UNIVERSITY OF CAMBRIDGE

Machine Learning for Chemists

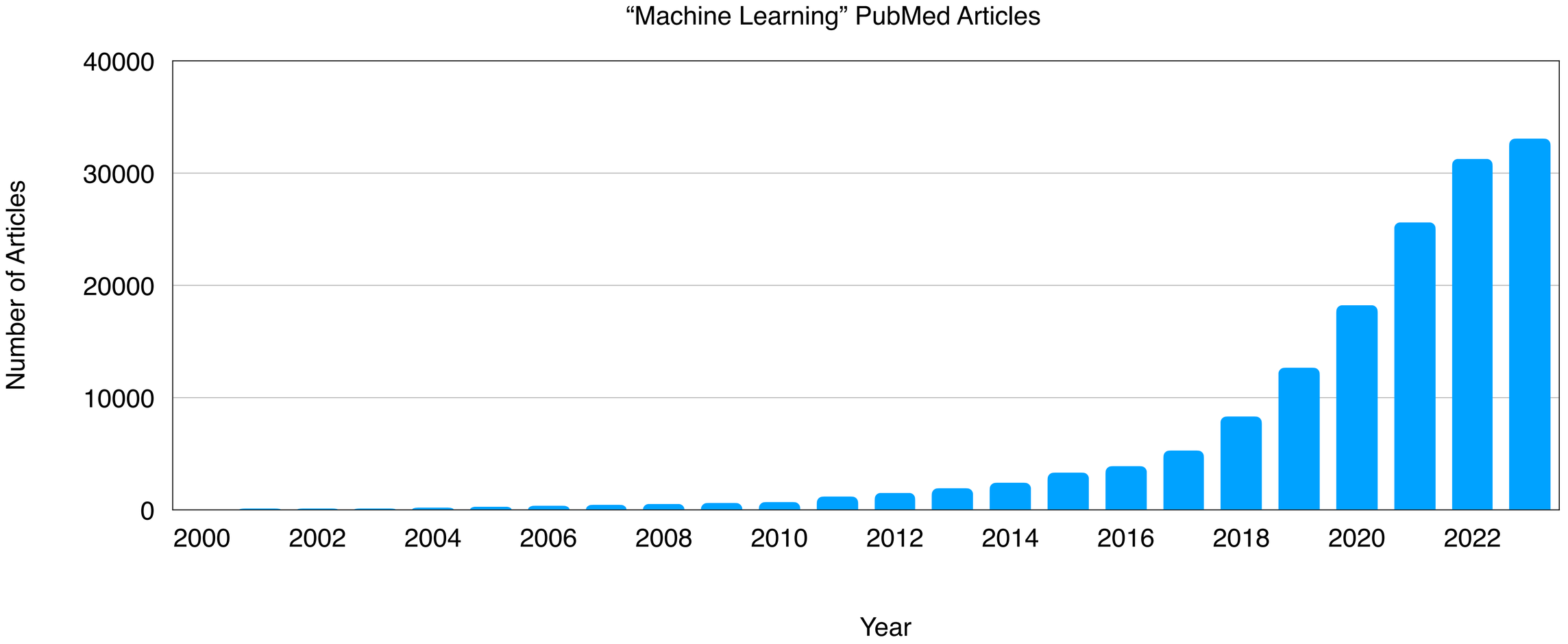
– Introduction –

---

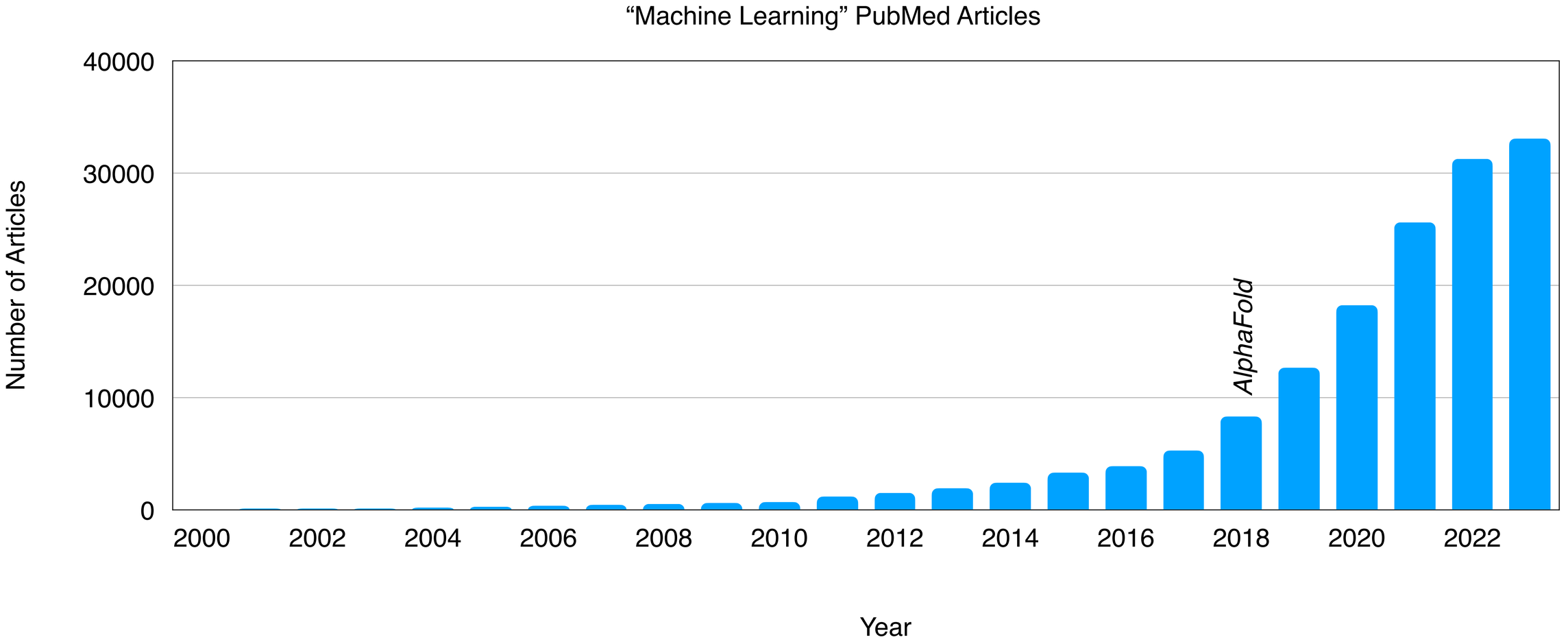
5 February 2024

The Next Trend: Machine Learning

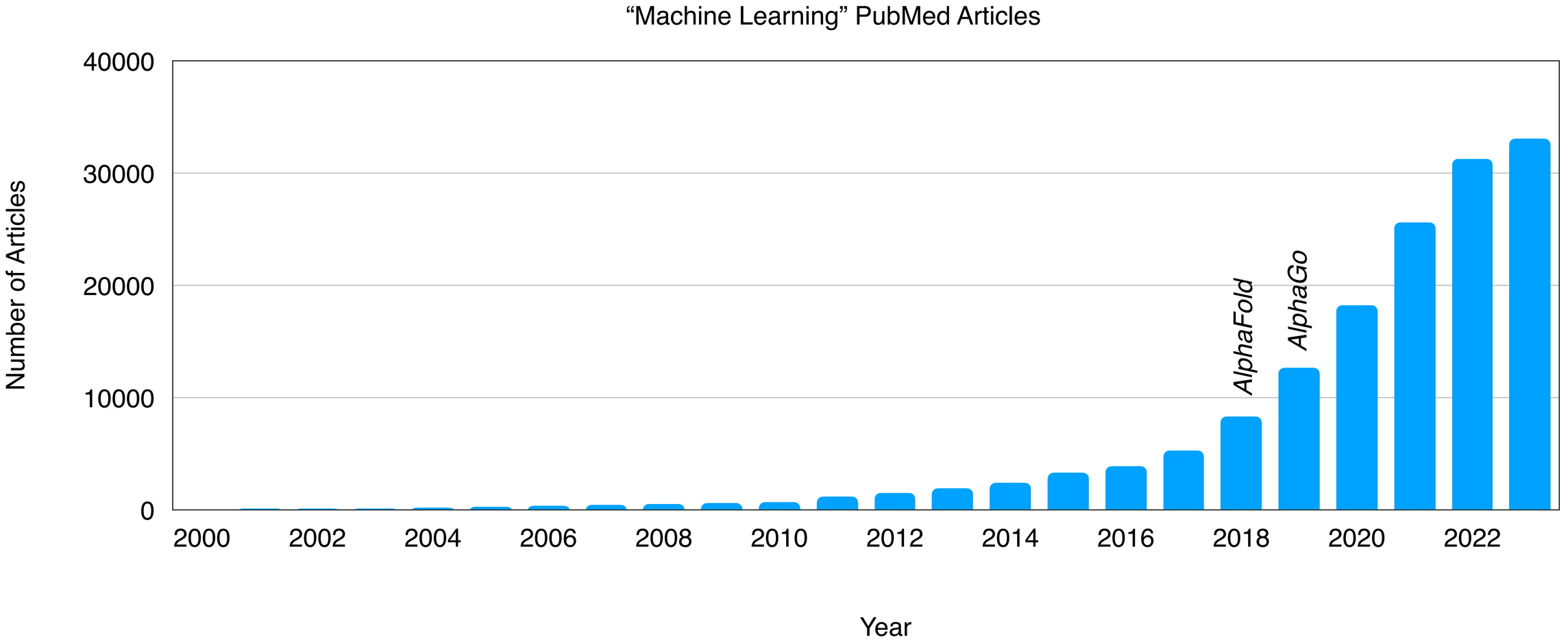
UNIVERSITY OF CAMBRIDGE



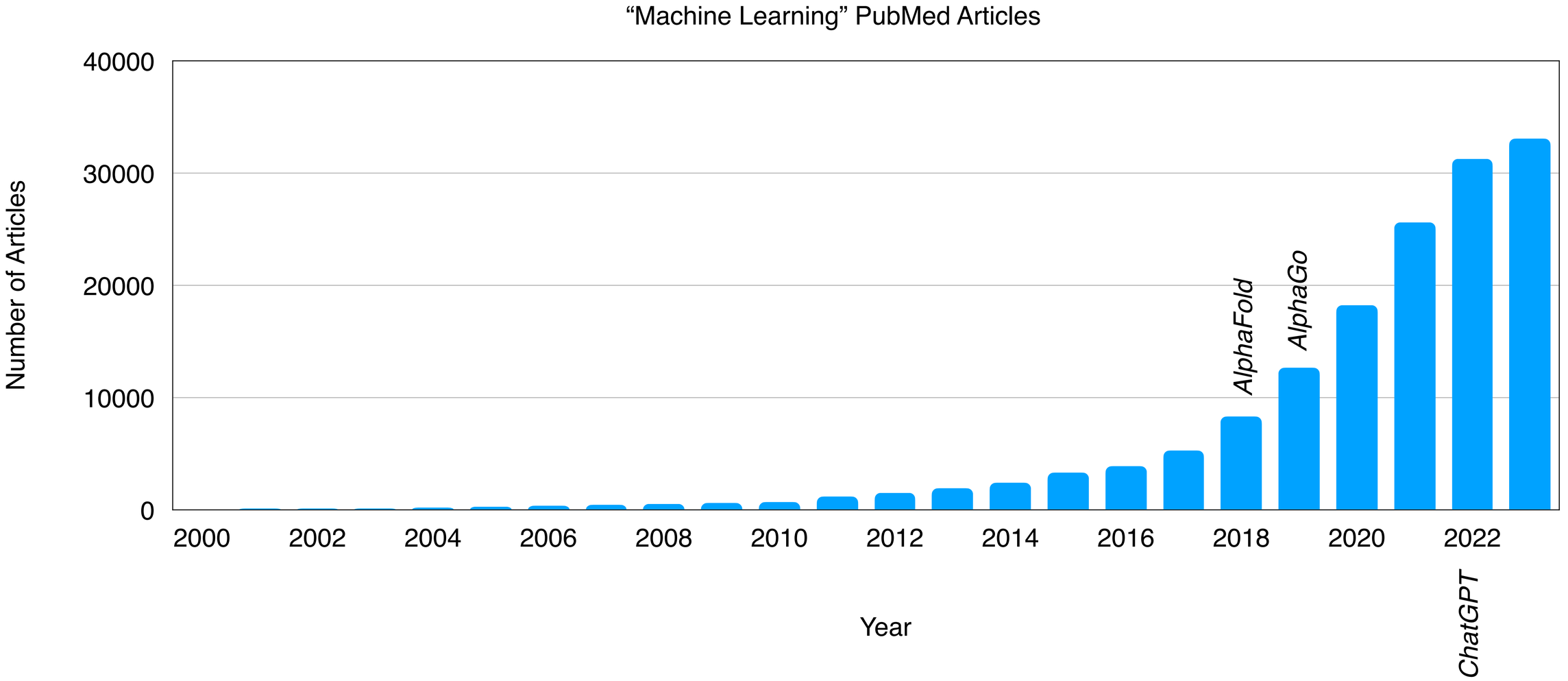
The Next Trend: Machine Learning



The Next Trend: Machine Learning



The Next Trend: Machine Learning



The Current New Toys

UNIVERSITY OF CAMBRIDGE

ChatGPT

OpenAI

Claude

Anthropic

Midjourney

Midjourney, Inc.

Leela Chess Zero

Based on Deepmind's AlphaGo Zero

Open-source (free to use)

What To Expect

UNIVERSITY OF CAMBRIDGE

“Big picture” reasoning behind common ML techniques

PCA, SVM, Random Forest, Gaussian Process, Feed Forward Neural Networks

Basic steps in running the algorithms

