2018 SKILLS BUILDING PROGRAM

WORLD BANK GROUP
Health, Nutrition & Population

BIG DATA, ARTIFICIAL INTELLIGENCE AND DECISION SCIENCE IN HEALTH AND NUTRITION

Big Data/Machine Learning Methods for Health-related Policy Questions

Introduction to Key Methods

In partnership with

datasciencedojo — data science for everyone

Burnet Institute — Medical Research. Practical Action.

UCL

techemergence

symphony

Wadhwani AI — AI FOR SOCIAL GOOD

THE UNIVERSITY OF SYDNEY

saMRC — advancinglife

Machine Intelligence Institute of Africa — *Transform Africa through Machine Intelligence*

UNIVERSITY OF COPENHAGEN

iDSI — Better decisions. Better health.

FPD

GHCC — Global Health Cost Consortium

UNDP

BILL & MELINDA GATES foundation

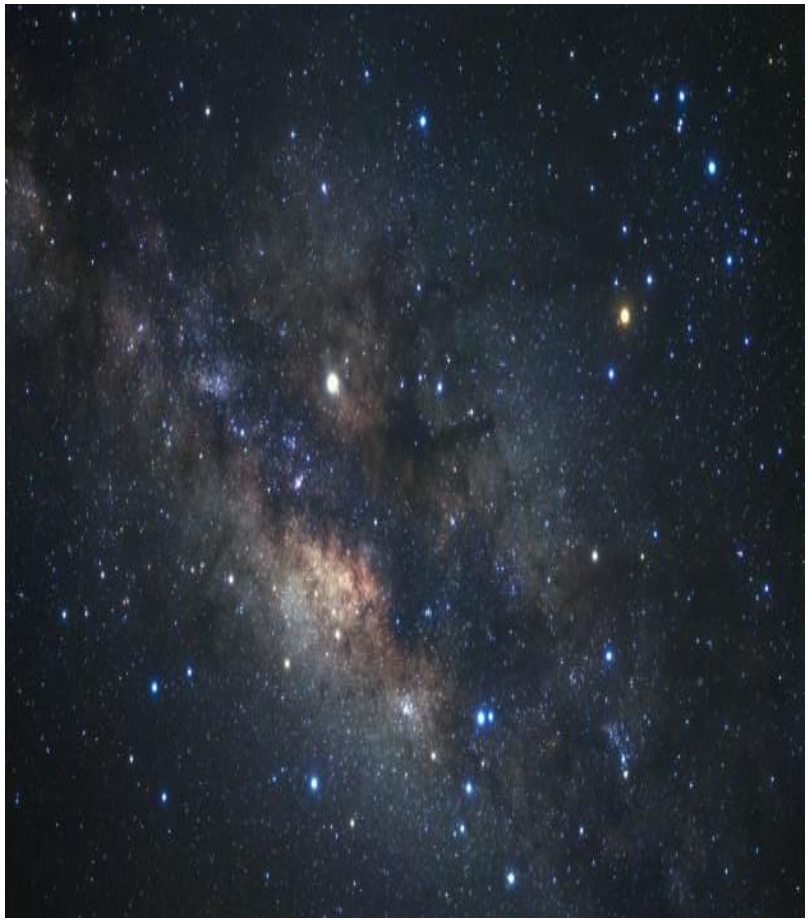
There's a Big Universe Out There…

With many unanswered questions.

Big data and machine learning helps us answer questions humans cannot easily answer.

For example:

- What are the chances that a patient will suffer a heart attack in the coming weeks? Can we prevent this?
- How do we solve the problem of meeting demand for healthcare services while having limited resources and staff?

