



Essentials of data science for managers

Volume 2: From artificial intelligence to business applications

Clément Levallois



ExpData Press

Essentials of data science for managers

*Volume 2: From Artificial Intelligence
to Business Applications*

Clément Levallois

 **ExpData Press**

Saint-Etienne

ESSENTIALS OF DATA SCIENCE FOR MANAGERS

by Clément Levallois

Copyright © 2018 Clément Levallois. All rights reserved.

Published by Peecho, Rokin 75-5, 1012KL Amsterdam, Netherlands

April 2018: first edition

Revision history for the first release:

2018-04-01: first release

From the same author:

Levallois, C. et al, eds. (2015) . *Twitter for Research*. Ecully: EMLYON Press.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and ExpData Press was aware of the trademark claim, the designations have been printed in caps or initial caps.

For Manon, Léon and Tristan

Table of Contents

- Preface 1
 - A textbook for managers..... 1
 - Is this textbook too technical or too easy for me? 1
- Machine learning, data science and artificial intelligence 2
 - 1. Explaining machine learning in simple terms 2
 - a. A comparison with classic statistics 2
 - b. An illustration: the case of GPUs 3
 - 2. Three families of machine learning 5
 - a. The **unsupervised** learning approach 5
 - b. The **supervised** learning approach..... 6
 - c. The **reinforcement** learning approach 9
 - d. When is machine learning useful? 12
 - 3. Machine Learning and Data Science 12
 - 4. Artificial intelligence 14
 - a. Weak vs Strong AI 14
 - b. Two videos to understand AI further 15
- 7 roads to data-driven value creation 16
 - 7 roads to data-driven value creation 16
 - 1. PREDICT 16
 - Prediction: The ones doing it..... 16
 - Prediction: the hard part 17
 - 2. SUGGEST 17
 - Suggestion: The ones doing it 18
 - Suggestion: the hard part 18
 - 3. CURATE 18
 - Curation: The ones doing it 19
 - Curation: the hard part 19
 - 4. ENRICH 19
 - Enrichment: The ones doing it 20
 - Enrichment: the hard part 20
 - 5. RANK / MATCH / COMPARE 20
 - Ranking / matching / comparing: The ones doing it 20
 - Ranking / matching / comparing: the hard part 21
 - 6. SEGMENT / CLASSIFY 21
 - Segmenting / classifying: The ones doing it 22
 - Segmenting / classifying: the hard part 22
 - 7. GENERATE / SYNTHETIZE(experimental!) 23
 - Generating: The ones doing it 23
 - Generating: the hard part 24
 - Combos!..... 24
- Index 25

Preface

A textbook for managers

The target reader for this book is a manager who needs to clearly understand what "data science", "big data", "artificial intelligence" so that they can:

- **leverage** these technologies to improve the efficiency of their existing business,
- **innovate** with new products and services and develop new business guidelines

The promise of this book is to bring you from a starting point with no knowledge of these technical concepts, to a point where you understand the concepts **and** you can develop "data centric" business projects: when "data" contributes to creating value for the customer and all stakeholders.

Is this textbook too technical or too easy for me?

If you are unsure, try this simple test: <http://bit.ly/essentials-1-test>

→ There are 20 topics you should be comfortable answering. See how you score. If the score is low, you should read first the introductory volume to this series:

"Essentials of data science for managers: Volume 1, from big data to APIs"

Machine learning, data science and artificial intelligence

1. Explaining machine learning in simple terms

a. A comparison with classic statistics

Let's [compare](#) machine learning to something we would call "regular statistics":

A basic method in statistics is to compute a regression line to identify a trend from a scatter plot.

To illustrate, we take some data about marketing budgets and sales figures in the corresponding period:

