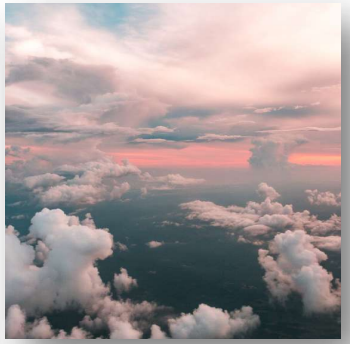
Introduction to Data Science

Lecture as part of TERI-NORCE Research School //
Funded by the Norwegian Ministry of Foreign Affairs



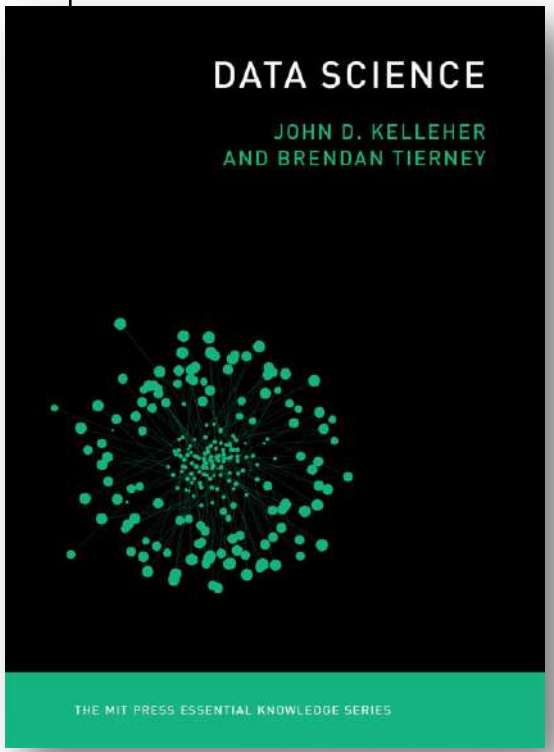
Dr. Michel d. S. **Mesquita**[1,2,3]

LinkedIn QR code

[1]NORCE Norwegian Research Centre, **Norway** / [2]Bjerknes Centre for Climate Research, **Norway** / [3]M2Lab.org, **Norway**

Main reference //
This lecture is based on the book shown below
'KT2018' hereafter

It provides an accessible and concise introduction to data science.



Kelleher and Tierney (2018) Data Science. Boston: The MIT Press Essential Knowledge Series. URL: https://mitpress.mit.edu/books/data-science

Data science /

«… a set of principles, problem definitions, algorithms, and processes for extracting nonobvious and useful patterns from large data sets»

Machine learning (ML) /

«… focuses on the design and evaluation of algorithms for extracting patterns from data»

Data mining /

«… generally deals with the analysis of structured data and often implies and emphasis on commercial applications»

KT2018

When to use data science? //
It has to be useful

When we deal with a large number of data examples or the patterns are too complex for humans to discover and extract manually

Actionable insight /

Insight = the pattern should give us relevant information about the problem that is not obvious

Actionable = the insight we get should be something we can use in some way

Brief history of data science //

Data science starts in the 1990s drawing knowledge from two historical fields: **data collection** and **data analysis**

Data gathering /

Earlier marks of solstices and sunrise

Transactional data: advent of writing brought record keeping, e.g. Mesopotamia ~3200 BC

Nontransactional data: demographic data, e.g. earliest census in pharaonic Egypt ~3000 BC

Relational data model: Edgar F. Codd in 1970 publishes a paper on a model, which lays the foundation for modern databases, such as the *structure query language (SQL)*

Big data: today we have volume, variety, velocity; *NoSQL databases*

Brief history of data science //

Data science starts in the 1990s drawing knowledge from two historical fields: **data collection** and **data analysis**

Thomas Bayes
(Wikipedia)

Data analysis /

Statistics: the branch of science that deals with the collection and analysis of data

Probability theory: 17th and 18th centuries, e.g. work of Blaise Pascal, Jakob Bernoulli, Thomas Bayes, and others. Probability distributions became the new tool to move beyond descriptive statistics

Statistical learning and modern data science: 19th century; new developments in probability theory facilitated statistical learning; e.g. Works by Pierre Laplace, Carl Friedrich Gauss, among others; method of least squares, which laid the foundation for linear regression and later artificial neural network

Data visualisation and exploratory data analysis: 1780-1820 William Playfair invented statistical graphics, such as the line chart, area chart, bar chart, pie chart; these led the foundation for modern methods, such as the t-distributed stochastic neighbord embedding (t-SNE) algorithm

Brief history of data science // The role of women

The Countess of Lovelace
(Wikipedia)

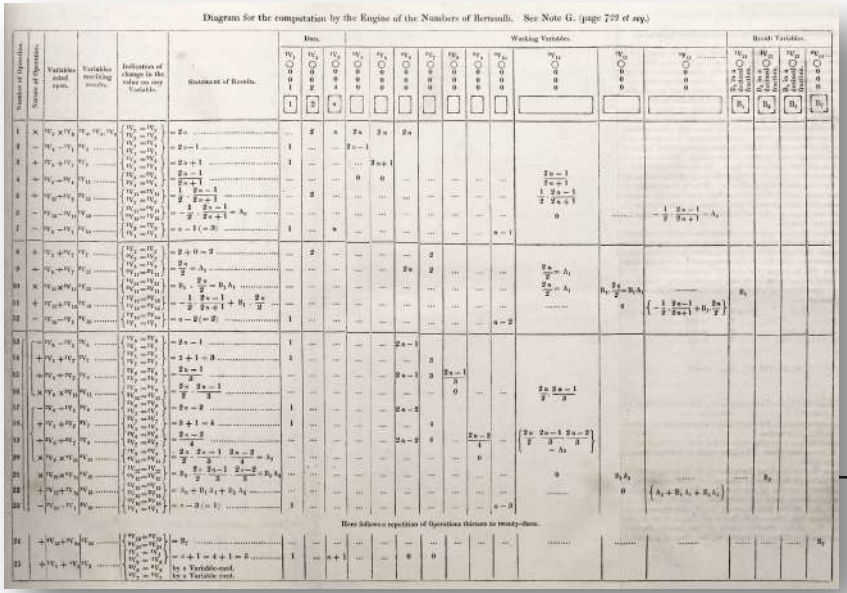
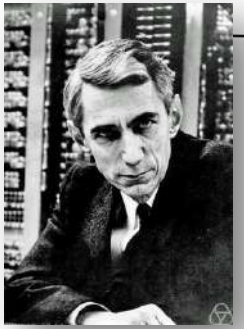


Diagram for the computation by the Engine of the Numbers of Bernoulli. See Note G. (page 739 et seq.)

