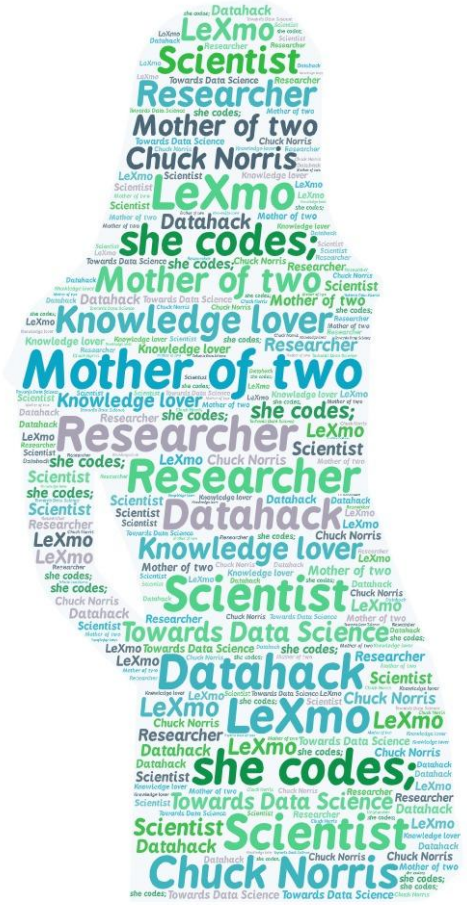
Introduction to Big Data

All you might want to dive into

By Dina Bavli

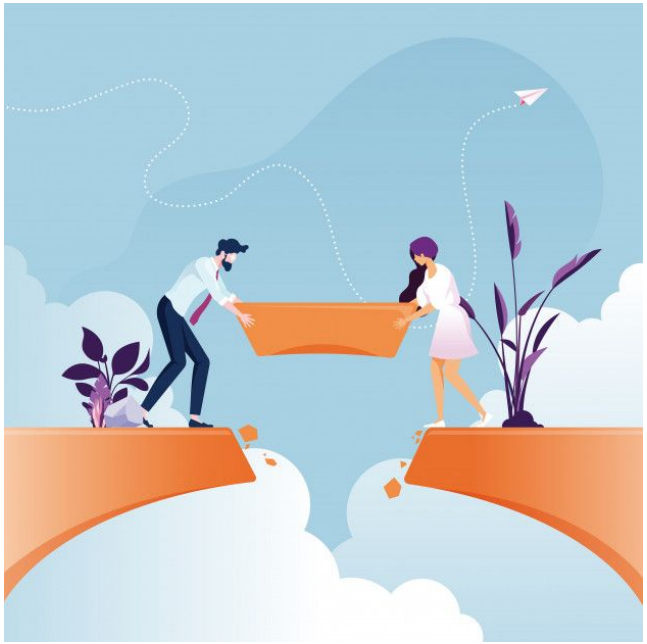


Who am I?

By Dina Bavli

Why am I here?

By Dina Bavli



By Dina Bavli

Why should you be here?

By Dina Bavli

Introducing the forest

Photo by Hatham on Unsplash

By Dina Bavli

What are we doing here?

By Dina Bavli

What are we going to do in the next two hours?

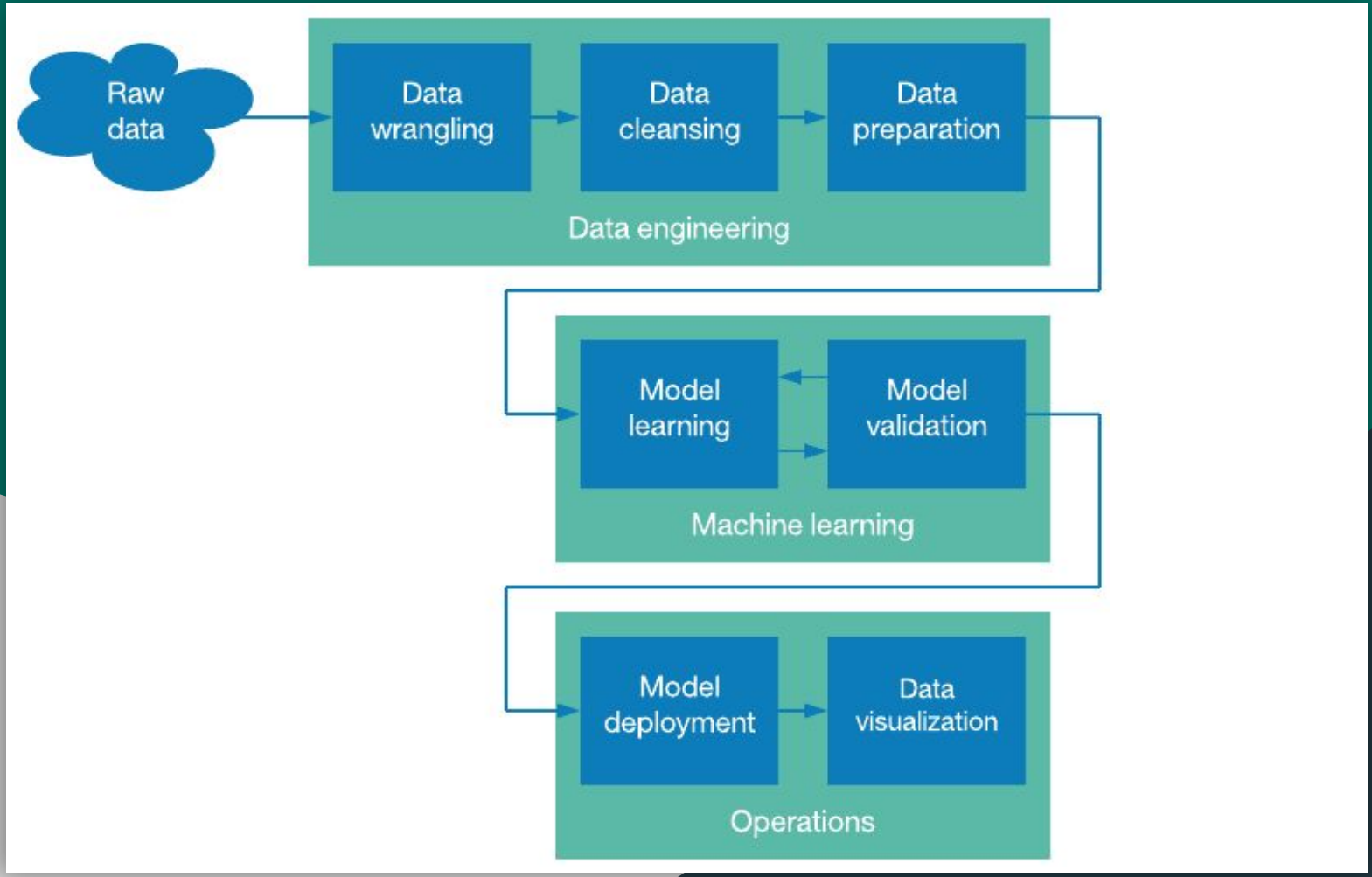
We will cover:

1. Data science pipeline.
2. Machine learning- a brief view of different types.
3. An intuitive introduction to deep learning, active learning, and reinforcement learning.
4. EDA - Exploratory Data Analysis or data exploration.
5. Data preprocessing- feature selection and feature engineering.
6. SNA- Social Network Analysis.

