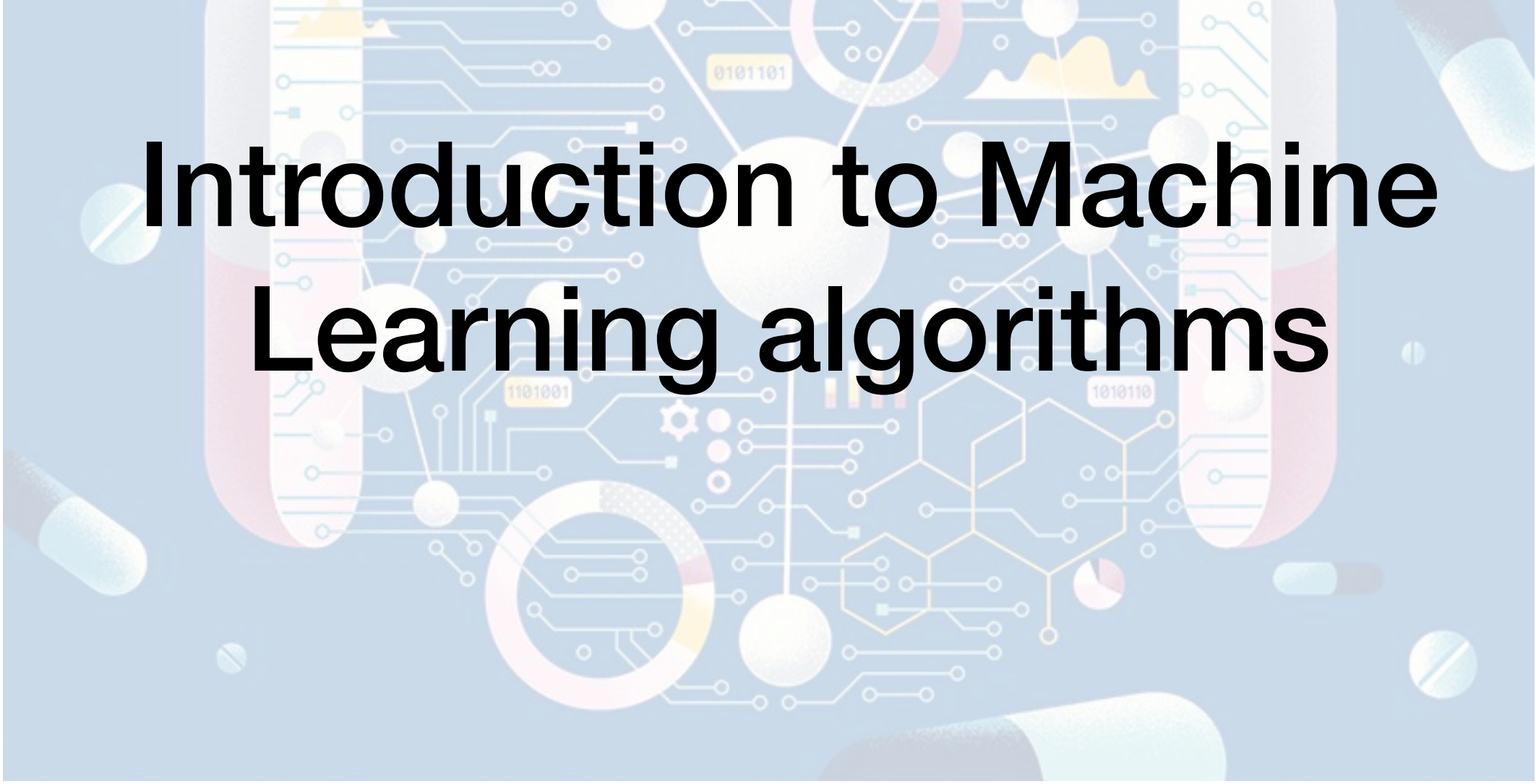


Introduction to Machine Learning algorithms

Dr. Fabien Plisson

Chemoinformatics in Drug Discovery

LANGEBIO, UGA CINVESTAV

October 15-18, 2019 - Irapuato, Mexico

Program

General Concept

Data Mining

Pre-processing

Exploratory Data Analysis

Predictive Modelling

Data Visualisation

Program

General Concept

Data Mining

Pre-processing

Exploratory Data Analysis

Predictive Modelling

Data Visualisation

What is Machine Learning?

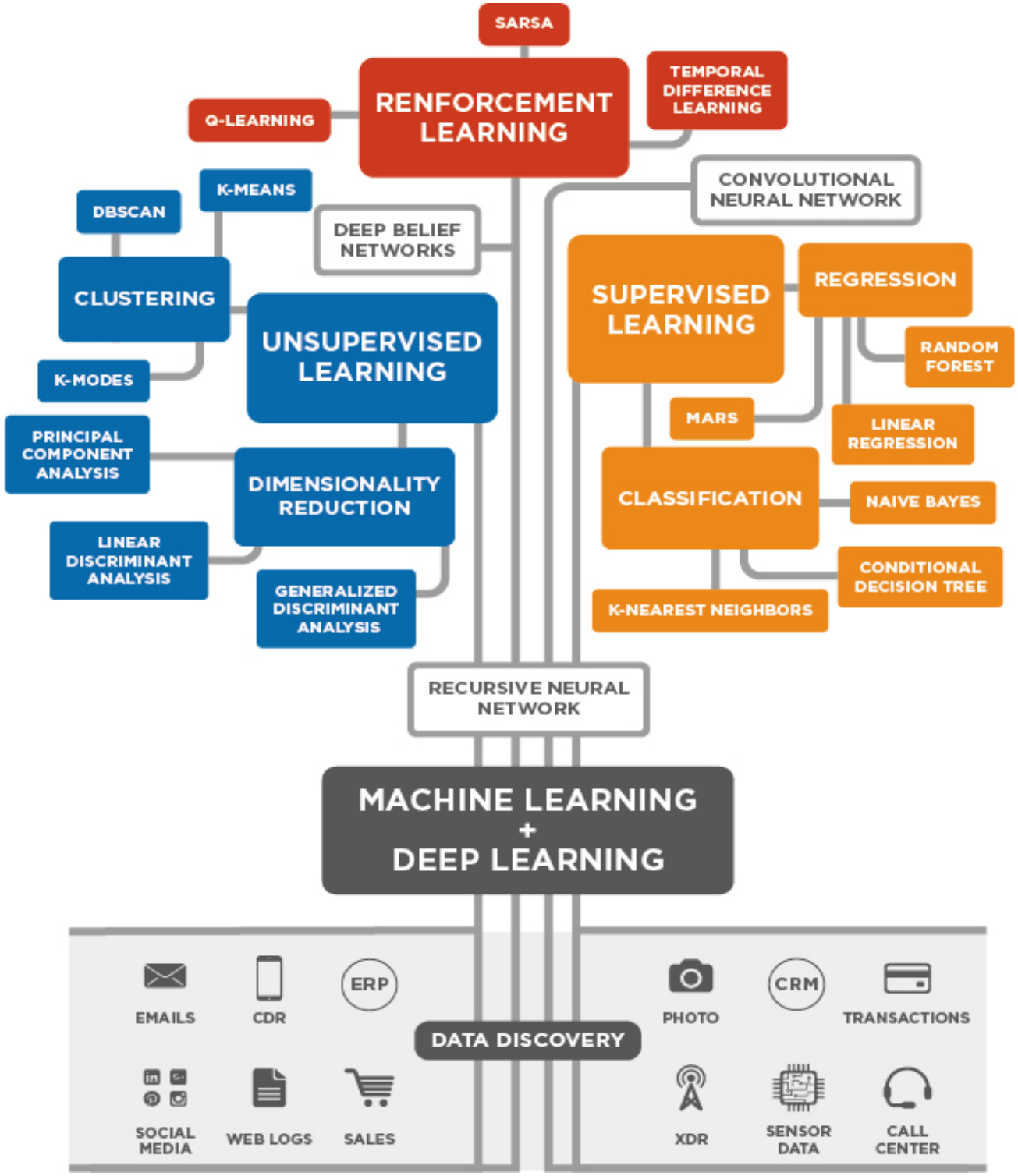
Use and development of algorithms **to identify patterns in the data that allow computers to learn**, and thus predict future behaviours. Computers are not explicitly programmed for such a task.

Supervised ML: The result is known, the computer looks for the best way to predict this result.

Unsupervised ML: The computer does not know the result, it groups objects based on similar patterns.

Reinforcement ML.

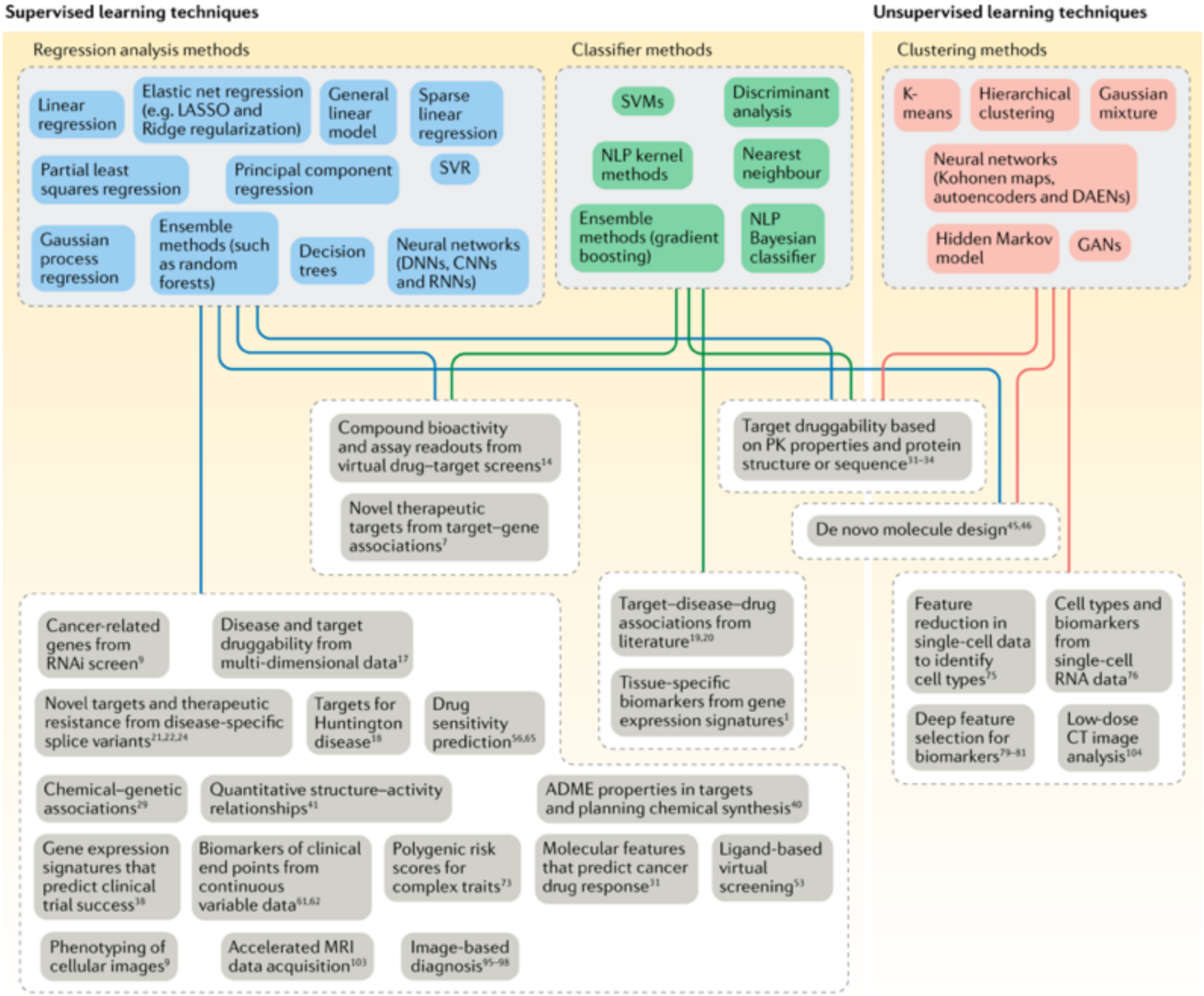
R python WEKA The University of Waikato julia


SARSA
RENFORCEMENT LEARNING
Q-LEARNING
TEMPORAL DIFFERENCE LEARNING
CONVOLUTIONAL NEURAL NETWORK
K-MEANS
DBSCAN
DEEP BELIEF NETWORKS
CLUSTERING
UNSUPERVISED LEARNING
SUPERVISED LEARNING
REGRESSION
RANDOM FOREST
K-MODES
MARS
LINEAR REGRESSION
PRINCIPAL COMPONENT ANALYSIS
DIMENSIONALITY REDUCTION
CLASSIFICATION
NAIVE BAYES
LINEAR DISCRIMINANT ANALYSIS
CONDITIONAL DECISION TREE
GENERALIZED DISCRIMINANT ANALYSIS
K-NEAREST NEIGHBORS
RECURSIVE NEURAL NETWORK
MACHINE LEARNING + DEEP LEARNING
EMAILS
CDR
ERP
PHOTO
CRM
TRANSACTIONS
DATA DISCOVERY
SOCIAL MEDIA
WEB LOGS
SALES
XDR
SENSOR DATA
CALL CENTER

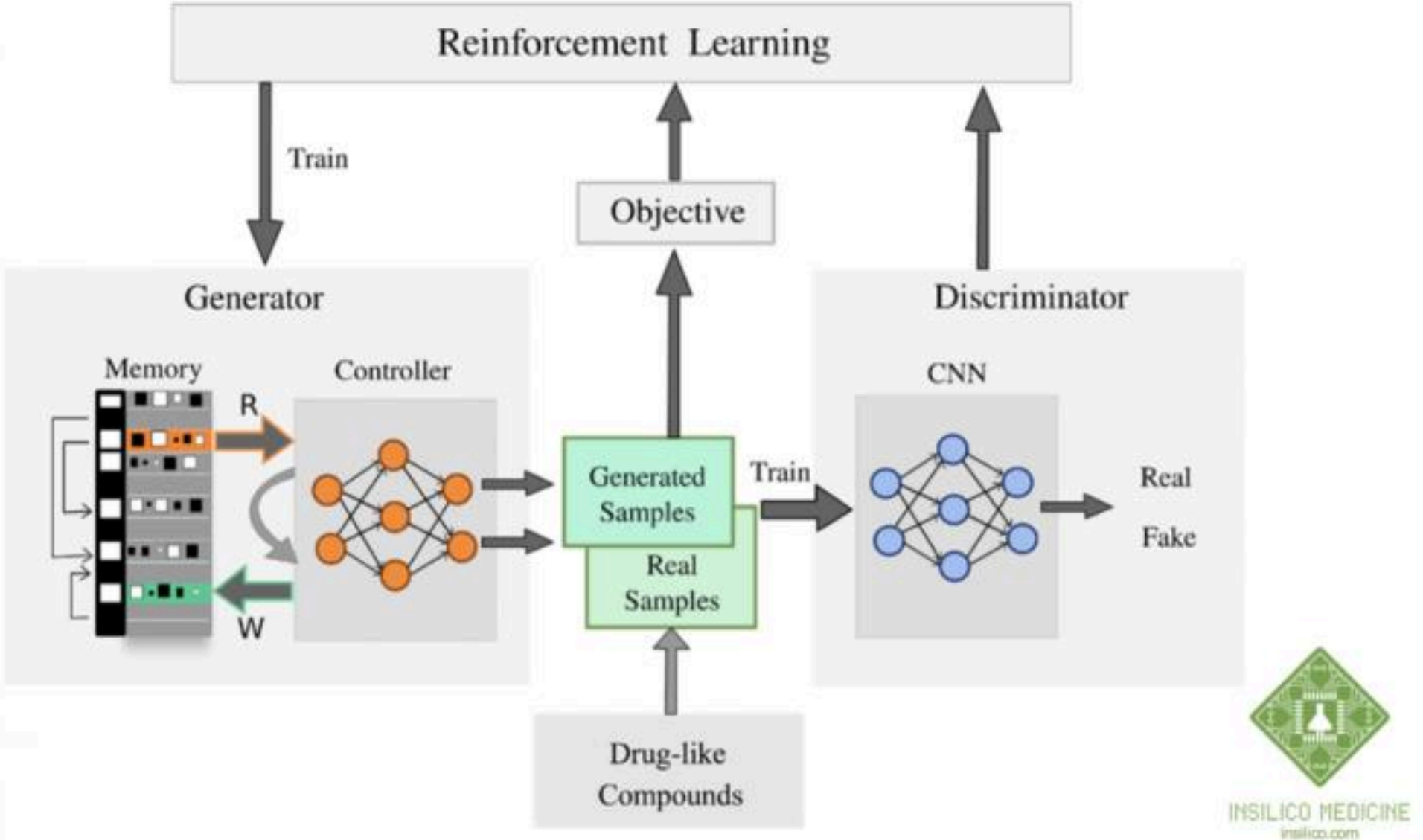
Applications

Database marketing
Handwriting recognition
Recommendation systems
Information retrieval
Object recognition in computer vision
Optical character recognition
Spam detection
Pattern recognition
Speech recognition
Bioinformatics
…

