Machine Learning with Python

Session 1: Introduction to Machine Learning, Data Exploration and Data Visualization using Python

Arghya Ray

What is machine learning?

Machine learning is the science (and art) of programming computers so that they can learn from the data.

“Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed.”

- Arthur Samuel, 1959

“A computer program is said to learn from experience **E** with respect to some task **T** and some performance measure **P**, if its performance on **T**, as measured by **P**, improves with experience **E**.” - Tom Mitchell, 1997

Why use Machine Learning?

Machine Learning is great for:

- Problems for which existing solutions require a lot of hand tuning or long lists of rules: *One Machine Learning algorithm can often simplify code and perform better.*
- Complex problems for which there is no good solution at all using a traditional approach: *the best Machine Learning techniques can find a solution.*
- Fluctuating environments: *a Machine Learning system can adapt to new data.*
- Getting insights about complex problems and large amount of data.

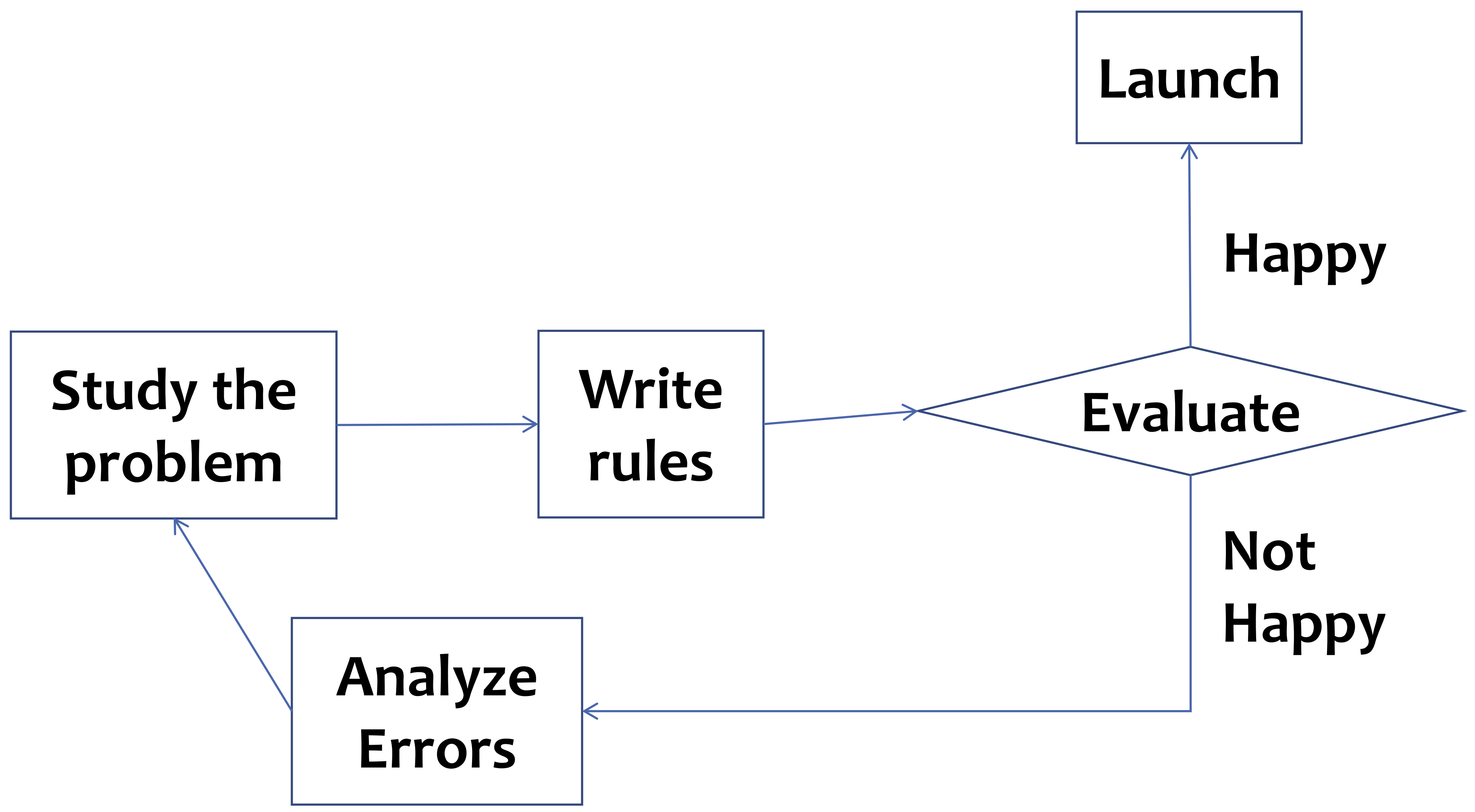


Figure 1. The traditional approach

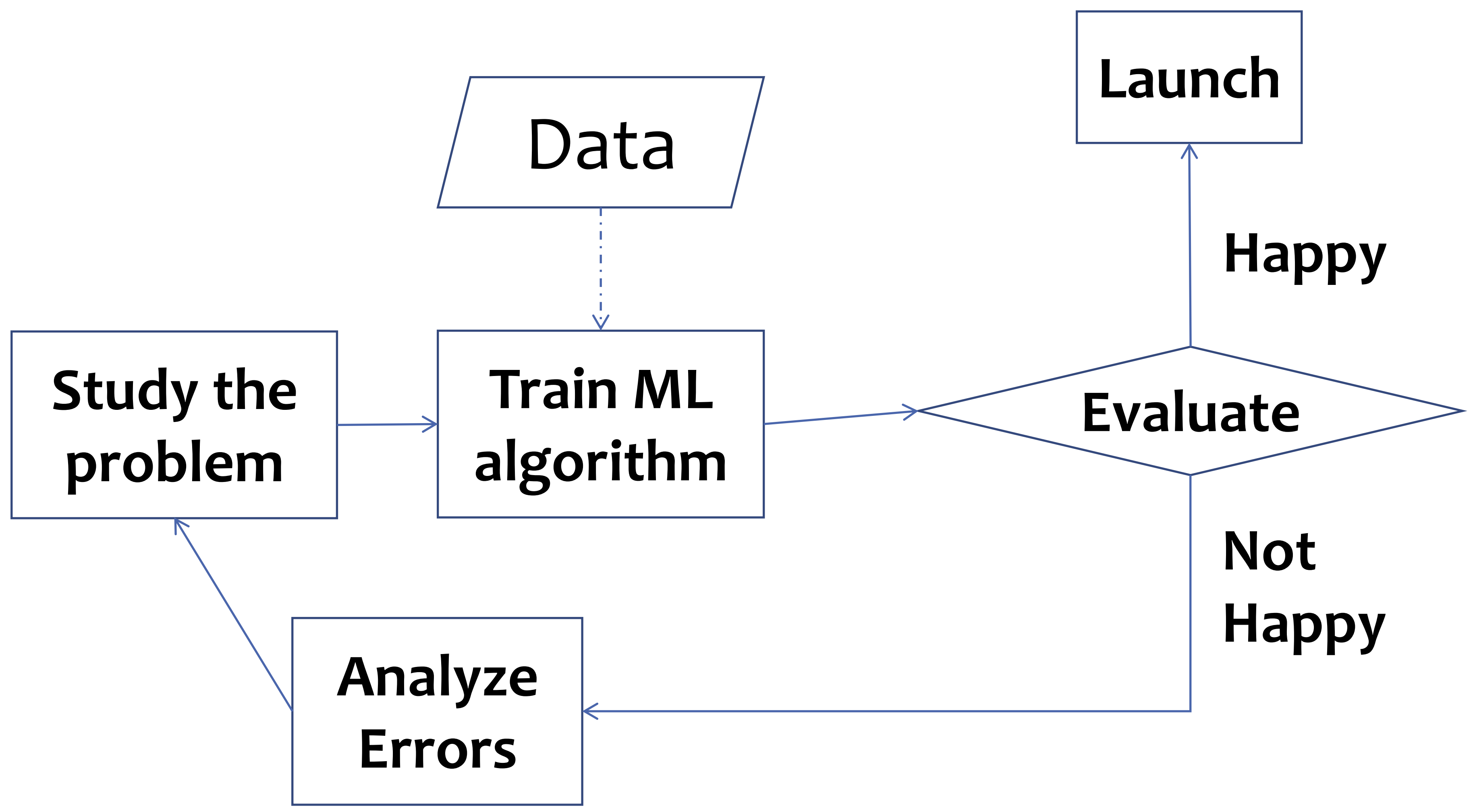


Figure 2. The Machine learning approach

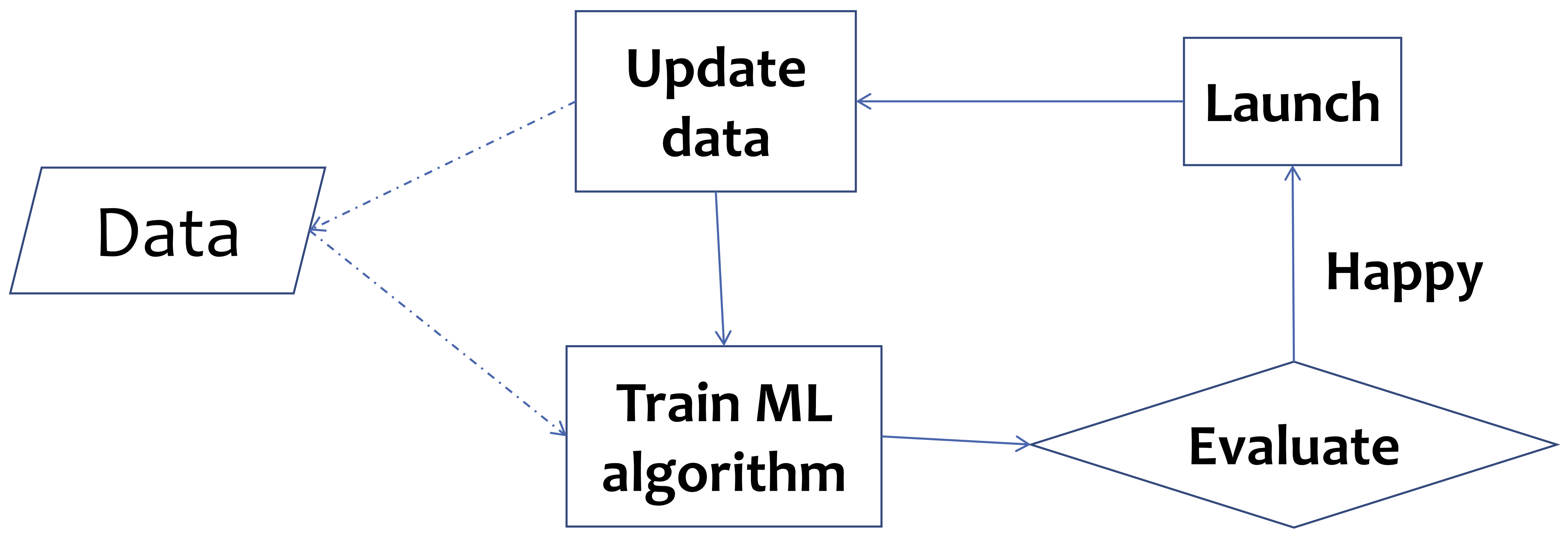


Figure 2. The machine learnng approach

Data Mining is a collection of techniques for efficient discovery of previously unknown, valid, novel, useful and understandable patterns in large databases. The patterns must be actionable so that they may be used in an enterprise's decision making.

Why do we need data-mining now?

