IN3020/IN4020 – Database Systems Spring 2021, Week 18.2

Data Science:
The bridge between data & science

M. Naci Akkøk, CEO, In-Virtualis, Assoc. Prof. UiO/Ifi (ASR), Assoc. Prof. OsloMet/CEET (Engineering)

Egor Kostylev, Assoc. Prof, UiO/Ifi, Analytical Solutions & Reasoning (ASR) research group

Sagar Sen, Senior Research Scientist, SINTEF Digital

Renée Wikestad, Principal Cloud Engineer & Specialist, Oracle

UiO : Institutt for informatikk
Det matematisk-naturvitenskapelige fakultet

Introduction

Key question – Why/how of newer DBMS´

When new types of applications become main-line, like data science or AI/ML applications or social media applications or IoT etc., they impose new requirements upon the underlying systems, like the infrastructure, and, of course, the DBMS.

How do the new generation DBMS´ answer the needs of data science?

The program

- Introduction (M. N. Akkøk, UiO/Ifi + OsloMet) – 10 min.
- Graph Neural Networks (E. Kostylev, UiO/Ifi) – 20 min.
- Explainable AI (S. Sen, SINTEF) – 20 min.
- AI/ML and the DBMS in practice (R. Wikestad, ORACLE) – 20 min.

- For discussion:
  What does Data Science require from DBMS´? – 20 min.

A very simple introduction to data science

Insight to the data science techniques, including AI/ML

UiO : Institutt for informatikk

Det matematisk-naturvitenskapelige fakultet

Statistics – the common base

- Understanding statistical techniques, how they relate to each other and where they are used
    - Mean, variance, standard deviation
    - Curve fitting and simple prediction
    - Anomalies, clustering (K-means), classification

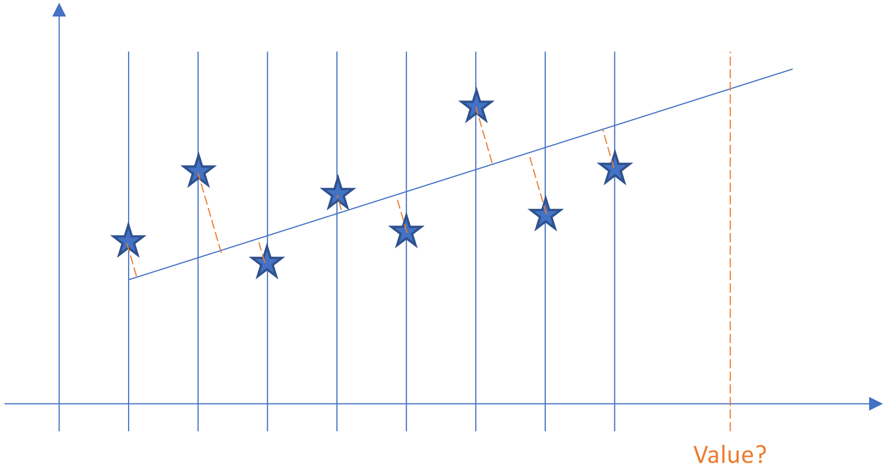
Statistics – towards data science

- Mean, variance, standard deviation
- Curve fitting and simple prediction
- Anomalies, clustering (K-means), classification

Mean. The average of all data-points.
Variance. A measure of the degree to which each data point deviates from the **mean**. A measure of the average distance from the average ☺
Standard deviation. Square root of the **variance**.
An indication the spreading of numbers from the **mean**.

