FUTURE EDUCATION MODENA

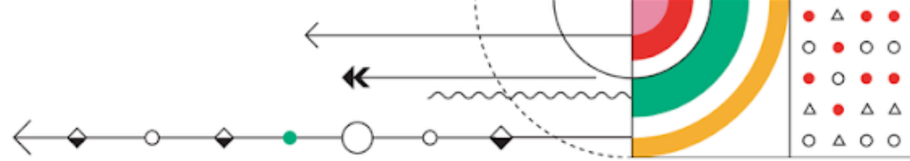
School of Data

I dati dalla sfida all'impatto

Lezione 1

Lorenzo Andreoli
Data Scientist @FEM

FUTURE
EDUCATION
MODENA



Cos’è FEM

Un centro internazionale per elevare ricerca, qualità e impatto dell'educazione in società

→ **Attività di ricerca** sulla frontiera dell’innovazione educativa

→ **Design e Sperimentazione** di esperienze educative altamente innovative

→ **Attività formative** e di divulgazione

→ **Incubazione** e accelerazione di *startup* educative

FUTURE EDUCATION MODENA

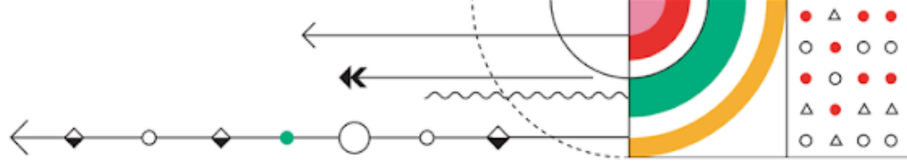
Su di me

LORENZO ANDREOLI

- Educational Data Scientist **@FEM**
- Mi piace utilizzare i dati per migliorare la vita delle persone
- Originario di **Reggio** ma ho sempre vissuto a Milano prima di trasferirmi per studio e lavoro a **Maastricht** e poi **Berlino**
- Alle medie non andavo benissimo ma poi mi son messo a studiare (non è mai troppo tardi)
- Ho trovato la mia passione all'intersezione tra matematica, storia, e scienze umane.
- Quando andavo alle medie il programma più utilizzato era MSN

Source

FUTURE
EDUCATION
MODENA



Ditemi di voi?

- Nome
- Materia preferita?
- App / programma preferito? E perché?
- Cosa vorresti imparare dalla School of Data?

