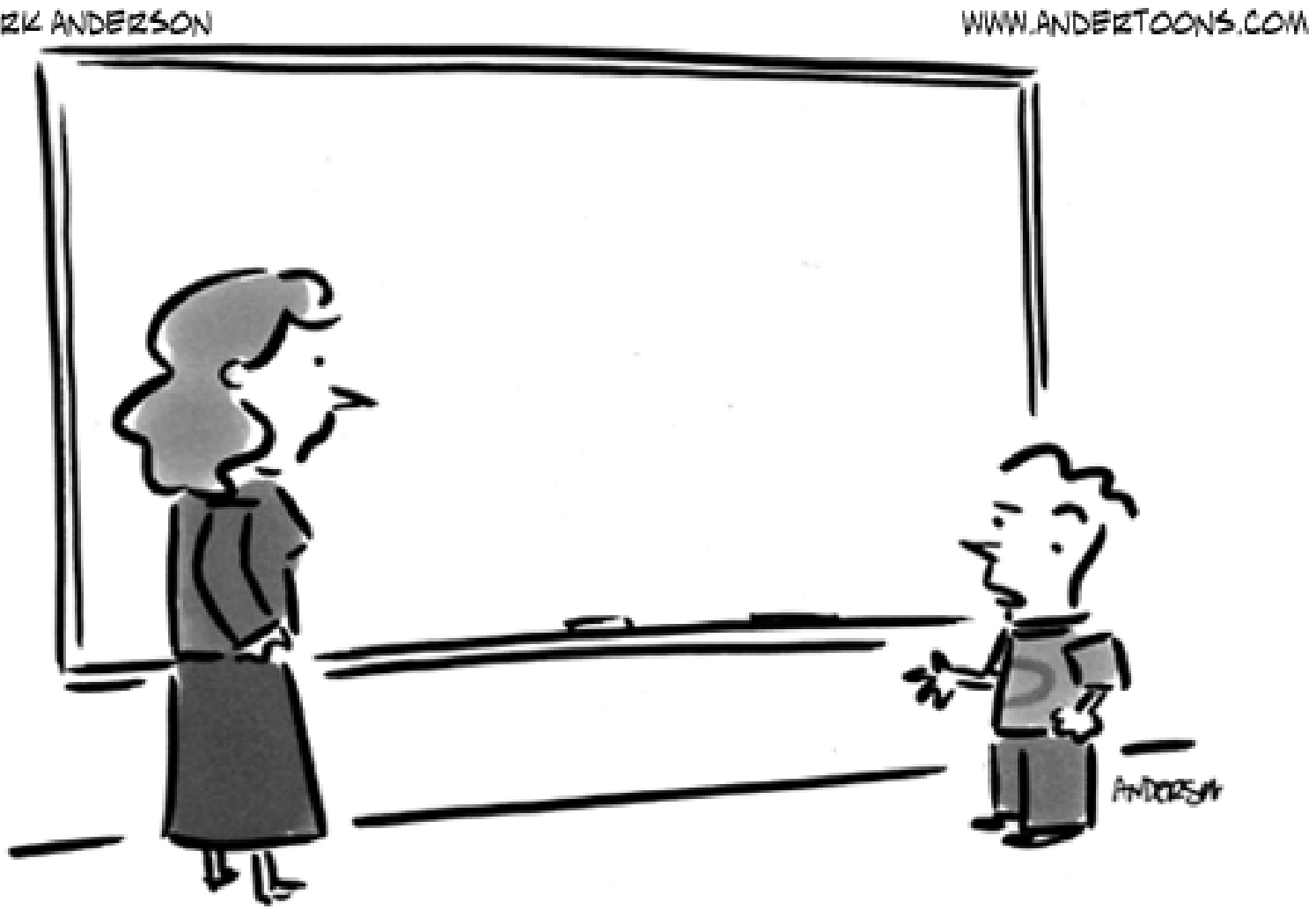
Introduction

Ludovic d'Estampes