Machine Learning in Drug Discovery



Machine Learning in Drug Discovery

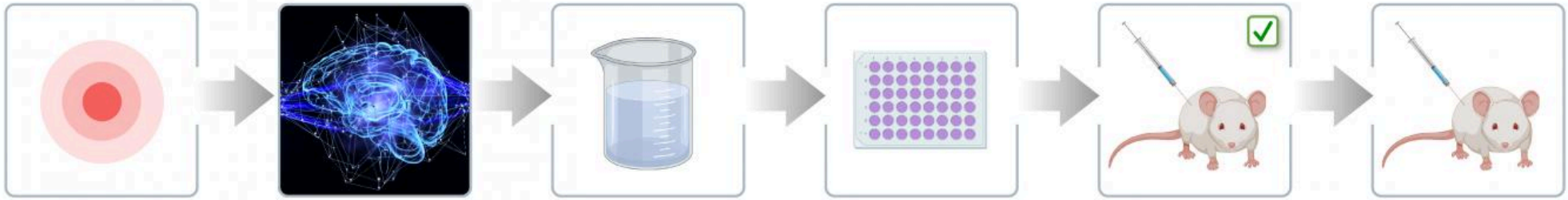


DEEP LEARNING ENABLES RAPID IDENTIFICATION OF POTENT DDR1 KINASE INHIBITORS

46 days

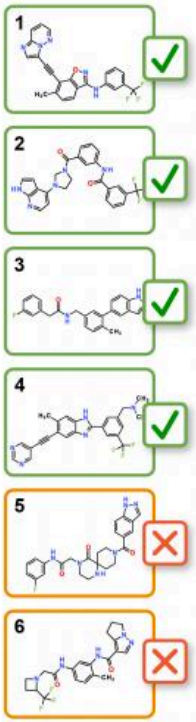
21 days

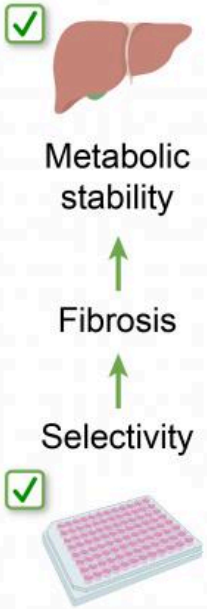
DDR1 | GENTRL | Rapid Synthesis | *in vitro* assays | *in vivo* assays | Pre-Clinical Trials



Target Nominated

GENERATIVE TENSORIAL REINFORCEMENT LEARNING





Mouse Pharmakinetics PK assays

nature biotechnology paper accepted

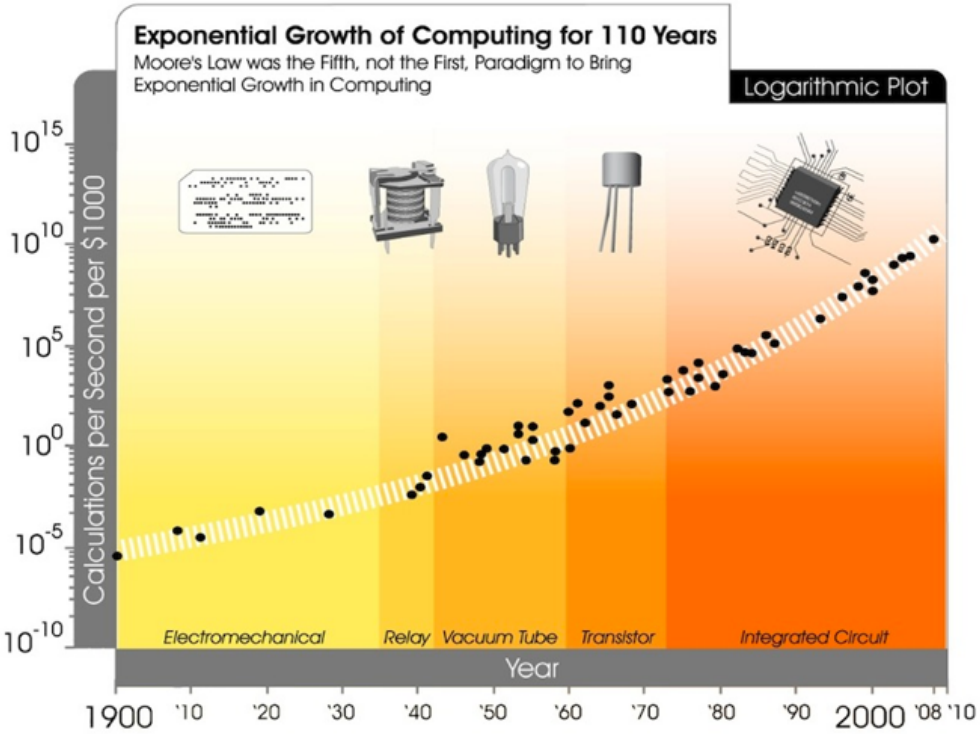
Good PK Properties Confirmed

Insilico Medicine
insilico.com

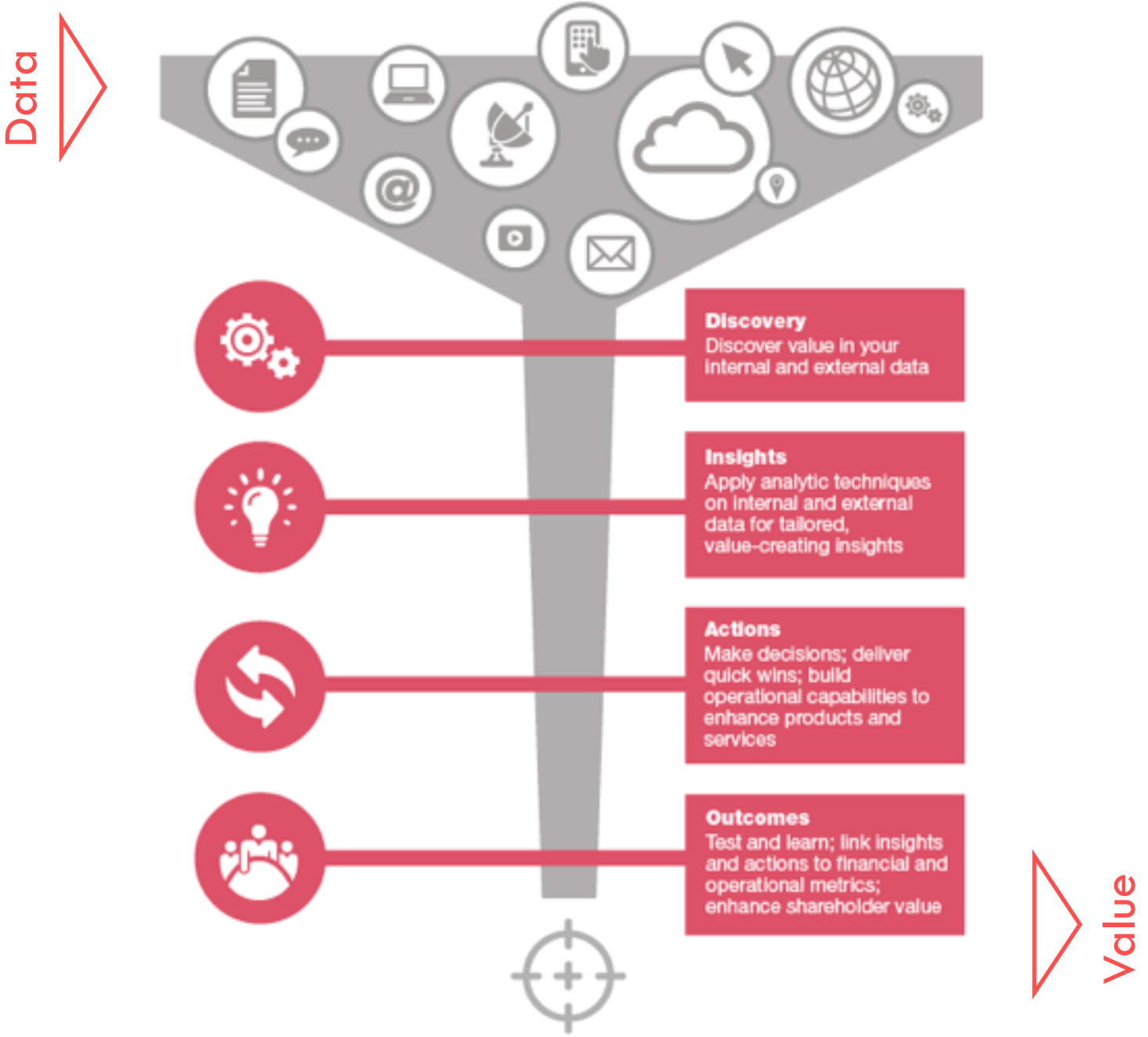
Traditional hit-lead generation: 2–3 years
GENTRL approach: <2 months

Driving Forces

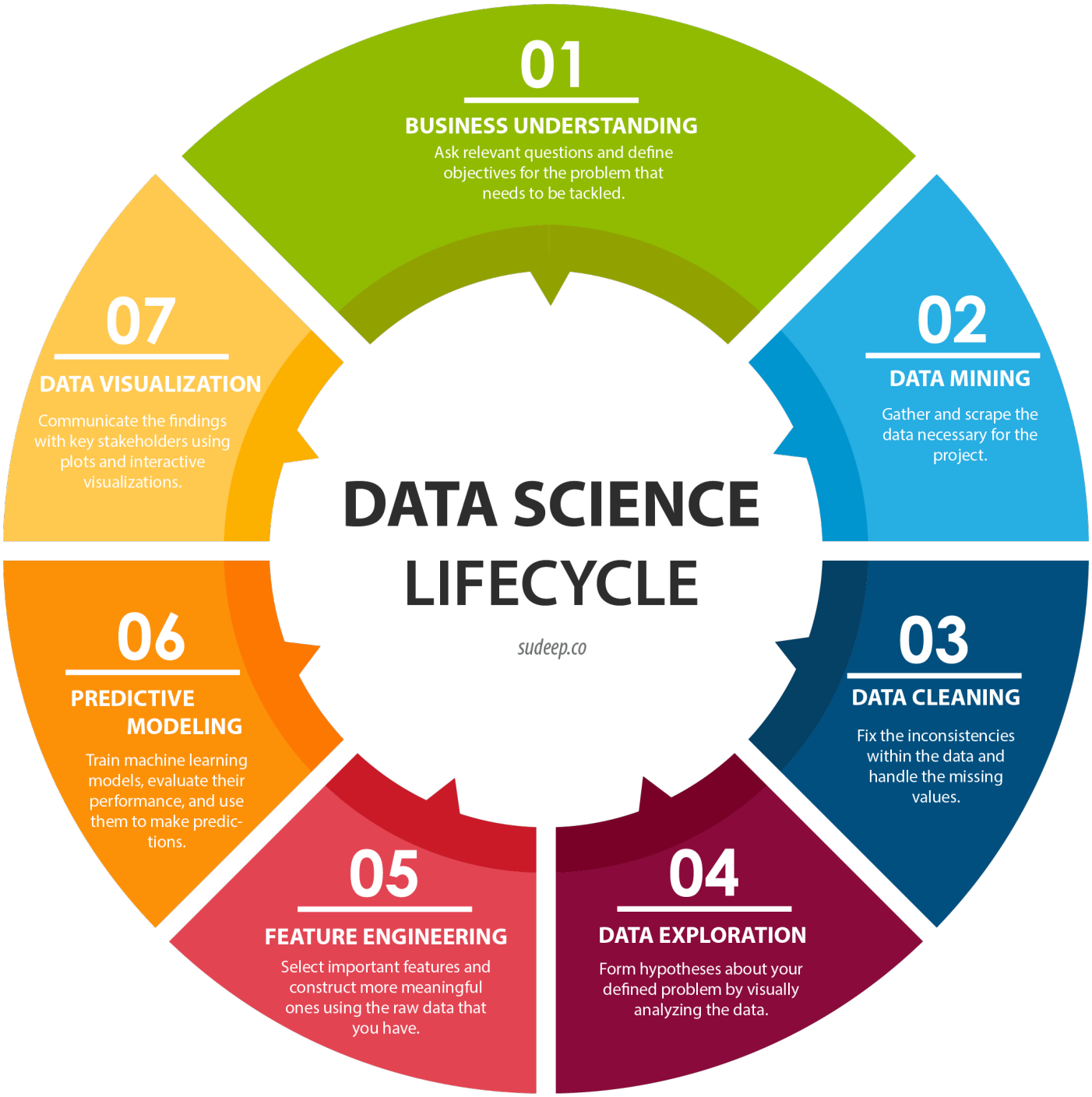
Data Deluge, Faster Computing & Cheaper Storage



General Idea



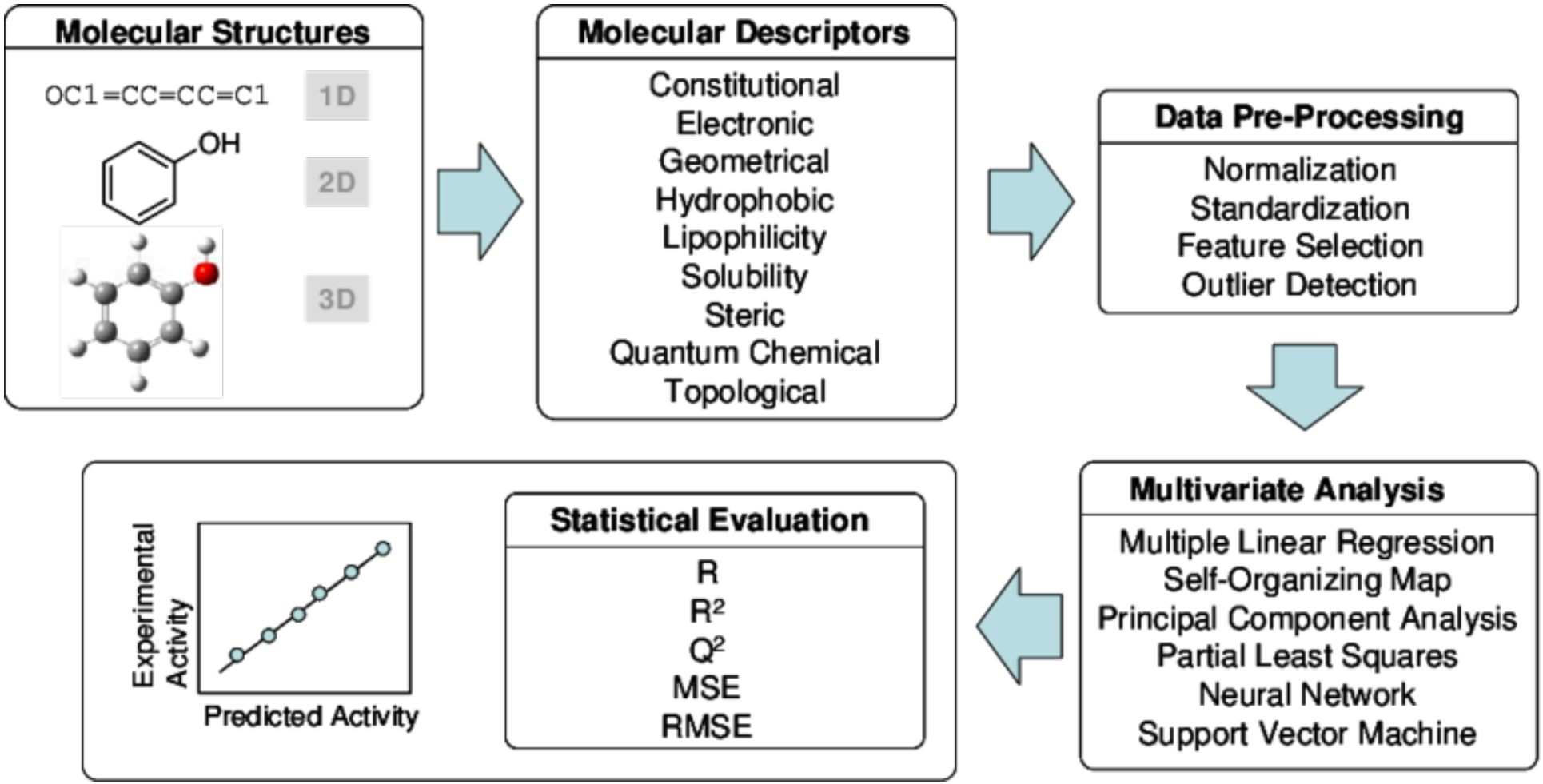
Protocol


01
BUSINESS UNDERSTANDING
Ask relevant questions and define objectives for the problem that needs to be tackled.
02
DATA MINING
Gather and scrape the data necessary for the project.
03
DATA CLEANING
Fix the inconsistencies within the data and handle the missing values.
04
DATA EXPLORATION
Form hypotheses about your defined problem by visually analyzing the data.
05
FEATURE ENGINEERING
Select important features and construct more meaningful ones using the raw data that you have.
06
PREDICTIVE MODELING
Train machine learning models, evaluate their performance, and use them to make predic-tions.
07
DATA VISUALIZATION
Communicate the findings with key stakeholders using plots and interactive visualizations.
DATA SCIENCE LIFECYCLE
sudeep.co

Step By Step

1. Identify Problem
2. Data Mining
3. Pre-processing
4. Exploratory Data Analysis
5. Feature Engineering
6. Predictive Modelling
7. Data Visualisation

Predicting biological activity based on Molecular Descriptors



Program

Machine Learning - General Concept

Data Mining

Pre-processing

Exploratory Data Analysis

Predictive Modelling

Data Visualisation

Data Mining

Collecting the relevant data to your problem.

