

MY JOURNEY OF AS A DATA SCIENTIST:

Using AI, Machine Learning and
Big Data to Solve Problems in
the Retail Industry

Amir Tavasoli



Data Scientist at Home Depot Canada



Agenda

- I. Data Analytics in modern world
- II. Start of Data Science as a profession
- III. Where Data Science stands today
- IV. Data Science in Retail Industry:
 1. Inventory Management
 2. Supply Chain Management
 3. Shelf Optimization
 4. Price Optimization
 5. Recommendation Engines
- V. Experimentation
- VI. How to to become a Data Scientist

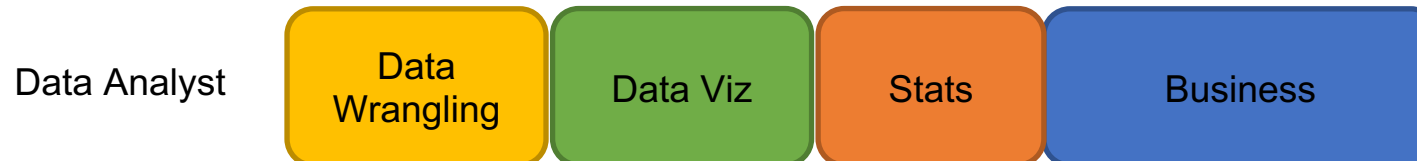
I. Data Analytics in Modern World

Data Analytics: How it all got started

- Data Analytics and using data in decision making have been focus of many organizations for years.
- At first, data analysts needed immense business knowledge and were using tools like excel or access to create simple models to help decision making.

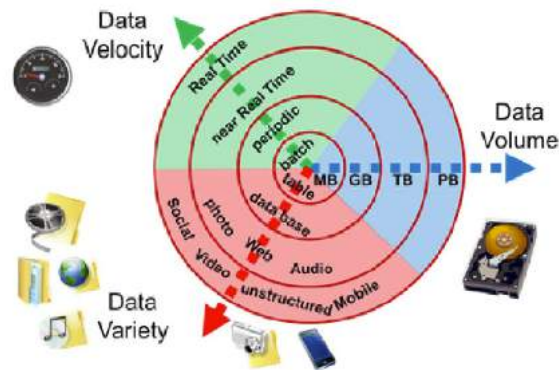


- As Data Analytics got momentum, business started asking more advanced questions like:
 - What is the effect of using this model in terms of dollars?
 - Can we test this?
 - Just tell me what is top 3 choices?
- This added **statistics and experimental design** as another important dimension to to data analyst job. Since these concepts were hard to explain, this led to addition of **data visualization** as a requirement for being a data analyst. They would use these to communicate key insights to the larger audience with easily interpretable data insights.

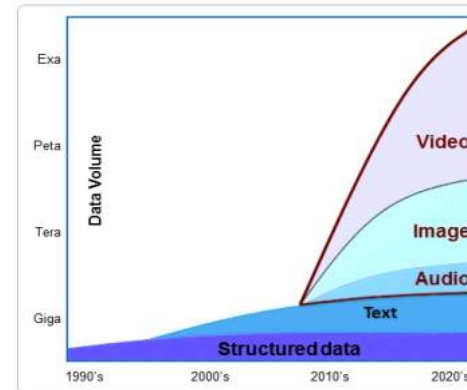


Explosion of Data

- Recently, all dimension of data (Volume, Velocity, and Variety) started to grow at an extremely fast pace.
- In 2020, every person will generate 1.7 megabytes in just a second!



Adopted from: https://en.wikipedia.org/wiki/Big_data



Adopted from: <https://digitalhumanities101.wordpress.com/2014/10/21/importance-of-image-data-and-image-processing-techniques-for-digital-humanities/>

- This created massive technological challenges on how to store, process, and visualize data:

Classic Technology	Age of Big Data
Store and Process data in Excel or MySQL	Using cloud-based SQL services like Google Big Query or In-House Hadoop Servers
Visualize data in Excel and Power Point	Using cloud-based Visualization services like Looker and Tableau
N/A	Use cloud-based NoSQL services to store and process images, audio, videos, etc.



Data Analytics in Age of Big Data

- The growth and change of data with this extreme pace added new challenges to data analytics in terms of storage and extraction. All datasets moved from local databases to cloud environments.
- Classic or modern data analysts at the time were not equipped with the means to handle this level of growth and change in data. This created a need for new roles to work with data analysts:
- **Data Engineering:** They extract, transform, load (ETL), and clean the data from diverse sources into one or more consolidated databases. They also build data pipelines, perform data validation, etc.
- **Technologies used by Data Engineers:** SQL, Big Data (Hadoop, Hive, Pig, Spark, etc.), Cloud Ecosystem for Big Data (AWS EMR, Redshift, Google's BigQuery, Dataproc, Dataflow, or Azure related services)

Data Engineer

Data Engineering

Business



Data Analytics in Age of Big Data (cont.)

- The role of data analytics evolved accordingly as well.
- Aside from business knowledge, now they needed to:
 - Work with Data Engineers to find, clean and source required data for enabling data-driven decision making
 - Use tools that can analyze large datasets to aid stakeholders with meaningful insights and help data-driven decision making as a result
 - Develop reproducible business reports leveraging a variety of visualization tools that are running in Cloud-based backends.
- They needed to learn new skills like Big Query SQL, Altryx, KNIME, Data Visualization (e.g., Tableau, Looker, SpotFire, etc.)

Modern Data Analytics
in Age of Big Data

Data Engineering

Data Visualization

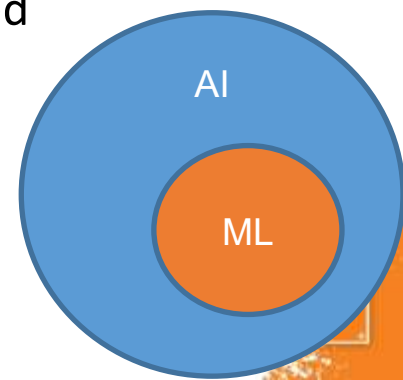
Business



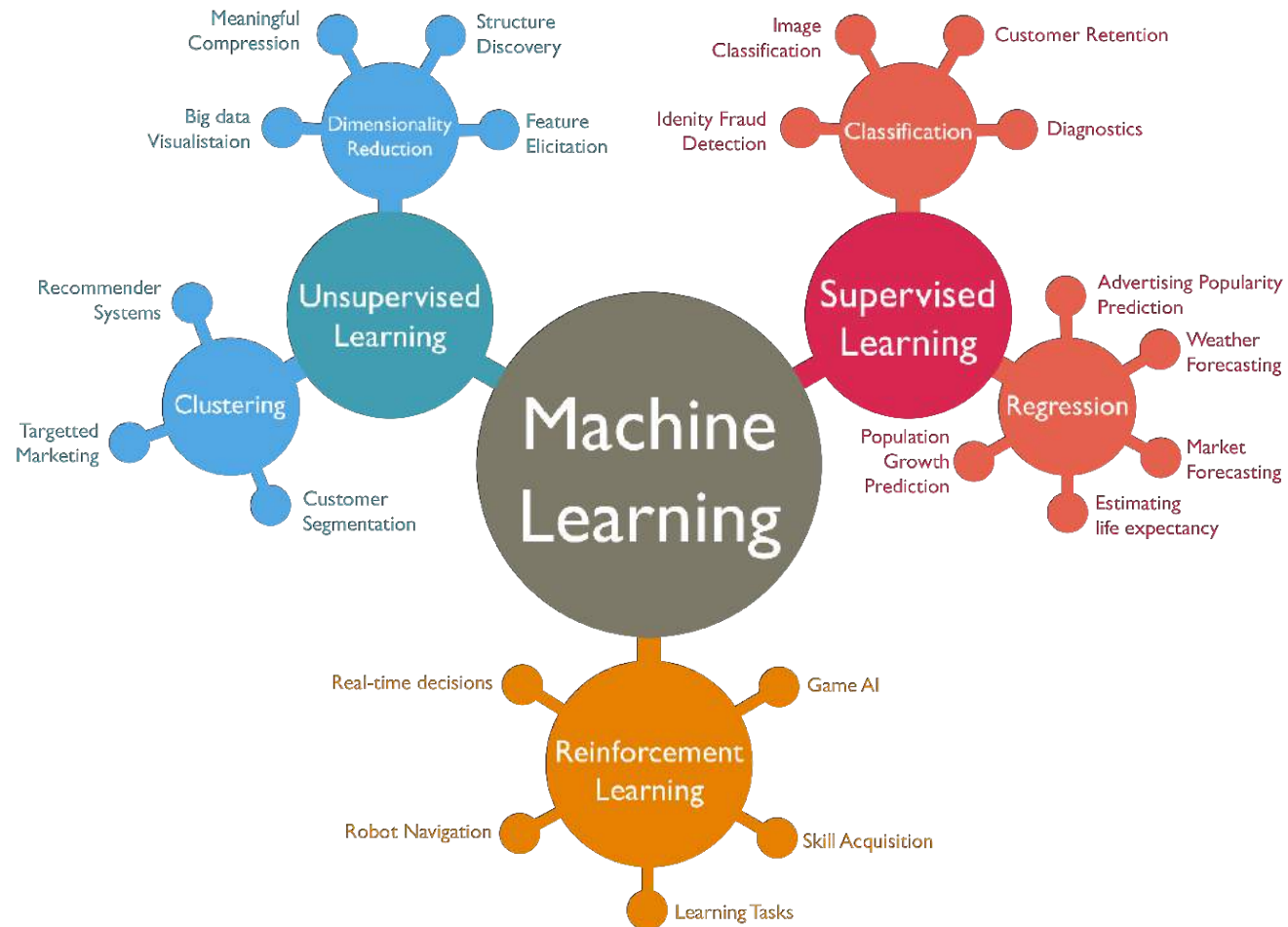
II. Start of Data Science as a profession

What is Machine Learning?

- With explosion of data, there was a need for more advanced algorithms than simple statistical models. These algorithms needed to be able take advantage of big data.
- This opened the door for Artificial Intelligence, Machine Learning, and Deep Learning to enter the field of data analytics.
- **Artificial Intelligence (AI)** is refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. The official definition of AI is that it refers to “the study of *“intelligent agents”*: any device that perceives its environment and takes actions that maximize its chance of successfully achieving its goal” (Wikipedia)
- For example, knowledge representation, natural language processing, and learning.
- **Machine Learning (ML)** is a subfield of AI that “deals with the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead.” (Wikipedia)
- Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.
- For example, image recognition, email filtering, and fraud detection.

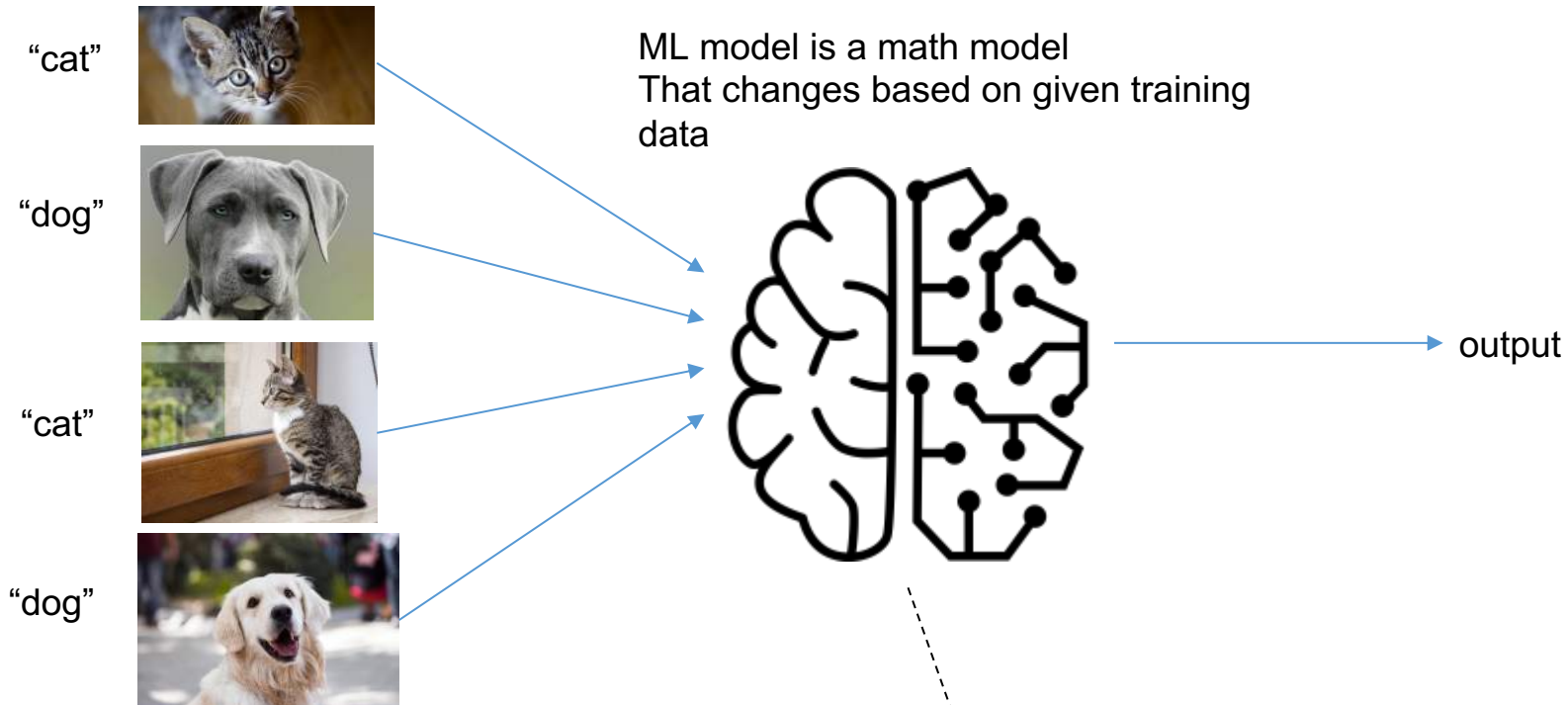


What is Machine Learning? (cont.)

