GBDi
Government Big Data Institute
สถาบันส่งเสริมการวิเคราะห์และบริหารข้อมูลขนาดใหญ่ภาครัฐ (สวข.)

depa

โครงการอบรมหลักสูตร Hands-on Data Science and Machine Learning

Introduction to Data Science and Analytic Thinking

18 พ.ค. 2565

ศักดิ์สิทธิ์ ศรีมะโรง

Data Scientist, GBDi

Course Overview (1)

1. Introduction to data science and analytic thinking

- ความรู้เบื้องต้นเกี่ยวกับการวิเคราะห์ข้อมูล รวมถึงการอภิปรายเทคนิคทั่วไปที่สำคัญ สำหรับการวิเคราะห์และแปลงข้อมูลให้ได้มาซึ่งสารสนเทศที่มีความหมายจากชุดข้อมูลต่างๆ อธิบายหลักการที่เกี่ยวข้องกับการดำเนินการเก็บรวบรวมข้อมูล การแปลงข้อมูล การทำความสะอาดข้อมูล (Data cleaning and integration) และการตัดสินใจจากข้อมูล โดยเน้นการคิดเชิงวิเคราะห์

2. Data science with basic Python programing or R programing

- สอนการเขียนโปรแกรมพื้นฐานเฉพาะเพื่อการวิเคราะห์ข้อมูล โดยการพัฒนาความรู้ความสามารถในการเขียนโปรแกรมภาษา Python หรือภาษา R ซึ่งเป็นภาษาที่ใช้กันอย่างแพร่หลาย

3. Introduction to statistics

- 3.1.Descriptive statistics: ความรู้เบื้องต้นเกี่ยวกับสถิติเชิงพรรณนา ได้แก่ ค่าเฉลี่ย มัธยฐาน ฐานนิยม ความแปรปรวน และค่าเบี่ยงเบนมาตรฐาน
- 3.2.Basic statistical inference : ความรู้เบื้องต้นเกี่ยวกับสถิติเชิงอนุมาน การประเมินค่าพารามิเตอร์ในประชากร การทดสอบสมมติฐาน ความคลาดเคลื่อนในการทดสอบสมมติฐาน การประมาณแบบช่วง การประเมินคุณสมบัติและการควบคุมความผิดพลาดของการทดสอบสมมติฐาน ตัวอย่างเนื้อหาเช่น Frequency Distributions, Central Tendency, Correlation, Hypothesis Testing, ANOVA

Course Overview (2)

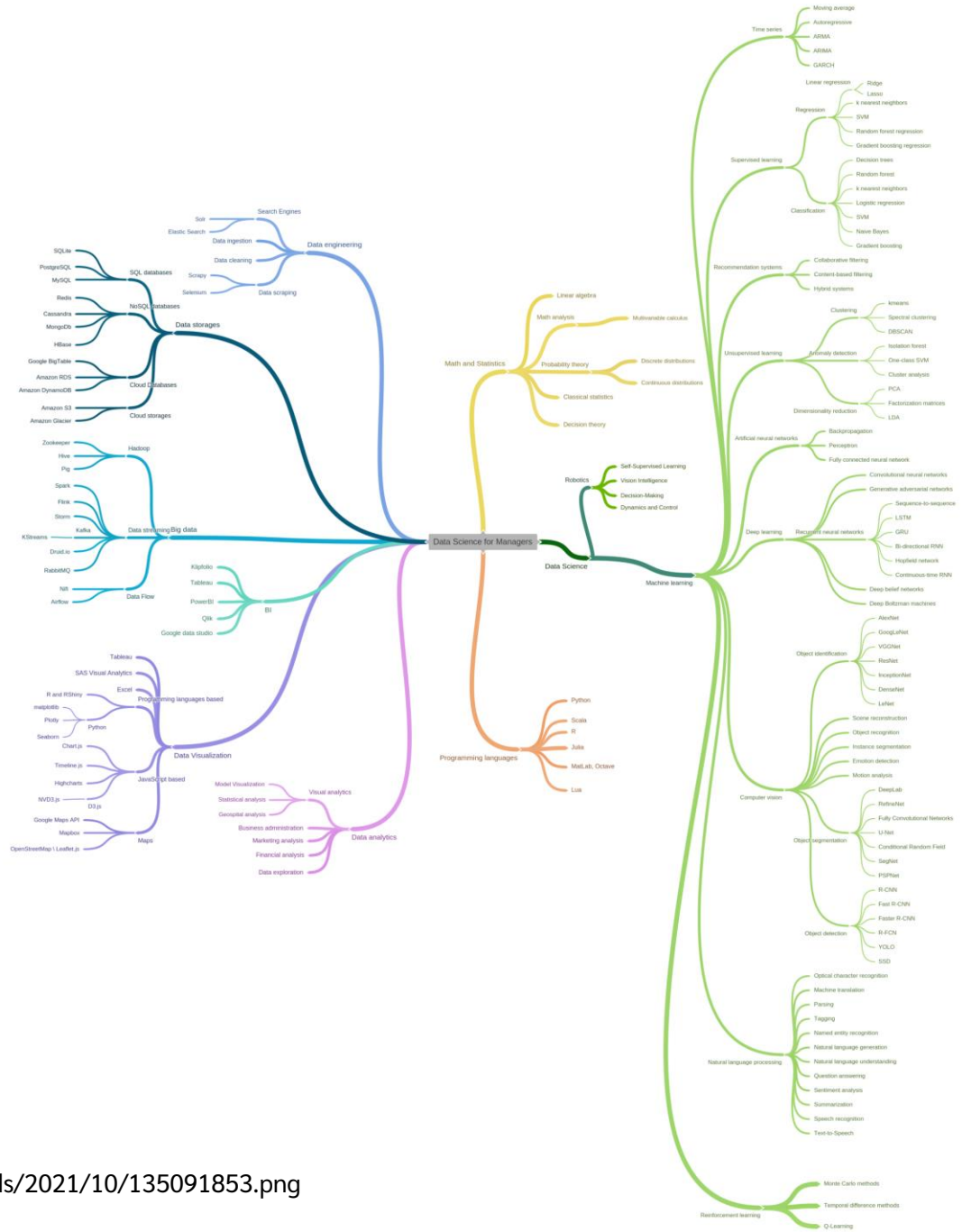
4. Exploratory data analysis

- การวิเคราะห์ข้อมูลเบื้องต้น ประกอบด้วย การตรวจสอบ ความผิดพลาดในการกรอกข้อมูลค่าสูญหาย (Missing value)ค่าผิดปกติ (Outlier)การแจกแจงของข้อมูล ความเท่ากันของความแปรปรวน (Homogeneity of variance or equality of variance) ความสัมพันธ์เชิงเส้นตรง ภาวะร่วมเส้นตรงพหุ (Multicollinearity) และส่วนเหลือ (Residual) รวมทั้งการนำเสนอข้อมูล การแสดงความถี่ การแสดงการเปรียบเทียบและแนวโน้ม การแสดงการจัดลำดับ

5. Basic machine learning

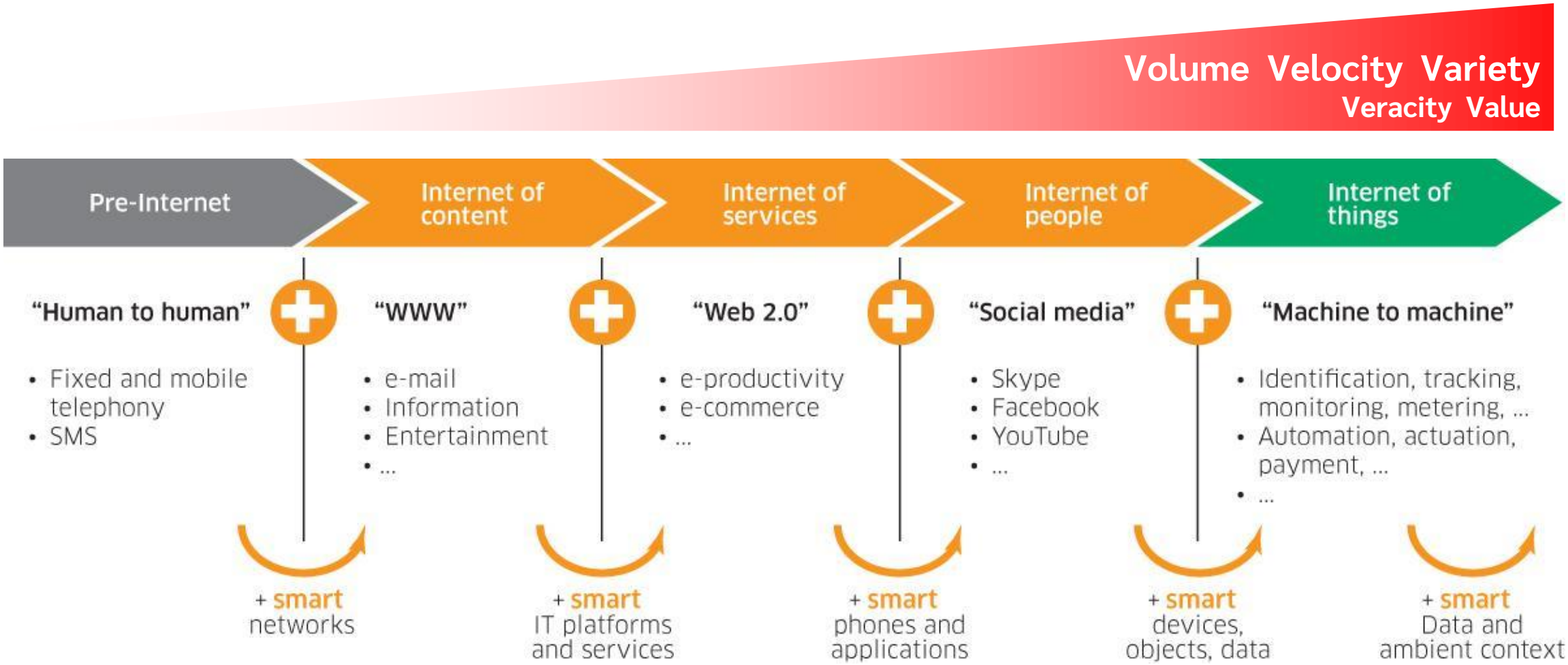
- ความรู้เบื้องต้นเกี่ยวกับการเรียนรู้ของเครื่อง ขั้นตอนที่จำเป็นสาหรับการสร้างแบบจำลองการเรียนรู้ของเครื่อง แนวคิดการเรียนรู้ด้วยเครื่องและประเภทการเรียนรู้แบบมีผู้สอน (supervised learning) และการเรียนรู้แบบไม่มีผู้สอน (unsupervised learning) การประเมินขั้นตอนวิธีการเรียนรู้ของเครื่อง การวิเคราะห์ที่ไม่ขึ้นกับขั้นตอนวิธี อัลกอริทึมสำหรับการจัดแบ่งประเภท เรียนรู้คลังโปรแกรมที่ใช้อย่างแพร่หลายของอัลกอริทึมสำหรับการเรียนรู้ด้วยเครื่อง การเตรียมประมูลผลข้อมูลเพื่อสร้างชุดข้อมูล(Data Set)ที่มีคุณภาพ การบีบอัดเพื่อลดมิติของข้อมูล การประเมินผลแบบจำลองและการปรับแต่งพารามิเตอร์
- 5.1.Linear regression (Generalized linear model, Linear regression, Poisson regression, Survival analysis)
- 5.2.Classification (Decision tree, Support vector machine, Random forest)
- 5.3.Clustering (kNN, Hierarchical clustering, k-mean clustering)
- 5.4.Co-occurrence analysis (Content-based filtering, Collaborative filtering)
- 5.5.Performance evaluation and cost-benefit (Accuracy, ROC, RMSE, Cost-benefit)

Contents

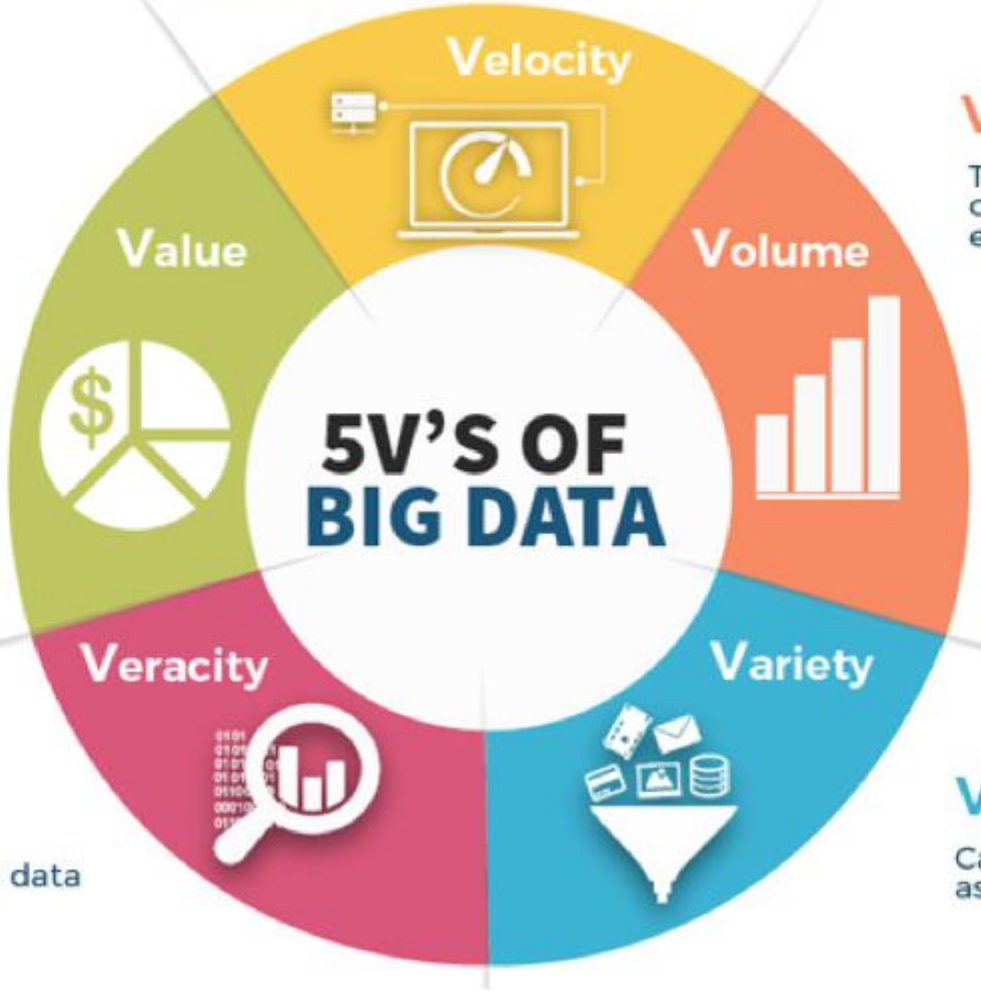


Data Science for Managers
https://www.datasciencecentral.com/wp-content/uploads/2021/10/135091853.png

Evolution of Digital & Internet



Source: https://twitter.com/nokiacem/status/581161700184887296/photo/1



Velocity

Speed at which data is emanating and changes are occurring between the diverse data sets

Volume

This refers to the sheer volume of data being generated every second.

Value

Having access to big data is all well and good but that's only useful if we can turn it into a value.

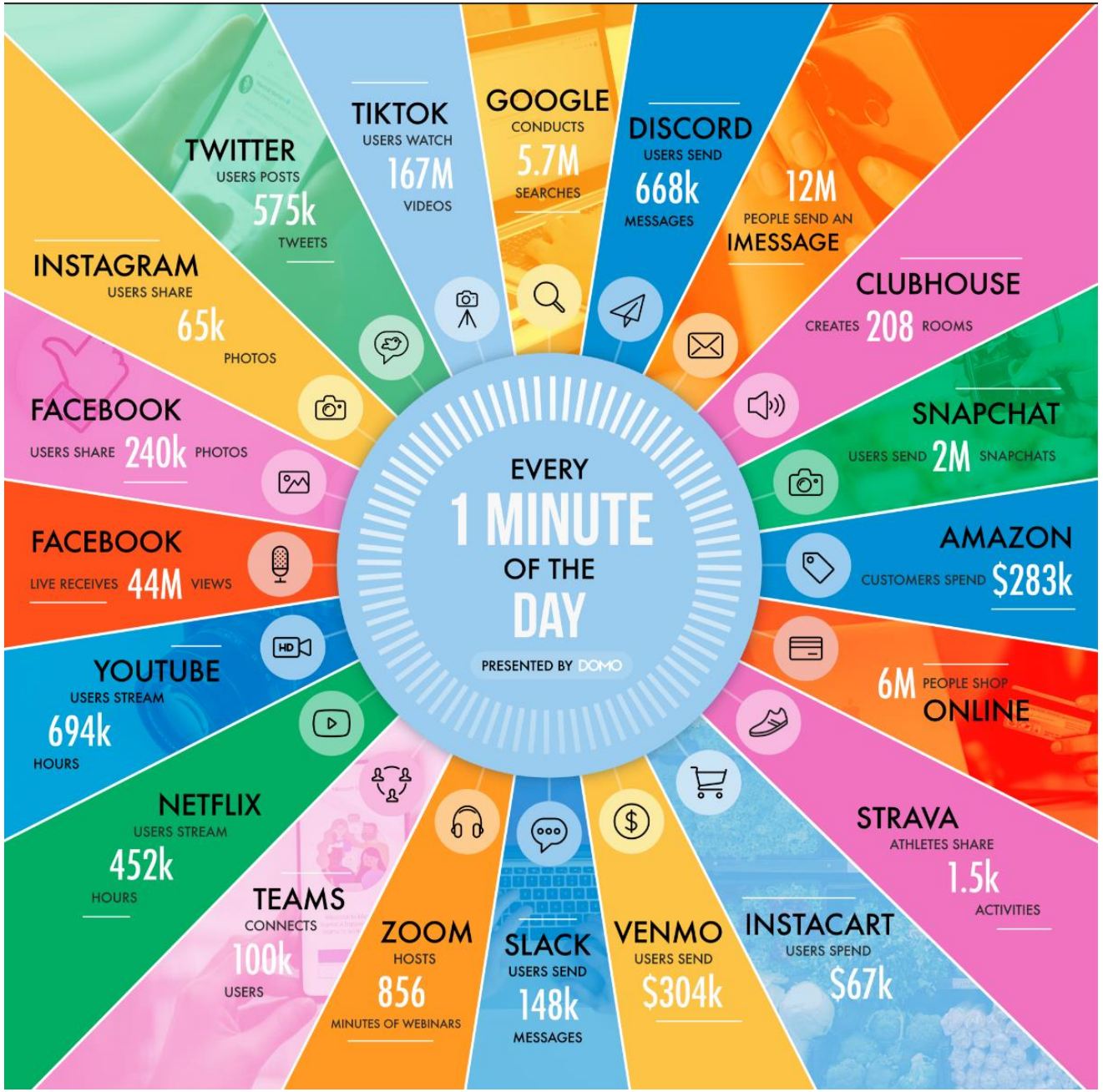
Variety

Can use structured as well as unstructured data.

Veracity

Data reliability and trust. Verifying and validating the data

https://www.techentice.com/the-data-veracity-big-data/



DOMO

Data Never Sleeps 9.0

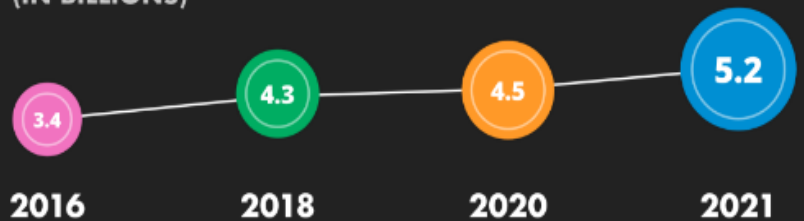
How much data is generated *every minute*?

The 2020 pandemic upended everything, from how we engage with each other to how we engage with brands and the digital world. At the same time, it transformed how we eat, how we work and how we entertain ourselves. Data never sleeps and it shows no signs of slowing down. In our 9th edition of the "Data Never Sleeps" infographic, we bring you a glimpse of how much data is created every digital minute in our increasingly data-driven world.

As of July 2021, the internet reaches 65% of the world's population and now represents 5.17 billion people—a 10% increase from January 2021. Of this total, 92.6 percent accessed the internet via mobile devices. According to Statista, the total amount of data consumed globally in 2021 was 79 zettabytes, an annual number projected to grow to over 180 zettabytes by 2025.

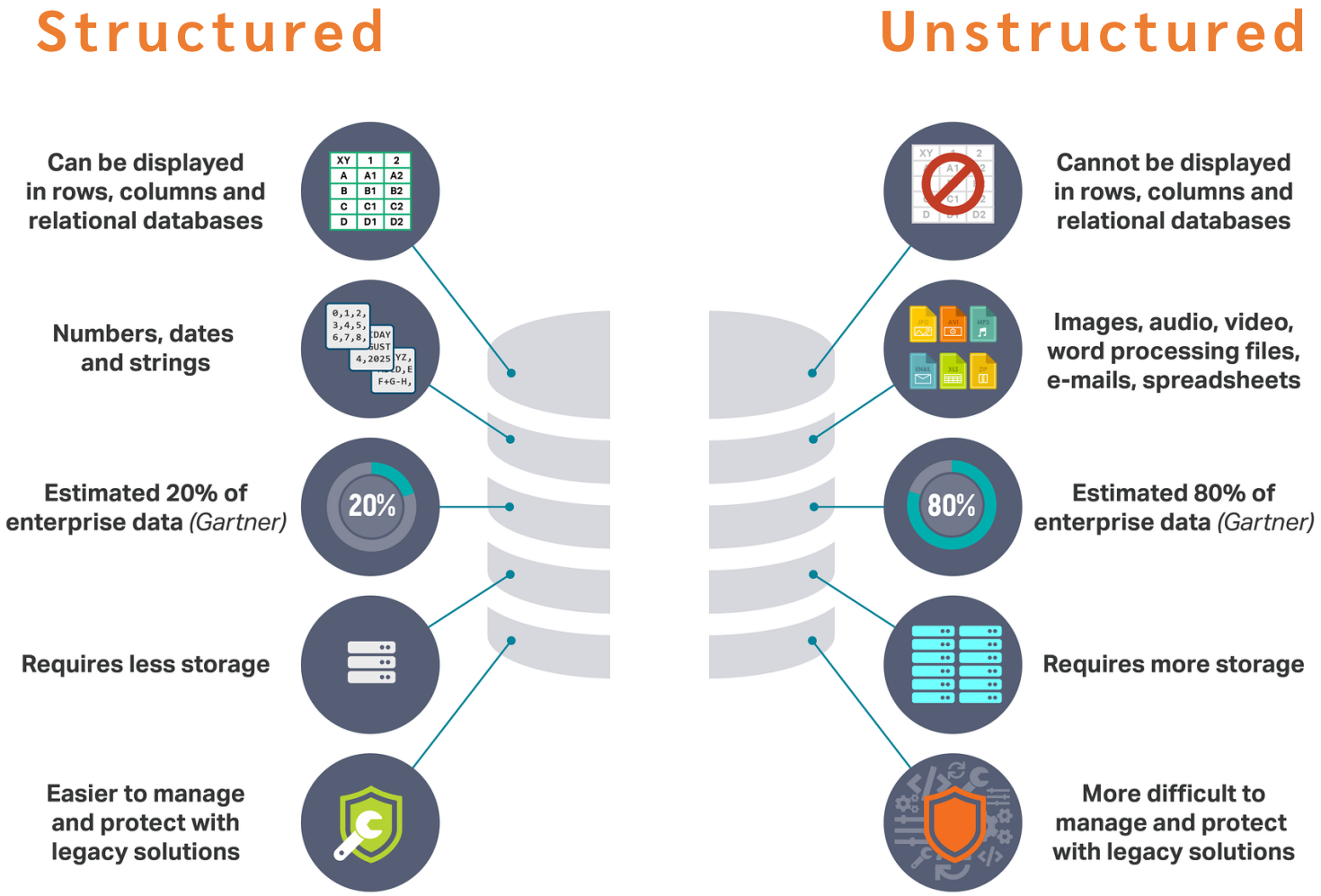
Global Internet Population Growth

(IN BILLIONS)



Source: https://www.visualcapitalist.com/from-amazon-to-zoom-what-happens-in-an-internet-minute-in-2021/

Big Data - *An umbrella term for all sorts of data*



Source: https://lawtomated.com/structured-data-vs-unstructured-data-what-are-they-and-why-care/photo/1

What is Big Data Analytics?