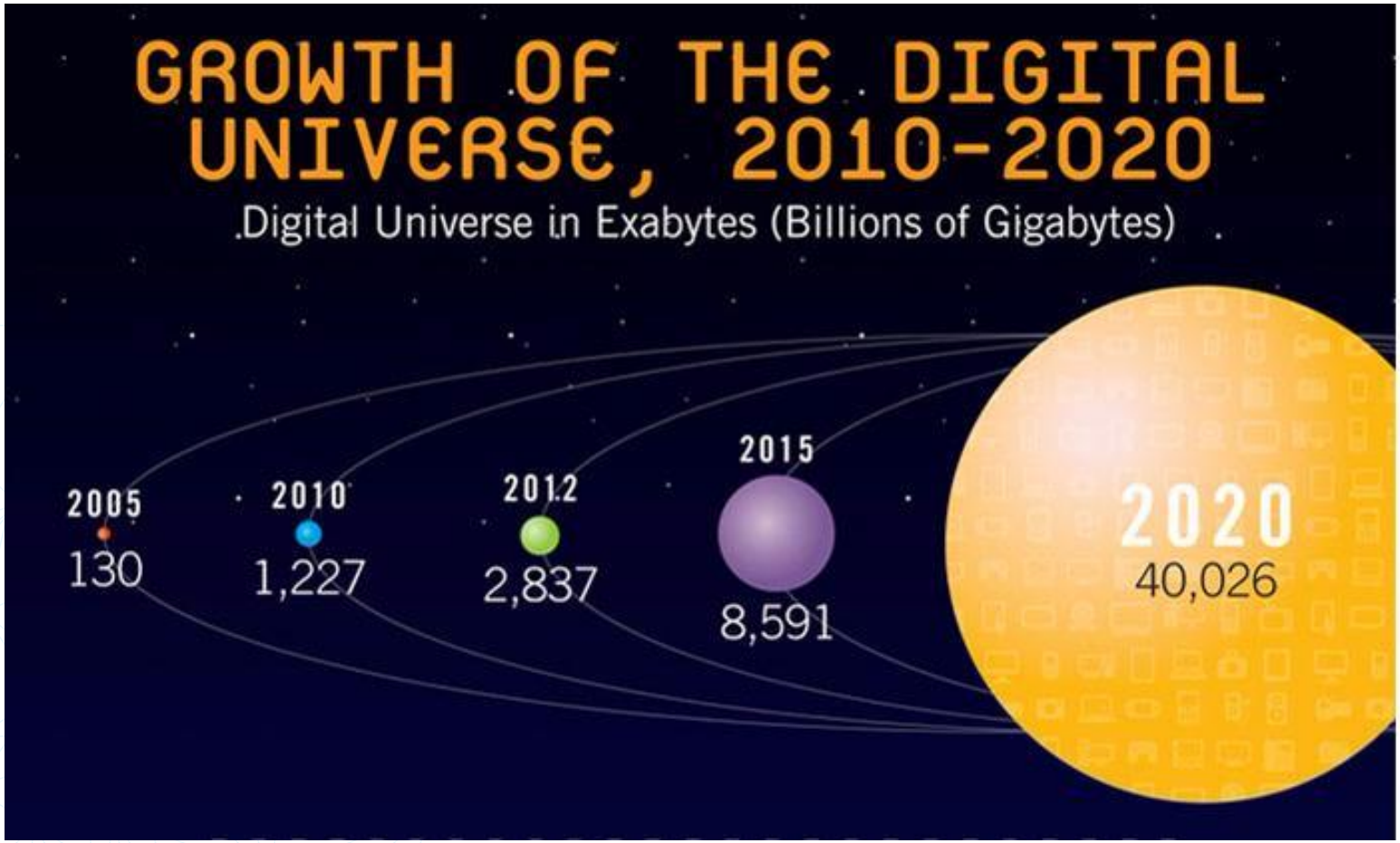
Mineure - Science des données pour l'ingénieur