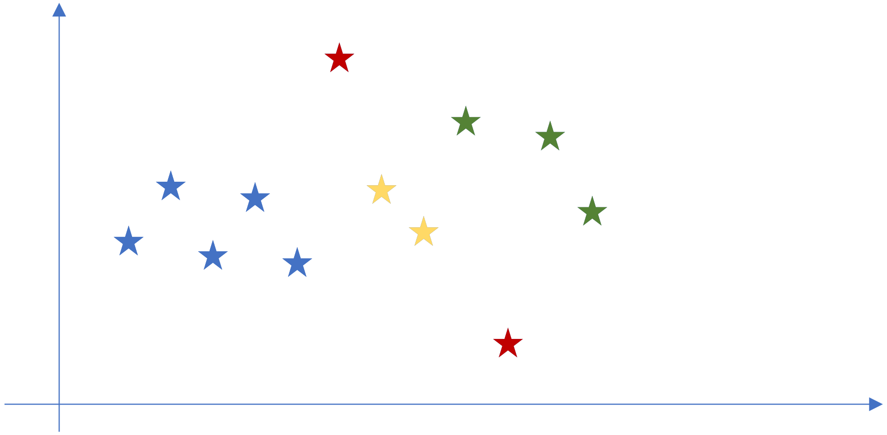
Statistics – towards data science

- Mean, variance, standard deviation
- Curve fitting and simple prediction
- Anomalies, clustering (K-means), classification

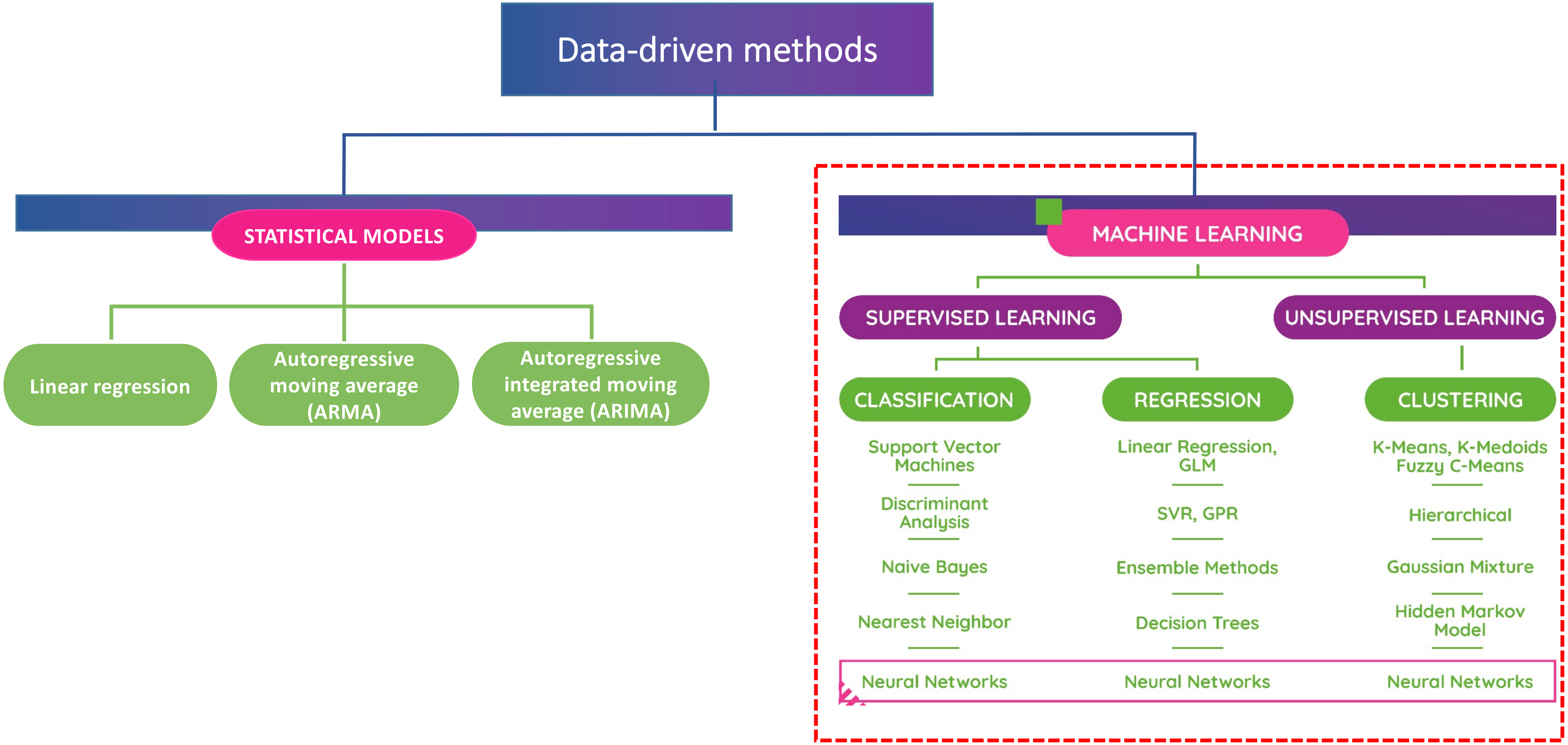


Statistics – towards data science

- Mean, variance, standard deviation
- Curve fitting and simple prediction
- Anomalies, clustering (K-means), classification



