DATA MINING

Introduction to Data Mining

Ashish Pujari

Lecture Outline

- Class Introductions
- Syllabus Review
- Introduction to Data Mining
- Data Mining Applications
- Software

Syllabus

- Week 1: Introduction to Data Mining
- Week 2: Dimensionality Reduction 1
- Week 3: Dimensionality Reduction 2
- Week 4: Cluster Analysis I
- Week 5: Cluster Analysis II
- Week 6: Association Rules Mining
- Week 7: Recommender Systems
- Week 8: Bayesian Networks
- Week 9: Graph Mining

Coursework

- Submissions
  - 4 bi-weekly assignments
  - 2 paper reviews
  - Quizzes
  - Final Project
- Class participation
  - Attendance
  - Team-work
  - In-class discussions
  - Zoom

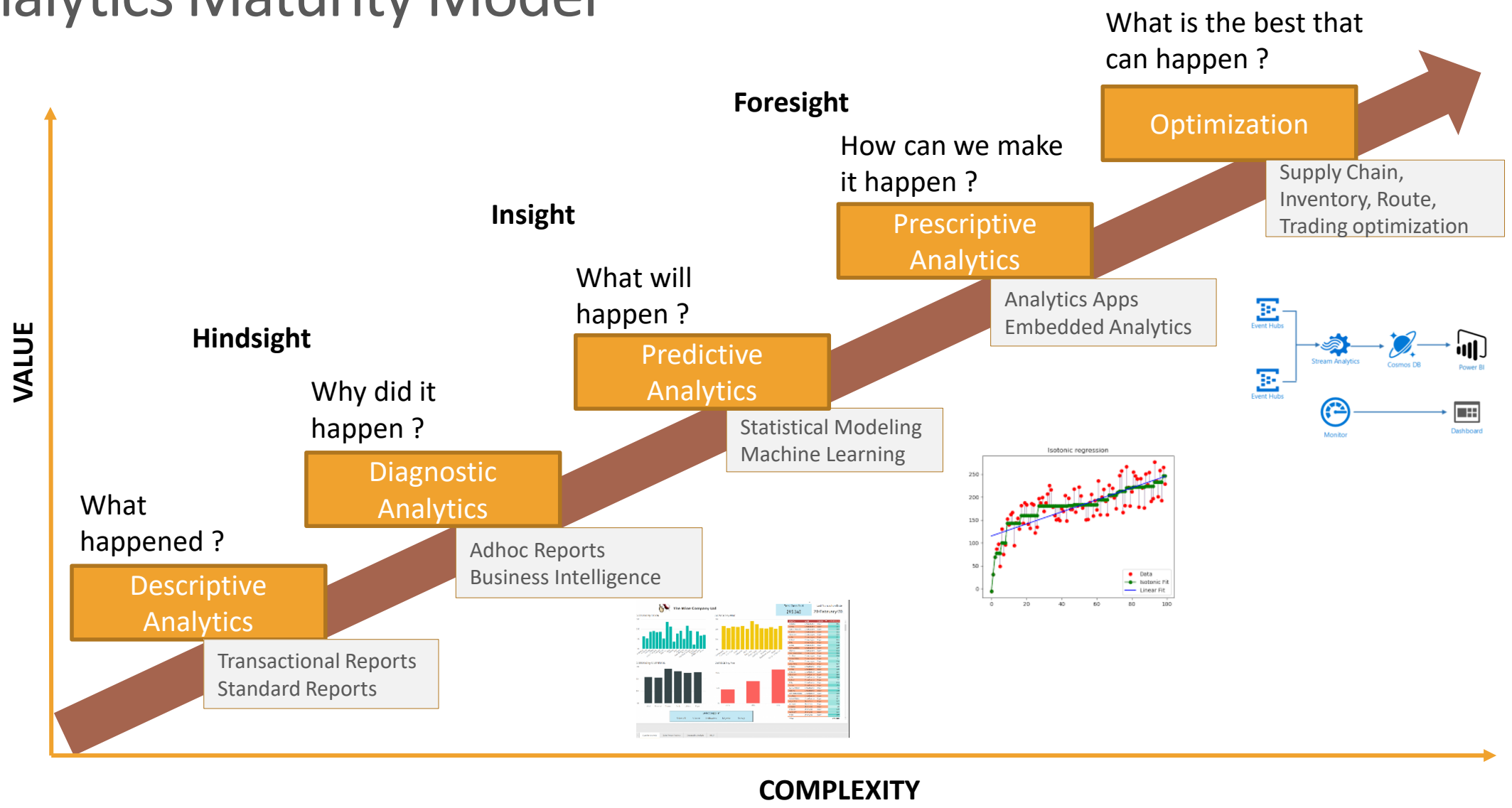
Resources

- Suggested Books
  - The Elements of Statistical Learning
  - Python Machine Learning
  - Hands on Machine Learning
  - Probabilistic Graphical Models
  - Machine Learning: A Probabilistic Perspective

- Websites
  - Google AI Blog
  - Face Book AI
  - KD Nuggets
  - Medium – Data Science

DATA MINING

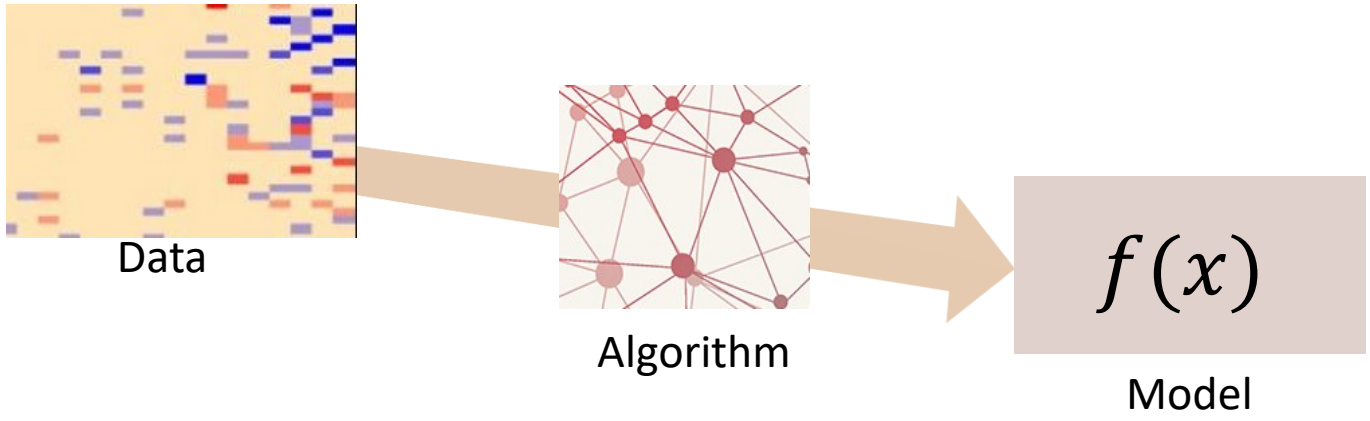
Analytics Maturity Model



Machine Learning

Field of AI that gives “computers the ability to learn without being explicitly programmed” - *Arthur Samuel*

“A computer program is said to **learn** from experience ***E*** with respect to some class of tasks ***T*** and performance measure ***P*** if its performance at tasks in ***T***, as measured by ***P***, improves with experience ***E***.” - *Tom M. Mitchell*

