Data Products

Prepared by Dr. Salimah Mokhtar

What comes to mind?

The next bus will arrive in 10 minutes.

The price of a hotel reservation for next week is $97.

What do these people have in common?

Learning Objectives:-

1. To explain what data product is.
2. To categorize types of data products.
3. To describe data products functions.
4. To express interfaces or interactions.
5. To be familiar with feature engineering.
6. To discuss on deployment / productization.

The Age of the Data Product

- The **information revolution** - driven as it is by networked communication systems and the Internet, is unique in that it has created a surplus of a valuable new material — **data** — and transformed us all into both **consumers** and **producers**.
- The sheer amount of data being generated is **tremendous**. Data increasingly affects every aspect of our lives.
- We have developed a reasonable **expectation** for products and services that are highly personalized and finely tuned to our needs, creating a market for a new information technology — the **data product**.

https://www.oreilly.com/library/view/data-analytics-with/9781491913734/ch01.html

What is Data Product?

Data science is about **insights** (not information, not technology)

- The best insights are **actionable**.

Data products are products that derive substantial **value** from **analytics** and whose primary objective is to **use data to facilitate an end goal**.

- **Data product**: Google Search, Google Analytics
- **Information product**: Apps on mobile, online courses

Data products are the reason data scientists are lately treated like rock stars.

What is Data Product?

A **data product** is an application or tool that uses data to help businesses improve their decisions and processes.

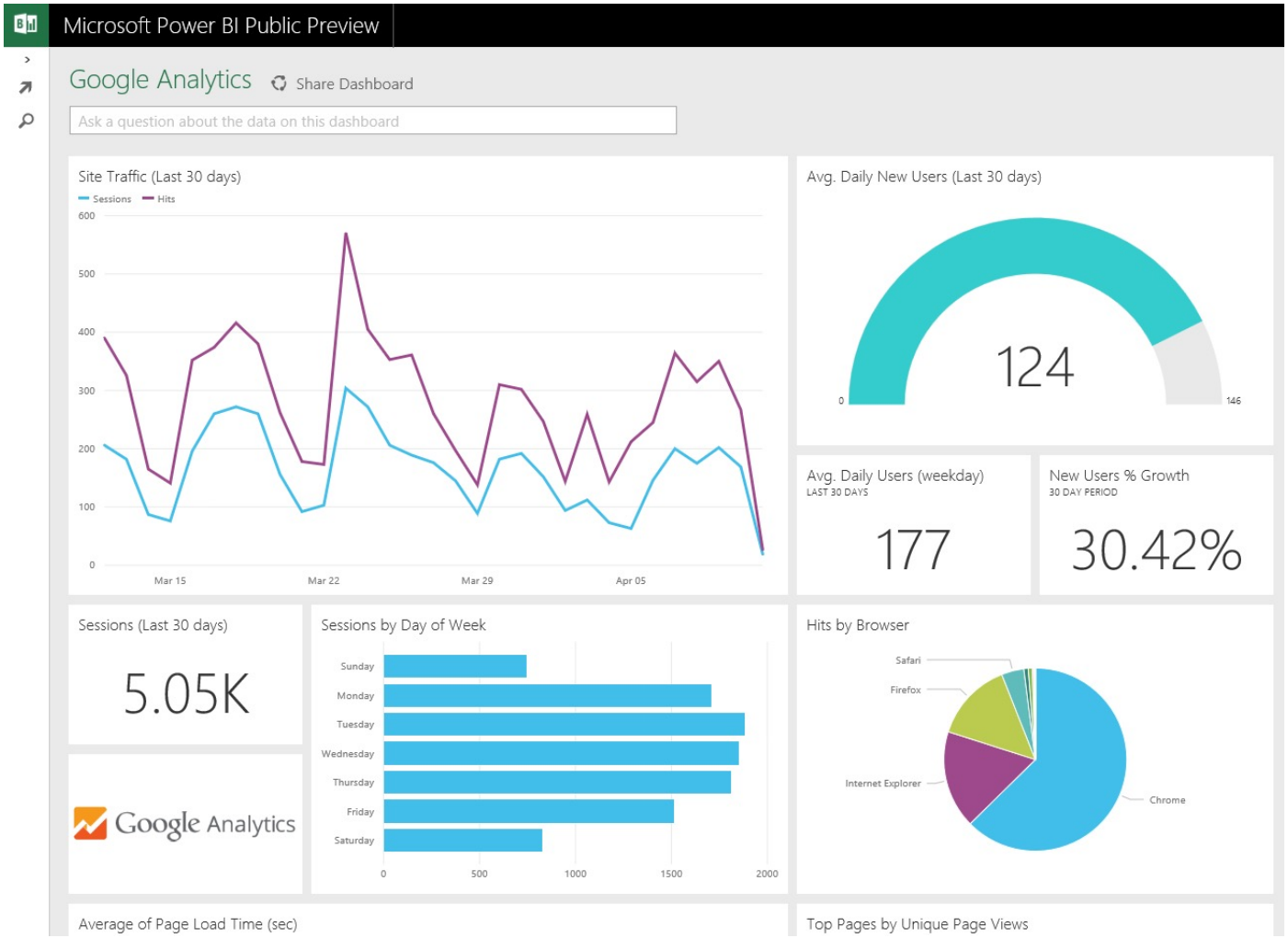
Data products that provide a friendly user interface can use data science to provide predictive analytics, descriptive data modeling, data mining, machine learning, risk management, and a variety of analysis methods to non-data scientists.

https://www.tableau.com/learn/whitepapers/turn-data-products-data-scientist-data-business-owner

Examples of Data Products

The **Bloomberg Terminal** is a computer software system provided by the financial data vendor Bloomberg L.P. that enables professionals in the financial service sector and other industries to access Bloomberg Professional Services through which users can monitor and analyze real-time financial market data and place trades on the electronic trading platform. (Wikipedia)

Salesforce Einstein integrates AI technology with Salesforce's Software-as-a-Service (SaaS) CRM. It uses data gathered on every user action to provide predictive analytics, natural language processing (NLP) capabilities, and machine learning to Salesforce customers.



Google Analytics is a web analytics service offered by Google that tracks and reports website traffic, currently as a platform inside the Google Marketing Platform brand.

> A data application acquires its value from the data itself, and creates more data as a result. It's not just an application with data; it's a data product.

"The future belongs to the companies and people that turn data into products." -- Mike Loukides, O'Reilly

Notice the Different!

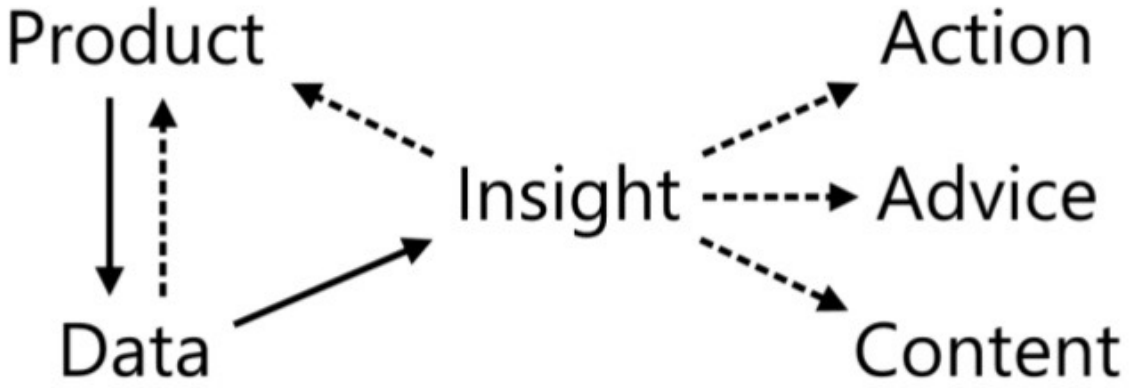
Information Product - It's any product or service that you can sell to people to provide them with information, usually about a specific topic.

e.g., Instruction manual, news service and online directory.

