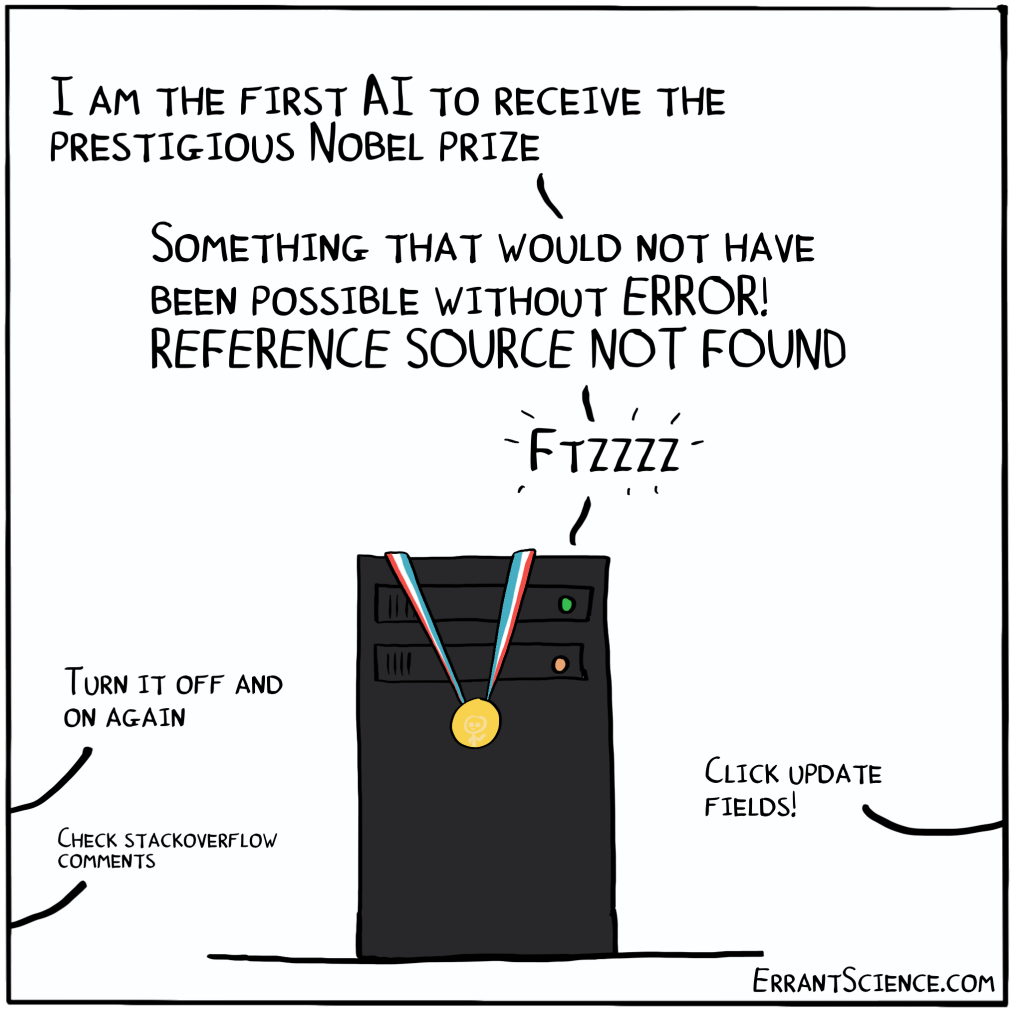
“Will an AI win the next Chemistry Nobel Prize and replace us?

Jeremy G. Frey

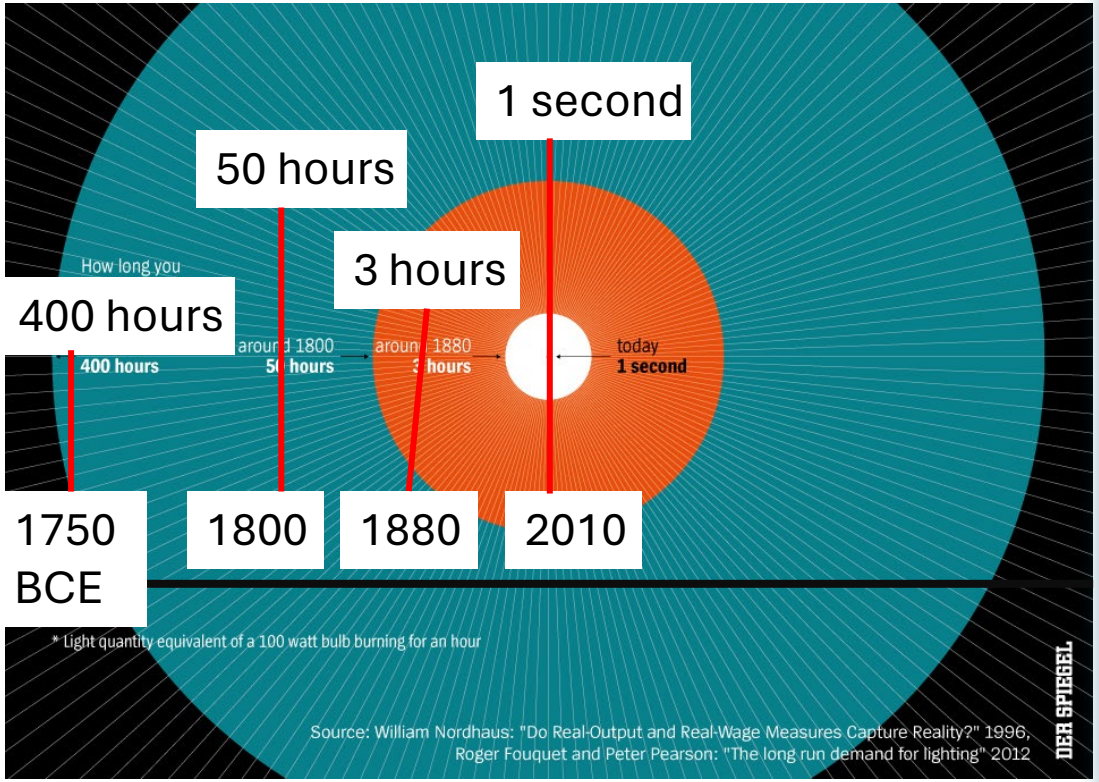
Head of Computational Systems Chemistry

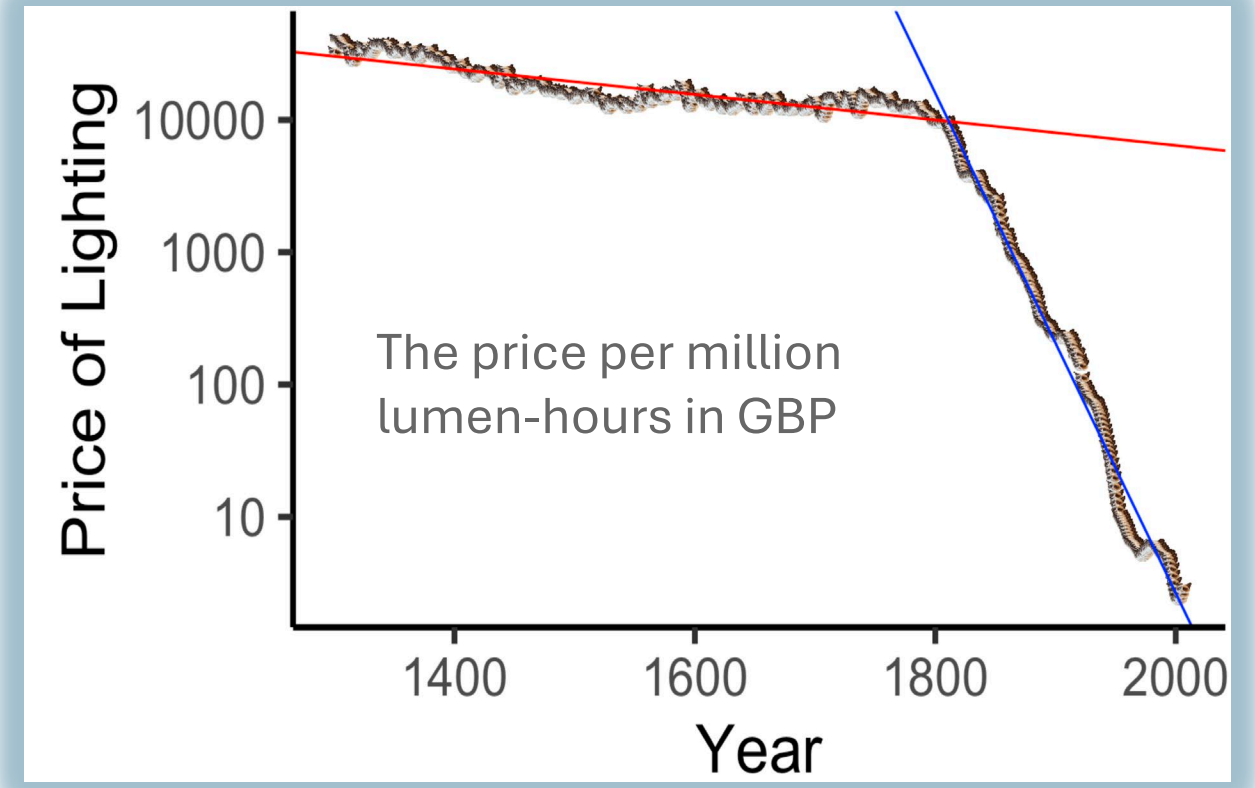
School of Chemistry, University of Southampton



The Transforming Power of Technological Advances

How long you had to work to get an hour of light

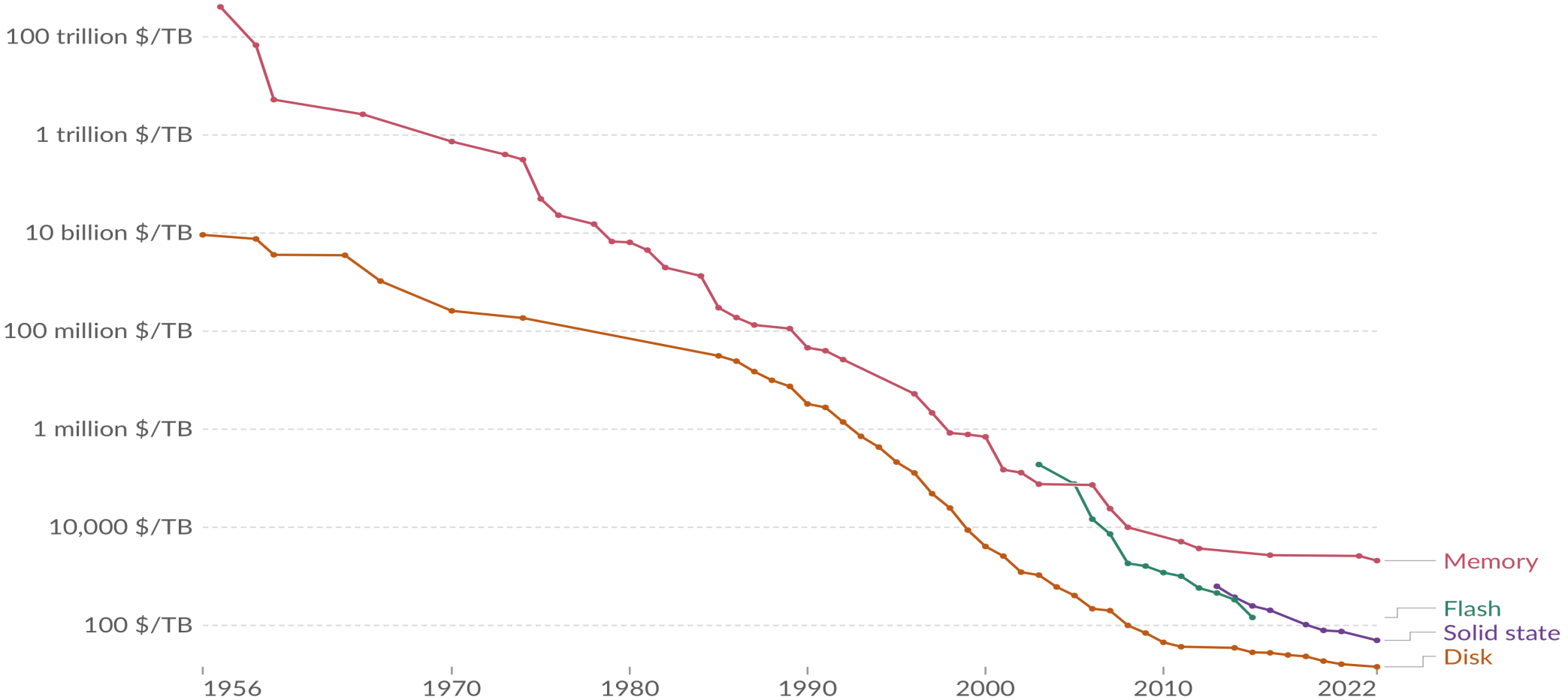




https://ourworldindata.org/grapher/the-price-for-lighting-per-million-lumen-hours-in-the-uk-in-british-pound

Historical cost of computer memory and storage

This data is expressed in US dollars per terabyte (TB). It is not adjusted for inflation.



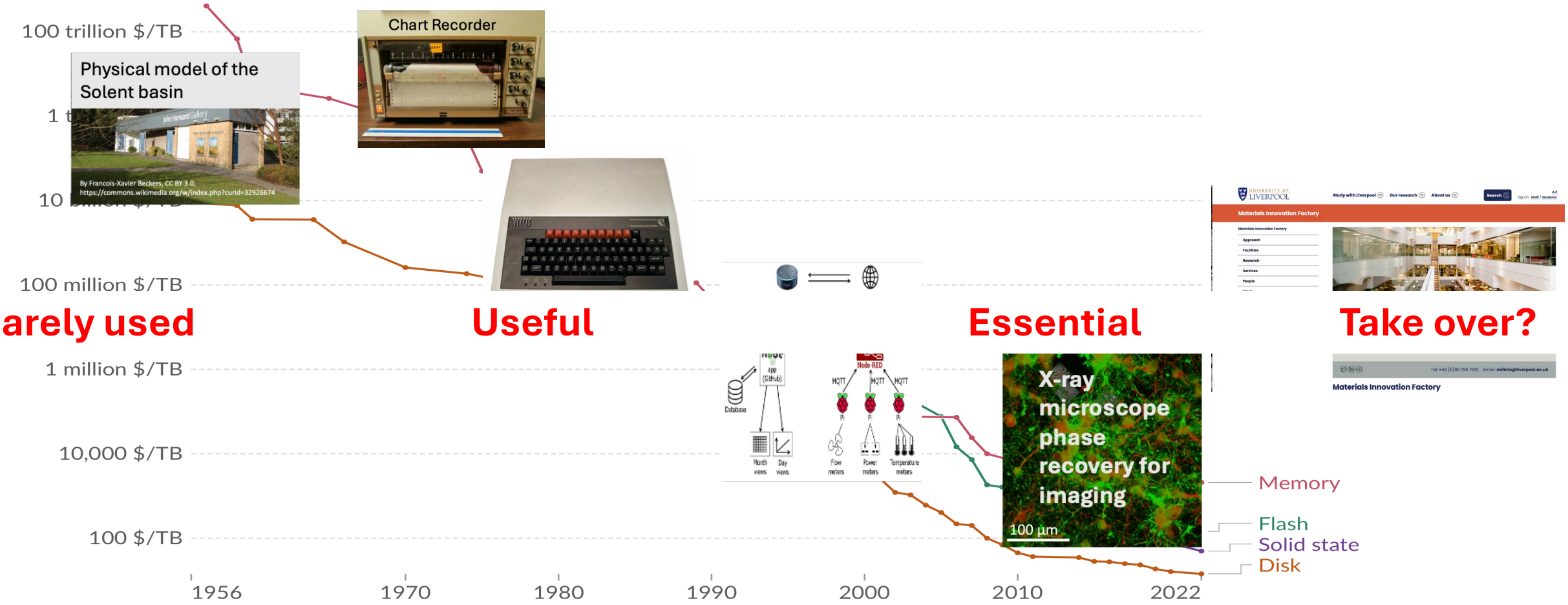
Data source: John C. McCallum (2022) OurWorldInData.org/technological-change | CC BY

Note: For each year, the time series shows the cheapest historical price recorded until that year.

Historical cost of computer memory and storage

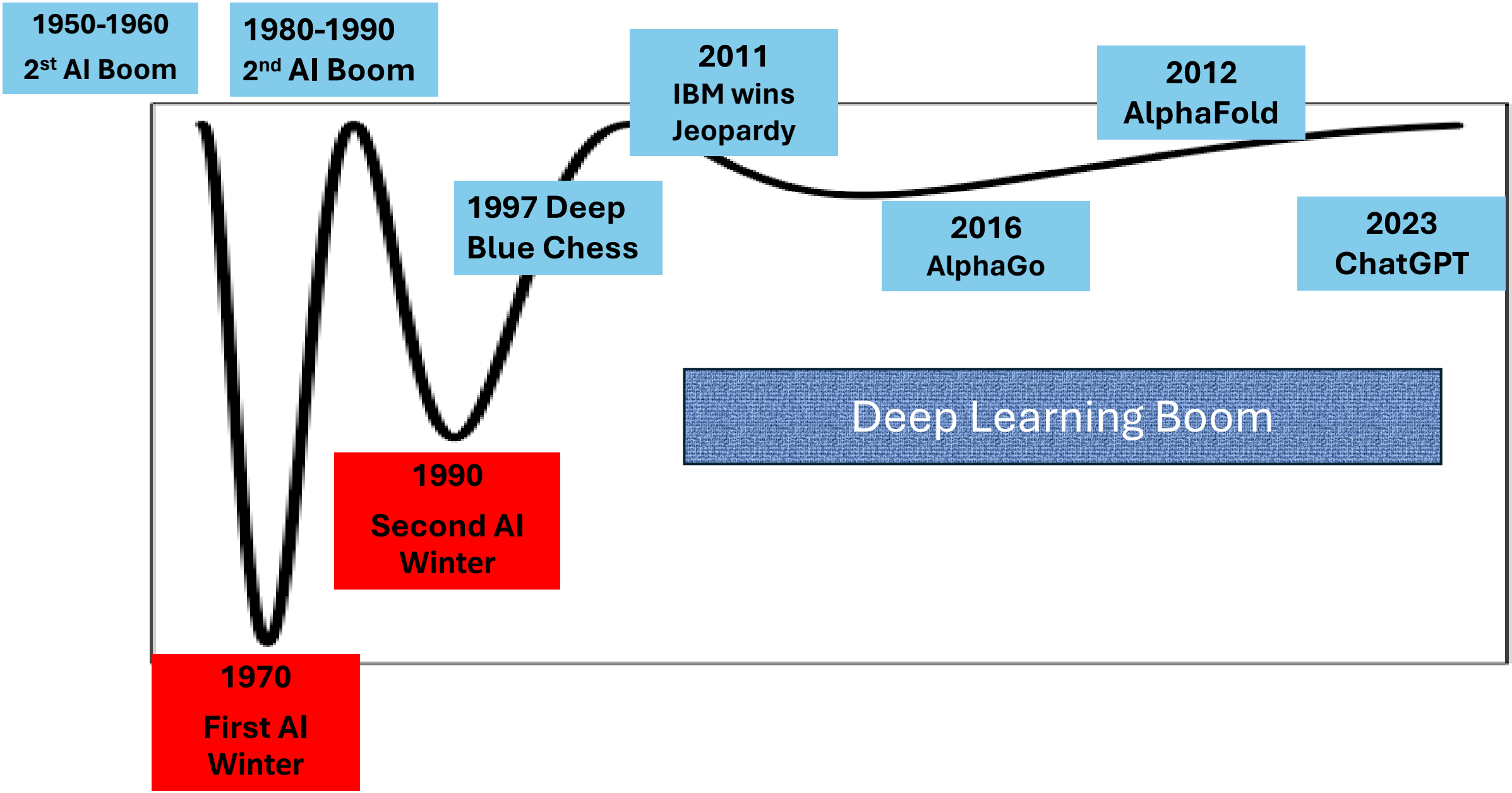
This data is expressed in US dollars per terabyte (TB). It is not adjusted for inflation.