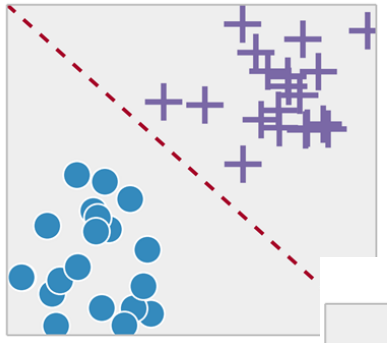


Types of Machine Learning

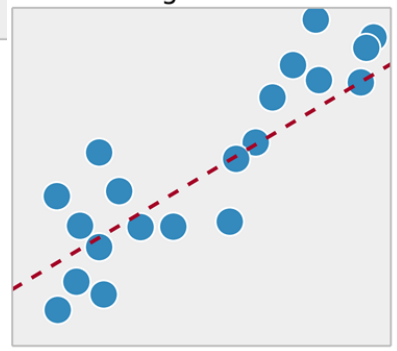
Machine Learning

1. Supervised Learning

Classification

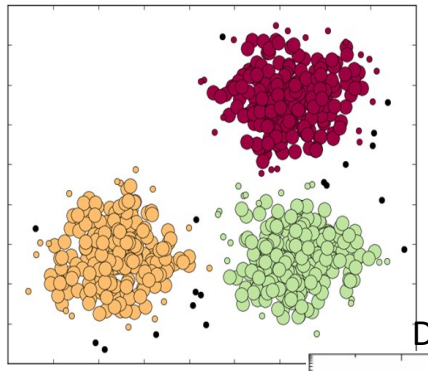


Regression

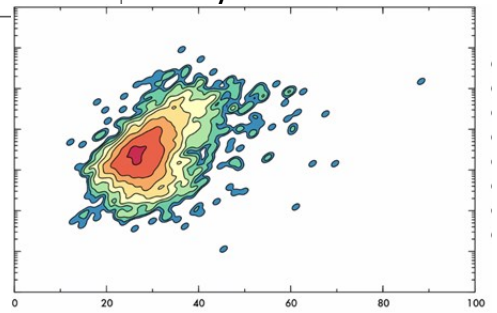


2.Unsupervised Learning

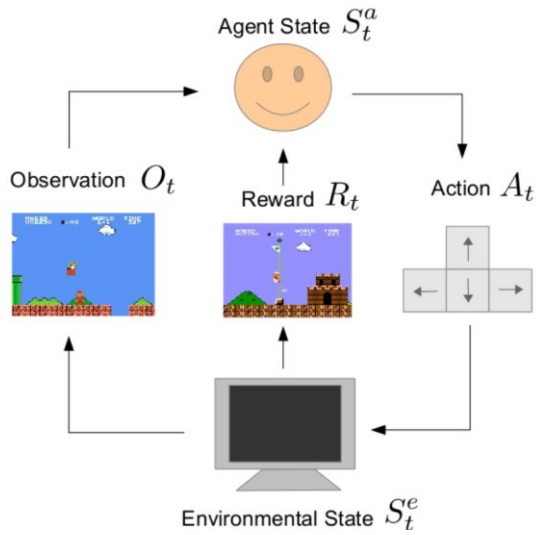
Clustering



Density Estimation



3.Reinforcement Learning



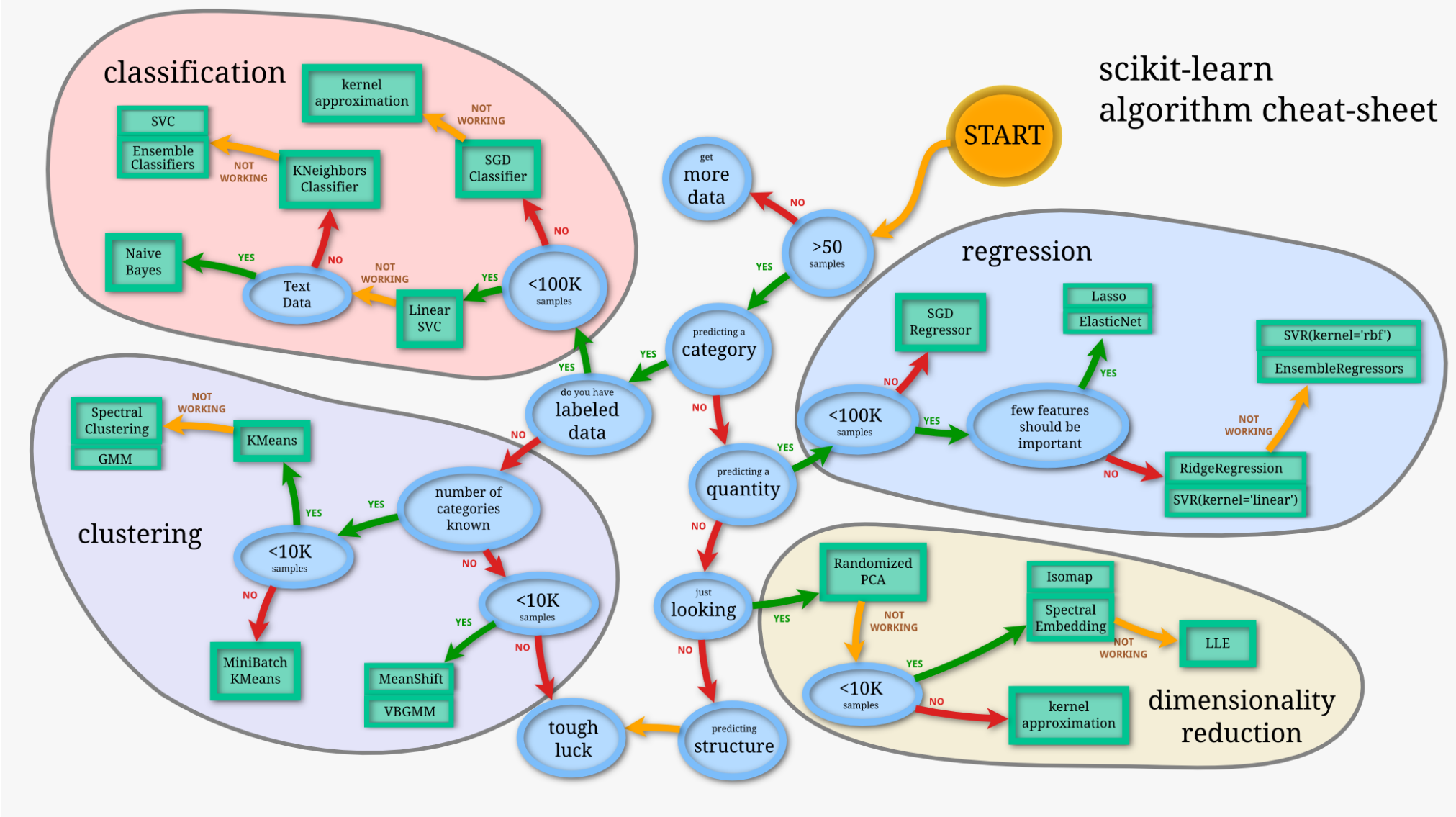
Supervised vs Unsupervised Learning

	Supervised	Unsupervised
Data	(x, y) where x is data, y is label	x - just data, no labels
Goal	Learn a function to map $x \rightarrow y$ Form a model for $\mathrm{P}(x\|y)$, where y is the label for x	Learn underlying hidden structure of the data Form a model for $\mathrm{P}(x)$, where x is an input vector
Methods	Regression, object detection, semantic segmentation, image captioning, etc.	Clustering, dimensionality reduction, feature learning, density estimation, etc.
Example	 "Cat"	 

What kind of questions can ML answer ?

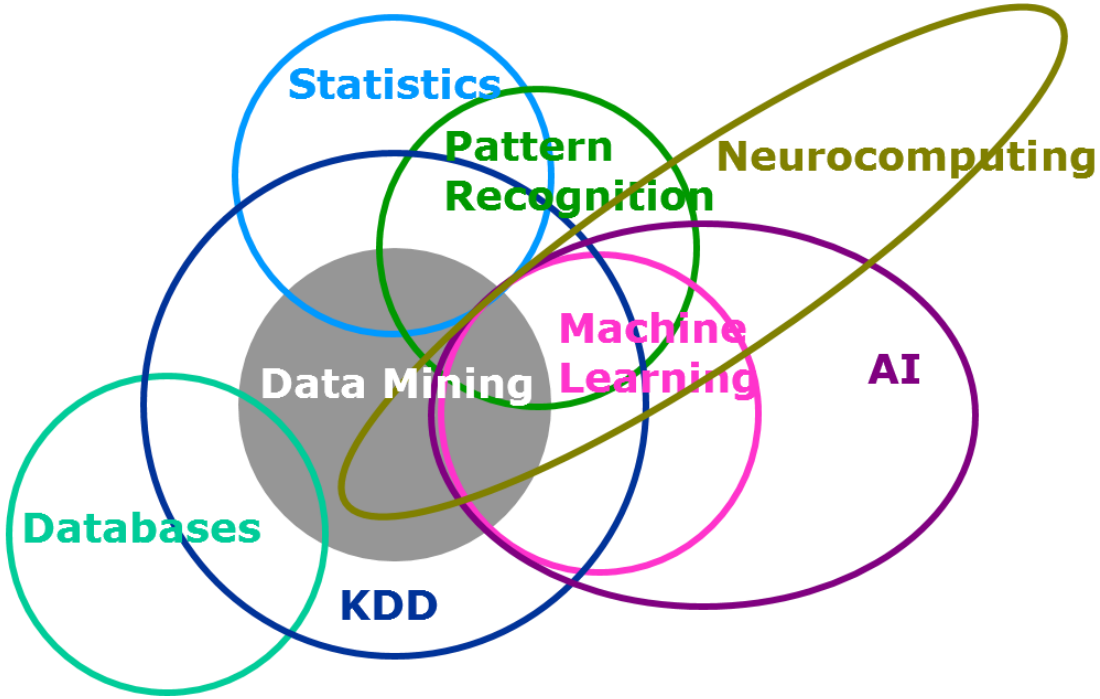
Questions	Description	Examples
Is this A or B?	**Classification**: Questions that have two or more possible answers.	Which animal is in this image? Will this customer click on the top link?
Is this an anomaly ?	**Anomaly Detection**: Questions about events that seems out of the ordinary.	Is this pressure reading unusual? Is this combination of purchases normal for this customer ?
How much/how many ?	**Forecasting/Regression**: When we are looking for a number	What will the temperature be next Tuesday? How many customers will we acquire next quarter ?
How is it organized ?	**Segmentation/Clustering**: Knowing if there a hidden structure in our data points.	Which shoppers have similar tastes in produce? Which viewers like the same kind of movies?
What should I do next?	**Recommendation**: Deciding the optimal future actions based on past information	Where should I place this ad on the webpage so that the viewer is most likely to click it ? How many shares of this stock should I buy ?

