Data

Umair Mehmood Imam

INTRODUCTION TO A.I - UMAIR M.I

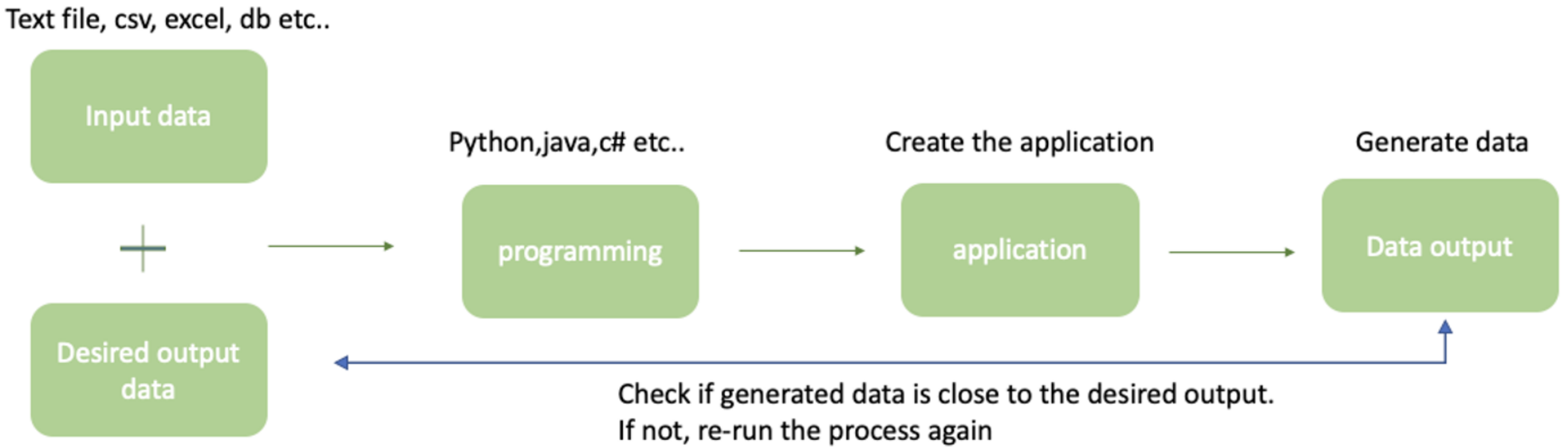
OSLO METROPOLITAN UNIVERSITY
STORBYUNIVERSITETET

Steps to design an A.I system

1. Identify the problem
2. Prepare the data
3. Choose the algorithms
4. Train the algorithms with the data
5. Run on a selected platform

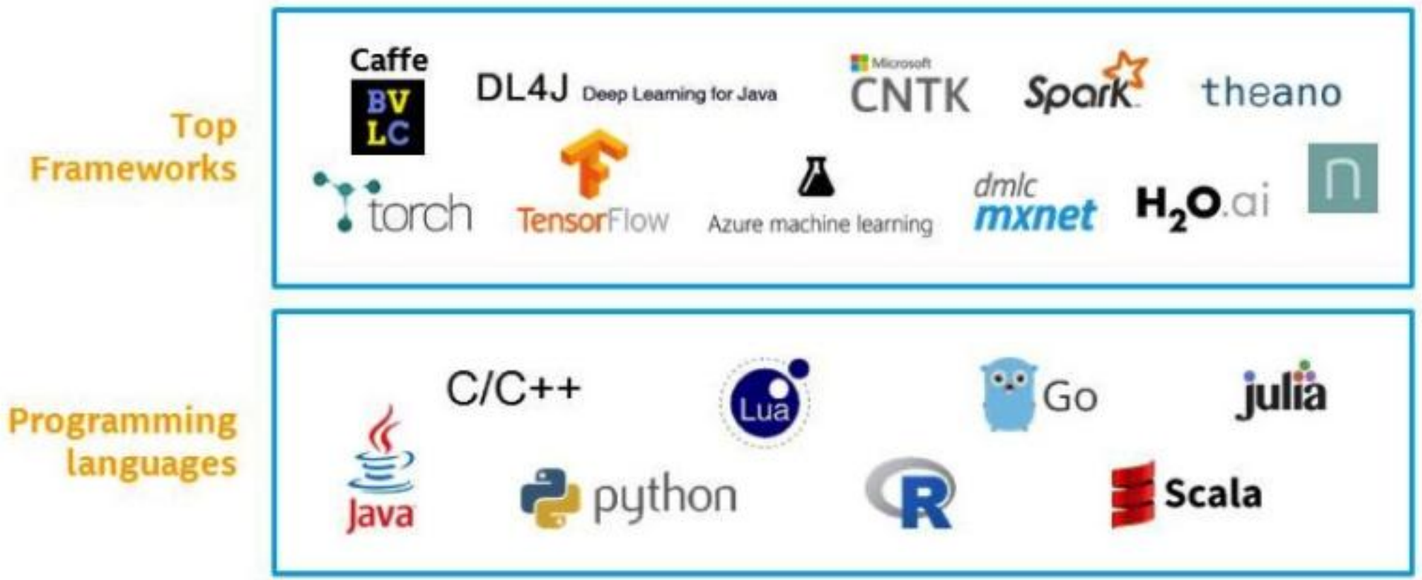
General software development

A.I based software development



A.I is about algorithms, tools and data

Algorithms and tools



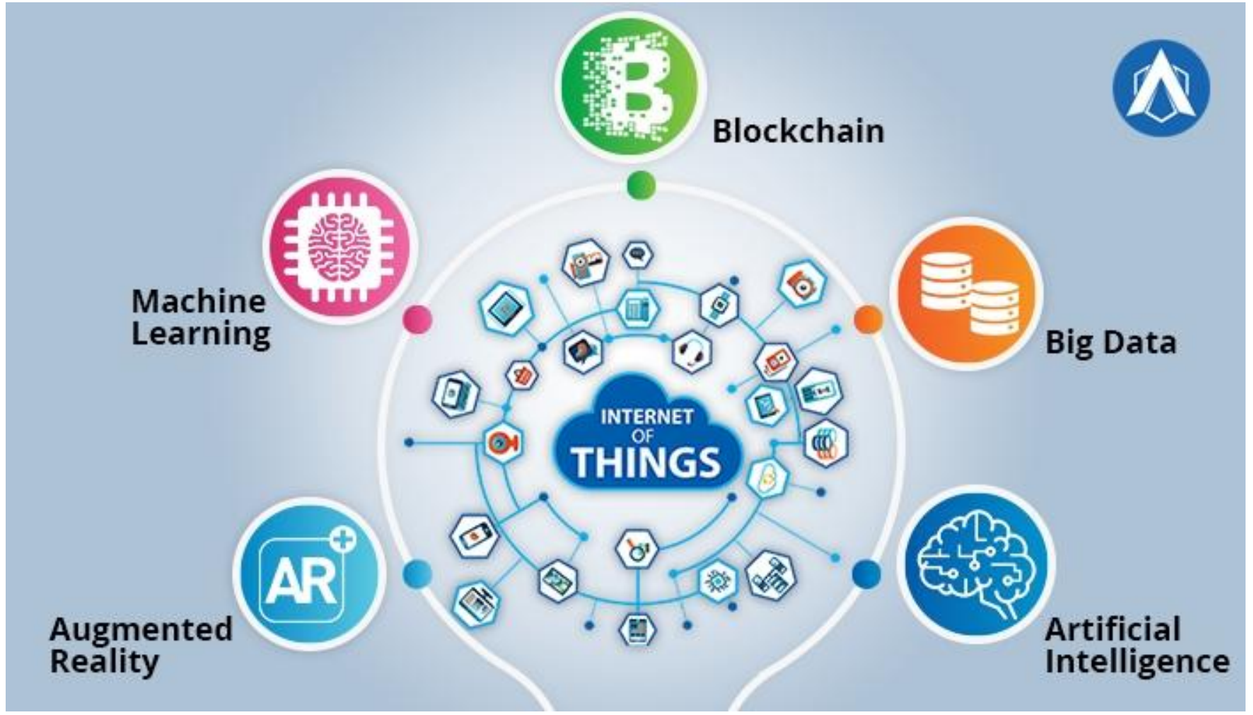
DATA

Open data sets. e.g Kaggle

Private data, e.g. any organizations data


The Economist
The world's most valuable resource
Data and the new rules of competition
Google