(Chemical and Biological) Data can be:
- curated and stored in **databases / repositories**
- **manual curation**

Common programs retrieve data from databases:

mongoDB MySQL ORACLE PostgreSQL

Queries are usually the language used to access the relevant information from database(s)

Public / Private Repositories

ZINC Substances Catalogs Tranches Biological More

ZINC15

Welcome to ZINC, a free database of commercially-available compounds for virtual screening. ZINC contains over 230 million purchasable compounds in ready-to-dock, 3D formats. ZINC also contains over 750 million purchasable compounds you can search for analogs in under a minute.

UniProt UniProtKB

BLAST Align Retrieve/ID mapping Peptide search

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein seque

UniProtKB
UniProt Knowledgebase
Swiss-Prot (551,987)
Manually annotated and reviewed.
TrEMBL (66,905,753)
Automatically annotated and not reviewed.

UniRef
Sequence clusters

UniParc
Sequence archive

Proteomes

Supporting data
Literature citations
Taxonomy
Subcellular locations
Cross-ref. databases
Diseases
XXX
Keywords

DBAASP V2.702

DATABASE OF ANTIMICROBIAL ACTIVITY AND STRUCTURE OF PEPTIDES

NEWS

Monomer: 13398

Dimer: 171

Multi Peptide: 175

MarinLit

A database of the marine natural products literature

Number of Articles: 3 4 0 6 1 Last Update : 10-10-2019

MarinLit is a database dedicated to marine natural products research. The database was established in the 1970s by Professors John Blunt and Murray Munro at the University of Canterbury, New Zealand. It was designed as an in-house system to fulfil the needs of the University of Canterbury Marine Group and has evolved to contain unique searchable features and powerful dereplication tools. The extremely comprehensive range of data contained along with these powerful features makes MarinLit the database of choice for marine natural products researchers.

Get started with MarinLit
- Introduction to the database
- How to access
- Contact us

Sign in
with your subscriber username and password or via your home institution

Data resources at EMBL-EBI

Genes, genomes & variation

European Nucleotide Archive
European Variation Archive
European Genome-phenome Archive
Ensembl
Ensembl Genomes
GWAS Catalog
Metagenomics portal

Gene & protein expression

ArrayExpress
Expression Atlas
PRIDE

Protein sequences, families & motifs

InterPro
Pfam
UniProt

Molecular structures

Protein Data Bank in Europe
Electron Microscopy Data Bank

Chemical biology

ChEMBL
ChEBI

Systems

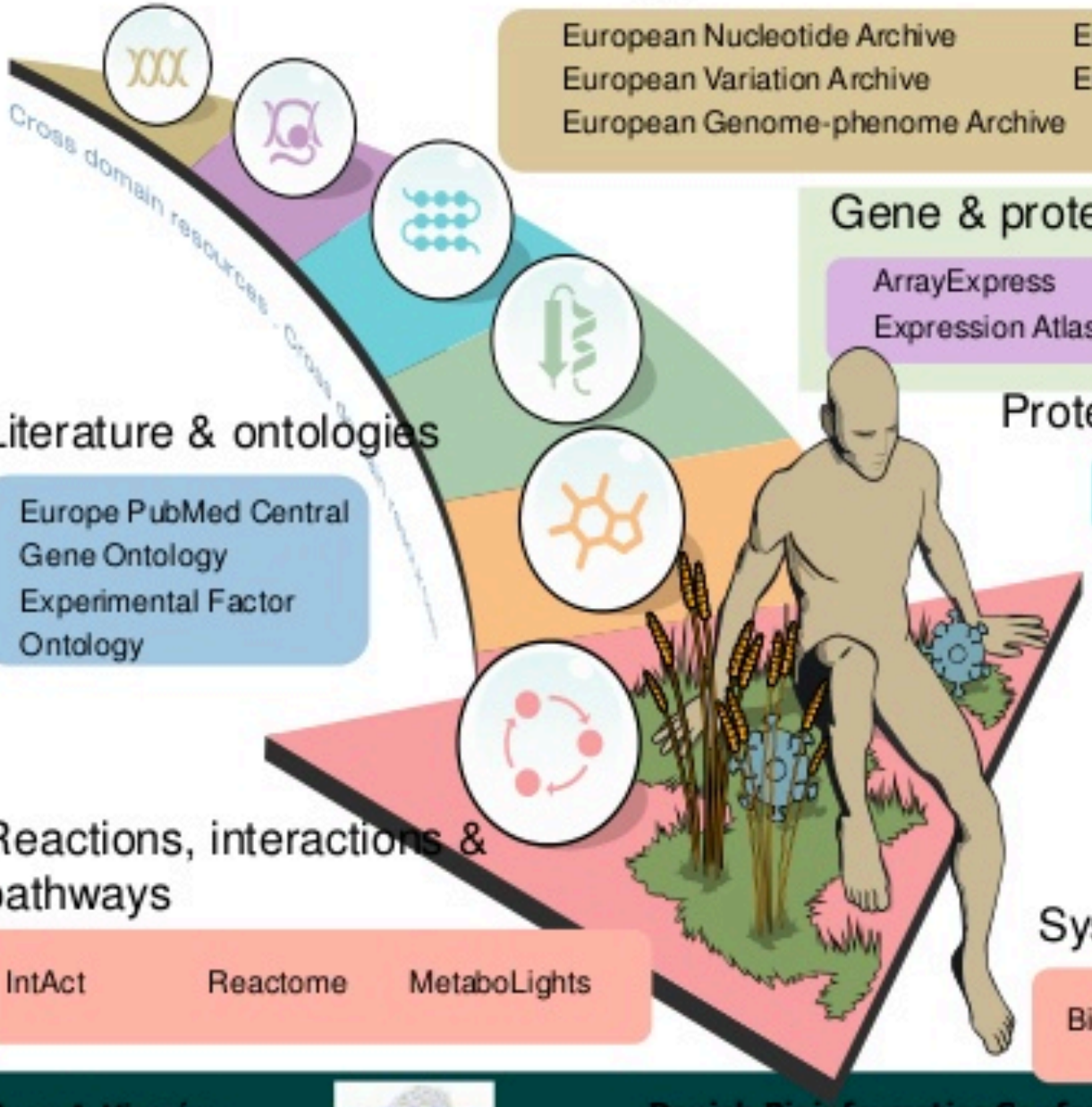
BioModels
Enzyme Portal
BioSamples

Literature & ontologies

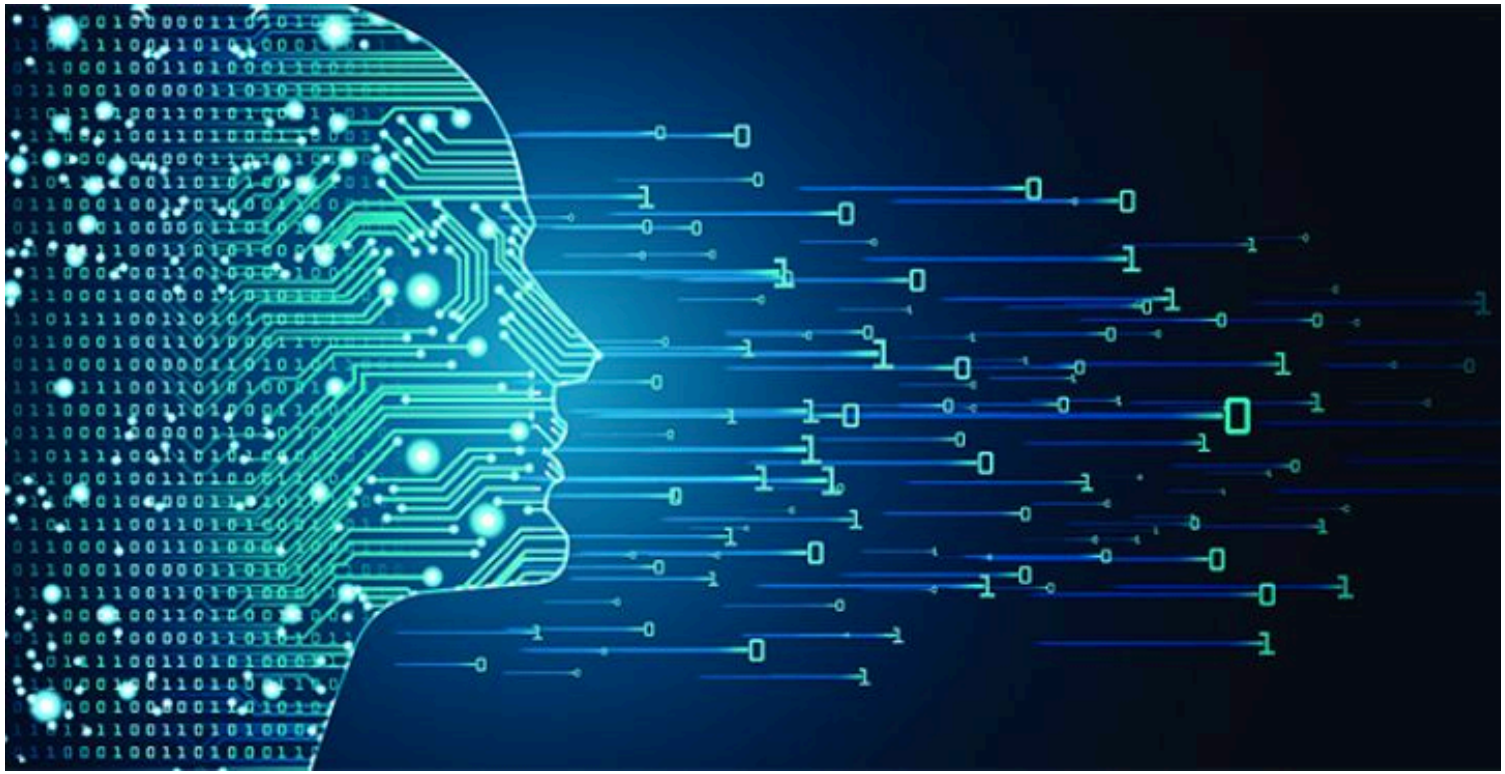
Europe PubMed Central
Gene Ontology
Experimental Factor Ontology