ML Algorithms



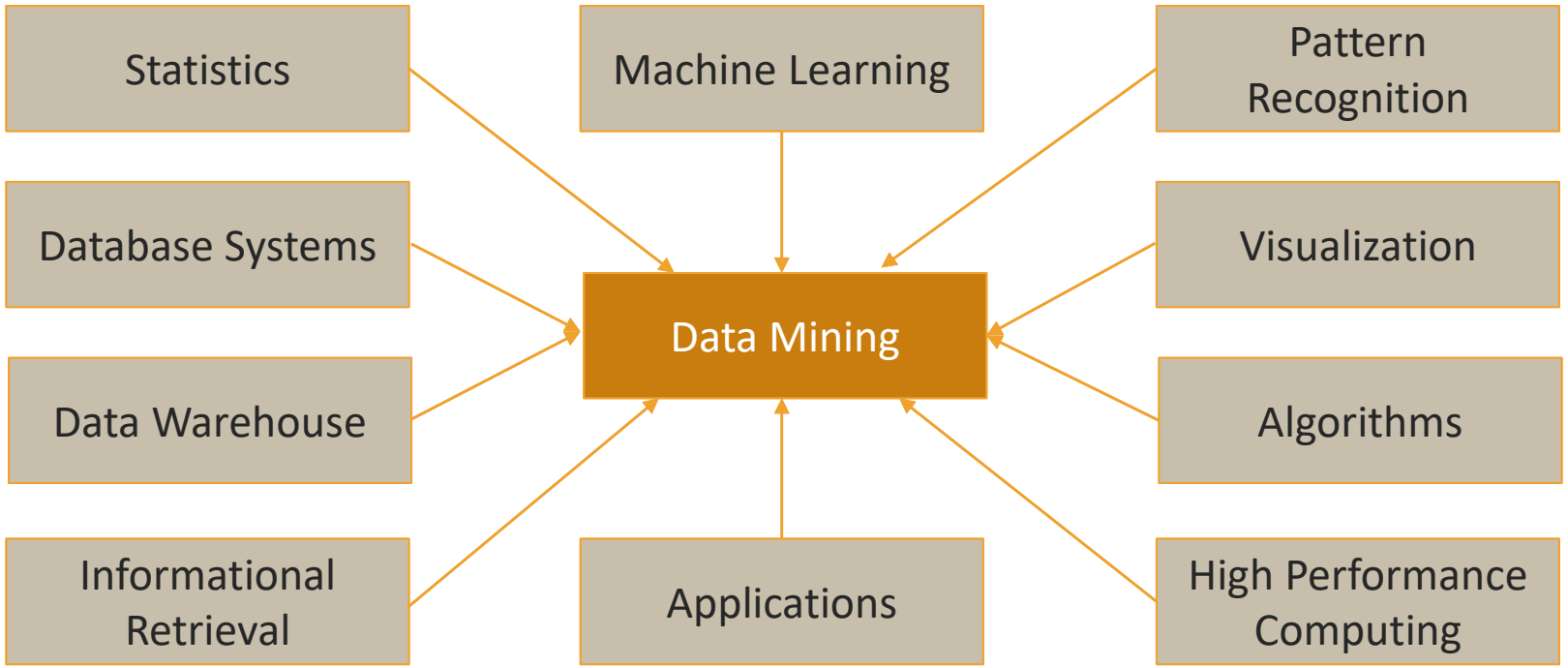
Data Mining

- Definitions
  - Systematic process of discovering patterns in data sets through the use of computer algorithms.
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- Alternative names
  - Knowledge discovery in databases (KDD)
  - Knowledge extraction
  - Pattern analysis
  - Business intelligence, etc.

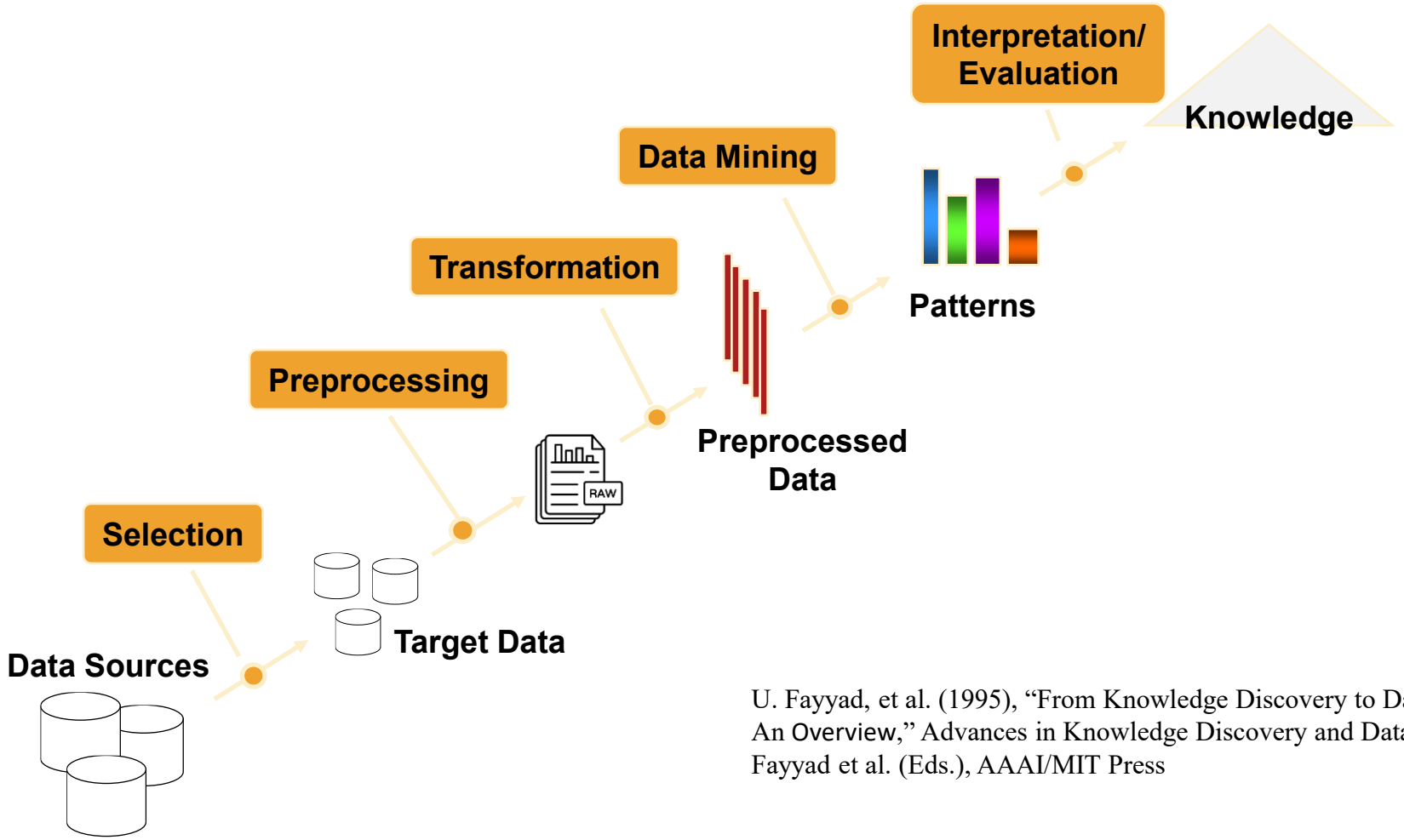


Source: SAS

Data Mining


Statistics
Machine Learning
Pattern Recognition
Database Systems
Visualization
Data Mining
Data Warehouse
Algorithms
Informational Retrieval
Applications
High Performance Computing