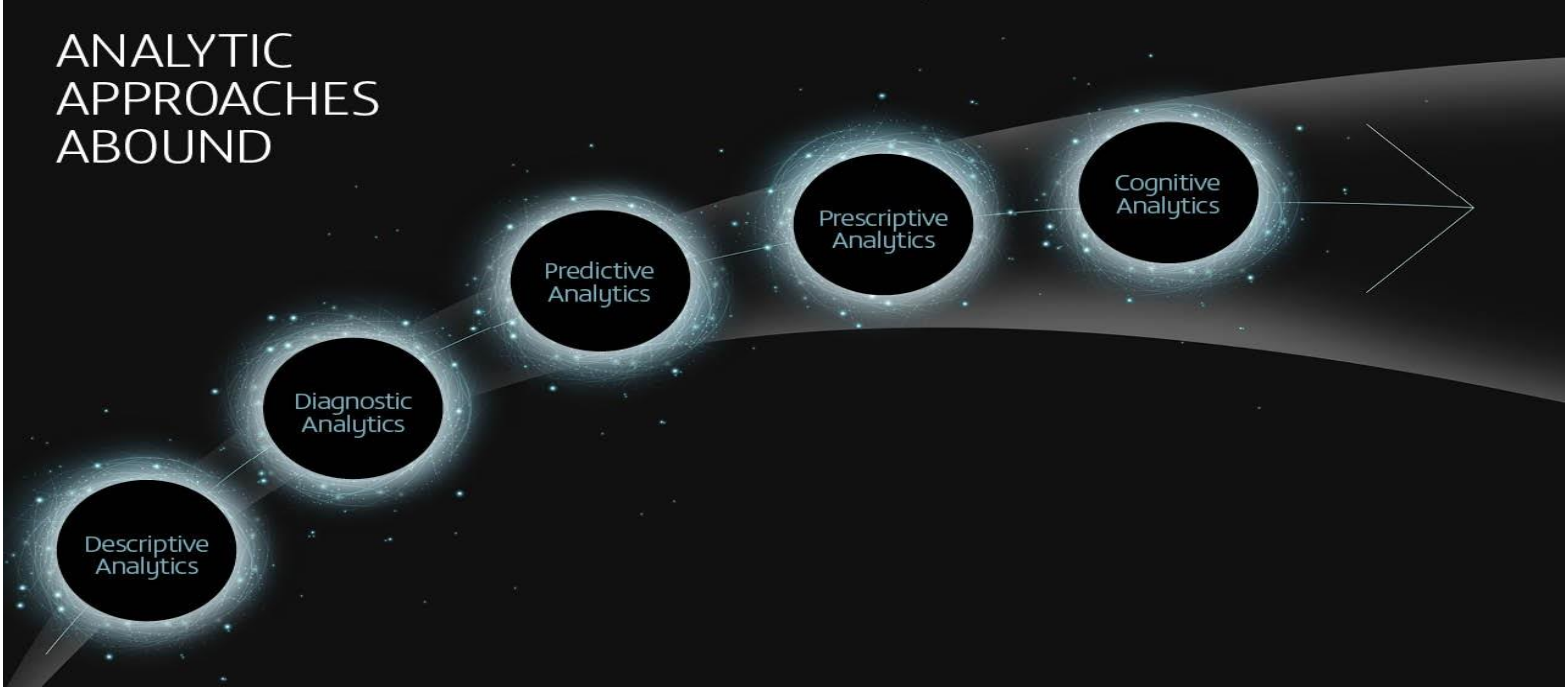
- Growth in data
- Decline in cost of processing
- Growth in data storage capacity
- Competitive environment
- Availability of various data-mining softwares

Data-Mining Applications:

- Prediction and Description
- Relationship Marketing
- Customer Profiling and Customer Segmentation
- Outlier Identification and Fraud Detection

Domains where data-mining is used:

- Astronomy
- Banking and Finance
- Business
- Crime Prevention
- Education
- Government
- Health-care
- Manufacturing
- Telecommunications
- Transportation



ANALYTIC APPROACHES ABOUND

Data Analytics has evolved over the years from **Descriptive** (*what has happened*) to **Diagnostic** (*why did it happen*) to **Predictive** (*what could happen*) to **Prescriptive** (*what action could be taken*).

The next big paradigm shift will be towards **Cognitive Analytics** which will exploit the massive advances in High Performance Computing by combining advanced Artificial Intelligence and Machine Learning techniques with data analytics approaches.

Descriptive Analytics

Descriptive analytics is the interpretation of historical data to better understand changes that have occurred in a business.

E.g., Year over year pricing changes, month over month sales growth, etc.

Predictive Analytics

“The purpose of predictive analytics is not to tell you what will happen in future. It cannot do that. In fact no analytics can do that. Predictive Analytics can only forecast what might happen in future, because all predictive analytics are probabilistic in nature,” – Dr. Michael Wu, Lithium Technologies.

The three keystones of predictive analytics are: (a) decision analysis and optimization; (b) transactional profiling; (c) predictive modeling.

Predictive Analytics exploits patterns in transactional and historical data to identify risks and opportunities,

Prescriptive Analytics

Prescriptive Analytics is an emerging discipline and represents a more advanced use of predictive analytics.

Prescriptive analytics goes beyond simply predicting options in the predictive model and actually suggests a range of prescribed actions and the potential outcomes for each action.

Dr. Wu said that “Since a prescriptive model is able to predict the possible consequences based on different choices of action, it can also recommend the best course of action for any pre-specified outcome.”

E.g., 1. Google’ self driving car

2. In the energy sector, gas producers and pipeline companies use prescriptive analytics to identify factors affecting the price of oil and gas.

Business Intelligence and Business Analytics:

In Industry, business intelligence is using BI tools to get some information. The data is used form data warehouse to get some answers from certain queries or generate reports. BI uses historical data.

E.g.: As far as banking loans are concerned, how does a customer behave – before marriage and after marriage.

What is the purchase pattern of customers on weekends in various cities.

