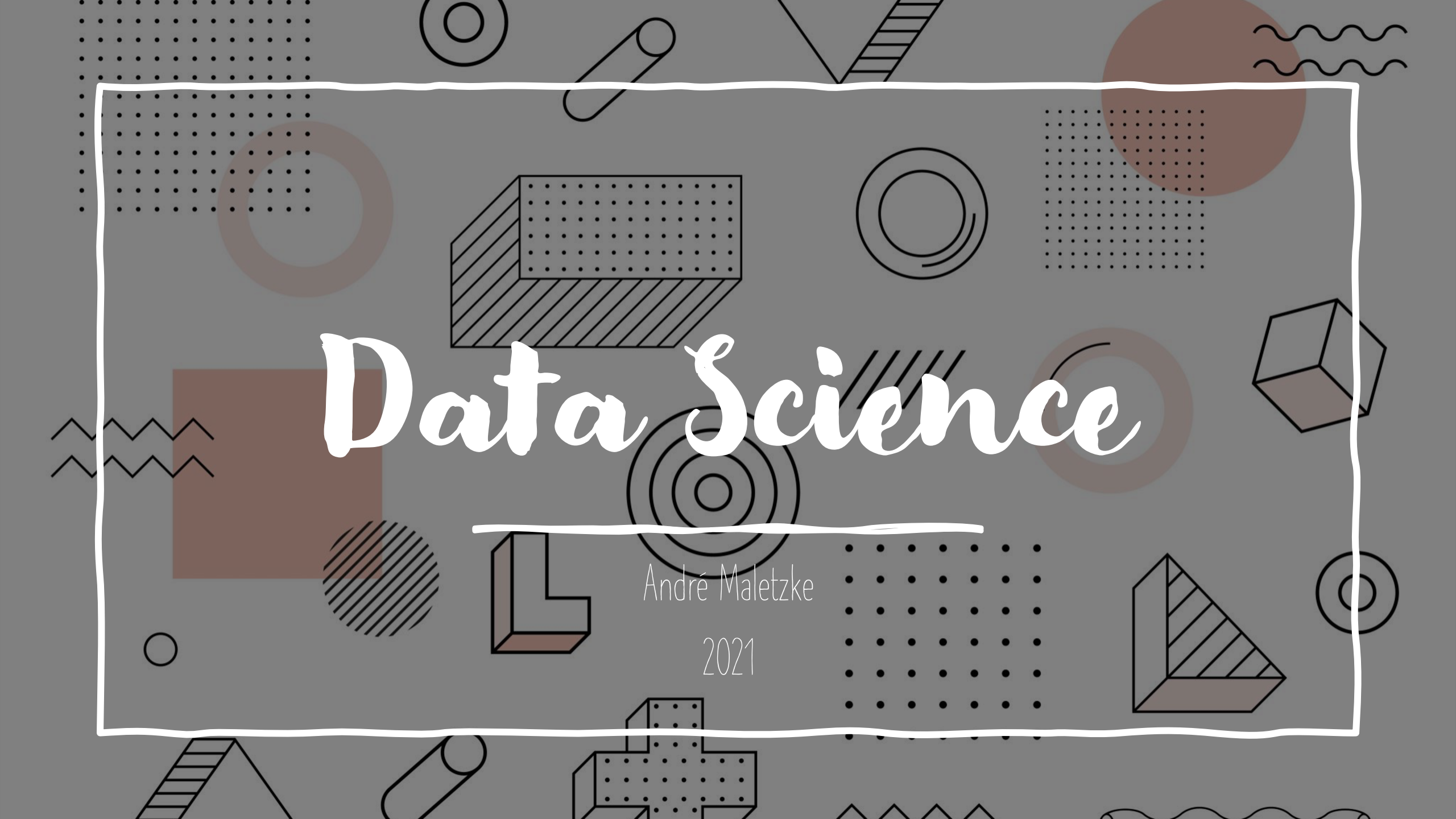
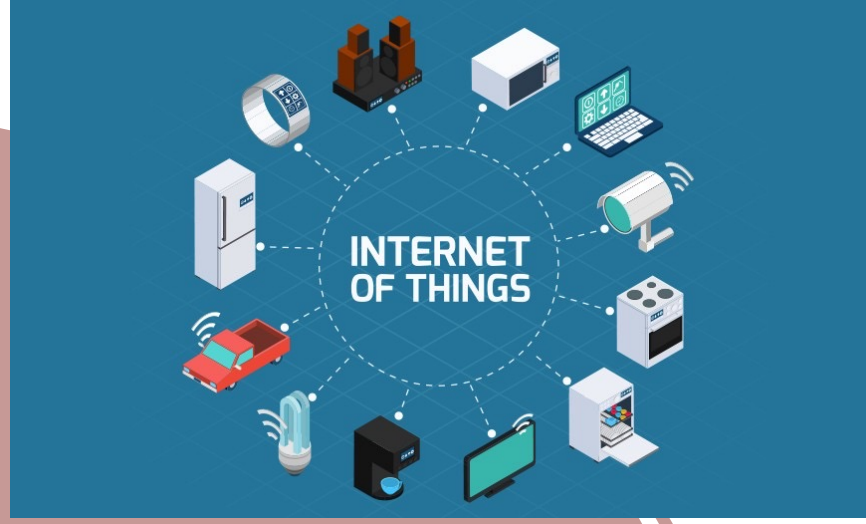
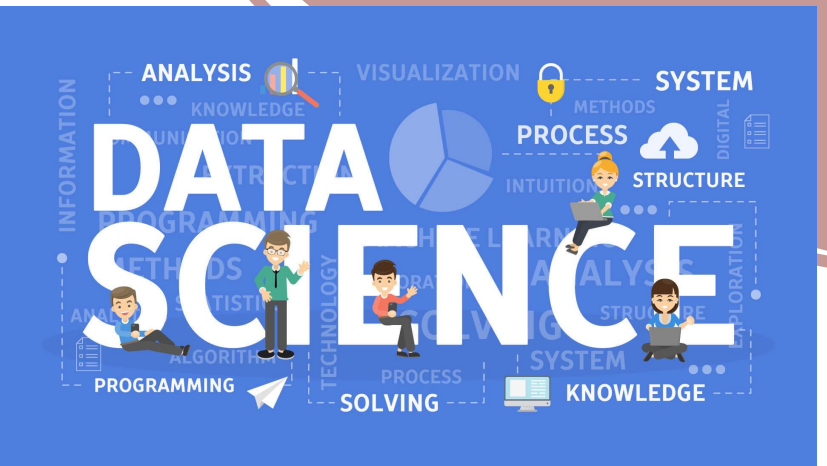

