

Introduction to Big Data

All you might want to dive into



Who am I?

By Dina Bavli

Why am I here?



By Dina Bavli

Why should you be here?

Introducing the forest



Photo by [Hatham](#) on [Unsplash](#)

By Dina Bavli

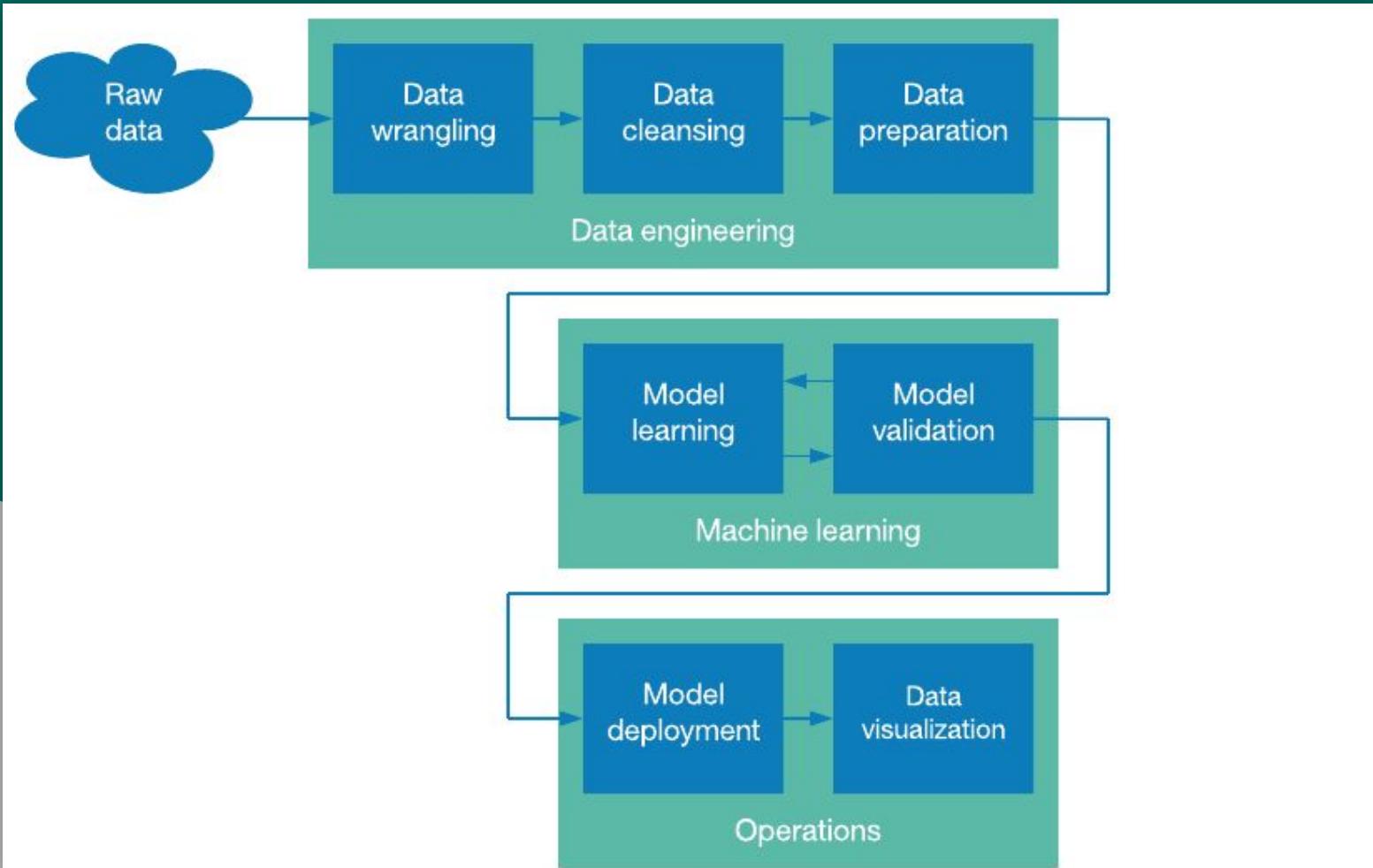
What are we doing here?

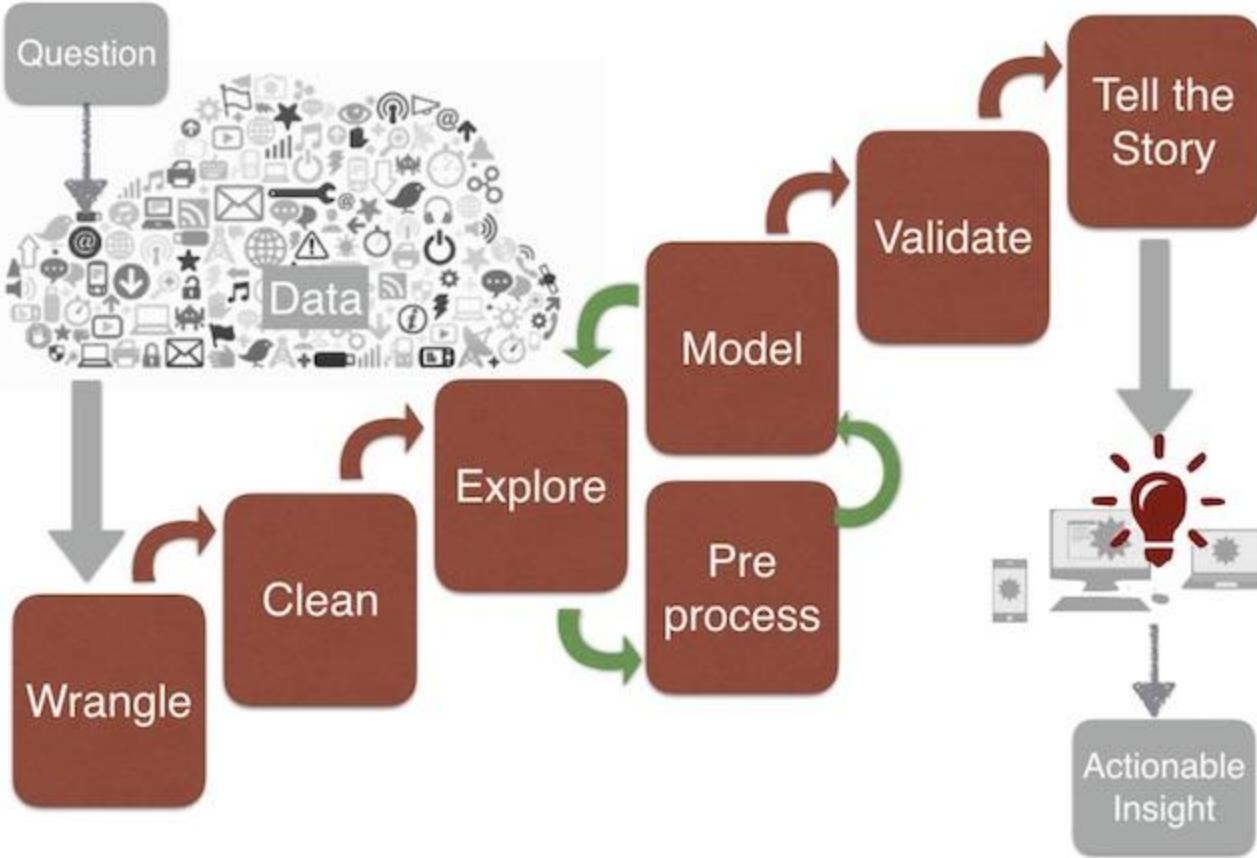
What are we going to do in the next two hours?

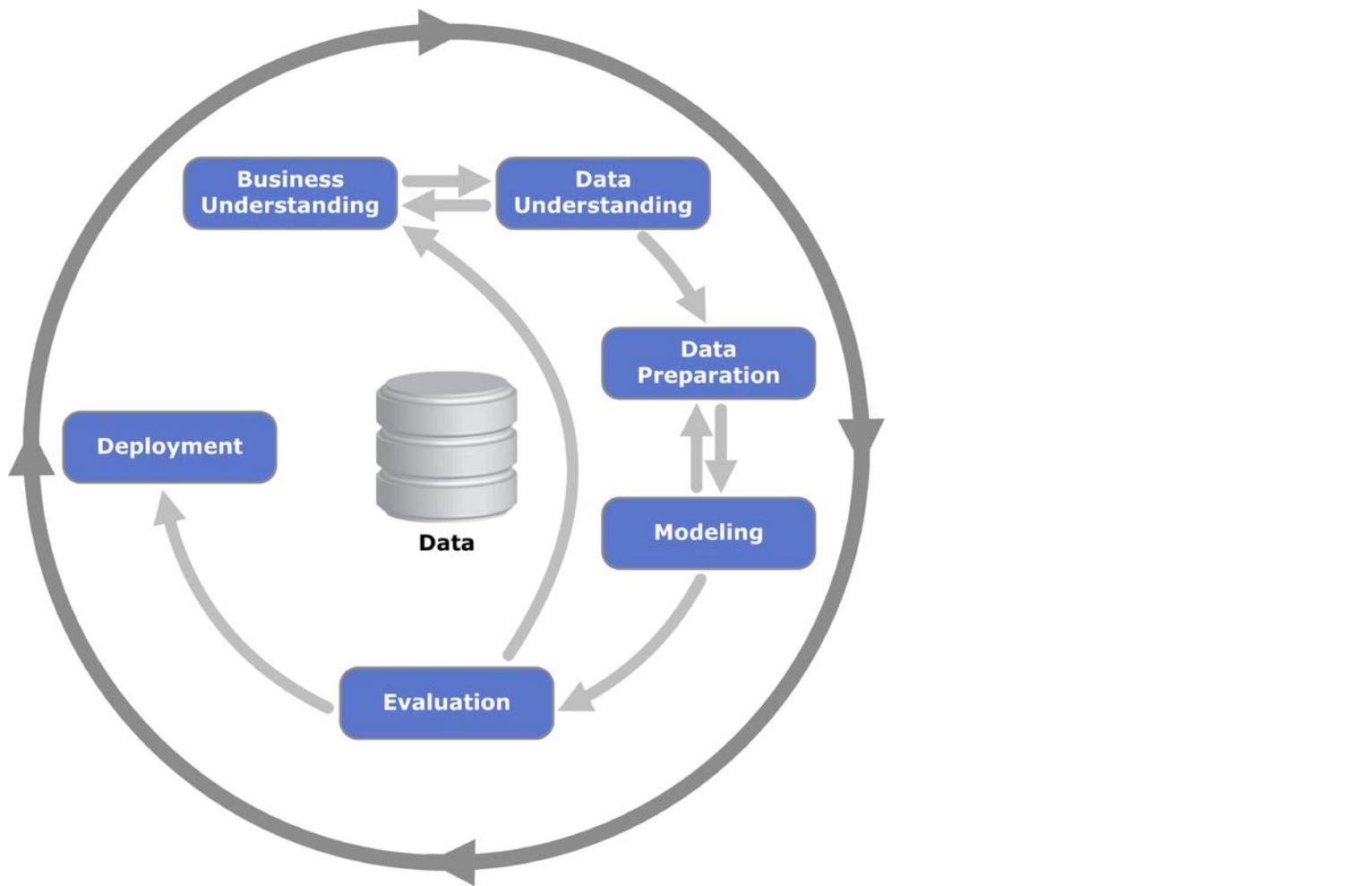
We will cover:

