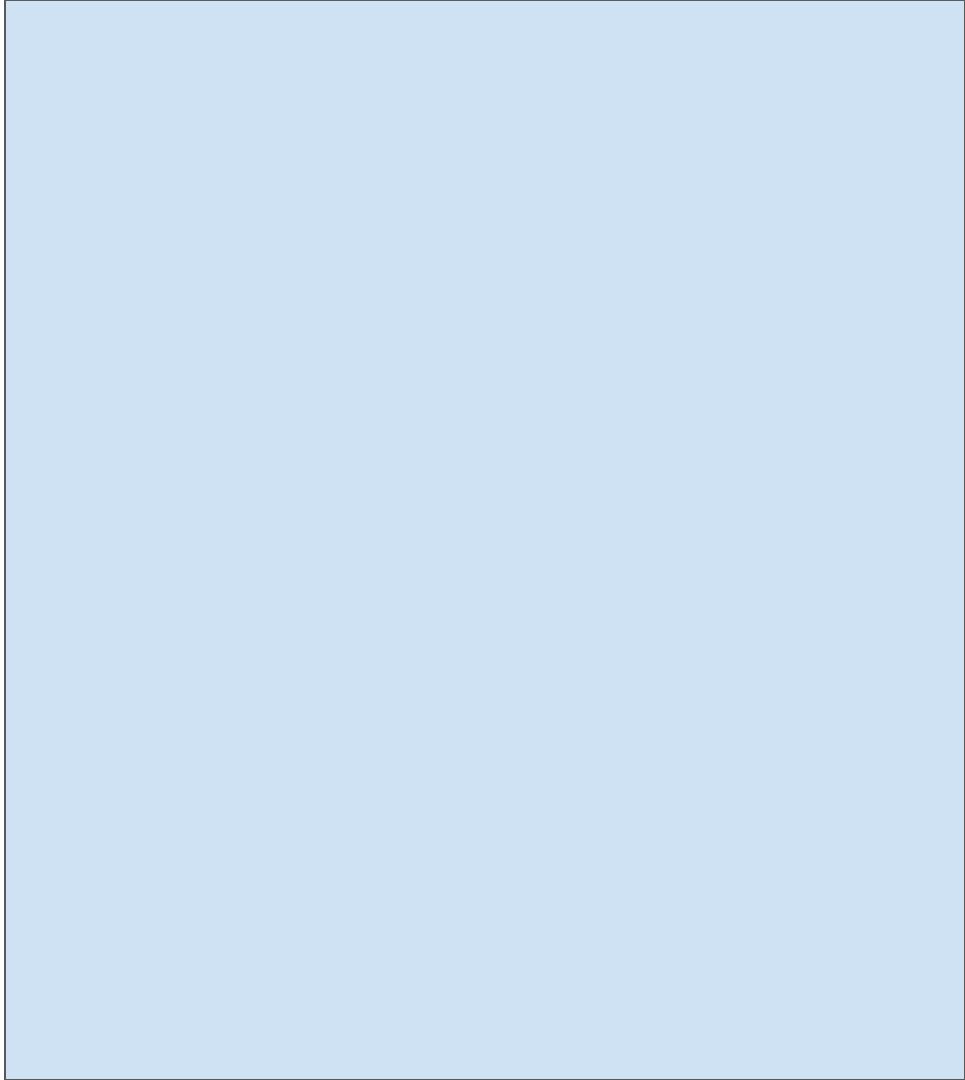
SUMMARY I

(CIi WiSe24/15-09.02-o)

Introduction

Cli WiSe24/01-20.10-x



Chemistry subject considerations

Central science

Chemistry is often referred to as the central science because it joins together physics and mathematics, biology and medicine, and the earth and environmental sciences. Knowledge of the nature of chemicals and chemical processes therefore provides insights into a variety of physical and biological phenomena.

The nature of this relationship is one of the main topics in the philosophy of chemistry and in scientometrics.

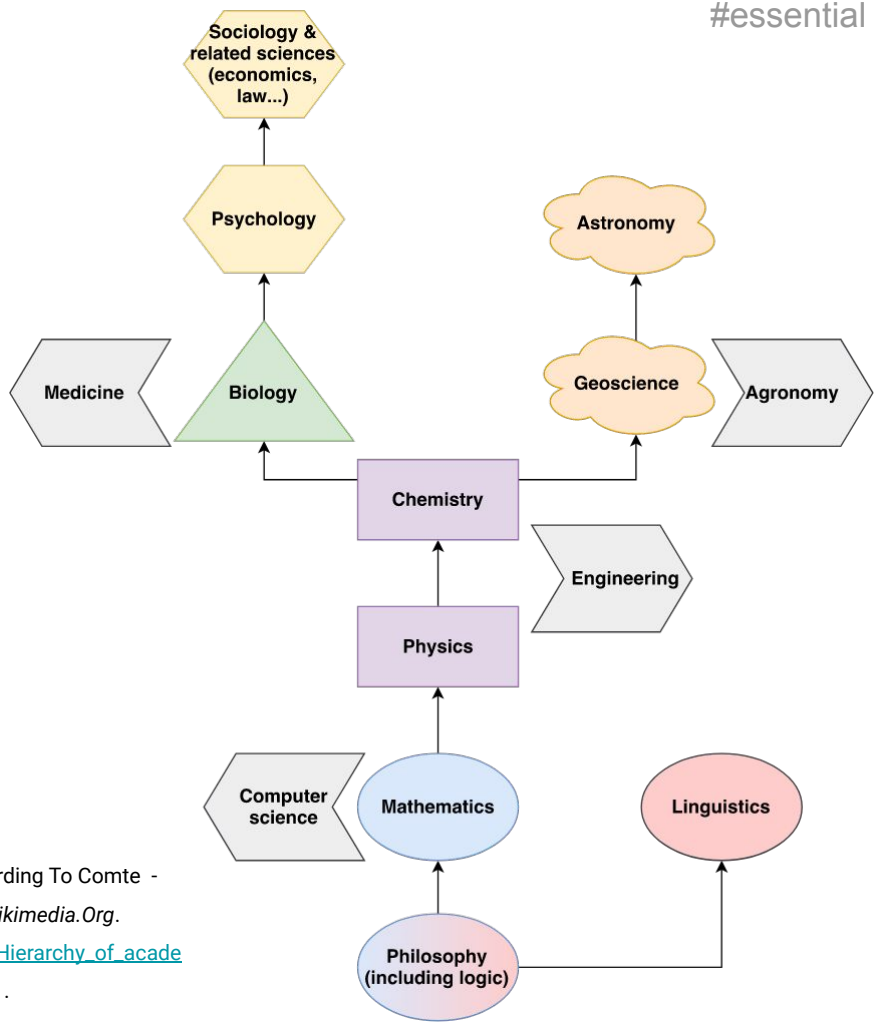
Popularized by its use in a textbook by Theodore L. Brown and H. Eugene LeMay, titled Chemistry: *The Central Science* (1977)

Chemistry, School, Current Students, and Why Chemistry. 2022. "Why Study Chemistry : School Of Chemistry". *School Of Chemistry*. https://chemistry.unimelb.edu.au/study/current-students/why-study-chemistry .

Auguste Compte (1798–1857)

Each Science came into being to seek the “Laws” of a particular level of facts which man experience in the world.

Next, each Science depends on the developments of its predecessor in a “hierarchy” or, better to say, in an “ordering” of Sciences by increasing complexity and decreasing generality.

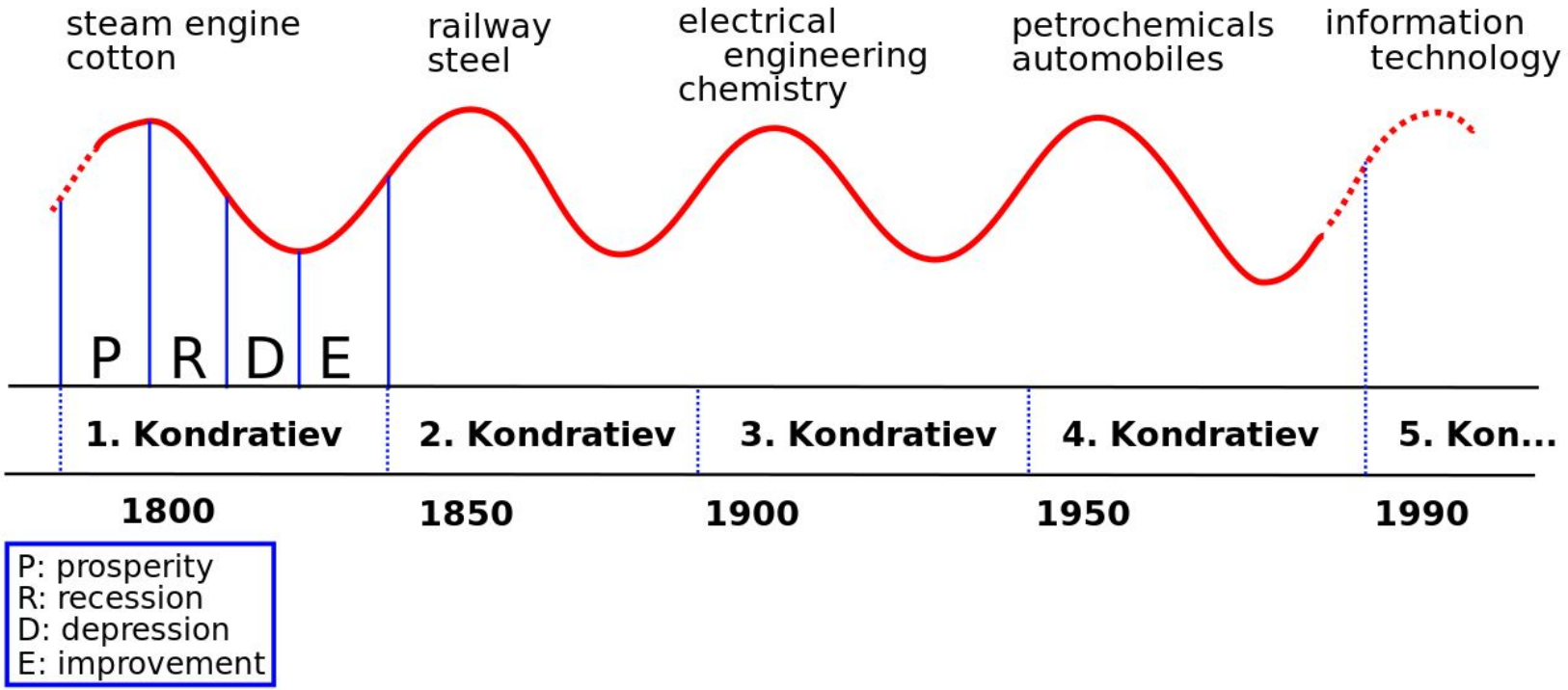


"Hierarchy Of Academics Disciplines According To Comte - Wikimedia Commons". 2017. *Commons.Wikimedia.Org*. https://commons.wikimedia.org/wiki/File:Hierarchy_of_academics_disciplines_according_to_Comte.svg .

Why a new age?

