Exploring practices in machine learning and machine discovery for heterogeneous catalysis

Ichi Takigawa

https://itakigawa.github.io/

Institute for Liberal Arts and Sciences, Kyoto University
Institute for Chemical Reaction Design and Discovery, Hokkaido University
RIKEN Center for Advanced Intelligence Project

This talk

Share **a viewpoint from** **the ML side (as I am an ML researcher, not a chemist)**
after >7 years struggling in heterogeneous catalyst design and discovery

With great people in chemistry!

Hokkaido University
Institute for Catalysis

Prof. Ken-ichi SHIMIZU

Prof. Takashi TOYAO

Prof. Satoru Takakusagi
Prof. Zen Maeno
Prof. Takashi Kamachi

Prof. Koji Tsuda (U Tokyo)

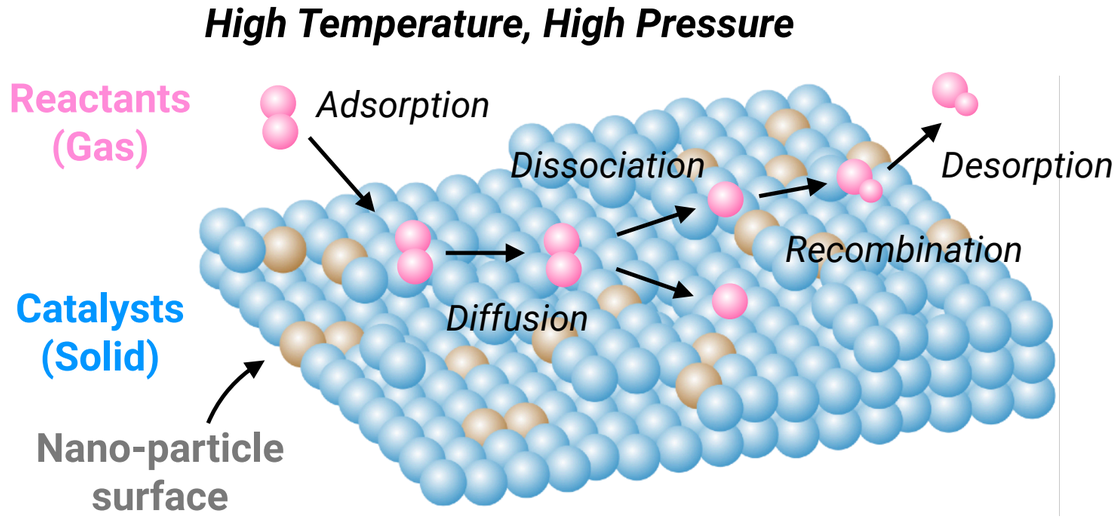
Shinya Mine
Takumi Mukaiyama
Motoshi Takao
Yuan Jing
Gang Wang
Duotian Chen
Kah Wei Ting
Taichi Yamaguchi
Koichi Matsushita
S.M.A.H. Siddiki

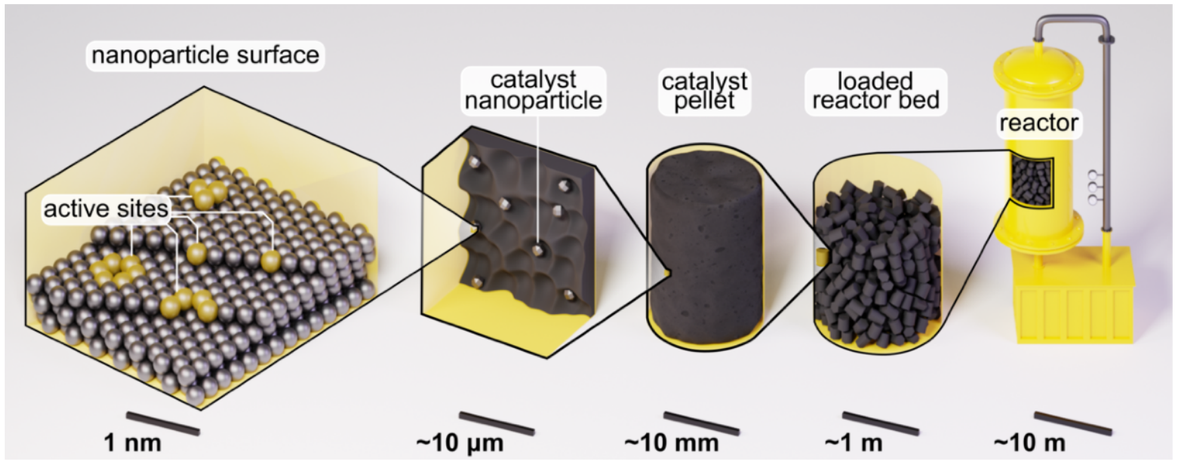
Keisuke Suzuki
Shoma Kikuchi

Heterogeneous catalysis

Gas-phase reactions on solid-phase catalyst surface (Heterogeneous catalysis)

Industrial Synthesis (e.g. Haber-Bosch), Automobile Exhaust Gas Purification, Methane Conversion, etc.



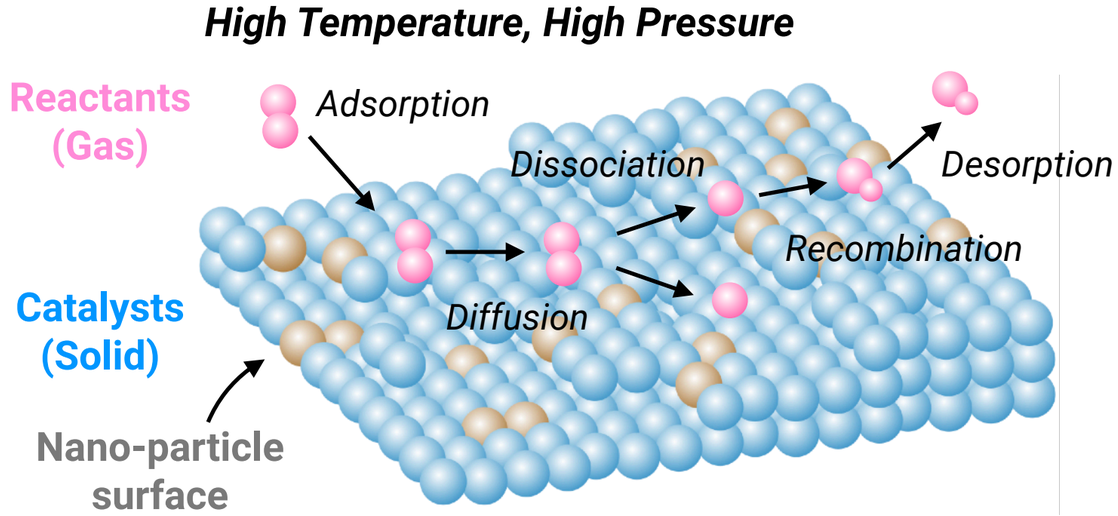


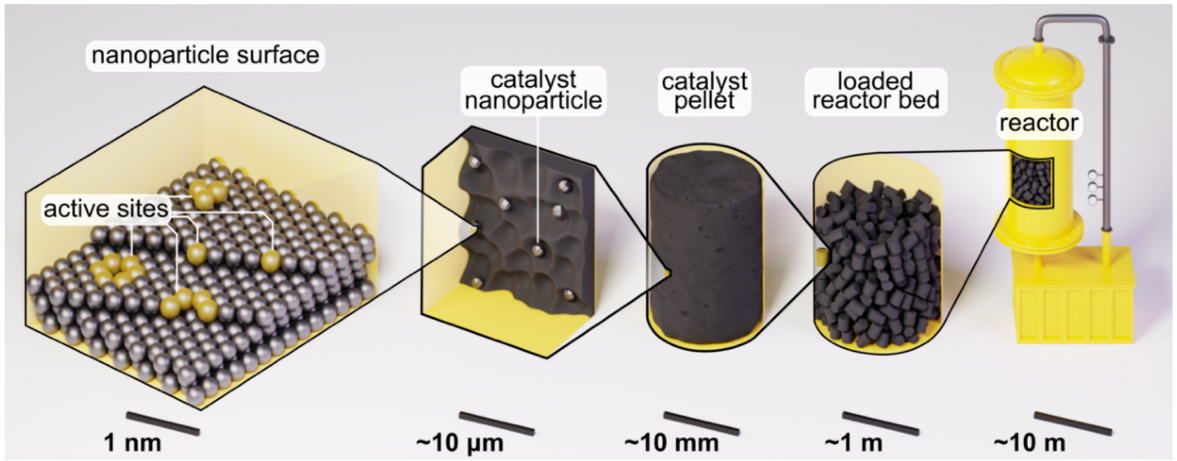
https://en.wikipedia.org/wiki/Heterogeneous_catalysis

Heterogeneous catalysis

Gas-phase reactions on solid-phase catalyst surface (Heterogeneous catalysis)

Industrial Synthesis (e.g. Haber-Bosch), Automobile Exhaust Gas Purification, Methane Conversion, etc.



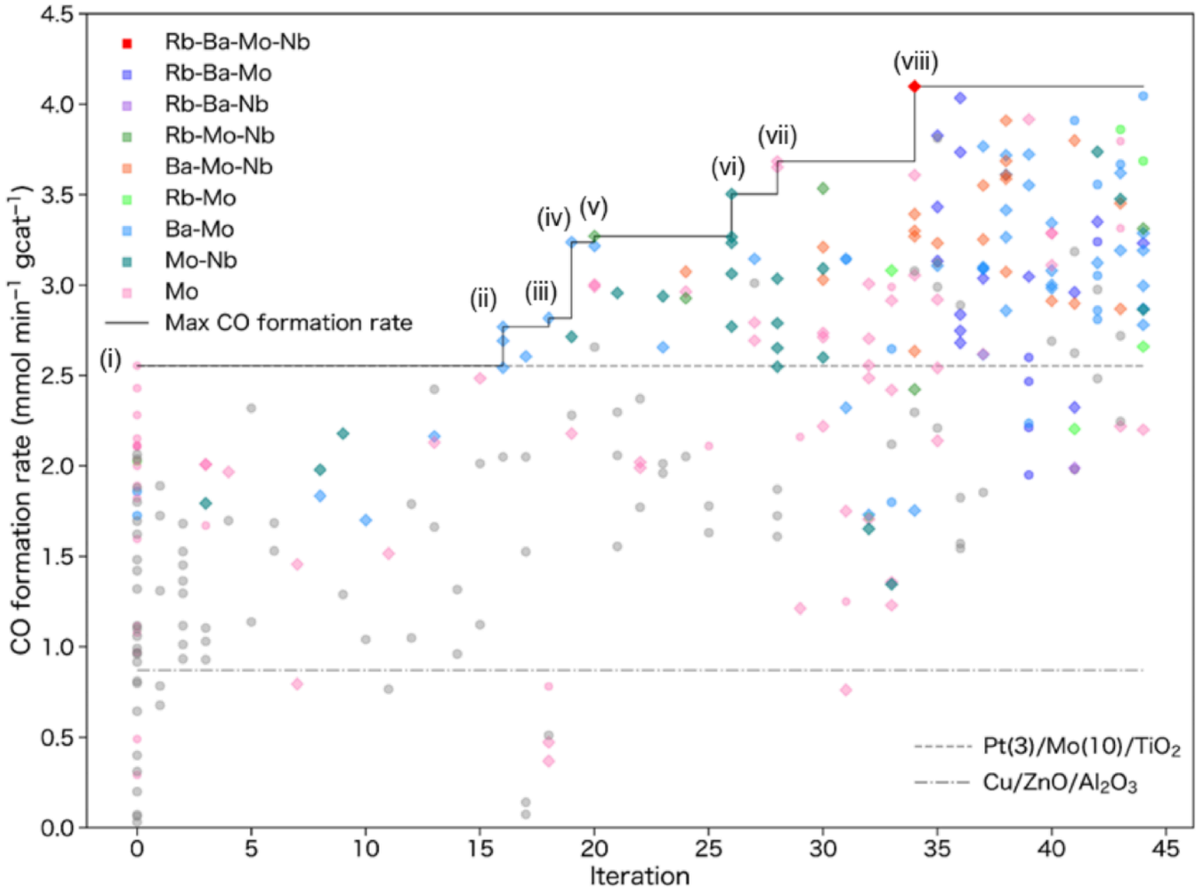


https://en.wikipedia.org/wiki/Heterogeneous_catalysis

Involves **devilishly complex too-many-factor processes.**
A solid surface shares its border with the external world.

God made the bulk; the **surface** was invented by the **devil** —— Wolfgang Pauli

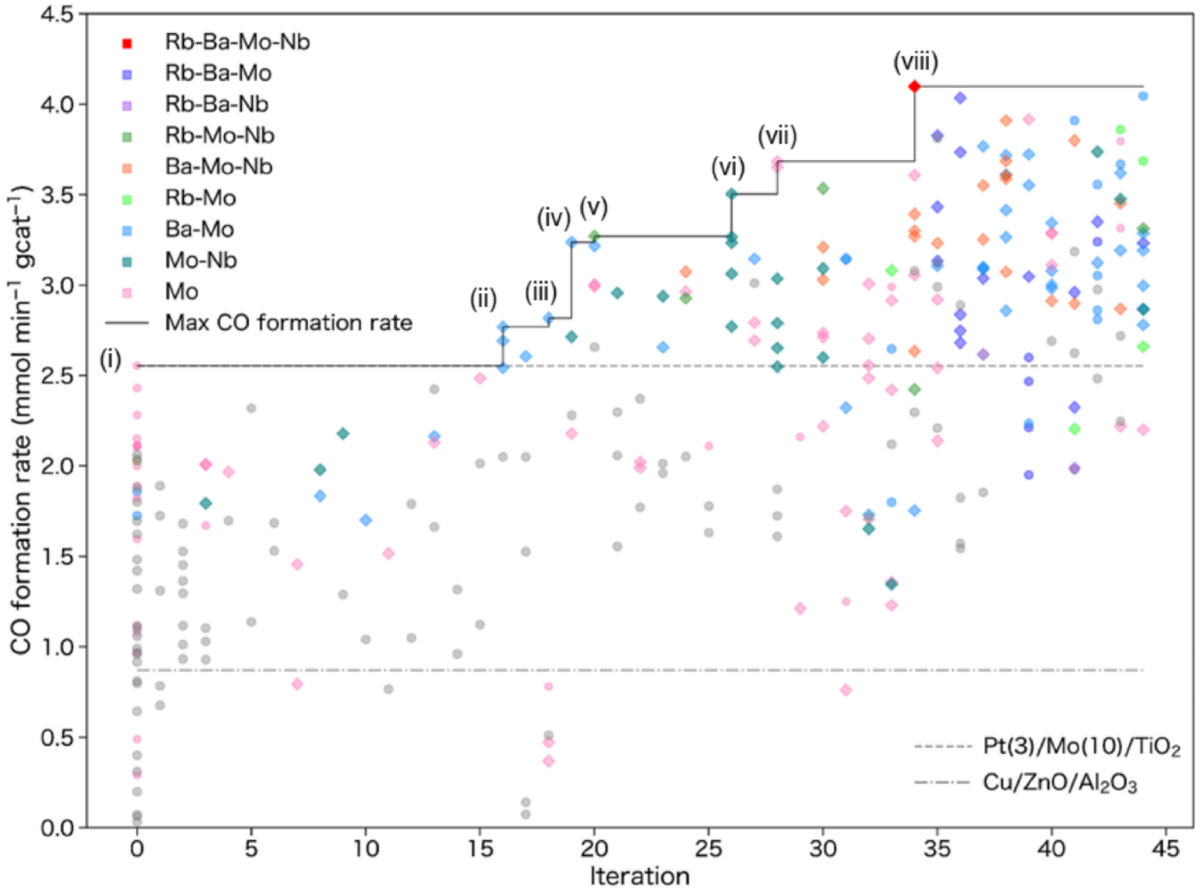
Our recent research: Results



Our Target:
$Pt(3)/X_1\text{-}X_2\text{-}X_3\text{-}X_4\text{-}X_5/TiO_2$ RWGS Calalyst