Marketing budget vs. Sales

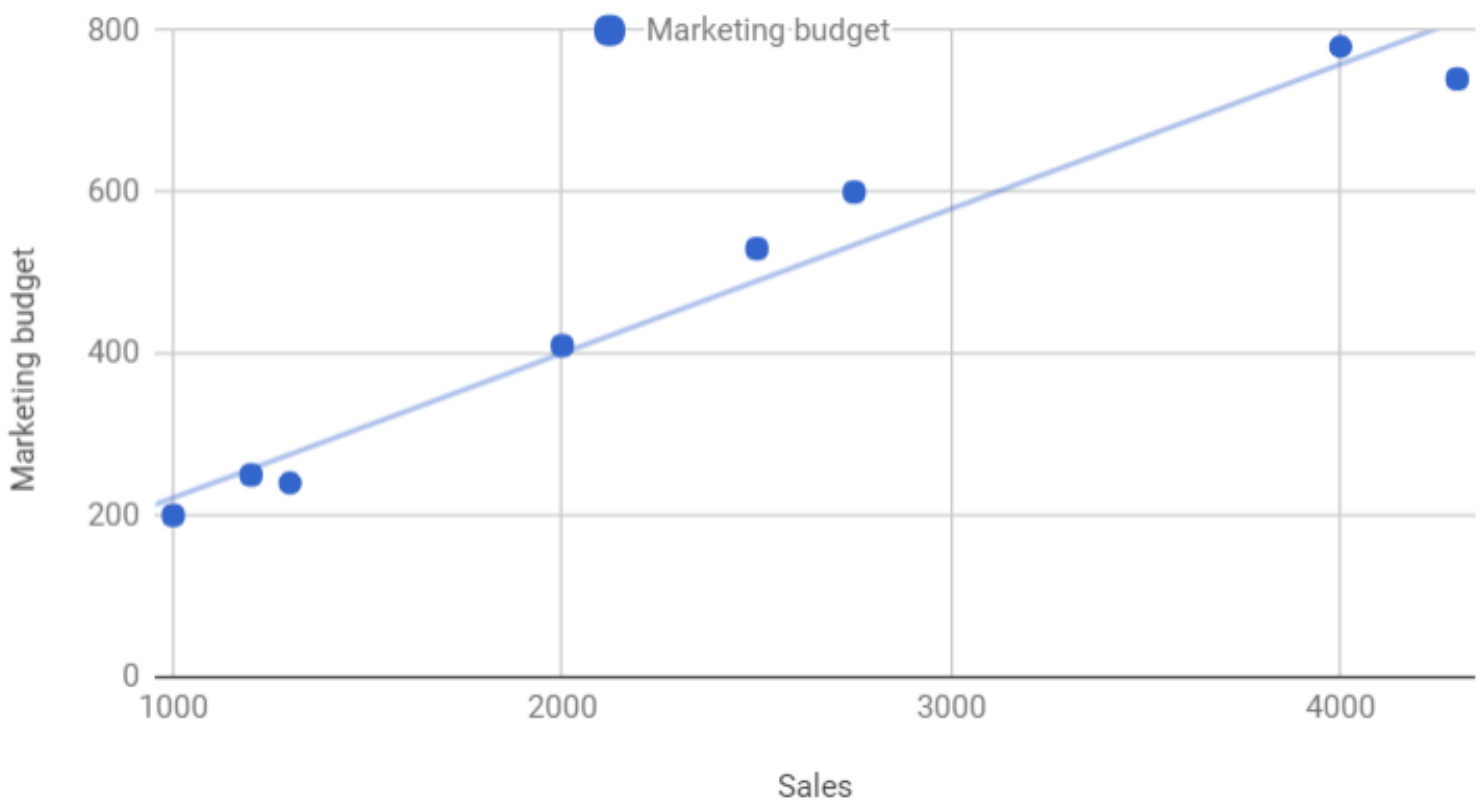


Figure 1. A linear regression

"Regular statistics" enables, among other things:

1. to find the numerical relation between the 2 series, based on a pre-established formal model (eg, [ordinary least squares](#)).
- we see that sales are correlated with marketing spendings. It is likely that more marketing spending causes more sales.
2. to predict, based on this model:

→ by tracing the line further (using the formal model), we can predict the effect of more marketing spending

"Regular statistics" is advanced by scientists who:

1. are highly skilled in mathematics

→ their goal is to find the exact mathematical expression defining the situation at hand, under rigorous conditions

→ a key approach is **inference**: by defining a **sample of the data** of just the correct size, we can reach conclusions which are valid for the entire dataset.

2. have no training in computer science / software engineering

→ they neglect how hard it can be to run their models on computers, in terms of calculations to perform.

→ since they focus on **sampling** the data, they are not concerned with handling entire datasets with related IT issues.

Machine learning does similar things to statistics, but in a slightly different way:

- there is an emphasis on getting the prediction right, not caring for identifying the underlying mathematical model
- the prediction needs to be achievable in the time available, with the computing resources available
- the data of interest is in a format / in a volume which is not commonly handled by regular statistics package (eg: images, observations with hundreds of features)

Machine learning is advanced by scientists who are typically:

1. highly skilled in statistics (the "classic" statistics we have seen above)
2. with a training or experience in computer science, familiar with working with unstructured / big data
3. working in environments (industry, military, ...) where the operational aspects of the problem are key determinants (unstructured data, limits on computing resources)

Machine learning puts a premium on techniques which are "computationally adequate":

- which need the minimum / the simplest algebraic operations to run: the best technique is worthless if it's too long or expensive to compute.
- which can be run in such a way that multiple computers work in parallel (simultaneously) to solve it.

(footnote: so machine learning, in my opinion, shares the spirit of "getting things done" as was [operations research in the early days](#))

The pursuit of improved models in traditional statistics is not immune to the notion of computational efficiency - it does count as a desirable property - but in machine learning this is largely a pre-requisite.

b. An illustration: the case of GPUs

A key illustration of the difference between statistics and machine learning can be provided with the use of graphic cards.

Graphic cards are these electronic boards full of chips found inside a computer, which are used for the display of images and videos on computer screens:



Figure 2. A graphic card sold by NVidia, a leading manufacturer

In the 1990s, video gaming developed a lot from arcades to desktop computers. Game developers created computer games showing more and more complex scenes and animations. (see [an evolution of graphics](#), and [advanced graphics games in 2017](#)).

These video games need powerful video cards (aka [GPUs](#)) to render complex scenes in full details - with calculations on light effects and animations **made in real time**.

This pushed for the development of ever more powerful GPUs. Their characteristics is that they can compute simple operations to change pixel colors, **for each of the millions of pixels of the screen in parallel**, so that the next frame of the picture can be rendered in milliseconds.

Millions of simple operations run in parallel for the price of a GPU (a couple of hundreds of dollars), not the price of

dozens of computers running in parallel (can be dozens of thousands of dollars)? This is interesting for computations on big data!

If a statistical problem for prediction can be broken down into simple operations which can be run on a GPU, then a large dataset can be analyzed in seconds or minutes on a laptop, instead of cluster of computers.

To illustrate the difference in speed between a mathematical operation run without / with a GPU:

► <https://www.youtube.com/watch?v=-P28LKWTzrI> (YouTube video)

The issue is: to use a GPU for calculations, you need to conceptualize the problem at hand as one that can be:

- broken into a very large series
- of very simple operations (basically, sums or multiplications, nothing complex like square roots or polynomials)
- which can run independently from each other.

Machine learning typically pays attention to this dimension of the problem right from the design phase of models and techniques, where statistics would typically not consider the issue, or only downstream: not at the design phase but at the implementation phase.

Now that we have seen how statistics and machine learning differ in their approach, we still need to understand how does machine learning get good results, if it does not rely on modelling / sampling the data like statistics does?

Machine learning can be categorized in 3 families of tricks:

2. Three families of machine learning

a. The unsupervised learning approach

This designates all the methods which take a fresh dataset and find interesting patterns in it, **without training on previous, similar datasets**.

The analogy is with a person doing a task for the first time:

→ she learns a new thing by applying clever heuristics, without having been training on the task before.

Example: in your wedding, how to sit people with similar interests at the same tables?

The set up:

- a list of 100 guests, and 3 tastes you know they have for each of them
- 10 tables with 10 sits each.
- a measure of similarity between 2 guests: 2 guests have similarity of 0% if they share 0 tastes, 33% if they share 1 taste, 66% with 2 tastes in common, 100% with three matching interests.
- a measure of similarity at the level of a table: the sum of similarities between all pairs of guests at the table (45 pairs possible for a table of 10).