Reactions, interactions & pathways

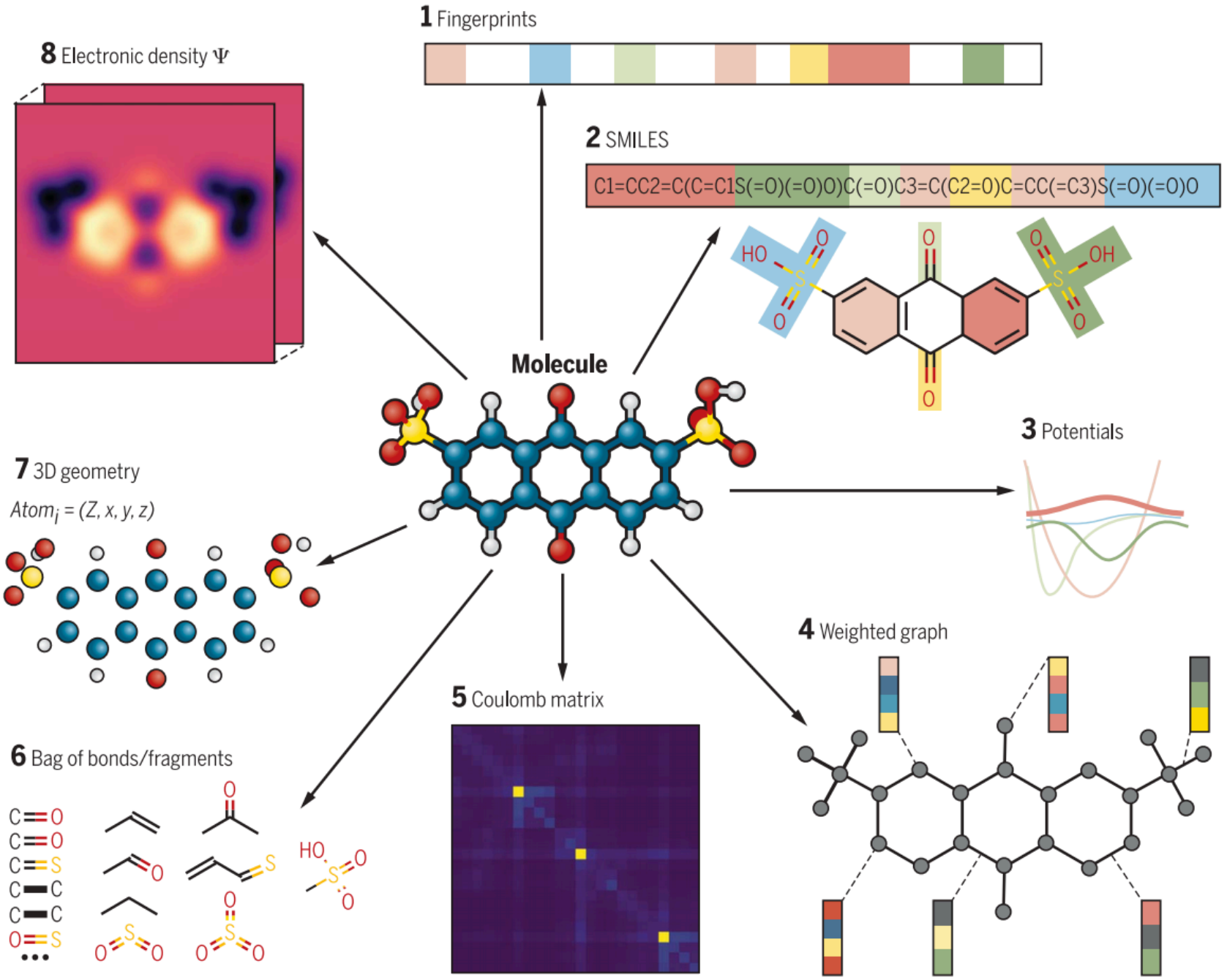
IntAct
Reactome
MetaboLights



Digitalization



Encoding Molecules



Encoding Molecules

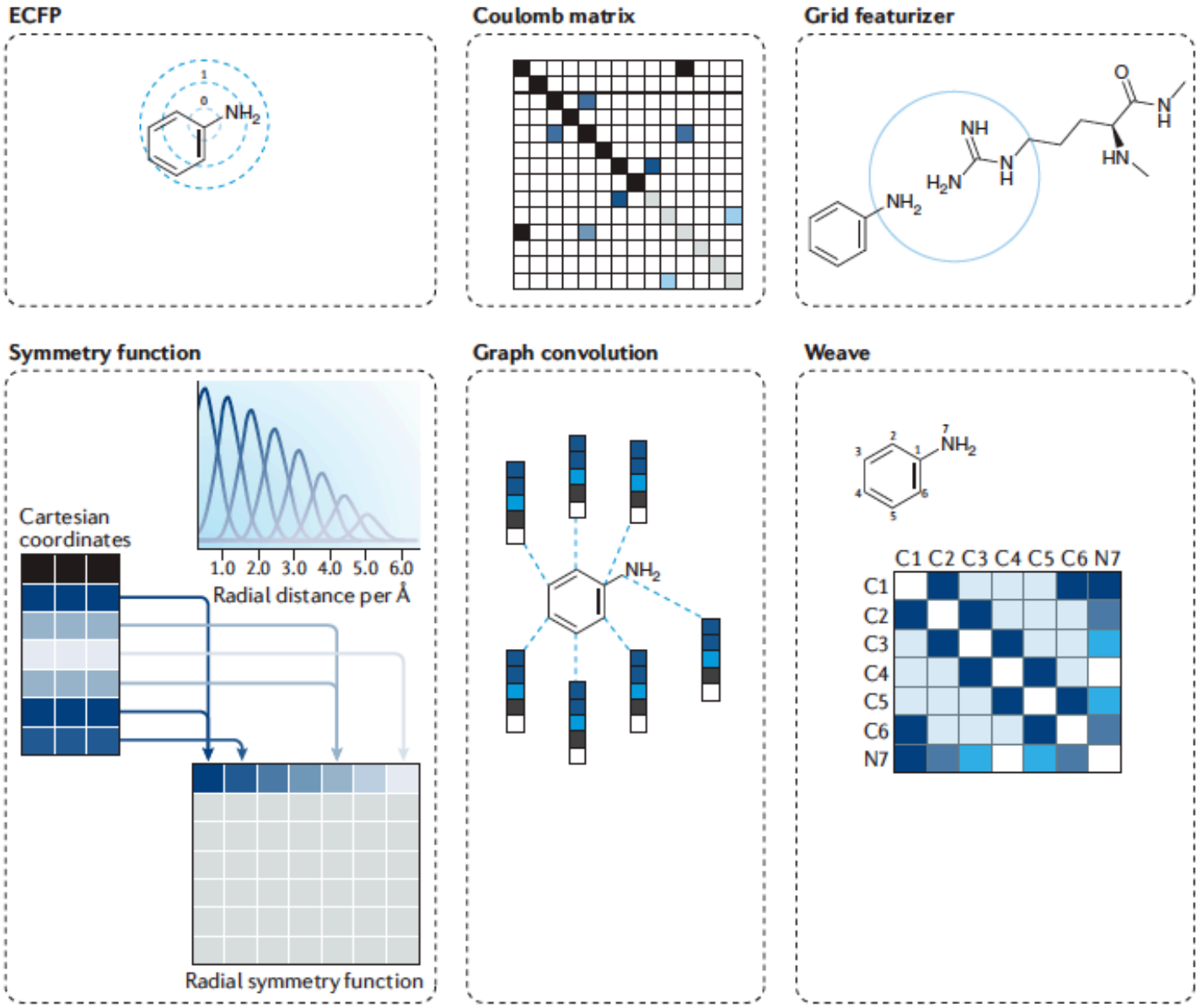


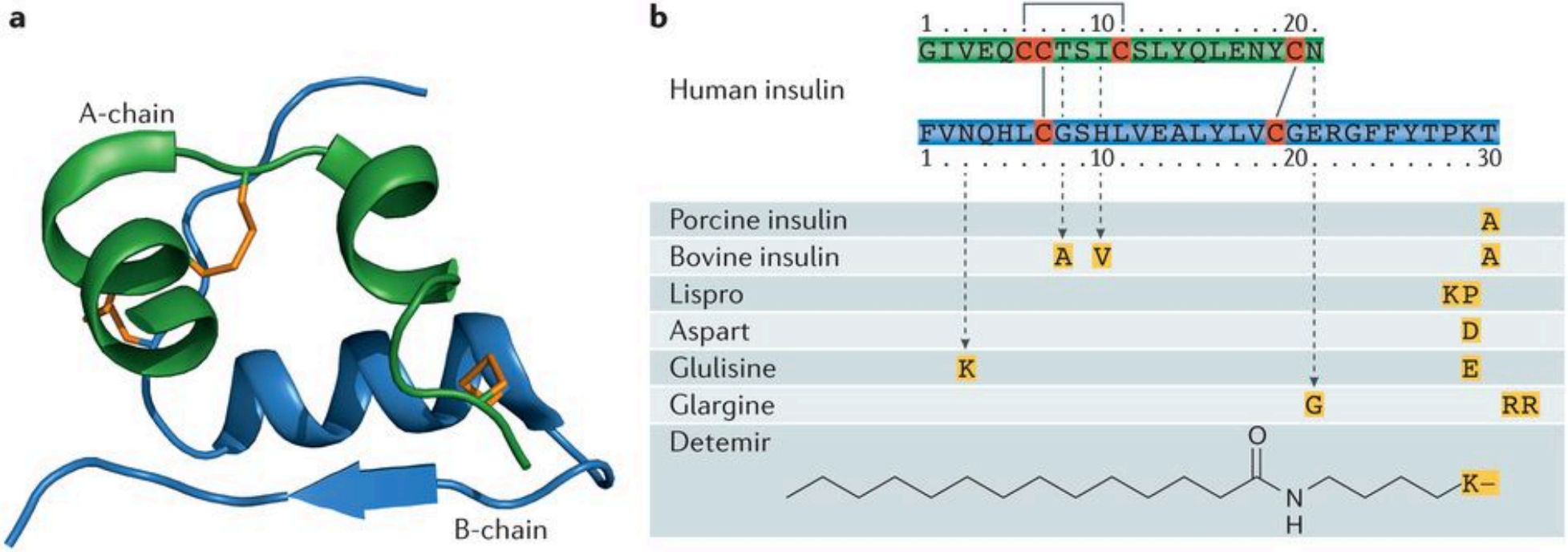
Fig. 3 | **The challenges of compound structure representation in machine learning models.**

Encoding Proteins

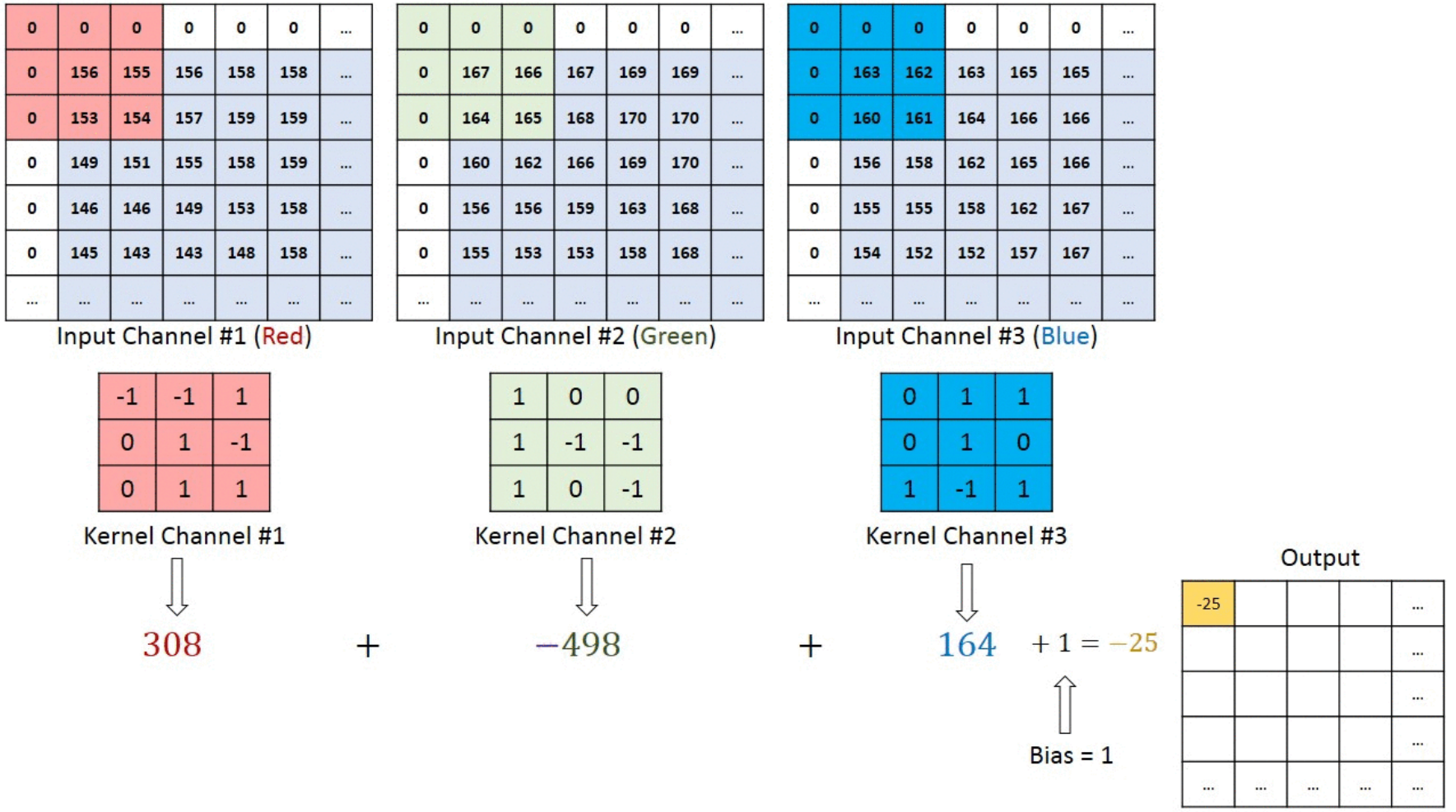
A protein P is represented by the linear chain given by its collapsed graph at amino acid level. A reduced molecular graph representation G(V,E,C) known as a *string signature* where;

V: residues a , E: contiguous in sequence, C: amino acid type

$$ {}^{h}\sigma(P) = \sum_{a \in A} {}^{h}\sigma(a) $$



Encoding Images



Program

General Concept

Data Mining

Pre-processing

Exploratory Data Analysis

Predictive Modelling

Data Visualisation

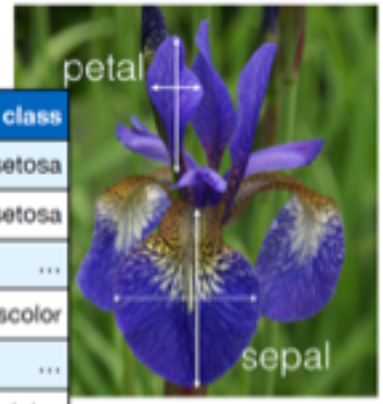
Nomenclature

IRIS

https://archive.ics.uci.edu/ml/datasets/Iris

Instances (samples, observations)

	sepal_length	sepal_width	petal_length	petal_width	class
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
...	...	...	...	...	...
50	6.4	3.2	4.5	1.5	veriscolor
...	...	...	...	...	...
150	5.9	3.0	5.1	1.8	virginica



Features (attributes, dimensions)

Classes (targets)

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1D} \\ x_{21} & x_{22} & \cdots & x_{2D} \\ x_{31} & x_{32} & \cdots & x_{3D} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{ND} \end{bmatrix}$$

$$\mathbf{y} = [y_1, y_2, y_3, \cdots y_N]$$

Feature Engineering

Process of creating, modifying or maintaining features making them more appropriate for their use in modelling and maximum information can be extracted out of them.

Pros

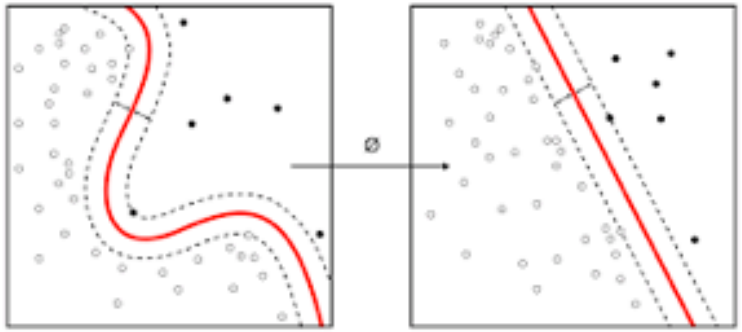
The better the variables that are analysed, the better and more reliable the results will be. *"Garbage In, Garbage Out"*

Cons

Time-consuming process

In-depth knowledge of features

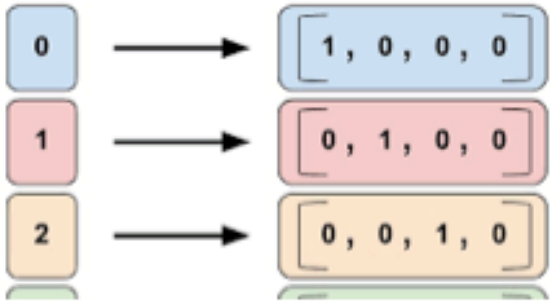
Compatibility features' appearance / model



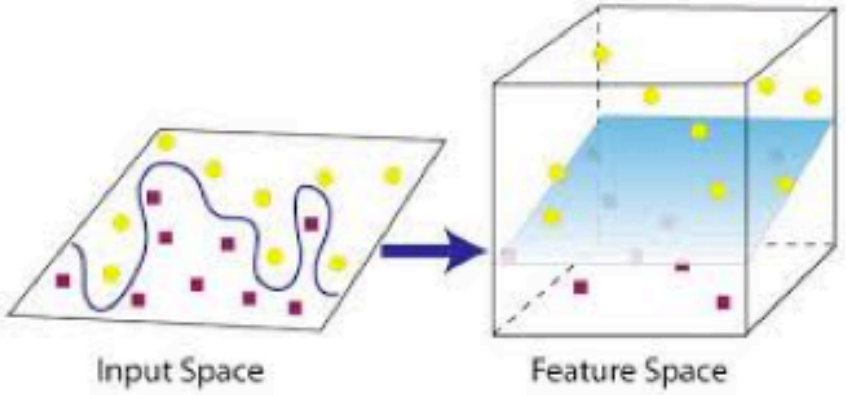
1. TRANSFORMATION

$$\frac{x - X_1}{max(X) - min(X)}$$

2. SCALING



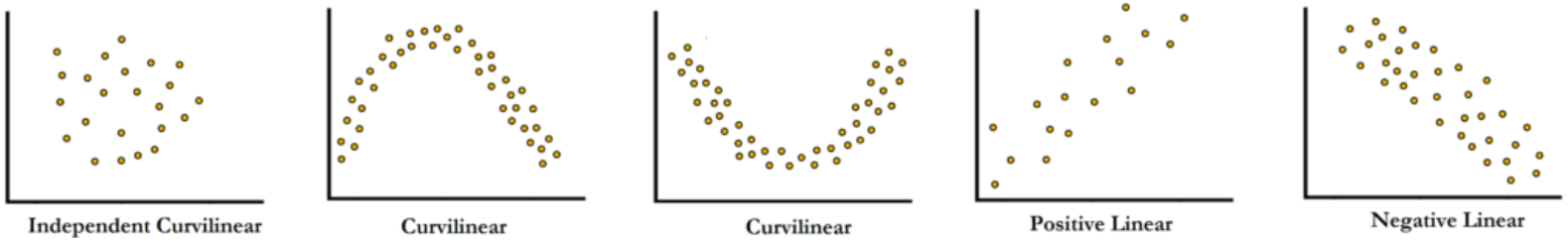
3. CONSTRUCTION



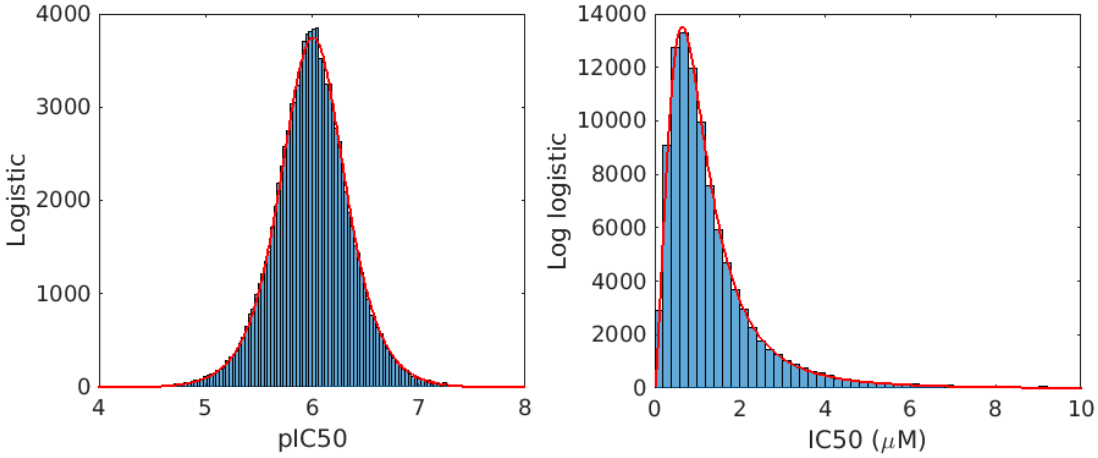
4. REDUCTION

1. Transformation

Replacement of a variable by a function that helps (often) to transform:
- non-linear relationships into linear relationships
- non-normal (skewed) distributions into normal distributions



1. Logarithm
2. Square / Cube Root
3. Arcsine
4. Box-Cox



2. Scaling

Many ML algorithms require feature scaling to prevent the model from giving greater weightage to certain attributes (variables) as compared to others.