Ada Lovelace's notes were labelled alphabetically from A to G. In note G, she describes an algorithm for the Analytical Engine to compute Bernoulli numbers. It is considered to be the first published algorithm ever specifically tailored for implementation on a computer, and Ada Lovelace has often been cited as the first computer programmer for this reason.
(Wikipedia)

Data science starts in the 1990s drawing knowledge from two historical fields: **data collection** and **data analysis**



Claude Shannon
(Wikipedia)

Data analysis (cont.) /

20th century

Karl Pearson developed modern hypothesis testing; R. A. Fisher developed statistical methods for multivariate analysis and maximum likelihood estimate

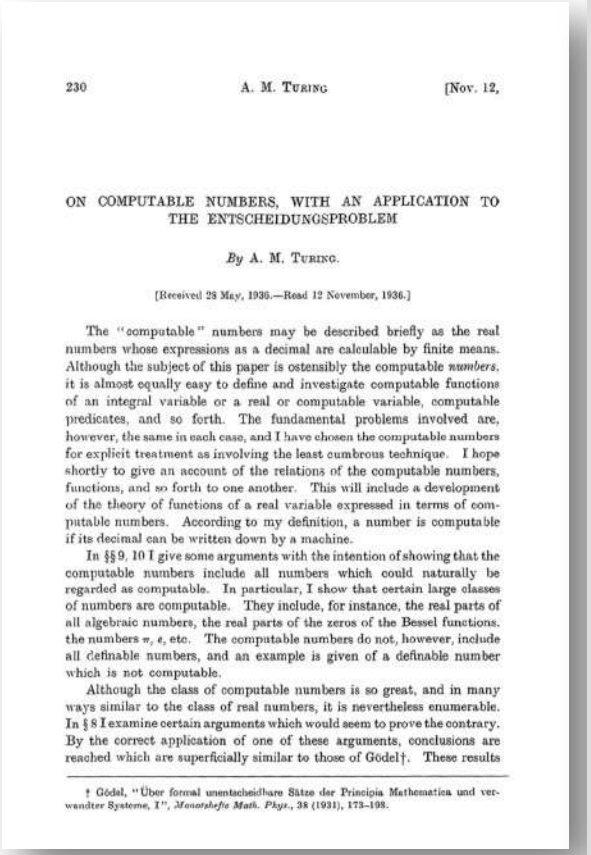
Alan Turing in the Second World War led the invention of the electronic computer

Warren McCulloch and Walter Pitts in 1943 proposed the first mathematical model of a neural network

Claude Shannon in 1948 published «A Mathematical Theory of Communication» and founded *information theory*

Brief history of data science // Alan Turing

Alan Turing
(Wikipedia)



230 A. M. TURING [Nov. 12,

ON COMPUTABLE NUMBERS, WITH AN APPLICATION TO THE ENTSCHEIDUNGSPROBLEM

By A. M. TURING.

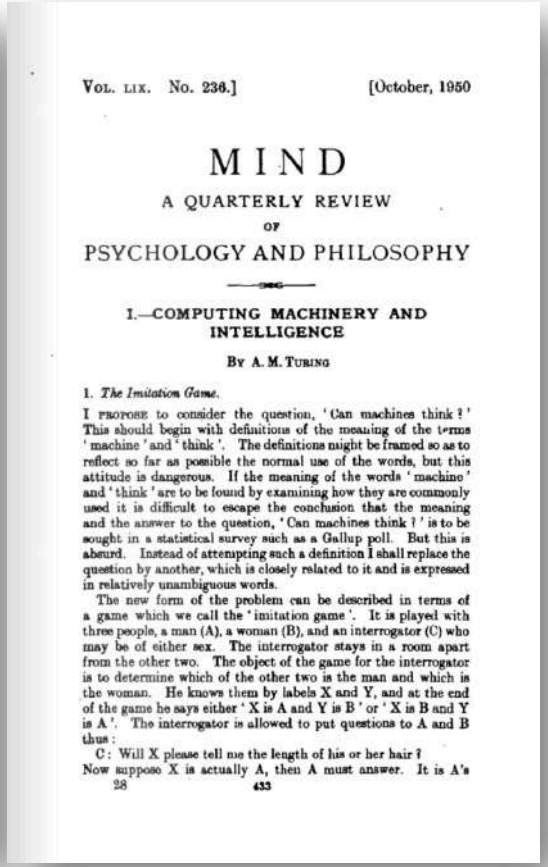
[Received 28 May, 1935.—Read 12 November, 1936.]

The "computable" numbers may be described briefly as the real numbers whose expressions as a decimal are calculable by finite means. Although the subject of this paper is ostensibly the computable *numbers*, it is almost equally easy to define and investigate computable functions of an integral variable or a real or computable variable, computable predicates, and so forth. The fundamental problems involved are, however, the same in each case, and I have chosen the computable numbers for explicit treatment as involving the least cumbrous technique. I hope shortly to give an account of the relations of the computable numbers, functions, and so forth to one another. This will include a development of the theory of functions of a real variable expressed in terms of computable numbers. According to my definition, a number is computable if its decimal can be written down by a machine.