Our World in Data

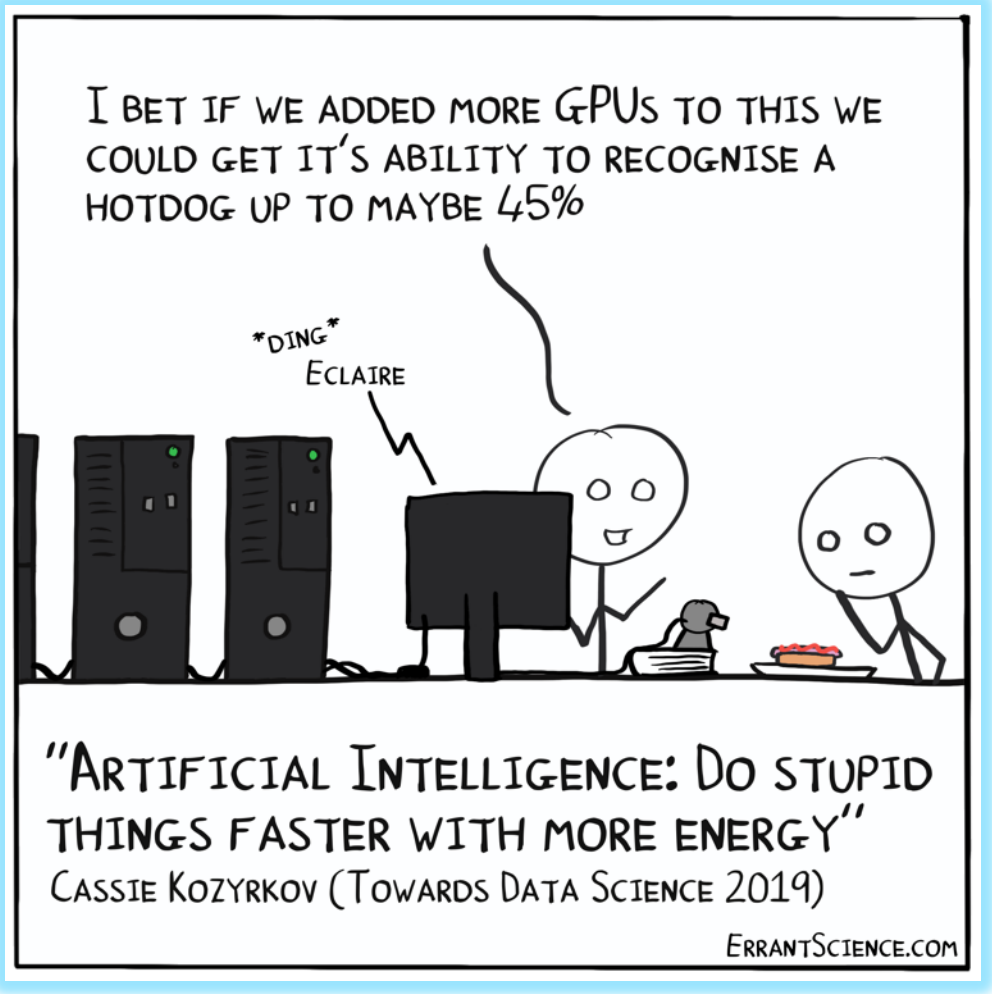


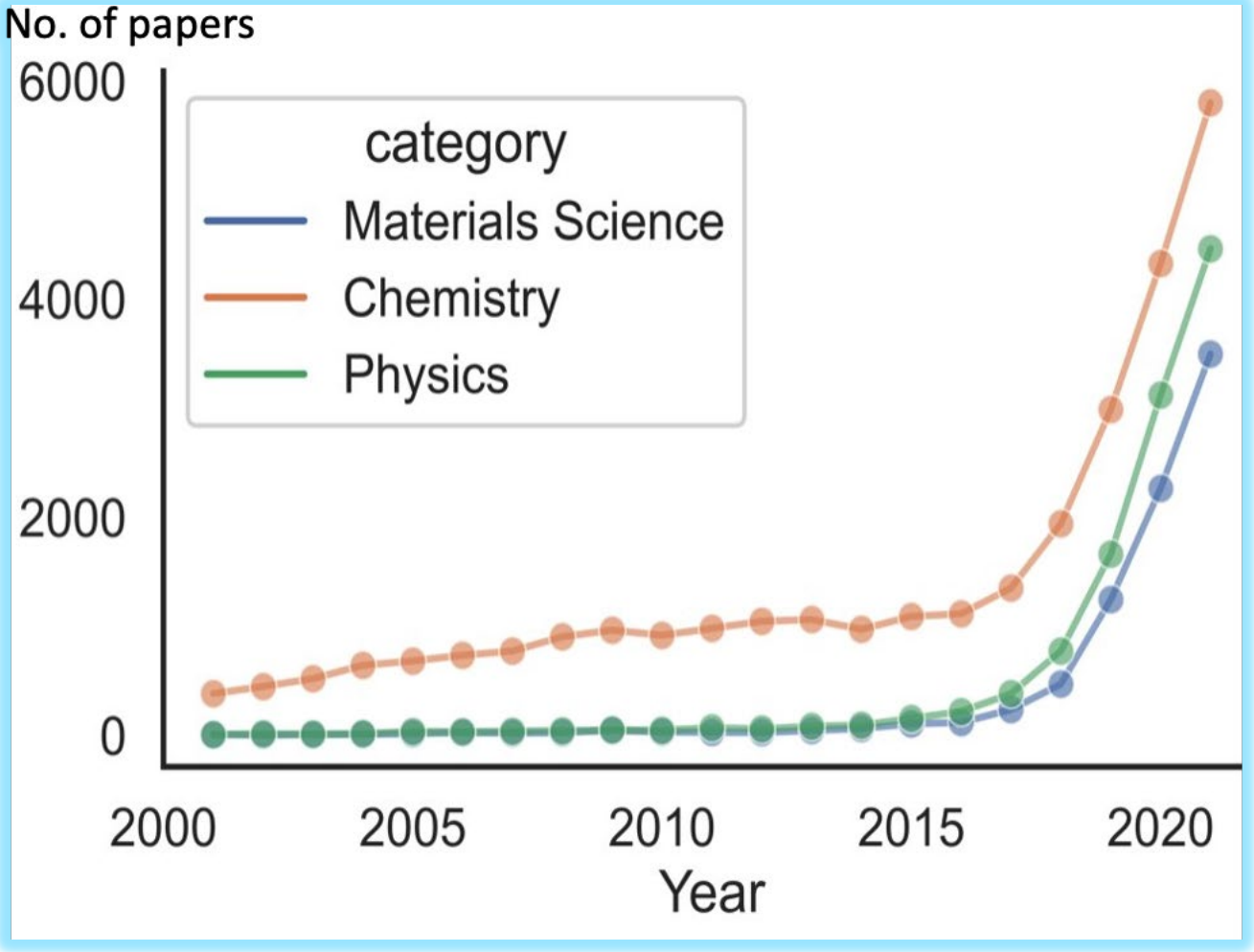
Data source: John C. McCallum (2022) OurWorldInData.org/technological-change | CC BY

Note: For each year, the time series shows the cheapest historical price recorded until that year.


1950-1960
2st AI Boom
1980-1990
2nd AI Boom
2011
IBM wins
Jeopardy
2012
AlphaFold
1997 Deep
Blue Chess
2016
AlphaGo
2023
ChatGPT
Deep Learning Boom
1990
Second AI
Winter
1970
First AI
Winter

Augmented Chemical Intelligence?





Chocolate rules the world

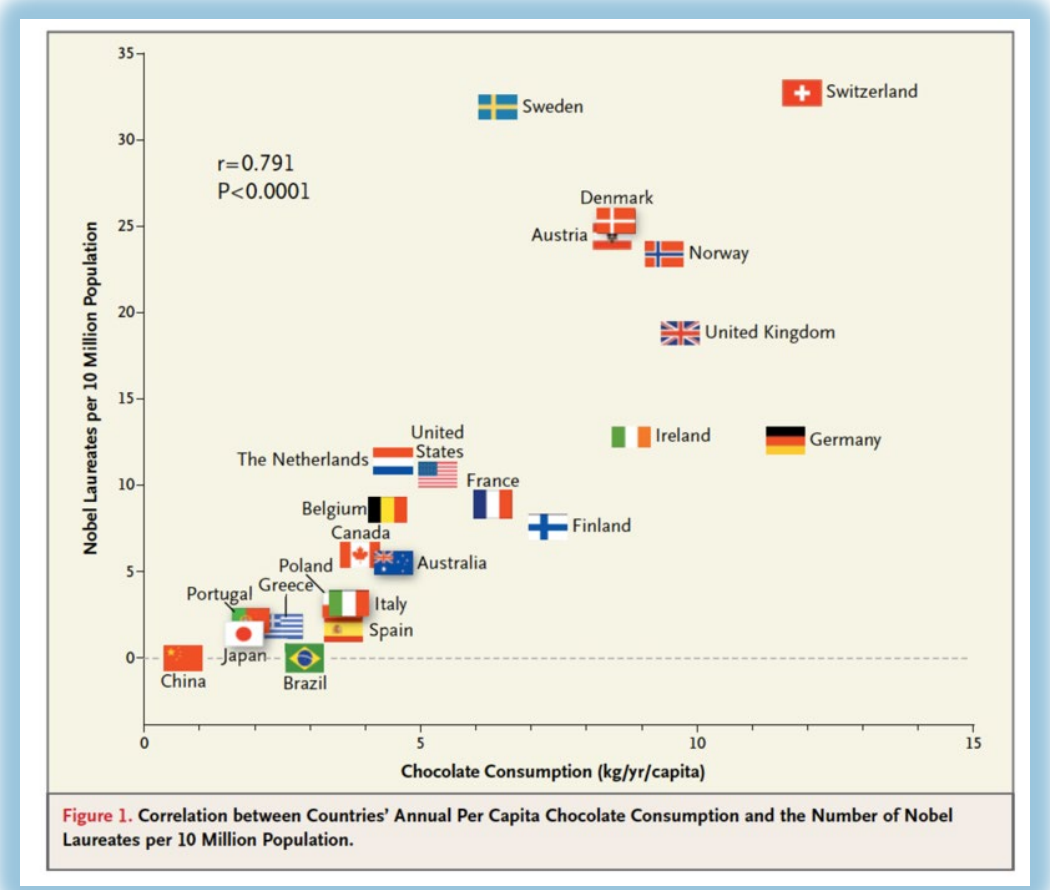
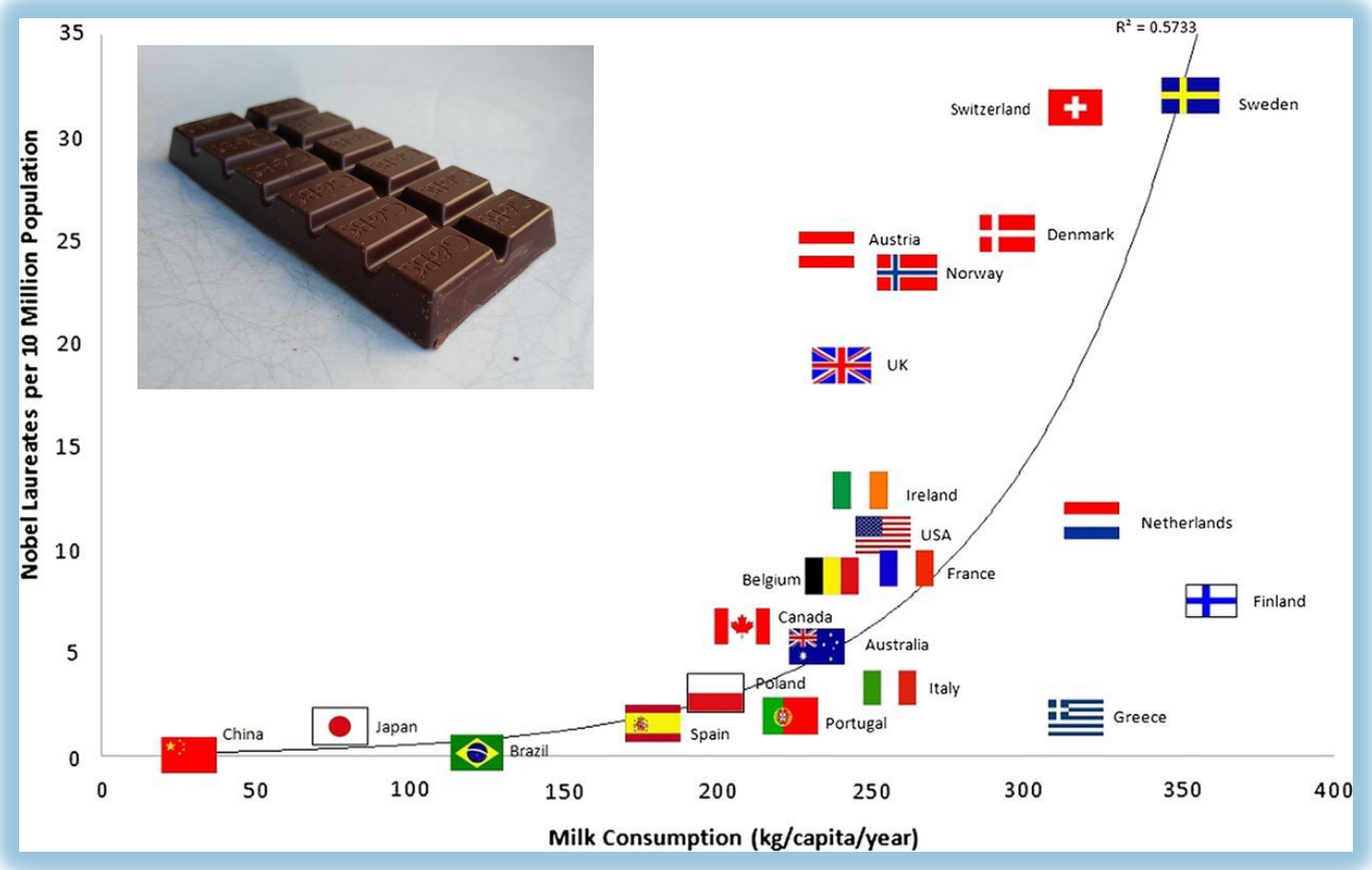


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.



Correlation and Causation – Causal reasoning and Explainable AI

What is special about Chemistry?

Models

- All models are wrong, but some are useful (Box)
- Different models are useful in different circumstances and for different audiences
- Sometimes more than one model is useful for the same problem

- In Chemistry sometimes the very models used at the same time are mutually inconsistent
- This all goes to make the concept of explanation in Chemistry qualitatively different from scientific disciplines

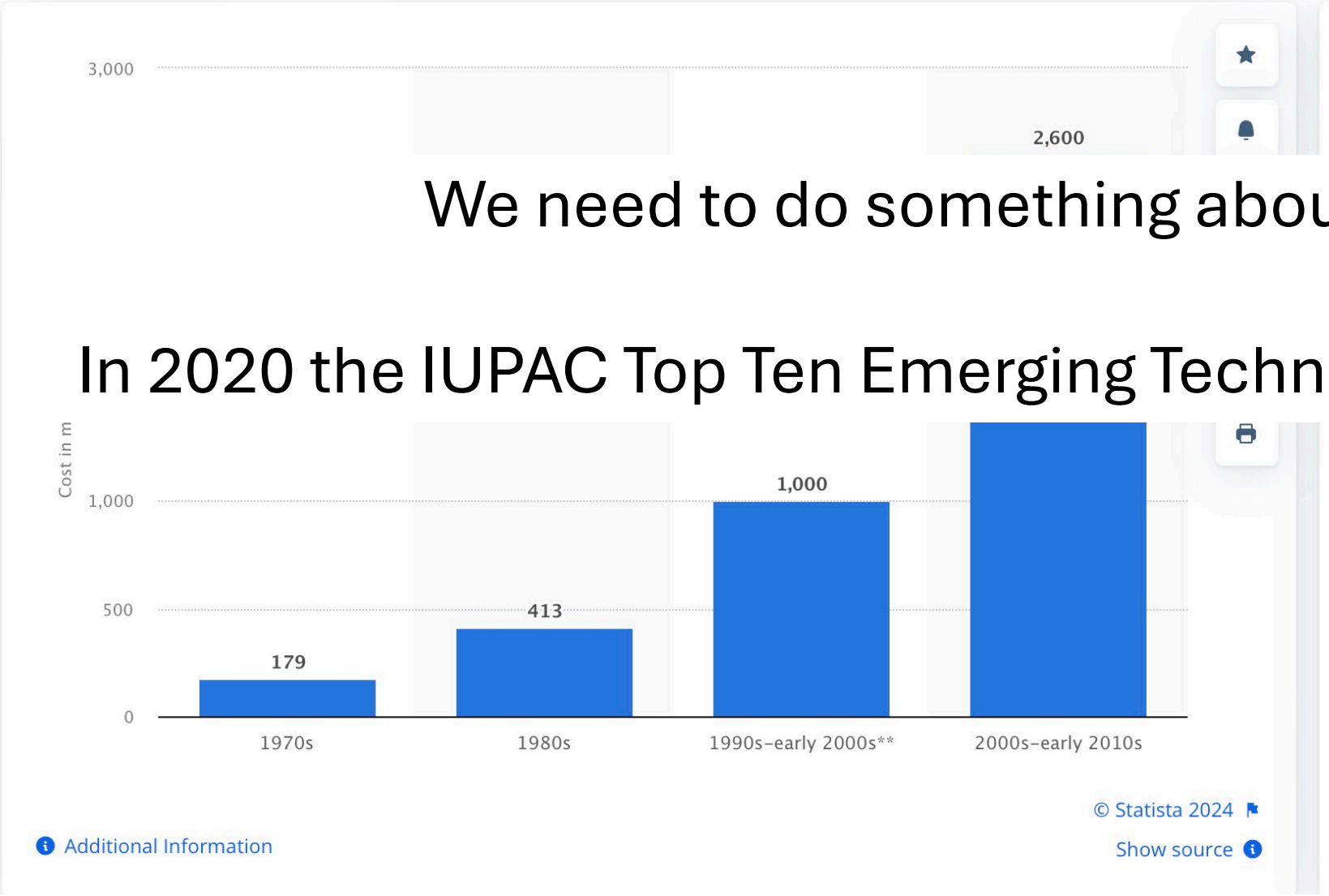
Models

The explanations freely flow back and forth between the rigorous mathematical style of the physical sciences to the more category style of historical biological and medical explanations.