- **Models** e.g. K-means requires Euclidean distance matrix
- **Gradient descent** (optimisation algorithm) - scaling reduce computing time
- **Feature Extraction** e.g. scaling prevents emphasising variables in PCA or LDA

1. Min-Max Scaling (**Normalisation**)
Scales ranges: [0, 1] or [-1, 1]

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

$$Xnormalized_{[-1,1]} = \frac{Xcurrent - ((X\,\max + X\,\min)/2)}{(X\,\max - X\,\min)/2}$$

2. Z score Normalisation (**Standardisation**)
Normal distribution properties
with mean = 0, standard deviation = 1

$$z = \frac{x - \mu}{\sigma}$$

3. Construction

Generating new variables (features) based on existing variables to improve reliability of (linear) predictive models or to reduce noise (random data points)

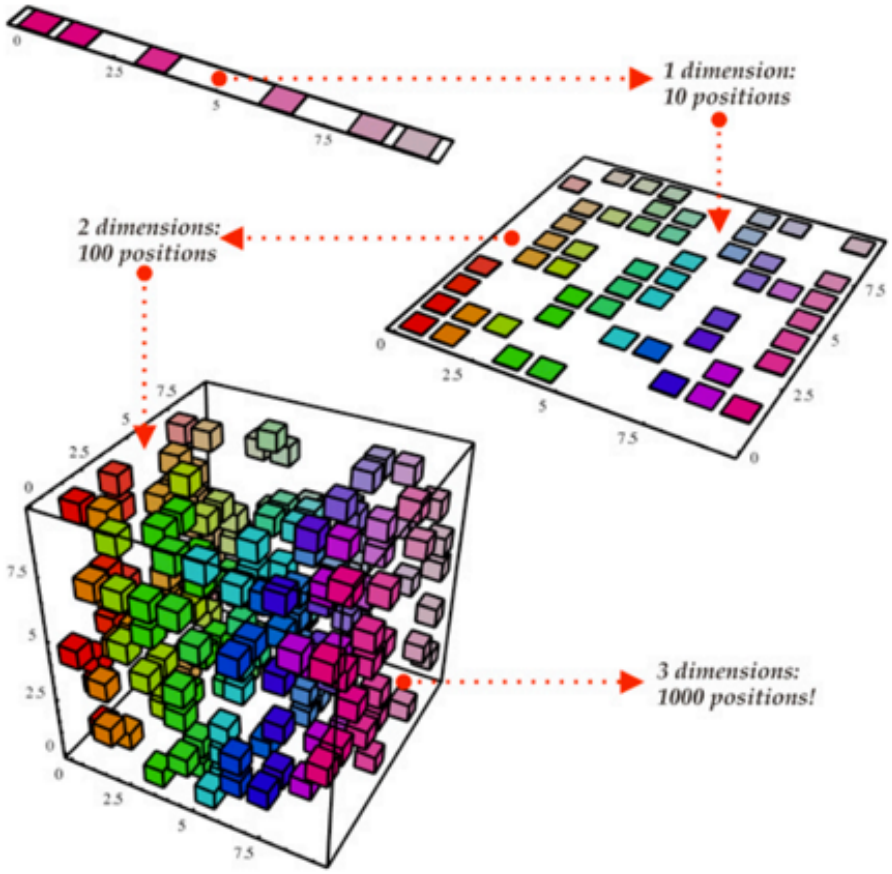
1. **Binning** e.g. People age, IQ levels correlations
2. **Encoding** e.g. One-hot encoding countries
3. **Derived Variables** e.g. Key Performance Indicators (KPI)

Derived Variables / Features

4. Reduction

Dimension(ality) reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables. It can be divided into **feature selection** and **feature extraction**.

1. **Unsupervised methods**
2. **Supervised methods**



Outlier Detection

Outliers are extreme values that deviate from other observations on data , they may indicate a variability in a measurement, experimental errors or a novelty. In other words, an outlier is an observation that diverges from an overall pattern on a sample.

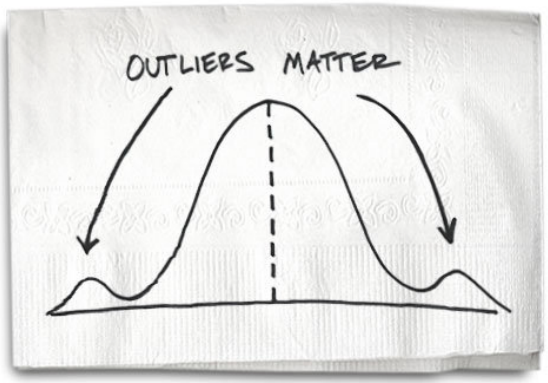
1. **Univariate outliers** can be found when looking at a distribution of values in a single feature space.

2. **Multivariate outliers** can be found in a n-dimensional space (of n-features).

3. **Point outliers** are single data points that lay far from the rest of the distribution.

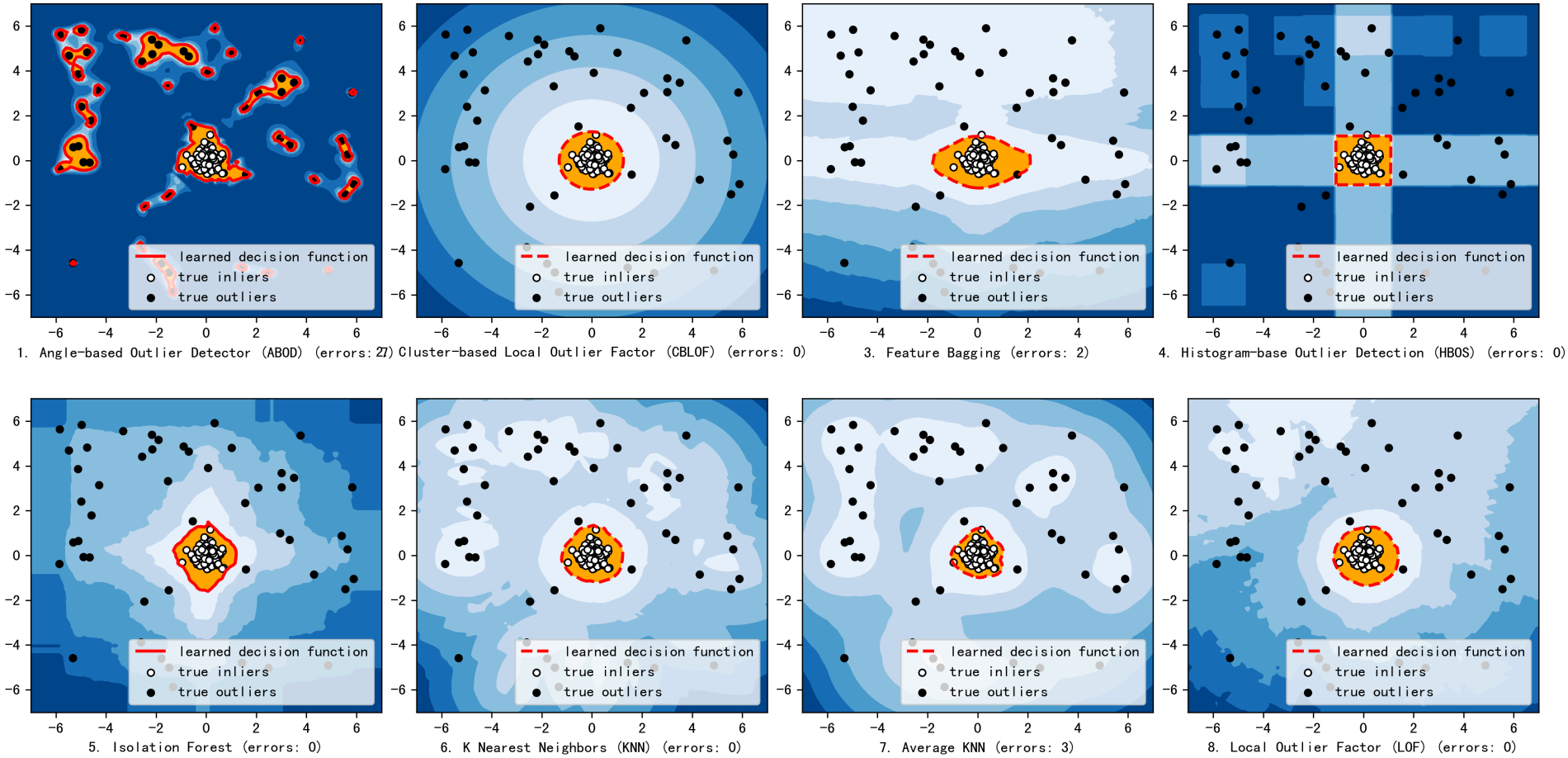
4. **Contextual outliers** can be noise in data
   e.g. *punctuation symbols / text analysis*
   background noise signal / speech recognition

5. **Collective outliers** can be subsets of novelties in data
   e.g. *signal that may indicate the discovery of new phenomena*



Detection Methods

- Z-Score or Extreme Value Analysis (**parametric**)
- Probabilistic and Statistical Modelling (**parametric**)
- Linear Regression Models (PCA, LMS)
- Proximity Based Models (**non-parametric**)
- Information Theory Models
- High Dimensional Outlier Detection Methods (high dimensional sparse data)



Program

General Concept

Data Mining

Pre-processing

Exploratory Data Analysis

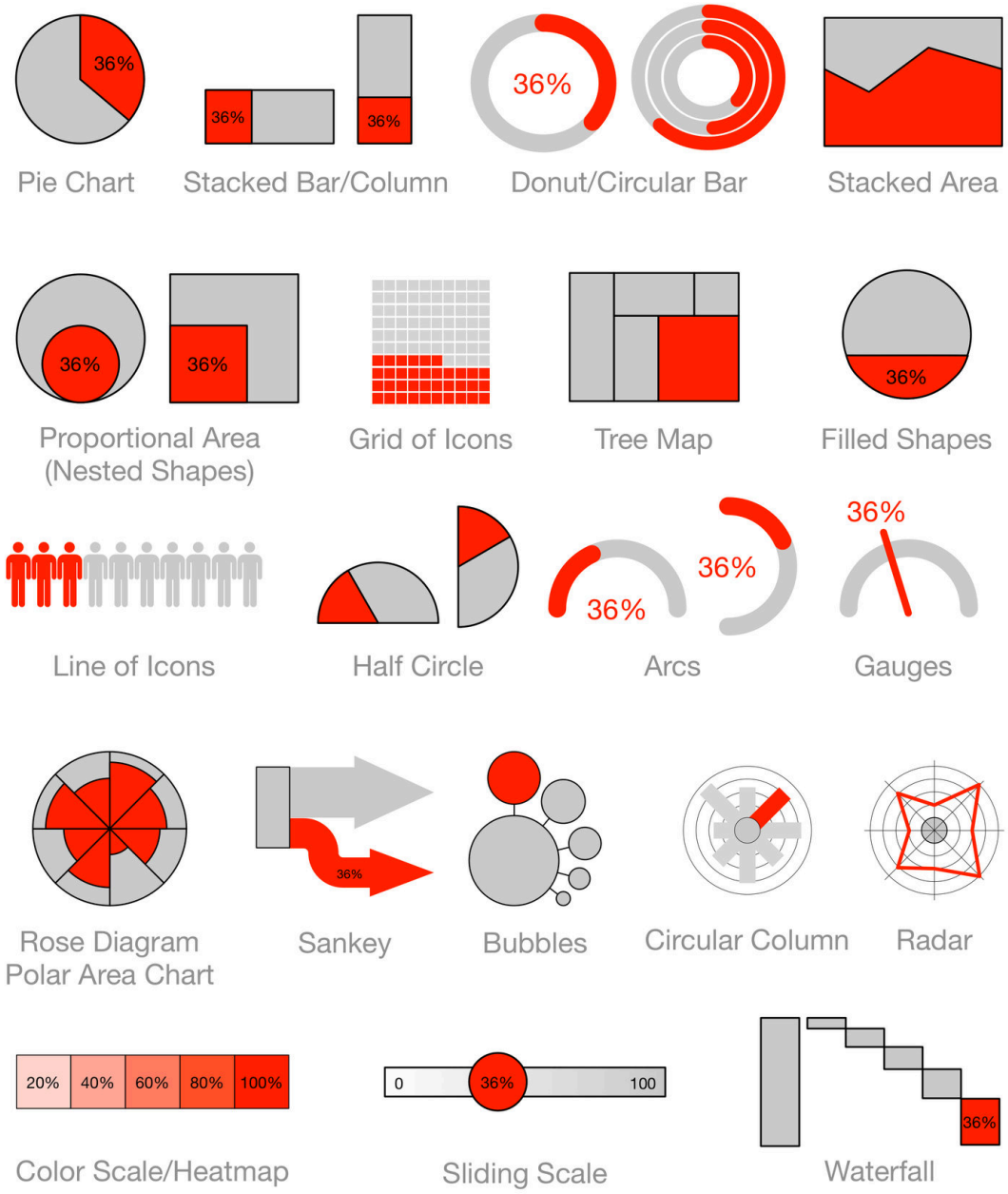
Predictive Modelling

Data Visualisation

Telling a story

Exploratory Data Analysis

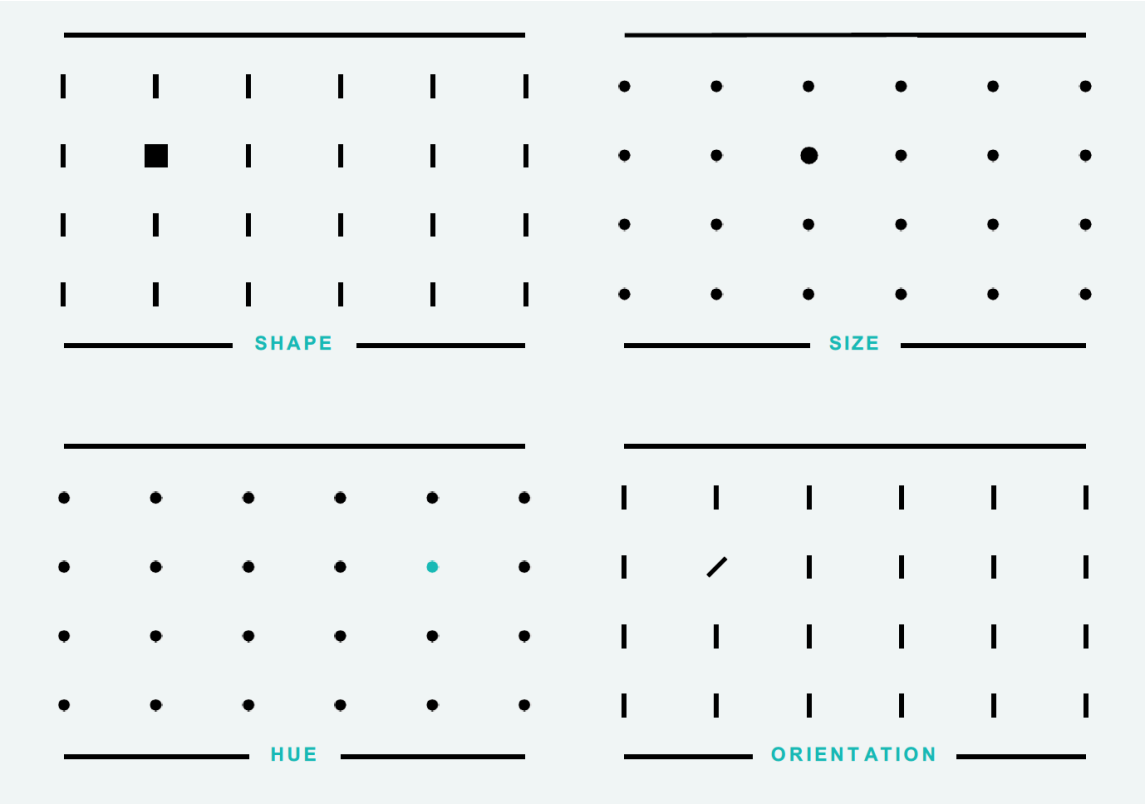
In statistics, **exploratory data analysis** (EDA) is an approach to analyzing **data** sets to summarize their main characteristics, often with visual methods



Our Brains Love Visuals

Through the visual system, the human brain quickly recognizes, stores and recalls images, seamlessly and subconsciously cementing ideas in long-term memory.

Our brain gathers information through pre-attentive processing of visual cues in our environment, which we unconsciously absorb and filter—within 250 milliseconds.

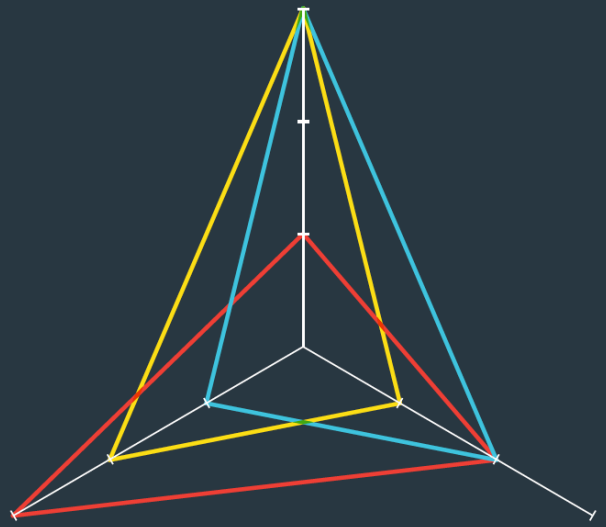


Notice that your eye is naturally drawn to these variations on the left.

