n,p)$.

Resolution 1/2

Since

$$\mathbb{P}(X = x|p) = \binom{n}{x} p^x (1-p)^{n-x},$$

Resolution 1/2

Since

$$\mathbb{P}(X = x|p) = \binom{n}{x} p^x (1-p)^{n-x},$$

$$\mathbb{P}(a < p < b \text{ and } X = x) = \int_a^b \binom{n}{x} p^x (1-p)^{n-x} \mathrm{d}p,$$

Resolution 1/2

Since

$$\mathbb{P}(X = x|p) = \binom{n}{x} p^x (1-p)^{n-x},$$

$$\mathbb{P}(a < p < b \text{ and } X = x) = \int_a^b \binom{n}{x} p^x (1-p)^{n-x} \mathrm{d}p,$$

and

$$\mathbb{P}(X = x) = \int_0^1 \binom{n}{x} p^x (1-p)^{n-x} \mathrm{d}p,$$

Resolution 2/2

then

$$\mathbb{P}(a < p < b | X = x) = \frac{\int_a^b \binom{n}{x} p^x (1-p)^{n-x} \mathrm{d}p}{\int_0^1 \binom{n}{x} p^x (1-p)^{n-x} \mathrm{d}p}$$
$$= \frac{\int_a^b \binom{n}{x} p^x (1-p)^{n-x} \mathrm{d}p}{\mathcal{B}(x+1, n-x+1)},$$

i.e., $p|x \sim \mathcal{B}(x+1, n-x+1)$.
(Beta distribution)

Pour se remettre dans le bain

Pour se remettre dans le bain

1. Quelle est la différence entre *statistical learning* et *machine learning* ?
2. Donner la définition d'un algorithme d'apprentissage.
3. Quels sont les quatre grands types d'apprentissage ?
4. Que définissent *fat data* et *tall data* ?
5. Comment compare-t-on les performances d'algorithmes d'apprentissage ?
6. Quelle notion est souvent cruciale au moment de choisir une bonne fonction de perte ? En donner la définition.
7. Donner quelques exemples de fonctions de perte.
8. Enoncer le théorème de Bayes.
9. Quel est le rôle de la distribution *a priori* ?
10. Donner la définition d'un modèle bayésien.
11. Quelles sont les quatre distributions importantes en bayésien ?

The posterior distribution

$$\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta)$$

The posterior distribution

$$\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta)$$

- Operates conditional upon the observations.

The posterior distribution

$$\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta)$$

- Operates conditional upon the observations.
- Integrates simultaneously prior knowledge and information brought by data.

The posterior distribution

$$\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta)$$

- Operates conditional upon the observations.
- Integrates simultaneously prior knowledge and information brought by data.
- Coherent and complete inferential scope and unique motor of inference.

The posterior distribution

$$\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta)$$

- Operates conditional upon the observations.
- Integrates simultaneously prior knowledge and information brought by data.
- Coherent and complete inferential scope and unique motor of inference.
- Usually known up to a constant! $m(\mathbf{x})$ may be intractable.

Prior distributions

There is no such thing as *the* prior distribution!

Prior distributions

There is no such thing as *the* prior distribution!

Usually encapsulates *prior* knowledge on θ.

Prior distributions

There is no such thing as *the* prior distribution!

Usually encapsulates *prior* knowledge on θ.

Vague priors (such as $\theta \sim \mathcal{N}(0, 100)$).

Prior distributions

There is no such thing as *the* prior distribution!

Usually encapsulates *prior* knowledge on θ.

Vague priors (such as $\theta \sim \mathcal{N}(0, 100)$).

Improper priors: $\int \pi(\theta) \mathrm{d}\theta = +\infty$.

Prior distributions

There is no such thing as *the* prior distribution!

Usually encapsulates *prior* knowledge on θ.

Vague priors (such as $\theta \sim \mathcal{N}(0, 100)$).

Improper priors: $\int \pi(\theta) \mathrm{d}\theta = +\infty$.