The ability to handle mutually inconsistent explanatory principles of a characteristic of Chemists and their training.

It is one of the reasons why a chemistry training is useful more widely than just in Chemical sciences.

Cost of developing a drug in the U.S. from the 1970s until today (in million U.S. dollars)*



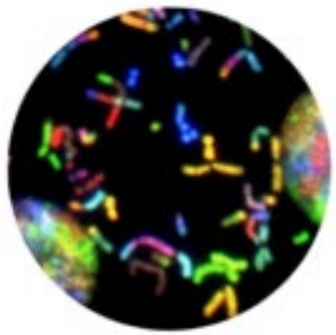
We need to do something about this!

In 2020 the IUPAC Top Ten Emerging Technologies included AI

IUPAC Top Ten Emerging Technologies in Chemistry 2021

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

Blockchain



Semisynthetic life

Superwettability

Artificial humus

RNA synthesis

Sonochemical coating

Chemo-luminescence

Sustainable ammonia

Targeted protein degradation

Single-cell metabolomics

IUPAC Top Ten Emerging Technologies in Chemistry 2022

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

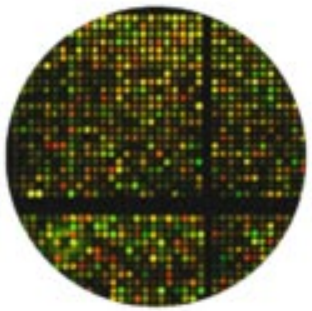
Sodium batteries

Nanozymes

Aerogels

Fluorescent sensors

Solar fuels



Nanoparticle megalibraries

Fibre-based batteries

Textile displays

Rational vaccines with SNA

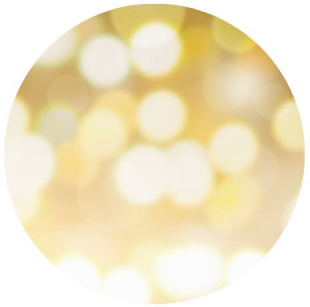
VR-enabled modelling

#IUPACTop2022

IUPAC Top Ten Emerging Technologies in Chemistry 2023

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

Wearable sensors



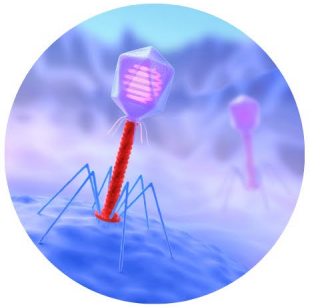
Photocatalytic hydrogen

Artificial muscles

GPT language models in chemistry

Synthetic electrochemistry

Chloride-mediated removal of ocean CO_2



Phage therapy

Biological recycling of PET

Depolymerisation

“Low-sugar” vaccinations

#IUPACTop2023

From QSAR to Machine Learning

Chemical Science

EDGE ARTICLE

View Article Online
View Journal | View Issue

CrossMark click for updates

Cite this: *Chem. Sci.*, 2016, **7**, 1600

QSAR analysis of substituent effects on tambjamin anion transporters†‡

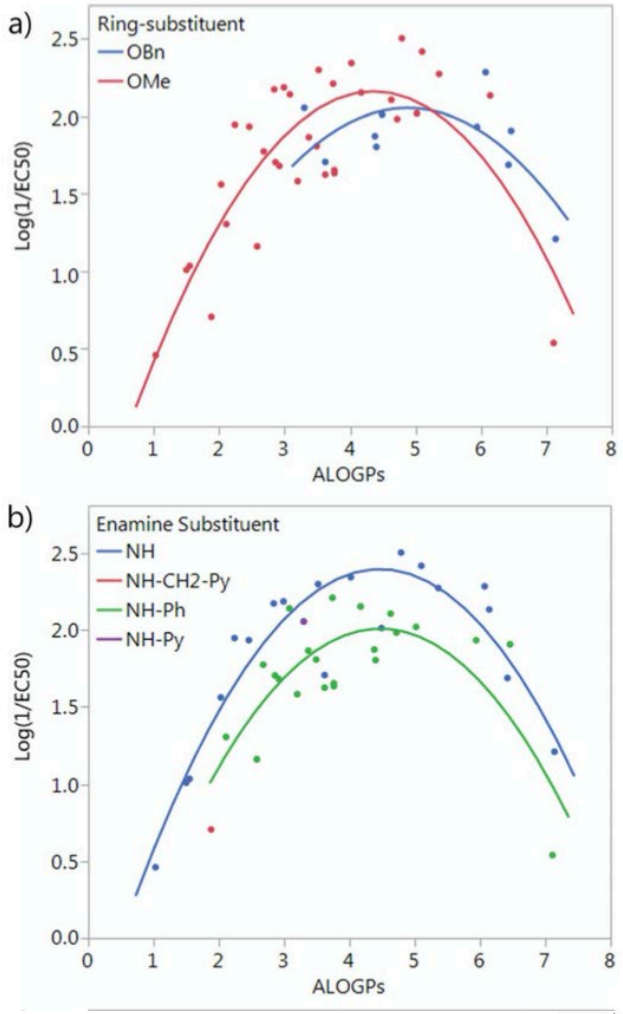
Nicola J. Knight,[a] Elsa Hernando,[b] Cally J. E. Haynes,§[a] Nathalie Busschaert,¶[a] Harriet J. Clarke,[a] Koji Takimoto,[c] María García-Valverde,[b] Jeremy G. Frey,*[a] Roberto Quesada*[b] and Philip A. Gale*[a]

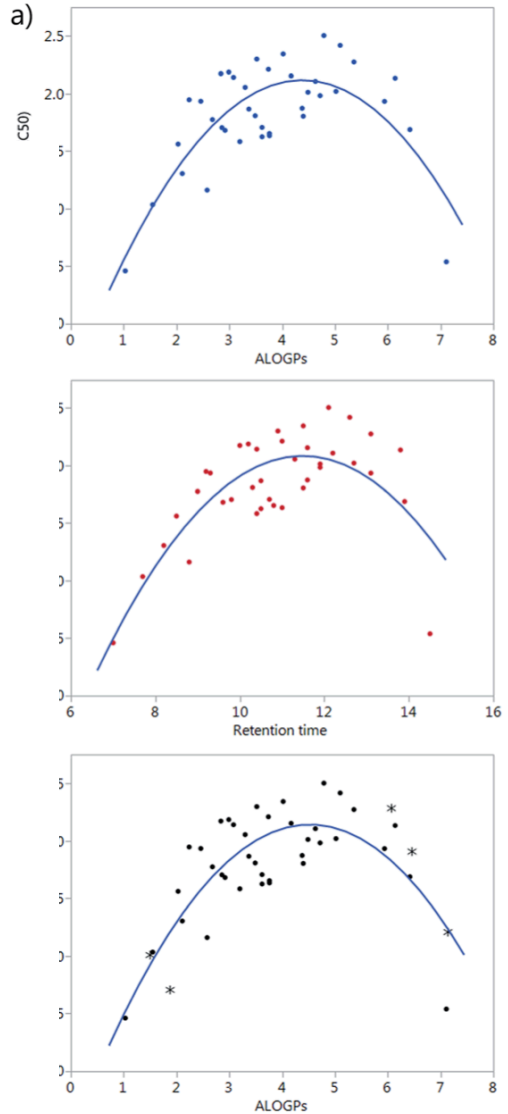
The transmembrane anion transport activity of 43 synthetic molecules based on the structure of mari alkaloid tambjamine were assessed in model phospholipid (POPC) liposomes. The anionophoric activ of these molecules showed a parabolic dependence with lipophilicity, with an optimum range transport efficiency. Using a quantitative structure–transport activity (QSAR) approach it was possible rationalize these results and to quantify the contribution of lipophilicity to the transport activity of the derivatives. While the optimal value of log *P* and the curvature of the parabolic dependence is a prope of the membrane (and so similar for the different series of substituents) we found that for relativ simple substituents in certain locations on the tambjamine core, hydrophobic interactions clea dominate, but for others, more specific interactions are present that change the position of t membrane hydrophobicity parabolic envelope.

Received 16th October 2015
Accepted 17th November 2015

DOI: 10.1039/c5sc03932k

www.rsc.org/chemicalscience





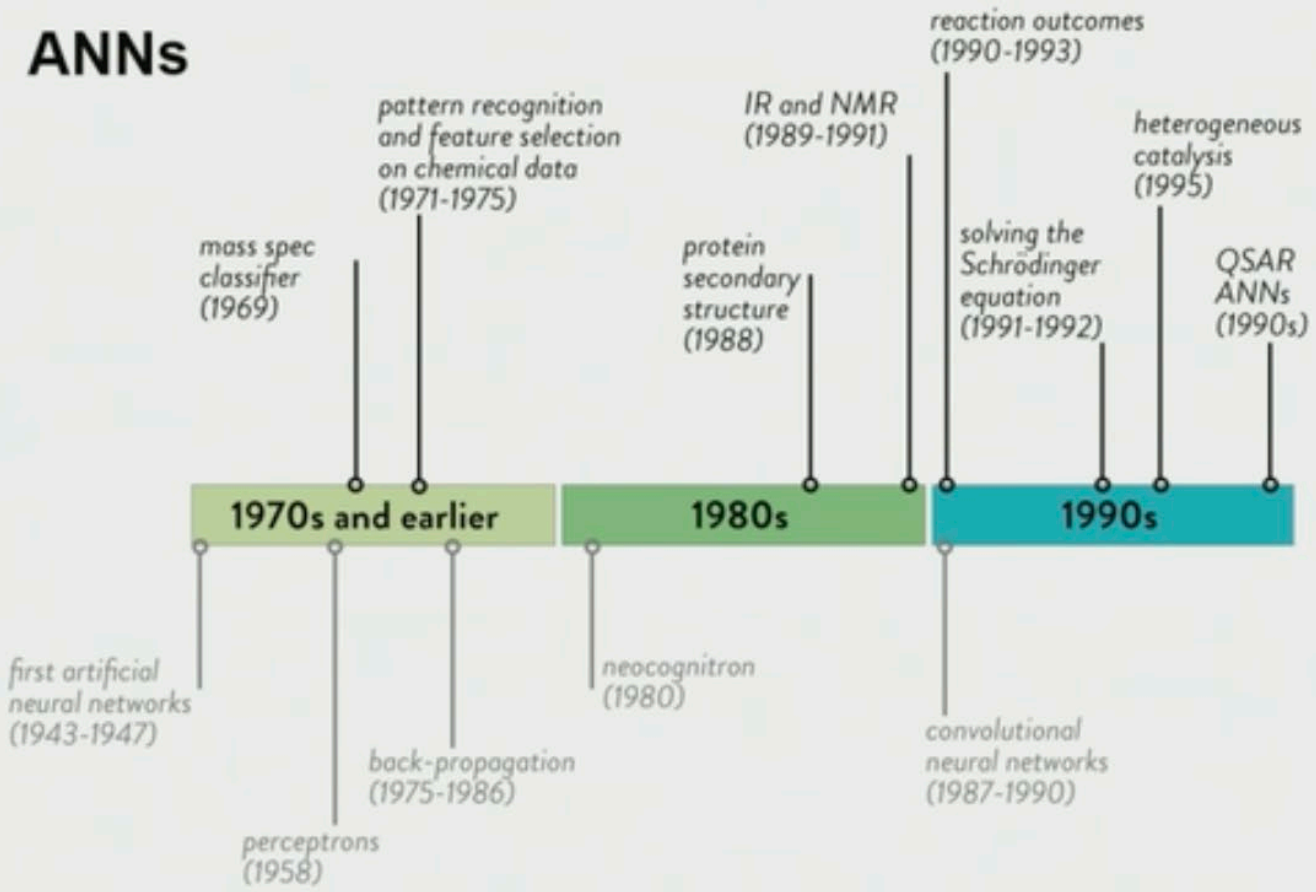
Plot of log(1/EC_{50}) *vs.* ALOGPs of the first set of 38 ; (b) plot of log(1/EC_{50}) *vs.* RT of the first 38 compounds; log(1/EC_{50}) *vs.* ALOGPs showing all 43 compounds, the additional molecules are highlighted by *.

Machine Learning in Chemistry

Relevant Data	Equation / Model	Name
Known	Known	Theory Physical Chemistry
Unknown	Known	Unproved theory
Known	Unknown	Statistical Modelling (QSAR)
Unknown	Unknown	Machine Learning

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

ENTER MACHINE LEARNING WITH ANNS



1992 San Francisco

Moscone Center
Room 220–222, Mezzanine Level

Chemical Applications of Neural Networks

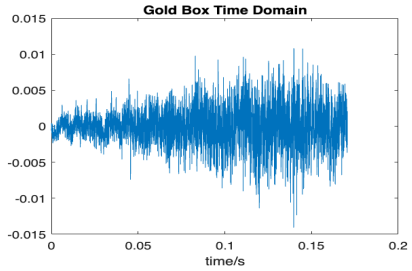
Sponsored by Division of Computers in Chemistry

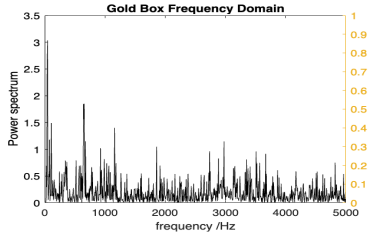
G. M. Maggiora, *Presiding*

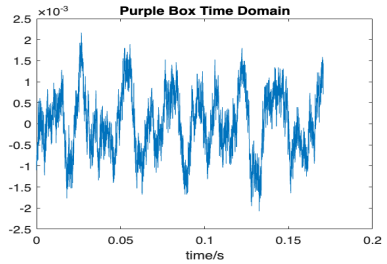
2:00 Neural networks as a tool for multispectral data interpretation. M. E. Munk, M. S. Madison, **E. W. Robb**

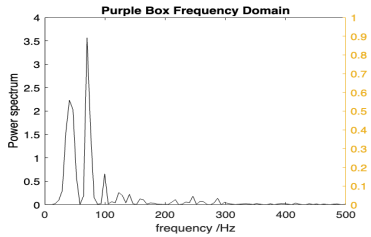
3:15 Simulation of carbon-13 NMR spectra by artificial neural networks. **P. C. Jurs**, L. S. Anker

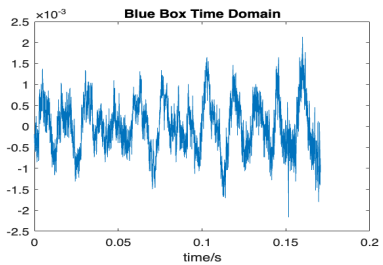
4:00 Neural networks as function mapping devices for relating chemical structure to chemical properties and biological activities. **D. W. Elrod**, G. M. Maggiora, R. G. Trenary

