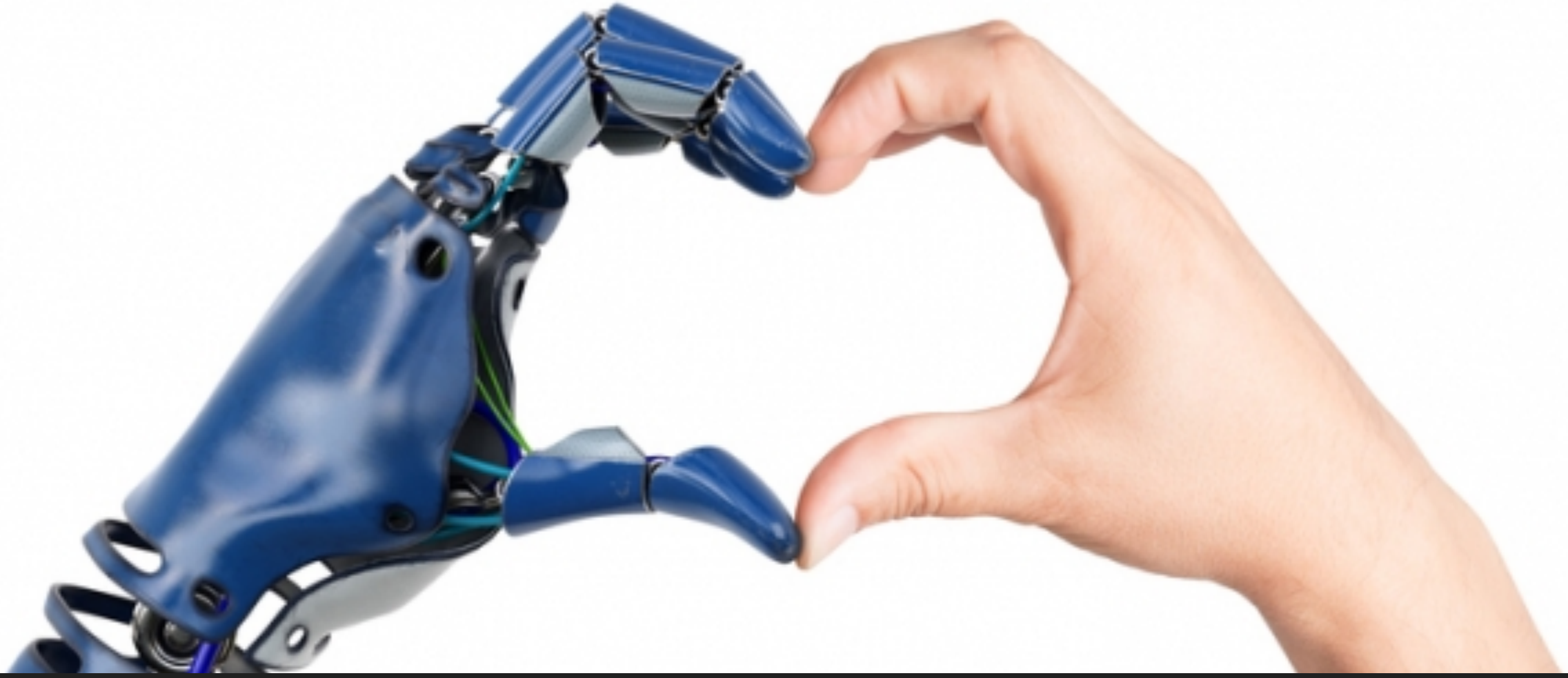


Models

Data Science Applications in R

Definition

Definition

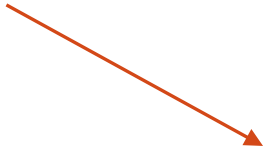
A MODEL DESCRIBES A SYSTEM USING MATHEMATICAL LANGUAGE

Definition

A MODEL DESCRIBES A **SYSTEM** USING MATHEMATICAL LANGUAGE

Definition

A MODEL DESCRIBES A **SYSTEM** USING MATHEMATICAL LANGUAGE



a set of things working together as parts of a mechanism or an interconnecting network; a complex whole.

Definition

A MODEL DESCRIBES A **SYSTEM** USING MATHEMATICAL LANGUAGE

MODELS HELP TO EXPLAIN A SYSTEM AND TO STUDY
THE RELATIONSHIPS THAT EXIST BETWEEN ITS ELEMENTS

Definition

A MODEL DESCRIBES A **SYSTEM** USING MATHEMATICAL LANGUAGE

MODELS HELP TO EXPLAIN A SYSTEM AND TO STUDY
THE RELATIONSHIPS THAT EXIST BETWEEN ITS ELEMENTS

THE PURPOSE OF THE MODELS IS TO PREDICT THE
BEHAVIOR OF THE SYSTEM UNDER STUDY

Models

MODELS ARE COMPOSED BY RELATIONS AND VARIABLES

Models

MODELS ARE COMPOSED BY RELATIONS AND VARIABLES

RELATIONS ARE DESCRIBED BY OPERATORS (MATHEMATICAL OR LOGICAL)

Models

MODELS ARE COMPOSED BY RELATIONS AND VARIABLES

RELATIONS ARE DESCRIBED BY OPERATORS (MATHEMATICAL OR LOGICAL)

