Introduction to artificial intelligence and its application to materials science

Luca M. Ghiringhelli

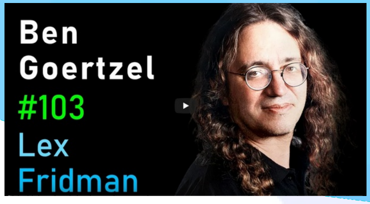
4th IKZ-FAIRmat winter school
Machine learning in materials science and crystal growth
January 23-25, 2023 – Berlin, Germany

Few bits of taxonomy

Artificial intelligence

Narrow AI: The theory and development of computer systems that perform tasks normally associated to human intelligence such as perceiving, *classifying*, *learning*, *abstracting*, *reasoning*, and/or acting

General AI: Full autonomy



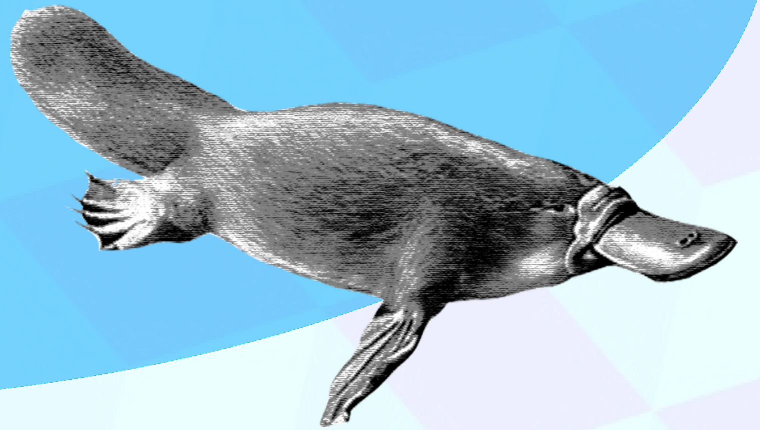
https://youtu.be/OpSmCKe27WE

Few bits of taxonomy (~~beware of~~ embrace the platypus)

Artificial intelligence

Narrow AI: The theory and development of computer systems that perform tasks normally associated to human intelligence such as perceiving, *classifying*, *learning*, *abstracting*, *reasoning*, and/or acting

General AI: Full autonomy



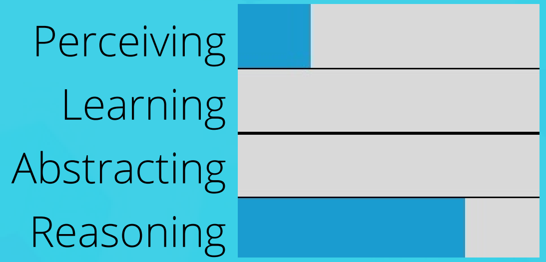
Few bits of taxonomy

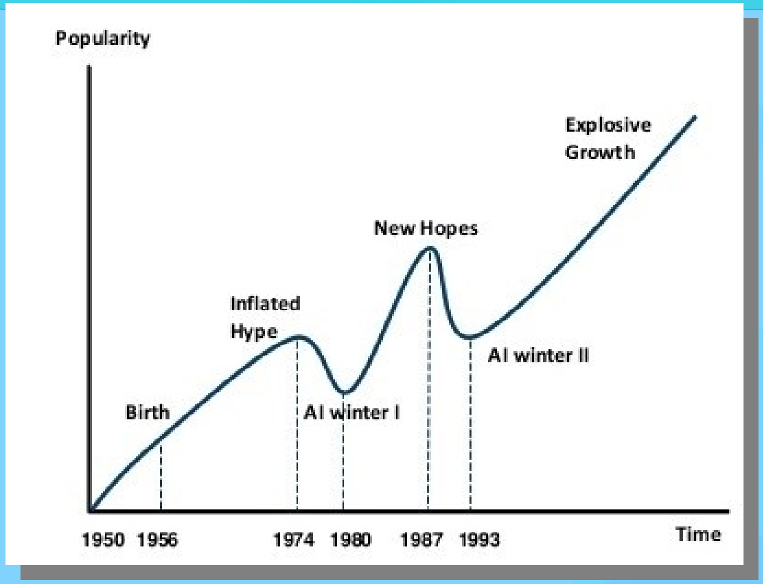
Artificial intelligence, the three waves

(The past)

Handcrafted reasoning

Expert systems, goal trees, if-then rules





John Launchbury, Director I2O, DARPA https://www.youtube.com/watch?v=-O01G3tSYpU

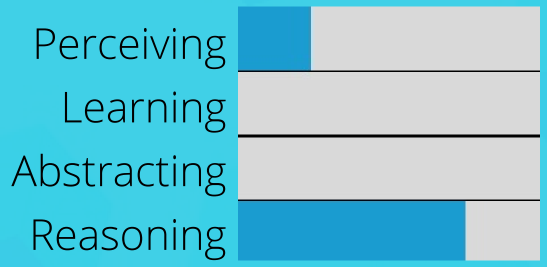
Few bits of taxonomy

Artificial intelligence, the three waves

(The past)

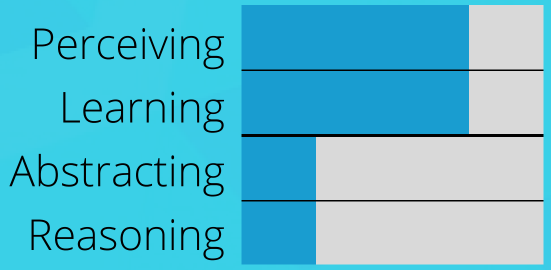
Handcrafted reasoning

Expert systems, goal trees, if-then rules



(The present)

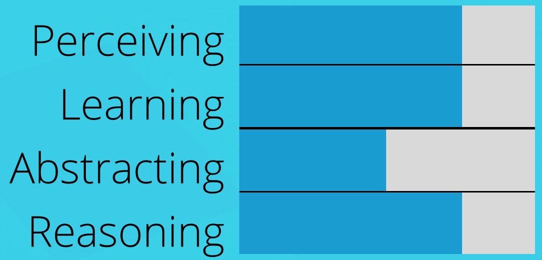
Statistical learning



(The future)

Contextual adaptation

E.g., Neural-symbolic learning and reasoning



John Launchbury, Director I2O, DARPA https://www.youtube.com/watch?v=-O01G3tSYpU

Few bits of taxonomy

Artificial intelligence

Exploratory analysis (Descriptive induction)

gives insight and may lead to hypotheses

"Finding the question is often more important than finding the answer" (John Tukey)

Confirmatory analysis (Predictive induction)

tools that one can use to test ideas

Few bits of taxonomy

Artificial intelligence

Exploratory analysis (Descriptive induction)

gives insight and may lead to hypotheses

- clustering
- dimension reduction
- *subgroup discovery*

Confirmatory analysis (Predictive induction)

tools that one can use to test ideas

- regression
- classification

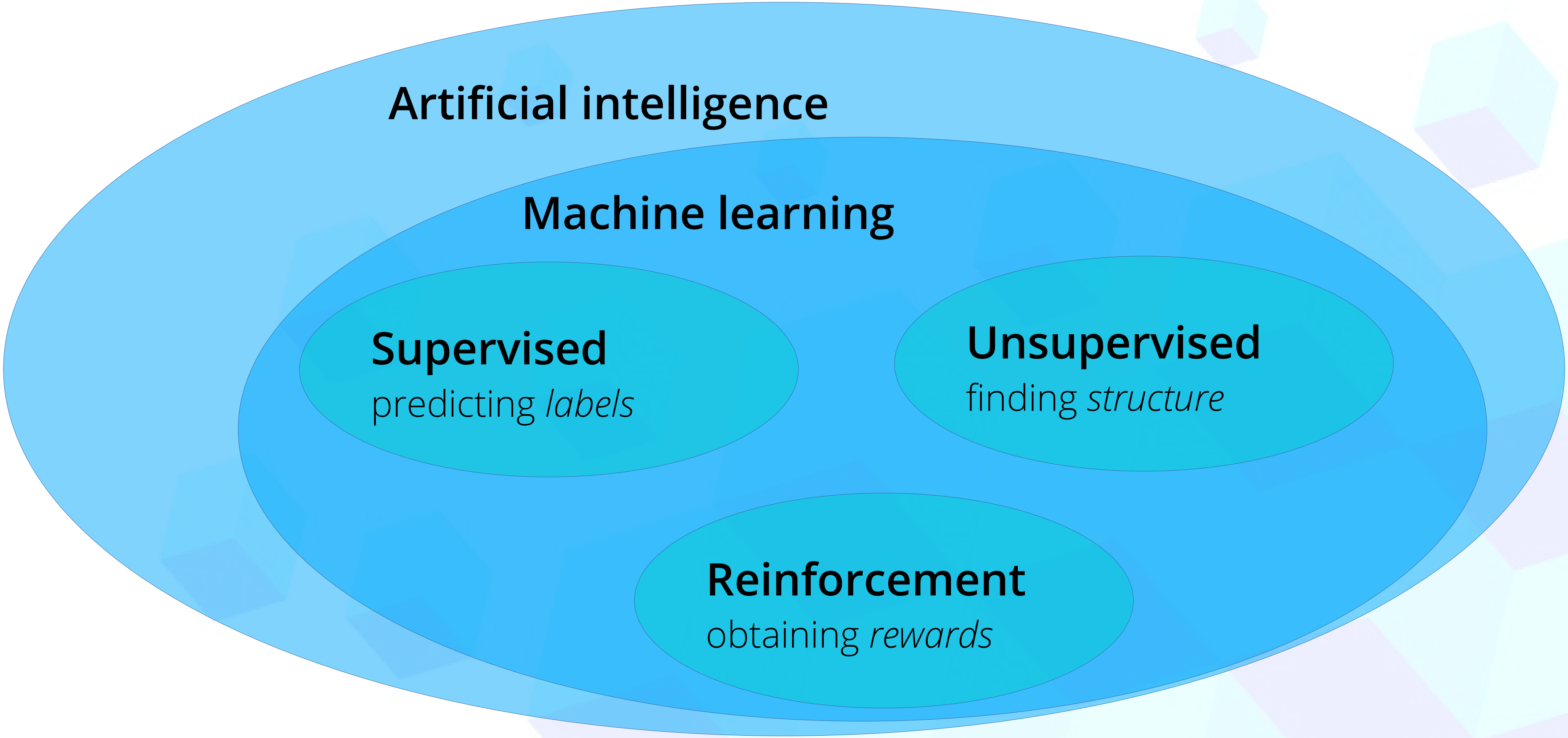
Few bits of taxonomy

Artificial intelligence

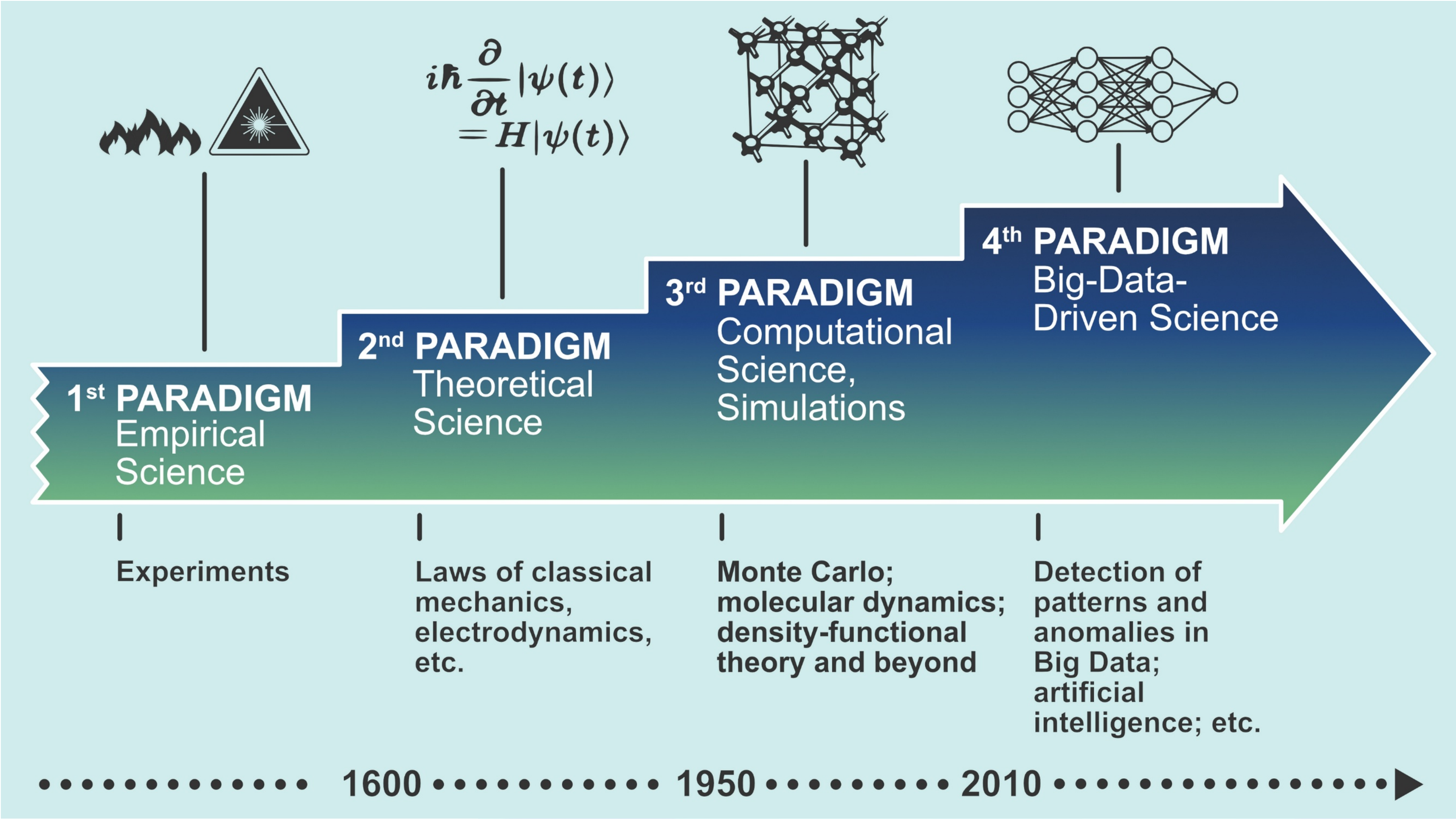
Machine learning

Statistical learning algorithms.
Learning = improving with more data.
Regularized regression