Taxonomy of data-driven methods



Data-driven methods

STATISTICAL MODELS

Linear regression

Autoregressive moving average (ARMA)

Autoregressive integrated moving average (ARIMA)

MACHINE LEARNING

SUPERVISED LEARNING

UNSUPERVISED LEARNING

CLASSIFICATION

- Support Vector Machines
- Discriminant Analysis
- Naive Bayes
- Nearest Neighbor
- Neural Networks

REGRESSION

- Linear Regression, GLM
- SVR, GPR
- Ensemble Methods
- Decision Trees
- Neural Networks

CLUSTERING

- K-Means, K-Medoids Fuzzy C-Means
- Hierarchical
- Gaussian Mixture
- Hidden Markov Model
- Neural Networks

REF.: Dr. Hossain Tabari, Research Scientist, KU Leuven, Department of Civil Engineering

Supervised vs. unsupervised learning

Supervised learning

Data: (x, y)

x is data, y is label.

Goal: Learn a function to map $x \longrightarrow y$

Use cases: Classification, regression, object detection, etc.

Unsupervised learning

Data: (x)

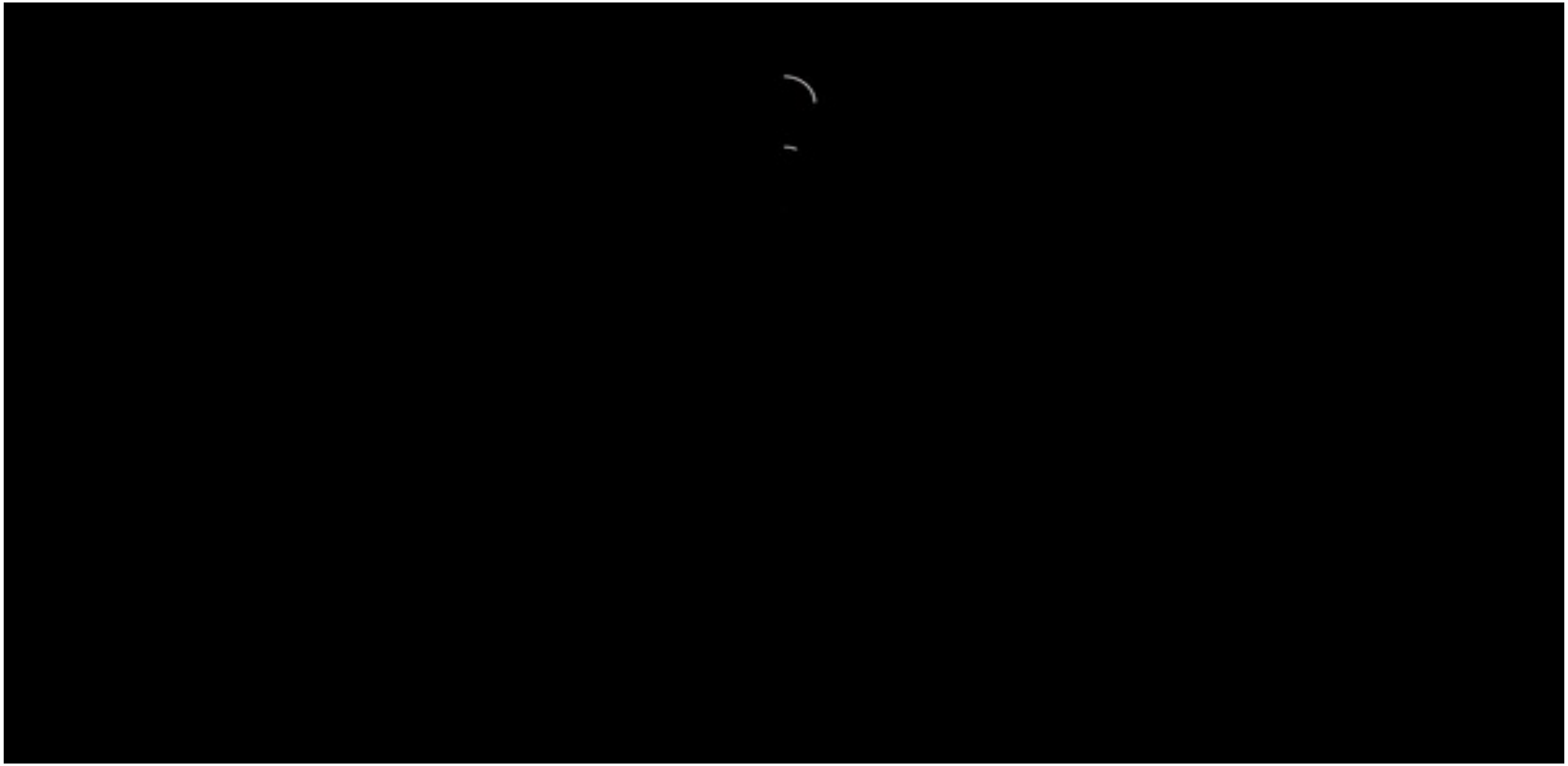
x is data, no labels!