Data Mining Process



U. Fayyad, et al. (1995), "From Knowledge Discovery to Data Mining: An Overview," Advances in Knowledge Discovery and Data Mining, U. Fayyad et al. (Eds.), AAAI/MIT Press

Steps of a Data Mining Process

1. Learning the business/application domain
2. Creating a target data set: data selection
3. Data cleaning and preprocessing
4. Data reduction and transformation
5. Choosing functions of data mining
6. Choosing the mining algorithm(s)
7. Data mining: discover patterns of interest
8. Pattern evaluation and knowledge presentation
9. Use of discovered knowledge

Patterns

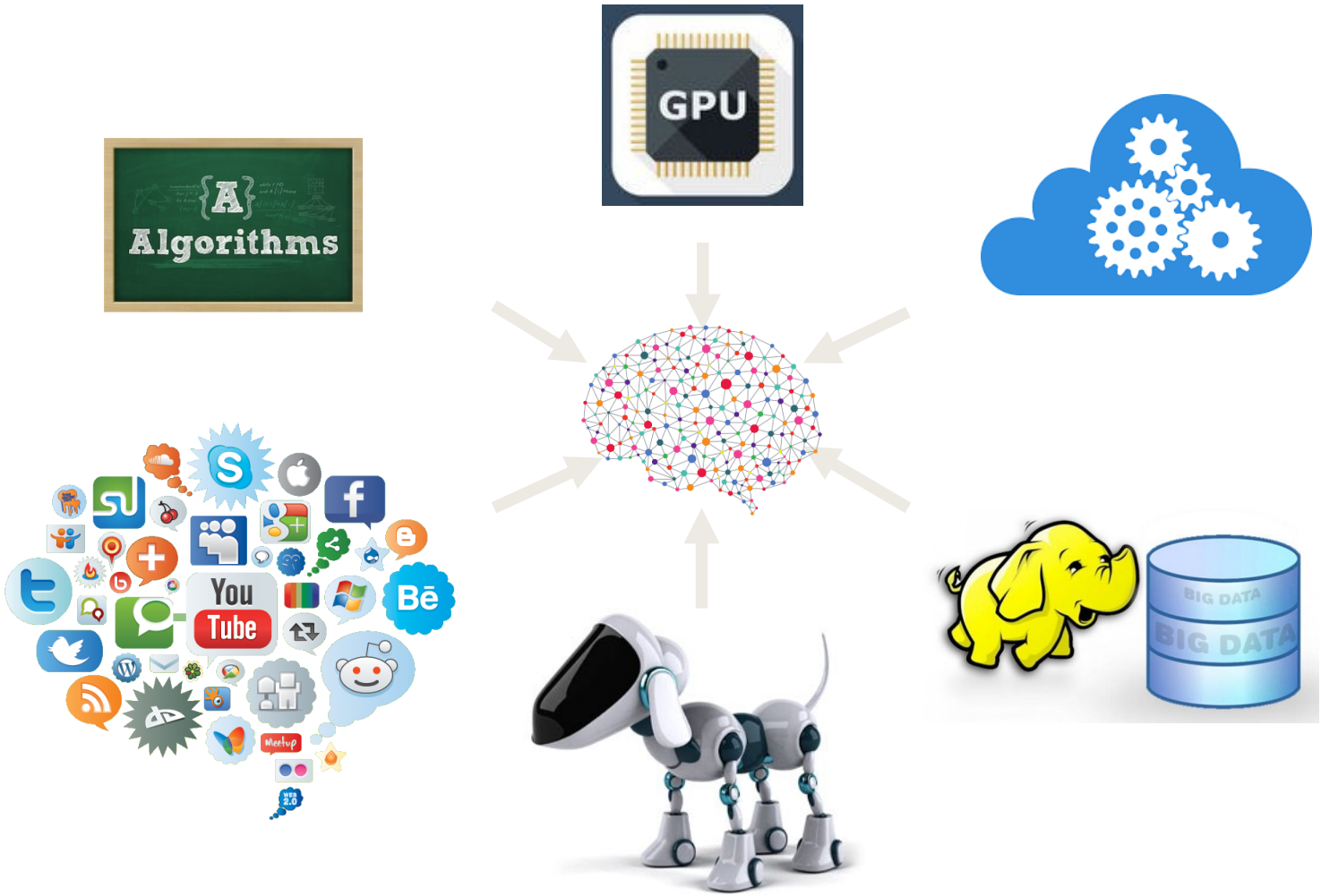
- Periodic Patterns
  - Seen repeating themselves after a certain lapse of time.
  - E.g., Time series data, biological sequences, spatiotemporal data, etc.
- Associative Patterns
  - Co-occurring groups of things that are complementary to each other.
  - E.g., Market basket, shopping carts, etc.
- Abnormal Patterns
  - Data has a clear deviation from normal behavior or appearance is not periodic.
  - E.g., Credit card/insurance fraud, health metrics, etc.
- Structural Patterns
  - Pathfinding in graphs or cluster identification
  - E.g., Market segmentation, real estate clustering, routing, etc.

Pattern Extraction Approaches

- Search
  - Finding all the interesting patterns
  - Heuristic vs. exhaustive search
  - Limited by compute resources - complexity
- Optimization
  - Searching for only interesting patterns
  - First generate all the patterns and then filter
  - Limited by compute resources - complexity
- Visualization
  - Use human perception to recognize patterns in large data sets
  - Perceive non-trivial patterns
  - Limited by data set size and high dimensionality

Programming Languages

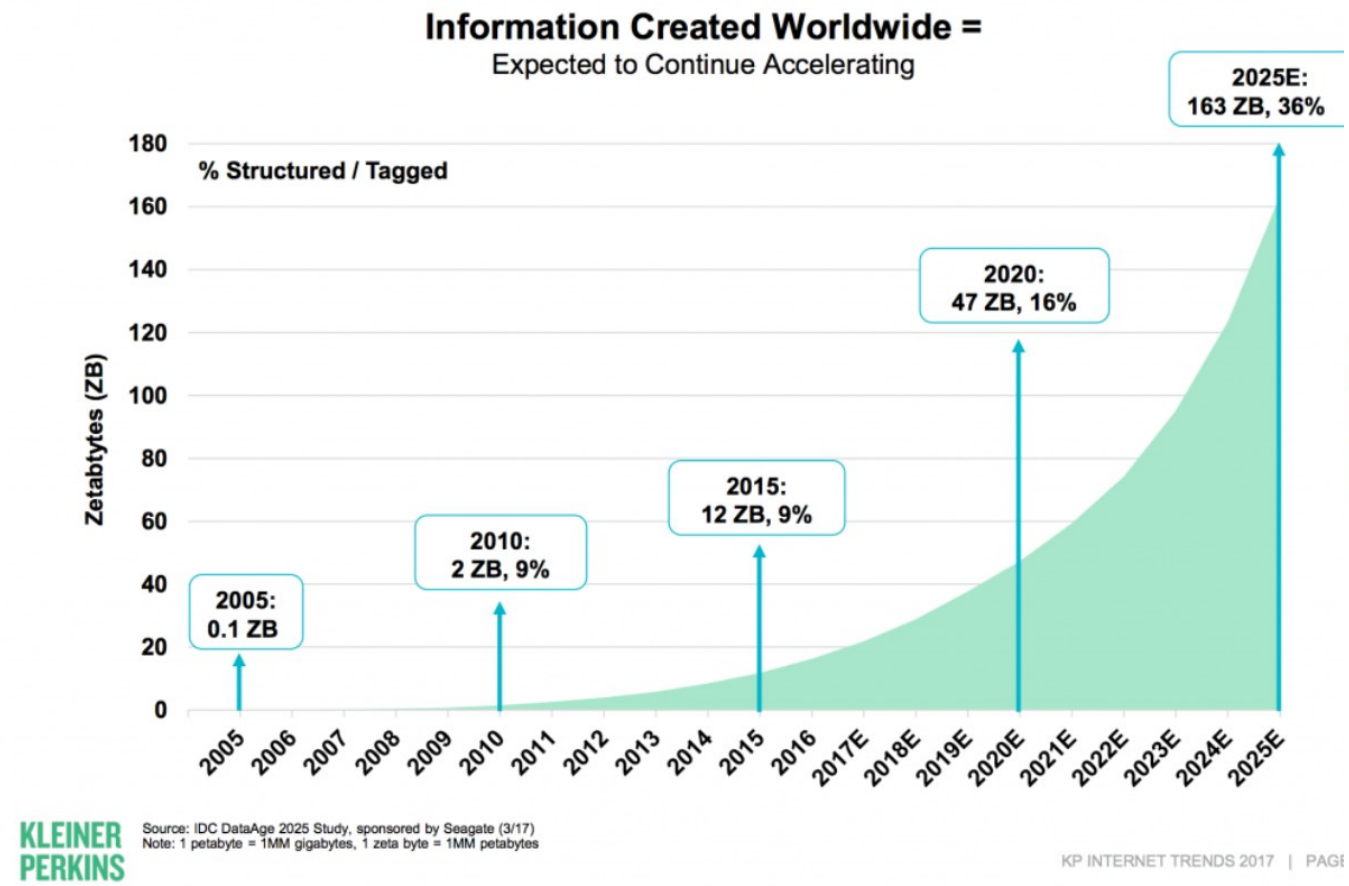
Techology Convergence



Data Mining Challenges

- Data Quality
- High Dimensionality
- Scalability
- Data Complexity
- Data Governance