A **knowledge product** is a result of human thought that has value.

e.g., Lesson learned report, a summary of best practices, thesis and journal articles.

Data products are created with data science workflows, specifically through the application of models, usually predictive or inferential, to a domain-specific dataset.

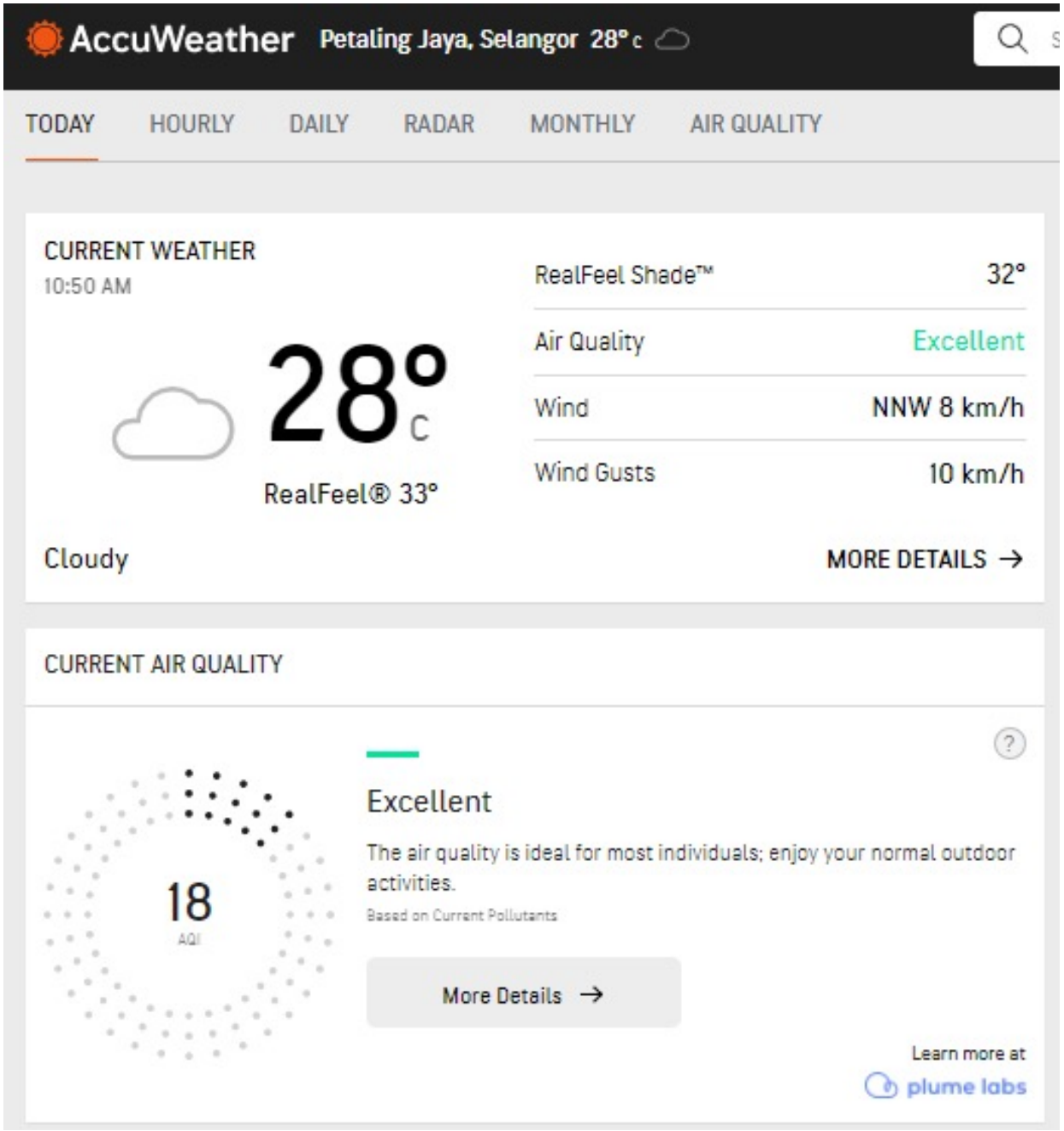


A data product is an **economic engine**.
It derives value from data and then produces more data, more value, in return.

3 Types of Data Products

1. Data as a Service

- **Data** itself is the product.
- These products are offered to users as either a **paid** for or **free service**.
- All data products that create **direct revenue** fall into this category.
- Companies offering this type of data product provide **data for specific interests** such as to-the-second accurate stock-market data or location-specific weather data.
- e.g., AcuWeather, Gro Intelligence


AccuWeather
Petaling Jaya, Selangor 28° c
TODAY
HOURLY
DAILY
RADAR
MONTHLY
AIR QUALITY
CURRENT WEATHER
10:50 AM
28° C
RealFeel® 33°
Cloudy
RealFeel Shade™ 32°
Air Quality Excellent
Wind NNW 8 km/h
Wind Gusts 10 km/h
MORE DETAILS →
CURRENT AIR QUALITY
18
AQI
Excellent
The air quality is ideal for most individuals; enjoy your normal outdoor activities.
Based on Current Pollutants
More Details →
Learn more at
plume labs

Gro Intelligence

Sign up to view more data

Getting the Food to Where It's Needed

The **West's growth** in **surplus food** more than **covers future deficits** expected in **Asia** and **sub-Saharan Africa**.

But the **tough logistical challenges** of **getting food** to **where it must go underlines** the **need** for **regions** to be somewhat **self-sufficient**.

The **Indian subcontinent** generally **grows enough food** for the **people** living in the region,

but the **population** is so **huge** that any **shock to supply** or **demand** immediately **affects world market balances**.

Europe and the states of the **former Soviet Union** expect **declining populations**.

That means their **agricultural exports** will be **rising** as **production** continues to **grow** and **demand shrinks**.

Sources: World Bank, FAO, Gro Intelligence

2. Data-Enhanced Products

- These are **data-based additional functions** which modify a traditional product to increase its value.
- Data products which **enhance** a physical or virtual product fall into this second category.
- The value of such a product is reflected in the change in revenue (price or quantity) of the enhanced product.
- Most **recommenders** fall in this category, as they improve the sales of products.
  - 35% of Amazon's revenue comes from recommendations - and why 75% of Netflix content is consumed based on recommendations.