Some helpful debugging tips

Feel free to email or come with questions!

What Isn’t Covered

UNIVERSITY OF CAMBRIDGE

Python - it’s flaws and intricacies

Rigorous mathematical proofs behind the ML techniques

A comprehensive review of ML techniques

Feel free to ask questions about things not covered!

The Basic Idea Behind Machine Learning

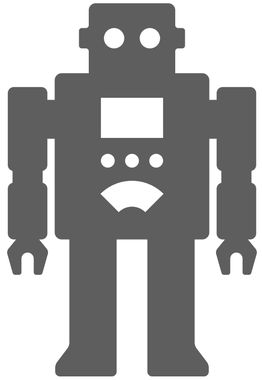
UNIVERSITY OF CAMBRIDGE

This is how we represent ML:

The Basic Idea Behind Machine Learning

UNIVERSITY OF CAMBRIDGE

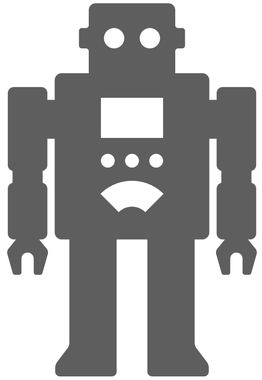
This is how we **should** represent ML:



The Basic Idea Behind Machine Learning

UNIVERSITY OF CAMBRIDGE

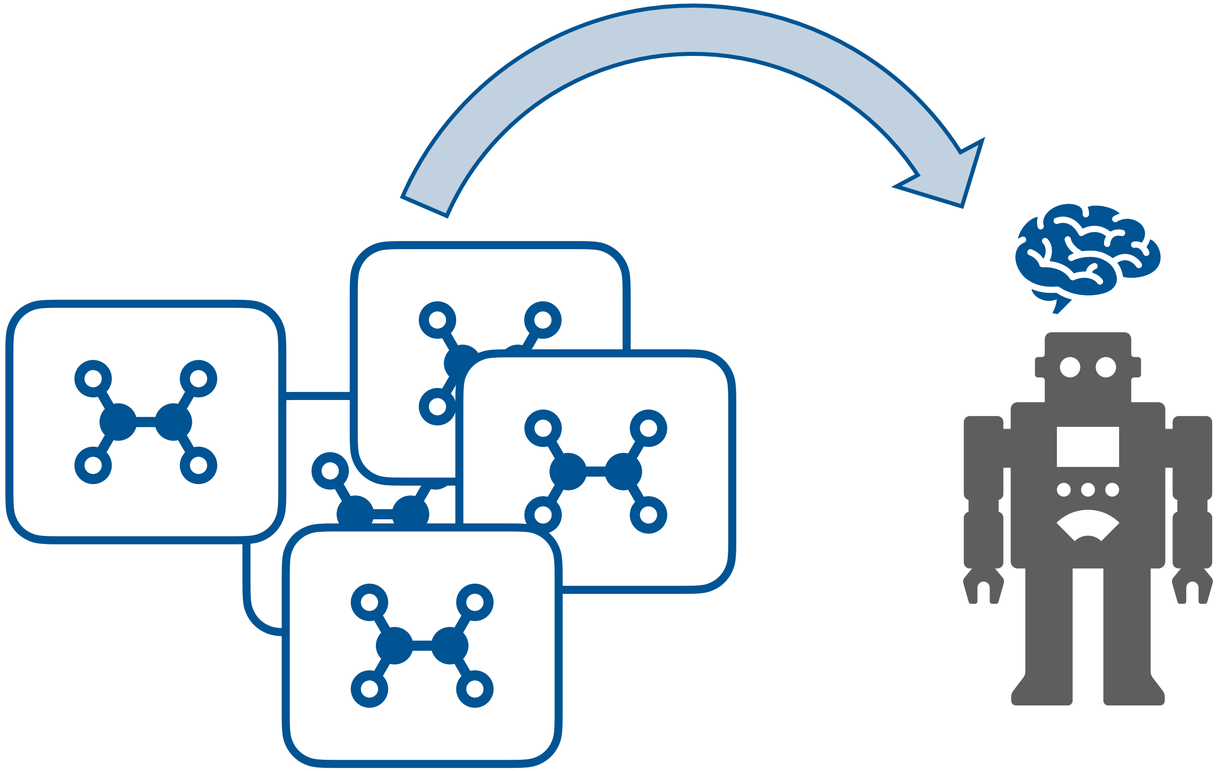
This is how we **should** represent ML:



Very good at *one* task (e.g., good at understanding chemistry / molecules)

How do we do that?