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.

Data science pipeline







Machine learning

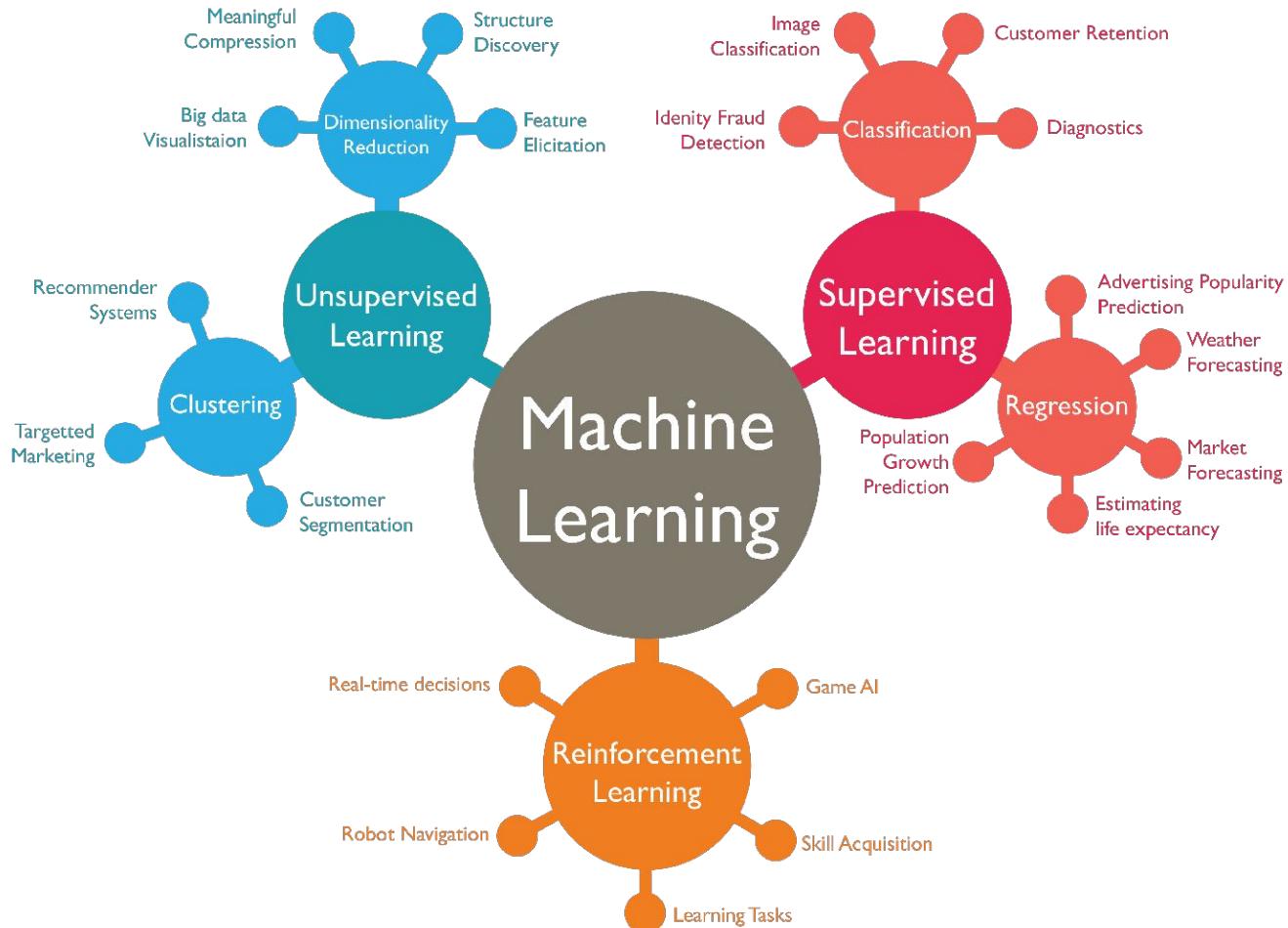
Feature Matrix (X)

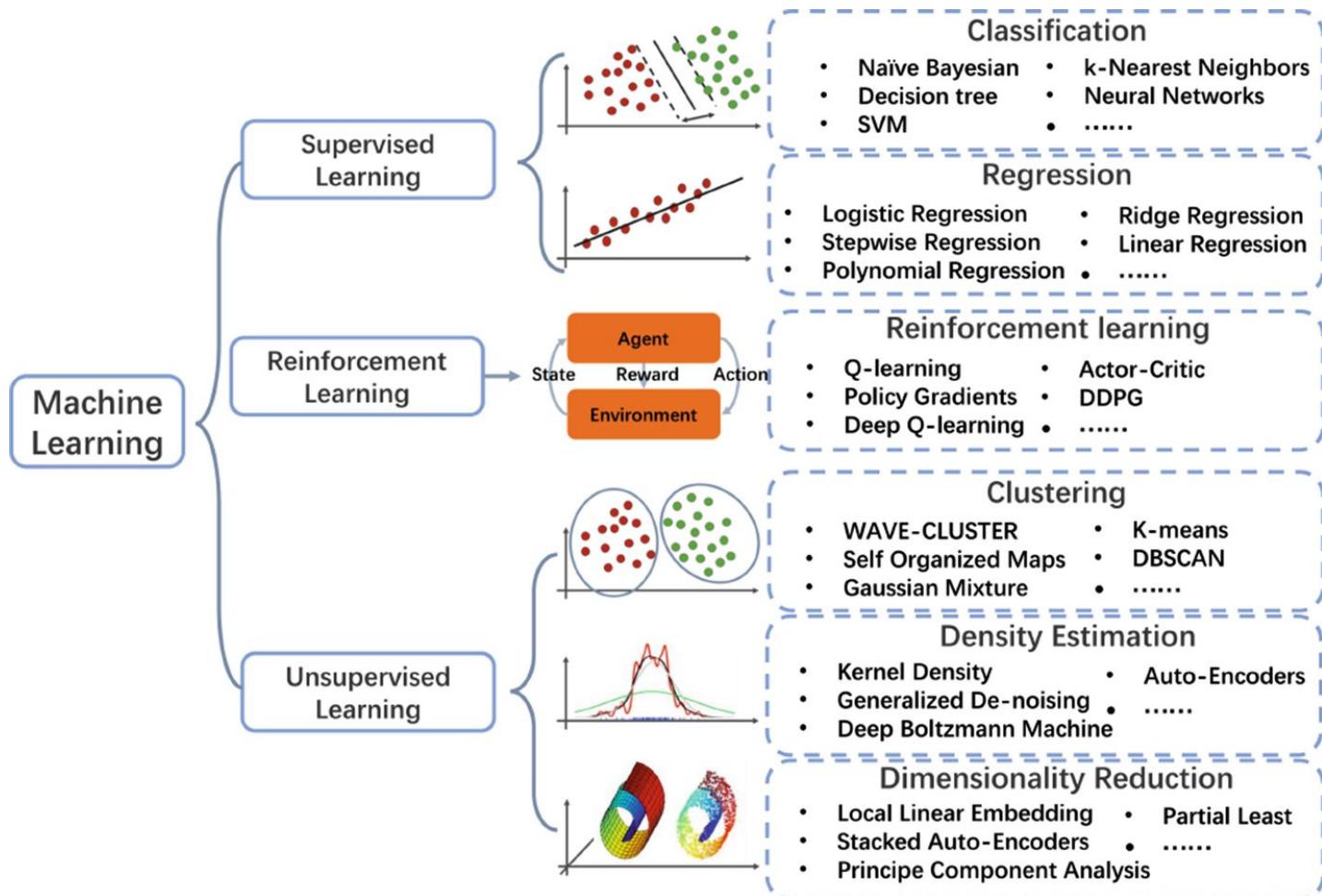
$n_features \rightarrow$

$\rightarrow n_samples$

Target Vector (y)