Goal: Learn some hidden or underlying structure of the data

Use cases: Clustering, feature or dimensionality reduction, etc.

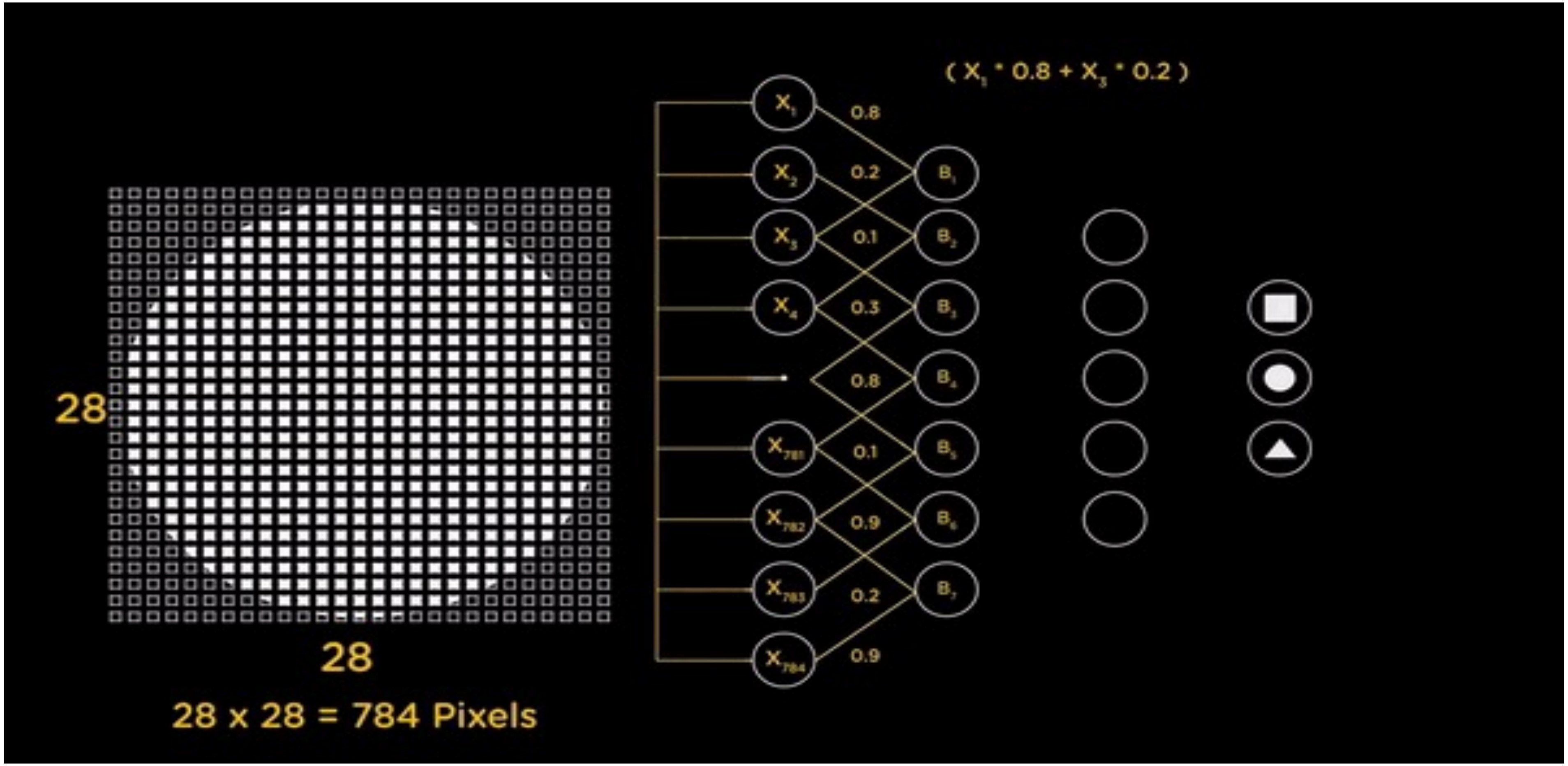
REF.: Dr. Hossain Tabari, Research Scientist, KU Leuven, Department of Civil Engineering