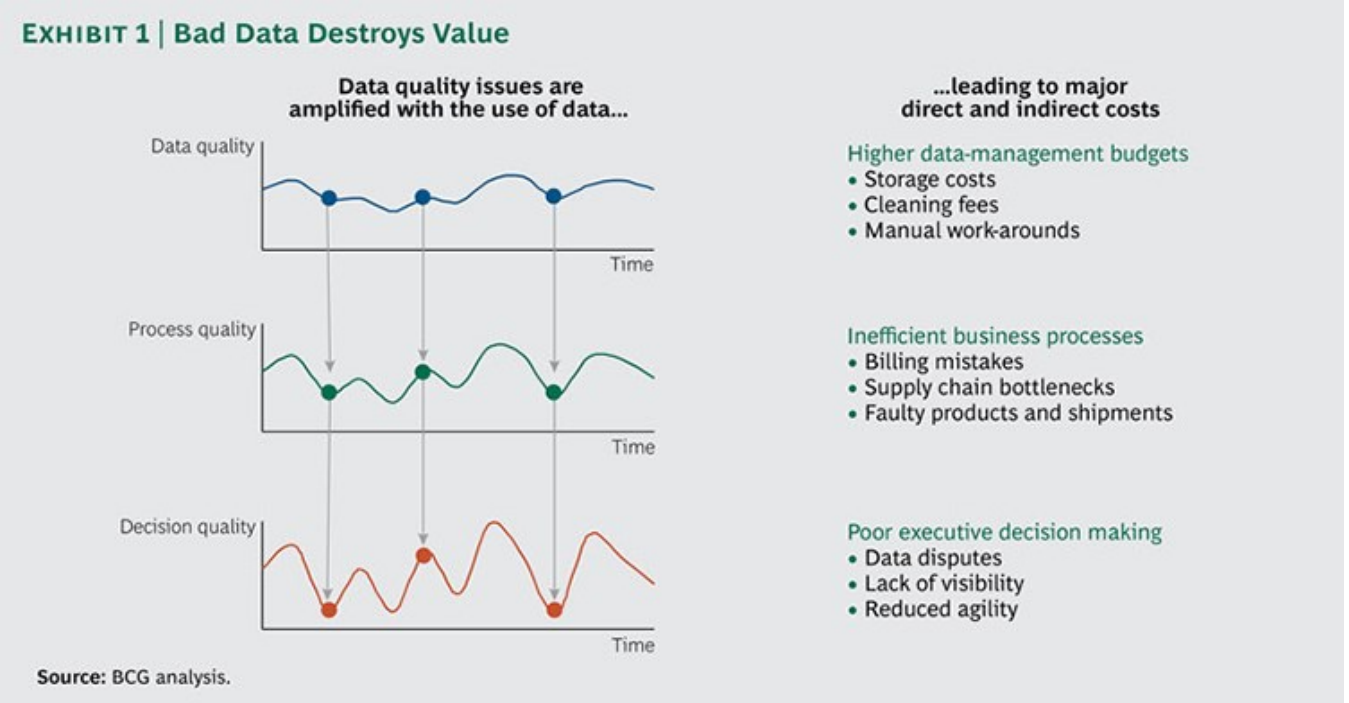
Big Data


Information Created Worldwide =
Expected to Continue Accelerating
% Structured / Tagged
Zetabytes (ZB)
2005: 0.1 ZB
2010: 2 ZB, 9%
2015: 12 ZB, 9%
2020: 47 ZB, 16%
2025E: 163 ZB, 36%
0
20
40
60
80
100
120
140
160
180
2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017E 2018E 2019E 2020E 2021E 2022E 2023E 2024E 2025E
KLEINER PERKINS
Source: IDC DataAge 2025 Study, sponsored by Seagate (3/17)
Note: 1 petabyte = 1MM gigabytes, 1 zeta byte = 1MM petabytes
KP INTERNET TRENDS 2017 | PAGE


60 SECONDS
facebook
1 Million Logging In
18.1 Million Texts Sent
YouTube
4.5 Million Videos Viewed
390,030 Apps Downloaded
347,222 Scrolling Instagram
87,500 People Tweeting
1.4 Million Swipes
188 Million Emails Sent
1 Million Views
41 Music Streaming Subscriptions
180 Smart Speakers Shipped
4.8 Million Gifs Served
41.6 Million Messages Sent
2.1 Million Snaps Created
$996,956 Spent Online
694,444 Hours Watched
3.8 Million Search Queries
Google
NETFLIX
tinder
twitch
GIPHY
amazon echo
Google Home
Facebook Messenger
WhatsApp

Data Quality

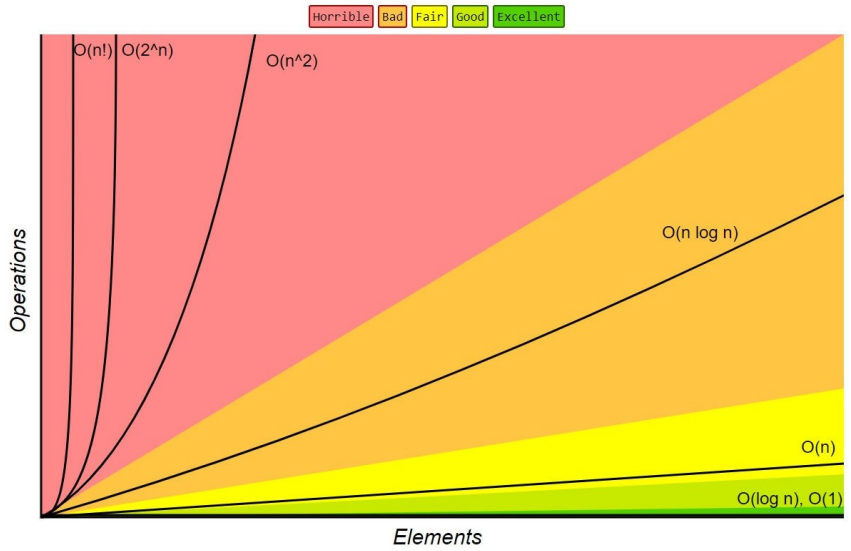
- Poor data quality caused by
  - Manual data entry
  - Measurement related errors
  - Duplicate data entry
  - Absence of well-defined standards
  - Inconsistent data formatting
  - Numeric approximations
  - Software and hardware constraints



IBM's estimate (2016) of the yearly cost of poor-quality data is $3.1 trillion in the US alone

Algorithmic Complexity

- Complexity is defined a numerical function T(n) - time versus the input size n.
- We want to define time taken by an algorithm in terms of its input without depending on the implementation details.