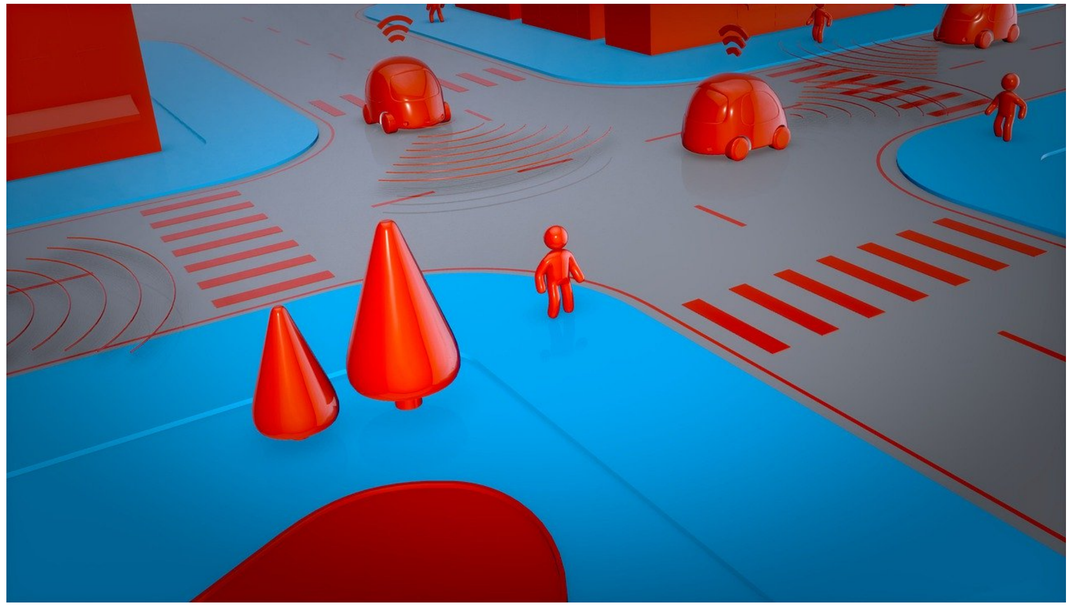
FUTURE EDUCATION MODENA

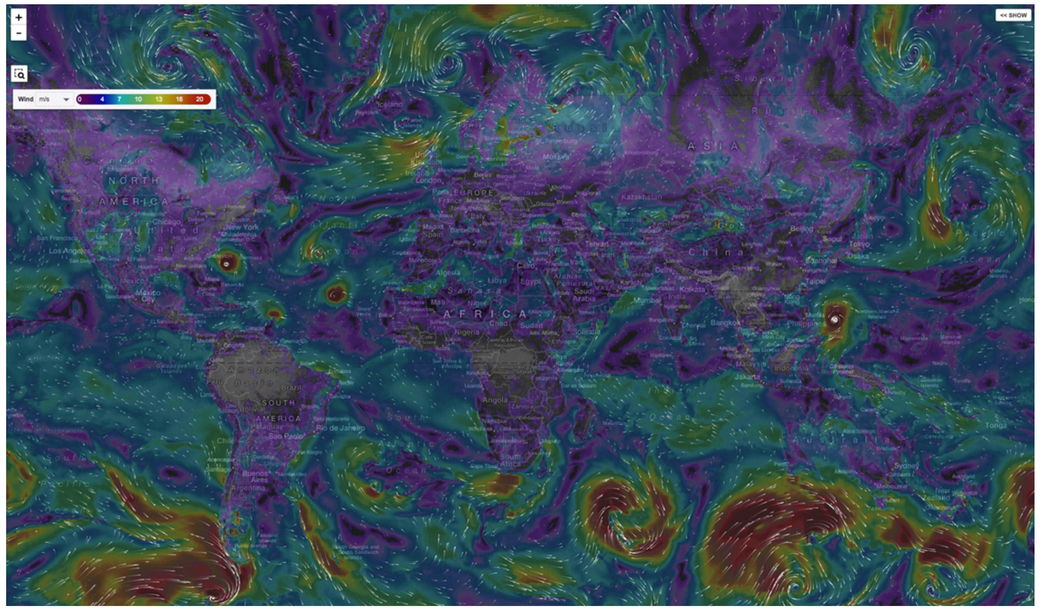
Il curricolo del percorso

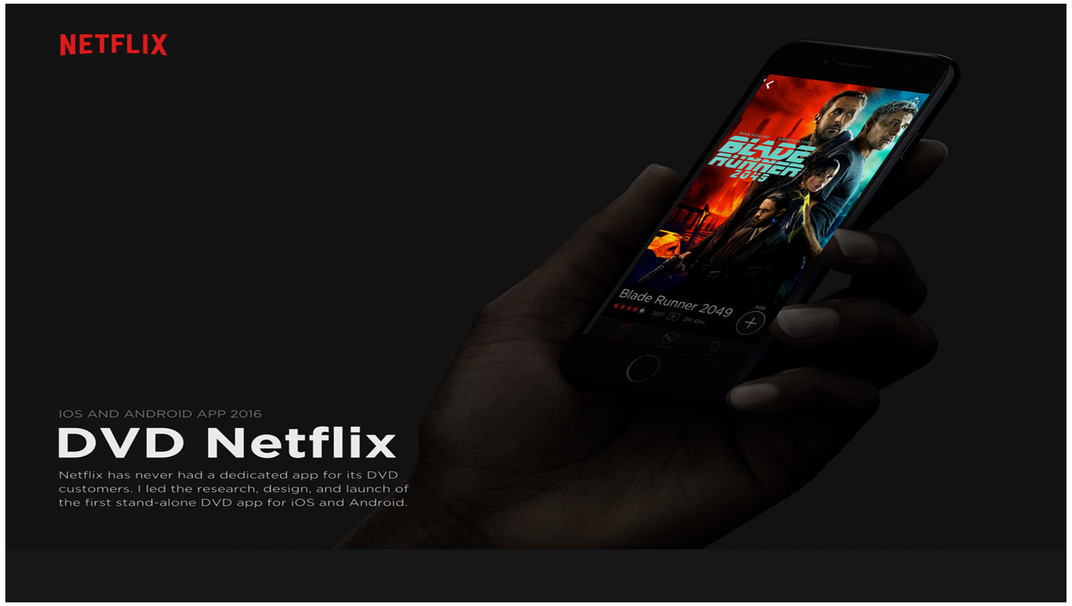
1 **Exploratory Data Analysis**	2 **Introduzione a Python e Data Science**	3 **Introduzione alla Statistica**	4 **Statistica Descrittiva**	5 **Data Visualization**
6 **Data Transformation & Data Cleaning**	7 **Data Visualization Avanzata**	8 **Data Humanities: Text Analysis**	9 **Open Maps: Geospatial Data**	10 **Preparazione alla Data Challenge**

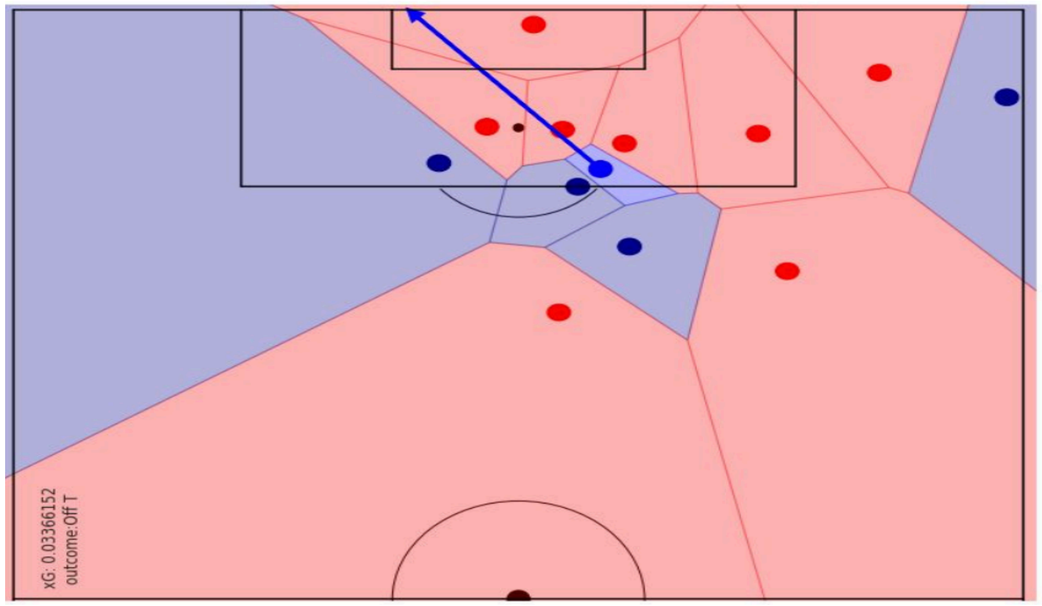
Data Challenge

In che ambito si utilizzano i dati e la Scienza dei Dati (Data Science)?




Wind m/s
0 4 7 10 13 16 20
<< SHOW
NORTH AMERICA
ASIA
AFRICA
SOUTH AMERICA
EUROPE


NETFLIX
Blade Runner 2049
IOS AND ANDROID APP 2016
DVD Netflix
Netflix has never had a dedicated app for its DVD customers. I led the research, design, and launch of the first stand-alone DVD app for iOS and Android.


xG: 0.0336152
outcome:Off T

Source

FUTURE EDUCATION MODENA

ANSAcom

Operazioni chirurgiche a distanza, effettuato test con 5G

Al San Raffaele con Vodafone su modello artificiale di laringe

MILANO 09 ottobre 2019 09:56 ANSACOM

L'ad di Vodafone Italia Aldo Bisio (D) durante la presentazione di 5G Healthcare Day di Vodafone - RIPRODUZIONE RISERVATA

Medicina E Ricerca

HOME ALIMENTAZIONE E FITNESS MEDI(

L'intelligenza artificiale fa la diagnosi come il pediatra

Un algoritmo basato su tecniche di apprendimento au

Il software italiano che ha cambiato il mondo della polizia predittiva

L'algoritmo di KeyCrime è diverso dagli altri software che cercano di prevenire il crimine, sotto accusa per le loro limitazioni. Ce lo racconta il suo ideatore