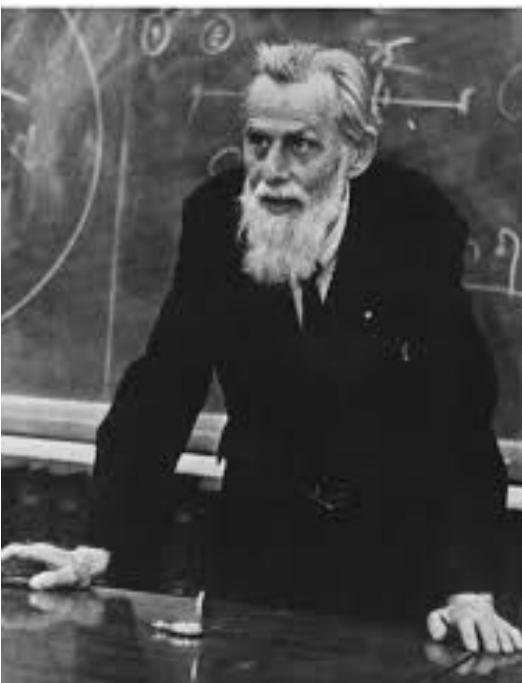
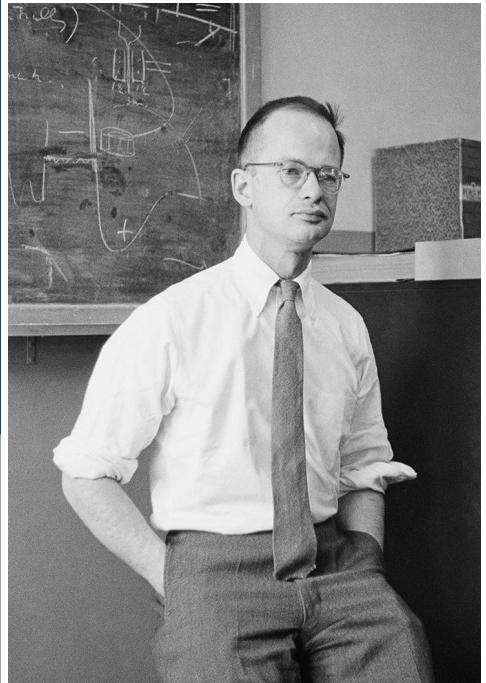
$\rightarrow n_samples$



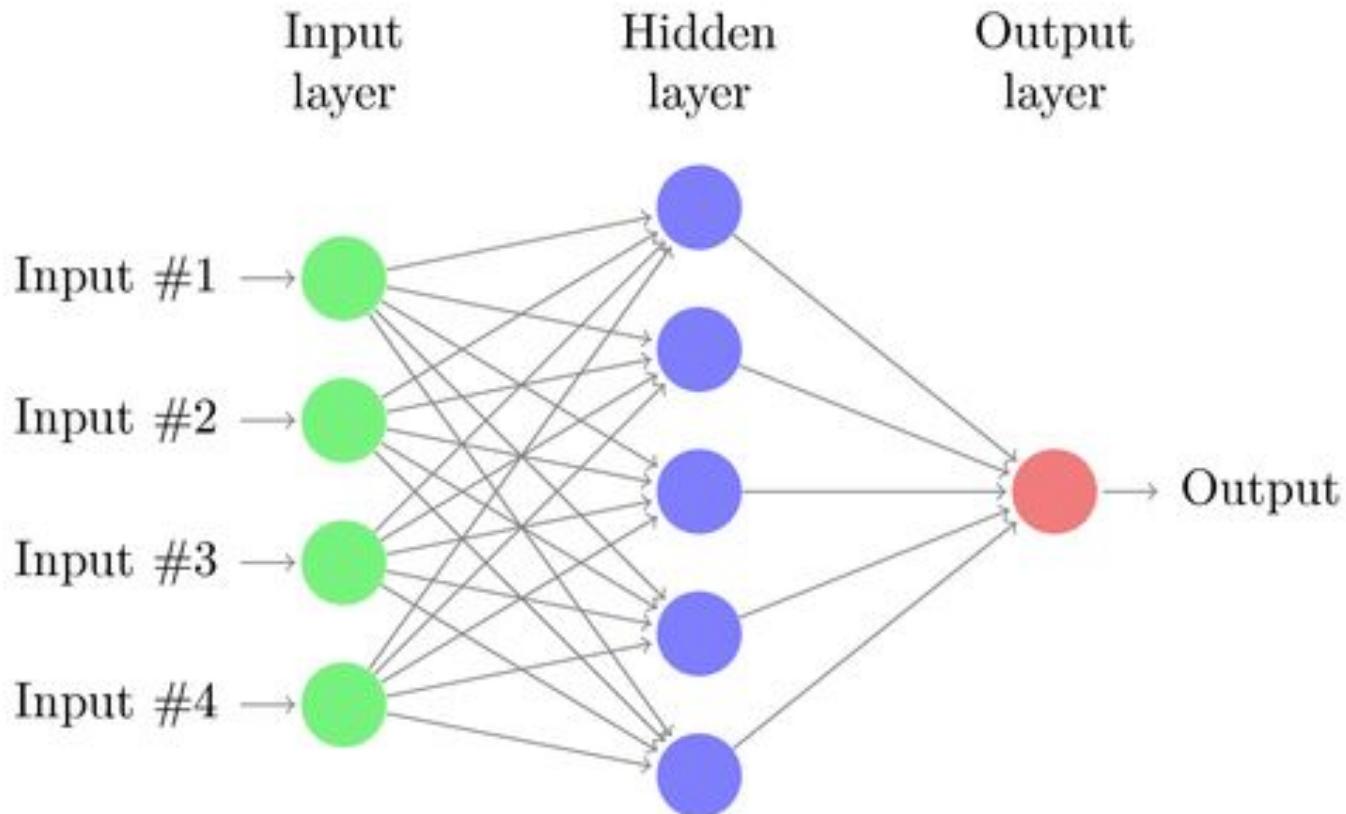


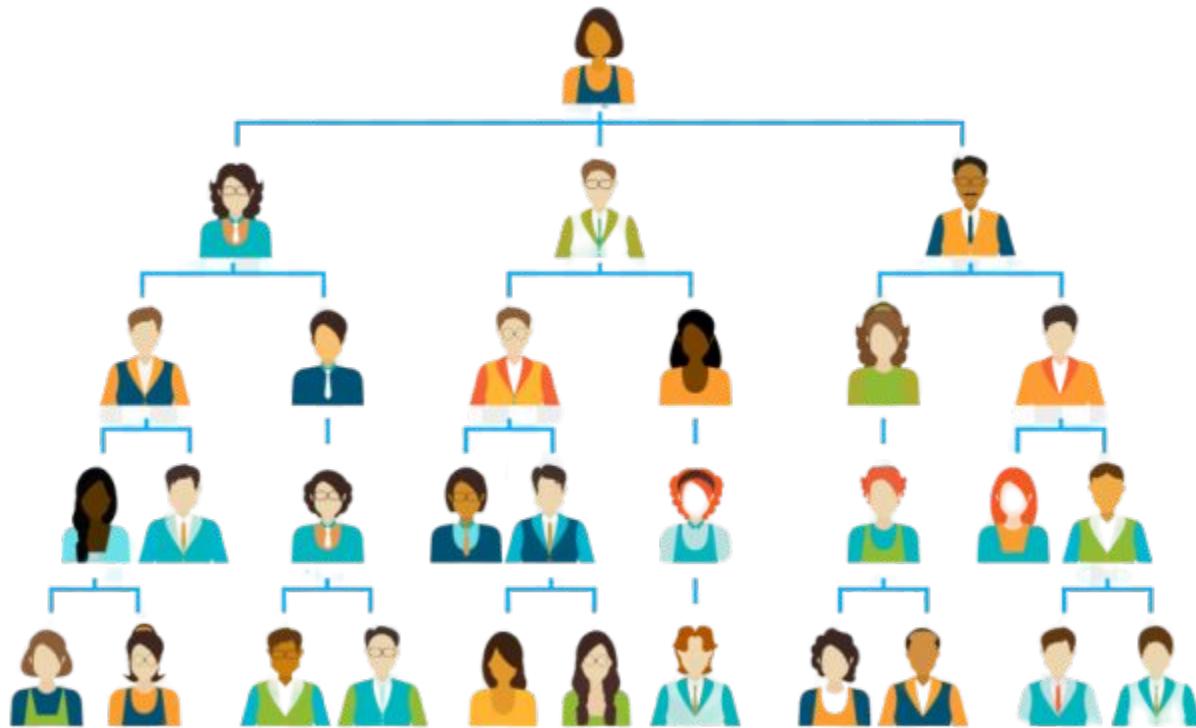
Into Deep

By Dina Bavli



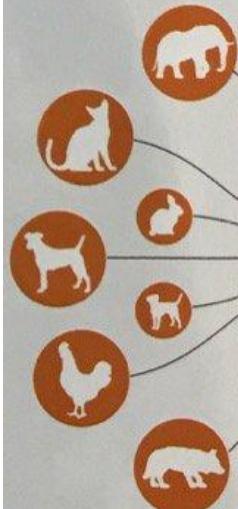
By Dina Bawli





TRAINING

During the training phase, a neural network is fed thousands of labeled images of various animals, learning to classify them.