January 2024

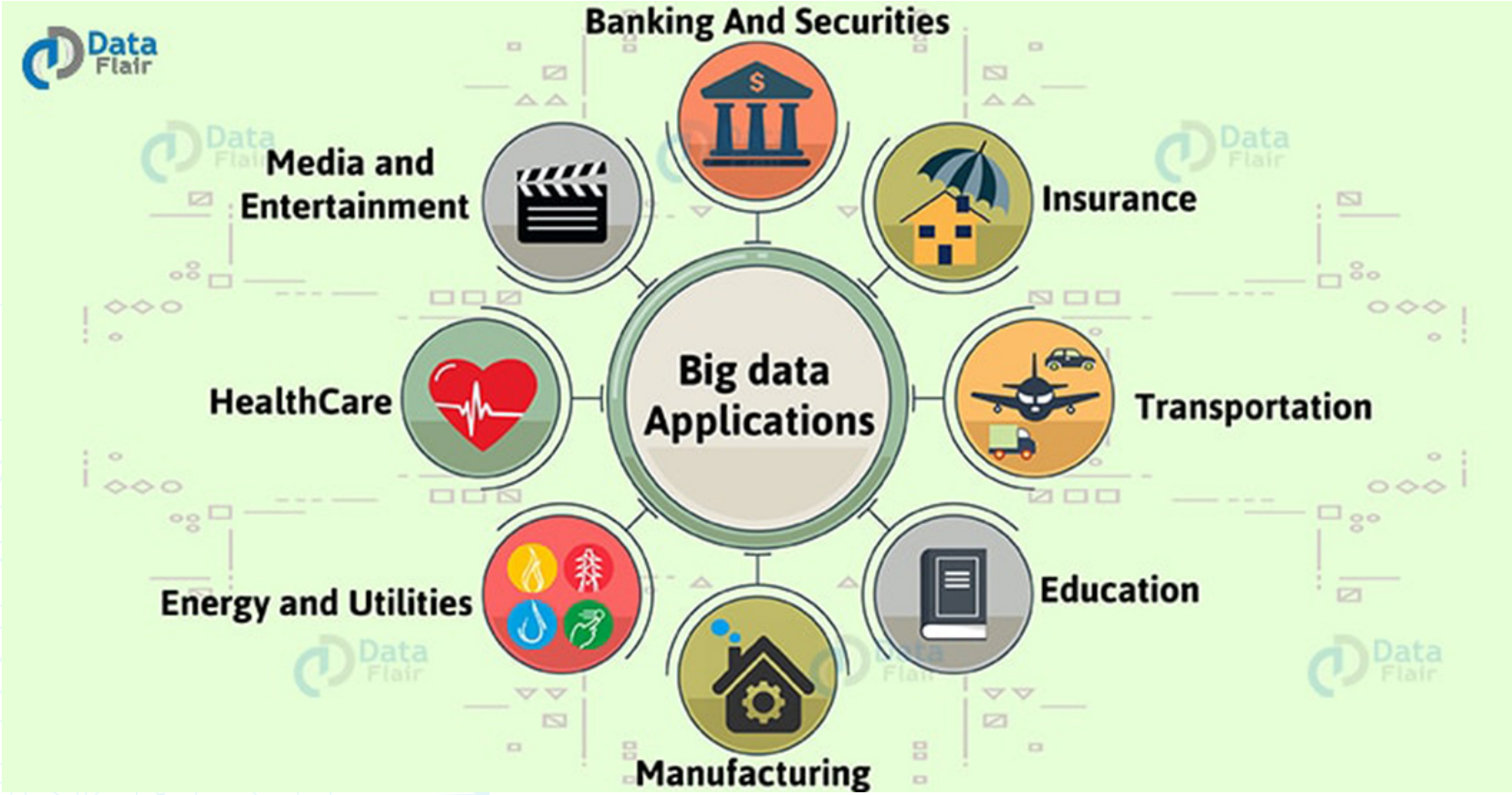


"Before I write my name on the board, I'll need to know how you're planning to use that data."

Data sources and quantities



Big data applications



What is machine Learning?

- Machine learning can be defined as computational methods using experience to improve performance or to make accurate predictions. (cf. M. Mohri, A. Rostamizadeh and A. Talwalkar: Foundations of Machine Learning)
- Experience refers to the past information available to the learner, which typically takes the form of electronic data collected and made available for analysis.
- It is also assumed that there is a link within the data and that our algorithms will help us find it. This is a strong assumption: by default, there is usually no link in the data.
- Tom Mitchell (1998) : *A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.*

What is the difference between machine learning and statistics?

