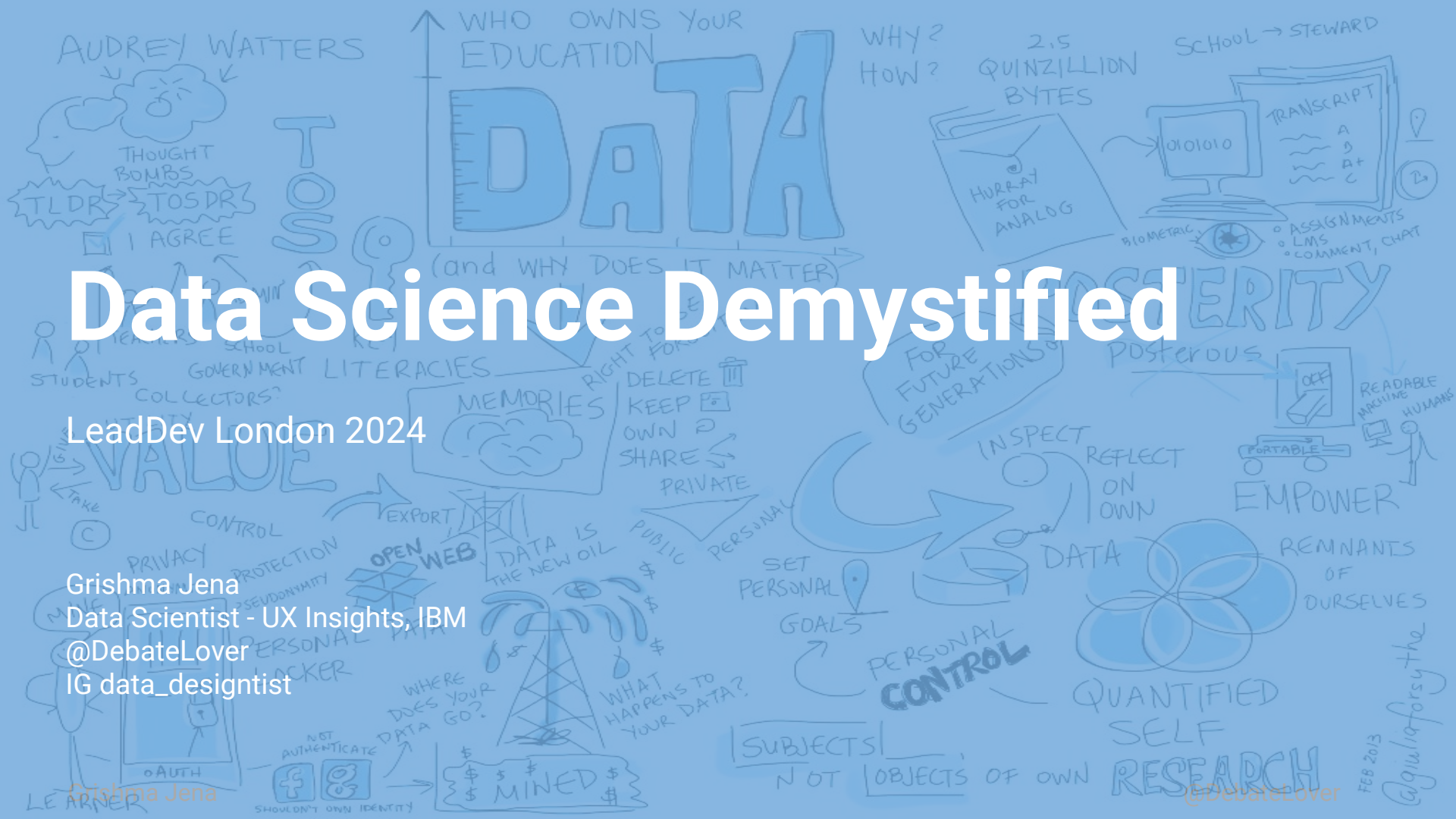


Data Science Demystified

LeadDev London 2024

Grishma Jena
Data Scientist - UX Insights, IBM
@DebateLover
IG data_designtist

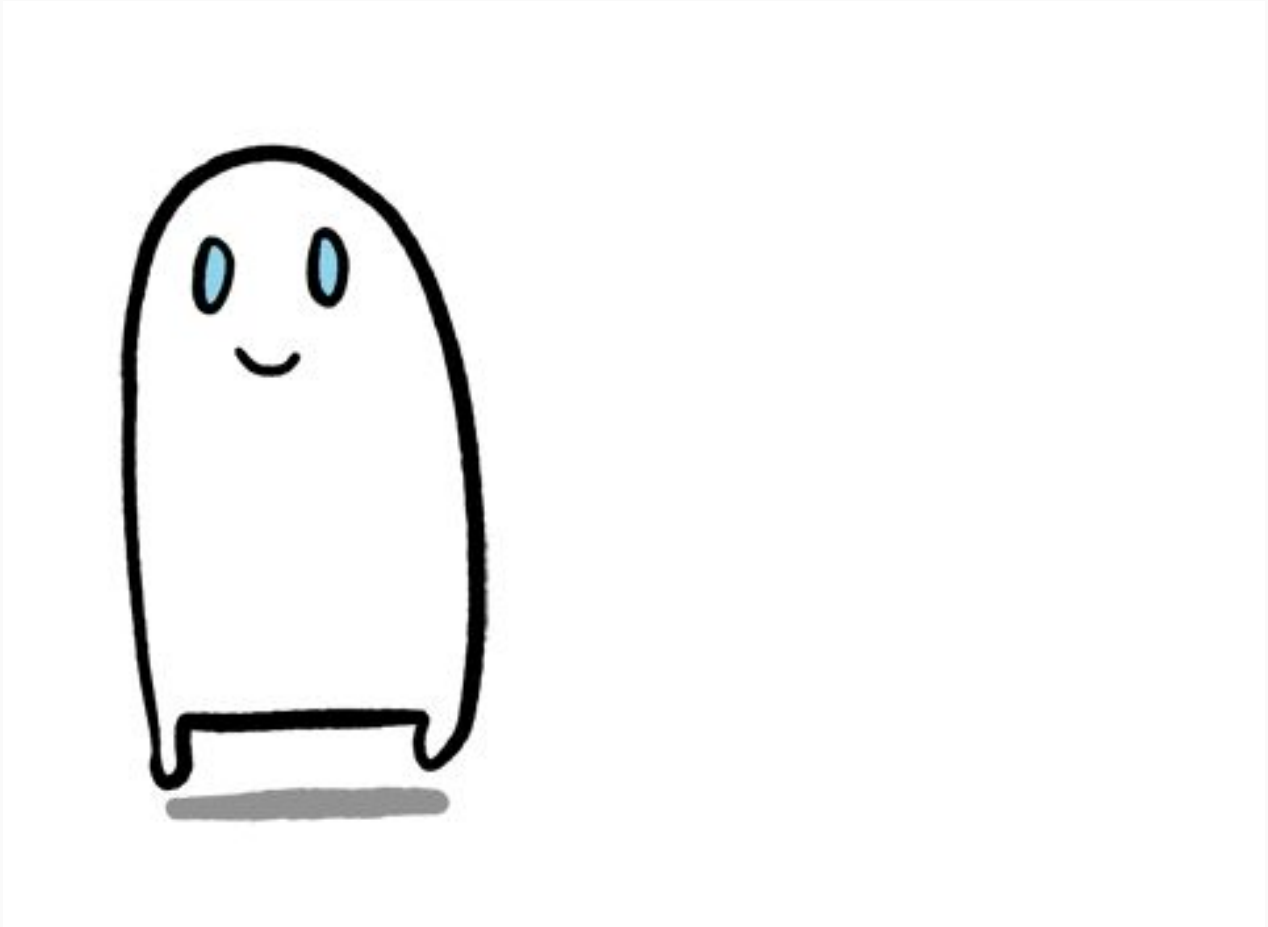
How much data was produced in 2023?