A possible solution using an unsupervised approach:

- on a computer, assign randomly the 100 guests to the 10 tables.
- for each table:

- measure the degree of similarity of tastes for the table

- exchange the sit of 1 person at this table, with the sit of a person at a different table.
- measure again the degree of similarity for the table: if it improves, keep the new sits, if not, revert to before the exchange

And repeat for all tables, many times, until no exchange of sits improves the similarity. When this stage is achieved, we say the model has "**converged**".

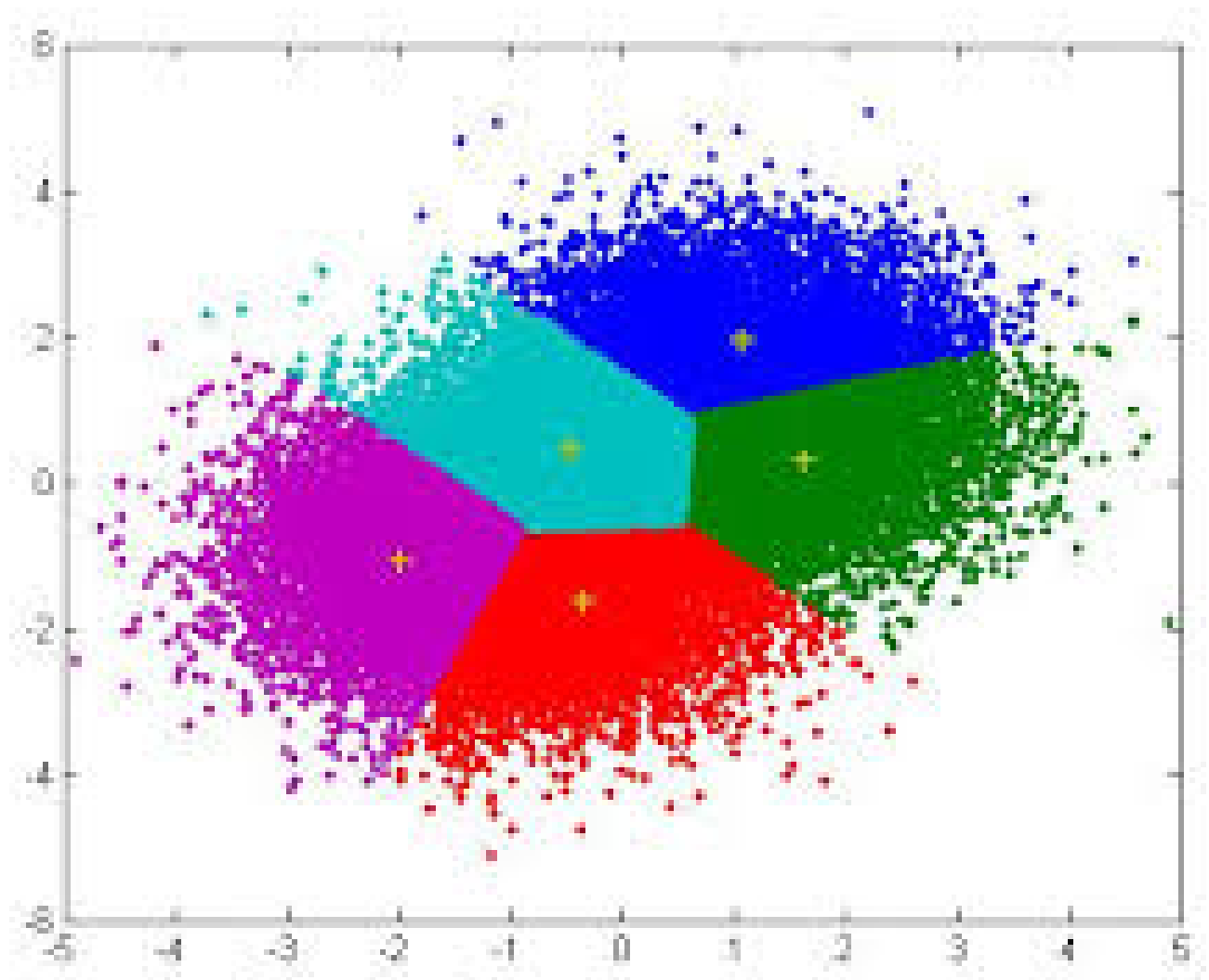


Figure 3. K-means, an unsupervised learning approach

b. The supervised learning approach

Take 50,000 or more observations, or data points, like:

****an image of a cat, with the caption "cat"**

****an image of a dog, with the caption "dog"**

****another image of a cat, with the caption "cat"**

etc....

- you need 50,000 observations of this kind, or more! It is called the **training set**
- this is also called a **labelled dataset**, meaning that we have a label describing each of the observation.

The task is: if we give our computer a new image of a cat without a label, will it be able to guess the label "cat"?

The method:

- take a list of random coefficients (in practice, the list is a vector, or a matrix)
- for each of the 50,000 pictures of dogs and cats:
 - apply the coefficients to the picture at hand (let's say we have a dog here)
 - If the result is "dog", do nothing, it works!
 - If the result is "cat", change slightly the coefficients.
 - move to the next picture
- After looping through 50,000 pictures the parameters have hopefully adjusted and fine tuned. This was the **training of the model**.

Now, when you get new pictures (the **fresh set**), applying the trained model should output a correct prediction ("cat" or "dog").

Supervised learning is currently the most popular family of machine learning.

Chihuahua or Muffin?



Figure 4. A hard test case for supervised learning

It is called **supervised** learning because the learning is very much constrained / supervised by the intensive training performed:

→ there is limited or no "unsupervised discovery" of novelty.

► <https://www.youtube.com/watch?v=4HCE1P-m1l8> (YouTube video)

Important take away on the supervised approach:

8 • **collecting large datasets for training is key.** Without these data, no supervised learning.

- supervised learning is not good at analyzing situations entirely different from what is in the training set.

c. The reinforcement learning approach

To understand reinforcement learning in an intuitive sense, we can think of how animals can learn quickly by **ignoring** undesirable behavior and rewarding desirable behavior.