Recommender

Highest Rated IMDb "Top 250" Titles

1 to 50 of 249 titles | Next »

View Mode: Compact | **Detailed**

Sort by: Popularity | Alphabetical | **IMDb Rating▼** | Number of Votes | US Box Office | Runtime | Year | Release Date

1. The Shawshank Redemption (1994)

R | 142 min | Drama

9.3 Rate this 80 Metascore

Two imprisoned men bond over a number of years, finding solace and eventual redemption through acts of common decency.

Director: Frank Darabont | Stars: Tim Robbins, Morgan Freeman, Bob Gunton, William Sadler

Votes: 2,008,037 | Gross: $28.34M

2. The Godfather (1972)

R | 175 min | Crime, Drama

9.2 Rate this 100 Metascore

The aging patriarch of an organized crime dynasty transfers control of his clandestine empire to his reluctant son.

Director: Francis Ford Coppola | Stars: Marlon Brando, Al Pacino, James Caan, Diane Keaton

Votes: 1,375,755 | Gross: $134.97M

3. Data as Insights

Insights here are defined as actionable, data-driven findings that create business value.

- These are products that **analyze data to provide insights to decision maker** within an organization.

  e.g., Google Analytics, Tableau, Cloudera / Hortonworks / MapR

- **Insights as a Service** is a software service that specifically delivers quality, actionable insights. Typically, such services are hosted in the cloud.

Data products: types

Type 1	Type 2	Type 3
Data as a Service	Data-enhanced Products	Data as Insights
› Weather data	› Autonomous driving	› Marketing planning

› Recommendations

An Example of Data Product based on Predictive Modeling

Recommender systems

A subclass of information filtering systems that are meant to predict the preferences or ratings that a user would give to a product. Recommender systems are widely used in movies, news, research articles, products, social tags, music, etc.

Amazon recommender system

Amazon examines items customers have purchased, and based on similar purchase behavior of other users, makes recommendations. In this case, order history data is combined with recommendation algorithms to make predictions about what customer might purchase in the future.

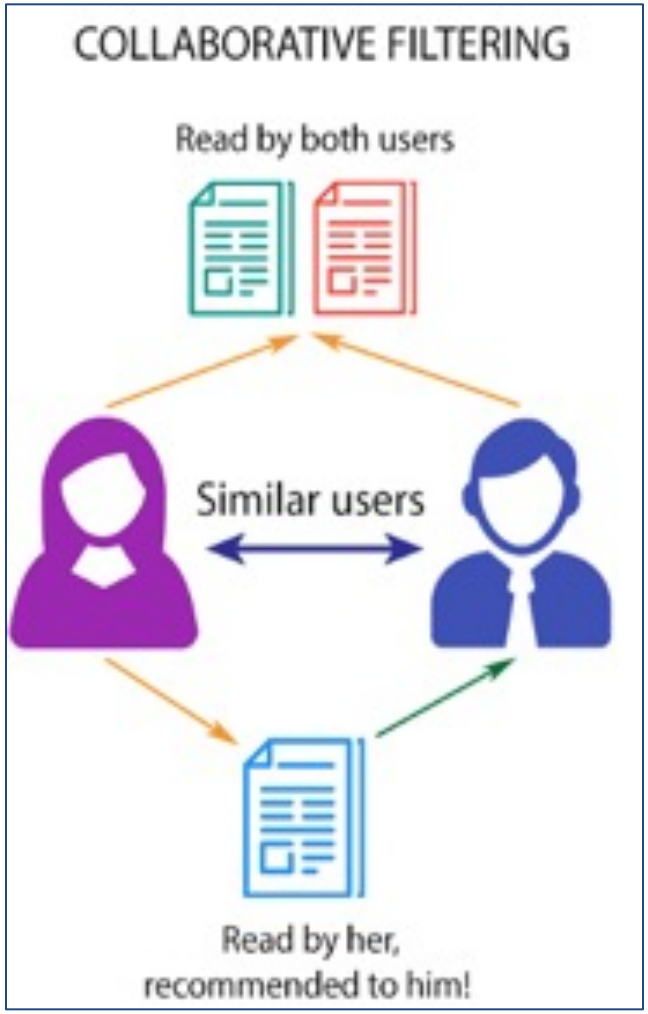
Recommendation Engines

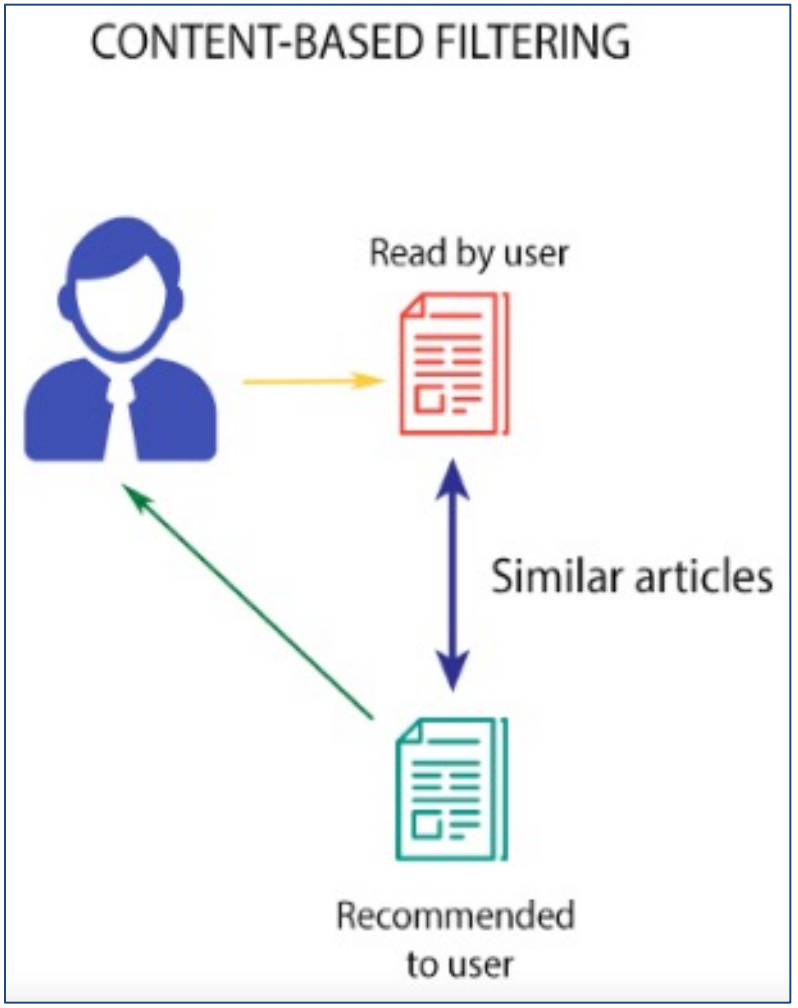
- **Recommendation engines** filter out the products that a particular customer would be interested in or would buy based on his or her **previous buying history**.
- The **more data** available about a customer the **more accurate** the recommendations.

https://www.iteratorshq.com/blog/an-introduction-recommender-systems-9-easy-examples/

Types of Recommender Systems

The TWO main types of recommender systems are:





Collaborative Filtering

A method of **making automatic predictions** (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating).