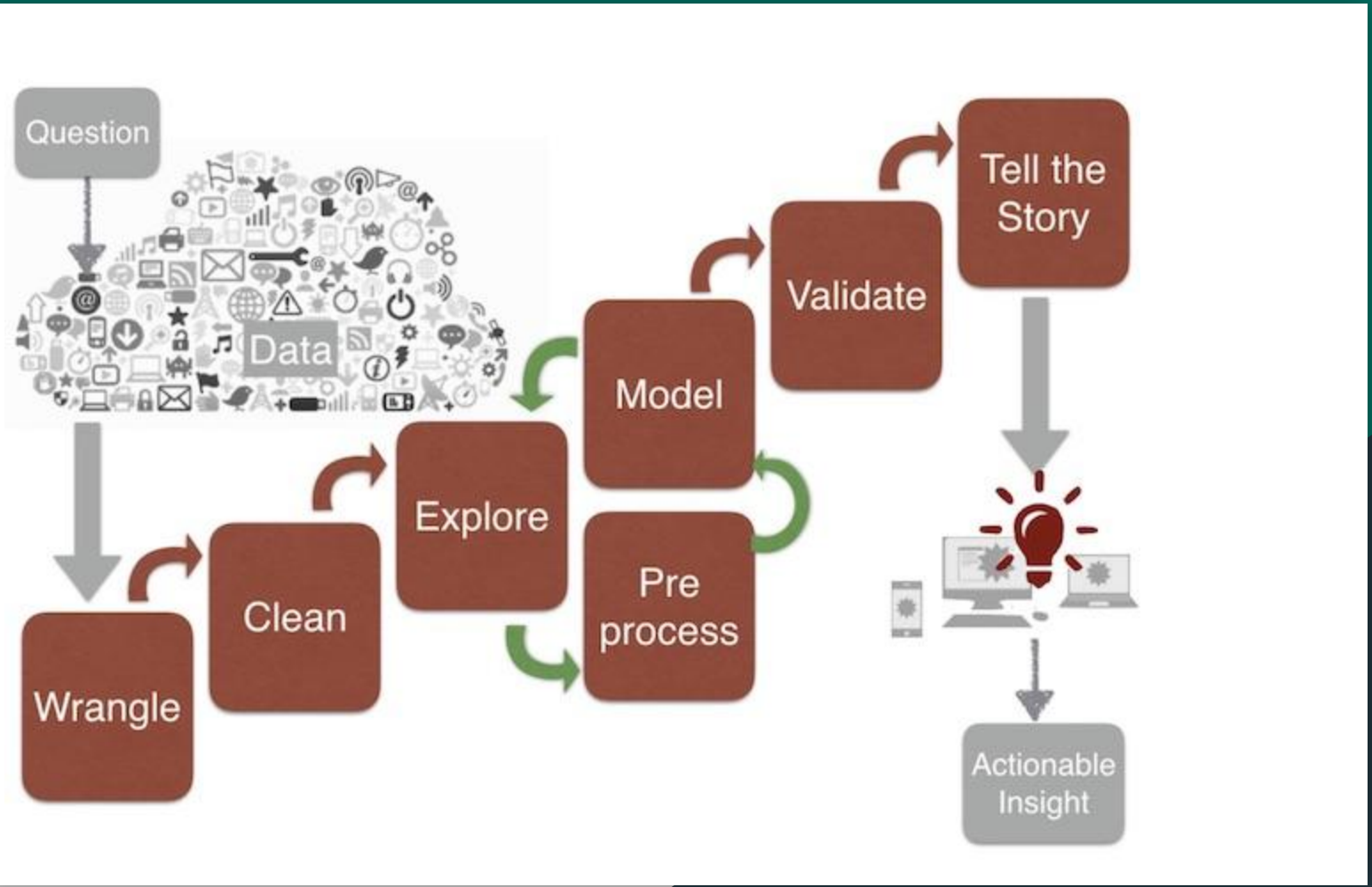
By Dina Bavli

Data science pipeline

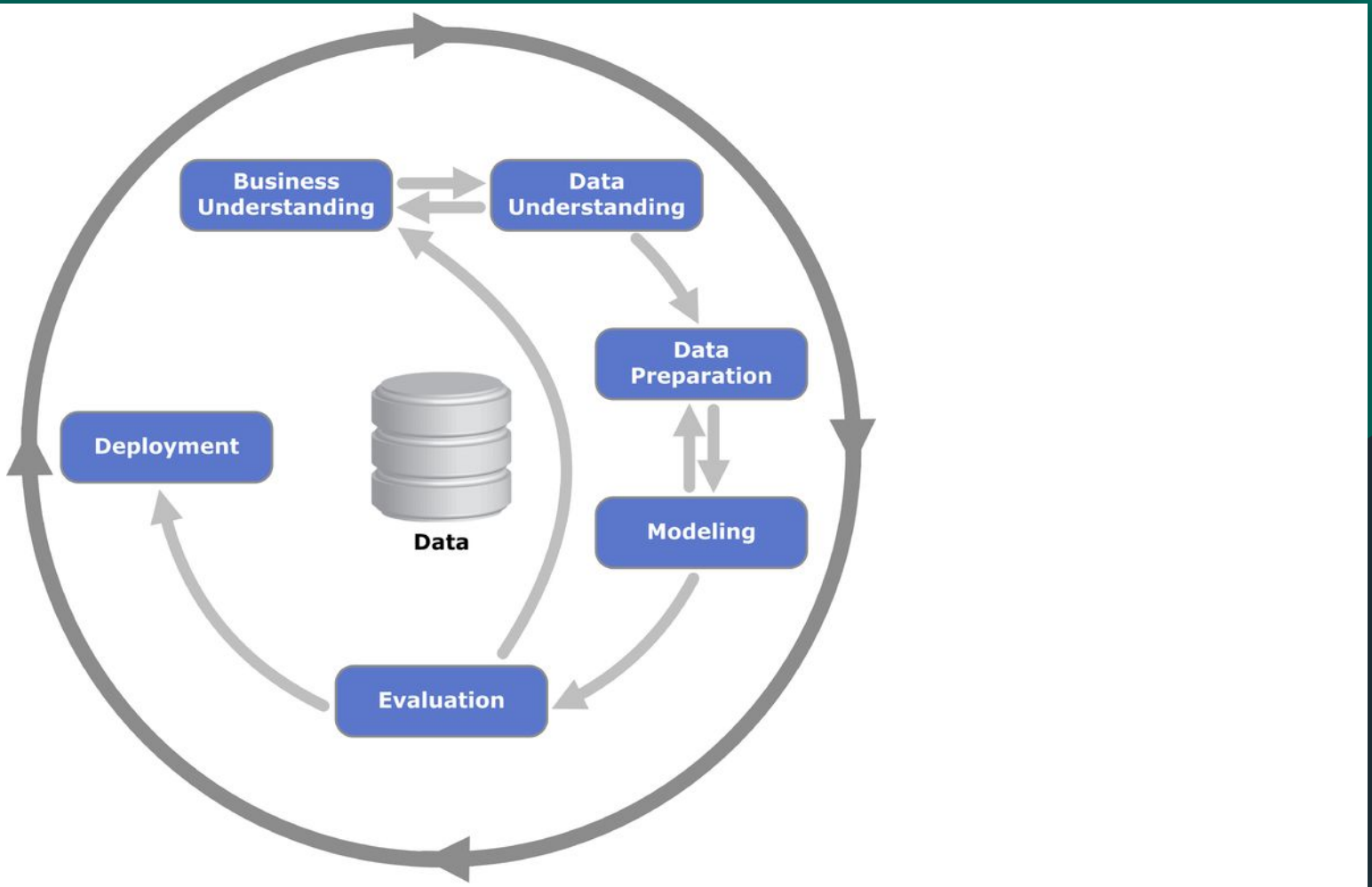
By Dina Bavli



By Dina Bavli



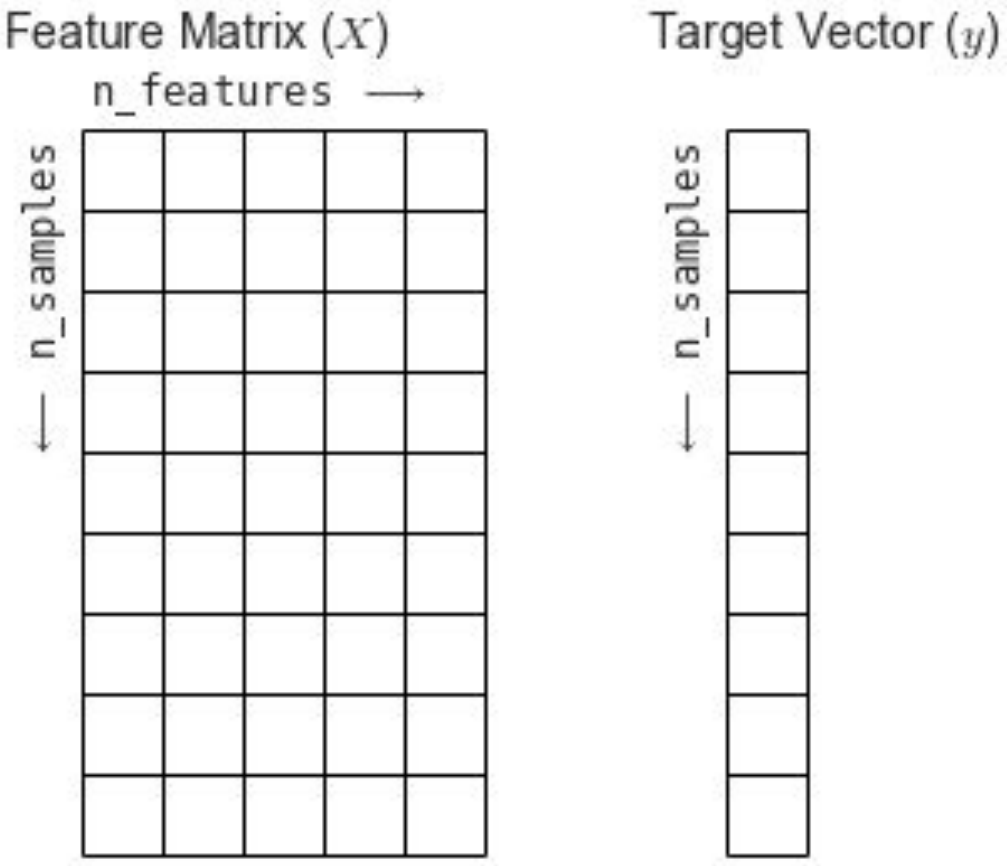
By Dina Bavli

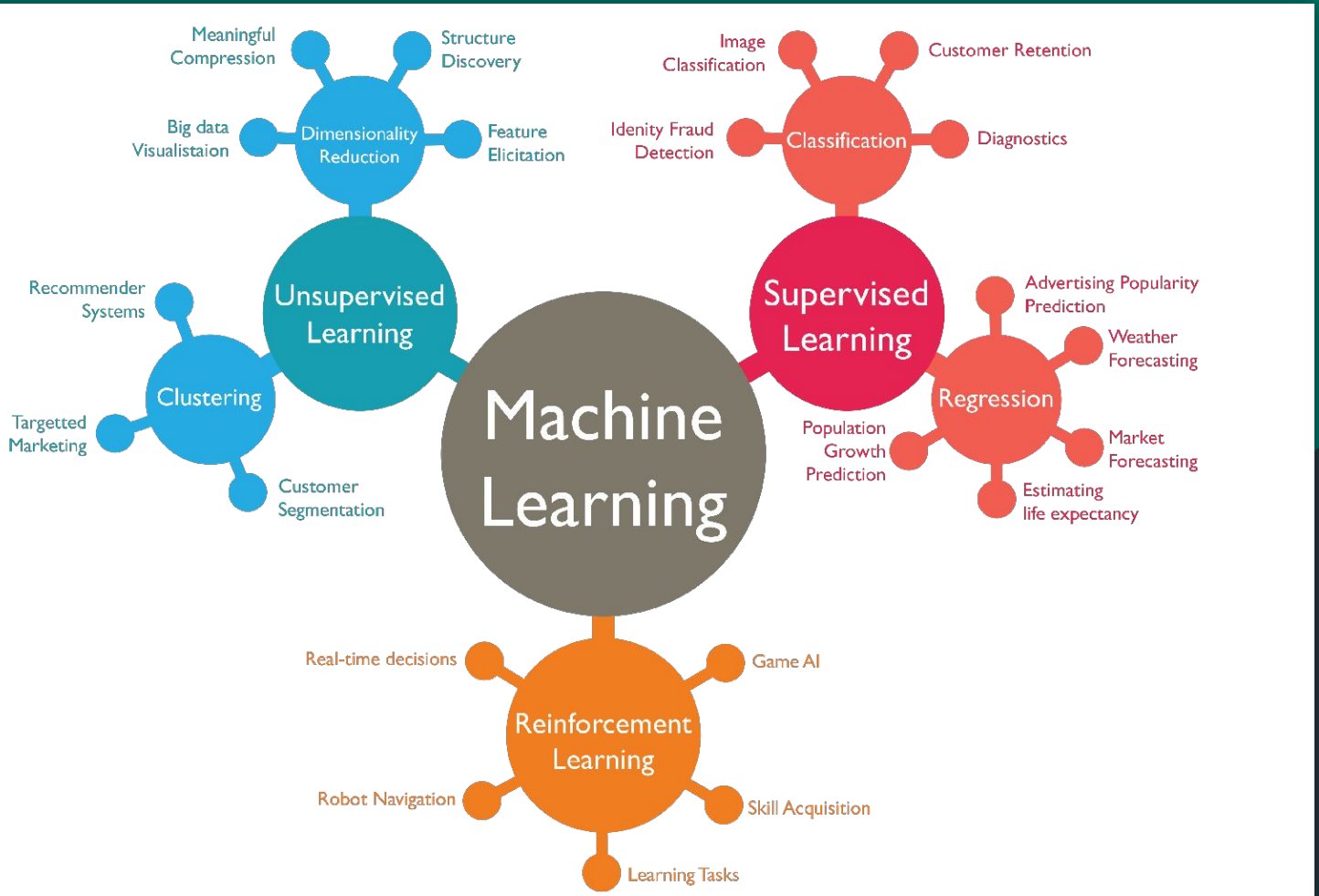


By Dina Bavli

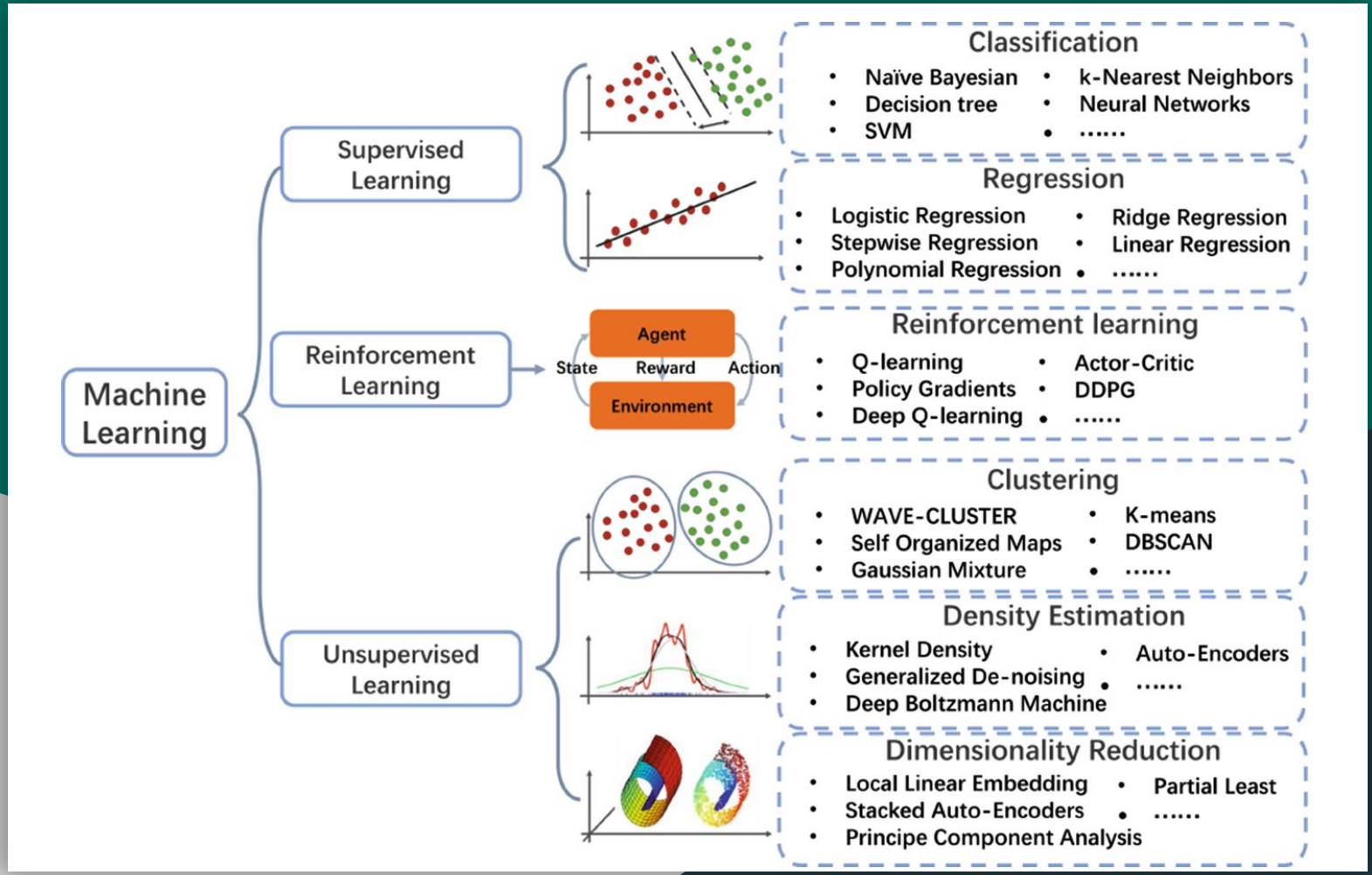
Machine learning

By Dina Bavli


Feature Matrix (X)
n_features ⟶
n_samples ↓
Target Vector (y)
n_samples ↓



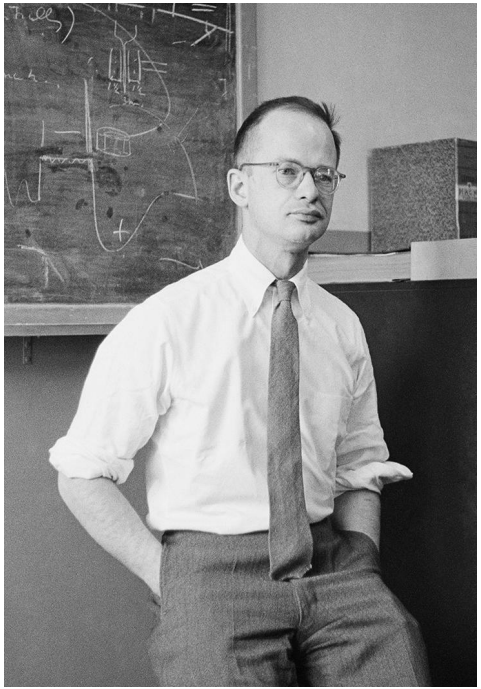
By Dina Bavli

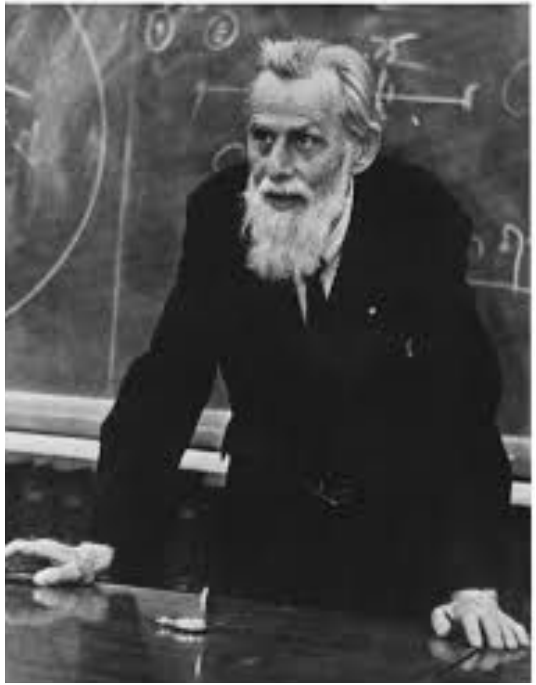


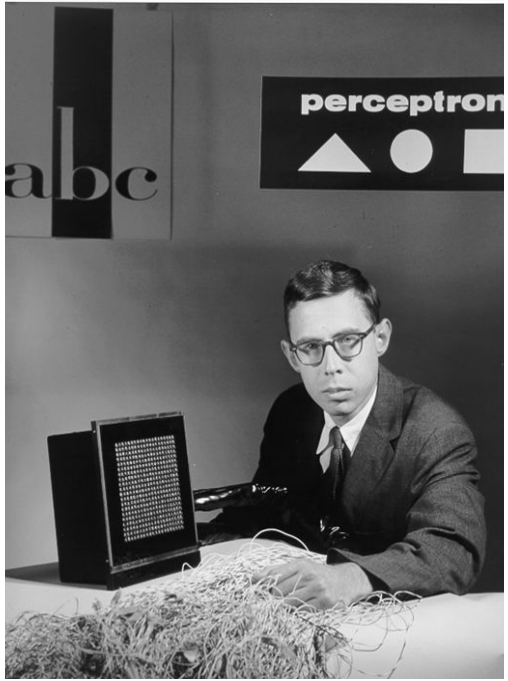
By Dina Bavli

Into Deep

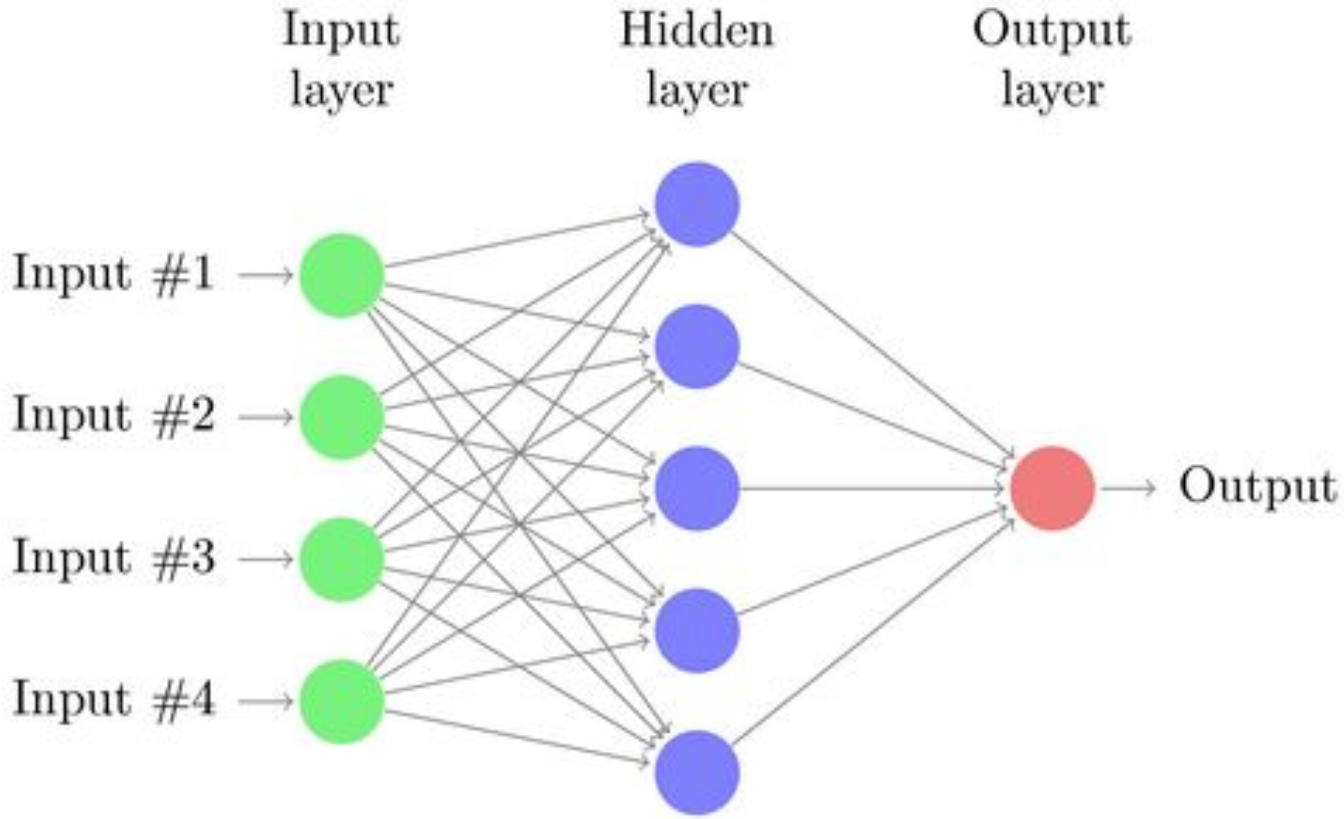
By Dina Bavli



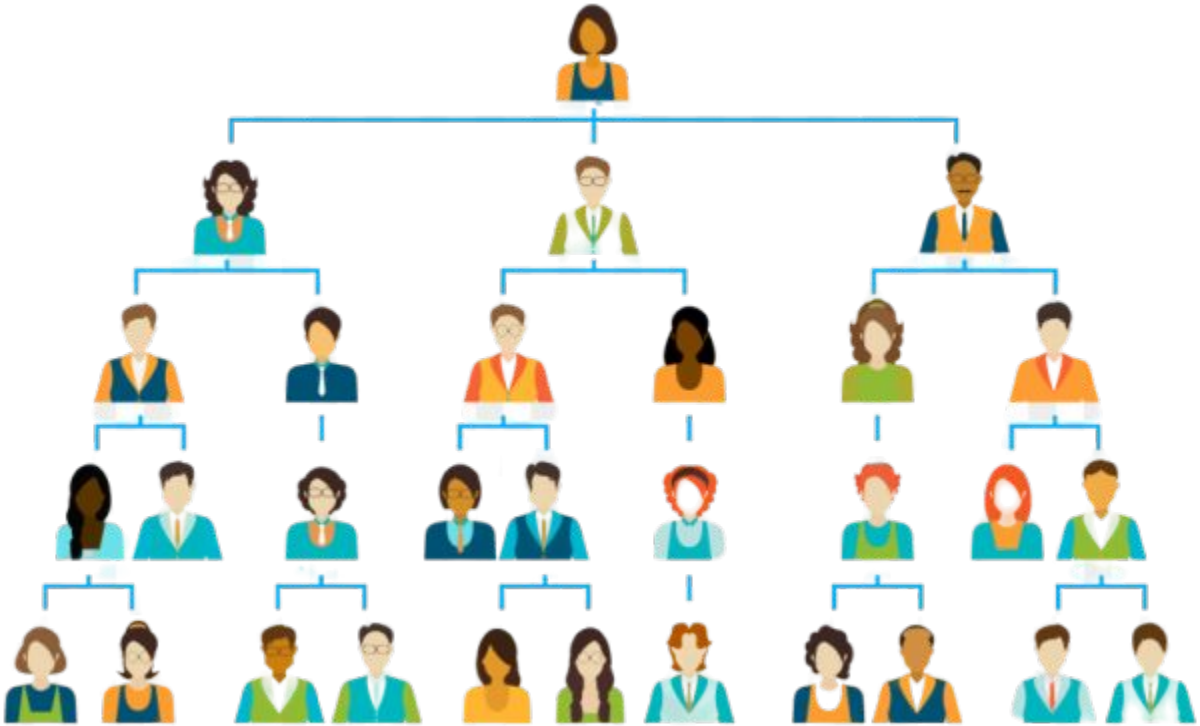




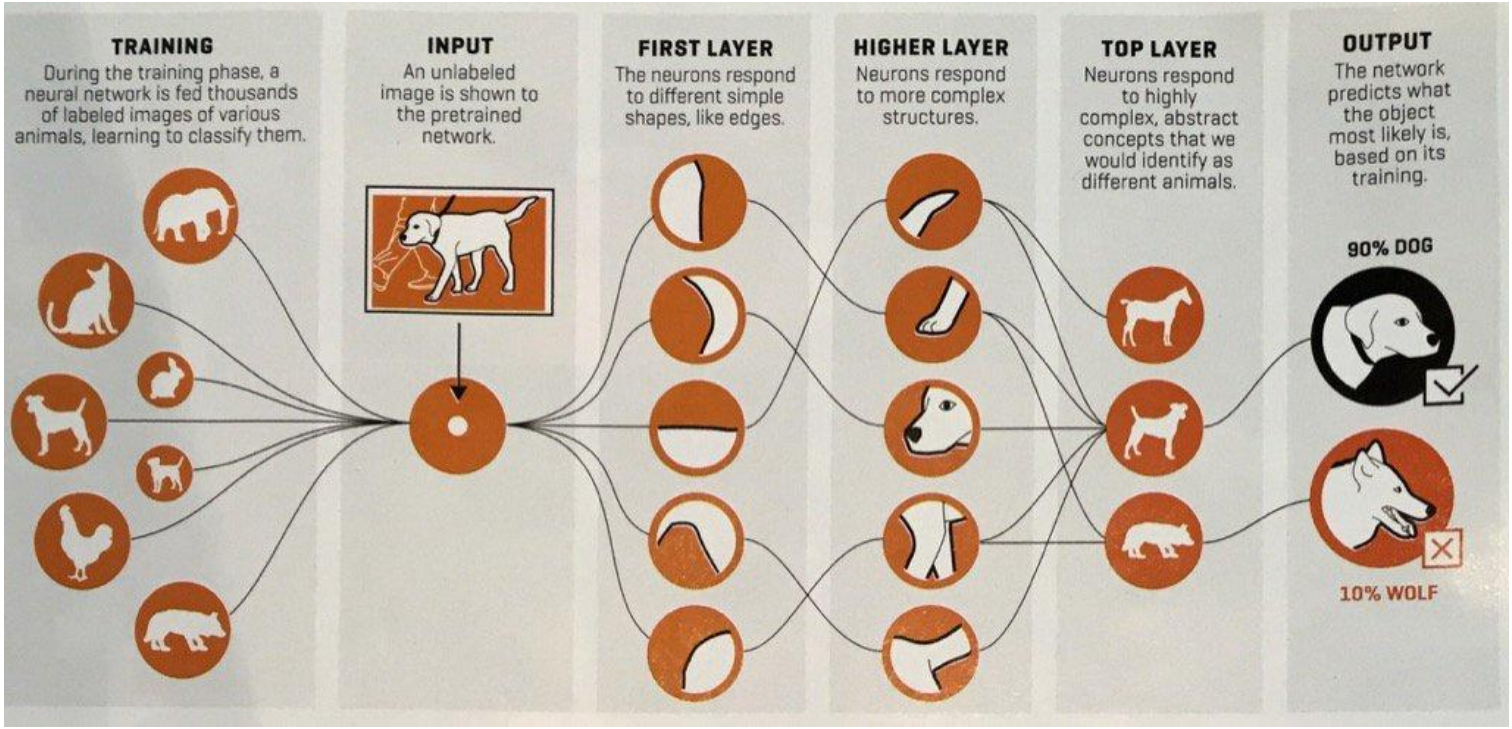
By Dina Bavli



By Dina Bavli