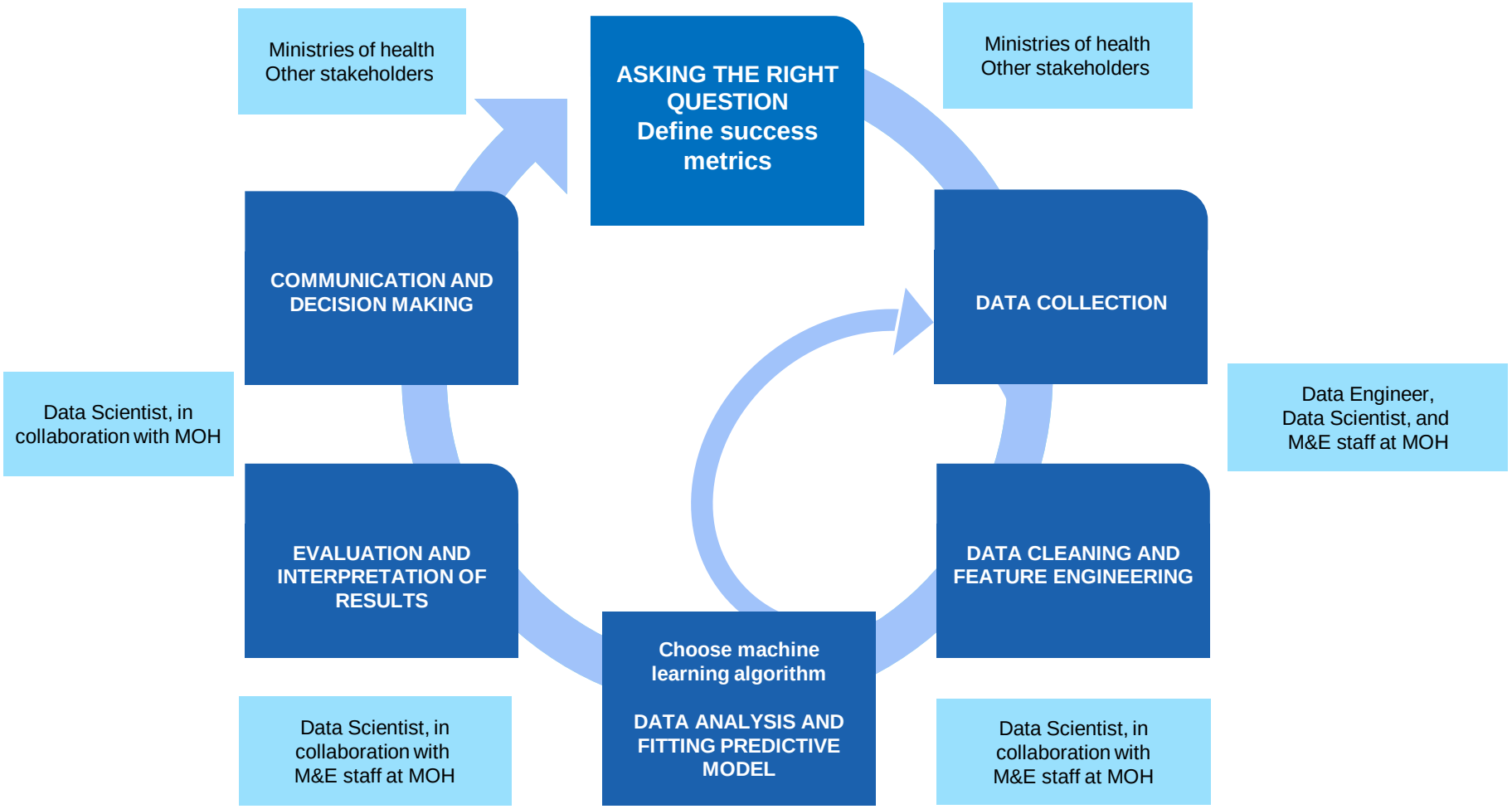


There is Method in the Madness!

These methods help guide the way and answer different questions:

- Regression
- Classification
- Reinforcement Learning
- Clustering
- Association Analysis
- Recommender Systems