Accelerated discovery of multi-elemental reverse water-gas shift catalysts using extrapolative machine learning approach. https://doi.org/10.26434/chemrxiv-2022-695rj

Our recent research: Results

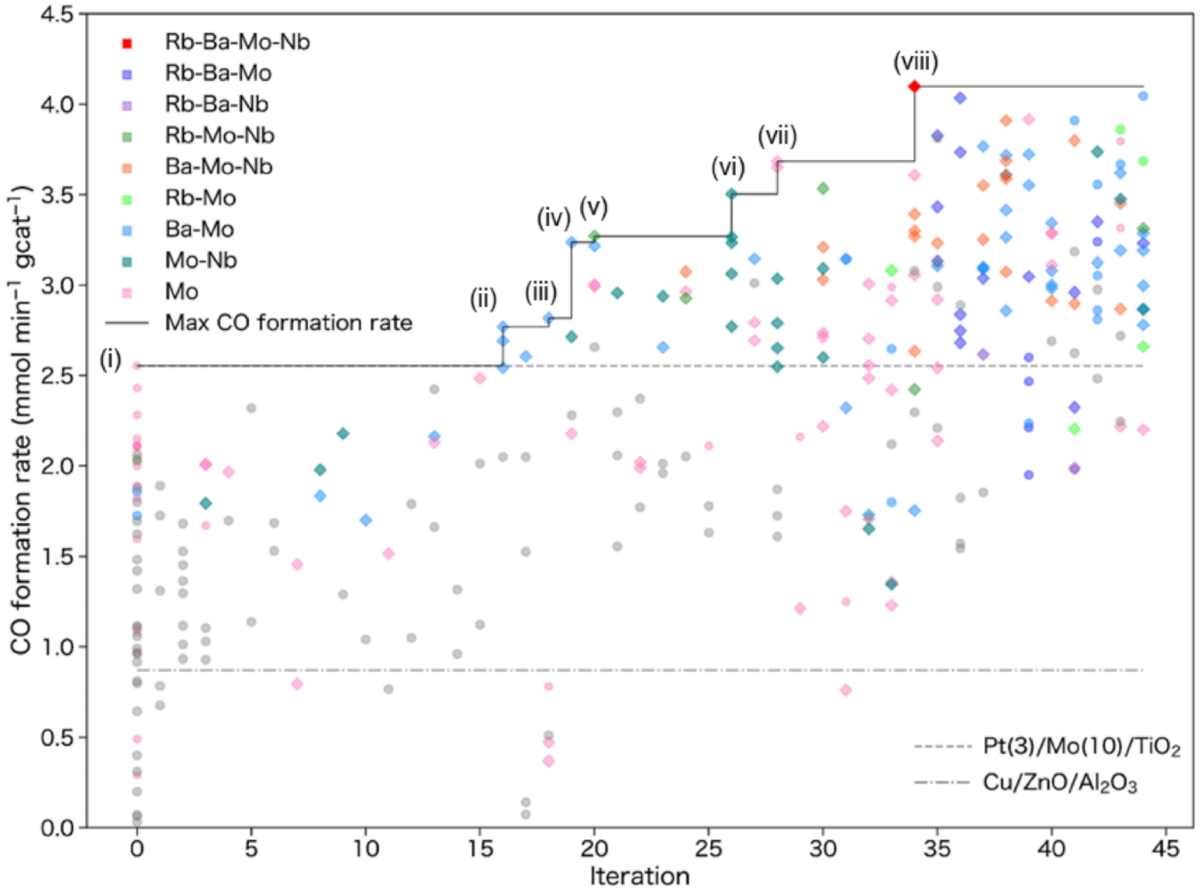


Our Target:
$Pt(3)/X_1\text{-}X_2\text{-}X_3\text{-}X_4\text{-}X_5/TiO_2$ RWGS Calalyst

- Discovered **more than 100 catalysts** better than the previously reported best catalyst.

Accelerated discovery of multi-elemental reverse water-gas shift catalysts using extrapolative machine learning approach. https://doi.org/10.26434/chemrxiv-2022-695rj

Our recent research: Results

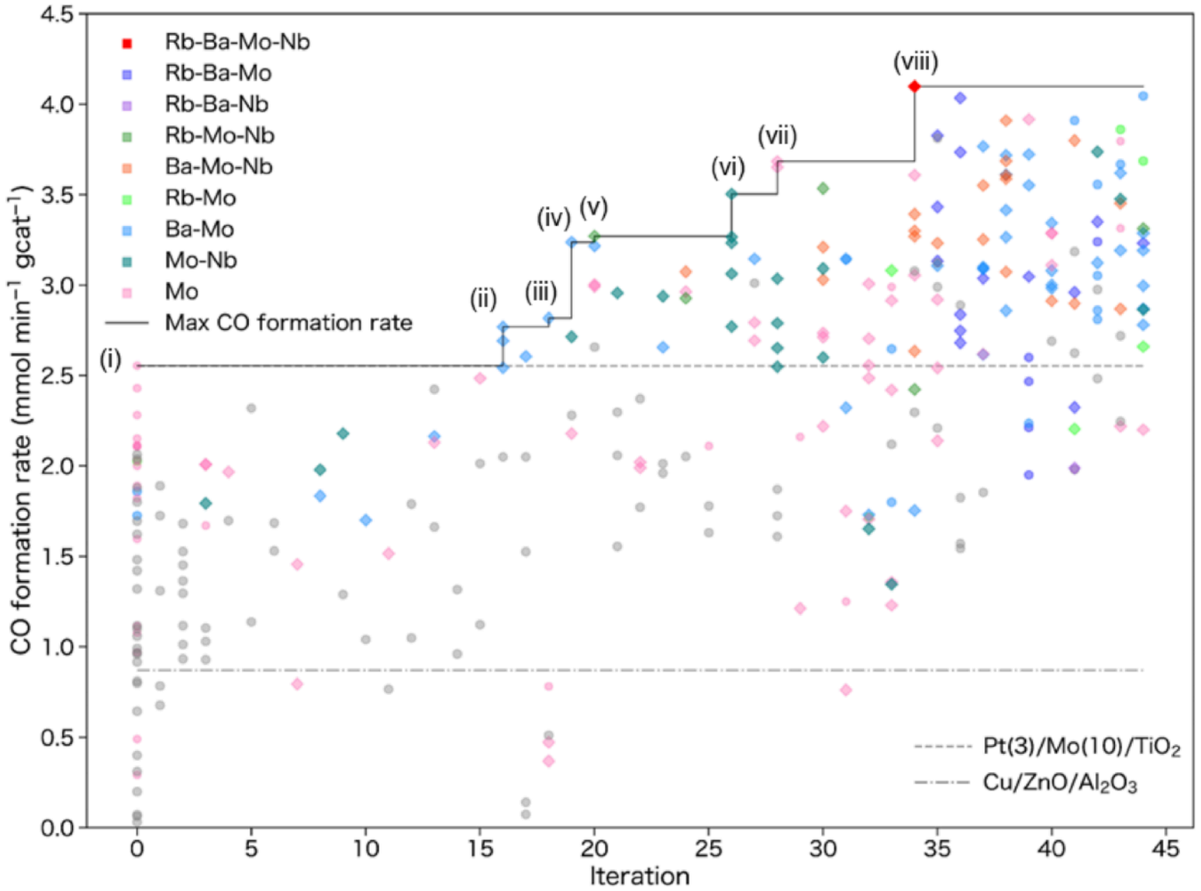


Our Target:
$Pt(3)/X_1\text{-}X_2\text{-}X_3\text{-}X_4\text{-}X_5/TiO_2$ RWGS Calalyst

- Discovered **more than 100 catalysts** better than the previously reported best catalyst.
- **300 catalysts tested in total** by **44 cycles** of ML prediction + experiment

Accelerated discovery of multi-elemental reverse water-gas shift catalysts using extrapolative machine learning approach. https://doi.org/10.26434/chemrxiv-2022-695rj

Our recent research: Results

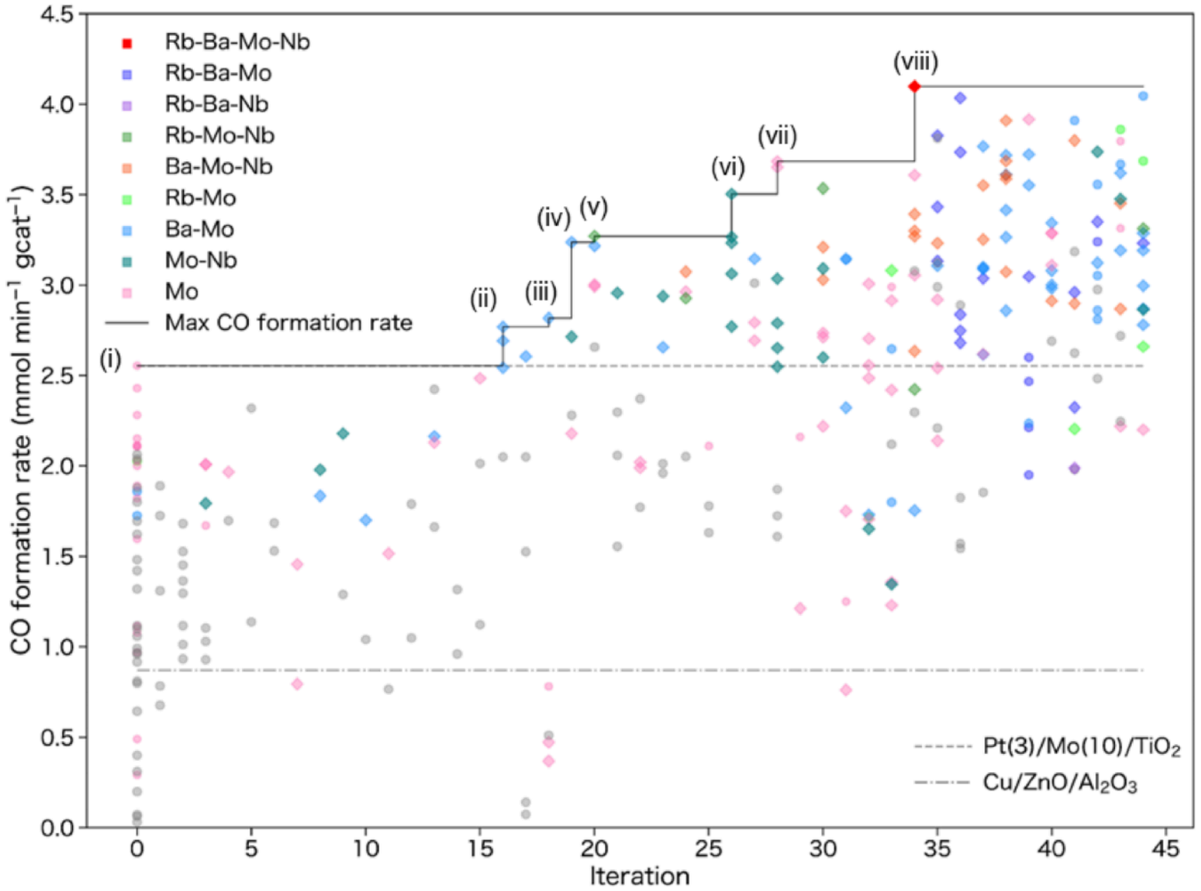


Our Target:
$Pt(3)/X_1\text{-}X_2\text{-}X_3\text{-}X_4\text{-}X_5/TiO_2$ RWGS Calalyst

- Discovered **more than 100 catalysts** better than the previously reported best catalyst.
- **300 catalysts tested in total** by **44 cycles** of ML prediction + experiment
- The optimal catalyst Pt(3)/Rb(1)-Ba(1)-Mo(0.6)-Nb(0.2)/TiO2 was hardly predictable by human experts

Accelerated discovery of multi-elemental reverse water-gas shift catalysts using extrapolative machine learning approach. https://doi.org/10.26434/chemrxiv-2022-695rj

Our recent research: Results



Our Target:
$Pt(3)/X_1\text{-}X_2\text{-}X_3\text{-}X_4\text{-}X_5/TiO_2$ RWGS Calalyst

- Discovered **more than 100 catalysts** better than the previously reported best catalyst.
- **300 catalysts tested in total** by **44 cycles** of ML prediction + experiment
- The optimal catalyst Pt(3)/Rb(1)-Ba(1)-Mo(0.6)-Nb(0.2)/TiO2 was hardly predictable by human experts
- Notably, Nb was **never used** in training.

Accelerated discovery of multi-elemental reverse water-gas shift catalysts using extrapolative machine learning approach. https://doi.org/10.26434/chemrxiv-2022-695rj

Our recent research: Method

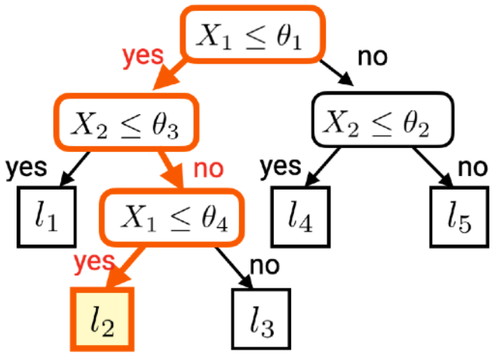
Decision tree ensembles (with UQ)
i.e. histogram on data-dependent partitions

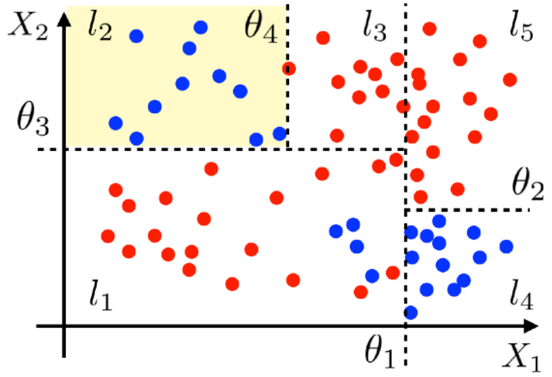
+

Abstracted (coarse grained) featurization of chemical compositions

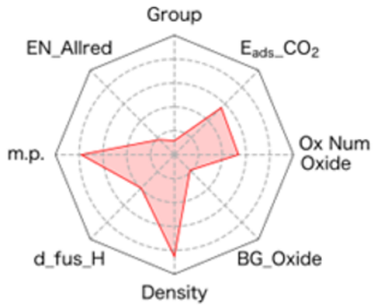
- ExtraTrees regressor
- Gradient Boosted Trees regressor

Input representations by **elemental features**
e.g. “composition-based feature vector (CBFV)”

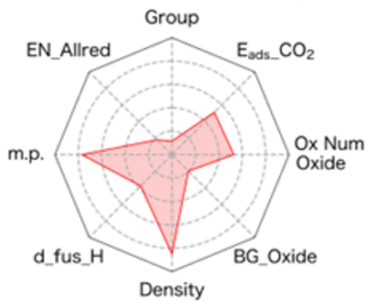




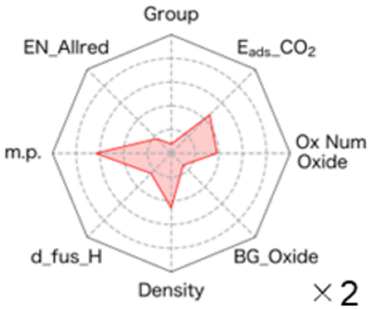
Pt(3)/**Ba(2)-Mo(1)-Tm(1)-Eu(0.5)-Dy(0.5)**/TiO_2