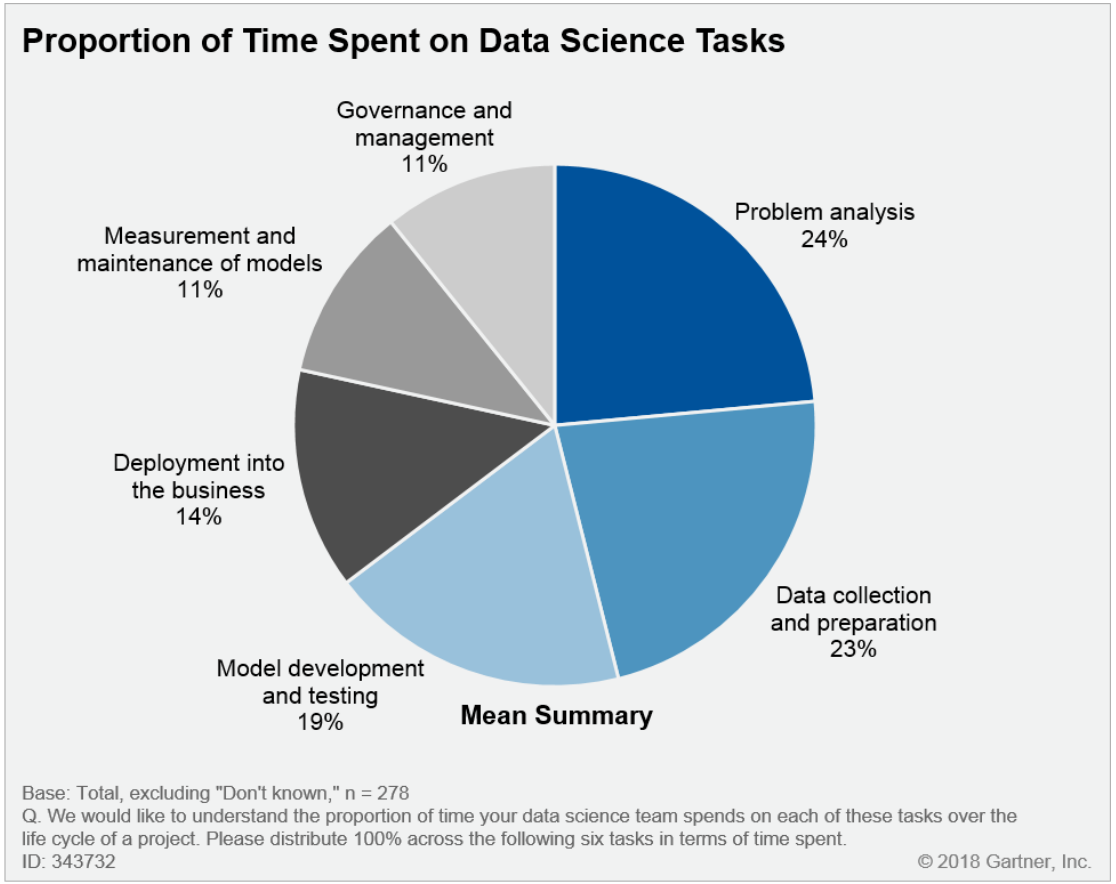


Source: Big-O Cheat Sheet, 2016.

Big O Notation	Definition, Examples
Constant Time: **O(1)**	Execution time is the same regardless of the input size. Examples: array: accessing any element, fixed-size stack: push and pop methods
Linear Time: **O(n)**	Execution time is directly proportional to the input size, i.e., time grows linearly as input size increases. Examples: array: linear search, traversing, find minimum
Logarithmic Time: **O(log n)**	Execution time is proportional to the logarithm of the input size Examples: binary search
Quadratic Time: **$O(n^2)$**	Execution is proportional to the square of the input size. Examples: bubble sort, selection sort, insertion sort

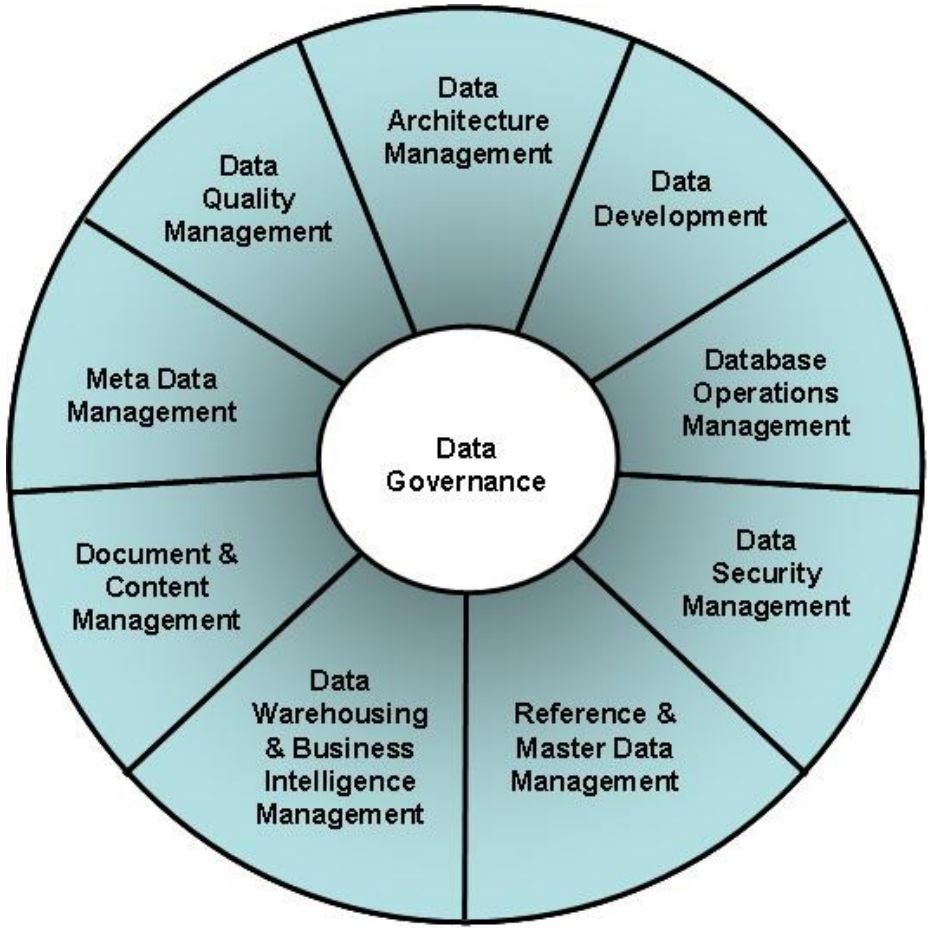
Data Preparation

A significant portion of the time spent in data science is for data collection, data understanding and preparation



Data Governance

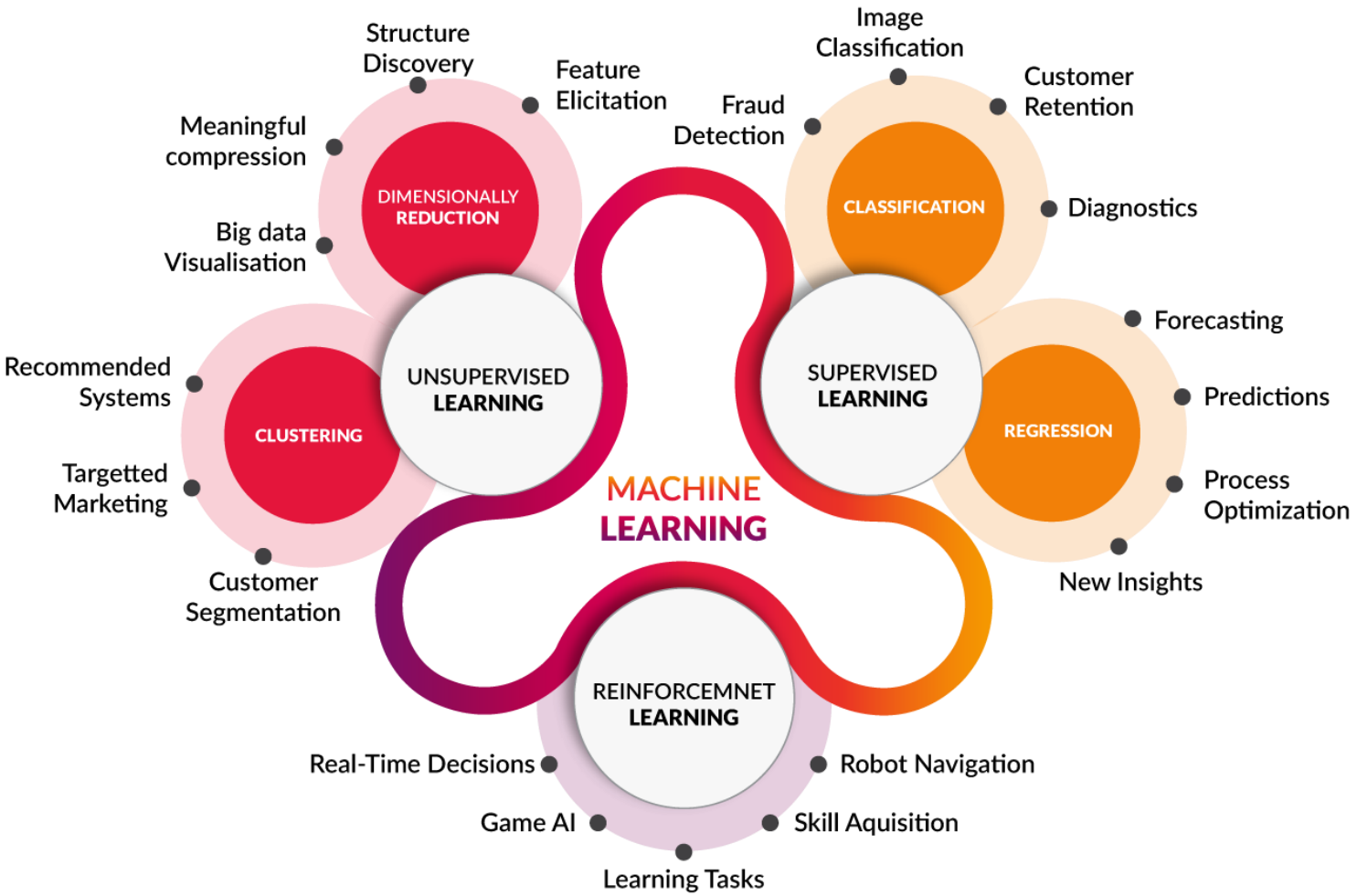
- Exercise of authority and control (planning, monitoring and enforcement) over the management of data assets
- Creation and enforcement of policies or standards for appropriate:
  - Data entry, update & use
  - Data quality control activities
  - Business metadata management
- Activities ensuring compliance with data governance policies