Kondratiev

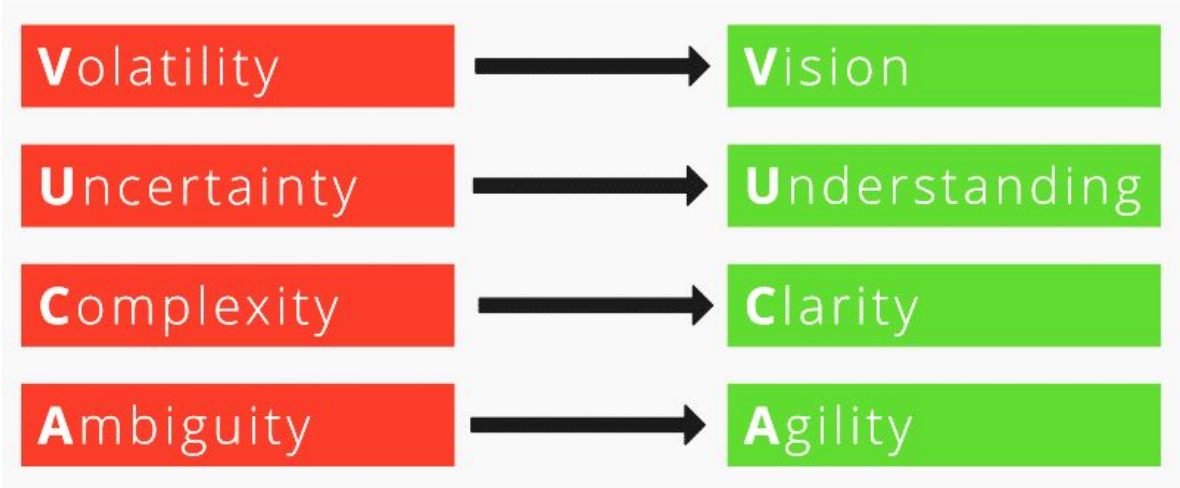
Changes follow a cycle



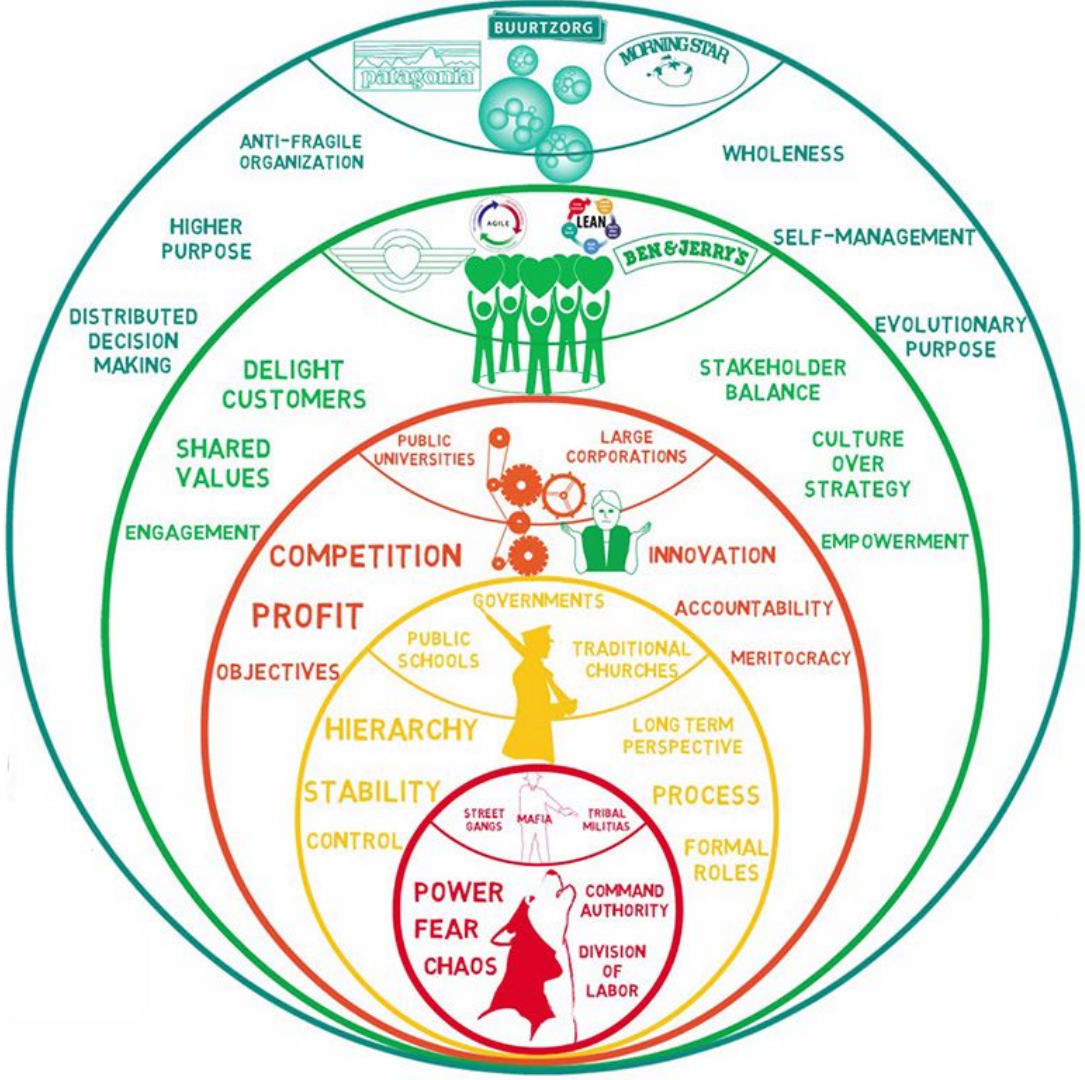
https://en.wikipedia.org/wiki/Kondratiev_wave

Challenges

- DC It is not yet a ‘ready product’, its is a evolving concept
- DC pay of at holistic scale
- VUCA circumstances
- Financing



Spiral dynamics



"Reinventing Management, Part 1: What Color Is Your Organization?". 2016. Enlivening Edge. https://enliveningedge.org/views/reinventing-management-part-1-what-color-is-your-organization/.

Digital Chemistry

Parts

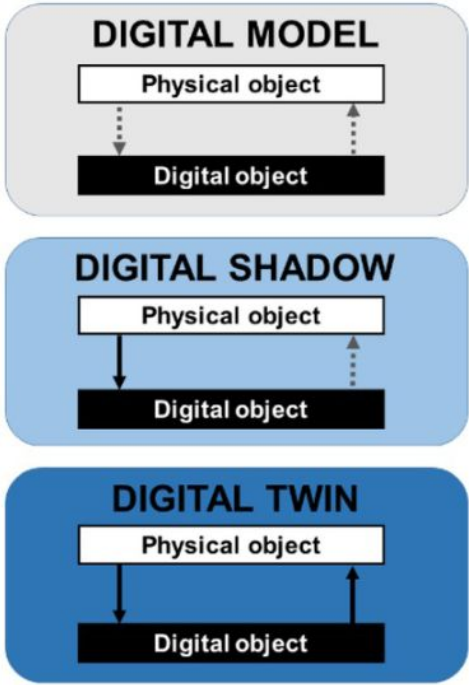
Digital Chemistry (DC) - based on 5 pillars

- Laboratory informatics
- Computational chemistry
- Cheminformatics
- AI and Data Science
- Robotics (Laboratory automation)

“A digital twin of chemistry can only be formed with the integration of all parts.”

Digital twins

The digital twin of chemistry



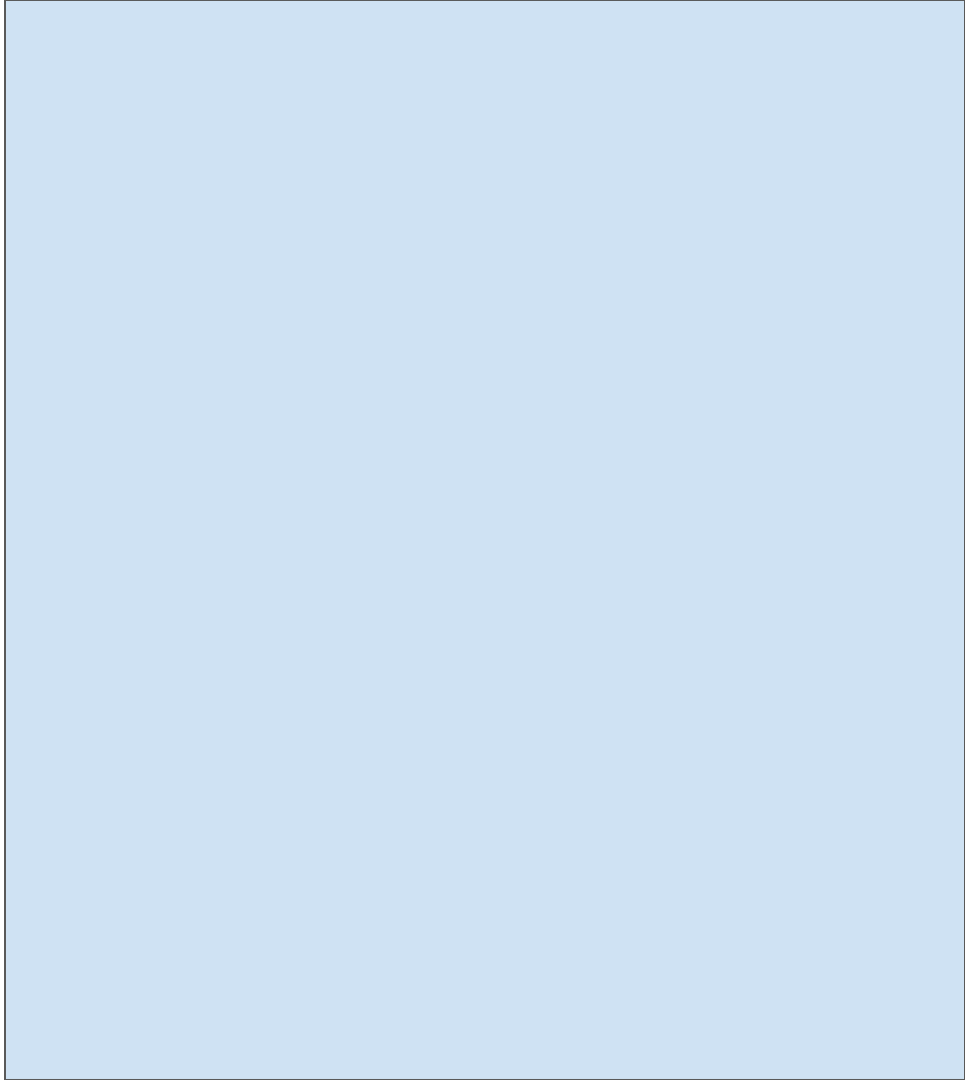
Thoughts:

- *"Virtual chemistry is claimed by education"*
- *Why just 'digital'?*

Nikula, Riku-Pekka, Marko Paavola, Mika Ruusunen, and Joni Keski-Rahkonen. 2020. "Towards Online Adaptation Of Digital Twins". *Open Engineering* 10 (1): 776-783. doi:10.1515/eng-2020-0088.

