

Data science et Big Data

Définitions et concepts - Nouveaux enjeux scientifiques

Emmanuel Bacry

Researcher at CNRS

Associate Professor at Ecole Polytechnique

Head of the "Data Science Initiative"

Centre de Mathématiques Appliquées

emmanuel.bacry@polytechnique.fr

<http://www.cmap.polytechnique.fr/~bacry>



- **Data Science** : Extraction of information from data
- Extraction can be difficult :
 - Just because the problem is complex !
 - Speed
 - Variety of the data
 - Volume (“Big” Data)
 - acquisition, cleaning, storage, analysis, visualization, ...
 - ...

- Data are everywhere
- Big or small data, with or without variety, ...

⇒ **Recent developments in Computer Science and Statistics**
from both academic and industrial research

- High computing power + adapted tools
 - Distributed computing
 - High Performance Computers (HPC)
 - Graphical Processor Units (GPU)
 - Massively Parallel Processors (MPP)
 - Field Programmable Gate-Array (FPGA)
- Statistical (Machine) learning (large dimension)
 - From examples : supervised learning
 - Unsupervised learning

Recommandation system

An example of supervised learning

More Ideas Based on Your Browsing History

You looked at



[Thriving in the Knowledge Age: New...](#) Paperback by John H. Falk
~~\$29.95~~

You might also consider



[Museum Administration: An Introduction](#) Paperback by Hugh H. Genoways
~~\$31.95~~ **\$28.75**



[Exhibit Labels: An Interpretive Approach](#) Paperback by Beverly Serrell
~~\$34.95~~ **\$27.85**

› [Find similar items](#)

Recommendations don't have to be about showing you more of the same...

Netflix, Amazon, Google News ...

Matrix completion

⇒ **Supervised learning**

Web Search Engines

An example of unsupervised learning

- Huge quantity of data $\simeq$ more than 1000 billion web pages
- Dynamic data
- Speed
- “Ranking” algorithm :
 - Sorting according to the audience (number of clicks, ...)
 - Sorting by the popularity (number of links, “Page Rank” algorithm)
 - **Sorting by relevance** (LDA : Latent Dirichlet Allocation, ...)
 $\implies$ **Unsupervised learning**

- Strong interdisciplinarity
- Academic world $\longleftrightarrow$ public institutions and/or private companies (data access and concrete applications)

Challenges in computer science

- New languages/frameworks MPP/GPU/FPGA
→ Hadoop, Spark, GraphLab, Cuda, ...
- New database system : noSQL, ...

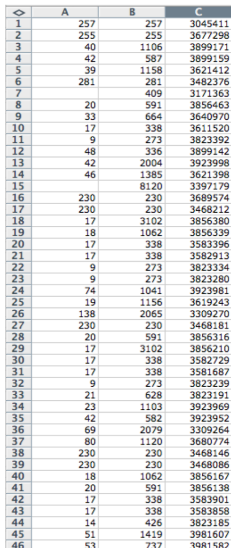
Challenges in Machine learning / Statistics

- Supervised learning relatively mastered :
 - Scalability ?
→ Reducing number of passes over the data
 - Large dimension
→ Reducing the dimension using parcimony hypothesis

Data science

Reducing the dimension and Parcimony

d = dimension



	A	B	C
1	257	257	3045411
2	255	255	3677298
3	40	1106	3899171
4	42	587	3899159
5	39	1158	3621412
6	281	281	3482376
7		409	3171363
8	20	591	3856463
9	33	664	3640970
10	17	338	3611520
11	9	273	3823392
12	48	336	3899142
13	42	2004	3923998
14	46	1385	3621398
15		8120	3397179
16	230	230	3689574
17	230	230	3468212
18	17	3102	3856380
19	18	1062	3856339
20	17	338	3583396
21	17	338	3582913
22	9	273	3823334
23	9	273	3823280
24	74	1041	3923981
25	19	1156	3619243
26	138	2065	3309270
27	230	230	3468181
28	20	591	3856316
29	17	3102	3856210
30	17	338	3582729
31	17	338	3581687
32	9	273	3823239
33	21	628	3823191
34	23	1103	3923969
35	42	582	3923952
36	69	2079	3309264
37	80	1120	3680774
38	230	230	3468146
39	230	230	3468086
40	18	1062	3856167
41	20	591	3856138
42	17	338	3583901
43	17	338	3583858
44	14	426	3823185
45	51	1419	3981607
46	53	737	3981582