Few bits of taxonomy



About terminology: (big-)data-driven science

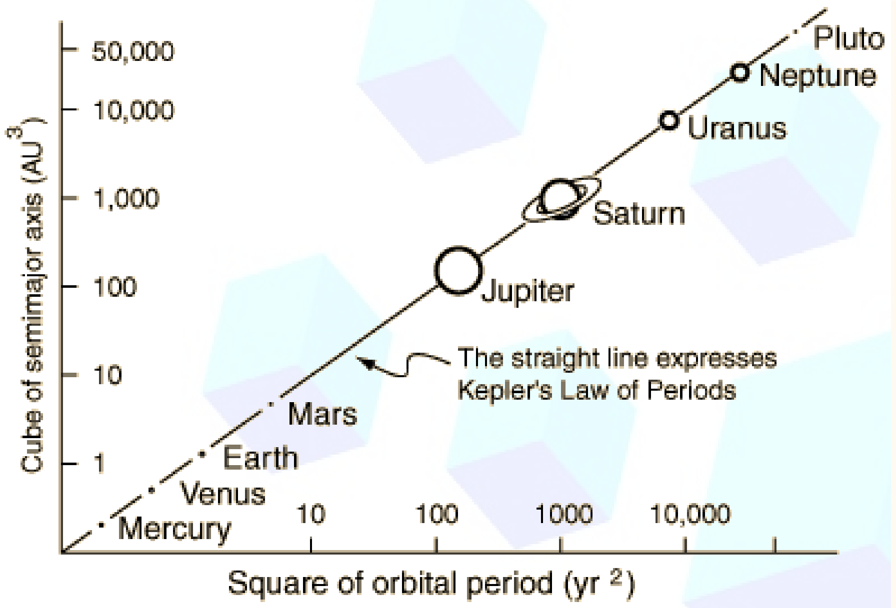


Jim Gray: The 4th Paradigm, Data Intensive Discovery, edited by Hey, Tansley, and Tolle, Microsoft Research (2007)

Science is always data driven

Suppose to know the trajectories of all planets in the solar system,
- from accurate observations (experiment), or
- by numerically integrating general-relativity equations

(i.e., the highest level of theory)



(Orbital period)^2 = C (orbit's major axis)^3

Data
(collected by
Tycho Brahe)

Statistical learning
(performed by
Johannes Kepler)

Physical law
(assessed by
Isaac Newton)

Science is always data driven

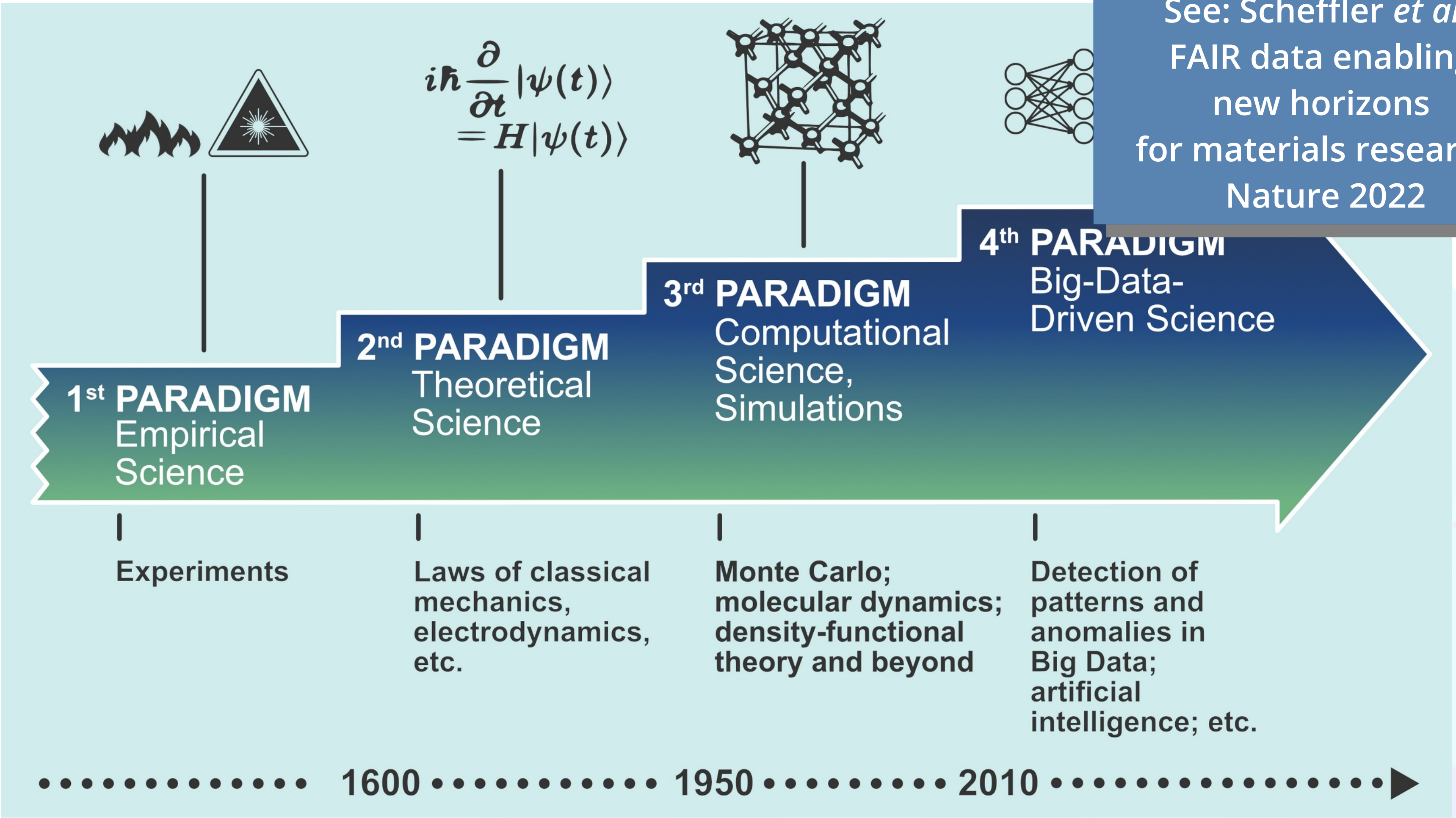
Mendeleev's **1871** periodic table

Reihen	Gruppe I. — R^2O	Gruppe II. — RO	Gruppe III. — R^2O^3	Gruppe IV. RH^4 RO^2	Gruppe V. RH^3 R^2O^5	Gruppe VI. RH^2 RO^3	Gruppe VII. RH R^2O^7	Gruppe VIII. — RO^4
1	H=1							
2	Li=7	Be=9.4	B=11	C=12	N=14	O=16	F=19	
3	Na=23	Mg=24	Al=27.3	Si=28	P=31	S=32	Cl=35.5	
4	K=39	Ca=40	—=44	Ti=48	V=51	Cr=52	Mn=55	Fe=56, Co=59, Ni=59, Cu=63.
5	(Cu=63)	Zn=65	$\boxed{—=68}$	$\boxed{—=72}$	As=75	Se=78	Br=80	
6	Rb=85	Sr=87	?Yt=88	Zr=90	Nb=94	Mo=96	—=100	Ru=104, Rh=104, Pd=106, Ag=108.
7	(Ag=108)	Cd=11			b=122	Te=125	J=127	
8	Cs=133	Ba=137	?Di=138	?Ce=140	—	—	—	— — — —
9	(—)	—	—	—	—	—	—	
10	—	—	?Er=178	?La=180	Ta=182	W=184	—	Os=195, Ir=197, Pt=198, Au=199.
11	(Au=199)	Hg=200	Tl=204	Pb=207	Bi=208	—	—	
12	—	—	—	Th=231	—	U=240	—	— — — —

Ga=69.7 Ge=72.6

About terminology: (big-)data-centric science

See: Scheffler *et al.*, FAIR data enabling new horizons for materials research, Nature 2022



Jim Gray: The 4th Paradigm, Data Intensive Discovery, edited by Hey, Tansley, and Tolle, Microsoft Research (2007)

Data vs algorithm driven breakthroughs

Year	Breakthroughs in AI	Datasets (First Available)	Algorithms (First Proposed)
1994	Human-level read-speech recognition	Spoken Wall Street Journal articles and other texts (1991)	Hidden Markov Model (1984)
1997	IBM Deep Blue defeated Garry Kasparov	700,000 Grandmaster chess games, aka "The Extended Book" (1991)	Negascout planning algorithm (1983)
2005	Google's Arabic- and Chinese-to-English translation	1.8 trillion tokens from Google Web and News pages (collected in 2005)	Statistical machine translation algorithm (1988)
2011	IBM Watson became the world Jeopardy! champion	8.6 million documents from Wikipedia, Wiktionary, Wikiquote, and Project Gutenberg (updated in 2010)	Mixture-of-Experts algorithm (1991)
2014	Google's GoogleNet object classification at near-human performance	ImageNet corpus of 1.5 million labeled images and 1,000 object categories (2010)	Convolutional neural network algorithm (1989)
2015	Google's Deepmind achieved human parity in playing 29 Atari games by learning general control from video	Arcade Learning Environment dataset of over 50 Atari games (2013)	Q-learning algorithm (1992)
Average No. of Years to Breakthrough		**3 years**	**18 years**

Source: V. Gadepally Artificial Intelligence and Machine Learning, https://youtu.be/t4K6lney7Zw

Statistical learning in practice

Logical flow-chart

(Annotated) Data !

Features / descriptors / representations

Training algorithm: parameters vs hyperparameters.

Training metrics

Model selection

Cross-validation metrics

Test

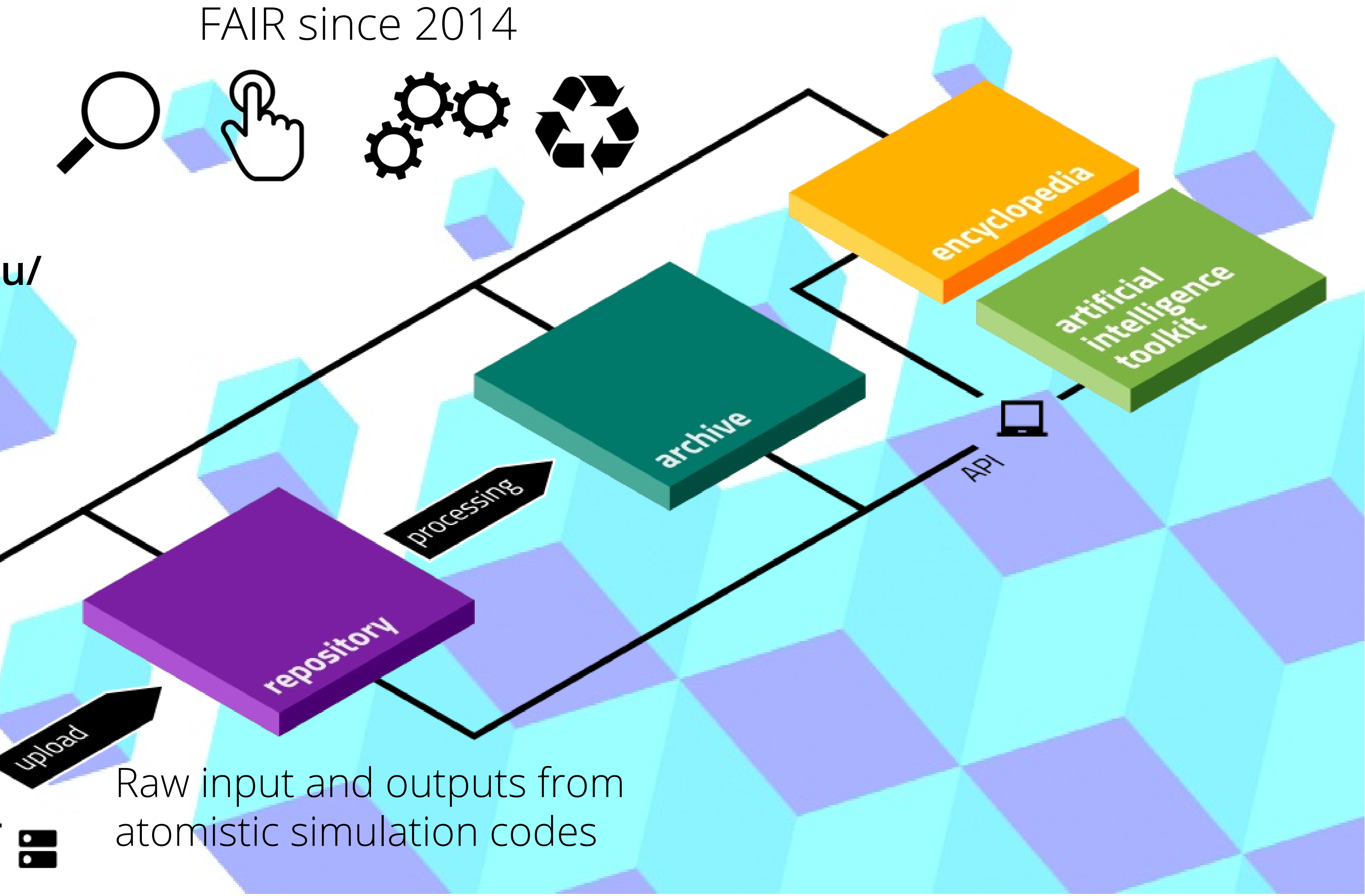
See the virtual AI full course
https://www.fair-di.eu/fairmat/outreach/materials
for video lectures and hands-on notebooks

The NOMAD Laboratory - https://nomad-lab.eu/

NOMAD
NOVEL MATERIALS DISCOVERY

https://nomad-lab.eu/

FAIR since 2014



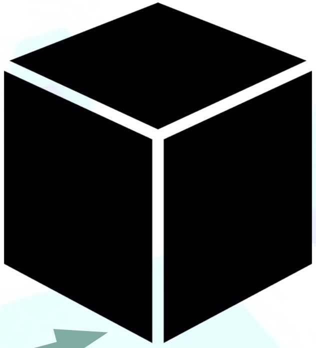
Raw input and outputs from atomistic simulation codes

NOVEL MATERIALS DISCOVERY

Metadata for computational materials science

- Uploader
- Date
- Location

Atomistic-simulation code



- Code name
- Version
- Libraries

Input structure
- Coordinates

  $Z_i \;\; x_i \, y_i \, z_i$
- Cell vectors
- (Topology)

Output
- Total energy
- Forces
- Electron density
- Electrostatic pot.
- El. band structure
- Self energy
- ...