Transactional systems

Cli WiSe24/02-27.10-x



Transactional systems

Lab

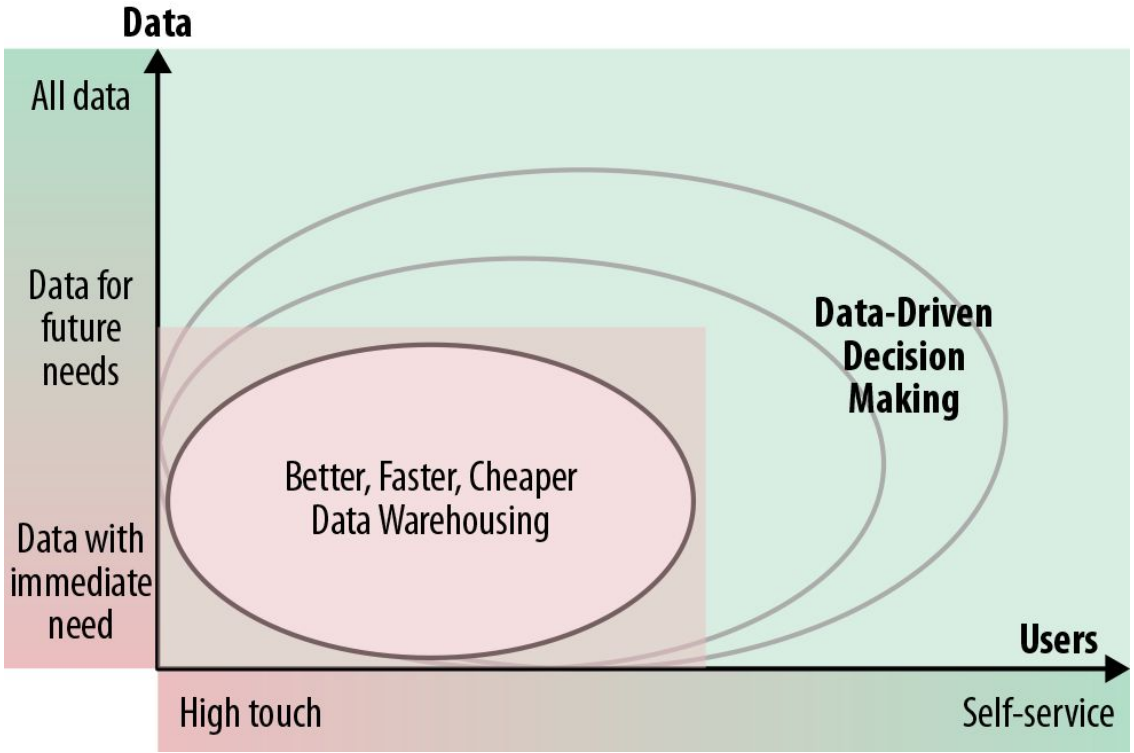
Systems discussed

- ELN - Electronic lab notebook
- LIMS - Laboratory information system
- LES - Laboratory execution system
- SDMS - Scientific data management system
- WIKI
- CDB - Chromatographic database
- PAT - Process analytic technologies

Data concepts

Architectures

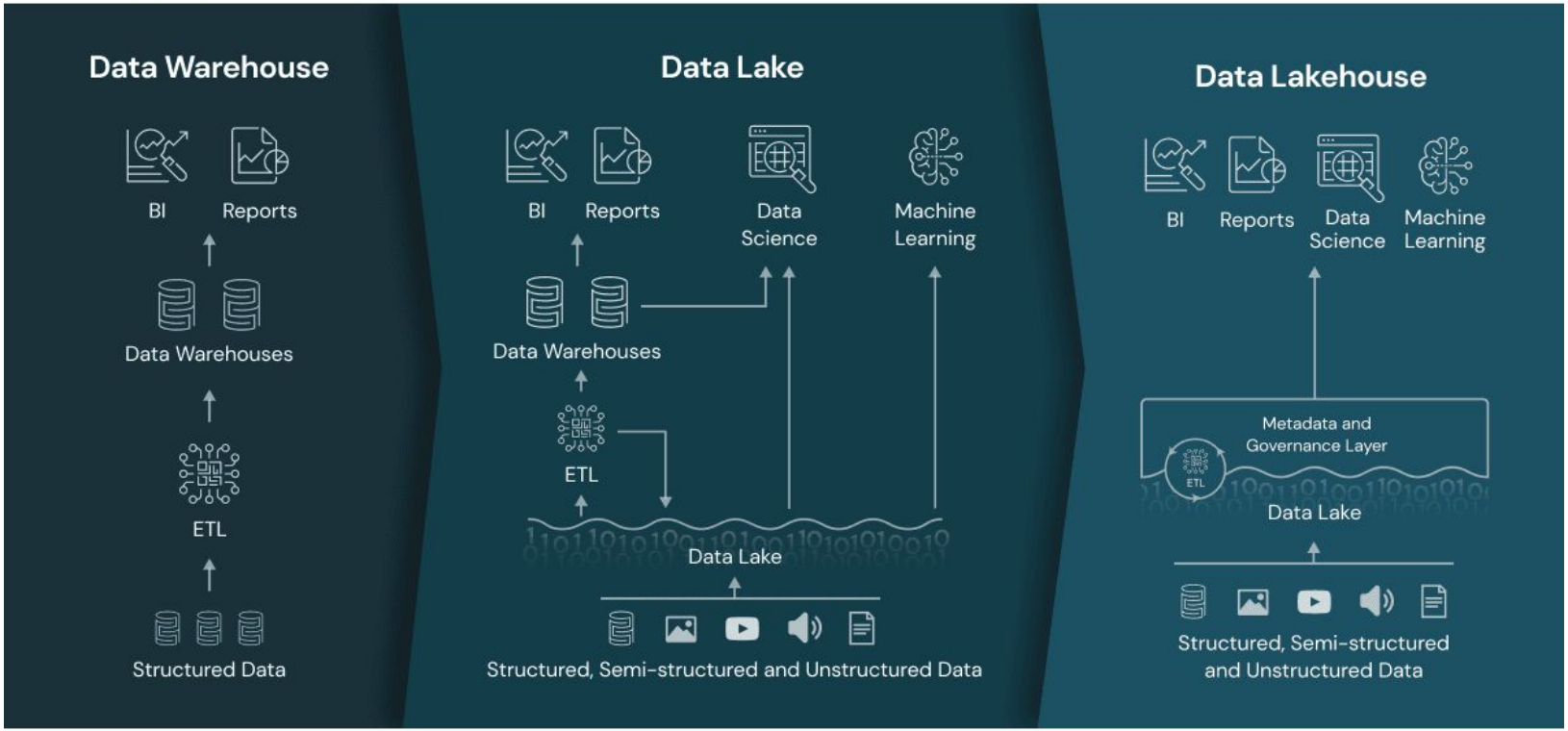
Value proposition



https://www.databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html

#essential

Evolution



Data Lakehouse - Data Lake problems

While suitable for storing data, data lakes lack some critical features:

- they do not support transactions,
- they do not enforce data quality,
- and their lack of consistency / isolation

… makes it almost impossible to mix appends and reads, and batch and streaming jobs.

For these reasons, many of the promises of the data lakes have not materialized, and in many cases leading to a loss of many of the benefits of data warehouses.

"What Is A Lakehouse?". 2020. *Databricks*.
https://www.databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html.

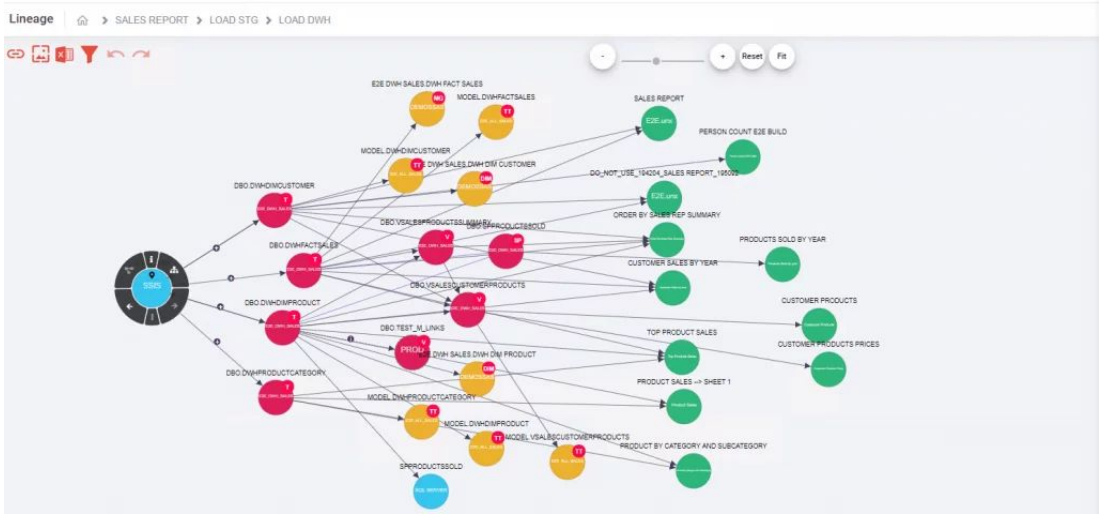
Pipelines

Data Lineage

Data lineage includes:

- the data origin,
- what happens to it, and
- where it moves over time.

Data lineage gives visibility while greatly simplifying the ability to trace errors back to the root cause in a data analytics

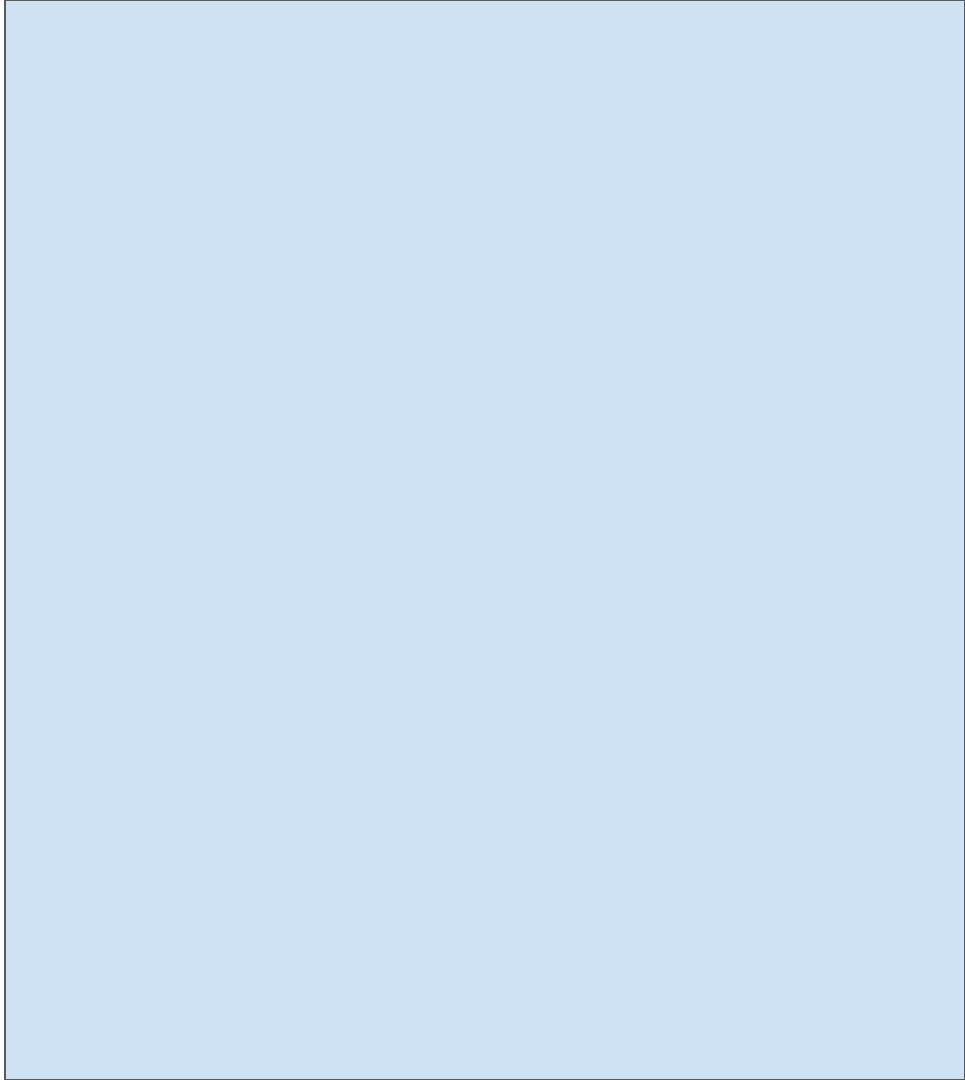


Provenance is defined as a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing.