In §§ 9, 10 I give some arguments with the intention of showing that the computable numbers include all numbers which could naturally be regarded as computable. In particular, I show that certain large classes of numbers are computable. They include, for instance, the real parts of all algebraic numbers, the real parts of the zeros of the Bessel functions, the numbers π, e, etc. The computable numbers do not, however, include all definable numbers, and an example is given of a definable number which is not computable.

Although the class of computable numbers is so great, and in many ways similar to the class of real numbers, it is nevertheless enumerable. In § 8 I examine certain arguments which would seem to prove the contrary. By the correct application of one of these arguments, conclusions are reached which are superficially similar to those of Gödel†. These results

† Gödel, "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme, I", *Monatshefte Math. Phys.*, 38 (1931), 173–198.



VOL. LIX. No. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING

1. *The Imitation Game.*

I PROPOSE to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, 'Can machines think?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

The new form of the problem can be described in terms of a game which we call the 'imitation game'. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either 'X is A and Y is B' or 'X is B and Y is A'. The interrogator is allowed to put questions to A and B thus:

C: Will X please tell me the length of his or her hair?

Now suppose X is actually A, then A must answer. It is A's

28 433

Turing (1936, 1950)

Papers by Alan Turing on the topics of computable numbers and artificial intelligence.

Data science starts in the 1990s drawing knowledge from two historical fields: **data collection** and **data analysis**

Nils Nilsson
(Wikipedia)

Data analysis (cont.) /

Evelyn Fix and Joeseph Hodges proposed a model for discriminatory analysis (classification or pattern-recognition problem), which became the basis for modern nearest-neighbor models

Establishment of the field of *artificial intelligence* at a workshop in Dartmouth College

The term *machine learning* was beginning to be used to describe programs that gave a computer the ability to learn from data

Nils Nilsson's book *Learning Machines* in 1965 showed how neural networks could be used to learn linear models for classification

Data science starts in the 1990s drawing knowledge from two historical fields: **data collection** and **data analysis**

Ensemble fatigue
(Benestad et al., 2017, Nature Climate Change)

Data analysis (cont.) /

Earl Hunt, Janet Marin, and Philip Stone in 1966 developed the concept-learning system framework, which was the progenitor of an important family of ML algorithms that induced decision-tree models

A number of independent researchers developed and published earlier versions of the *k-means clustering* algorithm

Today: some of the most important developments include *ensemble models*, where predictions are made using a set of models, and *deep-learning neural networks*, which have multiple layers of neurons (with applications in machine vision, natural-language processing,…)

The Emergence and Evolution of Data Science //

1990s: the term «data science» came to prominence in discussions of the need for statisticians to join with computer scientists to bring mathematical rigor to computational analysis of large data sets

1997: C.F. Jeff Wu's public lecture «Statistics = Data Science?»

2001: William S. Cleveland published an action plan for creating a university department in the field of data science

Also in 2001 //

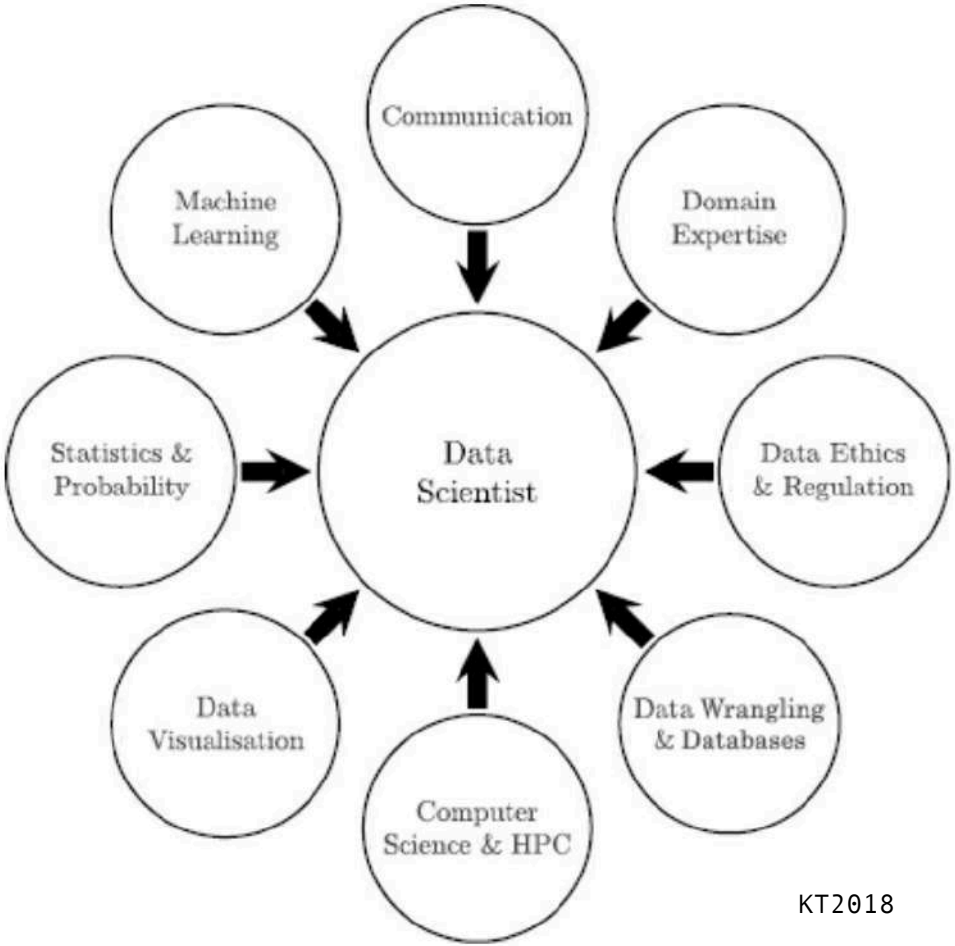
Leo Breiman published the paper «Statistical Modeling: the two cultures» /

His distinction between a statistical focus on models that explain the data versus an algorithmic focus on models that can actually predict the data highlights a core difference between statisticans and ML researchers

Today //

The role of a data scientist has become very broad

It is difficult for an individual to master all of the skill areas



KT2018

Datum /

Or piece of information is an abstraction of a real-world entity (person, object, or event)

The terms *variable, feature,* and *attribute* are often used interchangeably to denote an individual abstraction

Each entity is typically described by a number of attributes. For instance, a book might have the following attributes: author, title, topic, genre, publisher, price, etc.