Input model
- xc treatment / force field
- Relativity treatment
- Basis set
- Numerical integr. settings

NOMAD Metainfo

NOVEL MATERIALS DISCOVERY

- Code name
- Version
- Libraries

<u>Input structure</u>
- Coordinates
- Cell vectors
- (Topology)

<u>Input model</u>
- xc treatment / force field
- Relativity treatment
- Basis set
- Numerical integr. settings

<u>Output</u>
- Total energy
- Forces
- Electron density
- Electrostatic potential
- Electronic band structure
- Self energy

Type of variables (both features and targets)

Quantitative → Regression

Discrete

Continuous (including scalars, vectors/arrays and matrices/tensors)

Categorical (Qualitative) → Classification

(Binary)

Nominal

<u>According to S. S. Stevens "On the Theory of Scales of Measurement"</u> (Science, 1946, jstor 1671815):

Nominal	labels categories (order has no meaning)	[Classification]
Ordinal	defines a rank (intervals have no meaning)	[Classification]

Type of variables (both features and targets)

Quantitative → Regression

- Discrete
- Continuous (including scalars, vectors/arrays and matrices/tensors)

Categorical (Qualitative) → Classification

- (Binary)
- Nominal

<u>According to S. S. Stevens "On the Theory of Scales of Measurement"</u> (Science, 1946, jstor 1671815):

Nominal	labels categories (order has no meaning)	[Classification]
Ordinal	defines a rank (intervals have no meaning)	[Classification]
Interval	has values of equal intervals that mean something (arbitrary zero)	[Regression]
Ratio	same as interval except that the zero means: does not exist	[Regression]
Cardinal	used for counting ("how many?")	[Class./Regr.]

Statistical learning: features, descriptors, and more

Property = P (model$_1$, model$_2$, ...)

model$_1$(representation$_1$, representation$_2$, ...) model$_2$(...)

representation$_1$(descriptor$_1$, descriptor$_2$, ...) representation$_2$(...)

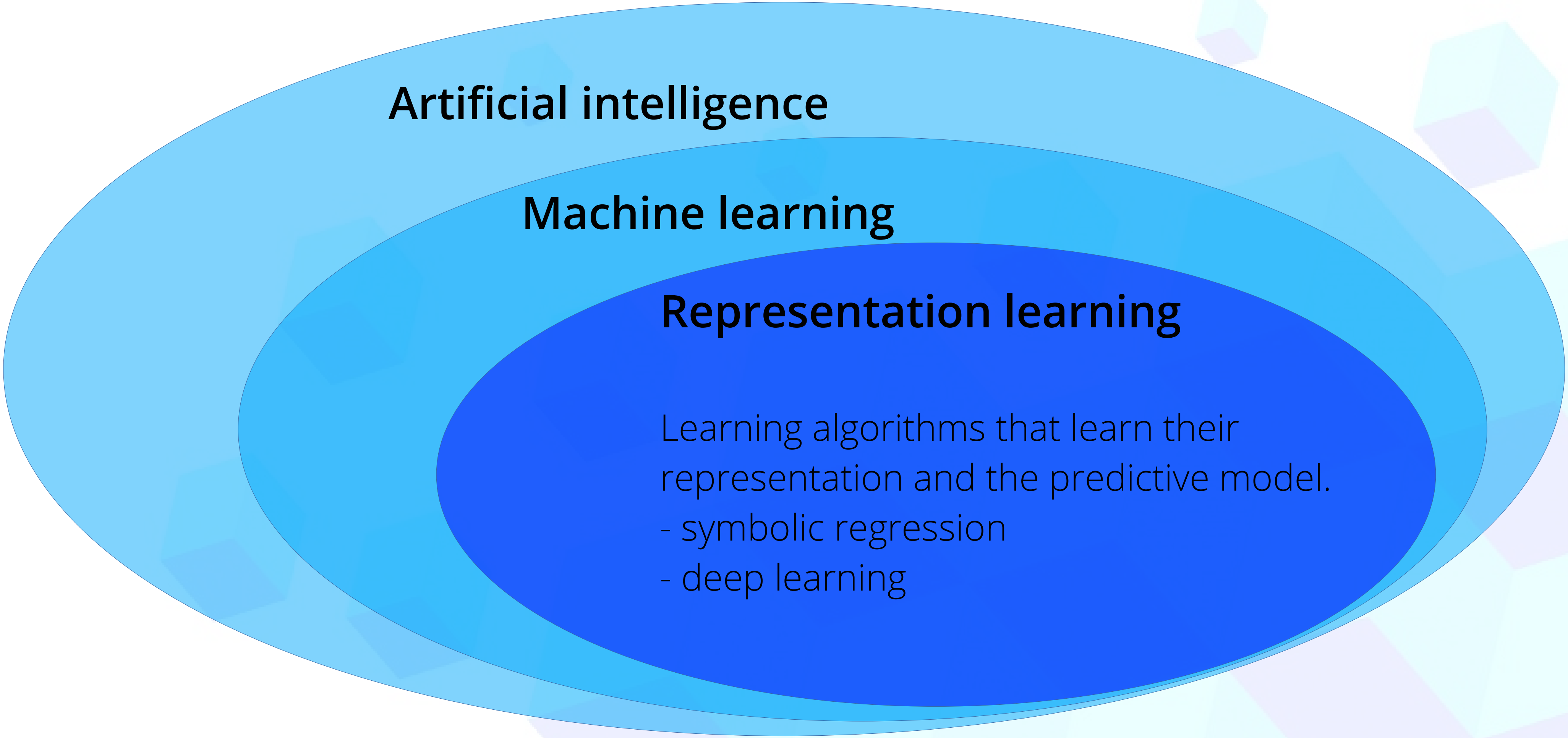
descriptor$_1$ [feature$_1$, feature$_2$, ...] descriptor$_2$[...]

feature$_1$(primary_feature$_1$, primary_feature$_2$) feature$_2$(...)

- Descriptors aka fingerprints
- Primary features in materials science:
  Coordinates and physical-chemical properties of chemical elements
  material / data point $i \rightarrow \{\boldsymbol{R}_i, Z_i\}$

Onat *et al.* JCP 2020 doi: 10.1063/5.0016005;
see also Ghiringhelli *et al.* PRL 2015 doi:10.1103/PhysRevLett.114.105503, Bártok et al., PRB 2013 doi: 10.1103/PhysRevB.87.184115

Few bits of taxonomy



The prima donna: the descriptor

(i) Invariance: descriptors should be invariant under symmetry operations: permutation of atoms and translation and rotation of structure.

(ii) Sensitivity (local stability): small changes in the atomic positions should result in *proportional* changes in the descriptor, and vice versa.

(iii) Differentiability: having continuous functions that are differentiable.

(iv) Global Uniqueness: the mapping of the descriptor should be unique for a given input atomic environment (i.e. the mapping is injective).

(v) Dimensionality: the dimension of the spanned hyper-dimensional space of the descriptor should be sufficient to ensure uniqueness, but not larger.

(vi) Scalability: ideally, descriptors should be easily generalized to any system or structure with a preference to have no limitations on number of elements, atoms, or properties.

(vii) Complexity: to have a low computational cost so the method can be fast enough to scale to the required size of the simulations and to be used in high-throughput screening of big-data.

(viii) Interpretability: features of the encoding can be mapped directly to structural or material properties for *easy* interpretation of results.

Onat *et al.* JCP 2020 doi: 10.1063/5.0016005;
see also Ghiringhelli *et al.* PRL 2015 doi:10.1103/PhysRevLett.114.105503, Bártok et al., PRB 2013 doi: 10.1103/PhysRevB.87.184115

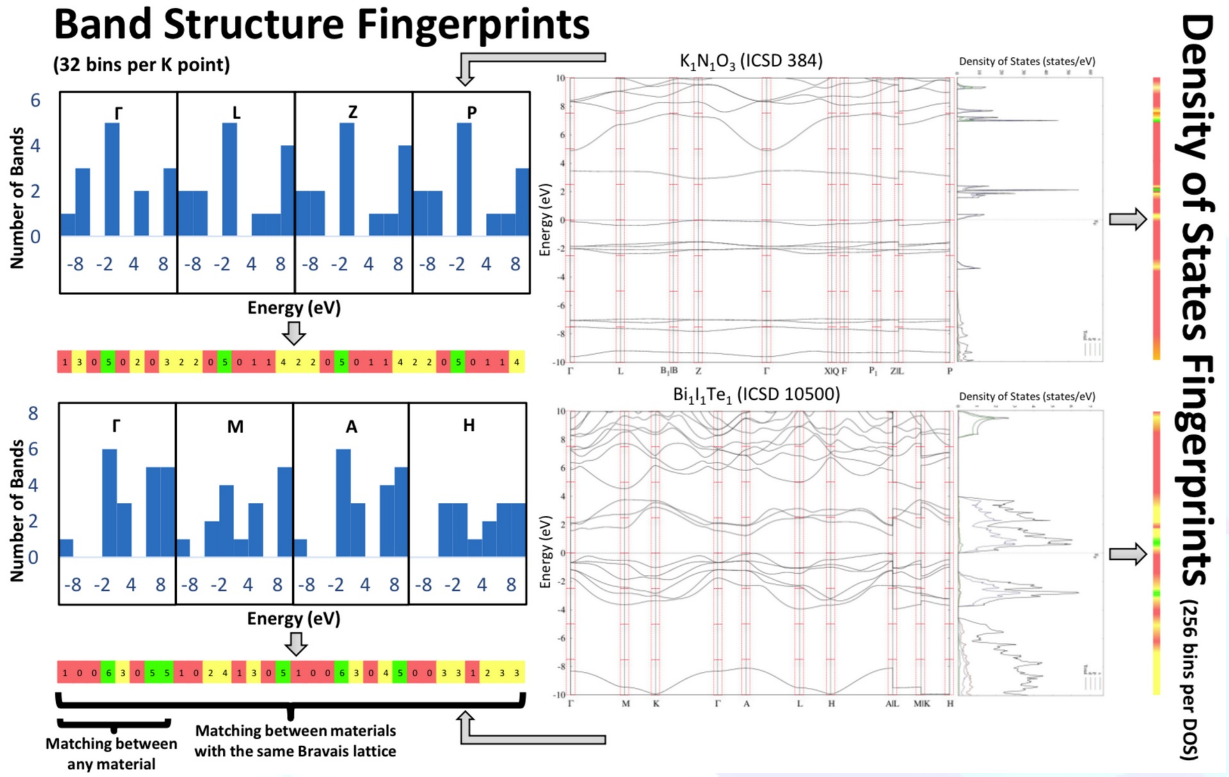
Statistical learning: features, descriptors, and more

Linear 4-blocks periodic polymers.

7 blocks: CH_2, SiF_2, $SiCl_2$, GeF_2, $GeCl_2$, SnF_2, $SnCl_2$

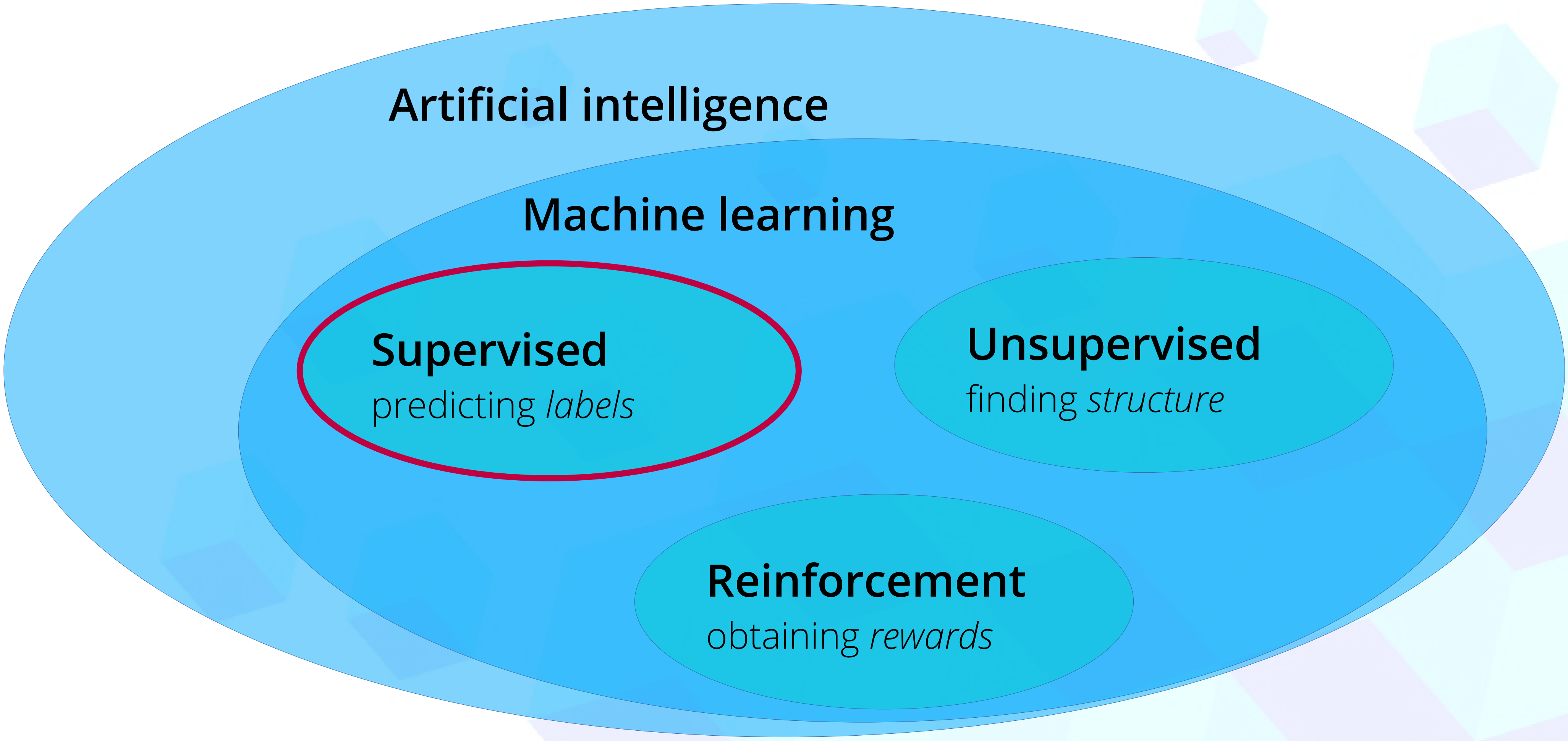
Descriptor: 20 dimensions [# building blocks of type *i*, of *ii* pairs, of *iii* triplets]

Pilania, Wang, ..., and Ramprasad, Scientific Reports 3, 2810 (2013). DOI: 10.1038/srep02810



Isayev, ..., and Curtarolo,
Chemistry of Materials 27, 735 (2015)
DOI: 10.1021/cm503507h

Few bits of taxonomy



Supervised statistical learning: prediction vs inference

Estimating f such that the target Y is expressed as:

$$Y = f(X) + \epsilon$$

Error not depending on X

$$X = (X_1, X_2, \ldots, X_p)$$

E.g., (multi)linear model:

$$Y = \sum_i^p c_i X_i + \epsilon$$

Prediction $\hat{Y} = \hat{f}(X)$

$$E(Y - \hat{Y})^2 = E[f(X) + \epsilon - \hat{f}(X)]^2 = \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}}$$

James, Witten, Hastie, Tibshirani, An Introduction to Statistical Learning, Springer (2013)

The prima donna: the descriptor

(i) **Invariance**: descriptors should be invariant under symmetry operations: permutation of atoms and translation and rotation of structure.
(ii) **Sensitivity** (local stability): small changes in the atomic positions should result in *proportional* changes in the descriptor, and vice versa.
(iii) **Differentiability**: having continuous functions that are differentiable.
(iv) **Global Uniqueness**: the mapping of the descriptor should be unique for a given input atomic environment (i.e. the mapping is injective).
(v) **Dimensionality**: the dimension of the spanned hyper-dimensional space of the descriptor should be sufficient to ensure uniqueness, but not larger.