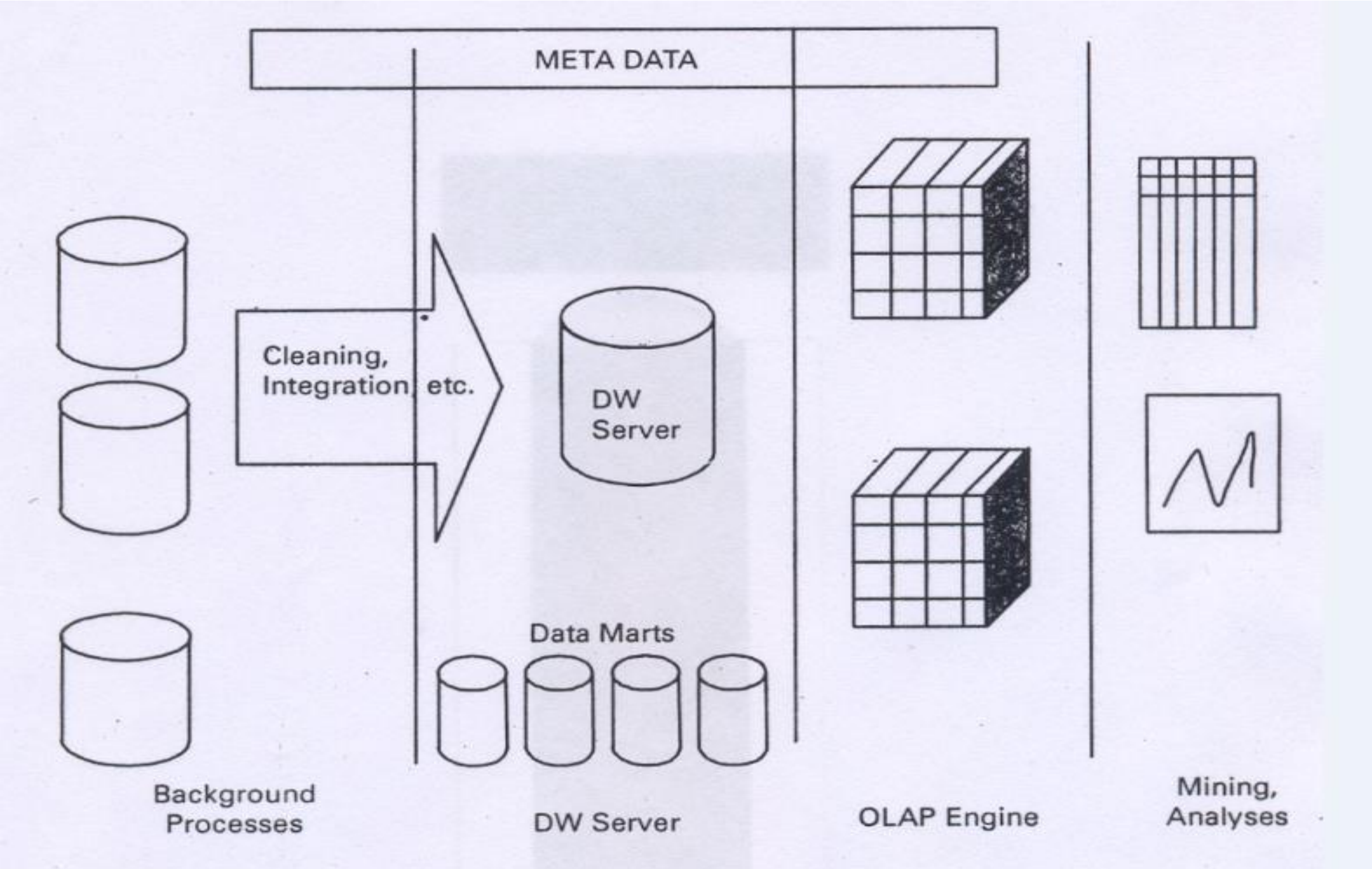