https://damadach.org/dama-dmbok-functional-framework/

APPLICATIONS

Machine Learning Applications



http://www.cognub.com/index.php/cognitive-platform/

CPG - Pricing Analytics

- Better pricing decisions through data
  - Discounts and promotions
  - Price thresholds and Competitive effects
  - Seasonality, External factors
- Value-based Pricing
  - Optimized pricing structure that maximizes profitability
  - Identifying key business value drivers
  - Assessing the value of the product against key value drivers
  - Assessing the comparative value vs. competitor products

Source: adaptly.com

Source: actionableinsights.online

Demand Forecasting, Assortment Optimization

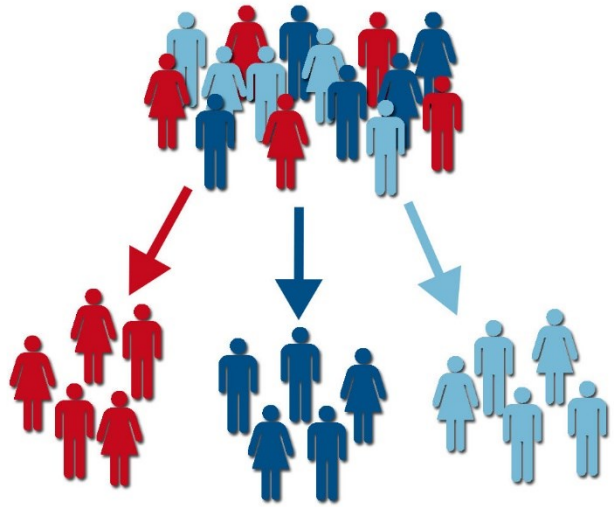
- Put the right products on every shelf at every outlet to satisfy ever-evolving customer demands
- Compare segment portfolios to ensure the distribution and market penetration of SKUs is optimized.
- CPG companies can identify which markets are right for each product, identify which products are winning in their markets and which products are due for retirement.

Media Mix Modeling (MMM)

- Create an ideal campaign that will drive engagements and sales
  - Measure impact of marketing and advertising campaigns to determine how various elements contribute to conversion.
  - Understand trends - seasonality, weather, holidays, brand authority, external influencers etc.
- Data-driven attribution
  - multi-touch attribution - tracks engagements throughout the consumer journey.

Marketing - Customer Segmentation

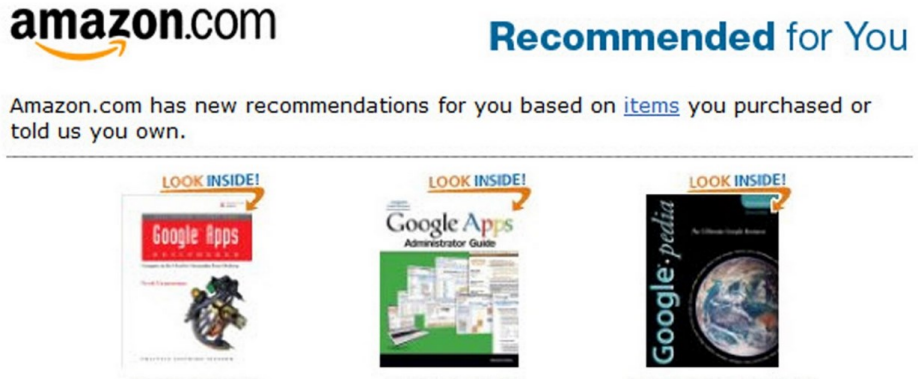
- Micro-segmentation
  - Provides a better overview of the industry/market
  - Seeks to identify the customers characteristics: where they are, who they are, how they live, and how they buy
  - Enables marketing on a granular level with personalized, customized messages
- Realtime Dynamic Segmentation
  - Track customer segments as they evolve
  - Activity-based data – website tracking information, purchase histories, call center data, mobile data, response to incentives
  - Social influence/sentiment data – product/company associations (e.g. likes or follows), online comments and reviews, customer service records

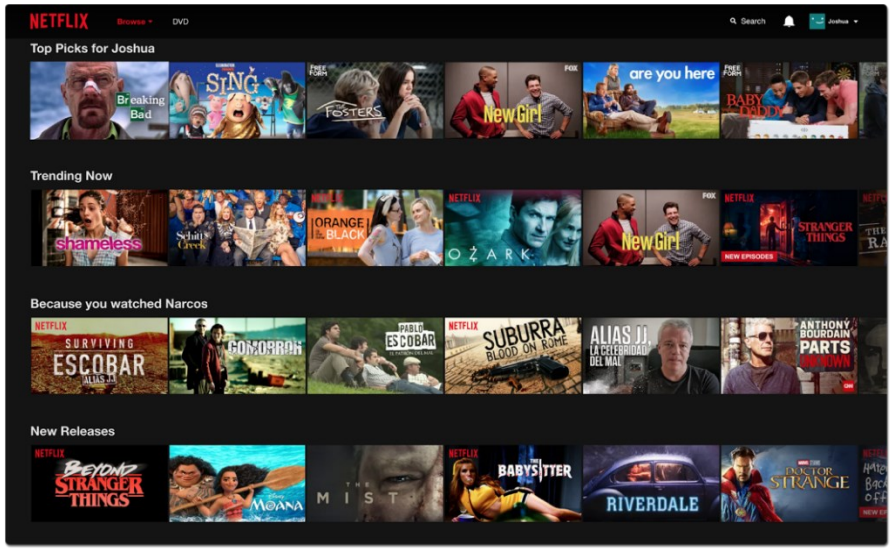


https://blog.reachforce.com/blog/how-big-data-is-changing-the-face-of-market-segmentation
https://www.fullsurge.com/resources/customer-segmentation-in-the-age-of-big-data

Recommender Systems

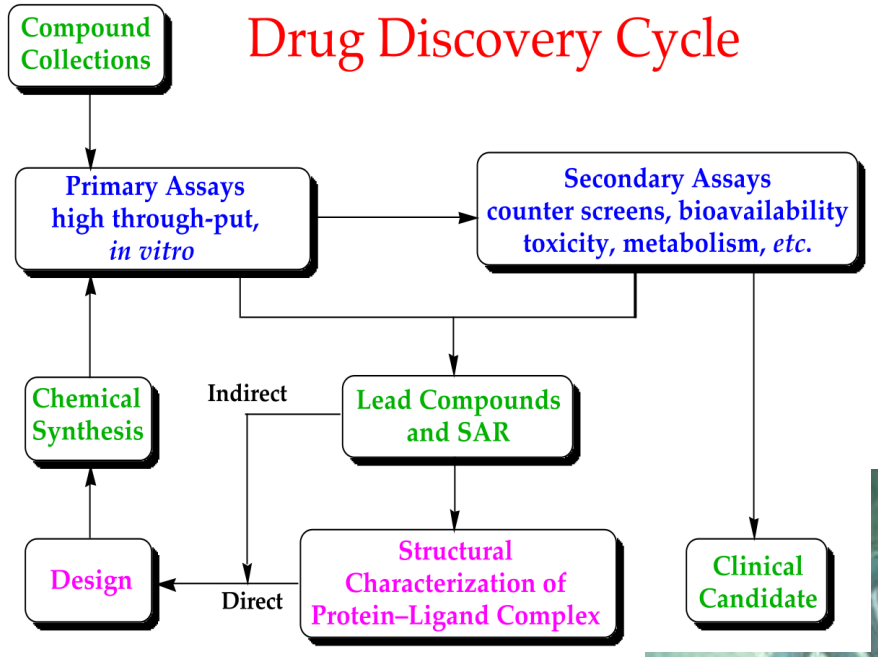
- Recommender Systems aim to help a user or a group of users to select items from a crowded item or information space.
- Types of Recommender Systems
  - Most Popular Items
  - Association and Market Basket Models
  - Content Filtering
  - Collaborative Filtering
  - Hybrid Models