VARIABLES ARE ABSTRACTION OF THE PARAMETERS OF THE SYSTEM

Models

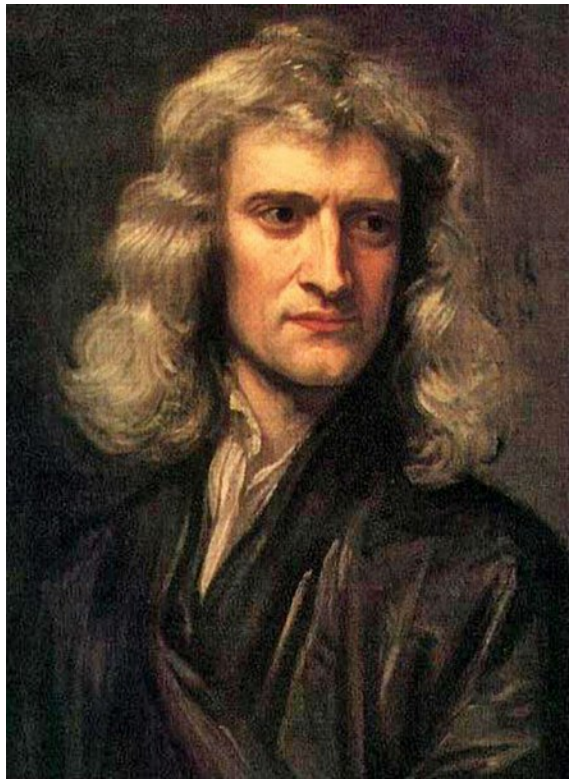
MODELS ARE COMPOSED BY RELATIONS AND VARIABLES

RELATIONS ARE DESCRIBED BY OPERATORS (MATHEMATICAL OR LOGICAL)

VARIABLES ARE ABSTRACTION OF THE PARAMETERS OF THE SYSTEM

VARIABLES CAN BE **MEASURED**

Popular Models



Isaac Newton

"Un corpo mantiene il proprio stato di quiete o di moto rettilineo uniforme, finché una forza non agisce su di esso".

"L'accelerazione di un corpo è direttamente proporzionale e ha la stessa direzione della forza netta agente su di esso, mentre è inversamente proporzionale alla sua massa".

"Per ogni forza che un corpo esercita su di un altro corpo , ne esiste istantaneamente un'altra uguale in modulo e direzione, ma opposta in verso, causata dal corpo che agisce sul corpo".

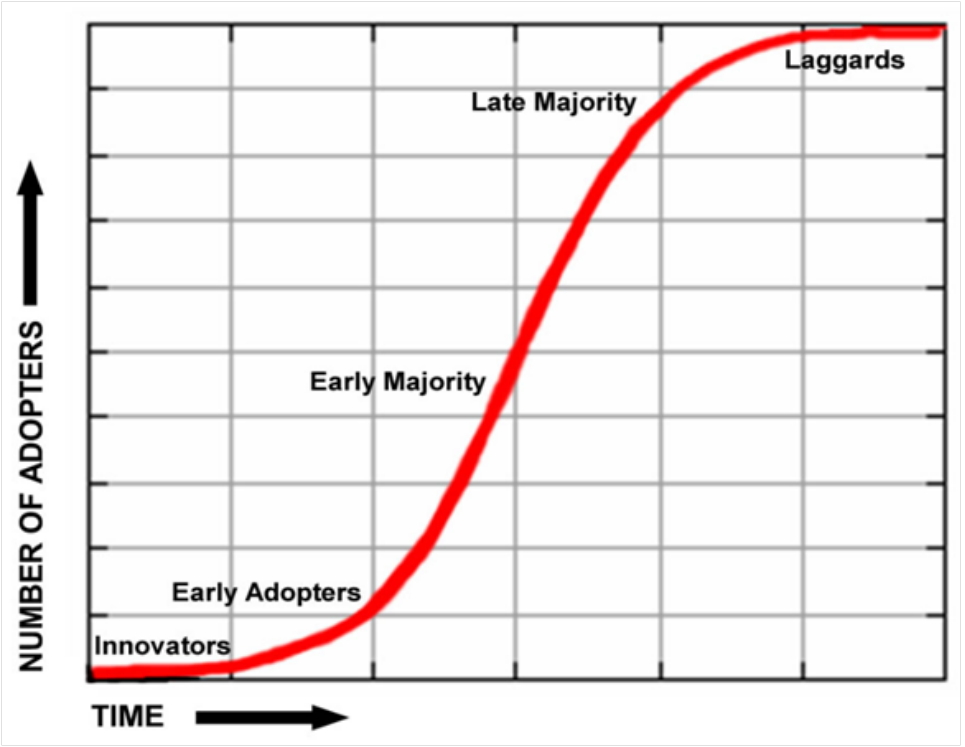
Popular Models

Vilfredo Pareto

“Nella maggior parte dei sistemi naturali circa il 20% delle cause provoca l'80% degli effetti.”.

Popular Models

Everett Rogers



How to Design a Model

Knowledge

Where to start?

Reality

How to Design a Model

Knowledge

Top/Down

Induction

Deduction

Bottom/Up

Reality

An actionable definition of knowledge

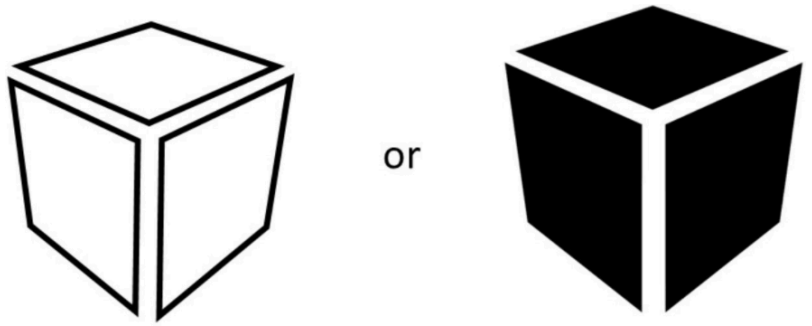
White Boxes, Black Boxes

WE HAVE TWO CATEGORIES OF SYSTEMS:

- White Box: All the information necessary to describe the behavior of the system is known
- Black Box: Only inputs and outputs of the system are known

All systems are placed in the white Box-Black box spectrum

White Boxes, Black Boxes



Examples?