By Dina Bavli

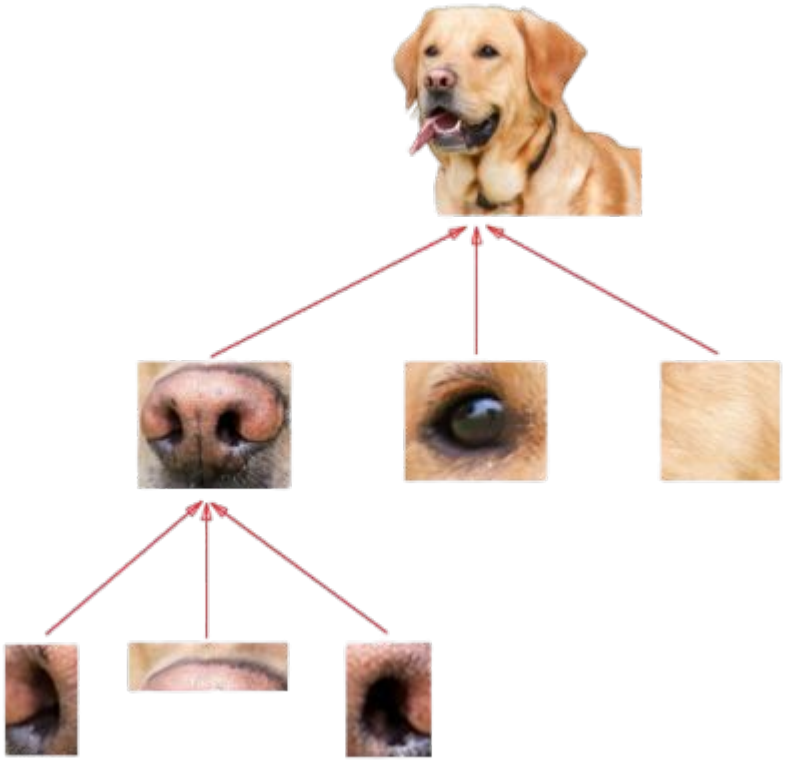


By Dina Bavli

Regular Pitbull-Mix Face: 0.8

Left Eye: 0.8 Right Eye: 0.7 Snout: 0.9 Mouth: 0.8

By Dina Bavli



By Dina Bavli

Smart Labeling

By Dina Bavli

Passive Machine Learning

Raw, unlabeled data

Unlabeled data passed to oracle

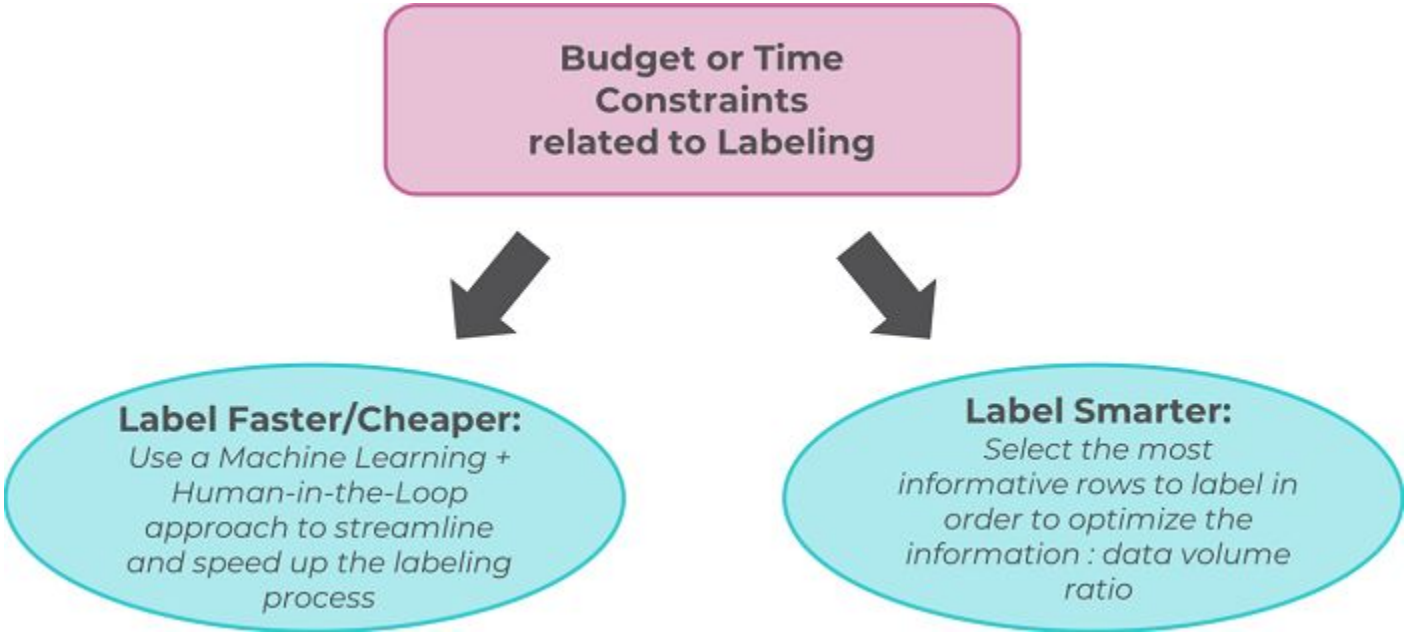
Oracle

Labelled Data

Machine Learning Model

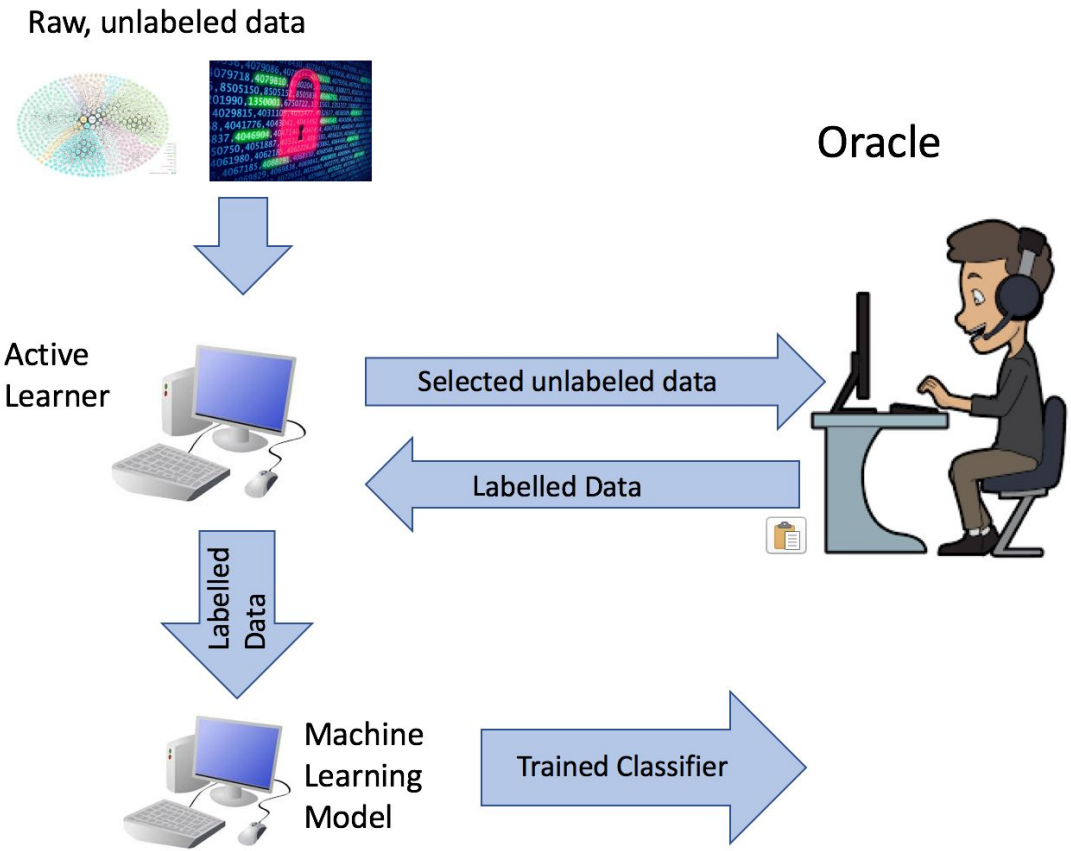
Trained Classifier

By Dina Bavli



By Dina Bavli

Active Machine Learning

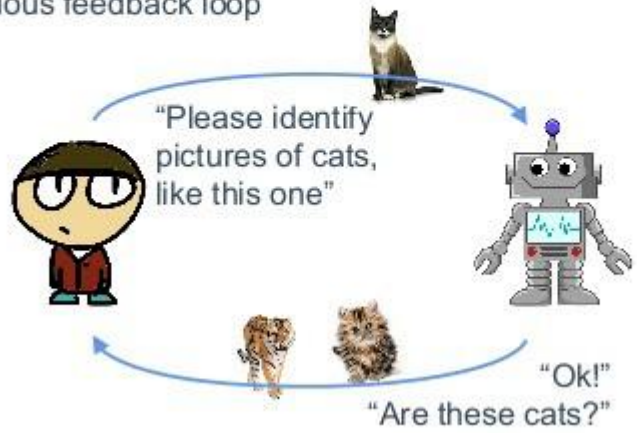


By Dina Bavli

Active Learning

Selecting the optimal data to manually label for Machine Learning

Often a continuous feedback loop

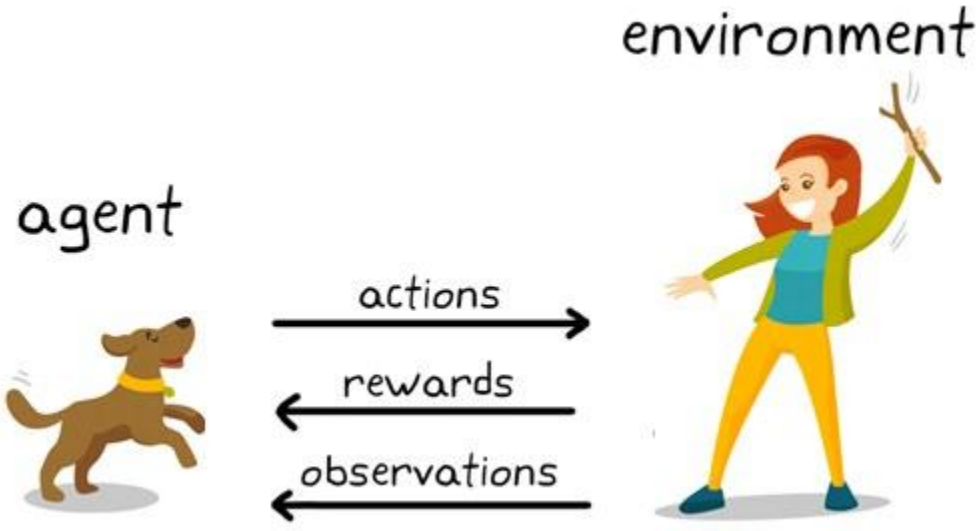


By Dina Bavli

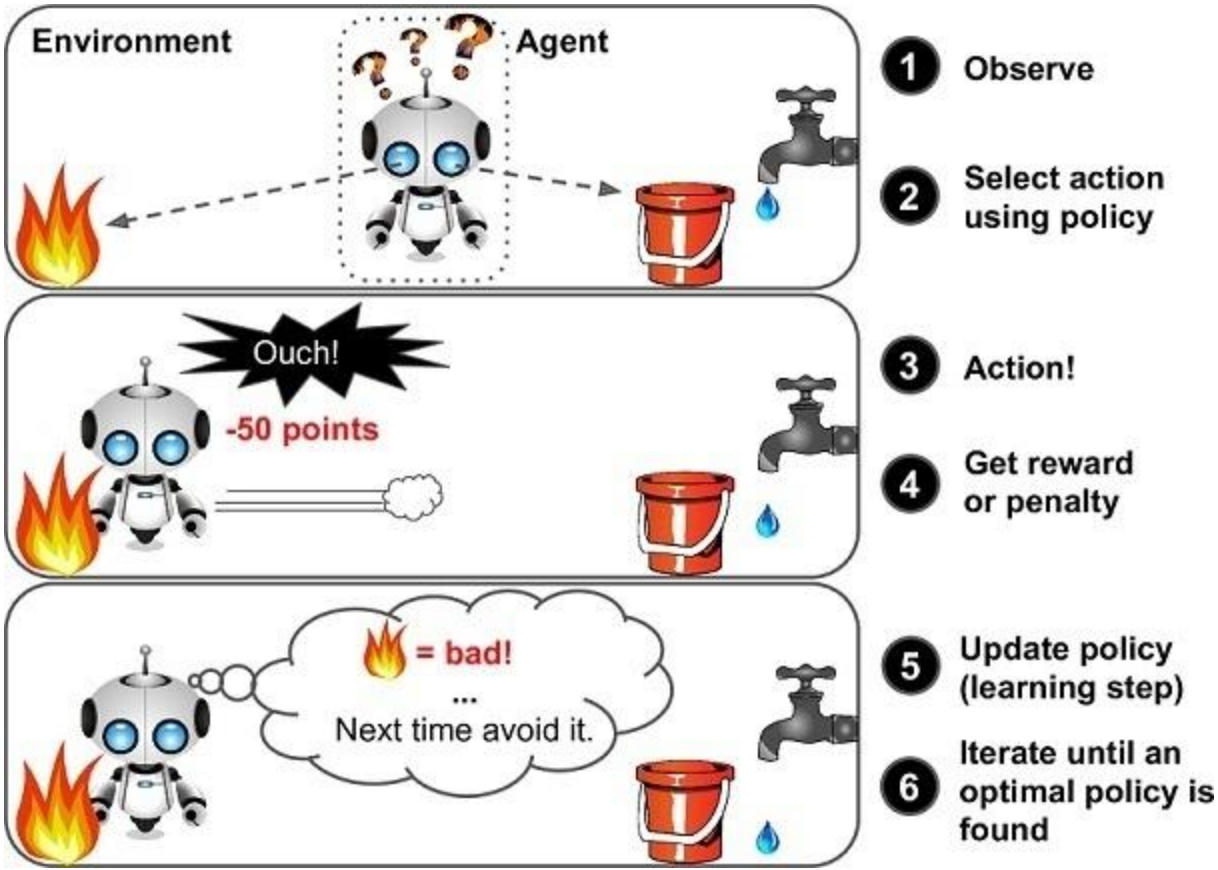