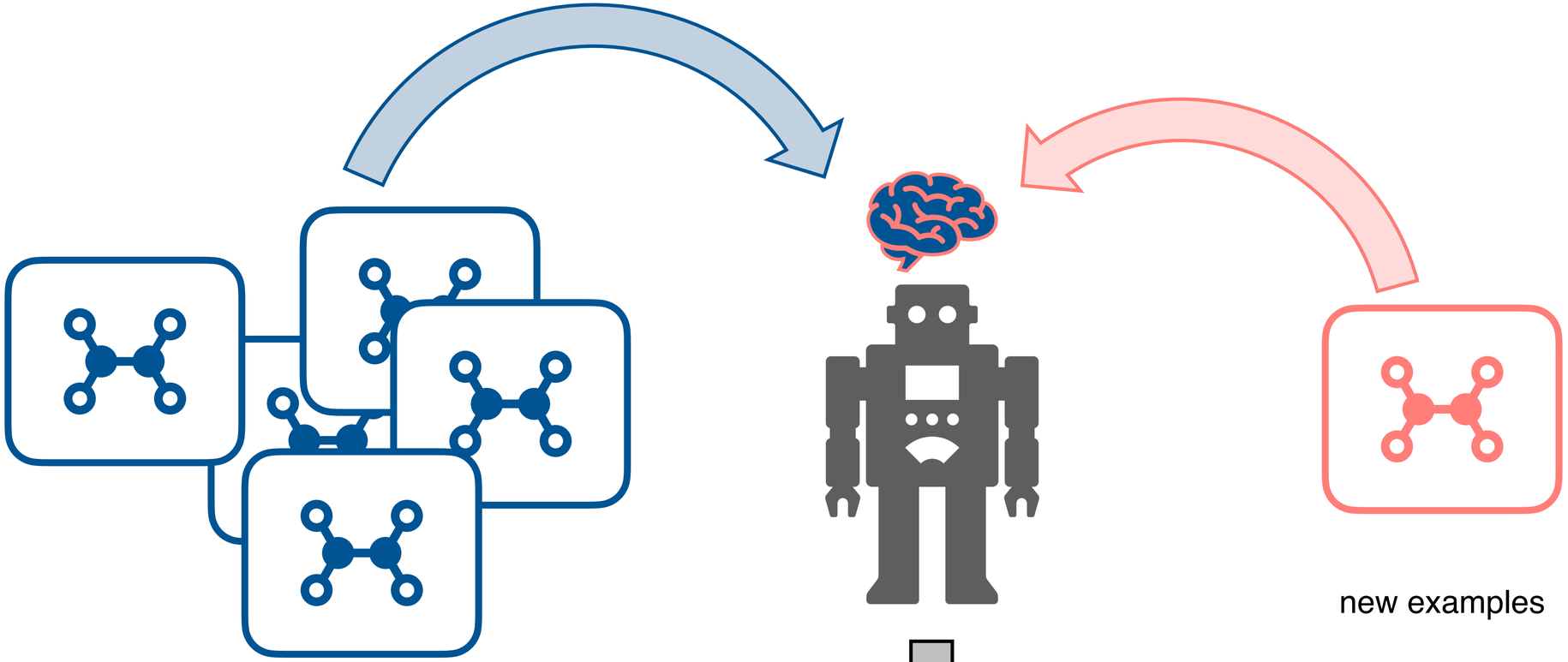
The Basic Idea Behind Machine Learning



examples

(train set)

The Basic Idea Behind Machine Learning



examples

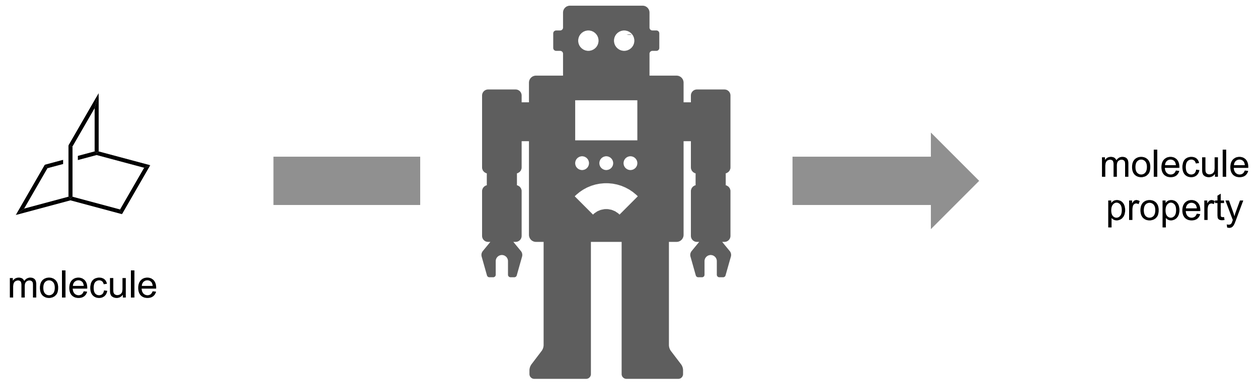
(train set)

evaluation of chemistry knowledge

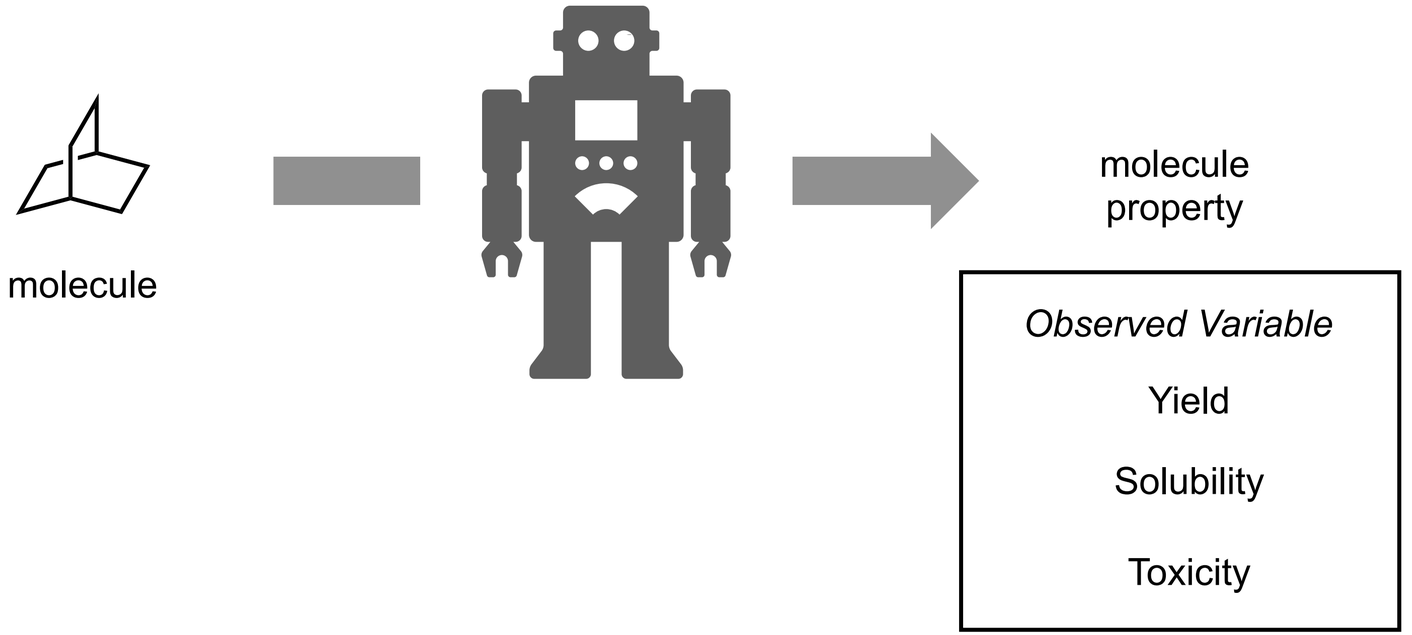
new examples

(test set)

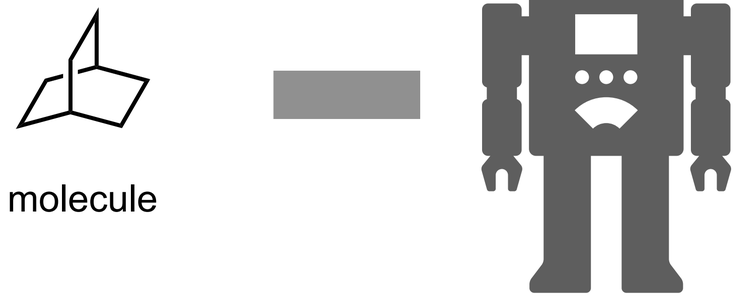
What is Machine Learning, Really?