(vi) **Scalability**: ideally, descriptors should be easily generalized to any system or structure with a preference to have no limitations on number of elements, atoms, or properties.
(vii) **Complexity**: to have a low computational cost so the method can be fast enough to scale to the required size of the simulations and to be used in high-throughput screening of big-data.
(viii) **Discrete Mapping**: always map to the same hyperdimensional space with constant size feature sets, regardless of the input atomic environment

(ix) **Interpretability**: features of the encoding can be mapped directly to structural or material properties for *easy* interpretation of results.

Onat *et al.* JCP 2020 doi: 10.1063/5.0016005; see also Ghiringhelli *et al.* PRL 2015 doi:10.1103/PhysRevLett.114.105503, Bártok et al., PRB 2013 doi: 10.1103/PhysRevB.87.184115

Supervised statistical learning: prediction vs inference

Estimating f such that the target Y is expressed as:

$$Y = f(X) + \epsilon$$

Error not depending on X

$$X = (X_1, X_2, \ldots, X_p)$$

E.g., (multi)linear model:

$$Y = \sum_{i}^{p} c_i X_i + \epsilon$$

Prediction $\hat{Y} = \hat{f}(X)$

$$E(Y - \hat{Y})^2 = E[f(X) + \epsilon - \hat{f}(X)]^2 = \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{Irreducible}}$$

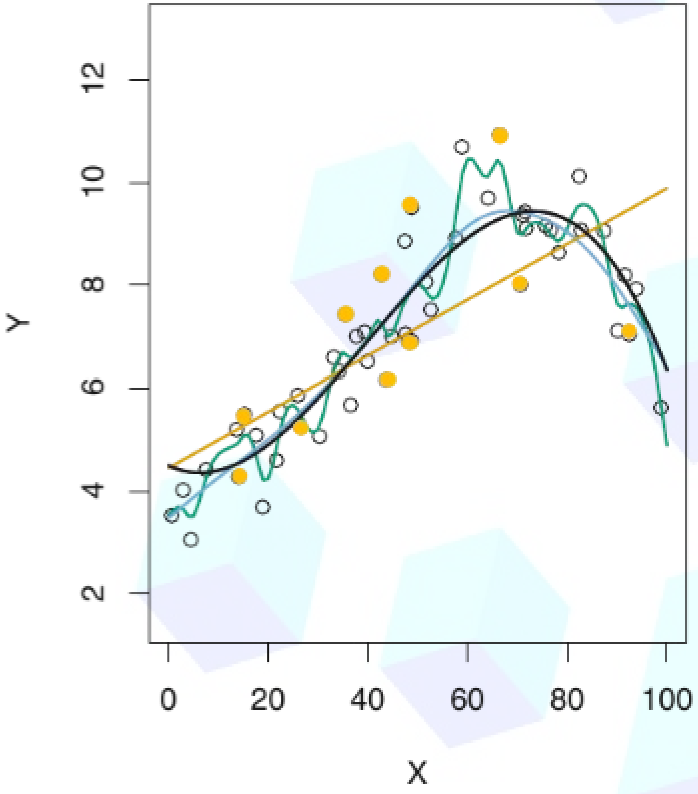
Inference (interpretation)

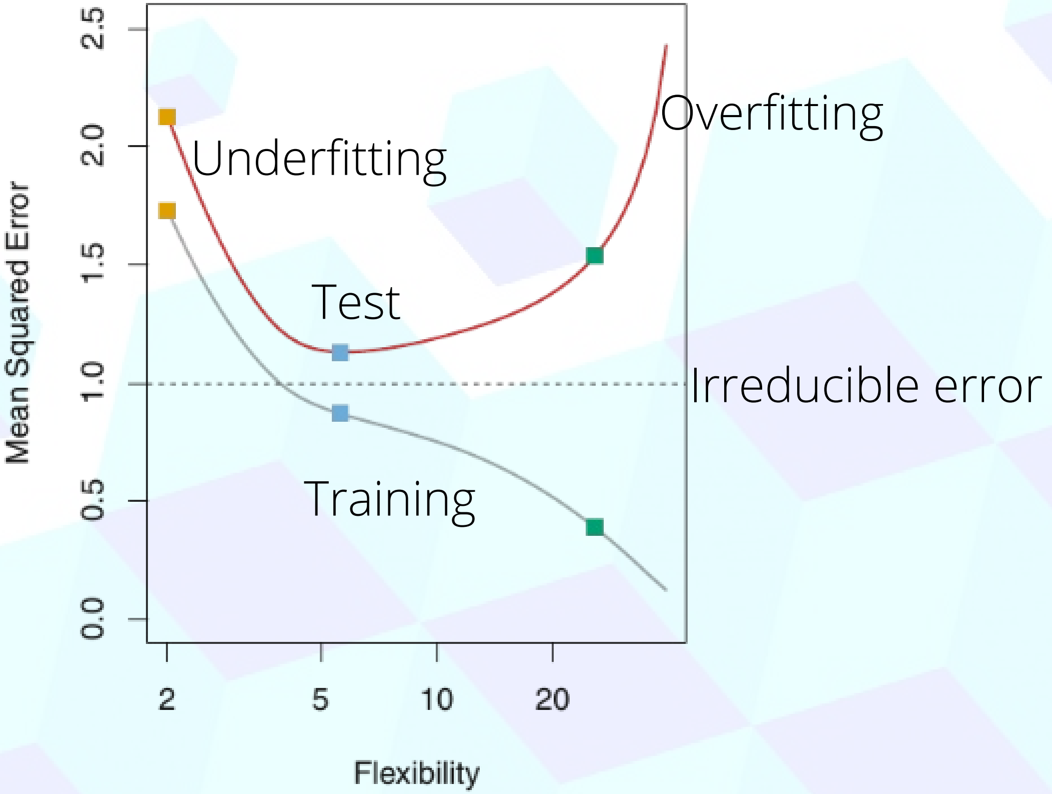
- Which features are associated with the target?
- What is the relationship between the target and each feature?
- Can the relationship between the target Y and each feature be adequately summarized using a linear equation, or is the relationship more complicated?

James, Witten, Hastie, Tibshirani, An Introduction to Statistical Learning, Springer (2013)

Supervised statistical learning: quality of the fit

The quality of the fit is measured on test data (not used for training)





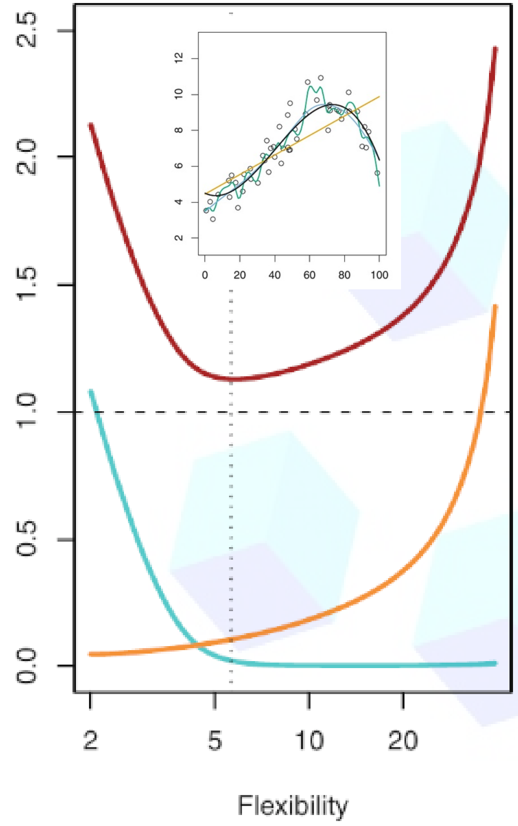
James, Witten, Hastie, Tibshirani, An Introduction to Statistical Learning, Springer (2013)

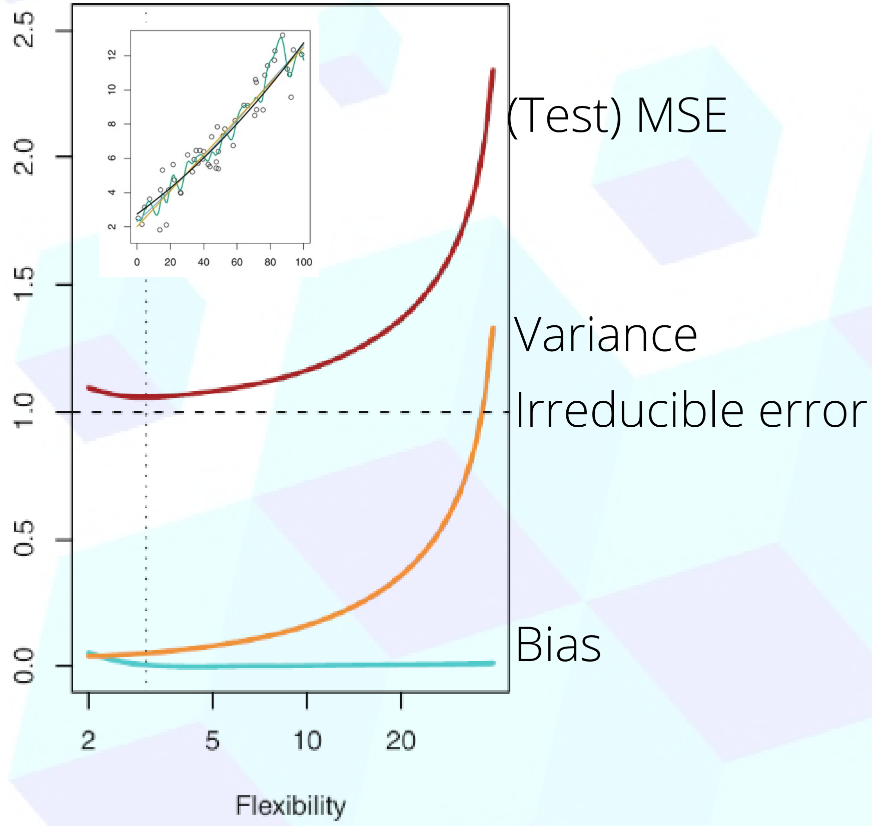
Supervised statistical learning: Bias/variance trade-off

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$

Depends on training sets

Inherent to model

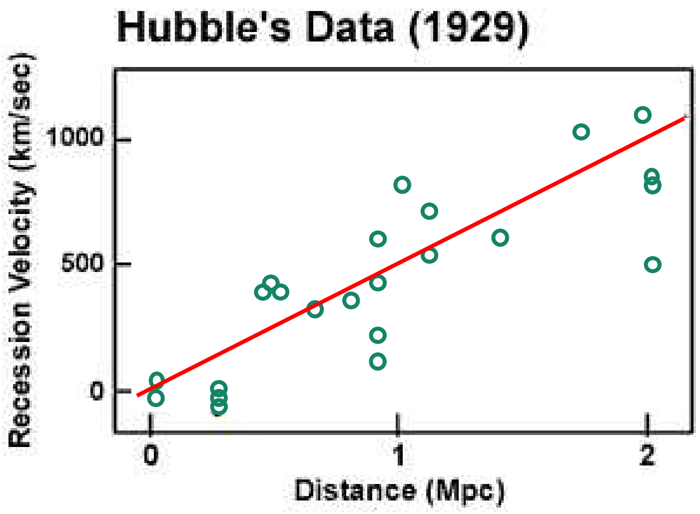




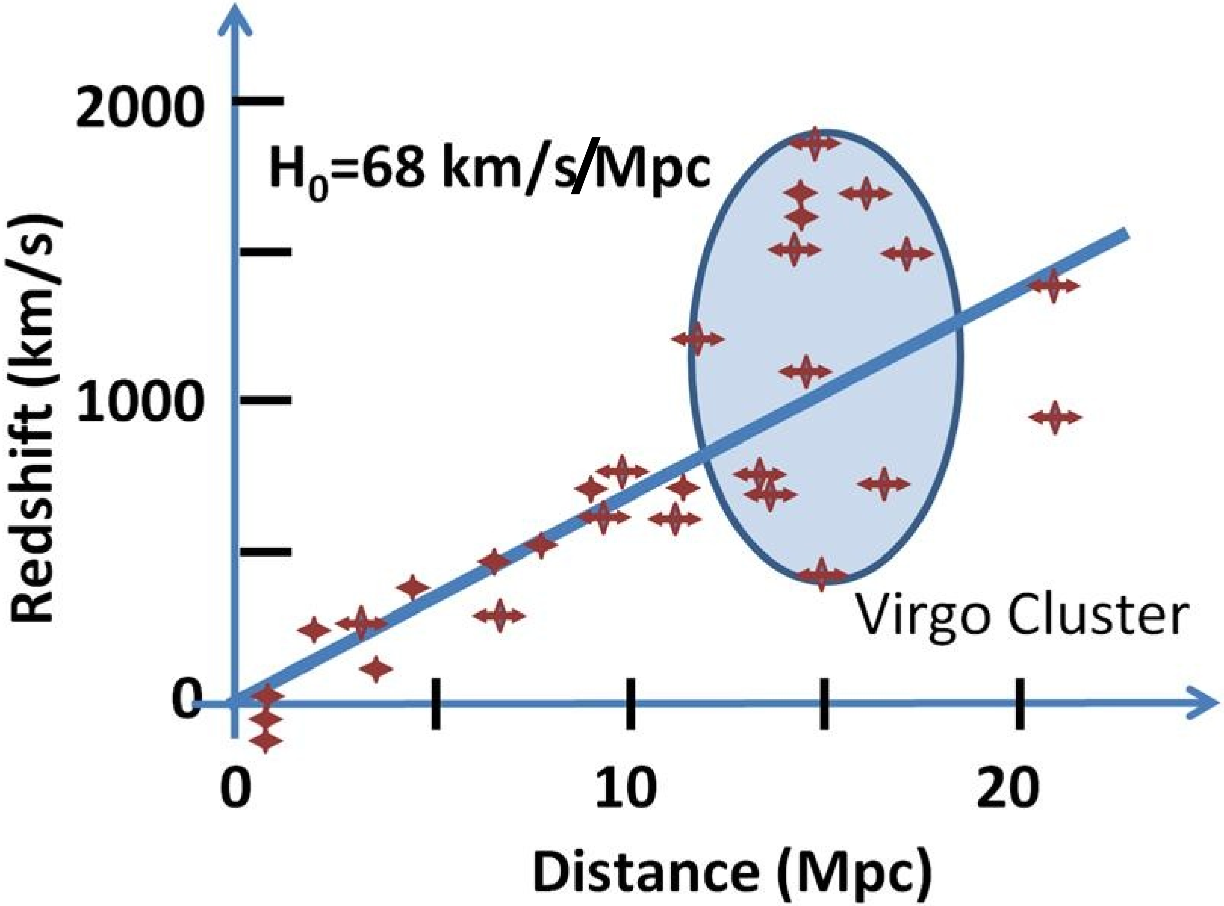
Do not snub linear fits!