Pt(3)/**Mo(1)-Ba(1)-Tb(1)-Ho(1)-Cs(0.5)**/TiO_2



Pt(3)/**Rb(1)-Ba(1)-Mo(0.6)-Nb(0.2)**/TiO_2



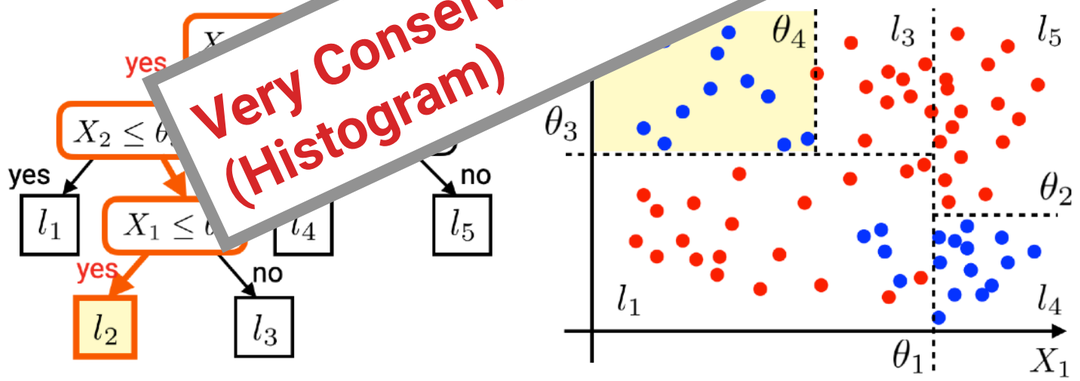
× 2

Our recent research: Method

Decision tree ensembles (with UQ)
i.e. histogram on data-dependent partitions

- ExtraTrees regressor
- Gradient Boosted T… r

Very Conservative Prediction (Histogram)



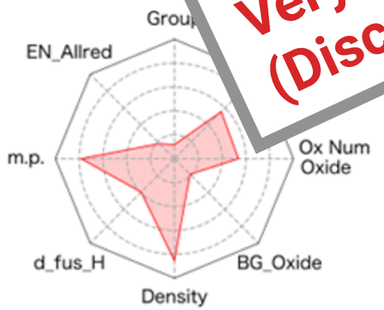
+

Abstracted (coarse grained) featurization of chemical compositions

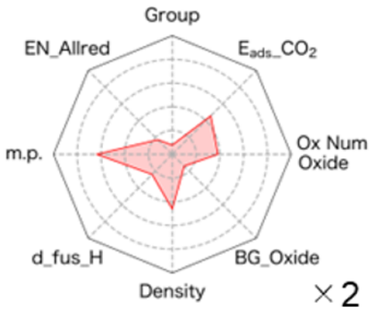
Input representations by **elemen… tures**
e.g. “composition-based f… (CBFV)”

Very Radical Representation (Discard specific details)

Pt(3)/**Ba(2)-Mo(1)-Tm(…**
Eu(0.5)-Dy(0.5)/TiO_2



Pt(3)/**Rb(1)-Ba(1)-Mo(0.6)-Nb(0.2)**/TiO_2



×2

Our recent research: Method

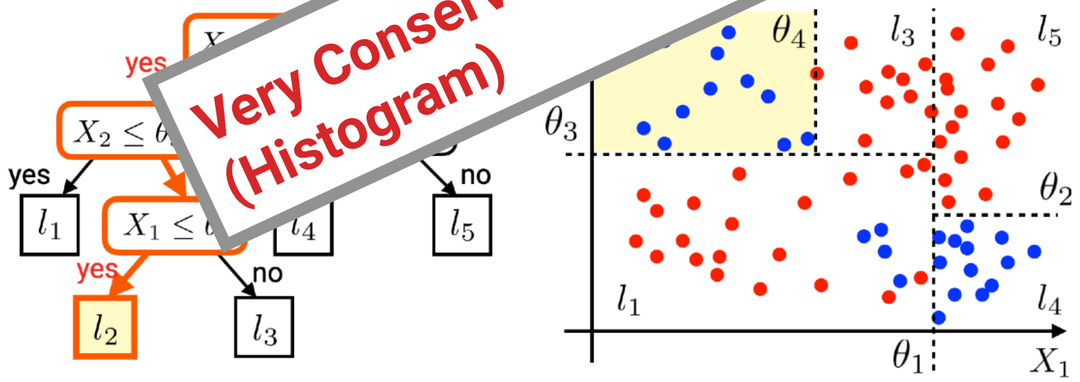
Decision tree ensembles (with UQ)
i.e. histogram on data-dependent partitions

+

Abstracted (coarse grained) featurization of chemical compositions

- ExtraTrees regressor
- Gradient Boosted T… r

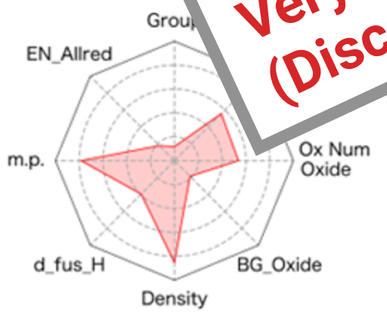
Very Conservative Prediction (Histogram)



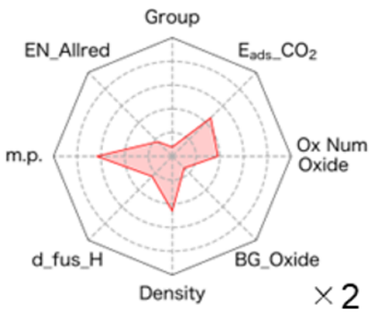
Input representations by **elemen… tures**
e.g. “composition-based f… (CBFV)”

Very Radical Representation (Discard specific details)

Pt(3)/**Ba(2)-Mo(1)-Tm(…**
Eu(0.5)-Dy(0.5)/TiO_2



Pt(3)/**Rb(1)-Ba(1)-Mo(0.6)-Nb(0.2)**/TiO_2



× 2

This talk will hopefully explain why we go for such a **standard** method choice
(even though I’m a ML researcher doing research also in GNNs and Transformers)

My prologue: Materials informatics?

At first I had an optimistic image of the unfamiliar field of "Materials Informatics"...
(after I worked in machine learning for bioinformatics for 10 years)

Step 1

We give all possible types of available data into ML

Step 2

ML becomes smarter than standard experts

Step 3

ML suggests more and more promising materials

Takeaways

Three lessons learned as I experienced **this illusion** being shattered…

Takeaways

Three lessons learned as I experienced **this illusion** being shattered…

1. The goals of ML and ‘materials/chemical science’ are **fundamentally different**. What we need here is not ML but a much harder problem of ‘machine discovery.’

Takeaways

Three lessons learned as I experienced **this illusion** being shattered…

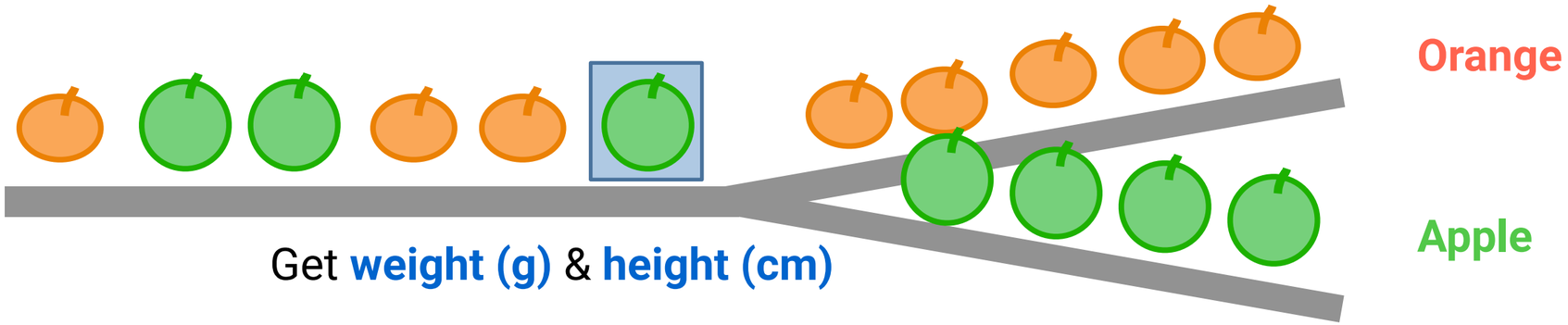
1. The goals of ML and ‘materials/chemical science’ are **fundamentally different**. What we need here is not ML but a much harder problem of ‘machine discovery.’
2. If we go for a hypothesis-free + off-the-shelf solution, exploration by **decision tree ensembles**, combined with UQ and abstracted (coarse grained) feature representations, will give a very strong baseline.

Takeaways

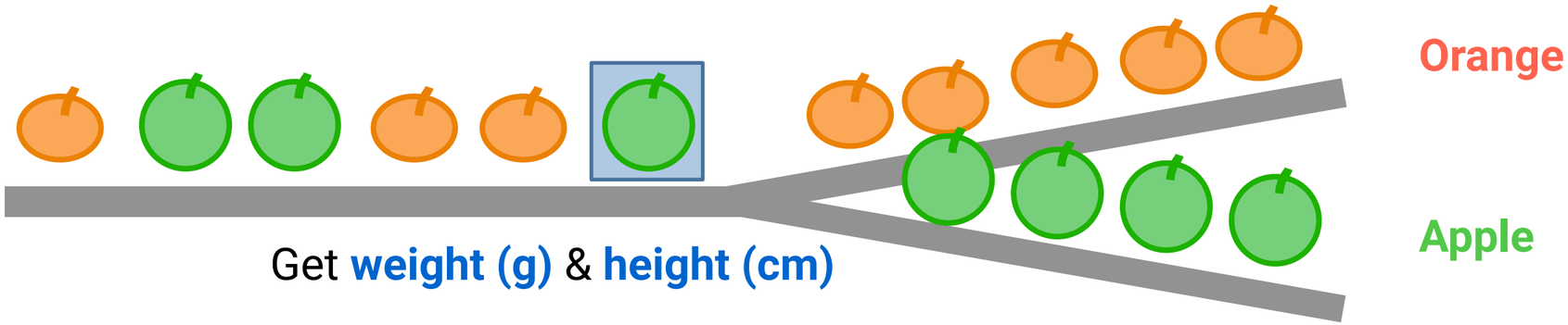
Three lessons learned as I experienced **this illusion** being shattered…

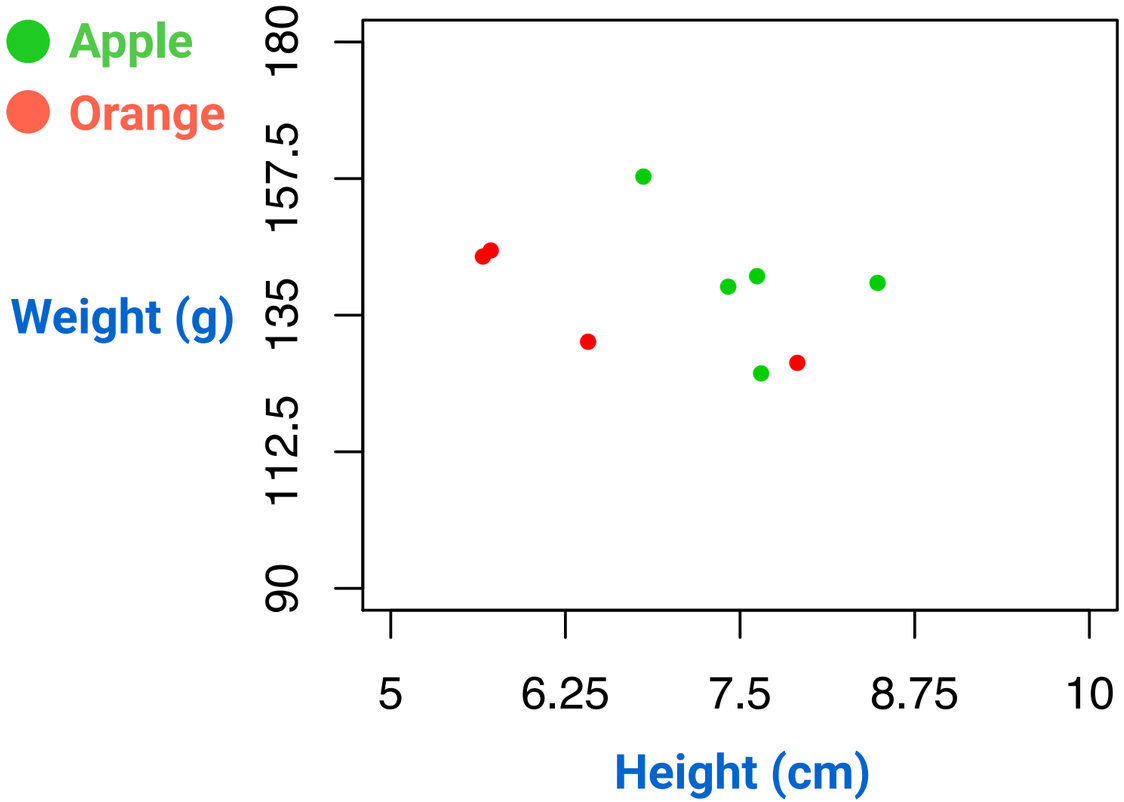
1. The goals of ML and ‘materials/chemical science’ are **fundamentally different**. What we need here is not ML but a much harder problem of ‘machine discovery.’
2. If we go for a hypothesis-free + off-the-shelf solution, exploration by **decision tree ensembles**, combined with UQ and abstracted (coarse grained) feature representations, will give a very strong baseline.
3. If we want more than that, we **can’t be hypothesis free**. Any strategies to narrow down the scope as well as domain expertise really matters.