This is easy and takes just seconds. The following video shows B.F. Skinner, main figure in psychology in the 1950s-1970s:

► <https://www.youtube.com/watch?v=TtfQlkGwE2U> (*YouTube video*)

Footnote: how does this apply to learning in humans? On the topic of learning and decision making, I warmly recommend [this book by Paul Glimcher](#), professor of neuroscience, psychology and economics at NYU:

(this is a very hard book to read as it covers three disciplines in depth. The biological mechanisms of decision making it describes can be inspiring to design new computational approaches.)



FOUNDATIONS OF —
**Neuroeconomic
Analysis**

PAUL W. GLIMCHER

OXFORD

Besides pigeons, reinforcement learning can be applied to any kind of "expert agents".

Take the case of a video game like Super Mario Bros:

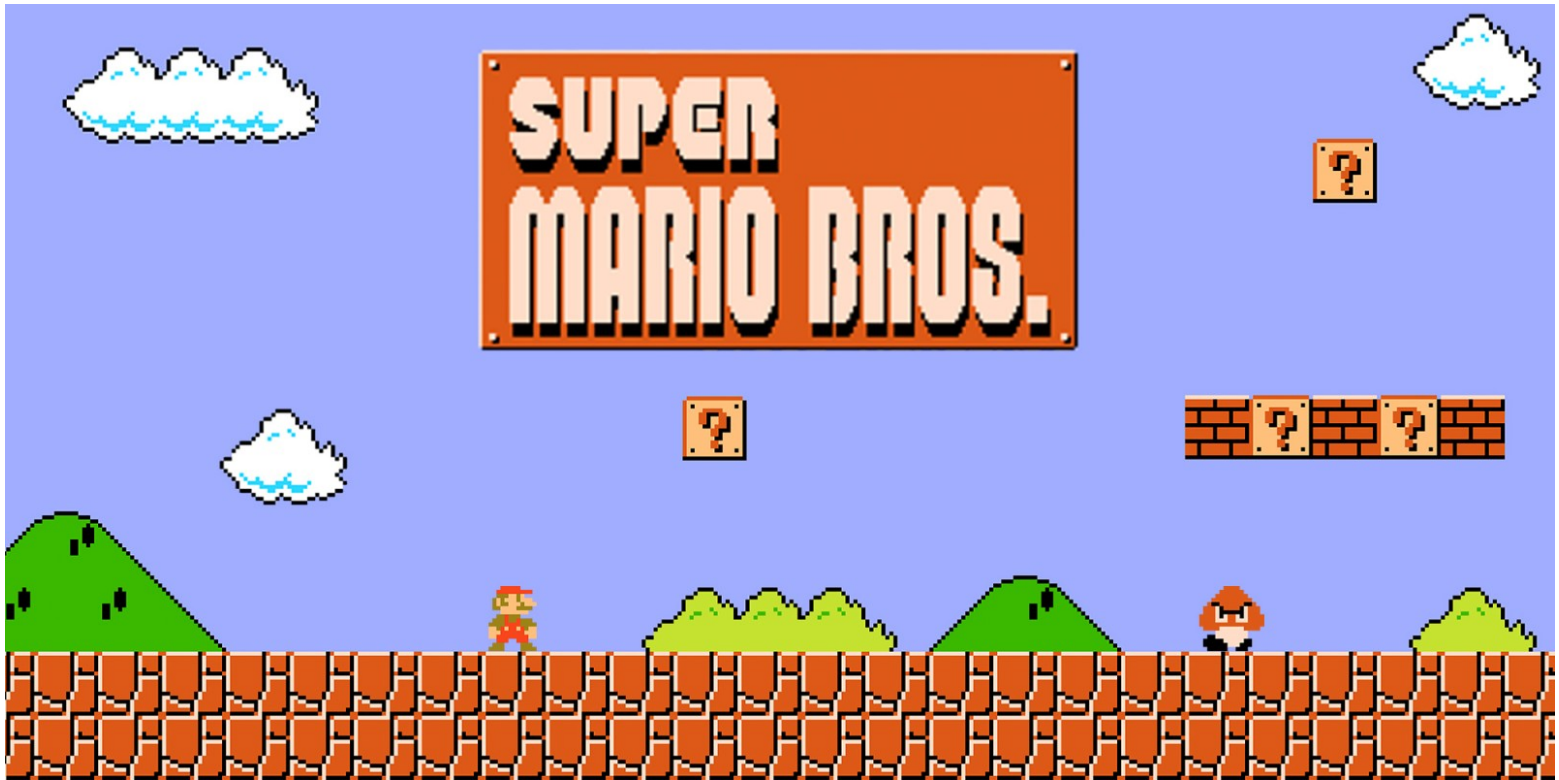


Figure 6. Mario Bros, a popular video game

Struture of the game / the task:

- Goal of the task: Mario should collect gold coins and complete the game by reaching the far right of the screen.
- Negative outcome to be avoided: Mario getting killed by ennemies or falling in holes.
- Starting point: Mario Bros is standing at the beginning of the game, doing nothing.
- Possible actions: move right, jump, stand & do nothing, shoot ahead.

Reinforcement learning works by:

1. Making Mario do a new random action ("try something"), for example: "move right"
2. The game ends (Mario moved right, gets hit by a ennemy)
3. This result is stored somewhere:
 - move right = good (progress towards the goal of the game)
 - walking close to an ennemy and getting hit by it = bad
4. Game starts over (back to step 1) with a a combination of
 - continue doing actions recorded as positive
 - try something new (jump, shoot?) when close to a situation associated with a negative outcome

After looping from 1. to 4. thousands of times, Mario completes the game, without any human player:

Reinforcement learning is perceived as corresponding to an important side of human learning / human intelligence (goal oriented, "trial and error").

d. When is machine learning useful?

Using machine learning can be a waste of resource, when well known statistics could be easily applied.