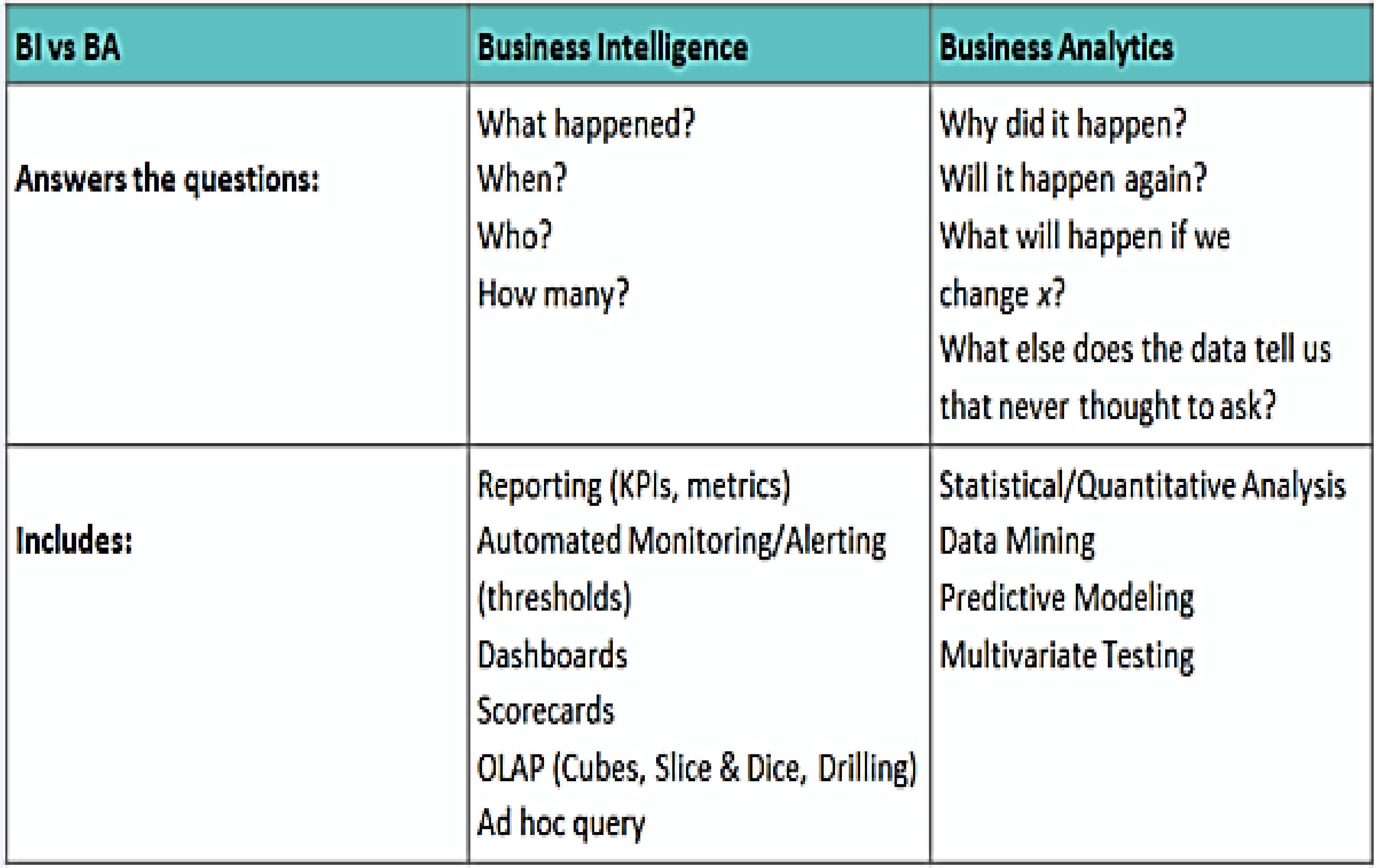