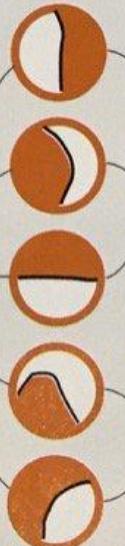
INPUT

An unlabeled image is shown to the pretrained network.



FIRST LAYER

The neurons respond to different simple shapes, like edges.



HIGHER LAYER

Neurons respond to more complex structures.



TOP LAYER

Neurons respond to highly complex, abstract concepts that we would identify as different animals.



OUTPUT

The network predicts what the object most likely is, based on its training.





Regular Pitbull-Mix Face: 0.8



Left Eye: 0.8



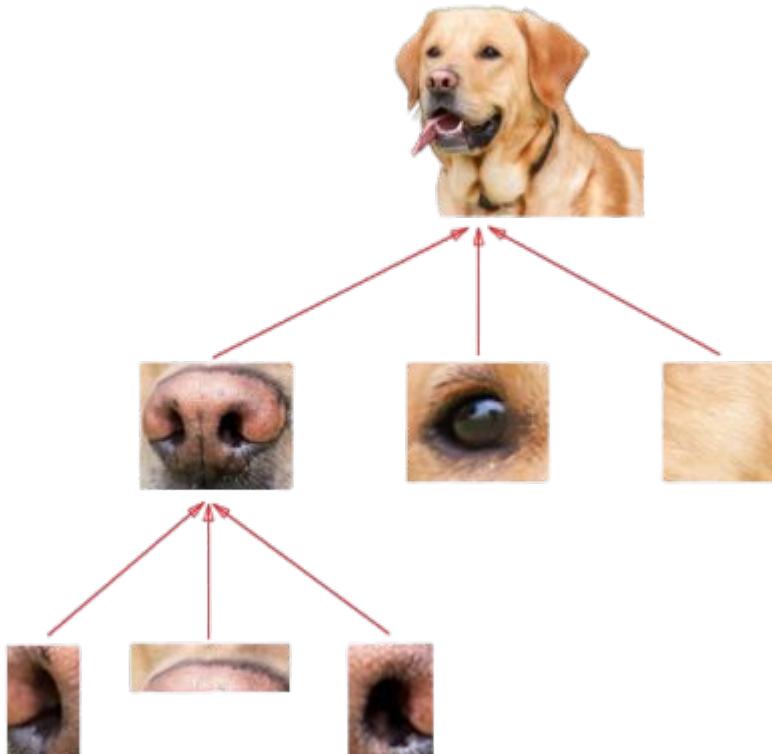
Right Eye: 0.7

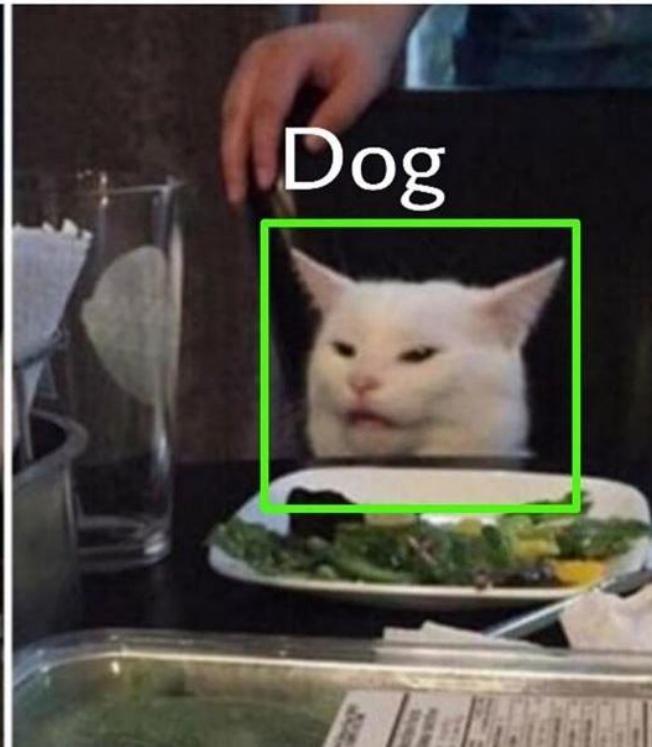


Snout: 0.9



Mouth: 0.8





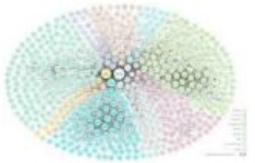
By Dina Bawli

Smart Labeling

By Dina Bavli

Passive Machine Learning

Raw, unlabeled data



Oracle



Machine
Learning Model



Unlabeled data
passed to oracle

Labelled Data

Trained Classifier

Budget or Time
Constraints
related to Labeling



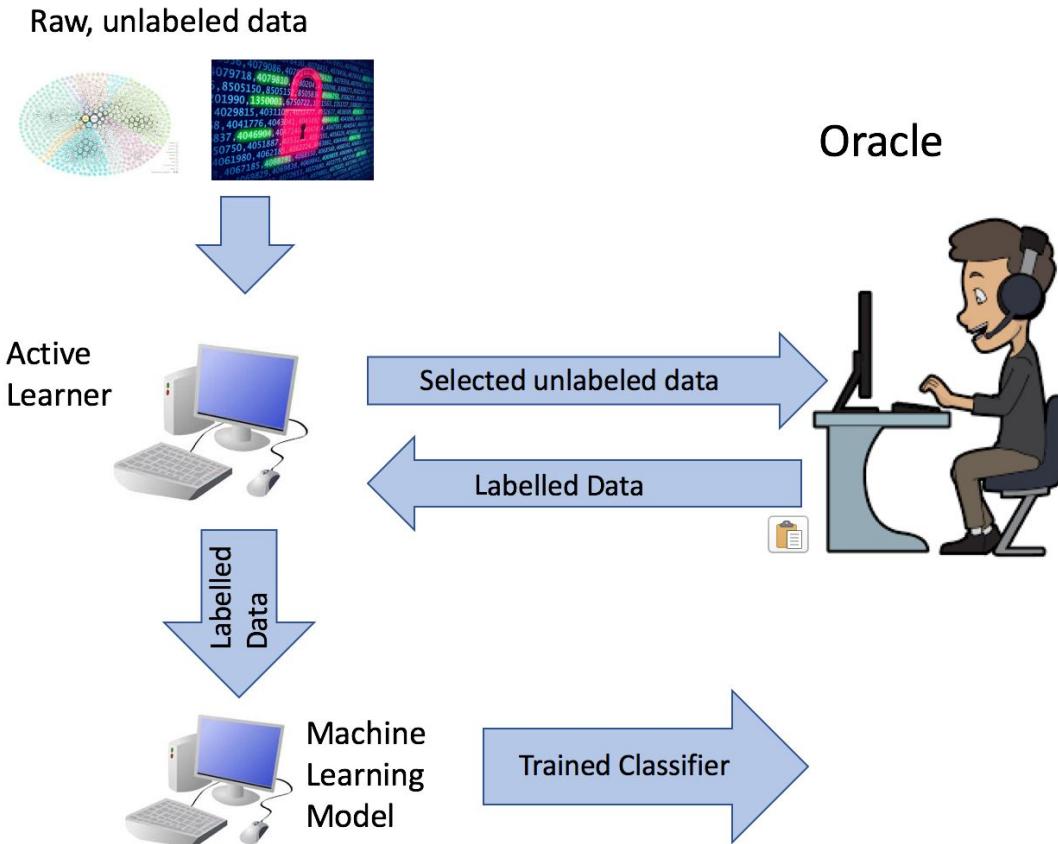
Label Faster/Cheaper:

*Use a Machine Learning +
Human-in-the-Loop
approach to streamline
and speed up the labeling
process*

Label Smarter:

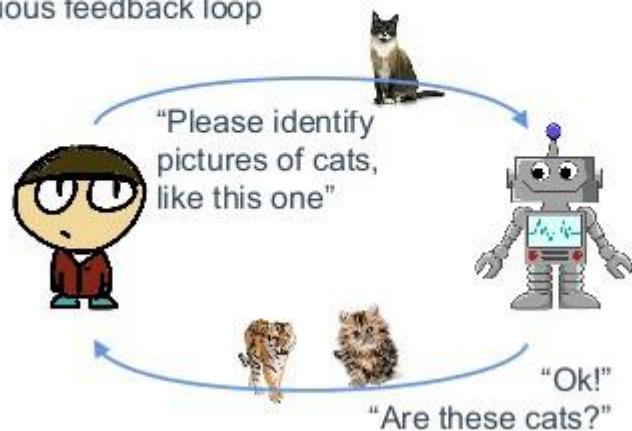
*Select the most
informative rows to label in
order to optimize the
information : data volume
ratio*