- According to Wasserman, a professor in both the Department of Statistics and in the Machine Learning Department at Carnegie Mellon

  "The short answer is: None. They are . . . concerned with the same question: how do we learn from data?"

- But Wasserman notes that if you look at some of the details, there is a "more nuanced" answer that reveals minor differences:
  - Statistics emphasizes formal statistical inference (confidence intervals, hypothesis tests, optimal estimators) in low dimensional problems.
  - Machine Learning emphasizes high dimensional prediction problems.

What is the difference between machine learning and statistics?

- Furthermore, ML is more focused on making accurate predictions; making good predictions trumps more formal considerations like testing assumptions, etc.
- Go a step further, we can also state that ML isn't just more focused on making predictions, but is more focused on building software systems that make predictions.
- So, some people argue that ML is more of an "engineering discipline" whereas statistics is more of a "mathematical" discipline.
- ML and statistics tend to favour different tools (Python, R)
- ML typically work on "bigger" data than statistics

What is the difference between machine learning and statistics?

Professor Rob Tibshirani created a glossary comparing several major terms in machine learning vs statistics.

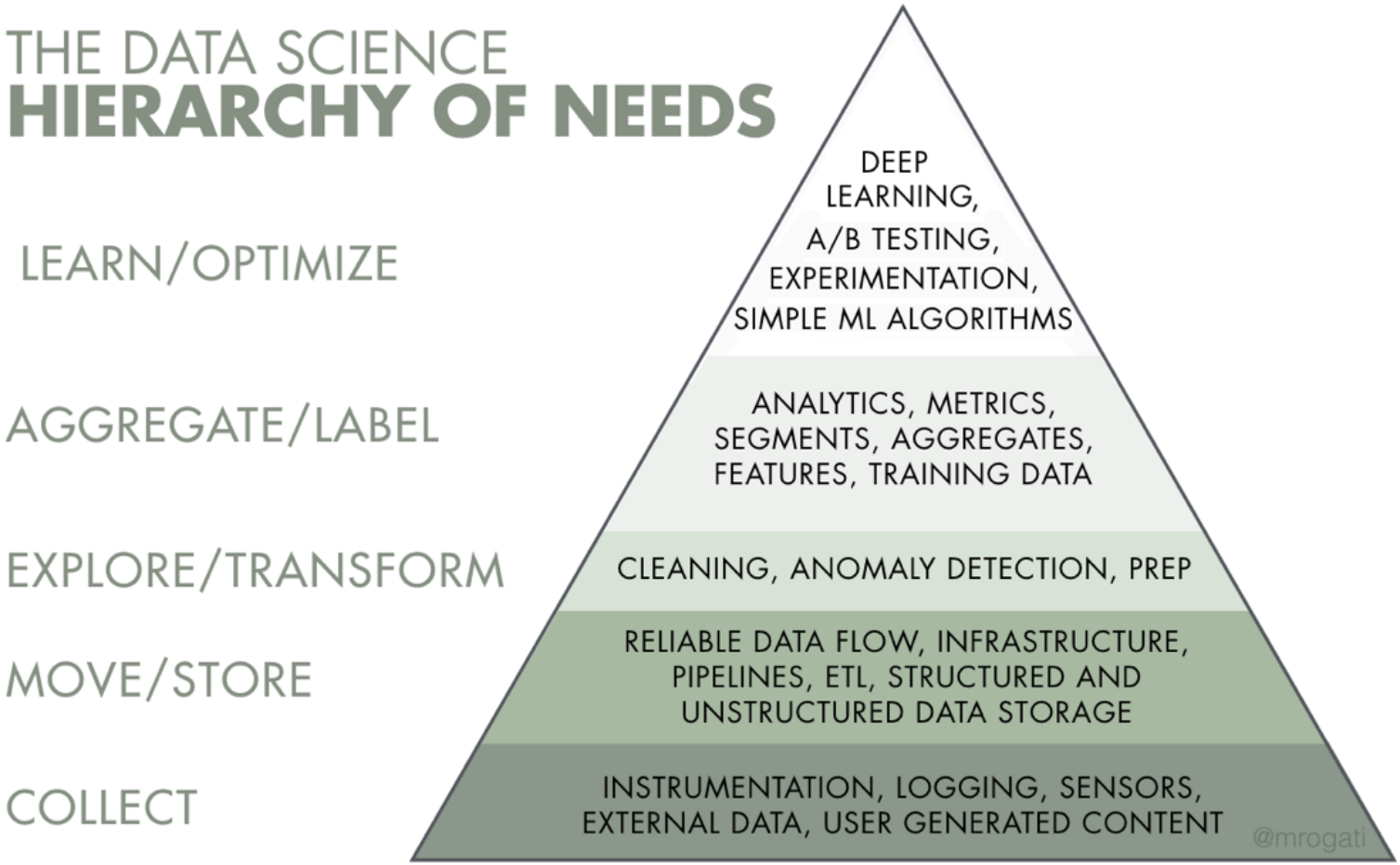
Machine Learning	Statistics
network,graphs	model
weights	parameters
learning	fitting
generalization test	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant $\$1,000,000$	large grant $\$50,000$
nice place to have a meeting: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