Source

FUTURE EDUCATION MODENA

ANSAcom

Operazioni chirurgiche a distanza, effettuato test con 5G

Al San Raffaele con Vodafone su modello artificiale di laringe

MILANO 09 ottobre 2019 09:56 ANSACOM

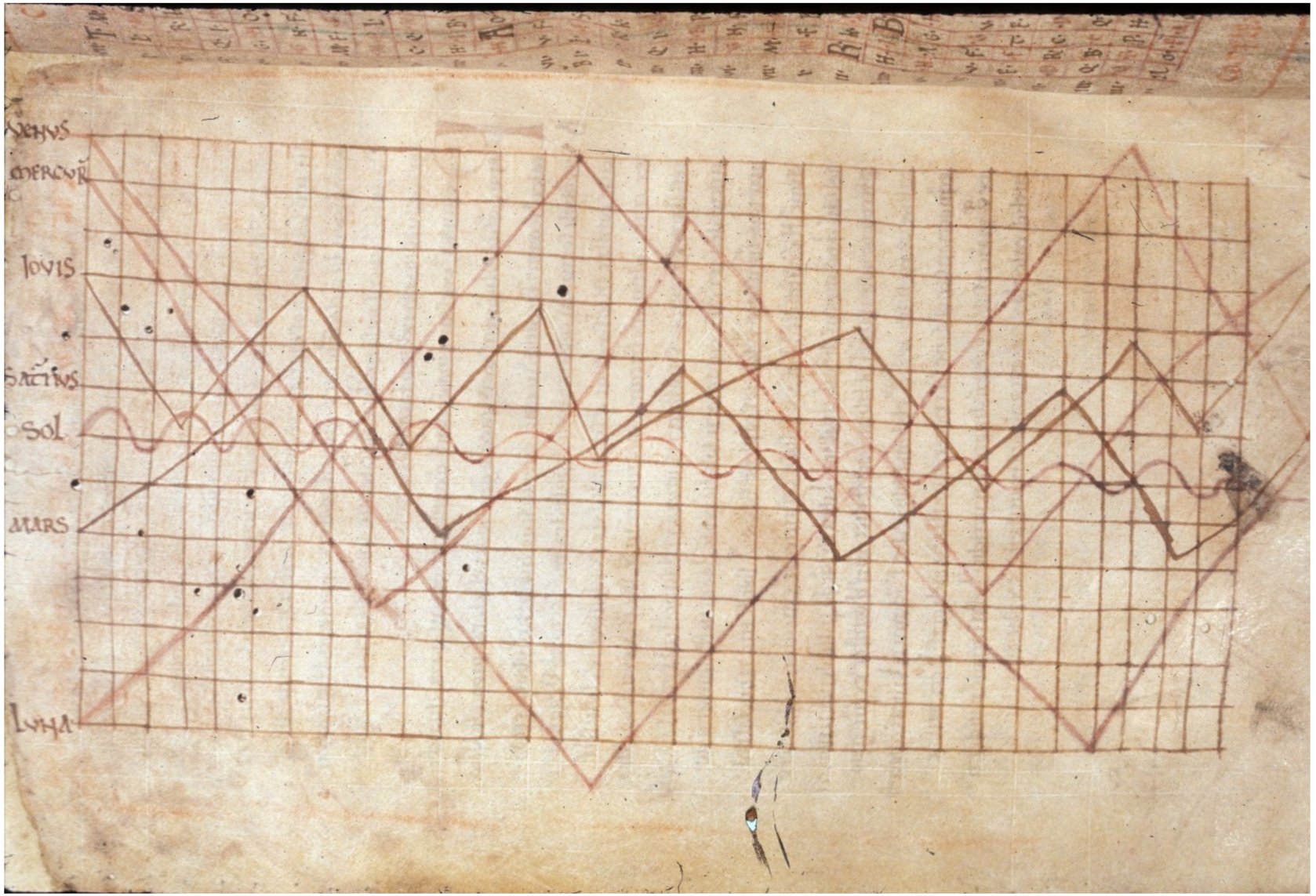
L'ad di Vodafone Italia Aldo Bisio (D) durante la presentazione di 5G Healthcare Day di Vodafone - RIPRODUZIONE RISERVATA

Medicina E Ricerca

HOME ALIMENTAZIONE E FITNESS MEDI(

L'intelligenza artificiale fa la diagnosi come il pediatra

Un algoritmo basato su tecniche di apprendimento au



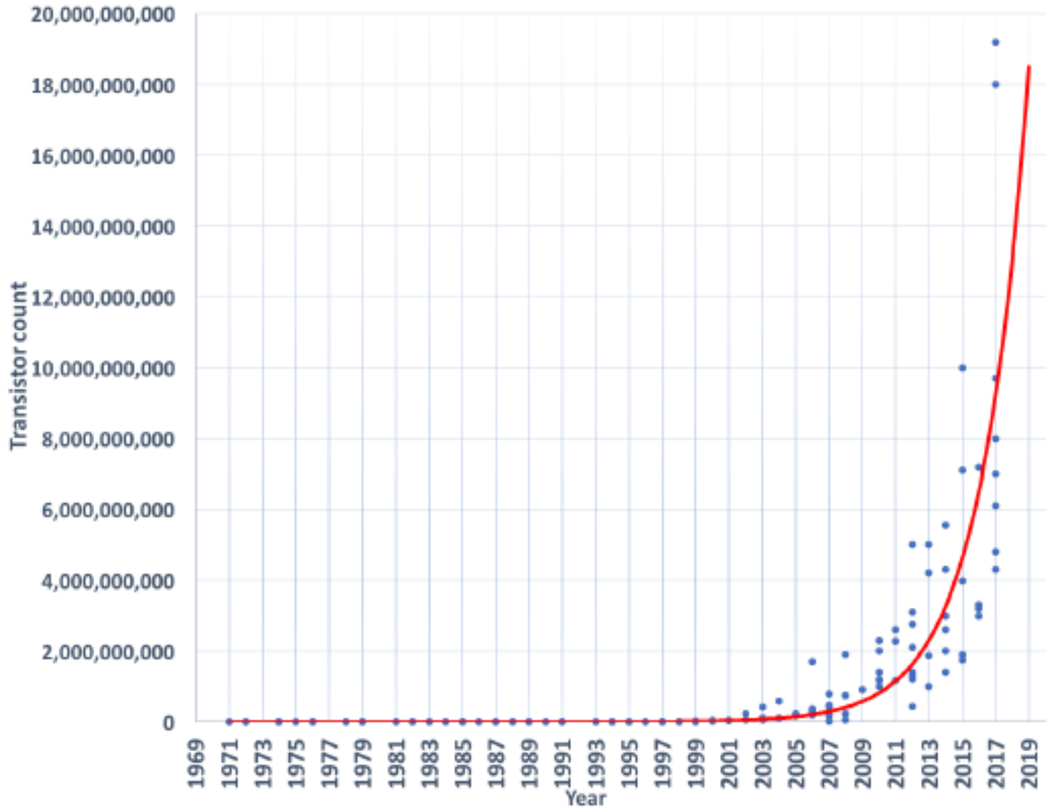
Source

FUTURE EDUCATION MODENA

Perché è importante studiare Data Science?

FUTURE EDUCATION MODENA

È veramente difficile prevedere il futuro



50 Years of Progress in Speech and Speaker Recognition Research

Sadaoki Furui, Non-member

ABSTRACT

Research in automatic speech and speaker recognition has now spanned five decades. This paper surveys the major themes and advances made in the past fifty years of research so as to provide a technological perspective and an appreciation of the fundamental progress that has been accomplished in this important area of speech communication. Although many techniques have been developed, many challenges have yet to be overcome before we can achieve the ultimate goal of creating machines that can communicate naturally with people. Such a machine needs to be able to deliver a satisfactory performance under a broad range of operating conditions. A much greater understanding of the human speech process is required before automatic speech and speaker recognition systems can approach human performance.

Keywords: Speech recognition, Speaker recognition, Statistical modeling, Robust recognition

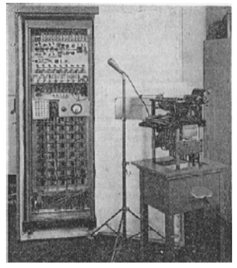


Fig.1: *Phonetic typewriter with parts exposed (H. Olsen and H. Belar, RCA Labs, 1956).*

1. INTRODUCTION

Speech is the primary means of communication between humans. For reasons ranging from technological curiosity about the mechanisms for mechanical realization of human speech capabilities to the desire to automate simple tasks which necessitate human-machine interactions, research in automatic speech and speaker recognition by machines has attracted a great deal of attention for five decades.

Based on major advances in statistical modeling of speech, automatic speech recognition systems today find widespread application in tasks that require human-machine interface, such as automatic call processing in telephone networks and query-based information systems that provide updated travel information, stock price quotations, weather reports, etc.

This paper reviews major highlights during the last five decades in the research and development of automatic speech and speaker recognition so as to provide a technological perspective. Although many technological progresses have been made, there still remain many research issues that need to be tackled.