Hints that "classic" statistical modelling (maybe as simple as a linear regression) should be enough:

- The dataset is not large (below 50k observations), supervised learning is not going to work
- The data is perfectly structured (tabular data)
- The data points have few features

Cases when "classic" statistics modelling is **necessary**:

- The question is about the relative contribution of independent variables to the determination of an outcome

3. Machine Learning and Data Science

Machine learning is a step in the longer chain of steps of data science.

The process was formalized as [kdd](#): "Knowledge Discovery in Databases":

Figure 1. Overview of the steps constituting the KDD process

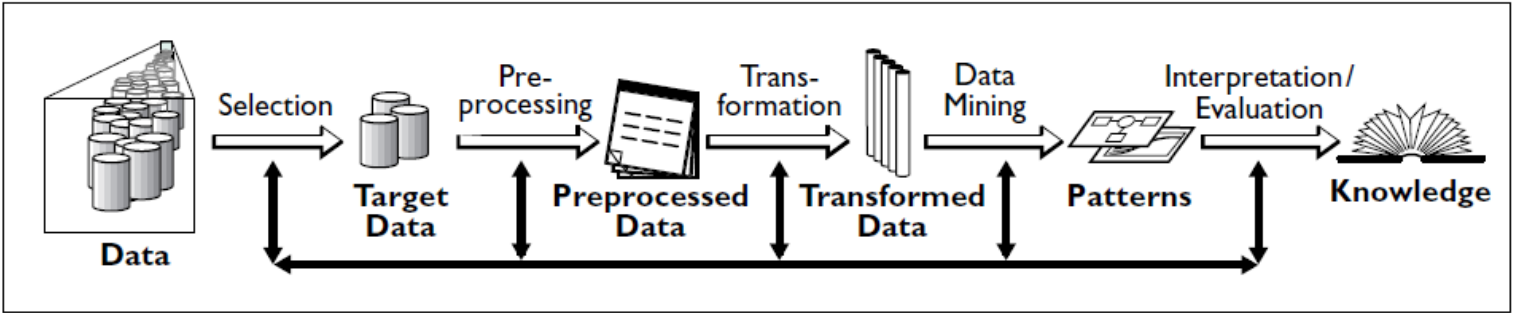


Figure 7. KDD - knowledge discovery in databases

More recent representations of the steps in data processing have been suggested, making room for the role of data visualization (see the lecture on the topic):

→ see [the version by Ben Fry \(source\)](#) and this one by Moritz Stefaner:

Workflow

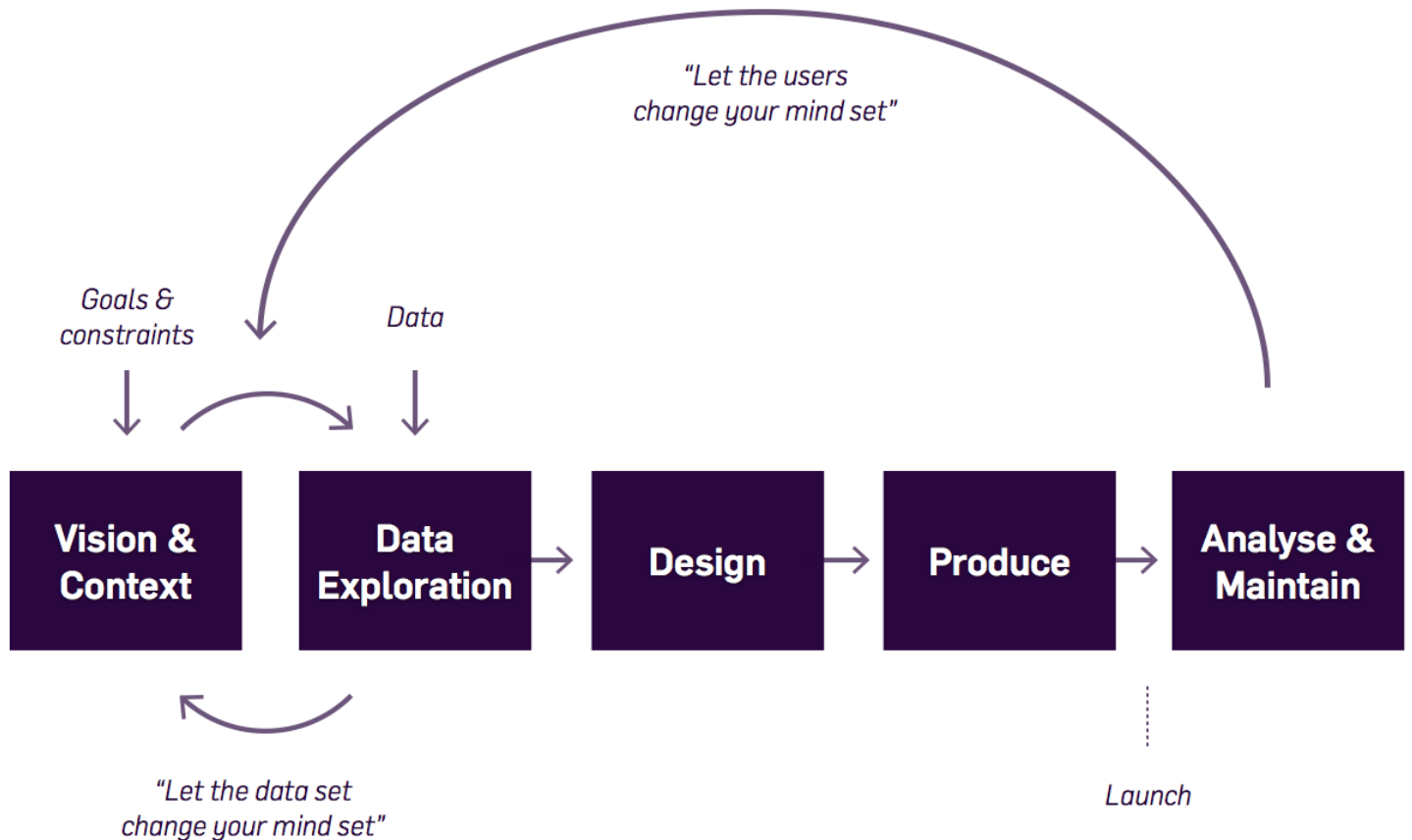


Figure 8. data visualization workflow by Moritz Stefaner

([source](#))

Machine learning is one of the techniques (along with traditional statistics) that intervenes at the step of "Data mining".