What is Machine Learning, Really?



What is Machine Learning, Really?



molecule

molecule
property

Computed Variable

HOMO-LUMO gap

Vibrational Energies

Stability

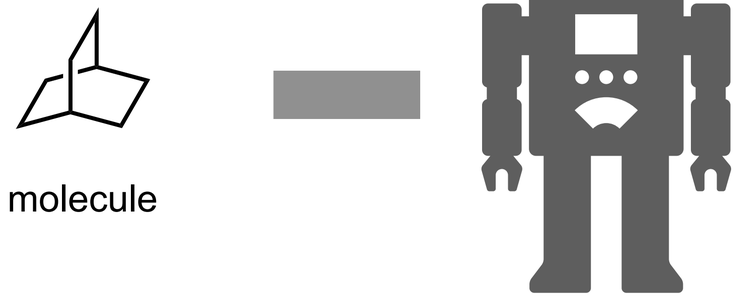
Observed Variable

Yield

Solubility

Toxicity

What is Machine Learning, Really?



molecule

molecule
property

Computed Variable

HOMO-LUMO gap

Vibrational Energies

Stability

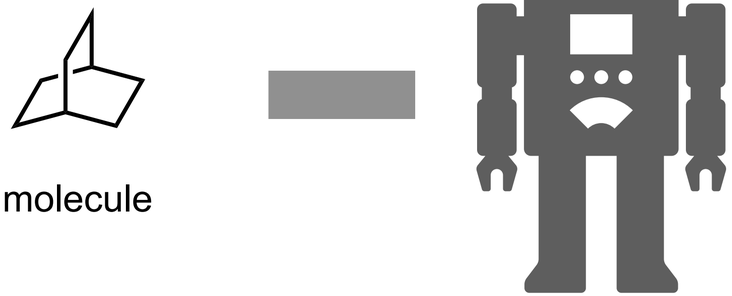
Observed Variable

Yield

Solubility

Toxicity

What is Machine Learning, Really?



molecule

Computer readable format

Made up of numbers

molecule property

Computed Variable

HOMO-LUMO gap

Vibrational Energies

Stability

Observed Variable

Yield

Solubility

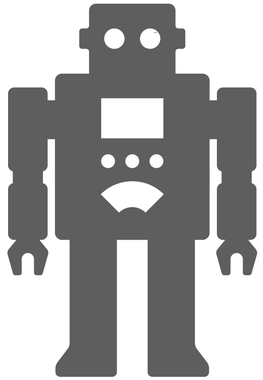
Toxicity

What is Machine Learning, Really?

molecule

Computer readable format

Made up of numbers



molecule property

Computed Variable

HOMO-LUMO gap

Vibrational Energies

Stability

Observed Variable

Yield

Solubility

Toxicity

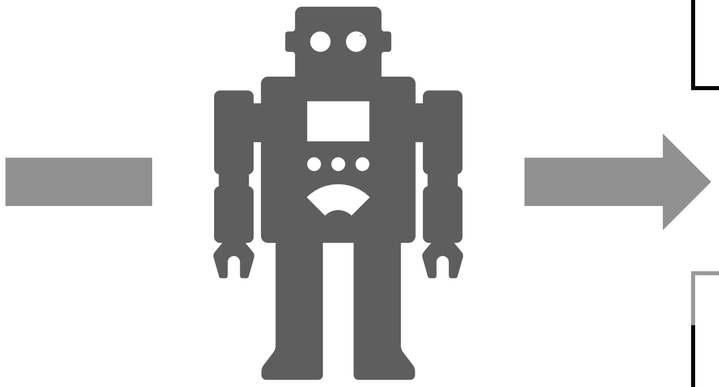
Choose a property

Choose a molecular representation

Choose an algorithm

What is Machine Learning, Really?

molecule



Computed Variable

HOMO-LUMO gap

Vibrational Energies

Stability

molecule property

Observed Variable

Yield

Solubility

Toxicity

Computer readable format

Made up of numbers

EASY!

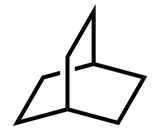
Choose a property

Choose a molecular representation

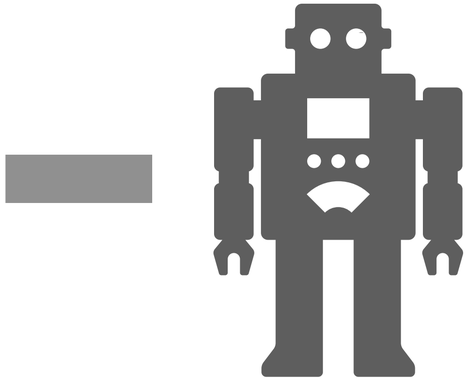
Choose an algorithm

What are you interested in?

What is Machine Learning, Really?



molecule



Computed Variable

HOMO-LUMO gap

Vibrational Energies

Stability

molecule property

Observed Variable

Yield

Solubility

Toxicity

Computer readable format

Made up of numbers

EASY!

Choose a property

What are you interested in?

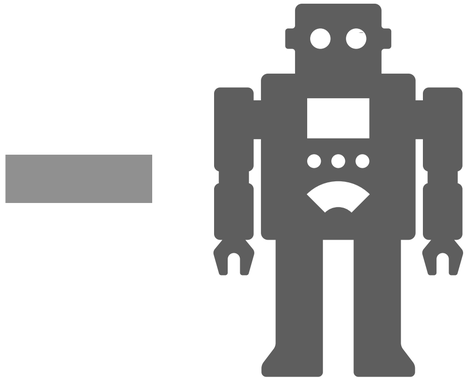
Choose a molecular representation

Choose an algorithm

Subsequent lectures

What is Machine Learning, Really?

molecule



Computed Variable

HOMO-LUMO gap

Vibrational Energies

Stability

molecule property

Observed Variable

Yield

Solubility

Toxicity

Computer readable format

Made up of numbers

EASY!

Choose a property

Choose a molecular representation

Choose an algorithm

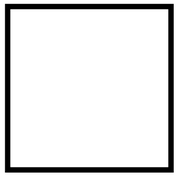
What are you interested in?

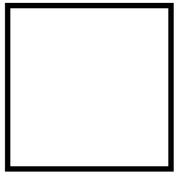
Subsequent lectures

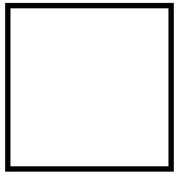
Something that stores information

UNIVERSITY OF CAMBRIDGE

Something that stores information

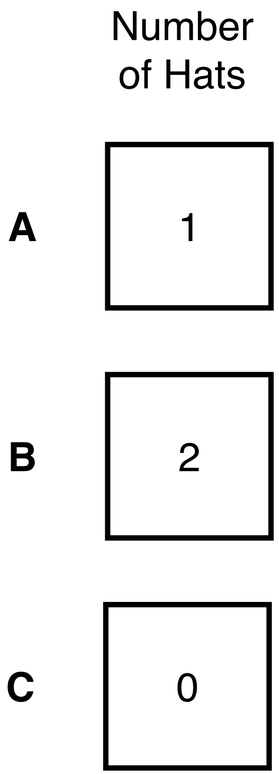
A 

B 

C 

UNIVERSITY OF CAMBRIDGE

Something that stores information



1D vectors

Something that stores information

	Number of Hats	Number of Cats
A	1	2
B	2	5
C	0	3

2D vectors

Something that stores information

	Number of Hats	Number of Cats	Favorite Number
A	1	2	e
B	2	5	π
C	0	3	$-\sqrt{2}$

3D vectors

UNIVERSITY OF CAMBRIDGE

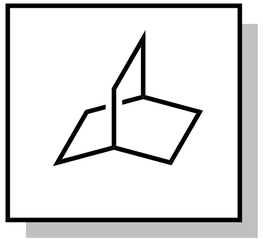
Something that stores information

	Number of Hats	Number of Cats	Favorite Number	Birth-date	Birth Month	Birth Year	Owns A Bike?	Criminal Record?	Lives in England	
A	1	2	e	15	1	1967	0	1	0	...
B	2	5	π	30	9	2001	0	0	0	...
C	0	3	$-\sqrt{2}$	22	5	1991	1	0	1	...