~120 zettabytes*

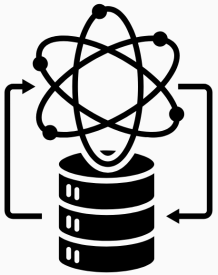
*1 zettabyte = 1 trillion gigabytes
Source: Edge Delta

“The future of business belongs to people who can make sense of large quantities of data.”

- *Hal Varian, Chief Economist at Google*



About me

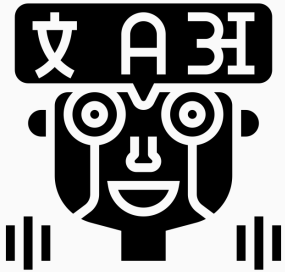


Created by Sugiman
from Noun Project

Data Scientist at IBM in San Francisco

Created by feri ulan taufiq
from Noun Project

User Experience Insights team



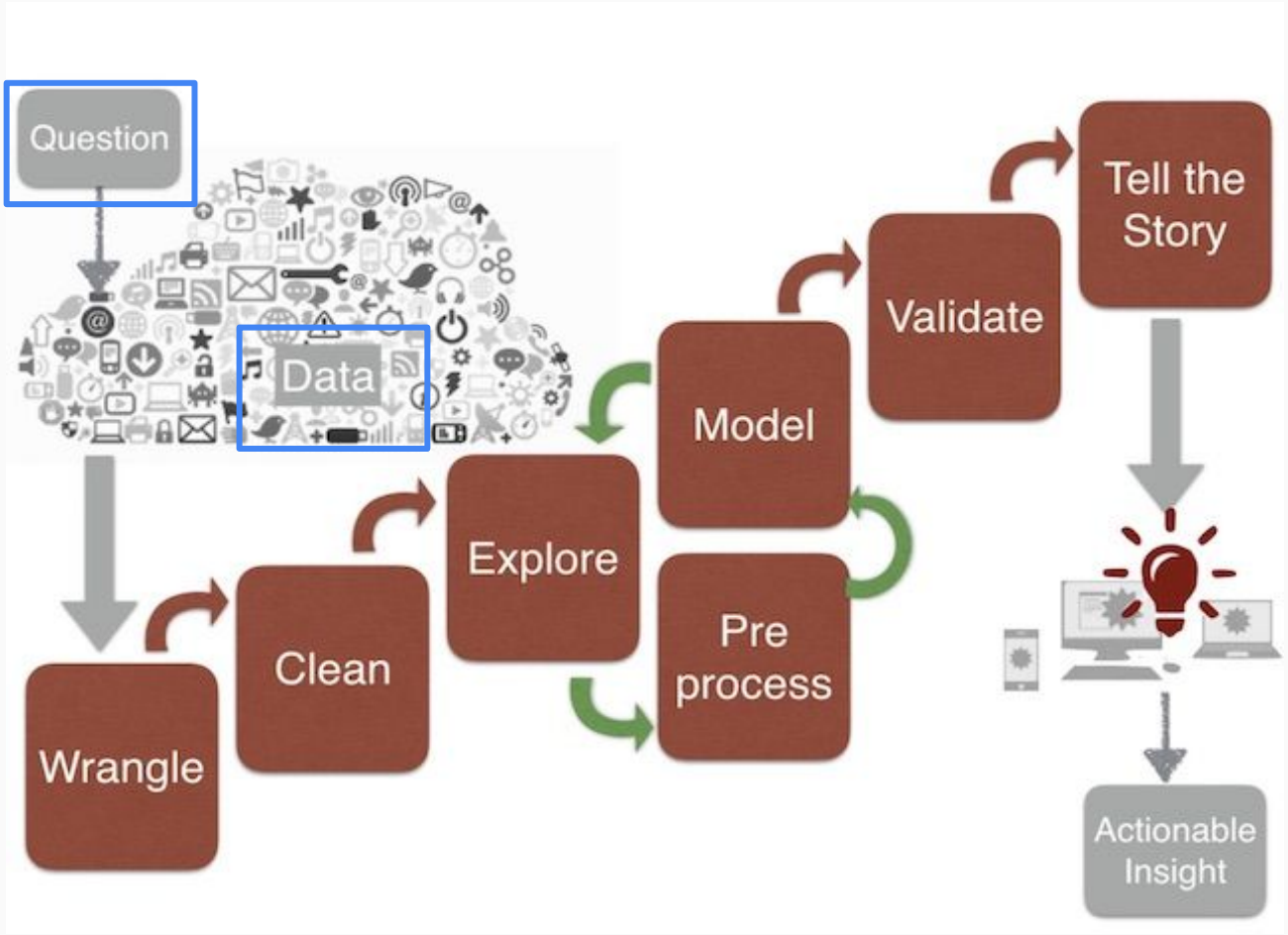
Created by Eucalyp
from Noun Project

Machine Learning and Natural Language Processing

Created by putrakali735
from Noun Project

Community building and tech advocacy


YOU MEAN TO TELL ME
DATA SCIENCE IS MORE THAN A BUZZWORD



Source: Wolfram