Manuscript received on January 5, 2006.
The author is with Department of Computer Science, Tokyo Institute of Technology. e-mail: furui@cs.titech.ac.jp

This paper is an extended version of the paper titled "50 years of progress in speech and speaker recognition" which was originally presented at SPECOM conference held at Patras, Greece, in October 2005.

2. SPEECH RECOGNITION

The progress of automatic speech recognition (ASR) technology in the past 50 years can be summarized as follows [63, 33, 24]:

2.1 1950s and 1960s

(1) ***General:*** The earliest attempts to devise ASR systems were made in 1950s and 1960s, when various researchers tried to exploit fundamental ideas of acoustic phonetics. Since signal processing and computer technologies were yet very primitive, most of the speech recognition systems investigated used spectral resonances during the vowel region of each utterance which were extracted from output signals of an analogue filter bank and logic circuits.

(2) ***Early systems:*** In 1952, at Bell Laboratories, Davis, Biddulph, and Balashek built a system for isolated digit recognition for a single speaker [11], using the formant frequencies measured/estimated during vowel regions of each digit. In an independent effort at RCA Laboratories in 1956, Olson and Belar tried to recognize 10 distinct syllables of a single speaker, as embodied in 10 monosyllabic words (Fig. 1) [57]. In 1959, at University College in England, Fry and Denes tried to build a phoneme recognizer to recognize four vowels and nine consonants [17]. By incorporating statistical information concerning allowable phoneme sequences in English, they

FUTURE EDUCATION MODENA

Cosa è un dato

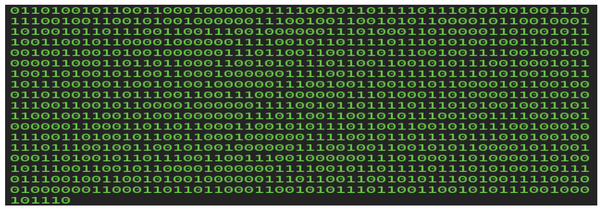
Un **dato** (dal latino *datum* che significa dono, cosa data) è una descrizione elementare, spesso codificata, di un'informazione, un'entità, di un fenomeno, di una transazione, di un avvenimento o di altro.

1 A B C D E F G H

2 1 2 3 4 5 6 7 8 9

3 

4 

5 

A DAY IN DATA

The exponential growth of data is undisputed, but the numbers behind this explosion – fuelled by internet of things and the use of connected devcies – are hard to comprehend, particularly when looked at in the context of one day

500m

tweets are sent every day

Twitter

4PB

of data created by Facebook, including

350m photos

100m hours of video watch time

Facebook Research

294bn

billion emails are sent

Radicati Group

320bn emails to be sent each day by 2021

306bn emails to be sent each day by 2020

3.9bn

people use emails

4TB

of data produced by a connected car

Intel

DEMYSTIFIYING DATA UNITS

From the more familiar 'bit' or 'megabyte', larger units of measurement are more frequently being used to explain the masses of data

Unit		Value	Size
b	bit	0 or 1	1/8 of a byte
B	byte	8 bits	1 byte
KB	kilobyte	1,000 bytes	1,000 bytes
MB	megabyte	$1{,}000^2$ bytes	1,000,000 bytes
GB	gigabyte	$1{,}000^3$ bytes	1,000,000,000 bytes
TB	terabyte	$1{,}000^4$ bytes	1,000,000,000,000 bytes
PB	petabyte	$1{,}000^5$ bytes	1,000,000,000,000,000 bytes
EB	exabyte	$1{,}000^6$ bytes	1,000,000,000,000,000,000 bytes
ZB	zettabyte	$1{,}000^7$ bytes	1,000,000,000,000,000,000,000 bytes
YB	yottabyte	$1{,}000^8$ bytes	1,000,000,000,000,000,000,000,000 bytes

*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.

463EB

of data will be created every day by 2025

IDC

95m

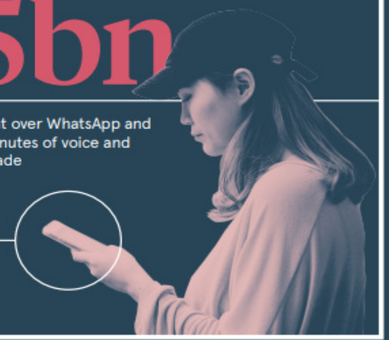
photos and videos are shared on Instagram

Instagram Business

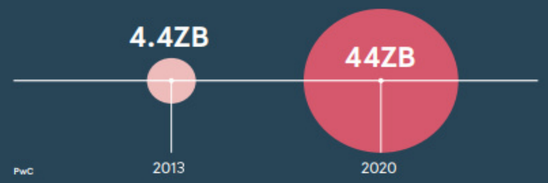
65bn

messages sent over WhatsApp and two billion minutes of voice and video calls made

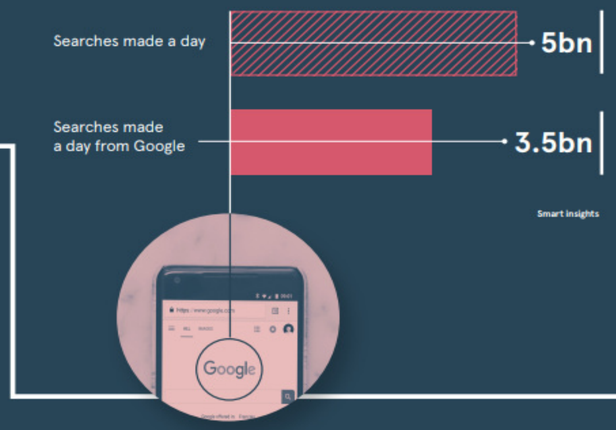
Facebook



ACCUMULATED DIGITAL UNIVERSE OF DATA



PwC

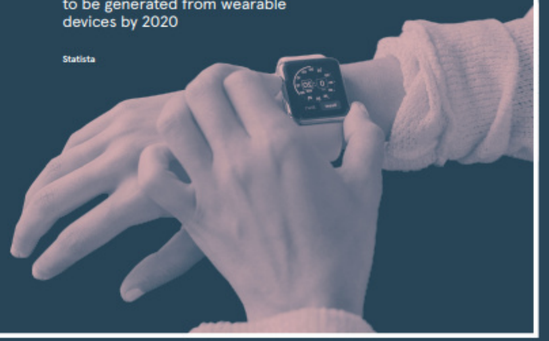


Smart Insights

28PB

to be generated from wearable devices by 2020

Statista



RACONTEUR

Cos’è la Scienza dei Dati (Data Science)