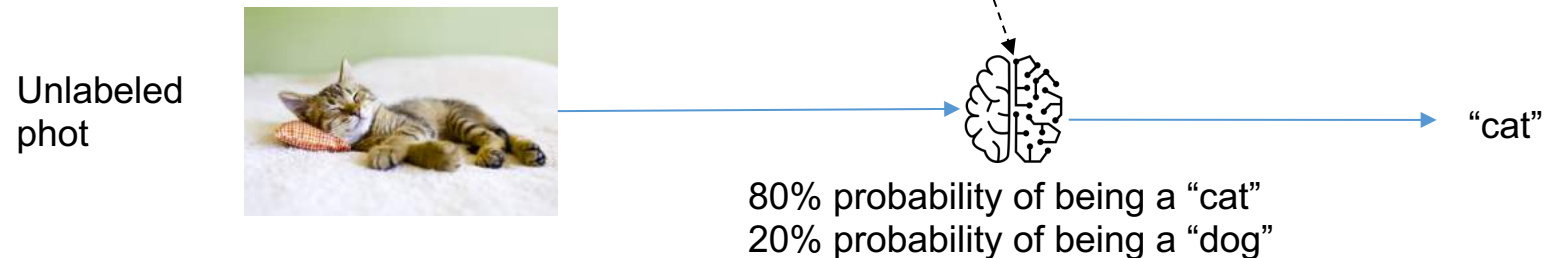


Machine Learning: Supervised learning example

- Stage 1: Train ML model with example

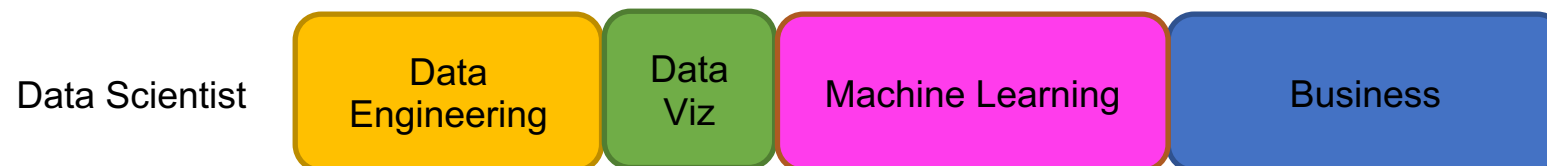


- Stage 2: The trained model is used for prediction



What is Data Science?

- Data Analysts job was already complex enough in this new age and there was a need for a new role that can handle:
 - **Machine Learning and AI:** They needed to be able to program and implement these advanced algorithms in new environments like cloud or Hadoop.
 - **Unstructured data:** They needed to be able to interpret and automate tasks that were related to textual, image, audio, and video data.
 - **Advanced Statistics:** They needed build advanced statistical and optimization models, e.g., price, shelf, or inventory optimization.
 - **Advanced Experimentations:** They needed to build advanced online experimentations that can capture effect of a change in a diverse environment that has thousands of user interaction per minute.
- Data Science today means art of building data driven solutions using advanced algorithms in the age of big data.
- **Technologies** used by data scientists: Data Engineering Tools (Cloud-based SQL, Hadoop, Spark), Programming (Python, Java), ML Platforms (TensorFlow, PyTorch, Keras), Visualization Tools (Tableau, Looker), Relevant Business Tools (SAP).

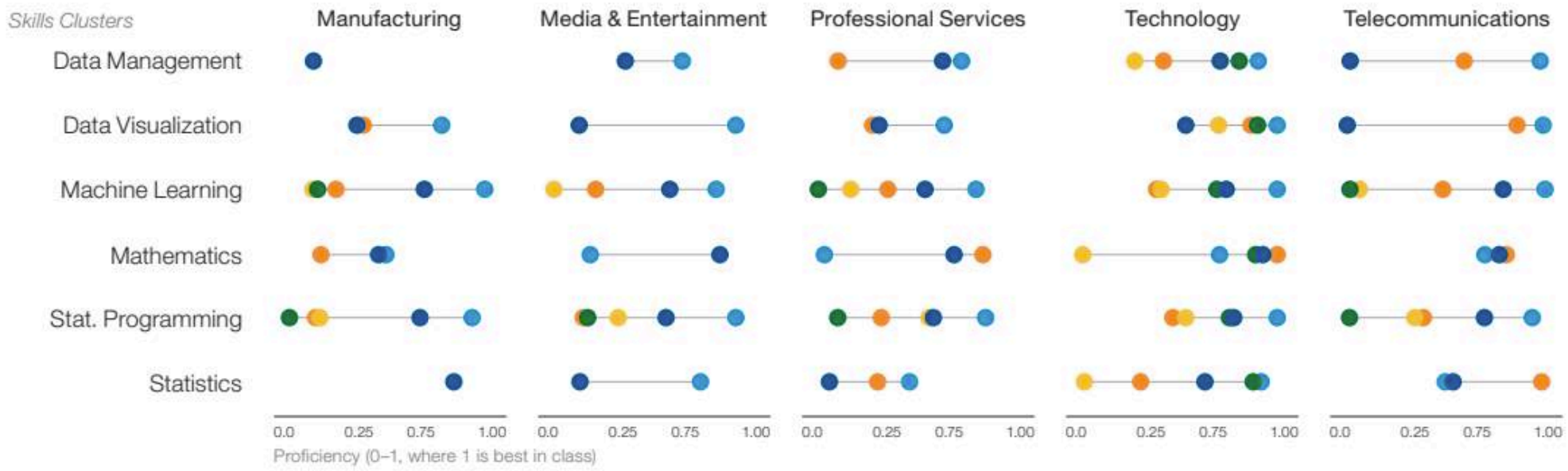
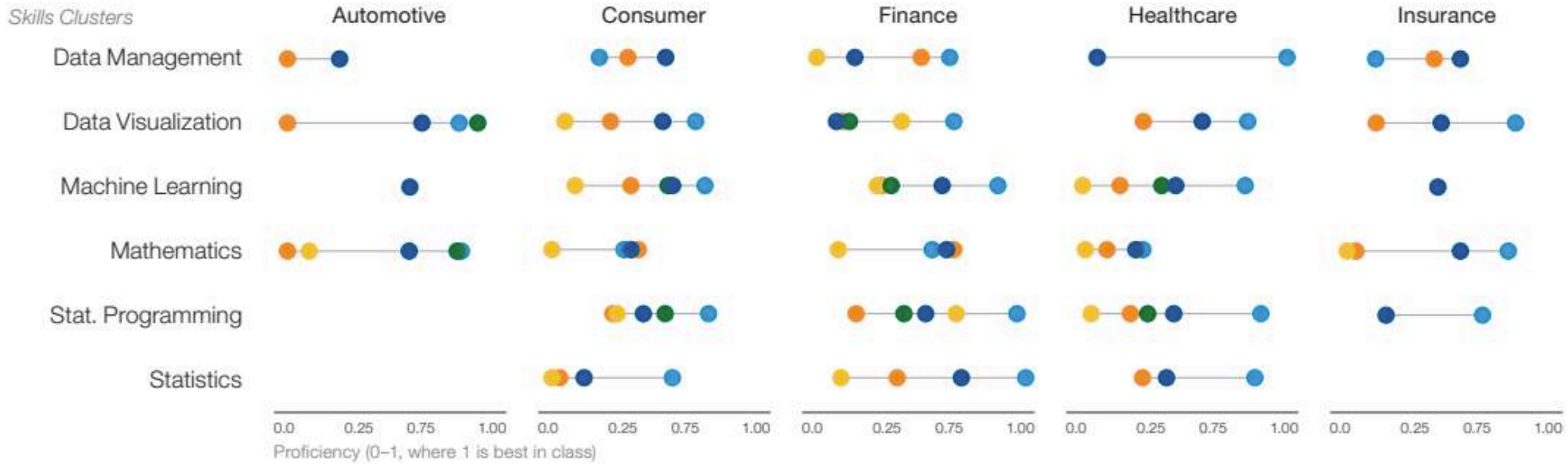


III. Where Data Science Stands Today

Why Data Science Matters?

- According to World Economic Forum report in 2018, Data Science and AI is the “the fourth industrial revolution” and is changing the labour market in a fundamental fashion.
- 97.2% of organizations are investing in big data and AI i.e. Data Science.
- 90% of data generated within any organization is not structured and you need data scientist to be able to turn that data into decision making material.
- While data science roles and skills form a relatively small part of the workforce, recent trends indicate that these are currently among the highest in demand roles in the labour market.
- The data science skillset is not fixed and is rapidly evolving as new opportunities in data analysis and further technological advances redefine the specific skills composition of data scientist roles.
- Recommended Read from World Economic Forum:
http://www3.weforum.org/docs/WEF_Data_Science_In_the_New_Economy.pdf

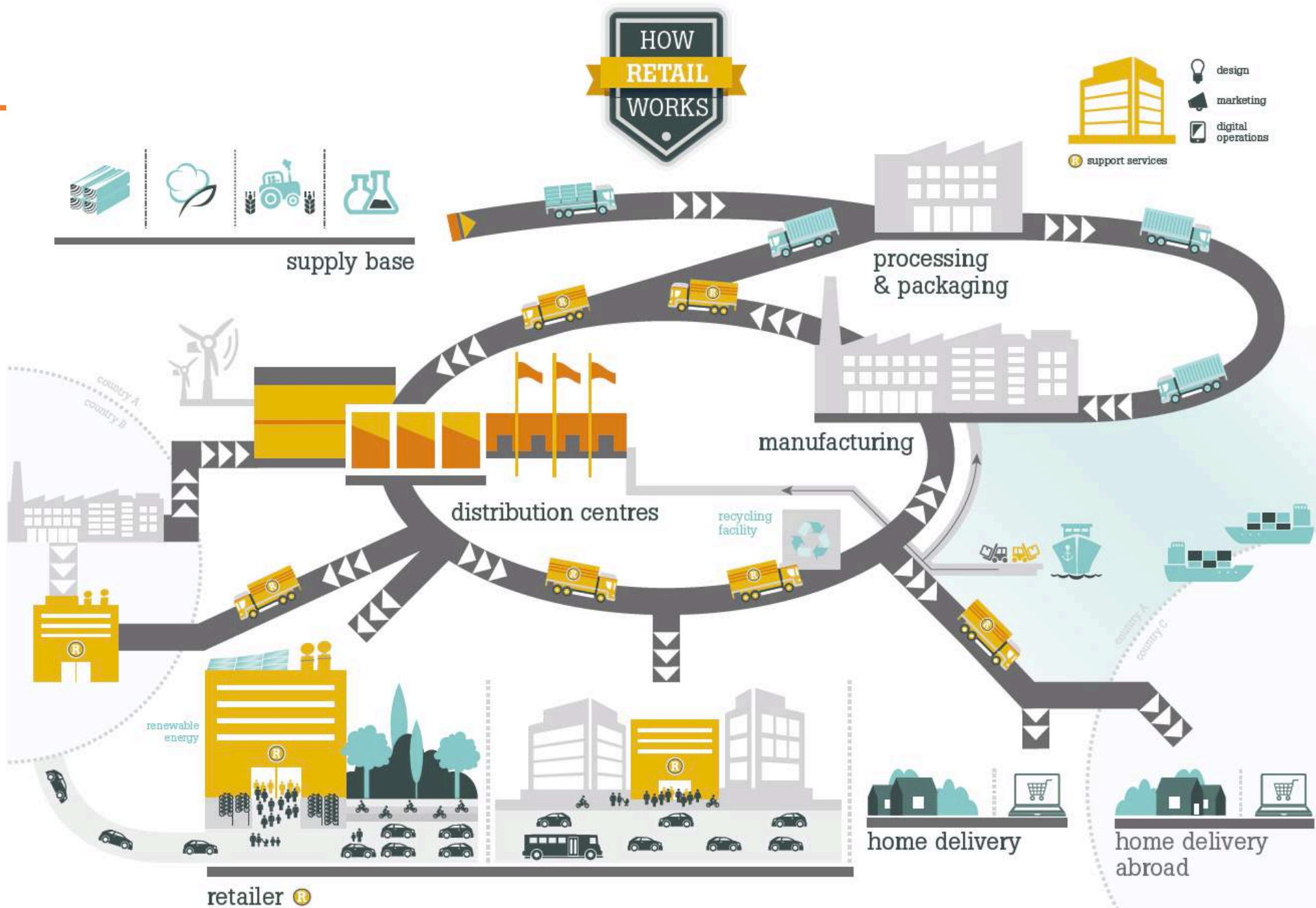
Data science skills proficiency, by industry and region