Data Science pipeline

Question and data sources

Formulate a question the stakeholder is trying to answer

Structured
CSV, database

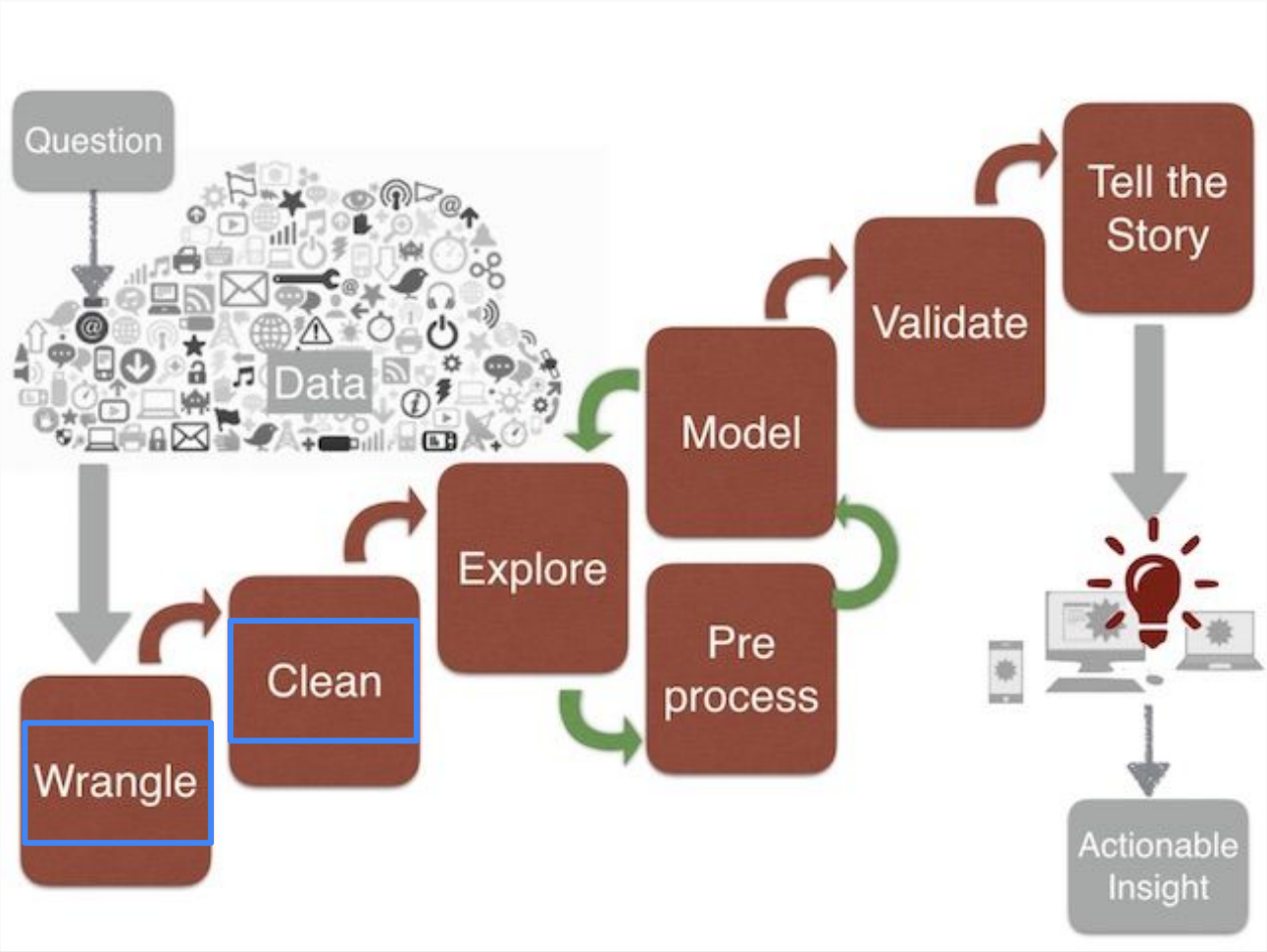
Unstructured
Audio, video, documents

Semi-structured
JSON, web pages

Data sources

Structured
CSV, database

Semi-structured
JSON, web pages

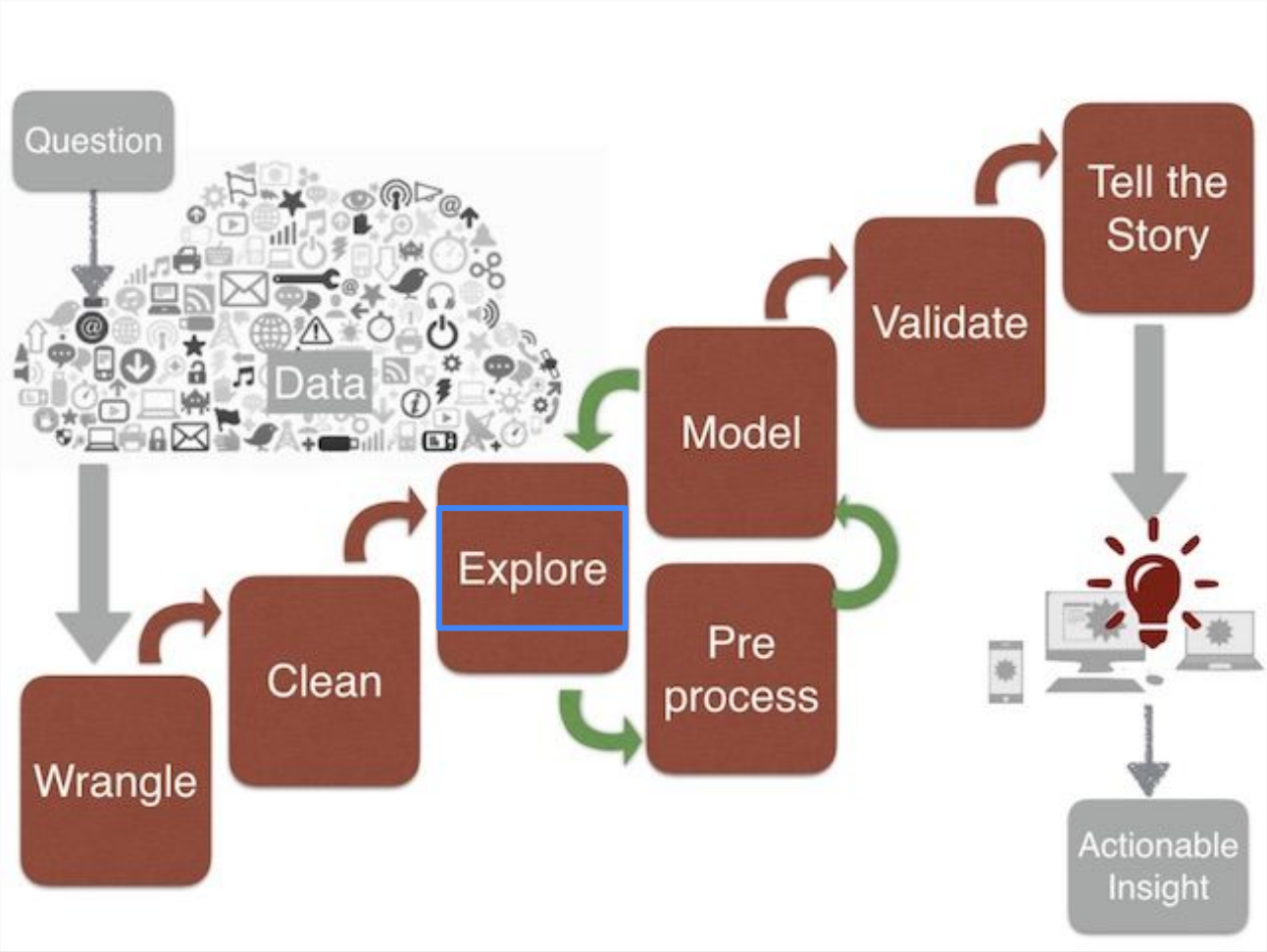


Data Science pipeline

Data preparation and cleaning

Data preparation - gathering, selecting, transforming data for easy access and analysis

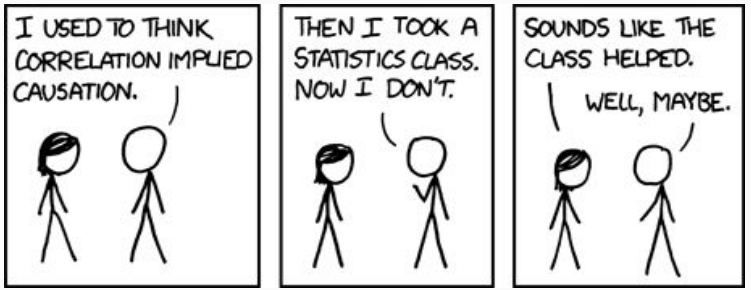
Standardize multiple sources NY=New York	Scaling or normalization	Discard or replace missing, NAs, negative values
Interpolate values	Convert categories to numbers Yes=1, No=0	Deduplicate records