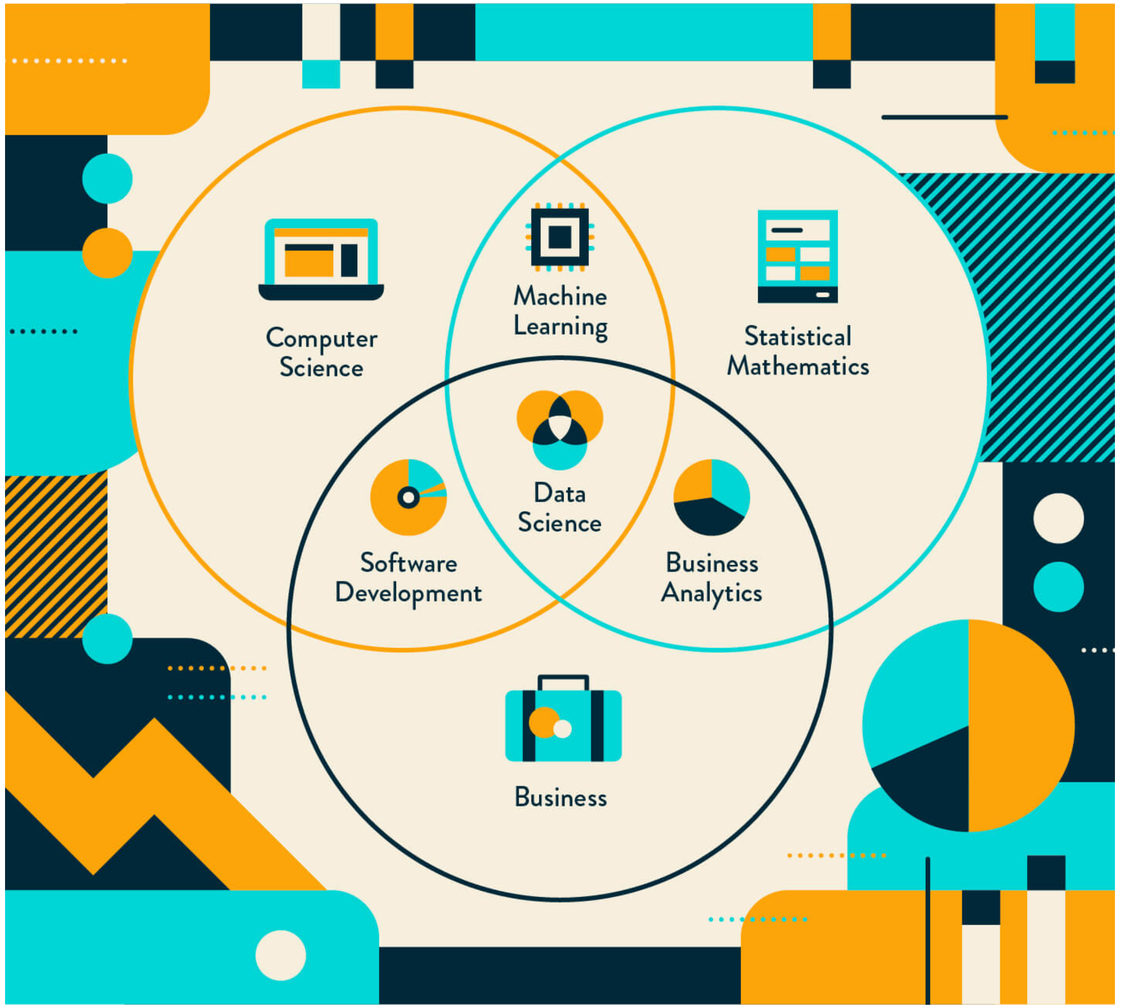
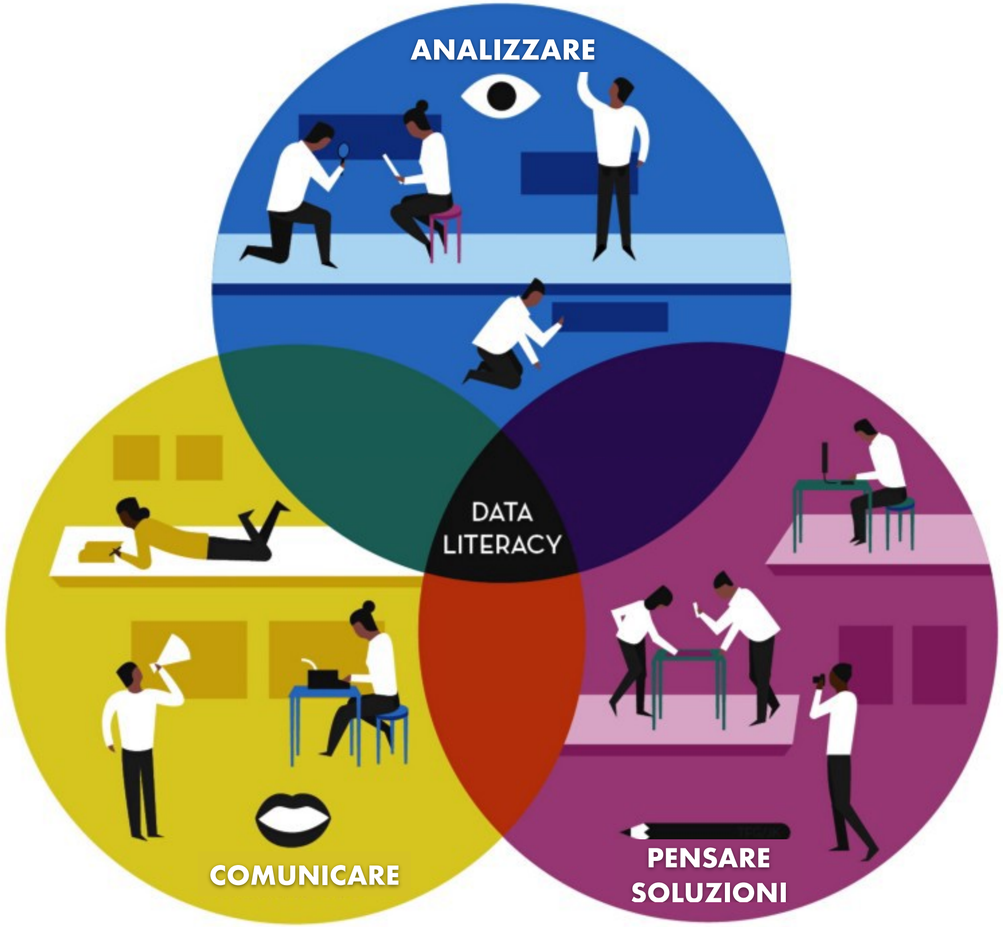