A prior on θ may depend on additional parameters: those are called hyperparameters.

Conjugacy and explicit calculus of posteriors

Conjugate priors are a specific family of distributions with nice analytical properties.

Conjugacy and explicit calculus of posteriors

Conjugate priors are a specific family of distributions with nice analytical properties.

A family $\mathcal{F}$ of probability distributions is *conjugate* for a likelihood $f(x|\theta)$ if for every $\pi \in \mathcal{F}$, the posterior distribution $\pi(\theta|x)$ also belongs to $\mathcal{F}$.

Conjugacy and explicit calculus of posteriors

Conjugate priors are a specific family of distributions with nice analytical properties.

A family $\mathcal{F}$ of probability distributions is *conjugate* for a likelihood $f(x|\theta)$ if for every $\pi \in \mathcal{F}$, the posterior distribution $\pi(\theta|x)$ also belongs to $\mathcal{F}$.

Only of interest when $\mathcal{F}$ is parameterized: switching from the prior to the posterior is reduced to an update of parameters.

Advantages

- Limited/finite information conveyed by data x

Advantages

- Limited/finite information conveyed by data x
- Preservation of the structure of the prior $\pi(\theta)$

Advantages

- Limited/finite information conveyed by data x
- Preservation of the structure of the prior $\pi(\theta)$
- Exchangeability

Advantages

- Limited/finite information conveyed by data x
- Preservation of the structure of the prior $\pi(\theta)$
- Exchangeability
- Allows for generation of "virtual observations"

Advantages

- Limited/finite information conveyed by data x
- Preservation of the structure of the prior $\pi(\theta)$
- Exchangeability
- Allows for generation of "virtual observations"
- Most importantly: tractability and simplicity

Exponential families

The family of distributions

$$f(x|\theta) = C(\theta)h(x)\exp(R(\theta)T(x))$$

is called an *exponential family*.

Exponential families

The family of distributions

$$f(x|\theta) = C(\theta)h(x)\exp(R(\theta)T(x))$$

is called an *exponential family*. When

$$f(x|\theta) = h(x)\exp(-\theta x - \psi(\theta))$$

the family is said to be *natural*.

Exponential families

The family of distributions

$$f(x|\theta) = C(\theta)h(x)\exp(R(\theta)T(x))$$

is called an *exponential family*. When

$$f(x|\theta) = h(x)\exp(-\theta x - \psi(\theta))$$

the family is said to be *natural*.

Main interest: allow for conjugate priors

$$\pi(\theta|\mu,\lambda) = K(\mu,\lambda)\exp(\theta\mu - \lambda\psi(\theta)), \quad \lambda > 0.$$

Classical exponential families and conjugate priors

Classical exponential families and conjugate priors

[Examples of exponential families]

Classical exponential families and conjugate priors

[Examples of exponential families]

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$

Classical exponential families and conjugate priors

[Examples of exponential families]

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Normal	Normal	Normal
$\mathcal{N}(\theta, \sigma^2)$	$\mathcal{N}(\mu, \tau^2)$	$\mathcal{N}(\rho(\sigma^2\mu + \tau^2 x), \rho\sigma^2\tau^2)$
		$\rho^{-1} = \sigma^2 + \tau^2$

Classical exponential families and conjugate priors

[Examples of exponential families]

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Normal $\mathcal{N}(\theta, \sigma^2)$	Normal $\mathcal{N}(\mu, \tau^2)$	Normal $\mathcal{N}(\rho(\sigma^2\mu + \tau^2 x), \rho\sigma^2\tau^2)$ $\rho^{-1} = \sigma^2 + \tau^2$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + x, \beta + 1)$

Classical exponential families and conjugate priors

[Examples of exponential families]

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Normal $\mathcal{N}(\theta, \sigma^2)$	Normal $\mathcal{N}(\mu, \tau^2)$	Normal $\mathcal{N}(\rho(\sigma^2\mu + \tau^2 x), \rho\sigma^2\tau^2)$ $\rho^{-1} = \sigma^2 + \tau^2$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + x, \beta + 1)$
Gamma $\mathcal{G}(\nu, \theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + \nu, \beta + x)$