Figure 2.12 Data Warehouse Architecture



BI vs BA	Business Intelligence	Business Analytics
Answers the questions:	What happened? When? Who? How many?	Why did it happen? Will it happen again? What will happen if we change x? What else does the data tell us that never thought to ask?
Includes:	Reporting (KPIs, metrics) Automated Monitoring/Alerting (thresholds) Dashboards Scorecards OLAP (Cubes, Slice & Dice, Drilling) Ad hoc query	Statistical/Quantitative Analysis Data Mining Predictive Modeling Multivariate Testing

Difference between predictive analytics and business intelligence:
Business Intelligence answers the question, "From what ZIP code does my most valuable customers come?"
Predictive analytics however answers, "How much revenue can I expect from customers in a particular ZIP code?"

Types of Machine Learning Systems

Broad Categories of Machine Learning Systems:

- ***Whether or not they are trained with human supervision***
    - Supervised
    - Unsupervised
    - Semi-supervised
    - Reinforcement
- ***Whether or not they can learn incrementally on the fly***
    - Online Learning
    - Batch Learning
- ***Whether they work by simply comparing new data points to known data points, or instead detect patterns in the training data and build a predictive model***
    - Instance based learning
    - Model based learning

Machine Learning systems can be classified according to the amount and type of supervision they get during training.

Supervised Learning:

Goal: Predict a single “target” or “outcome” variable.