LIFE CYCLE OF AI / DATA SCIENCE PROJECT IN HEALTH

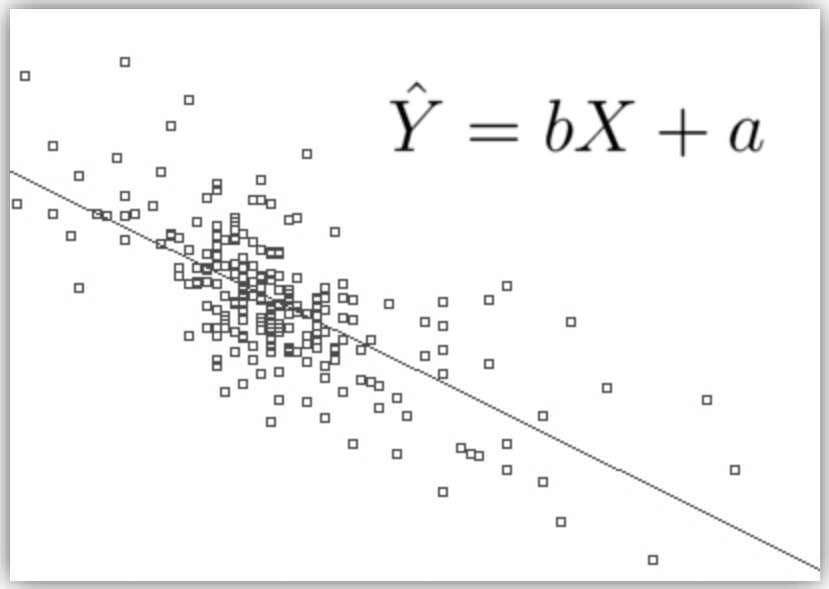


WORLD BANK GROUP
Health, Nutrition & Population

datasciencedojo
data science for everyone

Regression – Predicting a Numerical Value Problem

Maps the relationship between predictor variables to the outcome, with the outcome being an estimated numeric value.



Gribeco

Key methods:

- Linear regression
- Generalized linear regression
- Regression trees

Regression Problem

Problem: A government health agency needs a reliable estimate of the number of people who will likely need vaccinations in the coming months.

"How do we know if now is the right time to ramp up vaccine supply? What if we ramp up supply but our vaccines go to waste or expire due to there not being a 'real' (but rather perceived) demand? I would like to make a decision based on more certainty rather than going by general past trends and 'gut feeling'" – Head of Supply

Regression: A Quick Look at the Data

Check:

- Is the outcome quantitative (e.g. counts)?
- Are the predictor variables indicative of number of vaccines? Do they have a relationship with number of vaccines?

Patient ID	Visits in last 6 months	Age bracket	Vaccines
001	2	20-24	1
002	5	15-19	3
003	0	45-49	0

Regression and ROI

Using this method, the agency can minimize waste by avoiding over supply, or can avoid issues with under supply.

Using past data on 100,000 patients and their recent vaccination activity, the agency has predicted there will likely be 450 vaccinations (under 200 from what they expected!)

An estimated 200 vaccinations would have expired and gone to waste, costing the agency $$$

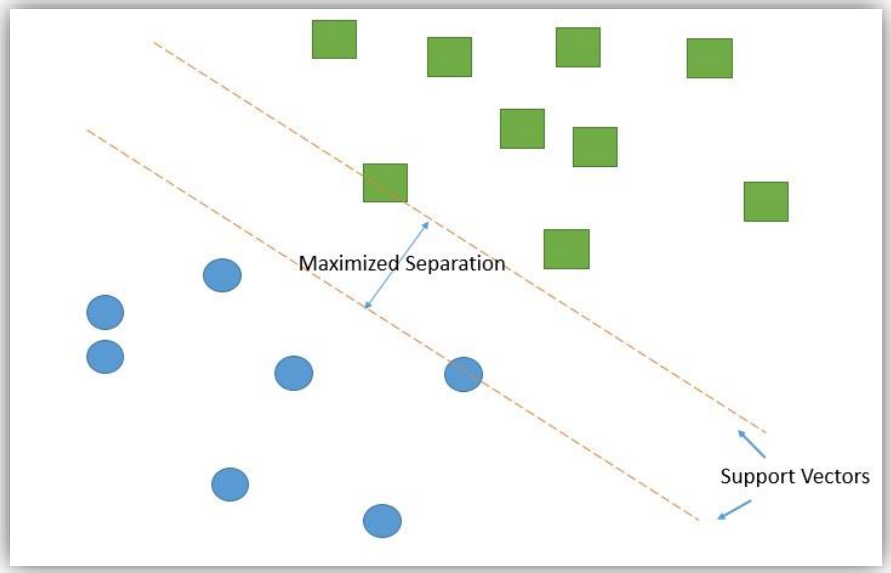
Healthcare policy on vaccine availability improves!

Classification – Finding the Class Type Problem

Determines which category a case or subject falls into based on their attributes or features, with the outcome being a class level/type.

Key methods:

- Support vector machine
- Logistic regression (binary outcome)
- Tree-based methods: Decision trees, boosted decision trees, random forest)
- Naïve Bayes



Qluong2016

Classification Problem

Problem: A doctor would like to know which patients are at high risk, medium risk and low risk of suffering a heart attack in the coming months.