Data Science
André Maletzke
2021


# Data + Science




# Data

- 2.5 quintillion bytes of data (every day)

  2,500,000,000,000,000,000

- This amount of data created will continue to grow
- 3.7 billion humans use the internet every day (src: Forbes.com)
- Google processes more than 40,000 searches every second
  - 3.5 billion searches each day

Source: https://www.corporatecomplianceinsights.com/data-data-data-everywhere/

Last update: 2018.

# Data is the new oil!

*"A NEW commodity spawns a lucrative, fast-growing industry, prompting antitrust regulators to step in to restrain those who control its flow. A century ago, the resource in question was oil. Now similar concerns are being raised by the giants that deal in data, the oil of the digital era."*

Source: theeconomist.com

The Economist

Menu | Weekly edition | Search

**Leaders**

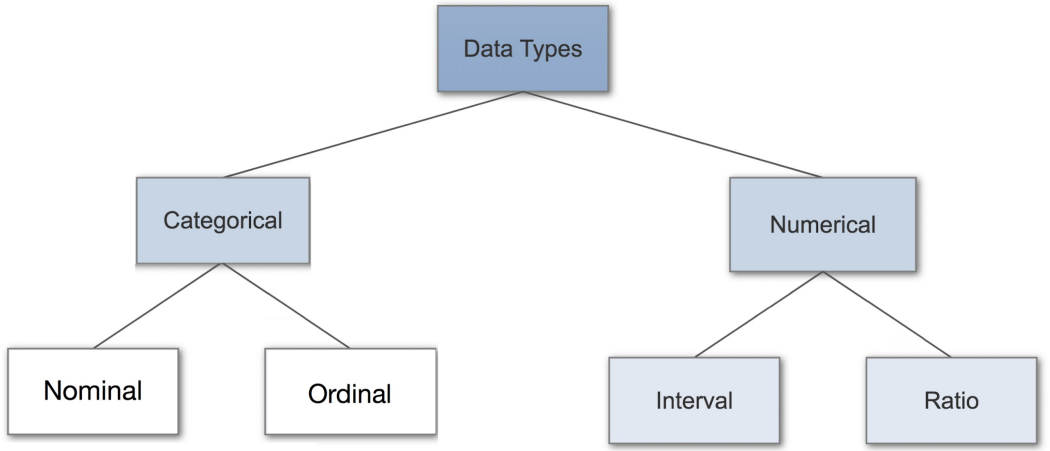
May 6th 2017 edition ›

Regulating the internet giants

## The world's most valuable resource is no longer oil, but data

The data economy demands a new approach to antitrust rules

# What is data?



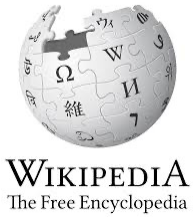

WIKIPEDIA
The Free Encyclopedia

**Data** are characteristics or information, usually numerical, that are collected through observation.[1] In a more technical sense, data are a set of values of qualitative or quantitative variables about one or more persons or objects

Although the terms "data" and "information" are often used interchangeably, these terms have distinct meanings. In some popular publications, data are sometimes said to be transformed into information when they are viewed in context or in post-analysis.

# Sciencia

**Science** (from the Latin word *scientia*, meaning "knowledge")[1] is a systematic enterprise that builds and organizes knowledge in the form of testable explanations and predictions about the universe.




# Scientific method

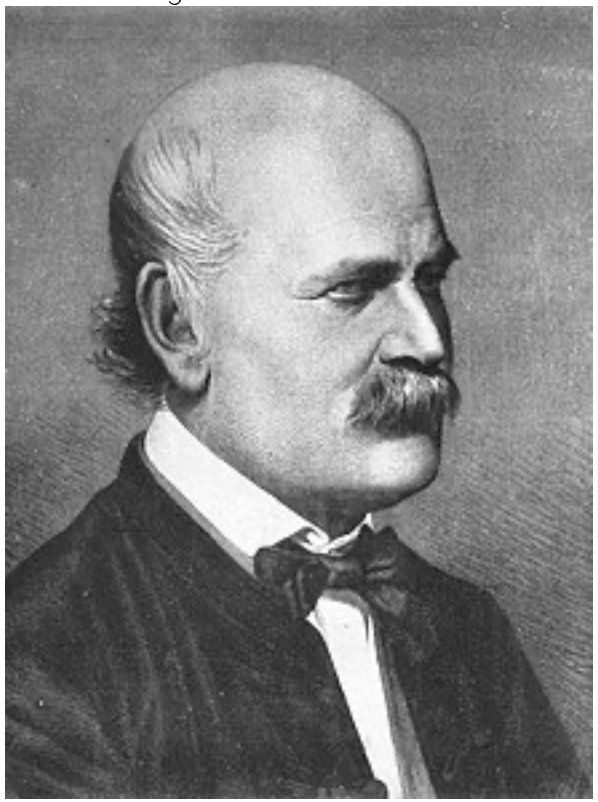
*Em sentido estrito, ciência refere-se ao sistema de adquirir conhecimento baseado no método científico bem como ao corpo organizado de conhecimento conseguido através de tais pesquisas.*

# Scientific method

Scientific method is the base of science!