White Boxes, Black Boxes

Systems vs Models

White Boxes, Black Boxes

What do models do for us?

Definitions

TO LEARN:

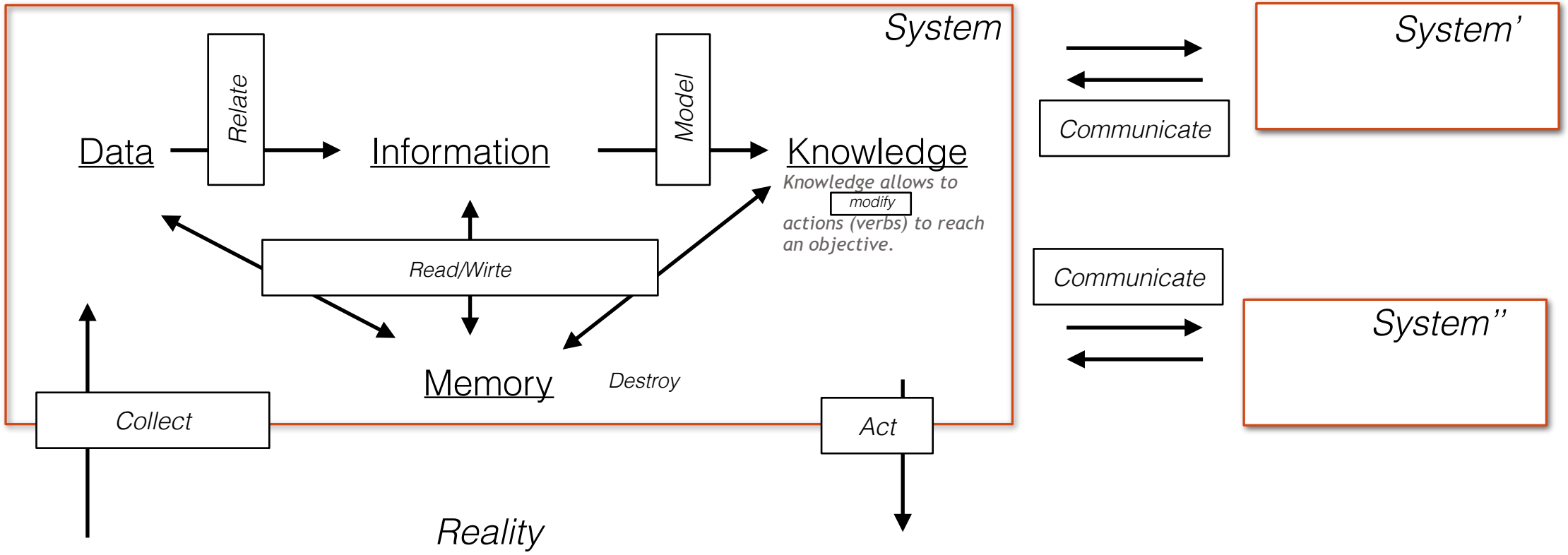
"to gain knowledge or understanding of or skill in by study, instruction, or experience"

(Merriam Webster English Dictionary, 2005)

MACHINE LEARNING:

- Built the model from data
- Many data but no much explanations of how the system works

A Process for Knowledge Collection



When ML is Usefull

- It is difficult to define a procedure to solve a problems (e.g. facial recognition)

When ML is Usefull

- It is difficult to define a procedure to solve a problems (e.g. facial recognition)
- The amount of knowledge necessary to solve some problems may be too large to be made explicit (e.g. medical diagnosis)

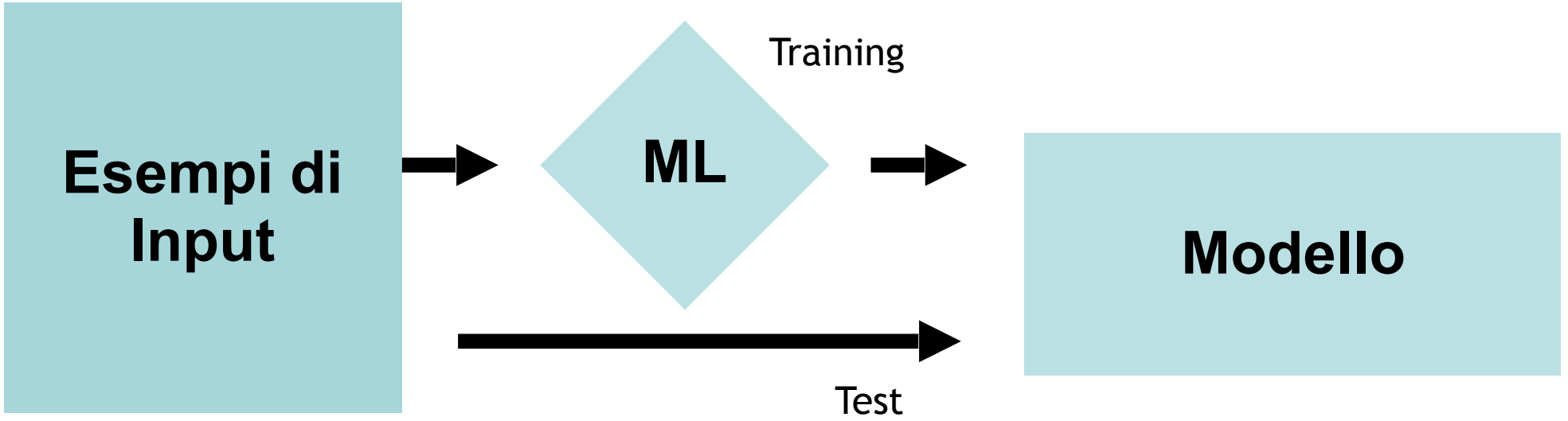
When ML is Usefull

- It is difficult to define a procedure to solve a problems (e.g. facial recognition)
- The amount of knowledge necessary to solve some problems may be too large to be made explicit (e.g. medical diagnosis)
- Some systems change over time and new knowledge must be constantly disclosed to have effective models. A continuous redesign of the model to solve the problem is impossible.