Typically, the **workflow of a collaborative filtering system** is:

1. Look for users who share the same rating patterns with the active user (the user whom the prediction is for).
2. Use the ratings from those like-minded users found in step 1 to calculate a prediction for the active user.

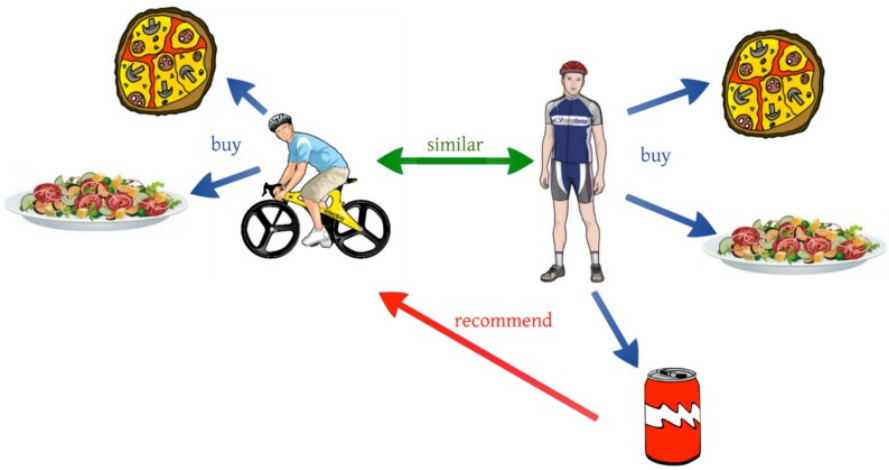
Two types of collaborative filtering techniques are used:

1. User-based collaborative filtering
2. Item-based collaborative filtering

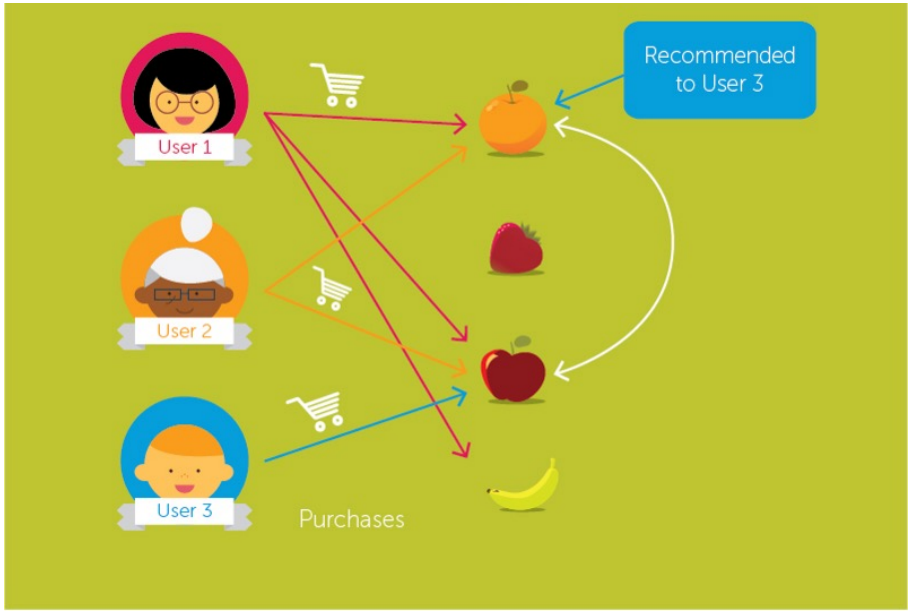
https://towardsdatascience.com/brief-on-recommender-systems-b86a1068a4dd

Two types of collaborative filtering techniques

User-based collaborative filtering

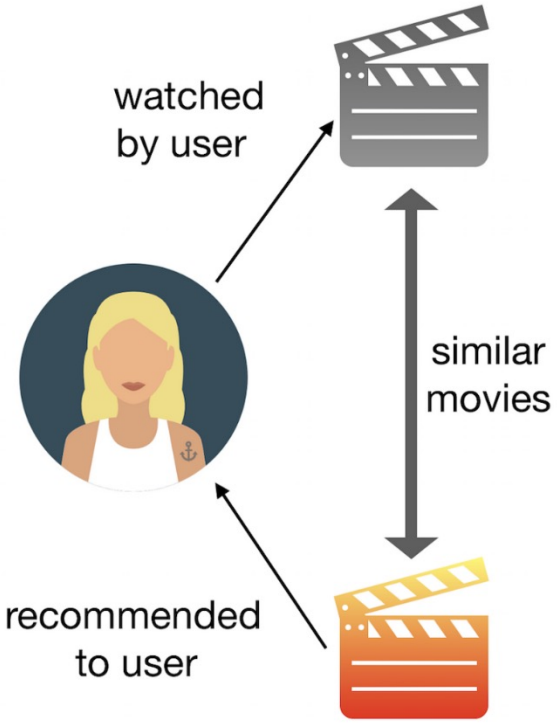


Item-based collaborative filtering



Content-based filtering

- This filtering is based on the **description**, or some **data** provided for that product.
- The system finds the **similarity** between products based on its context or description. The user's previous history is taken into account to find similar products the user may like.

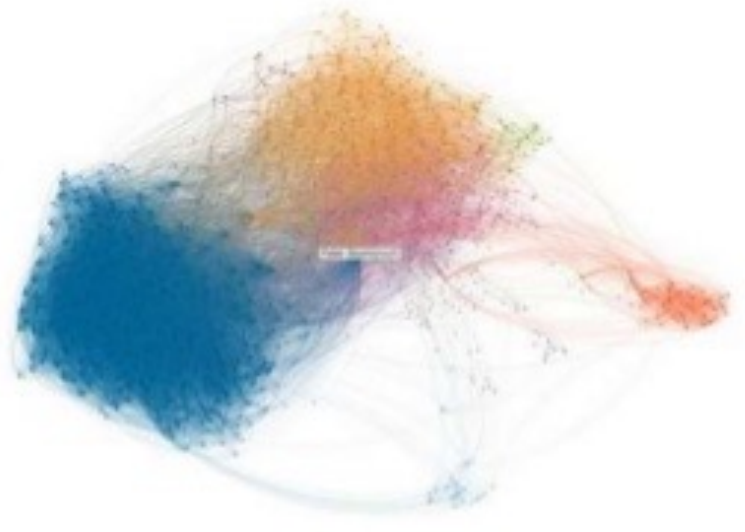


If a user likes movies such as 'Mission Impossible' then we can recommend him the movies of 'Tom Cruise' or movies with the genre 'Action'

Building Products from Data at LinkedIn

A few examples:

- People You May Know
- Skills and Endorsements
- Year in Review
- Network Updates Digest
- InMaps
- Who's viewed my profile
- Collaborative Filtering
- Groups You May Like
- and more…



Active vs Passive Data

- **Active / explicit data** – user needs to actively provide the data (e.g., User ratings and reviews).
- **Passive / implicit data** – data collection in which data is gathered automatically often without user knowledge (e.g., user clicks and views).