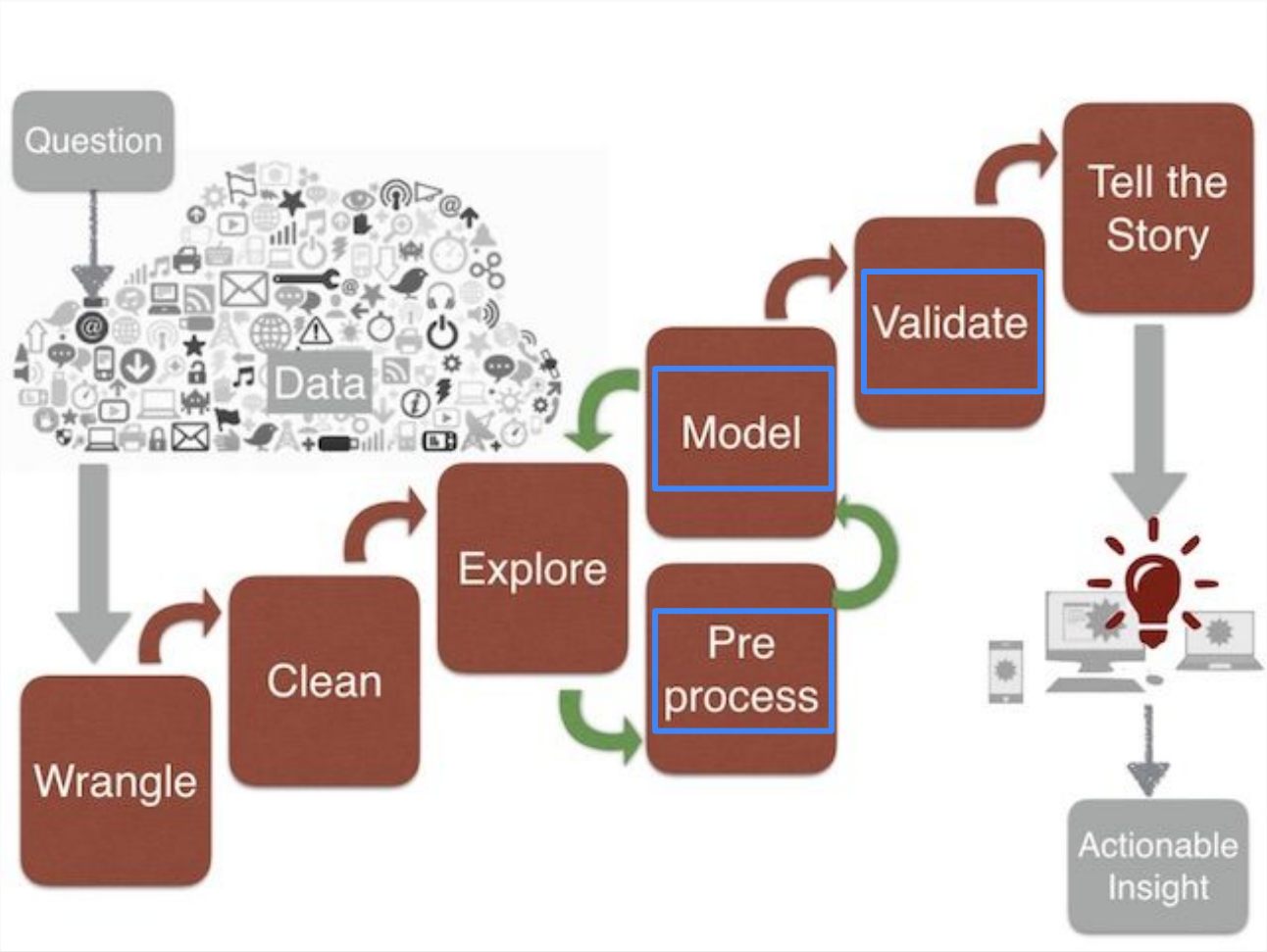
Data Science pipeline

Data exploration

- Initial investigation or Exploratory Data Analysis (EDA)
- Extract important variables
- Summarize characteristics
- Visualize properties of data using graphs
- Uncover initial patterns and points of interest
- Form hypotheses about defined problem