What makes data scientists important is that the steps of this kdd are highly interdependent.

You need individuals or teams who are not just versed in data mining:

→ because the shape of the data at the collection stage has a huge influence on the kind of techniques, and the kind of software, that can be used to discover knowledge.

The skills of a data scientist are often represented as the meeting of three separate domains:

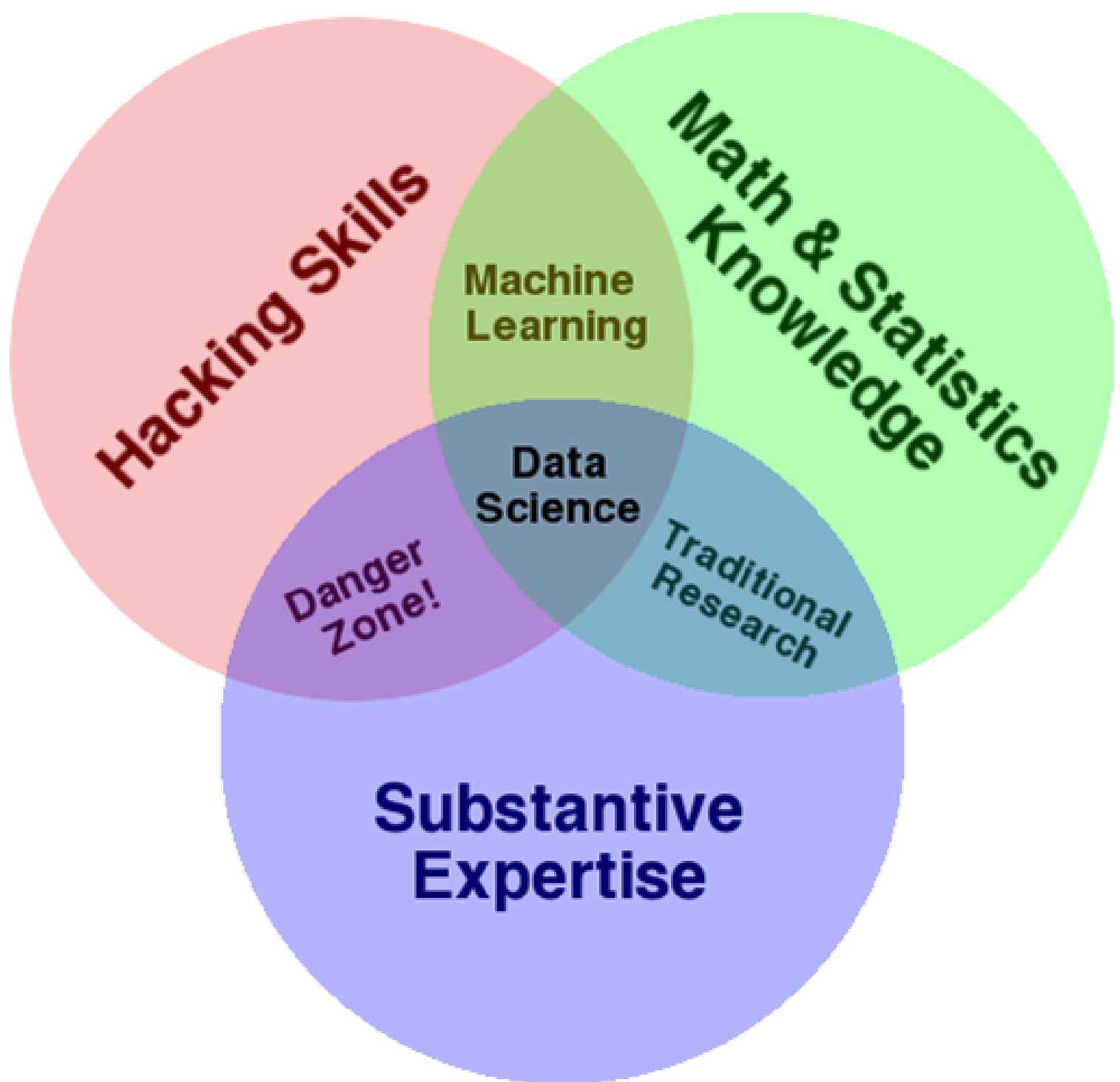


Figure 9. The Venn diagram of what is a data scientist

source: <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>

4. Artificial intelligence

a. Weak vs Strong AI

Weak AI designates computer programs able to outperform humans at complex tasks with a narrow focus (playing chess)

Weak AI is typically the result of applying expert systems or machine learning techniques seen above.

Strong AI is an intelligence that would be general in scope, able to set its own goal, and conscious of itself. Nothing is

close to that yet.

So AI is a synonymous with weak AI at the moment.

b. Two videos to understand AI further

Laurent Alexandre on the social and economic stakes of AI (in French):

► <https://www.youtube.com/watch?v=rJowm24piM4> (*YouTube video*)

John Launchbury, the Director of DARPA's Information Innovation Office (I2O) in 2017:

► <https://www.youtube.com/watch?v=-O01G3tSYpU> (*YouTube video*)

7 roads to data-driven value creation

7 roads to data-driven value creation



=== Not a closed list, not a recipe!

Rather, these are essential building blocks for a strategy of value creation based on data. ===

1. PREDICT




Prediction: The ones doing it

1. Predictive churn / default / ... (banks / telco)

2. Predicting crime 

3. Predicting deals 

4. Predictive maintenance 

Prediction: the hard part

1. Collecting data ([cold start problem](#))

2. Risk missing the long tail, algorithmic discrimination, stereotyping

3. Neglect of novelty

2. SUGGEST



Suggestion: The ones doing it

- ## 1. Amazon's product recommendation system



- ## 2. Google's "Related searches..."