« Classic » statistics

n large and d small

n = Number of observations

New Data

	AN	AI	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO	AS	AL	AM	AN	AO	AP	AO</
--	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	------

d = dimension

• d large $\Rightarrow$ several possibilities :

$\rightarrow d \ll n$

$\rightarrow d \simeq n$

$\rightarrow d \gg n$

$\rightarrow d$ increases with n

New Data

ID	AN	AI	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	BA	BB	BC	BD
1	PHI_MED_C	PHI_MED_L	PHI_MED_R	PHI_MED_N	PHI_MED_P	PHI_MED_S	PHI_MED_T	PHI_MED_U	PHI_MED_V	PHI_MED_W	PHI_MED_X	PHI_MED_Y	PHI_MED_Z	PHI_MED_1	PHI_MED_2	PHI_MED_3	PHI_MED_4	PHI_MED_5	PHI_MED_6	PHI_MED_7	PHI_MED_8
2																					
3																					
4	OLYCO THYR																				
5	LACTULOSE																				
6																					
7																					
8	RAMITONE																				
9	ELUERS																				
10	MUCON																				
11	GLUCALOSE																				
12	MOVICOL																				
13	LACTULOSE																				
14	DOMPERIDON																				
15																					
16	OSASTO																				
17	RAMITONE																				
18	CAMEFERNE																				
19	SOLIGEL																				
20	CHLORHEX																				
21	GLUCALOSE																				
22	GLUCALOSE																				
23	LACTULOSE																				
24	LACTULOSE																				
25	DOMPERIDON																				
26	LACTULOSE																				
27	OSASTO																				
28	OSASTO																				
29	OSASTO																				

d = dimension

- d large

- **Parcimony hypothesis**

→ Efficient recent algorithms

→ Black box ? (Regulation ?)

- Strong interdisciplinarity
- Link with public institutions and/or private companies (data access and concrete applications)

Challenges in computer science

- New languages/frameworks MPP/GPU/Cloud
→ Hadoop, Spark, GraphLab, Cuda, ...
- New database system : noSQL, ...

Challenges in Machine learning / Statistics

- Supervised learning relatively mastered :
 - Scalability ?
→ Reducing number of passes over the data
 - Large dimension
- **Unsupervised learning**
→ still many things to be done !

<http://polytechnique.fr/datascience>

- Teaching (3A, Master)
- Data Science Starter Program (twice a year, 10 weeks, 2 days/week)
- Federate different labs (CMAP, LIX)
- Develop strong links with public institutions/private companies
 - AXA, Orange, Thales, ETALAB, ...
 - Public health : APHP, HEGP, Tenon, **CNAMTS**, ...



- **SNIIRAM :**

- Accountant database (DCIR) +
- PMSI database +
- Hypocrate database (ALD) (Long -term illness) +
- ...

- **SNIIRAM $\simeq$ One of the largest health database in the world**

- 65 million persons
- $\simeq$ 500 Terabytes !

Potential high impact

- **On health**

- 2013 : link between 3rd generation contraceptive pill and pulmonary embolism risk
- 2014 : Cartography of 54 pathologies (HIV $\simeq$ 50 criteria)

- **On economics** (french public health budget $\simeq$ 170 billion euros/year)

- cartography to quantify the cost of different pathologies

- **Partnership over 3 years (2015-2017)**
Polytechnique-CNAMTS
- **CNAMTS “opens” SNIIRAM to researchers of the Data Science Initiative of Polytechnique**
- **Many research themes**
 - Identification of useful factors for analyzing medico-economical health pathways
 - Weak signal detection in pharmaco-epidemiology
 - Fraud detection
 - ...

⇒ **PreProcessing of the database**

- SNIIRAM : Oracle relational database with $\simeq$ 800 tables !
- Accountant oriented

→ Flatenning (User oriented, Institution oriented, ...)

⇒ **Event structuration**

- “Low level” structure $\simeq$ original SNIIRAM structure
- “High level” structure (made with experts)
 - Pathology definitions ?
 - Structuration of treatments (periodic, ...) ?
 - Structuration of medecines (same molecules, ...) ?
 - ...

⇒ **Designing “scalable” algorithms of Machine Learning**