When ML is Usefull

- It is difficult to define a procedure to solve a problems (e.g. facial recognition)
- The amount of knowledge necessary to solve some problems may be too large to be made explicit (e.g. medical diagnosis)
- Some systems change over time and new knowledge must be constantly disclosed to have effective models. A continuous redesign of the model to solve the problem is impossible.
- We have many examples of solutions to these problems

The Approach



What is the goal?

What to pay attention

- Learning Task:
  - What do you want to classify?
- Data and Assumption:
  - Which data do we have?
  - Whats their quality?
  - What Hypothesis and assumptions can we make about the problem?
- Representation:
  - What is the best representation of the observations (in terms of variables) to generate the model?
- Assessment:
  - How good is the model?

White Boxes, Black Boxes

Open the Black Box
Data-Driven Explanation of Black Box Decision Systems

DINO PEDRESCHI, University of Pisa, Italy
FOSCA GIANNOTTI, ISTI-CNR of Pisa, Italy
RICCARDO GUIDOTTI, ISTI-CNR & University of Pisa, Italy
ANNA MONREALE, University of Pisa, Italy
LUCA PAPPALARDO, ISTI-CNR of Pisa, Italy
SALVATORE RUGGIERI, University of Pisa, Italy
FRANCO TURINI, University of Pisa, Italy

ACM Reference Format:
Dino Pedreschi, Fosca Giannotti, Riccardo Guidotti, Anna Monreale, Luca Pappalardo, Salvatore Ruggieri, and Franco Turini. 2018. Open the Black Box Data-Driven Explanation of Black Box Decision Systems. 1, 1 (June 2018), 15 pages. https://doi.org/0000001.0000001

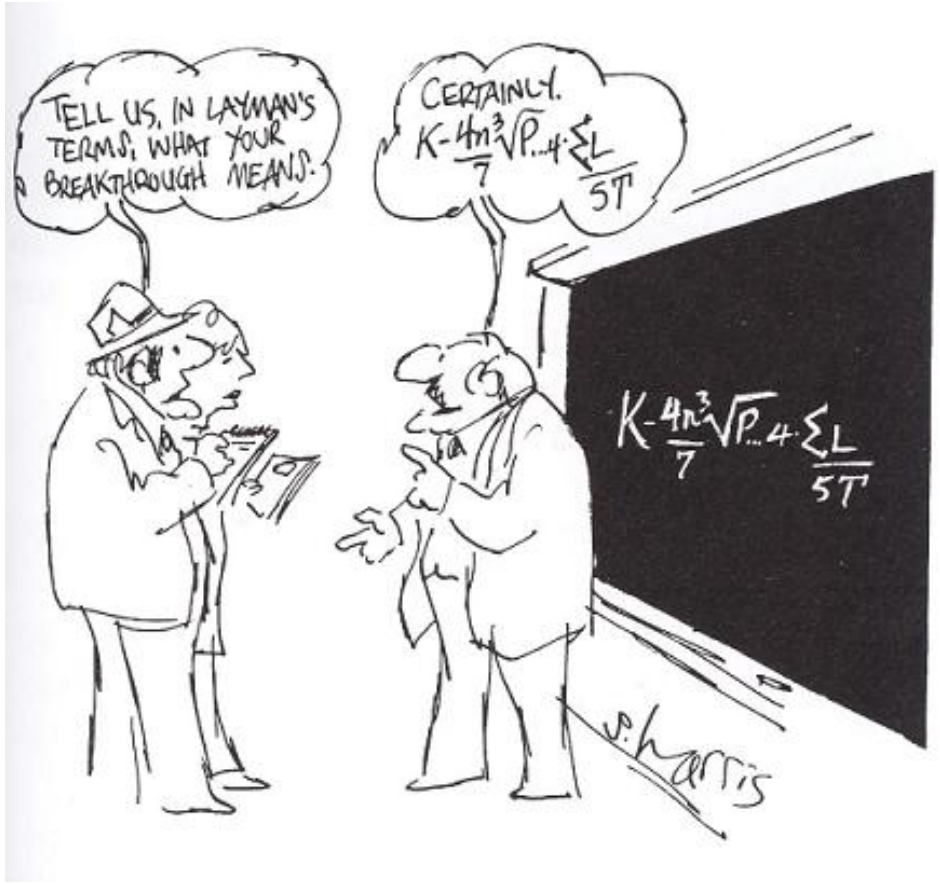
1 INTRODUCTION

The last decade has witnessed the rise of a black box society [33]. Ubiquitous obscure algorithms, often based on sophisticated machine learning models trained on (big) data, which predict behavioural traits of individuals, such as credit risk, health status, personality profile. Black boxes map user features into a class or a score without explaining why, because the decision model is either not comprehensible to stakeholders, or secret. This is worrying not only in terms of the lack of transparency, but also due to the possible biases hidden in the algorithms. Machine learning (ML) constructs predictive models and decision-making systems based on (possibly big) data, i.e., the digital traces of human activities (opinions, movements, lifestyles, etc.). Consequently, these models may reflect human biases and prejudices, as well as collection artifacts, possibly leading to unfair or simply wrong decisions. Many controversial cases have already highlighted that delegating decision-making to black box algorithms is critical in many sensitive domains, including crime prediction, personality scoring, image classification, personal assistance, and more (see box "The danger of black boxes".)