Asia Pacific Europe Latin America Middle East and Africa North America

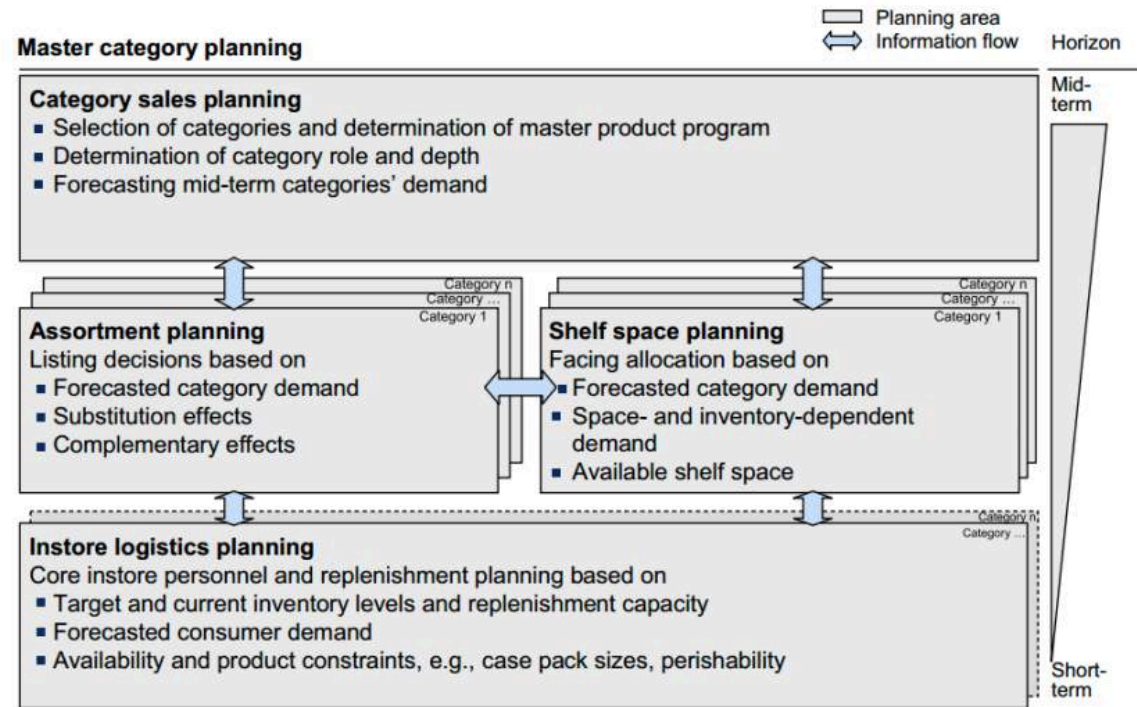


VI. Data Science in Retail Industry



Data Science Problem Space in Retail

- Data Science can be applied to almost every step of retailing that mentioned. The problem is space is inter-related to each other. For example:



Aguiar, M. (2015). The Retail Shelf Space Allocation Problem: New Optimization Methods Applied to a Supermarket Chain (Doctoral dissertation). Retrieved from <https://pdfs.semanticscholar.org/b6f1/9dcf4994dd0b1e8d4481cf27bc39c5491938.pdf>

- This problem space is further expanded by online retailing issues like recommender systems and online marketing.

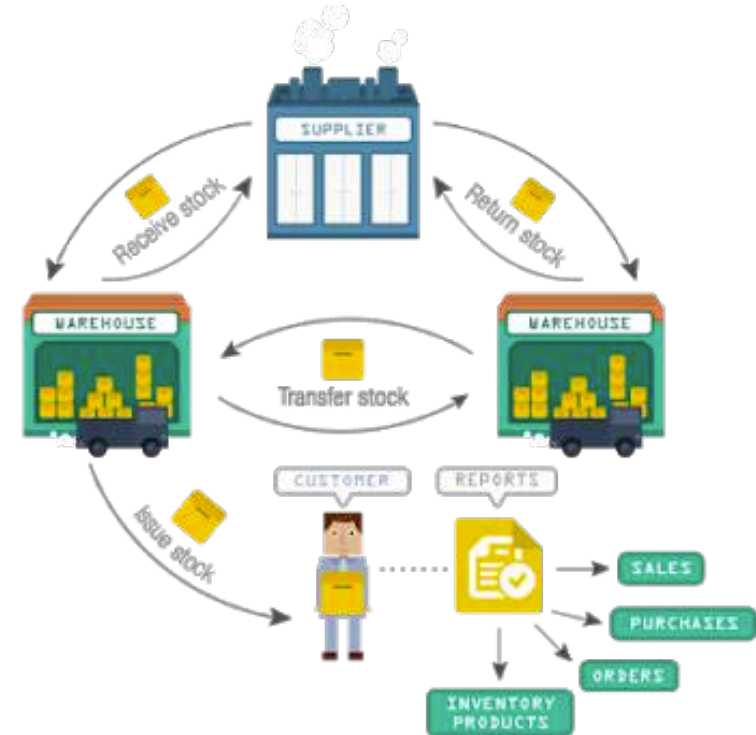
Data Science in Industry vs Academic Research

- Data science in industry and academic research both follow scientific process:
 1. Ask a question
 2. Build a Hypothesis
 3. Test Hypothesis with an experiment
 4. Analyse data and draw conclusion
 5. Report your results
- The first difference is that industry requires deep understand the business processes that are mostly not well documented. This understanding derives your hypothesis and experimentation and requires tons of human to human interactions and relationship building.
- Second difference is that every process that you touch in the industry affects hundreds of other processes within the industry while in academia you can separate the processes easier. This makes finding the impact of your models in industry very challenging.
- Moreover, in the industry you need to work with limitations of technology provided by IT departments.
- Furthermore, reporting of your results are mostly done through dashboarding which requires working with other data visualization experts and data analysts in your team.
- Finally, in the industry constant optimization of your process is the key to success. After the process is done and pipelines are in place, you have only done the first step of the process.

1. Inventory Management

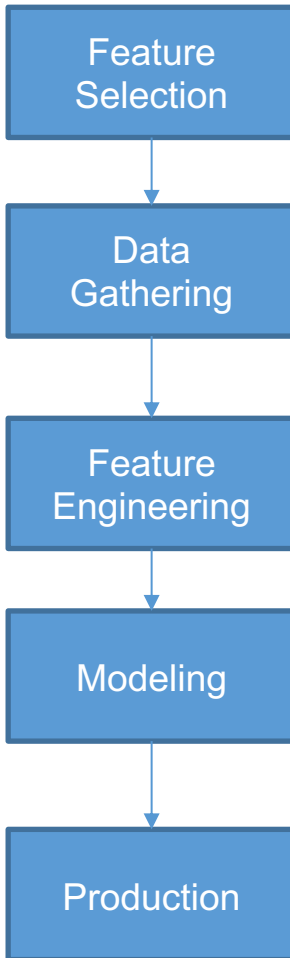
Data Science in Inventory Management

- In classic retailers' demand for a warehouse is determined using simple statistical models that have been modified based on retailers plans.
- ML models can be used to forecast demand for a certain warehouse with much higher accuracy.
- This forecast can be used to:
 - Optimize Inventory
 - Reduce Stockouts
 - Optimize the Warehouse Workforce



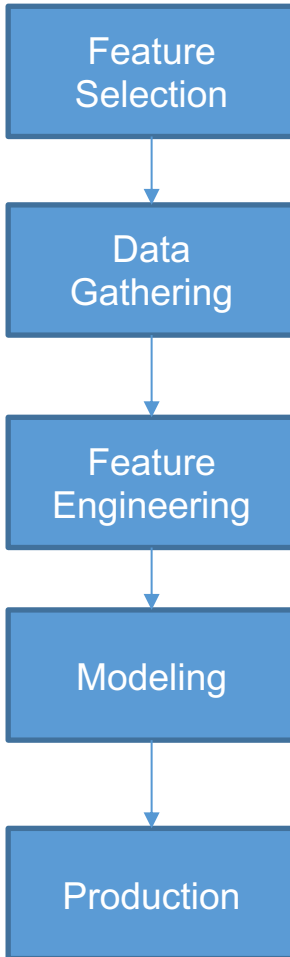
Data Science in Inventory Management (cont).

- Real world process of building such machine learning forecasting model:
- First step is to have many discussion with business. In these discussions, you can find that the following variables matter. This process is known as feature selection:
 - Weather: temperature, rain or snow, ...
 - Age of a product (New vs Old)
 - Type of product: hammer vs a dishwasher
 - Location of customers
 - Competition
 - ...
- Next step is to find relevant data for these variables. These datasets are mostly very “dirty” and require collaboration with other business patterns and data engineers.
- After that, these features need to be engineered into values that can be used by machine learning model. This step is known as feature engineering.



Data Science in Inventory Management (cont).

- This follows by choosing relevant ML models that can capture this data. In this case one can use models from simple linear regression to deep neural networks. These models capture the patterns that determine the relationship between these features and warehouse demand target.
- One common algorithm used for modeling these cases are decision trees: [demo](#)
- Then, historical tests will be done to find the model with the best performance. If the model with the better performance than current business process cannot be found, this process will be iterated on.
- After the model is built, this process need to be productionized so that it runs on regular basis and the output is fed into relevant databases.
- Now this data can be used by business in order make better decisions, in order to for example avoid stockouts within a warehouse.