No Labels- No Data

By Dina Bavli

Reinforcement Learning



By Dina Bavli



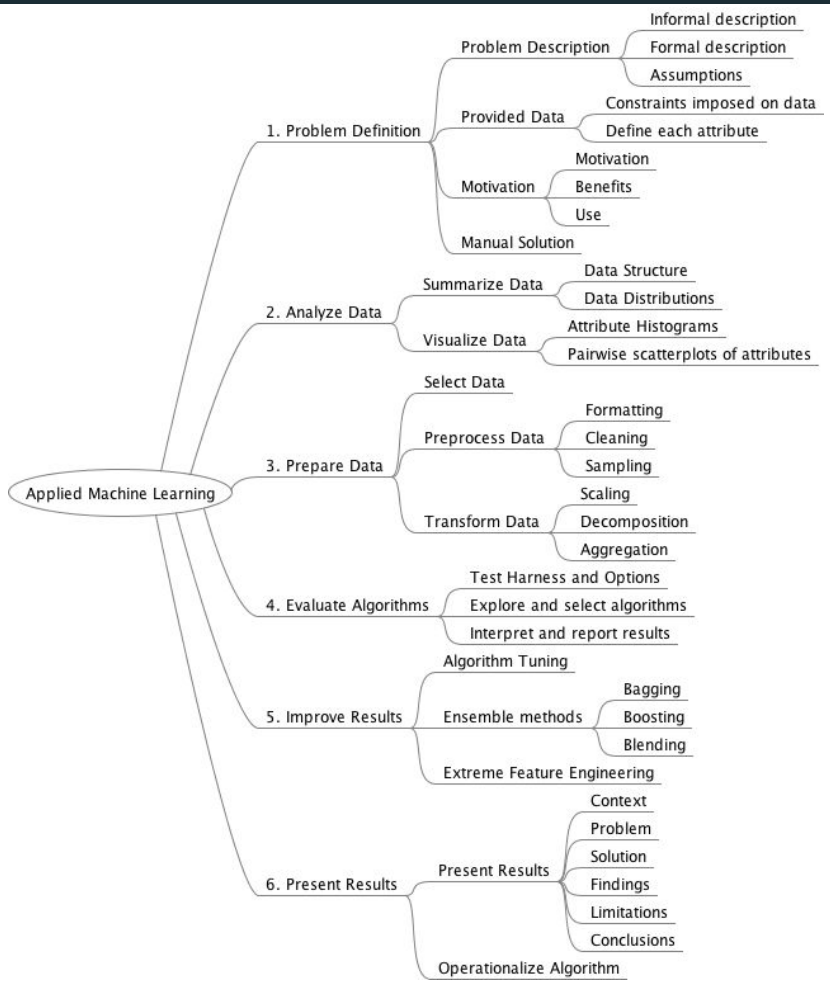
1. Observe
2. Select action using policy
3. Action!
4. Get reward or penalty
5. Update policy (learning step)
6. Iterate until an optimal policy is found

By Dina Bavli

By Dina Bavli

By Dina Bavli

Summary of Data Science Pipeline



By Dina Bavli

The White Rabbit put on his spectacles. **"Where shall I begin, please your Majesty?"** he asked.

"Begin at the beginning," the King said gravely, **"and go on till you come to the end: then stop."**

Begin at the beginning

By Dina Bavli

EDA-Data Exploration

By Dina Bavli

Why is EDA important?

GIGO- Garbage In, Garbage Out

Your analysis is as good as your data.