Source: XKCD



Data Science pipeline

Model building

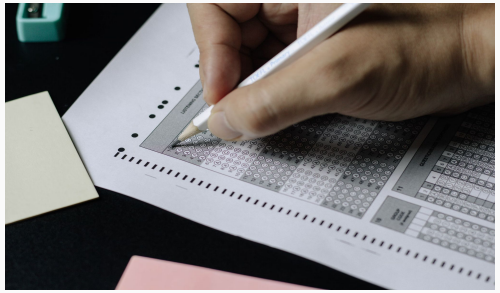
Feature engineering

Choose model and tune parameters

Train model on training data

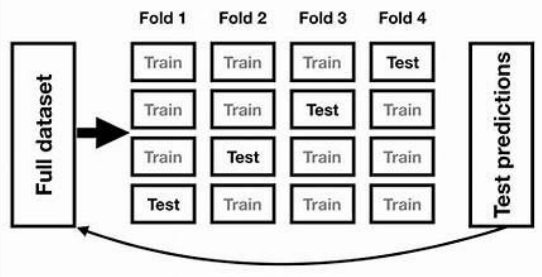
Model building and evaluation

Monitor against overfitting

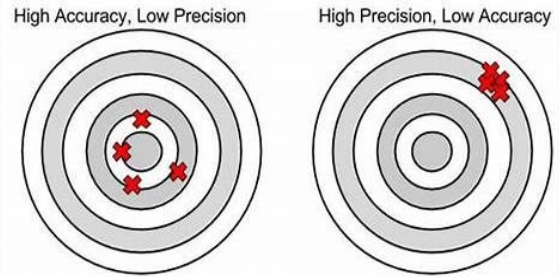


Evaluate model on unseen i.e. test data

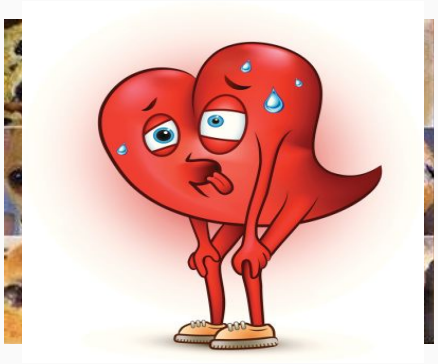
Source: XKCD



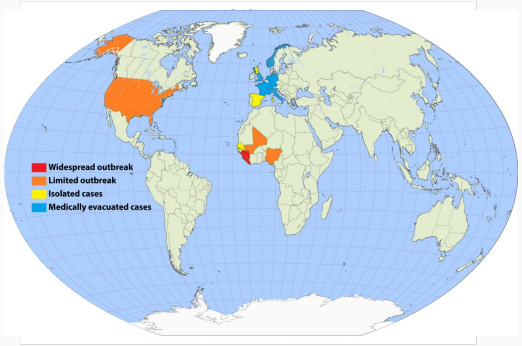
k-fold cross validation