Data set /

It consists of the data relating to a collection of entitites, with each entity described in terms of a set of attributes

In its most basic form a data set is organized in an *n*m* data matrix called the *analytics record*

KT2018

ID	Title	Author	Year	Cover	Edition	Price
1	*Emma*	Austen	1815	Paperback	20th	$5.75
2	*Dracula*	Stoker	1897	Hardback	15th	$12.00
3	*Ivanhoe*	Scott	1820	Hardback	8th	$25.00
4	*Kidnapped*	Stevenson	1886	Paperback	11th	$5.00

Choice of attributes /

How do you choose the most appropriate attributes?

Different attribute types /

Numeric, nominal and ordinal.

The data type of an attribute affects the methods we can use to analyze and understand the data, including both the basic statistics we can use to describe the distribution of values that an attribute takes and the more complex algorithms we use to ideentify the patterns of relationships between attributes

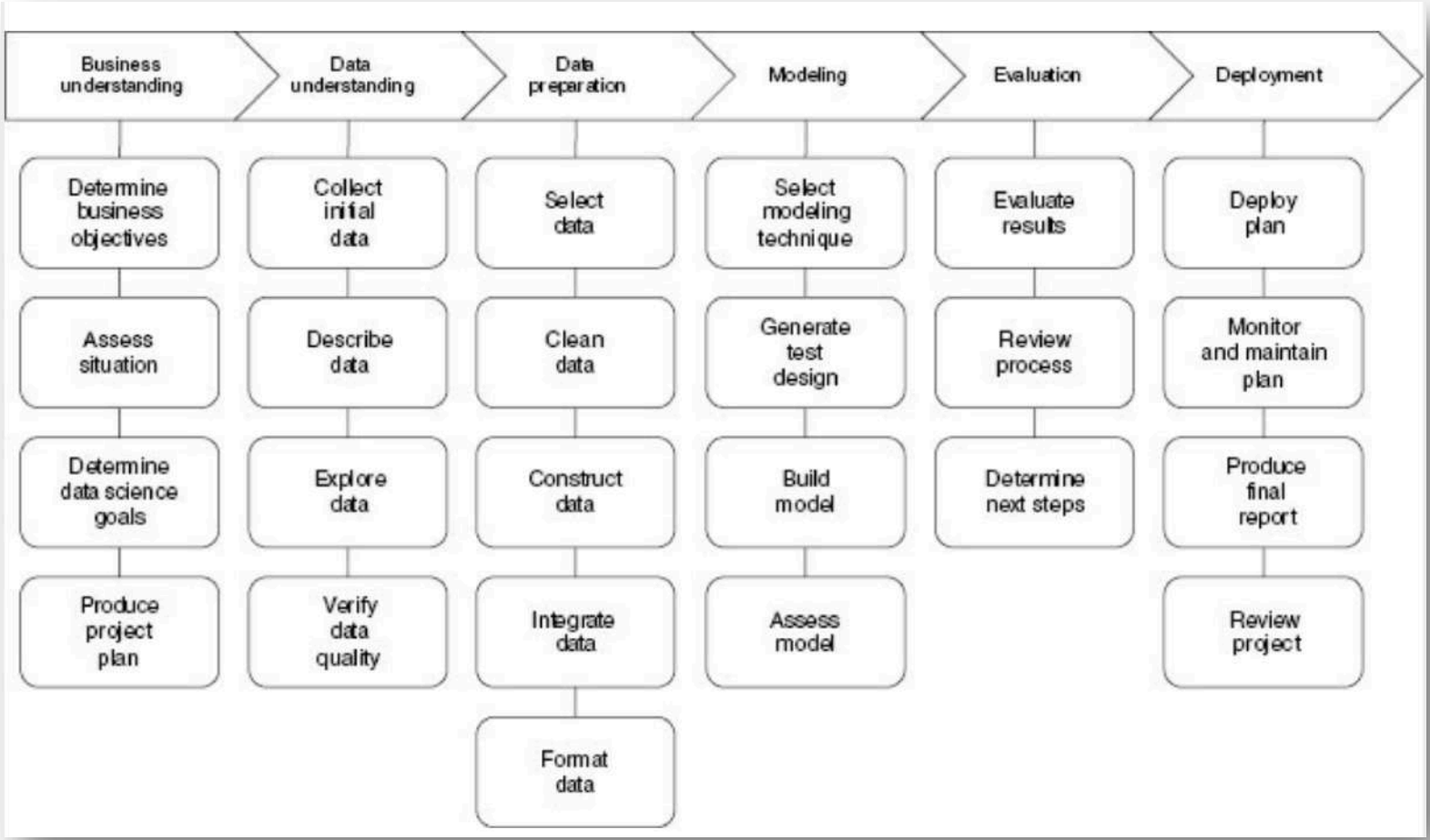
Structured data

- Can be stored in a table, and every instance in the table has the same structure (i.e., set of attributes)