White Boxes, Black Boxes

Implications?

Another Problem…



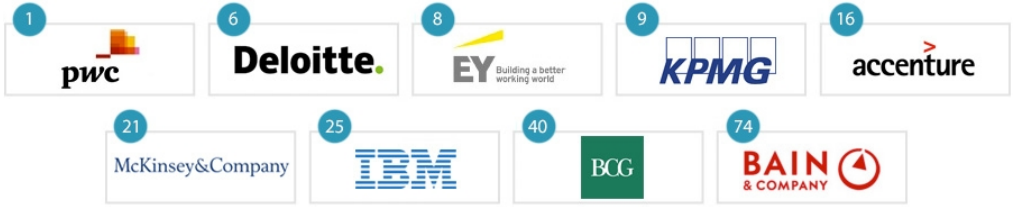
Another Problem…

ALL MODELS ARE WRONG BUT SOME ARE USEFUL

<< Now it would be very remarkable if any system existing in the real world could be *exactly* represented by any simple model. However, cunningly chosen parsimonious models often do provide remarkably useful approximations. For example, the law PV = RT relating pressure P, volume V and temperature T of an "ideal" gas via a constant R is not exactly true for any real gas, but it frequently provides a useful approximation and furthermore its structure is informative since it springs from a physical view of the behavior of gas molecules.
For such a model there is no need to ask the question "Is the model true?". If "truth" is to be the "whole truth" the answer must be "No". The only question of interest is "Is the model illuminating and useful?". >>

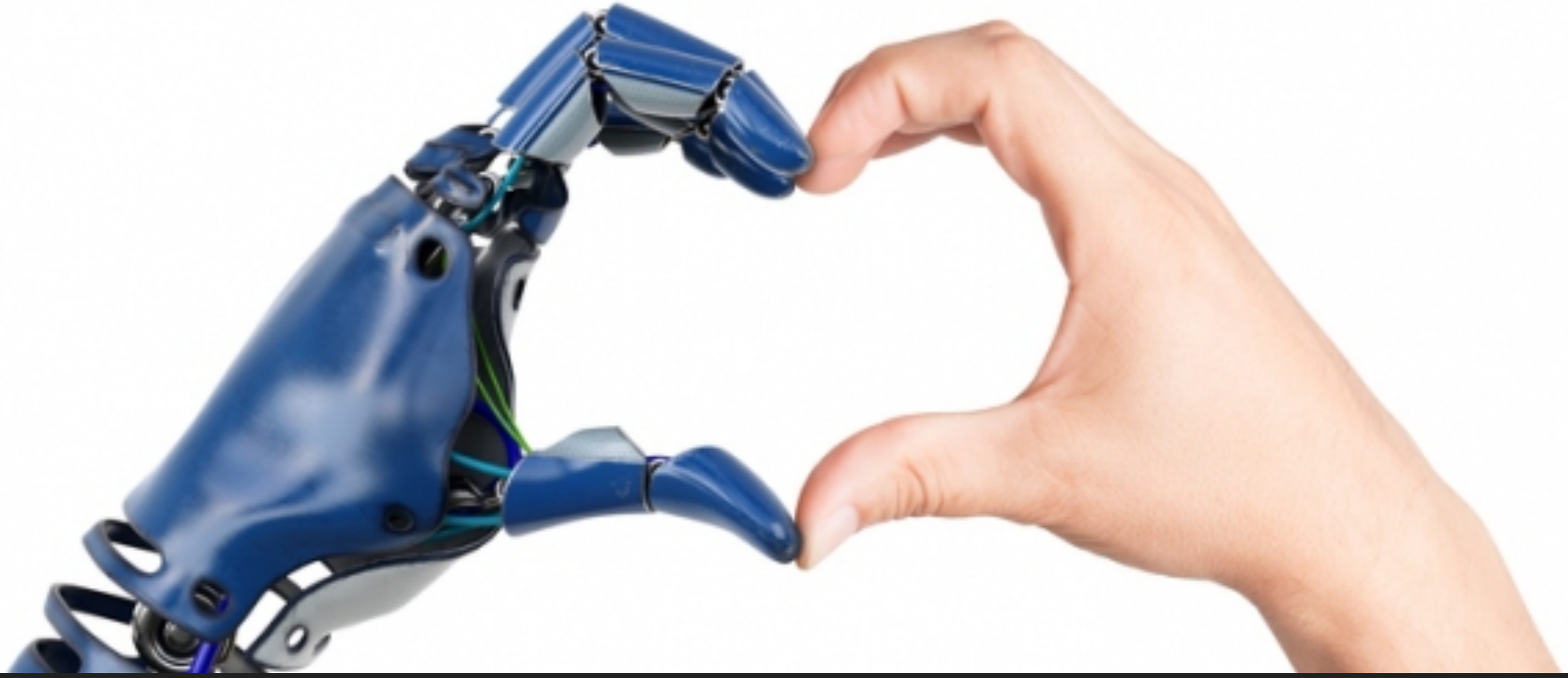
George E. P. Box

Modelling for Competitive Intelligence



LET'S THINKS ABOUT:

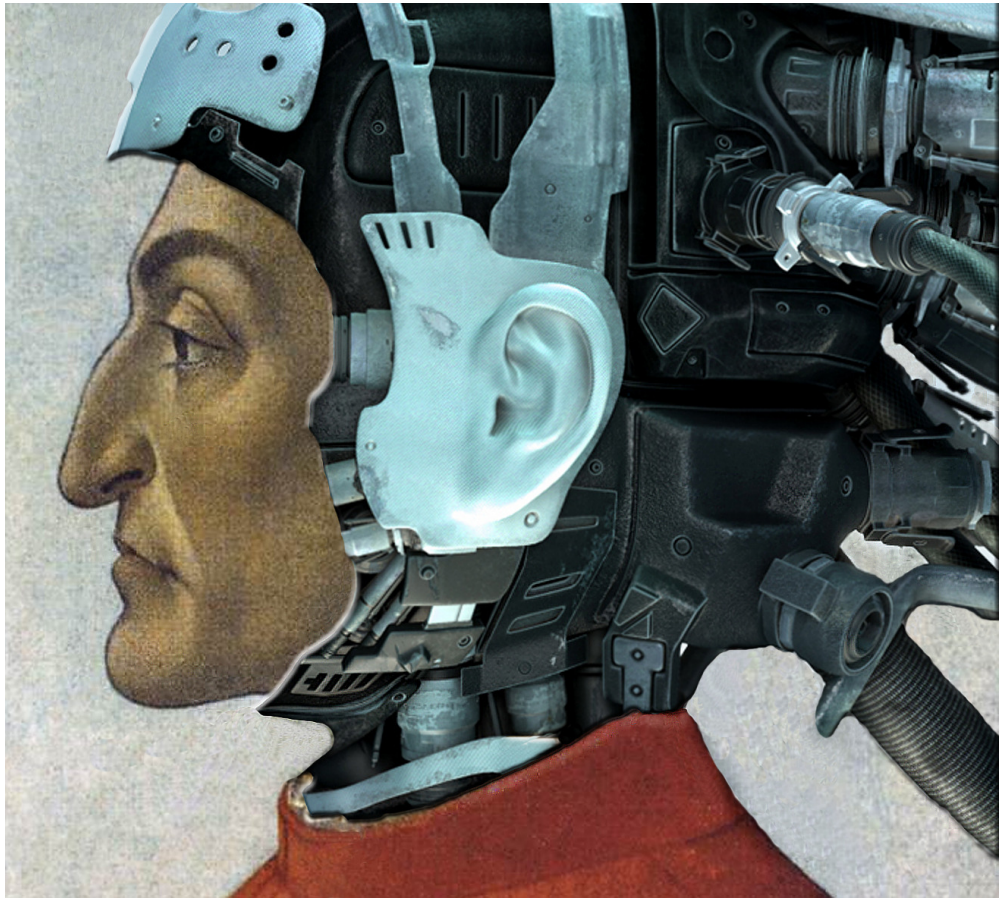
- WHICH IS THE GOAL OF COMPETITIVE INTELLIGENCE?
- WHY IS MACHINE LEARNING USEFUL?
- WHAT PROBLEMS CAN WE HAVE USING THESE METHODS?
- WHICH SOLUTIONS CAN WE ADOPT?