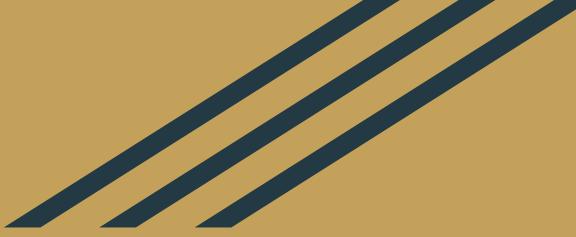
Active Machine Learning



Active Learning

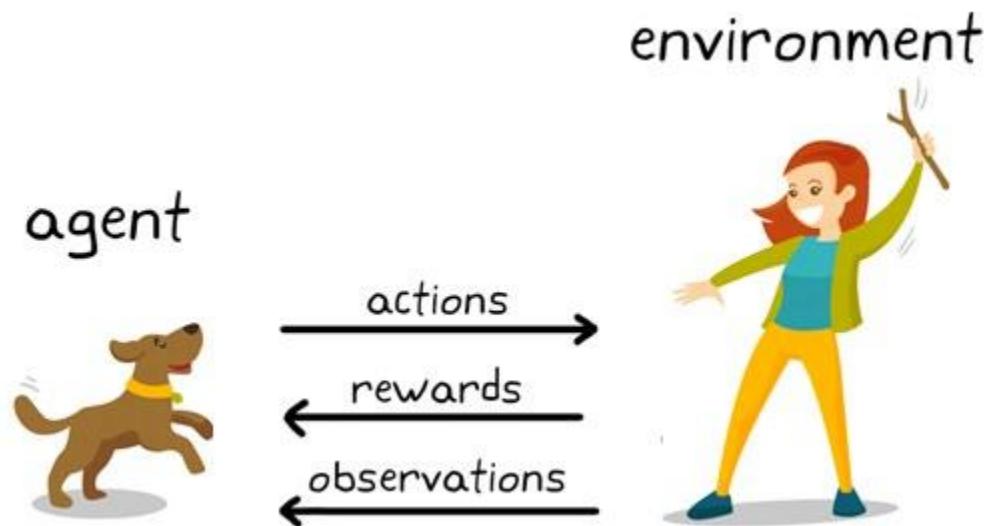
Selecting the optimal data to manually label for Machine Learning
Often a continuous feedback loop

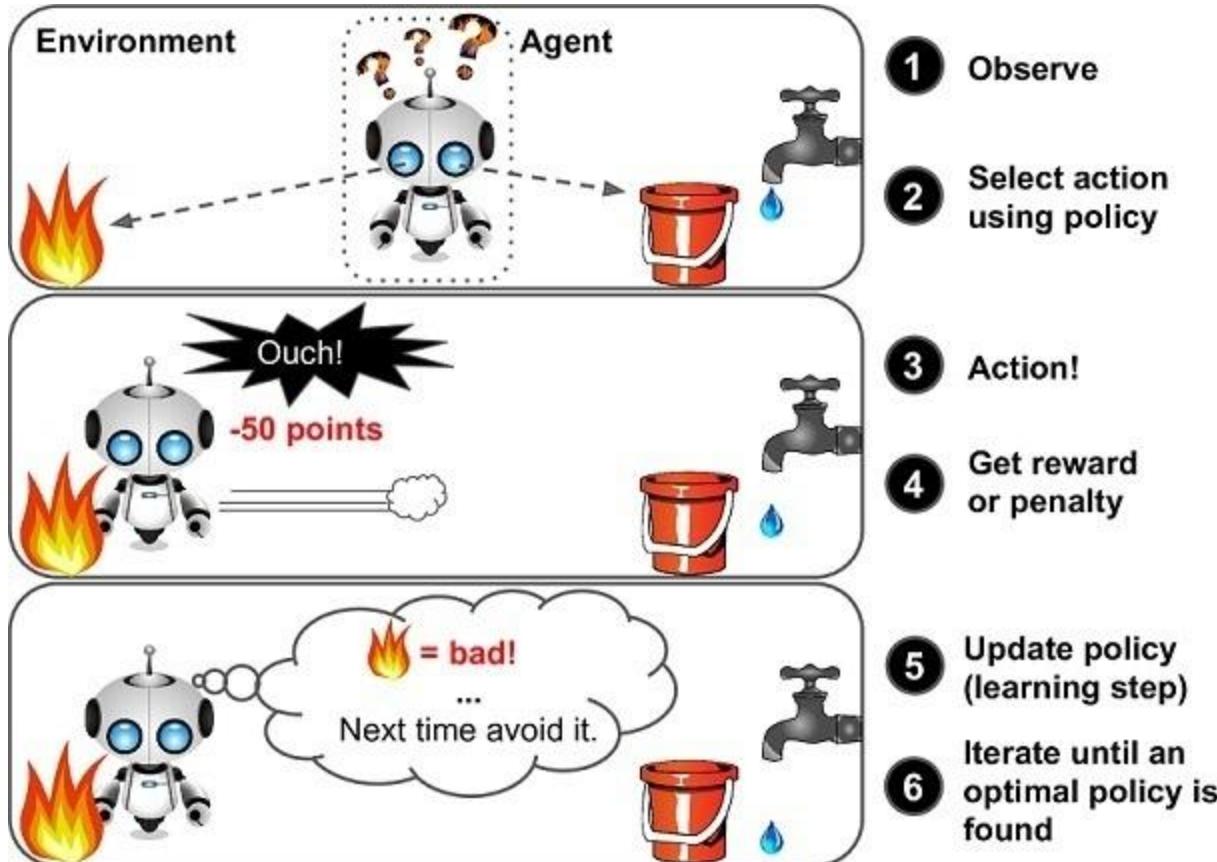




No Labels- No Data

Reinforcement Learning





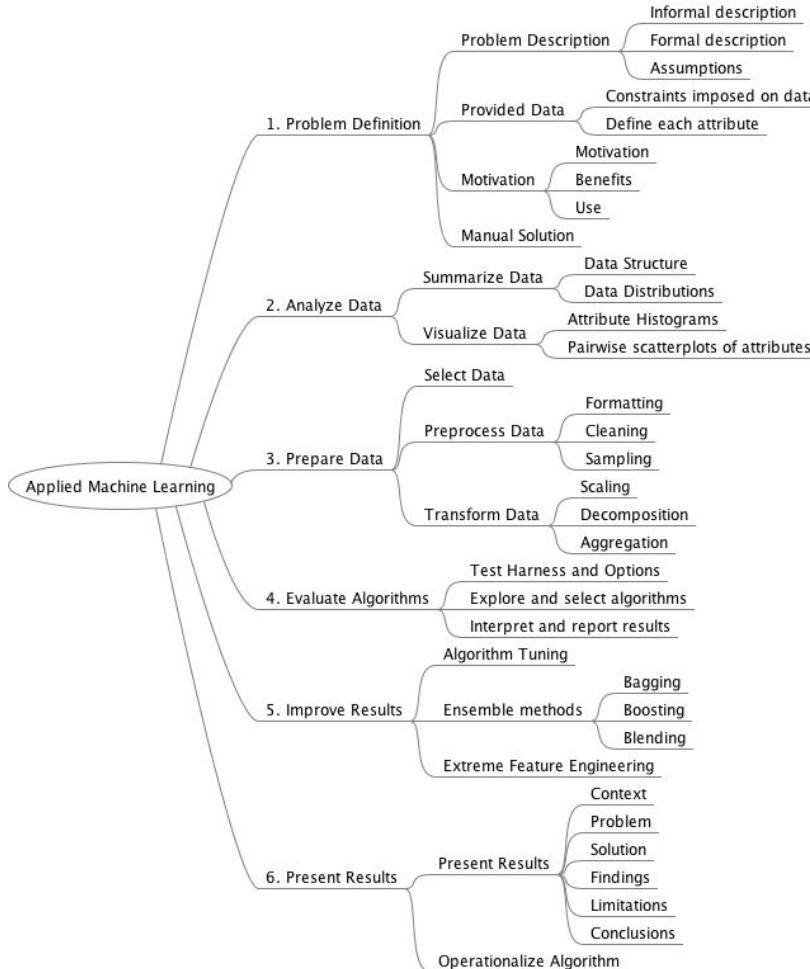


By Dina Bavli



By Dina Bavli

Summary of Data Science Pipeline



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 Bawli

EDA-Data Exploration

Why is EDA important?



GIGO- Garbage In, Garbage Out

Your analysis is as good as your data.

Exploratory Data Analysis does two main things:

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

Exploratory Data Analysis methods:

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

Commonly used plots for EDA:

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

Cleaning



jarmoluk

By Dina Bavli

What to clean?

Missing

Outliers

Features

Wrong cleaning

- Create bias
- Tilt the distribution
- Dismorf the data

Right cleaning

- Reduce costs
- Reduce process time
- Reduce overfitting

Features

Health insurance



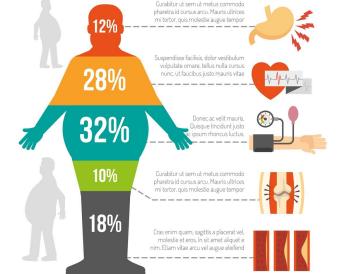
created by pch.vector - www.freepik.com

OBESITY INFOGRAPHICS

HARMFUL COMPONENTS OF FASTFOOD



HEALTH PROBLEMS CAUSED BY OBESITY



COMPARING OF LIFE AN OBESE MAN AND ATHLETE

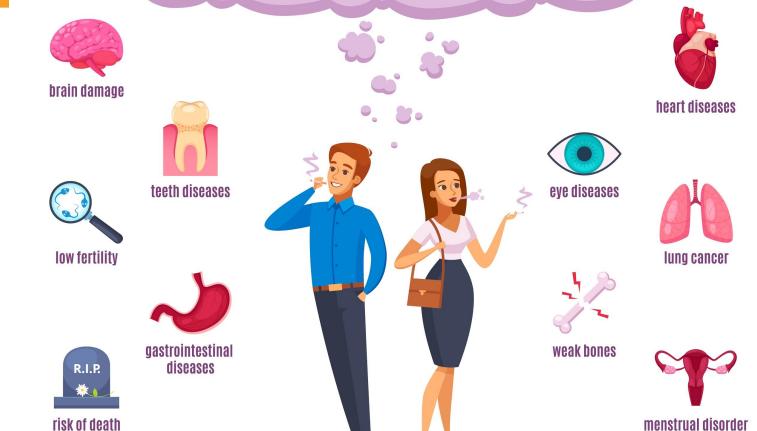


OVERWEIGHT AND OBESITY



Designed by macrovector / Freepik

DANGER OF SMOKING



WHAT CAN INCREASE THE DAMAGE



Designed by macrovector_official / Freepik

By Dina Bawli

Insurance equation= smoking + obesity + age^2

OBESITY INFOGRAPHICS



- LOREM IPSUM
CONSETETUR
ALIQUAM





12% Silueta: Problemas de salud mentales y emocionales.