# Data driven decisions... Historical notes

Dr. Ignaz Semmelweis



Source: en.wikipedia.org

Puerperal fever was common in mid-19th-century hospitals and often fatal

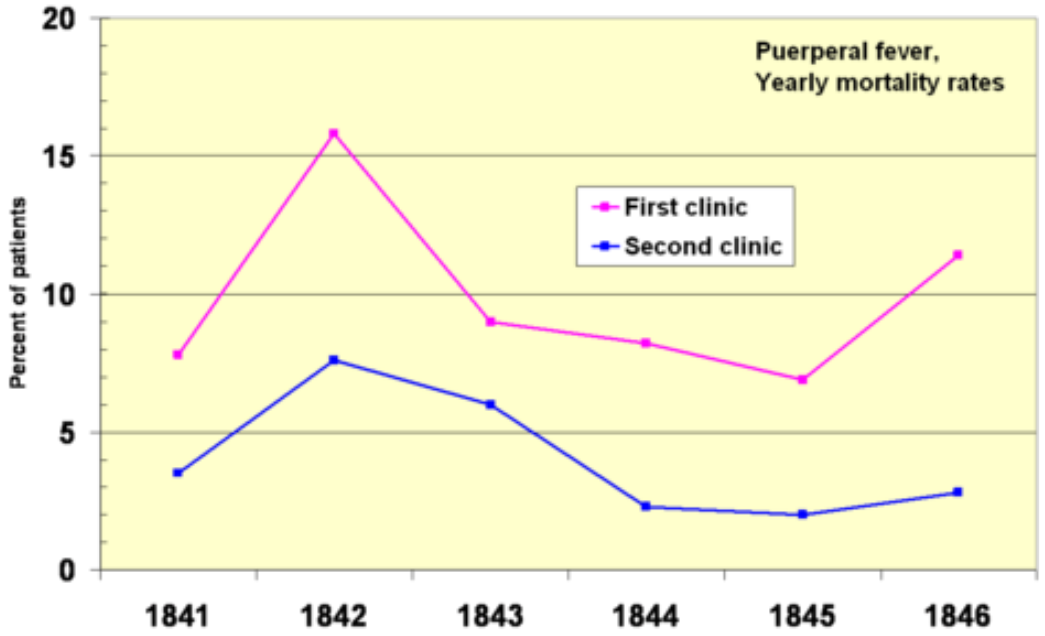
1846 – He was appointed assistant to Professor Johann Klein (First Obstetrical Clinic of the Vienna General Hospital )

One of his duties was check the medical **records**

**Observation:**

The First Clinic had an average maternal mortality rate of about **10%** due to puerperal fever. The Second Clinic's rate was considerably lower, averaging less than **4%**.

# Data driven decisions... Historical notes




## Why?

The two clinics used almost the same techniques

He excluded "overcrowding" as a cause (Second Clinic was always more crowded and yet the mortality was lower)

He eliminated climate as a cause because the climate was the same.

The only major difference was the individuals who worked there:

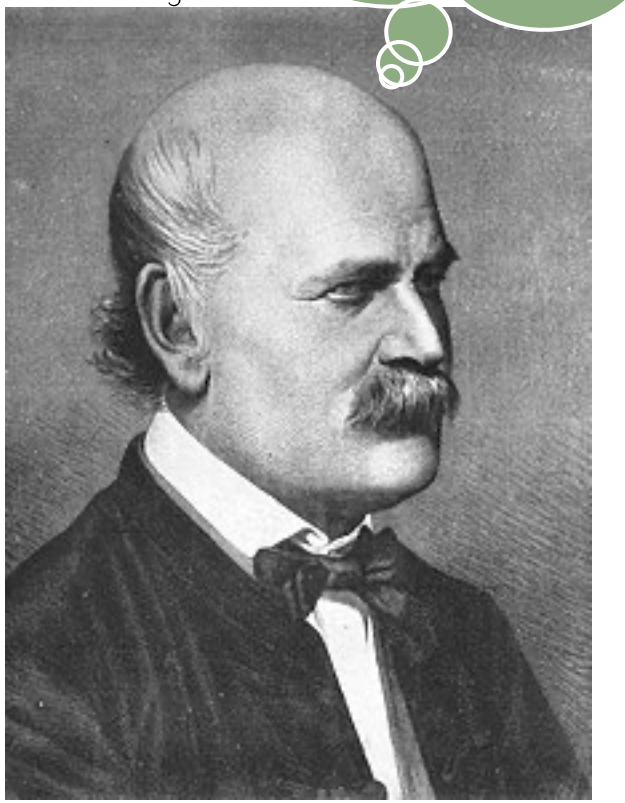
(1) The First Clinic was the teaching service for medical students

(2) The Second Clinic had been selected in 1841 for the instruction of midwives only

# Dat[illegible]decisions... Historical notes

**Theory of cadaverous poisoning**

Dr. Ignaz Sem[illegible]



Source: en.wikipedia.org

1847 - The death of Semmelweis' good friend (Jakob Kolletschka) - who had been accidentally poked with a student's scalpel while performing a *post mortem examination*

Semmelweis immediately proposed a connection between cadaveric contamination and puerperal fever