Hypothesis: the recession velocity of galaxies depends linearly on their distance.

$V = H_0 \cdot d$

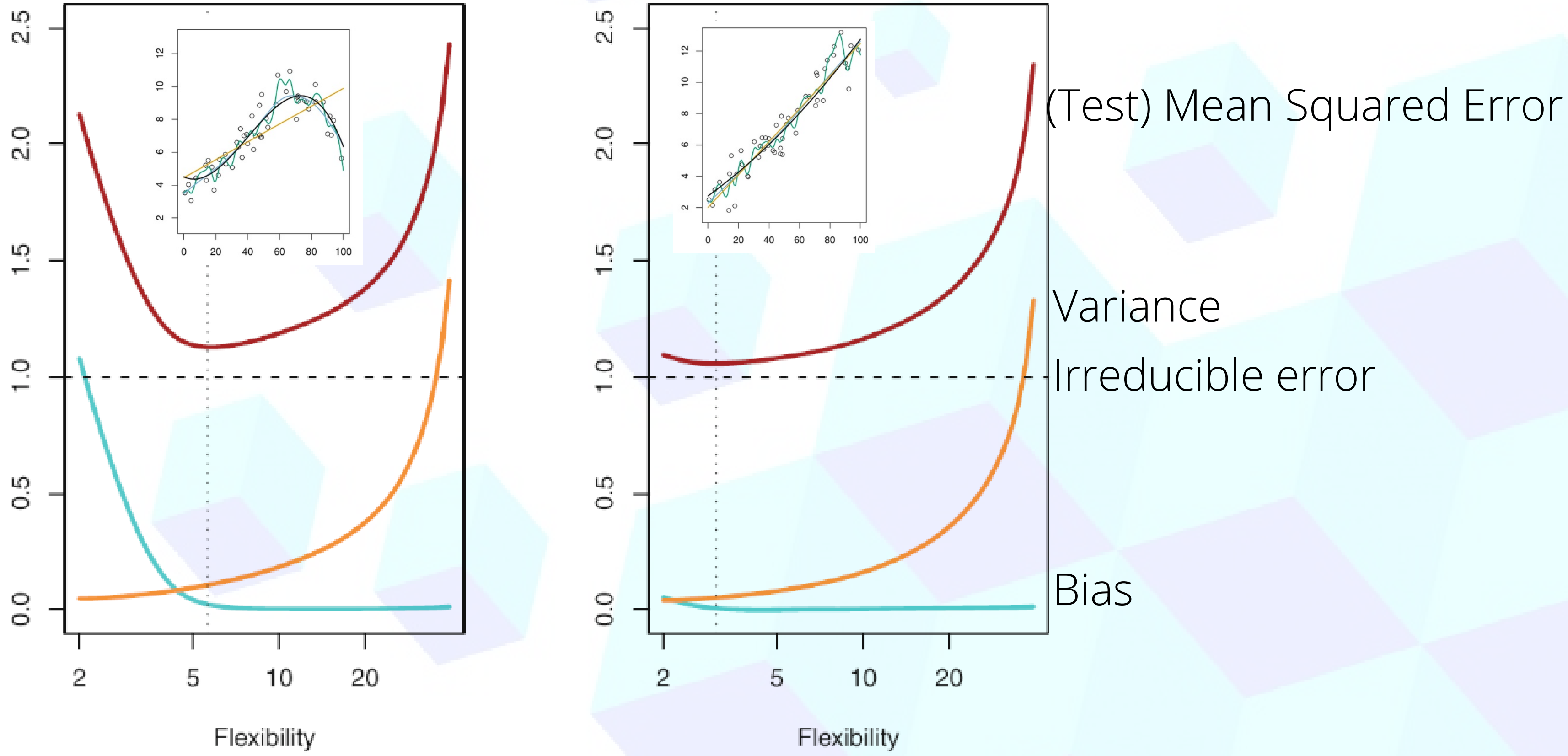


H_0 ~ 500 km/s/Mpc



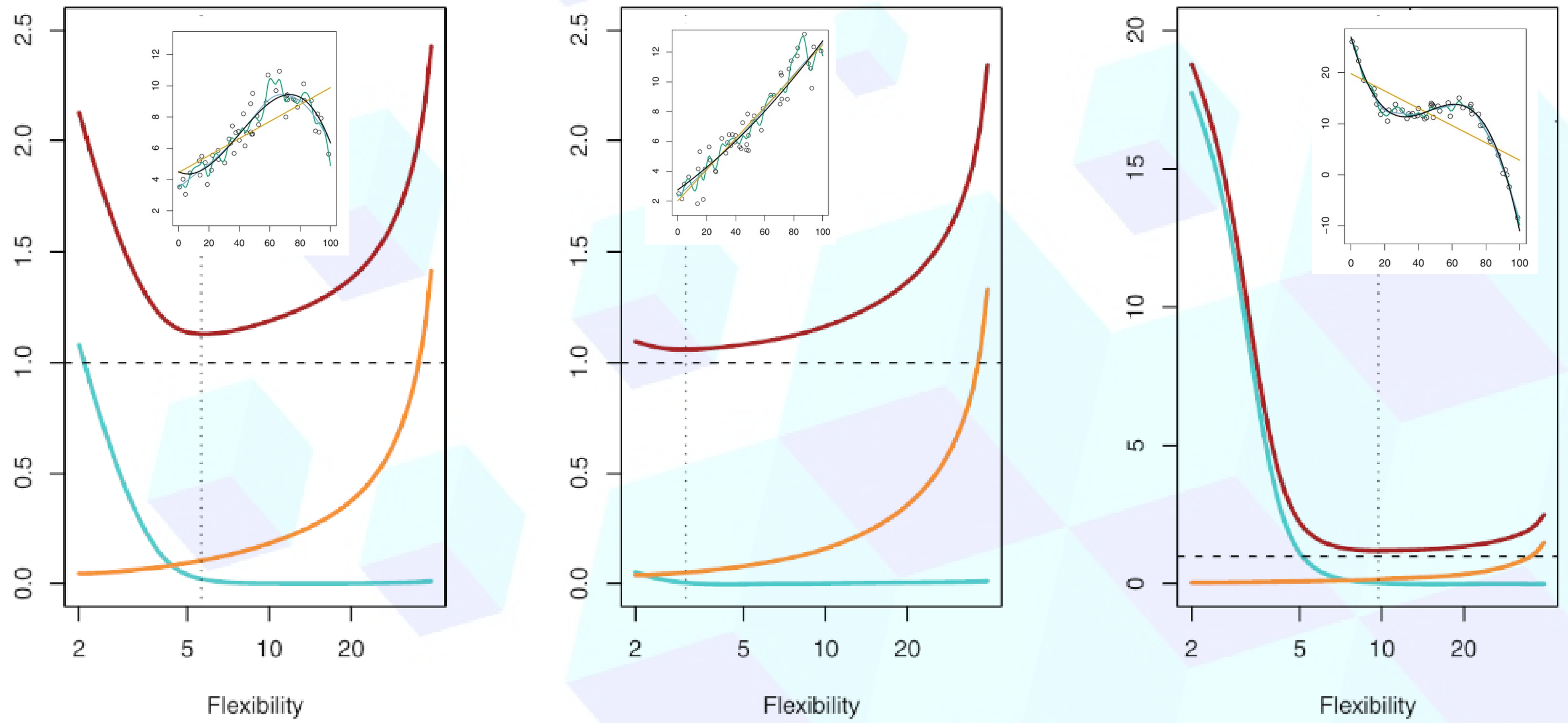
Supervised statistical learning: Bias/variance trade-off

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$



Supervised statistical learning: Bias/variance trade-off

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$



Supervised statistical learning: Learning metrics

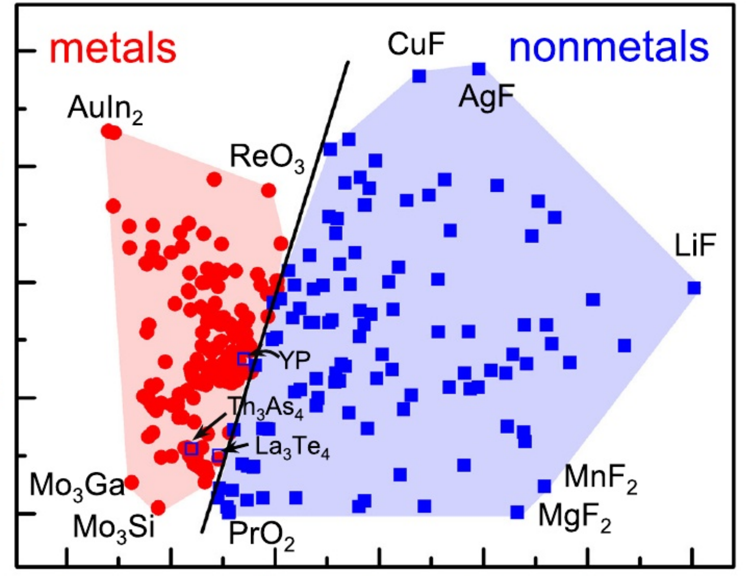
Regression: (Root) Mean Squared Error

Classification:

Misclassification error

$$\frac{1}{N}\sum_{i}^{N} I(y_i \neq \hat{y}_i)$$

1 if argument is true



The essence of learning: Regularization

$$\underset{\boldsymbol{c}\in\mathbb{R}^M}{\operatorname{argmin}} \sum_{j=1}^{N}\left(P_j - \sum_{l=1}^{M} d_{j,l}c_l\right)^2 = \underset{\boldsymbol{c}\in\mathbb{R}^M}{\operatorname{argmin}} \|\boldsymbol{P} - \boldsymbol{D}\boldsymbol{c}\|_2^2$$

ℓ_2 norm

p-norms: $\|\mathbf{x}\|_p = \left(\sum_i^N |x_i|^p\right)^{\frac{1}{p}}$

Regularization

Prefer "lower complexity" in the solution

$$\underset{\boldsymbol{c}\in\mathbb{R}^M}{\operatorname{argmin}} \|\boldsymbol{P} - \boldsymbol{D}\boldsymbol{c}\|_2^2 + \lambda\|\boldsymbol{c}\|_2^2$$

Supervised statistical learning: model selection

Hyperparameters

Let's assume a model class expressed as a sum over Gaussian (basis) functions:

$$P(\boldsymbol{d}) = \sum_{i=1}^{N} c_i \exp\left(-\|\boldsymbol{d}_i - \boldsymbol{d}\|_2^2 / 2\sigma^2\right)$$

$$\underset{\mathbf{c}\in\mathbb{R}^M}{\operatorname{argmin}} \sum_{i=1}^{N} (P(\boldsymbol{d}_i) - P_i)^2 + \lambda \sum_{i,j=1}^{N,N} c_i c_j \exp\left(-\|\boldsymbol{d}_i - \boldsymbol{d}_j\|_2^2 / 2\sigma^2\right)$$

Cross validation

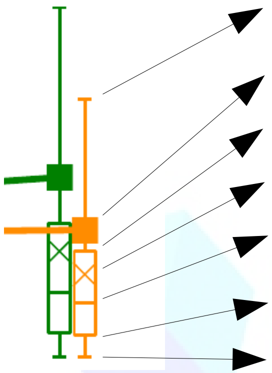
assessing the values of hyperparameters: model selection.

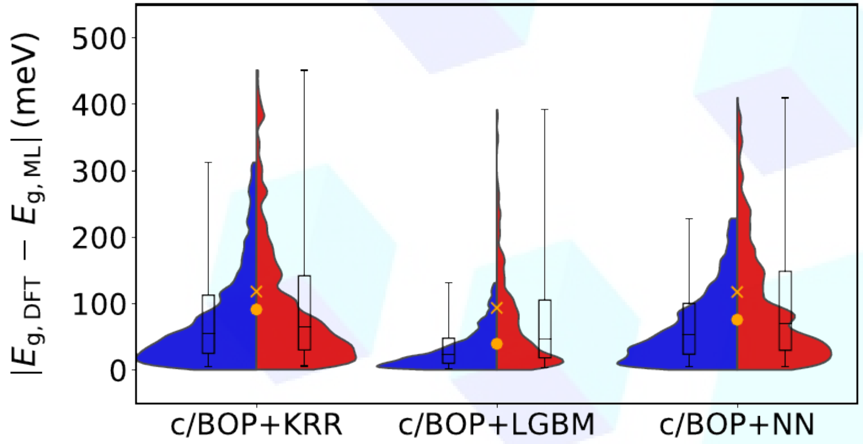
- Fix values of hyperparameters
- Split data into training and validation. Train and assess performance on test data
- Repeat over several training vs validation splits. Average performance over splits.
- Optimize hyperparameters (grid search, stochastic sampling)
  → select model with optimal values of the hyperparameters

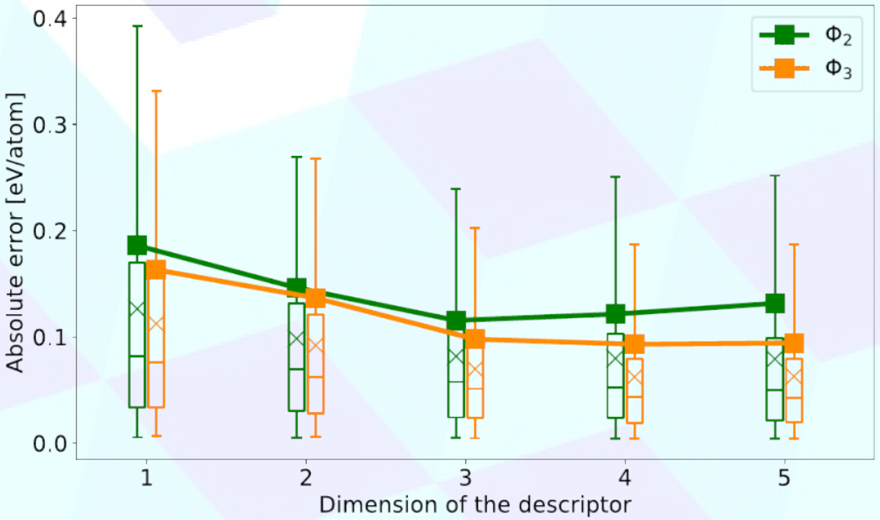
Regression: analysing results

The error metric for the analysis does not need to be the (root) mean squared error, (R)MSE.
Other interesting metrics: Mean absolute error (MAE), median (50th percentile), other percentiles.
Recommendation: use more statistic than just the center (average or median)

Box and violin plots







Sutton *et al.*, npj Comput Materials (2019) 10.1038/s41524-019-0239-3.