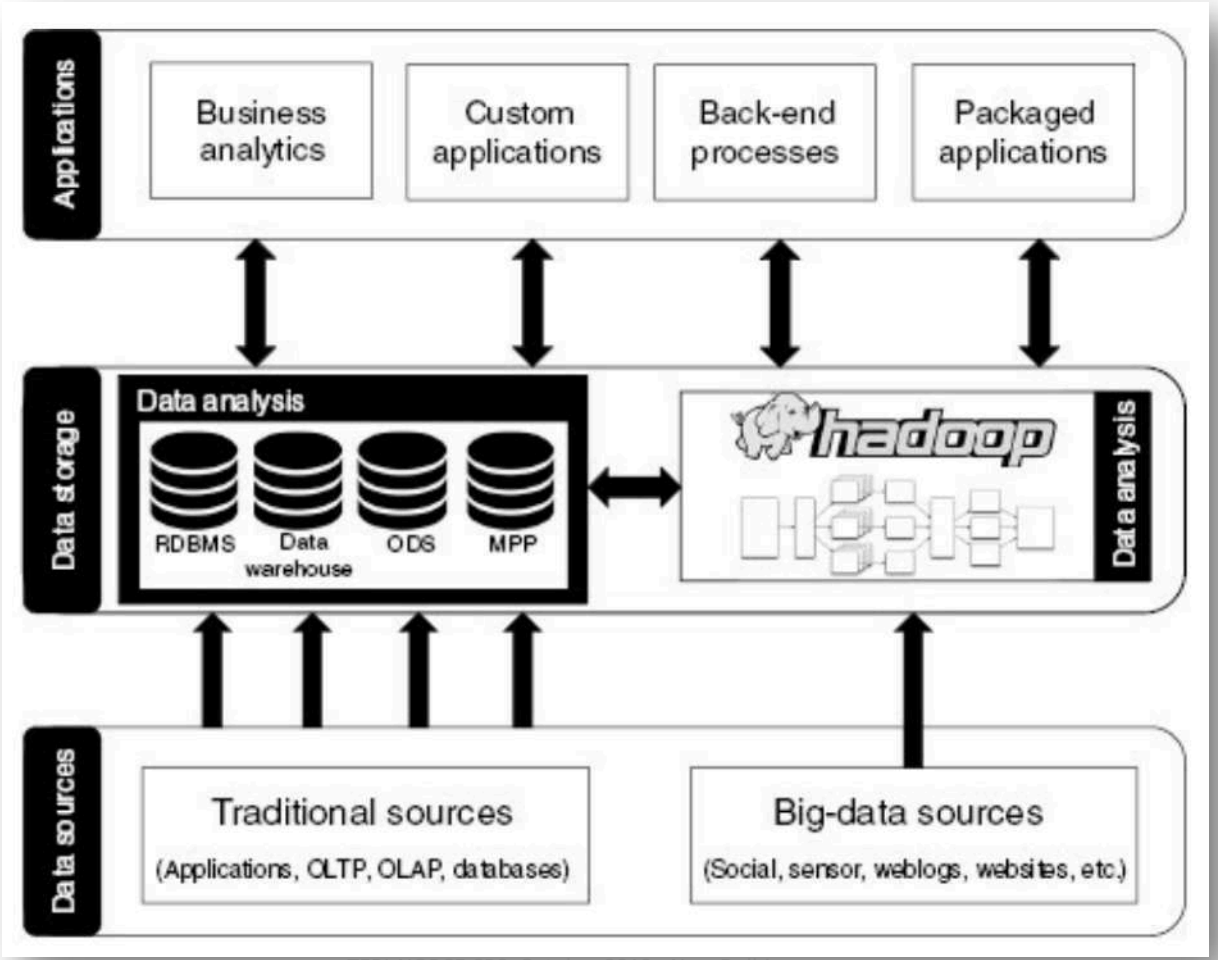
Unstructured data

- Each instance in the data set may have its own internal structure, and this structure is not necessarily the same in every instance

CRISP-DM Model //



Data Science ecosystem //

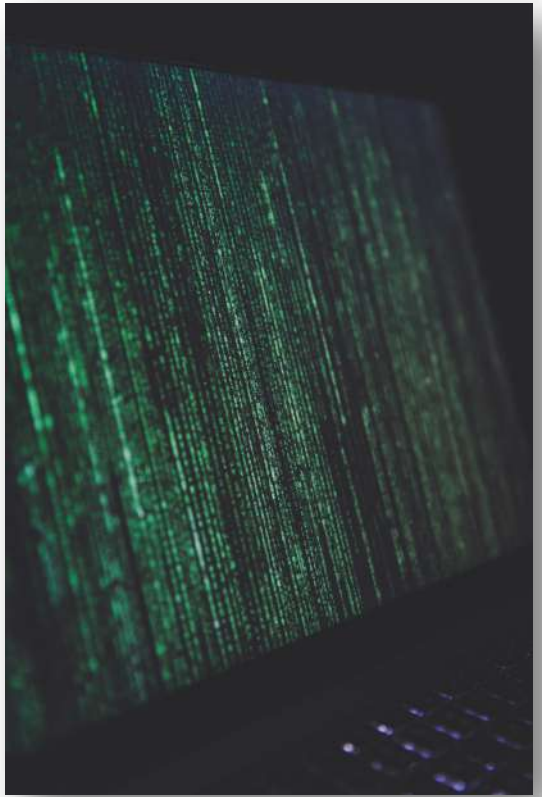


ML involves a two step process /

Identify useful patterns given some data (modelling). Some methods include:

- Decision trees
- Regression models
- Neural networks

Once a model is created, it is used for analysis



Supervised learning /

Learn a function that maps from the values of the attributes to the value of another (target) attribute.

«Supervised» because each of the instances in the data set lists both the input values and the output (target) value for each instance.

Each instance in the data set must be labeled with the value of the target attribute.