DATA IS THE NEW OIL


Blockchain
Machine Learning
Big Data
INTERNET OF THINGS
AR
Augmented Reality
Artificial Intelligence

Daily life of an A.I programmer

- 80% time spend on data (cleaning, preparing, labeling, analyzing etc)
- 5% on deployment (cloud/on premise)
- 15% on A.I development

AI Engineer

freelancermap

The role

- Machine learning, data science
- Software design
- Create and deploy machine learning algorithms

Background

- Degree in: Computer science, robotics, engineering, physics
- ML Coursera, AI Google Education
- MSc, PhD in related fields

Skills

- Data Science & Statistics
- Mathematics
- CI/CD & SDLC knowledge
- CS & Programming

Salary

Junior:	$ 57,000
Average:	$ 86,000
Top:	$ 114,000

Data pitfalls (problems which can occur with data)

- Assuming the data is clean
  - e.g spelling mistakes
- Outliers
  - Excluding outliers
  - Including outliers
- Ignroing seasonality
  - Easter vacations, summer holidays, black Friday etc.
- Context is critical
  - Ignoring size when reporting growth

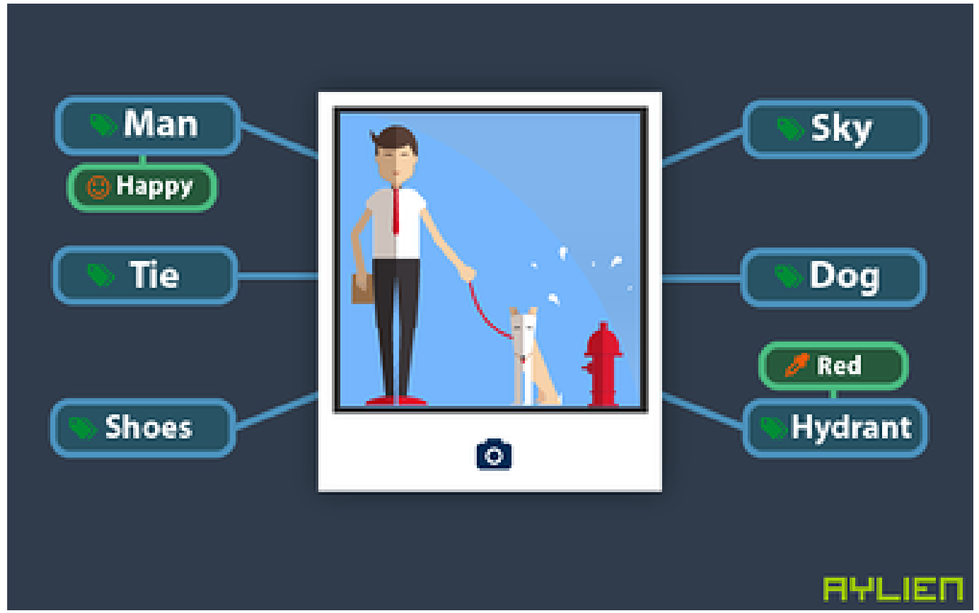
Countinued..

- Poor data Insights
- Not connecting with external data
- Lacking business understanding

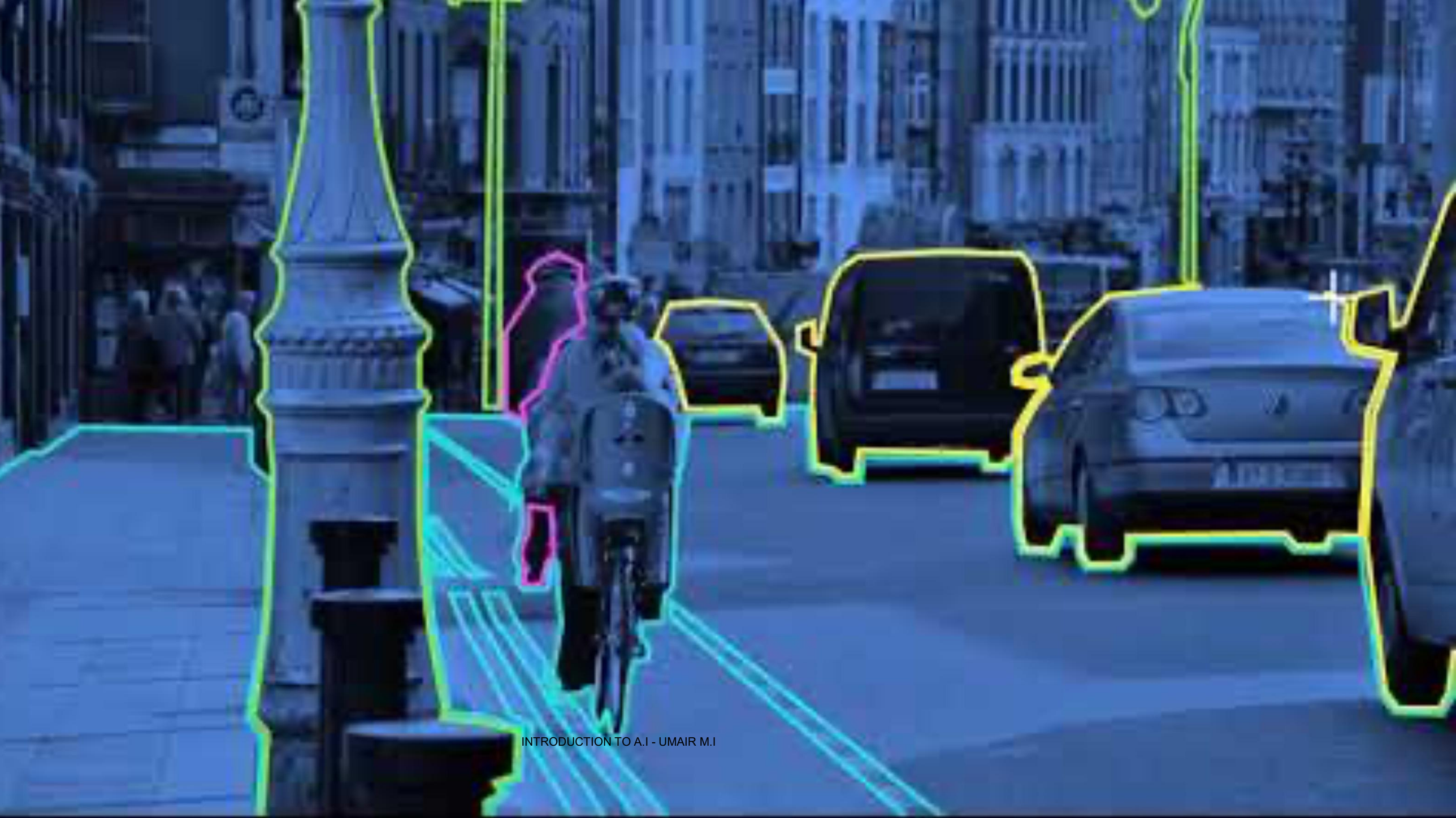
How to work with data ?