Metrics



Classification



Regression

Clustering

Anomaly detection

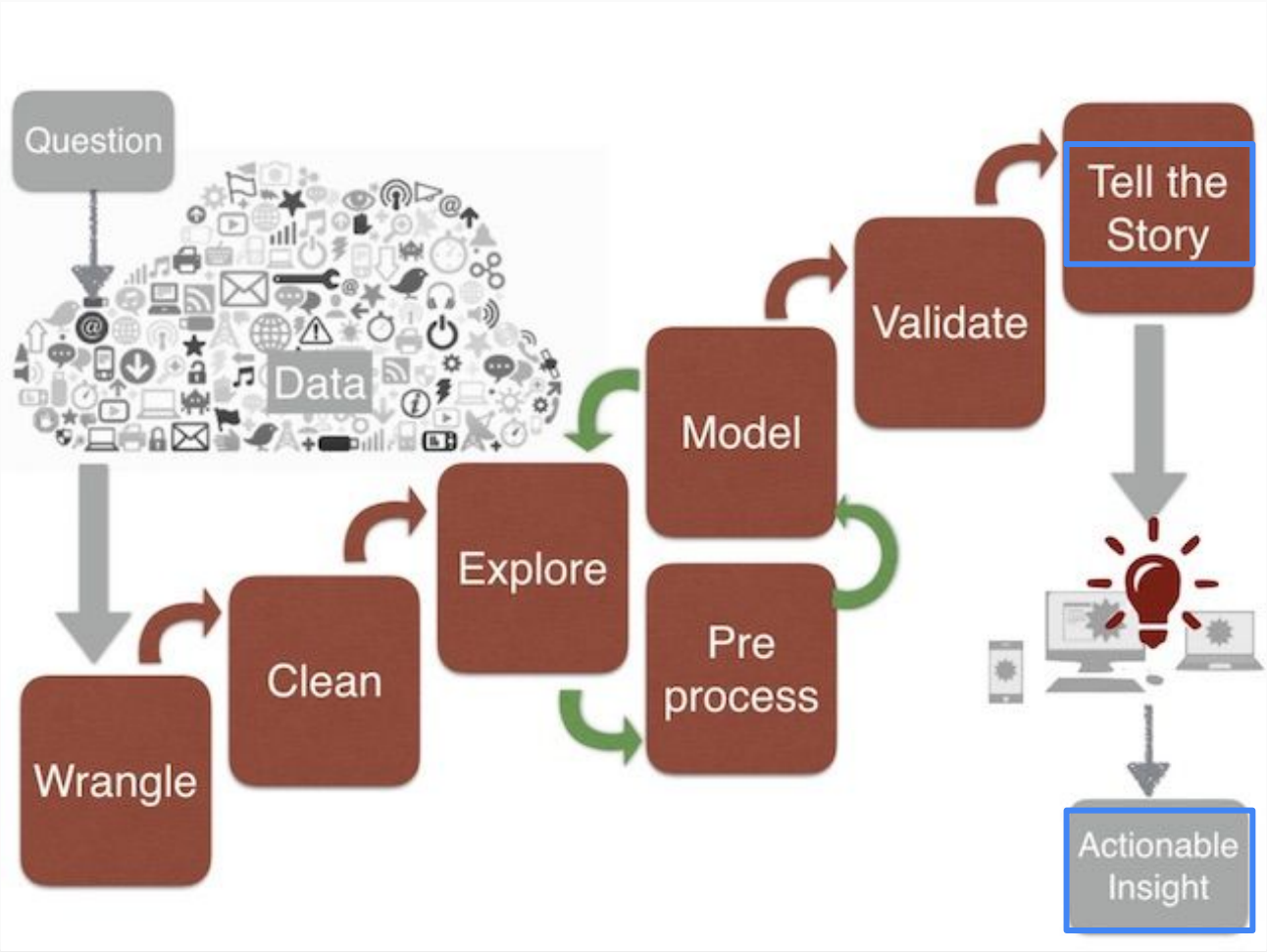
Reinforcement learning

Grishma Jena

@DebateLover

“Things get done only if the data we gather can inform and inspire those in a position to make a difference.”

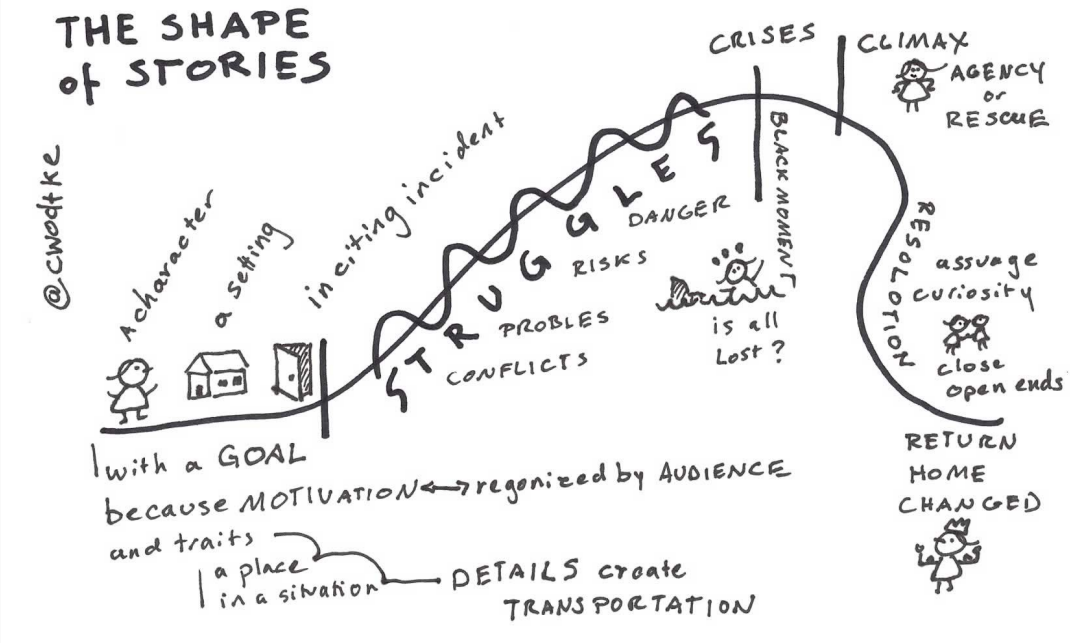
- *Dr. Mike Schmoker, Author*



Data Science pipeline

Data storytelling

Tell a compelling story with data using **characters**, **setting**, **conflict**, and **resolution**