Machine Learning converts data into predictions

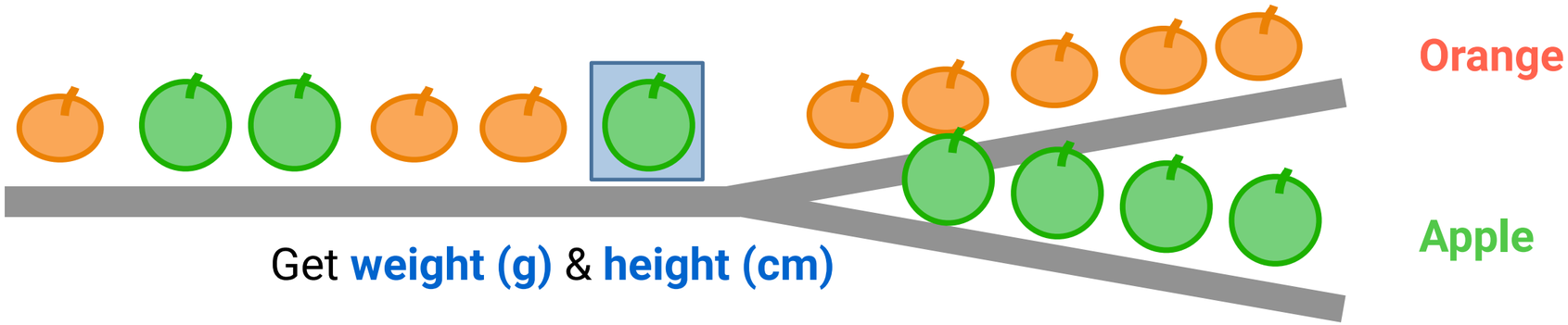


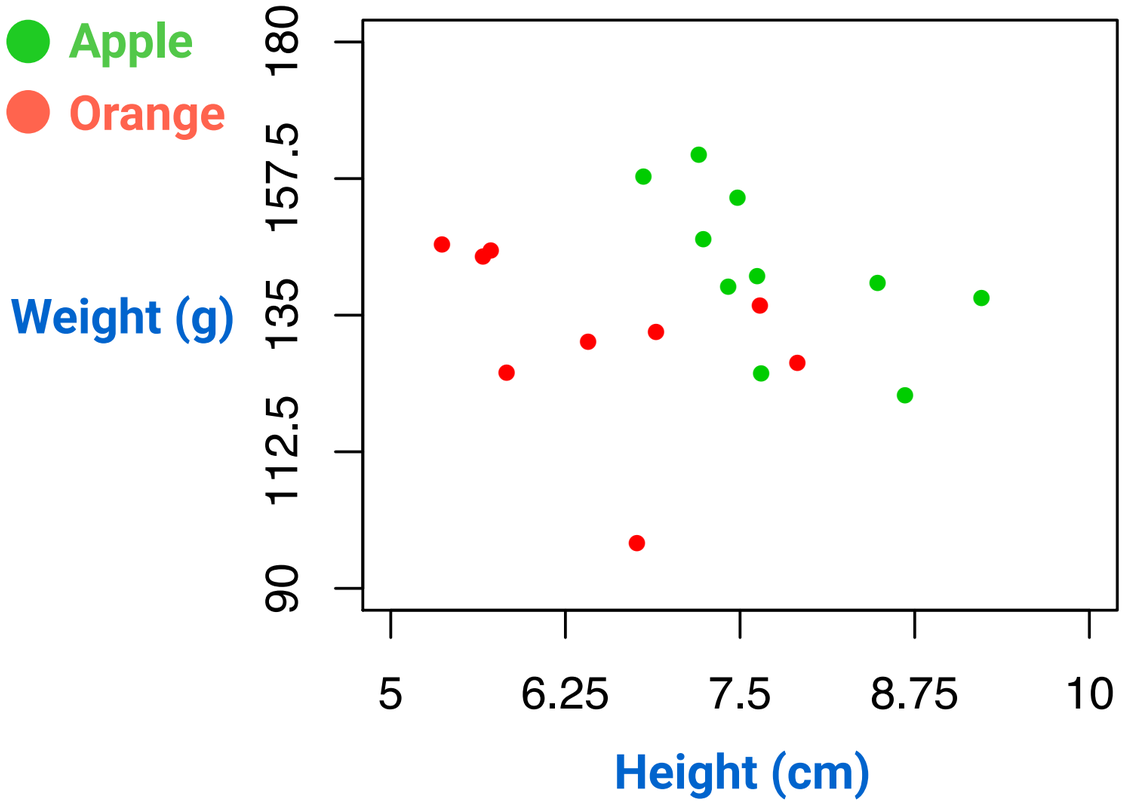
Machine Learning converts data into predictions



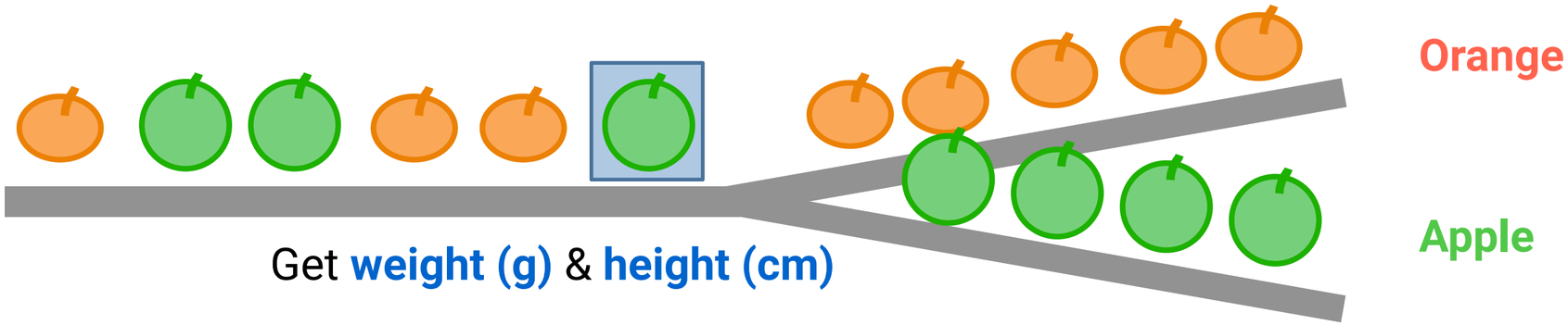


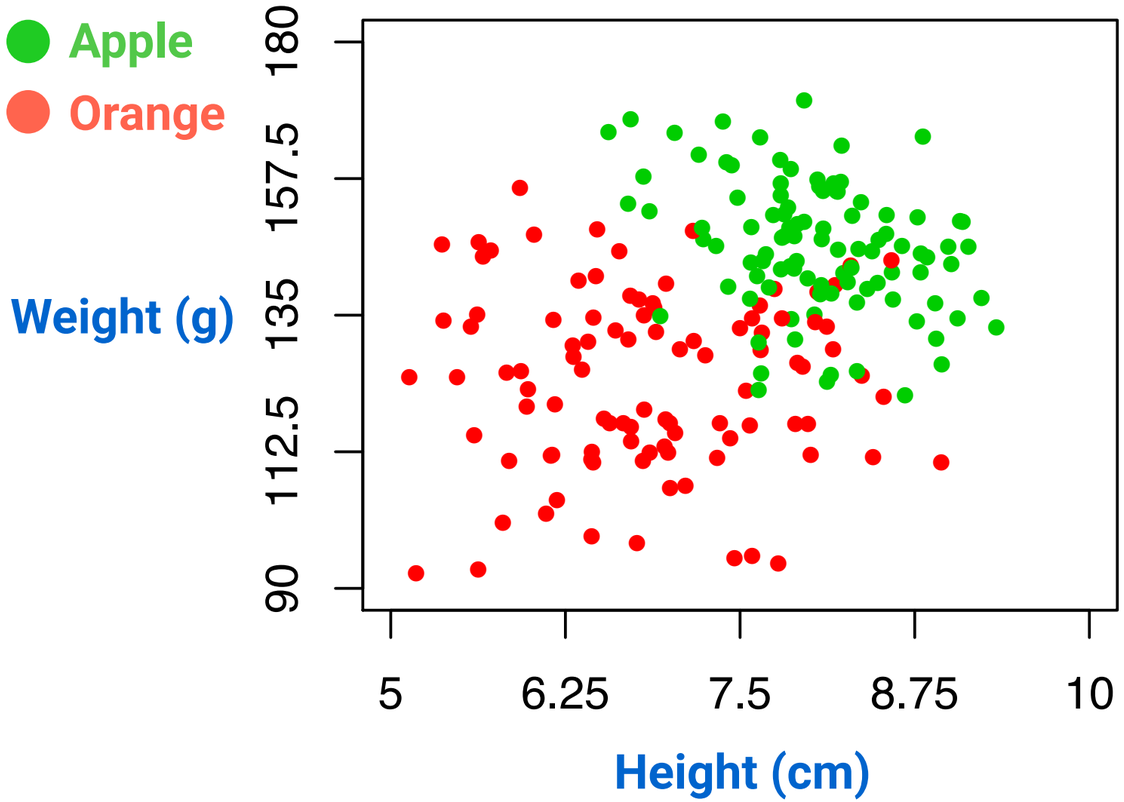
Machine Learning converts data into predictions



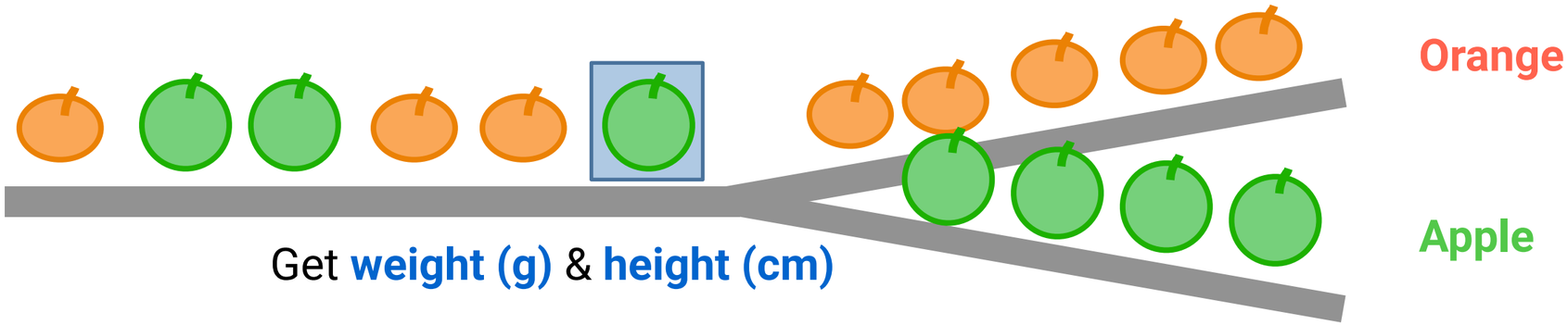


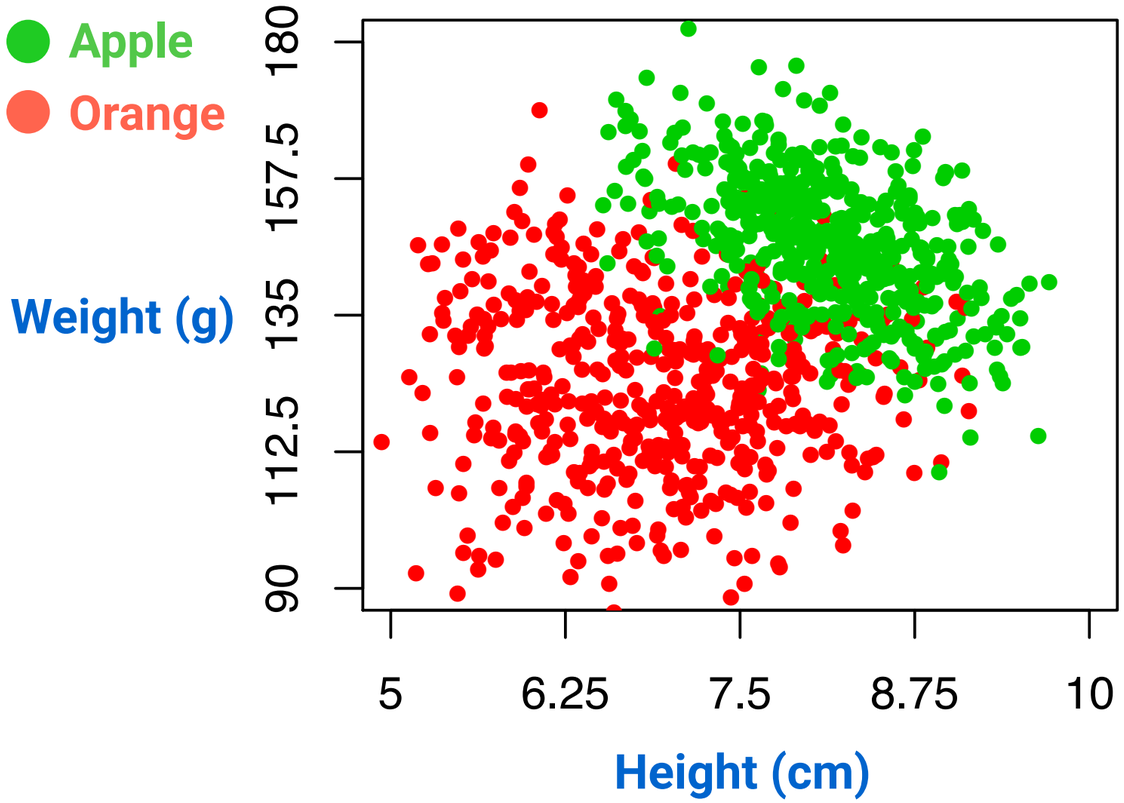
Machine Learning converts data into predictions



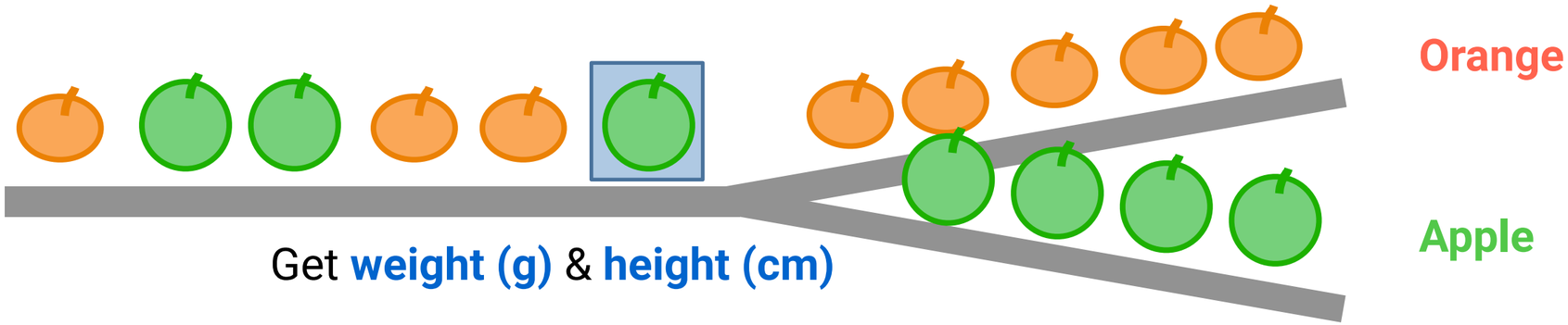


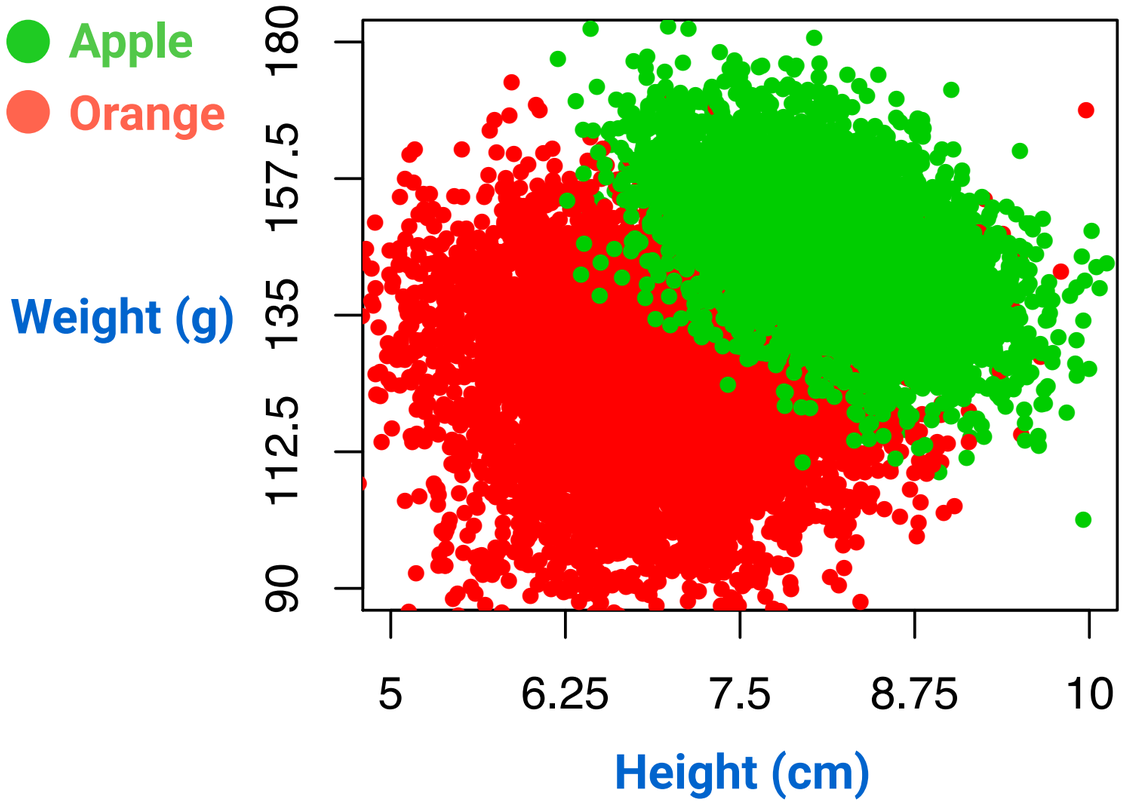
Machine Learning converts data into predictions