"I need to attend to each patient's specific needs, and not every patient needs to pay for a full check. I can do a standard diagnosis, but if I could know beforehand what level of risk he/she is at, that would help me decide if it is necessary to monitor the patient much more closely or not" – Doctor

Classification: A Quick Look at the Data

Check:

- Is the outcome categorical or a class type?
- Are the class types clearly defined or is there a lot of overlap? If there's overlap and too many categories, is classification useful to your problem, or should you score risk level instead of using classes?

Patient ID	Cholesterol level	Blood pressure level	Hours exercise pw	Smoker	Weight	Height	Age bracket	Risk level
001	2	3	0.25	0	180.20	62.00	25-29	Medium
002	1	1	2.50	0	132.52	65.20	30-34	Low
003	3	3	1.0	1	190.01	66.28	45-49	High

Classification and ROI

Using this method, the doctor can save time and pin point which patients need close monitoring, which patients need some monitoring, and which patients don't require much monitoring or further checks.

Using past data on 500 patients and their attributes, the doctor has determined only 2 of his/her patients are at high risk.

Not all patients need to go through the expense of detailed checks and monitoring.

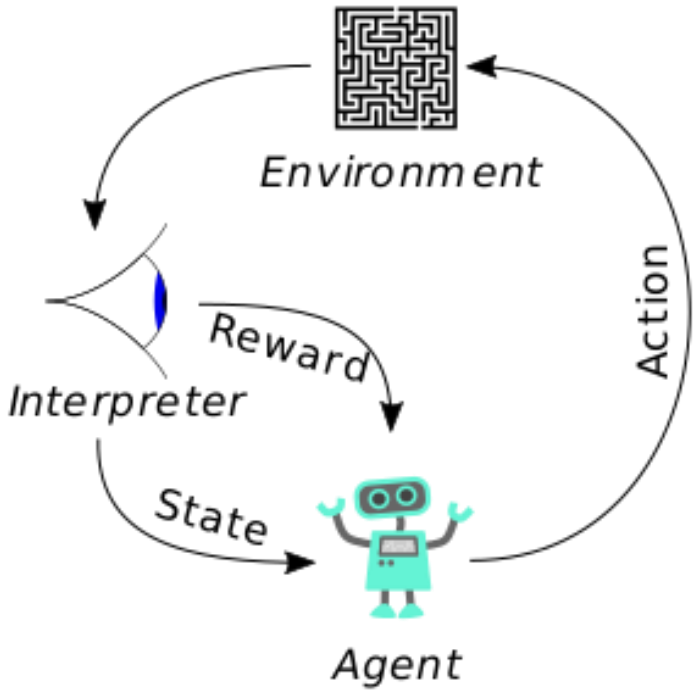
Healthcare policy on affordable healthcare improves!

Reinforcement – Mastering a Specific Task Problem

Takes the course of action that maximizes reward or the end goal, with the outcome being a score.

Key methods:

- Q-learning
- State-Action-Reward-State-Action (SARSA)
- Deep Q Network
- Deep Deterministic Policy Gradient (DDPG)



Megajuice

Reinforcement Problem

Problem: A home care organization does not have enough in-home nurses to attend to the amount of elderly people registering for assistance in their homes. The organization is open to robotic assistance.

"We just don't have enough people to do everyday tasks that require 24/7 assistance, such as sorting groceries into compartments, picking up objects around the home, unscrewing lid tops, and the like. Surely these menial tasks can be outsourced or automated? Our staff are far too busy as it is" – Home Care Manager

Reinforcement: A Quick Look at the Data

Check:

- Are you trying to reach an optimal outcome, given a situation or state?
- Is your outcome to have an agent achieve a specific task or goal?

Q-table example:

	UP	DOWN	LEFT	RIGHT
0	0	0	0	0
1	0	0	0	0
2	0	2.0	2.0	0
3	0	0	0	0

Reinforcement and ROI

Using this method, the home care organization can fully utilize the limited number of nurses they have to work on the tasks that are most critical, while assistive robots work on the menial tasks.