VALUE OF VISUALIZATION

- Academic/Scientific
- Marketing
- Editorial

APPEAL Well-designed information is stimulating, attractive and engaging. These qualities pique interest even before information is processed. Aesthetics are not superficial; they are how you get people's attention.



COMPREHENSION The brain is pre-wired to automatically interpret relationships between objects, allowing for instant comprehension with minimal effort. Representing these relationships visually, as opposed to merely describing them, means that your message is understood quickly, clearly and with significantly greater joy.

RETENTION Visualizations trigger us to pull information from our long-term memory, allowing for rapid connections to already stored information, which help to cement the concept in the brain.

Chart Suggestions—A Thought-Starter

What would you like to show?

Comparison

- Among Items
  - Two Variables per Item
    - Variable Width Column Chart
  - One Variable per Item
    - Many Categories
      - Table or Table with Embedded Charts
    - Few Categories
      - Many Items
        - Bar Chart
      - Few Items
        - Column Chart
- Over Time
  - Many Periods
    - Cyclical Data
      - Circular Area Chart
    - Non-Cyclical Data
      - Line Chart
  - Few Periods
    - Single or Few Categories
      - Column Chart
    - Many Categories
      - Line Chart

Relationship

- Two Variables
  - Scatter Chart
  - Bubble Chart
- Three Variables
  - Bubble Chart

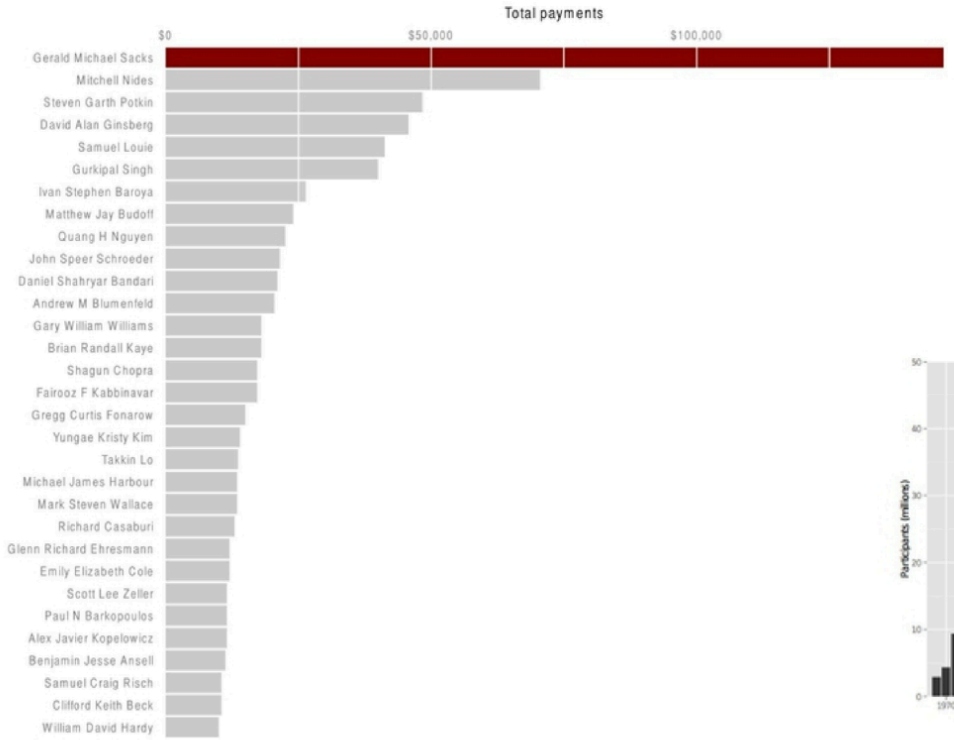
Distribution

- Single Variable
  - Few Data Points
    - Column Histogram
  - Many Data Points
    - Line Histogram
- Two Variables
  - Scatter Chart
- Three Variables
  - 3D Area Chart

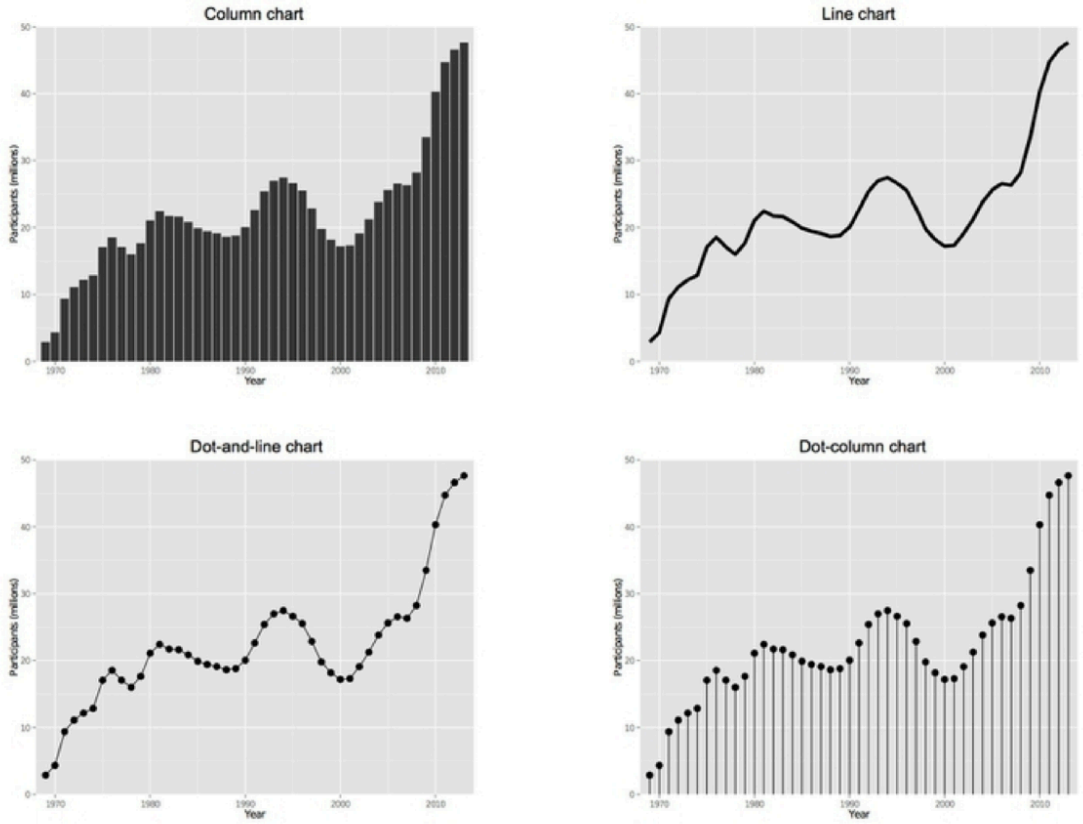
Composition

- Changing Over Time
  - Few Periods
    - Only Relative Differences Matter
      - Stacked 100% Column Chart
    - Relative and Absolute Differences Matter
      - Stacked Column Chart
  - Many Periods
    - Only Relative Differences Matter
      - Stacked 100% Area Chart
    - Relative and Absolute Differences Matter
      - Stacked Area Chart
- Static
  - Simple Share of Total
    - Pie Chart
  - Accumulation or Subtraction to Total
    - Waterfall Chart
  - Components of Components
    - Stacked 100% Column Chart with Subcomponents