By Dina Bavli

Exploratory Data Analysis does two main things:

1. Understanding variables and relationships between them.
2. It helps clean up the dataset.

By Dina Bavli

Exploratory Data Analysis methods:

- Univariate visualization
- Bivariate visualization
- Multivariate visualization
- Dimensionality reduction

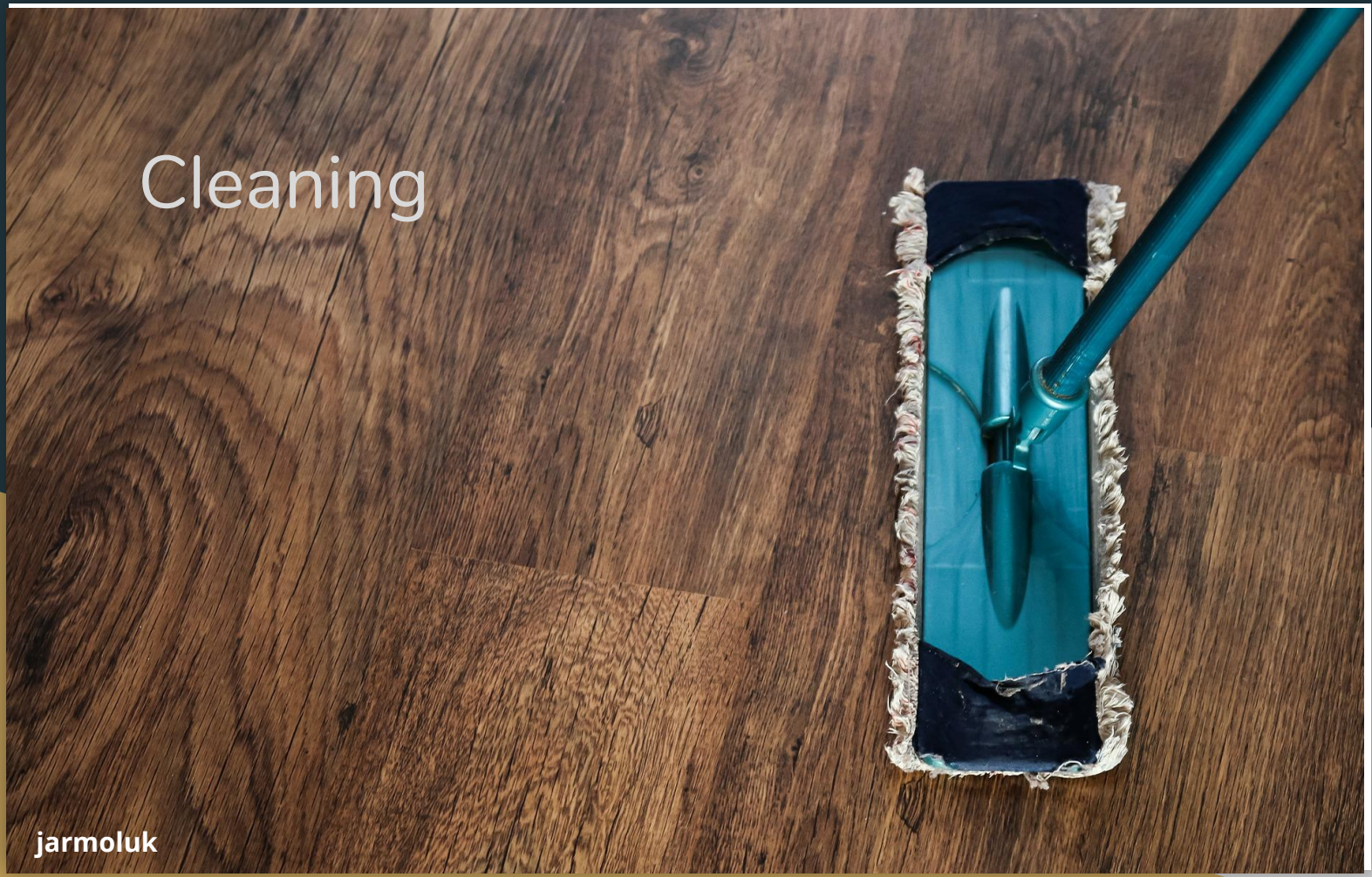
By Dina Bavli

Commonly used plots for EDA:

- Histogram
- Scatter plots
- Maps
- Feature correlation plot (heatmap)
- Time series plots

By Dina Bavli

Cleaning



jarmoluk

By Dina Bavli

What to clean?

Missing

Outliers

Features

By Dina Bavli

Wrong cleaning

- Create bias
- Tilt the distribution
- Dismorf the data

Right cleaning

- Reduce costs
- Reduce process time
- Reduce overfitting

By Dina Bavli

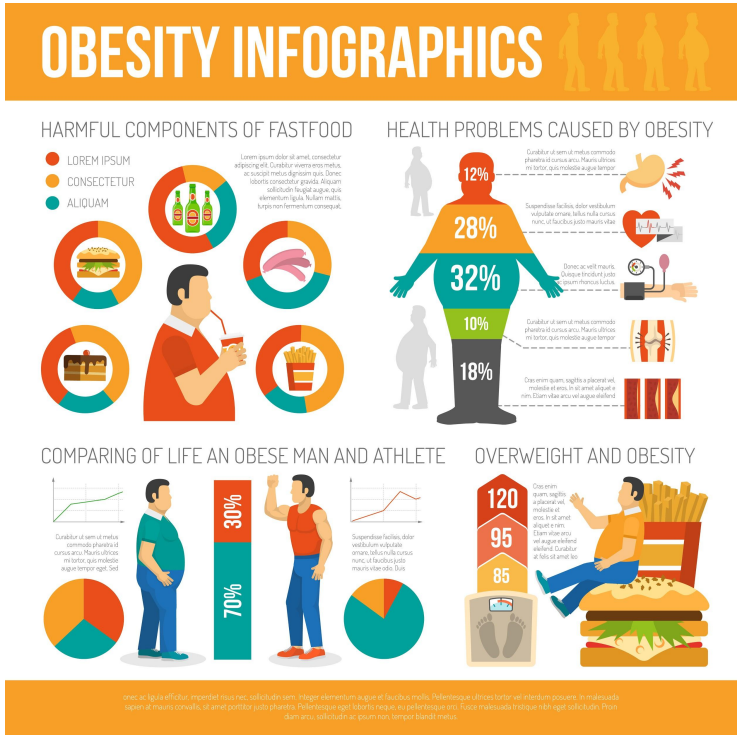
Features

By Dina Bavli

Health insurance

created by pch.vector - www.freepik.com

By Dina Bavli

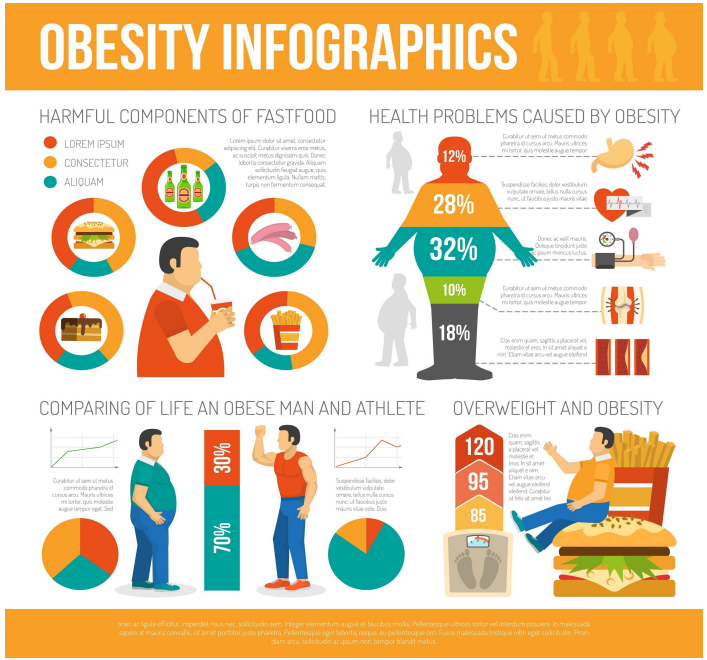


Designed by macrovector / Freepik

Designed by macrovector_official / Freepik

By Dina Bavli

Insurance equation= smoking + obesity + age^2



Designed by macrovector / Freepik

Designed by macrovector_official / Freepik

By Dina Bavli

Features and Influence

By Dina Bavli

Topic	G	A	R	P
Abortion	35	8	9	3
Catholic Church	27	9	553	7
George W. Bush	4	0	8	68
Israel	4	18	623	12
Michael Jackson	2	42	4	0

Table 4: A list of topics and the occurrence of issues associated with them in **A**ge, **G**ender, **R**eligion, and **P**olitics. An occurrence > 5 indicates it is an issue relevant to that topic.

Rosenthal, S., & McKeown, K. (2016, November). Social proof: The impact of author traits on influence detection. In *Proceedings of the First Workshop on NLP and Computational Social Science* (pp. 27-36).
https://www.aclweb.org/anthology/W16-5604.pdf

By Dina Bavli

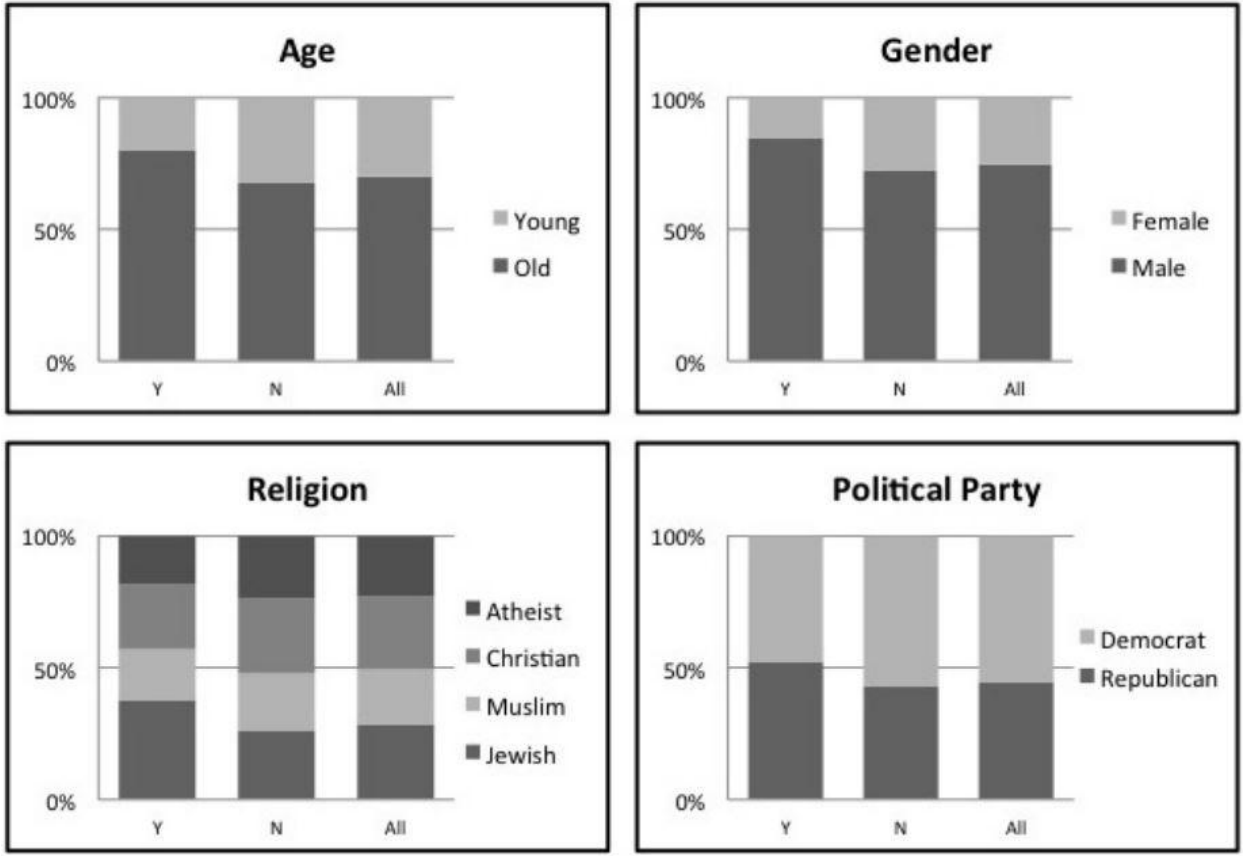


Figure 1: Breakdown of the binary features by influence (Y/N) and overall (All) in the training set.

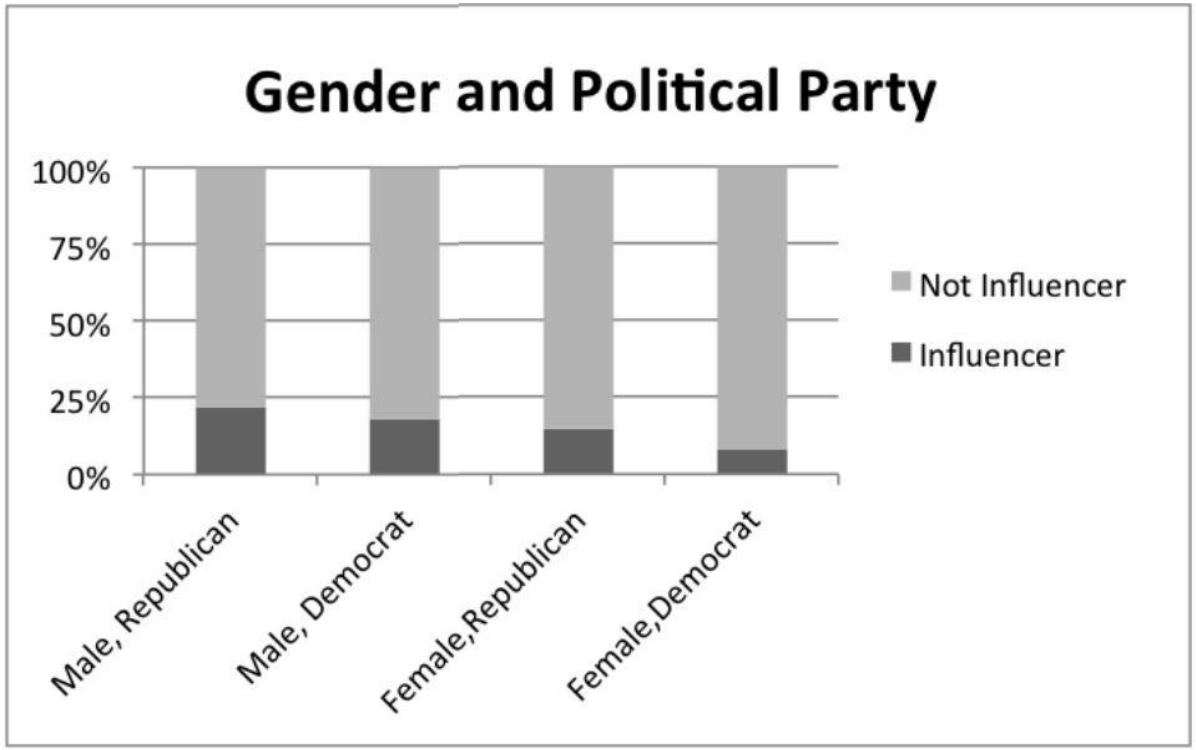
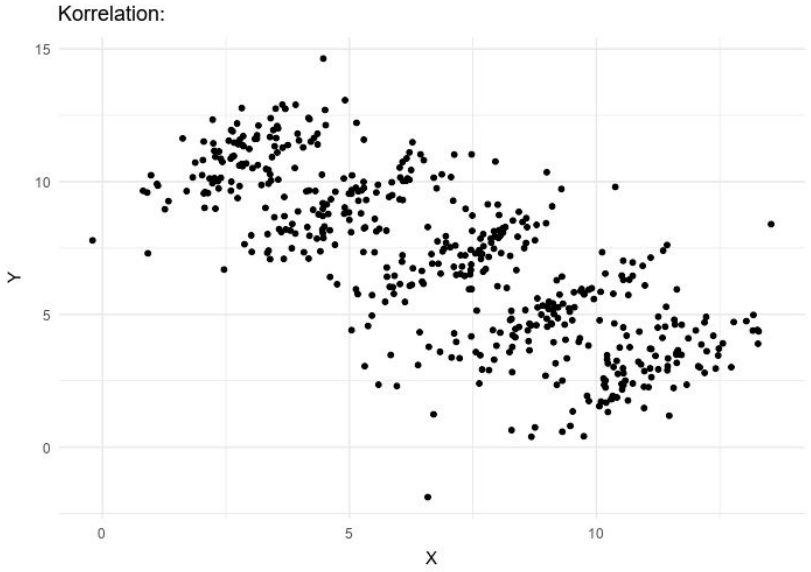


Figure 3: The breakdown of influencers and non-influencers in the training data based on the binary combination feature of gender and political party.