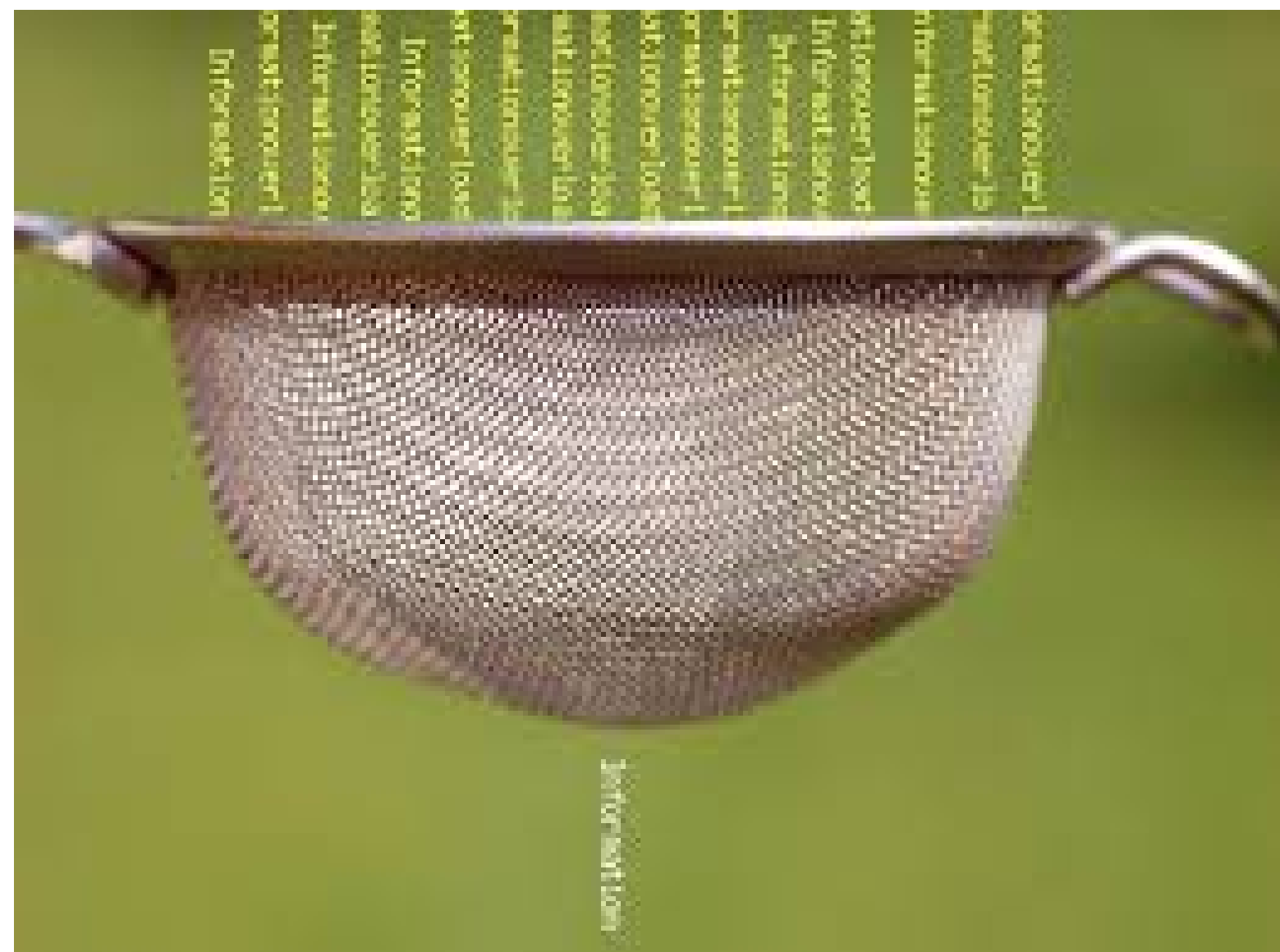
- ### 3. Retailer's personalized recommendations



Suggestion: the hard part

1. The [cold start problem](#), managing serendipity (see review: [paying version](#), free version not available) and "filter bubble" effects (review: [paying version](#), [free version here](#)).
2. Finding the value proposition which goes beyond the simple "you purchased this, you'll like that"

3. CURATE



Curation: The ones doing it

1. Clarivate Analytics curating metadata from scientific publishing


2. Nielsen and IRI curating and selling retail data


3. IMDb curating and selling movie data



Curation: the hard part

1. Slow progress: curation needs human labor to insure high accuracy, it does not scale the way a computerized process would.
2. Must maintain continuity: missing a single year or month hurts the value of the overall dataset disproportionately.
3. Scaling up / right incentives for the workforce: the workforce doing the curation should be paid fairly, which is [not the case yet](#).
4. Quality control

4. ENRICH



Enrichment: The ones doing it

1. Selling methods and tools to enrich datasets



2. Selling aggregated indicators



3. Selling credit scores

Enrichment: the hard part

1. Knowing which cocktail of data is valued by the market

2. Limit replicability

3. Establish legitimacy

5. RANK / MATCH / COMPARE



Ranking / matching / comparing: The ones doing it

1. Search engines ranking results



2. Yelp, Tripadvisor, etc... which rank places



3. Any system that needs to filter out best quality entities among a crowd of candidates

Ranking / matching / comparing: the hard part

1. Finding emergent, implicit attributes (imagine: if you rank things based on just one public feature: not interesting nor valuable)

2. Insuring consistency of the ranking (many rankings are less straightforward than they appear)

3. Avoid gaming of the system by the users (for instance, [companies try to play Google's ranking of search results at their advantage](#))

6. SEGMENT / CLASSIFY

Chihuahua or Muffin?