Comparison

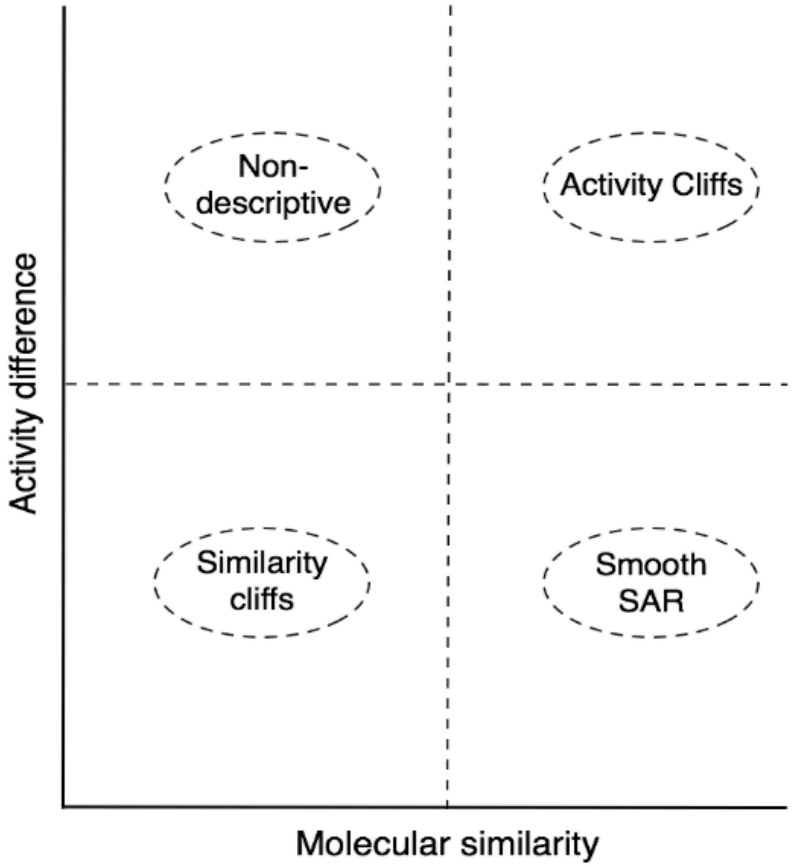


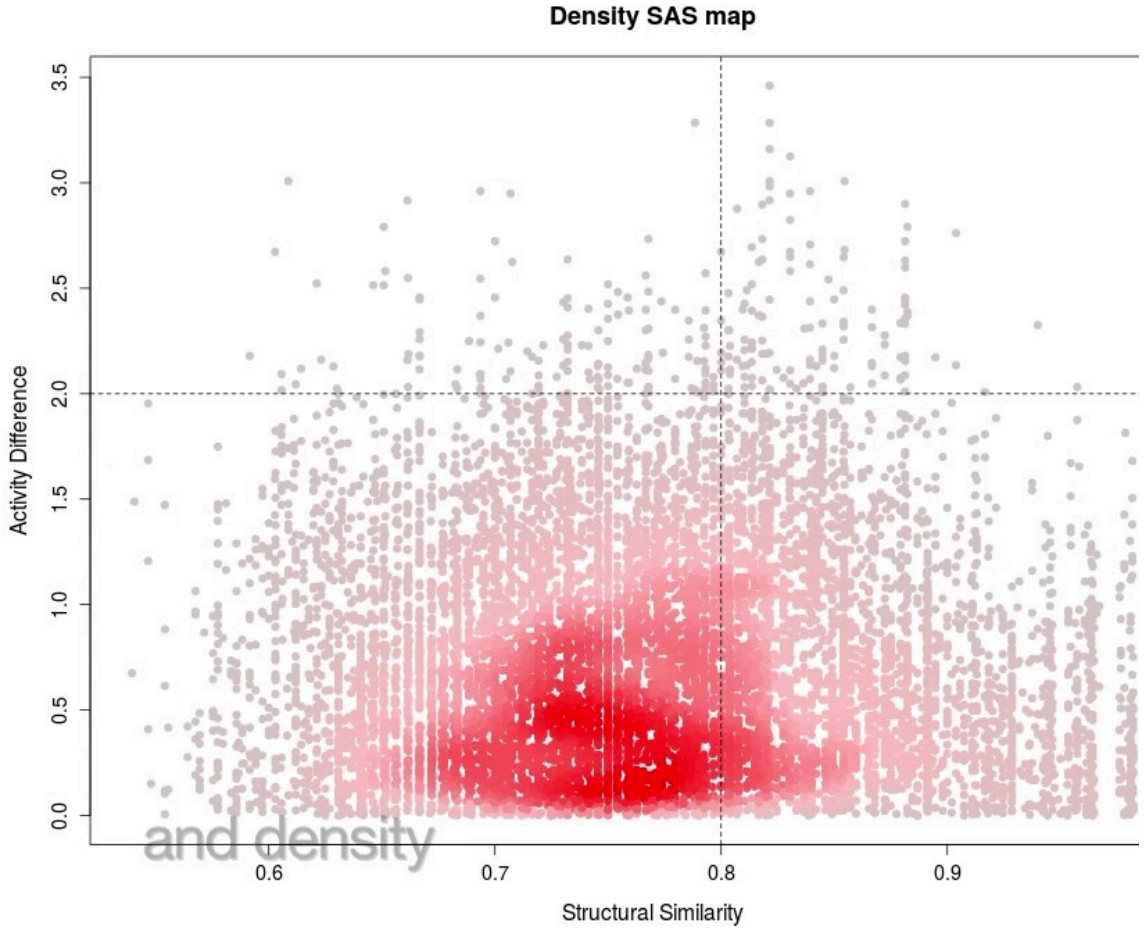
Pfizer paid doctors



Gvt food stamps 1969-2014

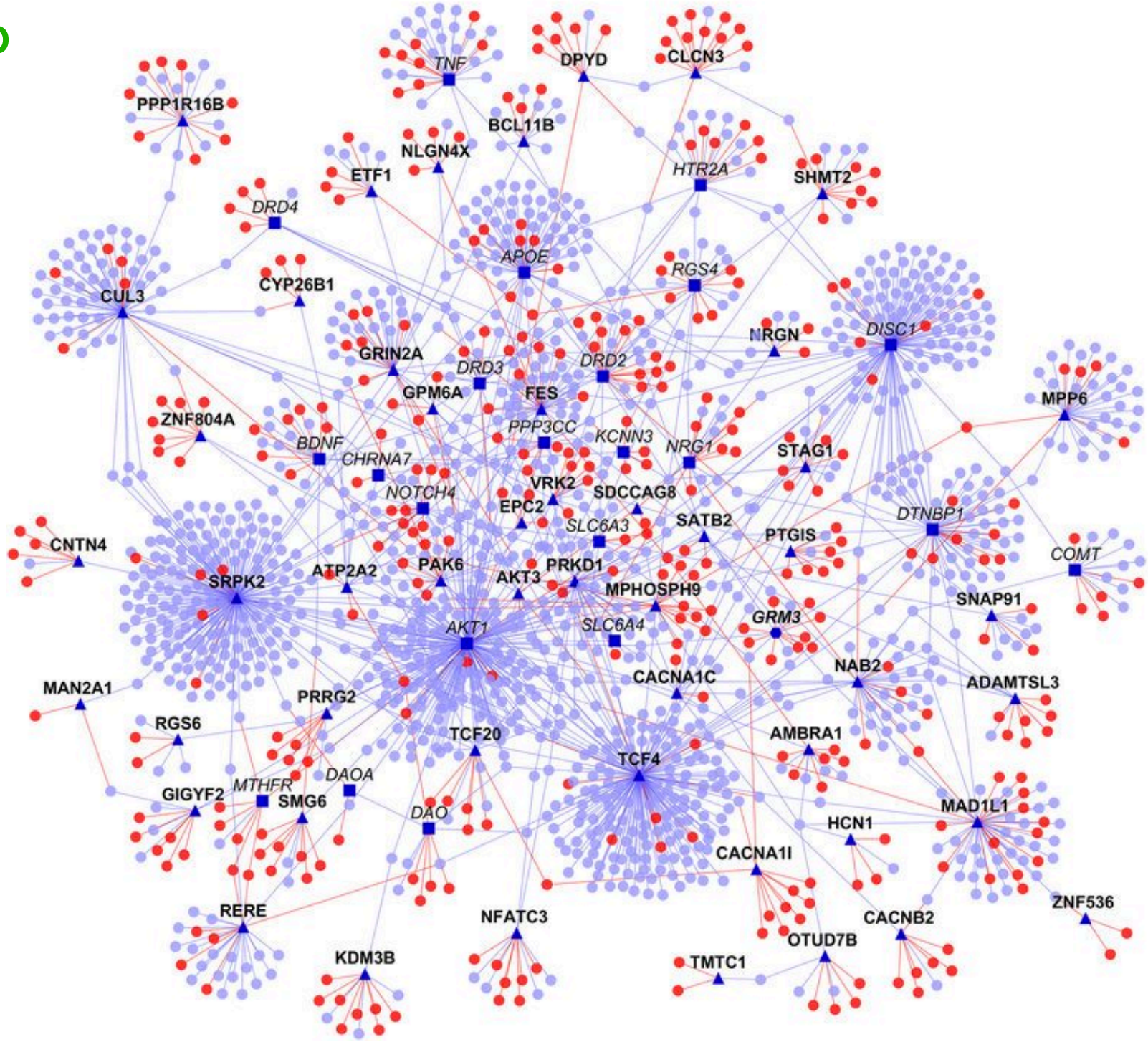
Relationship



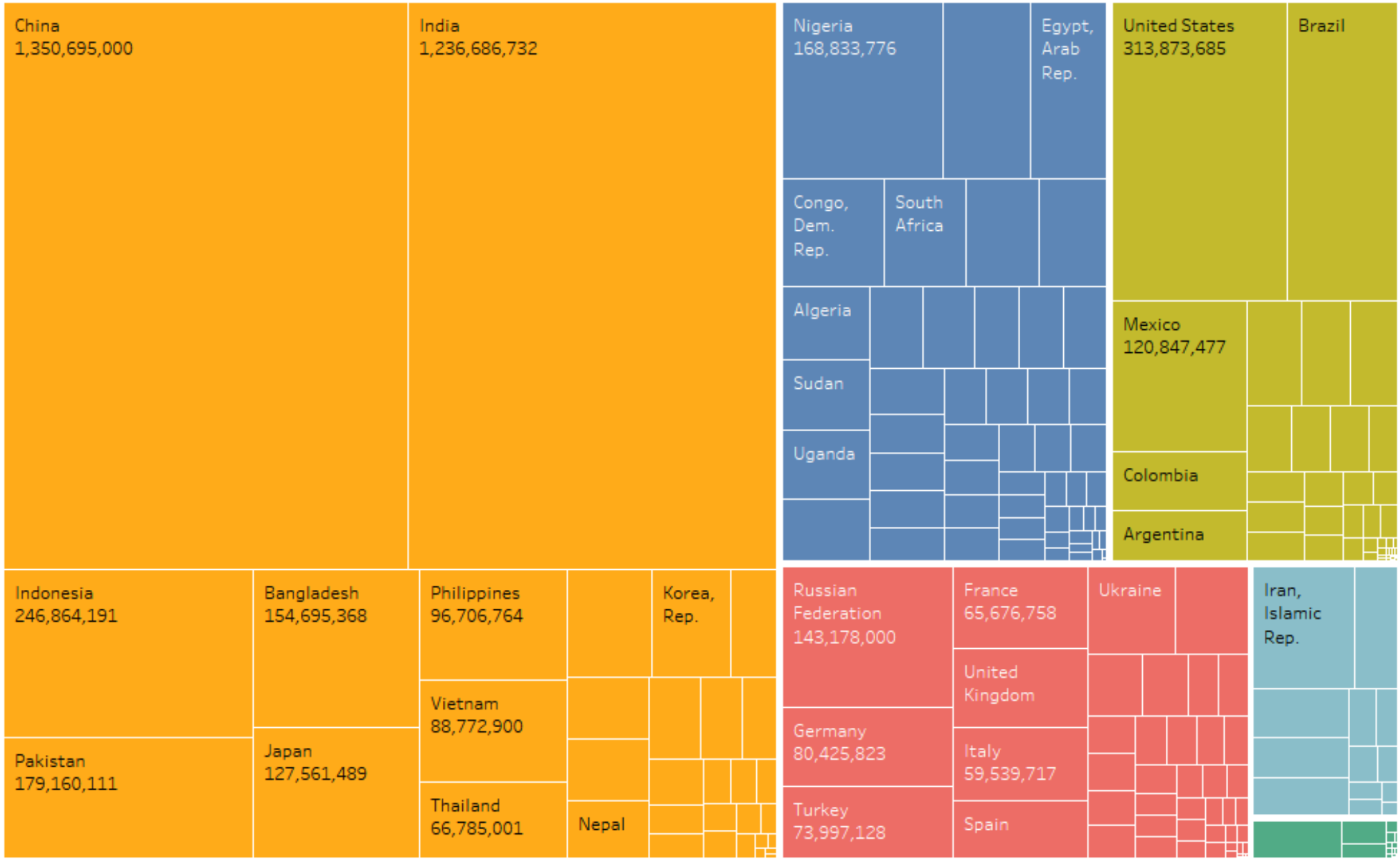


https://www.difacquim.com/d-tools/

Relationship



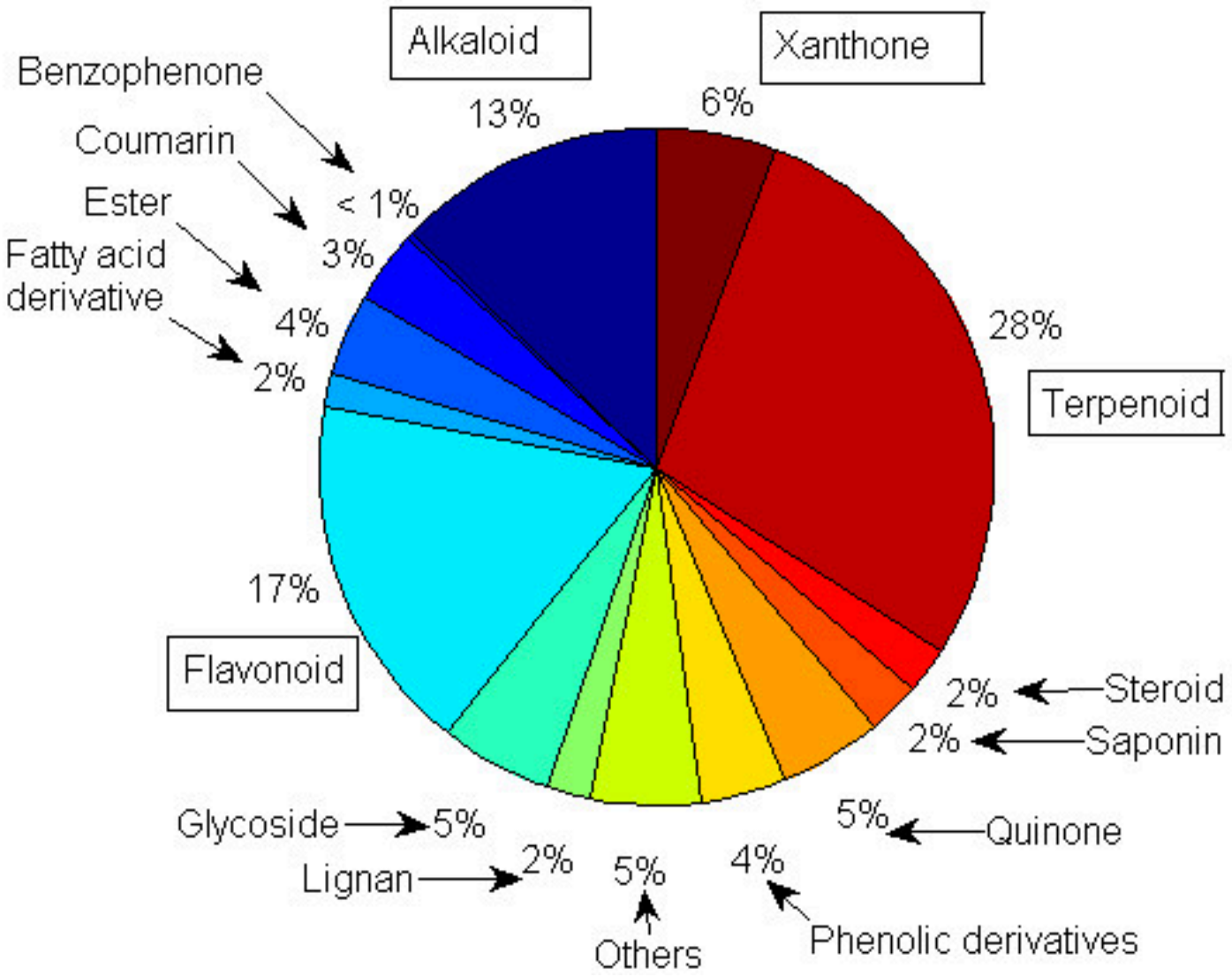
Composition



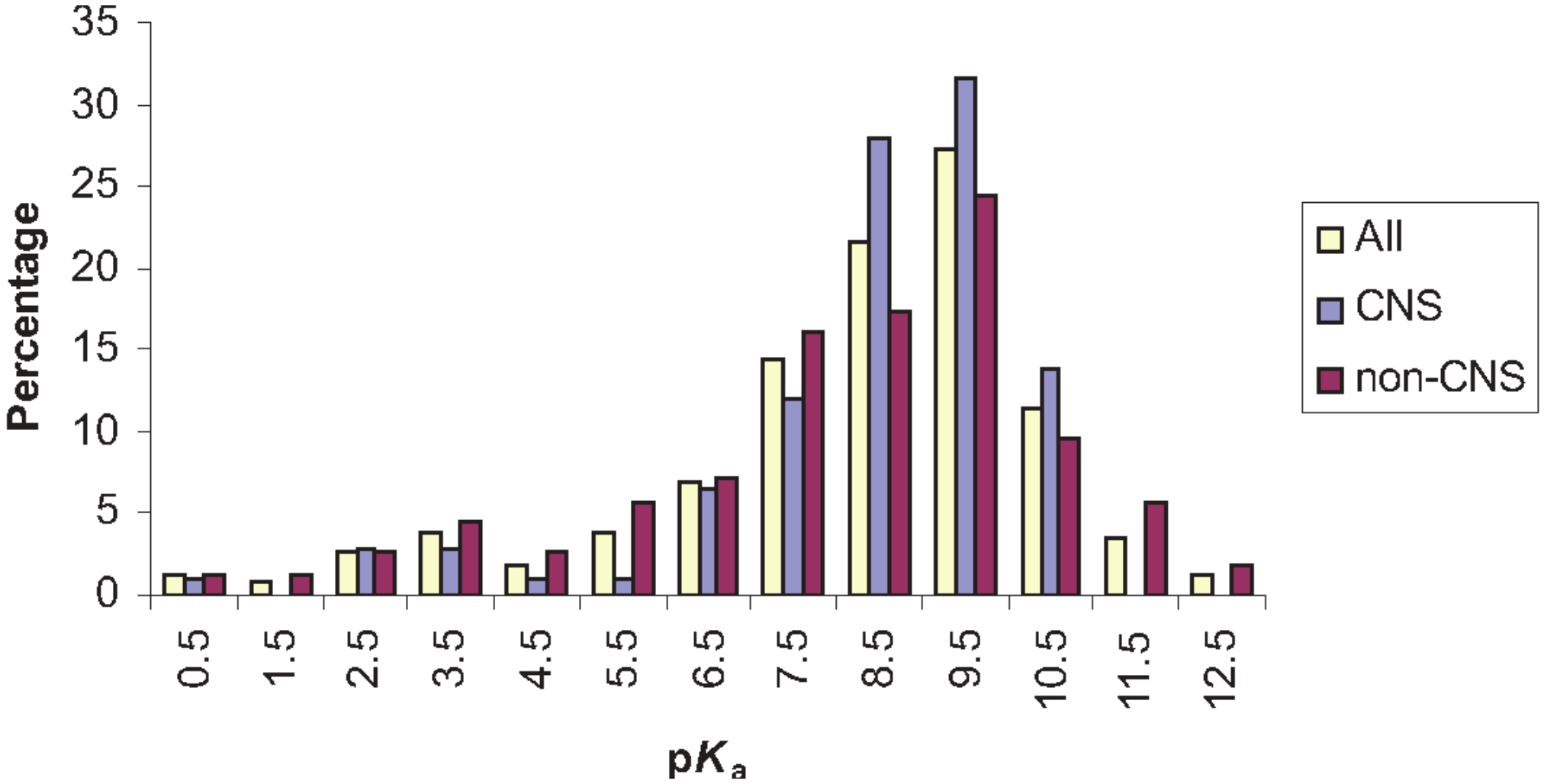
Region

Asia | Africa | The Americas | Europe | Middle East | Oceania

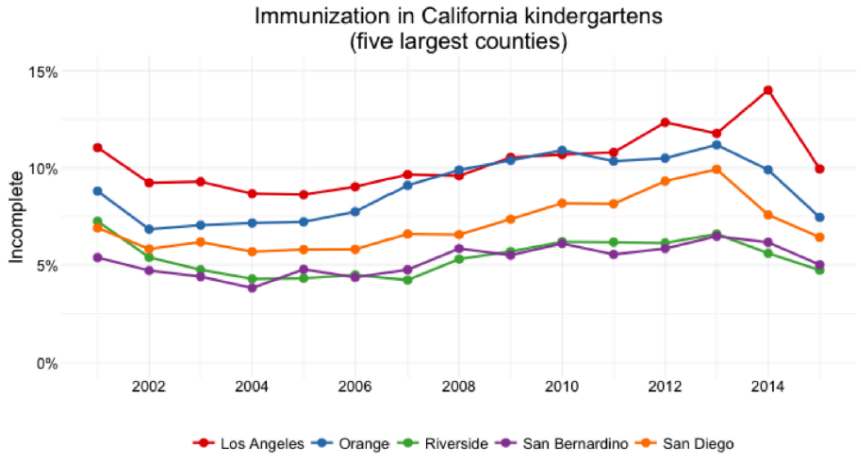
Composition

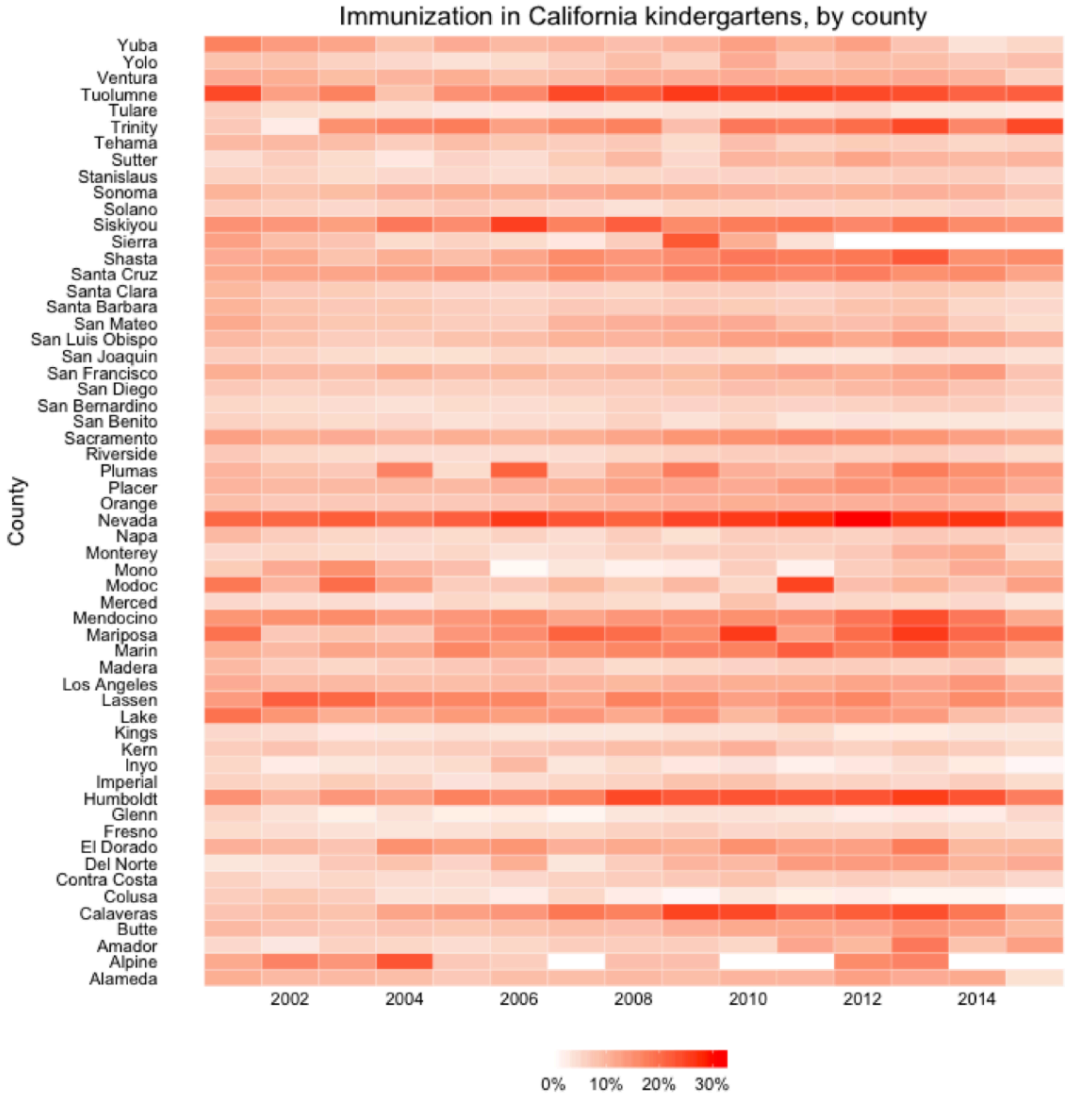


Distribution



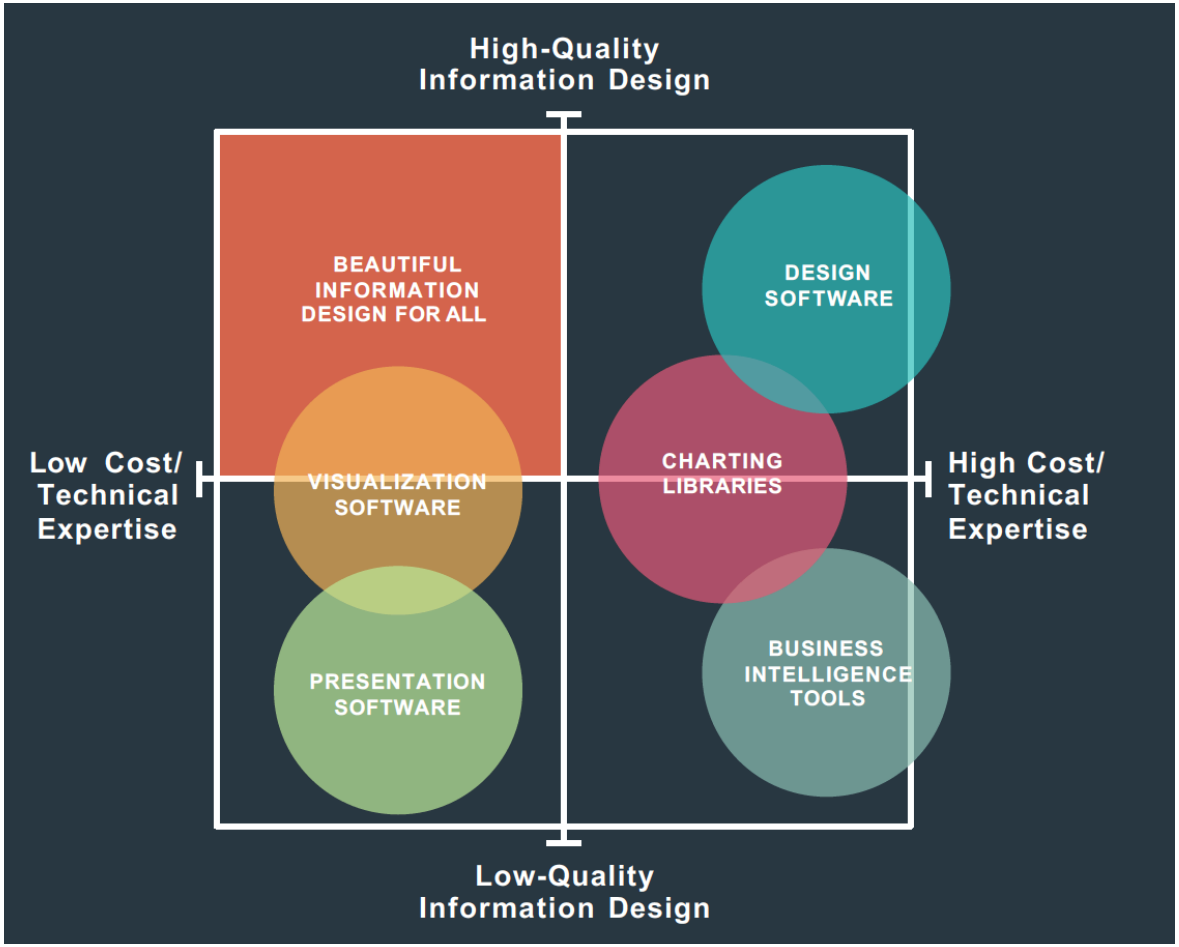
Distribution



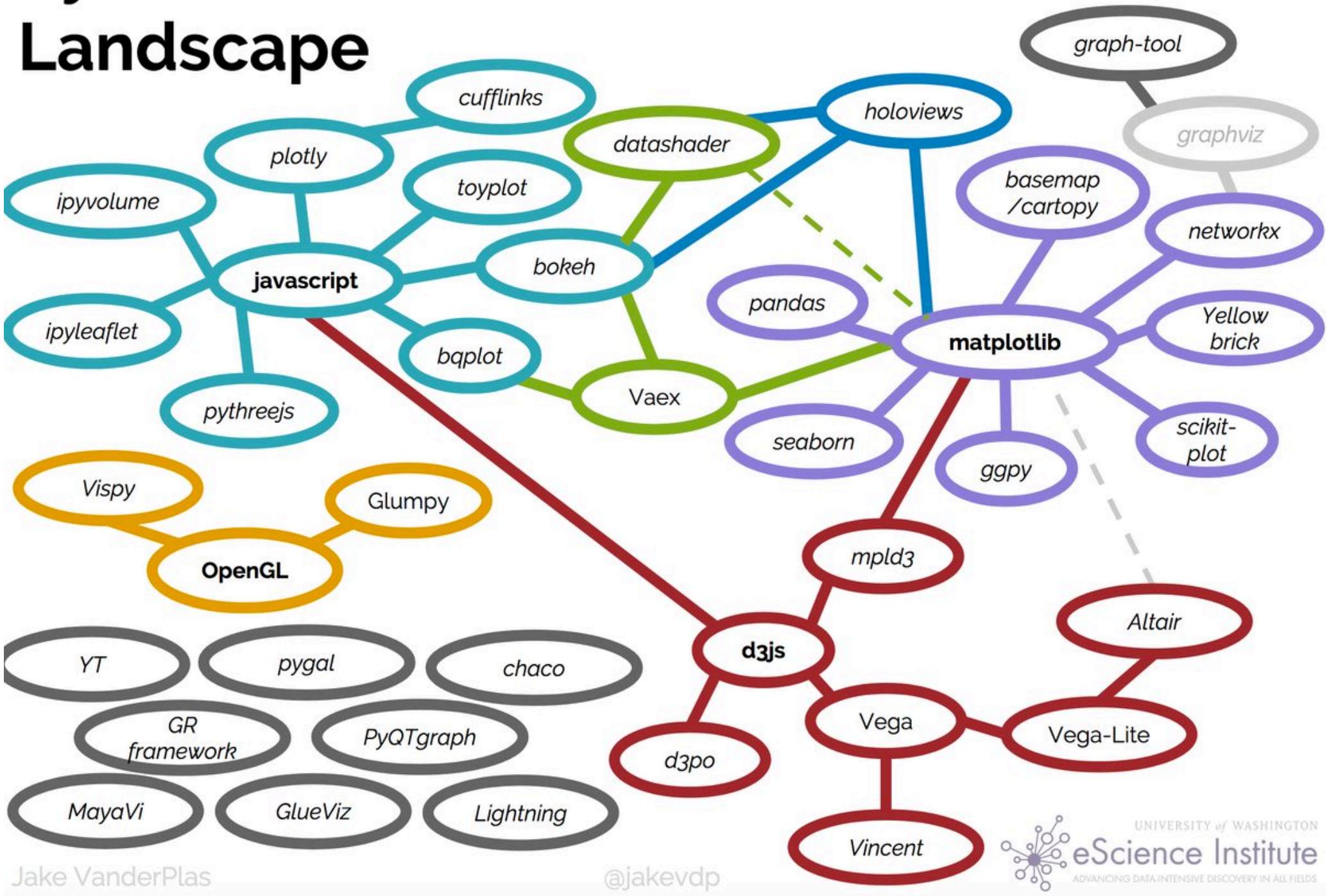


Design Tools

The traditional tools and design programs available are incomplete, challenging to master or limited in their aesthetic.Visualization software is an emerging field, providing better design tools that are easier to use.

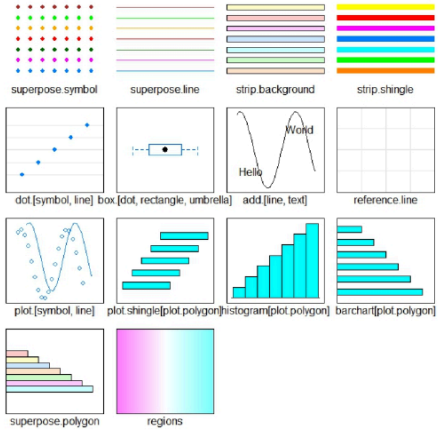


Python's Visualization Landscape

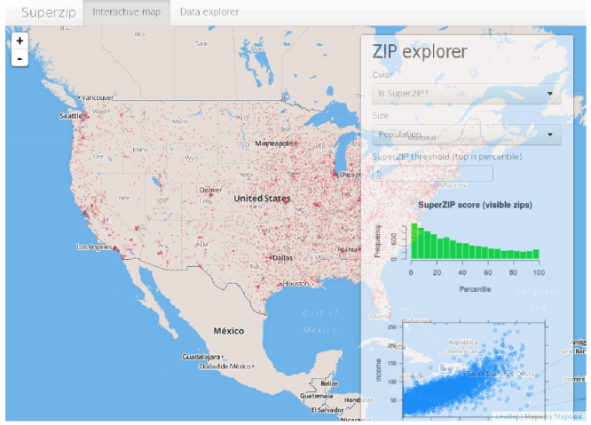


R Data Viz Packages

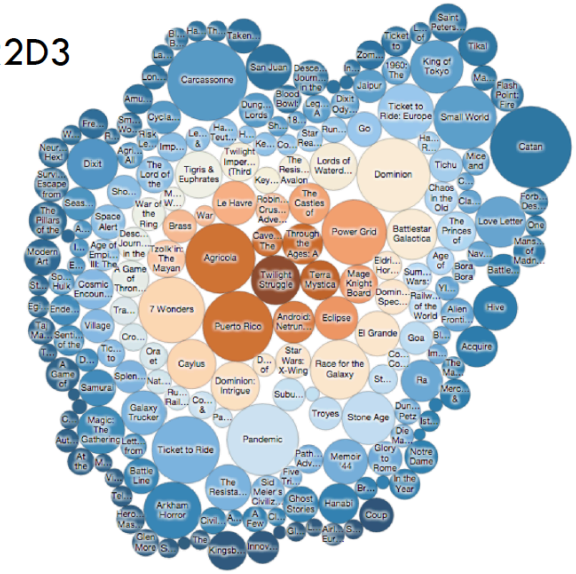
Lattice



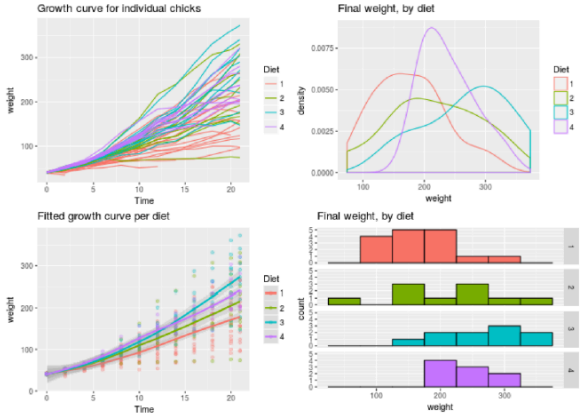
Rshiny



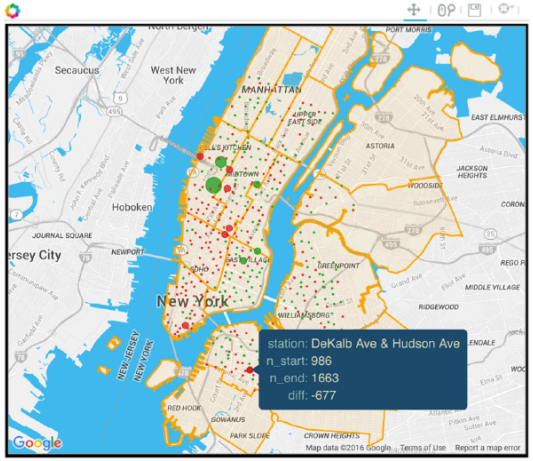