***Dr. Semmelweis proposed that the medical students carried "cadaverous particles" on their hands***

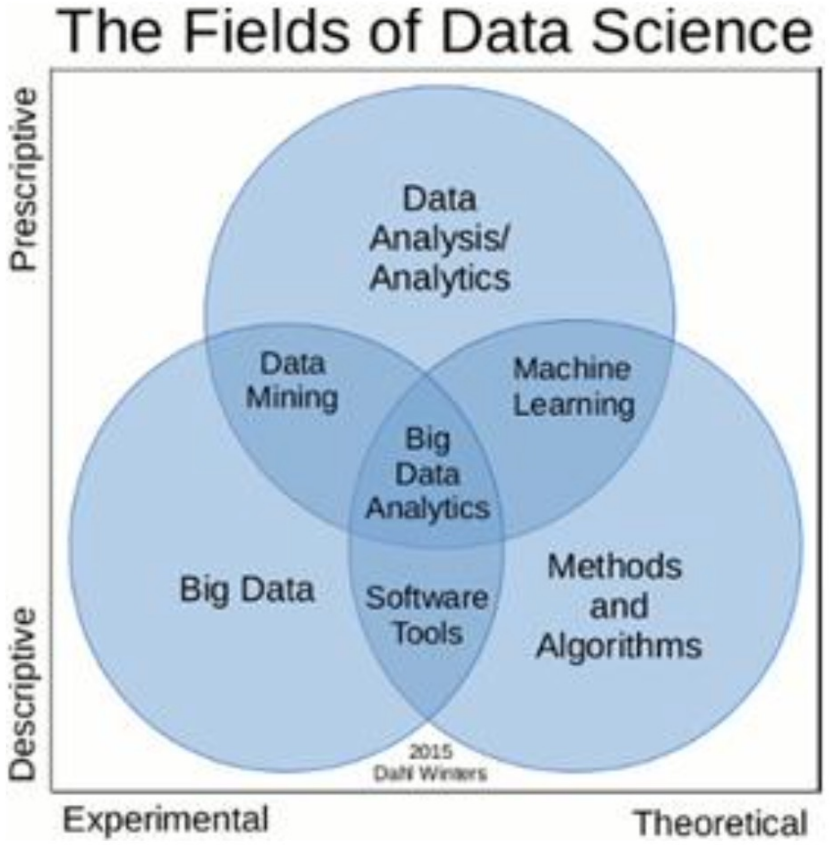
He instituted a policy of using a solution of chlorinated lime for washing hands between autopsy work and the examination of patients.

# Data driven decisions... Historical notes


Source: en.wikipedia.org

The result was the mortality rate in the First Clinic declined 90%

# What is Data Science?




Source: Class notes of Prof. Andre C. P. L. F. de Carvalho (2016)

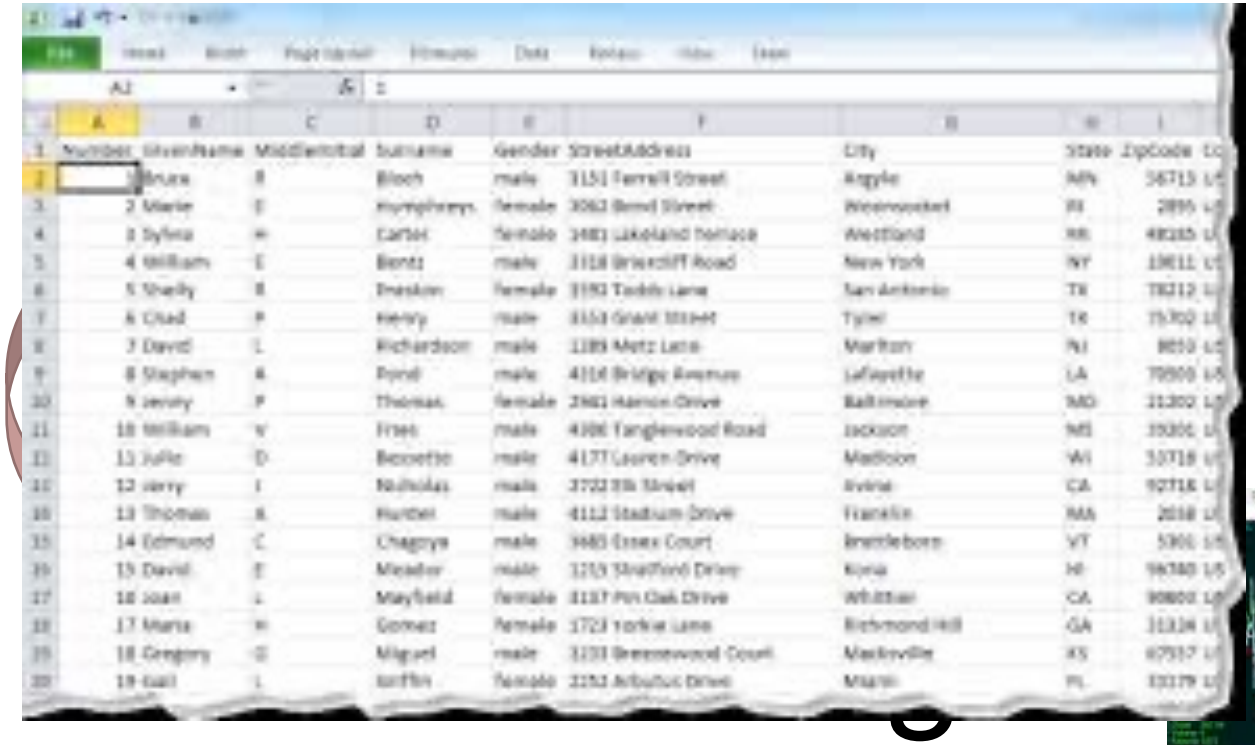
The purpose of computing is insight, not numbers.