https://en.wikipedia.org/wiki/Data_lineage
https://www.octopai.com/
https://getmanta.com/data-lineage-done-right-demo

Chemistry examples, REST

Cli WiSe24/03-03.11-o



Fundamentals

Chemistry

Chemistry Building blocks

- Experiment
- Analytics
- Models

LIVE-Demo
100 Milliarden Milliarden Teilchen

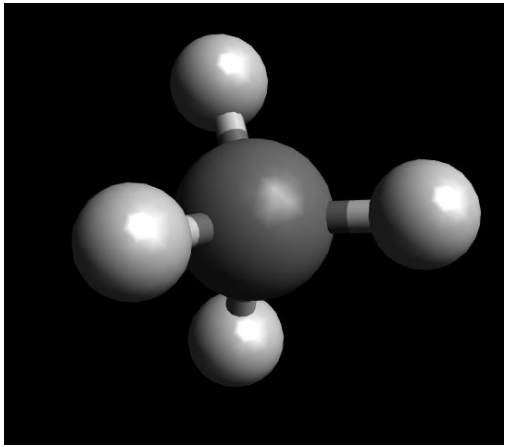
Der Kochsalzkristall etwa 0,1781 Quintillionen Teilchen, oder 178,1 Milliarden Milliarden Teilchen

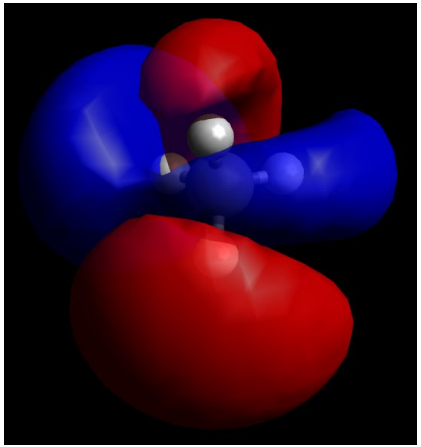
https://chat.openai.com/share/019adbff-1c74-4761-9071-4be8b0f0b895

Model

Chemistry

Ball and Stick or Wave?





$6{,}023 \times 10^{23}$

Model

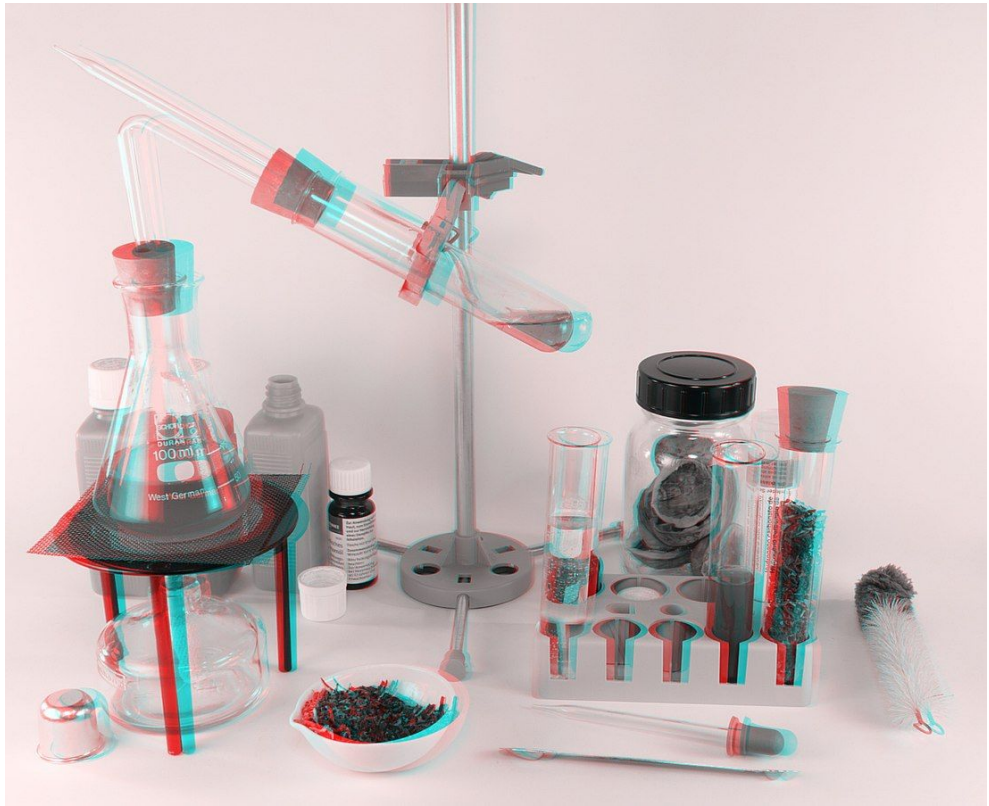
A model is an **abstract** description of a concrete system using concepts and language.

AOT, Edward Zalta (1981): Some objects (the ordinary concrete ones around us) *exemplify properties*, while others (abstract objects like numbers, and what others would call "nonexistent objects", like the round square and the mountain made entirely of gold) merely *encode* them.

Experiment

Chemistry

Only the experiment is valid!



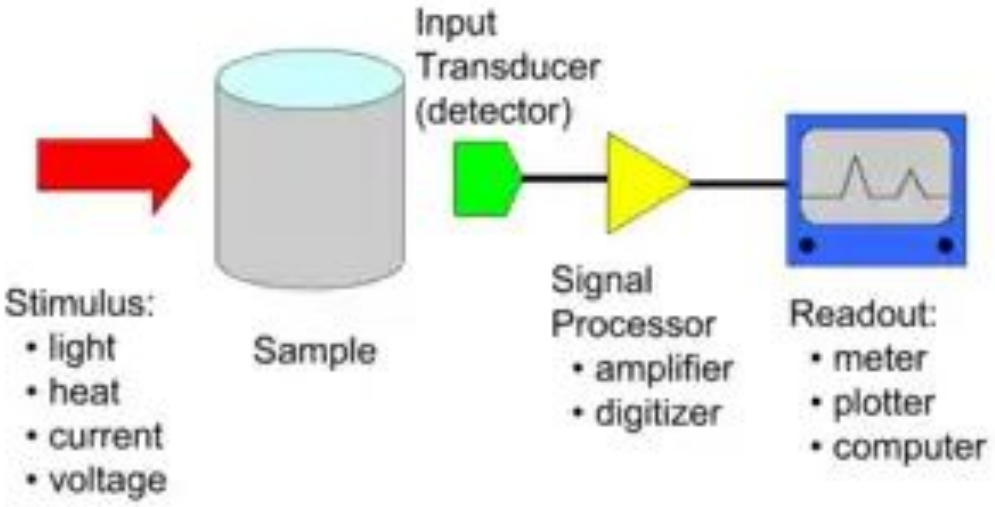
Analytics

Chemistry

Analytics

Analytical chemistry studies and uses instruments and methods to **separate, identify**, and **quantify matter**. In practice, separation, identification or quantification may constitute the entire analysis or be combined with another method. Separation isolates analytes. Qualitative analysis identifies analytes, while quantitative analysis determines the numerical amount or concentration.

Analytics



Block diagram of an analytical instrument showing the stimulus and measurement of response

Analytic process

Laboratory

Laboratory processes

Experiments: Benchtop

Testing: There are three phases of laboratory testing: Pre-analytical (pre-testing phase) Analytical (testing phase) Post-analytical (post-testing or reporting phase)

Management: For the most part, lab management involves certain lab-keeping chores such as maintaining instruments, restocking consumables, scheduling, giving technical advice, and keeping records of certain lab activities or incidences in the lab.

Testing

Risk-based thinking for chemical testing, Siu-kay Wong

https://www.researchgate.net/publication/314109558_Risk-based_thinking_for_chemical_testing .

Descriptions, features, fingerprints

(Cli WiSe24/04-10.11-o)

Digital twin of reaction and molecule

SMILES

Simple descriptor

SMILES

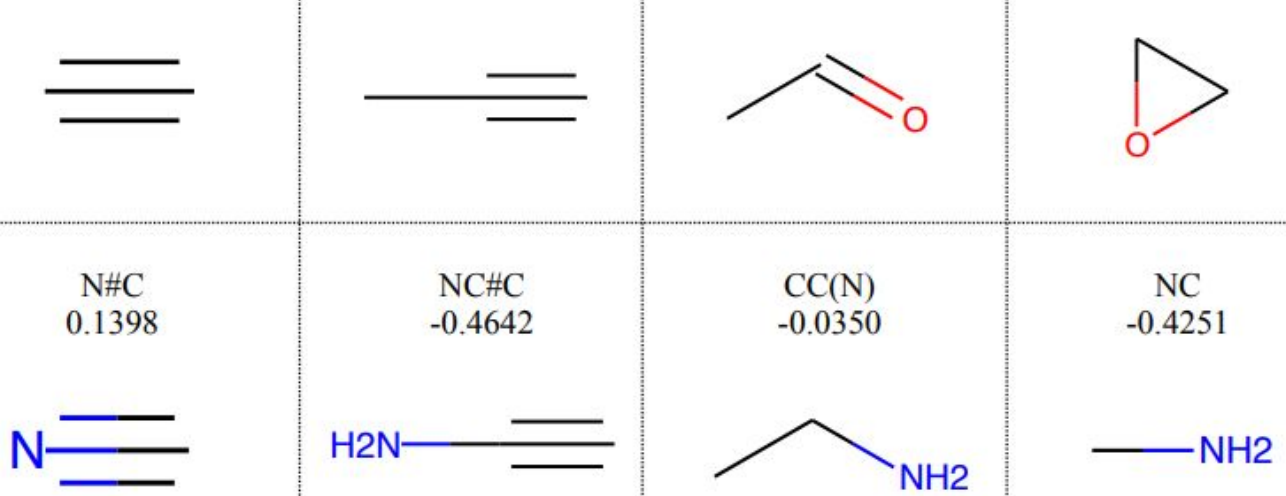
The *simplified molecular-input line-entry system* (SMILES) is a specification in the form of a line notation for describing the structure of chemical species using short ASCII strings. It is a *string* obtained by printing the symbol nodes encountered in a depth-first tree traversal of a chemical graph.

- The original SMILES specification was initiated in the 1980s. It has since been modified and extended.
- Typically, a number of equally valid SMILES strings can be written for a molecule. For example, `CCO`, `OCC` and `C(O)C` all specify the structure of ethanol.
- Can contain - but not enforces - stereochemistry
- From the view point of a formal language theory, SMILES is a word.

https://archive.epa.gov/med/med_archive_03/web/html/smiles.html

ML example: SMILES and pKa

C#C 0.2494	CC#C 0.6395	CC=O 0.2052	C1CO1 0.0166
N#C 0.1398	NC#C -0.4642	CC(N) -0.0350	NC -0.4251



https://arxiv.org/pdf/2012.09712.pdf

Fingerprints, features, and descriptors (Introduction)

Descriptors

Molecular descriptors are **numerical values that describe various properties of a molecule**, such as its size, shape, polarity, or reactivity.

They are often calculated from the 2D or 3D representation of a molecule and can be used to compare different molecules or to predict various molecular properties.

Examples of molecular descriptors include molecular weight, logP, number of hydrogen bond donors or acceptors, polar surface area, etc.

Features

Molecular features are **binary or categorical variables that represent specific molecular substructures or patterns within a molecule**.