2. Supply Chain Management

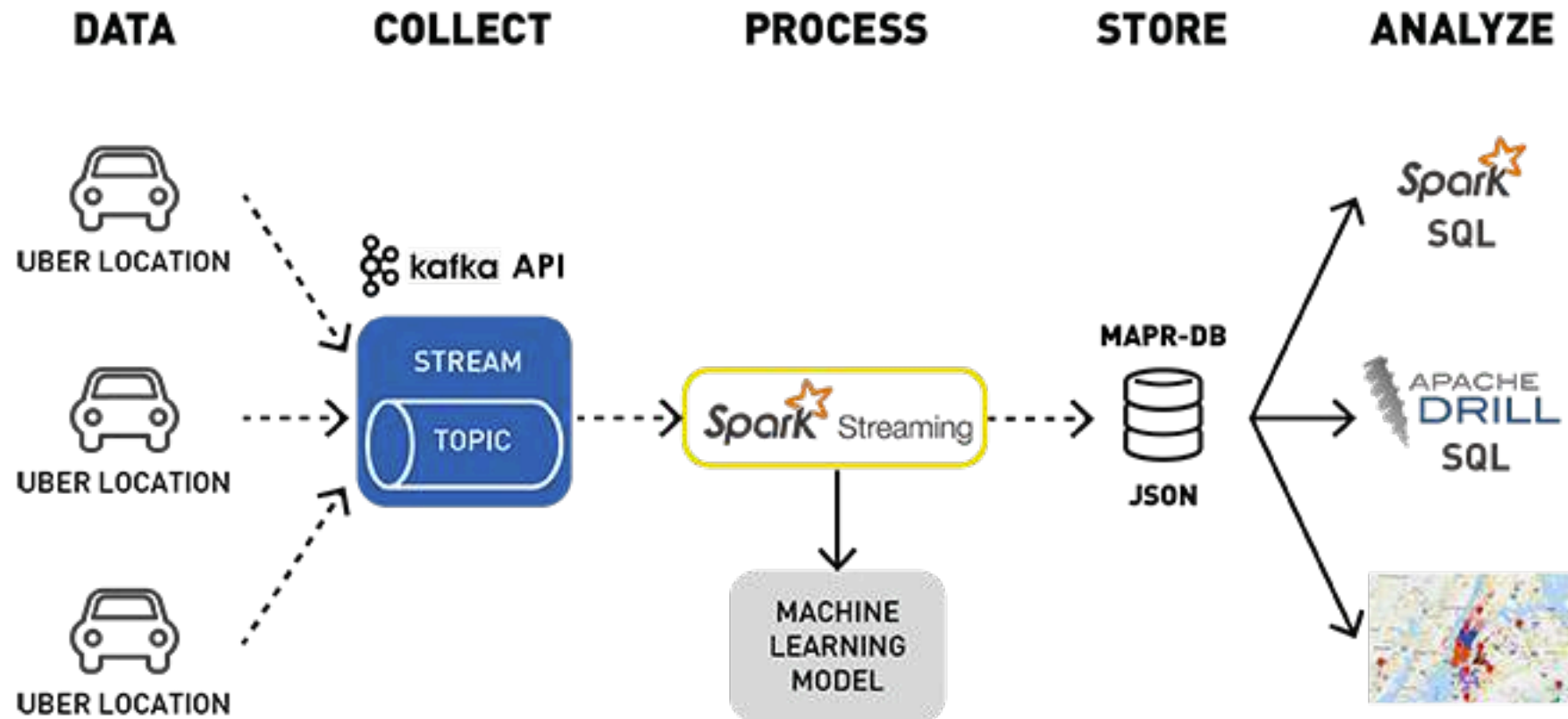
Supply Chain Optimization



- One major problem in supply chain is **how long does it take for trucks to arrive at a warehouse**. In classic retailing these times are pre-determined using excel and are kept the same for months before there was a requirement to change them.
- ML can be used to find how long service would require for delivery of a volume at the given docking center in a warehouse.
- Drivers send in reports and sensor data from trucks tracks the routes and account hours logged. Distribution centers send in their capacity data, the number of trucks at the dock, and time for loading and unloading. The ML model uses these features and can then accurately determine how long service would require for delivery of a volume at the given docking center.
- This challenge is like inventory optimization in a sense that data scientist need to go through the process of feature selection, data gathering, feature engineering, modeling, and production.
- The main difference is that datasets related to this problem are "live" datasets. The models should be able to ingest the data and make a decision in a matter of seconds.
- This data can be used to ensure the docking is ready before the arrival of truck.

Real-Time Machine Learning

- In supply chain management of modern retailing using streaming data is getting more and more common. A good example of such datasets is Uber:



3. Shelf Optimization

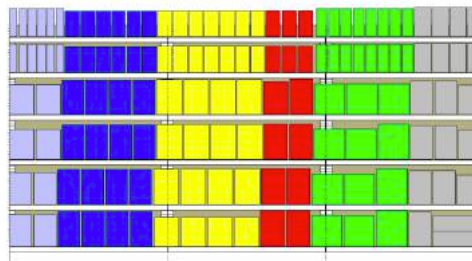
Shelf Optimization

- What goes on shelf and how many and in which order is done manually in classic retailing. One person fills up a shelf using a software in and makes it look “as good as possible” based on their business knowledge.

- They decide to put more expensive stuff first
- They decide to put same brands beside each other
- ...



- Then, they put these as what is known as planograms and send them to store to implement.



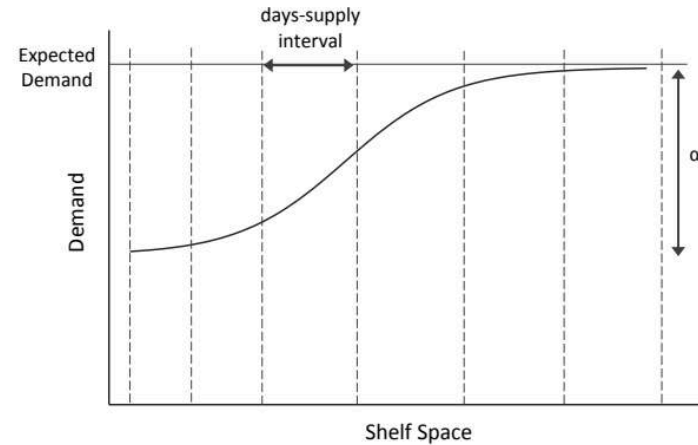
- Store try to make their shelves as close to possible to a planogram.

Shelf Optimization: A Decision Support Approach

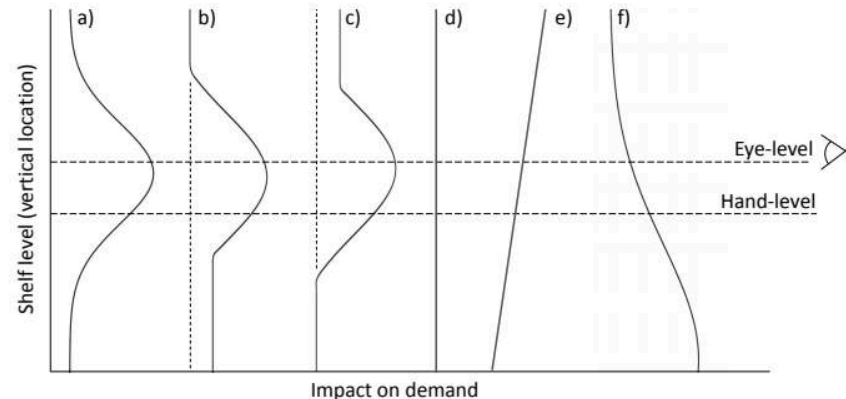
- This problem is different from other discussed problem in a sense that it is almost impossible to find reliable data on how the stores have implemented their planograms and what is the performance of these planograms are.
- We need a different approach to modeling than other problems that mentioned. The model needs to be built on what is best practice and the generated data is sent to stores based on these models to support their decision making while making a shelf.
- This is a very soft approach to help the decision in lack of more available data.
- This approach can be enhanced dramatically if technologies like image recognition is advanced to a level that we can have 100% view what is available on shelf and how customers are interacting with it.

Shelf Optimization: Space Elasticity

- This a multi-dimension optimization problem. One dimension is that shelf space that is given to a certain product

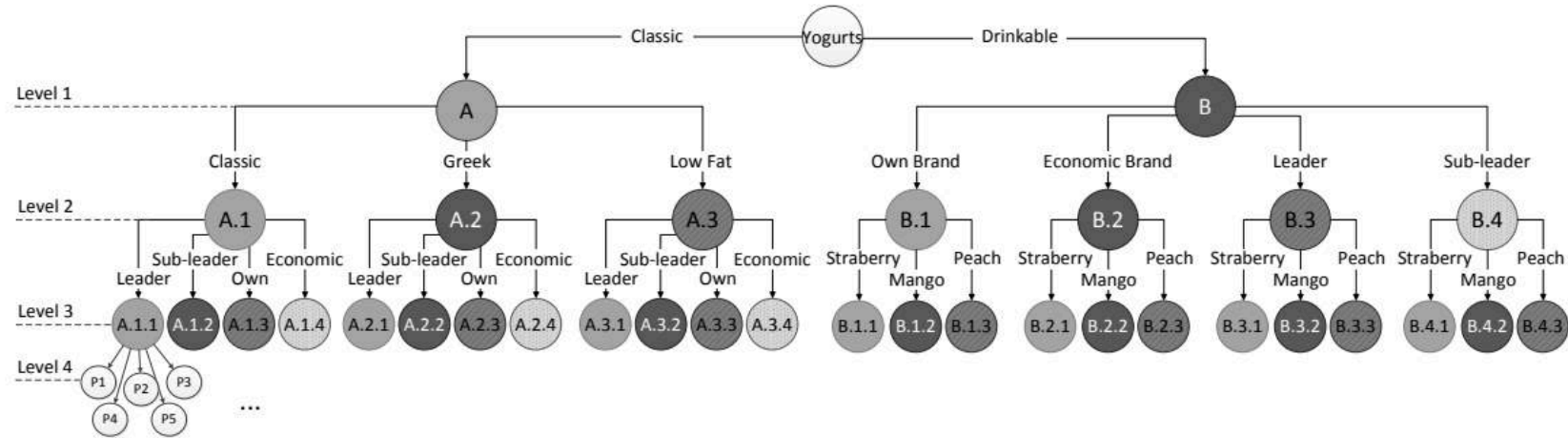


- The other dimension is the placement of product on shelf.

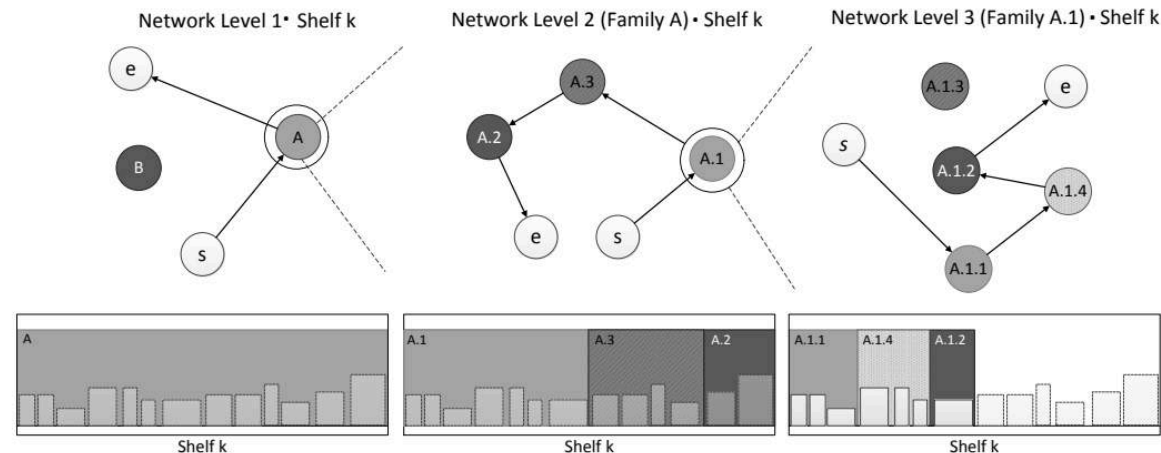


Shelf Optimization: Clustering

- First all products can be clustered into their known “families”:

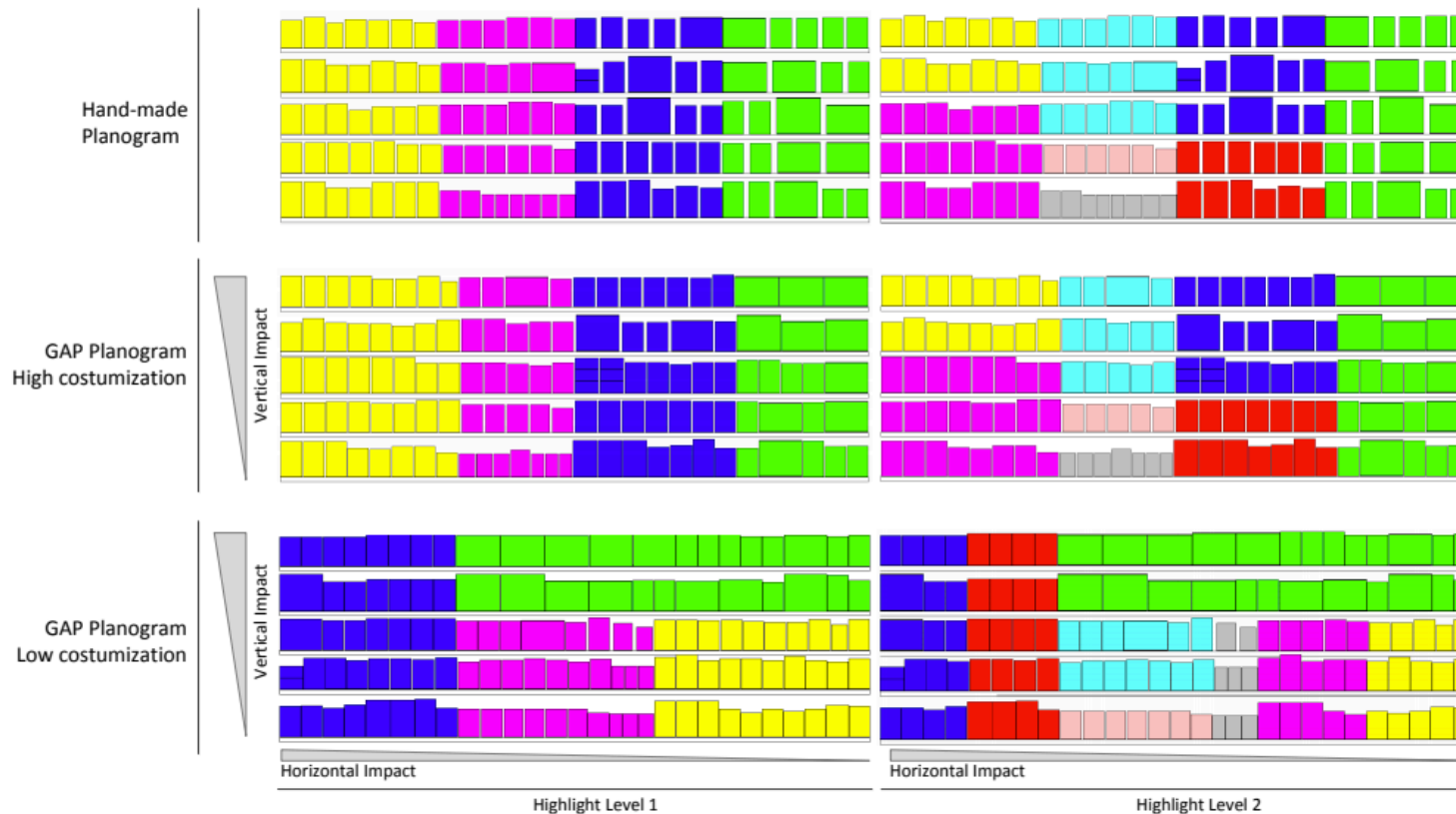


- Then you can optimize their placement using graph theory:



Shelf Optimization: Decision Support System

- After the optimized shelf is generated, it can be used a decision support for stores to put products on shelf.