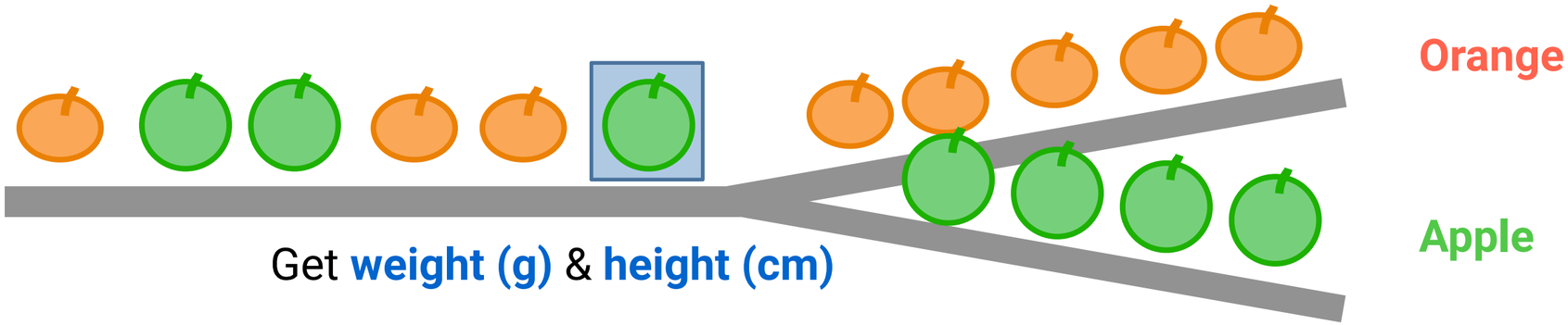


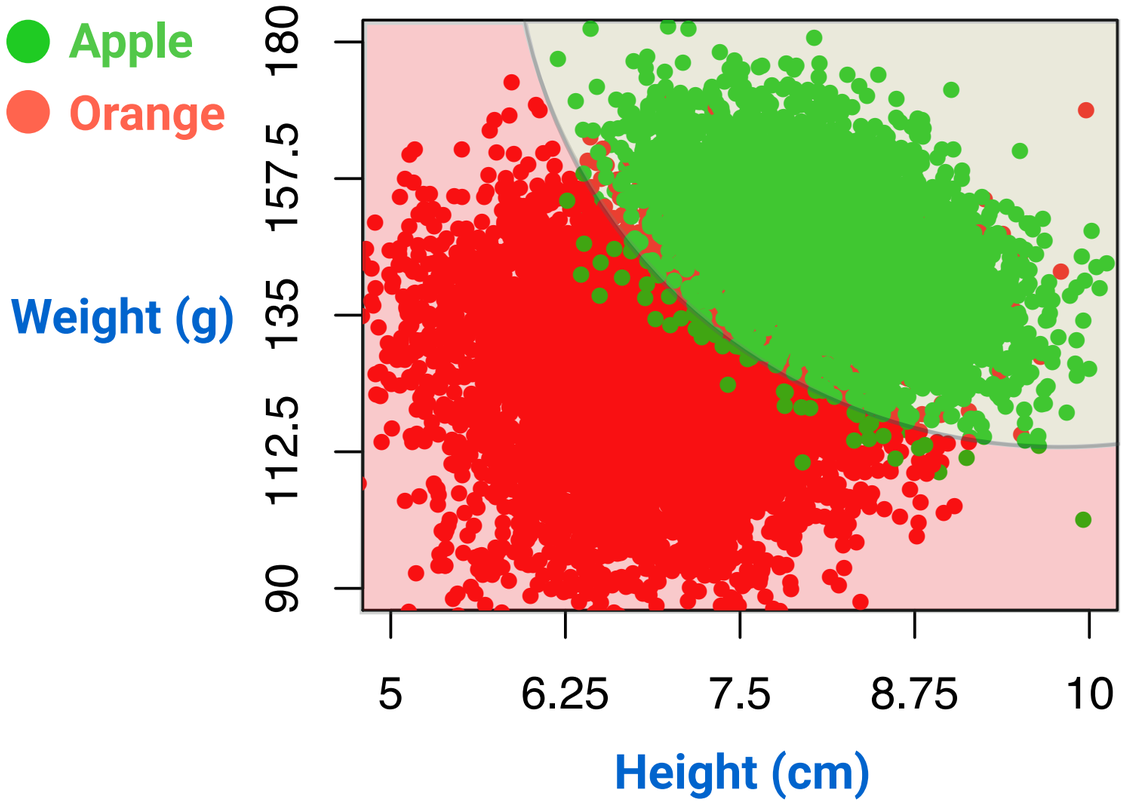
Machine Learning converts data into predictions



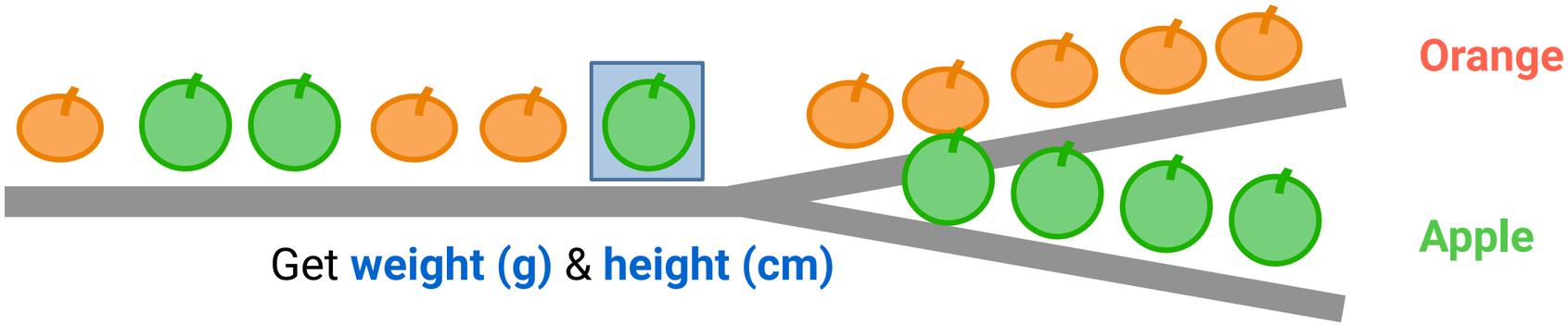


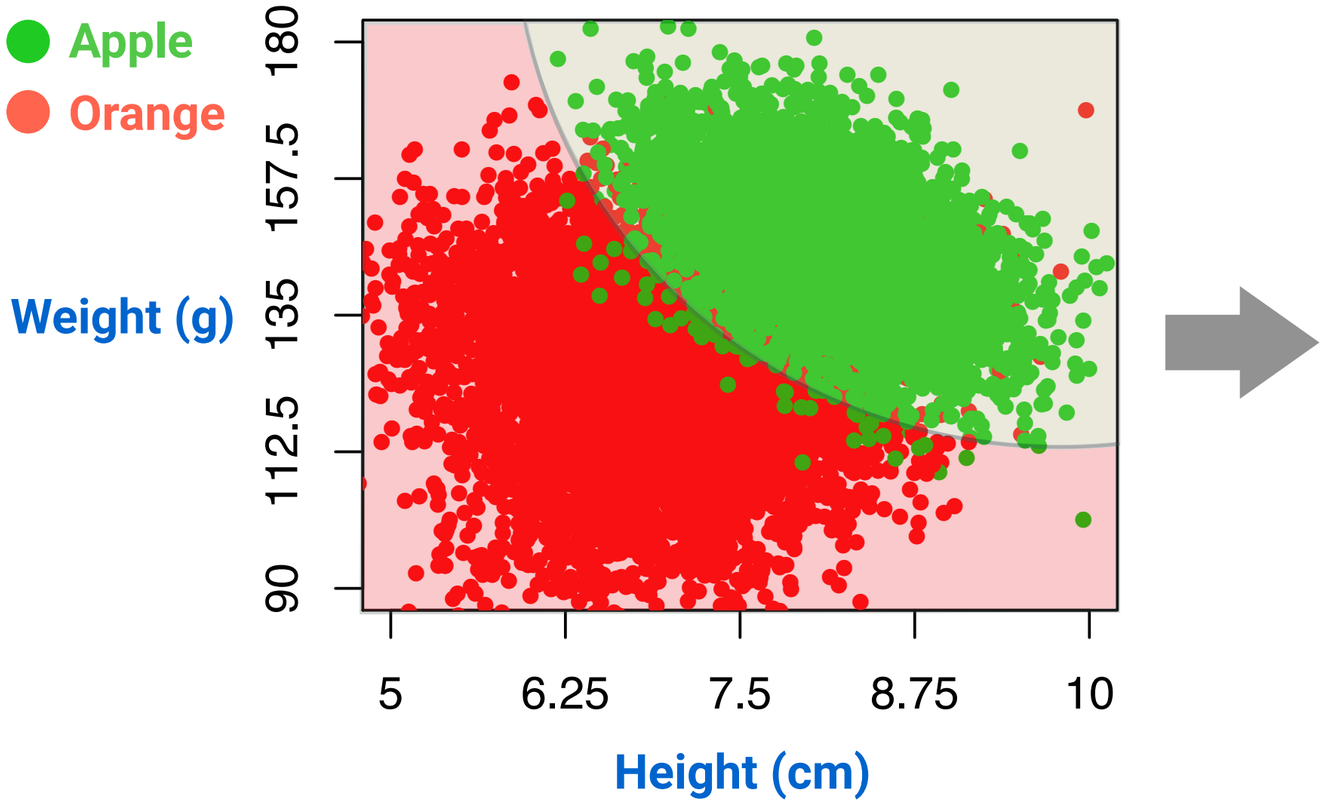
Machine Learning converts data into predictions



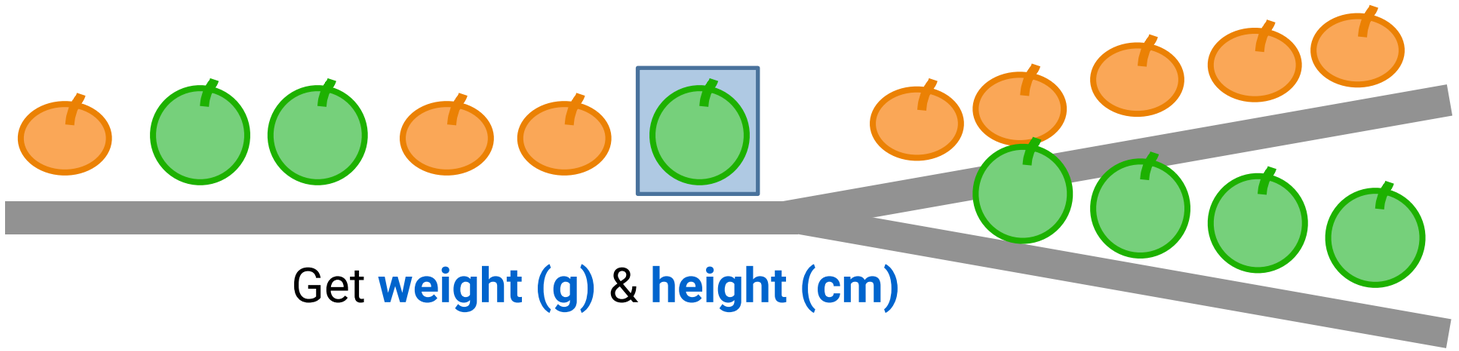


Machine Learning converts data into predictions

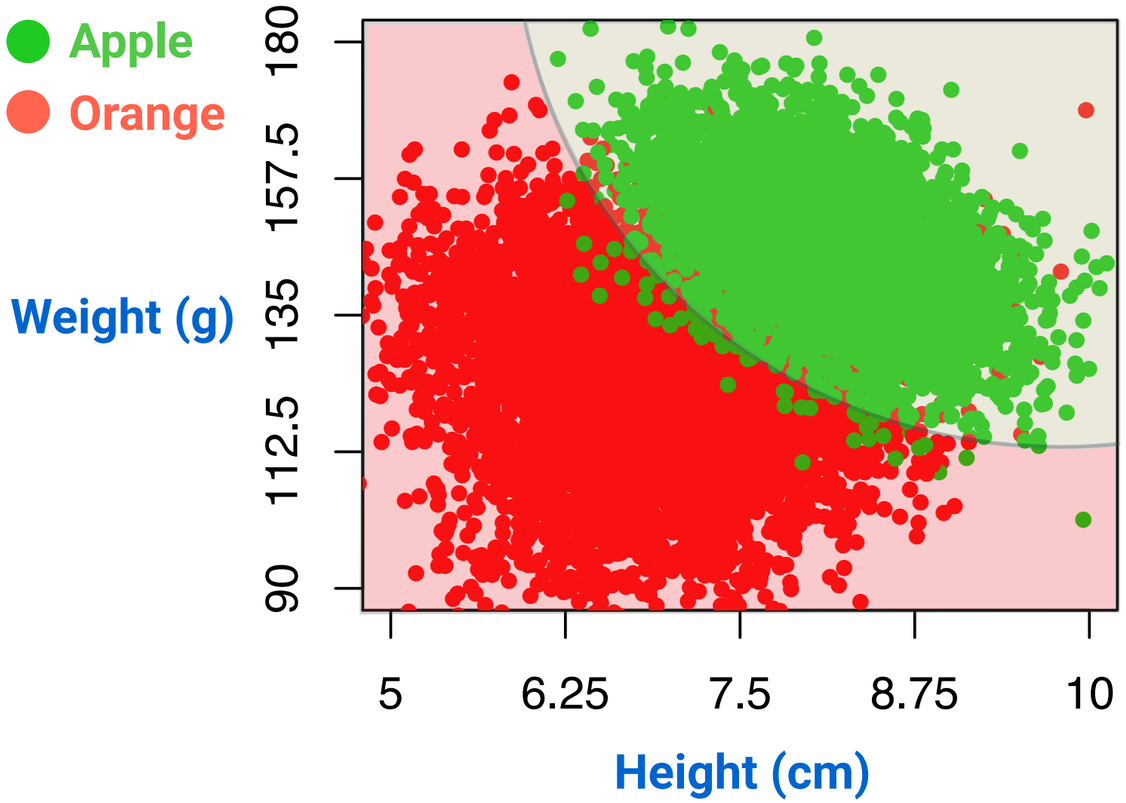




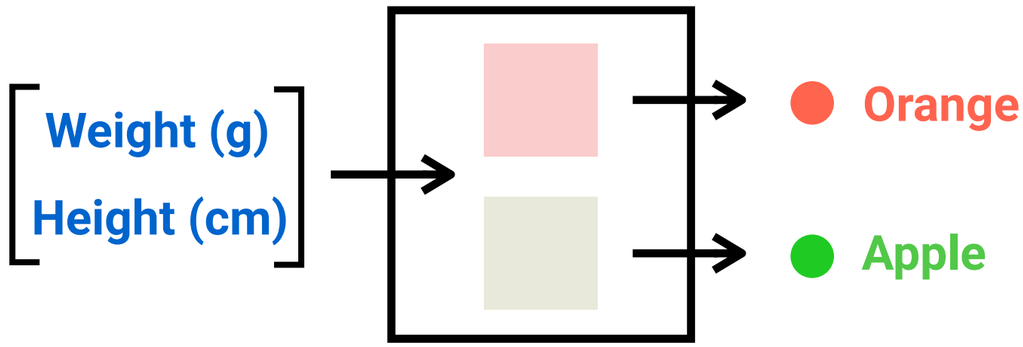
Machine Learning converts data into predictions