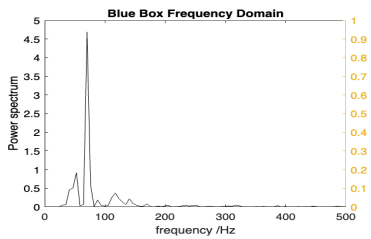

Gold Box Time Domain
time/s

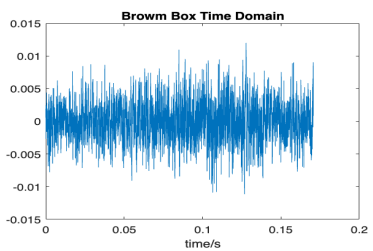

Gold Box Frequency Domain
Power spectrum
frequency /Hz

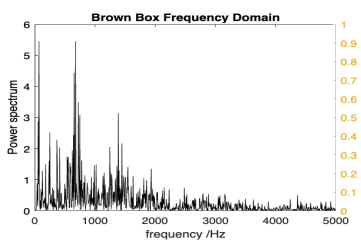

Purple Box Time Domain
time/s

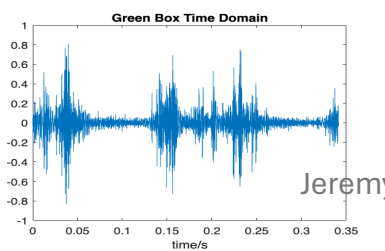

Purple Box Frequency Domain
Power spectrum
frequency /Hz

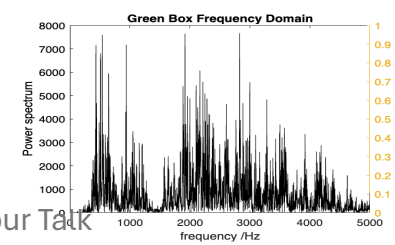

Blue Box Time Domain
time/s

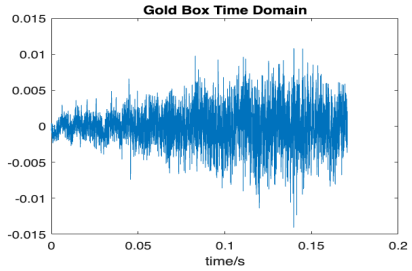

Blue Box Frequency Domain
Power spectrum
frequency /Hz

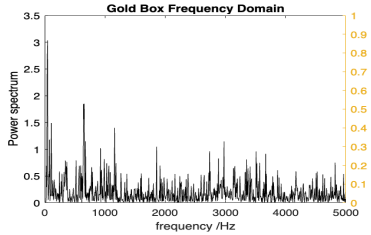

Browm Box Time Domain
time/s


Brown Box Frequency Domain
Power spectrum
frequency /Hz

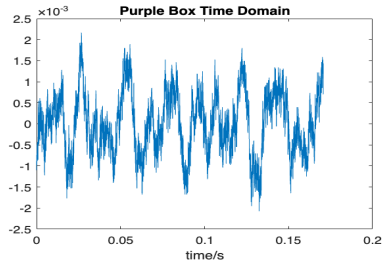

Green Box Time Domain
time/s

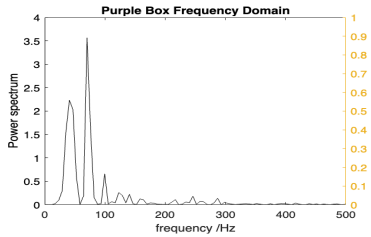

Green Box Frequency Domain
Power spectrum
frequency /Hz



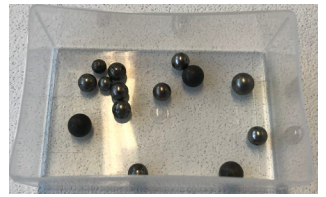


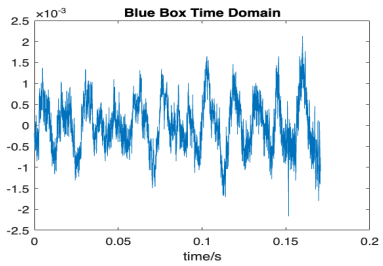
Gold

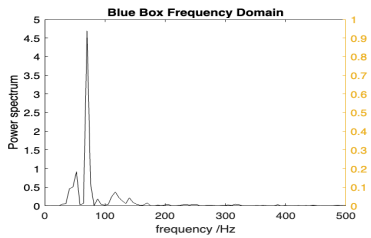




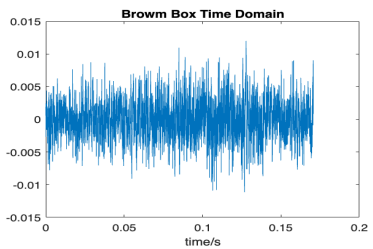
Blue

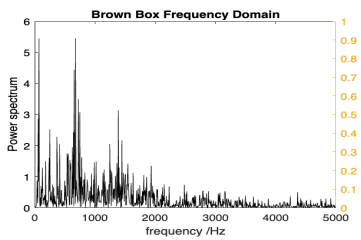




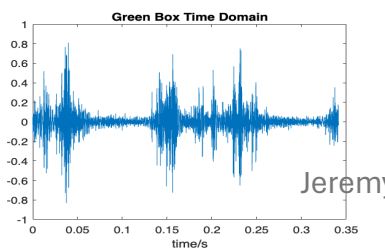


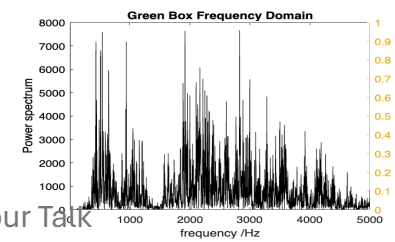
Purple





Brown





Green

Machine Learning: Data-driven Modelling

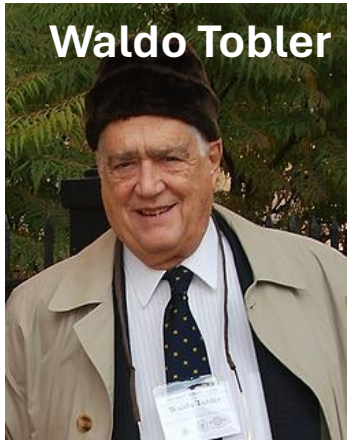
Data	$\{x_n,\ t_n\}_{n=1}^{N} \qquad \{x_n\}_{n=1}^{N}$
Function Approximator	$t = f(x, \theta) + v$
Parameter Estimation	$E_0 = \sum_{n=1}^{N} \{\| t_n - f(x_n; \theta) \|\}^2$
Prediction	$\hat{t}_{N+1} = f\left(x_{N+1}, \hat{\theta}\right)$
Regularization	$E_1 = \sum_{n=1}^{N} \{\| t_n - f(x_n) \|\}^2 + r(\|\theta\|)$
Modelling Uncertainty	$p\left(\theta \middle\| \{x_n, t_n\}_{n=1}^{N}\right)$
Probabilistic Inference	$E[g(\theta)] = \int g(\theta)\, p(\theta)\, d\theta = \frac{1}{N_s} \sum_{n=1}^{N_s} g\left(\theta^{(n)}\right)$
Sequential Estimation	$\theta(n-1\|n-1) \longrightarrow \theta(n\|n-1) \longrightarrow \theta(n\|n)$

Kalman & Particle Filters; Reinforcement Learning

Mahesan Niranjan (UoS) | AI4SD Summer School 2022 | AI4SD 2 / 5

https://shorturl.at/cfuvZ

https://www.ai4science.network/ai3sd-online-seminar-series/ai4sd-machine-learning-summer-school-2022/



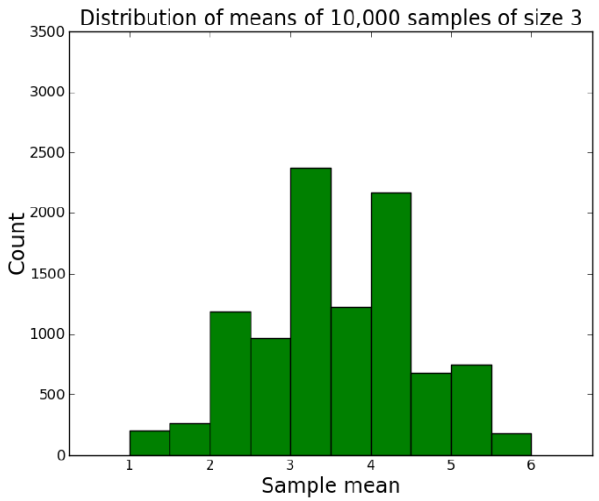
A map of chemical space?

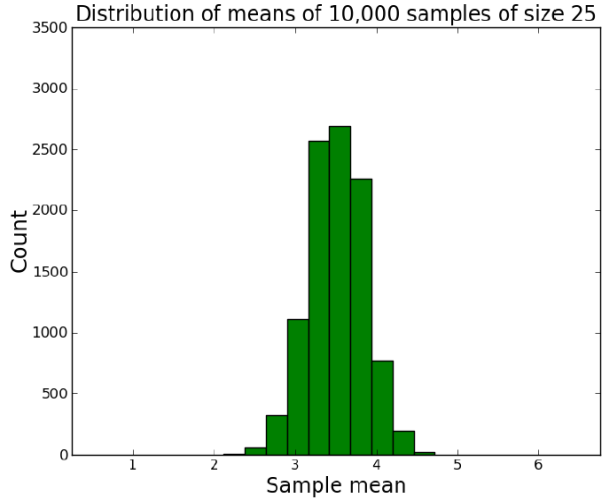
The First Law of Geography
”Everything is related to everything else, but near things are more related than distant things”.

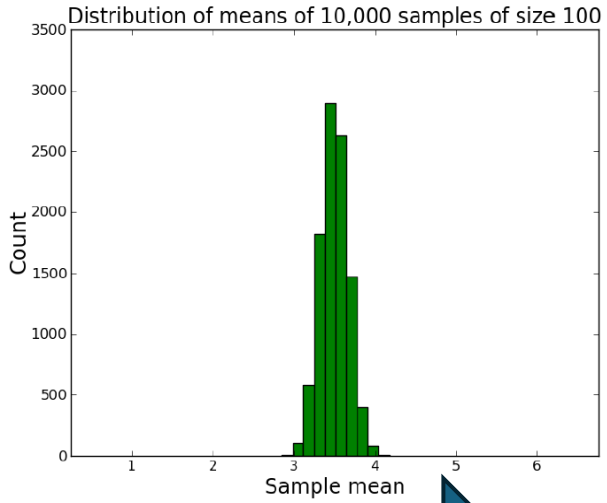
- “*Chemical* Space is big. You just won't believe how vastly, hugely, mind-bogglingly big it is. I mean, you may think it's a long way down the road to the *synthetic* chemist’s *lab*, but that's just peanuts to space.”

  — *With apologies to* **Douglas Adams,** The Hitchhiker's Guide to the Galaxy

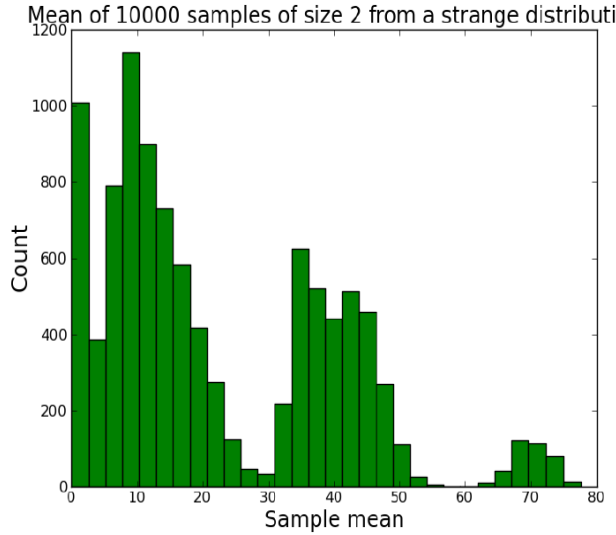
In Praise of High Dimensionality
The Central Limit Theorem