- Data labeling / annotation
- Data annonimization
- Synthetic data
- Data preperation
  - Data cleansing + feature engineering
- Data wrangling
- Data mining
- Data warehousing
  - ETL programming
- Data Engineering
  - Infrastructure, big data, cloud etc..

1. Data Labeling / annotation



OSLOMET




SIGN
SIGN
SIGN
SIGN
PERSON
PERSON
CAR
CAR
CAR
CAR
CAR
CAR
CAR
CAR
CAR
PERSON
PERSON
PERSON
PERSON


Roads

French authorities use AI to help recoup €200 million in unpaid tax from one department alone

Undeclared swimming pools and property extensions can be captured by AI imagery

Pools are among the most common 'undeclared' improvements captured by AI Francois BOIZOT / Shutterstock

https://www.connexionfrance.com/practical/french-authorities-use-ai-to-help-recoup-200-million-in-unpaid-tax-from-one-department-alone/701469

Forbes

INNOVATION > CLOUD

Meta Invests $14 Billion In Scale AI To Strengthen Model Training

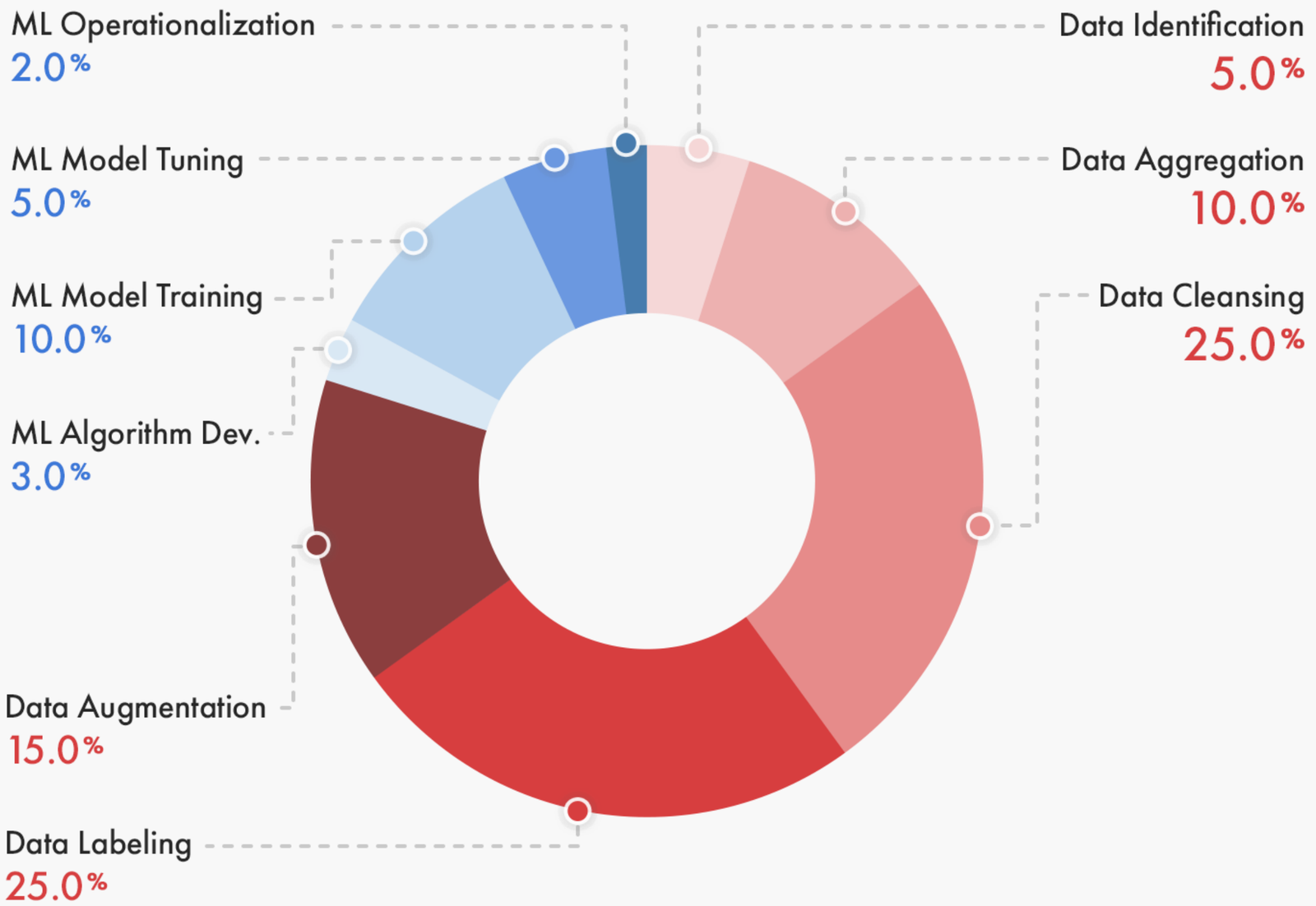
By Janakiram MSV, Senior Contributor. ⓘ I cover emerging technologies wit...