- Has definite outcomes or goals.
- Predict answers for new unknown values

Training data, where target value is known

Score to data where value is not known

Methods: Classification and Prediction

Unsupervised Learning:

Goal: Segment data into meaningful segments; detect patterns.

- No target (outcome) variable to predict or classify
- It makes sense of data from observations.

Methods: Association rules, data reduction & exploration, visualization, Anomaly detection

Semi-Supervised Learning:

Some algorithms can deal with partially labeled training data, usually a lot of unlabeled data and a little bit of labeled data.

E.g.: Some photo-hosting services, such as Google Photos, are good examples of this.

Reinforcement Learning:

The learning system, called an agent in this context, can observe the environment, select and perform actions, and get rewards in return (or penalties in the form of negative rewards). It must then learn by itself what is the best strategy, called a policy, to get the most reward over time.

Machine Learning systems can be classified based on whether or not the system can learn incrementally from a stream of incoming data.

Batch Learning/Offline Learning:

- In batch learning, the system is incapable of learning incrementally: it must be trained using all the available data. This will generally take a lot of time and computing resources, so it is typically done offline.
- First the system is trained, and then it is launched into production and runs without learning anymore; it just applies what it has learned.
- If you have a lot of data and you automate your system to train from scratch every day, it will end up costing you a lot of money. If the amount of data is huge, it may even be impossible to use a batch learning algorithm.

Incremental Learning/Online Learning:

- In online learning, you train the system incrementally by feeding it data instances sequentially, either individually or by small groups called mini-batches.
- One important parameter of online learning systems is how fast they should adapt to changing data: this is called the ***learning rate***.
- If you set a high learning rate, then your system will rapidly adapt to new data, but it will also tend to quickly forget the old data (you don’t want a spam filter to flag only the latest kinds of spam it was shown).
- Conversely, if you set a low learning rate, the system will have more inertia; that is, it will learn more slowly, but it will also be less sensitive to noise in the new data or to sequences of non-representative data points.
- If bad data is fed into the system, the system’s performance will gradually decline.

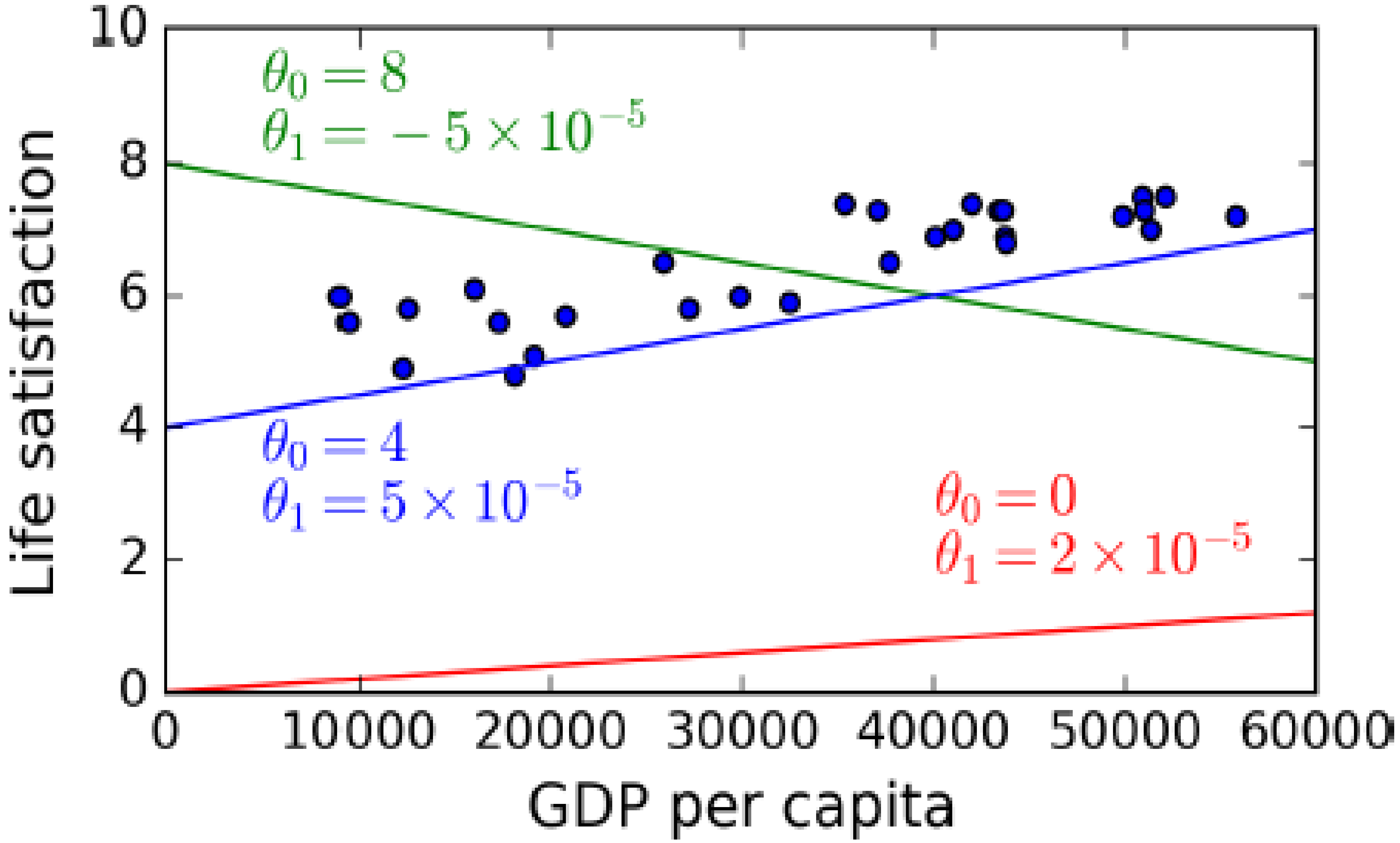
One more way to categorize Machine Learning systems is by how they generalize.
Having a good performance measure on the training data is good, but insufficient; the true goal is to perform well on new instances.