Classical exponential families and conjugate priors

[Examples of exponential families]

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Normal $\mathcal{N}(\theta, \sigma^2)$	Normal $\mathcal{N}(\mu, \tau^2)$	Normal $\mathcal{N}(\rho(\sigma^2\mu + \tau^2 x), \rho\sigma^2\tau^2)$ $\rho^{-1} = \sigma^2 + \tau^2$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + x, \beta + 1)$
Gamma $\mathcal{G}(\nu, \theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + \nu, \beta + x)$
Binomial $\mathcal{B}(n, \theta)$	Beta $\mathcal{B}e(\alpha, \beta)$	Beta $\mathcal{B}e(\alpha + x, \beta + n - x)$

Classical exponential families and conjugate priors

[Examples of exponential families]

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Normal $\mathcal{N}(\theta, \sigma^2)$	Normal $\mathcal{N}(\mu, \tau^2)$	Normal $\mathcal{N}(\rho(\sigma^2\mu + \tau^2 x), \rho\sigma^2\tau^2)$ $\rho^{-1} = \sigma^2 + \tau^2$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + x, \beta + 1)$
Gamma $\mathcal{G}(\nu, \theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + \nu, \beta + x)$
Binomial $\mathcal{B}(n, \theta)$	Beta $\mathcal{B}e(\alpha, \beta)$	Beta $\mathcal{B}e(\alpha + x, \beta + n - x)$
Negative binomial $\mathrm{Neg}(m, \theta)$	Beta $\mathcal{B}e(\alpha, \beta)$	Beta $\mathcal{B}e(\alpha + m, \beta + x)$

Classical exponential families and conjugate priors

[Examples of exponential families]

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Normal $\mathcal{N}(\theta, \sigma^2)$	Normal $\mathcal{N}(\mu, \tau^2)$	Normal $\mathcal{N}(\rho(\sigma^2\mu + \tau^2 x), \rho\sigma^2\tau^2)$ $\rho^{-1} = \sigma^2 + \tau^2$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + x, \beta + 1)$
Gamma $\mathcal{G}(\nu, \theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + \nu, \beta + x)$
Binomial $\mathcal{B}(n, \theta)$	Beta $\mathcal{B}e(\alpha, \beta)$	Beta $\mathcal{B}e(\alpha + x, \beta + n - x)$
Negative binomial $\mathrm{Neg}(m, \theta)$	Beta $\mathcal{B}e(\alpha, \beta)$	Beta $\mathcal{B}e(\alpha + m, \beta + x)$
Multinomial $\mathcal{M}_k(\theta_1, \ldots, \theta_k)$	Dirichlet $\mathcal{D}(\alpha_1, \ldots, \alpha_k)$	Dirichlet $\mathcal{D}(\alpha_1 + x_1, \ldots, \alpha_k + x_k)$

Classical exponential families and conjugate priors

[Examples of exponential families]

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Normal $\mathcal{N}(\theta, \sigma^2)$	Normal $\mathcal{N}(\mu, \tau^2)$	Normal $\mathcal{N}(\rho(\sigma^2\mu + \tau^2 x), \rho\sigma^2\tau^2)$ $\rho^{-1} = \sigma^2 + \tau^2$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + x, \beta + 1)$
Gamma $\mathcal{G}(\nu, \theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + \nu, \beta + x)$
Binomial $\mathcal{B}(n, \theta)$	Beta $\mathcal{B}e(\alpha, \beta)$	Beta $\mathcal{B}e(\alpha + x, \beta + n - x)$
Negative binomial $\mathrm{Neg}(m, \theta)$	Beta $\mathcal{B}e(\alpha, \beta)$	Beta $\mathcal{B}e(\alpha + m, \beta + x)$
Multinomial $\mathcal{M}_k(\theta_1, \ldots, \theta_k)$	Dirichlet $\mathcal{D}(\alpha_1, \ldots, \alpha_k)$	Dirichlet $\mathcal{D}(\alpha_1 + x_1, \ldots, \alpha_k + x_k)$
Normal $\mathcal{N}(\mu, 1/\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + 1/2, \beta + (\mu - x)^2/2)$