28% Corazón: Problemas de salud mentales y emocionales.

32% Riñón: Problemas de salud mentales y emocionales.

10% Silueta: Problemas de salud mentales y emocionales.

18% Riñón: Problemas de salud mentales y emocionales.

Silueta: Problemas de salud mentales y emocionales.

Corazón: Problemas de salud mentales y emocionales.

Riñón: Problemas de salud mentales y emocionales.



DANGER OF SMOKING



WHAT CAN INCREASE THE DAMAGE



Designed by macrovector / Freepik

Designed by macrovector official / Freepik

By Dina Bawli

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 **Age**, **Gender**, **Religion**, and **Politics**. 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>

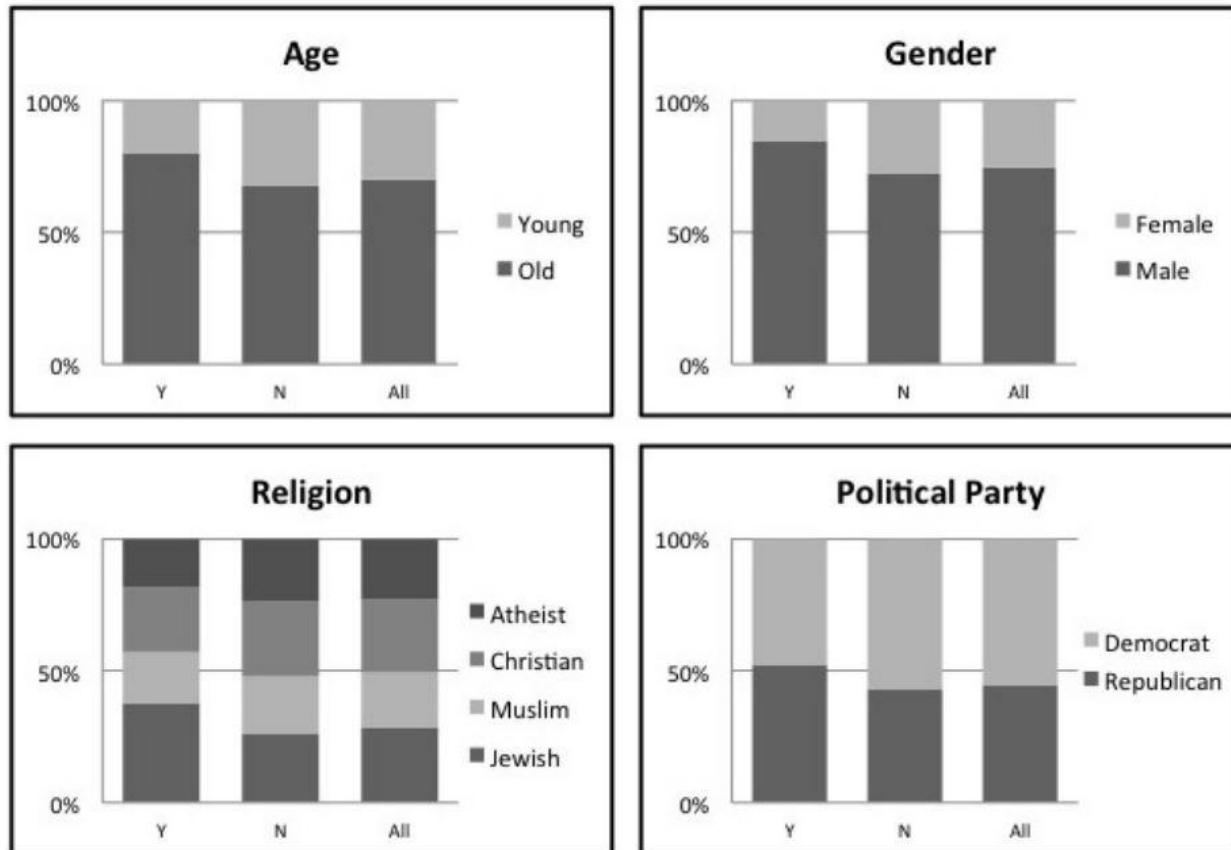


Figure 1: Breakdown of the binary features by influence (Y/N) and overall (All) in the training set.

Gender and Political Party

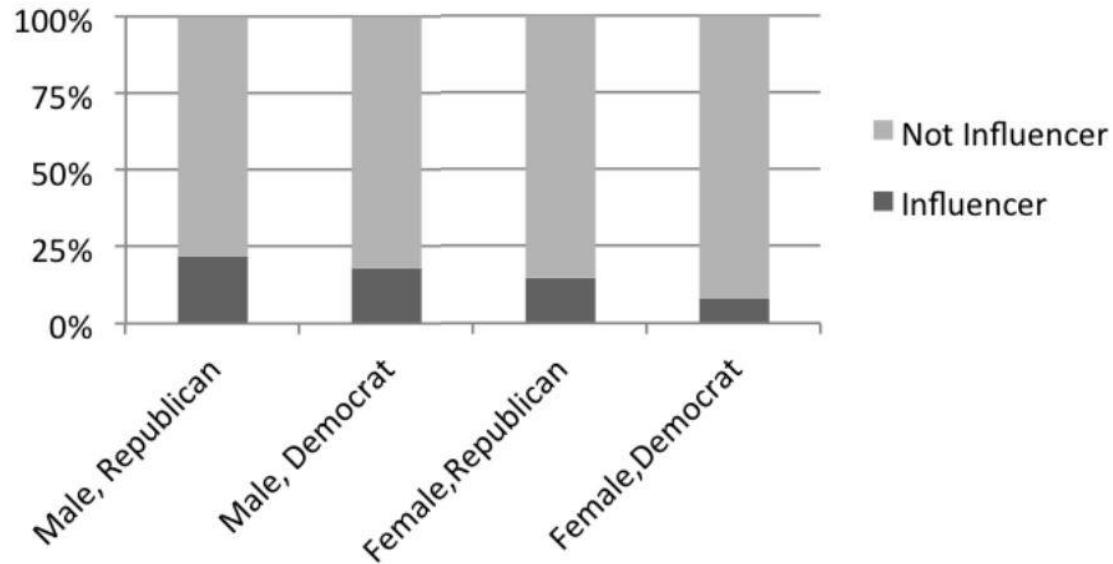
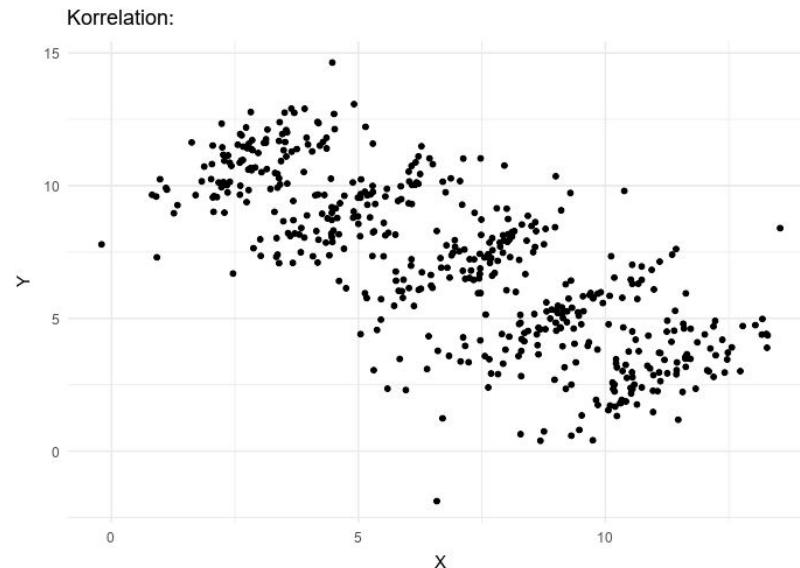


Figure 3: The breakdown of influencers and non-influencers in the training data based on the binary combination feature of gender and political party.