Objectives of the DATA's minor

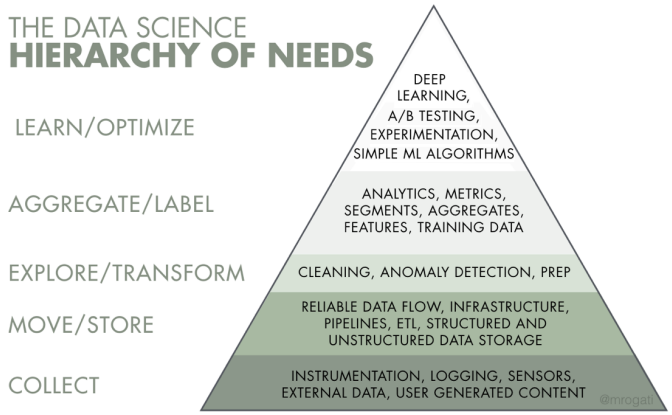
The general objective of this minor is not to train "data scientists" but to complete the IENAC training in order to:

- provide IENAC with data science skills that enable them to work with "data scientists" while providing their air transport skills;
- prepare for the evolution of the engineering profession due to the increasing use of big data;
- enable work in innovation on the use of big data in the aviation field.

Hierarchy of needs

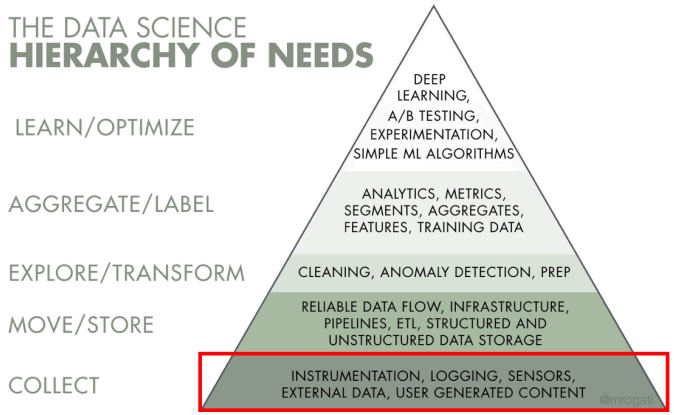


Hierarchy of needs



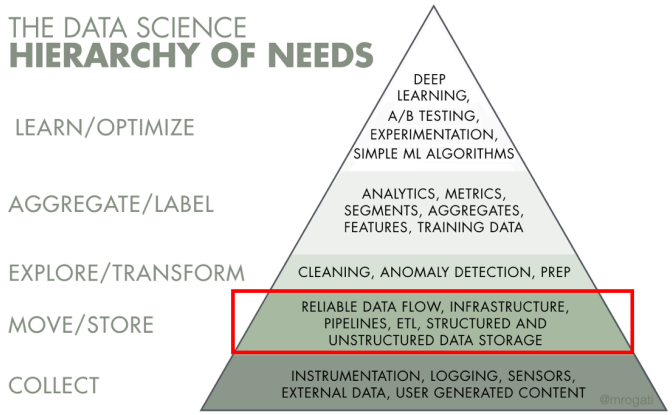
Teachers: Richard Alligier, Nicolas Couellan, Ludovic d'Estampes, David Gianazza, Laurent Lapasset, Paul Rochet