Follow Author

Published Jun 23, 2025, 01:03am EDT

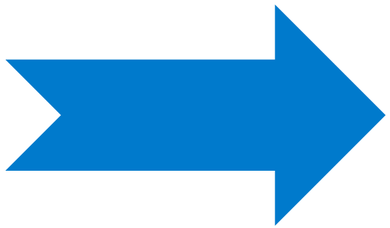
OSLOMET

Percentage of Time Allocated to Machine Learning Project Tasks



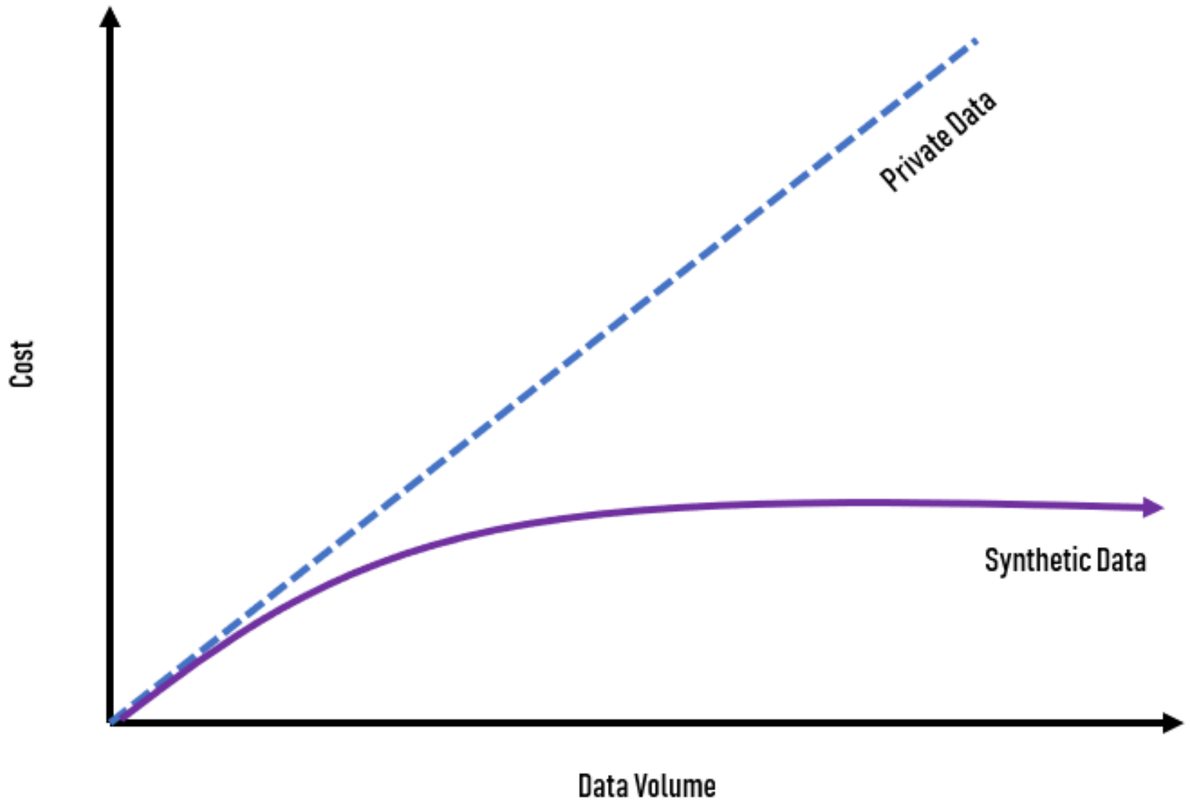
2. Data annonimization

id : 345619
name : Bob Evans
email : bob@gmail.com
phone : 734-576-893



id : 5461827
name : Jan Novak
email : ws45@sd7r.com
phone : 234-903-485

3. Synthetic data

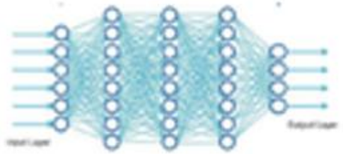


AI-generated **synthetic populations** of customers and their behavior

NAME	ZIP	AGE	GENDER	ITEM	EUR	DATE	TIME
Mary	1220	25y	female	Book	12€	4/2/19	8:12
John	2320	72y	male	Pizza	34€	4/2/19	18:12
...							
Kevin	8329	18y	male	Swim	6€	4/4/19	10:02

actual

mostly AI

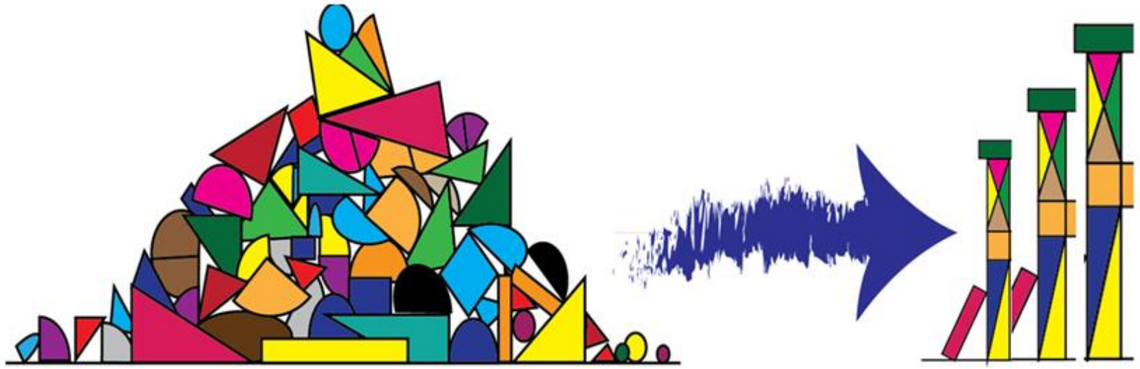


NAME	ZIP	AGE	GENDER	ITEM	EUR	DATE	TIME
Bob	3729	82y	male	Beer	6€	4/2/19	15:32
Sue	8022	24y	female	Sushi	12€	4/2/19	21:32
...							
Kim	3923	29y	female	Amazon	36€	4/4/19	12:32

synthetic

4. Data preperation(cleansing + feature engineering)

Feature engineering, also known as feature creation, is the process of constructing new features from existing data to train a machine learning model.



For example

- Character recognition
  - features may include histograms counting the number of black pixels along horizontal and vertical directions, number of internal holes, stroke detection and many others.
- Speech recognition
  - features for recognizing phonemes can include noise ratios, length of sounds, relative power, filter matches and many others.
- Spam detection
  - features may include the presence or absence of certain email headers, the email structure, the language, the frequency of specific terms, the grammatical correctness of the text.
- Computer vision
  - there are a large number of possible features, such as edges and objects.

Feature Extraction

- There usually have some meaningful features inside existing features, you need to extract them manually
- Some examples
  - Location
    - Address, city, state and zip code …. (categorical or numeric)
  - Time
    - Year, month, day, hour, minute, time ranges, …. (numeric)
    - Weekdays or weekend (binary)
    - Morning, noon, afternoon, evening, … (categorical)
  - Numbers
    - Turn age numbers into ranges (ordinal or categorical)

H2O.ai

OSLOMET

Consider a dataset with 2 patterns

Pattern 1: ^+*B*++*T*+*C*+

Pattern 2: +*R*++*T*+*C*

Extraction of features

Patterns	Features extracted
^+B++T+C+	^+B++, ++T+, +C+
+R++T+C	+R++, ++T+, +C

Representing patterns as a vector

Patterns	^+B++	++T+	+C+	+R++	+C
^+B++T+C+	1	1	1	0	0
+R++T+C	0	1	0	1	1

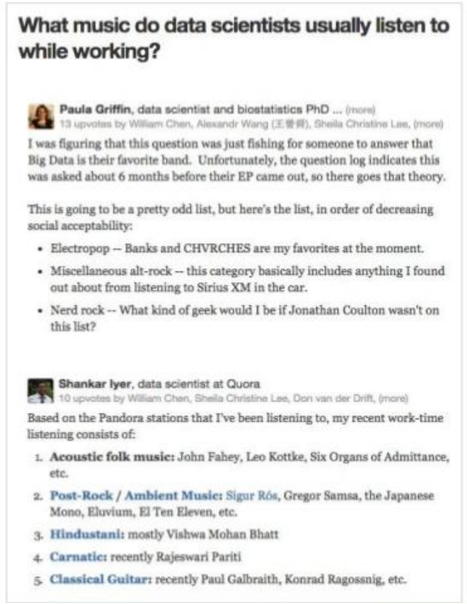
* Numbers in the cell indicate presence of that particular feature

Feature Engineering Example - Quora Answer Ranking

Quora

What is a good Quora answer?

- truthful
- reusable
- provides explanation
- well formatted
- ...



What music do data scientists usually listen to while working?

Paula Griffin, data scientist and biostatistics PhD ... (more)
13 upvotes by William Chen, Alexandr Wang (王誉霖), Sheila Christine Lee, (more)

I was figuring that this question was just fishing for someone to answer that Big Data is their favorite band. Unfortunately, the question log indicates this was asked about 6 months before their EP came out, so there goes that theory.