Ouyang, Ahmetcik *et al.*, J Phys Materials (2019) DOI: 10.1088/2515-7639/ab077b.

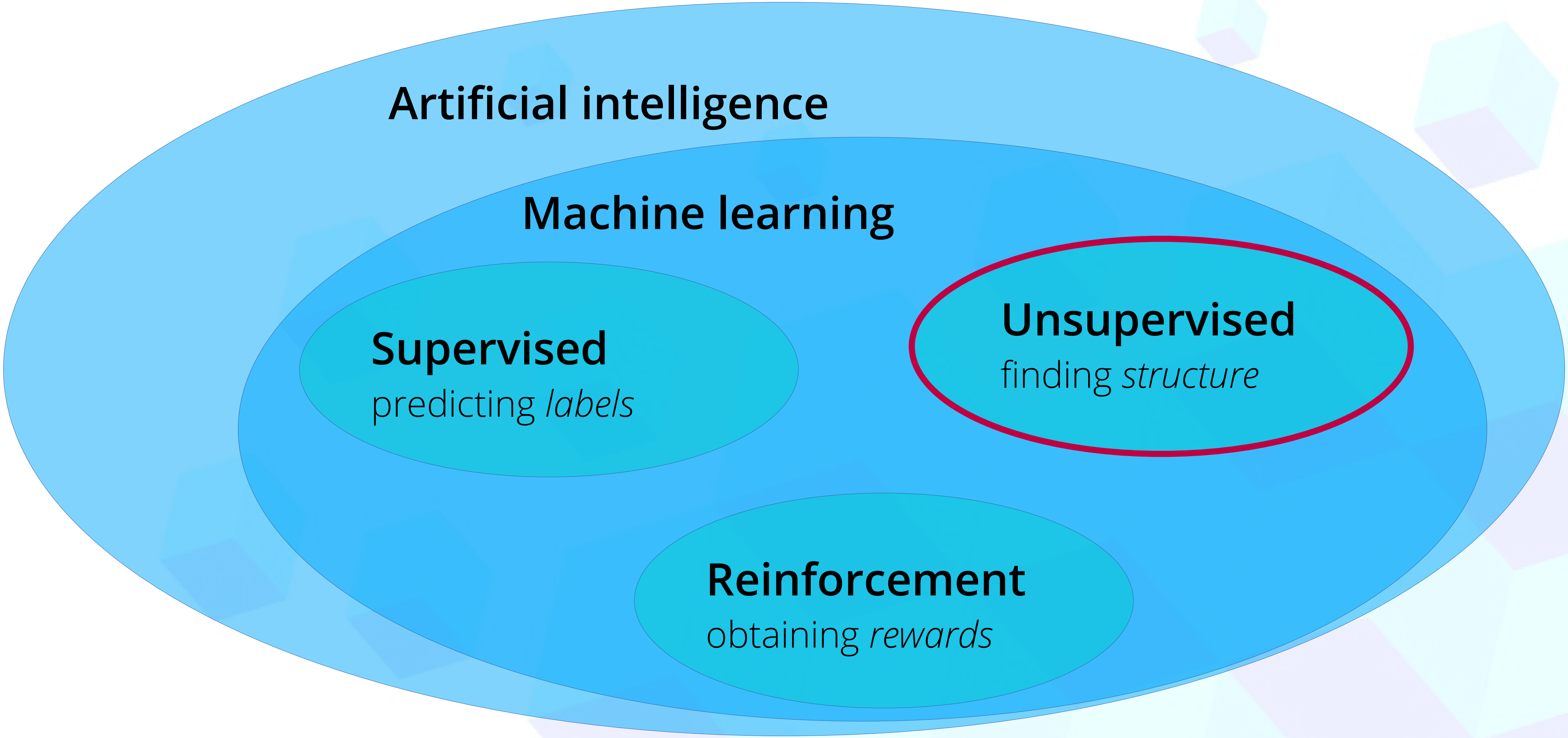
Classification: analysing results

		Predicted class	
		Positive	Negative
Actual class	Positive	True Positives (TP)	False Negatives (FN) Type I errors
	Negative	False Positives (FP) Type II errors	True Negatives (TN)

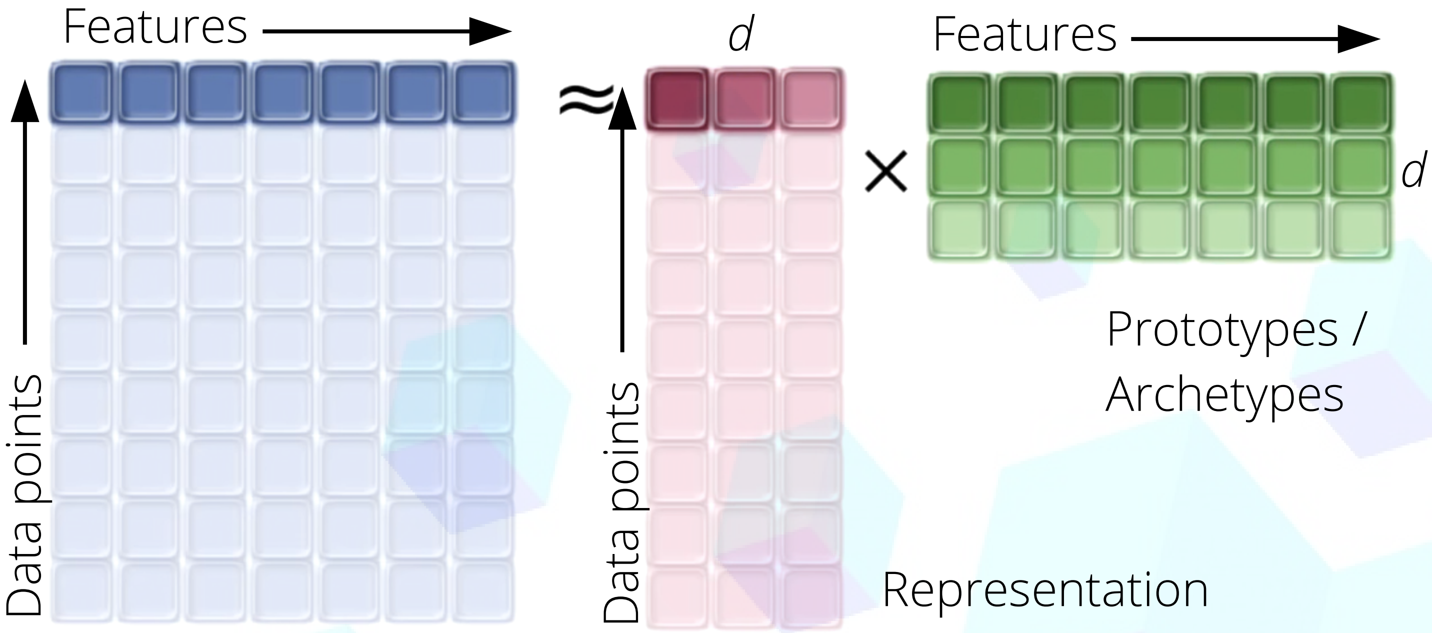
E.g.
Positive: is a metal

Accuracy = (TP + TN) / All
Precision = TP / (TP+FP)
Sensitivity = TP / (TP+FN)
Specificity = TN / (TN+FP)

Few bits of taxonomy

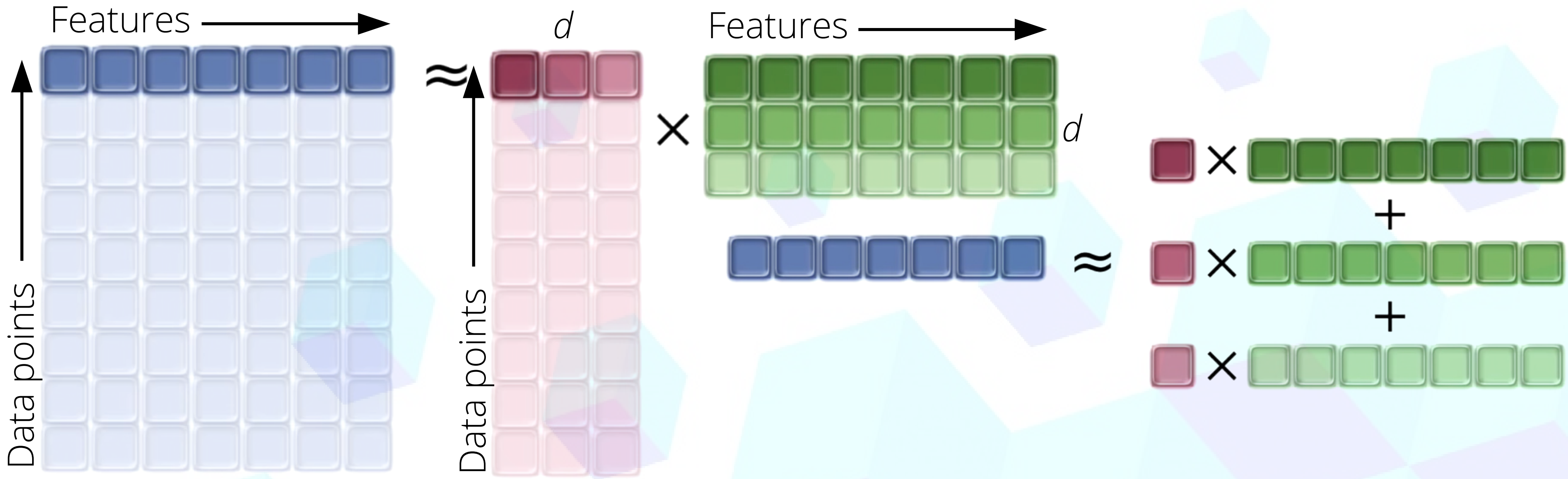


Unsupervised learning looking from space



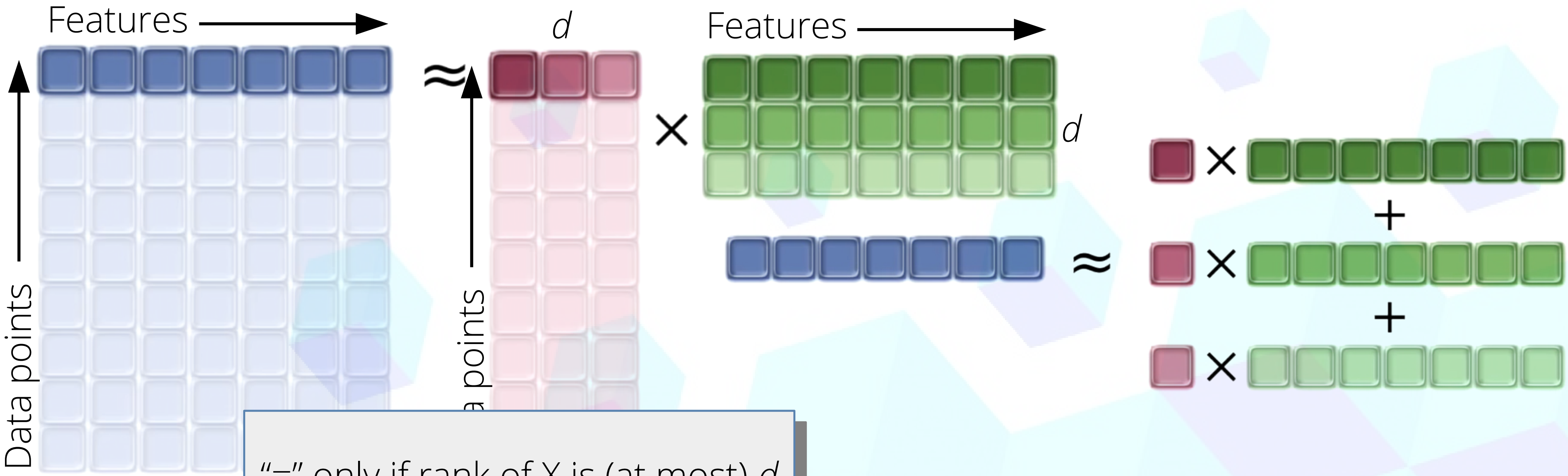
Leland McInnes, A Bluffer's Guide to Dimension Reduction https://youtu.be/9iol3Lk6kyU
Udell, Horn Zadell, Boyd, Generalized low rank models 2016 arXiv: 1410.0342

Unsupervised learning looking from space



Leland McInnes, A Bluffer's Guide to Dimension Reduction https://youtu.be/9iol3Lk6kyU
Udell, Horn Zadell, Boyd, Generalized low rank models 2016 arXiv: 1410.0342