- Examines large and different types of data to uncover hidden patterns, correlations and other insights.
- Largely used by companies to facilitate their growth and development.
- Majorly involves applying various data mining algorithms on the given set of data, which will then aid them in better decision making.

https://www.edureka.co/blog/big-data-analytics

Benefit of Big Data Analytics

1. Making smarter and more efficient organization

New York Police Department is utilizing data patterns, scientific analysis, and technological tools to prevent the occurrence of crime

Use historical arrest patterns and then maps them with events such as federal holidays, paydays, traffic flows, rainfall

https://www.edureka.co/blog/big-data-analytics

Benefit of Big Data Analytics

2. Optimize business operations

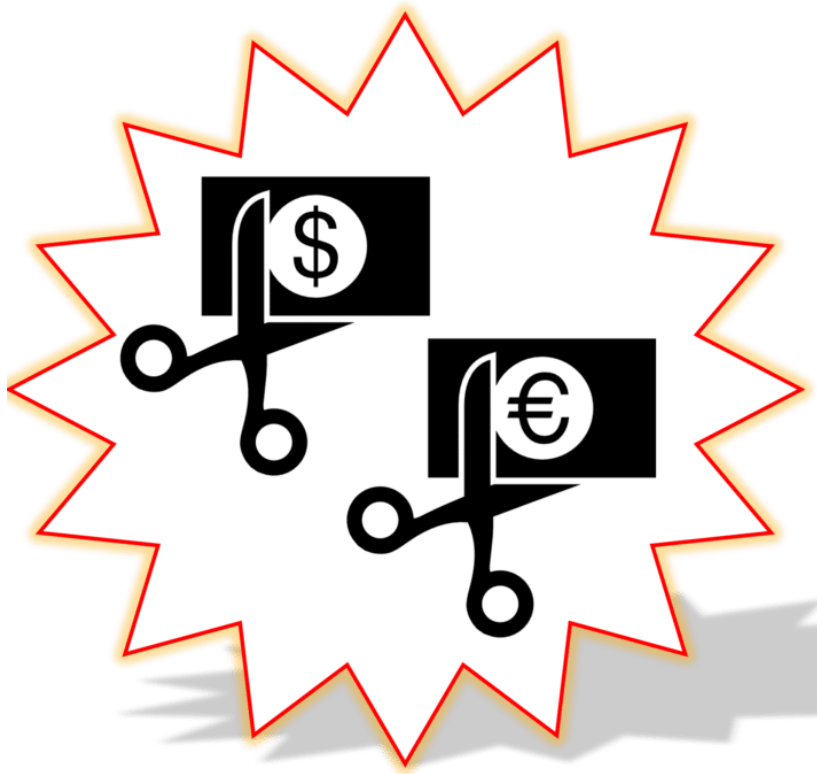
- Analysing all the clicks of every visitor on a website
- Studying the paths leading them to buy products
- Customer Satisfaction

Amazon uses customer click-stream data and historical purchase data of more than 300 million customers and each user is shown customized results on customized web pages.

https://www.edureka.co/blog/big-data-analytics

Benefit of Big Data Analytics

3. Reduce cost



Parkland Hospital uses analytics and predictive modelling to identify high-risk patients and predict likely outcomes once patients are sent home. As a result, Parkland reduced 30-day readmissions for patients with heart failure, by 31 percent, saving $500,000 annually.

https://www.edureka.co/blog/big-data-analytics

Benefit of Big Data Analytics

4. Generate new products

Big Data tools are used to operate Google's Self Driving Cars. The Toyota Prius is fitted with cameras, GPS as well as powerful computers and sensors to safely drive on the road without the intervention of human beings.

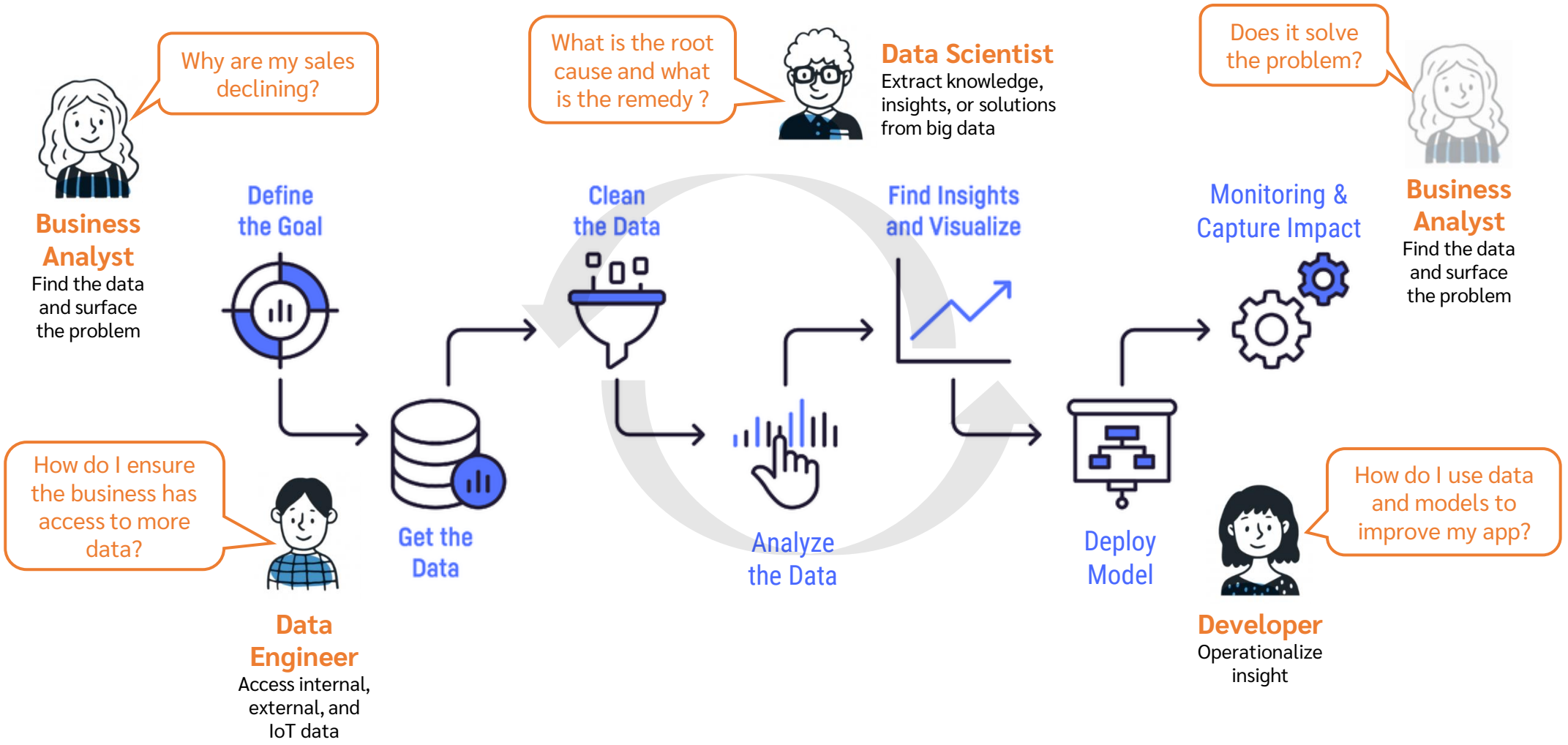
Netflix launched the seasons of its TV show House of Cards based on the user reviews, ratings and viewership.

NETFLIX

A smart yoga mat has sensors embedded in the mat will be able to provide feedback on your postures, score your practice, and even guide you through an at-home practice.

https://www.edureka.co/blog/big-data-analytics

Big Data Analytics Process & Personas



Source: https://namitkabra.wordpress.com/2016/12/05/the-4-personas-for-data-analytics/
https://blog.dataiku.com/2019/07/04/fundamental-steps-data-project-success

Answering Business Questions

Who are the most profitable customers?

A straightforward database query, if “profitable” can be defined clearly

Is there really a difference between the profitable customers and the average customer?

Statistical hypothesis testing

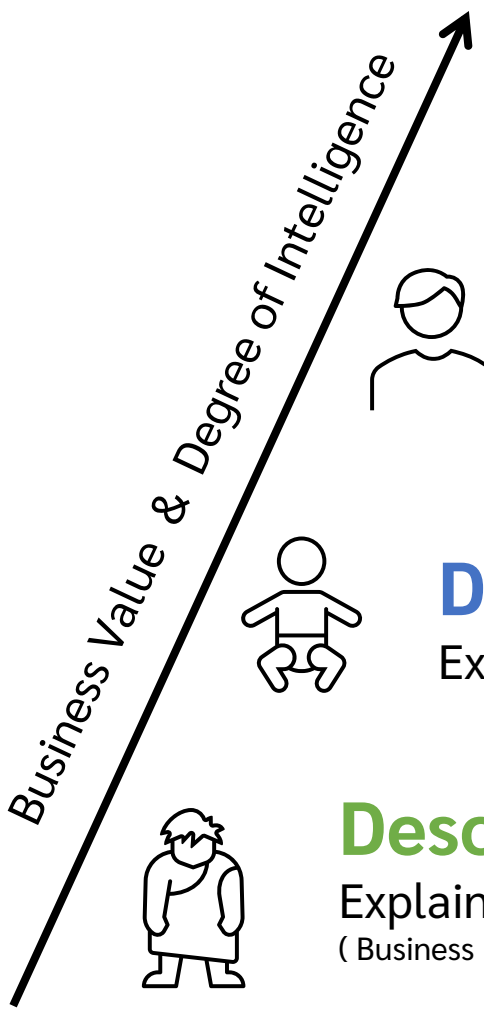
But who really are these customer? Can I characterize them?

Automated pattern finding

Will some new customer be profitable? How much revenue can I expect?

Predictive model of profitability

Levels of Analytics



Prescriptive

Recommend an action based on the forecast.

Optimization
What is the best that can happen?

Predictive

Forecasts what might happen.

Predictive Modeling
What will happen next?

Forecasting
What if these trends continue?

Diagnostic

Explains why it happened.

Statistical Analysis
Why is this happening?

Descriptive

Explains what happened?
(Business Intelligence: BI)

Query drilldown & Alerts
What exactly is the problem?

Ad-hoc reports
How many, How often, Where?

Standard reports
What happened?

Data Scientist

python R rapidminer

Data Analyst

X tableau Power BI

Data Scientist must-have skills

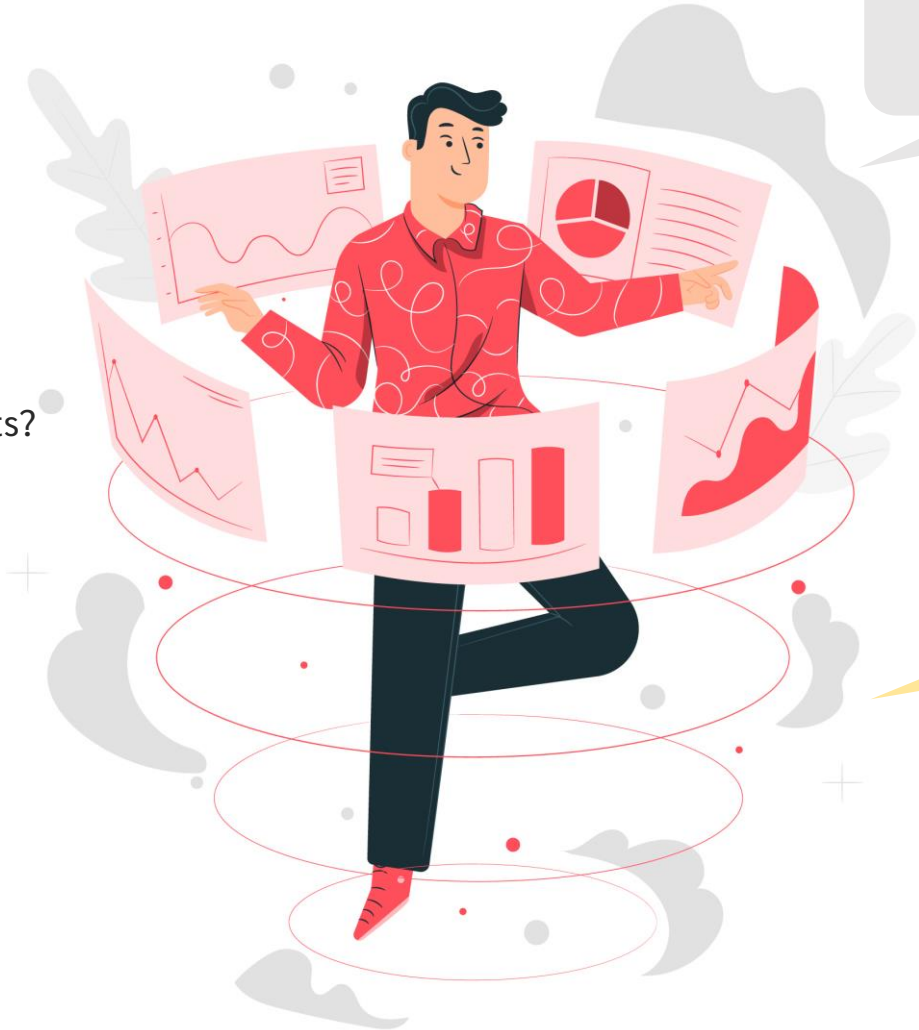
A data scientist's work typically involves making sense of messy data.

Computer Programming

- What are variables and constants? What is meant by datatype?
- What is meant by loops/conditional statements?
- What is meant by input/output/functions etc?
- What is meant by client/server/Databases/API /hosting/deployment etc

Statistics

- mean, median, mode, stranded deviation/variance, percentiles etc.
- distribution/probability/Bayes theorem etc.
- Statistical tests like — hypothesis testing, ANOVA, chi square, p-value etc.



Machine Learning Algorithms

- What is machine learning process?
- How to evaluate models?
- Regression, Classifications, Clustering
- Supervised VS. Unsupervised
- Time-series

Communication

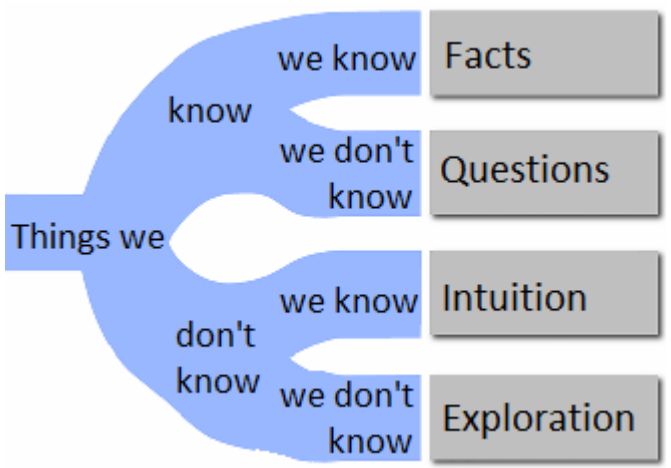
Storytelling - communicating your findings to an audience of (usually) non-data scientists.

Data Science Process (Lifecycle)

Business Objective | Data Collection | Data Cleansing | Explore Data (EDA) | Model Building | Model Evaluation | Model Deployment

"Getting the right question is the key to getting the right answer."

– Jeff Bezos



Goal

Source: USJournal

Analytics Topics

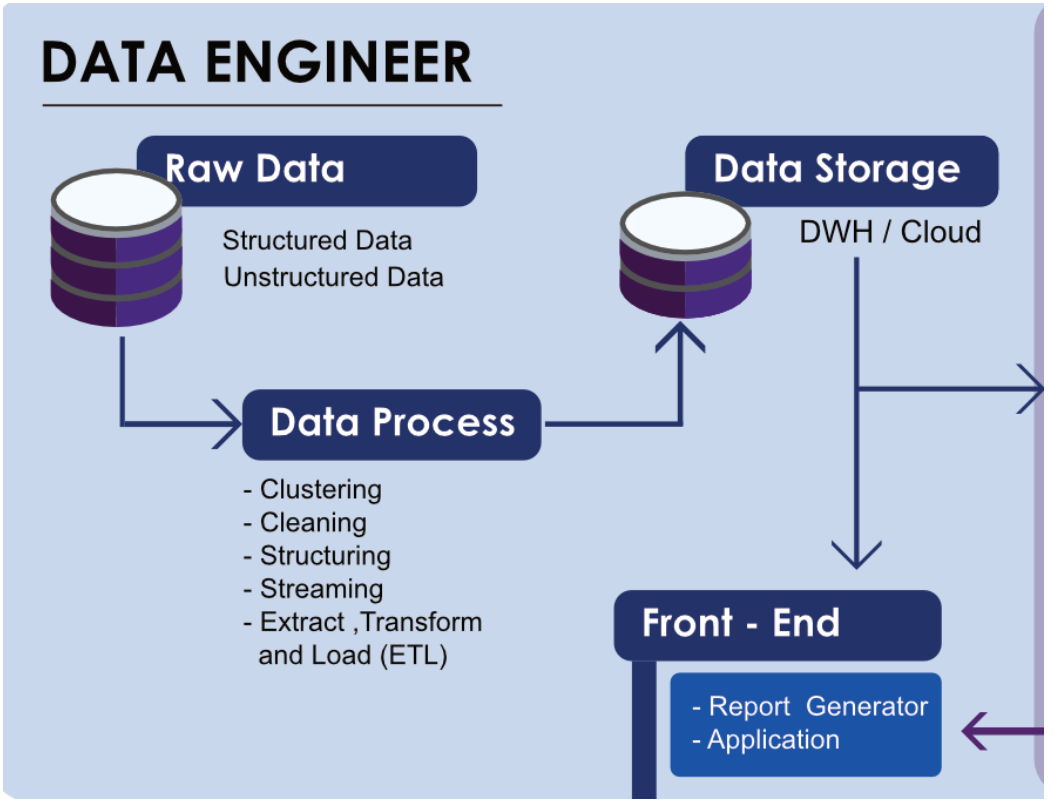
- Predicting Lifetime Value (LTV)
- Wallet share estimation
- Customer churn analysis
- Customer segmentation
- Product mix
- Cross selling
- Up selling
- Product recommendation
- Channel optimization
- Discount targeting
- Reactivation likelihood
- Target market
- Adwords optimization and ad buying
- Call center message optimization
- Call center volume forecasting
- Credit Scoring
- Treasury or currency risk
- Fraud detection
- Accounts Payable Recovery
- Anti-money laundering
- Lead prioritization
- Sales Script Analysis
- Demand forecasting
- Resume screening
- Employee churn
- Training recommendation
- Talent management

Data Science Process (Lifecycle)

Business Objective | **Data Collection** | Data Cleansing | Explore Data (EDA) | Model Building | Model Evaluation | Model Deployment

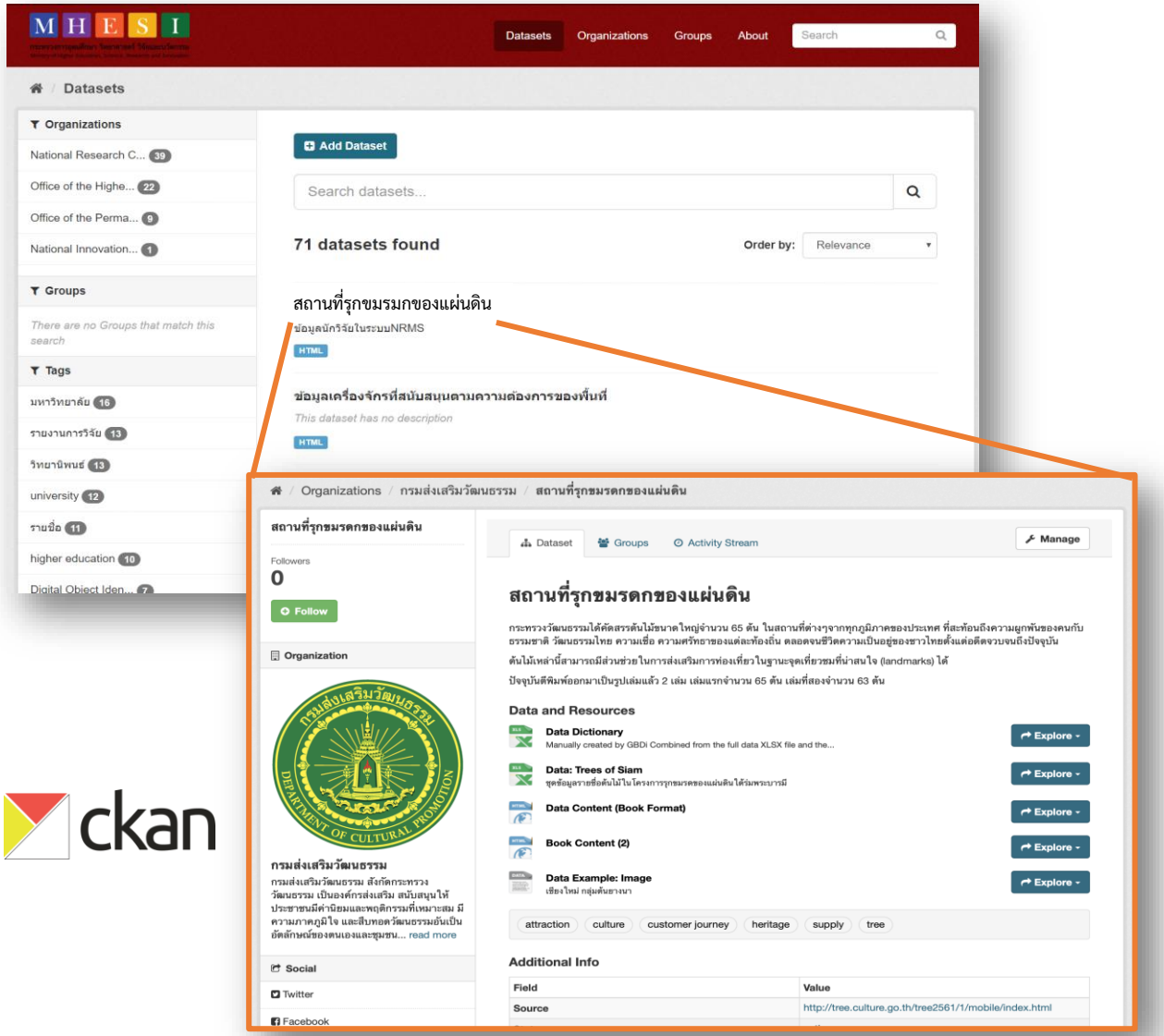
BASIS FOR COMPARISON	PRIMARY DATA	SECONDARY DATA
Meaning	Primary data refers to the first hand data gathered by the researcher himself.	Secondary data means data collected by someone else earlier.
Data	Real time data	Past data
Process	Very involved	Quick and easy
Source	Surveys, observations, experiments, questionnaire, personal interview, etc.	Government publications, websites, books, journal articles, internal records etc.
Cost effectiveness	Expensive	Economical
Collection time	Long	Short

https://researchguides.ben.edu/c.php?g=282050&p=4036581



https://medium.com/@info_46914/data-engineer-e863b37af79

Data Catalog



- Access the data catalog via data portal website.
- Search/Browse with one or multiple keywords relevant to problem
- Inspect metadata and preview the data dictionary of the dataset
- If the data from the datasets found matches the user's need, then retrieve/request the data via the method specify in the data catalog.