R2D3



Ggplot2



Rbokeh



Program

General Concept

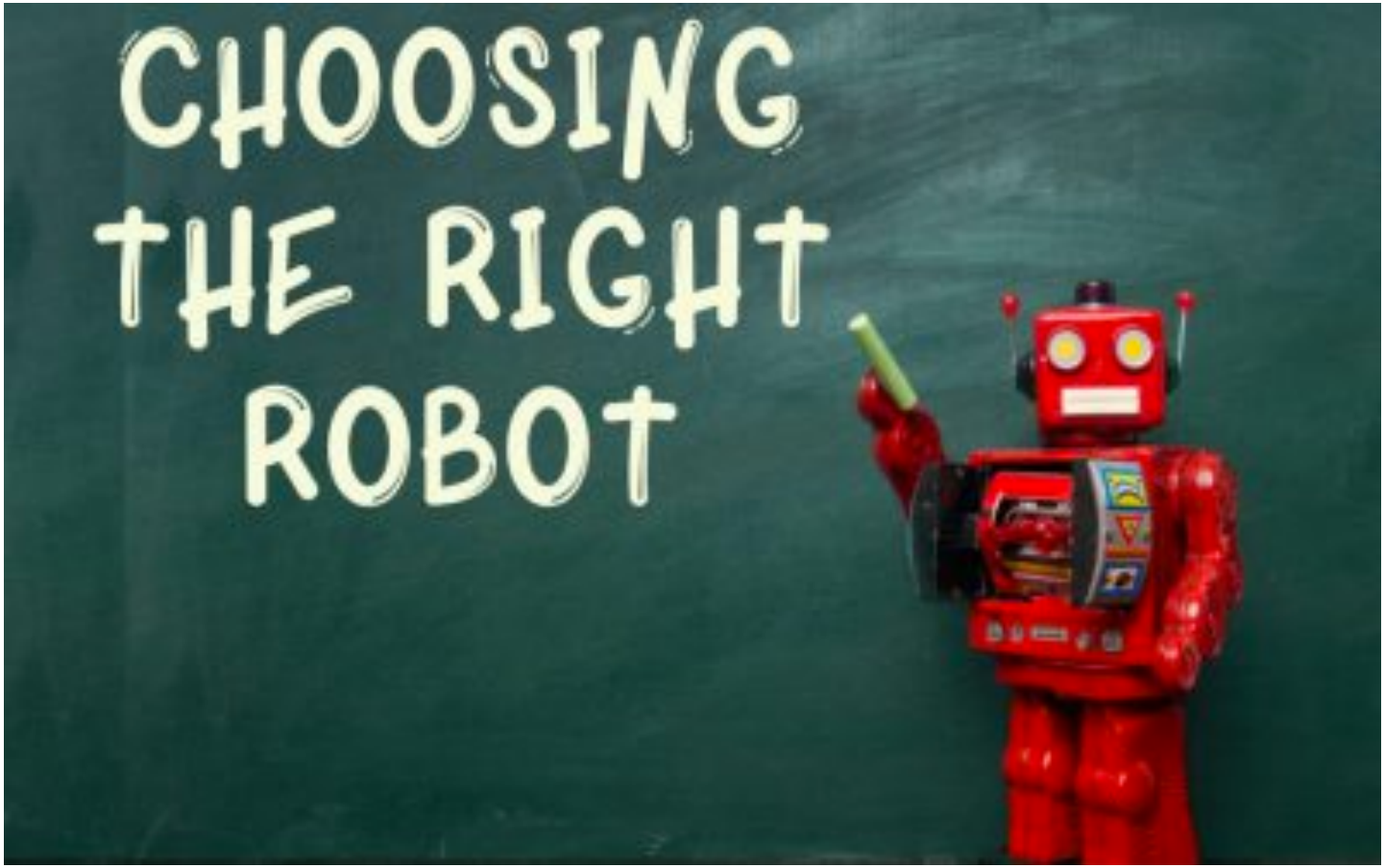
Data Mining

Pre-processing

Exploratory Data Analysis

Predictive Modelling

Data Visualisation


CHOOSING
THE RIGHT
ROBOT

Podcasts

Linear Machines 101

More or Less: Behind Stats

CodeBreaker

FT Tech Tonic

Data Skeptic

Linear Digressions

Partially Derivatives

The O'Reilly Data Show

Websites

Data Science Central

Open Data Science

Machine Learning Mastery

R Views

No Free Hunch

Simply Statistics

FlowingData

Medium

Online Videos

Siraj Raval / Udacity

Brandon Foltz / Stats 101

Derek Kane

Google Developers

Marginal Revolution Uni

ODSC DataCamp

DataSchool

Re.Work Events

Coursera / EdX

References

1. Dealing with missing data: Key assumptions and methods for applied analysis. Marina Soley
2. Data Scientist: The Sexiest Job of the 21st Century. Davenport T, Patil D. Harvard Business Review, 2012.
3. Deconstructing Data Science: Breaking The Complex Craft Into It's Simplest Part. Goldstein A. The Mission. 2017.
4. About Feature Scaling and Normalization. Sebastian Raschka; 2014.
5. Alberto Cairo, The Functional Art
6. Nathan Yau https://flowingdata.com/