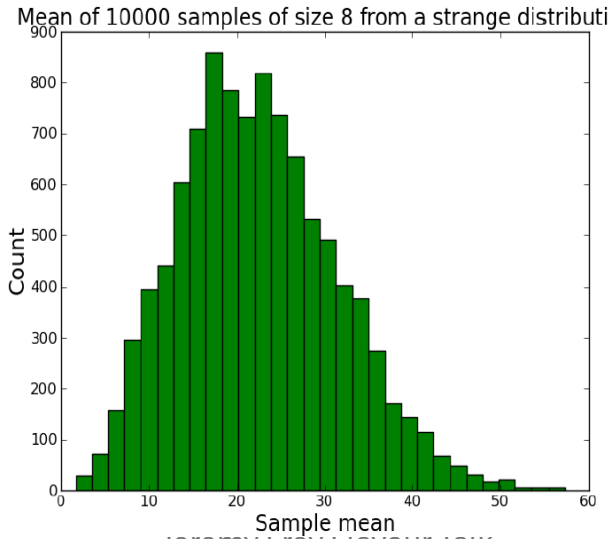


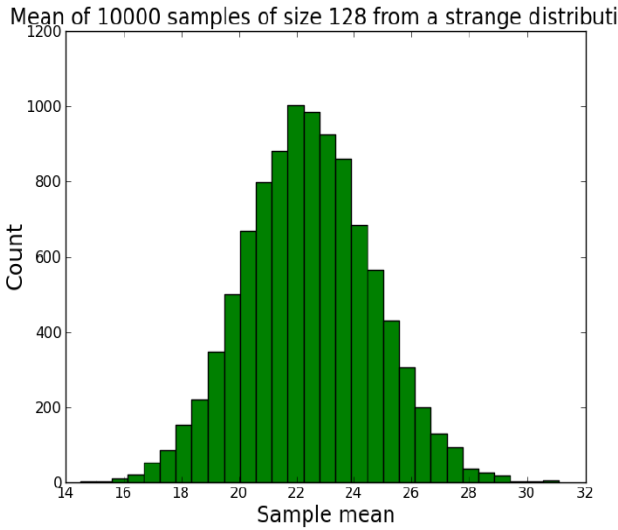




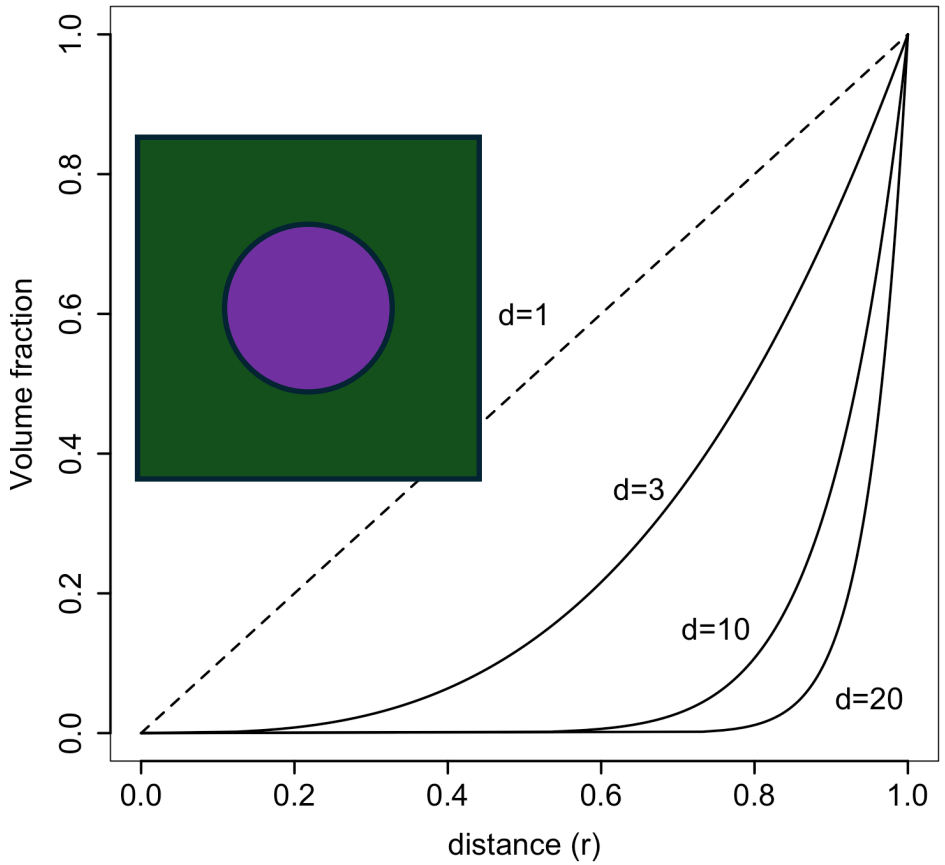
Increasing sample size



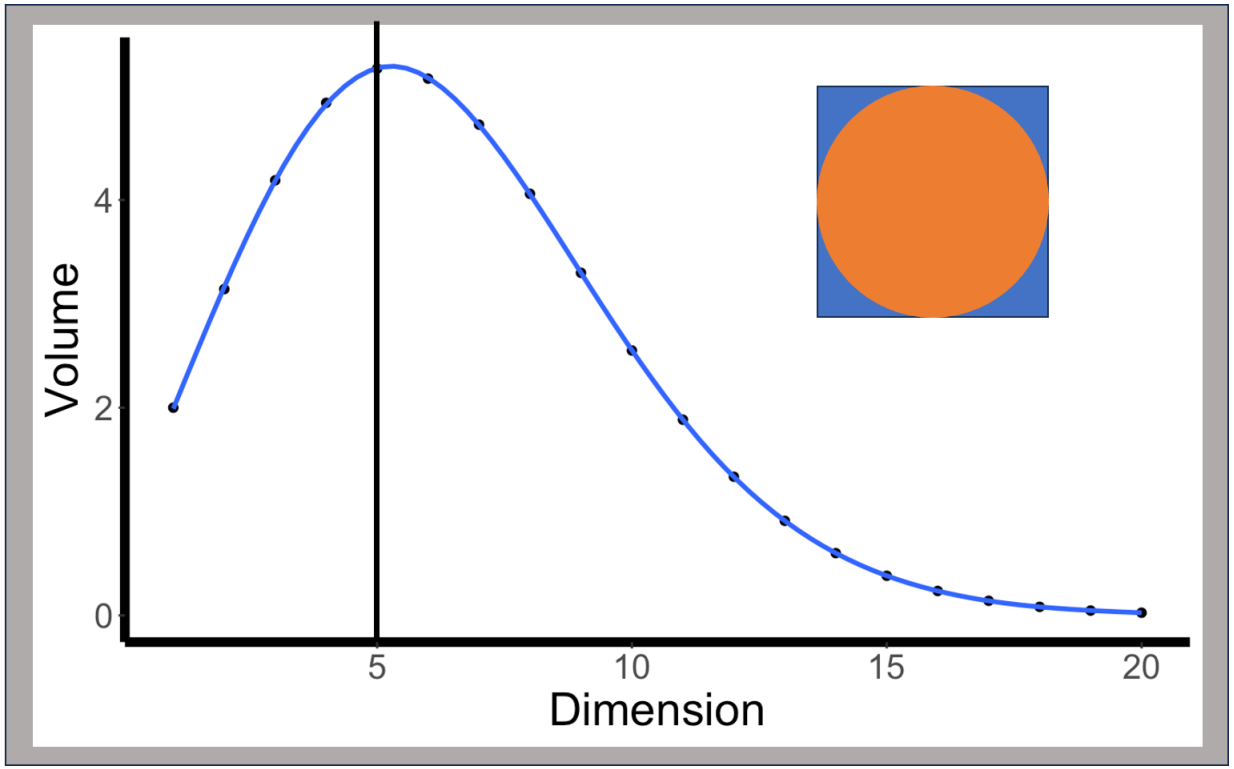




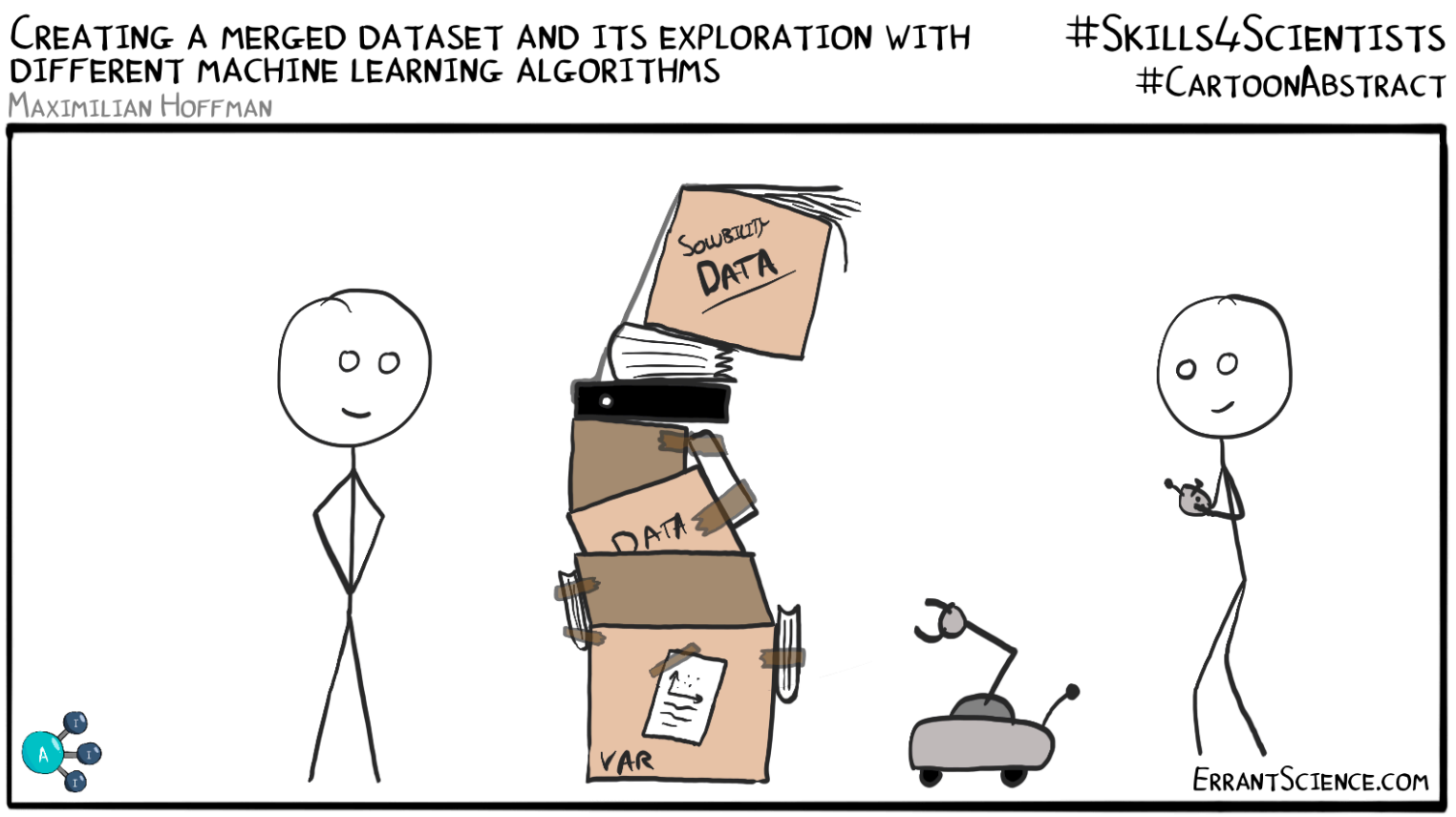
The Curse of High Dimensionality - Where are all the data points?



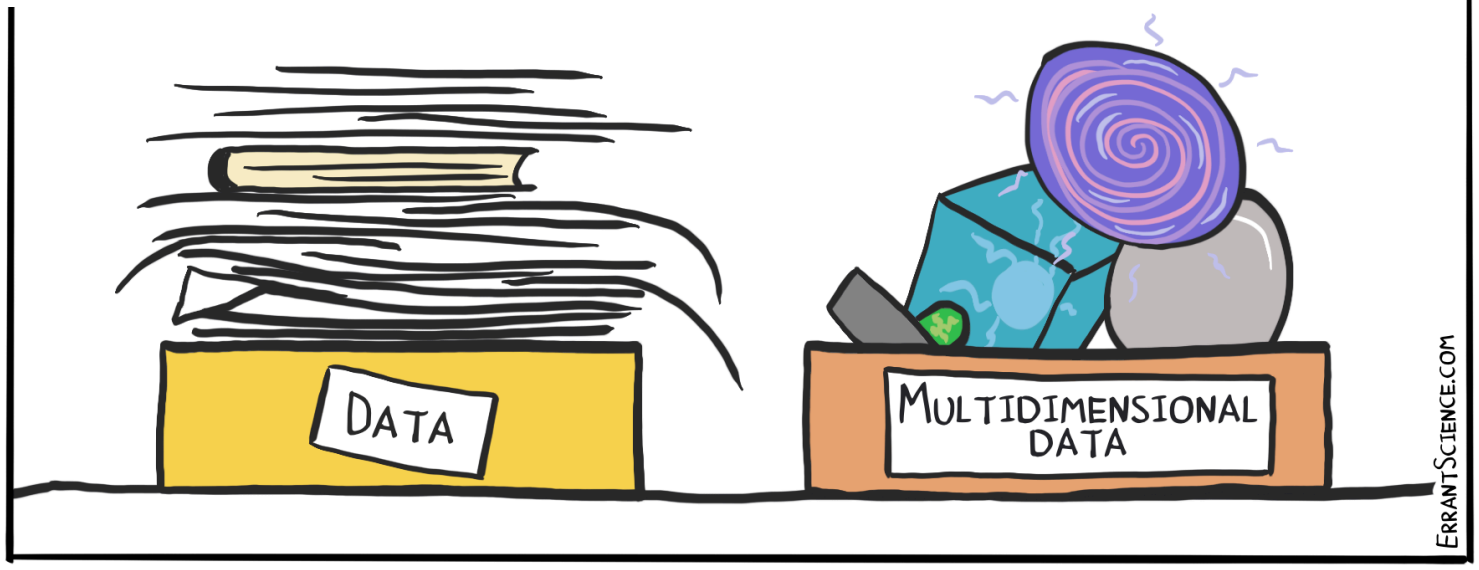
In high dimensions all the data is at the edge!



Volume of the sphere around a data point vs the volume of the space (cube)

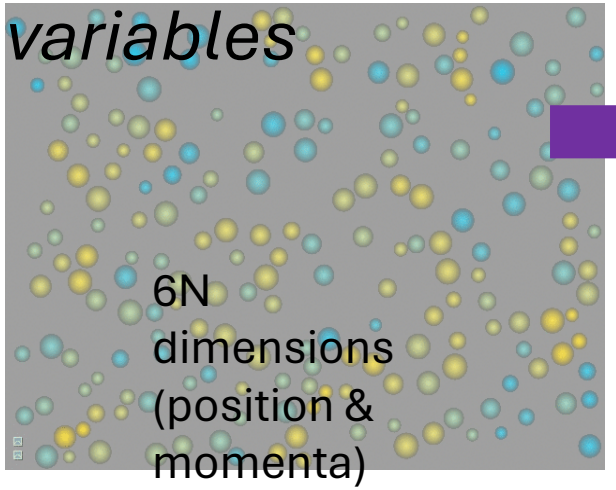


Dimensional Reduction

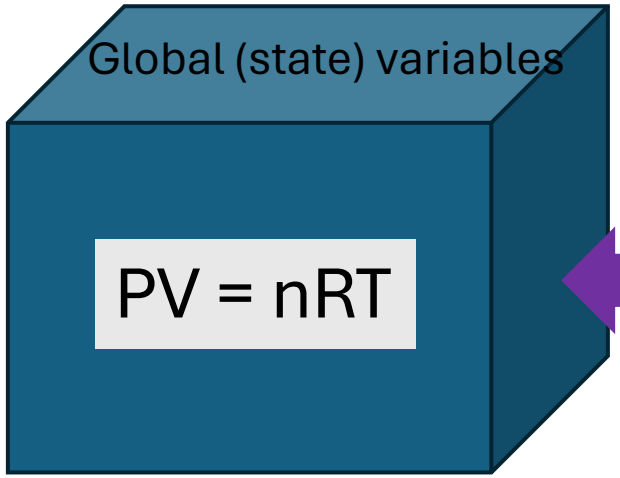


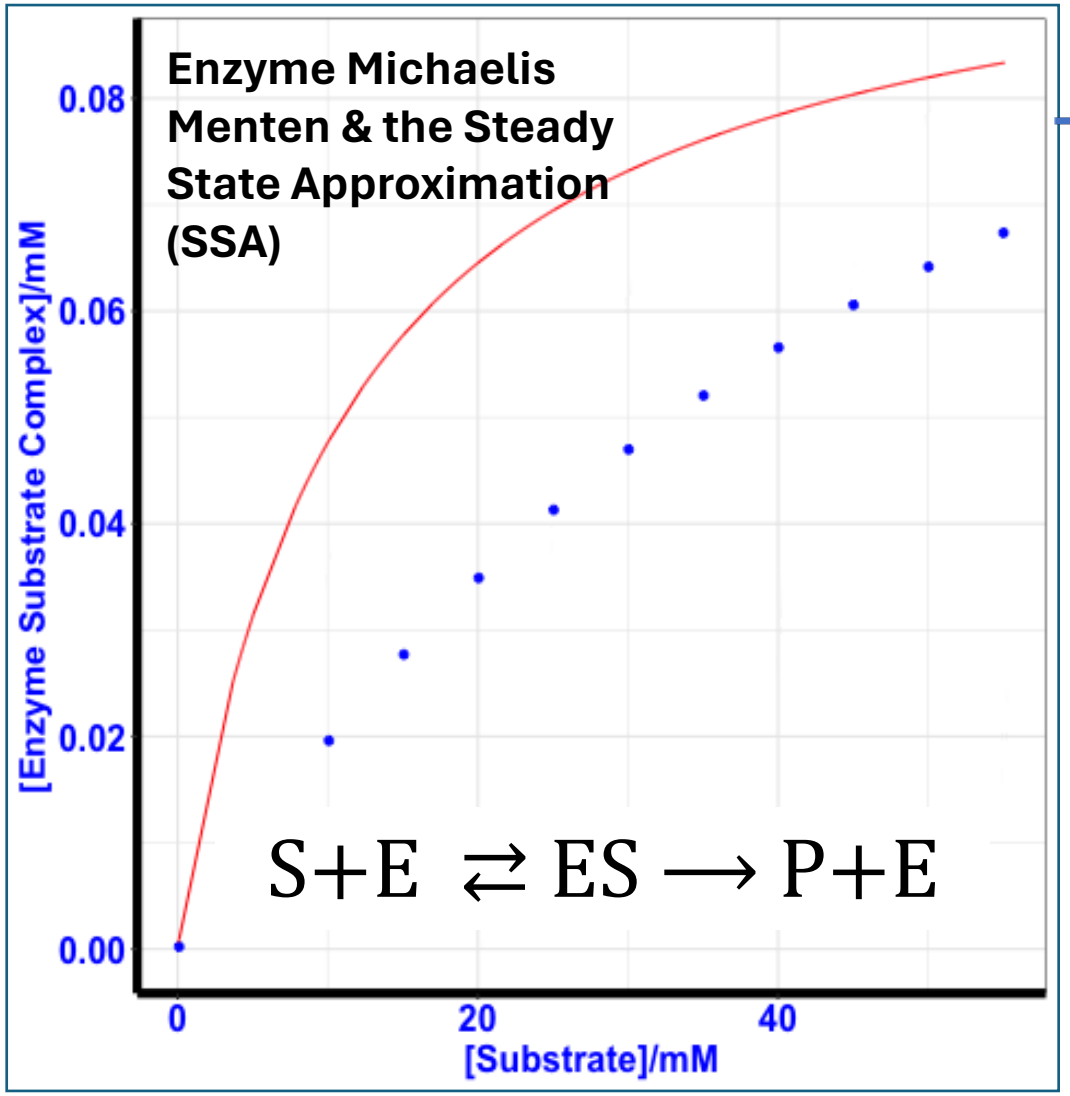
08/03/2024

individual particles to collective variables

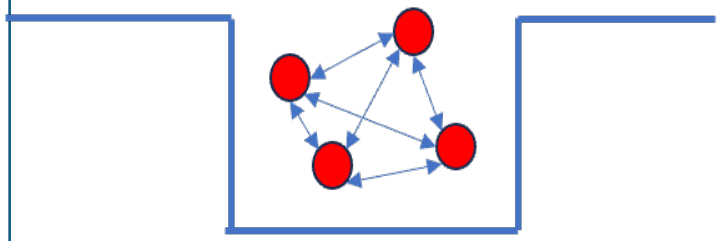


Statistical Thermodynamics



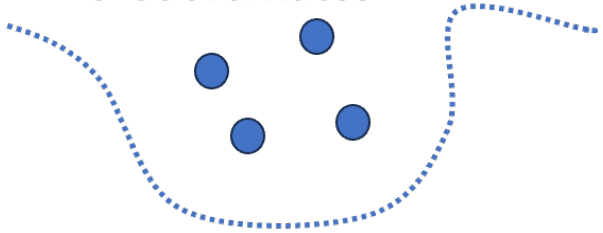


Schrodinger View
3N Coordinates



Formally Equivalent

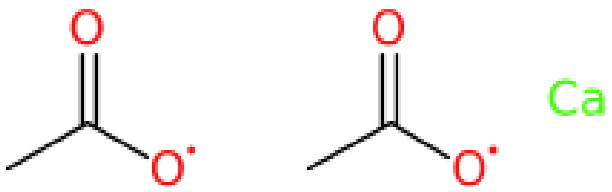
DFT View
3 Coordinates



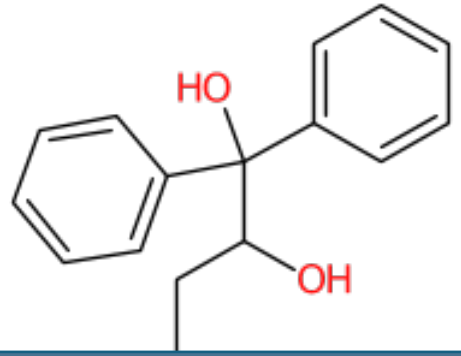
Kohn-Sham particle (non-interacting)
Effective potential

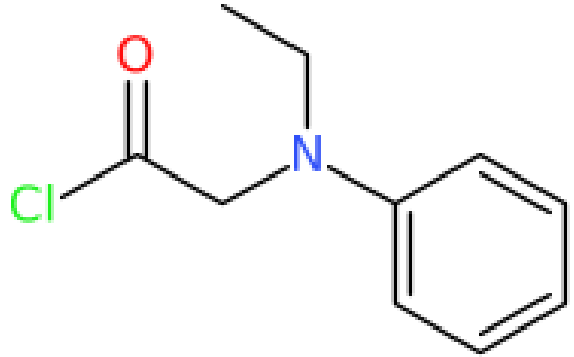
Revive the 1950's – The Wisswesser Alphabet

WLN notation uses uppercase alphabetic characters, numerals and specific punctuation

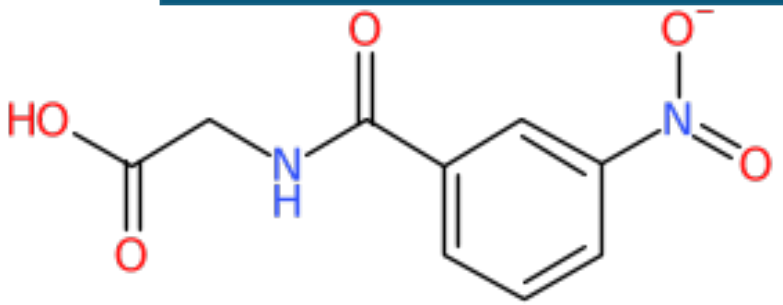


OV1 &OV1 &-CA-

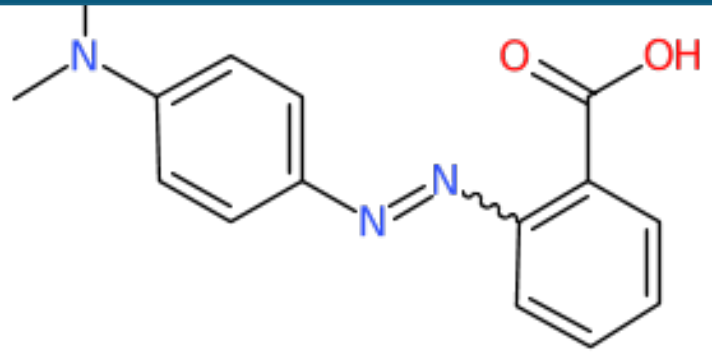




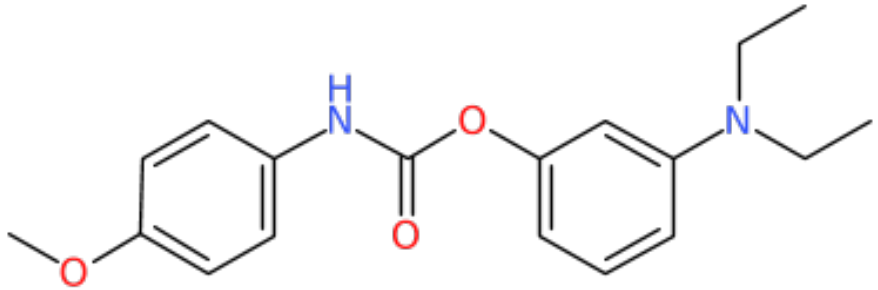
Getting molecules into the Computer - We can do better than SMILES