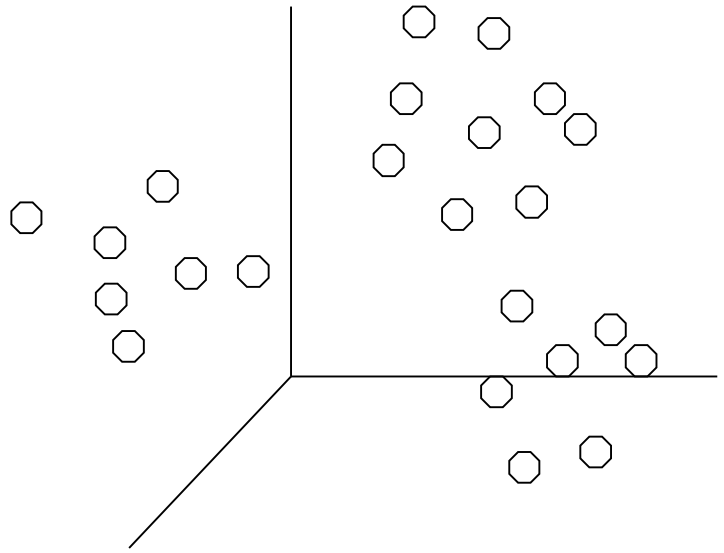
Trained on thousands of action-state data, the assistive robot is able to perform to a high standard.

Patients who required 24/7 care are now able to get their everyday needs met.

Healthcare policy on in-home care improves!

Clustering – Discovering Underlying Similarities and Differences Problem

Groups subjects or cases together based on how close or distant their values/attributes are.



Key methods:

- K-means/modes
- Mean-shift
- Hierarchical
- Gaussian mixture model (GMM)

Clustering Problem

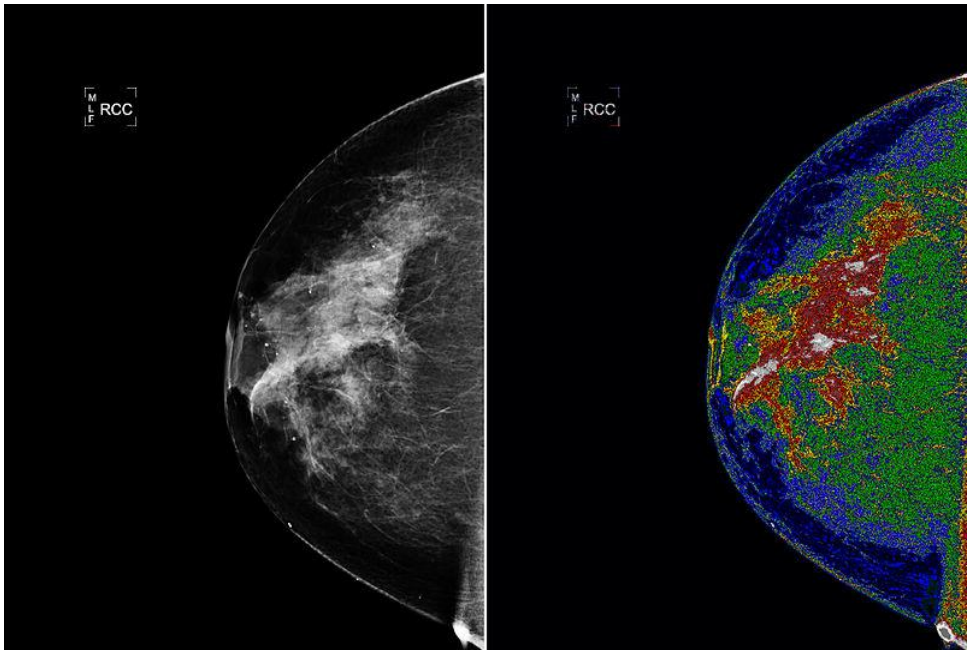
Problem: An oncologist needs to automatically detect signs of different types of breast cancer from mammogram images.

"If I had to sit and visually inspect every single mammogram of all our patients, it would probably take me years! In my field of work, I cannot afford to accidentally overlook or miss anything. It's essential that our screening systems detect and surface different types of breast cancer quickly" – Oncologist

Clustering: A Quick Look at the Data

Check:

- Are you trying to separate out a group of interest?
- Are you trying to identify each group of interest?



The image on the right shows the region of interest in white.
NASA Goddard Space Flight Center

Clustering and ROI

Using this method, the oncologist can quickly separate harmful invasive cancers from non-invasive type cancers for fast diagnosis and treatment.