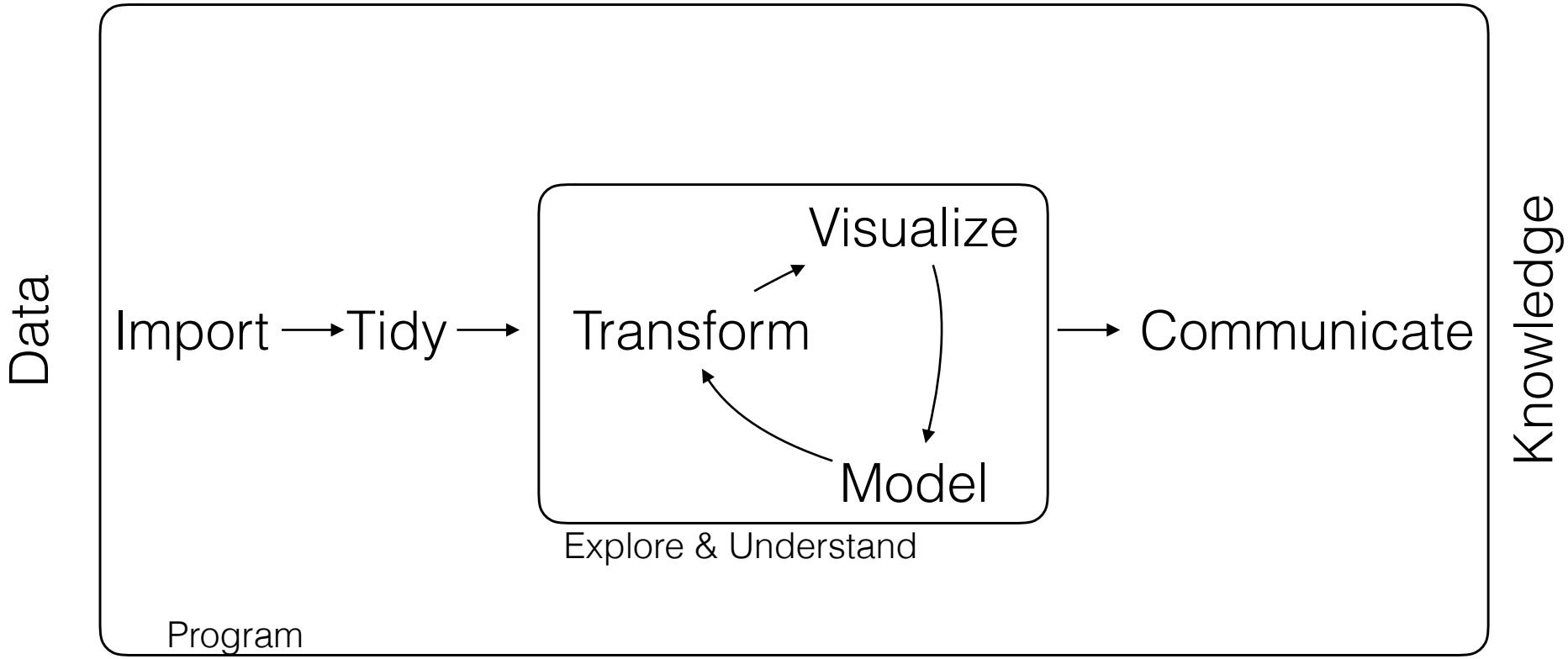
Machine Learning for Text Mining

Machine Learning for Text Mining

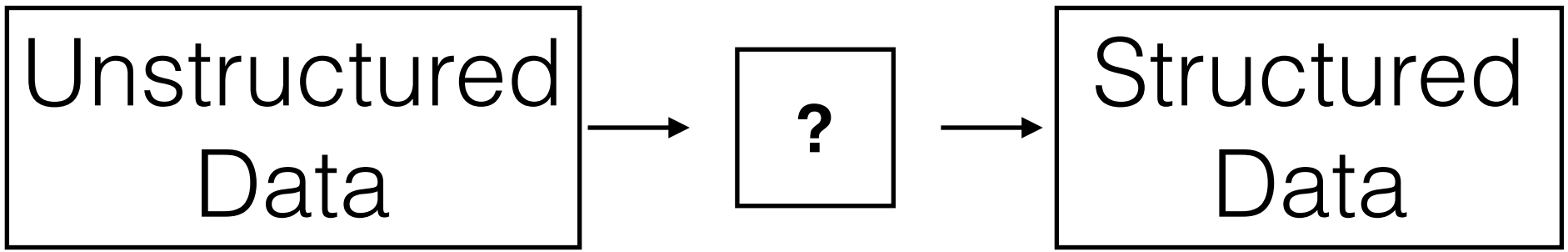
Texts are **dark**-boxes

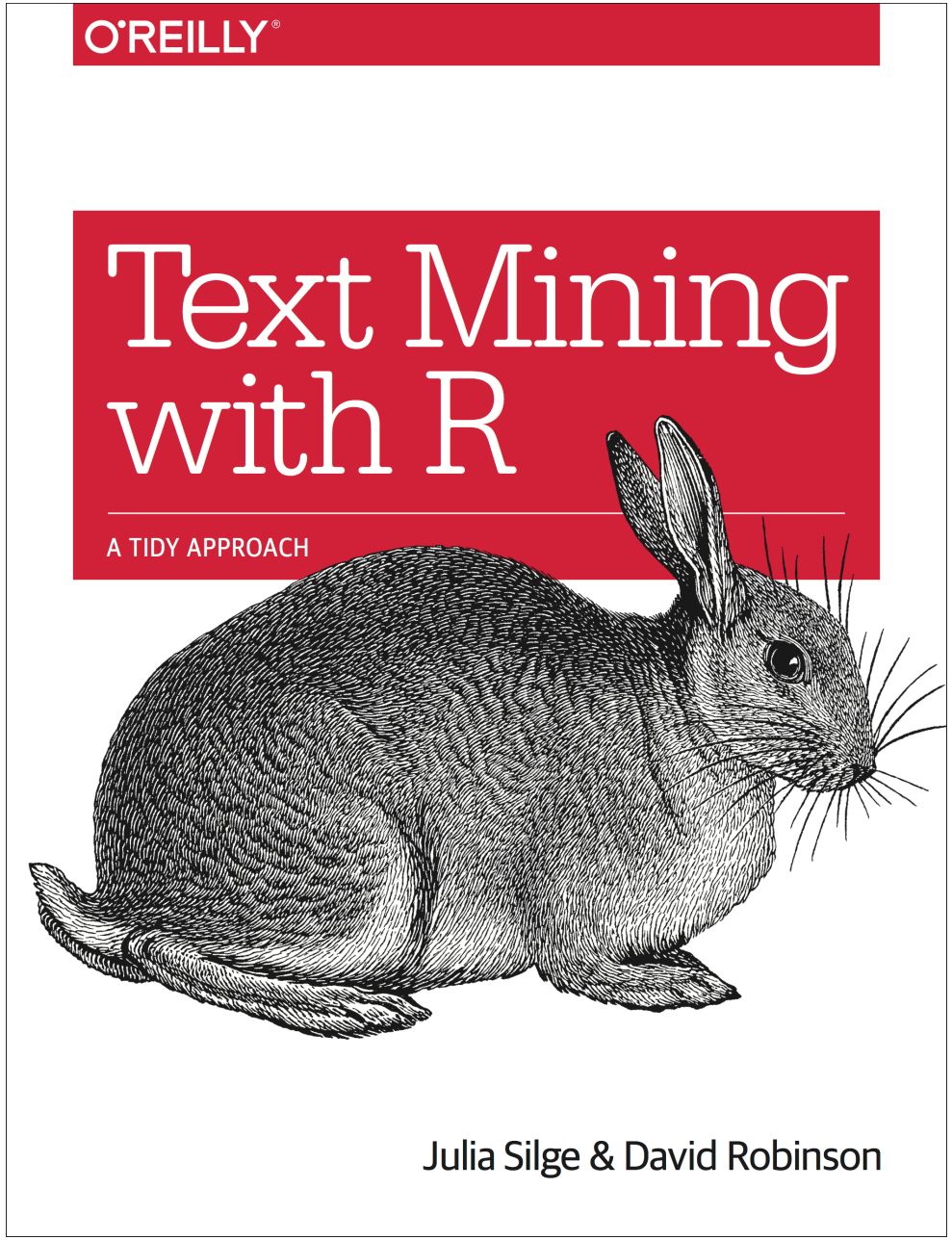


But we have to do other things before modelling….


Data
Import
Tidy
Visualize
Transform
Model
Explore & Understand
Communicate
Knowledge
Program


If data == text?


Unstructured Data
?
Structured Data


O'REILLY
Text Mining with R
A TIDY APPROACH
Julia Silge & David Robinson

The tidy text format

Tidy data has a specific structure:

- Each variable is a column
- Each observation is a row
- Each type of observational unit is a table

We thus define the tidy text format as being a table with one-**token**-per-row.