They are often used to create fingerprints or other binary representations of molecules, which can be used for similarity searching, machine learning, or other applications.

Examples of molecular features include the presence or absence of specific chemical functional groups, substructure fragments, or pharmacophores.

Why fingerprints (“1D chemical descriptors”)?

Molecular fingerprints are a way of

- encoding the structure of a molecule
- They are essential tools for mapping the chemical space
- Example: Input layer in ANN or Similarity Maps Using Fingerprints

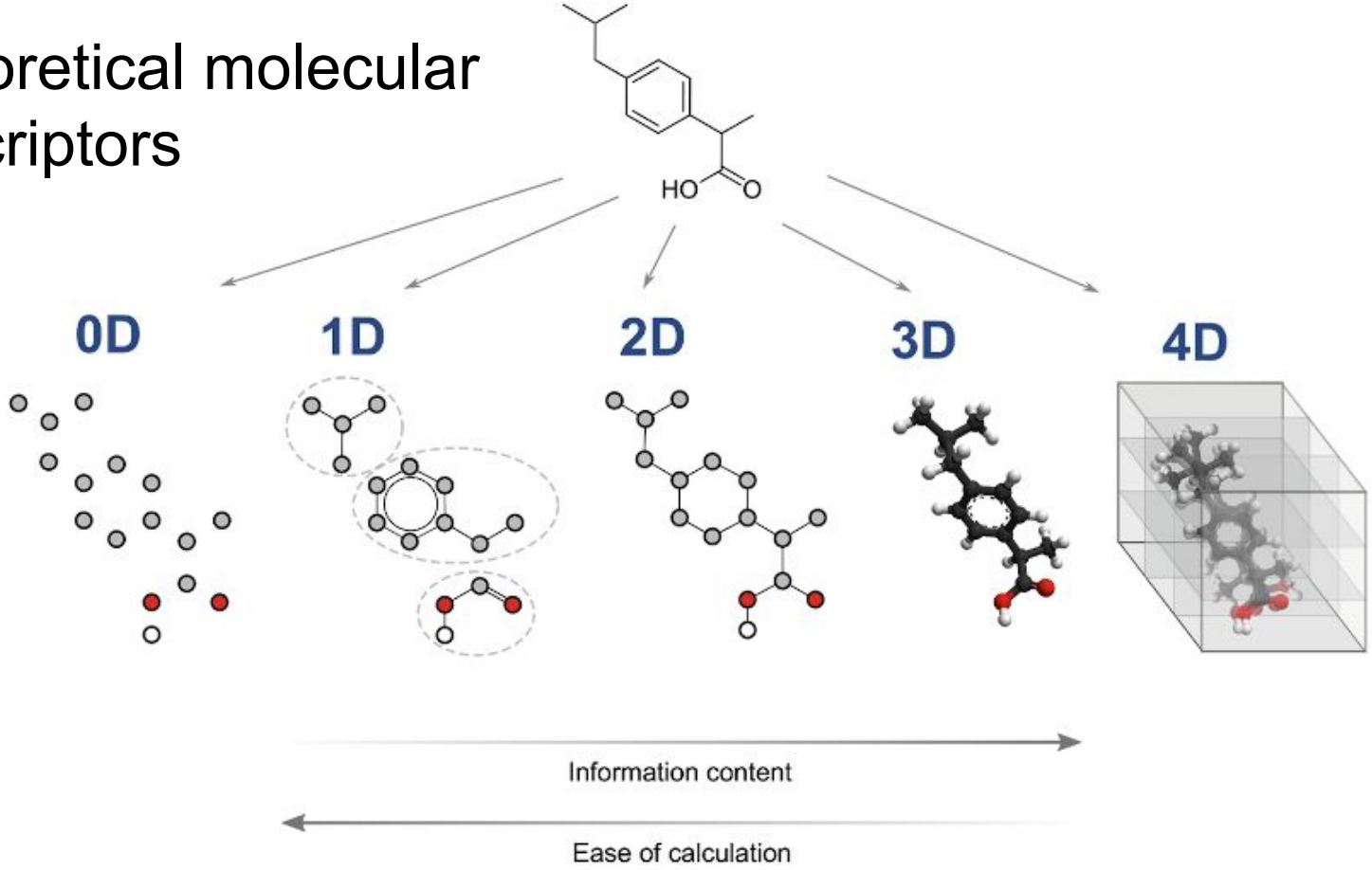
Descriptors and features

The transformation of data with the help of cheminformatic into descriptors and features *introduces implicit domain knowledge* to data science in chemistry.

The molecular descriptor is the result of a logic and mathematical procedure that *transforms chemical information* encoded within a symbolic representation of a molecule *into a useful number* or the result of some standardized experiment.

Also SMILES is a descriptor.

Theoretical molecular descriptors



https://link.springer.com/protocol/10.1007/978-1-4939-7899-1_1

Descriptor examples

This list contains an excerpt from *classic molecular descriptors* (of ~2000). Mostly they are the result of calculations within MM, MD, and QC.

1) Boiling point (**BP**)
2) Melting point (**MP**)
3) Heat capacity at T constant (**CT**)
4) Heat capacity at P constant (**CP**)
5) Entropy (**S**)
6) Density (**DENS**)
7) Enthalpy of vaporization (**HVAP**)
8) Standard enthalpy of vaporization (**DHVAP**)

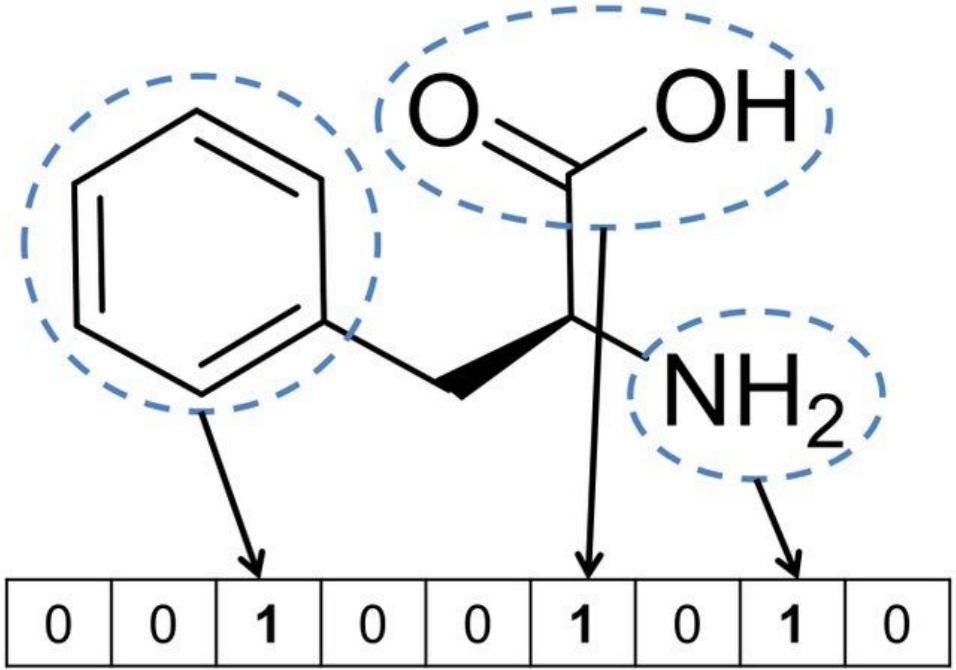
...

9) Enthalpy of formation (**HFORM**)
10) Standard enthalpy of formation (**DHFORM**)
11) Motor octane number (**MON**)
12) Molar refraction (**MR**)
13) Acentric factor (**AcenFac**)
14) Total surface area (**TSA**)
15) Octanol-water partition coefficient (**LogP**)
16) Molar volume (**MV**)
17) Log water solubility (**logSw**)
18) Total surface area (**TSA**)

...

Structural descriptors. Another class of descriptors plays an important role in organic chemistry. Examples are ring counts, aromaticity, or graph diameter.

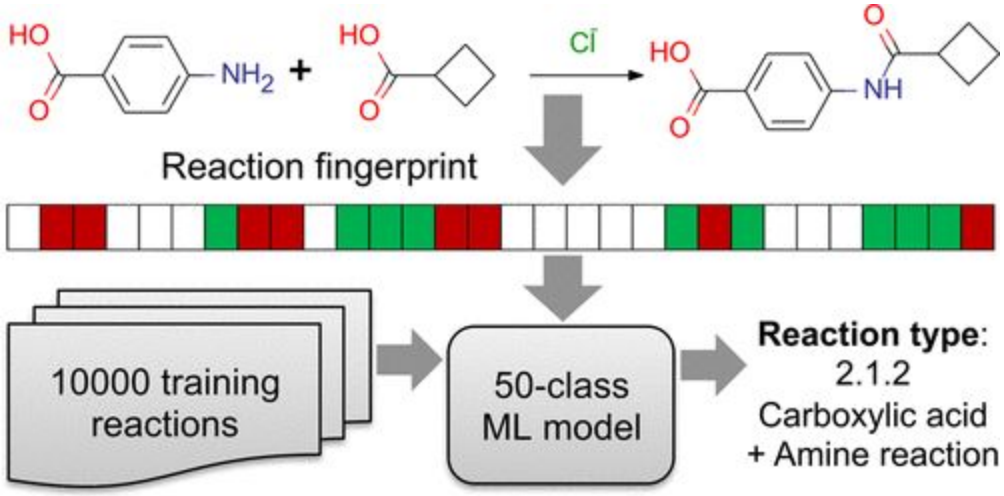
Fingerprint example



https://www.researchgate.net/publication/315513438_Chemical_Data_Formats_Fingerprints_and_Other_Molecular_Descriptions_for_Database_Analysis_and_Searching/figures?lo=1