This is going to be a pretty odd list, but here's the list, in order of decreasing social acceptability:

- Electropop -- Banks and CHVRCHES are my favorites at the moment.
- Miscellaneous alt-rock -- this category basically includes anything I found out about from listening to Sirius XM in the car.
- Nerd rock -- What kind of geek would I be if Jonathan Coulton wasn't on this list?

Shankar Iyer, data scientist at Quora
10 upvotes by William Chen, Sheila Christine Lee, Don van der Drift, (more)

Based on the Pandora stations that I've been listening to, my recent work-time listening consists of:

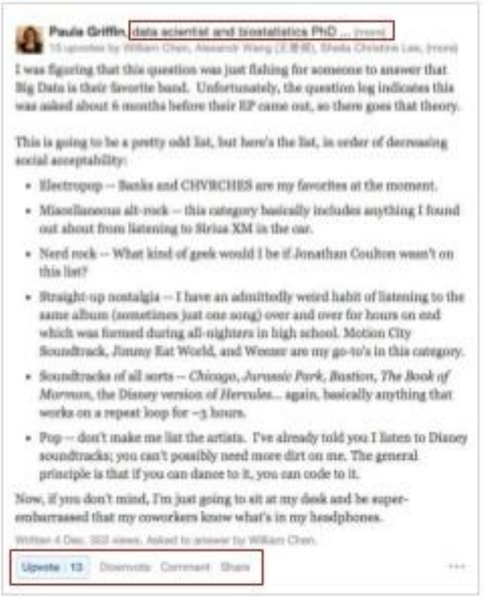
1. **Acoustic folk music:** John Fahey, Leo Kottke, Six Organs of Admittance, etc.
2. **Post-Rock / Ambient Music:** Sigur Rós, Gregor Samsa, the Japanese Mono, Eluvium, El Ten Eleven, etc.
3. **Hindustani:** mostly Vishwa Mohan Bhatt
4. **Carnatic:** recently Rajeswari Pariti
5. **Classical Guitar:** recently Paul Galbraith, Konrad Ragossnig, etc.

Feature Engineering Example - Quora Answer Ranking

Quora

How are those dimensions translated into features?

- Features that relate to the answer quality itself
- Interaction features (upvotes/downvotes, clicks, comments...)
- User features (e.g. expertise in topic)



Paula Griffin, data scientist and biostatistics PhD ... (more)
15 upvotes by William Chen, Alexandr Wang (王誉霖), Sheila Christine Lee, (more)

I was figuring that this question was just fishing for someone to answer that Big Data is their favorite band. Unfortunately, the question log indicates this was asked about 6 months before their EP came out, so there goes that theory.

This is going to be a pretty odd list, but here's the list, in order of decreasing social acceptability:

- Electropop -- Banks and CHVRCHES are my favorites at the moment.
- Miscellaneous alt-rock -- this category basically includes anything I found out about from listening to Sirius XM in the car.
- Nerd rock -- What kind of geek would I be if Jonathan Coulton wasn't on this list?
- Straight-up nostalgia -- I have an admittedly weird habit of listening to the same album (sometimes just one song) over and over for hours on end which was formed during all-nighters in high school. Motion City Soundtrack, Jimmy Eat World, and Weezer are my go-to's in this category.
- Soundtracks of all sorts -- *Chicago*, *Jurassic Park*, *Bastion*, *The Book of Mormon*, the Disney version of *Hercules*... again, basically anything that works on a repeat loop for ~3 hours.
- Pop -- don't make me list the artists. I've already told you I listen to Disney soundtracks; you can't possibly need more dirt on me. The general principle is that if you can dance to it, you can code to it.

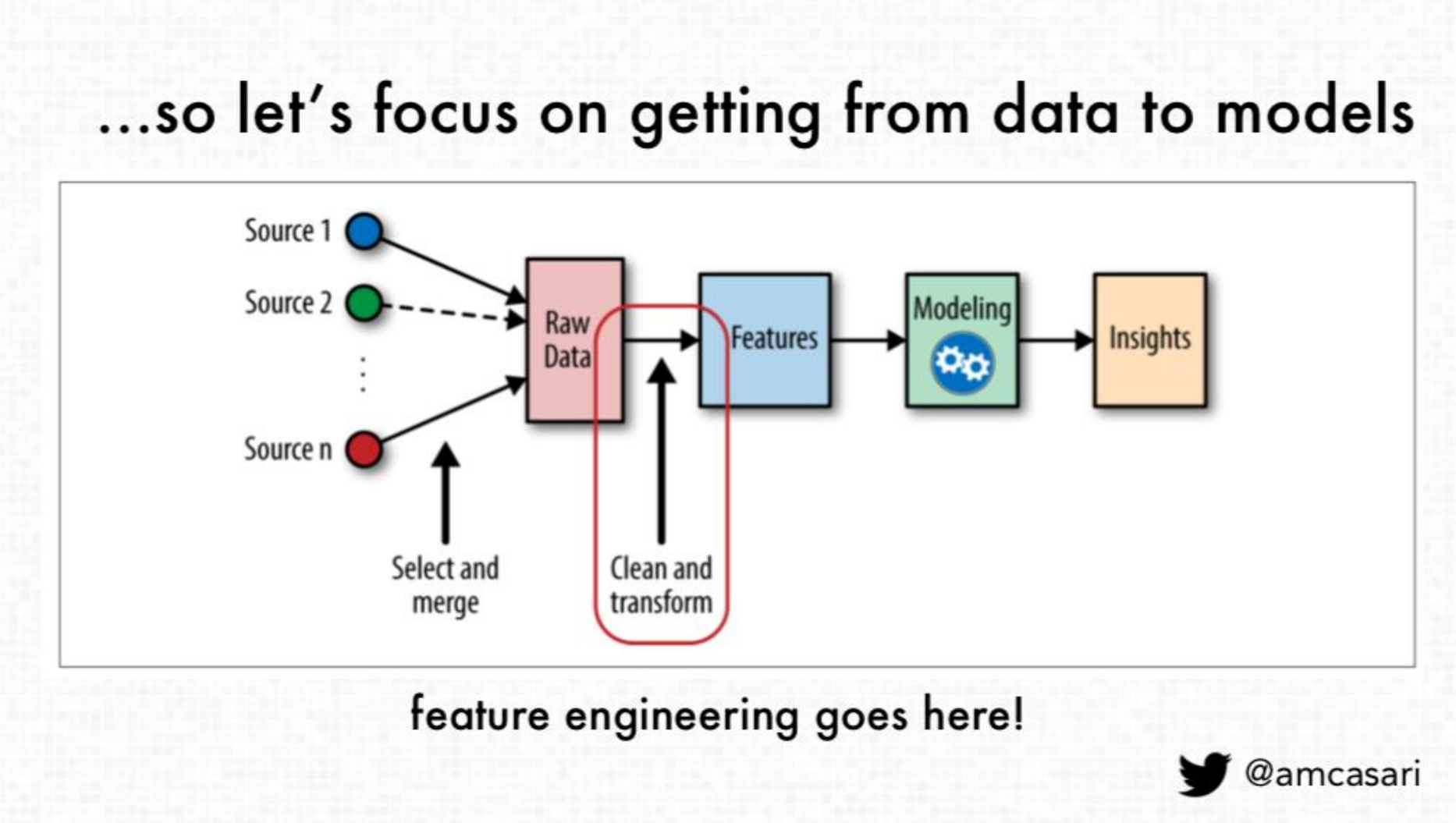
Now, if you don't mind, I'm just going to sit at my desk and be super-embarrassed that my coworkers know what's in my headphones.