0 = No 1 = Yes

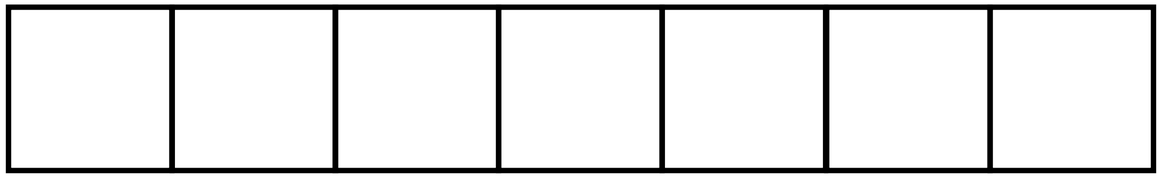
Nth dimensional vector

Representing Molecules As Numbers

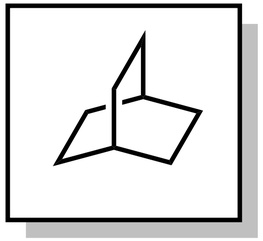


Presence / Absence “important” features

Molecular Fingerprints



NOTE: This is an oversimplified version of fingerprints!



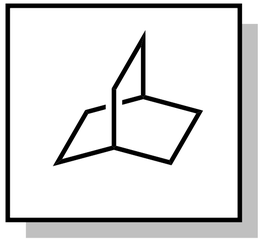
Presence / Absence “important” features

Molecular Fingerprints

OH	CH_2	CH_3	bridge-head C	CHO	Ph	CH_2-CH_2
0	1	0	1	0	0	1

NOTE: This is an oversimplified version of fingerprints!

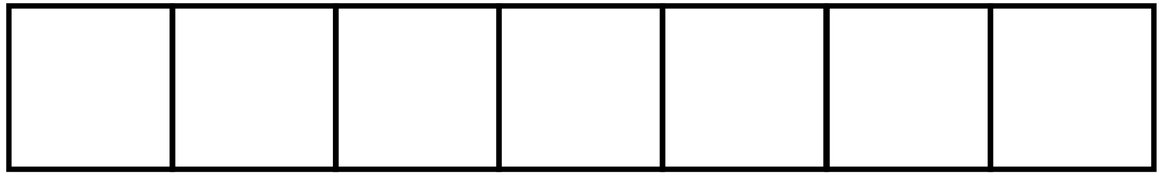
Representing Molecules As Numbers



Chemical Properties

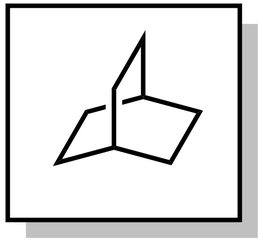
Typically User Defined

Often incorporates quantum chemical properties



NOTE: This is an oversimplified version of featurization!

Representing Molecules As Numbers



Chemical Properties

Typically User Defined

Often incorporates quantum chemical properties

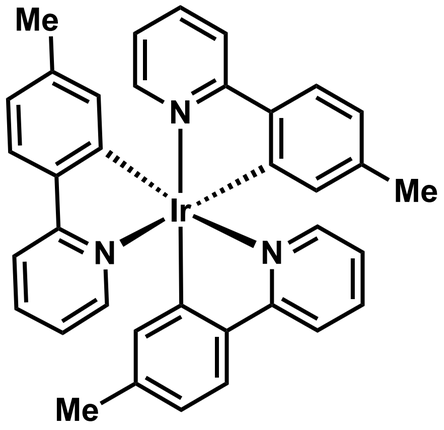
logP	HOMO	LUMO	gap	Num Atoms	Num Hs	Dihedral Angle 1
2.2	-10	-5	5	22	14	120

NOTE: This is an oversimplified version of featurization!

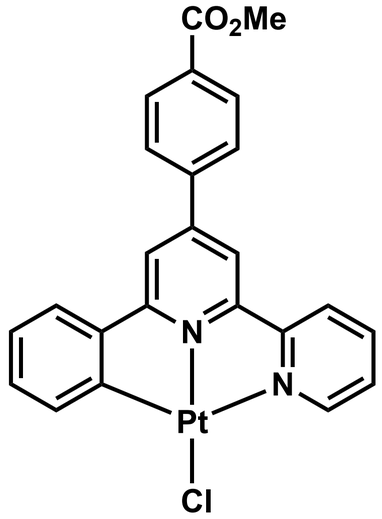
Examples of Helpful Machine Learning

UNIVERSITY OF CAMBRIDGE

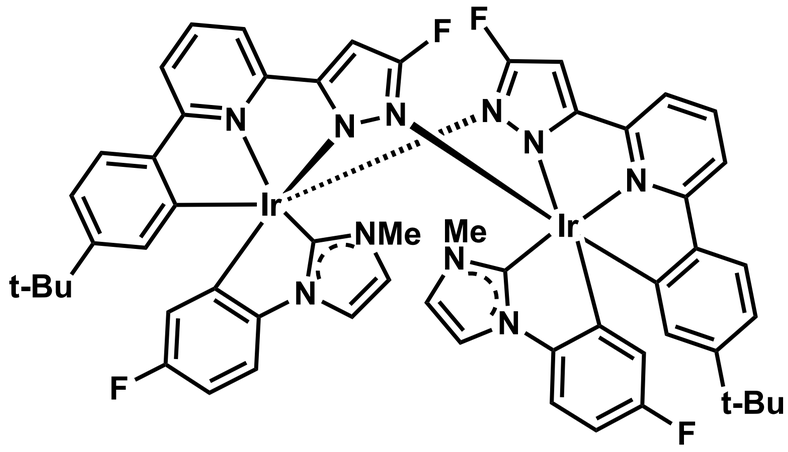
Recent OLED Molecules:



$Ir(mppy)_3$



PtL_2Cl



$[Ir\text{-}2]_2$

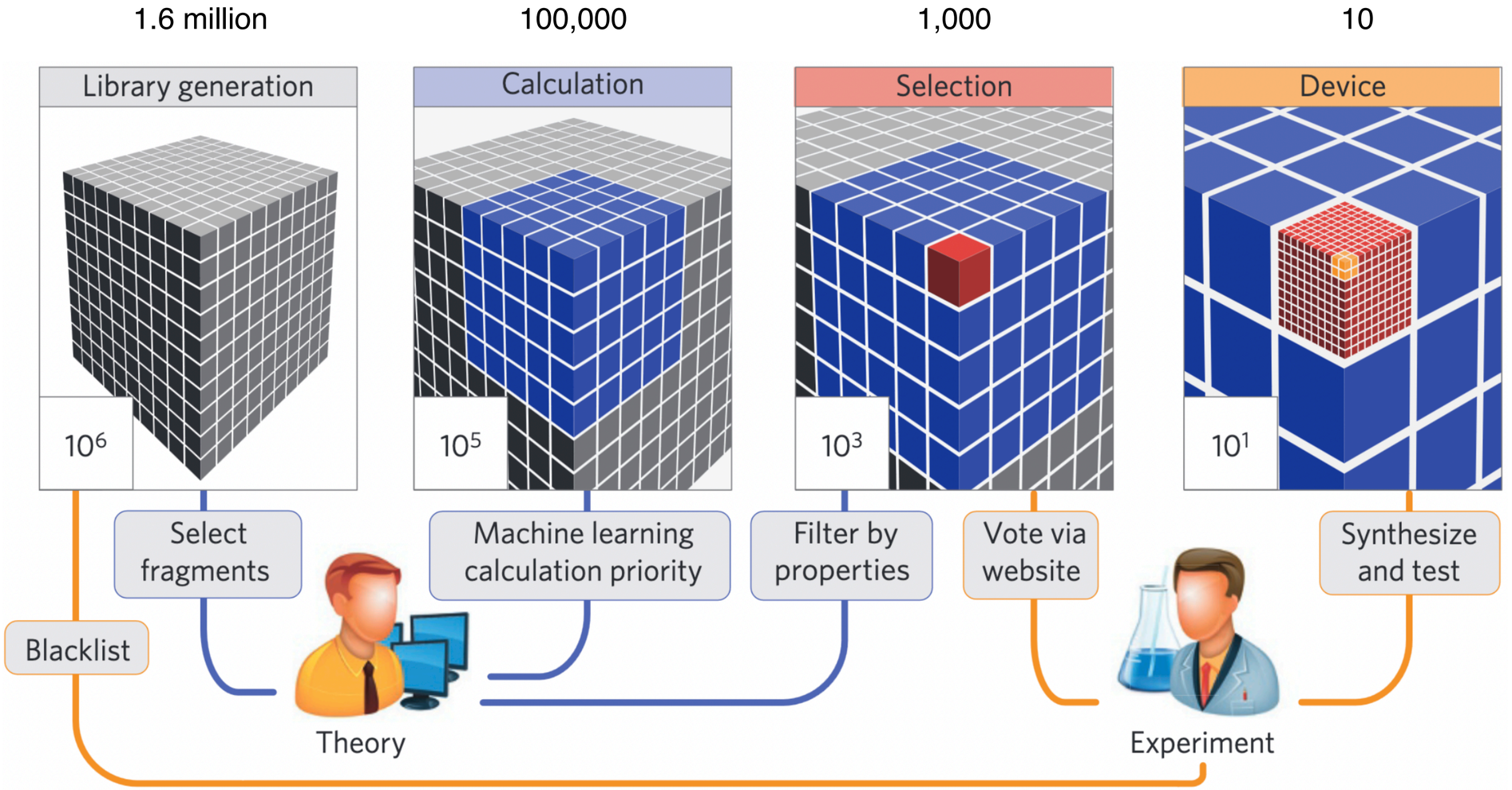
Expensive heavy metals

Can we identify new non-heavy metal containing OLED molecules?