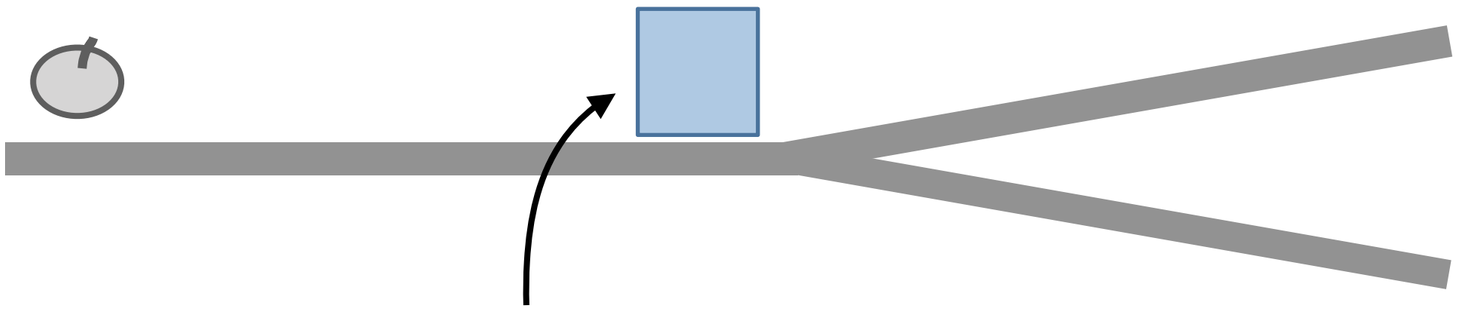
Get **weight (g)** & **height (cm)**



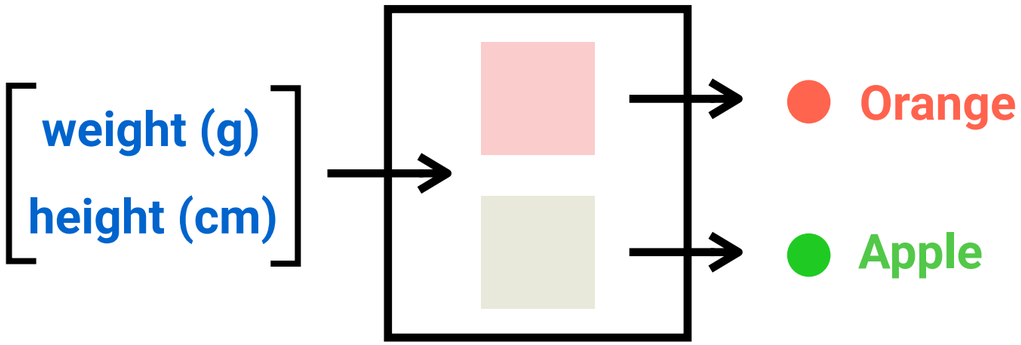
Computer program for prediction

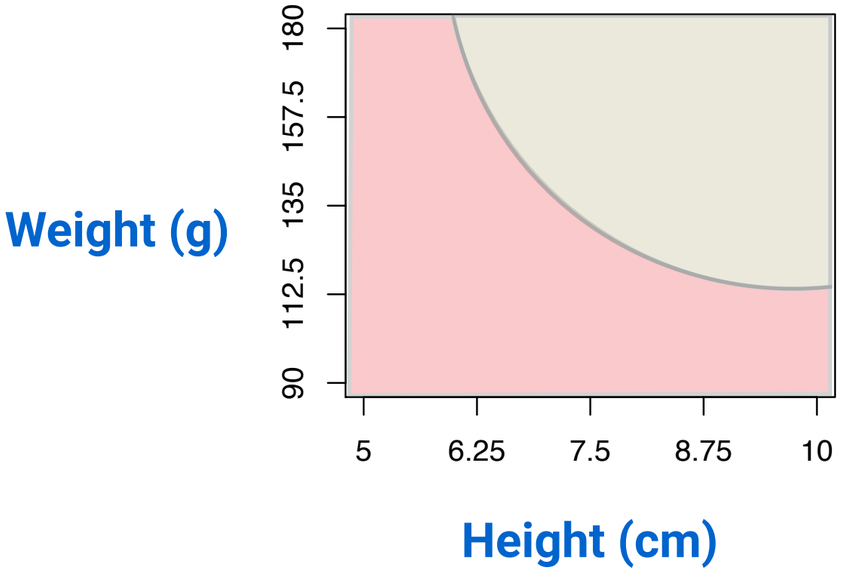


Machine Learning converts data into predictions



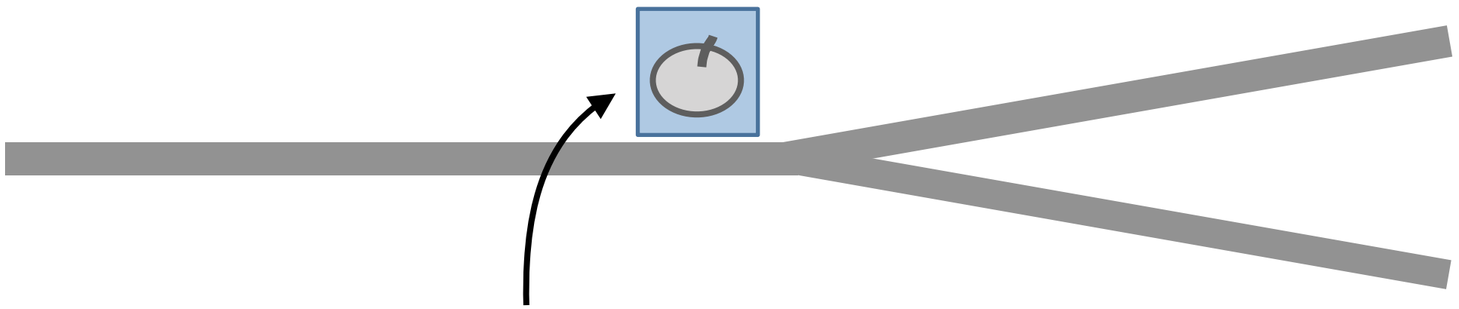
The computer program we got from training



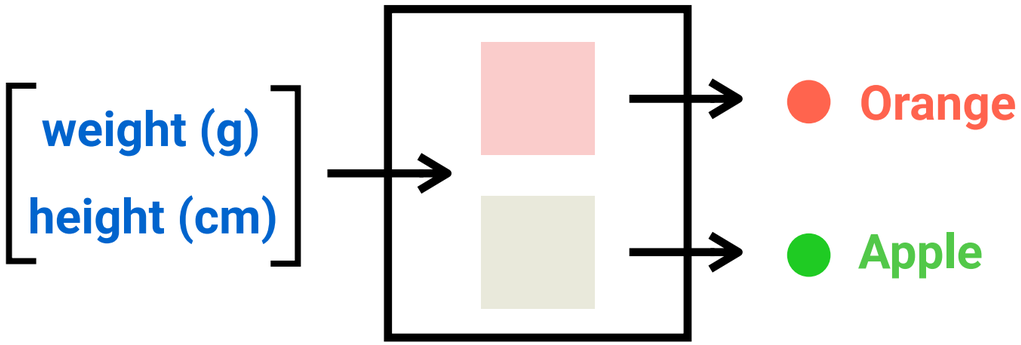


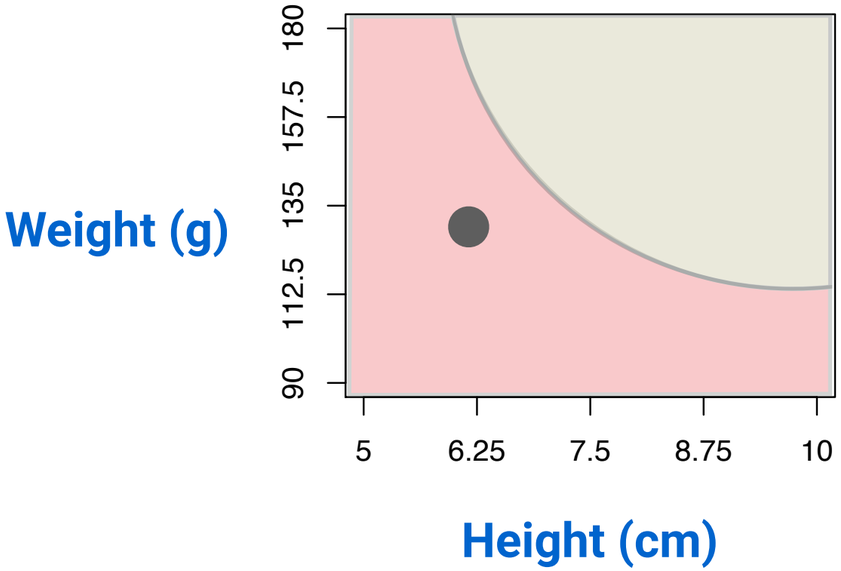
This program can make prediction
for **different examples** than the ones shown in training!

Machine Learning converts data into predictions



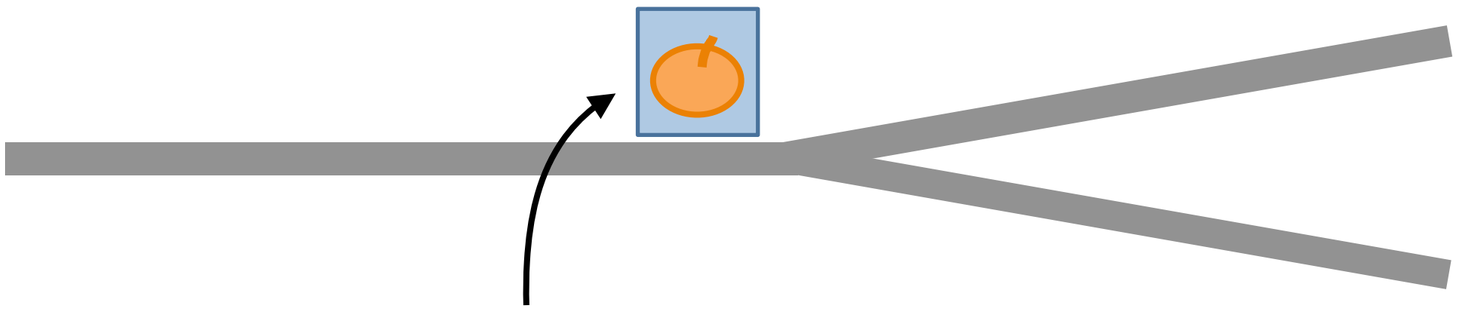
The computer program we got from training



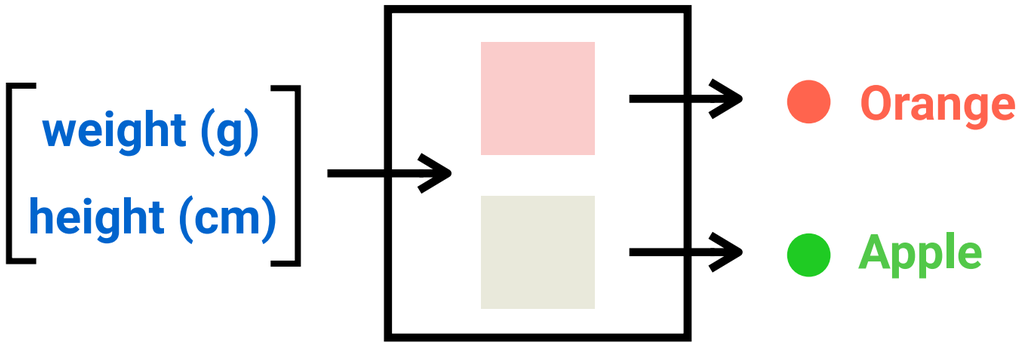


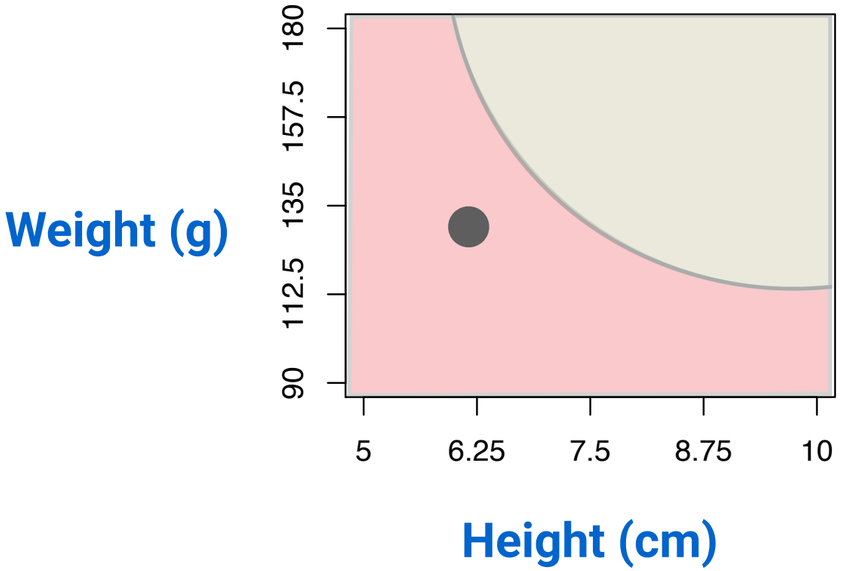
This program can make prediction
for **different examples** than the ones shown in training!