Artificial neural networks



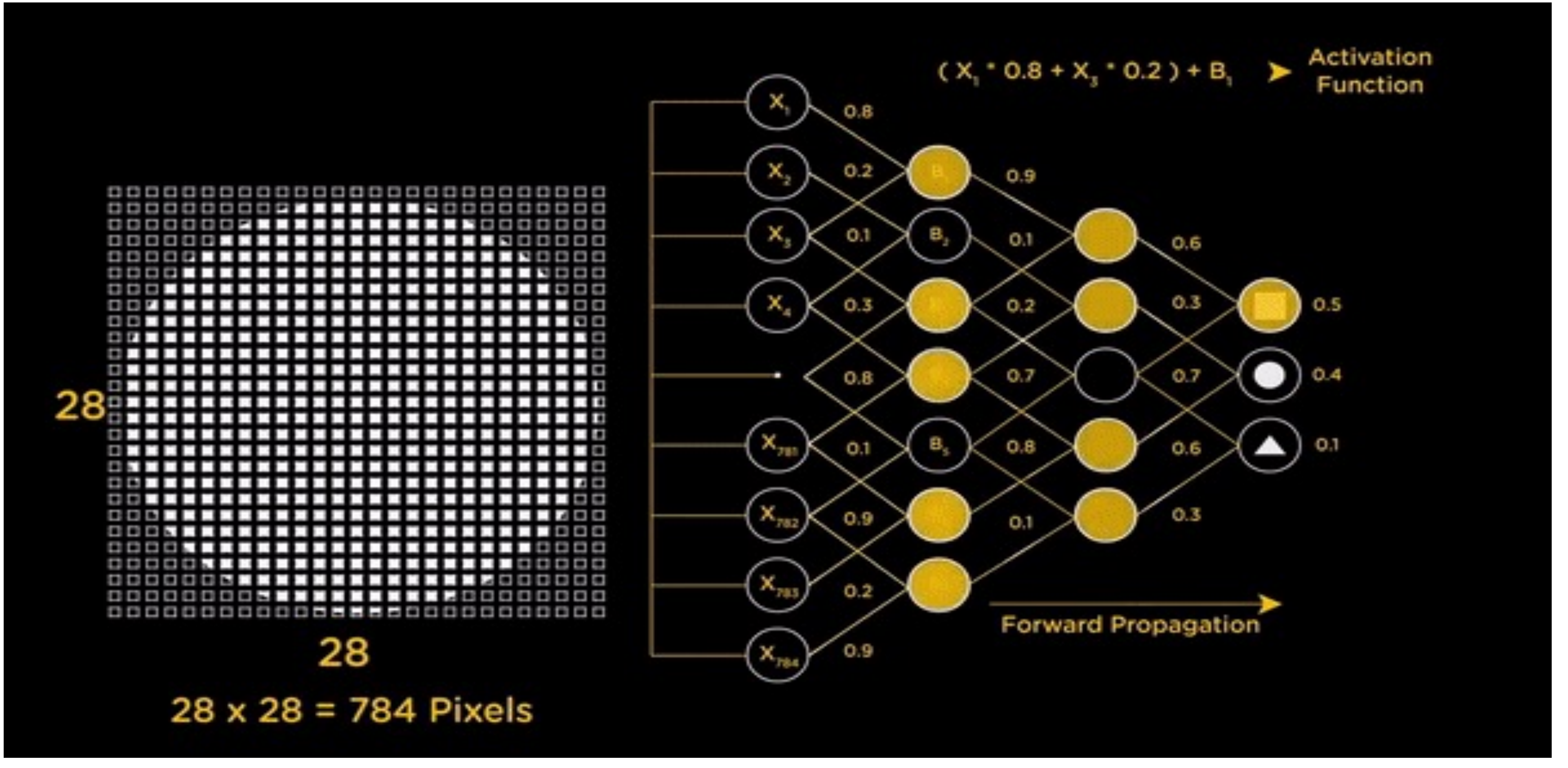
REF.: Dr. Hossain Tabari, Research Scientist, KU Leuven, Department of Civil Engineering