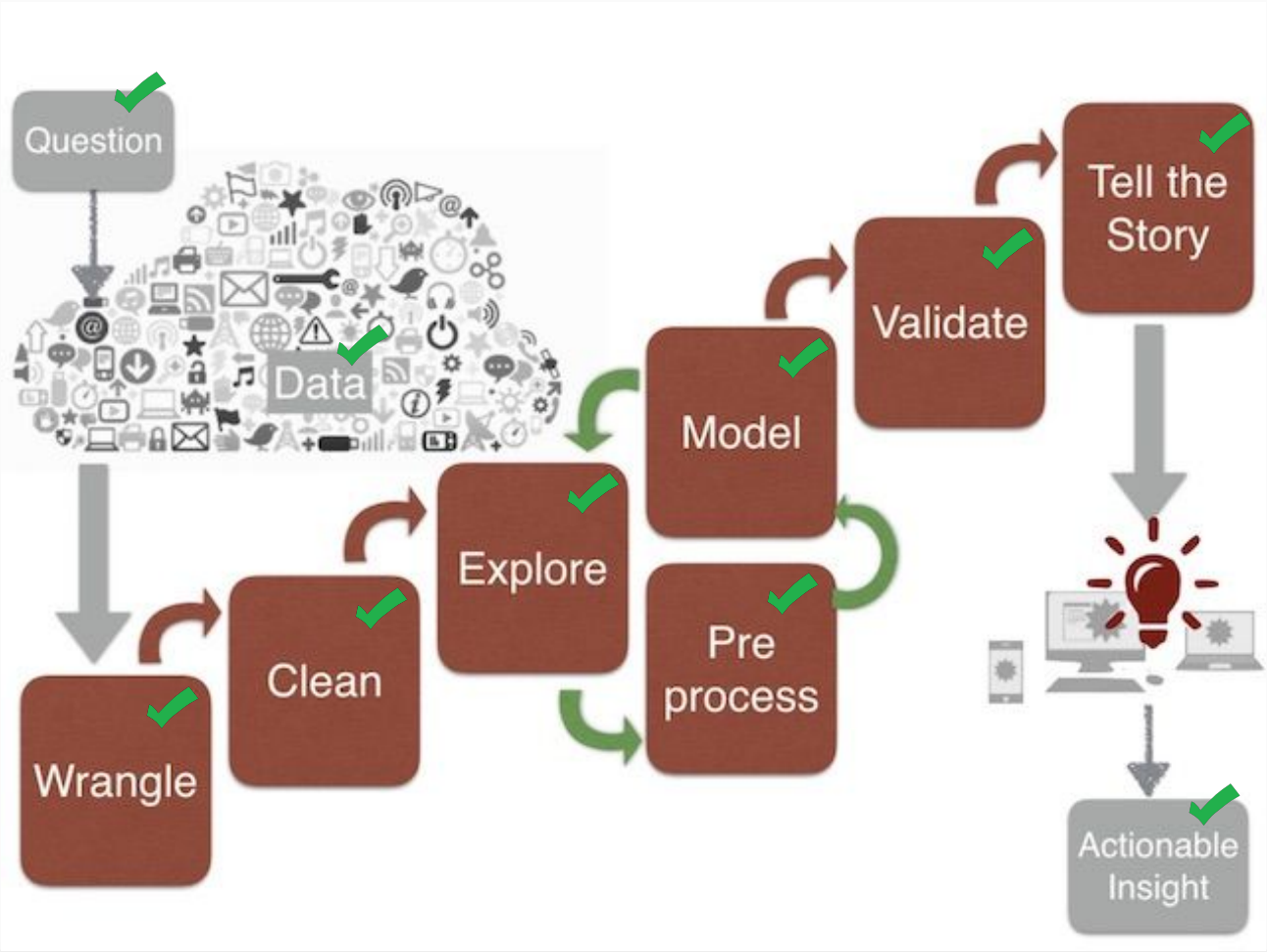
Source: Christina Wodtke's blog

Source: Narrative Science

Data storytelling

Data Exploration: Numbers and dashboards; time-consuming; generic report; can't understand right info fast

Data Storytelling: Communicate like a human; instant explanation; personalized to needs; understand important things now



Data Science pipeline

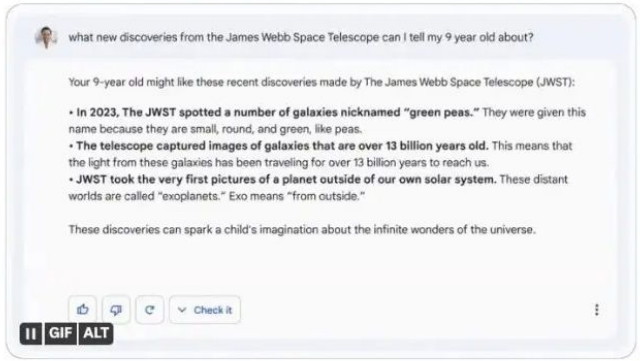
MICROSOFT / WEB / TL;DR

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

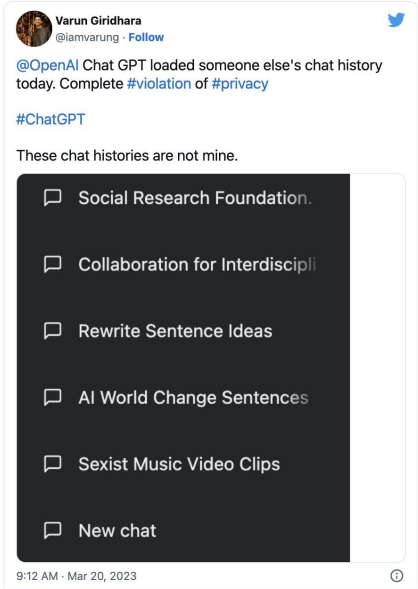
Amazon scraps secret AI recruiting tool that showed bias against women

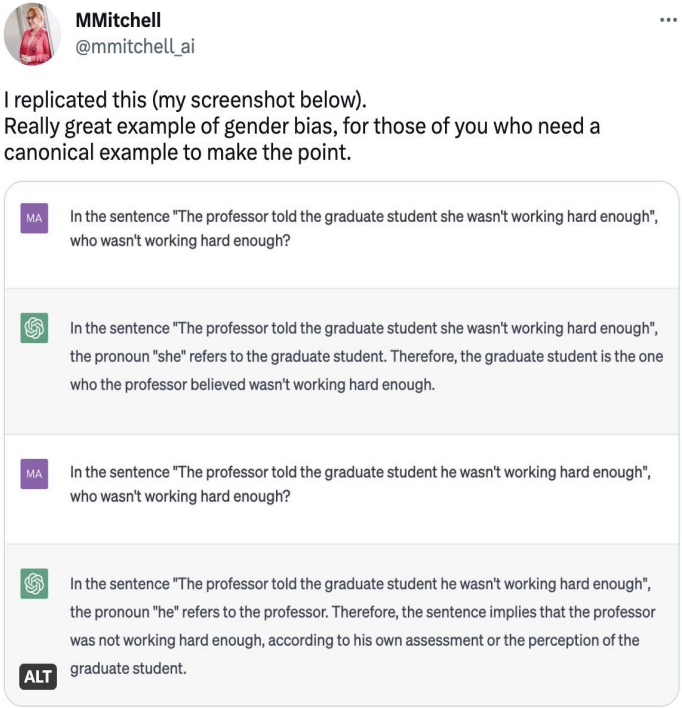
Samsung Bans Staff's AI Use After Spotting ChatGPT Data Leak

- Employees accidentally leaked sensitive data via ChatGPT
- Company preparing own internal artificial intelligence tools



Thermostats, Locks and Lights: Digital Tools of Domestic Abuse





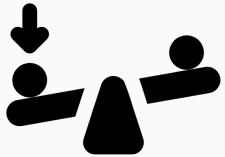
“All data is created by people. And all people create data...Today we divorce people from their data, and that gives companies license to forget about the people behind the data...It allows us to divorce ourselves from the responsibility of what that data can do.”