WNR CVM1VQ



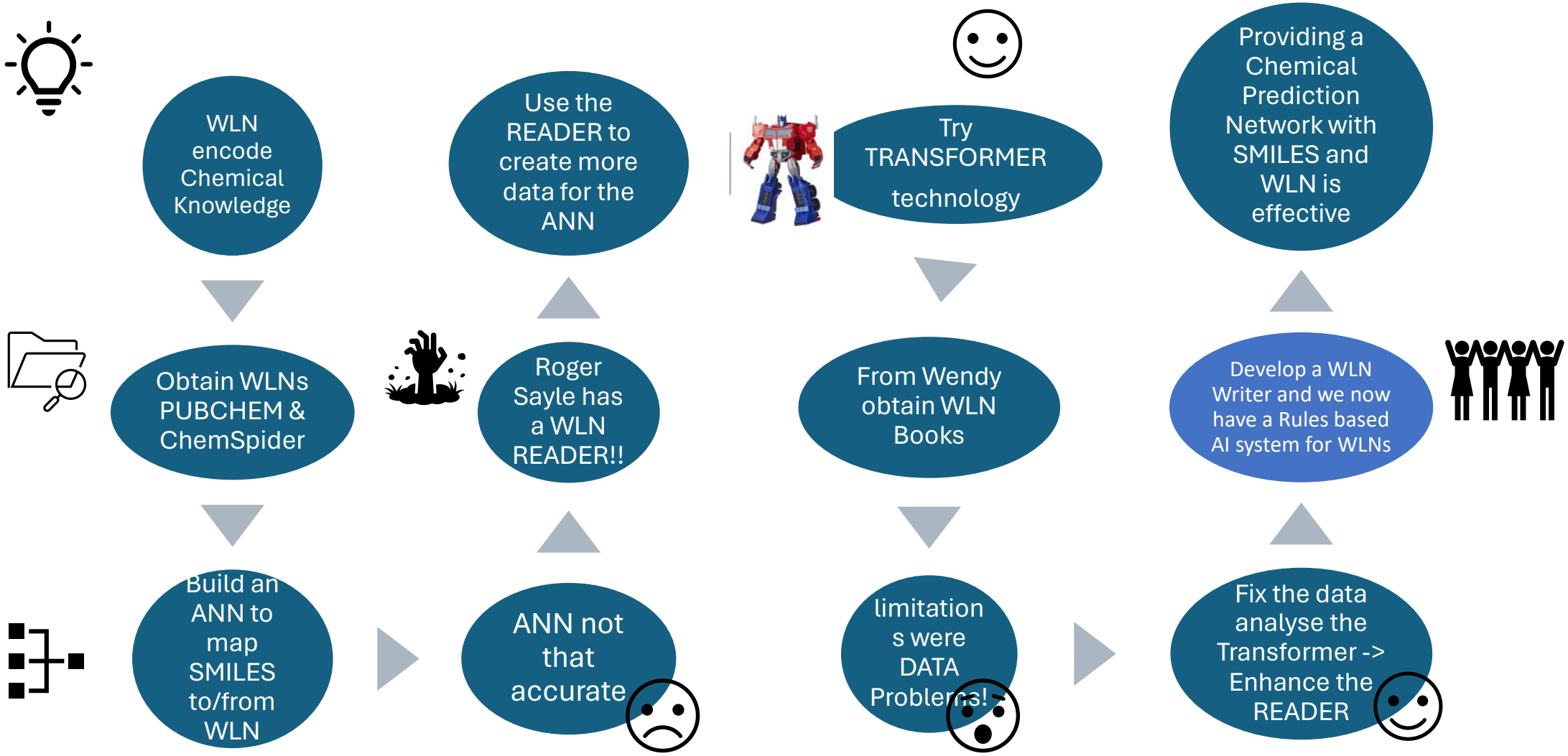
QVR BNUNR DN1&1



2N2&R COVMR DO1

CDKDepict for all image generation

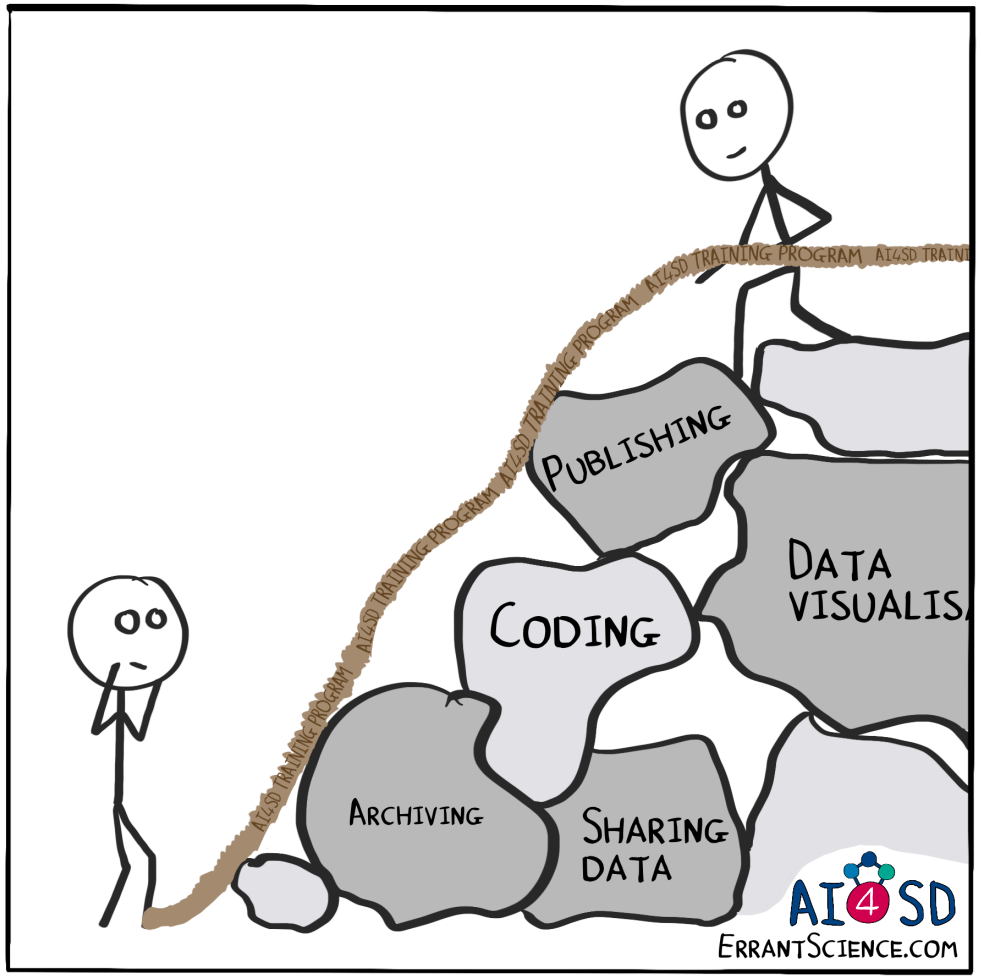
A story of the success of Augmented Intelligence



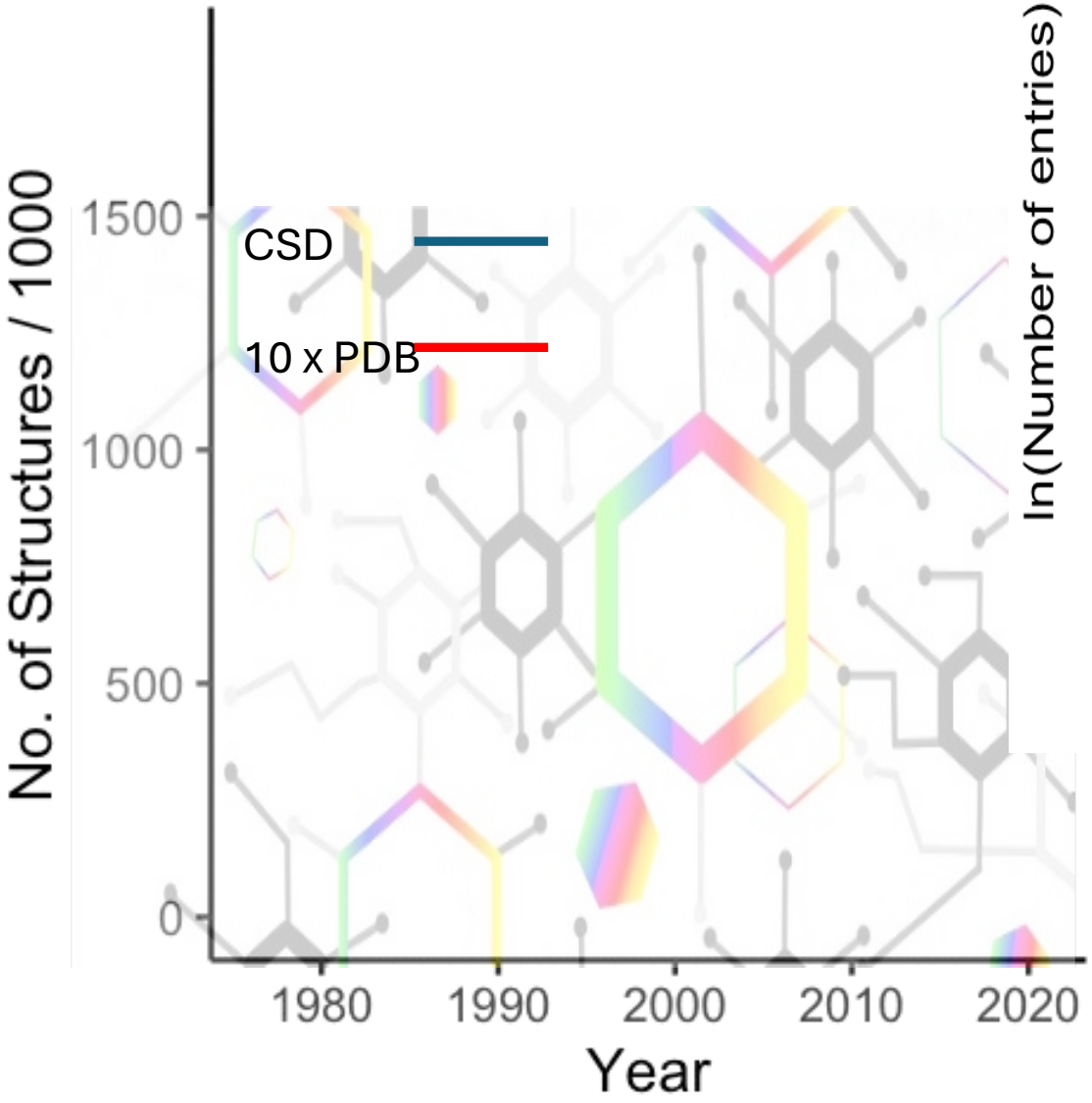
https://github.com/Mblakey/wisswesser

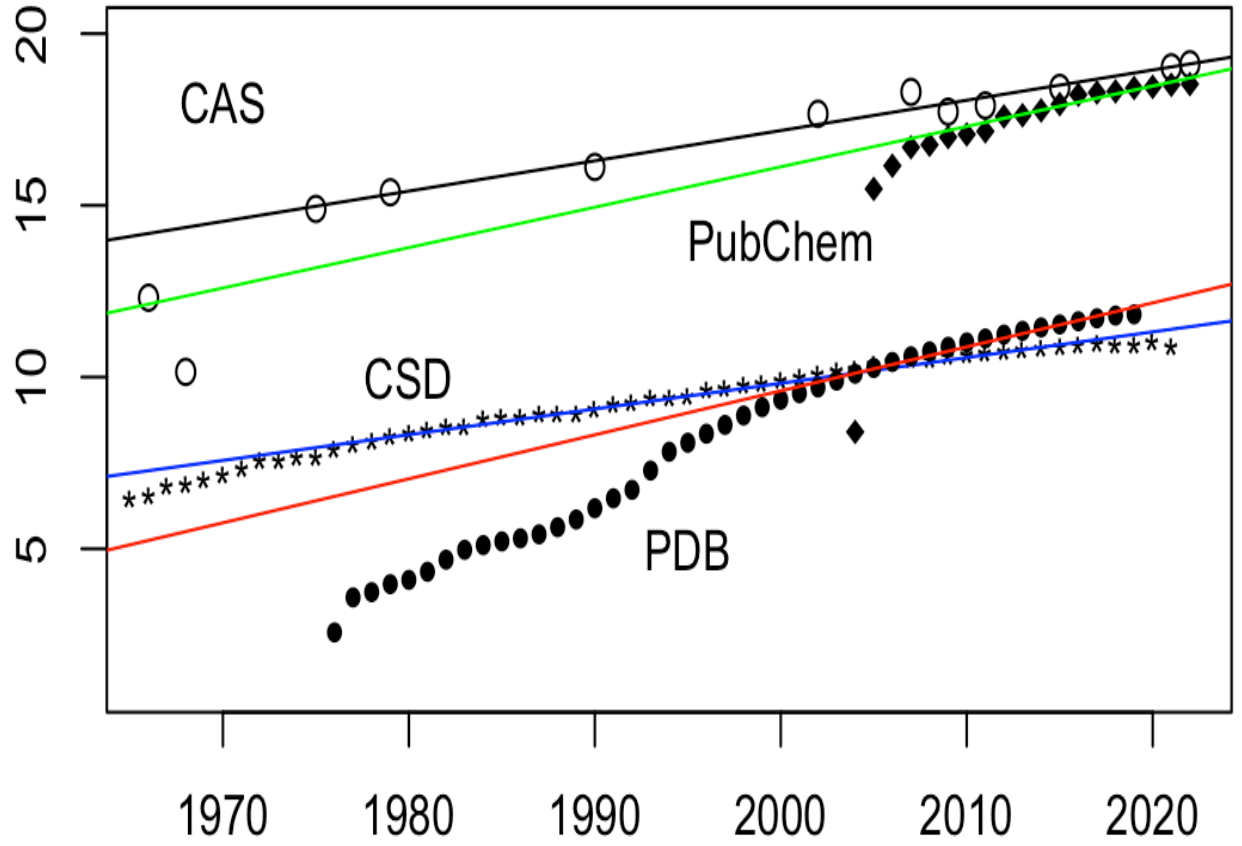
Deep Learning

- AlphaGo,
- We have more data and we can use more data…
- But we rarely have enough in Chemistry
- Need also to include chemical insight
- AlphaFold….



Data, Data everywhere but not enough to model





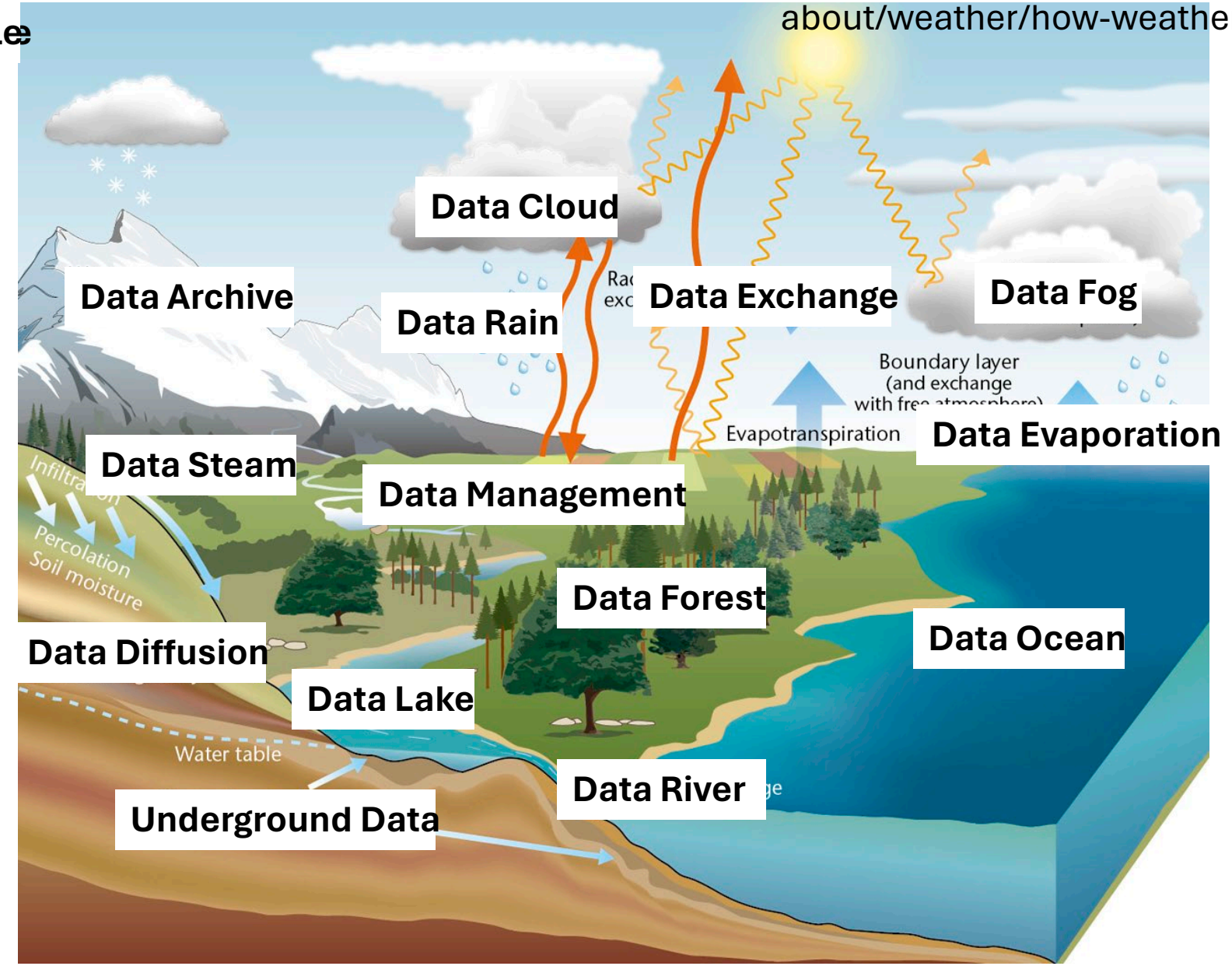
Rapid increase in the number of *available* crystal structures

Small proportion of the number of structures determined.

Data-logical Cycle

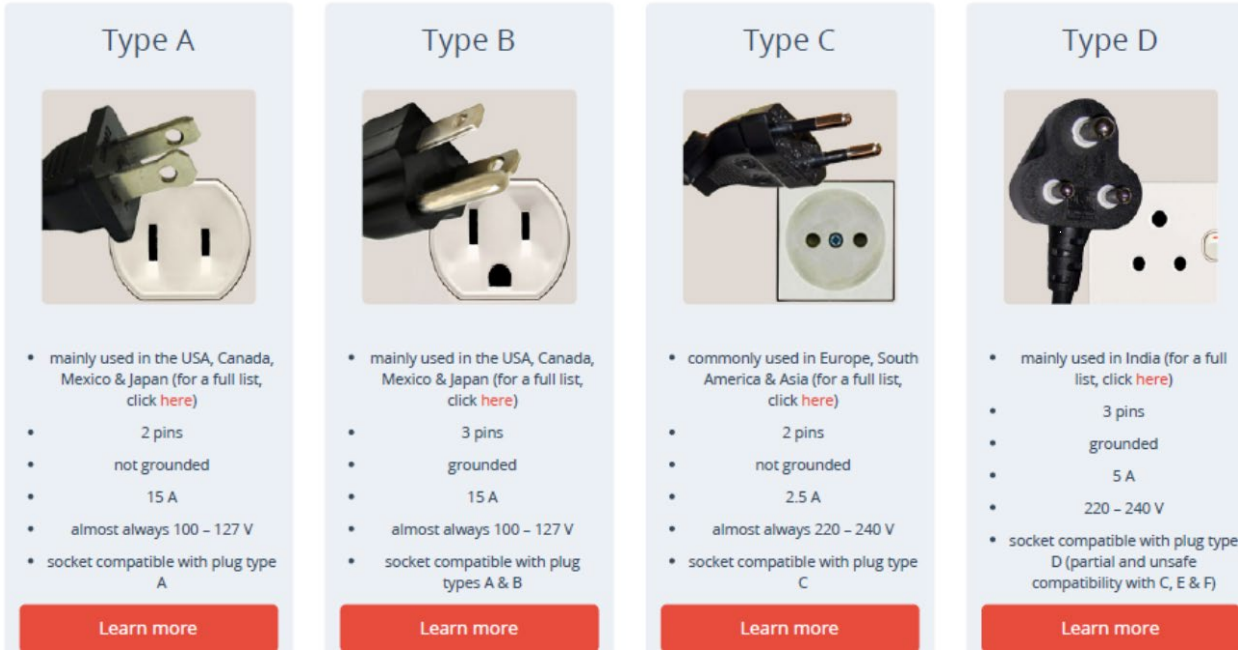
https://www.metoffice.gov.uk/weather/learn-about/weather/how-weather-works/water-cycle

Comets



Solar Wind Erosion

Standards – as much time converting as extracting….