Adv. Mater. **2021**, *33*, 2005630.

Examples of Helpful Machine Learning

ML-assisted library prioritization of new OLED materials



1.6 million compounds is too many to experimentally validate **AND** to perform quantum simulations upon.

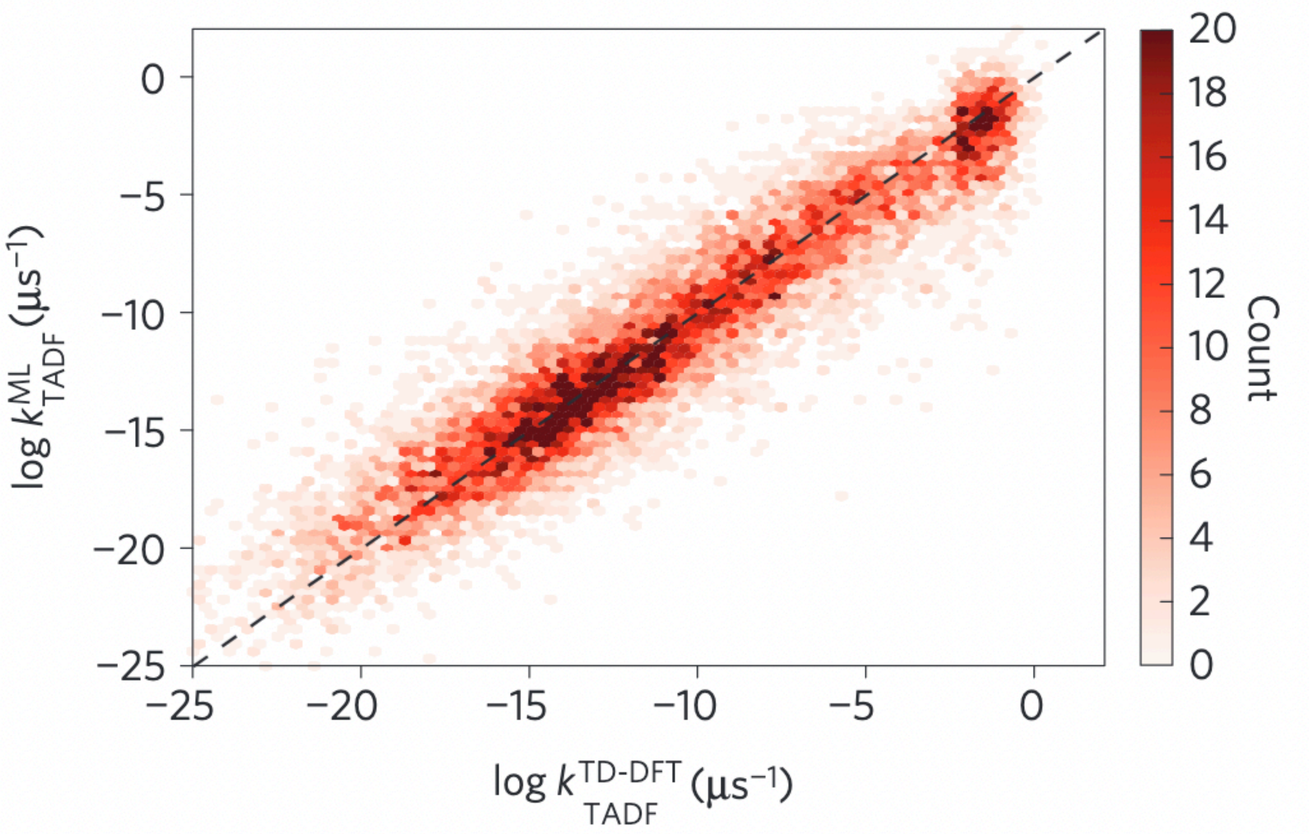
Gómez-Bombarelli *et al. Nat. Mater.* **2016**, *15*, 1120.

Examples of Helpful Machine Learning

UNIVERSITY OF CAMBRIDGE

Randomly chose 40,000 library compounds for DFT computation of the rate of thermally activated delayed fluorescence (k_{TDAF}).

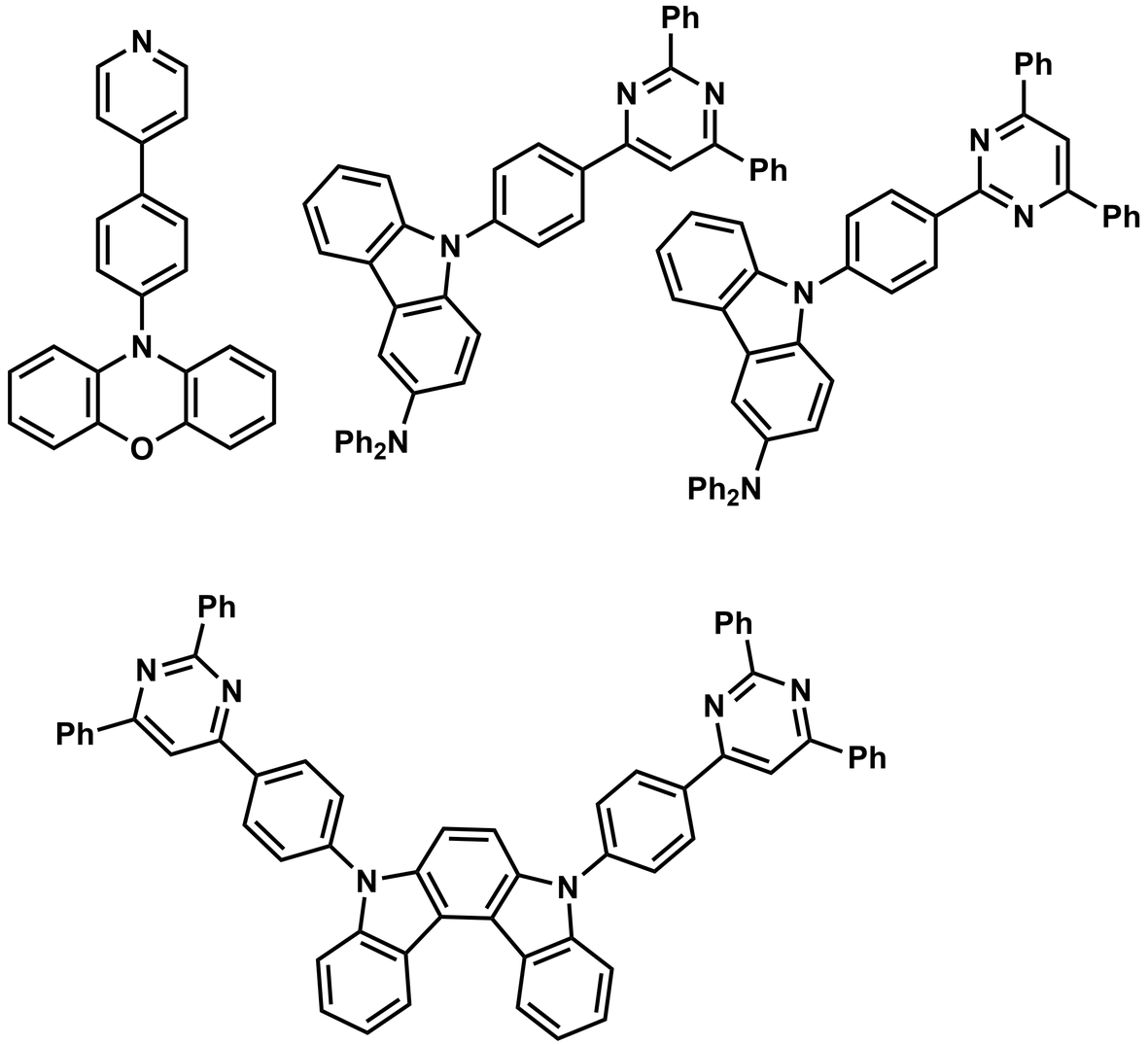
Developed a ML model that could accurate predict k_{TDAF} at fraction of time but with same level of accuracy.