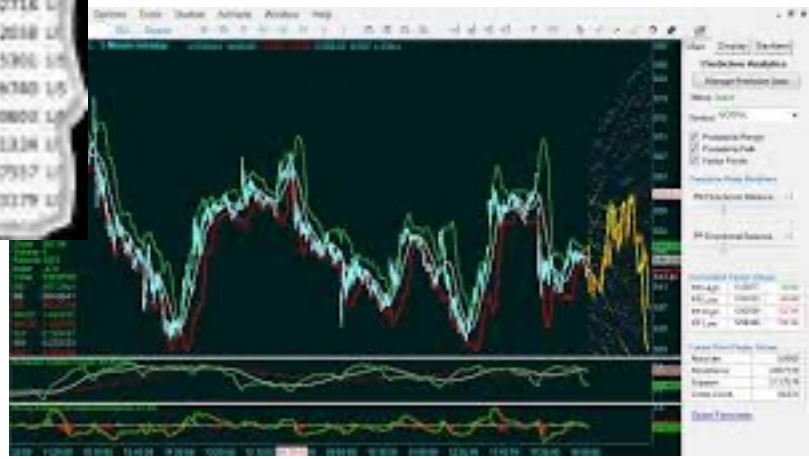
– Richard W. Hamming

*Like any emerging field, it hasn't been completely defined yet …*

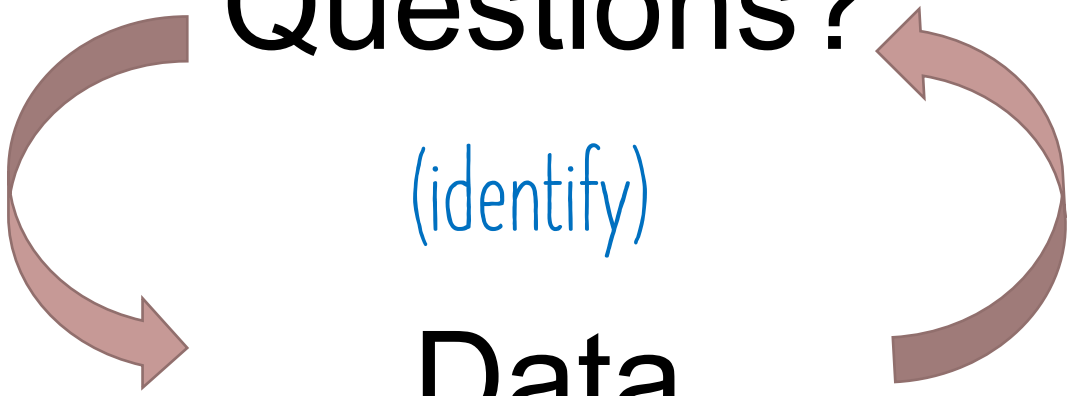
*Steven S. Skiena*

# Data Science

# Data Science




- Missing values
- Inconsistencies
- Redundancies
- Data from different sources
- Different formats

Real world data challenges!

*"…people usually spend about 60–70% of their time just on gathering and cleaning the data."*

Vanshika Goel
Source: medium.com

# Data Science

# Questions about the problem

- Does the past represent the future?
- What do I want to model?
- How will the model be used?
- What data do I need? Or What data do I have?
- How hard to get the data?

# Data Science problems

## Classification

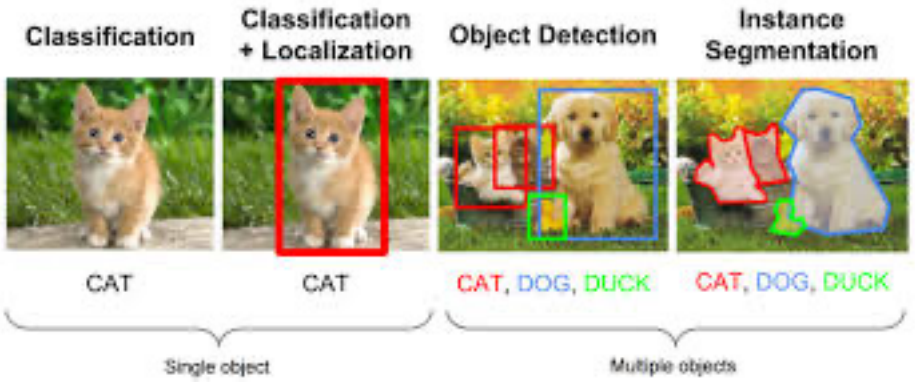



Image classification

## Regression


Stock price prediction

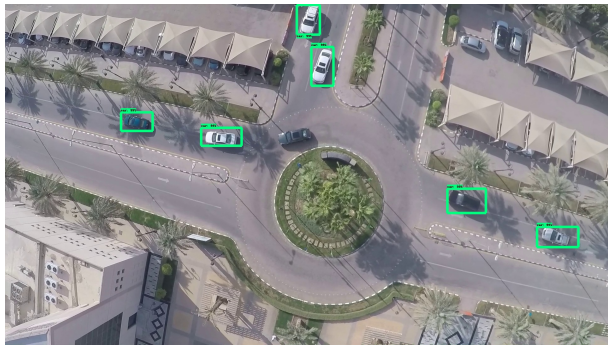
# Data Science problems

## Clustering or ranking


How (and when) do we best present that movie recommendation to the user in a way that maximizes viewership and monthly subscriber loyalty?

Source: https://becominghuman.ai/how-netflix-uses-ai-and-machine-learning-a087614630fe

# Data Science problems

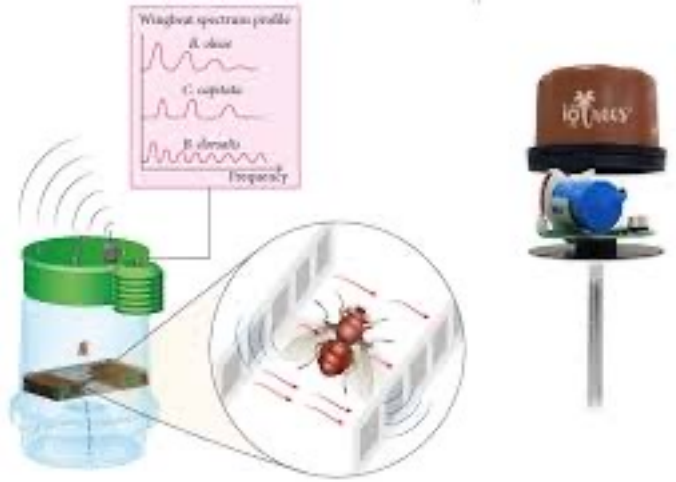


Counting cars from aerial images.
Source: Amar et al., (2019)

**Counting**

**Regression**
**Classification**
**Clustering**
**Counting**


Precision agriculture technologies



Insect pests

**Classification**
**Counting**

# Data Science problems

## Counting/Quantification

Positive ratings


Positive ratings

Counting

consumers' opinion (Gao and Sebastiani, 2015)