Machine Learning converts data into predictions



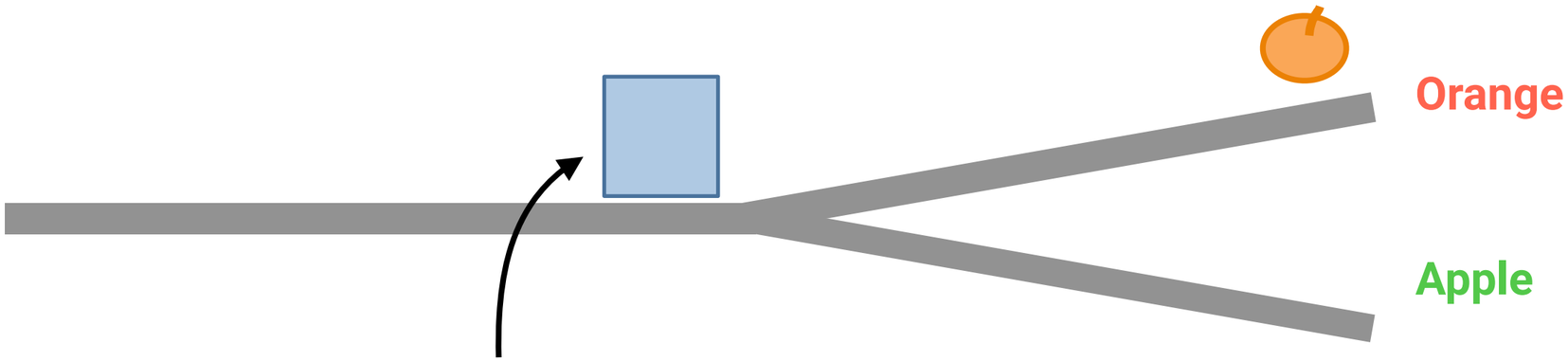
The computer program we got from training



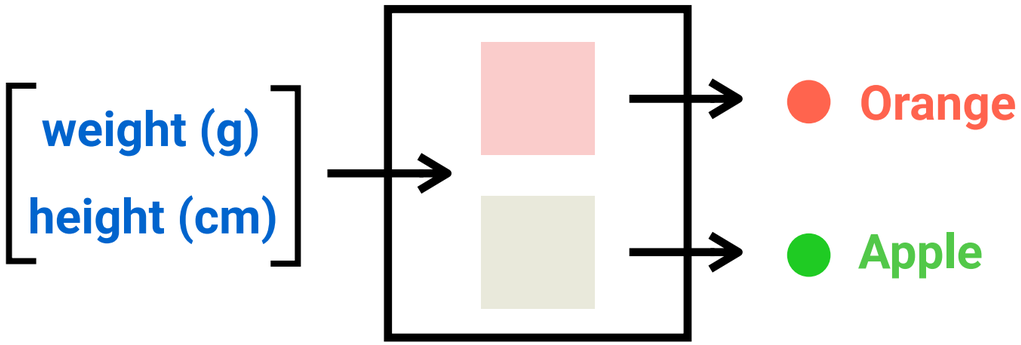


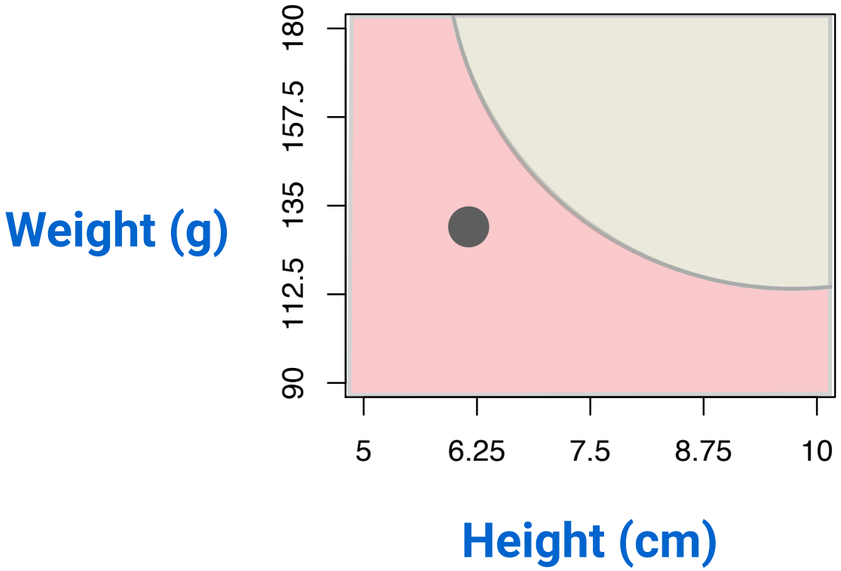
This program can make prediction
for **different examples** than the ones shown in training!

Machine Learning converts data into predictions



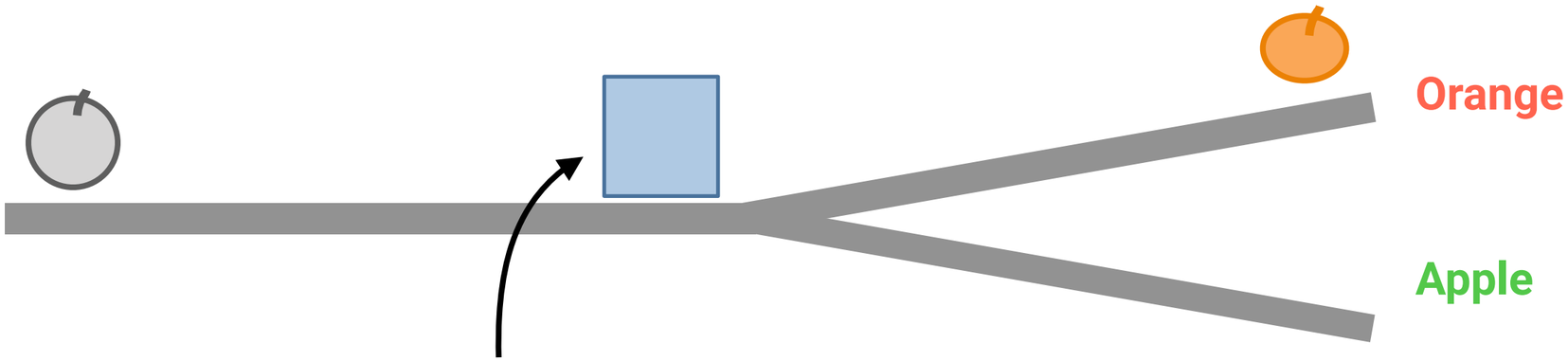
The computer program we got from training



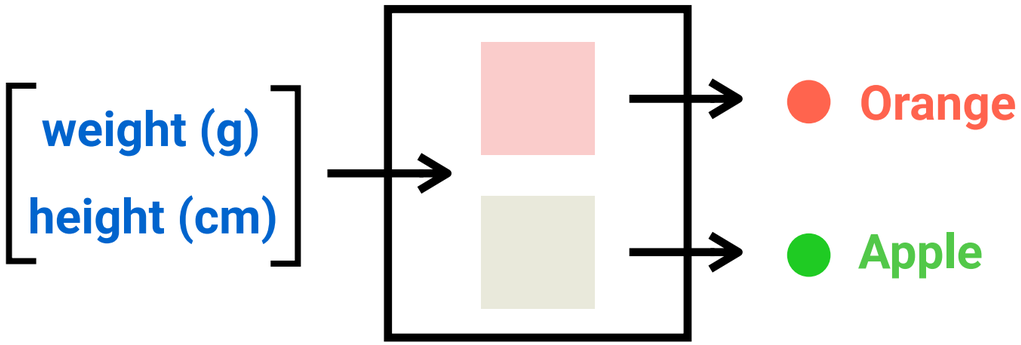


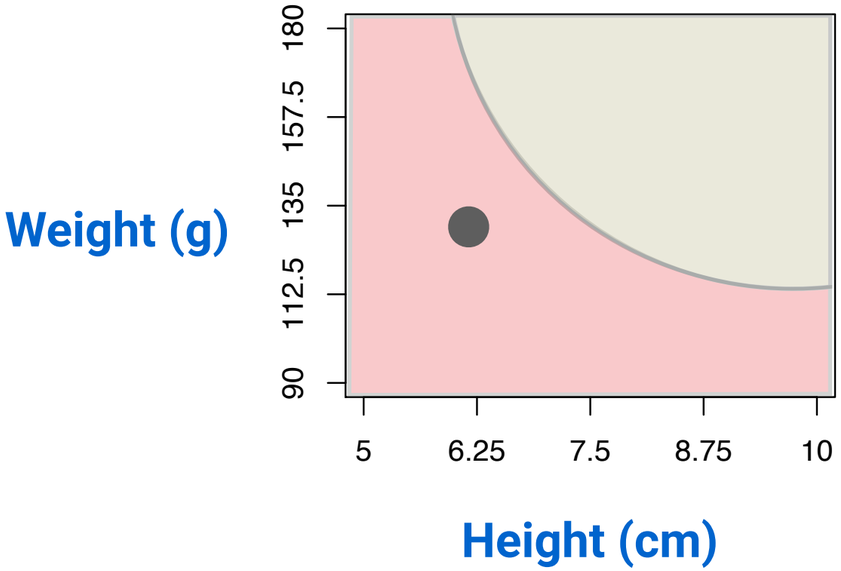
This program can make prediction
for **different examples** than the ones shown in training!

Machine Learning converts data into predictions



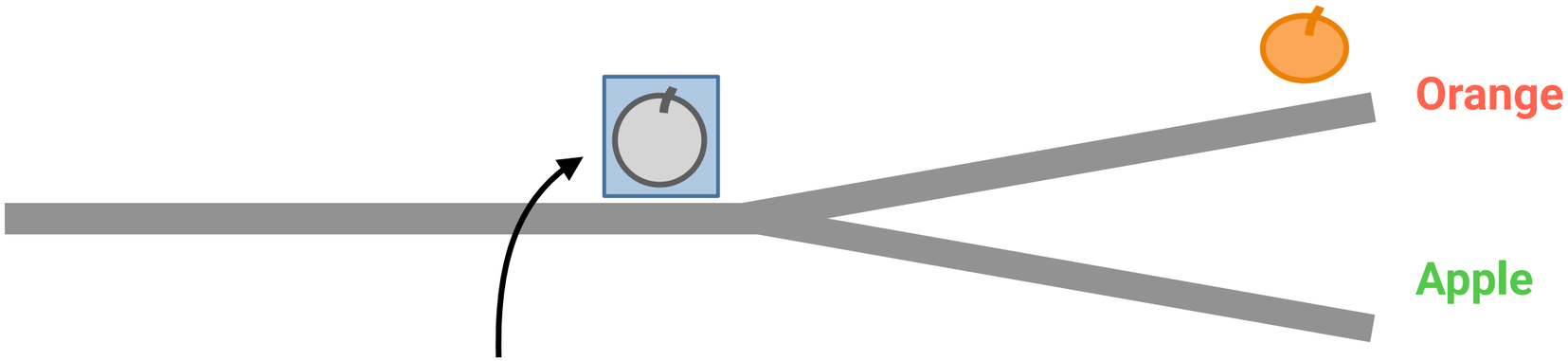
The computer program we got from training



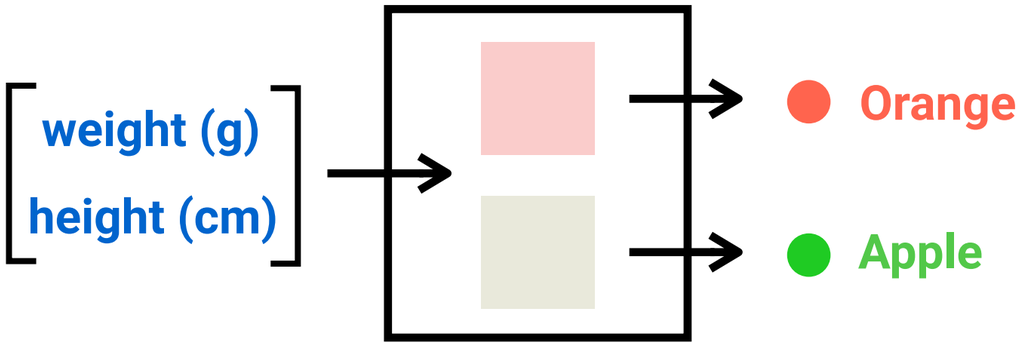


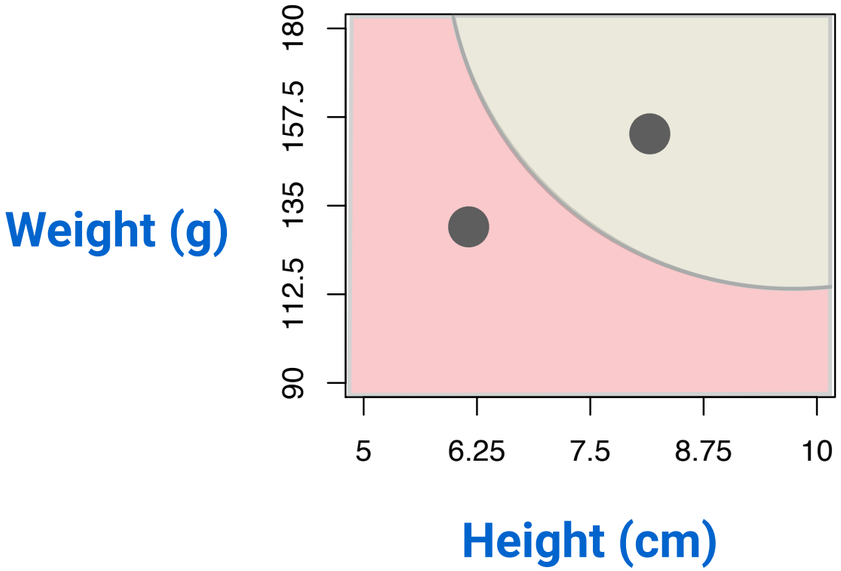
This program can make prediction
for **different examples** than the ones shown in training!

Machine Learning converts data into predictions



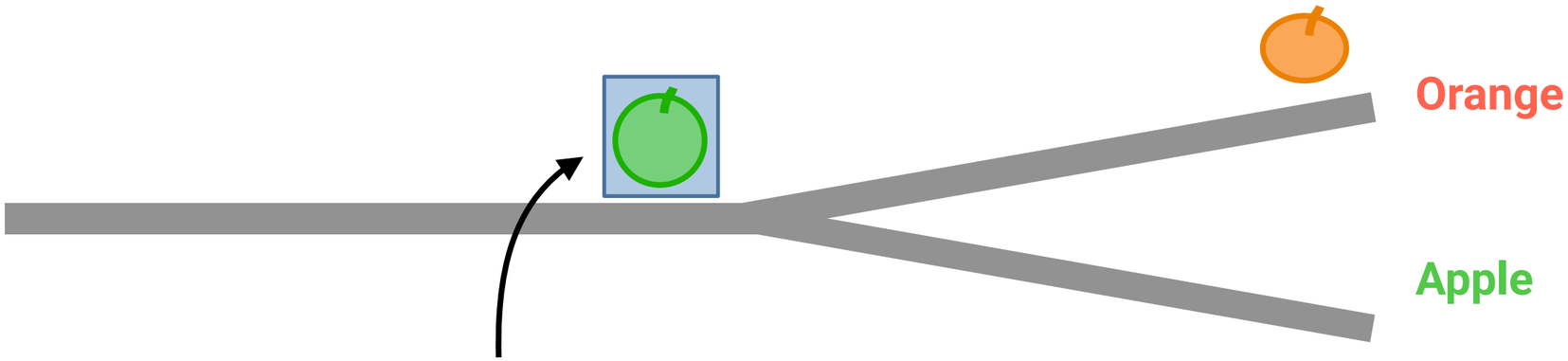
The computer program we got from training



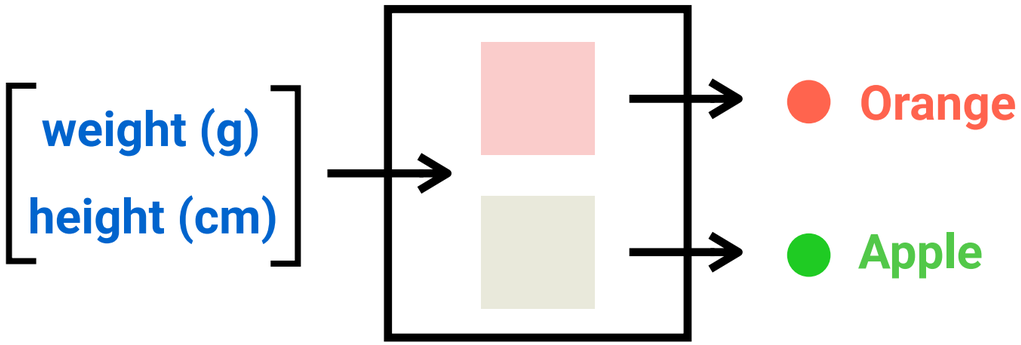


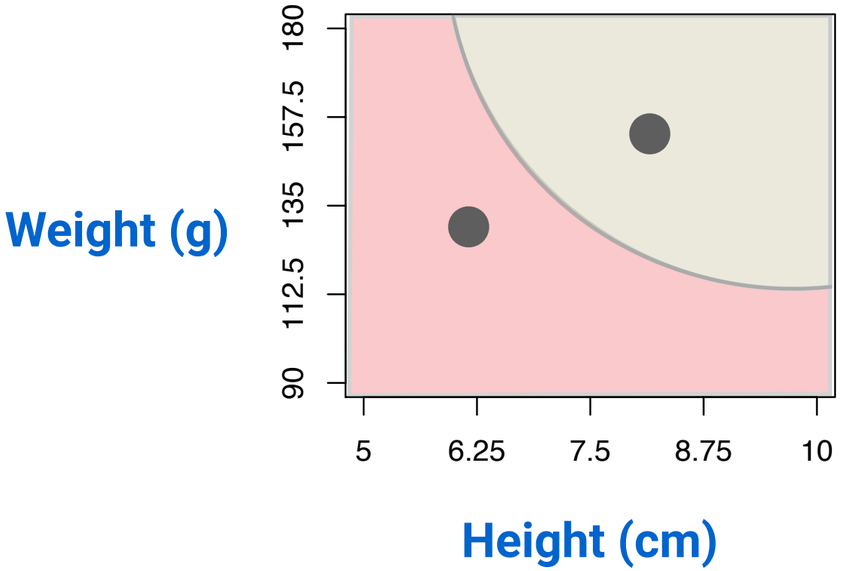
This program can make prediction
for **different examples** than the ones shown in training!