Gómez-Bombarelli *et al. Nat. Mater.* **2016**, *15*, 1120.

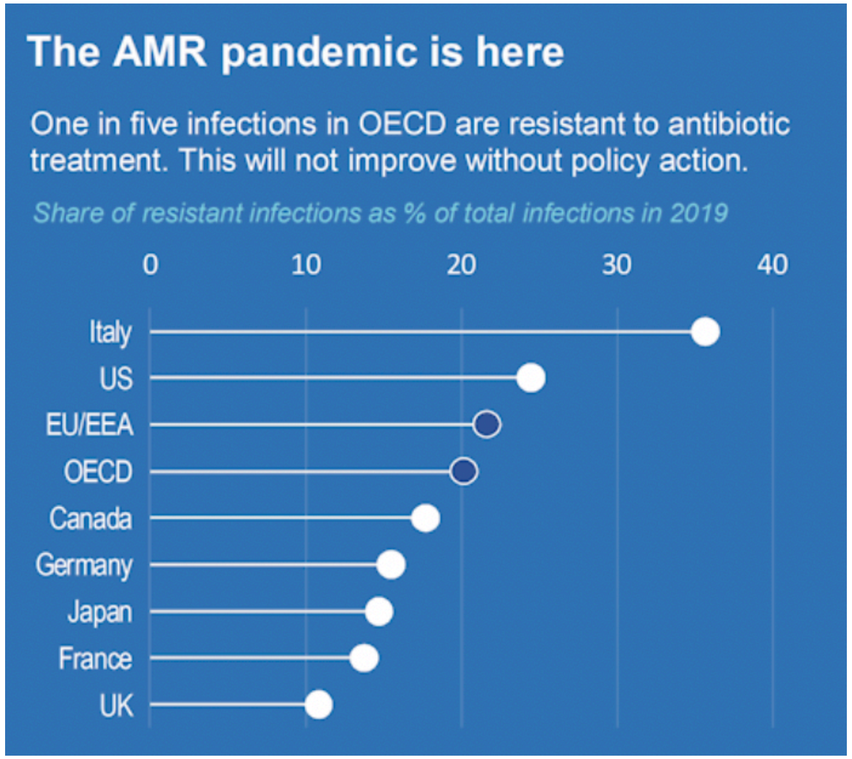
Examples of Helpful Machine Learning

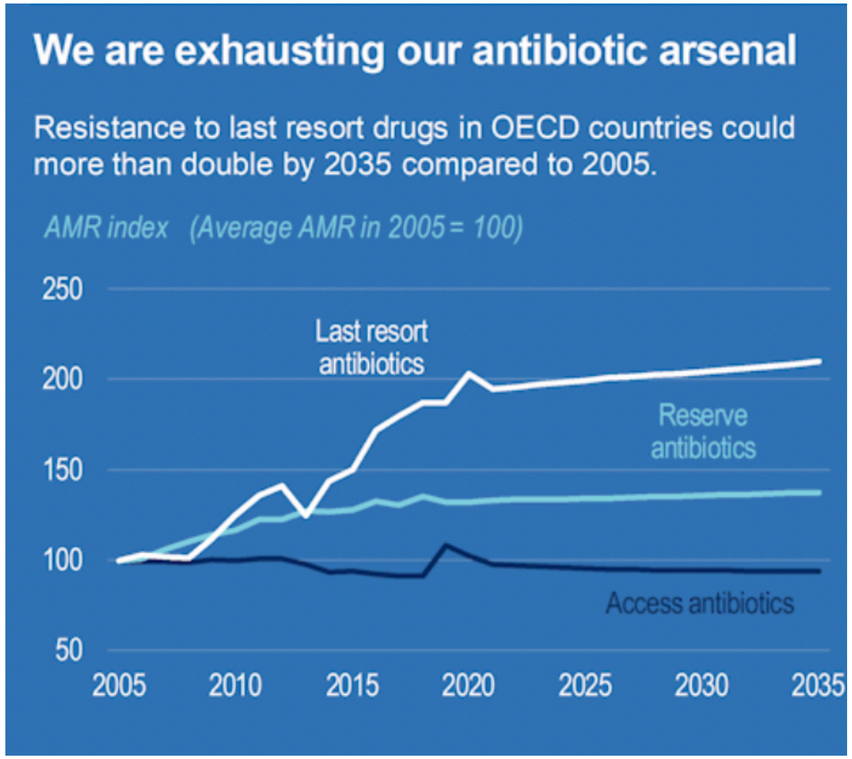
Experimentally validated promising compounds.

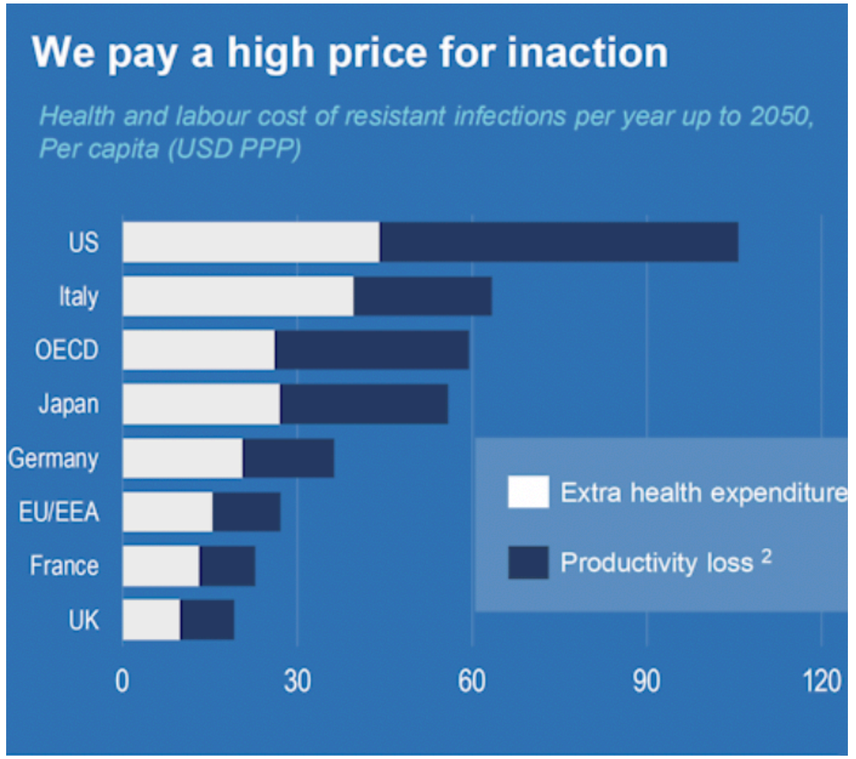


Gómez-Bombarelli *et al. Nat. Mater.* **2016**, *15*, 1120.

Examples of Helpful Machine Learning





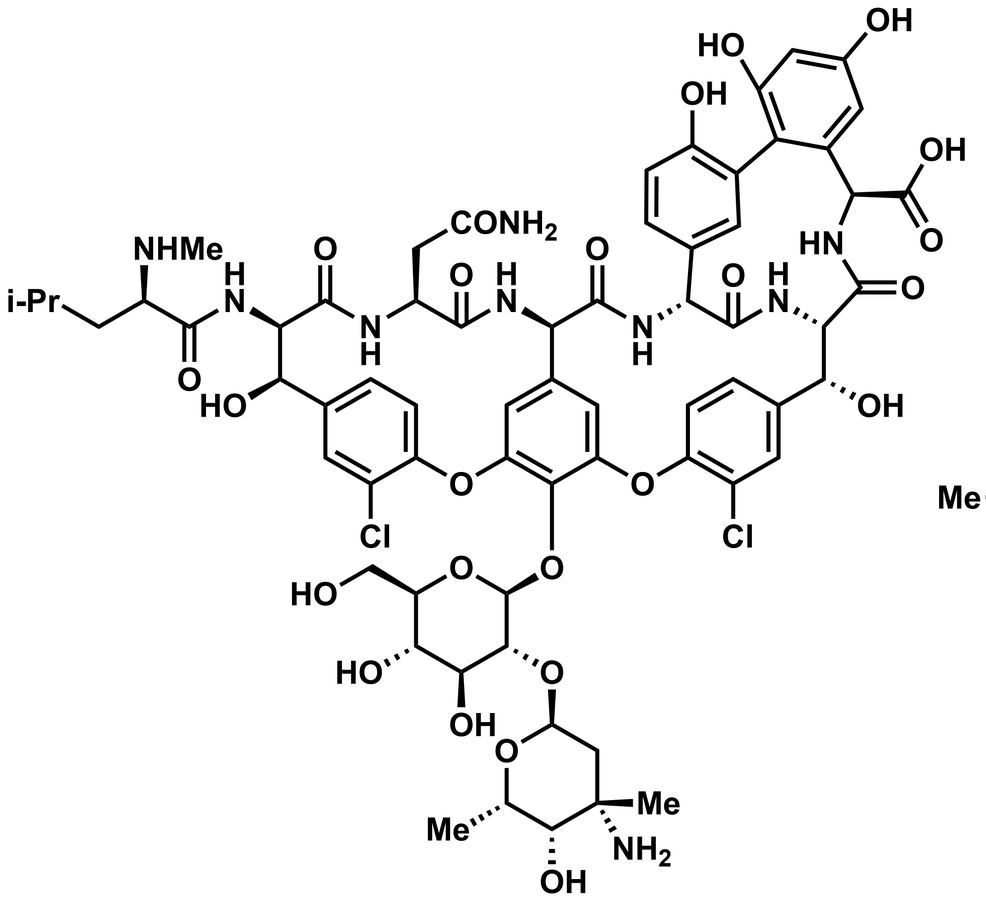


Need to discovery new antibiotics.