Written 4 Dec. 353 views. Asked to answer by William Chen.

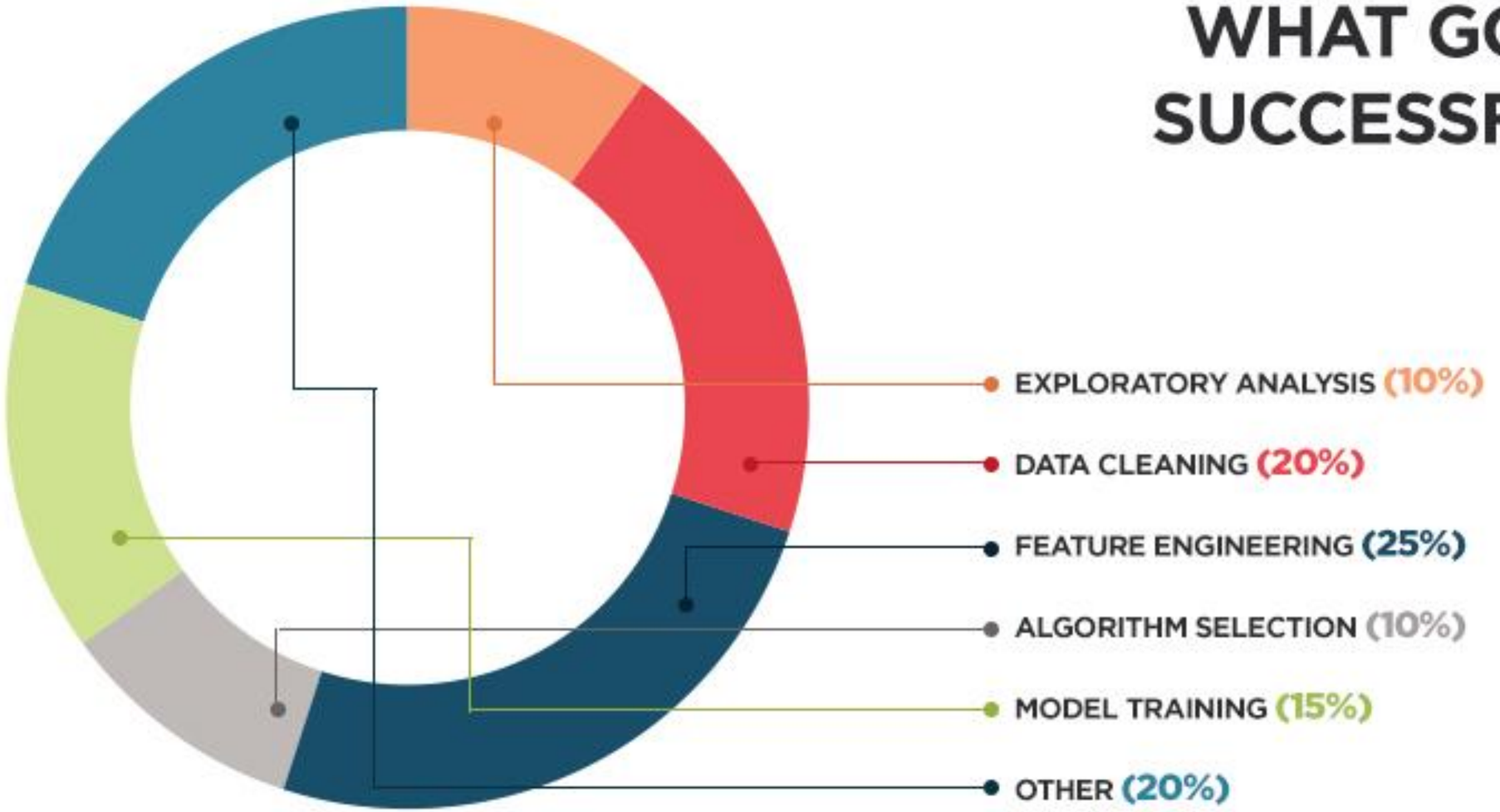
Upvote · 15 Downvote Comment Share

Data preperation(cleansing + feature engineering)

OSLOMET



WHAT GOES INTO A SUCCESSFUL MODEL



DATA SCIENCE PRIMER
by EliteDataScience

Feature engineering -> Transformations

Transformations

- Create new features out of one or more of the existing data.

client_id	joined	income	credit_score
46109	2002-04-16	172677	527
49545	2007-11-14	104564	770
41480	2013-03-11	122607	585
46180	2001-11-06	43851	562
25707	2006-10-06	211422	621

client_id	joined	income	credit_score	join_month
46109	2002-04-16	172677	527	4
49545	2007-11-14	104564	770	11
41480	2013-03-11	122607	585	3
46180	2001-11-06	43851	562	11
25707	2006-10-06	211422	621	10

Feature engineering -> Aggregations

Aggregations

- Performed across data and usually calculates statistics

client_id	joined	income	credit_score
46109	2002-04-16	172677	527
49545	2007-11-14	104564	770
41480	2013-03-11	122607	585
46180	2001-11-06	43851	562
25707	2006-10-06	211422	621

client_id	joined	income	credit_score	join_month	log_income	mean_loan_amount	max_loan_amount	min_loan_amount
46109	2002-04-16	172677	527	4	12.059178	8951.600000	14049	559
49545	2007-11-14	104564	770	11	11.557555	10289.300000	14971	3851
41480	2013-03-11	122607	585	3	11.716739	7894.850000	14399	811
46180	2001-11-06	43851	562	11	10.688553	7700.850000	14081	1607
25707	2006-10-06	211422	621	10	12.261611	7963.950000	13913	1212
39505	2011-10-14	153873	610	10	11.943883	7424.050000	14575	904
32726	2006-05-01	235705	730	5	12.370336	6633.263158	14802	851
35089	2010-03-01	131176	771	3	11.784295	6939.200000	13194	773
35214	2003-08-08	95849	696	8	11.470529	7173.555556	14767	667
48177	2008-06-09	190632	769	6	12.158100	7424.368421	14740	659

Please read on your own time

- https://towardsdatascience.com/feature-engineering-in-python-part-i-the-most-powerful-way-of-dealing-with-data-8e2447e7c69e