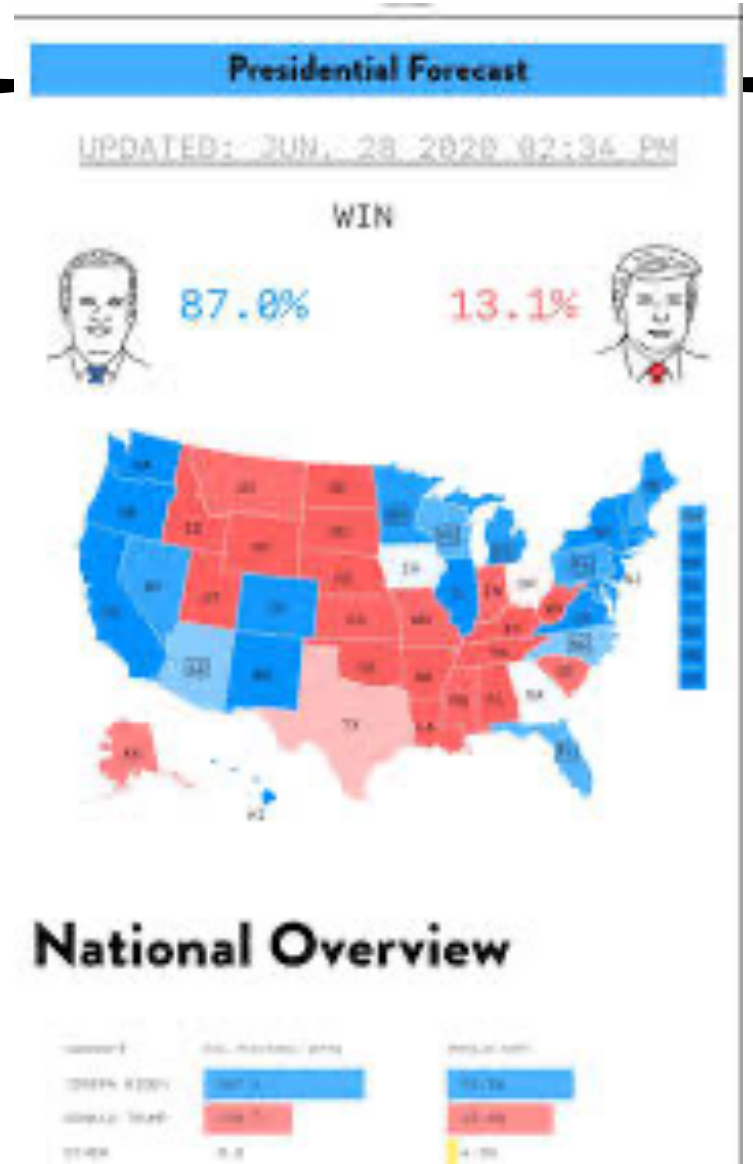
## Where does the data come from?


## How does the data look like?

- Rating
- Reviews
- Others

# Data Science problems

Estimate the prevalence of support for different political candidates (Hopkins and King, 2010)



# Data Science problems




## Credit score prediction

Who will get the credit card?

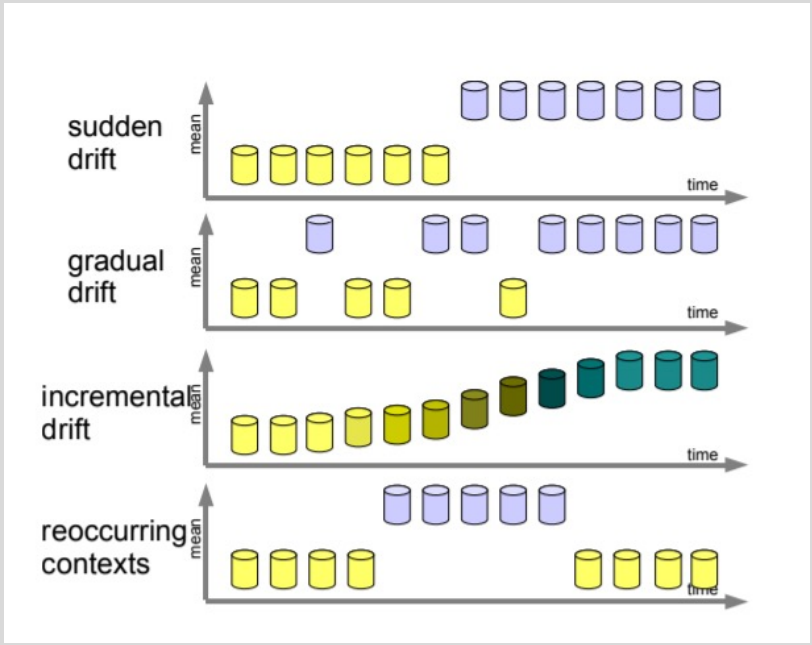
Who will get a bank loan?

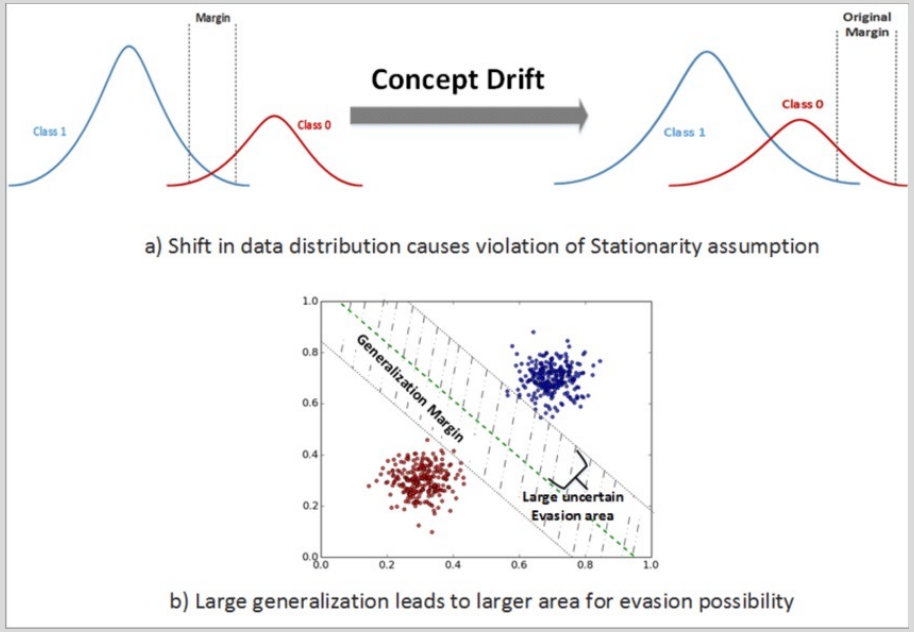
Etc.