Reaction fingerprints

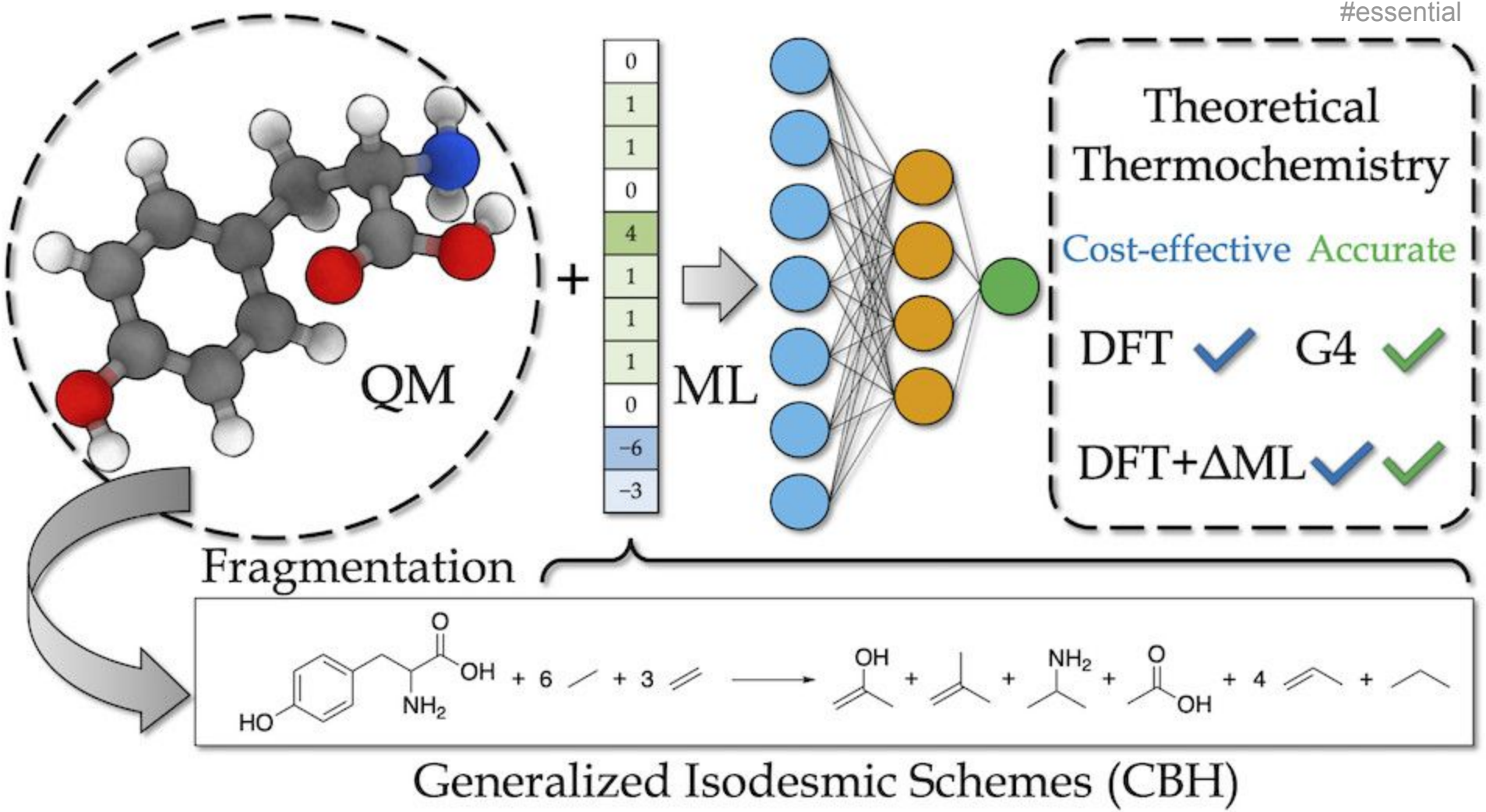
#essential



https://pubs.acs.org/doi/10.1021/ci5006614

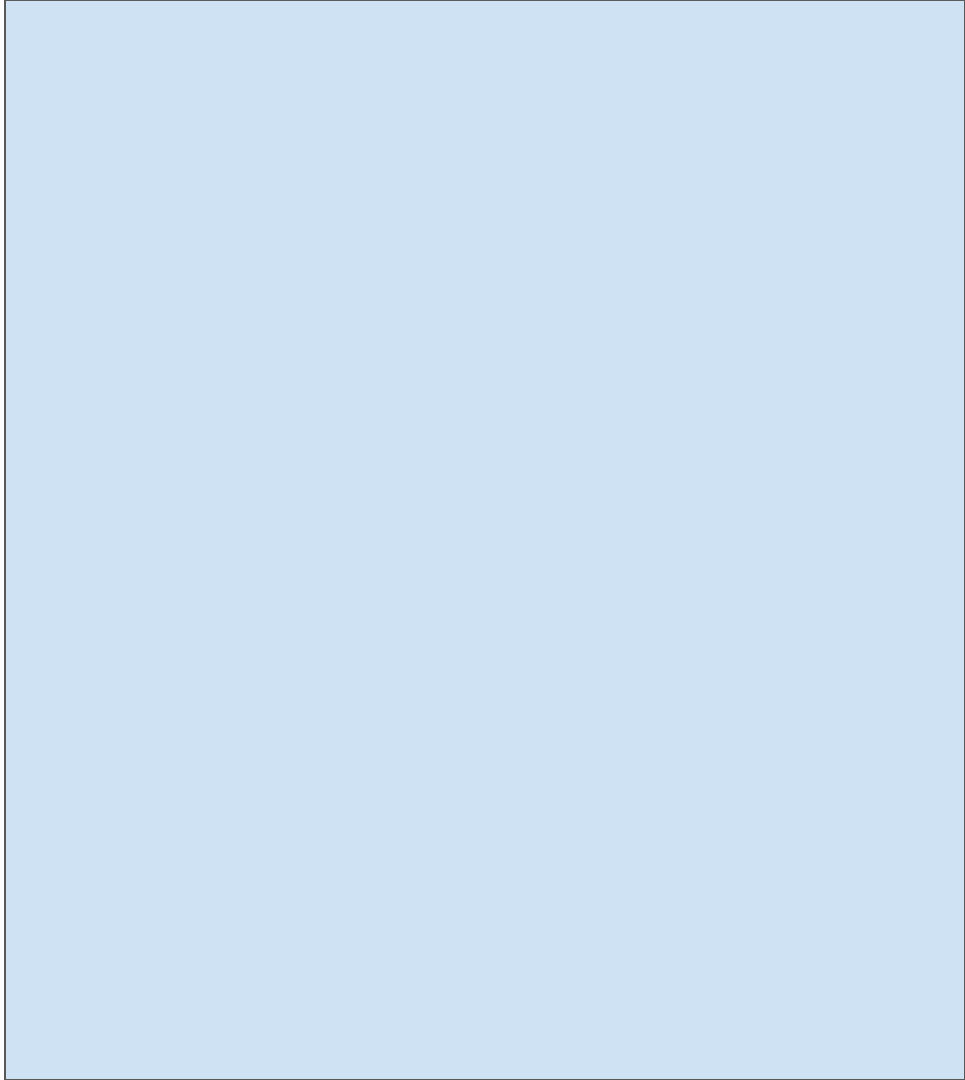
Molecular descriptors

QM Scope


QM
0
1
1
0
4
1
1
1
0
−6
−3
ML
Theoretical
Thermochemistry
Cost-effective Accurate
DFT
G4
DFT+ΔML
Fragmentation
HO
NH2
OH
O
+ 6
+ 3
+ 4
Generalized Isodesmic Schemes (CBH)

Molecule representations, Data files, cartridges

(Cli WiSe24/06-24.11-o)



Chemical file format examples

MOLFILE

An **MDL Molfile** is a file format for holding information about the atoms, bonds, connectivity and coordinates of a molecule.

The molfile consists of some header information, the Connection Table (CT) containing atom info, then bond connections and types, followed by sections for more complex information.

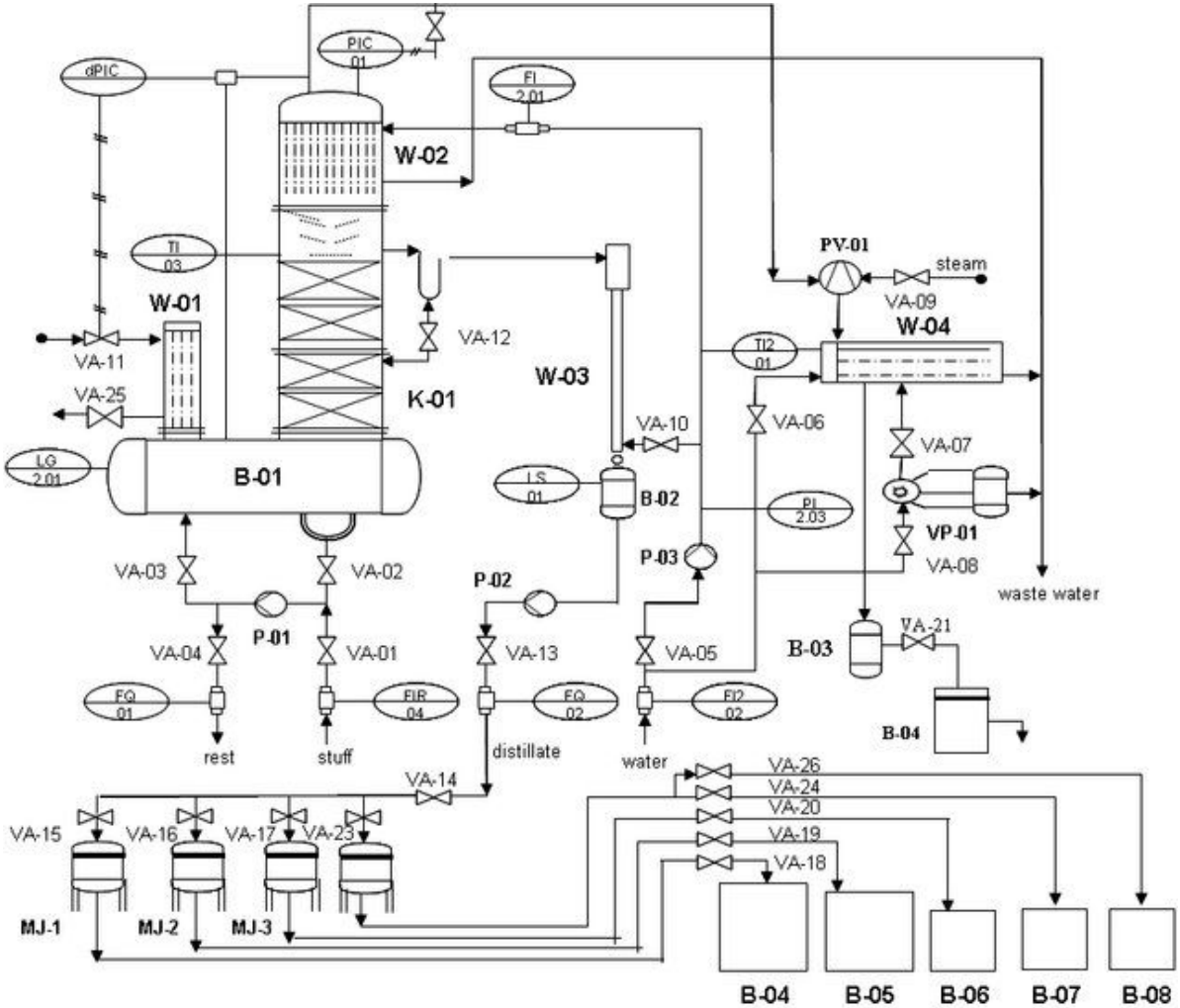
ctab

Filename extension	`.mol`
Internet media type	`chemical/x-mdl-molfile`
Type of format	chemical file format

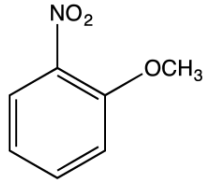
→ https://en.wikipedia.org/wiki/Chemical_file_format

Chemical table file (CT File) is a family of text-based chemical file formats that describe molecules and chemical reactions. One format, for example, lists each atom in a molecule, the x-y-z coordinates of that atom, and the bonds among the atoms.