4. Price Optimization

Price Optimization

- In classic retailing price is set manually using excel looking at historical data. This process gets repeated a few times a year for the highest impact products.
- Modern retailers use ML and AI to determine the prices, as a result, is very hard to compete with how prices are changed from modern retailers like amazon:

Its



Instant Pot Duo 80 7-in-1 Electric Pressure Cooker, Steamer, Saute, Yogurt Maker, and Warmer, 8-Qt, Stainless Steel/Black

by Instant Pot
★★★★☆ 52,309 ratings | 450 answered questions
Amazon's Choice for "instant pot 8 qt"

List Price: ~~CDN\$ 159.95~~
Price: **CDN\$ 125.82** ✓prime FREE One-Day
You Save: **CDN\$ 34.13 (21%)**

You could get 5% back at Amazon.ca, Whole Foods Market stores, grocery stores, and restaurants for 6 months upon approval for the Amazon.ca Rewards Mastercard. See terms and learn more.

New & Used (30) from **CDN\$ 87.06** ✓prime FREE Shipping

Size: **8Quart**

6Quart 8-Qt **8Quart**

Style: **Instant Pot**

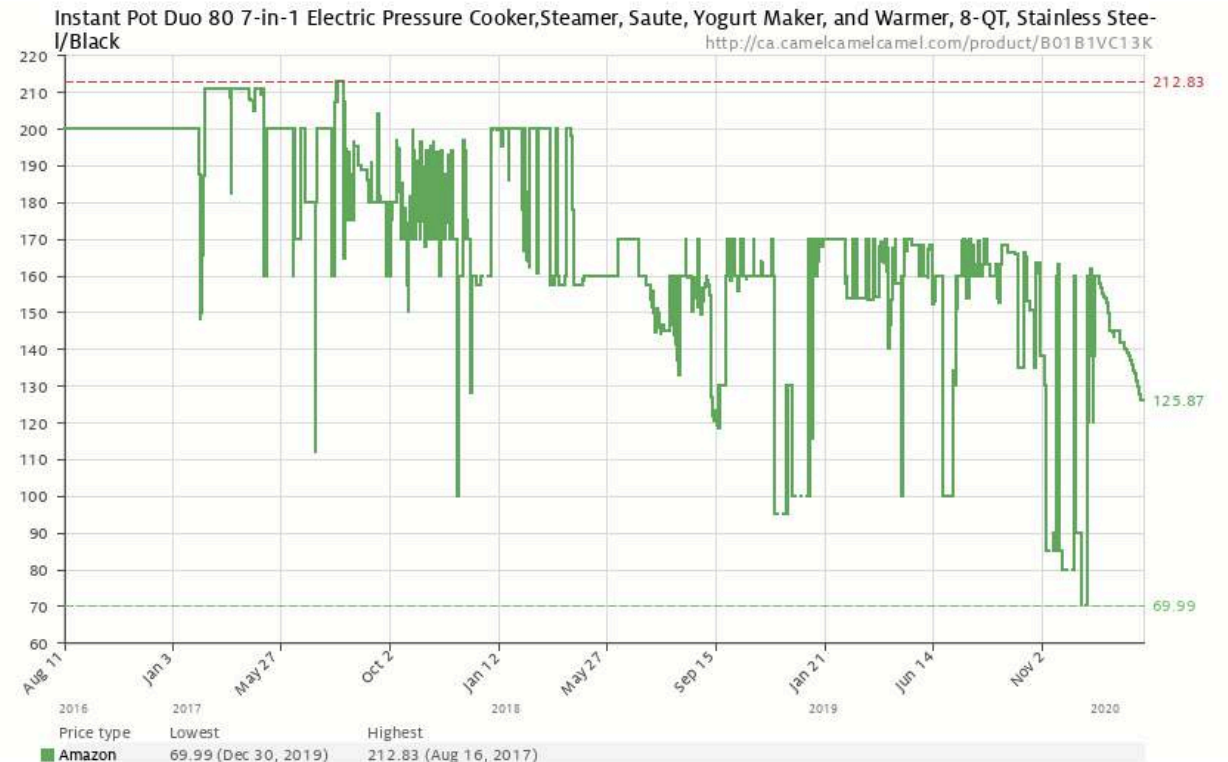
Instant Pot + Cooking Fast Cookbook

+ Indian Cookbook Duo Crisp

+ Fast & Healthy Cookbook

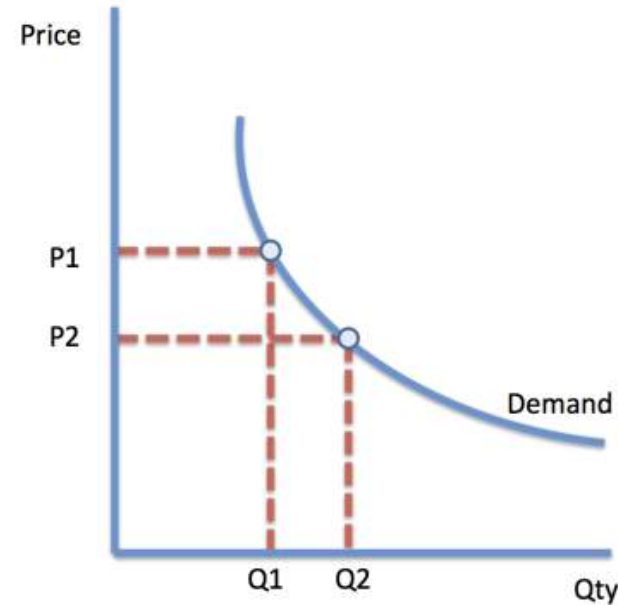


Amazon Price History



Modeling Price Elasticity

- Classic Economic models



- Using machine learning regression trees one can build a demand model and use price as a predictor to the model, “then formulate a price optimization model to maximize revenue [...] using demand predictions from the regression trees as inputs.” see:

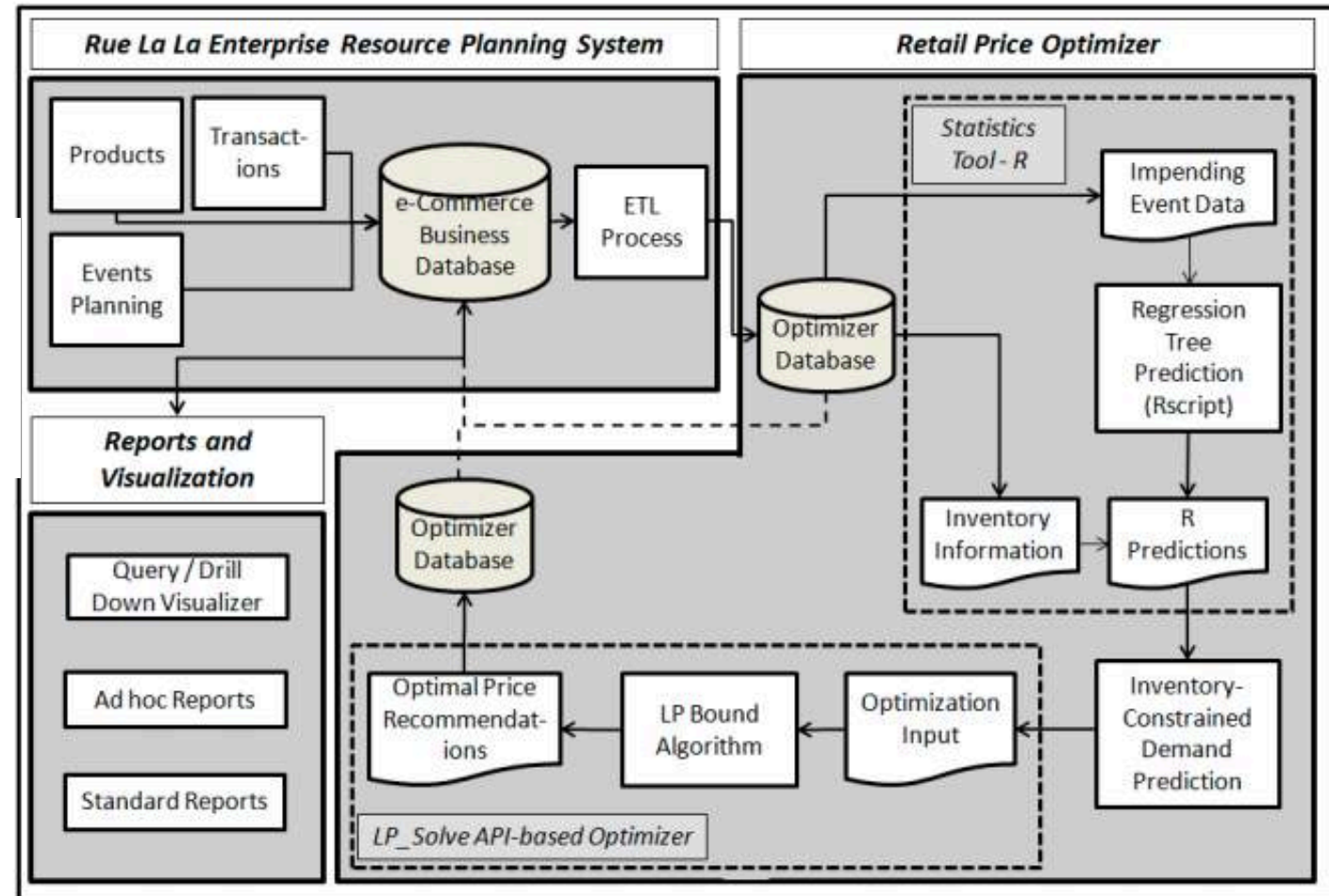
Ferreira, K. J., Hong, B., Lee, A., & Simchi-levi, D. (2013). *Analytics for an Online Retailer: Demand Forecasting and Price Optimization*. (2012), 1–41.

Modeling Price Elasticity (cont.)