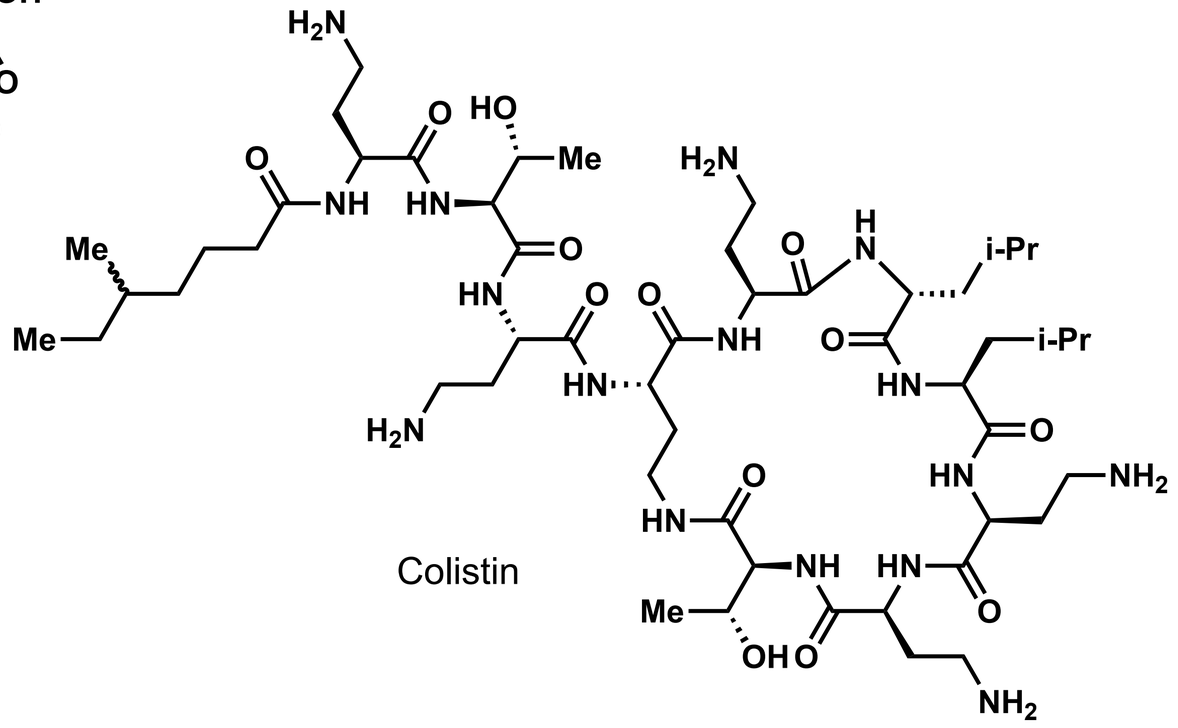
OECD **2023**, OECD Health Policy Studies, OECD Publishing, Paris.

UNIVERSITY OF CAMBRIDGE

Antimicrobial Peptides:



Vancomycin



Colistin

Can we identify new antimicrobial peptides?

Examples of Helpful Machine Learning

UNIVERSITY OF CAMBRIDGE

Test Set (Size)	Avg Number of Correct Predictions*	Avg ML Accuracy*
Unseen sequences from train data (~97)[a]	N/A	90.2%
Sequences annoted as antibacterial peptides in Swiss-Prot (39)[b]	26.0	66.7%
Swiss-Prot sequences that excluded peptides containing signal + antibacterial sequences (24)[b]	22.0	91.7%

* = Average over 3 different ML techniques

[a] = Accuracy = correct predictions / total predictions

[b] = Accuracy = correct predictions / (correct prediction + false positives)

Best average accuracy shown.

Lata *et al. BMC Bioinformatics* **2007**, *8*, 1.

UNIVERSITY OF CAMBRIDGE

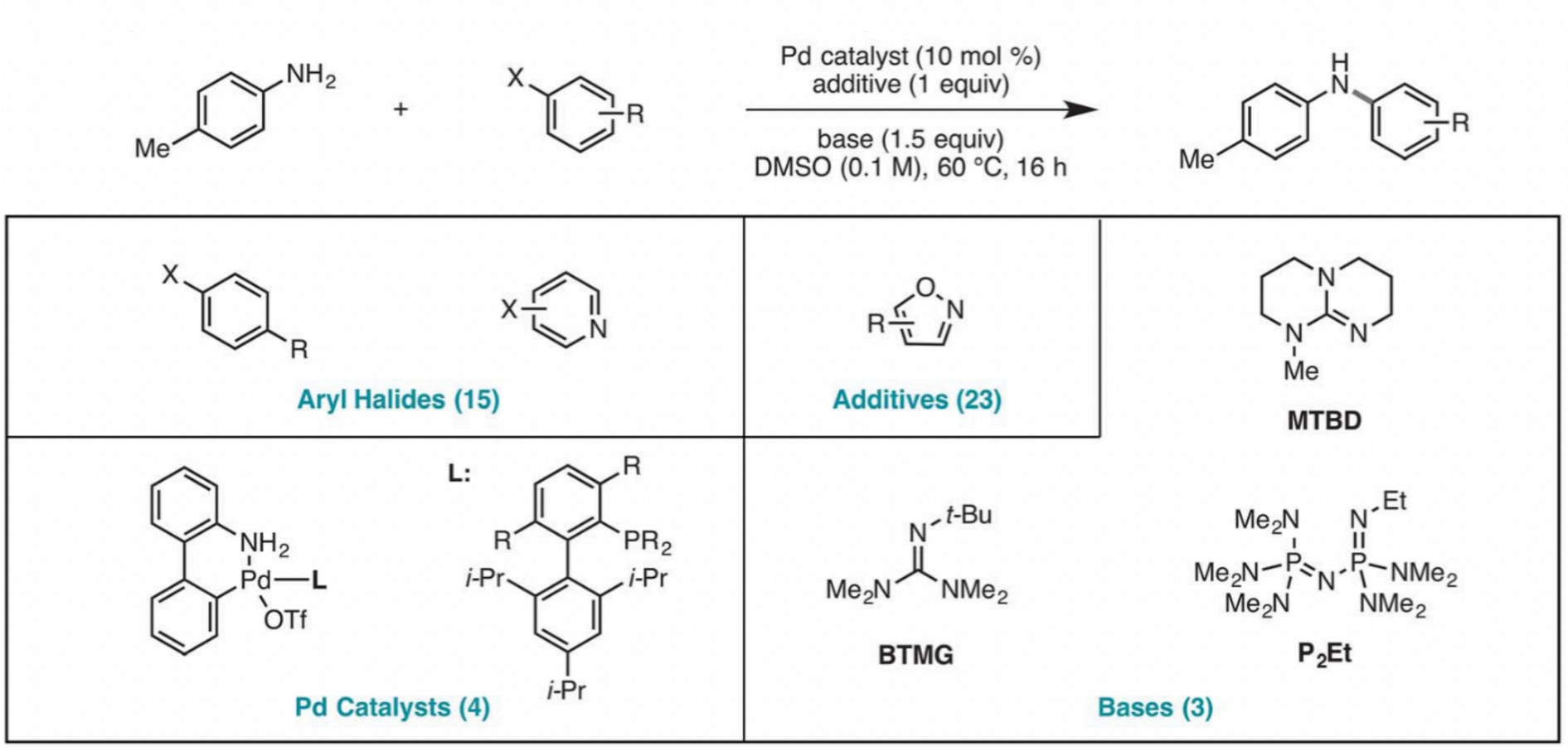
REPORT

ORGANIC CHEMISTRY

Predicting reaction performance in C–N cross-coupling using machine learning

Derek T. Ahneman,[1] Jesús G. Estrada,[1] Shishi Lin,[2]
Spencer D. Dreher,[2]* Abigail G. Doyle[1]*

Let’s Meet The Dataset



https://github.com/doylelab/rxnpredict/blob/master/data_table.csv