Image source

FUTURE
EDUCATION
MODENA

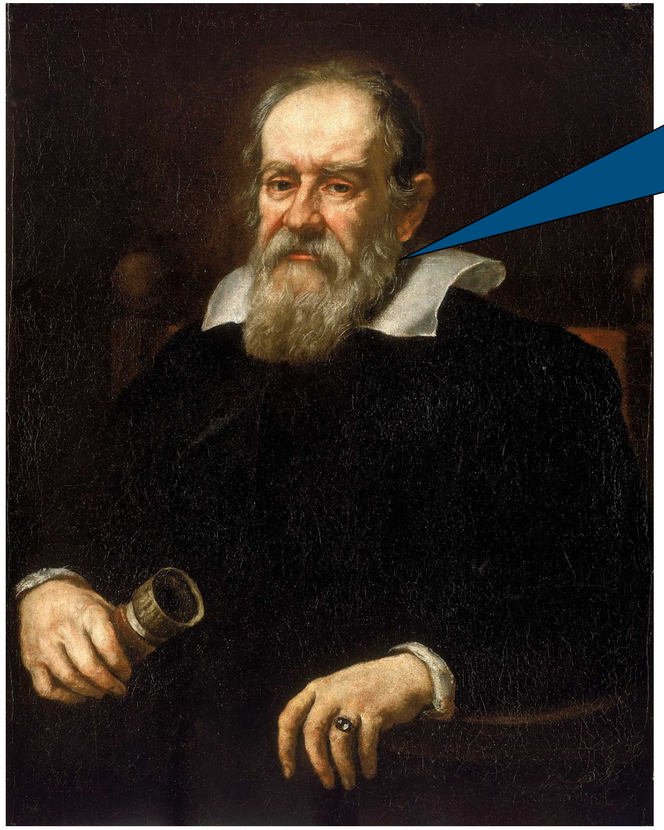
Cosa faremo in questi workshop

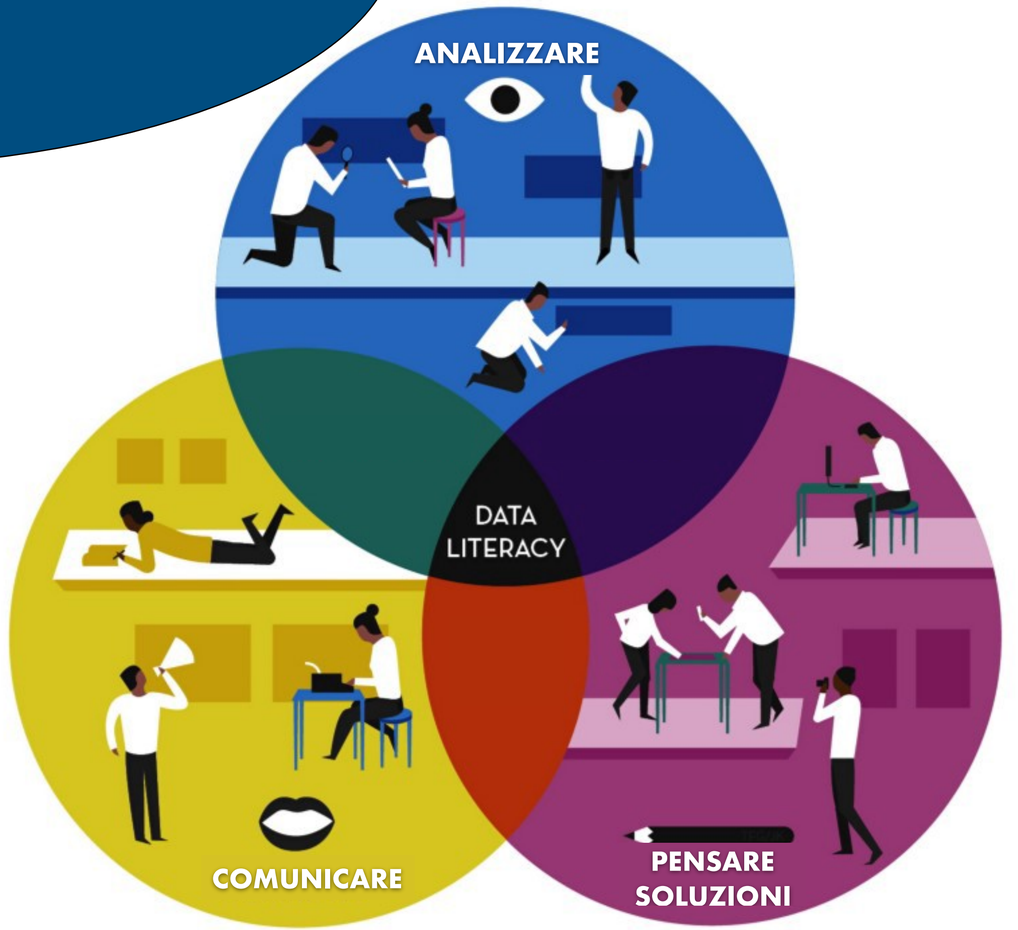


Graphic by Jehoaddan Kulakoff.

FUTURE EDUCATION MODENA

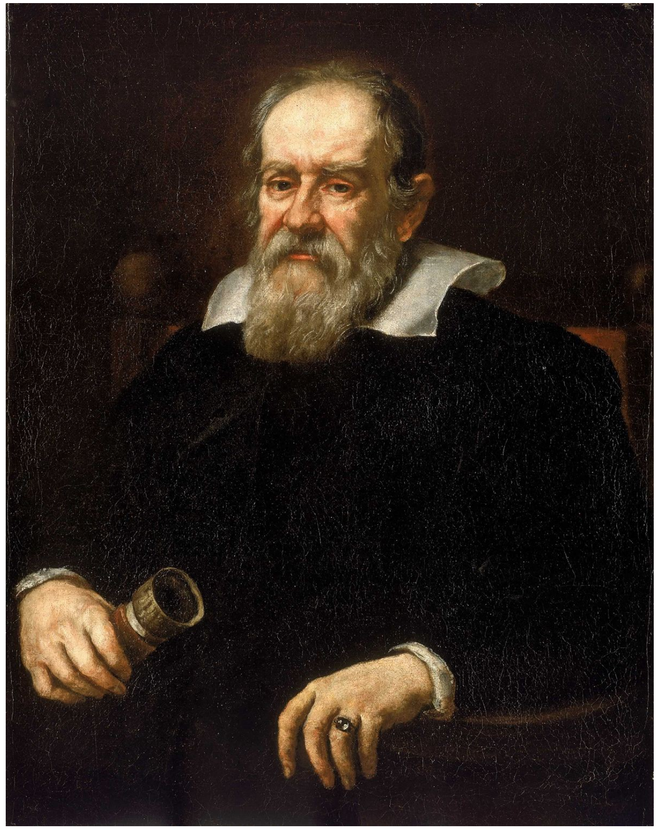
State parlando di qualcosa molto simile a quello che ho detto io





Il metodo scientifico

Un ipotesi scientifica sta alla base del metodo scientifico (e di molto altro)



Galileo Galilei

1. Le basi di un'ipotesi

Perché un'ipotesi diventi un'ipotesi scientifica, dev'essere **comprovata o smentita da sperimentazioni o da osservazioni meticolose** *(es. se lo zucchero causa la carie, allora le persone che mangiano tante caramelle sono più a rischio di avere carie.)*

2. Testare un'ipotesi

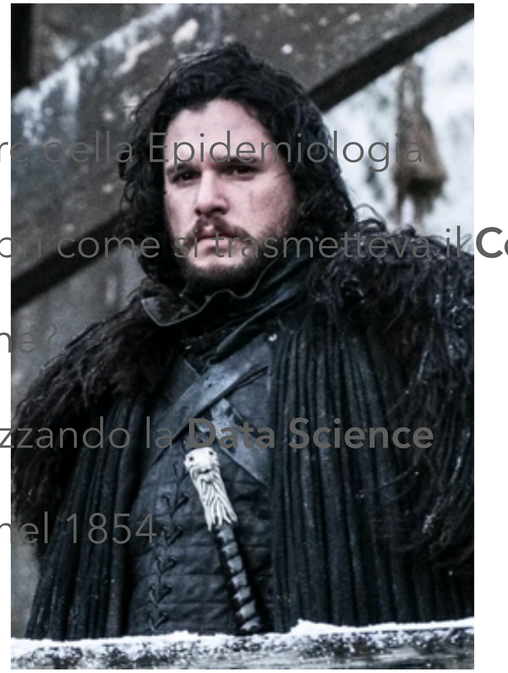
È importante notare che tutte le ipotesi devono **essere verificabili.** Il tratto primario di un'ipotesi è che può essere testata, e i test ripetuti.