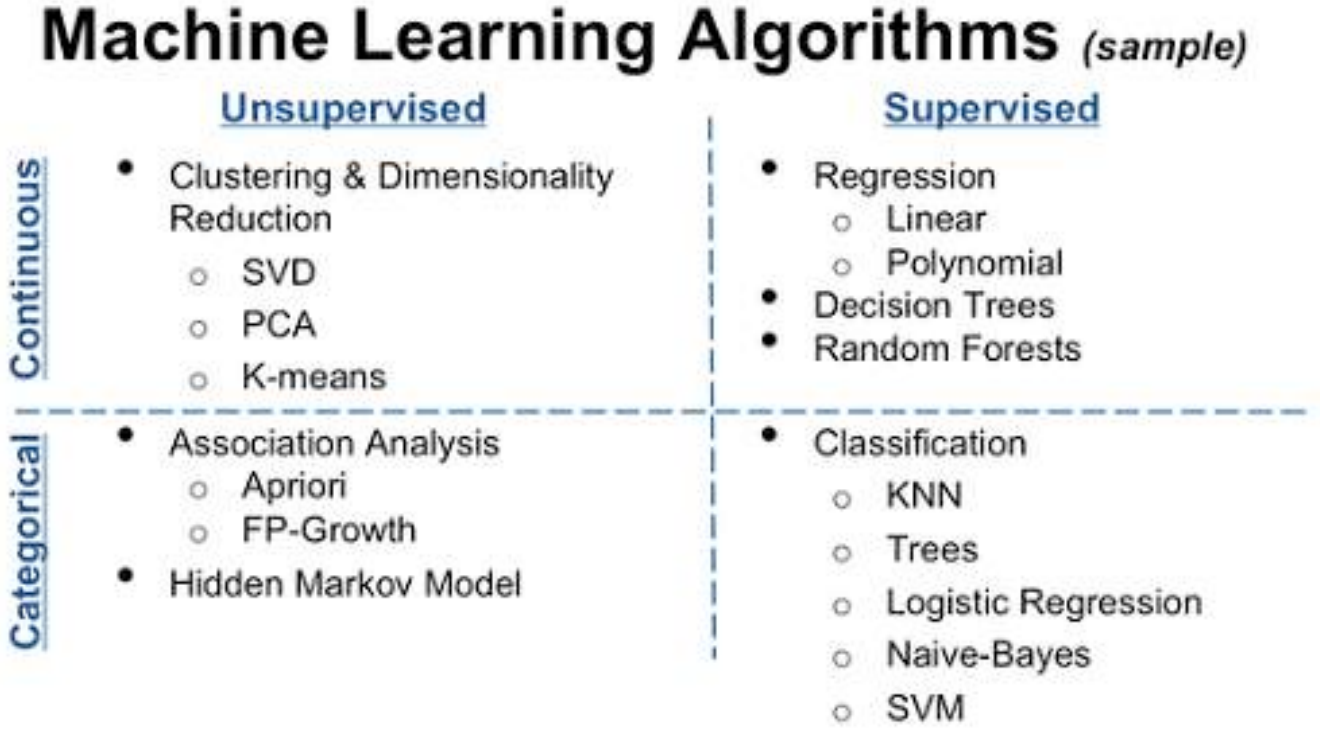
Unsupervised learning /

There is no target attribute. The algorithm has the more general task of looking for regularities in the data.

- Cluster analysis
- Euclidean distance

«Supervised» because each of the instances in the data set lists both the input values and the output (target) value for each instance.

Each instance in the data set must be labeled with the value of the target attribute.



Source: https://tinyurl.com/yydmyo8x

Activity //
Q1 and Q2

Hypothetical dataset

Ten years of accumulated All India Rainfall in mm for the monsoon season and pre-monsoon DJF mean ENSO values for the corresponding year

All_India_Rainfall (mm)	El_Nino_Southern_Oscillation
1030	-1.50
950	0.50
980	0.80
1099	-1.80
1100	-1.80
1140	-1.90
999	0.90
950	0.85
988	0.88

Definitions /

Let x_i be the value of the explanatory variable for the ith datapoint, and Y_i the random variable representing the response for the same datapoint

Assumptions /

$$E(Y_i) = \alpha + \beta x_i$$

The mean of the response variable Y_i depends on the value of x_i of the explanatory variable in a linear fashion

$$Y_i = \alpha + \beta x_i + \epsilon_i$$

The variation of the response variable Y_i about the mean is represented by a random variable ϵ_i

$$E(\epsilon_i) = 0$$

Note that for the mean of Y_i to be as equation 1 above, ϵ_i has to be zero

$$E(\epsilon_i) \sim N(0, \sigma^2)$$

The variance of ϵ_i is the same for all values of the explanatory variable (error variance). Also the different ϵ_i (i=1,2,…,n), are independent of each other

Alternative way to write the model /

$$aX + b \sim N(a\mu + b, a^2\sigma^2)$$

Property of distributions. Also, assuming that Y_i is a fixed quantity $\alpha + \beta x_i$ plus a random variable $N(0, \sigma^2)$, then:

$$Y_i = N(\alpha + \beta x_i, \sigma^2)$$

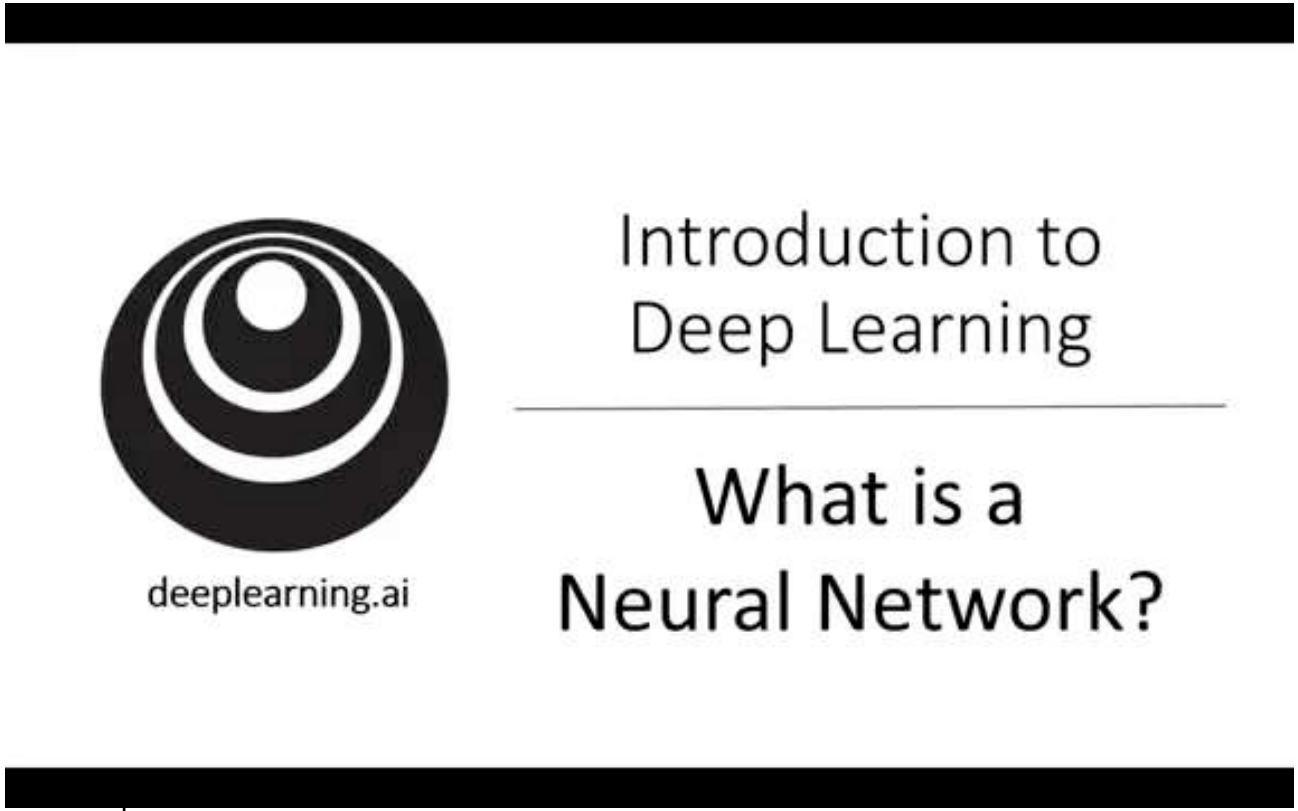
This is the simple linear regression model

Sometimes we need to transform the data. Provided that x>1, we can use the following transformation alternatives:

$$\ldots,\ x^{-2},\ x^{-1},\ x^{-1/2},\ \log x,\ x^{1/2},\ x^{1},\ x^{2},\ldots.$$

Reduce the high values in a dataset relative to low values

Stretch out high values relative to low ones



Deeplearning.ai
Video by Andrew Ng
URL: https://youtu.be/n1l-9lIMW7E

MATLAB Tech Talks
Video by Shyamal Patel
URL: https://www.youtube.com/watch?v=3cSjsTKtN9M

Example //
Recognising objects with Convolutional Neural Networks
Michel's solution, based on video from Deeplearning.ai

contributed articles

DOI:10.1145/2500499

Big data promises automated actionable knowledge creation and predictive models for use by both humans and computers.

BY VASANT DHAR

Data Science and Prediction

USE OF THE term "data science" is increasingly common, as is "big data." But what does it mean? Is there something unique about it? What skills do "data scientists" need to be productive in a world deluged by data? What are the implications for scientific inquiry? Here, I address these questions from the perspective of predictive modeling.

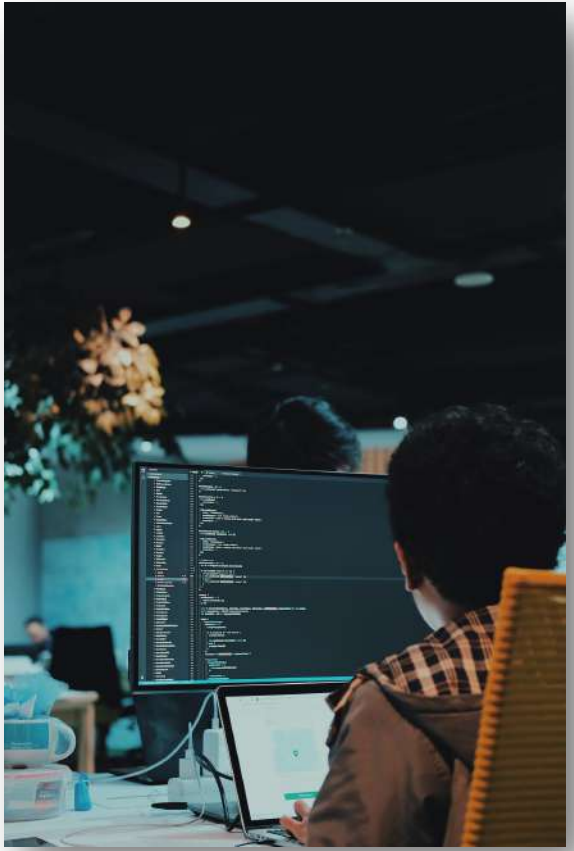
The term "science" implies knowledge gained through systematic study. In one definition, it is a systematic enterprise that builds and organizes knowledge in the form of testable explanations and predictions.[1] Data science might therefore imply a focus involving data and, by extension, statistics, or the systematic study of the organization, properties, and analysis of data and its role in inference, including our confidence in the inference. Why then do we need a new term like data science when we have had statistics for centuries? The fact that we now have huge amounts of data should not in and of itself justify the need for a new term.

The short answer is data science is different from statistics and other existing disciplines in several important ways. To start, the raw material, the "data" part of data science, is increasingly heterogeneous and unstructured—text, images, video—often emanating from networks with complex relationships between their entities. Figure 1 outlines the relative expected volumes of unstructured and structured data from 2008 to 2015 worldwide, projecting a difference of almost 200 petabytes (PB) in 2015 compared to a difference of 50PB in 2012. Analysis, including the combination of the two types of data, requires integration, interpretation, and sense making that is increasingly derived through tools from computer science, linguistics, econometrics, sociology, and other disciplines. The proliferation of markup languages and tags is designed to let computers interpret data automatically, making them active agents in the process of decision making. Unlike early markup languages (such as HTML) that emphasized the display of information for human consumption, most data generated by humans and computers today is for consumption by computers; that is, computers increasingly do background work for each other and make decisions automatically. This scalability in decision making has become possible because of big data that serves as the raw material for the creation of new knowledge; Watson, IBM's "Jeopardy!" champion, is a prime illustration of an emerging machine intelligence fueled by data and state-of-the-art analytics.