Segmenting / classifying: The ones doing it

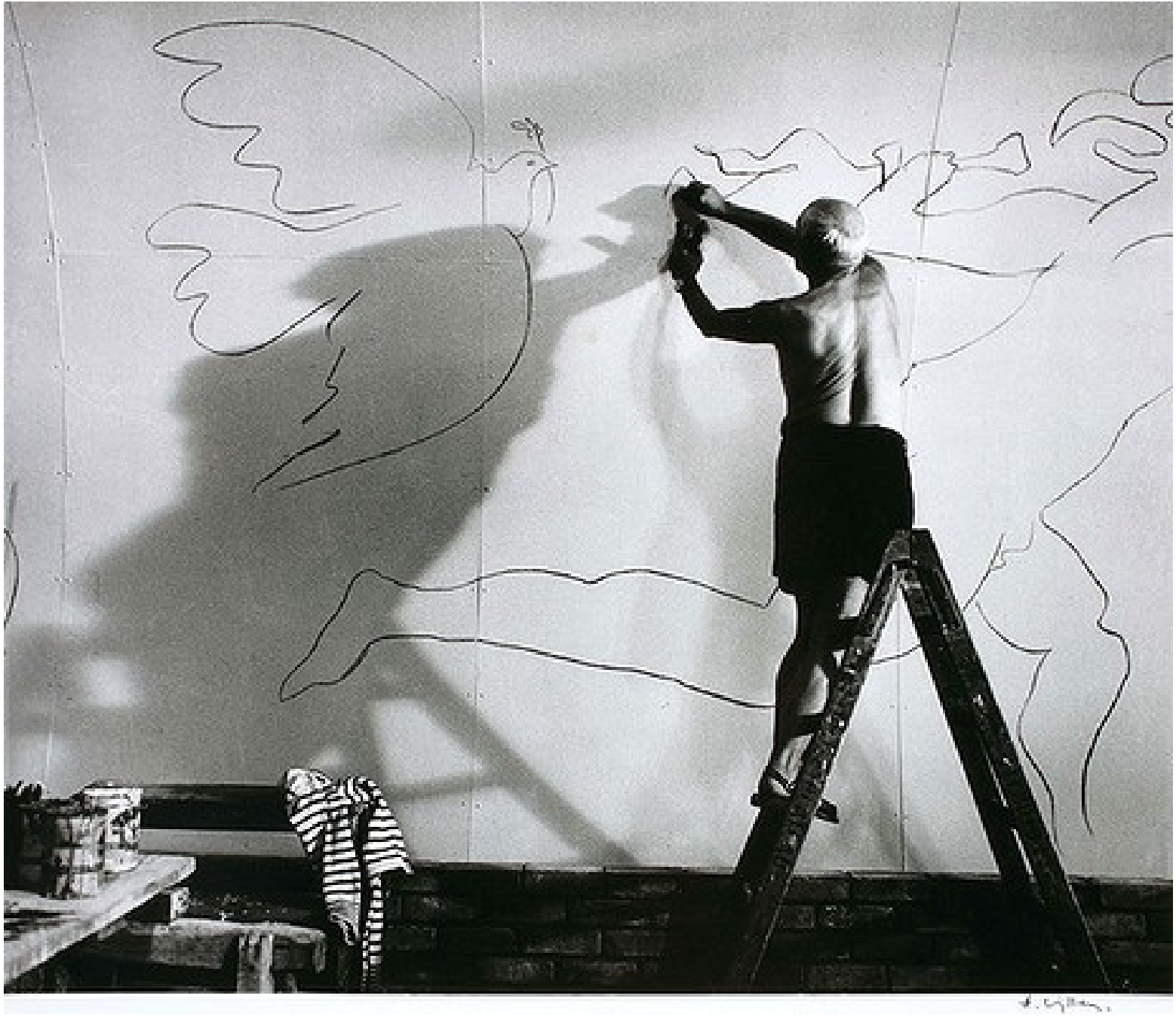
1. Tools for discovery / exploratory analysis by segmentation

2. Diagnostic tools (spam or not? buy, hold or sell? healthy or not?)  medimsight
Medical Imaging Cloud Platform

Segmenting / classifying: the hard part



1. Evaluating the quality of the comparison
2. Dealing with boundary cases
3. Choosing between a pre-determined number of segments (like in the k-means) or letting the number of segments emerge





7. GENERATE / SYNTHETIZE(experimental!)



Generating: The ones doing it

(click on the logos to get to the relevant web page)


1. Intelligent BI with [Aiden](#)  [aiden.ai](#)
2. [wit.ai](#), the chatbot by FB  [wit.ai](#)

- 3. Virtual assistants company 
- 4. Image generation 
- 5. Close-to-real-life speech synthesis 
- 6. Generating realistic car models from a few parameters by Autodesk: 

A video on the generation of car models by Autodesk:

► <https://www.youtube.com/watch?v=25xQs0Hs1z0> (YouTube video)

Generating: the hard part

- 1. Should not create a failed product / false expectations
- 2. Both classic (think of ) and frontier science: not sure where it's going

Combos!



Figure 10. Combinations

ABOUT THE AUTHOR

Clément Levallois is professeur agrégé de l'école normale supérieure and Associate Professor at em **lyon business school**, where he conducts research projects in data mining, data visualization and network analysis in various fields of social sciences. His teaching activities center on the transmission of a digital culture to students and executive participants.

Clément Levallois is a Java coder and an active supporter of Gephi, the leading software for network visualization. In a previous academic life, he researched the history of economics and biology in post-war U.S.A.

His past and current projects can be seen at <http://clementlevallois.net>, and he can be reached on Twitter at @seinecle.

ESSENTIALS OF DATA SCIENCE FOR MANAGERS

Volume 2: From artificial intelligence to business applications

Managers can hardly ignore the opportunities afforded by “big data”, an expression often used in relation with “data science” or “artificial intelligence”. But how to find the time to learn these complex notions, for the specific purpose of using them in a business context? This book offers a clear and complete presentation of the concepts and technologies a manager should know in order to make use of them in a professional context.

This volume (Volume 2: From Artificial Intelligence to Business Applications) is the second stage of the learning path. Simple definitions of artificial intelligence are provided, with examples illustrating what use case each type of AI can address. The second part of this volume introduces seven families of business applications which put data science and AI at work to create value for customers and other parts of an organization. This volume is the second in a series started with “Essentials of Data Science for Managers, Volume 1: From big data to APIs”.

Clément Levallois is professeur agrégé de l'école normale supérieure and Associate Professor at em lyon business school, where he conducts research projects in data mining, data visualization and network analysis in various fields of social sciences. His teaching activities center on the transmission of a digital culture to students and executive participants.