Data Science Process (Lifecycle)

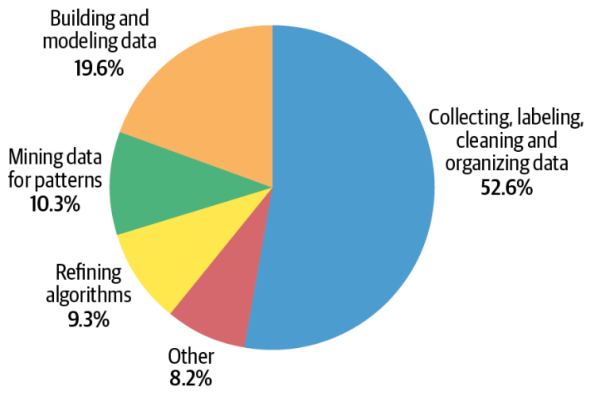
Business Objective | Data Collection | **Data Cleansing** | Explore Data (EDA) | Model Building | Model Evaluation | Model Deployment

Data cleansing or **Data wrangling** is the process of detecting and correcting corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

https://riccosmartdata.com/data-cleansing-or-data-cleaning/

#	ID	Name	LastName	Birthday	Join date	Gender	Tel	E-mail
1	1111	ธนวัฒน์	เจริญด้วยทรัพย์	1-1-1987	01/01/2019	M	860147805	username @ riccoprint.com
2	1112	วัชสัณห์	เจริญด้วยจิตใจ	07/07/1987	01/01/2019	M	087-0147-805	username2@riccoprint.com
3	1113	พลภัฒม์	กำไลอันประเสริฐ	1987/07/07	01/01/2019	M	0810147808	username.riccoprint.com
4	1114	มนวรรธน์		05/07/2530	01/01/2019	F	0820147805	username3@ riccoprint.com
5	1115	ขนมพันธ์	ผูกพันกับหนังสือ	09/12/1988	01/01/2019	A	0830147805	username4@riccoprint.com

Missing values | Formats | Invalid values | Formats | Formats



*"The **most time-consuming** part of a Data Science project is data cleaning and organizing."*

Data Science Process (Lifecycle)

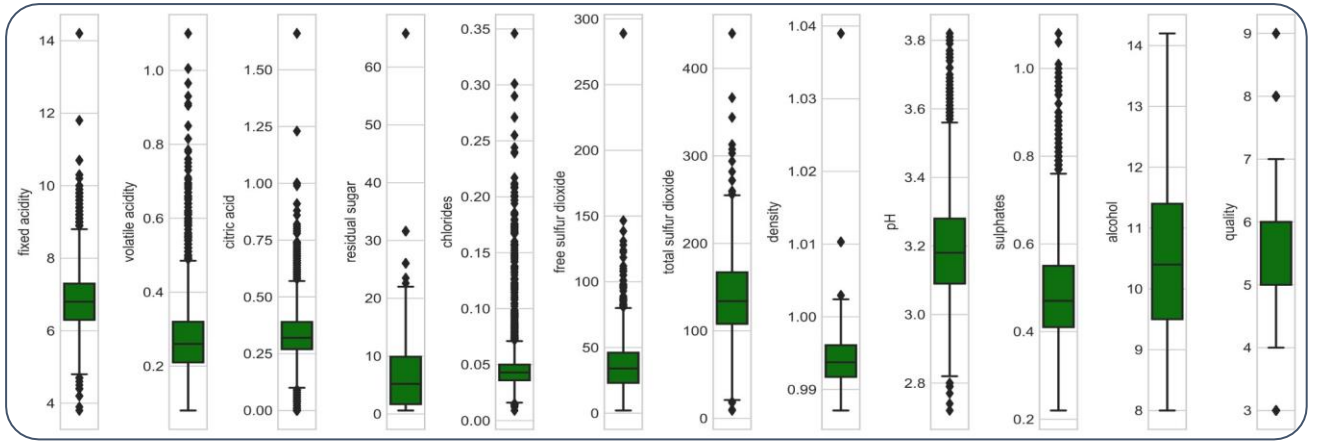
Business Objective | Data Collection | Data Cleansing | **Explore Data (EDA)** | Model Building | Model Evaluation | Model Deployment

Process of performing initial investigations on data to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