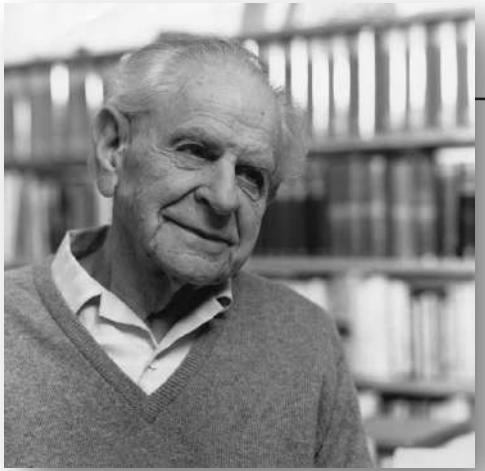
» key insights

- Data science is the study of the generalizable extraction of knowledge from data.
- A common epistemic requirement in assessing whether new knowledge is actionable for decision making is its predictive power, not just its ability to explain the past.
- A data scientist requires an integrated skill set spanning mathematics, machine learning, artificial intelligence, statistics, databases, and optimization, along with a deep understanding of the craft of problem formulation to engineer effective solutions.

In which ways is data science different from statistics?



How much do you agree/disagree with Karl Popper's argument for evaluating a theory and scientific progress?



Karl Popper
(Wikipedia)

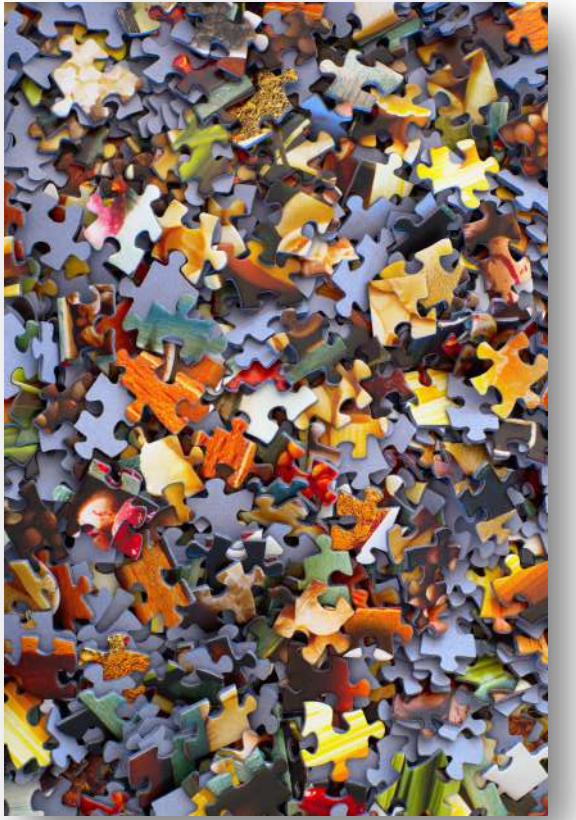
What powerful capability is machine learning able to give us?

«…simpler models are more likely to hold up on future observations than more complex ones, all else being equal» (p.68)

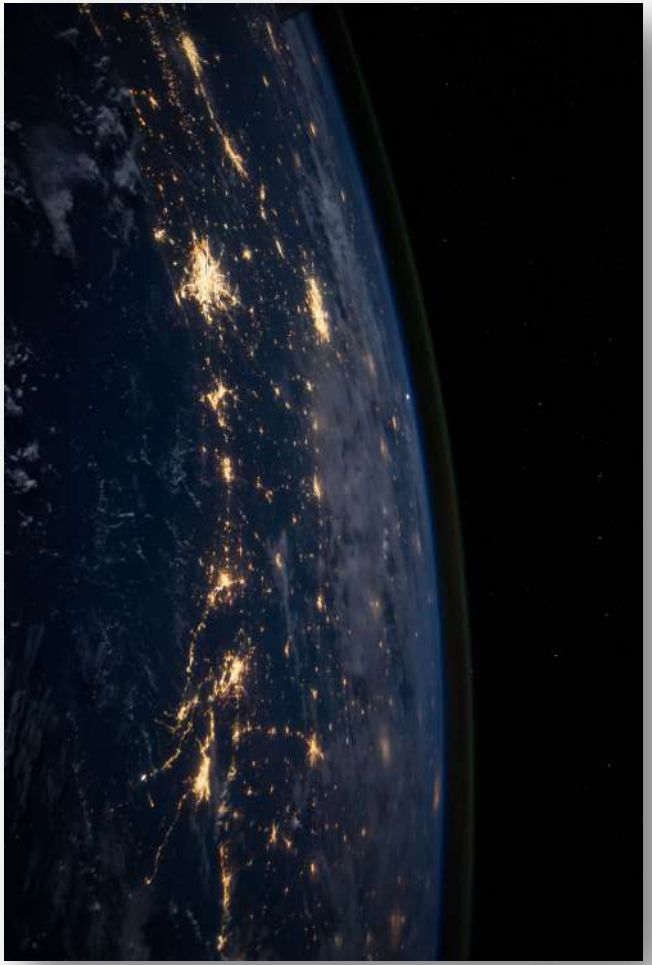
How much do you agree/disagree with this statement?

Why are problem formulation skills so important?

Where do errors come from?



How does the Internet contribute to inexpensive large-scale randomized experiments?



Why is earth science considered to be one of the «data-starved areas of inquiry»?

What should universities do to improve their students' data science skills?

How would you like to integrate data science in your own research work?

Summary and future work //

In conclusion //

Data science has become a growing field, thanks to many past researchers in the field. Climate science will benefit even more in the years to come, as more and more researchers and students learn how to apply different algorithms to solve climate problems.

We hope this Climate Research School can inspire you as well! In the next days, you will be learning several tools that can become part of your professional toolbox!

Thank you for your attention!
Feel free to get in touch with us //

mido@norceresearch.no
+47 4760 2340