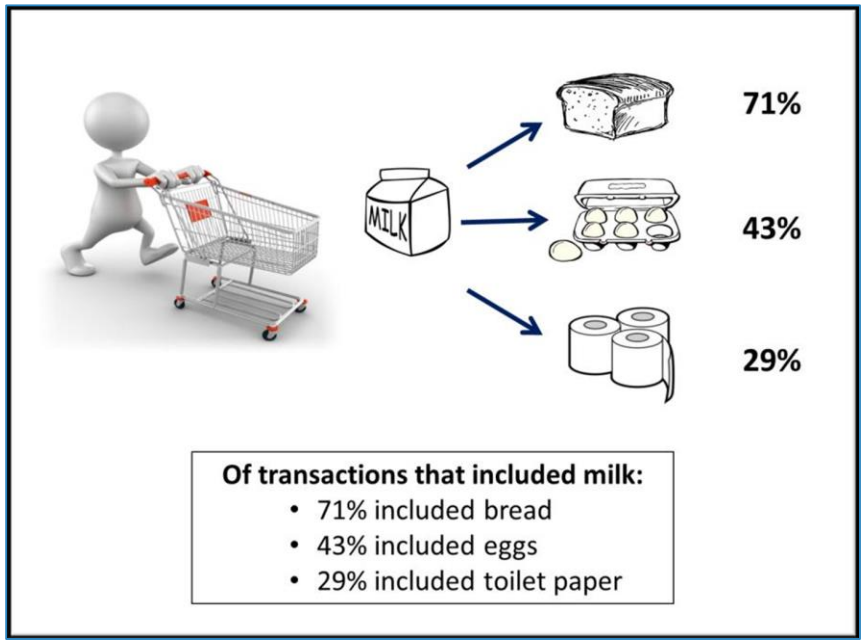
Trained on hundreds of mammograms, cancers are spotted almost immediately.

Patients are treated more quickly, preventing the disease from further developing.

Healthcare policy on fast diagnosis and treatment improves!

Association Analysis – Discovering Relationships Problem

Uncovers relationships in the form of association rules with frequent items.



Your behavior is being predicted, not by studying you, but by studying others.

Association Analysis Problem

Problem: A researcher wants to find out if certain genes have association with disabilities or cancers.

"It is difficult to know which genes cause disabilities. Is this hereditary? Can we study the occurrence of other genome sequences?" – Researcher

Association: A Quick Look at the Data

Check:

- Are you trying to predict the occurrence of an item based on the occurrences of other items?
- Is there a dependency where C always pairs with G, for example?

DNA 1	A	G	T	C	G	T	C	G	A	A	T	A	C	A
DNA 2	G	T	A	G	G	T	T	A	G	C	C	C	A	C
DNA 3	T	C	A	G	C	C	T	T	G	G	C	C	A	G
DNA 4	T	G	A	C	A	C	A	T	G	C	A	G	T	A

Rule discovered > C always pairs with G

Association Analysis and ROI

Using multiple genome sequences a researcher can identify genes that occur in disability.

By comparing genes from blood relatives to find the exact difference:

- Helps doctors diagnose the cause of the condition and decide the treatment

Helps create policies in health innovation, science and research!

Recommender Systems – Choosing the Best Option Problem

Suggests the best option based on its similarity to a person's preferences.

Key methods:

- Cosine similarity
- Pearson's correlation
- Euclidean similarity

Recommender Systems Problem

Problem: A network of fertility clinics has discovered patients are not finding the right clinic that meets their needs. From user surveys, women seeking fertility clinics were not happy with their clinic.

"Every patient has their own specific needs and preferences; there is no one-size fits all when it comes to fertility. If we could build an app that automatically suggests which clinic is the best at servicing their need, then that would help ensure they get their needs met" – Clinic Customer Service

Recommender Systems: A Quick Look at the Data

Check:

- Are you trying to score user ratings?
- Are you trying to match user preferences with a set list of options?

Pearson's example:

Alice	5	3	4	4	?	
Sara	3	1	2	3	3	sim=0.85
Anna	4	3	4	3	5	sim=0.90
Donna	3	3	1	5	4	sim=0.70
Evi	1	5	5	2	1	sim=0.79

Recommender Systems and ROI

Using this method, the network of fertility clinics can match each patient's nuances to the right clinic.

Using user ratings on which hospital is best on different types of fertility treatment, the recommender system has resulted in a significant increase in positive clinic reviews.

Patient needs are met and their experience improves.

Healthcare policy for providing adequate services improves!

From Analytics/ML Translator to Data Science Engineer

Analytics/ML translators are just as important as data scientists and data engineers as they understand the business problem!

- Identify applications and use cases
- Convey the needs of the business to data scientists and engineers and vice versa
- Generate user buy-in
- Educate the business on high level concepts of analytics/ML