Non-informative priors: Jeffreys priors

Based on Fisher information:

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \log \mathcal{L}(\theta|x)}{\partial \theta^2}\right].$$

Non-informative priors: Jeffreys priors

Based on Fisher information:

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \log \mathcal{L}(\theta|x)}{\partial \theta^2}\right].$$

Jeffreys prior is defined as

$$\pi^\star(\theta) \propto \det(I(\theta))^{1/2}.$$

Non-informative priors: Jeffreys priors

Based on Fisher information:

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \log \mathcal{L}(\theta|x)}{\partial \theta^2}\right].$$

Jeffreys prior is defined as

$$\pi^\star(\theta) \propto \det(I(\theta))^{1/2}.$$

Pros & Cons:

Non-informative priors: Jeffreys priors

Based on Fisher information:

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \log \mathcal{L}(\theta|x)}{\partial \theta^2}\right].$$

Jeffreys prior is defined as

$$\pi^\star(\theta) \propto \det(I(\theta))^{1/2}.$$

Pros & Cons:

- Relates to information theory

Non-informative priors: Jeffreys priors

Based on Fisher information:

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \log \mathcal{L}(\theta|x)}{\partial \theta^2}\right].$$

Jeffreys prior is defined as

$$\pi^\star(\theta) \propto \det(I(\theta))^{1/2}.$$

Pros & Cons:

- Relates to information theory
- Parameterization invariant

Non-informative priors: Jeffreys priors

Based on Fisher information:

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \log \mathcal{L}(\theta|x)}{\partial \theta^2}\right].$$

Jeffreys prior is defined as

$$\pi^\star(\theta) \propto \det(I(\theta))^{1/2}.$$

Pros & Cons:

- Relates to information theory
- Parameterization invariant
- Suffers from dimensionality curse

Non-informative priors: Jeffreys priors

Based on Fisher information:

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \log \mathcal{L}(\theta|x)}{\partial \theta^2}\right].$$

Jeffreys prior is defined as

$$\pi^\star(\theta) \propto \det(I(\theta))^{1/2}.$$

Pros & Cons:

- Relates to information theory
- Parameterization invariant
- Suffers from dimensionality curse
- Depends on data: incoherence with the likelihood principle

Example

If $x \sim \mathcal{B}(n, \theta)$, Jeffreys' prior is

$$\pi(\theta) \propto \mathcal{B}e(1/2, 1/2).$$

If $n \sim \mathrm{Neg}(x, \theta)$, Jeffreys' prior is

$$\pi(\theta) \propto \theta^{-1}(1-\theta)^{-1/2}$$

Non-informative priors: Laplace priors

With a finite set $\{\theta_1, \ldots, \theta_p\}$, uniform prior $\pi(\theta_i) = 1/p$.

Non-informative priors: Laplace priors

With a finite set $\{\theta_1, \ldots, \theta_p\}$, uniform prior $\pi(\theta_i) = 1/p$.

Continuous extension: $\pi(\theta) \propto 1$. This is no longer a probability distribution yet if $\int f(x|\theta)\mathrm{d}\theta < +\infty$, the posterior is well-defined as a probability distribution. Modeling is crucial. Weakness: lack of reparameterization invariance.

Pour se remettre dans le bain

- Expliquer l'intérêt de la conjugaison, et en donner la définition.
- Quel est l'intérêt d'utiliser une vraisemblance issue d'une famille exponentielle naturelle ?
- Donner des exemples de lois conjuguées.
- Donner deux exemples de méthodes de construction de priors non-informatifs. Quelles sont les limites de ces méthodes ?