5. Data wrangling

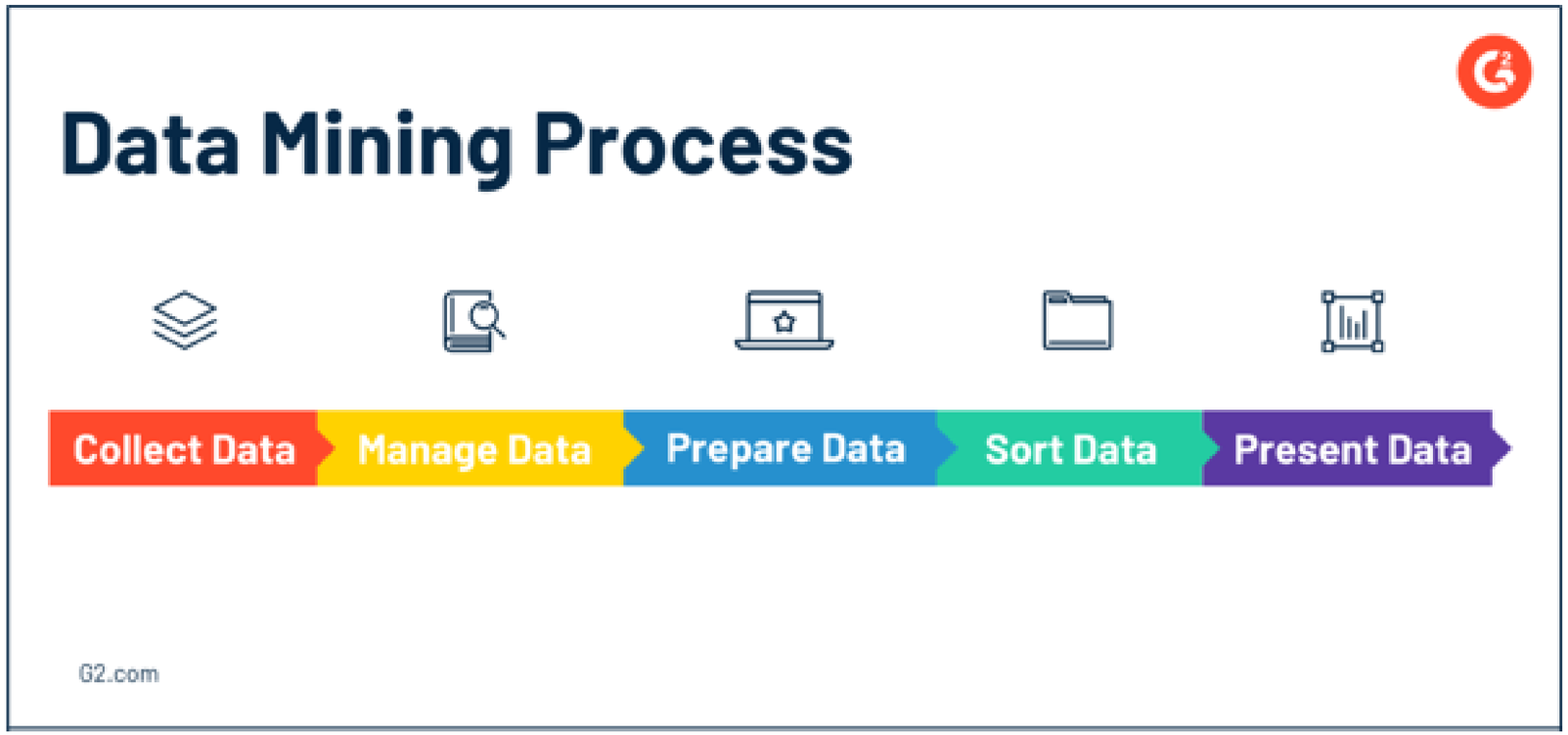
The process of transforming and mapping **data** from one "raw" **data** form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics.

Tools:

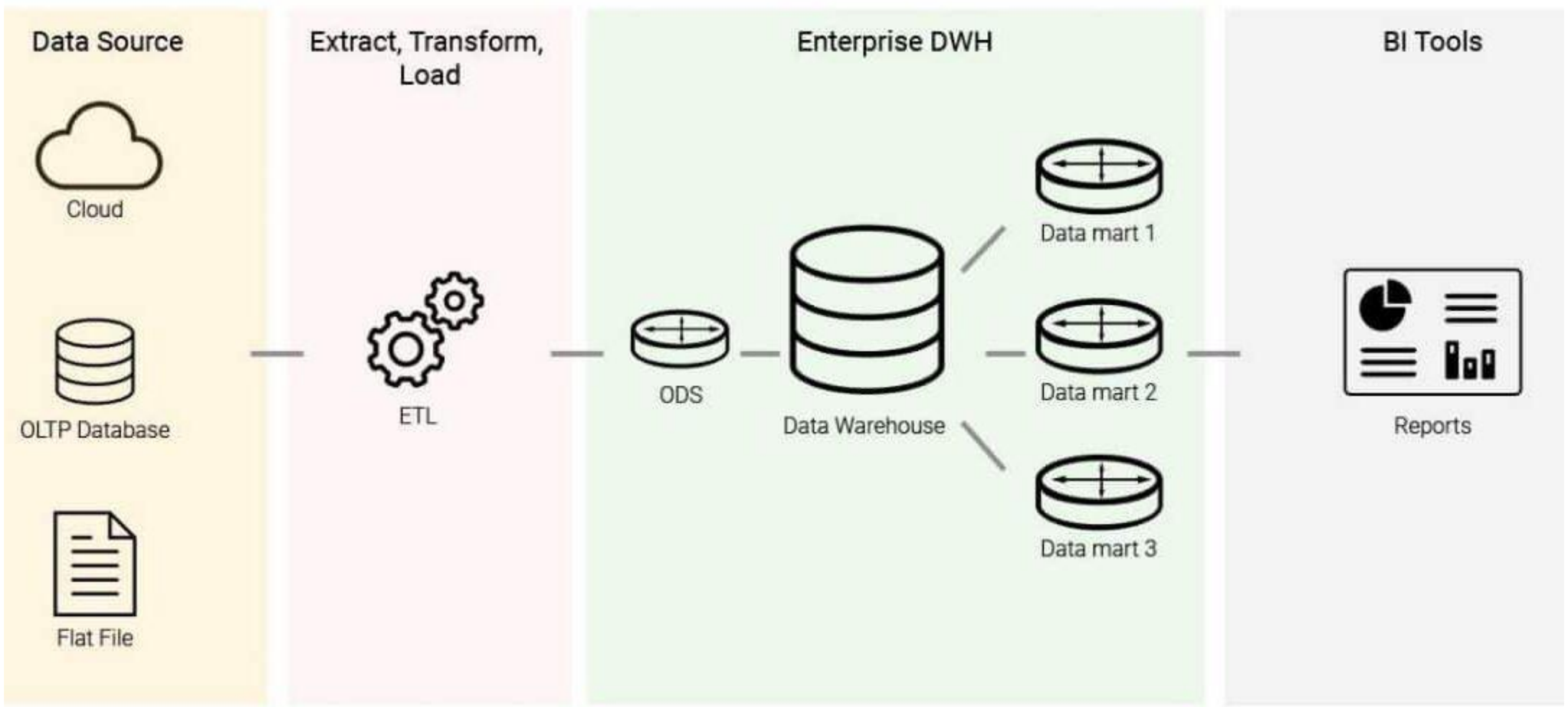
Python, Excel, Tablua, Google data prep

6. Data Mining

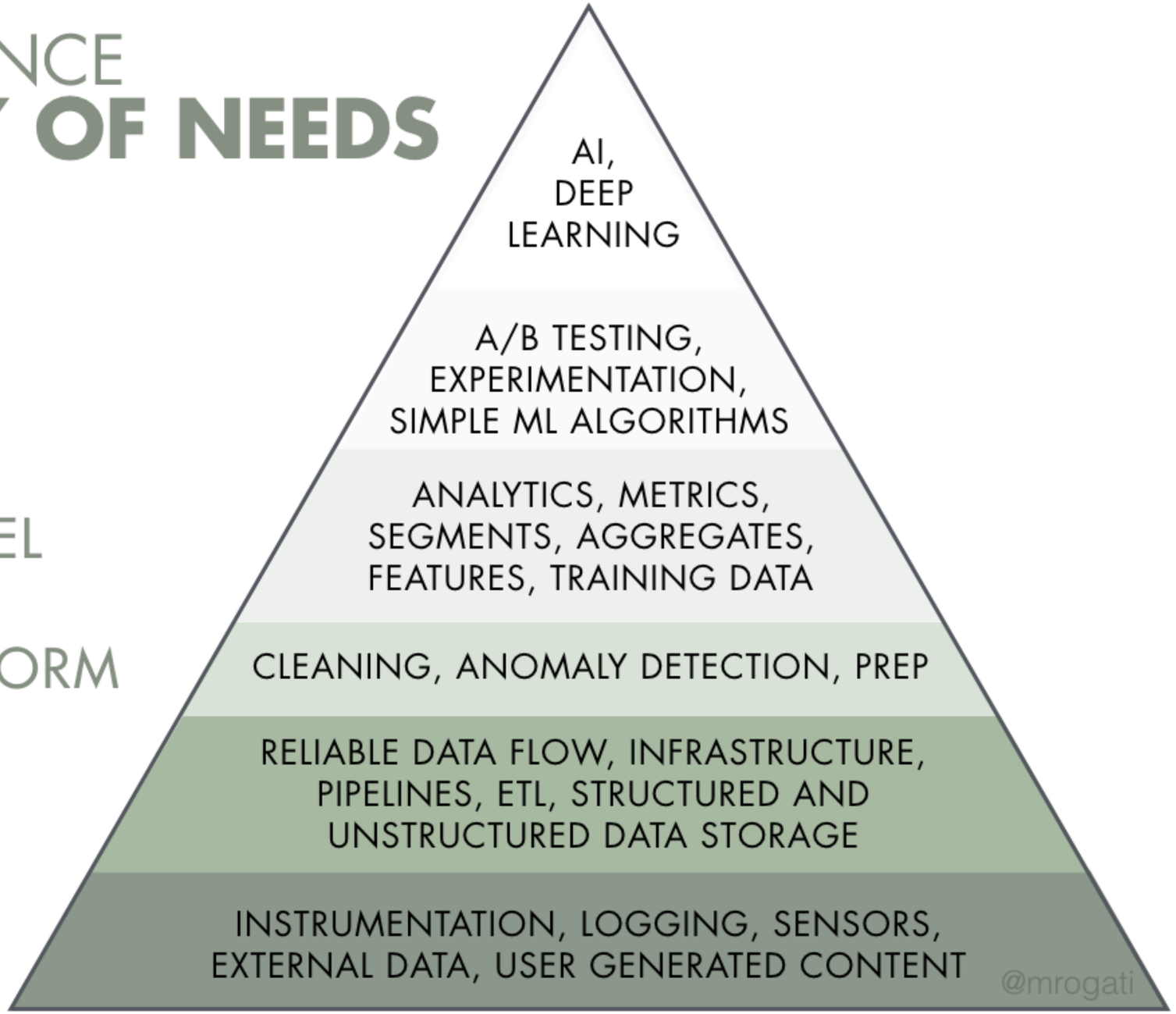
The process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.



7. Data warehousing



THE DATA SCIENCE **HIERARCHY OF NEEDS**



LEARN/OPTIMIZE

AGGREGATE/LABEL

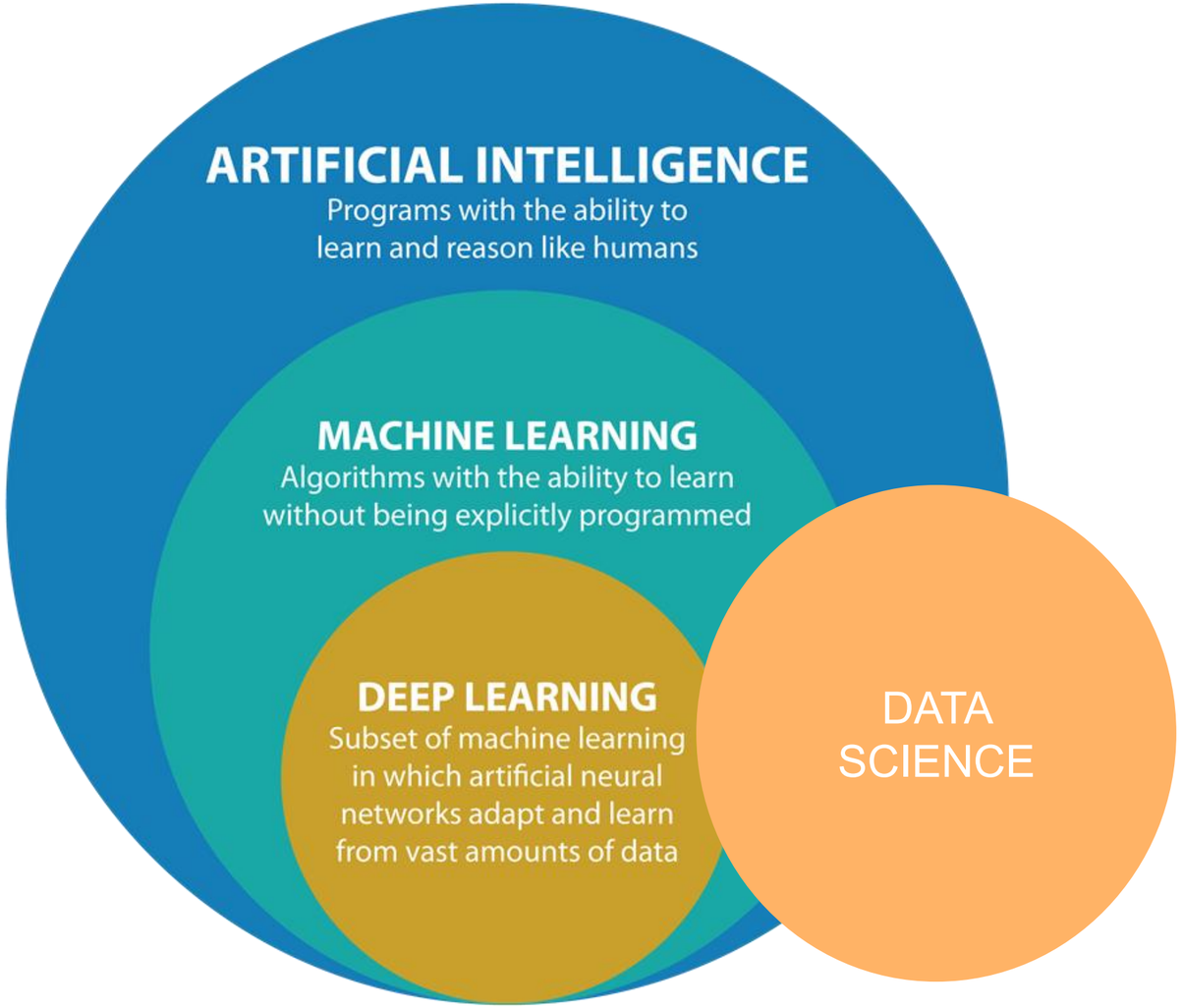
EXPLORE/TRANSFORM

MOVE/STORE

COLLECT

by Monica R

This has given rise to a field related to AI,



Data Science vs Artificial Intelligence

Factors	Data Science	Artificial Intelligence
Scope	Involves various underlying data operations	Limited to the implementation of ML algorithms
Type of Data	Structured and unstructured	Standardized in the form of embeddings and vectors
Tools	R, Python, SAS, SPSS, TensorFlow, Keras, Scikit-learn	Scikit-learn, Kaffe, PyTorch, TensorFlow, Shogun, Mahout
Applications	Advertising, Marketing, Internet Search Engines	Manufacturing, Automation, Robotics, Transport, Healthcare