Transactional industrial systems (introduction)


dPIC
PIC 01
FI 2.01
W-02
TI 03
W-01
VA-11
VA-25
VA-12
K-01
W-03
PV-01
steam
VA-09
W-04
TI2 01
VA-06
VA-07
VA-10
LG 2.01
B-01
LS 01
B-02
PI 2.03
VP-01
VA-08
waste water
P-03
P-02
VA-03
VA-02
P-01
VA-04
VA-01
VA-13
VA-05
B-03
VA-21
FQ 01
FIR 04
FQ 02
FI2 02
B-04
rest
stuff
distillate
water
VA-14
VA-26
VA-24
VA-20
VA-19
VA-18
VA-15
VA-16
VA-17
VA-23
MJ-1
MJ-2
MJ-3
B-04
B-05
B-06
B-07
B-08

Unit operation sequence



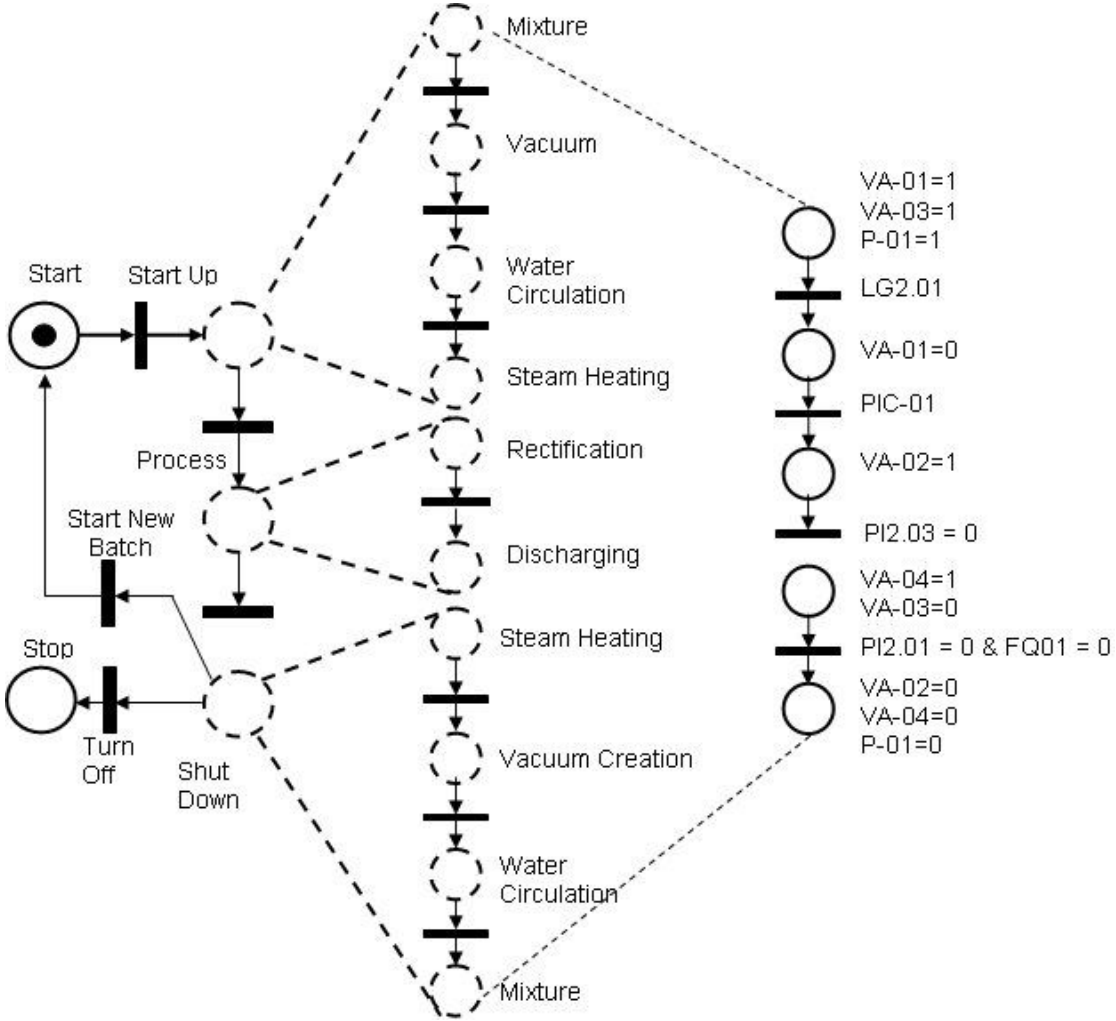
https://www.fr.de/frankfurt/montag-gift-regnete-11268840.html

Model

https://en.wikipedia.org/wiki/O-Nitroanisole

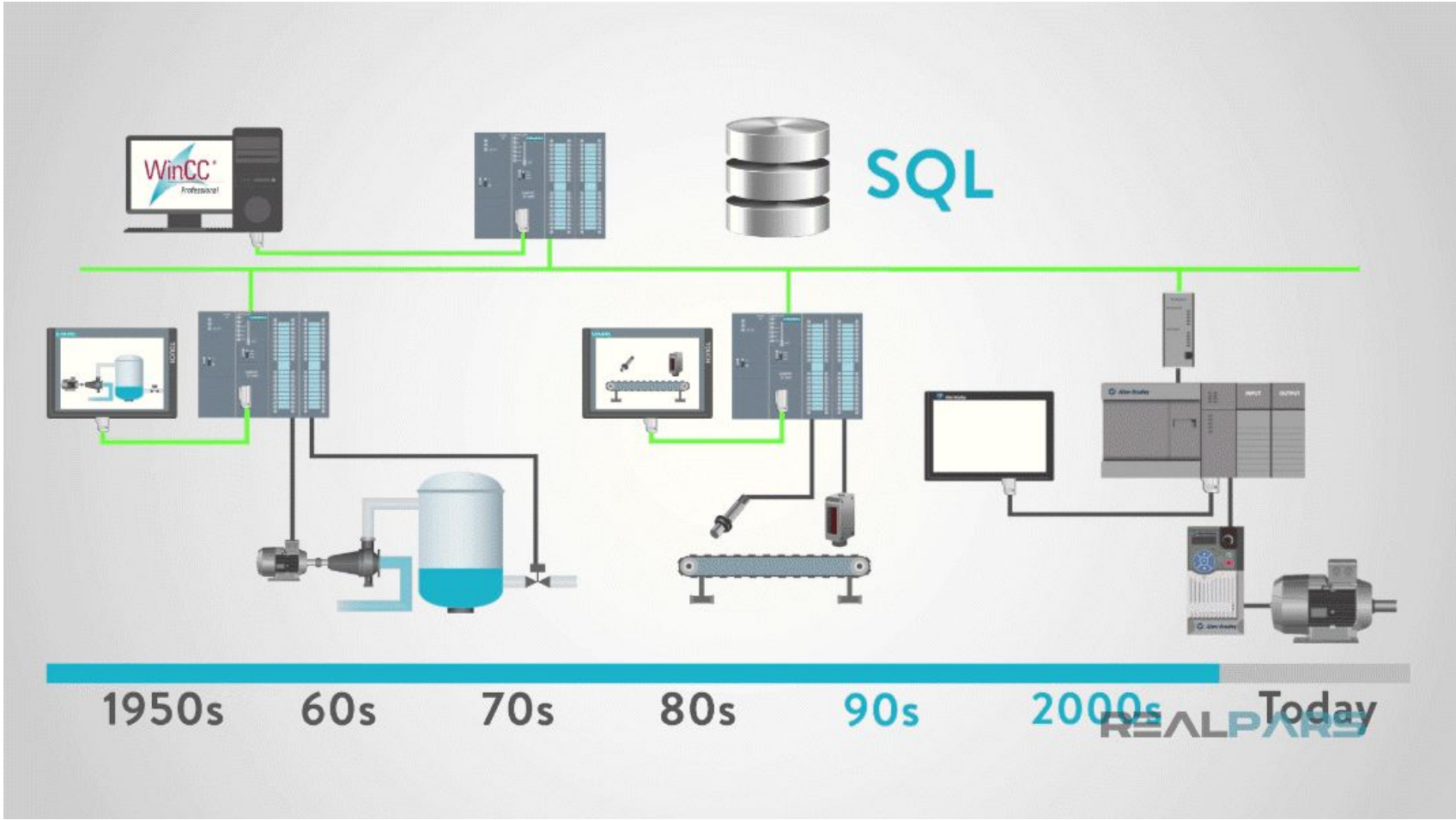
SCADA and production architectures

#essential



https://www.researchgate.net/publication/224347928_Intelligent_component_based_batch_control_using_IEC61499_and_ANSIISA_S88/figures?lo=1 // Idilia Batchkova

#essential



https://realpars.com/scada/

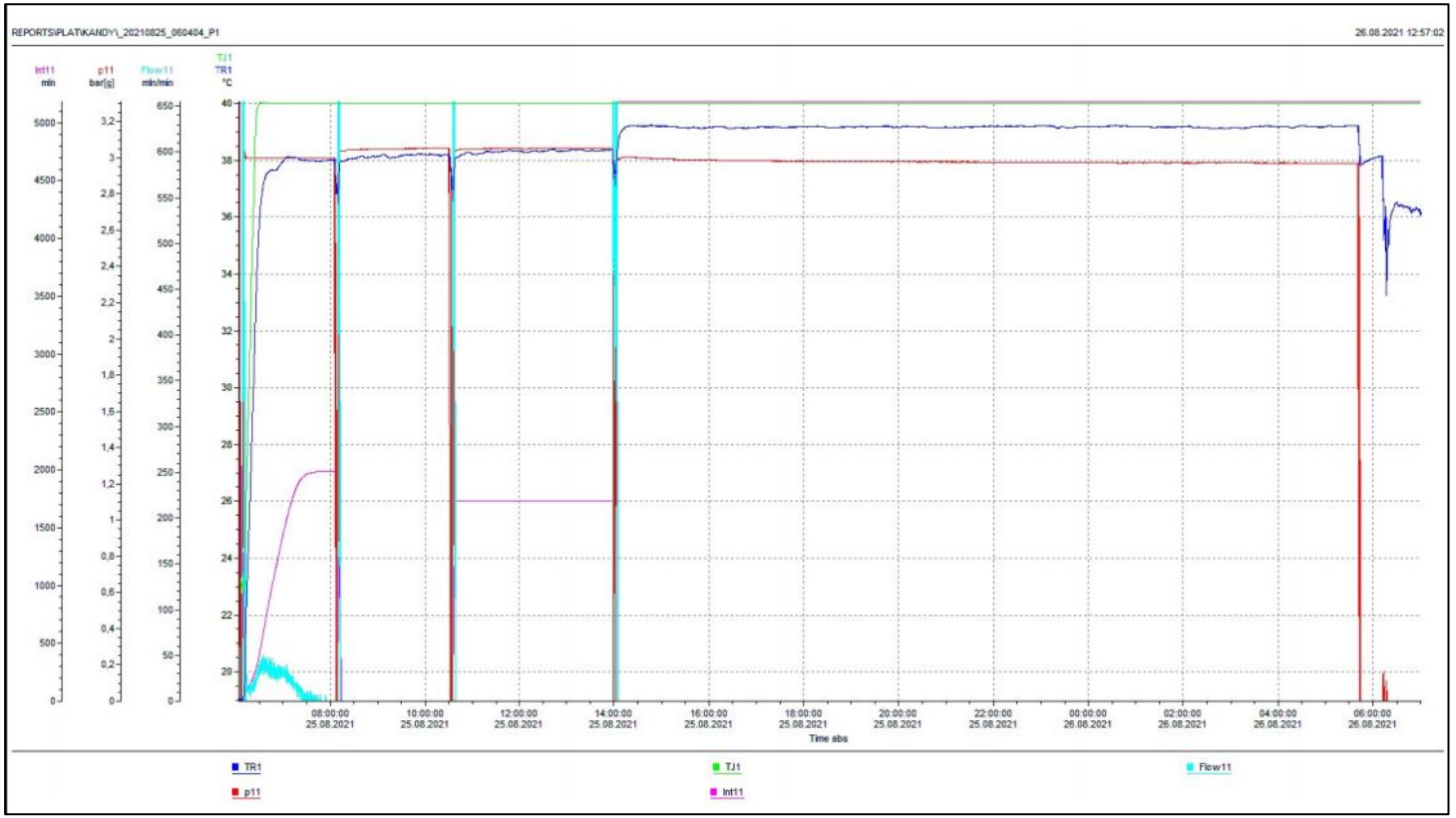
Simatic

SIMATIC is a series of programmable logic controller and automation systems, developed by Siemens. Introduced in 1958, the series has gone through four major generations, the latest being the SIMATIC S7 generation. The series is intended for industrial automation and production.