Data Products Functions

FIVE broad groups of data products functions:

1. Raw data,
2. Derived data,
3. Algorithms,
4. Decision support and
5. Automated decision-making

https://towardsdatascience.com/designing-data-products-b6b93edf3d23

(1) Raw Data

- Starting with **raw data**, we are collecting and **making available data as it is** (perhaps we're doing some small processing or cleansing steps).
- The user can then choose to use the data as appropriate, but most of the work is done on the user's side.

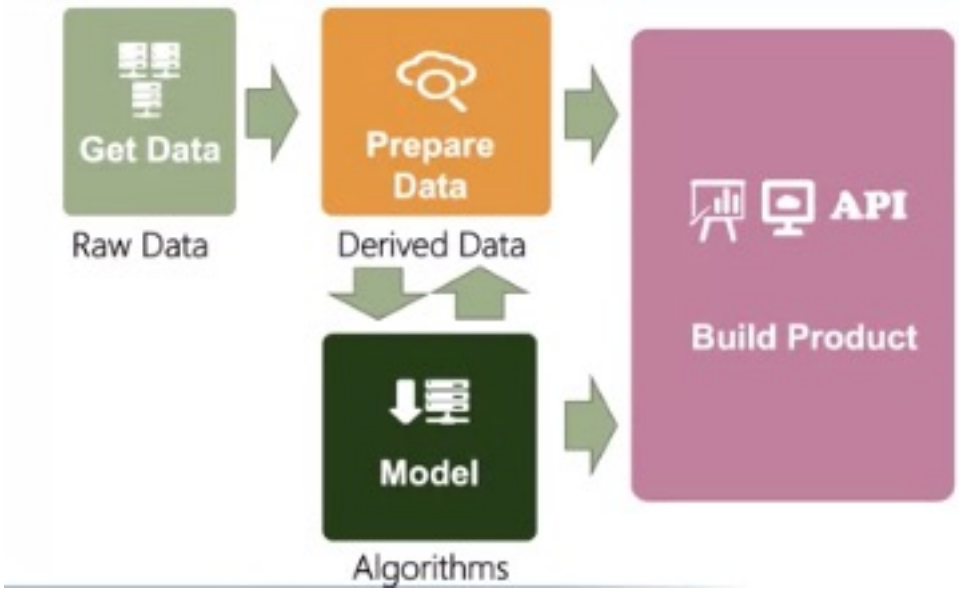
http://Gnip.com - the official reseller of Twitter's data(tweets)

(2) Derived Data

- In providing users with **derived data**, we are doing some of the processing on our side.
- In the case of customer data, **add additional attributes** like assigning a customer segment to each customer, or we could add their likelihood of clicking on an ad or of buying a product from a certain category.

What is Derived Data?

- A **derived data** element is a data element derived from other data elements using a mathematical, logical, or other type of transformation, e.g., arithmetic formula, composition, aggregation.



(3) Algorithms

Algorithms, or **algorithms-as-a-service**.

- We are given some data, we run it through the algorithm - be that machine learning or otherwise - and we return information or insights.
- A good example is **Google Image**: the user uploads a picture and receives a set of images that are the same or similar to the one uploaded.
- Behind the scenes, the product extracts features, classifies the image and matches it to stored images, returning the ones that are most similar.

(4) Decision Support

- Providing **information** to the user to **help them with decision-making** but we are not taking the decision ourselves.
- **Analytics dashboards** such as Google Analytics, Flurry, or WGSN would fall into this category.
- Give the user relevant information in an **easy-to-digest format** to allow them to take better decisions.
- In the case of **Google Analytics**, that could mean changing the editorial strategy, addressing leaks in the conversion funnel, or doubling down on a given product strategy.

The **important thing to remember** here is as follows:

- while user have taken design-decisions in data collection, derivation of new data, in choosing what data to display and how to display it, the users are still tasked with **interpreting** the data themselves.
- Users are **in control of the decision** to act (or not act) on that data.

(5) Automated Decision-Making

Here we outsource all of the intelligence within a given domain.

- **Netflix** product recommendations or **Spotify's** Discover Weekly would be common examples.
- **Self-driving cars** or **automated drones** are more physical manifestations of this closed decision-loop.
- We allow the **algorithm** to do the work and present the user with the final output (sometimes with an explanation as to why the AI chose that option, other times completely opaque).

Notes

- Generally speaking, these product types are listed in terms of increasing complexity. More specifically, they are listed in terms of increasing internal complexity and (should have) less complexity on the user's side.

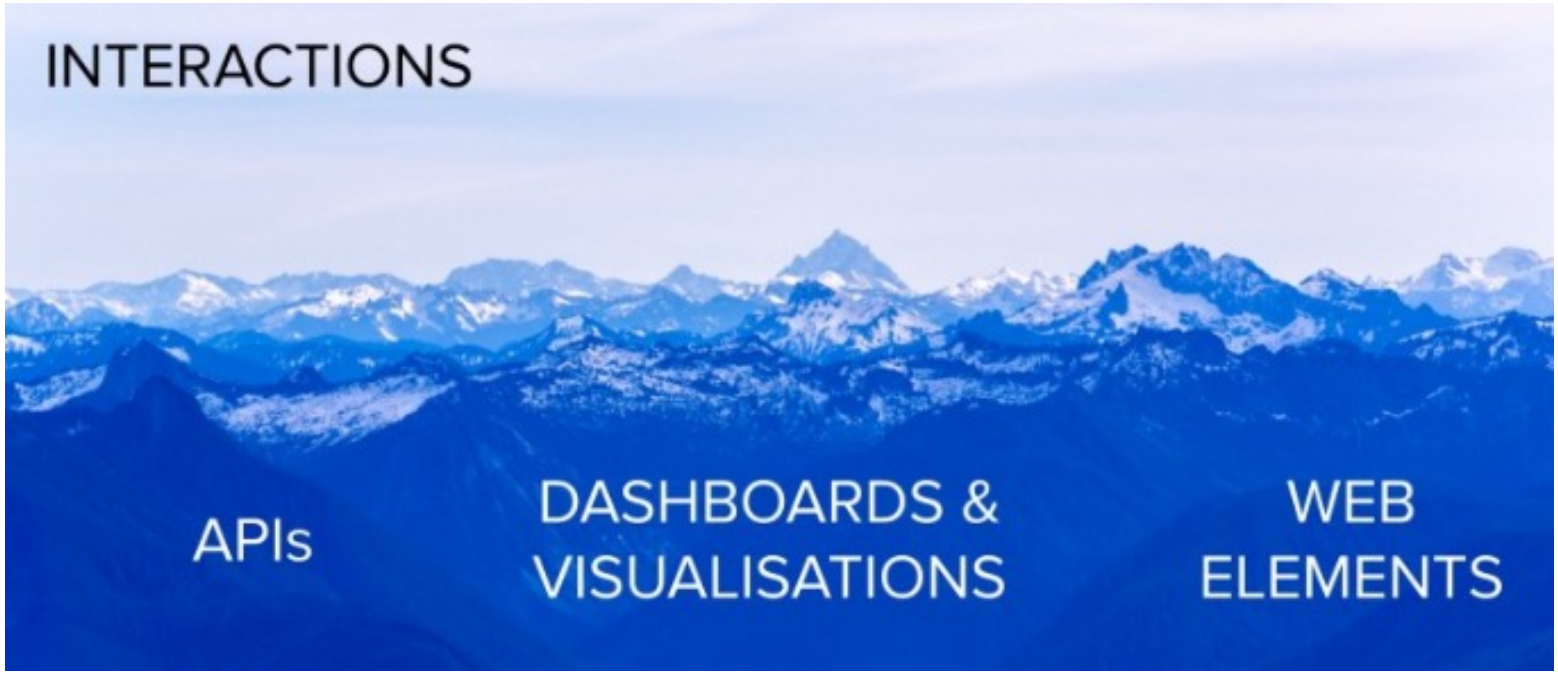
- The more computation, decision-making or "thinking" the data product does itself, the less thinking required by the user.