3. Da ipotesi a teoria

Sebbene teorie e ipotesi vengano molto spesso confuse tra loro, **le teorie sono i risultati di ipotesi comprovate**. Le ipotesi sono delle idee, le teorie invece spiegano i risultati ottenuti testando queste idee.

Data Scientist Challenge:
Fermare la diffusione del colera

John Snow



- Padre della Epidemiologia
- Scoprì come si trasmetteva il **Colera**
- Come?
- Utilizzando la **Data Science**
- Ma nel 1854

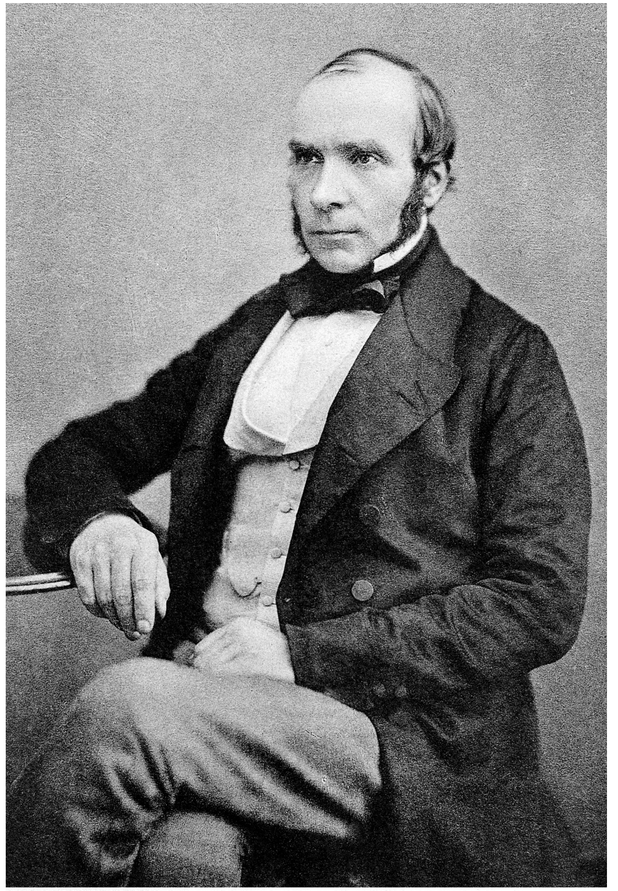


Image source

Londra 1850: non un posto pulito

A COURT FOR KING CHOLERA.

Image source

John Snow’s cholera map

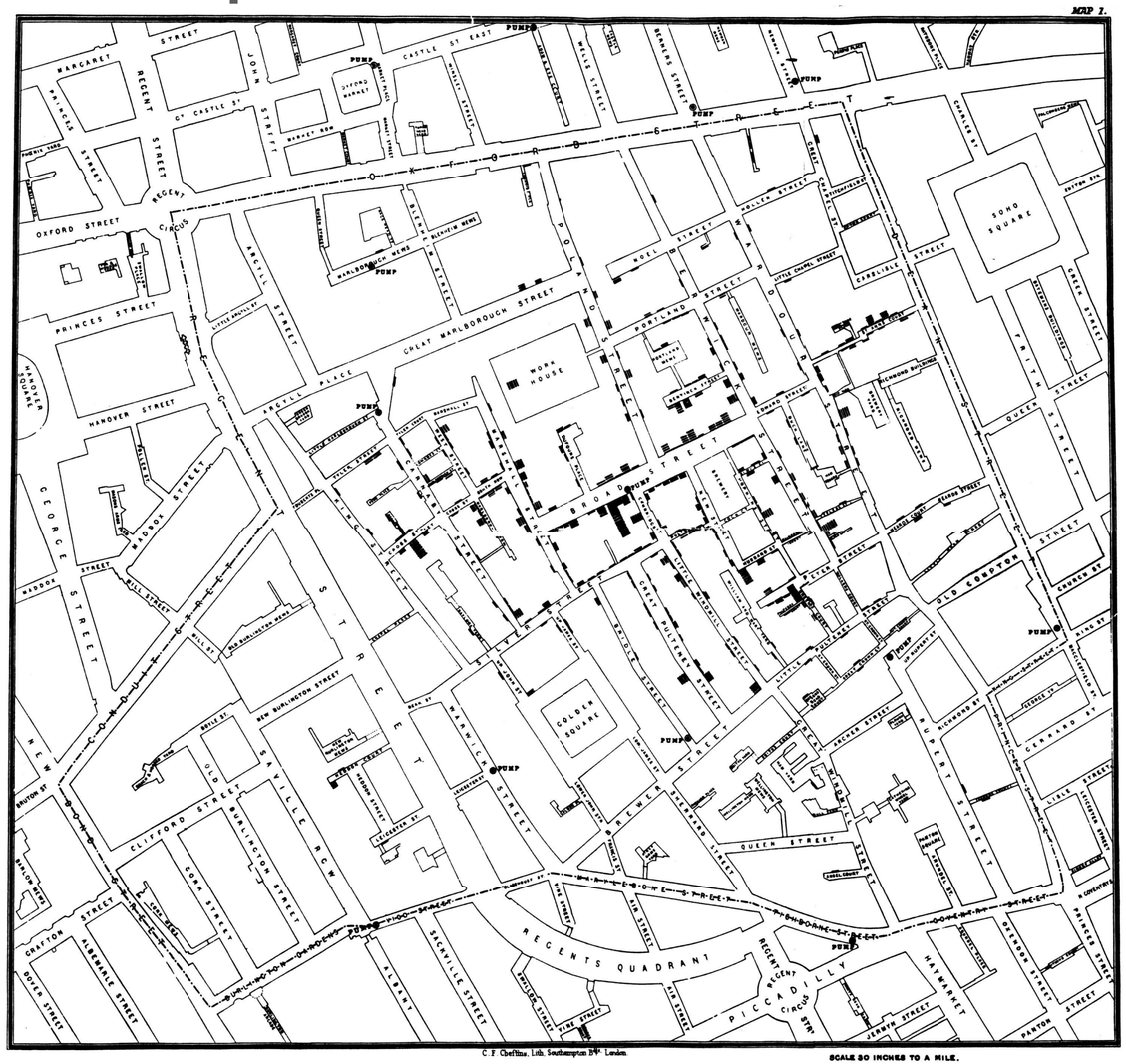
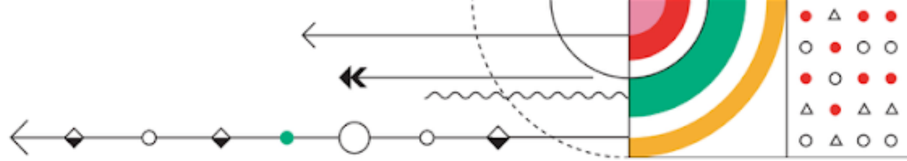


Image source

FUTURE EDUCATION MODENA



La sfida

Voi siete lo scienziato John Snow e dovete capire come si diffonde il **Colera**

Indicazioni:

1. Non era chiaro come si diffondesse, l'idea più diffusa era che si diffondesse per via aerea (Dottrina miasmatico-umorale)
2. Non esistevano gli acquedotti ne le fognature, tutto finiva per strada

Il vostro compito:

1. Formare una ipotesi investigativa su come si diffonda il colera
2. Osservare le prove presenti sulla mappa, quali indizi possono essere i più utili?