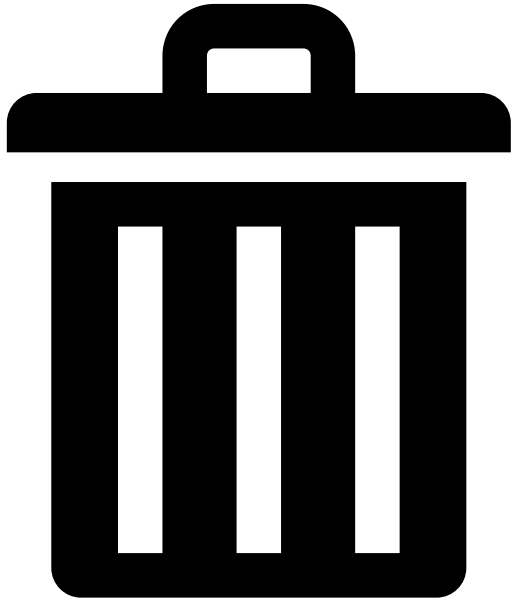


https://www.st-andrews.ac.uk/media/environmental-health-and-safety-services/Electrical-Adapters.pdf

Available from Amazon
Other suppliers are available

(DATA) Standards

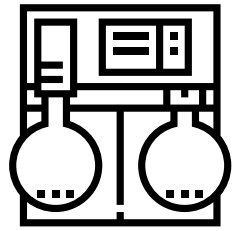
OECD Principles of QSAR (ML) Modelling



- Add a Principle ‘0’ – Characterize the data
- … the original OECD principles did not call out a specific principle to capture the importance of data aggregation and curation prior to modelling,
- Transparency in Modeling through Careful Application of OECD’s QSAR/QSPR Principles via a Curated Water Solubility Data Set,
- C. N. Lowe, *et al*, Chemical Research in Toxicology 2023 36 (3), 465-478

Making Haystacks to find needles

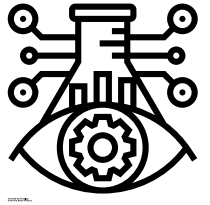
We need to be more insightful, creative, efficient, environmentally conscious…

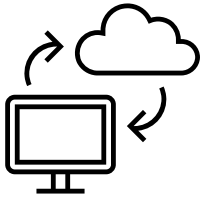


Which requires many good experiments

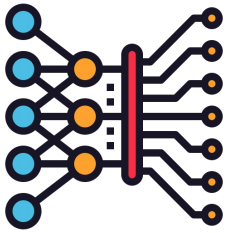
We can't make everything

Need to reduce number of experiments



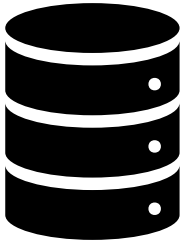


Need help in deciding what to make



We need a good Predictive Model

Needs access to lots of quality data

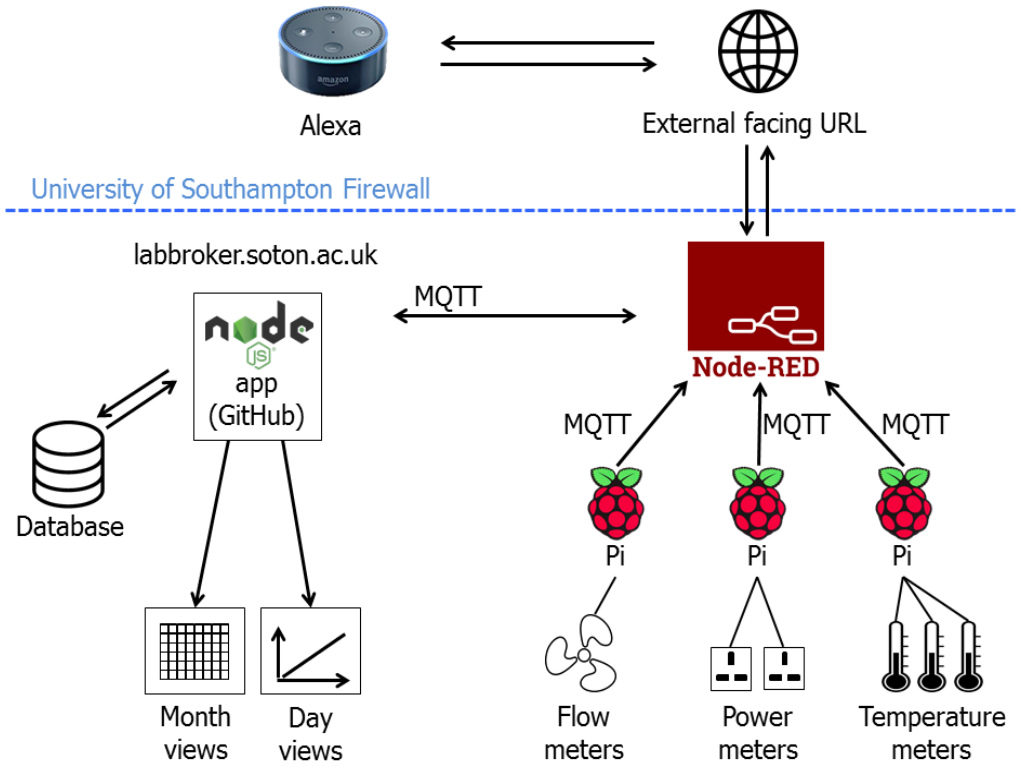


Highlight some relevant types of Machine Learning

Transfer Learning

Re-enforcement Learning (including Human in the Loop)

Smart Labs - Talk to Lab

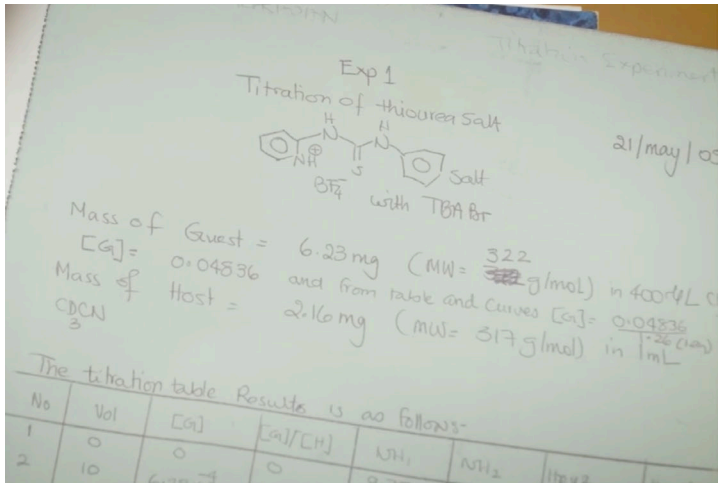


AI needs IA

Information Architecture

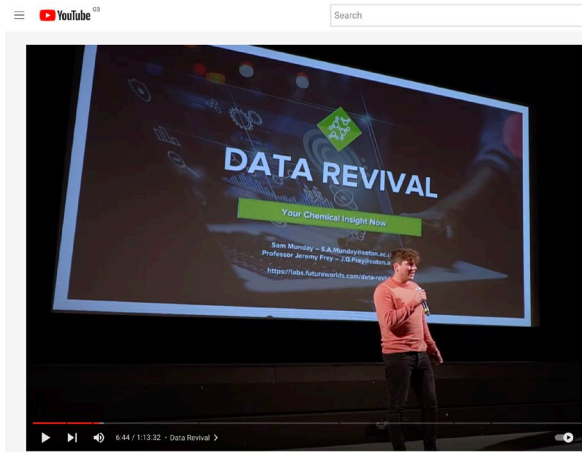
There is a need to restructure labs

DATA REVIVAL *- Recover old inaccessible data*



https://www.data-revival.com

UoS stored Chemistry notebooks
~ 2000 Chemists years
~ 10 Tb scanned data

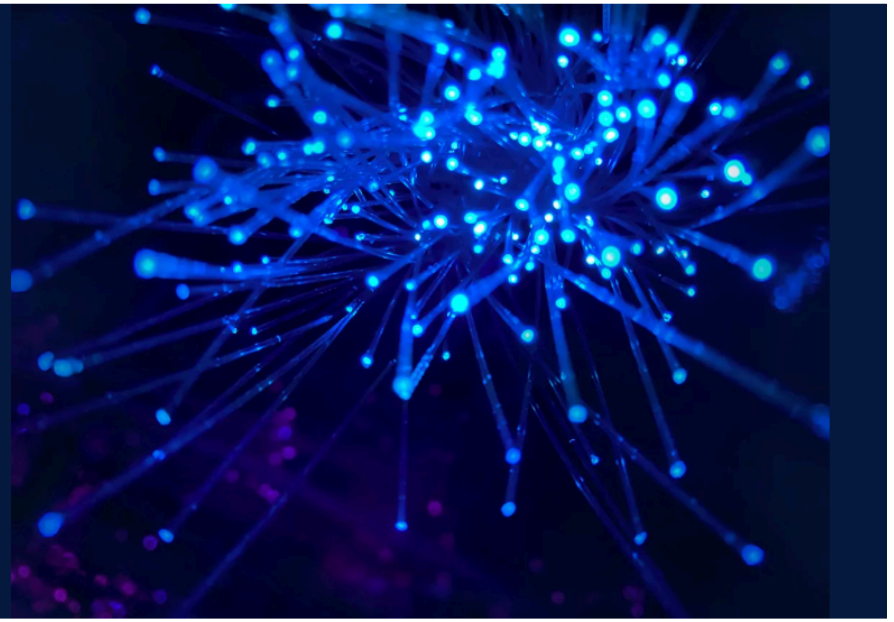


https://www.youtube.com/watch?v=kJalCEEYHQY&t=365s

PSDI
PHYSICAL SCIENCES DATA INFRASTRUCTURE

Home Background News an

Physical Sciences Data Infrastructure



An Integrated Data Infrastructure for the Physical Sciences

PSDI aims to accelerate research in the physical sciences by providing a data infrastructure that brings together and builds upon the various data systems researchers currently use.

Connecting Digital Research Infrastructures for the Physical Sciences

The aim of PSDI is to enable researchers in the physical sciences to handle data more easily by connecting the different data infrastructures they use. PSDI will connect and enhance existing infrastructure in Physical Sciences.

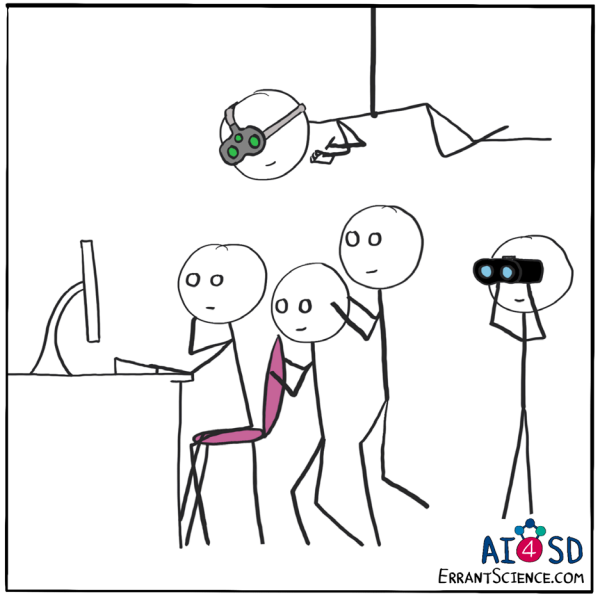
Through PSDI researchers will be able to:

- Find and Access to reference quality data from commercial and open sources
- Combine data from different sources
- Share data, software and models including experimental and simulation data
- Use AI to explore data
- Learn how to make the results of their research open and FAIR

Let's Have a Chat

2000's e-Science Programme recognised a growing need for the global (virtual) equivalent of the "Tea Room"

Covid mitigation strategy built on these online collaborations



AI & ML
ChatGPT
Large Language Models

Large Language Models

Very powerful

Dilemma for teaching

Can be used securely and on well defined data

Need careful prompting

May well be able to do much of the heavy lifting in data science

7342.pdf (6 pages)

Water surface is acidic

Victoria Buch*, Anne Milet†, Robert Vácha‡, Pavel Jungwirth‡§, and J. Paul Devlin¶

PNAS

ScienceDaily®
Your source for the latest research news

Mobile: iPhone Android Web Follow: Facebook Twitter Google+

HEALTH | PHYSICAL/TECH | ENVIRONMENT | SOCIETY/EDUCATION | QUIRKY

Latest Headlines | Health & Medicine | Mind & Brain | Space & Time | Matter & Energy | Computers & Math | Plan

Featured Research — *from universities, journals, and other organizations*

Rewrite the textbooks on water's surface tension: Air-water interface is negatively charged by the adsorption of hydroxide ions

Date: March 19, 2014

Source: University of Melbourne

Summary: Researchers in Australia make significant differen water used by the next g biophysicists and engine latest investigations hav the physical Chemistry c charged by the adsorptic

Share This

Related Topics

Matter & Energy
- Nature of Water
- Chemistry
- Inorganic Chemistry
- Thermodynamics
- Biochemistry
- Nuclear Energy

e Moléculaire, Unité Mixte de Recherche
tute of Organic Chemistry and
cular Systems, 16610 6 Prague,

December 19, 2006)

used. First, constant volume and
io molecular dynamics simulations
water with a pair of H_3O^+ and OH^-
ls, the simulation was carried out for
boundaries and a slab with two open
oundaries. This "on-the-fly" tech-
modeling of proton transfer systems
on transfer and transitions between
ter forms are automatically included
e. However, the high computational
y of the model system, the duration

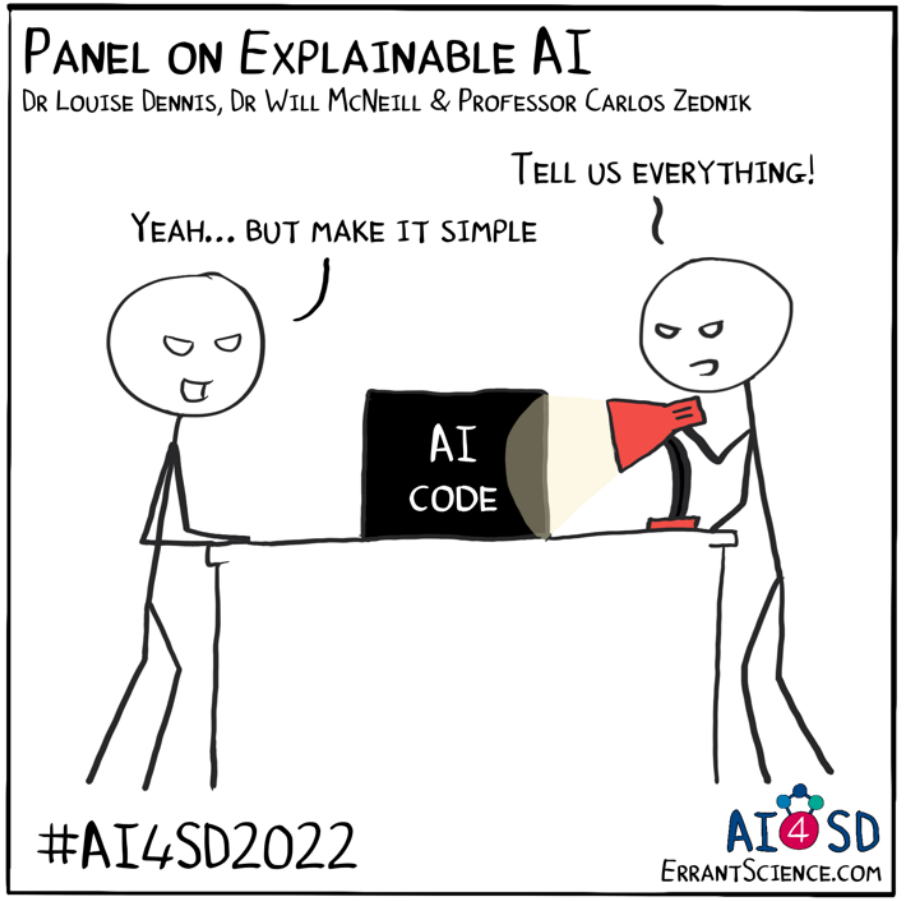
PAPER www.rsc.org/faraday_d | Faraday Discussions

The surface of neat water is basic

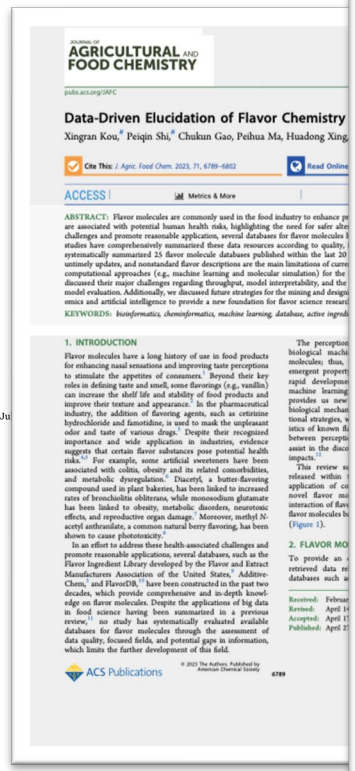
James K. Beattie,* Alex M. Djerdjev and Gregory G. Warr

Received 28th March 2008, Accepted 28th May 2008
First published as an Advance Article on the web 18th September 2008
DOI: 10.1039/b805266b

Ethical & Explainable AI
Needed for scientific discovery



Trends in Food Science & Technology 138 (2023) 738–751



30 June 2023; Accepted 21 Ju

Food Chemistry: Molecular Sciences 4 (2022) 100090

Contents lists available at ScienceDirect

Food Chemistry: Molecular Sciences

journal homepage: www.sciencedirect.com/journal/food-chemistry-molecular-sciences

ChemTastesDB: A curated database of molecular tastants

Cristian Rojas [a,*], Davide Ballabio [b], Karen Pacheco Sarmiento [a], Elisa Pacheco Jaramillo [a], Mateo Mendoza [a], Fernando García [c]

[a] Grupo de Investigación en Quimiometría y QSAR, Facultad de Ciencia y Tecnología, Universidad del Azuay, Av. 24 de Mayo 7-77 y Hernán Malo, Cuenca, Ecuador
[b] Milano Chemometrics and QSAR Research Group. Department of Earth and Environmental Sciences, University of Milano-Bicocca, P.za della Scienza 1-20126, Milano, Italy
[c] Facultad de Ciencias Económicas, Universidad Nacional de Córdoba. Centro de Investigaciones en Ciencias Económicas, Grupo vinculado CIECS – UNC – CONICET, Córdoba, Argentina

ARTICLE INFO

Keywords:
ChemTastesDB
Database
Tastes
Chemical space
Foodinformatics

ABSTRACT

The purpose of this work is the creation of a chemical database named *ChemTastesDB* that includes both organic and inorganic tastants. The creation, curation pipeline and the main features of the database are described in detail. The database includes 2944 verified and curated compounds divided into nine classes, which comprise the five basic tastes (sweet, bitter, umami sour and salty) along with four additional categories: tasteless, non-sweet, multitaste and miscellaneous. *ChemTastesDB* provides the following information for each tastant: name, PubChem CID, CAS registry number, canonical SMILES, class taste and references to the scientific sources from which data were retrieved. The molecular structure in the HyperChem (*.hin*) format of each chemical is also made available. In addition, molecular fingerprints were used for characterizing and analyzing the chemical space of tastants by means of unsupervised machine learning. *ChemTastesDB* constitutes a useful tool to the scientific community to expand the information of taste molecules and to assist *in silico* studies for the taste prediction of unevaluated and as yet unsynthetized compounds, as well as the analysis of the relationships between molecular structure and taste. The database is freely accessible at https://doi.org/10.5281/zenodo.5747393.