- Typically, raw data, derived data and algorithms have **technical users**.
- Most often they tend to be **internal products** in an organization but counter-examples would include Ad Exchanges, or API suites.
- Decision support and automated decision-making products tend to have a more **balanced mix** of technical and non-technical users.

Data Interactions

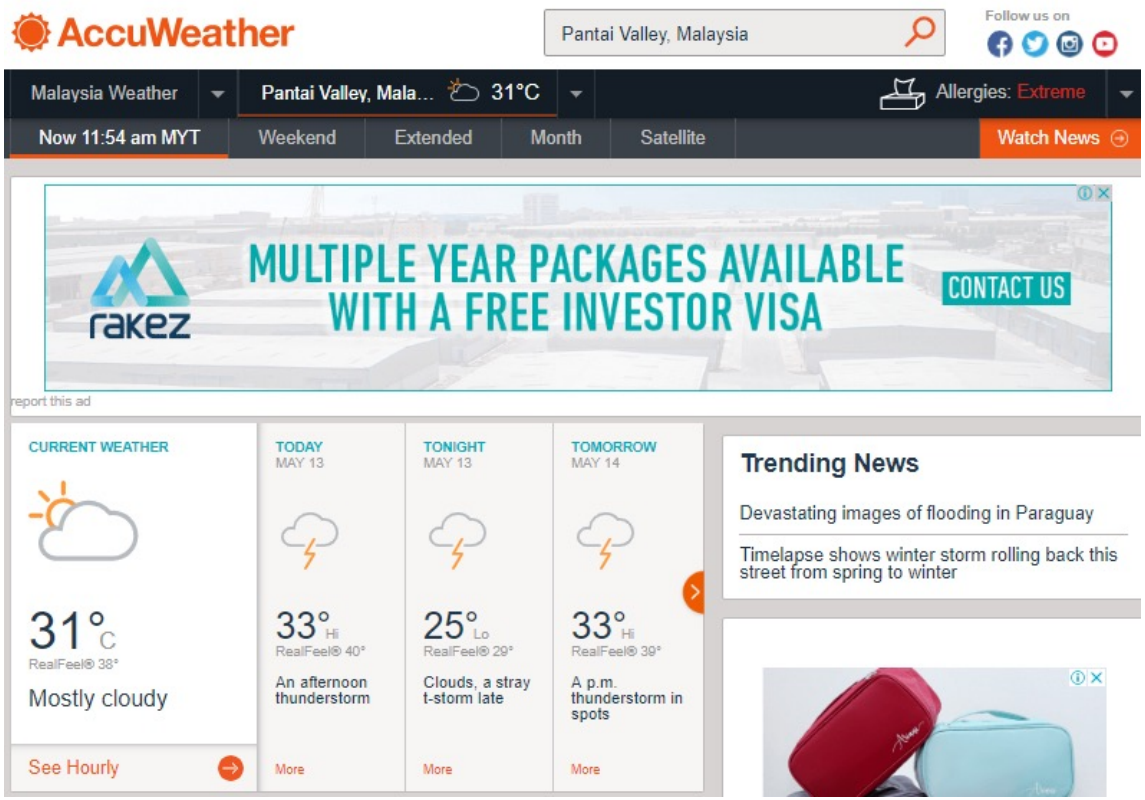
Each of these data products can be presented to our users in a variety of ways.

- What are these **interfaces** or **interactions**?



API

- **APIs** are the **de-facto standard** for building and connecting modern applications.

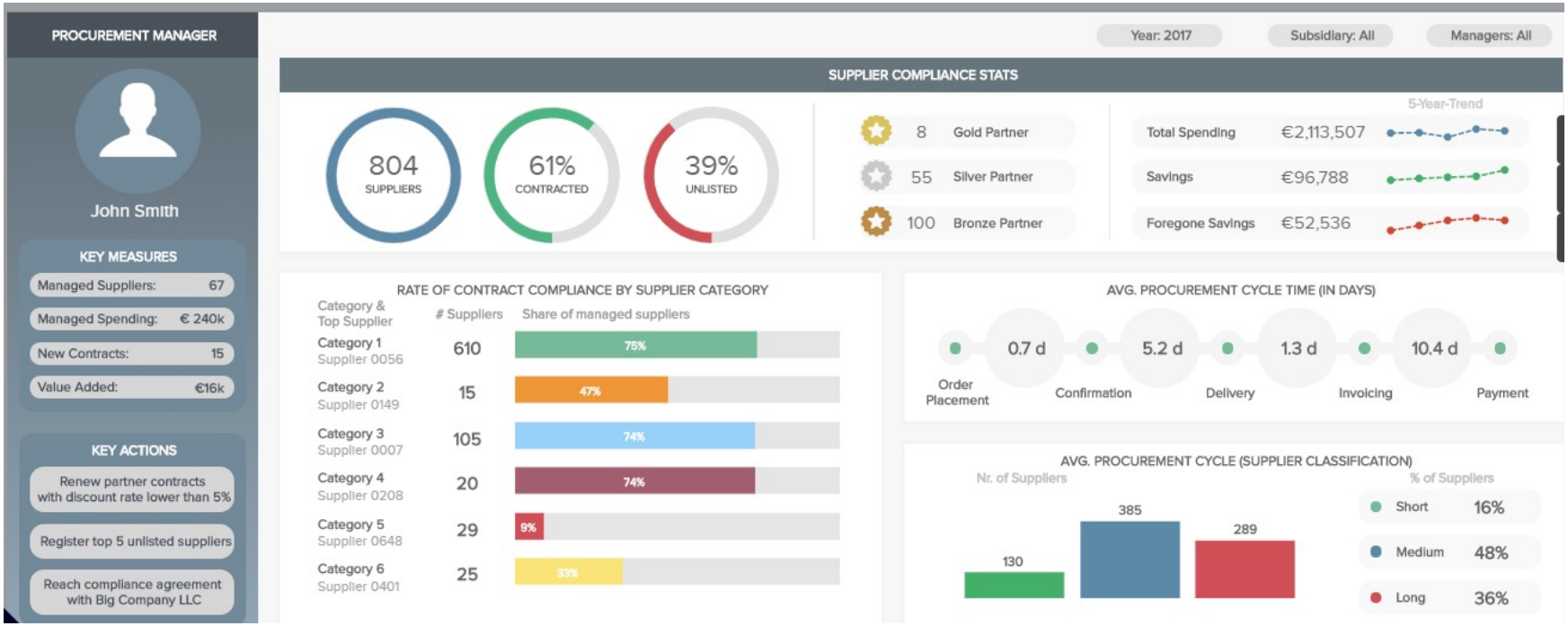


https://www.accuweather.com/en/my/pantai-valley/787644/weather-forecast/787644

AccuWeather,* via its self-service portal, offers both an API product with up-to-the-minute weather data and an API product with daily weather data.