Simpson's paradox

Need to take into account when combining features



Pace~svwiki / CC BY-SA

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

Homophily, Social Proof, Echo Chambers and Filter Bubbles

Influence by Topic

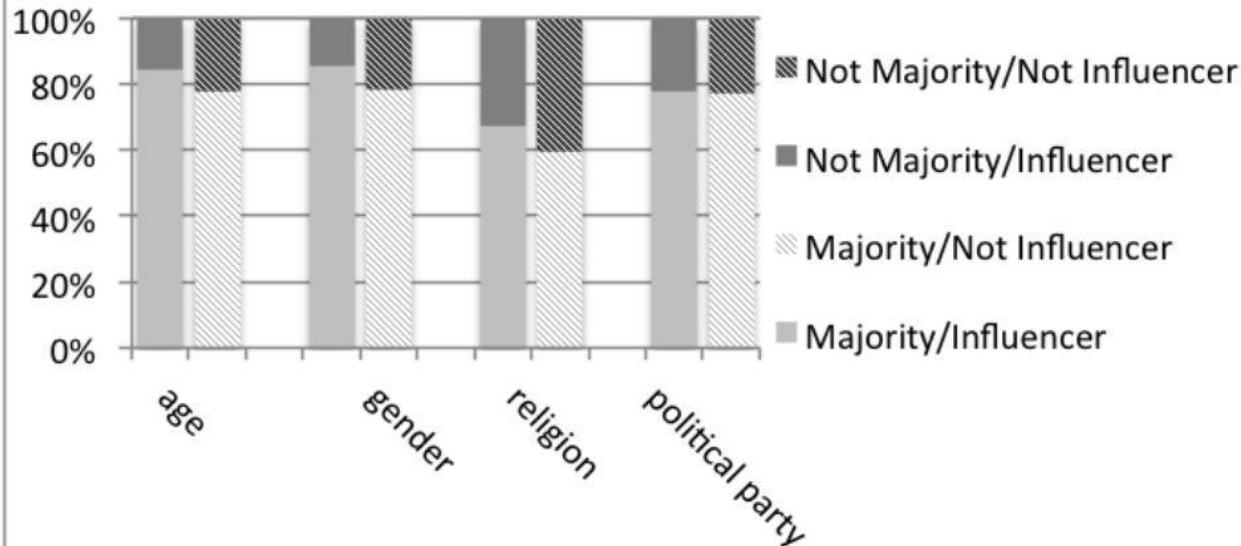
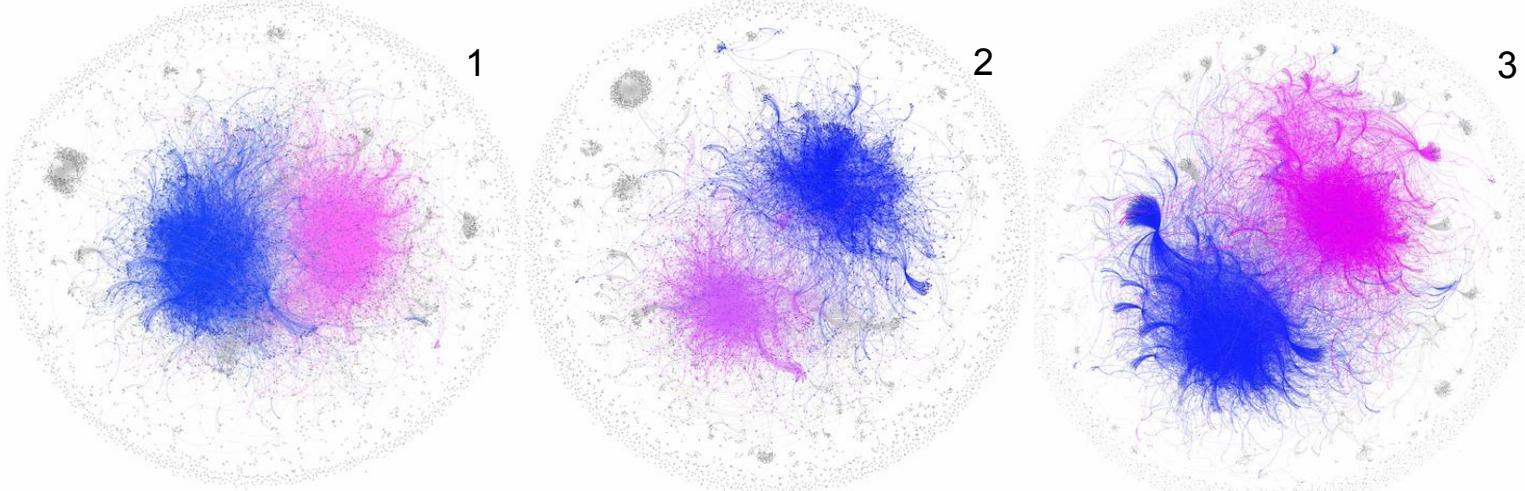


Figure 2: The breakdown of the users being in the majority within their document for each author trait with topic being taken into account.

Identifying Types of Influencers using SNA



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>

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

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

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>

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.

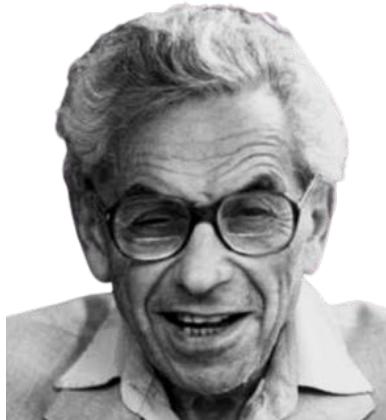
SNA- can be applied to security

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

What do those have in common?

And how does it connect to SNA?





By Dina Bawli

6°

Degrees of separation



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/>



Image by [Free-Photos](#) from [Pixabay](#)

By Dina Bawli



Documents

Author Graph

Term Graph (Experimental)

Search Definition

Add Term...

Default Group (3)

3D Printing

Fused Deposition Modeling

metal 3D printing

laser (8)

EBM

electron beam melting

laser

Laser 3D Printer

selective laser melting

Selective Laser Sintering

SLM

SLS

medical applications (13)

3D Printed Anatomical Model

3D Printed Implants

3D Printing In Healthcare

3D Printing In Medicine

3D printing medicine

3D Printing Within Healthcare

biomaterials

bioprinting

Implants

Medical 3D Printing

orthopaedic implants

Results (2695)

Sort by: Relevance

Add relevant terms for better results

- Direct Metal Laser Sintering
- Ti 6Al 4V
- Metal 3D Printing Technology
- Medical 3D Printing Market
- Fused Deposition Modelling
- Metal 3D Printer

Next

Article Emerging Technology and Applications of 3D Printing in the Medical Field.

Neil J Mardis

2018 | 6 citations | Missouri Medicine

3D printing technology evolved in the 1980s, but has made great strides in the last decade from both a cost and accessibility standpoint. While most printers are employed for commercial uses, medical 3D printing is a growing application which serves to aid physicians in the diagnosis, therapeutic planning...

Article 3D printing technologies

Dimitrios Mitsouras, Peter C Liacouras

2017 | 14 citations | 3D Printing in Medicine

Article Three-Dimensional Printing: The Future of Digital Dentistry World ...?? A Brief Description

Nandebam Premita

2018 | Journal of Orofacial & Health Sciences

Article Shaping the future: recent advances of 3D printing in drug delivery and healthcare

Sarah J Trenfield, Athereer Awad, Christine M Madia, Grace B Hatton, Jack Firth, Alvaro Goyanes, Simo...

Analytics & Filters

Relevance per Year



Enter text to filter by

Displayed Results

Classifications

Domains

Authors

Journals

Conference Proceedings

Institutes of Authors

Regions of Authors

Countries of Author Institutes

Emails

Mentioned Persons



Documents

Author Graph

Term Graph (Experimental)

Authors (2430)

Size by: Document Count

Color by: Recent Region

Highlight: Name



Analytics & Filters

Conference Proceedings

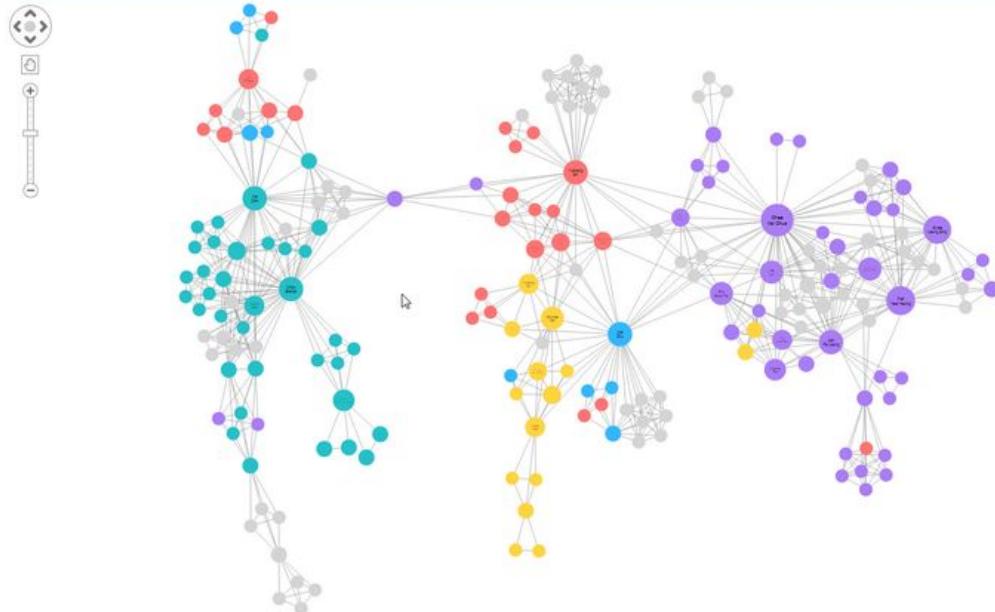
Institutes of Authors

Regions of Authors

Type to filter...