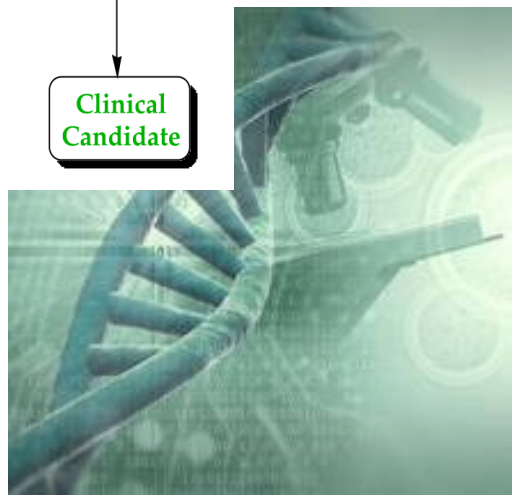




Biotechnology and Genomics

- Genomics
  - Decoding an entire genome cost around $10 million in 2007; today $1,000 per genome (NIH)
  - Millions of people can have their genomes sequenced
- Drug - Discovery, Recycling, Safety, Fraud
  - Automated screening of millions of compounds for test in preclinical trials
  - Models needed to analyze massive virtual libraries of compounds that amount to terabytes and potentially petabytes of data

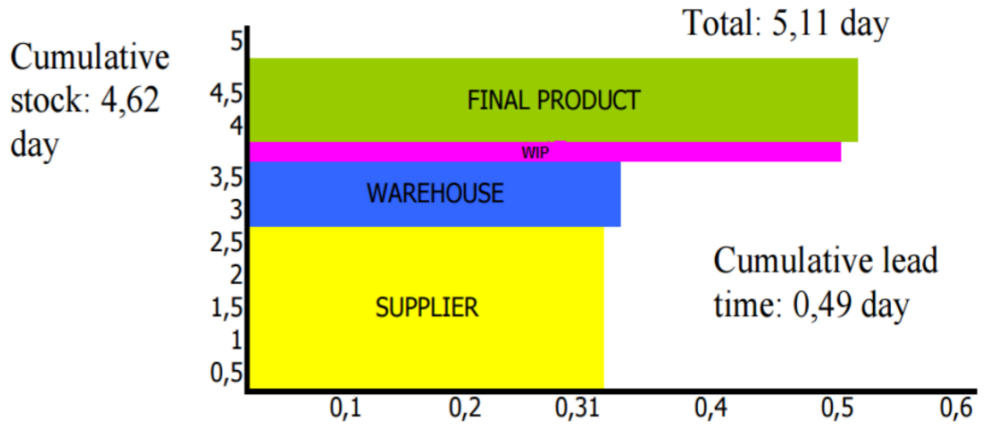




https://www.fiercebiotech.com/special-report/10-reasons-why-biotech-needs-big-data

Supply Chain Analytics

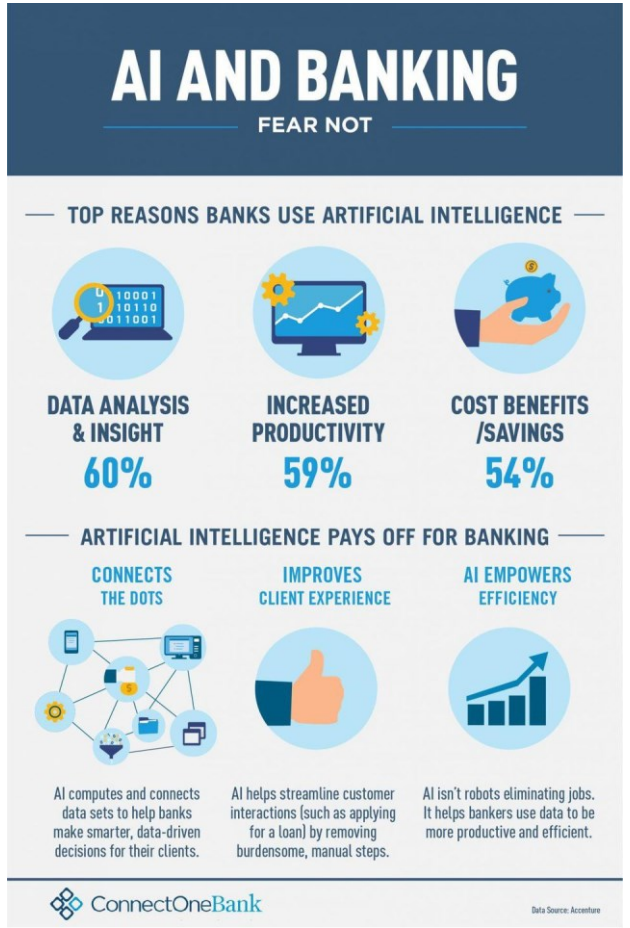
- Demand forecasting
- Inventory planning and management
  - Threshold, replenishment etc..
- Routing and optimization
- Quality management
  - IoT and Sensor Analytics
  - Equipment failure
- Space optimization



SC RESPONSE MATRIX (SOURCE (Alaca & Ceylan 2011))

Banking

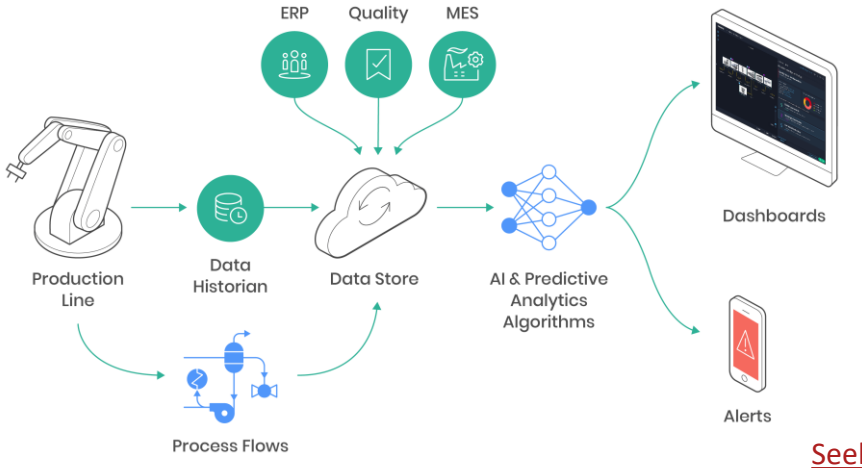
- Several banks have replaced older statistical modeling approaches with machine learning techniques and in some cases, experienced:
  - 10 % increases in sales of new products
  - 20 % savings in capital expenditures
  - 20 % increases in cash collections
  - 20 % declines in churn.
- Devised new recommendation engines for clients in retailing and in small and medium-sized companies.
- Built microtargeted models that more accurately forecast who will cancel service or default on their loans, and how best to intervene.



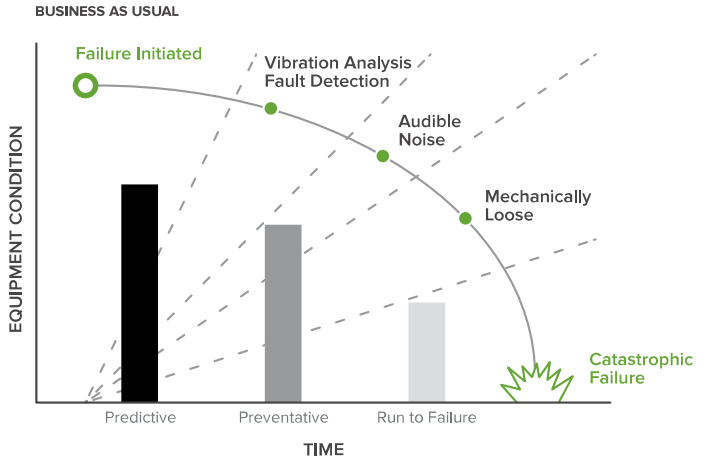
https://www.mckinsey.com/industries/high-tech/our-insights/an-executives-guide-to-machine-learning

Asset Management - Predictive Maintenance

- Comprehensive picture of machine health
  - Monitor the location and health of assets across your construction sites, factories, etc. in one dashboard.
- Increase availability of mission-critical assets
  - Reduce unplanned downtime. Flag equipment that needs repair and take proactive measures to avoid breakdowns.
- Improve key service processes.
  - Arm technicians with the insights they need to arrive on site prepared to complete jobs quickly. Reduce administrative burdens, ensuring more time is spent on revenue-generating activities.



Seebo



Splunk