Unsupervised learning looking from space



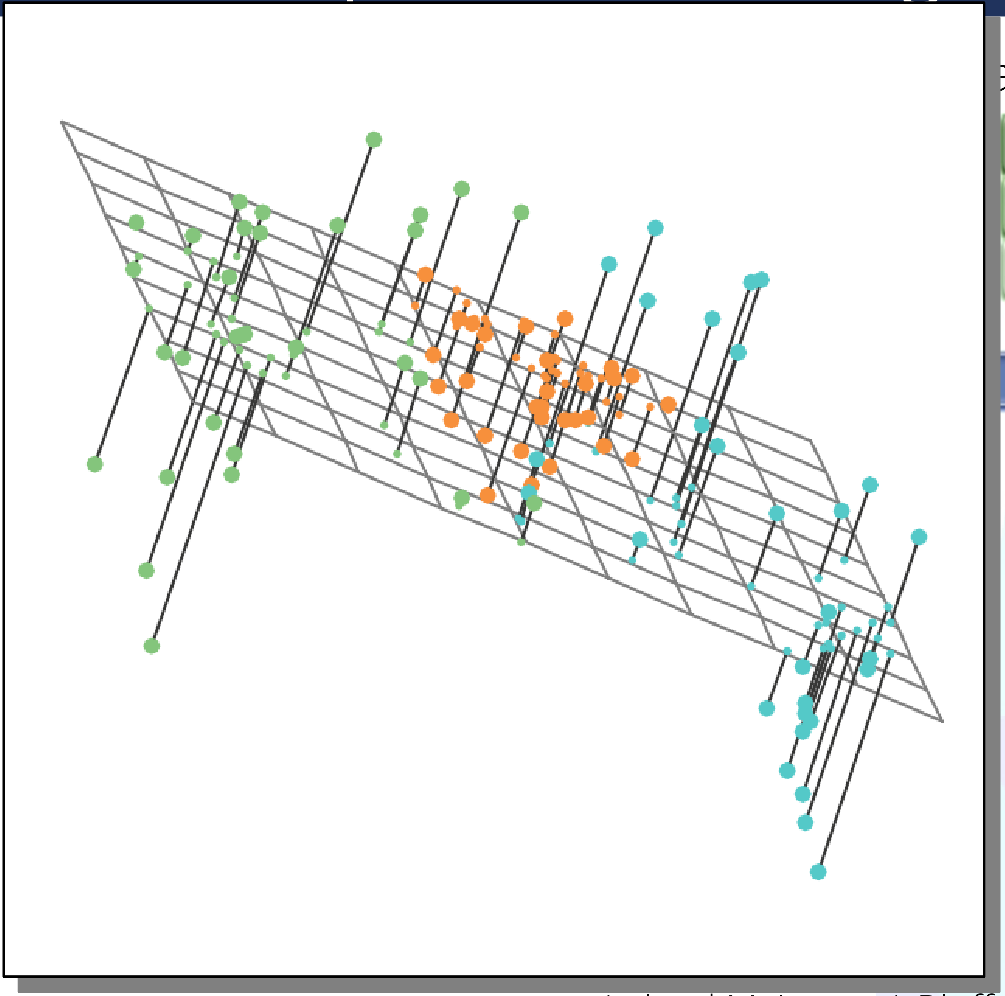
"=" only if rank of X is (at most) *d*

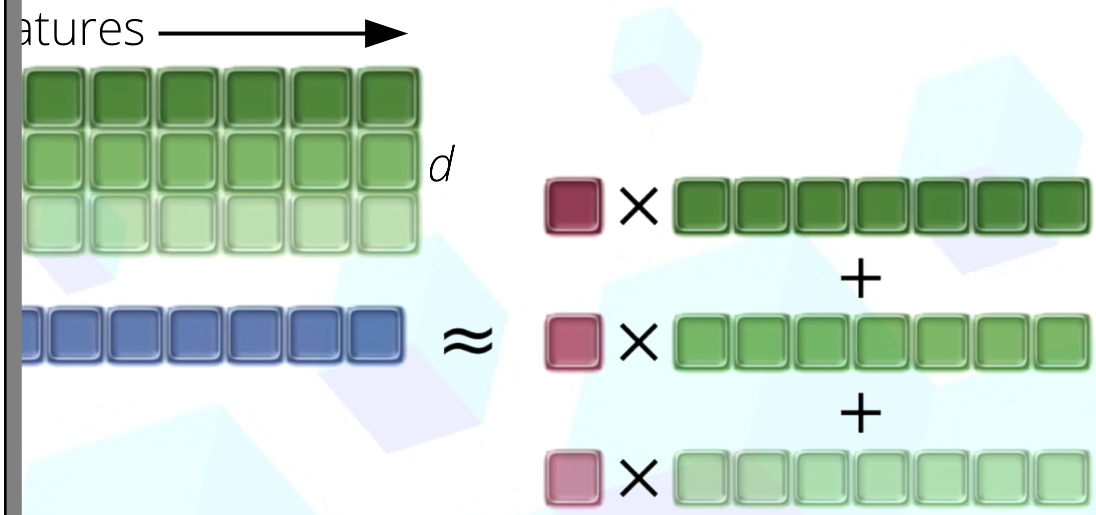
$$X \approx UV$$

minimize $\sum_{i=1}^{N} \sum_{j=1}^{D} \left(X_{ij} - (UV)_{ij} \right)^2$

Leland McInnes, A Bluffer's Guide to Dimension Reduction https://youtu.be/9iol3Lk6kyU
Udell, Horn Zadell, Boyd, Generalized low rank models 2016 arXiv: 1410.0342

Unsupervised learning looking from space



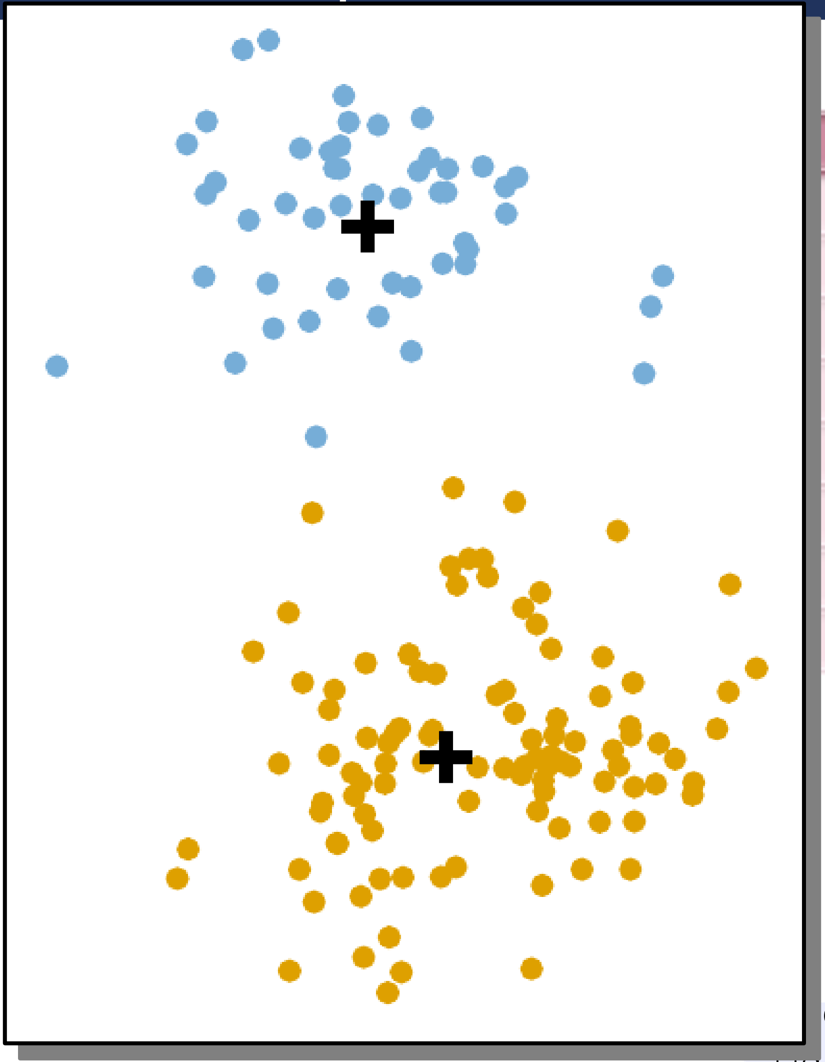


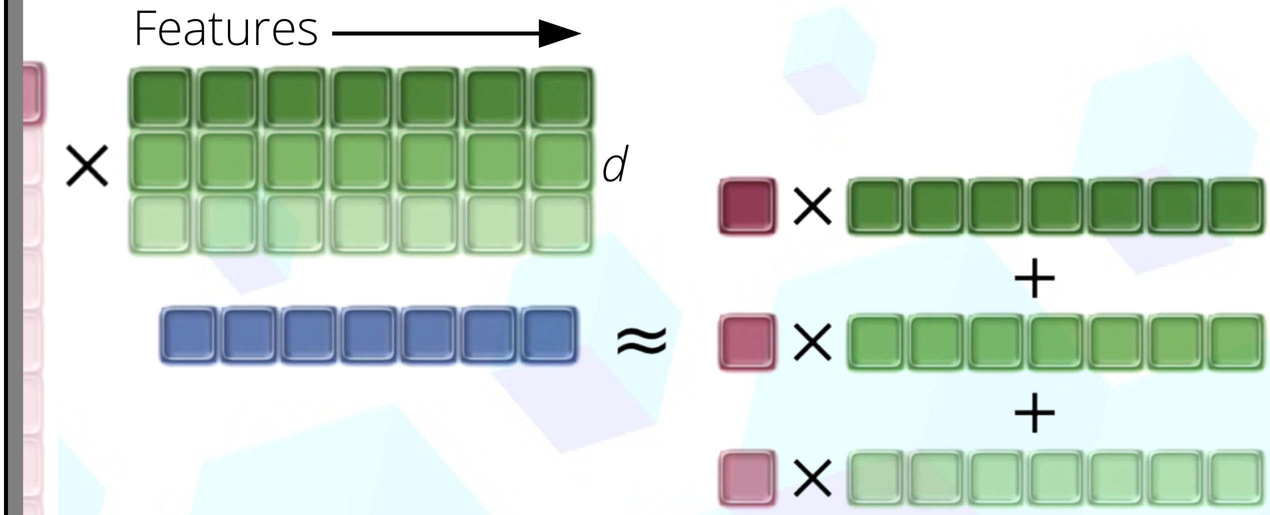
As many *d* as necessary:

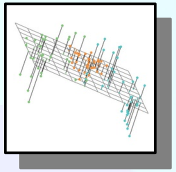
Dimension reduction

Leland McInnes, A Bluffer's Guide to Dimension Reduction https://youtu.be/9iol3Lk6kyU
Udell, Horn Zadell, Boyd, Generalized low rank models 2016 arXiv: 1410.0342

Unsupervised learning looking from space





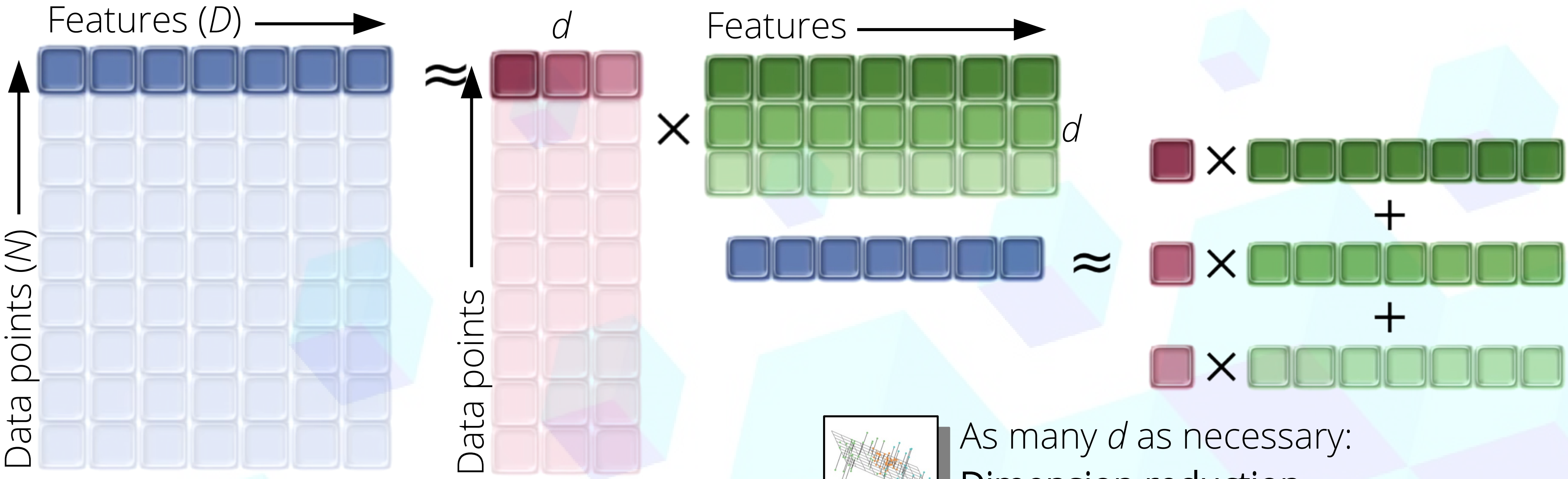


As many *d* as necessary:
Dimension reduction

Only one *d*:
(centroid-based) **cluster analysis**

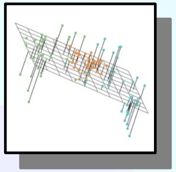
...es, A Bluffer's Guide to Dimension Reduction https://youtu.be/9iol3Lk6kyU
...dell, Horn Zadell, Boyd, Generalized low rank models 2016 arXiv: 1410.0342