Artificial neural networks



REF.: Dr. Hossain Tabari, Research Scientist, KU Leuven, Department of Civil Engineering

Artificial neural networks



REF.: Dr. Hossain Tabari, Research Scientist, KU Leuven, Department of Civil Engineering

Topic for discussion: What does data science require of database management systems?

What are the characteristics of data science applications?

Characteristics of Data Science

- Data science applications are data-driven or data-intensive

- Remember Big Data characterization?

Volume	Large amounts of data (maybe not always?)
Velocity	Fast! Real time? Ingestion, access, processing?
Variety	Different sources, formats, granularity
Value	Quality, completeness, fitness-to-purpose

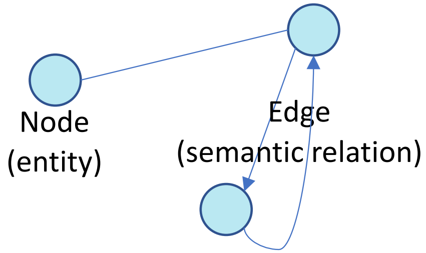
Example: Computer aided molecular design (CAMD)

- Computational molecular modeling & simulation
- Very complex and time consuming

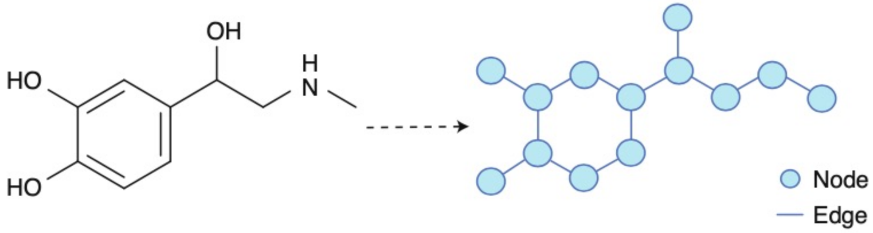
- Some optimization solutions, directions:
  - Efficient data representation: Representing molecular structure & free-energy data using (extended) property graphs
  - Using graph structure in machine learning (Egor & Sagar)
  - Processing (search/retrieval) optimization

Technology #03 (Using Extended Property Graph)