Tokens

- **Tokens**= meaningful unit of text

- **Tokenizations**= process of splitting text into tokens.

For tidy text mining, the token that is stored in each row is most often a single word, but can also be an n-gram, sentence, or paragraph.

N-grams

A contiguous sequence of n words.

Example (**n**=2):

The rain in Spain falls mainly on the plain

The rain, rain in , in Spain, Spain falls, falls mainly …

Skip-grams

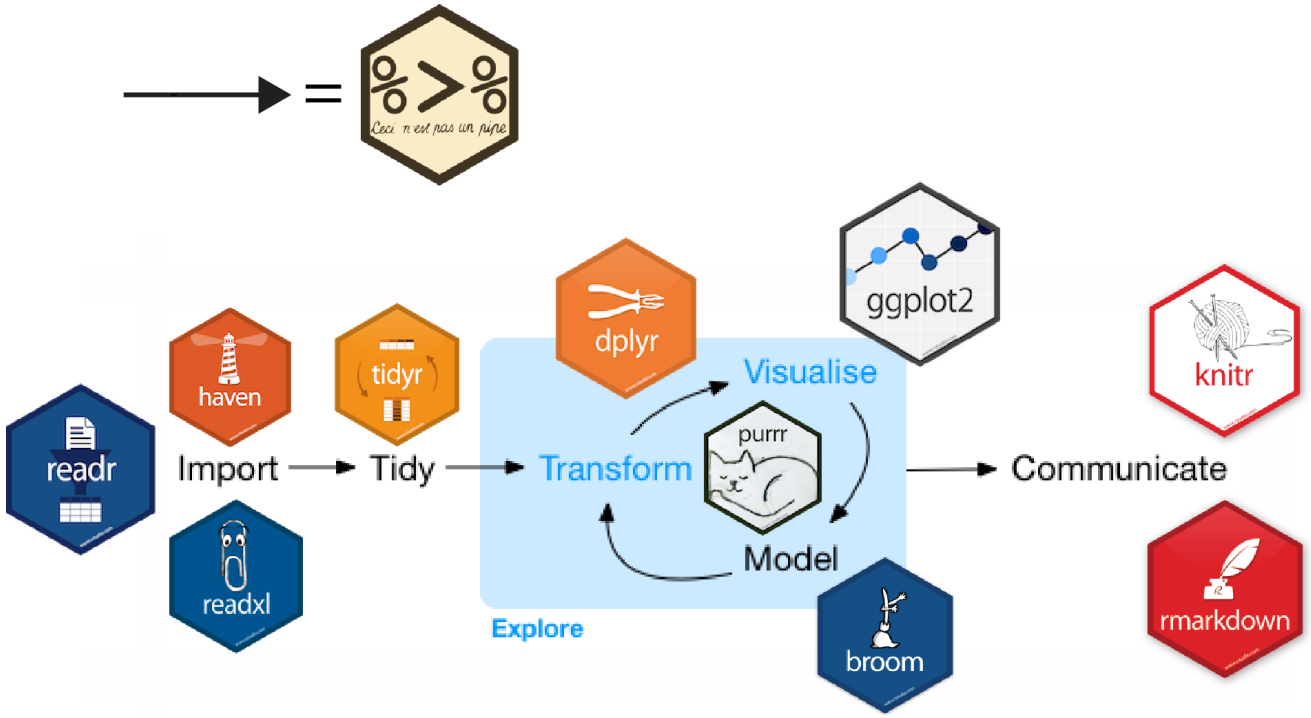
A generalization of n-grams in which the components (typically words) need not be consecutive in the text under consideration, but may leave gaps that are skipped over.

Example (**n**=2, **k**=1):

The rain in Spain falls mainly on the plain

The in, rain Spain, in falls, Spain mainly, falls on, mainly the, and on plain.

Advantages of tidy format



data.frames

factors

strings

Other popular structures

String: Text can, of course, be stored as strings, i.e., character vectors, within R, and often text data is first read into memory in this form.

Corpus: These types of objects typically contain raw strings annotated with additional metadata and details.

Document-term matrix: This is a sparse matrix describing a collection (i.e., a corpus) of documents with one row for each document and one column for each term.

Other popular structures

Tidytext approach doesn't expect a user to keep text data in a tidy form at all times during an analysis. The package includes functions to tidy() objects from popular text mining R packages such as:

- tm (Ingo Feinerer and Meyer 2008)
- quanteda (Benoit and Nulty 2016).

This allows, for example, a workflow where importing, filtering, and processing is done using dplyr and other tidy tools, after which the data is converted into a document-term matrix for machine learning applications. The models can then be re-converted into a tidy form for interpretation and visualization with ggplot2.

Text transformation processes

Tokenisation: split the text in meaningful units

Lemmatisation: extract the lemma (root form) of the word

POS tagging: assign a Part of Speech to each word

Machine Learning for Text Analysis

From: U.S. Bank <service@usbank.com>
Subject: **Customer Service**
Date: December 8, 2008 5:25:15 AM PST
To: undisclosed-recipients:;

This is a reminder that your U.S. Bank Account needs to be verified.
To continue using your card, please verify your account immediately.

To verify your account, please click the link below, log in and follow the provided steps:

http://www4.usbankv.com/internetBanking/?LoginRouter

Regards,
U.S. Bank

Machine Learning for Text Analysis

Machine Learning for Text Analysis

ESISTE LA TENDENZA NATURALE DEI LETTORI AD ANALIZZARE L'INTENTO EMOTIVO DEL CONTENUTO DI UN TESTO, PER INFERIRE SE QUESTO È POSITIVO O NEGATIVO.

Some readings