- Distinguish Attributes
- Univariate Analysis
- Multivariate Analysis
- Abnormal & Missing Values
- Outliers
- Feature Engineering

```
In [6]: df.describe()
Out[6]:
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
count	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000
mean	6.854788	0.278241	0.334192	6.391415	0.045772	35.308085	138.360657	0.994027	3.188267	0.489847	10.514267
std	0.843868	0.100795	0.121020	5.072058	0.021848	17.007137	42.498065	0.002991	0.151001	0.114126	1.230621
min	3.800000	0.080000	0.000000	0.600000	0.009000	2.000000	9.000000	0.987110	2.720000	0.220000	8.000000
25%	6.300000	0.210000	0.270000	1.700000	0.036000	23.000000	108.000000	0.991723	3.090000	0.410000	9.500000
50%	6.800000	0.260000	0.320000	5.200000	0.043000	34.000000	134.000000	0.993740	3.180000	0.470000	10.400000
75%	7.300000	0.320000	0.390000	9.900000	0.050000	46.000000	167.000000	0.996100	3.280000	0.550000	11.400000
max	14.200000	1.100000	1.660000	65.800000	0.346000	289.000000	440.000000	1.038980	3.820000	1.080000	14.200000



Data Science Process (Lifecycle)

Business Objective → Data Collection → Data Cleansing → Explore Data (EDA) → **Model Building** → Model Evaluation → Model Deployment

Data Preparation

- Missing Values
- Data Types
- One-hot Encoding
- Ordinal Encoding
- Cardinal Encoding
- Handle Unknown Levels
- Target Imbalance
- Remove Outliers

Transformation

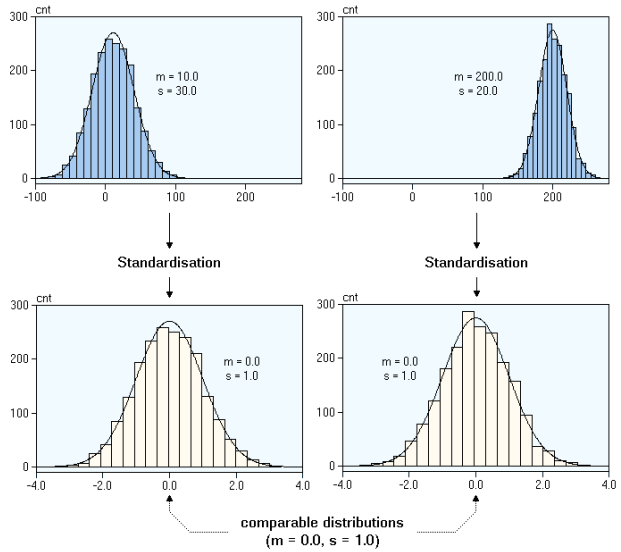
- Normalize
- Feature Transform
- Target Transform

Feature Engineering

- Feature Interaction
- Polynomial Features
- Group Features
- Bin Numeric Feature
- Combine Rare Levels
- Create Clusters

Feature Selection

- Feature Selection
- Remove Multicollinearity
- Principal Component Analysis
- Ignore Low Variance



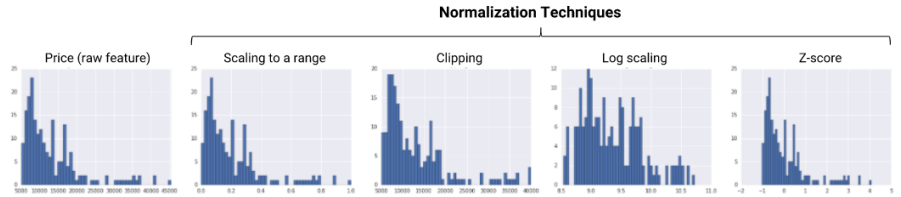
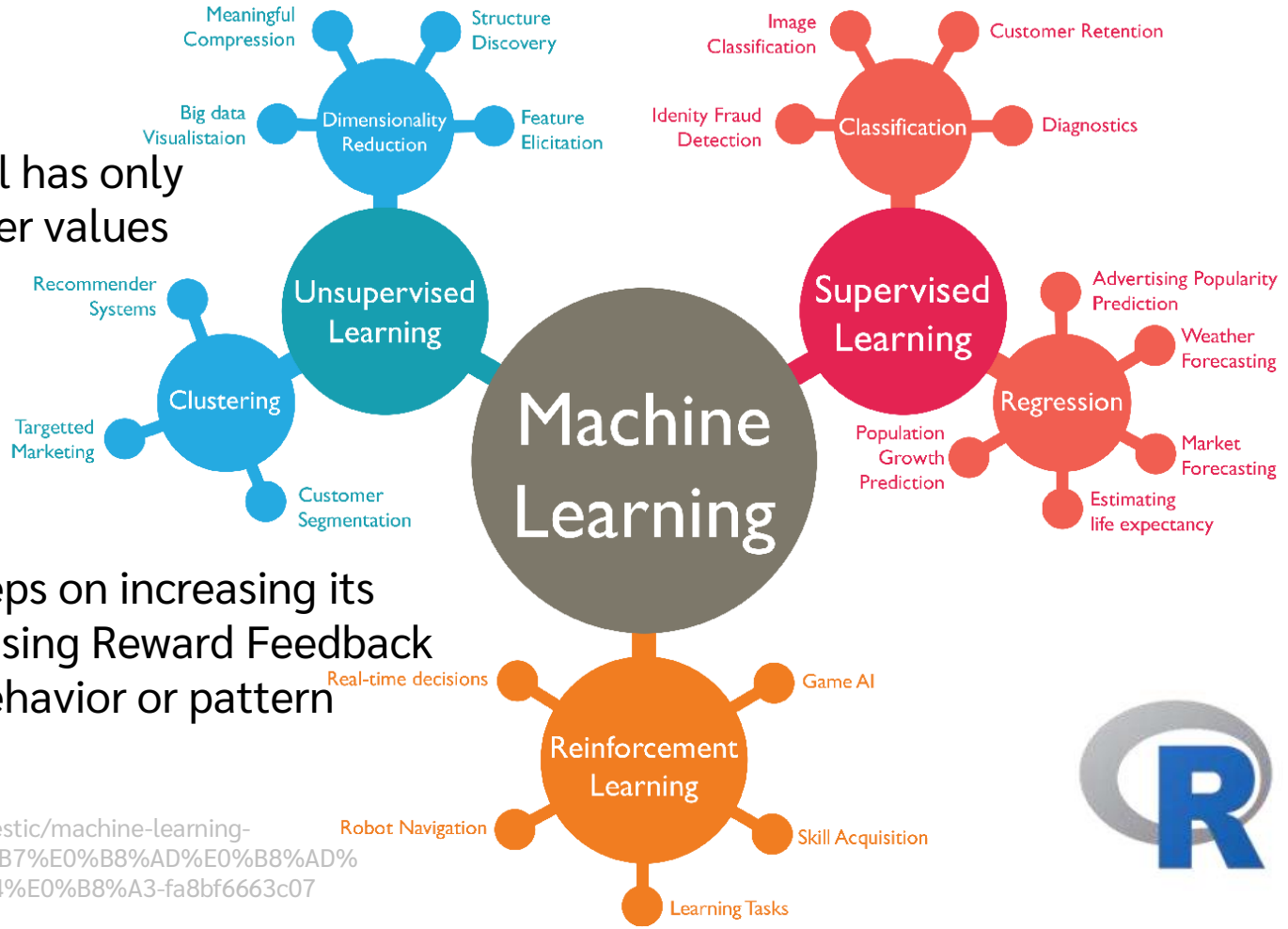


Figure 1. Summary of normalization techniques.

Data Science Process (Lifecycle)

Business Objective → Data Collection → Data Cleansing → Explore Data (EDA) → **Model Building** → Model Evaluation → Model Deployment



Training model has only input parameter values

The model keeps on increasing its performance using Reward Feedback to learn the behavior or pattern

Model is getting trained on a labelled dataset.

User ID	Gender	Age	Salary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	1
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	1
15728773	Male	27	58000	1
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	1
15727311	Female	35	65000	0
15570769	Female	26	80000	1
15606274	Female	26	52000	0
15746139	Male	20	86000	1
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1

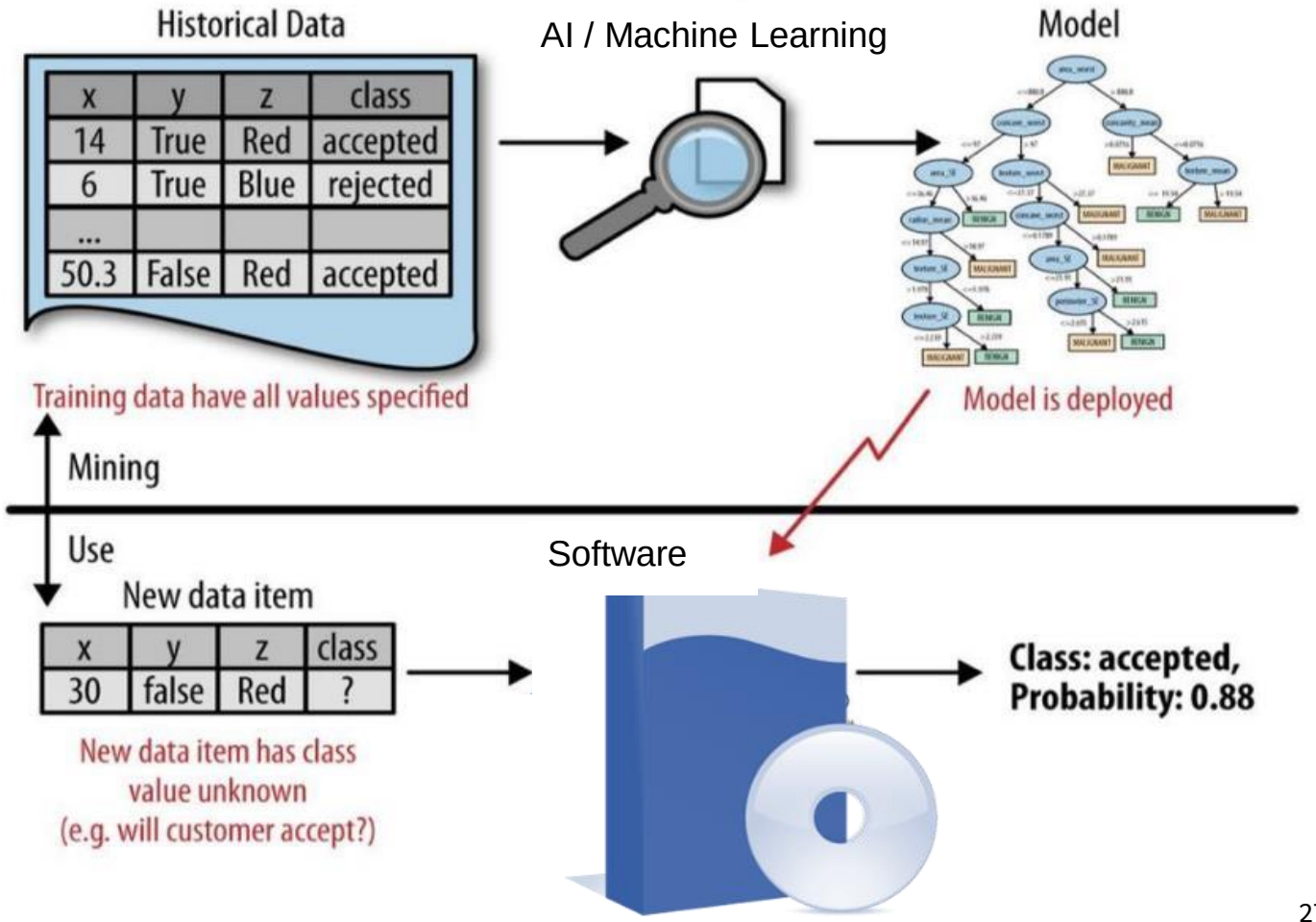
https://medium.com/investic/machine-learning-%E0%B8%84%E0%B8%B7%E0%B8%AD%E0%B8%AD%E0%B8%B0%E0%B9%84%E0%B8%A3-fa8bf6663c07

Data Science Process (Lifecycle)

Business Objective | Data Collection | Data Cleansing | Explore Data (EDA) | **Model Building** | Model Evaluation | Model Deployment

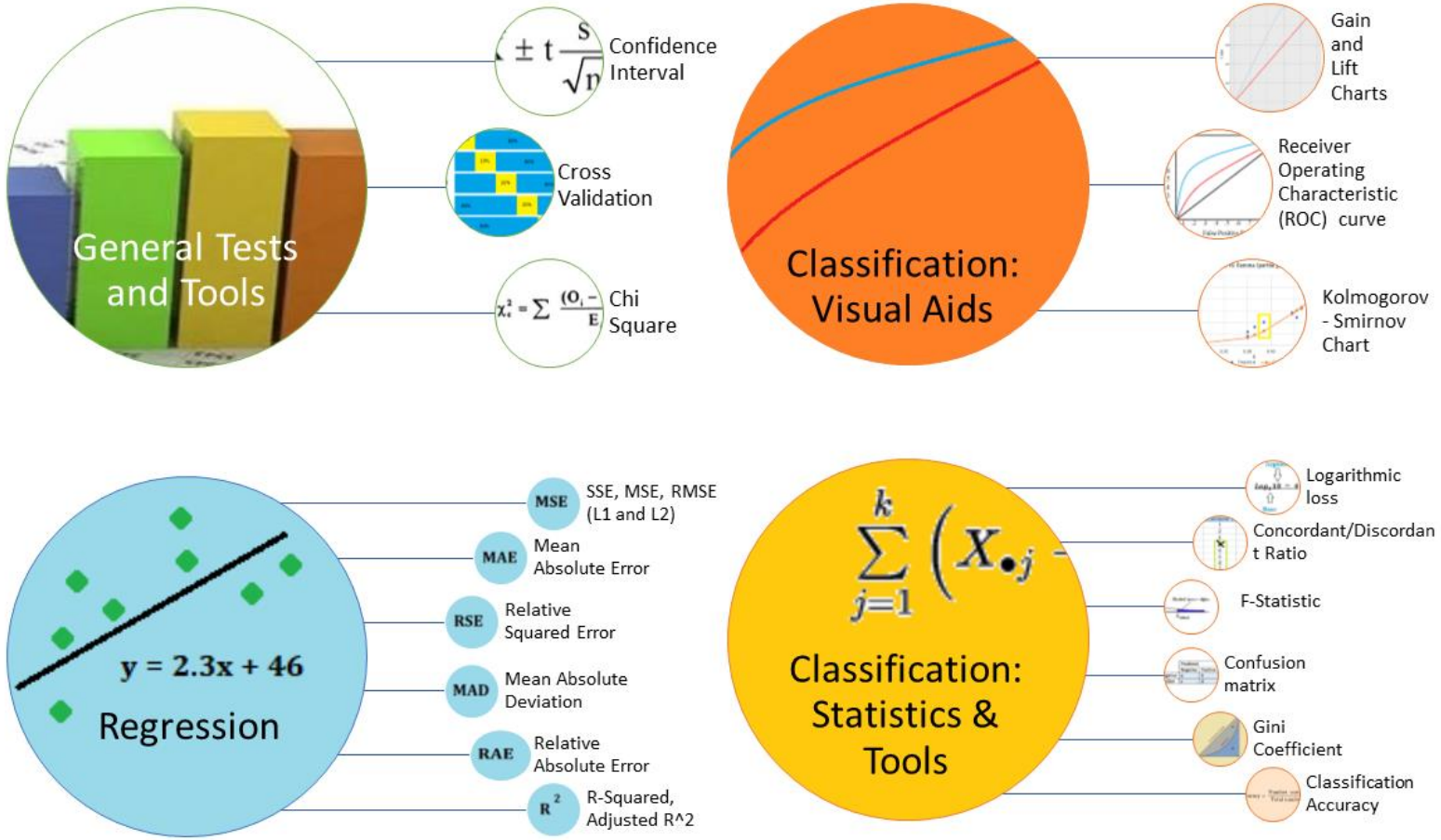
Predictive Analytics

(Supervised Machine Learning: Classification)



Data Science Process (Lifecycle)

Business Objective → Data Collection → Data Cleansing → Explore Data (EDA) → Model Building → **Model Evaluation** → Model Deployment

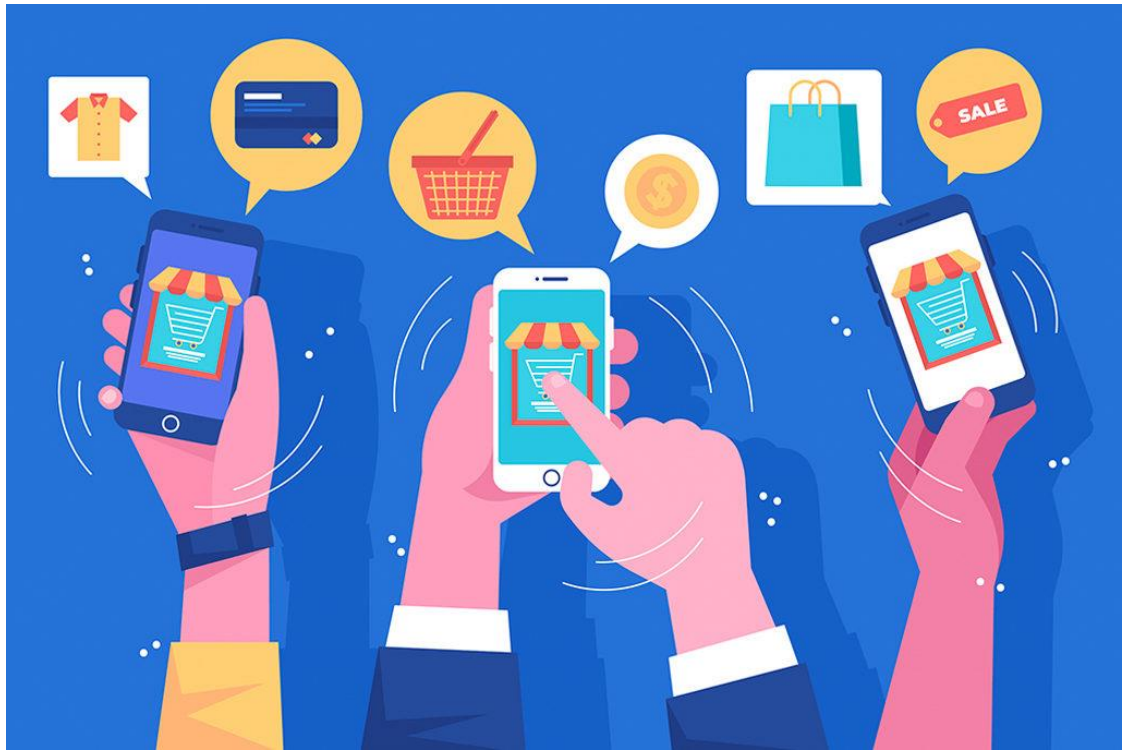


https://www.datasciencecentral.com/profiles/blogs/model-evaluation-techniques-in-one-picture

Data Science Process (Lifecycle)

Business Objective → Data Collection → Data Cleansing → Explore Data (EDA) → Model Building → Model Evaluation → **Model Deployment**

- To integrate a machine learning model into an existing production environment



docker

gradio

kubernetes

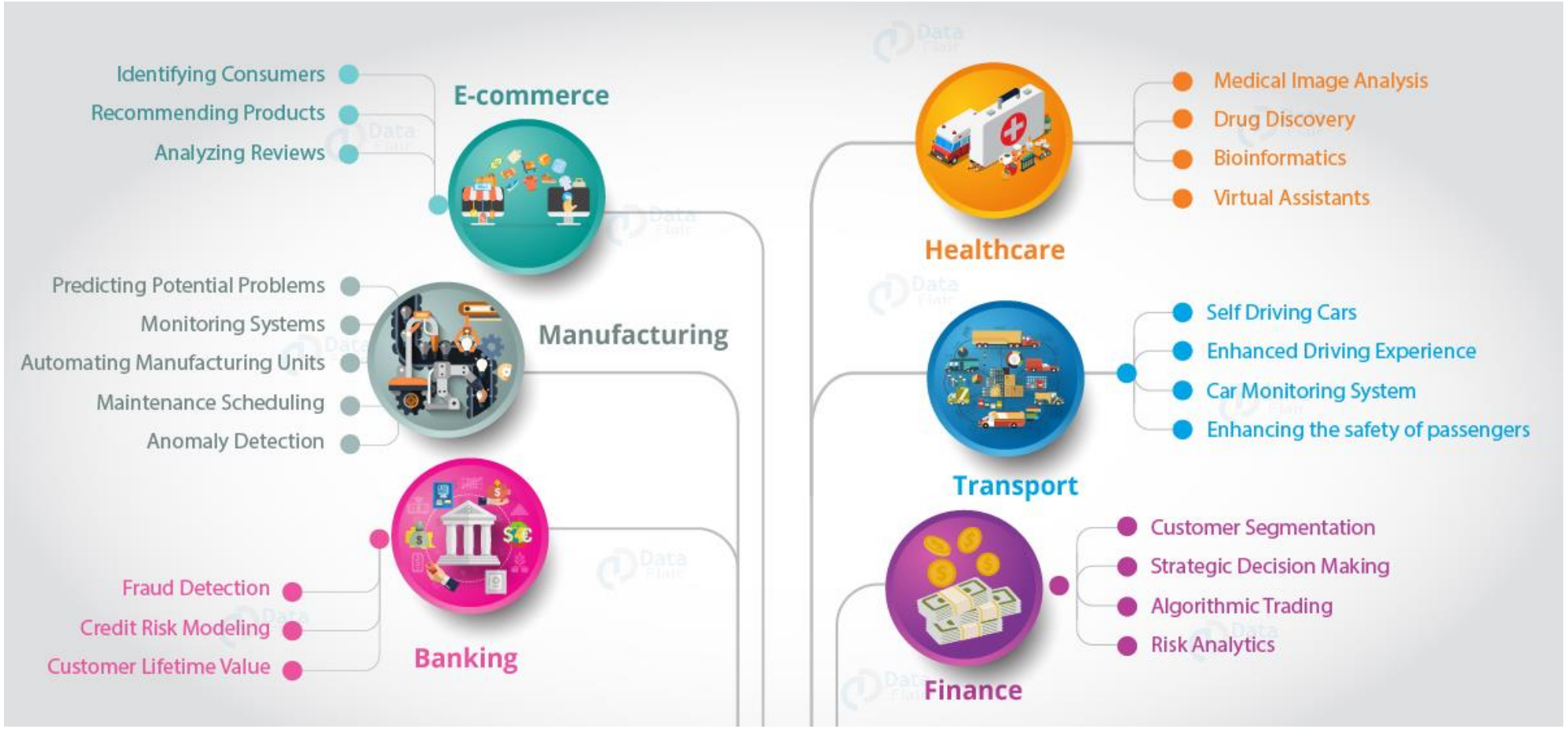
Amazon SageMaker

TensorFlow

mlflow

https://towardsdatascience.com/five-technologies-to-deploy-your-machine-learning-models-bddaad69e0d4

Data Science Applications



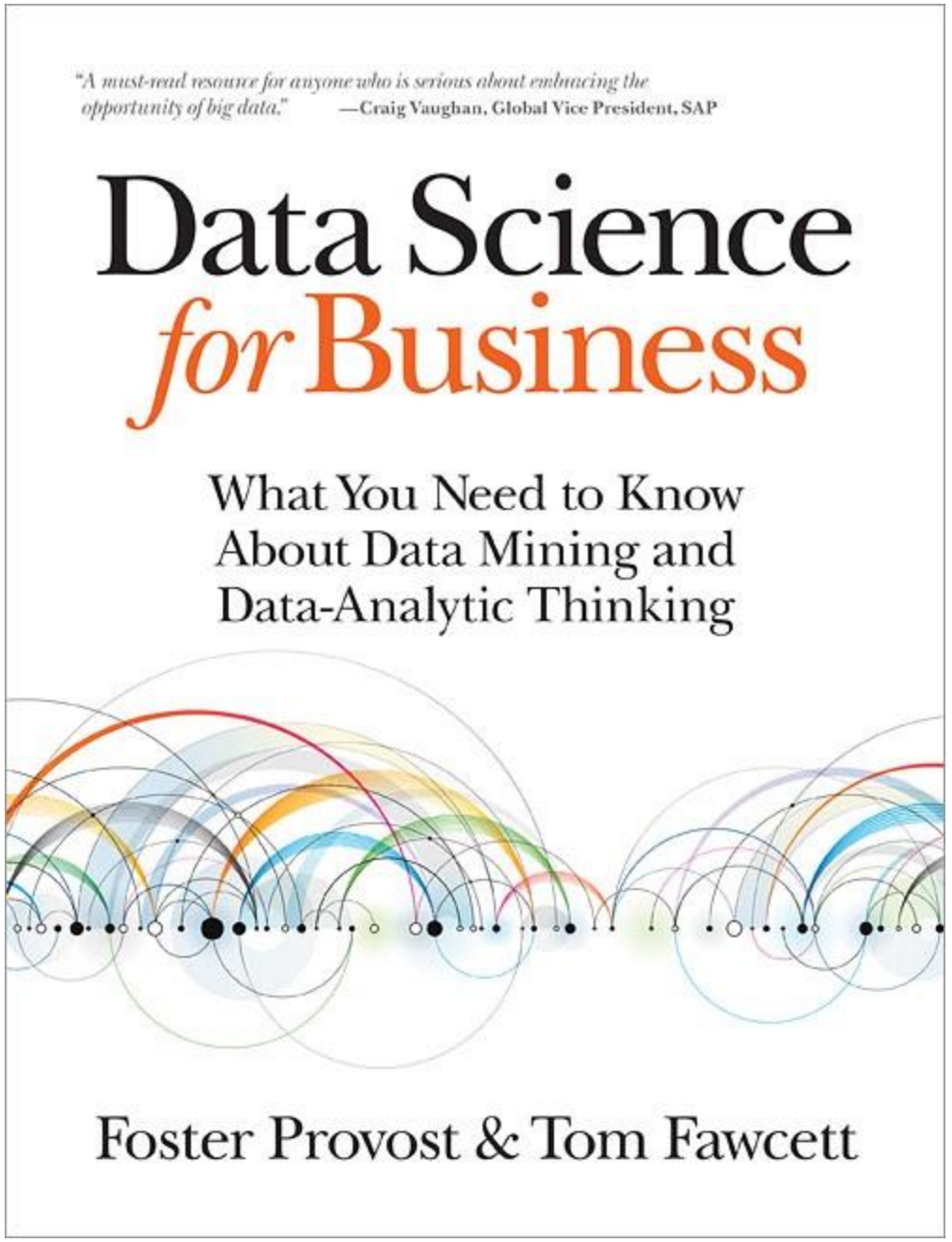
Telecommunication

Entertainment

https://data-flair.training/blogs/data-science-applications/

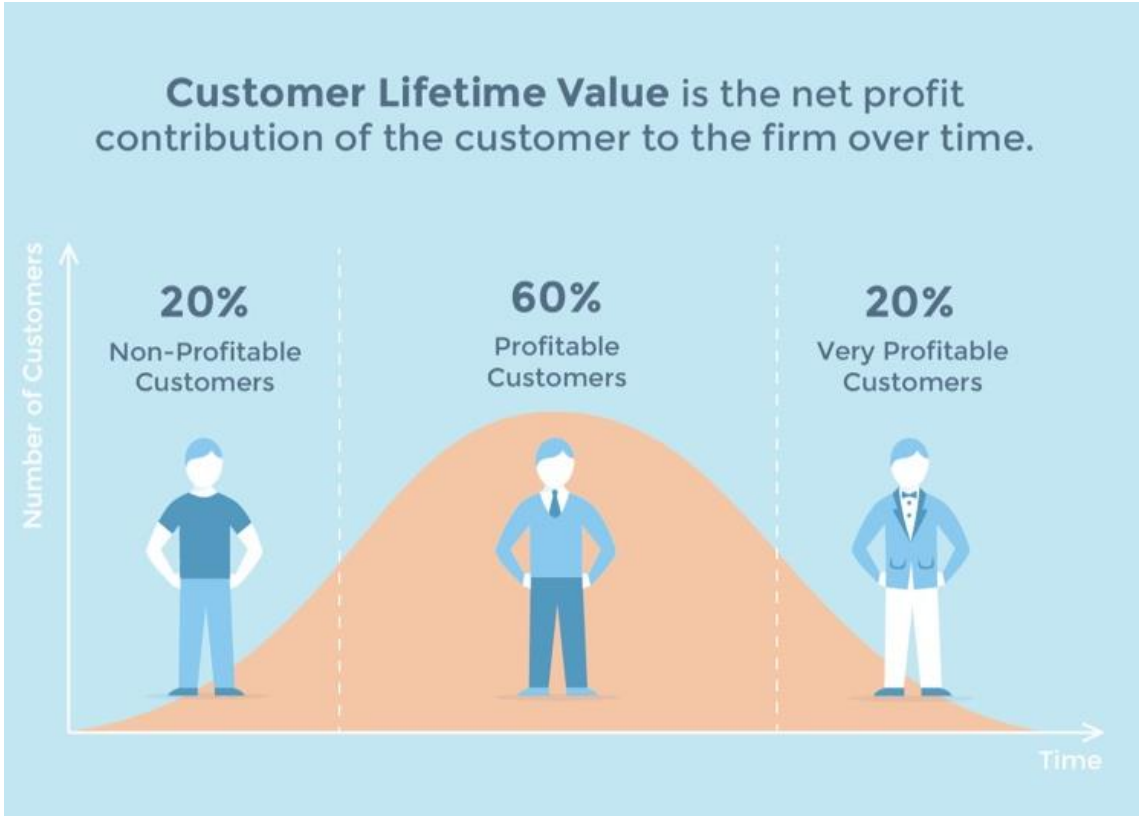
Intermission

10:30 – 10:45


"A must-read resource for anyone who is serious about embracing the opportunity of big data." —Craig Vaughan, Global Vice President, SAP
Data Science
for Business
What You Need to Know
About Data Mining and
Data-Analytic Thinking
Foster Provost & Tom Fawcett

Real world use cases:

Predict Customer Lifetime Values (CLV)

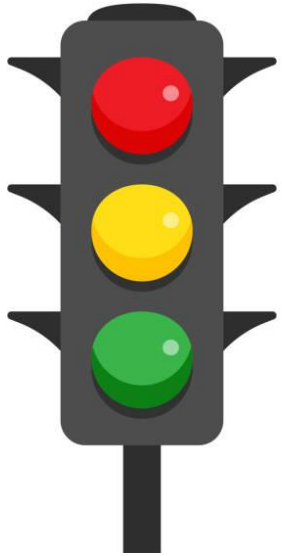


- Preserve specials for high-value customers
- Design marketing campaigns to attract more high-value customers.
- Pick acquisition channels that yield highest value customers.
- Create campaign that appeal to the highest value customers.
- Convert medium-value customers to high-value.

Real world use cases:

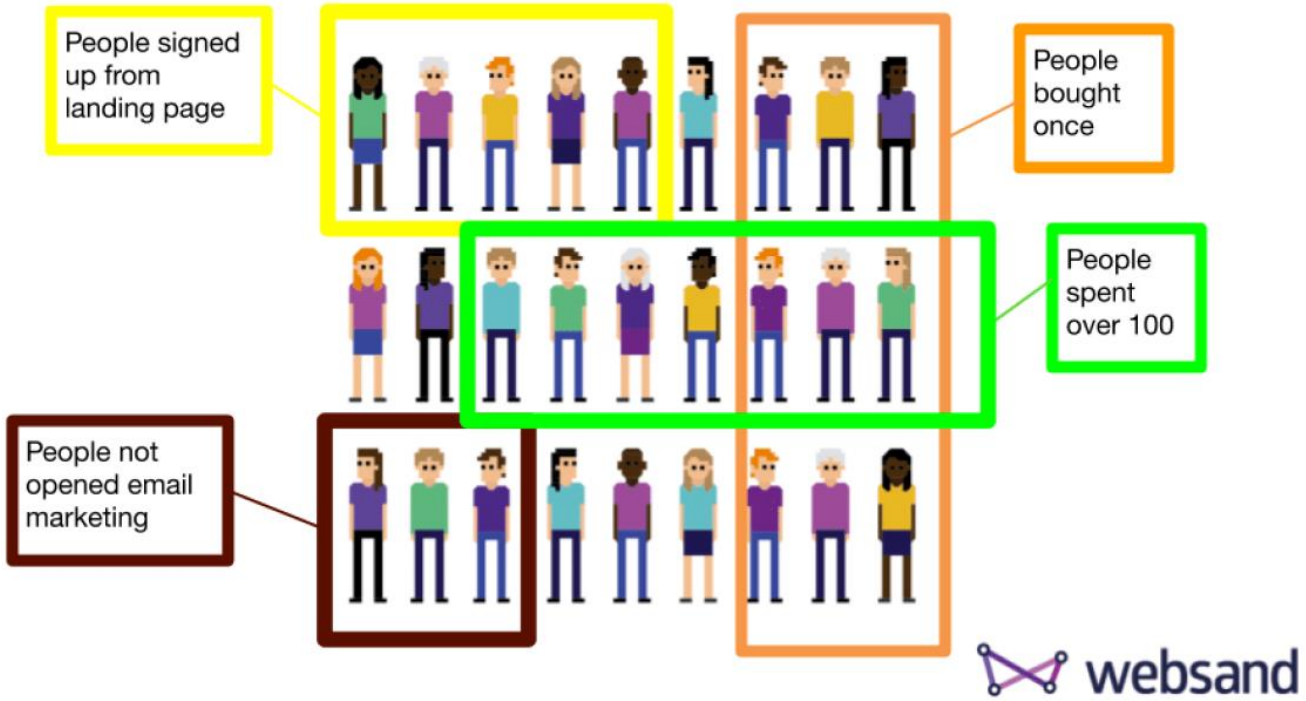
Credit Scoring

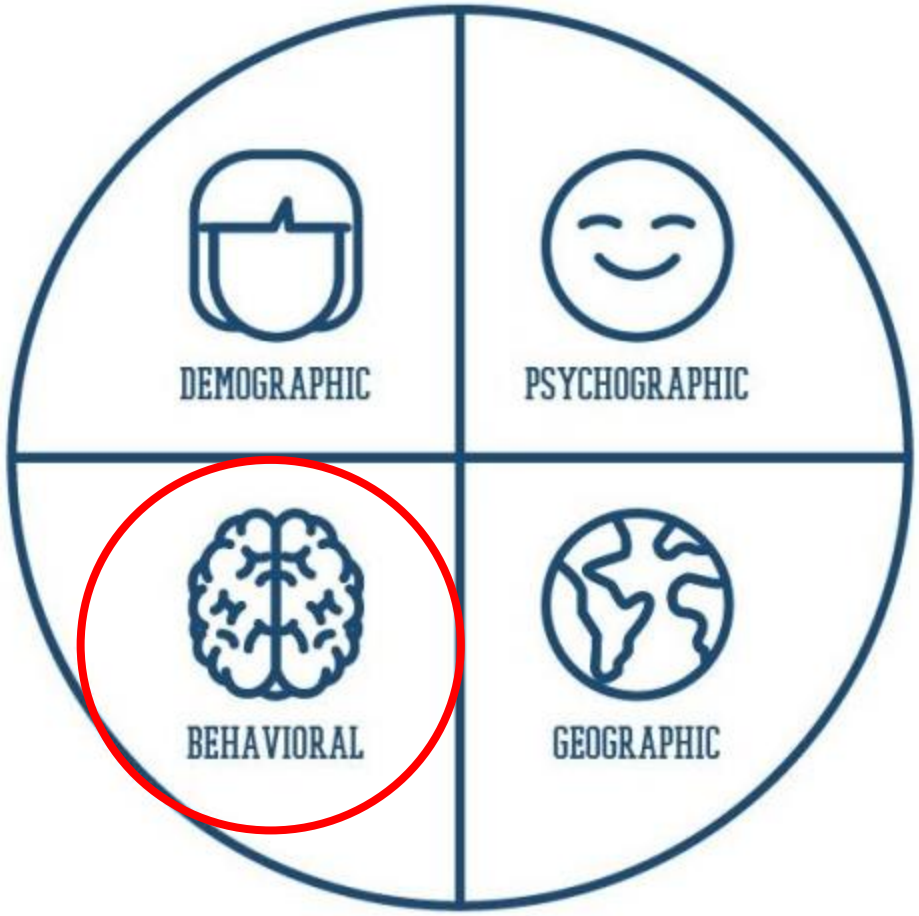
- **First Union Bank** deployed a value predicting system that assign green/yellow/red flag to each customer, based on their predicted lifetime value
- Service representatives were instructed to waive fee for green customers and not waive for red customers. For yellow customers, they can make their own judgement.
- This strategy generated over 100 million in incremental revenue.



Real world use cases:

Customer Segmentation

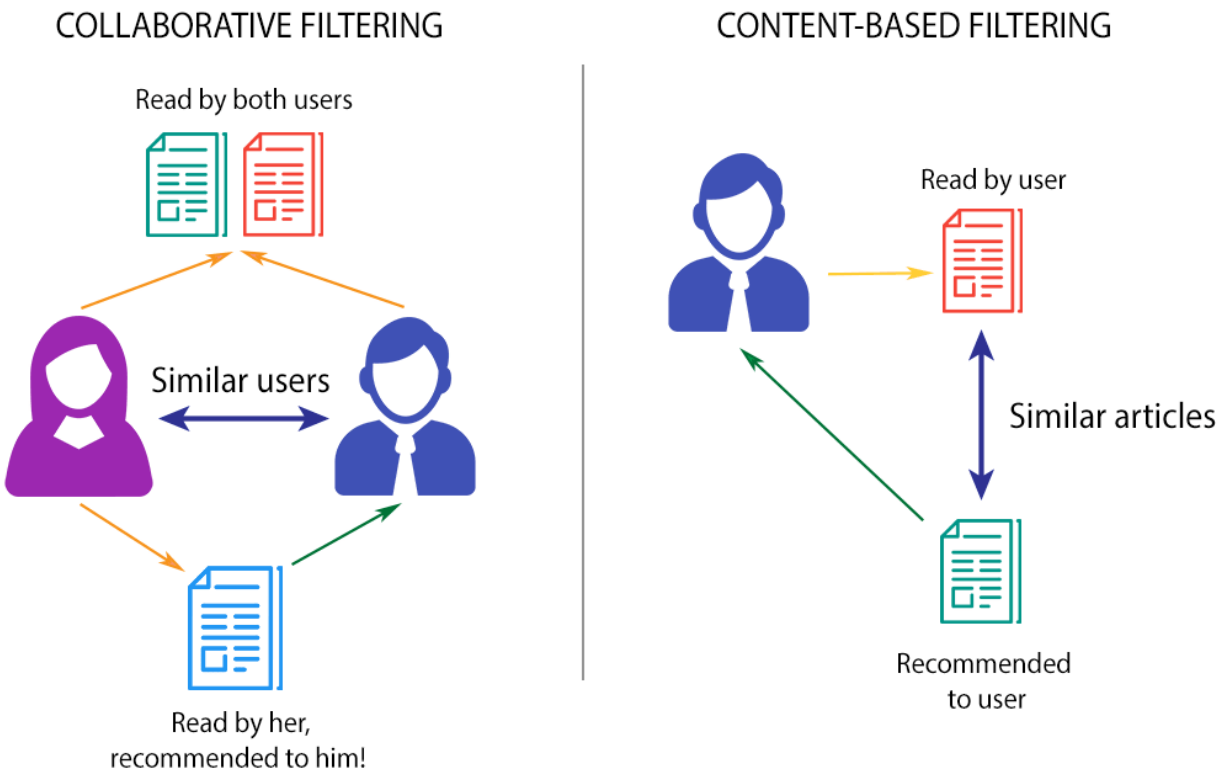




- The better you understand, who are the customers at your store, the easier it is to tailor to their needs.
- WHAT you sell and HOW you sell it.

Real world use cases:

Product Recommendation



- Benefits users in finding items of their interest.
- Help item providers in delivering their items to the right user.
- Identity products that are most relevant to users.
- Personalized content.
- Help websites to improve user engagement.

<u>Algorithms</u>

- *KNN: K-Nearest Neighbors*
- *Matrix Factorization*
- *Association rules*

Real world use cases:

Market Basket Analysis (Association Rules)

https://th.linkedin.com/pulse/ขนตอนการหากฎความสมพนธ-association-rules-eakasit-pacharawongsakda

Rule No.	Frequent itemset	Confidence	Lift
1	Apple ⇒ Cereal	100%	1.33
2	Beer ⇒ Eggs	100%	1.33
3	Eggs ⇒ Beer	100%	1.33
4	Beer, Cereal ⇒ Eggs	100%	1.33
5	Cereal, Eggs ⇒ Beer	100%	1.33
6	Cereal ⇒ Apple	67%	1.33
7	Beer ⇒ Cereal, Eggs	67%	1.33
8	Eggs ⇒ Beer, Cereal	67%	1.33
9	Beer ⇒ Cereal	67%	0.89
10	Cereal ⇒ Beer	67%	0.89
11	Cereal ⇒ Eggs	67%	0.89
12	Eggs ⇒ Cereal	67%	0.89
13	Cereal ⇒ Beer, Eggs	67%	0.89
14	Beer, Eggs ⇒ Cereal	67%	0.89

รูปที่ 8 แสดงกฎความสัมพันธ์ทั้งหมดที่สร้างได้พร้อมทั้งค่า confidence และ lift

- What product are often purchased together?
- Does product brands matter?
- How are the demographics of neighborhoods effecting what customers are buying?
- Where should each product be shelved to maximize the sales?

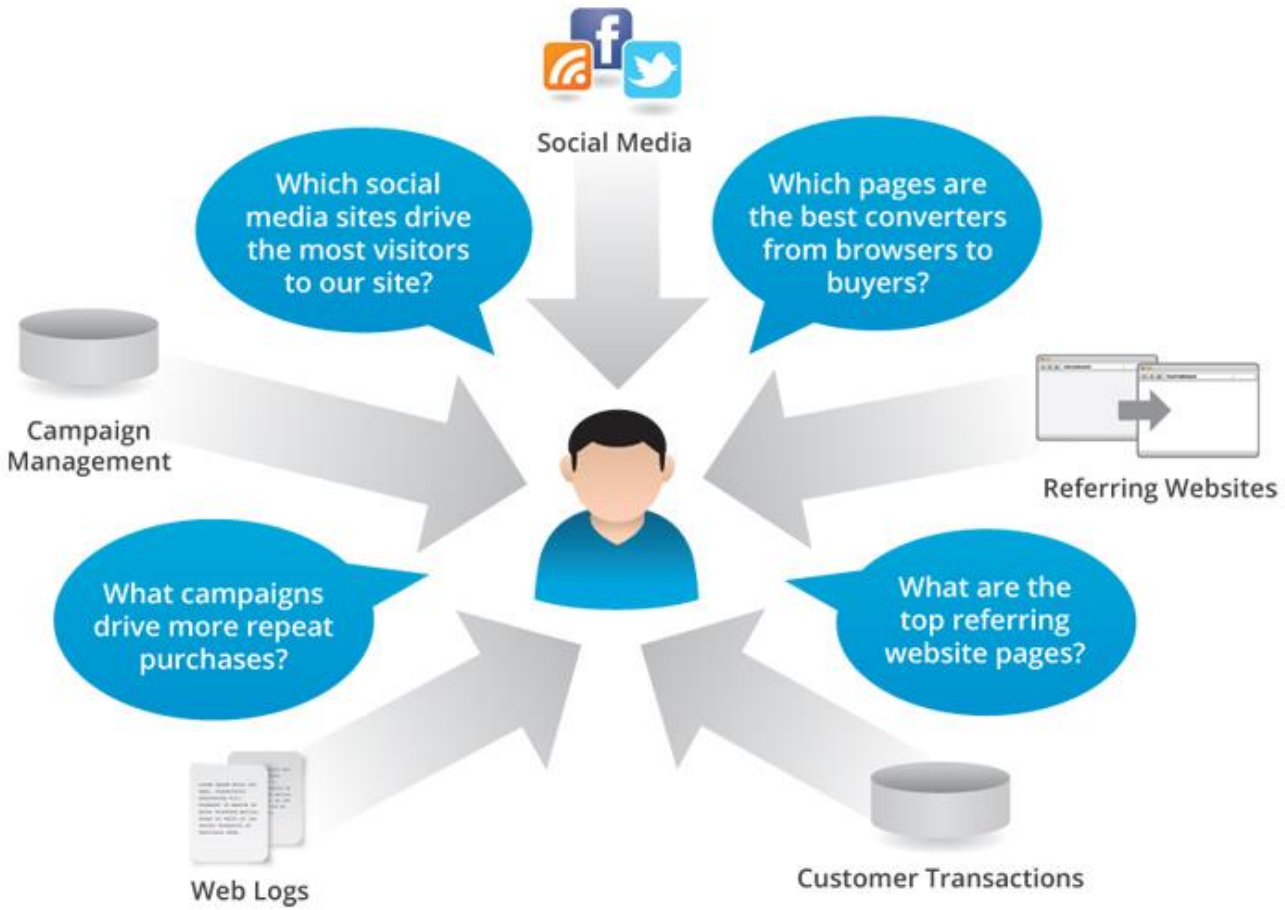
Real world use cases:

Inventory Prediction

- Stock the right products
- Purchasing pattern
- Predict purchasing trends and manage efficient stock

Real world use cases:

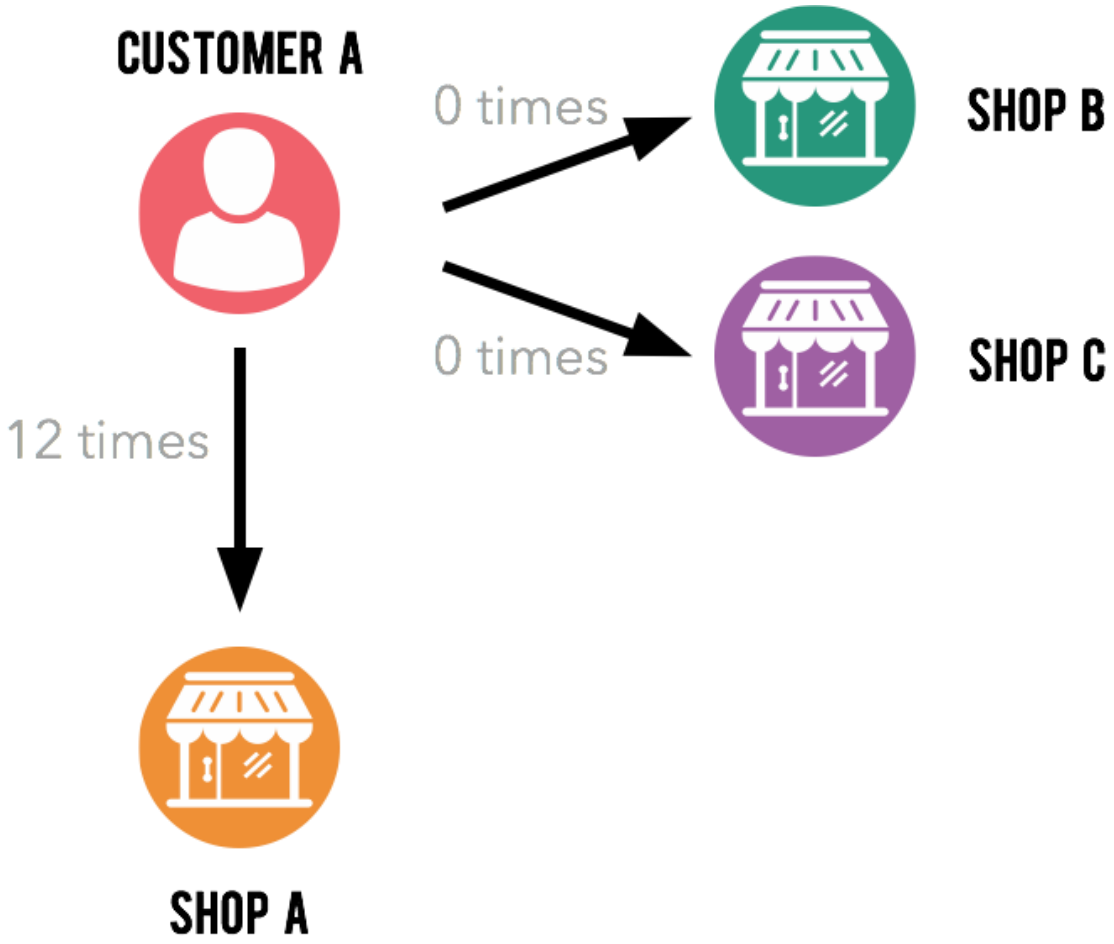
Campaign Analytics



- Tracking customer interactions (Google Analytics)
- Linking marketing activities and conversions across systems
- Linking social media marketing insights

Real world use cases:

Recommender Engine For Credit Card Users



- Predict shop preference from previous behaviors.
- Swipe time prediction.
- Win back inactive customers by predicting deal preference from demographics.

Real world use cases:

Debtor Segmentation

- Why can’t they pay?
- How late will they pay in the upcoming cycle?
- What actions have been working and how to maximize collection volume?

Real world use cases:

Sale Forecasting

- Predict short-term and long-term performance
- Sales planning
- Demand forecasting
- Inventory controls
- Supply chain management
- Continuous improvement
- Companies can base their forecasts on
  - Past sales data: sale trends over the last 5 years and factor analysis to see what impacts sales
  - Industry-wide comparisons
  - Economic trends.

GBDi

Real world use cases:

Anomaly Detection

TELECOM

Detect no aiming abuse, revenue fraud, service disruptions

BANKING

Flag abnomally high purchases/deposits, detect cyber intrusions

FINANCE & INSURANCE

Detect and prevent out of pattern or fraudulent spend, travel expenses

HEALTHCARE

Detect fraud in claims and payments; events from RFID and mobiles

MANUFACTURING

Detect abnormal machine behavior to prevent cost overruns

TRANSPORTATION

Ensure external communications to the vehicle are not intrusion

SOCIAL MEDIA

Detect compromised accounts, bots that generate fake reviews

NETWORKING

Detect intrusion into networks, prevent theft of source code or IP

SMART HOUSE

Detect energy leackage, standartazi smart sensor datasets

VIDEO SURVEILLANCE

Detect or track objects and persons of interest in monotonous footage

SQUADEX

https://squadex.com/insights/top-machine-learning-use-cases-business/

Real world use cases:

Sentimental Analysis

- Use text analytics to gather citizen concerns and sentiments
- Assess current situations to create a set of keywords to gather social media post relate to topics of interest
- Determine the sentiment with respect to topics or the overall contextual polarity of t post/comment

Real world use cases:

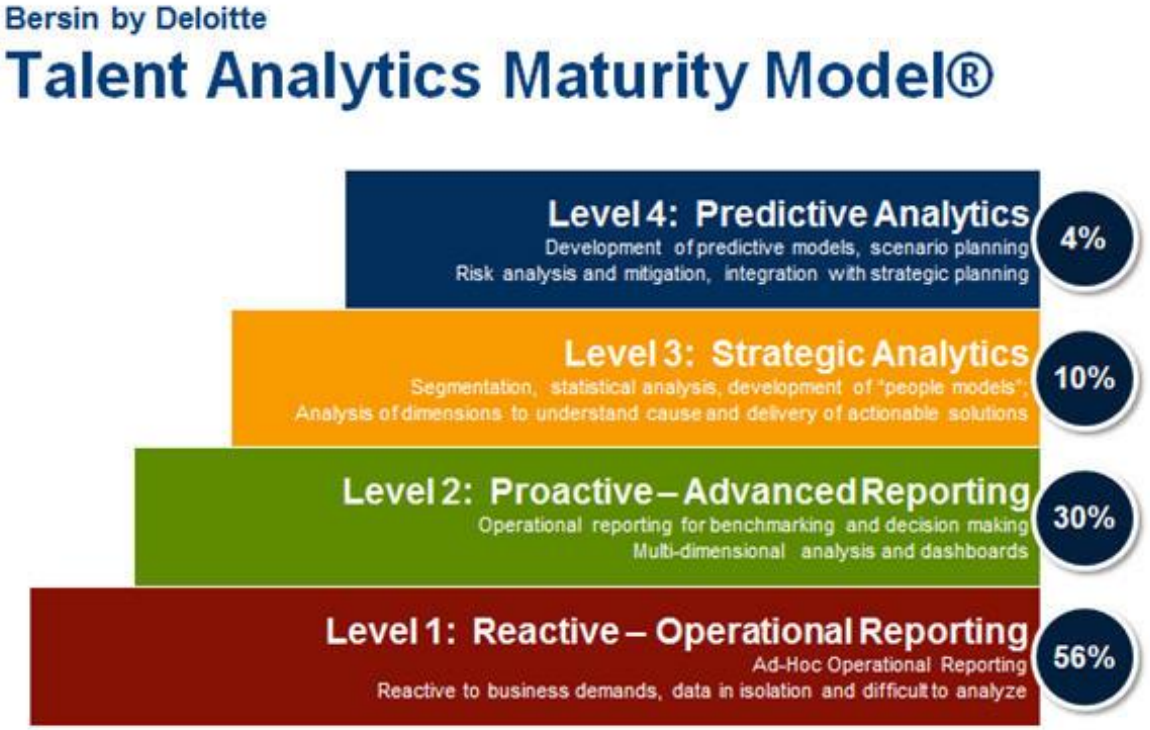
Influencer Analysis

- An influencer is an individual who has above-average impact on a specific niche process.
- On the social network, an influencer can refer to the most shaping a discussion about topic

Real world use cases:

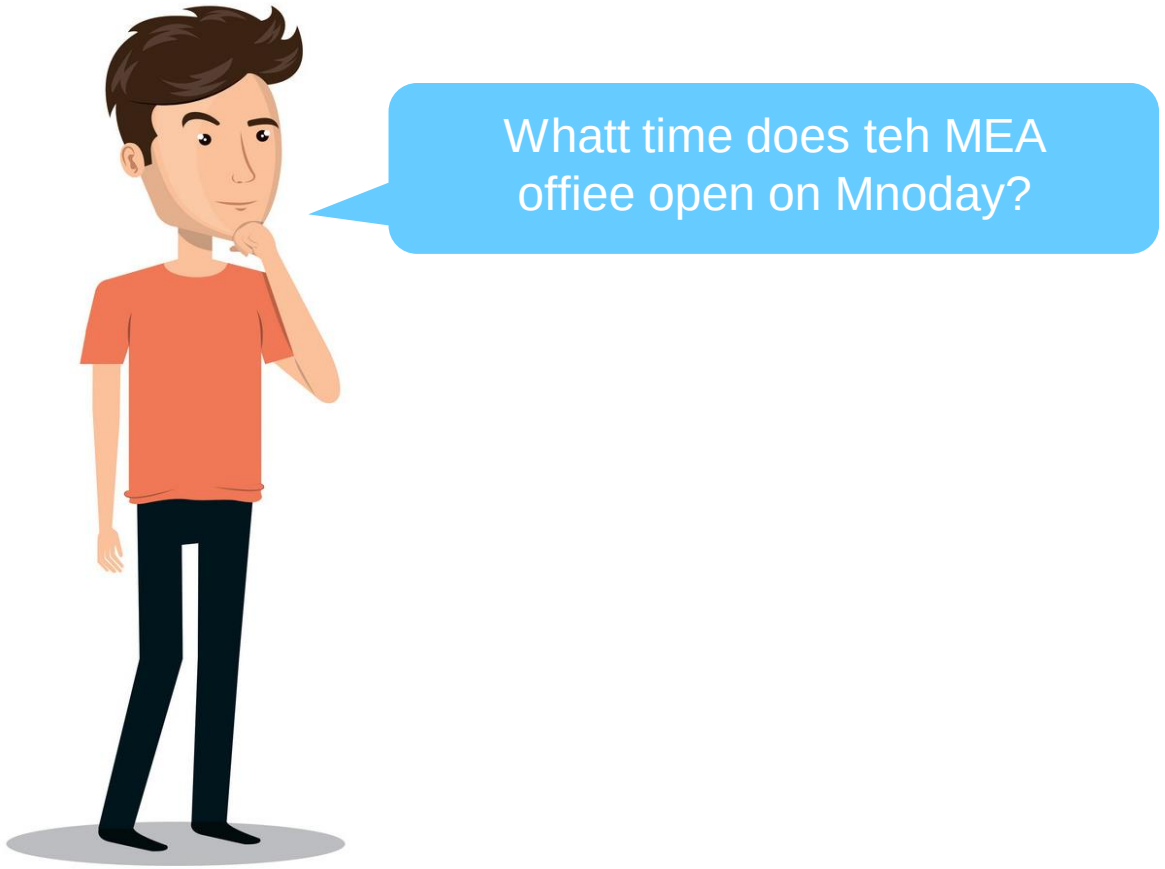
People Analytics

- Talent acquisition, retention, placement, promotion, compensation, succession planning.
- Analyzing the skills and attributes of high performers,
- Then build a template with quality hiring factors.
- Sources - digital ‘thought prints’ of prospects on social media
- Statistical analysis of productivity and turnover
    - Old indicators (such as GPA and education) are less critical than factors like experience.



Credit:Forbe

Chatbot



Chatbot

What time does the MEA office open on Monday?

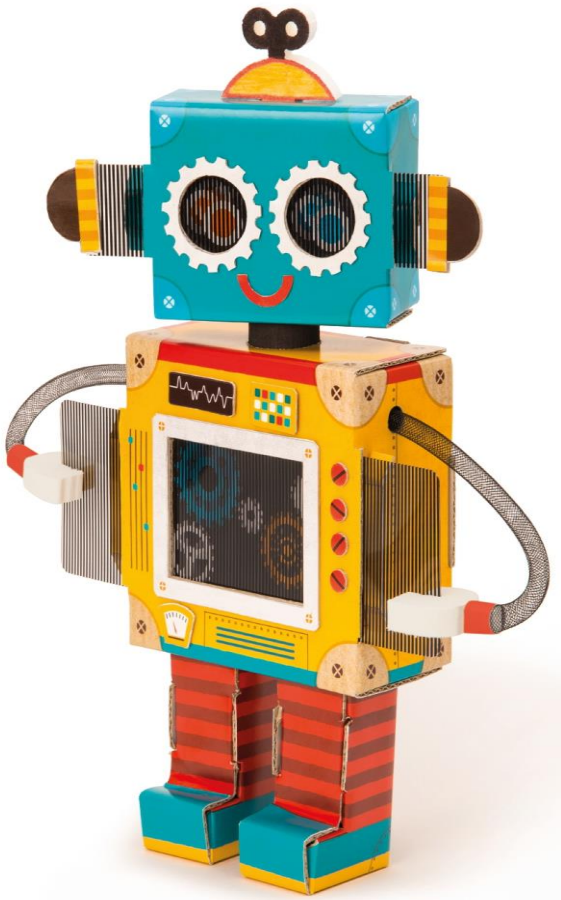
Spell Correction

Word Segmentation

POS Tagging

Entity Recognition

Intent Classification



Chatbot

What | time | does | the | MEA | office | open | on | Monday?

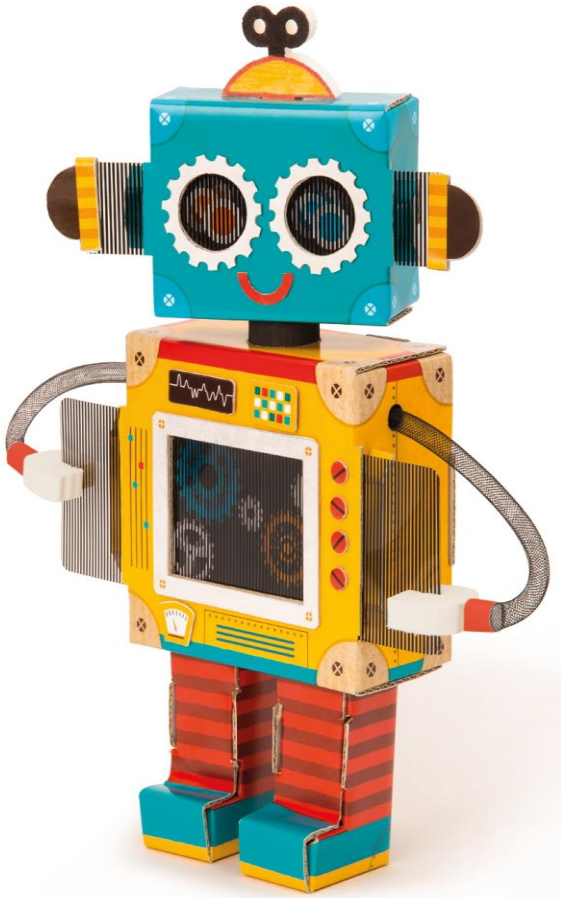
Spell Correction

Word Segmentation

POS Tagging

Entity Recognition

Intent Classification



Chatbot

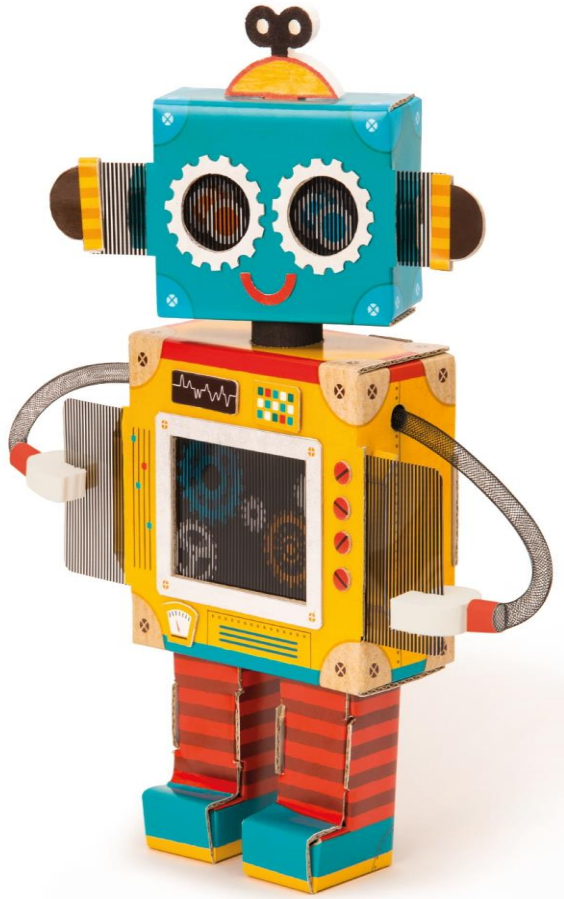
Spell Correction

Word Segmentation

POS Tagging

Entity Recognition

Intent Classification



Chatbot

What | time | does | the | MEA | office | open | on | **Monday?**

Date

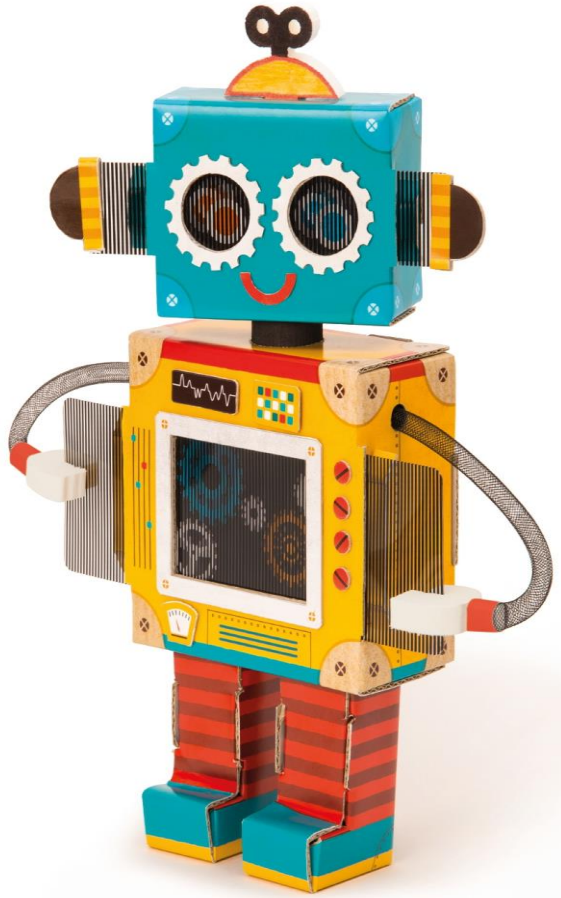
Spell Correction

Word Segmentation

POS Tagging

Entity Recognition

Intent Classification



Chatbot

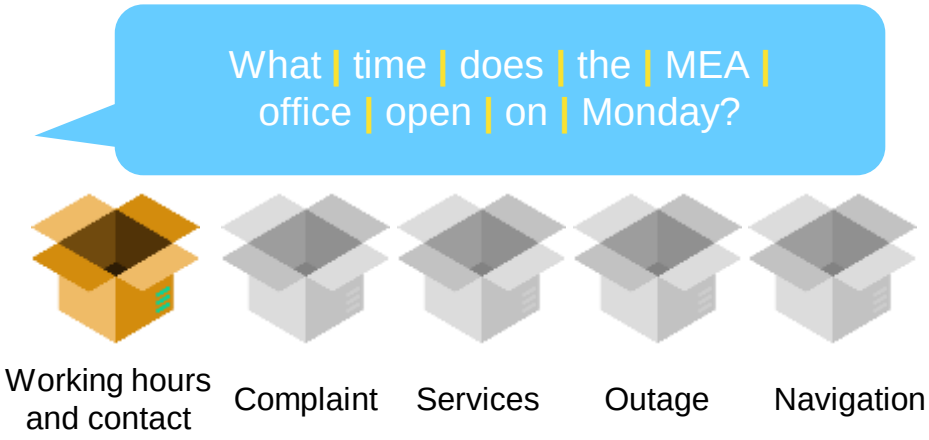
Spell Correction

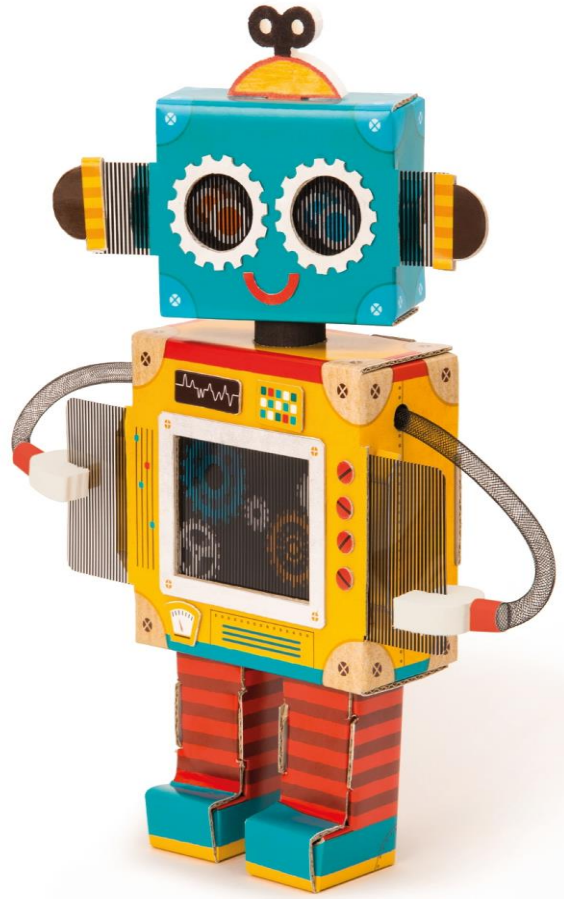
Word Segmentation

POS Tagging

Entity Recognition

Intent Classification

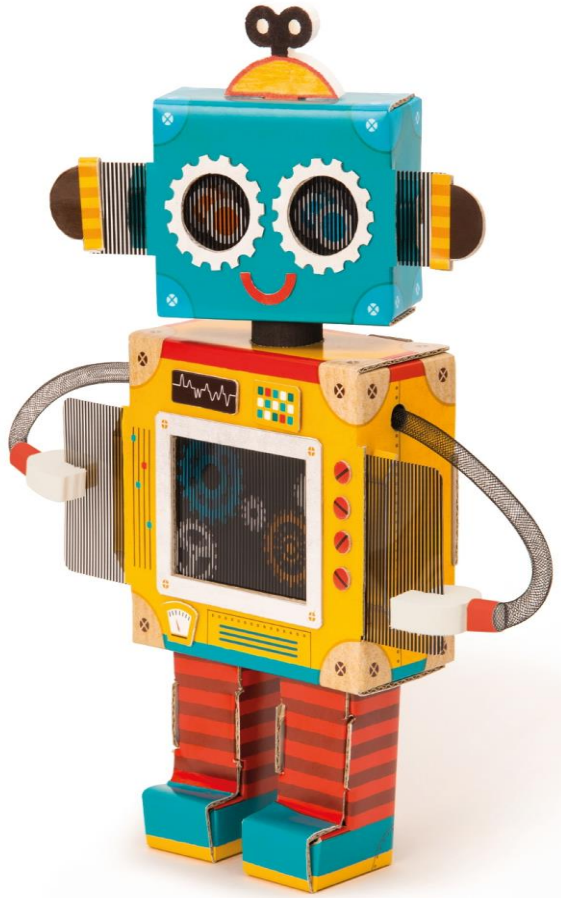




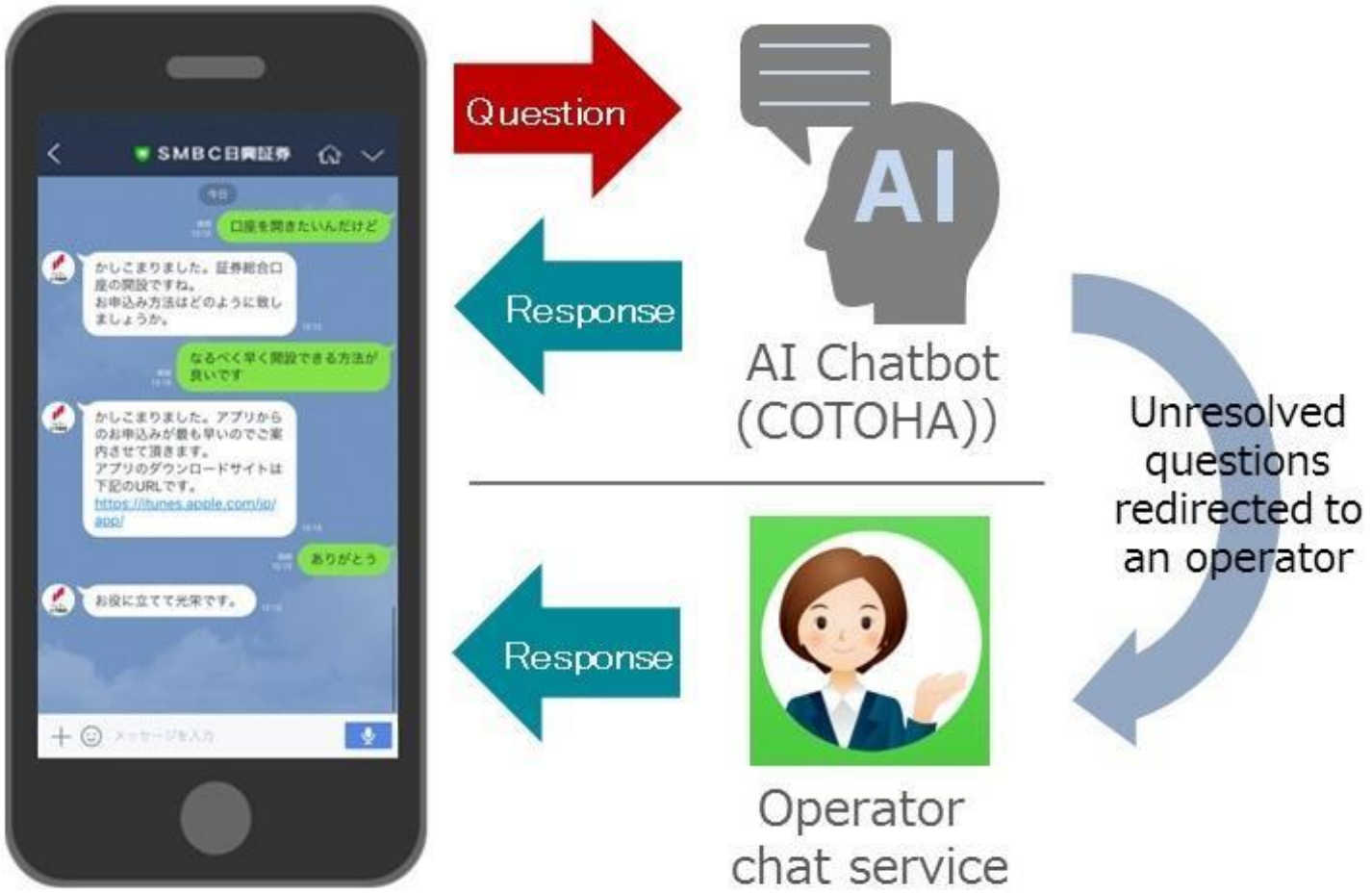
Chatbot

Whatt time does teh MEA offiee open on Mnoday?

MEA office is open on Monday from 08.00 to 16.00



Chatbot



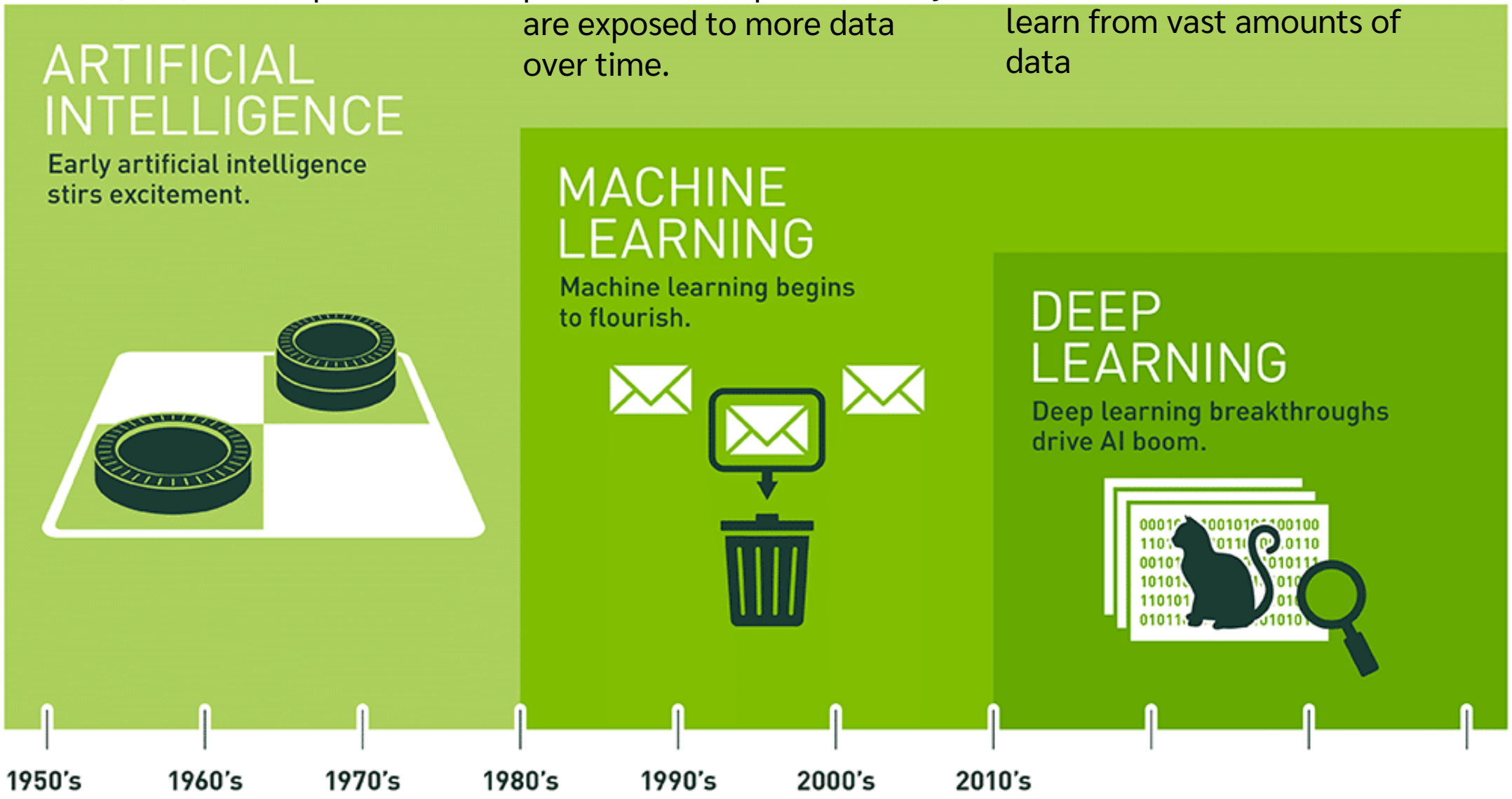
https://disruptive.asia/smbc-ntt-com-ai-chatbot/

AI - Artificial Intelligence

Program that can sense, reason, act, and adapt.

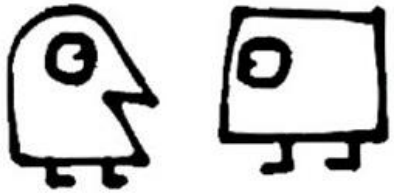
Algorithms whose performance improve as they are exposed to more data over time.

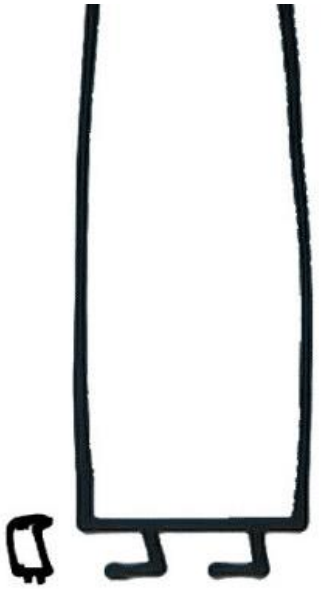
Subset of ML in which multilayered neural networks learn from vast amounts of data



https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

3 stages of AI





Narrow AI

Dedicated to assist with or take over specific tasks

General AI

Takes knowledge from one domain, transfers to other domain

Super AI

Machines that are an order of magnitude smarter than humans

NOW

FUTURE →

https://datakeen.co/en/what-is-artificial-intelligence/

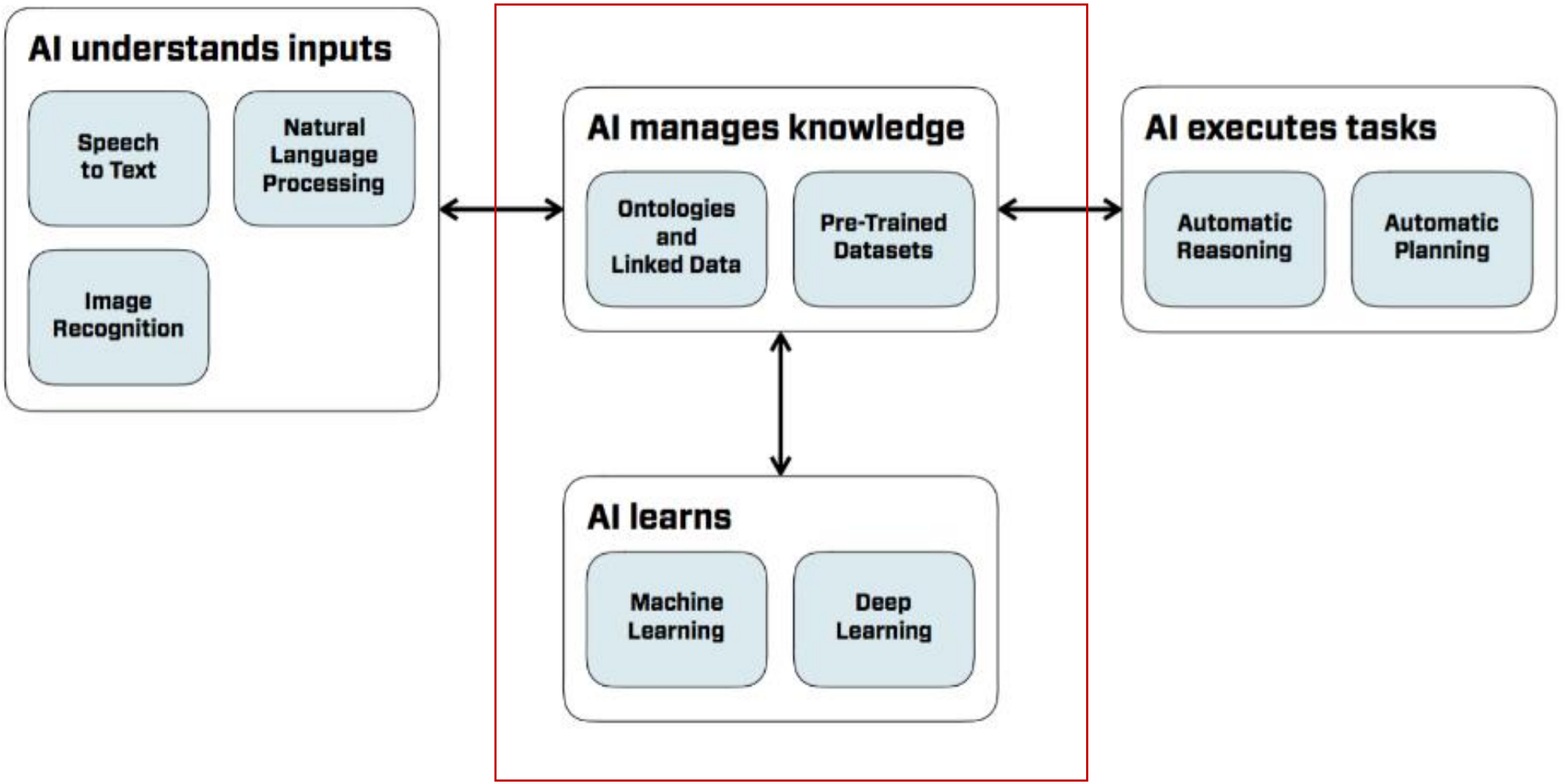
Credit: Chris Noessel

Narrow AI

Dedicated to assist with or take over specific tasks

Narrow AI

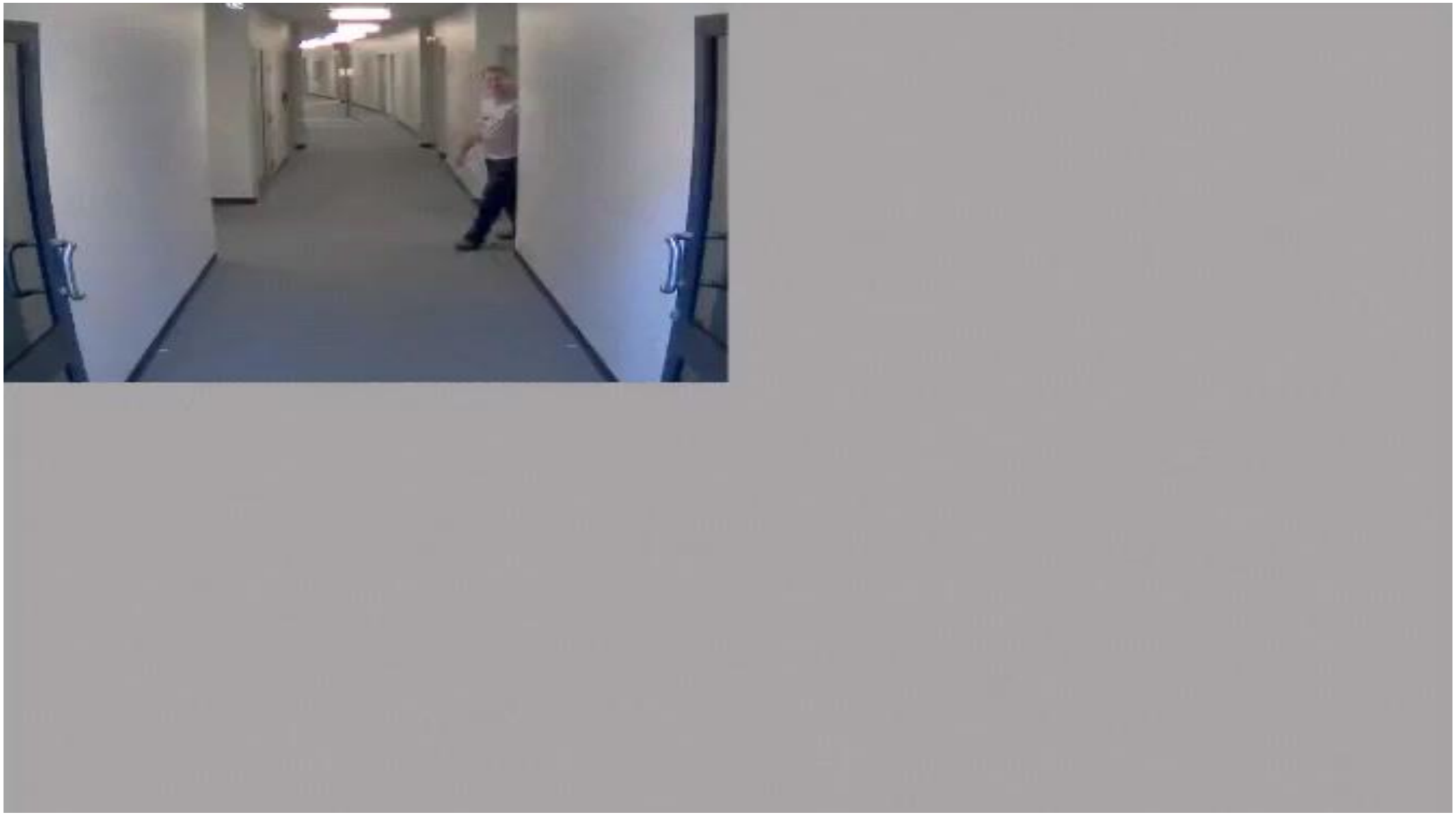
Dedicated to assist with or take over specific tasks



Credit: nicolamattina.it

Narrow AI

Dedicated to assist with or take over specific tasks



Narrow AI

Dedicated to assist with or take over specific tasks

Video Analytics for Profiling

Narrow AI

Dedicated to assist with or take over specific tasks

Recognize the emotions of students

Narrow AI

Dedicated to assist with or take over specific tasks

Facebook apologizes after its AI software labels Black men 'primates' in a video featured on the platform

https://www.bangkokpost.com/world/2176343/facebook-mistakenly-labels-black-men-primates

Facebook users in recent days who watched a British tabloid video featuring Black men were shown an auto-generated prompt asking if they would like to “keep seeing videos about Primates,” according to the New York Times.

Facebook uses artificial intelligence and facial recognition software to automatically analyse and categorise videos uploaded to their platform.

Civil rights activists have long complained of the implicit bias and inaccuracy of facial recognition software in recognising people who are not caucasian.

General AI

Takes knowledge from one domain, transfers to other domain

- Common Sense
- Background Knowledge
- Transfer Learning
- Abstraction
- Causality
- Consciousness

Super AI

Machines that are an order of magnitude smarter than humans

- AI as performant as all humans combined

"Hey Siri" "Hey Cortana" "Alexa" "OK Google" "Hi Bixby"

Narrow AI

Dedicated to assist with or take over specific tasks

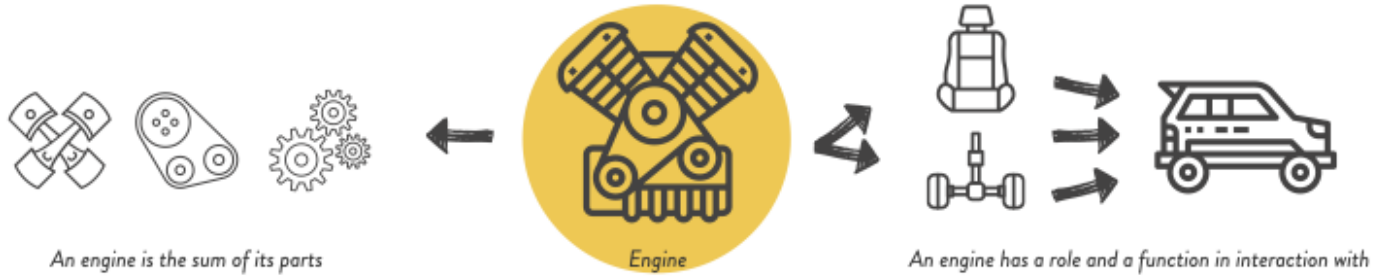
General AI

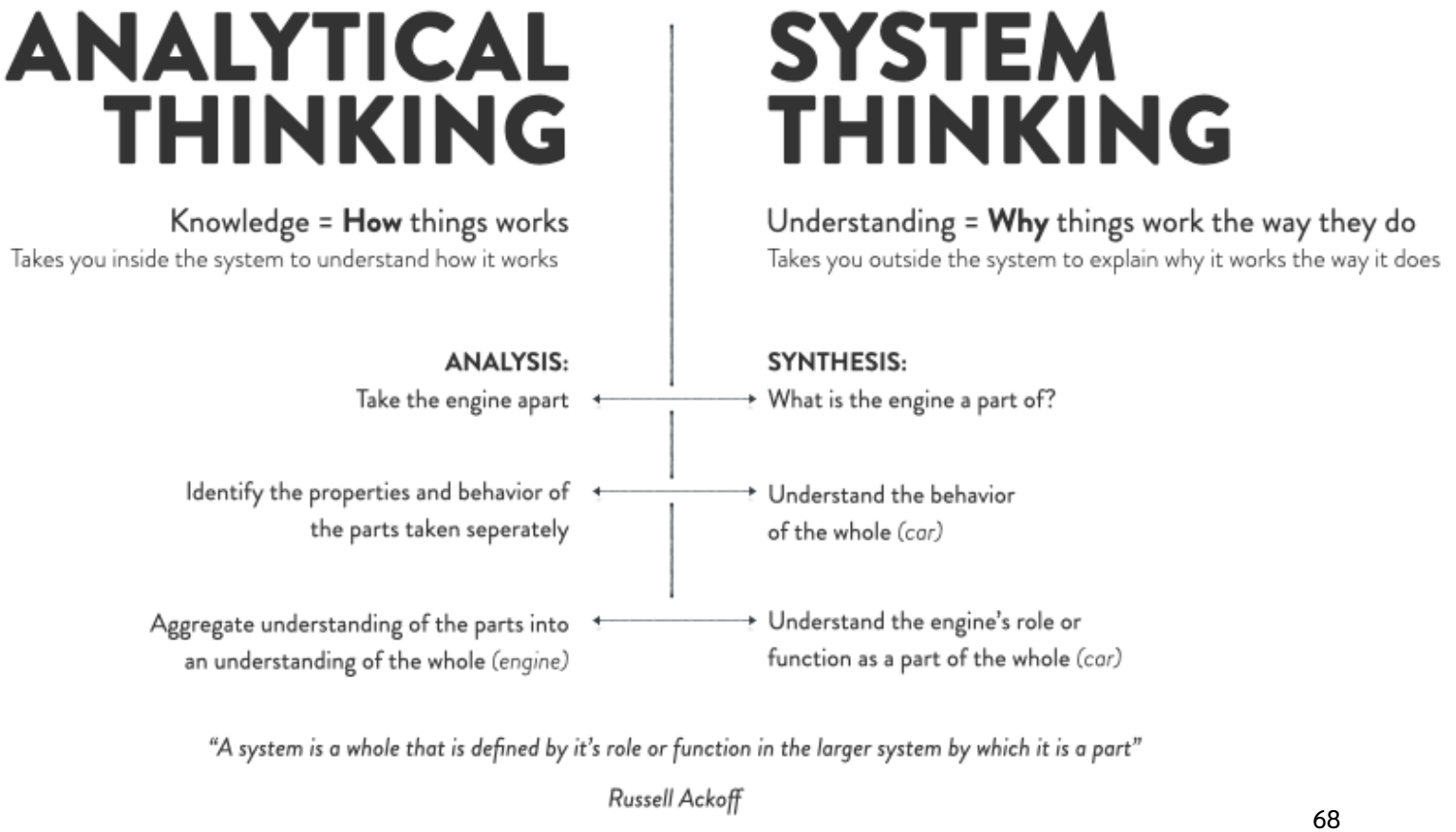
Takes knowledge from one domain, transfers to other domain

Analytical Thinking

Analytical Thinking

- Analytical skills refer to the ability to collect and analyze information, problem-solve, and make decisions.
- There are many types of analytical skills, including **communication, creativity, critical thinking, data analysis, and research**.
- You use analytical skills when detecting patterns, brainstorming, observing, interpreting data, and making decisions based on the multiple factors and options available to you.