Unsupervised learning looking from space

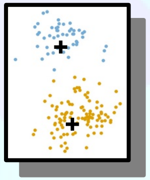


$$X \approx UV$$

minimize $$\sum_{i=1}^{N}\sum_{j=1}^{D}\left(X_{ij} - (UV)_{ij}\right)^2$$



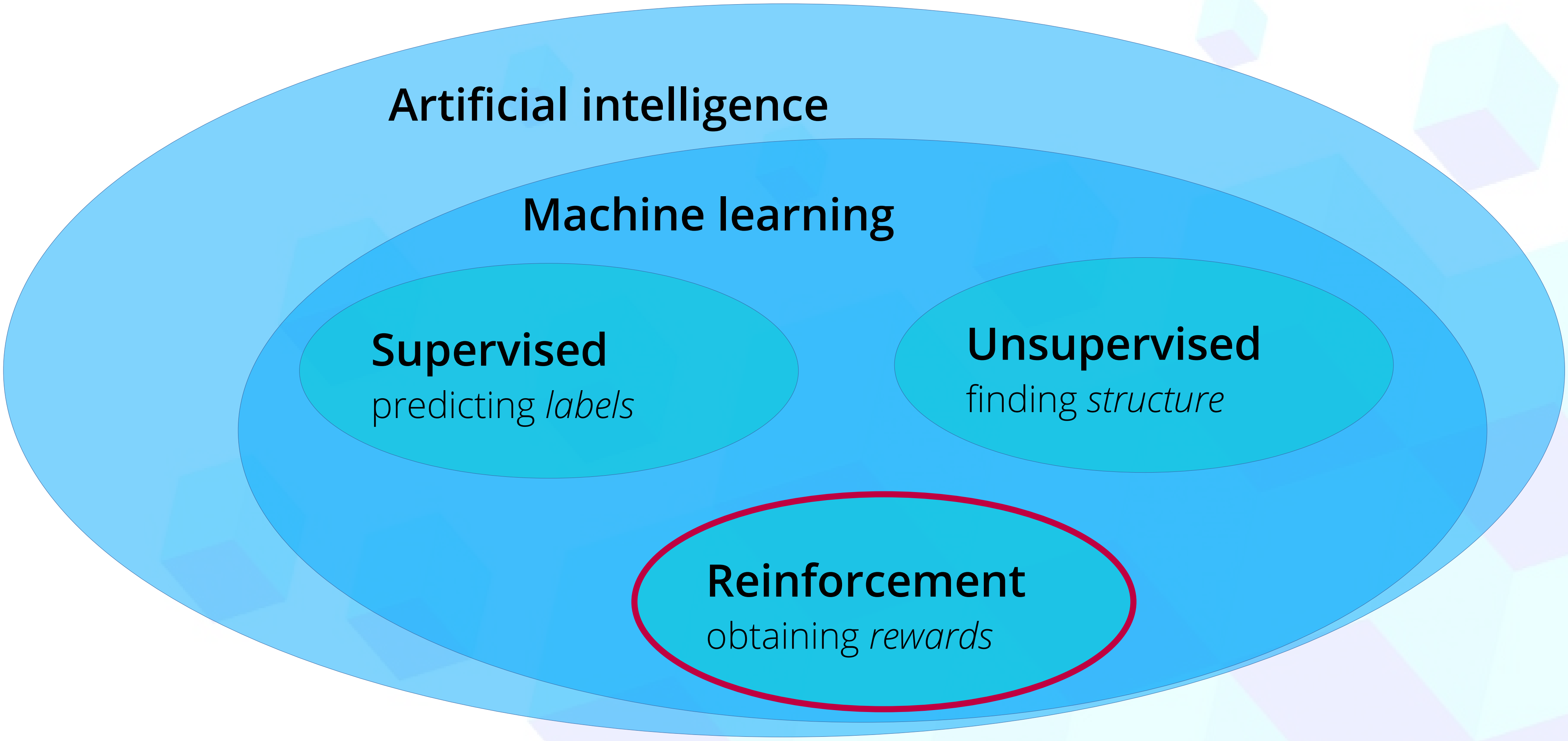
As many d as necessary:
Dimension reduction



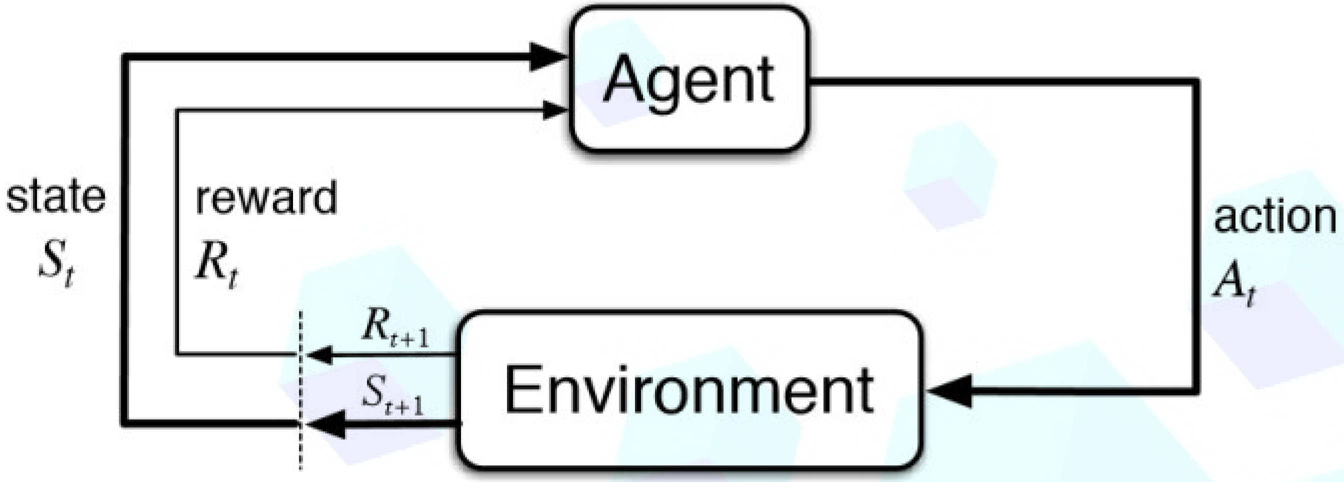
Only one d:
(centroid-based) **cluster analysis**

Leland McInnes, A Bluffer's Guide to Dimension Reduction https://youtu.be/9iol3Lk6kyU
Udell, Horn Zadell, Boyd, Generalized low rank models 2016 arXiv: 1410.0342

Few bits of taxonomy



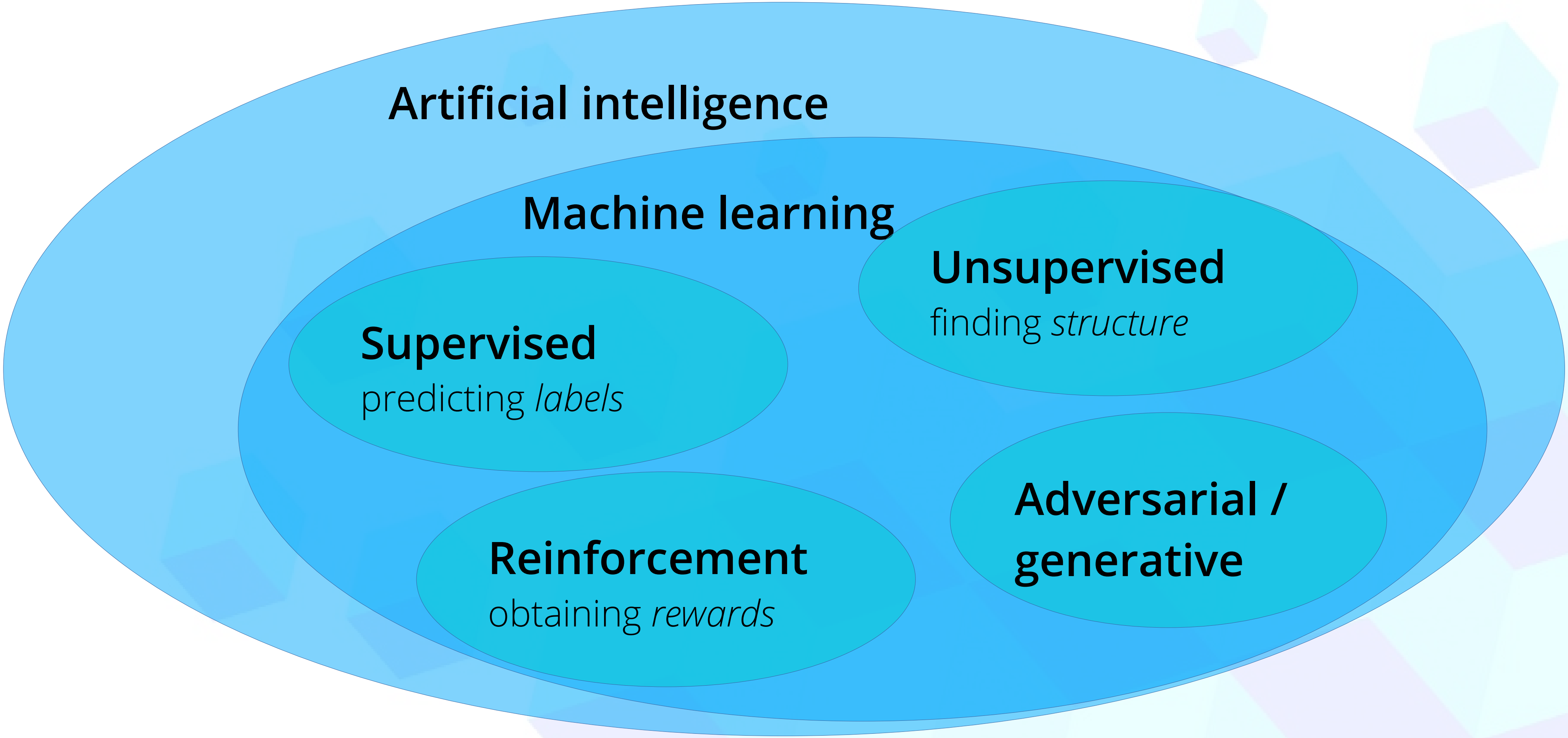
Reinforcement learning



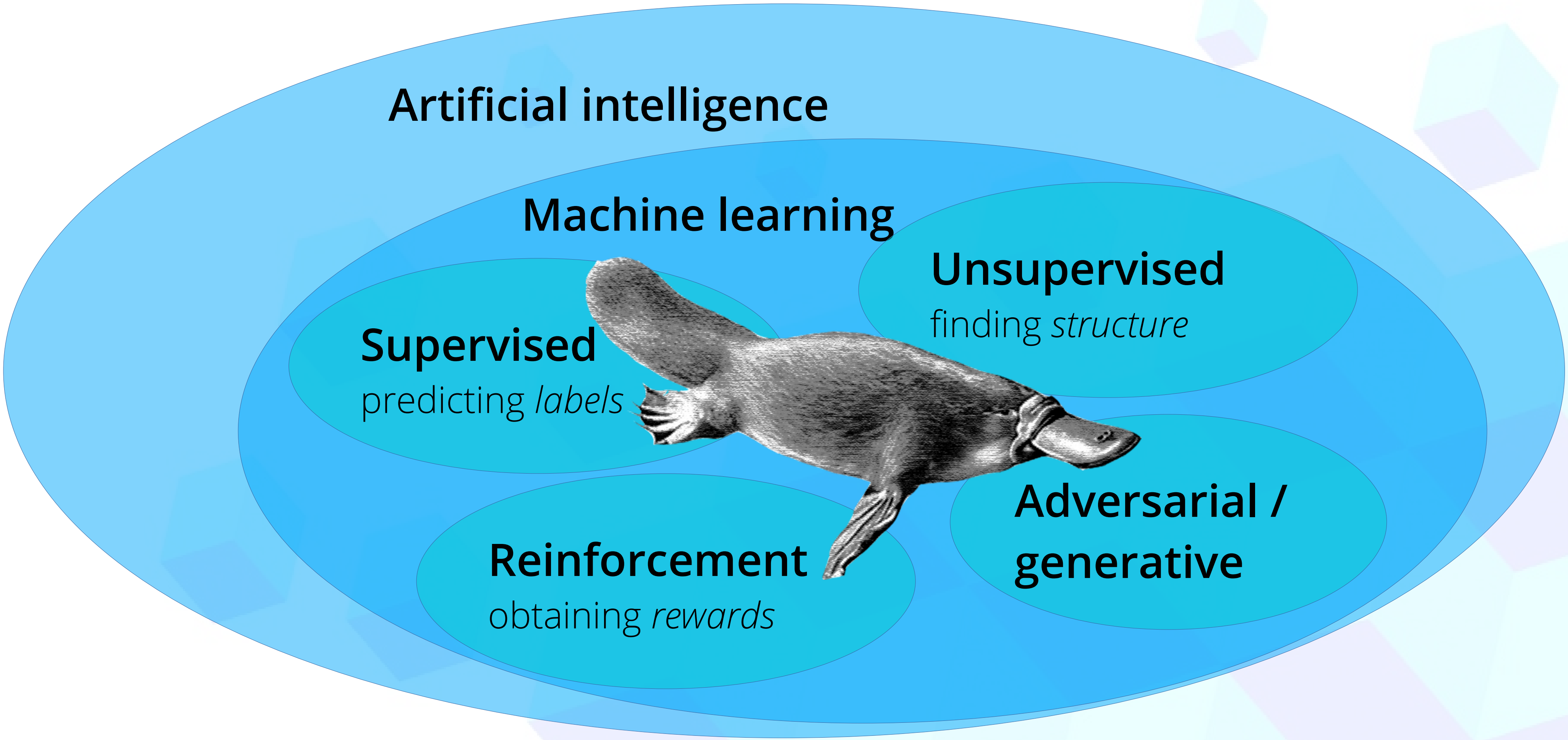
"A supervised learning starting which builds its own training dataset starting from no data".
The *agent* changes its *state* by taking *actions*: i.e., by *exploring* or *exploiting* the environment, on the basis of expected *reward* and *policy*. On the basis of the feedback from the *environment* (actual *rewards* or *punishments*), the model is updated.

See e.g.:
Jørgensen et al., JCP 2019 doi: https://doi.org/10.1063/1.5108871

Few bits of taxonomy



Few bits of taxonomy



Applications in materials science, an overview

Materials analysis/predictions

Predicting properties from composition (& structure) – materials informatics
Predicting properties from (ensemble of) configurations – surrogate models
Interpreting measurements, e.g., microscopy data

Materials design

Structure-oriented design
Elements-oriented design
Inverse design (Workflow design / Active learning)

The challenges

Small data
Reliability/Accountability

Applications in materials science, an overview

Materials analysis/predictions

Predicting properties from composition (& structure) – materials informatics
Predicting properties from (ensemble of) configurations – surrogate models
Interpreting measurements, e.g., microscopy data

Materials design

Structure-oriented design
Elements-oriented design
Inverse design (Workflow design / Active learning)

The challenges

Small data
Reliability/Accountability

I - An AI may not injure a human being or, through inaction, allow a human being to come to harm.

II - An AI must obey orders given it by human beings except where such orders would conflict with the First Law.

Applications in materials science, an overview

Materials analysis/predictions

Predicting properties from composition (& structure) – materials informatics

Predicting properties from (ensemble of) configurations – surrogate models

Interpreting measurements, e.g., microscopy data

Materials design

Structure-oriented design

Elements-oriented design

Inverse design (Workflow design / Active learning)

The challenges

Small data

Reliability/Accountability

Interpretability/Explainability

AI algorithms in the future lectures

Infrastructure

NOMAD and FAIRmat	Markus Scheidgen

When (initial) data are scarce

Active and reinforcement learning	Bjørk Hammer
Multi-fidelity learning	Gian-Marco Rignanese
Bayesian inference and optimization	Kentaro Kutsukake
Experiment design	Sergei Kalinin

Learning and exploiting interatomic potentials

Neural-network based	Daniel Schwalbe Koda
Kernel based	Volker Deringer