– Ovetta Sampson, Microsoft

Ethics and responsibility in Data Science

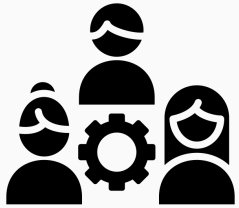
Created by Wan Ikhsan from Noun Project

Explore attacks or misuse



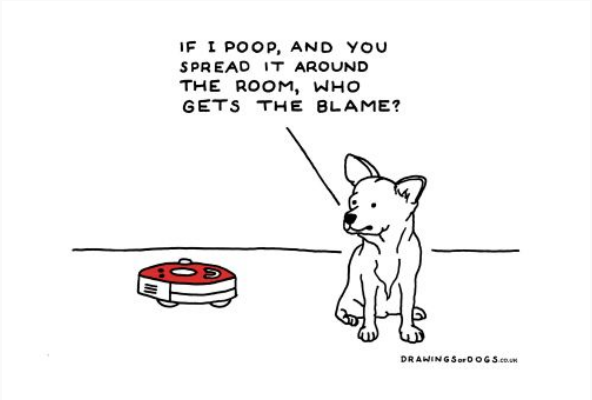
Created by komsiatun from Noun Project

Is the data fair and representative? What biases exist?



Created by Rusma Ratri Handini from Noun Project

Diverse team



Created by SeeMoo from Noun Project

Explicit consent + data protection

Created by Indriani from Noun Project

Fairness over time, for all

Created by Magicon from Noun Project

Shut down and redress if harm

Fairness, accountability and transparency in algorithms is important

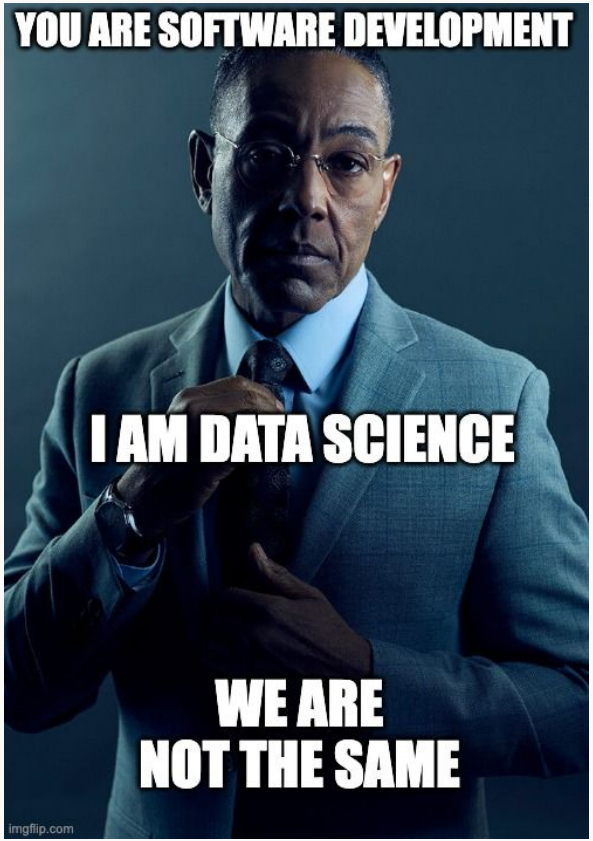
Checklist by Mike Loukides, Hilary Mason, DJ Patil

Data Science is unlike Software Development

Traditional agile principles don't apply to experimental and indeterministic nature of Data Science

"Data is the language your customers are using to talk back to you."

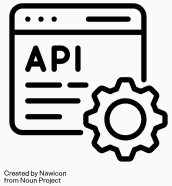
– Rochelle King, VP Netflix, ex-Spotify



Best practices

Data and Infrastructure

Data contracts



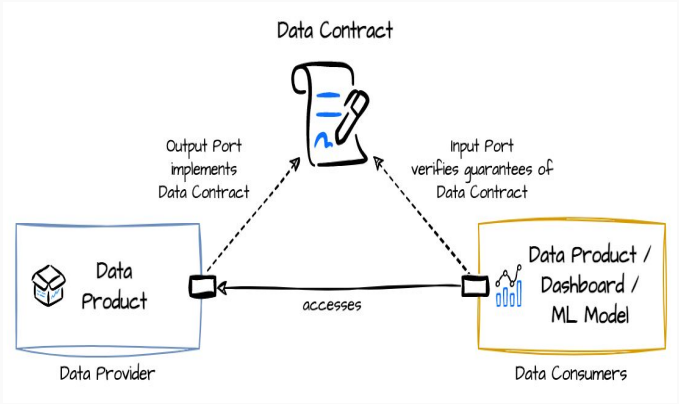
Vendors with APIs

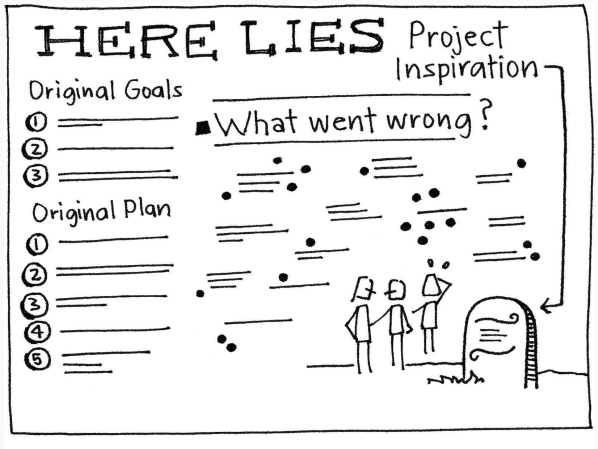
Governance and documentation

Culture

Pre-mortem

Encourage failure





www gjena.github.io

grishmajena

DebateLover

data_designtist

Thank you