Instance-based Learning:

- Possibly the most trivial form of learning is simply to learn by heart.
- If you were to create a spam filter this way, it would just flag all emails that are identical to emails that have already been flagged by users—not the worst solution, but certainly not the best. Instead of just flagging emails that are identical to known spam emails, your spam filter could be programmed to also flag emails that are very similar to known spam emails. This requires a measure of similarity between two emails.

Model-based Learning:

- Another way to generalize from a set of examples is to build a model of these examples, then use that model to make predictions..



Supervised vs Unsupervised Learning

- Deciding whether to issue a loan to an applicant based on demographic and financial data (with reference to a database of similar data on prior customers).
- In an online bookstore, making recommendations to customers concerning additional items to buy based on the buying pattern in prior transactions.
- Identifying a network data packet as dangerous (virus, hacker attack) based on comparison to other packets whose threat status is known.
- Identifying segments of similar customers.
- Predicting whether a company will go bankrupt or not based on comparing its financial data to those of similar bankrupt and non-bankrupt firms.
- Automated sorting of mail by zip-code scanning.
- Estimating the repair time required for an aircraft based on a trouble ticket.
- Printing of custom discount coupons at the conclusion of a grocery store checkout based on what you just bought and what others have bought previously.

Main Challenges of Machine Learning:

- Insufficient quantity of training data
- Non-representative training data → sampling noise. Sampling bias
- Poor quality data → standard format, 5,10,15,30, 225,
- Irrelevant features → P, Tw, GN, ExS (Feature Selection, Feature extraction)
- Overfitting the training data (it means that the model performs well on the training data, but it does not generalize well)
  - Overfitting happens when the model is too complex relative to the amount and noisiness of the training data
  - Constraining a model to make it simpler and reduce the risk of overfitting is called **regularization**.
  - Possible solutions:
    - To simplify the model by selecting one with fewer parameters, by reducing the number of attributes in the training data or by constraining the models.
    - To gather more training data
    - To reduce the noise in the training data (e.g., fix data errors and remove outliers)
- Underfitting the training data (it occurs when your model is too simple to learn the underlying structure of the data)
  - Possible solutions:
    - Selecting a more powerful model, with more parameters
    - Feeding better features to the learning algorithms (feature engineering)
    - Reducing the constraints on the model (e.g., reducing the regularization hyperparameters)

Testing and Validating

The content of the slides are prepared from different textbooks.

References:

- Links:
  - https://www.sas.com/en_in/insights/big-data/what-is-big-data.html
  - https://www.oracle.com/big-data/what-is-big-data/
  - https://www.w3schools.com/python/python_variables_multiple.asp

- Predictive Analytics for Dummies, By Anasse Bari, Mohamed Chaouchi, & Tommy Jung, Copyright 2016, John Wiley & Sons, Inc.
- Introduction to Data Mining with Case Studies, By G.K. Gupta. Copyright 2014 by PHI Learning Private Limited.

Thank you..