- This how an end-to-end solution from Rue-La-La looks like:

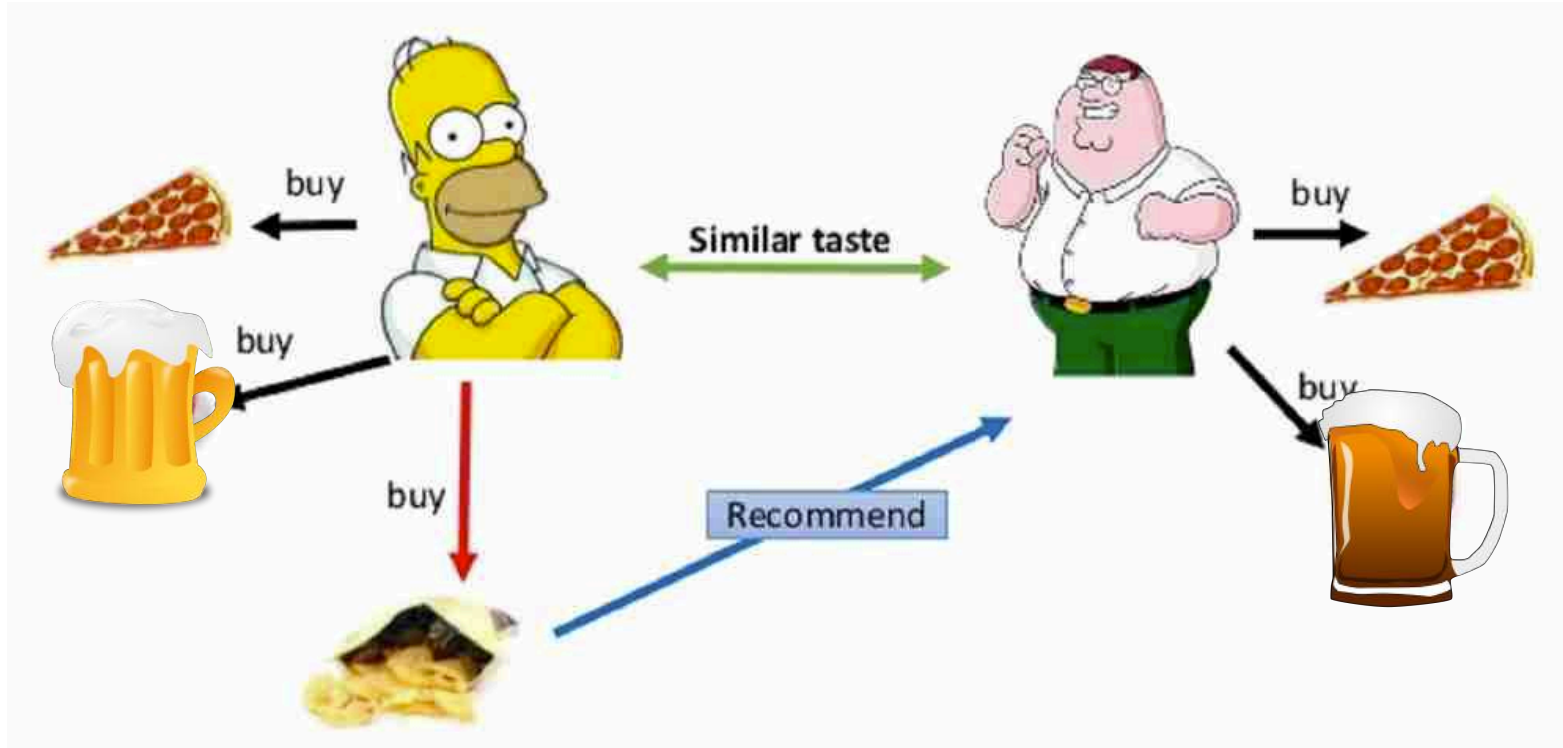


Summary of features used to develop demand prediction model



5. Recommendation Engines

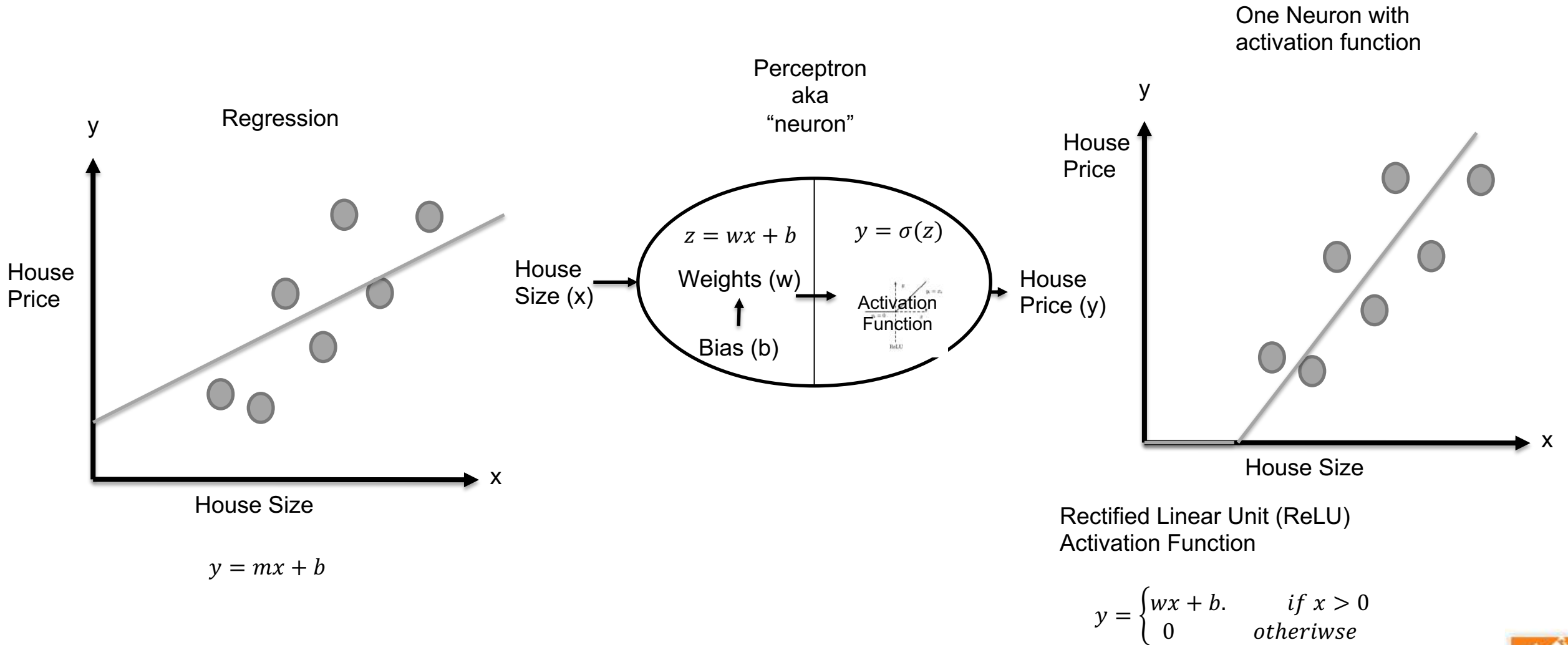
Recommendation Engines



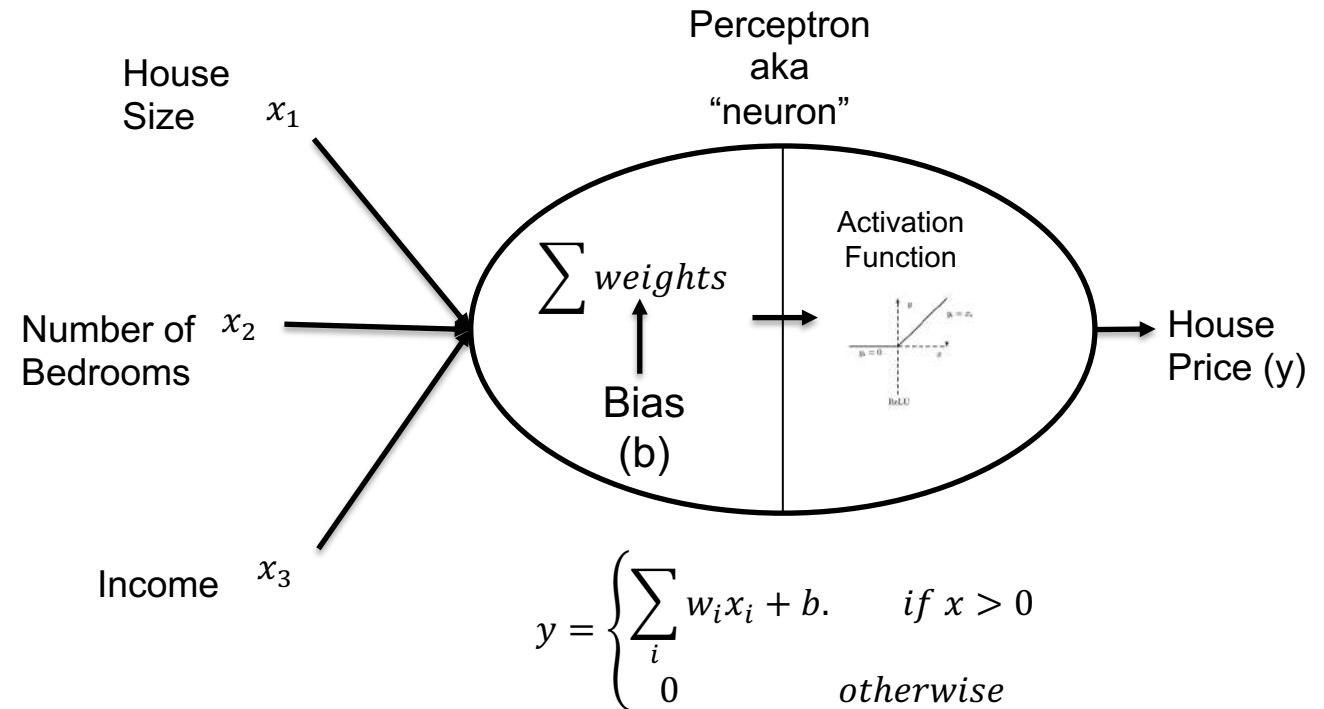
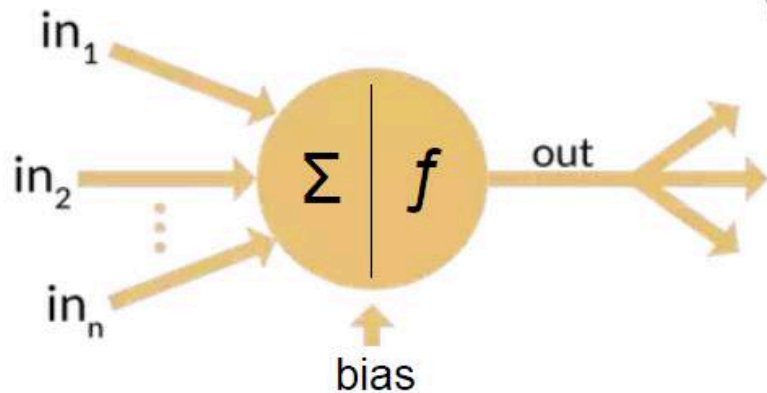
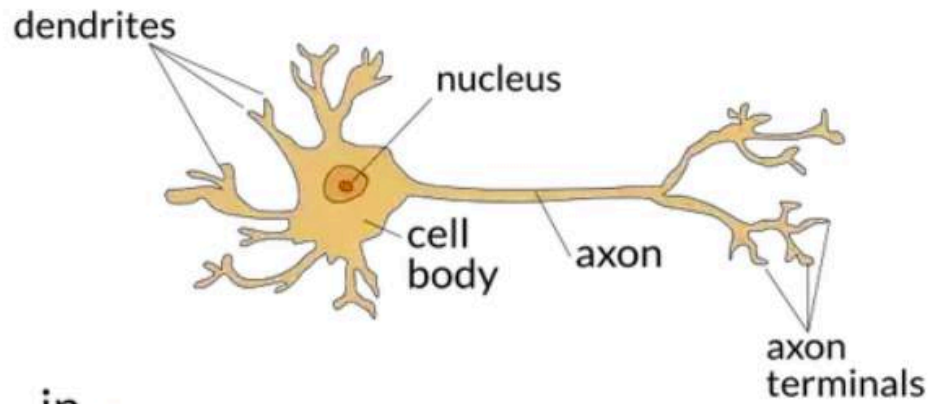
Recommendation Engines

- There are three types of recommendation engines:
 1. **Collaborative Filtering:** This model makes recommendations by learning from user-product historical interactions, either explicit (e.g. user's previous ratings) or implicit feedback (e.g. browsing history). (user to item)
 2. **Content-based:** This model is based primarily on comparisons across products' and users' auxiliary information. A diverse range of auxiliary information such as texts, images and videos can be considered. (item to item)
 3. **Hybrid:** This model refers to recommender system that integrates two or more types of recommendation strategies.
- Neural Networks are mostly used to represent hybrid models due to complexity of these models which need to consider thousands of variables with millions of interaction between them.