The name SIMATIC is a registered trademark of Siemens. It is a portmanteau of “Siemens” and “Automatic”.

Simatic PCS 7 V9.1

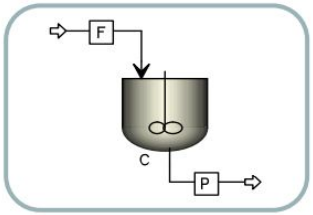
Original probe



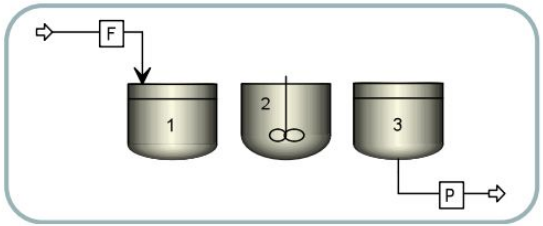
Batches

Batch = Recipe + Execution (t)

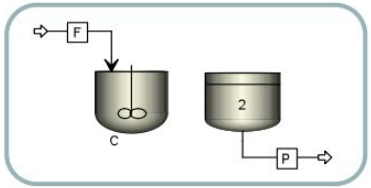
Continuous processes: manufacture of commodities.

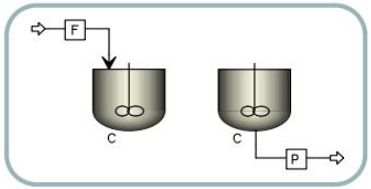


Batch processes: specialty chemicals, pharmaceuticals.



Semicontinuous processes: hybrids of batch and continuous.

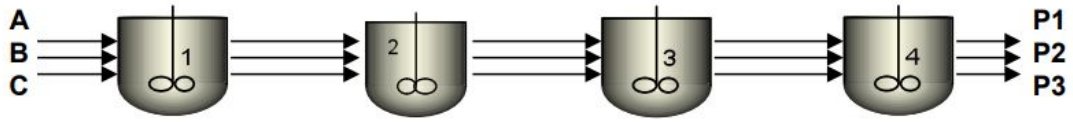




*** Fed-batch**

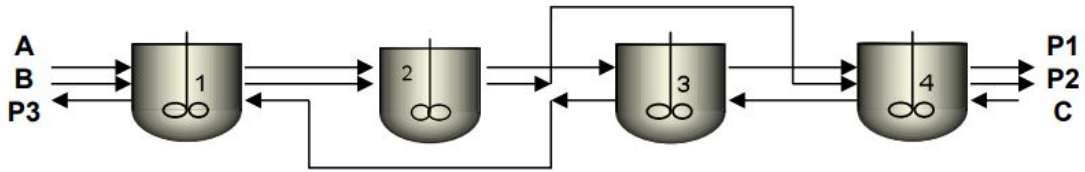
https://ocw.unican.es/pluginfile.php/2724/course/section/2514/topic_9.pdf

FLOWSHOP (or multiproduct) plants in which all products require all stages following the same sequence of operations.



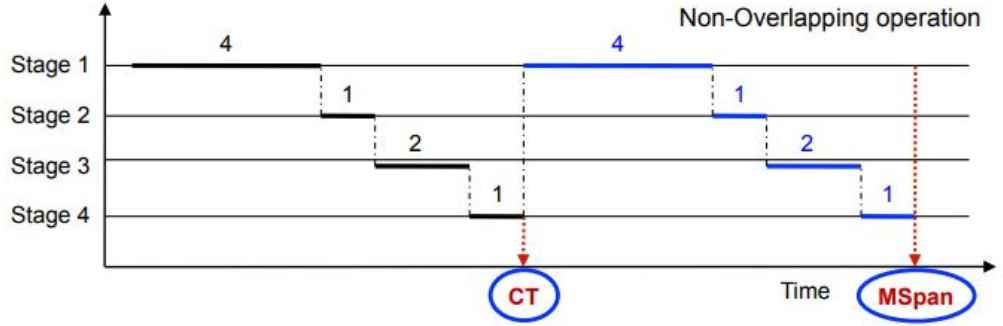
Flowshop plant

JOBSHOP (or multipurpose) plants where not all products require all stages and/or follow the same sequence.

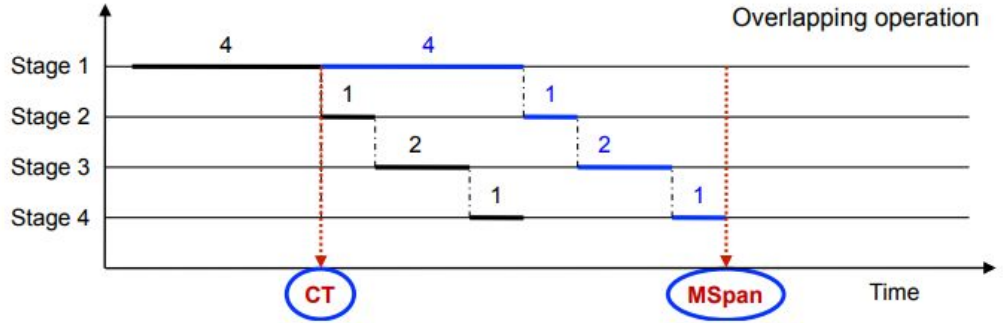


Jobshop plant

https://ocw.unican.es/pluginfile.php/2724/course/section/2514/topic_9.pdf



- Cycle time (CT) = 4 + 1 + 2 + 1 = 8 hrs.
- Makespan (2 batches) = 16 hrs. Poor equipment use.



- Cycle time (CT) = max {4, 1, 2, 1} = 4 hrs.
- Makespan (2 batches) = 12 hrs.

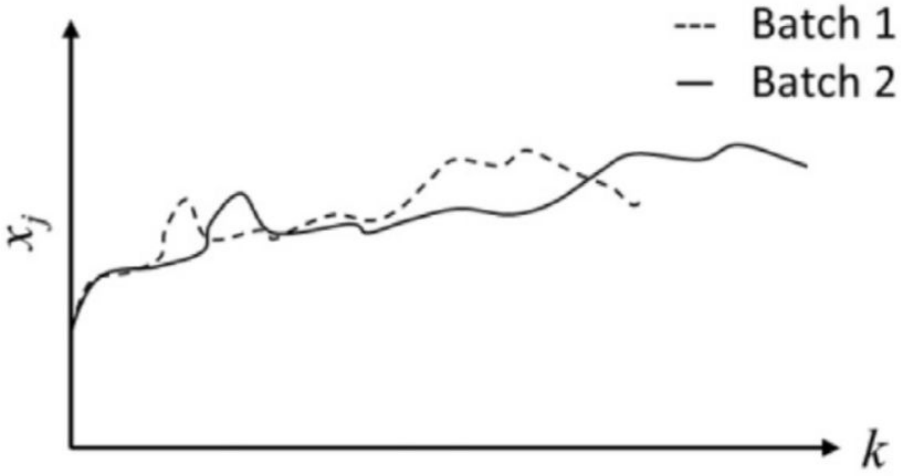
RECIPE

1. Mix 4 hrs.
2. Mix 1 hr.
3. Centrifuge 2 hrs.
4. Dry 1 hr.

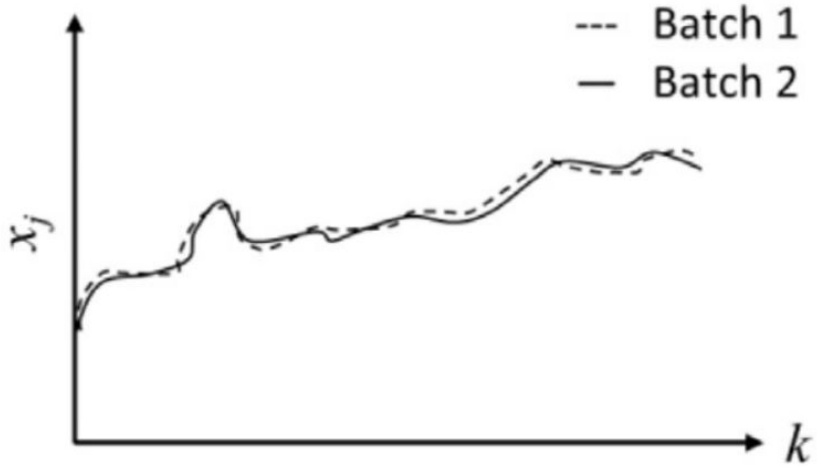
https://ocw.unican.es/pluginfile.php/2724/course/section/2514/topic_9.pdf

Batch analytics

Batch data alignment I



(a)



(b)

Data-driven methods for batch data analysis – A critical overview and
mapping on the complexity scale

Industrial Data Science for Batch Manufacturing Processes

Data-driven methods

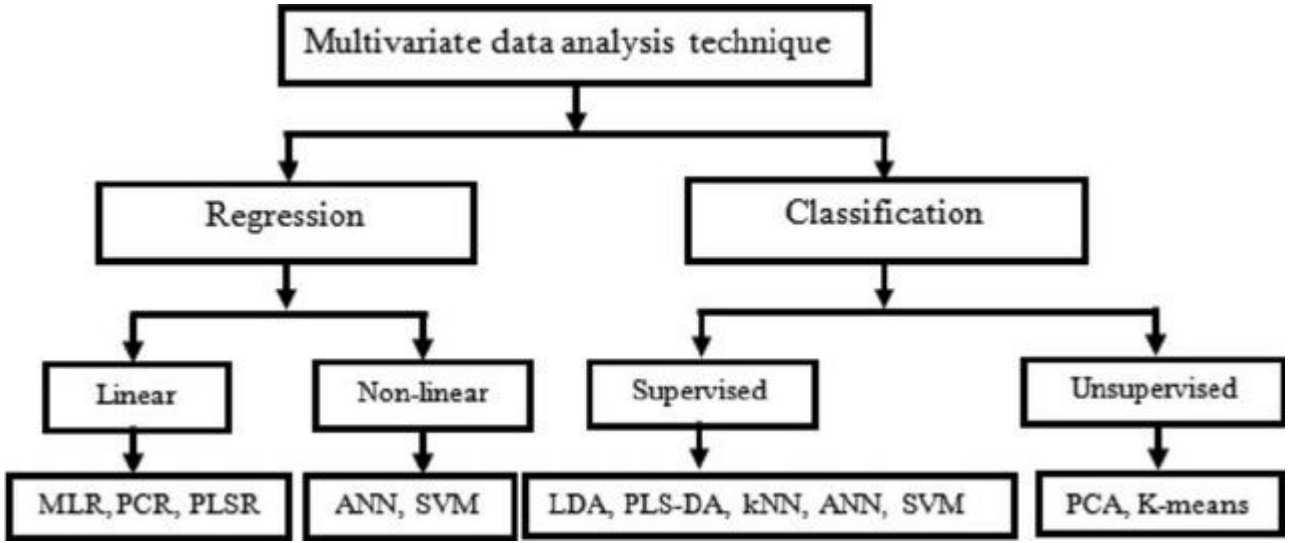
Data-driven methods can allow process engineers to quickly monitor and troubleshoot industrial batch processes in a manufacturing environment.

Therefore, our focus will be to present the intuition behind machine learning methods and their industrial applications pragmatically.

Most of the analysis will be done with point-and-click com-mercial software (JMP Pro, SAS Institute Inc).

Machine learning methods also available in open-source packages: Python (scikit-learn, pycaret, scikit-fda, pyphi) as well as other commercial software (e.g. SEEQ Trendminer Aspen ProMV, and SIMCA).

Multivariate data analysis techniques (MVDA)



https://www.researchgate.net/profile/Gamal-Elmasry