FUTURE EDUCATION MODENA

Tutto più chiaro

Finding the water pump with Cholera contamination

This chart takes a closer look at the area of cholera outbreak that John Snow famously mapped at 1854 in London. How many people died of Cholera in every 5 meters around each water pump? Pump No.1 (you can find its location on the map) has the densest and highest number of deaths within the reach of 100 meters.

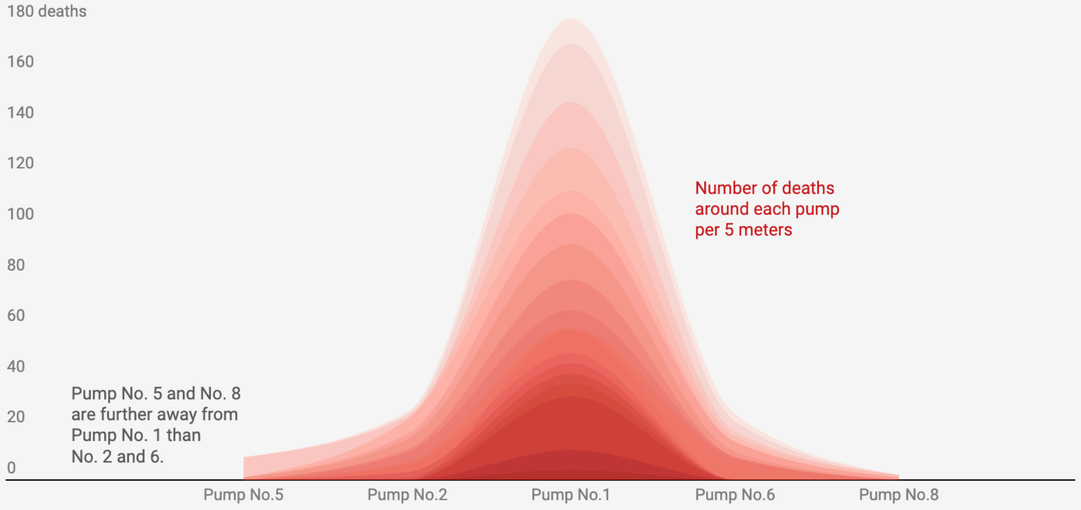


Chart: Defne Altiok • Source: Cholera Pumps and Deaths, Jon Snow (1854), digitised by Robin Wilson • Get the data

Articolo originale qui

Altra mappa

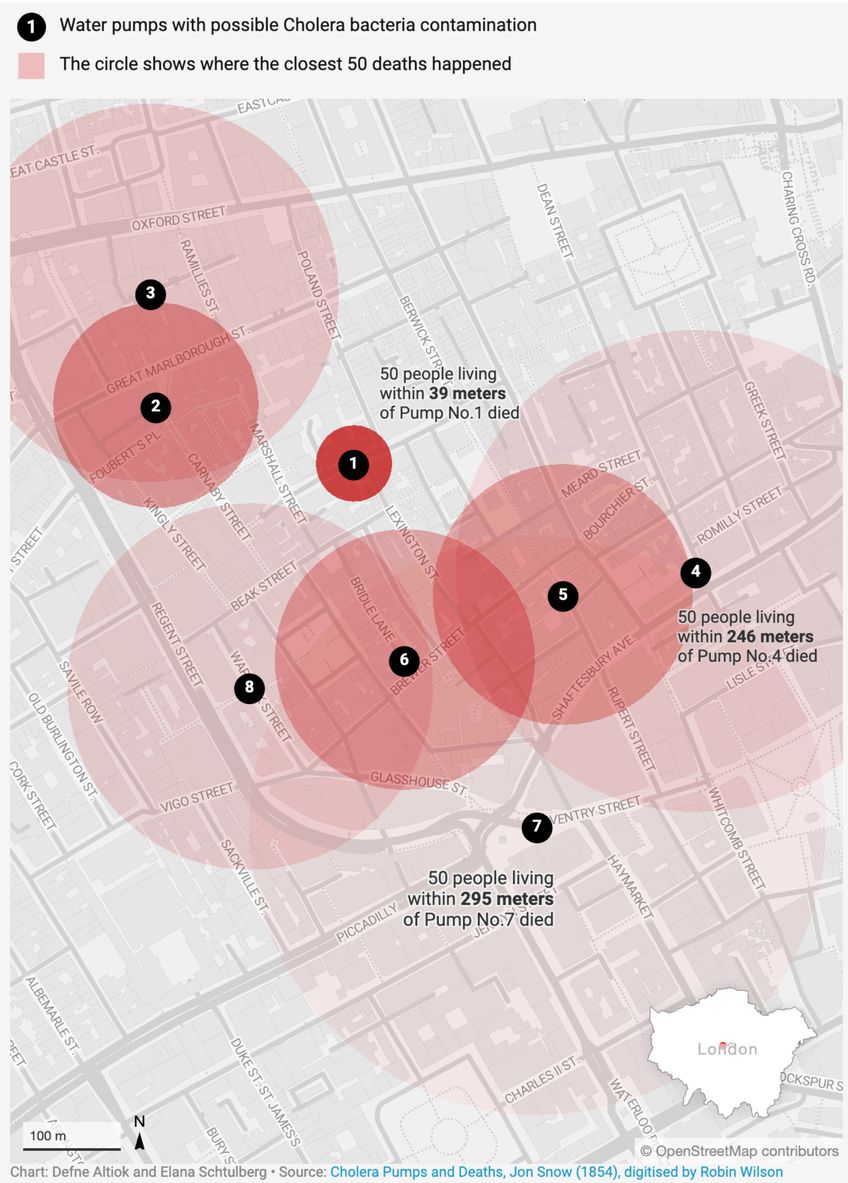
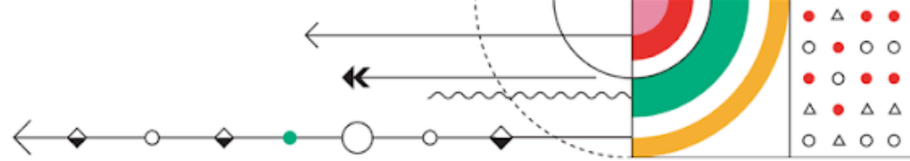


Chart: Defne Altiok and Elana Schtulberg • Source: Cholera Pumps and Deaths, Jon Snow (1854), digitised by Robin Wilson

Il pensiero computazionale

FUTURE EDUCATION MODENA



Pensiero Computazionale

identificare, analizzare, implementare
e verificare le possibili soluzioni ad un problema

Scomporre

Scomporre un problema e riconoscerne gli schemi e gli elementi principali

Astrarre

Astrarre il problema in un formato che ci permette di usare un esecutore (computer, essere umano, macchina) per risolverlo

Automatizzare

Programmare/automatizzare la risoluzione del problema consistente in una sequenza accuratamente descritta dei passaggi da compiere.

Aprite Github

Grazie

Prima Lezione Data Science for Innovation 2020

Lorenzo Andreoli
Data Scientist @FEM

lorenzo@fem.digital

www.fem.digital