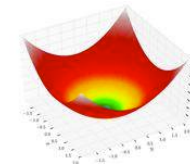
Neural Networks



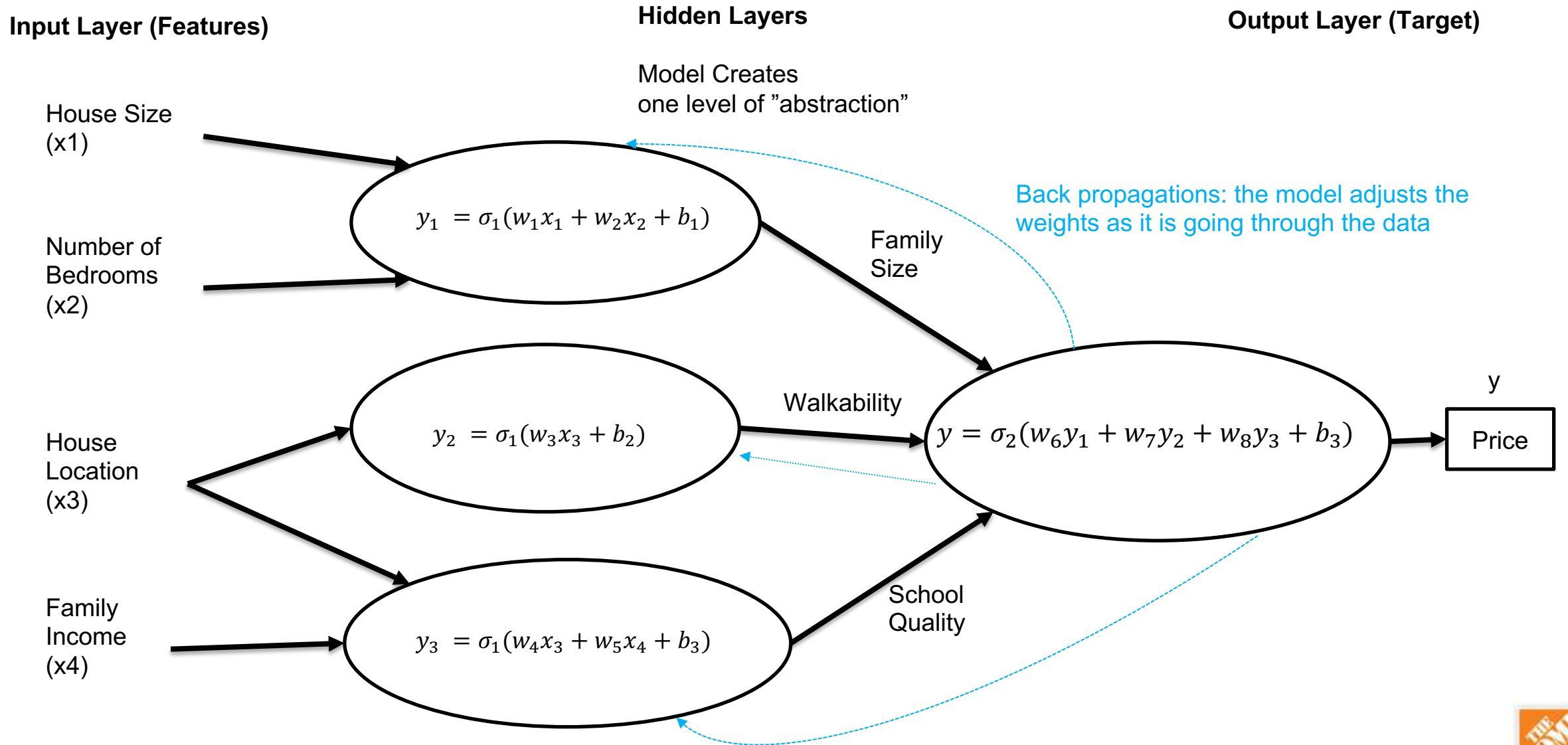
Neural Networks (with multiple inputs)



Learning Process: The weights are determined by using optimization function like gradient decent, i.e. try different weights on a slope to find the minimum error that fits the data

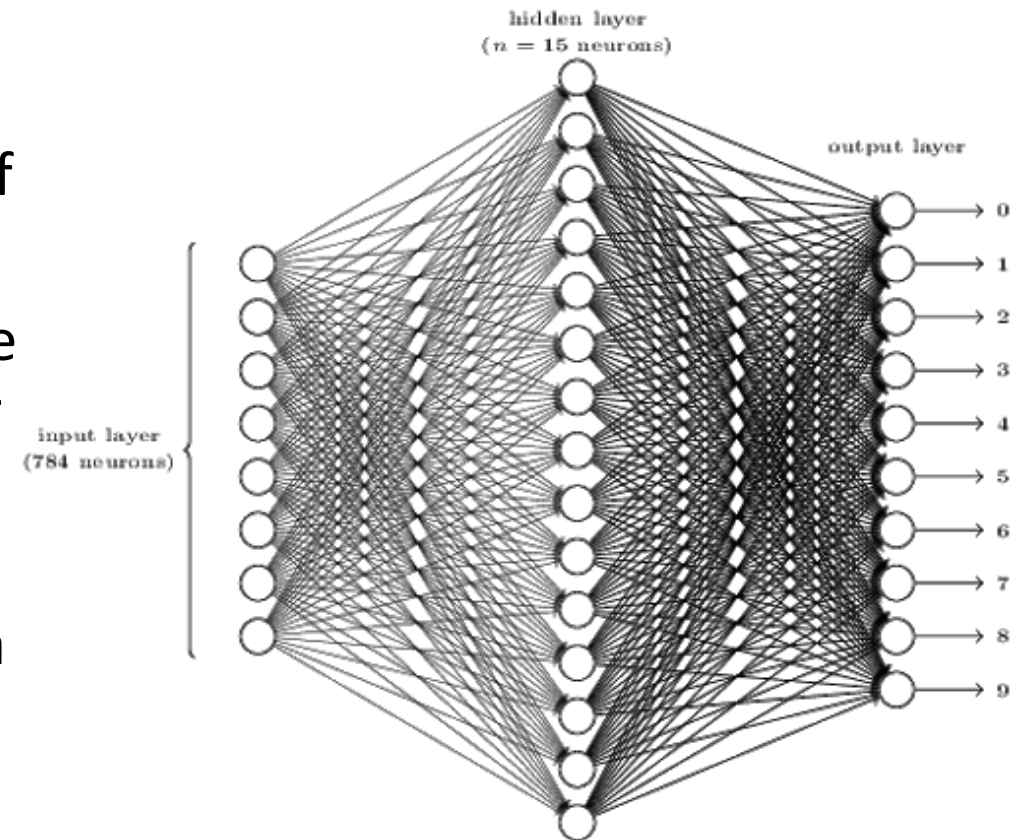


Multi-layer Neural Network



Deep Neural Network

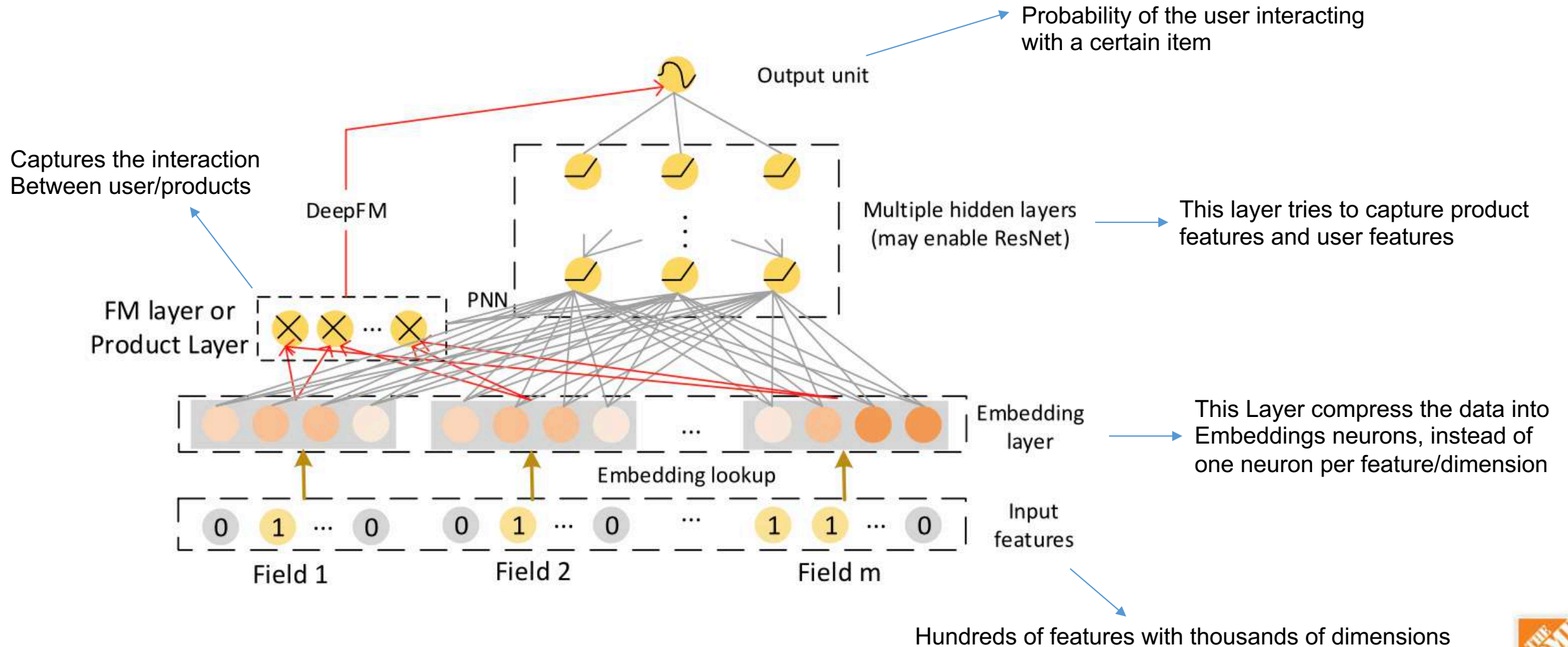
- Deep learning is a phrase used for complex neural networks.
- These networks can be made from millions of neurons.
- These deep learning networks have elaborate structures where the neural network designer guides how information flows through the model.
- These networks require huge amount of data to be trained properly.
- These networks are almost impossible to interpret in terms of “why” certain suggestions have been made from the model.



Hybrid Model for Recommendation




- Our goal was to build a hybrid model that considers all aspects of user interaction with products and all other information we have available about the user and product itself.
- The flexibility of deep neural networks makes it possible to combine several neural building blocks together to complement one another and form a more powerful hybrid model.
 - They can capture non-linear relationships
 - They can combine thousands of features with thousand of dimensions
 - Cloud environments fully supports the implementation of these models
- As a result, all state-of-the-art models from Facebook's DLRM, Microsoft's DeepFM and xDeepFM, Google, and ... are being build using deep neural network as well.

Neural Based Recommendation Engines (DeepFM)








Example: Email Recommendation

- At home depot we have built a deep learning hybrid model the following features: Product dimensions, customer information, price, basket size, online clickstream data, location, etc.

Time	Product	Photo
2019	DEWALT 20V MAX Li-Ion Cordless Drill/Driver and Impact Combo Kit (2-Tool) w/ (2) Batteries 1.3Ah, Charger and Bag	
2019	Home Decorators Collection Wheatfield Oak 12mm x 6.26-inch x 54.45-inch Laminate Flooring	
2019	DEWALT Screwdriver Set with Recess (37-Piece)	
2019	Home Decorators Collection Providence Pine 12 mm Thick x 6.26-inch Wide x 54.45-inch Length Laminate Flooring	
2019	DEWALT 20V MAX XR Li-Ion Cordless Brushless Hammer Drill/Impact Combo Kit (2-Tool) w/ (2) Batteries 2Ah and Charger	
2019	Lifeproof Gainsboro Oak 12mm Thick x 8.03-inch W x 47.64-inch L Laminate Flooring	

Which one did he click on in 2020?

Product	Photo	Model Score	Clicked?
GLACIER BAY Peyton 24-inch W 2-Door Freestanding Vanity in White With Top in White		35%	No
Ram Board 38-inch x 50-ft. Ram Board		59%	No
Leviton Single Pole Decora Light Switches in White (10-Pack)		12%	No
HDX 16-inch x 12-inch Multi-Function Microfibre Cloth		64%	No
Featherlite 6 ft. fiberglass Cross Step Ladder (300 lb). Capacity		90%	Yes


Online Marketing

- Facebook is using similar recommender systems (Facebook's DLRM) to market products to customers:

Suggested for You

Mississauga News
February 20 at 1:30 PM · 🌐

Home Depot holding job fairs at 7 stores in Mississauga and Brampton with more than 200 positions to fill.




MISSISSAUGA.COM
More than 200 jobs up for grabs at Mississauga and Brampton Home Depot job fairs

👍 22 4 Comments · 35 Shares

👍 Like 💬 Comment ➦ Share

The Home Depot
Sponsored · 🌐

Our selection of these latest styles, priced to let you renovate the flooring of your dreams on your budget.



Wickford Vinyl Flooring
NOW \$2.48 / sq. ft. [Shop Now](#)


Bramston Vinyl
NOW \$2.48 / sq. ft. [Shop Now](#)

👍❤️ 14 2 Shares


👍 Like 💬 Comment ➦ Share

Dyson
Sponsored · 🌐

Shop the widest selection of Dyson technology, direct from Dyson. Free shipping and returns on all machines.