- Most types of work require analytical skills. You use them to solve problems that may not have obvious solutions or have several variables.





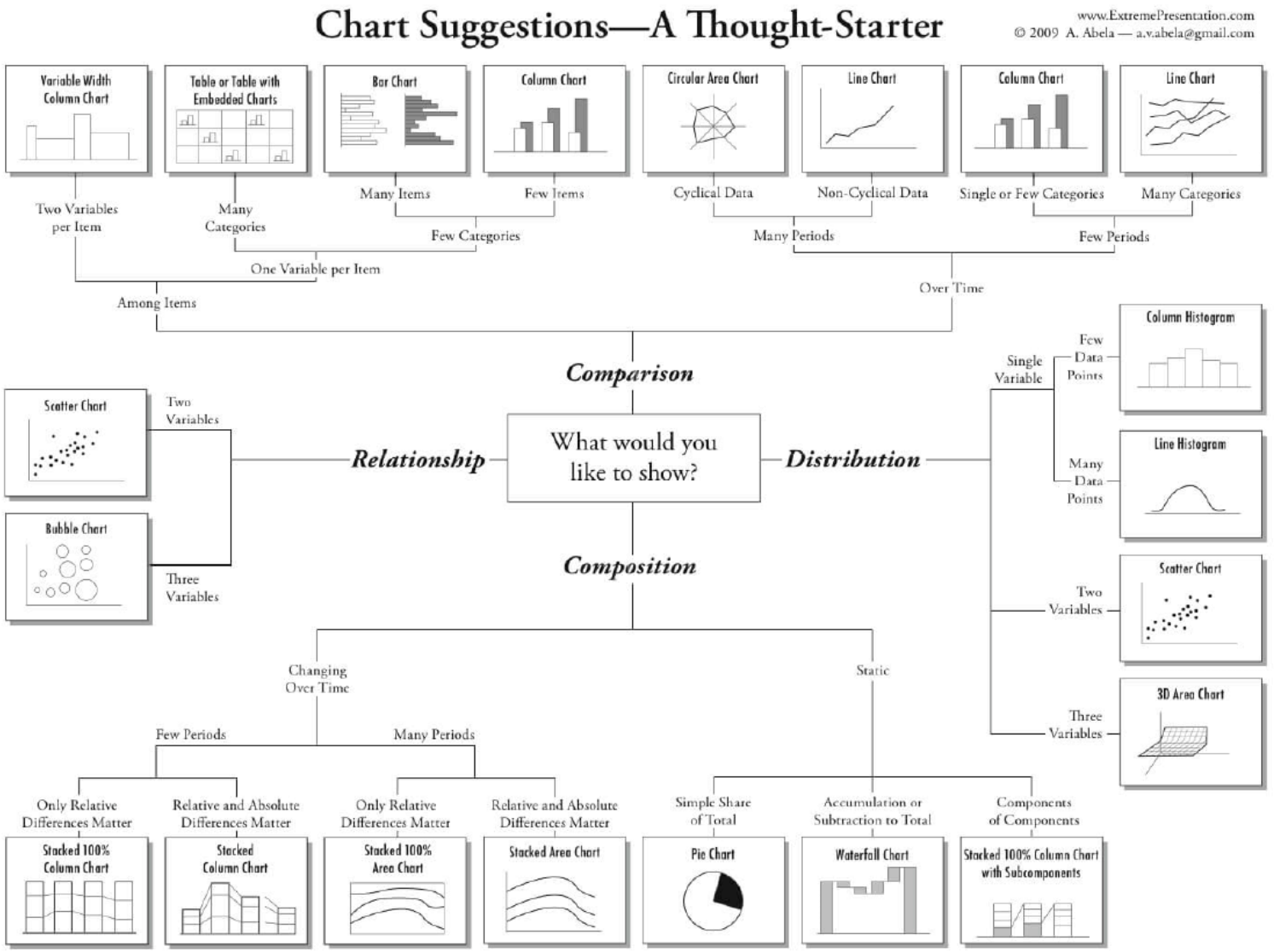
1. Communication

- Analysis only goes so far if you can't share and implement your findings.
- Analytical communication skills include:
  - Problem sensitivity
  - Active listening
  - Reporting
  - Surveying
  - Teamwork
  - Oral communication
  - Written communication
  - Conducting presentations

Analytical Thinking

1. Communication

- Data Visualization



2. Creativity

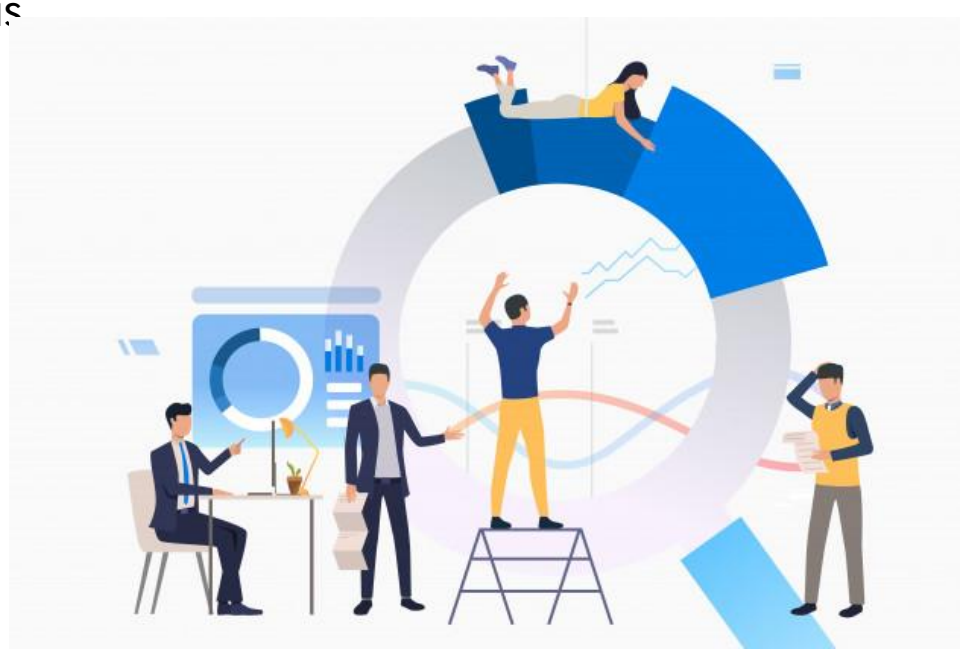
- Open-Mindedness
- The obvious solution is not always the best option.
- Creative skill sets include:
  - Budgeting
  - Brainstorming
  - Collaboration
  - Optimization
  - Predictive modeling
  - Restructuring
  - Strategic planning
  - Integration

3. Critical Thinking

- Evaluating information and then making a decision based on your findings.
- Critical Thinking skills include:
  - Process management
  - Big data analytics
  - Business intelligence
  - Case analysis
  - Causal relationships
  - Classifying
  - Comparative analysis
  - Correlation
  - Decision-making
  - Diagnostics
  - Data interpretation
  - Prioritization

4. Data Analysis

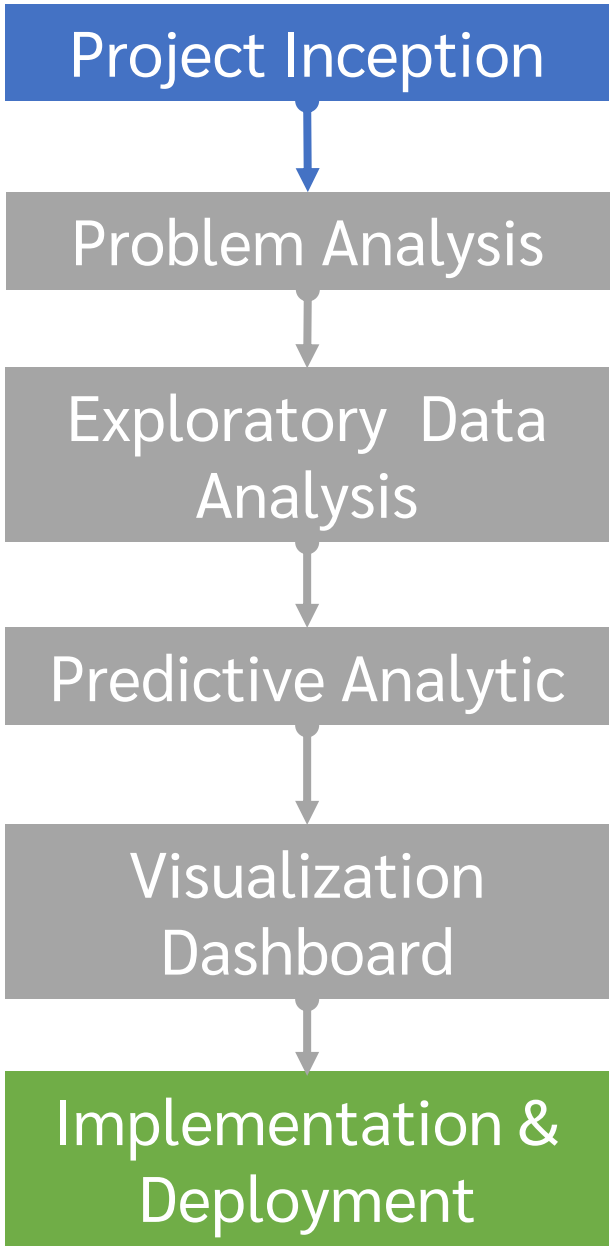
- Able to examine a large volume of data and identify trends in that data.
- Go beyond just reading and understanding information to make sense of it by highlighting patterns for top decision-makers.
- Data Analysis skills include:
  - Business analysis
  - Strengths, weaknesses, opportunities, and threats (SWOT) analysis
  - Descriptive analysis
  - Financial analysis
  - Policy analysis
  - Predictive analytics
  - Prescriptive analytics
  - Process analysis
  - Qualitative analysis
  - Quantitative analysis
  - Return on investment (ROI) analysis



5. Research

- Learn more about a problem
- Able to collect data and research a topic. This can involve reviewing spreadsheets, researching online, collecting data, and looking at competitor information.
- Analytical research skills include:
  - Investigation
  - Metrics
  - Data collection
  - Prioritization
  - Checking for accuracy

Data Project Design Workshop



Data Scientists และคณะทำงาน Big Data ของแต่ละกลุ่มงาน ร่วมกำหนดโจทย์ที่เหมาะสม และตั้งโครงการพัฒนาและทดสอบโมเดลคณิตศาสตร์นำร่องที่เหมาะสม

ทีม Data Scientists สำรวจข้อมูลที่มีอยู่ในปัจจุบันตามโจทย์นำร่องที่กำหนดเพื่อประเมินความพร้อมและแปลงจากความต้องการในเชิงปัญหาให้เป็นข้อกำหนดในเชิงข้อมูลและระบบ

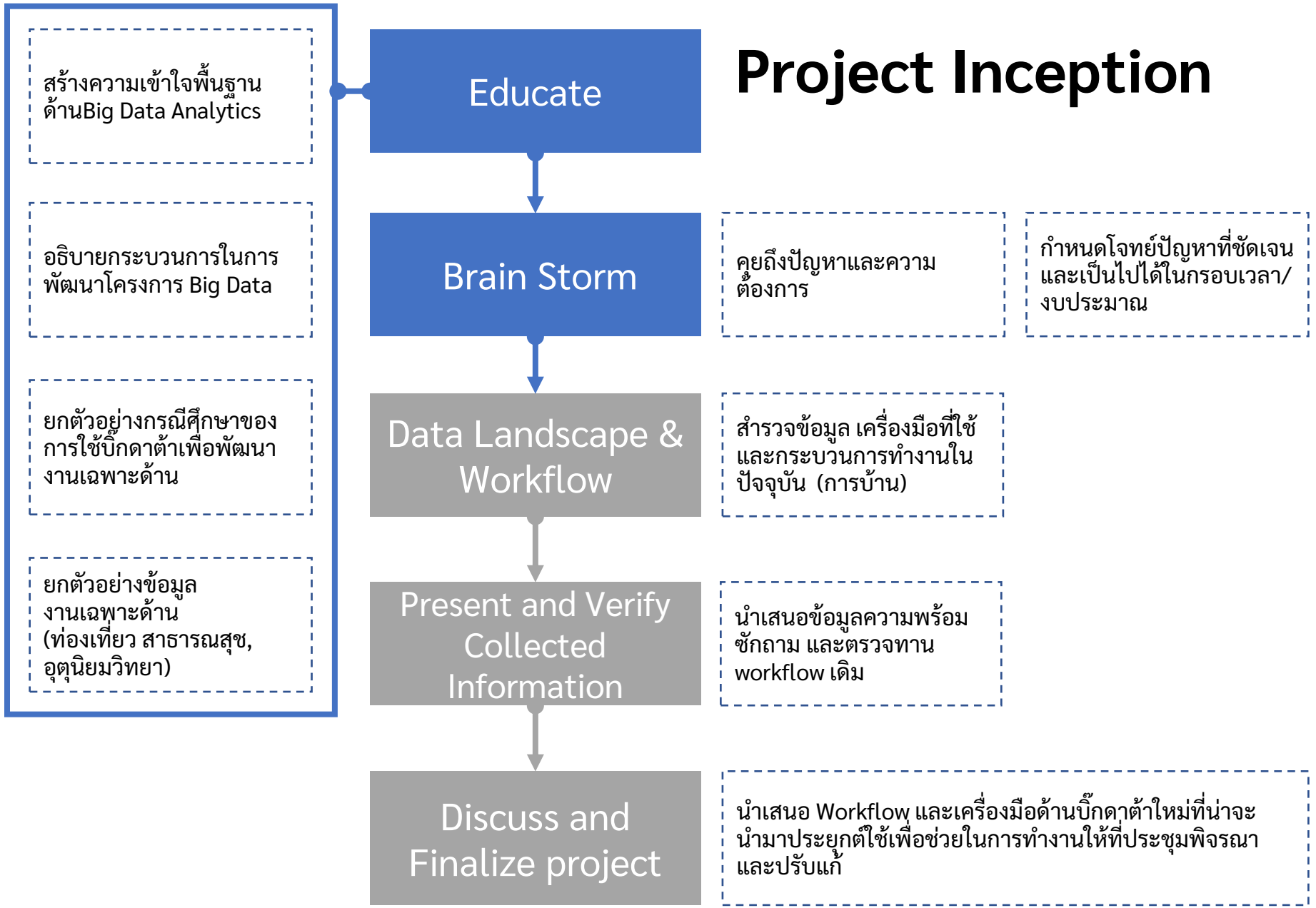
สำรวจการกระจายตัวของข้อมูลเพื่อทำความเข้าใจข้อมูล และหาความสัมพันธ์ระหว่างตัวแปรในข้อมูล ในขั้นตอนนี้ทีมงานจะต้องการตัวอย่างข้อมูลจริง ระบุข้อมูลที่ต้องการเพิ่มเติม และเริ่มเตรียมข้อมูลจริง

ทีมงานนำข้อมูลที่จัดเตรียมในเบื้องต้นมาใช้ในการสร้างแบบจำลองหรือโมเดลทางคณิตศาสตร์เพื่อการทำนาย โดยใช้เทคนิคและอัลกอริธึ่มต่างๆ และทดสอบความแม่นยำของโมเดลคณิตศาสตร์

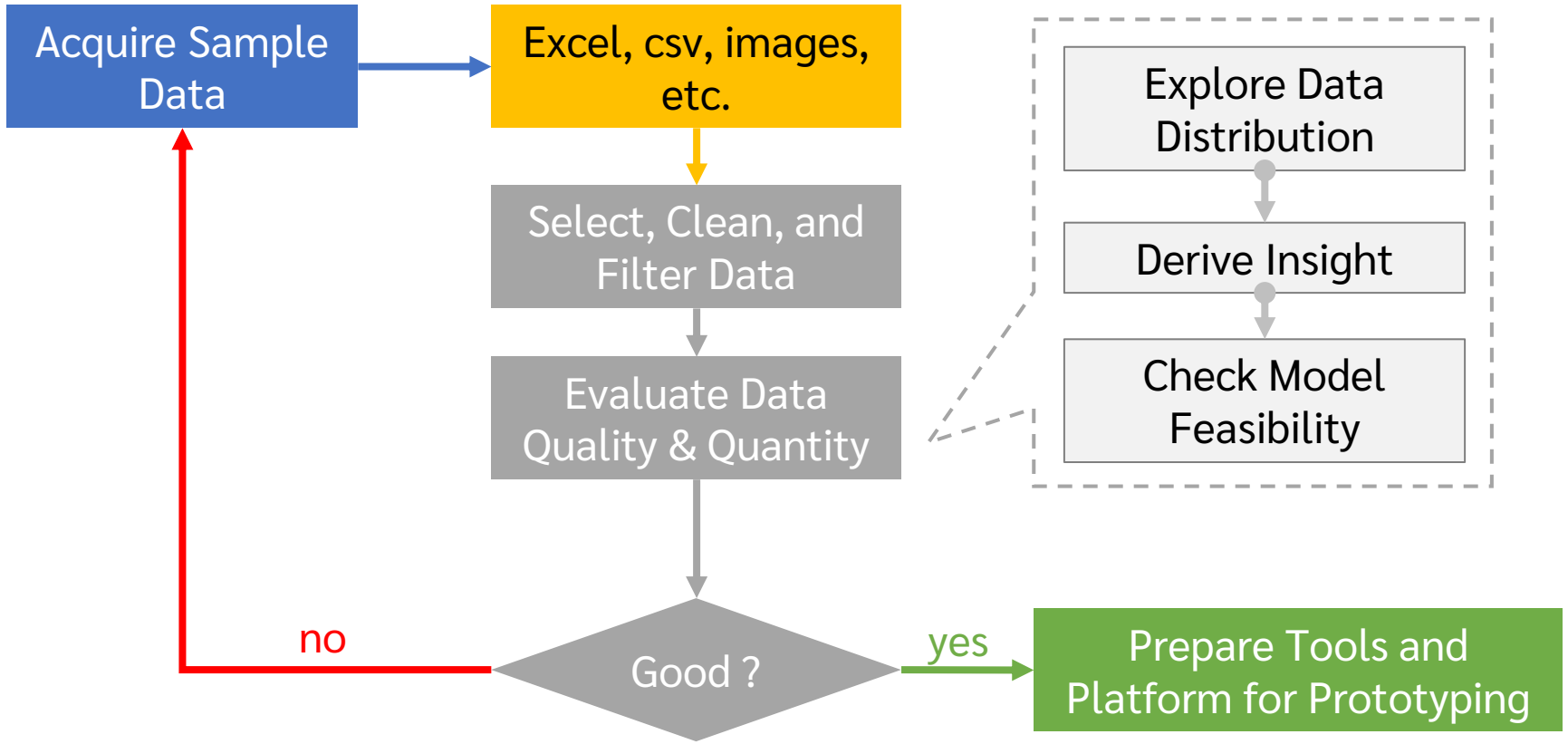
ออกแบบวิธีการแสดงผลโดยเลือกมิติของข้อมูลที่เหมาะสมบน Interactive Dashboard เพื่อให้คณะทำงานทดลองใช้และสื่อสารกับทีมผู้บริหาร และ ผู้ปฏิบัติ ให้สามารถนำเอาความเข้าใจดังกล่าวไปแปลงเป็นแผนการพัฒนาต่อยอด

หลังจากผลลัพธ์เป็นที่พอใจแล้ว นักพัฒนาระบบเริ่มพัฒนาโปรแกรมตามรูปแบบของโมเดลคณิตศาสตร์ที่วางไว้ และตั้งค่าให้โปรแกรมให้ประมวลผลโมเดลแบบอัตโนมัติตามความถี่ที่วางแผนไว้ จากนั้นติดตั้งระบบซอฟท์แวร์เพื่อการใช้งานจริง

Project Inception



Exploratory Data Analysis (EDA)



jupyter RStudio rapidminer WEKA

Predictive Analytics

Data Mining

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

Text Analytics

The process of deriving high-quality information from **text**. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning.

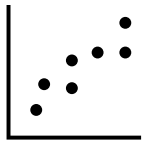
Machine Learning / Deep Learning

The science of getting computers to learn from data without having to be explicitly programmed by humans. Machine model can teach themselves to grow and change when exposed to new data.

Big Data Technology

Technology designed to manage and process extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions.

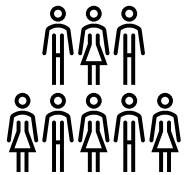
Example Tasks



Regression

produce a model that, given an individual, estimates the value of the particular variable specific to that individual.

- *How much will a given housing unit cost ?*



Clustering

group individuals in a population by their similarity (not driven by any specific purpose).

- *Do our customers form natural groups or segments?*

Co-occurrence Grouping

find associations between entities based on transactions involving them.

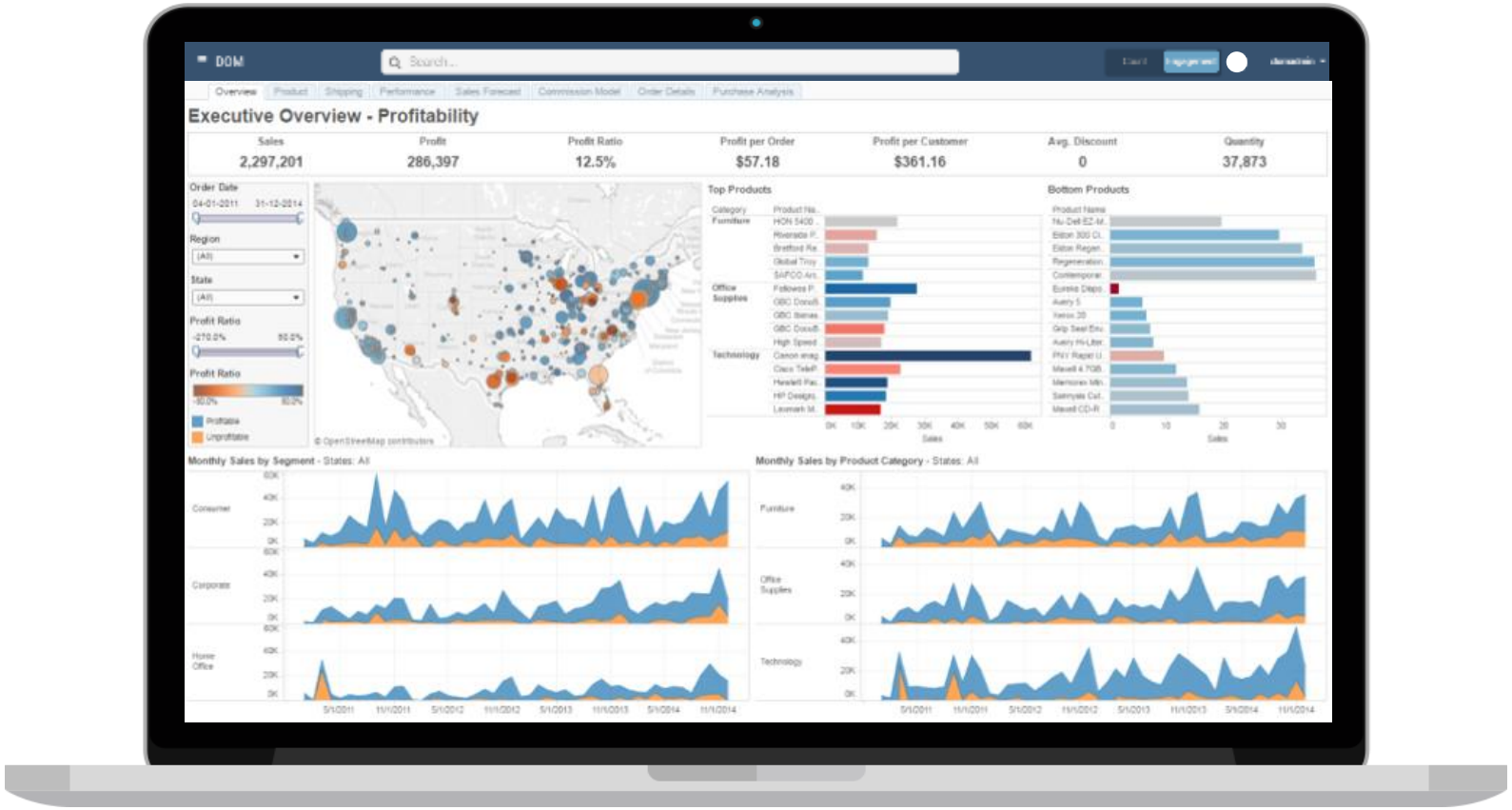
- *What items are commonly purchased together?*

Profiling

characterize the typical behavior of an individual, group, or population.