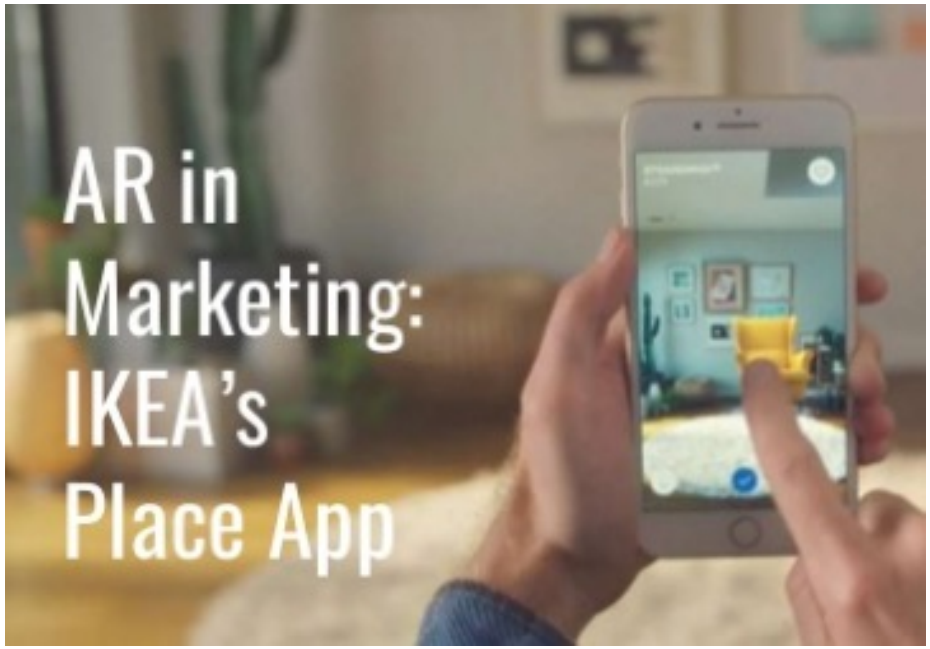
Dashboards & Visualizations

- **Dashboards** are a **data visualization** tool that allow all users to understand the analytics that matter to their business.



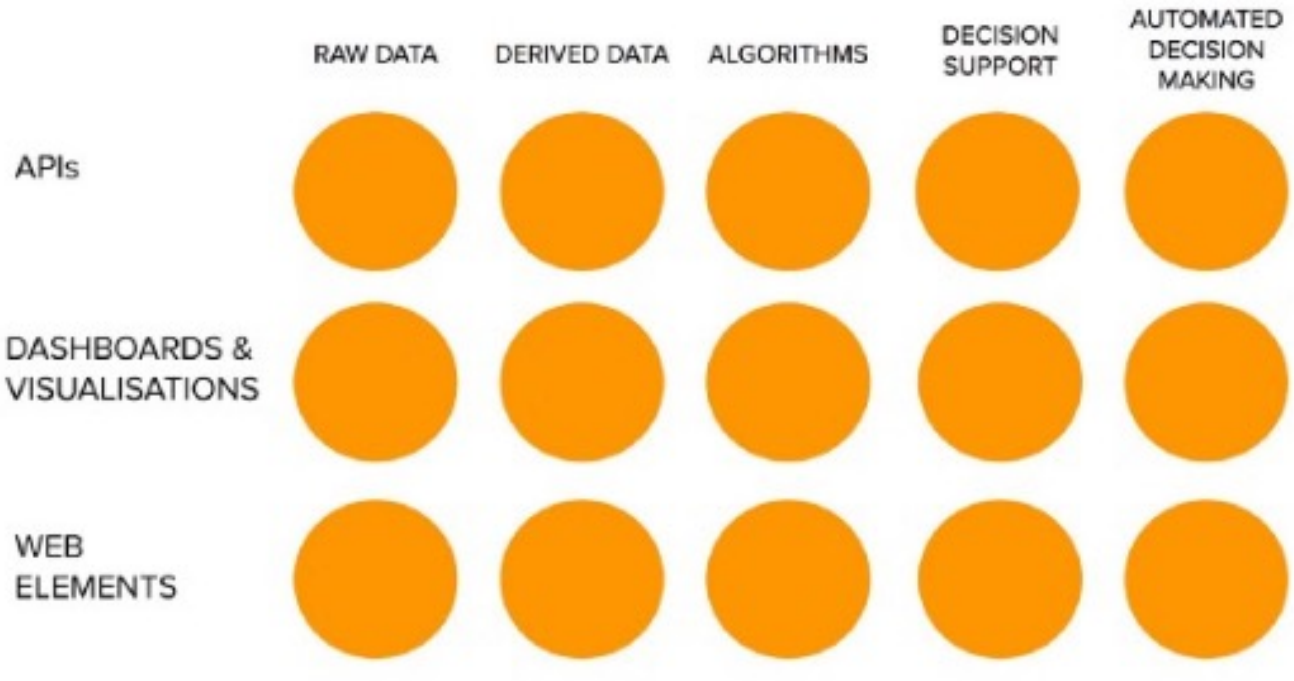
Web Elements

- More recently, these **interfaces** have been broadly extended to include **voice**, **robotics** and **augmented reality**, amongst others.



Data Product Matrix

Different products require different approaches.



Moving diagonally from the top-left circle (Raw data-API) toward the bottom-right circle (Automated decision-making-Web elements) is to move from **technical, engineering-driven products** towards those that are more typical **software products** (i.e., products that are more intuitive to product managers).

https://towardsdatascience.com/designing-data-products-b6b93edf3d23

Art of Data Science

- Data science builds models that work on large datasets, from thereon one makes predictions.
- The **art of data science** is to figure out which feature to use when:
  - If you look at datasets, it is rows of data stored in the table, every column is called a **feature** and the model that we build needs to shortlist the features.
  - Based on the features, one makes predictions and shortlisting of features is called **feature selection**.

https://www.analyticsindiamag.com/data-enabled-products-defining-future-data-science/

Feature Engineering

Feature engineering is the process of using domain knowledge of the data to create features that make machine learning algorithms work.

THREE main tasks in feature engineering:-

- Feature transformation
- Feature generation
- Feature selection

Feature Transformation

- Constructing new features from existing features; this is often achieved using mathematical mappings.
- For example, the BMI index is a feature obtained through feature transformation using a mathematic formula.

Feature Generation

- Generating new features that are often not the result of feature transformation.
- For example, one generates new usable features for images from the pixels of the images (as the pixels are not usable features).
- Many domain specific ways for defining features also belong in the feature generation category.
- Feature generation methods can be generic automatic ones, in addition to domain specific ones.
- Patterns mined from given data can also be used to generate new features. Sometimes the terms ``feature extraction" and "feature construction" are used for feature generation.

Feature Selection

- Selecting a small set of features from a very large pool of features. The reduced feature set size makes it computationally feasible to use certain machine learning and data analytic algorithms.
- **Feature selection** may also lead to improved quality on the result of those algorithms.
- Feature selection has traditionally been focused on the classification problem, but it is also needed for other data analytic problems.