Dyson Cinetic Big Ball Multi Floor
\$699.99 [Shop Now](#)



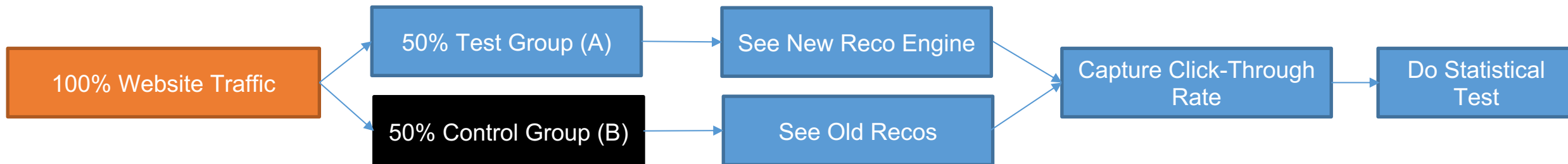
Dyson 360 Heurist
\$1,199.99

👍 Like 💬 Comment ➦ Share

V. Experimentation

Experiment Design

- One major challenge in data science is measuring the impact of changes that are made due to AI and ML.
- Major retailers have departments designated to experiment design.
- In some cases, the experiment can be designed before-hand and the effected groups can be grouped into two or more groups and A/B tests can be done to see the effect of these changes.

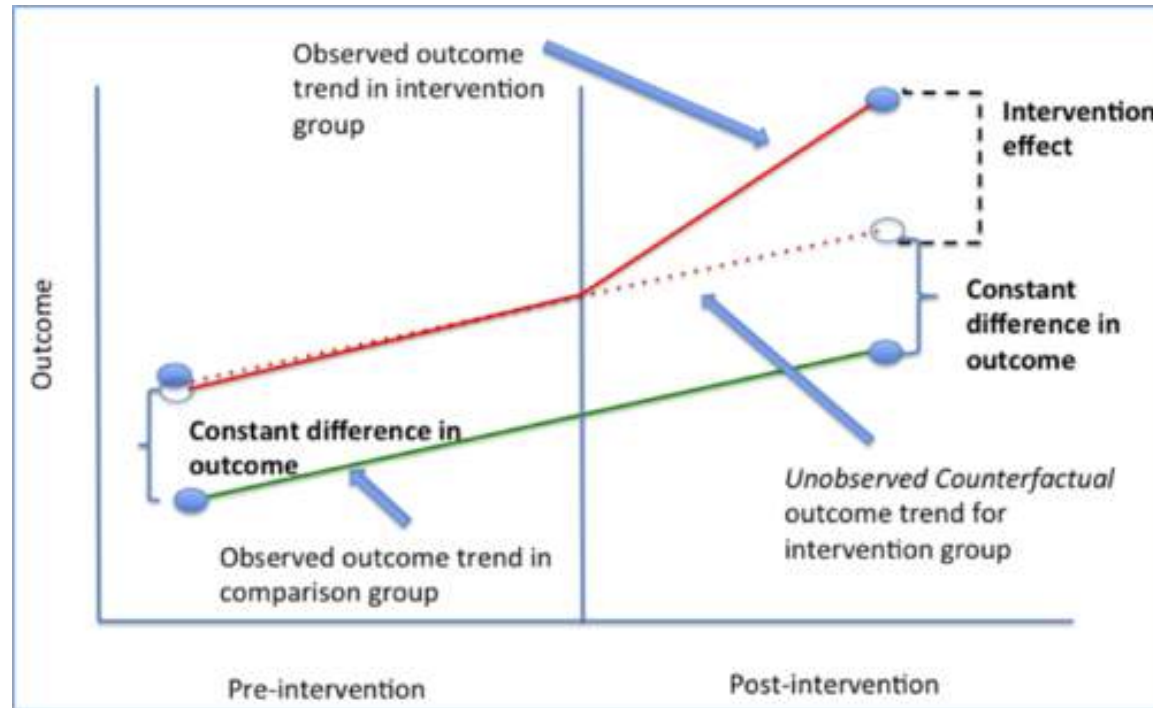


- The main challenges in business is finding similar groups since the sales are constantly changing.



Experiment Design (cont.)

- Since the business is constantly changing through time, this will require more advanced techniques in finding how the experiment have changed the results.
- One such technique is known as difference in differences which accounts for some changes that happen though time.

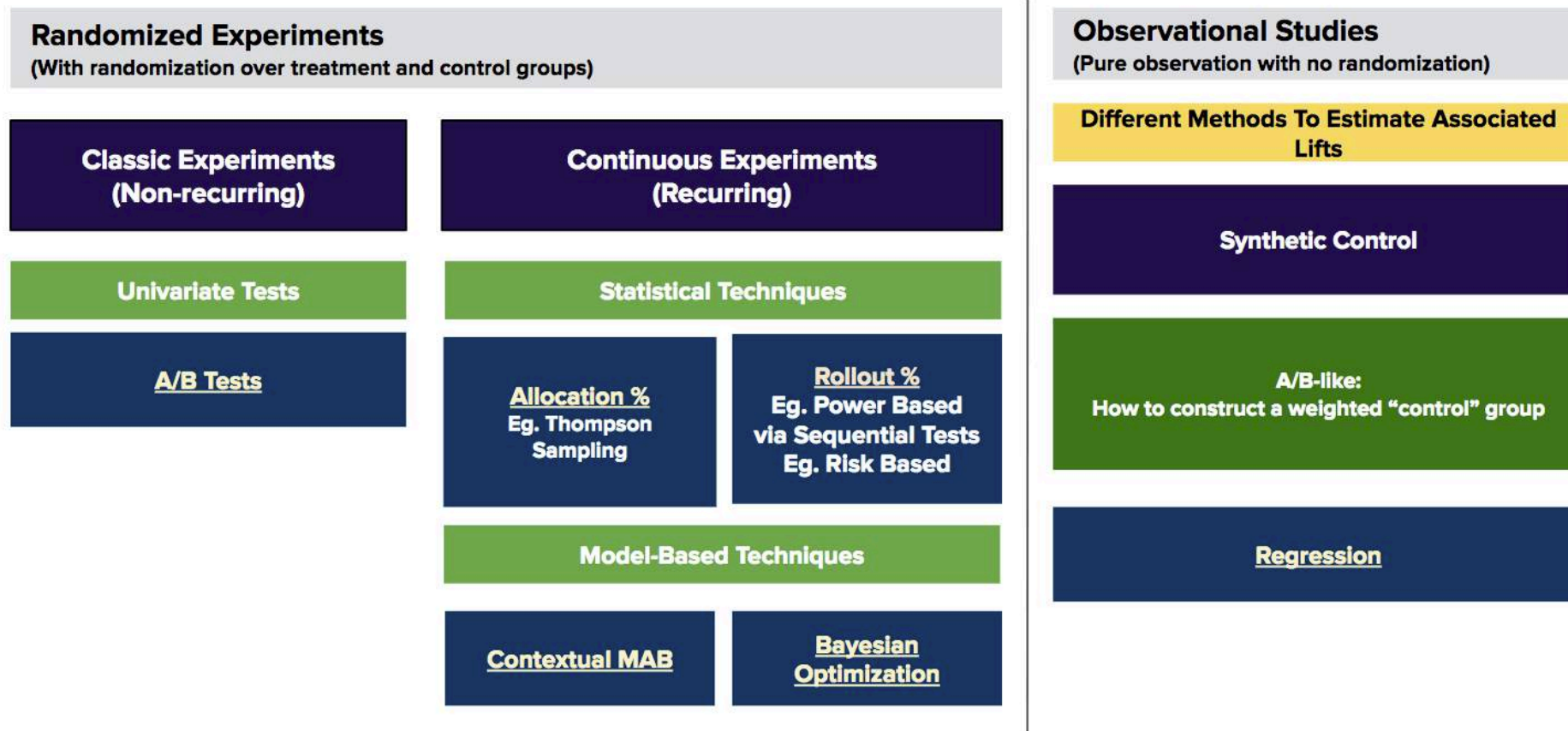


- When the test is not designed beforehand, the only available data to compare with would be historical data.
- In these cases, similar techniques can be used to compare the data with its historical counterparts while considering the growth that has happened throughout time.

Experiment Design (cont.)

- In modern retailers, experimentation is done routinely through automated platforms. A good similar example is Uber's experimentation platform

Overview of data generation, modeling and interpretation in statistical perspectives



VI. How to become a Data Scientist

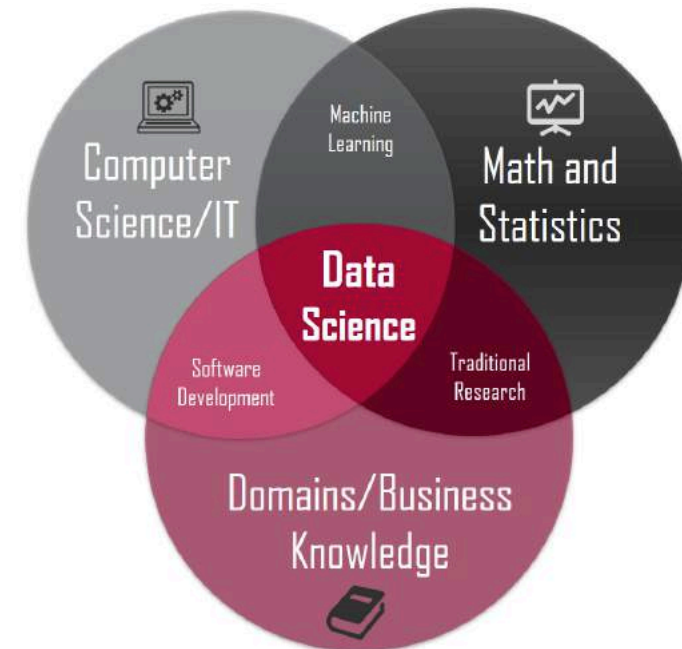
What you need to know as a Data Scientist?

- These are skills that you need according to World Economic Forum.

Change in rank of skills genome of selected emerging roles

rank	2015	change in rank	2018
1	Machine Learning	[new]	TensorFlow
2	Python (Programming Language)	[-1]	Machine Learning
3	Apache Spark	[+1]	Deep Learning
4	Deep Learning	[new]	Keras
5	Algorithms	[-2]	Apache Spark
6	Java	[new]	Natural Language Processing (NLP)
7	Big Data	[new]	Computer Vision
8	Hadoop	[-6]	Python (Programming Language)
9	Data Science	[-]	Data Science
10	C++	[new]	Amazon Web Services (AWS)

From: http://www3.weforum.org/docs/WEF_Data_Science_In_the_New_Economy.pdf

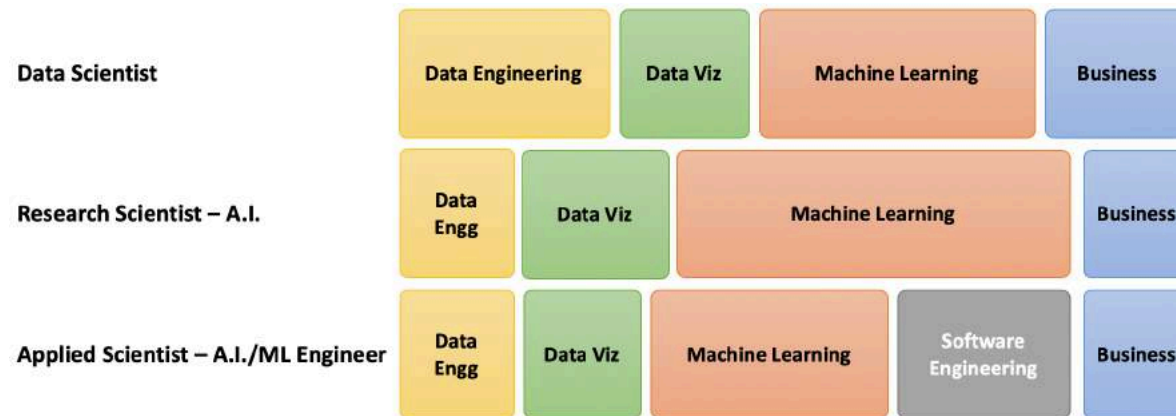


From: <https://www.datasciencesociety.net/data-science-career-path-after-college/>

- Python, Java, Spark are becoming more base skills and skills like TensorFlow, Keras, ... are getting more importance i.e. there is need for data scientist to become more and more specialized.

Specialization in Data Science

- Data Science itself is becoming more and more specialized as well.
- In 2020, data science market have evolved into three main branches:



- These specialization trend will continue in near future and data scientist need to specialize in certain aspect of the work, e.g., computer vision engineers, supply chain engineers, etc.

Thank you