Simpson's paradox

Need to take into account when combining features



A trend appears in several different groups of data but disappears or reverses when these groups are combined.

By Dina Bavli

Homophily, Social Proof, Echo Chambers and Filter Bubbles

By Dina Bavli

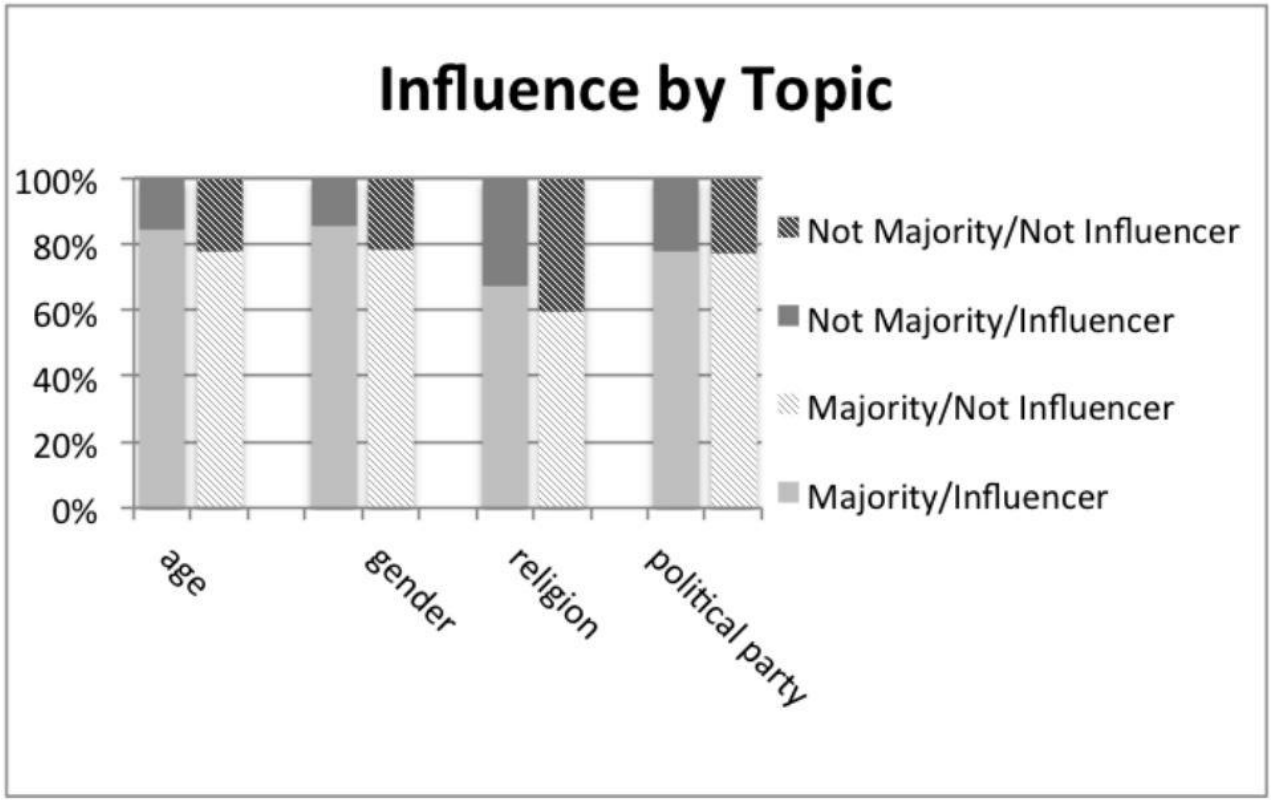
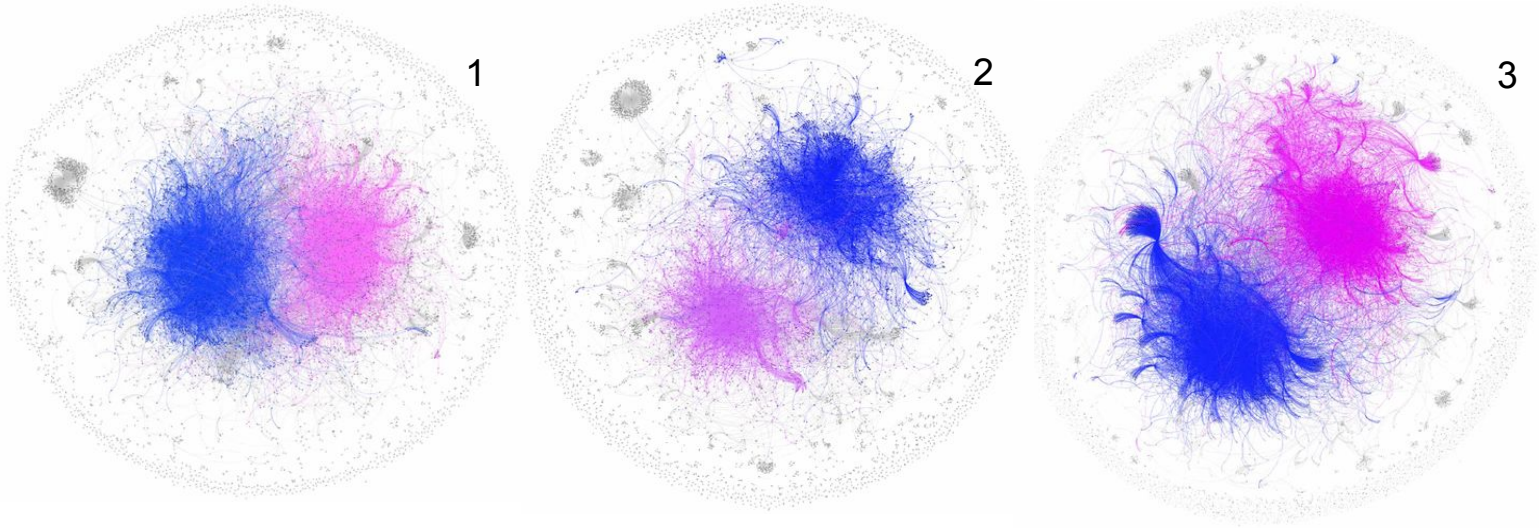


Figure 2: The breakdown of the users being in the majority within their document for each author trait with topic being taken into account.

By Dina Bavli

Identifying Types of Influencers using SNA

By Dina Bavli



Identifying four types of influencers using a matrix of participation, visibility, and political position.

Recuero, R., Zago, G., & Soares, F. (2019). Using social network analysis and social capital to identify user roles on polarized political conversations on Twitter. *Social Media+ Society*, *5*(2), 2056305119848745. https://journals.sagepub.com/doi/full/10.1177/2056305119848745

By Dina Bavli

What is a social network?

Points	Lines
People	Followers, friendships, kin
Companies	Acquire, trade, chain of supply
Disease	Chain of infection
Web pages	links
Countries	Flight routes, trade
Articles	Citation, writers

By Dina Bavli

Points and line have formal names that vary in different disciplines:

Points	Lines	Discipline
Vertices	Edges, arcs	Math
Nodes	Linked, edges	Computer Science
Sites	Bonds	physics
Actors	Ties, relations	Sociology

https://gawron.sdsu.edu/python_for_ss/course_core/book_draft/Social_Networks/Social_Networks.html#what-are-networks

By Dina Bavli

Metrics used:

- **Modularity-** that measures the strength of division of a network into modules (also called groups, clusters or communities)
- **Indegree-** the number of mentions and retweets received by a user, high indegree indicates on visibility, and
- **Outdegree-** the number of users that someone has retweeted or mentioned in a given network, that shows participation.

Recuero, R., Zago, G., & Soares, F. (2019). Using social network analysis and social capital to identify user roles on polarized political conversations on Twitter. *Social Media+ Society*, *5*(2), 2056305119848745.https://journals.sagepub.com/doi/full/10.1177/2056305119848745

By Dina Bavli

Types of influencers found:

Users with a clear political position:

1. **Opinion leaders**- with high indegree, which means a lot of mentions and retweets.
2. **Activists**- with high outdegree, which means the number of users that someone has retweeted or mentioned in a given network.

Users without a clear political position:

3. **Informational influencers**- with high indegree, usually news outlets
4. **News clippers**- with high outdegree.

By Dina Bavli

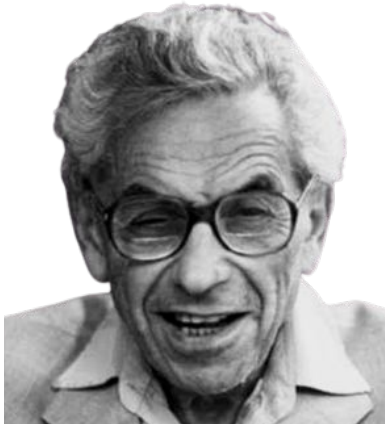
SNA- can be applied to security

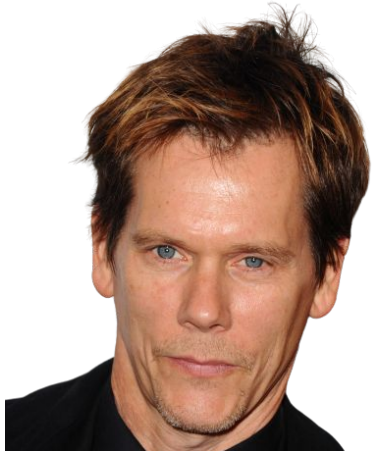
Phillips, E. F., Nurse, J. R., Goldsmith, M. H., & Creese, S. J. (2015). Applying social network analysis to security. https://www.researchgate.net/publication/274663734_Applying_Social_Network_Analysis_to_Security

By Dina Bavli

What do those have in common?

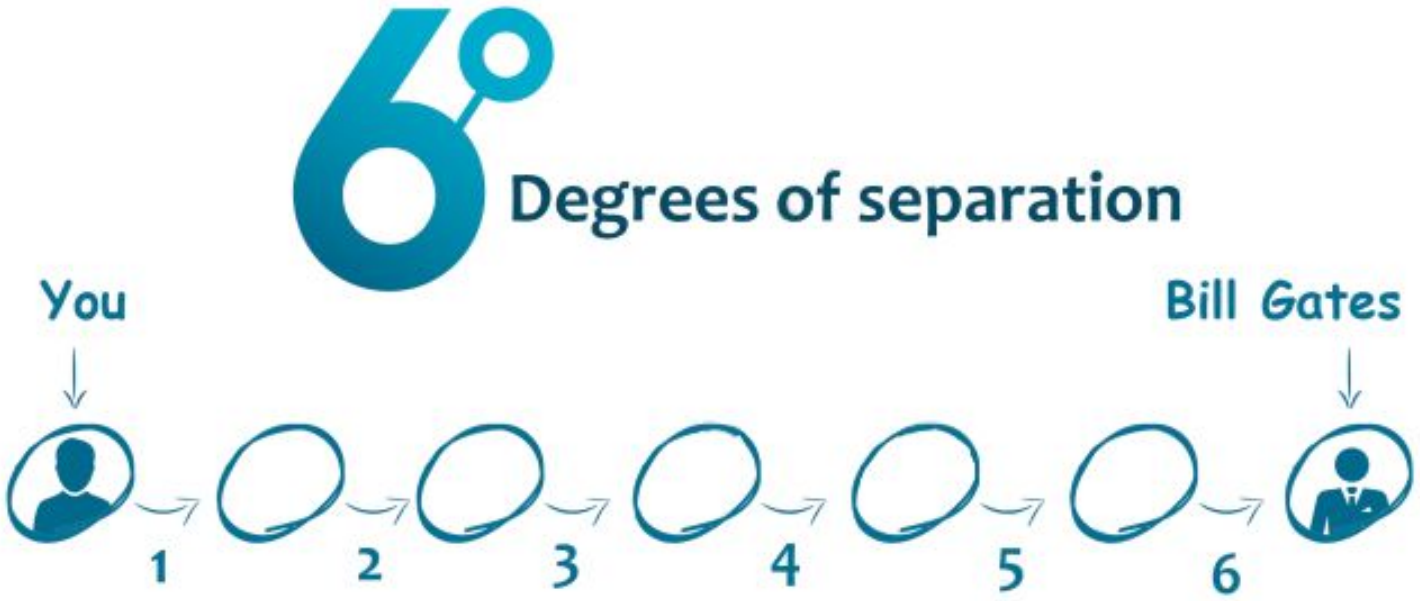
And how does it connect to SNA?