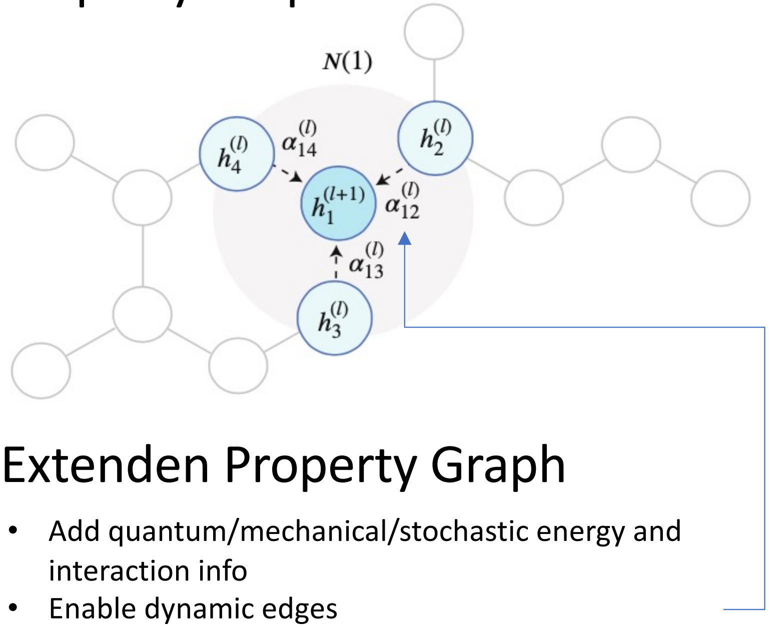
- Graph



- Graph representation of molecular structure is more natural & effective



- Property Graph



- Extenden Property Graph
  - Add quantum/mechanical/stochastic energy and interaction info
  - Enable dynamic edges
  - Also adding folding info (3D geometry),
  - and affinity info (see next slide)

REF.: Dr. M. Naci Akkøk, CEO, In-Virtualis AS

Technology #04 (Folding, Affinity & VR)

Molecules “fold” due to the distribution of pull-and-push forces.

When folded, some areas with specific surpluss “charge” (positive or negative) are exposed,

where other structures (with proper “charges”) may “dock”.

Much easier to see & refine in VR.

REF.: Dr. M. Naci Akkøk, CEO, In-Virtualis AS

Example: Computer aided molecular design (CAMD)

- Computational molecular modeling & simulation
- Very complex and time consuming – obvious need for HPC

- Some optimization solutions, directions:
  - Efficient data representation: Representing molecular structure & free-energy data using (extended) property graphs
  - Using graph structure in machine learning (Egor & Sagar)
  - Processing (search/retrieval) optimization

Topic for discussion:
This is what we briefly saw last time

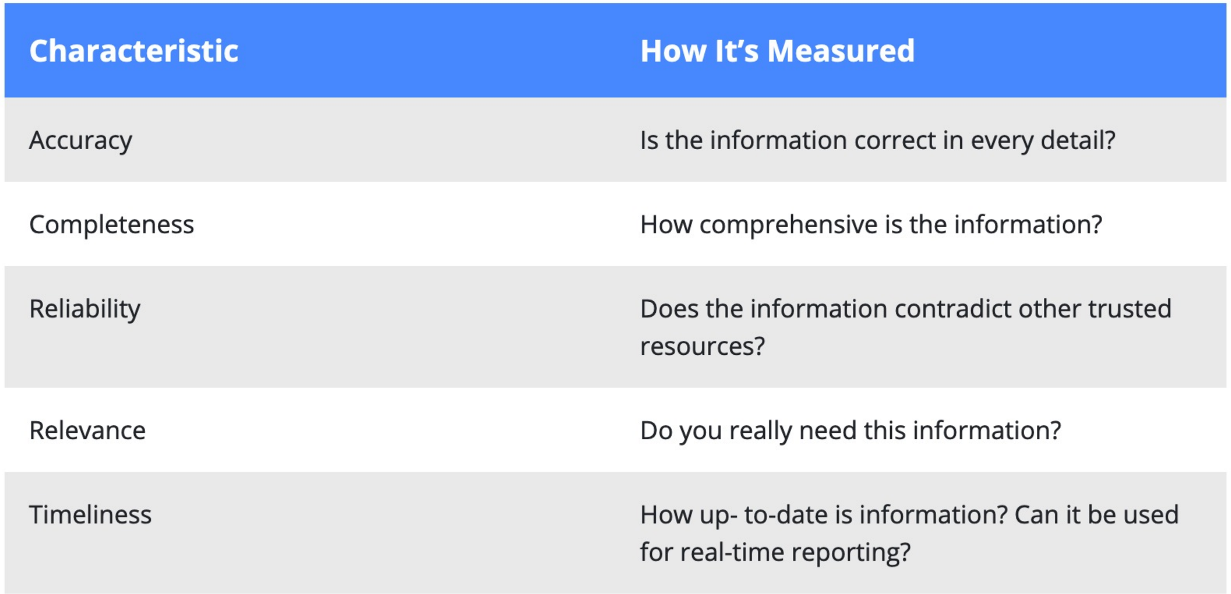
Date Quality Management (DQM)

- **Data Quality Management** is for ensuring the quality of the data (the information) we are pulling into the database or data we already have in the database.