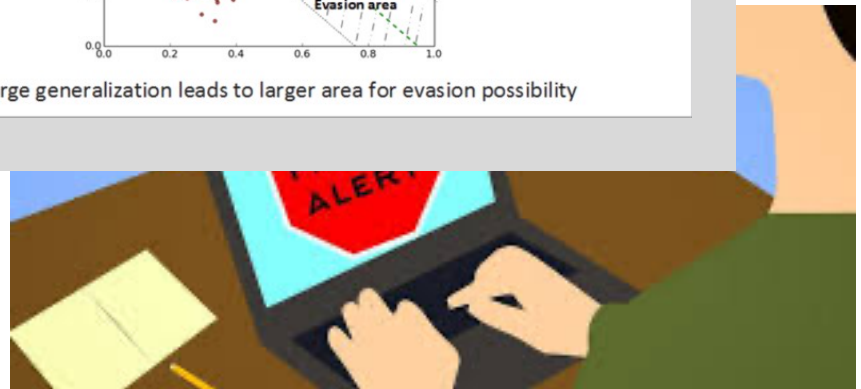
# Data Science problems

MODEL → PREDICTION → EXPLANATION → HUMAN

Sec




# It's everywhere!

# How to ask the right questions to get right answers

# Startups based on Data Science

## BigML

Headquartered in Oregon, BigML was founded in 2011 with the goal of designing and building a platform that would make machine learning accessible to everyone.

## DataRobot

DataRobot is a Boston-based business that has created AI technology and return-on-investment enablement services for businesses competing in what it describes as today's "intelligence revolution".

## Skin Analytics

Skin Analytics is a UK start-up that was founded by Neil Daly in 2012. The start-up has developed and tested an AI platform that aims to detect skin cancer at an early stage, to help those who need medical help get it sooner while giving peace of mind to those who don't need medical assistance.

## Dataiku

Dataiku is a New York-headquartered start-up that was founded by Clément Stenac, Florian Douetteau, Marc Batty and Thomas Cabrol in 2013. It has built a centralised data platform that supports businesses working with analytics at scale and enterprise AI.


FADCIL

# Data Science

# Data Science

Birdie is an **innovative startup** that helps companies understand millions of consumers' opinions

crunchbase news

birdie

# Data Science COVID-19

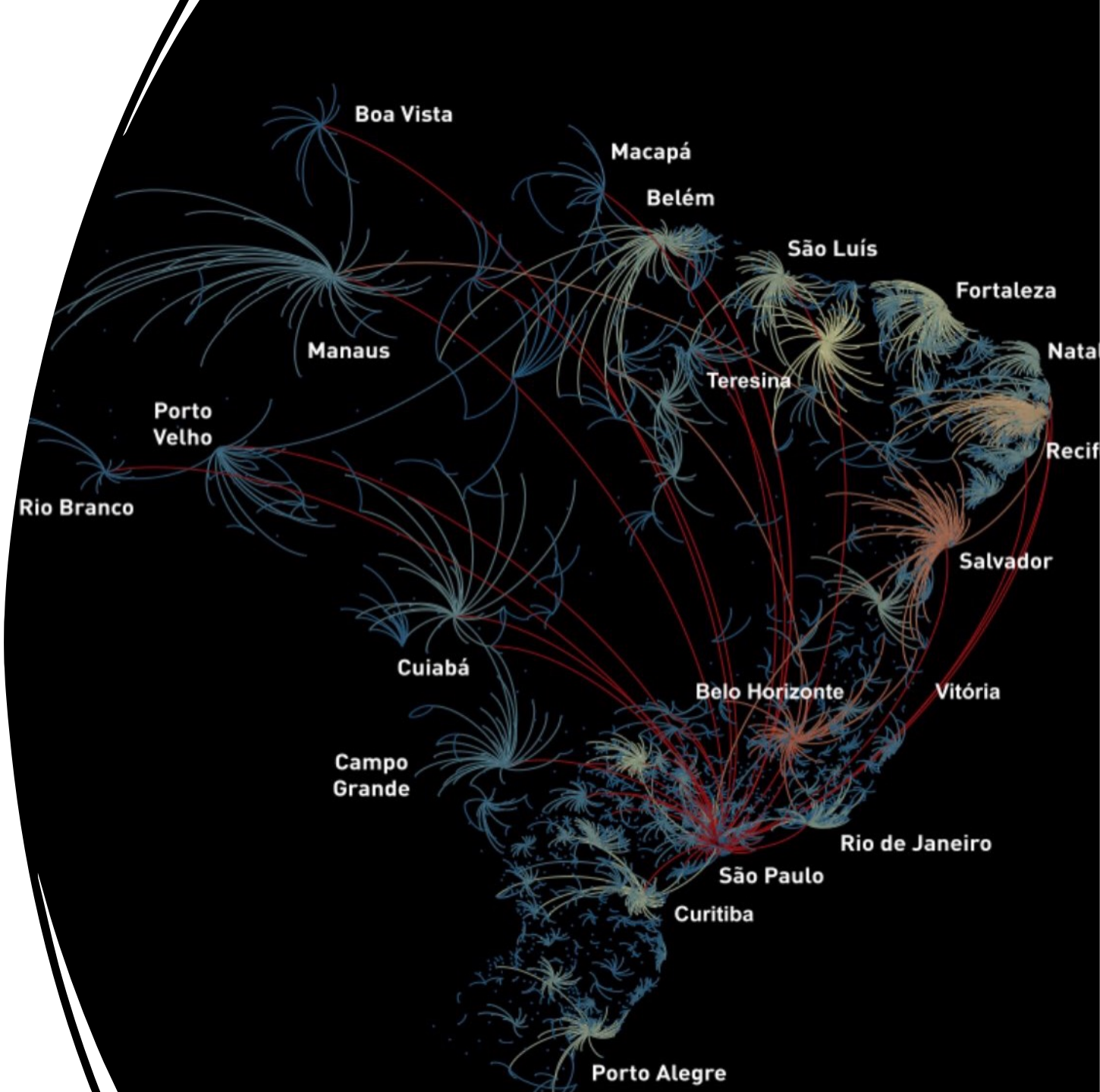
scientific reports

Check for updates

OPEN

## The impact of super-spreader cities, highways, and intensive care availability in the early stages of the COVID-19 epidemic in Brazil