- *What is the typical service usage of this customer segment ?*
- *Used to establish behavior norms for anomaly detection (fraud detection)*

Visualization Dashboard



tableau

Power BI

SAP Business Objects

pentaho

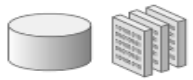
etc.

Implementation and Deployment

Data Input

Messaging, and Web Services

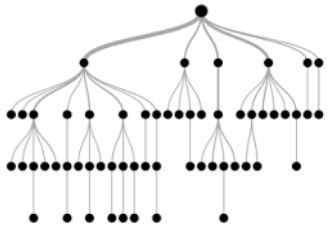
EDW, OLAP



Social Media, Weblogs

Machine Devices, Sensors

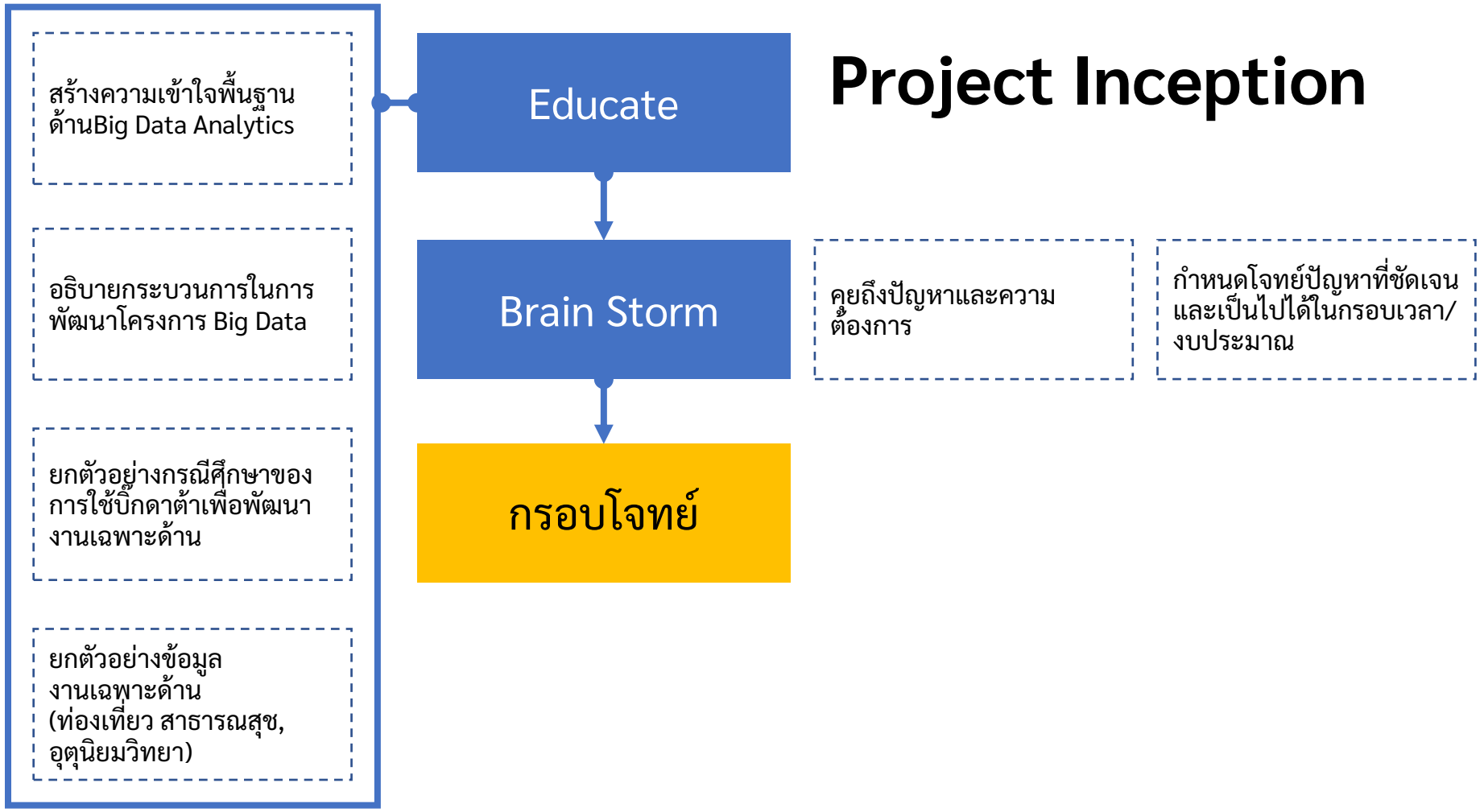
Predictor Software



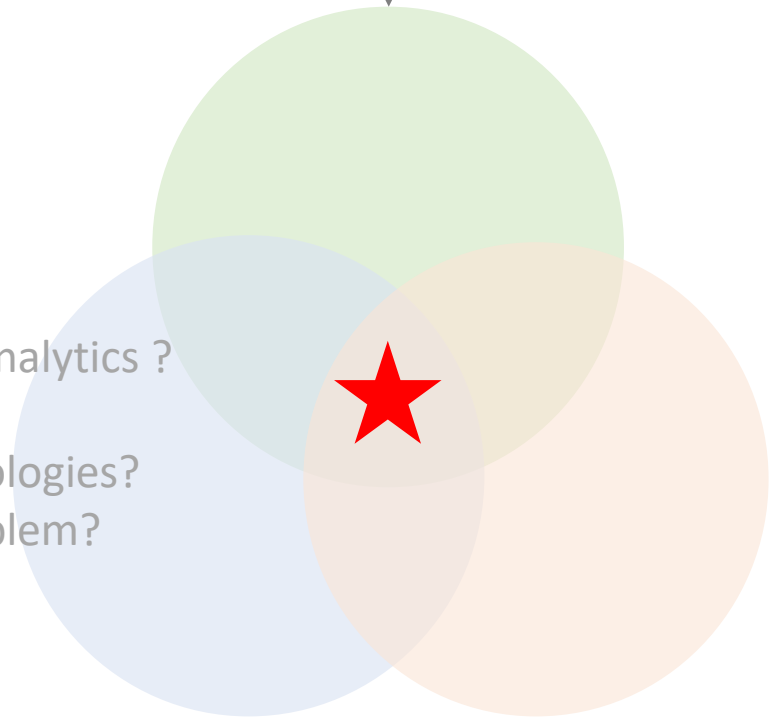
IT Infrastructure

Visualization

Project Inception



Pain points



Feasibility

What is technically & organizationally feasible ?

Can those needs be met using data analytics ?
Do we have or can we acquire data?
Do we have or can we acquire technologies?
Do we have people to tackle the problem?

Desirability

What do people desire?

Share stories about pain points at work.
Do you have real needs to
get rid of those pain points ?

Viability

What can be financially viable?

The cost of acquiring data ?
The time needed to acquire data ?

The analytic solutions that emerge should hit the overlap of these three lenses:
They need to be Desirable, Feasible and Viable

Project Design Canvas

GBDi

Ref: Design Thinking from SEE Program at Babson College

Name of the Project

Sentence description

Brief Concept

What is it and why is it good ?

Needs being met

List 1-3 needs

Visual Concept

Diagram that describe what does it look like ?

Values Created

Economic, Social, reputational, community, etc.

Team

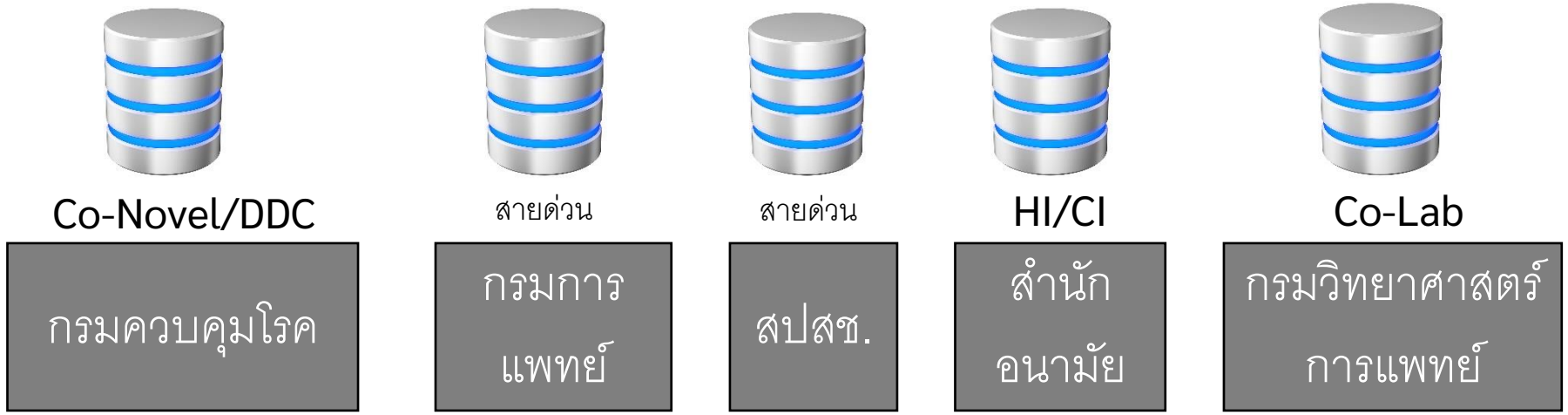
Who should be put in this project ?

Resources at hand

Data and Data format

CO-Link:
Covid Data Linkage

ตัวอย่างความจำเป็นของการมี Single Portal เพื่อบริหารและบูรณาการข้อมูล



โรงพยาบาลต้องการ

- ไม่ต้องค้นหา lab ซ้ำ ไม่ต้องซักอาการซ้ำ ไม่ต้องหาที่ดูแลผู้ป่วยซ้ำ ไม่ต้องตามคนไปรับส่งผู้ป่วยซ้ำ ไม่ต้องถามว่าเตียงว่างไหม อยากดูจากระบบเดียวได้ เพราะหน้างานยุ่งมากอยู่แล้ว
- รับ-ส่งผู้ป่วย / ย้ายผู้ป่วย จากเบาไปหนัก หรือ จากหนักไปเบา ได้สะดวก
- ให้มีทีมคอยช่วยเหลือโรงพยาบาลที่ต้องการ รับ-ส่ง ข้อมูลเพื่อบริการผู้ป่วยทั้ง Journey ผ่าน APIs
- ให้เอกชนที่อยากเข้ามาช่วย (ที่ได้รับการรับรองจาก สธ) ทั้งปัจจุบันและอนาคต สามารถนำข้อมูลนี้ ไปช่วยกันทำงานต่อให้โรงพยาบาลได้

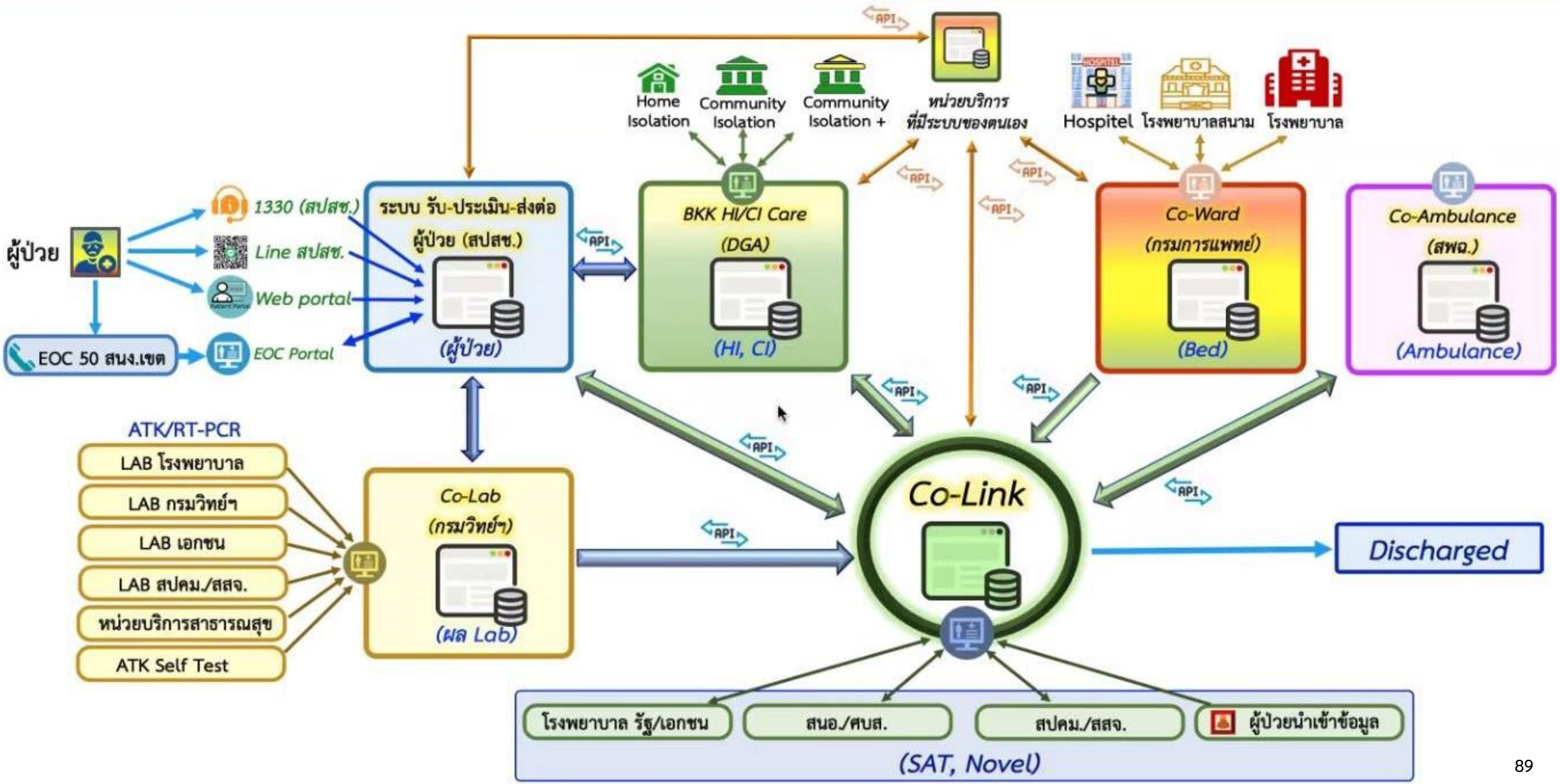
เริ่มต้น

ระบบการบูรณาการข้อมูลเพื่อการควบคุมโรคและการรักษาพยาบาล

GBDi

- ผลตรวจ Covid-19 เน้นกลุ่มที่ ผล เป็น positive
- การจัดการเตียง (ใครมารับใคร ส่งไปที่ไหน)
- ทะเบียนผู้ติดเชื้อ (กรม คร ต้องติดตามเพื่อสอบสวนโรค. รพ หรือ ศูนย์ อื่นๆ ต้องมารับตัว)
- การครองเตียง refer/discharge
- ความต้องการวัคซีน (บูรณาการข้อมูลลงทะเบียนจากระบบต่างๆ)

ตัวอย่างความจำเป็นของการมี Single Portal เพื่อบริหารและบูรณาการข้อมูล



Travel Link:
Thai Tourism Data Platform

Travel Link The Thai Tourism Data Platform

ขับเคลื่อนเศรษฐกิจการท่องเที่ยวด้วยข้อมูลเปิดภาครัฐ (Open Government Data)

- ชุดข้อมูลสถานที่ท่องเที่ยว (สป.ก.ท่องเที่ยว, ททท, กรมการท่องเที่ยว, อพท, ก. วัฒนธรรม, ฯลฯ)
- ชุดข้อมูลโรงแรม ที่พัก (กรมการปกครอง และ สำนักงานสถิติแห่งชาติ)
- ชุดข้อมูลการเดินทางโดยสารสาธารณะ
- ชุดข้อมูลพฤติกรรมนักท่องเที่ยว

Provided by MoU 22 หน่วยงาน

- ชุดข้อมูลที่เกี่ยวข้องกับการท่องเที่ยว จากแหล่งต่าง ๆ เช่น Google, Reddit, Pantip.com, Social Media

Data Collection and Integration

- **มาตรฐานข้อมูล (Data Standard)**
- **Data Catalog**
- **Open Data Service APIs**
- **AI Service APIs**
  - ระบบค้นหาสถานที่ท่องเที่ยว
  - ระบบแนะนำสถานที่ท่องเที่ยว
- **Data Analytics Dashboards**
- **Data Analytics Models**

หน่วยงานภาครัฐ: มีเครื่องมือสนับสนุนการขับเคลื่อน (Implement) นโยบาย

ภาคเอกชน: มีข้อมูลและระบบปัญญาประดิษฐ์ช่วยกระตุ้นธุรกิจ

ประชาชน: ได้รับความสะดวกสบายจากข้อมูลและระบบปัญญาประดิษฐ์

** Developed by GBDi*

Travel Link Website

- Data Catalog
- Data and AI service API
- Dashboard ข้อมูลท่องเที่ยว

3rd Party Applications

(หน่วยงานภายใต้ MOU, SMEs, Startups)

Data Sources *Data Integration and Analytics Services* *Presentations*

Attraction Data and AI services

บริการค้นหาสถานที่ท่องเที่ยวอัจฉริยะ (Smart Search)

- ค้นหาด้วยคำสำคัญ (Keyword Search)
- ค้นหาด้วยสถานที่ที่คล้ายกัน (Similar Search)

บริการแนะนำสถานที่ท่องเที่ยว และประมาณค่าใช้จ่าย
(Attraction Recommendation and Budget Estimation)

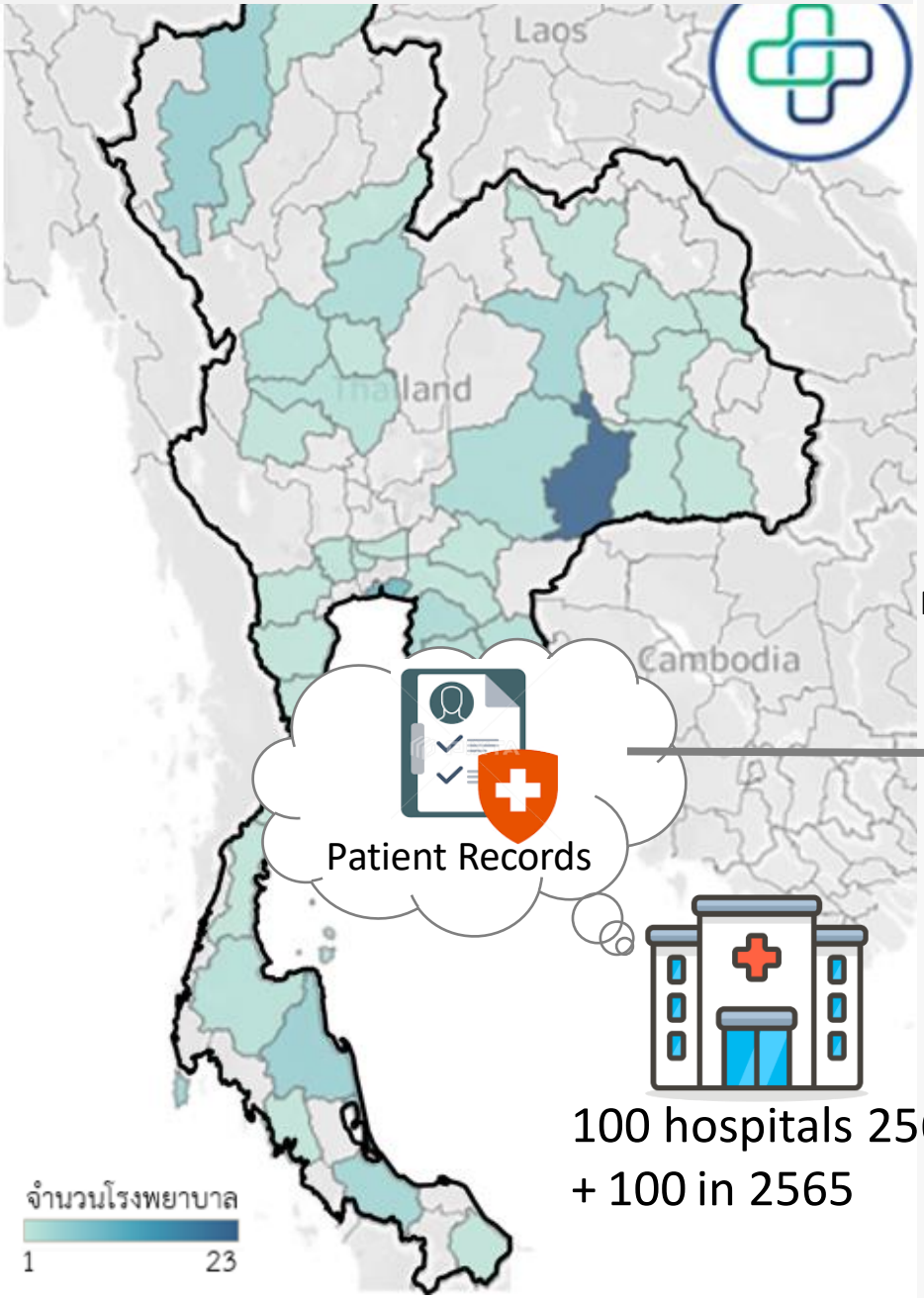
GBDi

HEALTH LINK

Health Information Exchange System

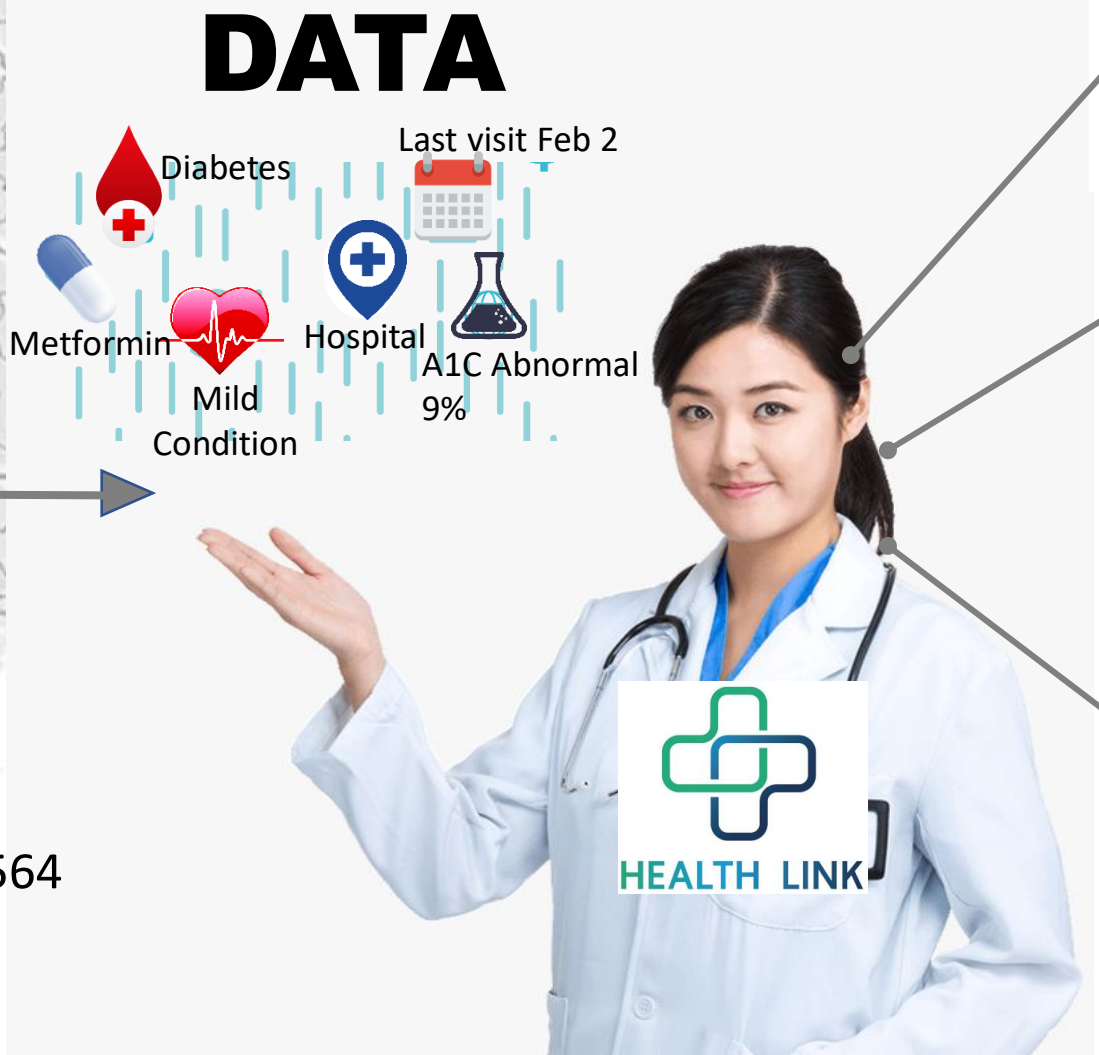
https://www.healthlink.go.th/

รวบรวมข้อมูลผู้ป่วยจากโรงพยาบาล



100 hospitals 2564
+ 100 in 2565

ส่งข้อมูลผู้ป่วยถึงมือแพทย์ทุกที่อย่างรวดเร็ว



GBDi

ประชาชนจะ ....

ประหยัดเวลาเดินทางไปขอประวัติ และค่าใช้จ่ายในการตรวจซ้ำโดยไม่จำเป็น

ลดความเสี่ยงโดยแพทย์ตรวจสอบประวัติเบื้องต้นคนไข้ได้ทันที กรณีรักษาฉุกเฉิน

ลดความเสี่ยงในการไปโรงพยาบาล โดยรับคำปรึกษาผ่าน Telemedicine ซึ่งแพทย์สามารถเปิดดูประวัติการรักษาได้อย่างสะดวก