- Any application that relies on data
  - Will need data of high quality
    (high enough with respect to its purpose)
  - And will be negatively affected by low quality data (may fail or give wrong information)

Data Quality?

Data is of high quality, if the data is **fit for the intended purpose** of use and if the data **correctly represent the real-world construct** that the data describes. *Ref. Profisee*



Characteristic	How It's Measured
Accuracy	Is the information correct in every detail?
Completeness	How comprehensive is the information?
Reliability	Does the information contradict other trusted resources?
Relevance	Do you really need this information?
Timeliness	How up- to-date is information? Can it be used for real-time reporting?

Ref. Syncsort

Content & format **CORRECTNESS**:

Key fields and other relevant fields are non-empty and are in the right format, and of the right type.

Content of field makes sense with respect to its format and expected use.

Ref. practice @ Oracle

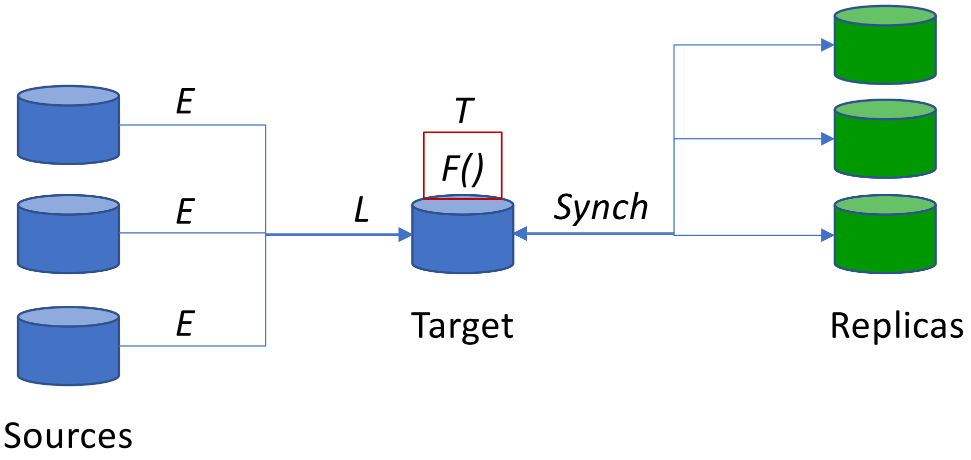
FUNCTIONS: Profiling, auditing, visualization, parsing & standardization, matching & merging , case-based clean-up, address/format verification.

Ref. EDQ sheet @ Oracle

DQM is related to ETL/ELT and DRM tools

- **Data Quality Management** is practically always related to **Extract-Transform-Load** or **Extract-Load-Transform**(*) tools and **Data Replication Management** tools.

- Typical process



(*) ELT usually faster

AI, ML and Data Science
New areas of applications that demand new types of DBMS

Yet another category of data intensive applications!

AI/ML in DQM

- One main area is introducing AI/ML to automate and improve the DQM process
  - Learn and automate profiling
  - Do the necessary corrections
  - Learn and automate transformations
  - Learn and improve ingestion performance
  - …

AI/ML in DBMS

- One main area is introducing some AI/ML into the DBMS
  - Self-install
  - Self-tune (through learned and optimized indexing etc.)
  - Self-repair
  - …

A DBMS for AI/ML

- How would you best represent a deep learning (neural) network in a DBMS?
  - Suggestion: Graph Neural Networks (Explainable AI)

- How would you design your DBMS to improve performance in the case of AI/ML and Data Science applications that require large amounts of data fast?
  - Suggestion: Embed the algorithms into the DB and let them run in the database (so that you don´t have to pull the data out first)

The Life Science Cluster

enabled by

In-Virtualis

Contact: naci@in-virtualis.com
nacia@ifi.uio.no
mehmetna@oslomet.no

+47 47026879

Join us

for the first MoMS/CAMD Seminar
June 9th at 09:00-11:00

Register at https://www.in-virtualis.com/events
(on-line, soon to be published)

UiO : Institutt for informatikk
Det matematisk-naturvitenskapelige fakultet