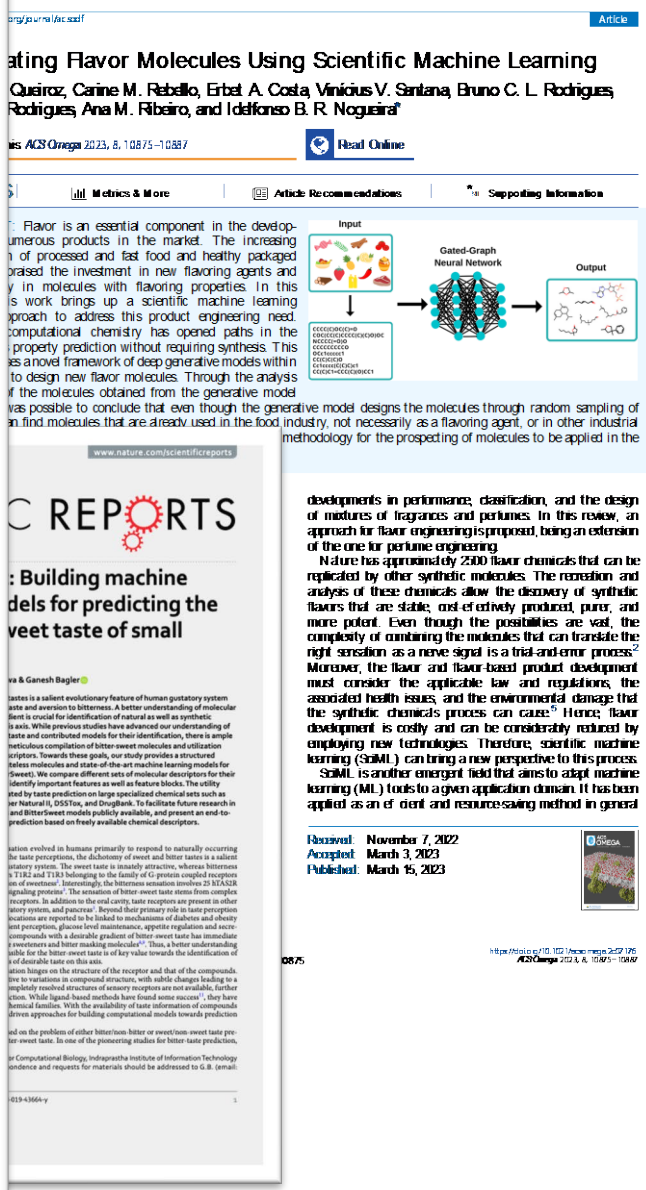
1. Introduction

The sensation of taste plays an important role in the food chemistry field, since it is closely related to the development and selection of food products and food intake. Throughout history, there has been a strong interest in understanding the mechanism by which gustatory sensation is perceived by humans (Damodaran & Parkin, 2017). The extraordinary developments in foodinfomatics (computational food chemistry) and bioinformatics (computational biochemistry) have provided the necessary tools to study the receptor/ligand binding interaction. In order to achieve a particular taste, it is now understood that the structure of the receptors and the specific features of the tastant ligands to interact with receptors must be analyzed (Chandrashekar et al., 2006; Rojas et al., 2016a). A molecular tastant is a water-soluble chemical compound (ligand) able to interact with the chemosensory receptors to produce a taste sensation (Di Lorenzo et al., 2009). The taste-receptor cells (TRCs) are located in the gustatory papillae of the tongue and palate epithelium, which react to tastants by means of receptor-ligand interactions along with other mechanisms. These additional mechanisms are associated with the opening of ion channels or through secondary messenger channels associated with nucleotides or phosphorylated inositol (Damodaran & Parkin, 2017; Di Lorenzo et al., 2009; Wong, 2018). Evidence suggests that there are five basic tastes (sweet, bitter, umami, sour and salty), which are also known as "*taste modalities*" or "*receptor-mediated tastes*" (Chandrashekar et al., 2006; Morini et al., 2011).

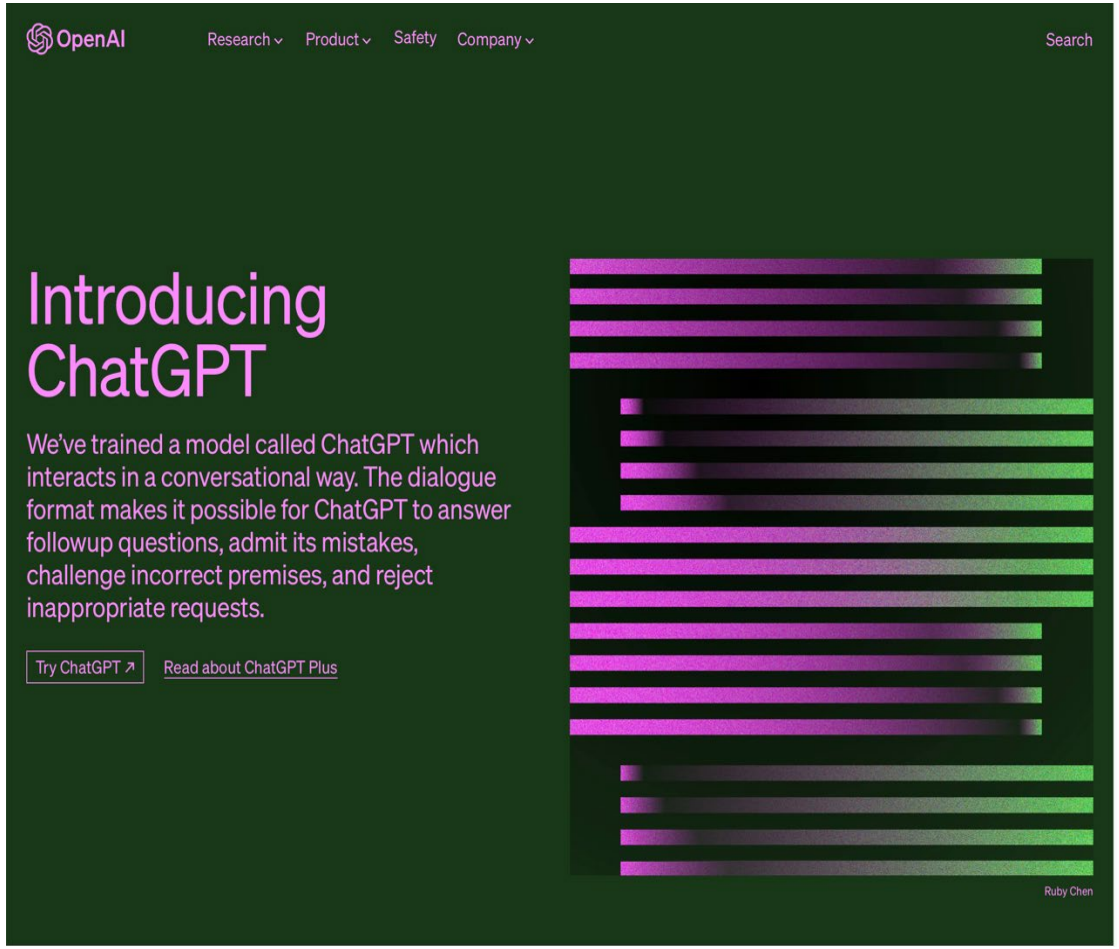
Among the basic tastes, sweetness is probably the most important, since sweeteners evoke a pleasant sensation in several foods and medicines (Chandrashekar et al., 2006; Damodaran & Parkin, 2017). *Sucrose* is used as a standard to quantify the relative sweetness (RS) of new sweet-tasting molecules (Rojas et al., 2016a; Rojas et al., 2016b). The sweet taste chemoreceptor is a G-protein coupled receptor (GPCR) of class C made up of T1R2/T1R3 subunits (Chandrashekar et al., 2006; Morini et al., 2011). In contrast to the pleasant sensation of sweetness, bitterness may be related to the protection of humans from the consumption of toxic compounds (Chandrashekar et al., 2006; Di Lorenzo et al., 2009), although in some foods or products it is perceived as a

* Corresponding author.
E-mail address: crojasvilla@gmail.com (C. Rojas).

https://doi.org/10.1016/j.fochms.2022.100090
Received 23 December 2021; Received in revised form 17 February 2022; Accepted 18 February 2022
Available online 21 February 2022



ChatGPT – will an AI win the Chemistry Nobel Prize?



As an AI language model, I cannot predict the future… …. it is possible for an AI to contribute significantly to chemistry research that could be awarded a Nobel Prize

However, it is important to note that the Nobel Prize is awarded to individuals or groups of individuals, not to machines or algorithms….

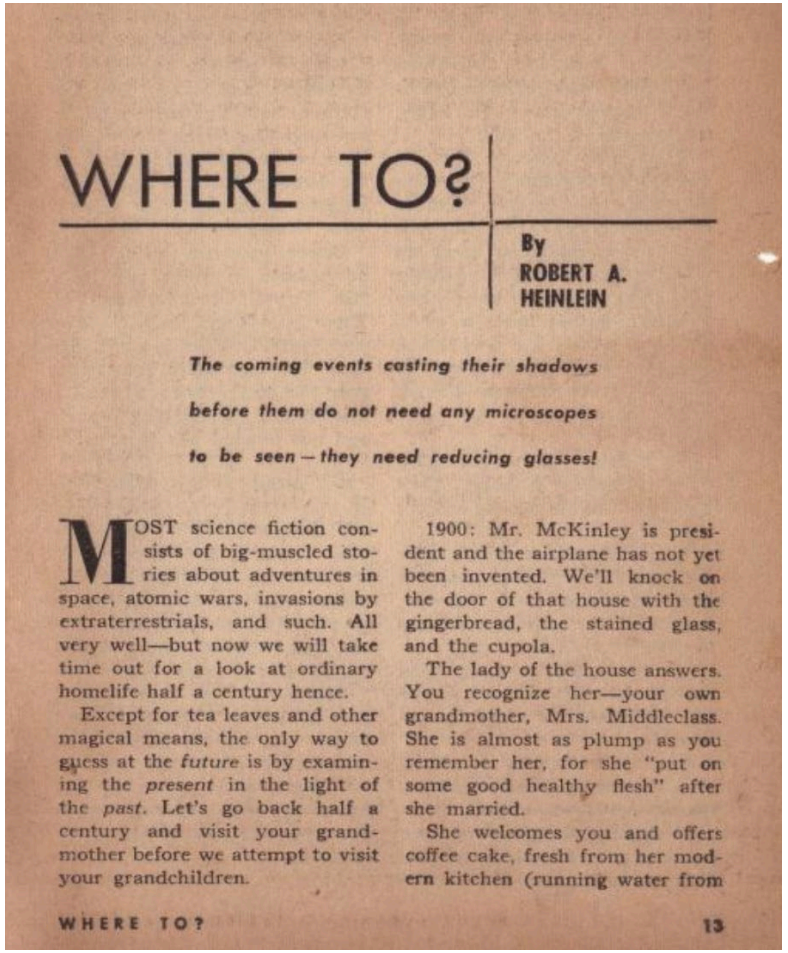
Even if AI plays a critical role… the prize would likely be awarded to the human scientists who developed and applied the AI methods.

Further thoughts from the 1950's

The Future
(*is* ours to see)

***When chemistry becomes a discipline*, mathematical chemists will design new materials, predict their properties, and tell engineers how to make them — without ever entering a laboratory.**

Robert A. Heinlein, "Where to?" 1950



WHERE TO?

By ROBERT A. HEINLEIN

The coming events casting their shadows before them do not need any microscopes to be seen — they need reducing glasses!

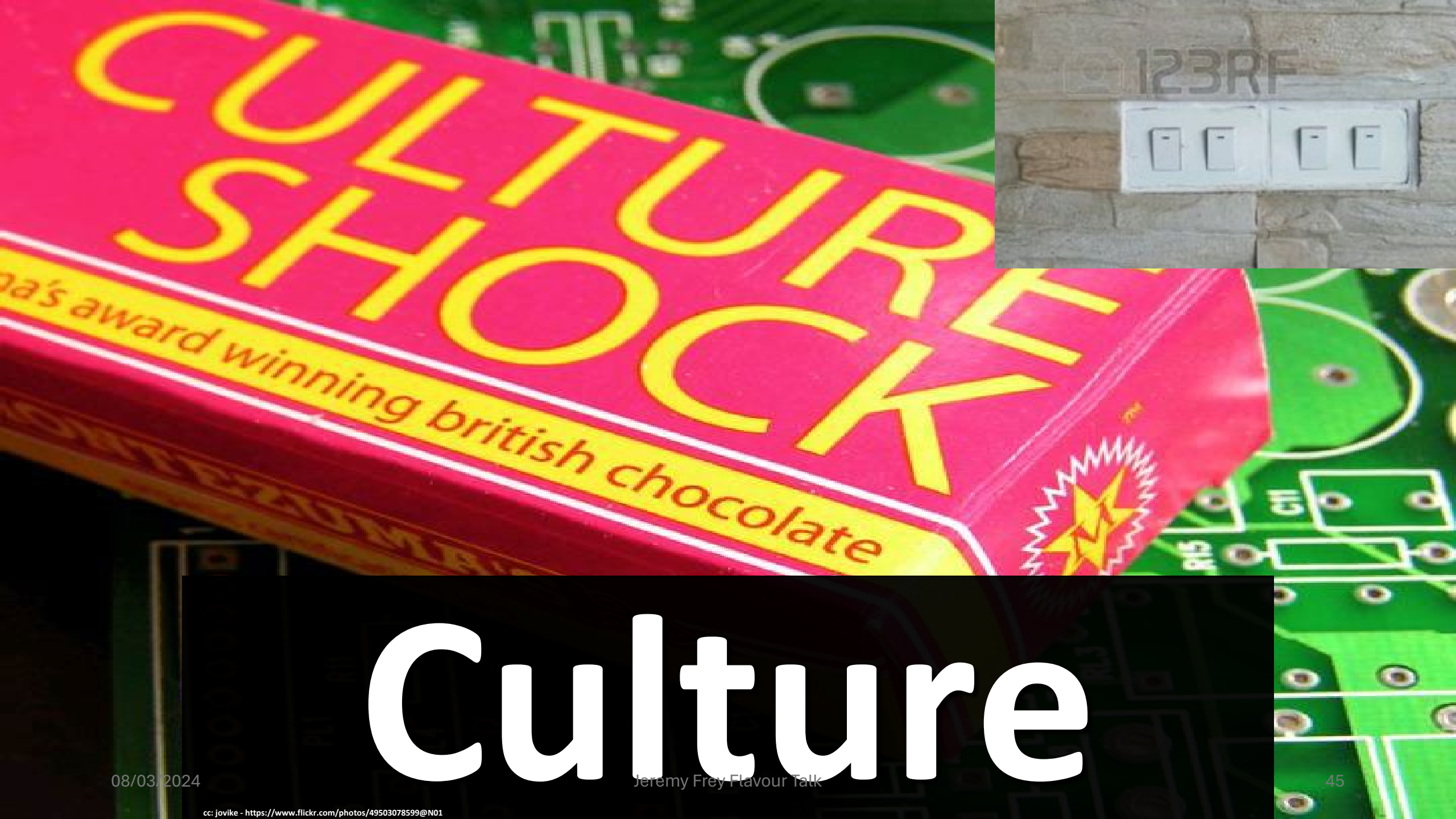
MOST science fiction consists of big-muscled stories about adventures in space, atomic wars, invasions by extraterrestrials, and such. All very well—but now we will take time out for a look at ordinary homelife half a century hence.

Except for tea leaves and other magical means, the only way to guess at the *future* is by examining the *present* in the light of the *past*. Let's go back half a century and visit your grandmother before we attempt to visit your grandchildren.

1900: Mr. McKinley is president and the airplane has not yet been invented. We'll knock on the door of that house with the gingerbread, the stained glass, and the cupola.

The lady of the house answers. You recognize her—your own grandmother, Mrs. Middleclass. She is almost as plump as you remember her, for she "put on some good healthy flesh" after she married.

She welcomes you and offers coffee cake, fresh from her modern kitchen (running water from



Culture

Thank you for listening

Thanks to all the brilliant members of my research group and my excellent colleagues, that made this work possible

Trust me Mort - no electronic communications superhighway, no matter how vast and sophisticated, will ever replace the art of the schmooze

All I am saying is that now is the time to develop the technology to deflect an asteroid