Miguel A. L. Nicolelis[1,2,3,4,5,6]✉, Rafael L. G. Raimundo[7], Pedro S. Peixoto[8] & Cecilia S. Andreazzi[9]



BrainCare

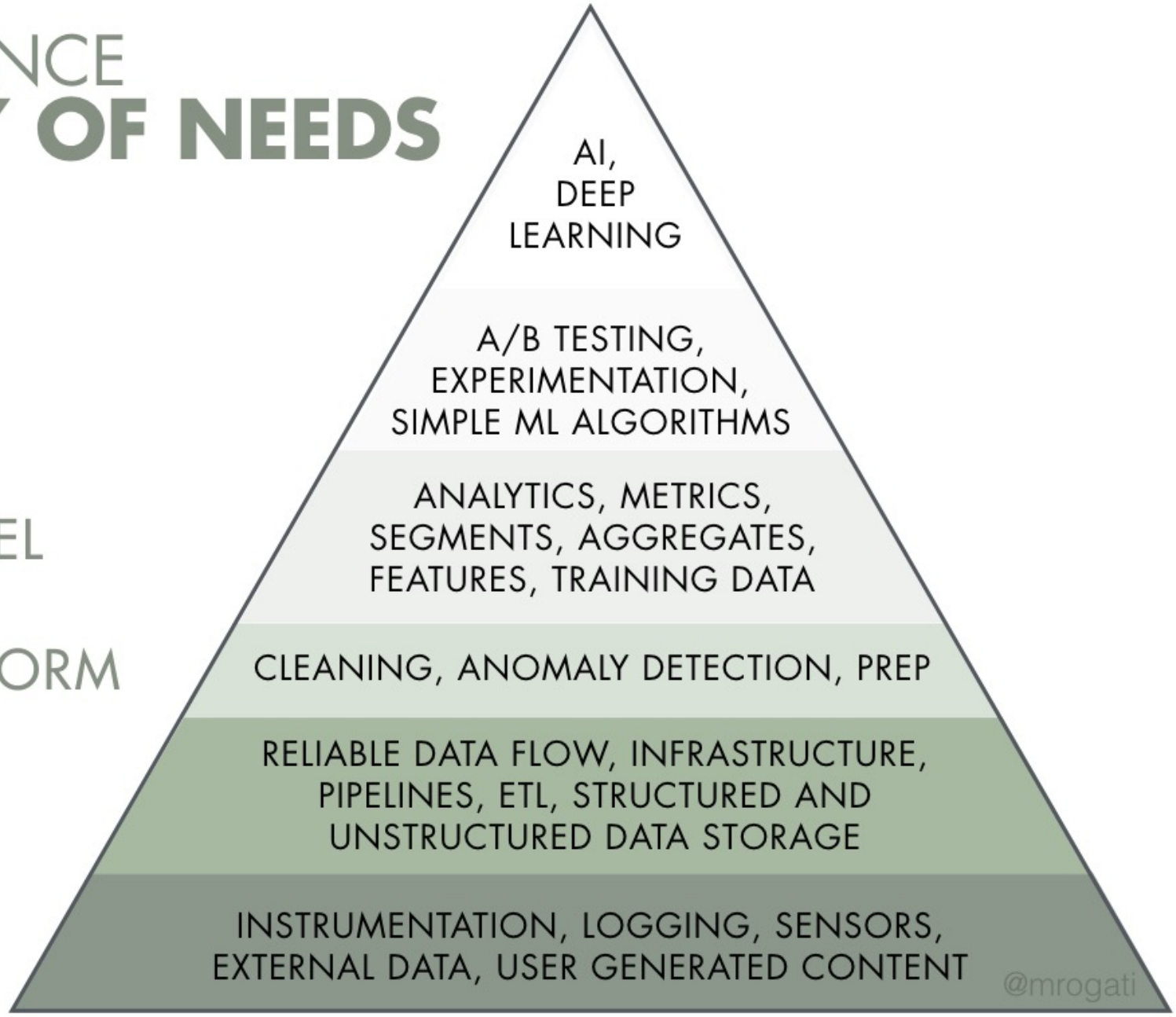
# Invenção brasileira pode trazer novo sinal vital para triagem do paciente

Criado por Sérgio Mascarenhas, sensor da Brain4care traz método não invasivo para monitorar problemas neurológicos e agora parte para conquistar uma fatia estimada de US$ 1 bilhão no mercado dos EUA

Por **Laura Pancini**
Publicado em: 22/11/2021 às 11h18
Tempo de leitura: **3 min**


Sensor da Brain4care é totalmente não invasivo e envia relatório em tempo real para aplicativo (Brain4care/Divulgação)

# THE DATA SCIENCE HIERARCHY OF NEEDS




LEARN/OPTIMIZE

AGGREGATE/LABEL

EXPLORE/TRANSFORM

MOVE/STORE

COLLECT

Source: https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007

# Jobs!

glassdoor

## Hot Companies Hiring Data Scientists

November 11, 2019 | Posted by Jillian Kramer

# Inspiration

- Intro to Data Science by Steve Brunton (2019) / University of Washington
  - eigensteve.com
- Ciência de Dados: Introdução e Motivação por Francisco Rodrigues (2020) / Universidade de São Paulo
  - https://www.youtube.com/watch?v=Im2lagDGDAU&t=636s
- Notas de aula do Prof. Andre Carvalho (2016) / Universidade de São Paulo