Feature Selection

A **dataset** about customers

- To find out: **what product customers are most likely to buy**.
- Have to figure out - **which features are important** in making those decisions.
- Might decide that **age of customer** is the feature to be included in the analysis, **gender** is a feature to be included but by some reason the post code they are residing is not a feature to be included this process.
- This is called **feature selection** and once we have our features we build different types of models that fall into a couple of different types of categories.

Notes

- **Automatic feature engineering** is about generic approaches for automatically generating a large number of features and selecting an effective subset of the generated features in the process.
- **Feature analysis and evaluation** is about evaluating the usefulness of features and feature sets. This is sometimes included as part of **feature selection**.

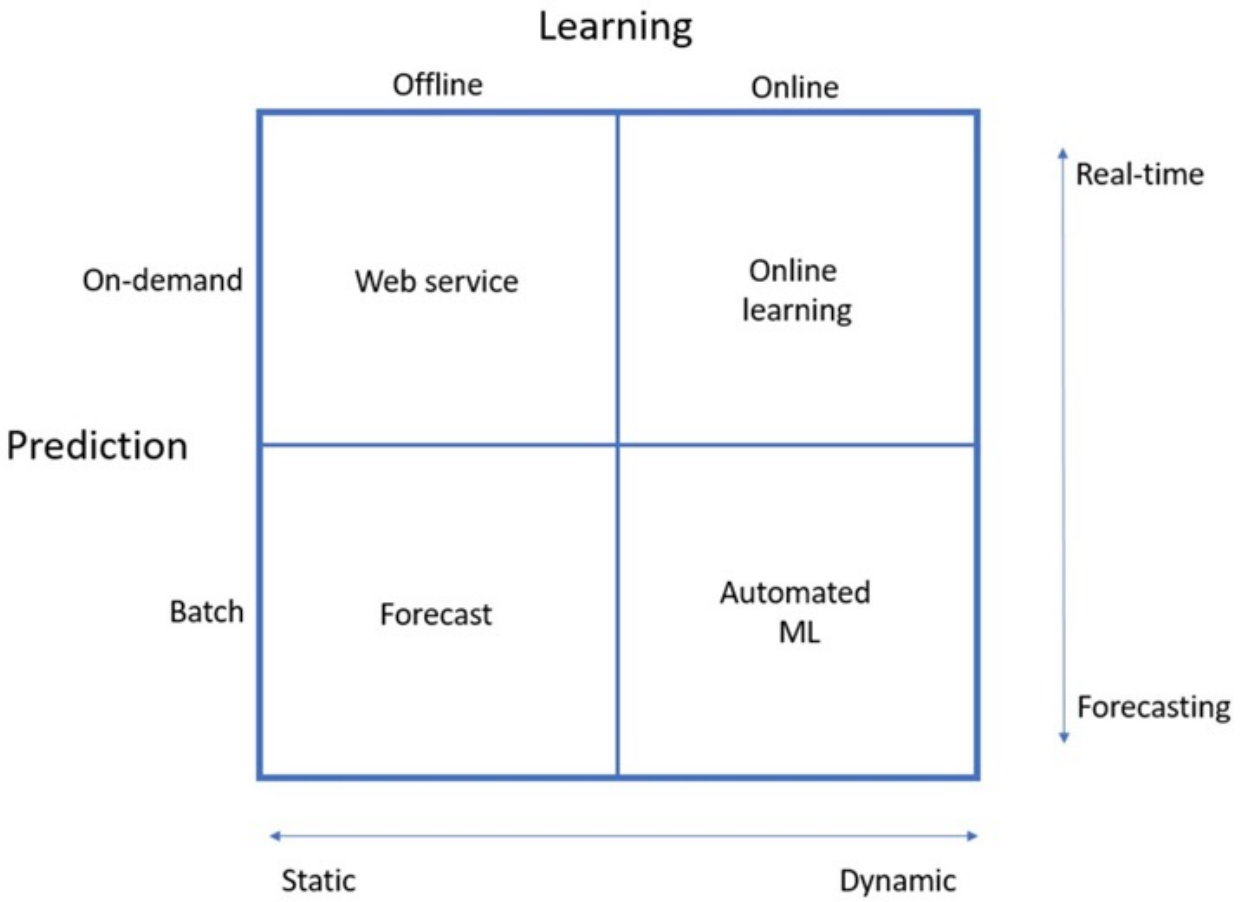
Productization

Any successful data science project must end with productization /deployment.

- This is the stage where trained models are deployed as application that can be easily accessed by the end users.

Productionalizing Machine Learning Models

FOUR different ways of productionalizing machine learning models.



Batch Prediction

The **simplest** form of machine learning workflow is the batch prediction.

- This is typically seen in academia and places like Kaggle.
- You take a static dataset, run your model on it, and output a forecast. How do you productionalize something like that?
- On Kaggle, you save your predictions to a CSV file that you submit through an online form.

Web Service

- The most common type of machine learning workflow is a **simple web service**.
- The web service takes in some parameters and gives out a prediction straight away.
- This is way **more agile** than the batch prediction scheme.
- The difference from batch predictions, apart from running in **near real-time**, is that it handles a single record at a time, instead of processing all the data at once.

Online Learning

- Emerging now is real-time streaming analytics, also known as hot path analytics.
- This works very well with the lambda architecture that's so popular in big data systems.
- The **input data** in this case would be a stream of events, and the model would be placed right in the firehose, so to speak, running the model on the data as it enters the system.
- The model would typically be running as a service on a Spark cluster or something similar. This is very useful for sensor data.

Automated ML

- Tendency to think of machine learning models as something you train, deploy and forget, but that's often not good enough.
- Online learning means that your model learns, improves and updates itself while in production.
- This obviously requires some engineering, but the payoff is a dynamic model.
- An even more sophisticated version of this is **automated machine learning**. Instead of updating the model, you can run an entire machine learning pipeline online in production that comes up with entirely new models on the go.

Datapreneurs

The entrepreneurs focused on data science and related topics like Business Intelligence, Business Analytics, Predictive Modeling, Machine learning etc.

The datapreneurs are classified in 4 areas:

i. Data Products,
ii. Data Science Services,
iii. Data Science Training, and
iv. Data Science Communities

Examples for the 4 areas:

i. Jim Goodnight & John Sall (SAS), Christian Chabot (Tableau)
ii. Gurjeet Singh (Ayasdi), Carlos Guestrin (Apple)
iii. Andrew Ng (Coursera), Sebastian Thrun (Udacity)
iv. Anthony Goldbloom (Kaggle), Gregory Piatetsky-Shapiro (Kdnuggets)

https://www.kdnuggets.com/2015/09/top-datapreneurs-data-science-analyticsvidhya.html

Try the Demo & Watch the Video

- https://demos.datasciencedojo.com/

Beyond Analytics — Building Data Products for Data Natives

https://databricks.com/session/beyond-analytics-building-data-products-for-data-natives (21 minutes)

So You Want to Build a Data Product? - http://on.wsj.com/1qbvL87

https://www.oreilly.com/library/view/data-analytics-with/9781491913734/ch01.html