Hierarchy of needs



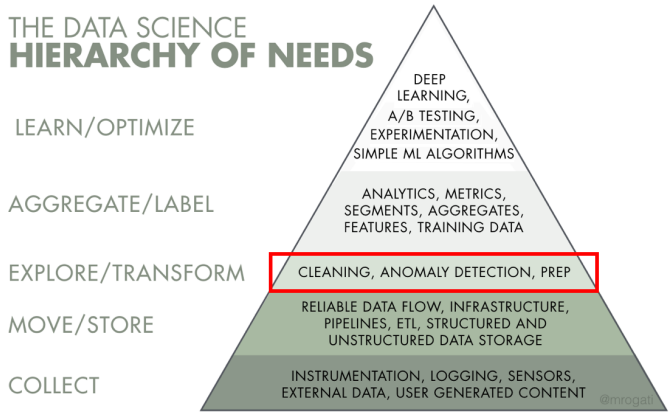
- Courses: Big Data (Laurent)

Hierarchy of needs



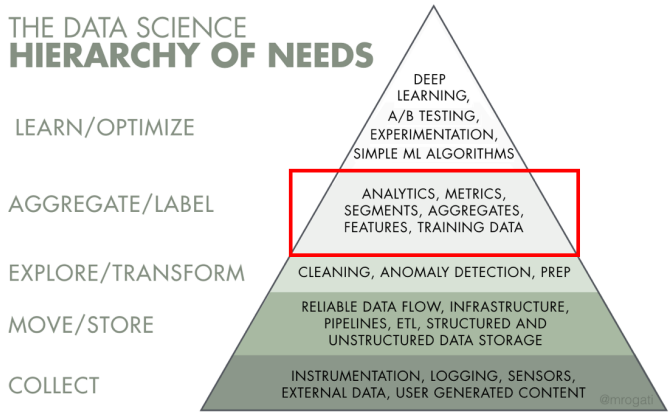
- Courses: Cloud (Laurent), Hadoop-Spark (Laurent), Kubernetes (Laurent), OpenStack (Laurent)

Hierarchy of needs



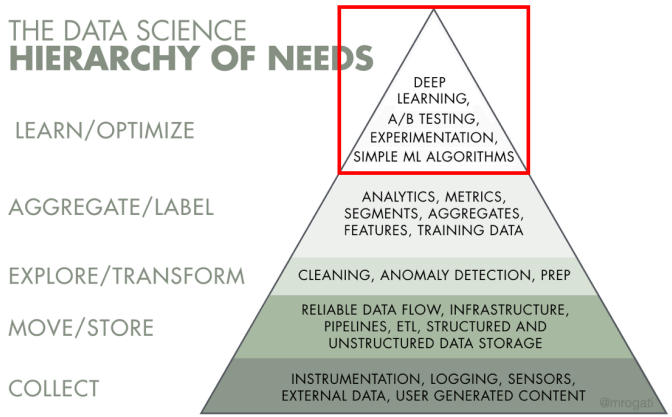
- Courses: Data preparation and exploration (Ludovic), Multidimensional analysis (Ludovic)
- Labs: Data munging/wrangling, Data visualisation, PCA and HAC

Hierarchy of needs



- Courses: Principles and methodology of Machine Learning (David)
- Labs: Risk and bias variance

Hierarchy of needs



- Courses: Linear regression models (Paul), CART-random forest (Richard), Gradient boosting (Richard), Support vector machine (Nicolas), Optimisation for machine learning (Nicolas), Virtualisation (Laurent), Docker (Laurent), Neural networks (David), Deep learning (Nicolas), NLP (Laurent)
- Labs: Linear regression models, CART-random forest, Gradient boosting, Support vector machine, Virtualisation, Docker, Neural networks, Deep learning, NLP

Conferences and Evaluation

Conferences

4 business driven conferences

Evaluation

- Reverse pedagogy: presentation of methods not studied or extension of methods studied (two sessions): coeff. 3.3
- Exam across all methods: coeff. 3.3
- Typical project "challenge" on wind turbines: coeff. 3.4
- Mandatory attendance at conferences