By Dina Bavli

By Dina Bavli



By Dina Bavli

Links

“Six Degrees of Kevin Bacon” also known as The Oracle of Bacon https://oracleofbacon.org/

The Erdös Number Project https://oakland.edu/enp/compute/

By Dina Bavli

SOCRATES
Empowering Knowledge

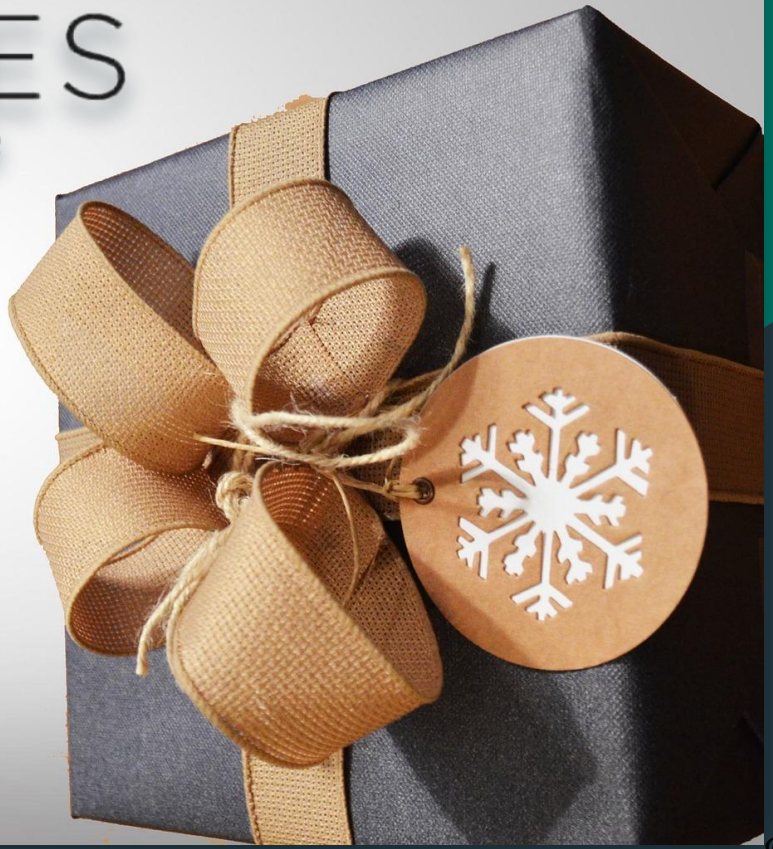
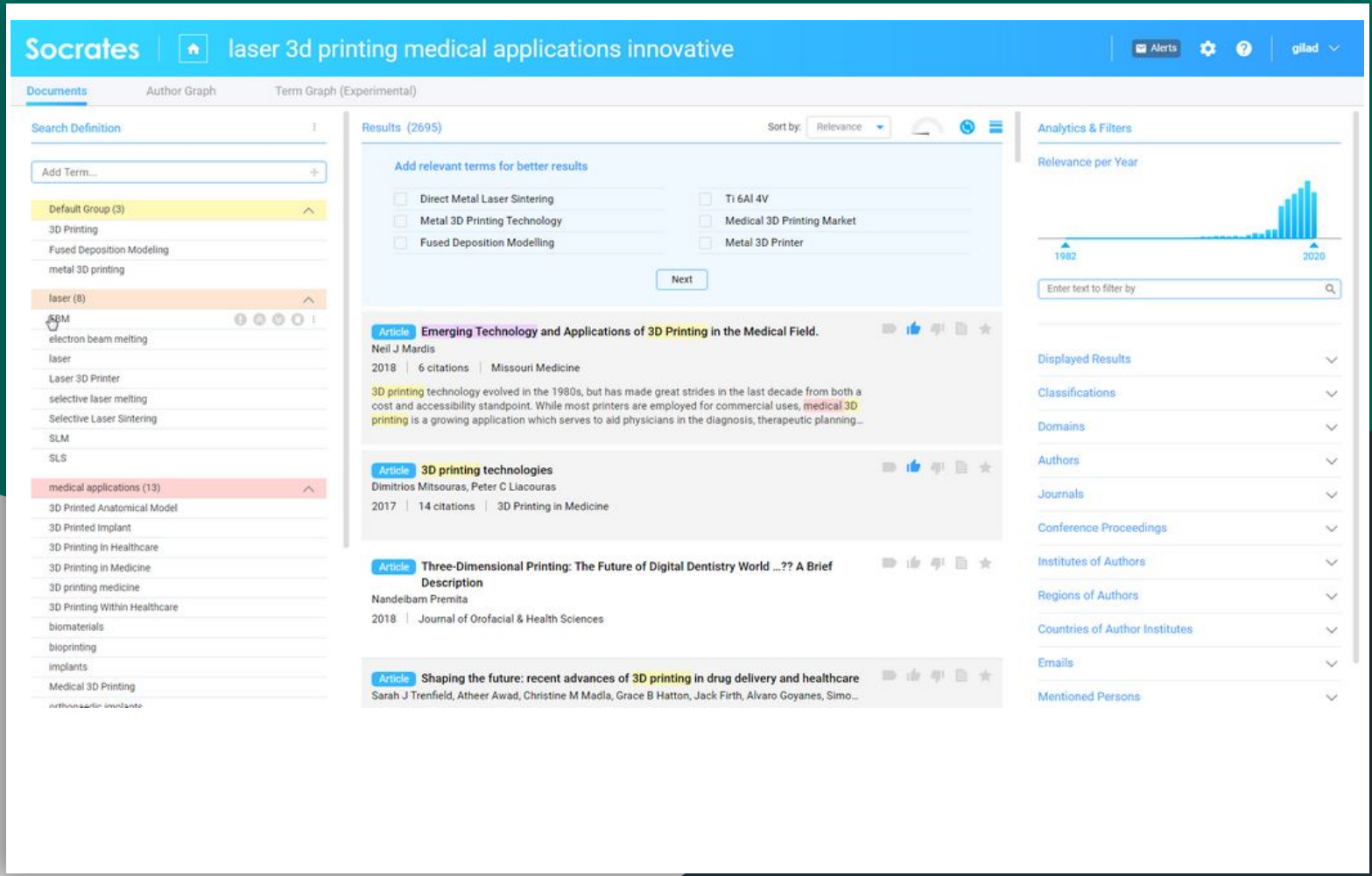
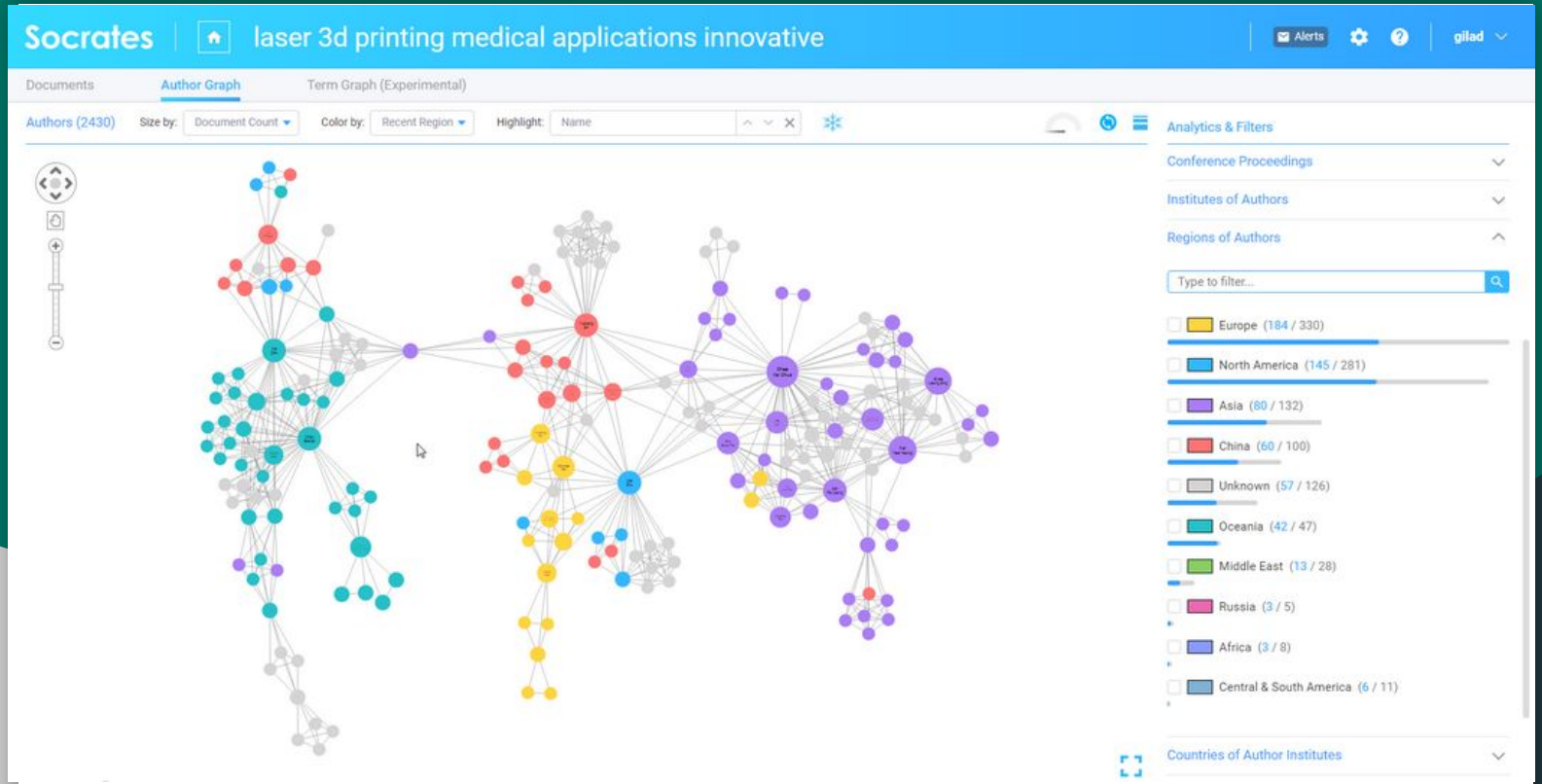