 Europe (184 / 330) North America (145 / 281) Asia (80 / 132) China (60 / 100) Unknown (57 / 126) Oceania (42 / 47) Middle East (13 / 28) Russia (3 / 5) Africa (3 / 8) Central & South America (6 / 11)

Countries of Author Institutes





Documents

Author Graph

Term Graph (Experimental)

Authors (384)

Size by: Score

Color by: Recent Region

Highlight: Name



Analytics & Filters





Documents

Author Graph

Term Graph (Experimental)

Organization Graph (Experimental)

Report Editor

Authors (1497)

Size by: Document Count

Color by: Recent Region

Highlight: Name



Profile Summary

Leonid
ChepelevCiprian
N IonitaAshish
GuptaWaleed
AlthobaityDimitrios
MitsourasFrank
J RybickiNicole
WakeAmir
ImanzadehGerald
T GrantKanako
K Kumamaru

View Profile



Nicole Wake



Institutes

2018 New York University, United States

Research Interests

3D Printing Special Interest Group

Medical 3D Printing

3D Printed Model

North America

3D Printed Anatomic Models

3D Printed Anatomical Model

Magnetic Resonance Imaging

Socrates | Author Profile: Nicole Wake

[Overview](#)[Documents](#)[Author Graph](#)[Term Graph \(Experimental\)](#)**Nicole Wake**[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

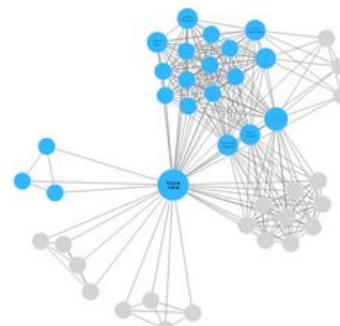
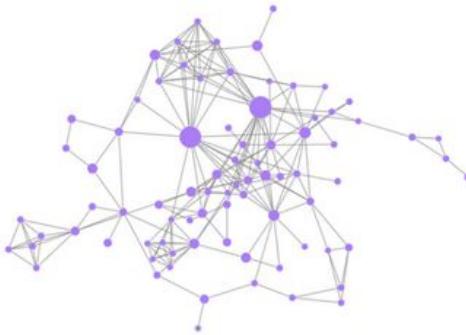
3D Printing Anatomical Models

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

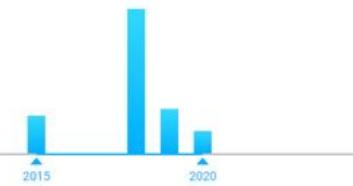
2. Radiological Society of North America (RSNA) 3D printing Special Interest Gr...

Leonid Chepelev, Nicole Wake, Justin Ryan, Waleed Althobaiti, 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, Andv M Christensen, Peter C Liacouras, Maureen Schickel, To...

[Publications per Year](#)*By Dina Bawli*

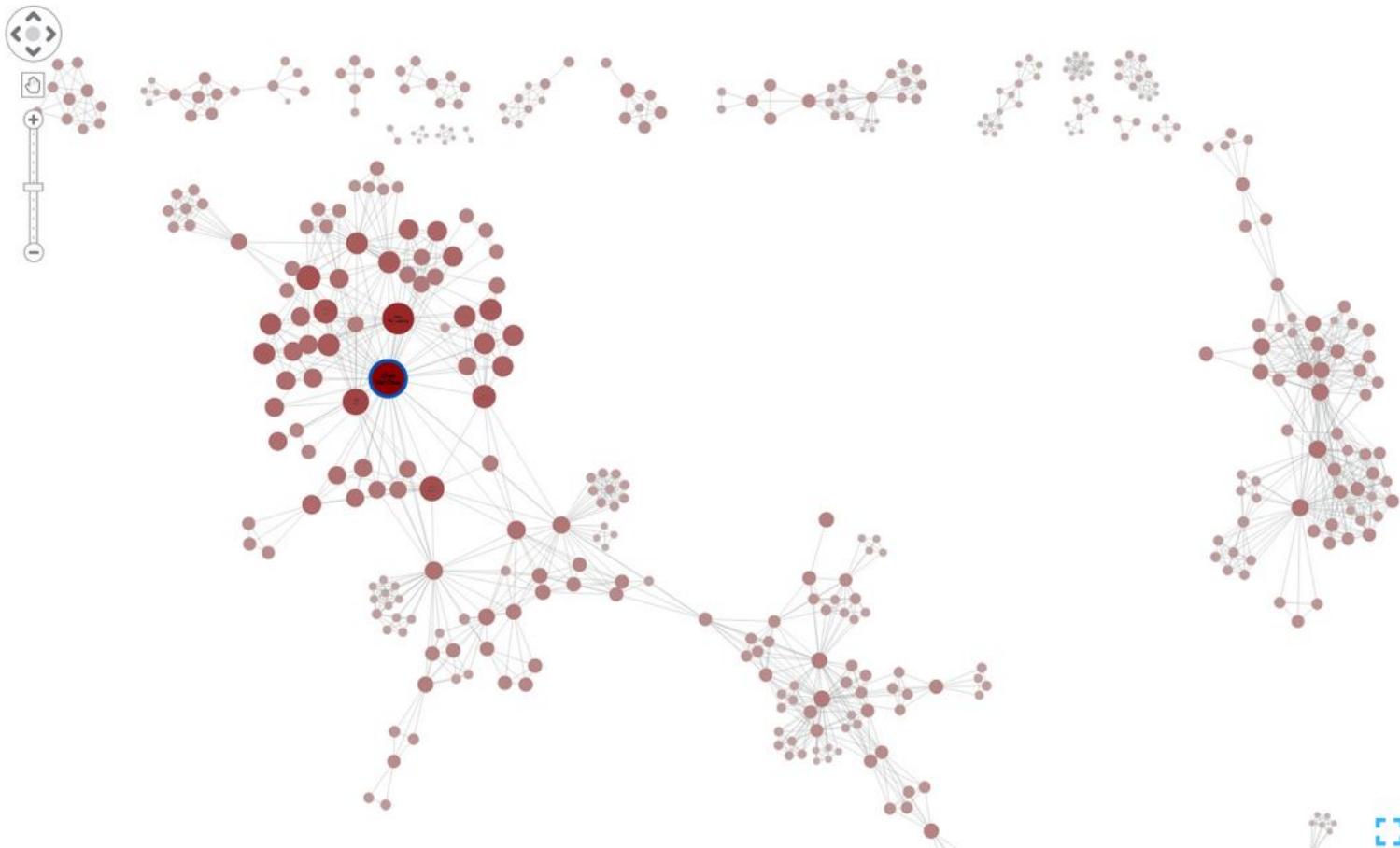
Authors (2439)

Size by: Similarity

Color by: Similarity

Highlight: wake

1 ⌂ ⌄ ⌁ ⌃



By Dina Bavli



Need more relevant terms to
your search? No problem



Documents

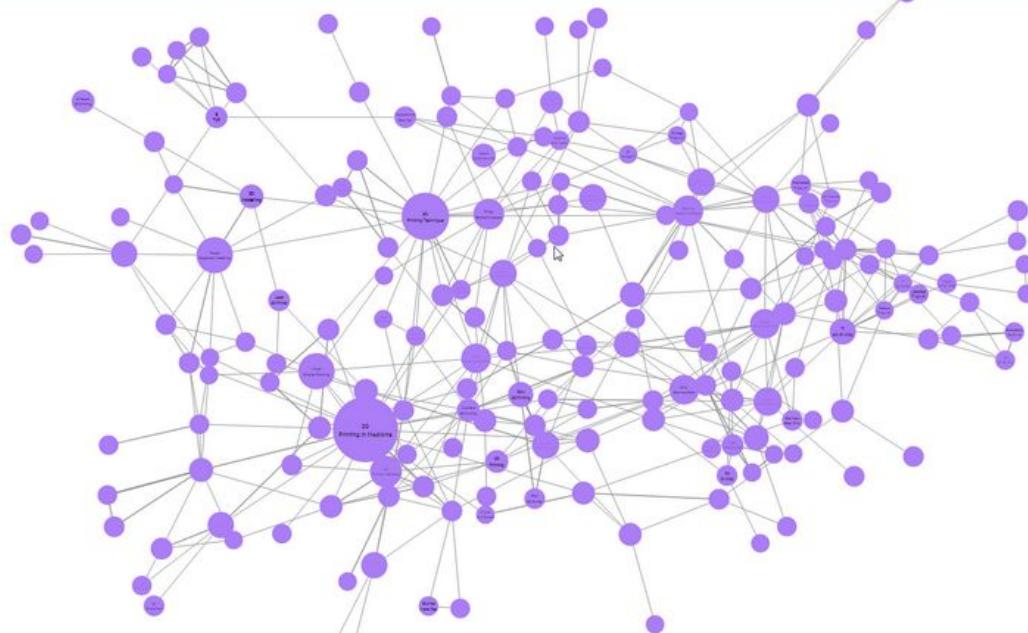
Author Graph

Term Graph (Experimental)

Terms (250)

Highlight:

Name



Analytics & Filters



1982

2020

Enter text to filter by

Displayed Results

Classifications

Domains

Authors

Journals

Conference Proceedings

Institutes of Authors

Regions of Authors

Countries of Author Institutes

Emails

Mentioned Persons

Mentioned Organizations

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](#)

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

TF-IDF

Term Frequency — Inverse Document Frequency

TF, DF, IDF

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

$$tf(t,d) = \text{count of } t \text{ in } d / \text{number of words in } d$$

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.

Huh?! What?!

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**.

Inverse Document Frequency

$$\text{idf}(t) = N/\text{df}$$

$\text{idf}(t)$ = 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**).

$$\text{idf}(t) = \log(N/\text{df})$$

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

$$\text{idf}(t) = \log(N/(\text{df} + 1))$$

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

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

There is a lot more to learn about
NLP

But that is another story and
shall be told another time.

Michael Ende

“ quotefancy

By Dina Bavli



Image by [athree23](#) from [Pixabay](#)



<https://www.linkedin.com/in/dina-bavli-502430158/>



<https://github.com/dinbav>

By Dina Bavli