Image by Free-Photos from Pixabay

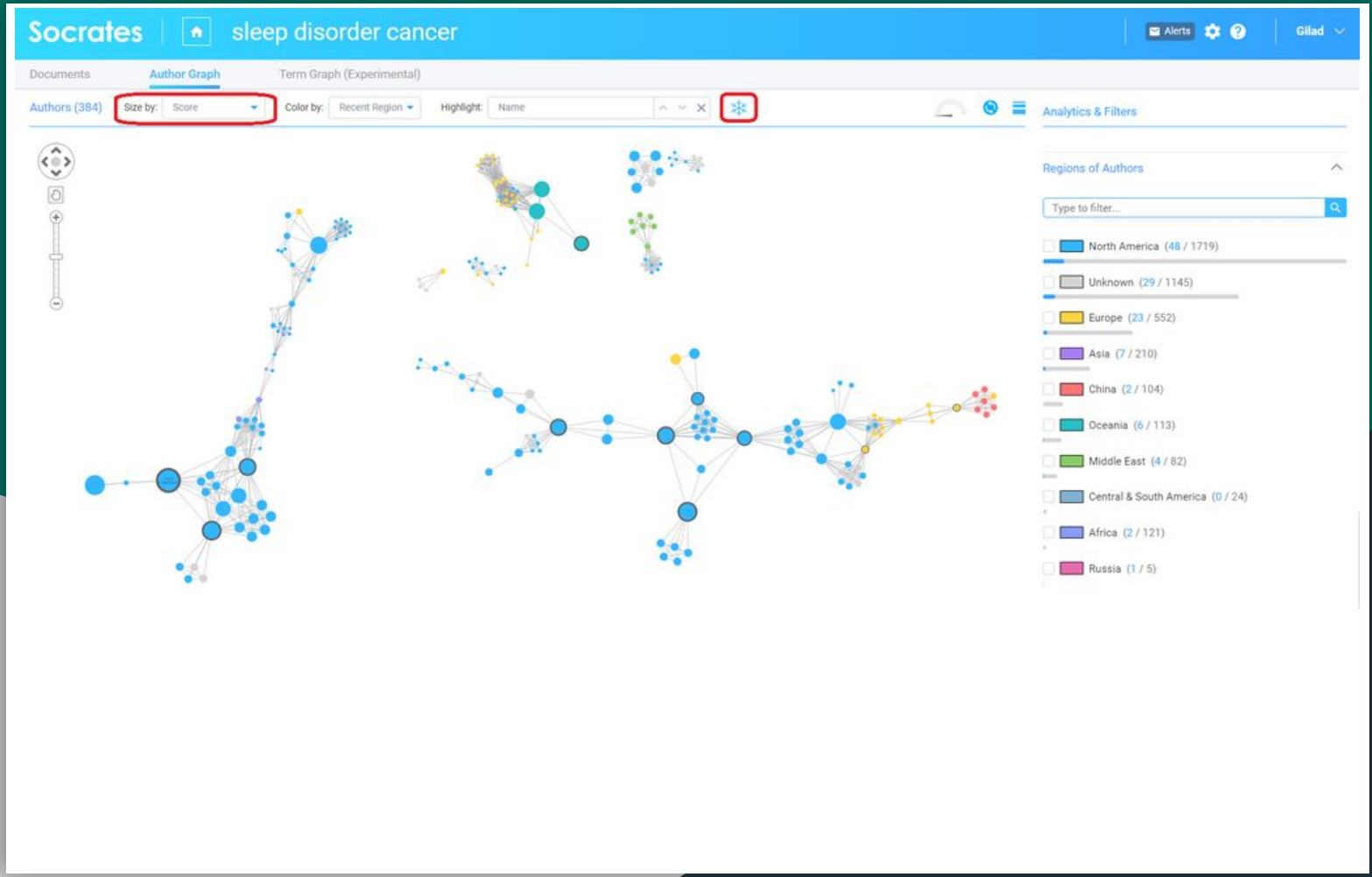
By Dina Bavli



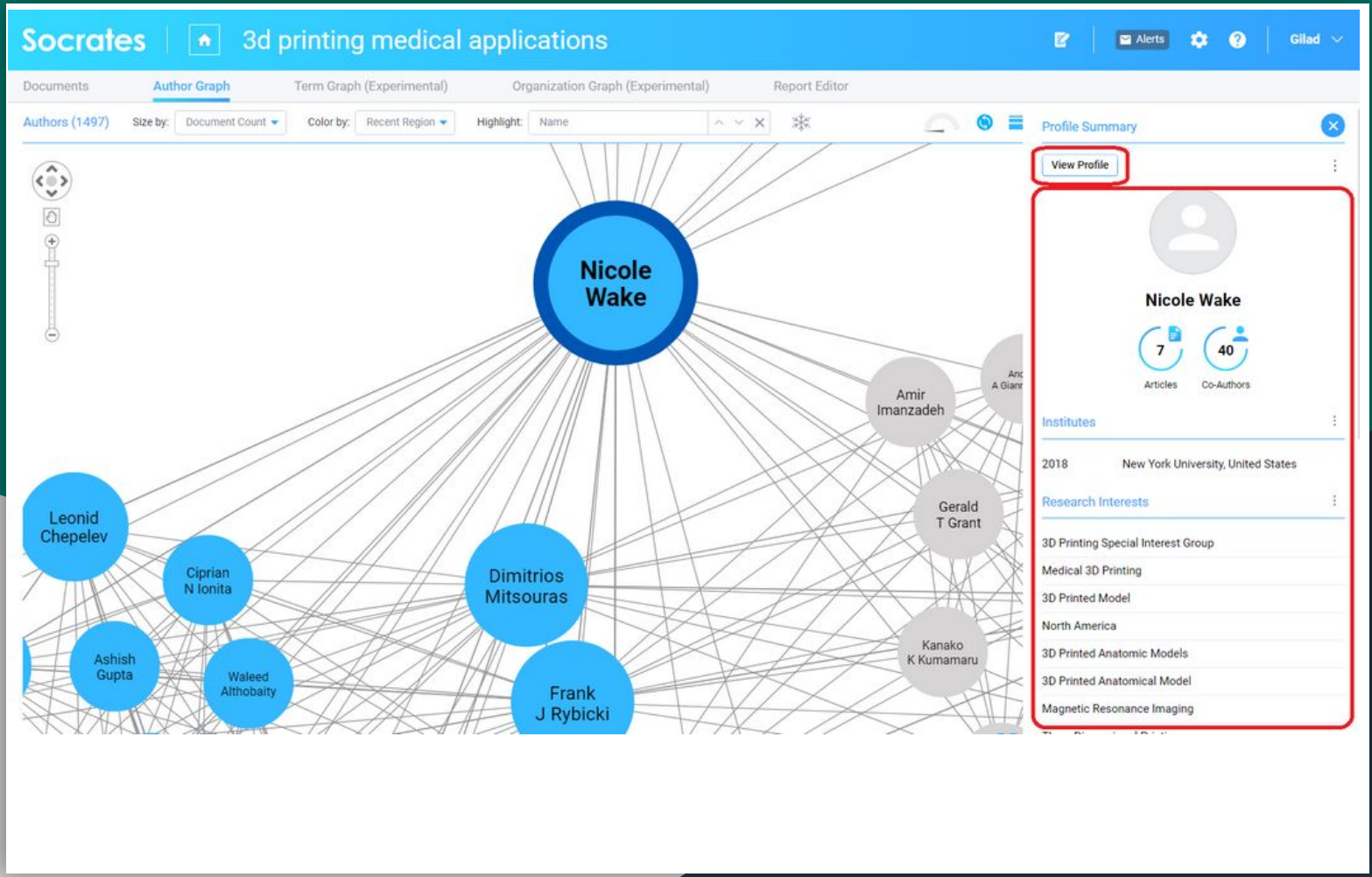
By Dina Bavli



By Dina Bavli



By Dina Bavli



By Dina Bavli

Socrates | Author Profile: Nicole Wake

Overview | Documents | Author Graph | Term Graph (Experimental)

Nicole Wake

7 Articles

40 Co-Authors

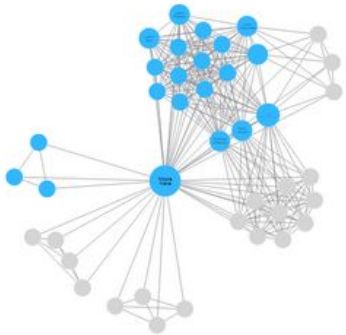
Institutes

2018 New York University, United States

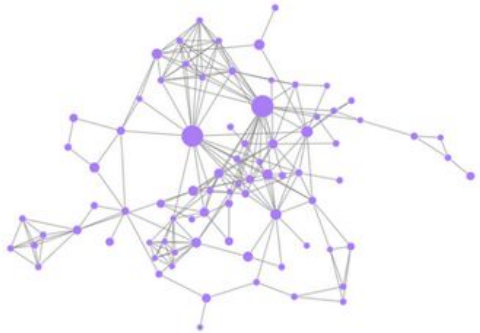
Research Interests

- 3D Printing Special Interest Group
- Medical 3D Printing
- 3D Printed Models
- North America
- 3D Printed Anatomical Model
- 3D Printing Anatomical Models
- Magnetic Resonance Imaging
- Medical Imaging Data
- Three Dimensional Printing
- 3D Printer

Author Graph



Term Graph (Experimental)



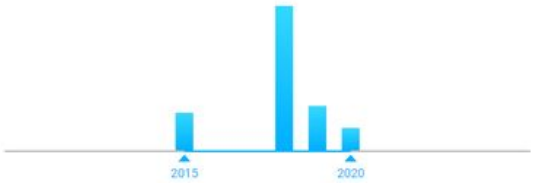
Publications

1. Principles of three-dimensional printing and clinical applications within the a...
Sarah Bastawrous, Nicole Wake, Dmitry Levin, Beth Ripley
Article 2018 | 4 citations | Abdominal Radiology

2. Radiological Society of North America (RSNA) 3D printing Special Interest Gr...
Leonid Chepelev, Nicole Wake, Justin Ryan, Waleed Althobaity, Ashish Gupta, Elsa Arribas, Lu...
Article 2018 | 34 citations | 3D Printing in Medicine

3. Creating patient-specific anatomical models for 3D printing and AR/VR: a sup...
Nicole Wake, Amy E Alexander, Andy M Christensen, Peter C Liacouras, Maureen Schickel, To...

Publications per Year

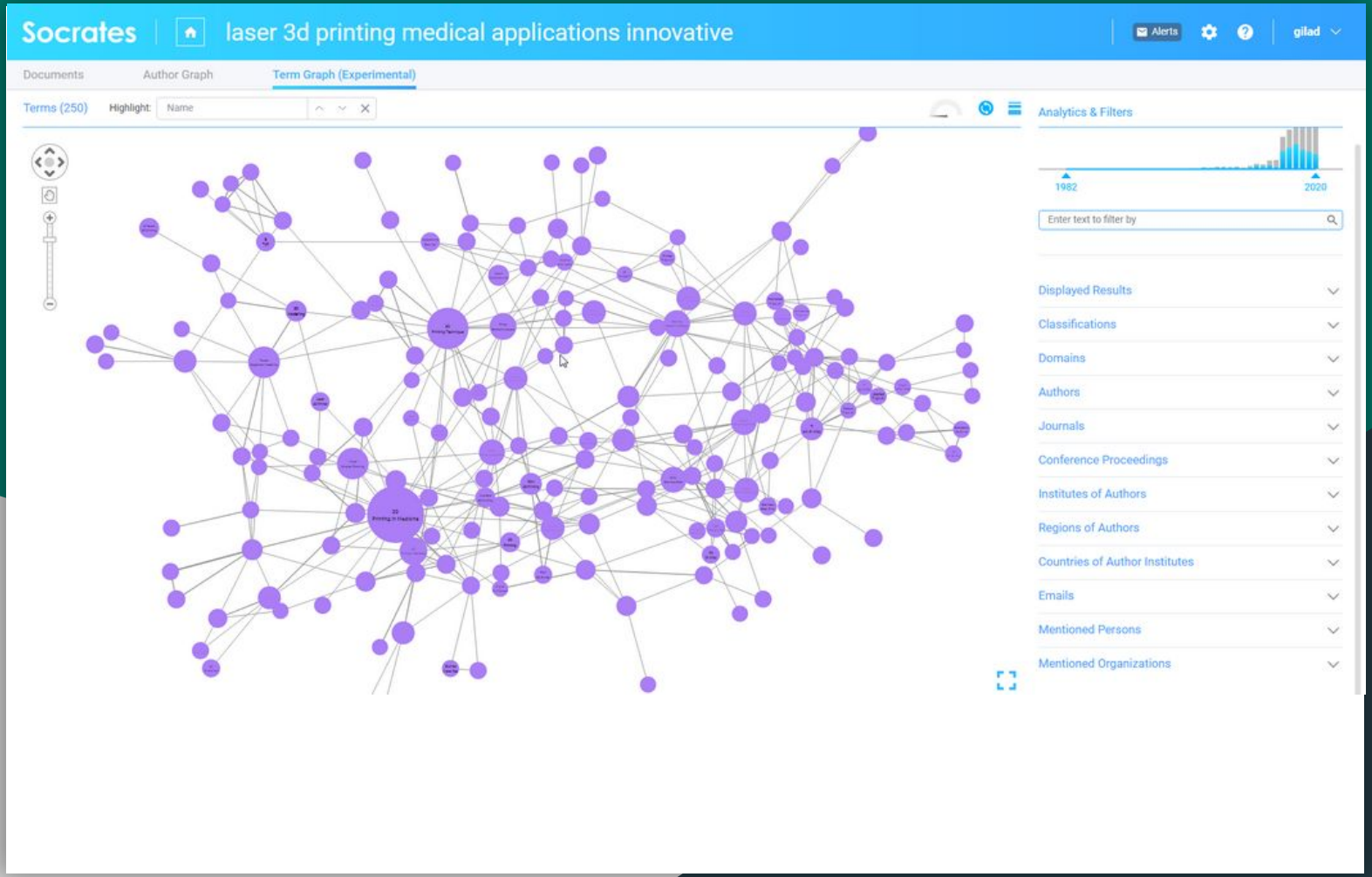


By Dina Bavli

By Dina Bavli

Need more relevant terms to your search? No problem

By Dina Bavli



By Dina Bavli

Search Definition

Add Term...

- Default Group (3)
  - 3D Printing
  - Fused Deposition Modeling
  - metal 3D printing
- laser (8)
- ! medical applications (13)
- innovation (6)
  - Disruptive technology
  - Emerging technology
  - Market disruption
  - New technology
  - Novel approach
  - State of the art

Results (1565 of 2695)

Sort by: Relevance

Article **Metallic powder-bed based 3D printing of cellular scaffolds for orthopaedic implants: A state-of-the-art review on manufacturing, topological design, mechanical properties and biocompatibility.**

Xipeng Tan, Yu Jun Tan, C S L Chow, Shu Beng Tor, Wai Yee Yeong

2017 | 160 citations | Materials Science and Engineering: C

Abstract Metallic cellular scaffold is one of the best choices for orthopaedic implants as a replacement of human body parts, which could improve life quality and increase longevity for the people needed. Unlike conventional methods of making cellular scaffolds, three-dimensional (3D) printing or additive manufacturing opens up new possibilities to fabricate those customisable intricate designs with highly interconnected pores. In the past decade, metallic powder-bed based 3D printing methods emerged and the techniques are becoming increasingly mature recently, where selective laser melting (SLM) and selective electron beam melting (SEBM) are the two representatives. Due to the advantages of good dimensional accuracy, high build resolution, clean build environment, saving materials, high customisability, etc., SLM and SEBM show huge potential in direct customisable manufacturing of metallic cellular scaffolds for orthopaedic implants. Ti-6Al-4 V to date is still considered to be the optimal materials for producing orthopaedic implants due to its best combination of biocompatibility, corrosion resistance and mechanical properties. This paper presents a state-of-the-art overview mainly on manufacturing, topological design, mechanical properties and biocompatibility of cellular Ti-6Al-4V scaffolds via SLM and SEBM methods. Current manufacturing limitations, topological shortcomings, uncertainty of biocompatible test were sufficiently discussed herein. Future perspectives and recommendations were given at the end.

Visit Website | View PDF

Electron Beam Melting | Ti 6Al 4V | Orthopaedic Implant | Selective Laser Melting | Selective Laser Sintering | 3D Printing Technique | 3D Printing Method | SLM | Mechanical Properties | Metal Powder

By Dina Bavli

If it looks like a duck

Image by pixel2013 from Pixabay NOT ducks- thus are geese

By Dina Bavli

If it looks like a duck- it's NLP

Image by pixel2013 from Pixabay NOT ducks- thus are geese

By Dina Bavli

Natural Language Processing

By Dina Bavli

TF-IDF

Term Frequency — Inverse Document Frequency

By Dina Bavli

TF, DF, IDF

TF (Term Frequency) is the frequency counter for a term t in document d.

tf(t,d) = count of t in d / number of words in d

DF (Document Frequency) is the count of occurrences of document in the document set N the term appear in.

term t IDF (Inverse Document Frequency) is the inverse of the frequency of the document which measures the informativity of term t.

By Dina Bavli

Huh?! What?!

By Dina Bavli

Word
Book
Library

By Dina Bavli

TF, DF, IDF

TF (Term Frequency) - how many times a specific **word** appears in the **book**?

tf(t,d) = how many times a specific **word** appears in the **book**? Divided with how **long** the **book** is?

DF (Document Frequency)-in how many **books** in the **library** a specific **word** appears?

IDF (Inverse Document Frequency) is the inverse of the frequency of the **book** which measures the informativity of the **word**.

By Dina Bavli

Inverse Document Frequency

idf(t) = N/df

idf(t) = All the Words in the **Library** Divided by how many **books** in the **library** a specific **word** appears?

Problem in case of a large corpus (**Library)**.

idf(t) = log(N/df)

Problem in case of a word that's not in vocabulary (cannot divide by 0)

idf(t) = log(N/(df + 1))

By Dina Bavli

TF-IDF

Term Frequency — Inverse Document Frequency

In a collection of documents, statistically measure how important a word is.

tf (Term Frequency) - how many times a specific word appears in the **book**?
Multiply by :
idf(t) = All the Words in the **Library** Divided by how many **books** in the **library** a specific **word** appears? (after taking into account a large **library** or a missing **word**)

The basic version of TF-IDF

$$tf\text{-}idf(t, d) = tf(t, d) * \log(N/(df + 1))$$

By Dina Bavli

There is a lot more to learn about NLP

By Dina Bavli

By Dina Bavli

Image by athree23 from Pixabay

https://www.linkedin.com/in/dina-bavli-502430158/ https://github.com/dinbav

By Dina Bavli