Machine Learning converts data into predictions



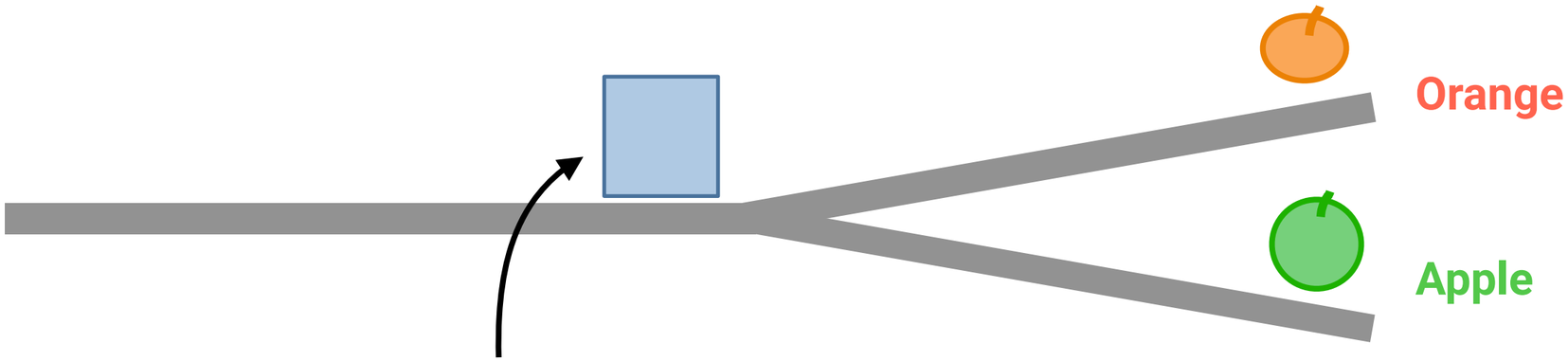
The computer program we got from training



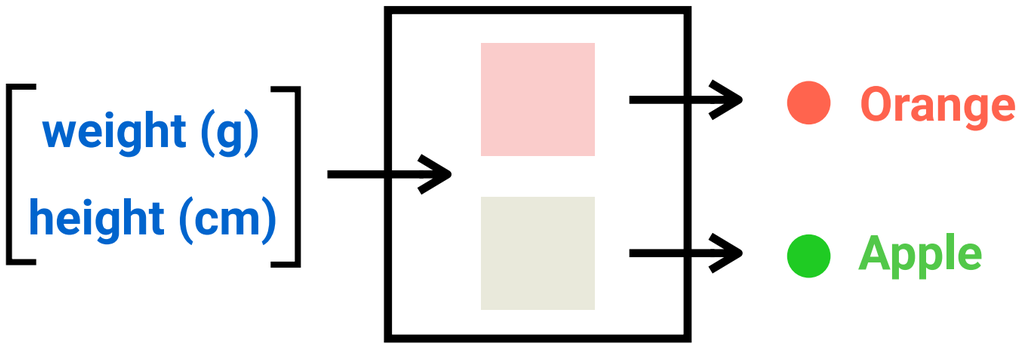


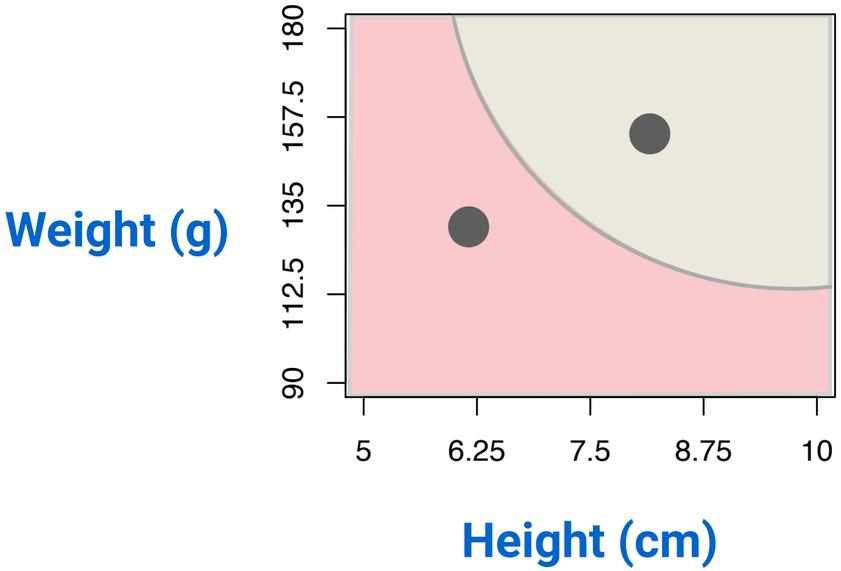
This program can make prediction
for **different examples** than the ones shown in training!

Machine Learning converts data into predictions



The computer program we got from training





This program can make prediction
for **different examples** than the ones shown in training!

ML = a new (lazy) way of computer programming!

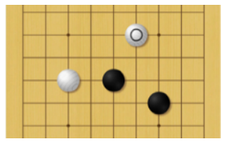
Synthesize a program (input-output function) just by giving input-output examples!

→ ***Object Recognition*** →

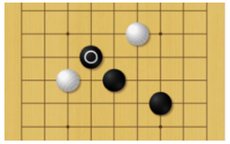
→ ***Speech Recognition*** → “ありがとう”

J’aime la musique → ***Machine Translation*** → I love music

Simple is better than → ***Language Model*** → Simple is better than complex



→ ***Game Play*** →



ML = a new (lazy) way of computer programming!

Synthesize a program (input-output function) just by giving input-output examples!

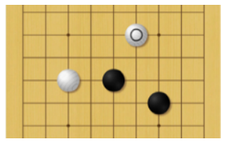
N.B. This does not mean that we also “understood” the input-output relationship.

→ ***Object Recognition*** →

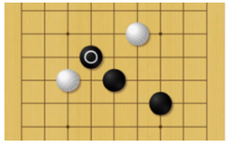
→ ***Speech Recognition*** → “ありがとう”

J’aime la musique → ***Machine Translation*** → **I love music**

Simple is better than → ***Language Model*** → **Simple is better than complex**

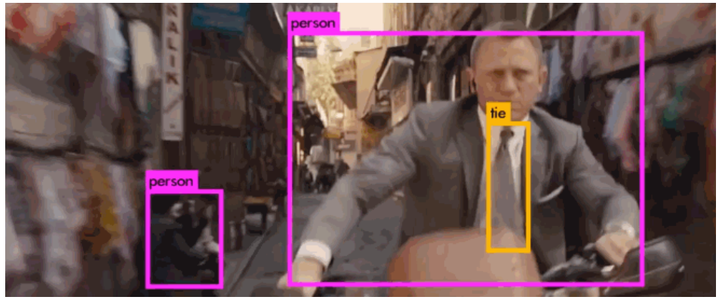


→ ***Game Play*** →



Very powerful technology if we use it in the right place

Image Recognition



Translation

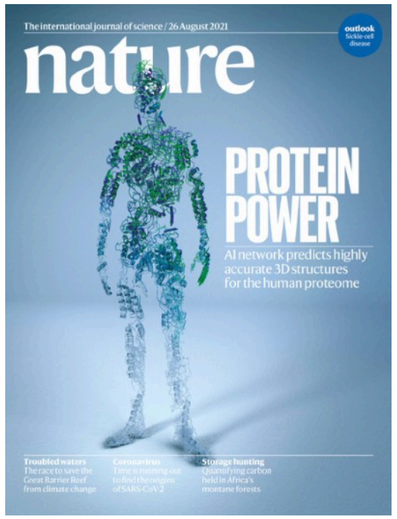
DeepL

Image/Video Conversion

“Deep Fake”

AlphaGo

AlphaFold2



AlphaTensor

OpenAI
ChatGPT

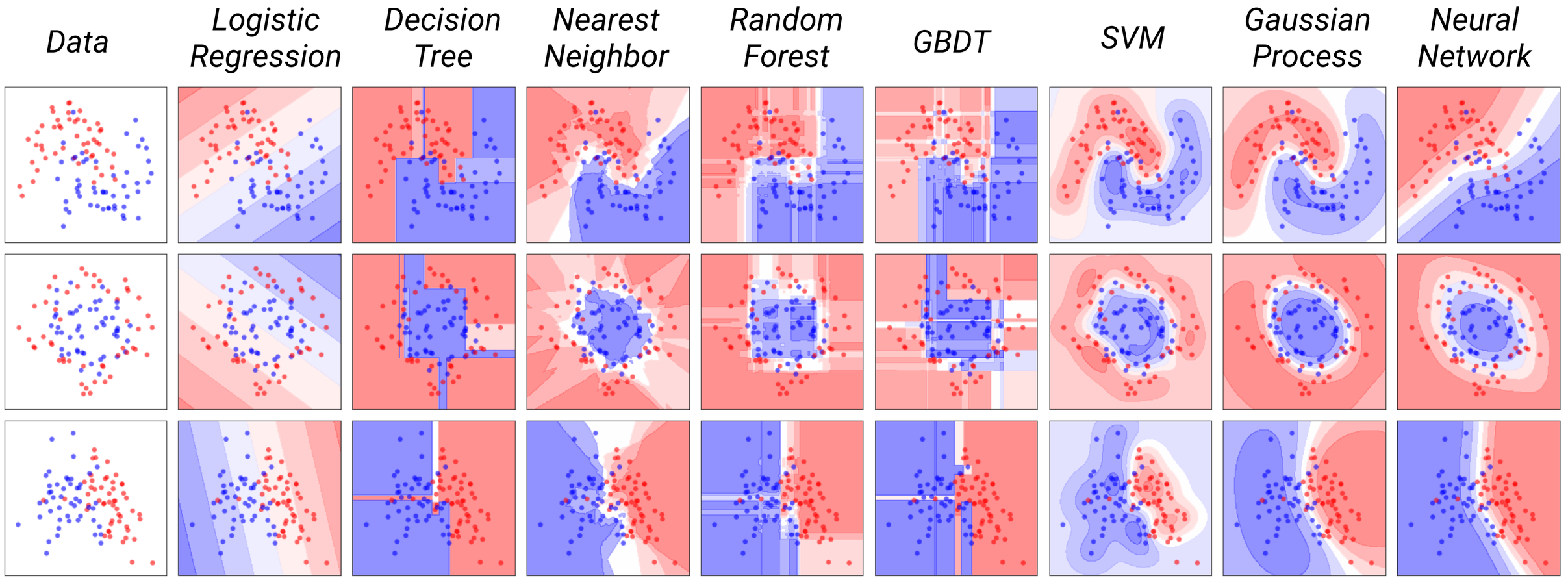
OpenAI
DALL·E 2

Midjourney

Stable Diffusion

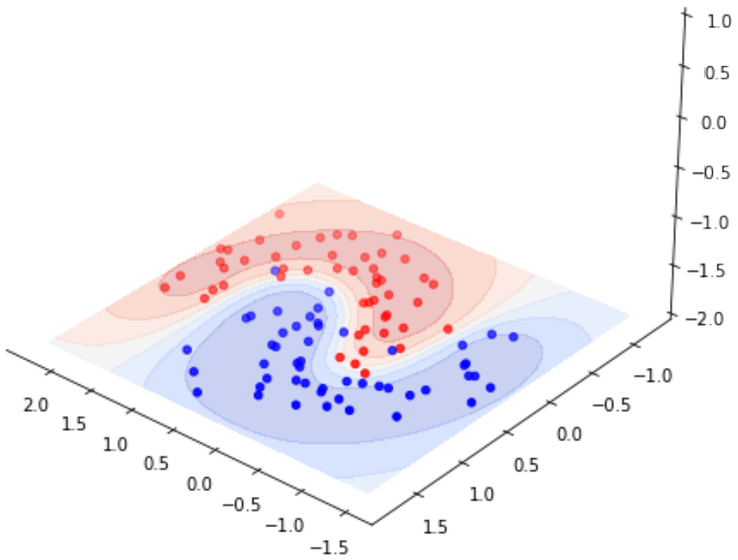
ML models are not unique even for the same dataset

There are as many ML models as there are ways to draw the boundary…



But all the inner workings are just function fitting to data

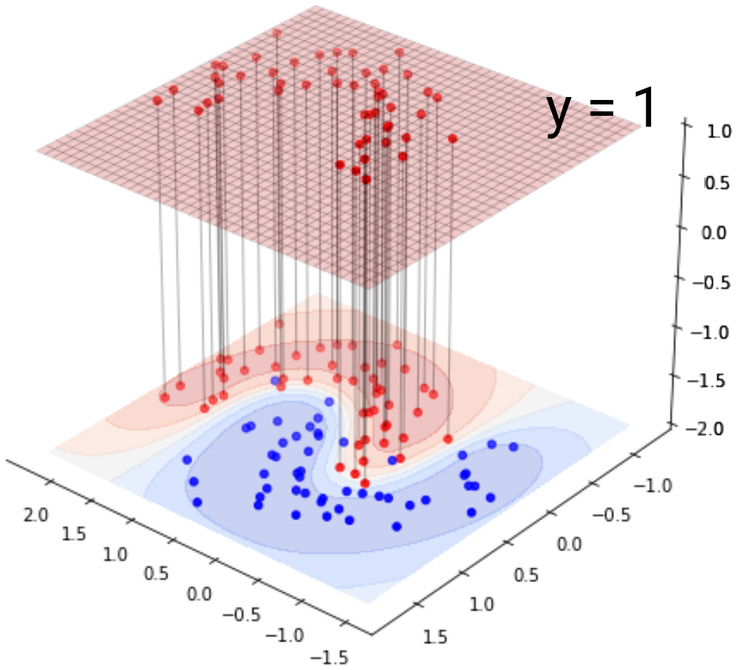
Every model just tries to fit a different type of functions to given data



Classification Setup

But all the inner workings are just function fitting to data

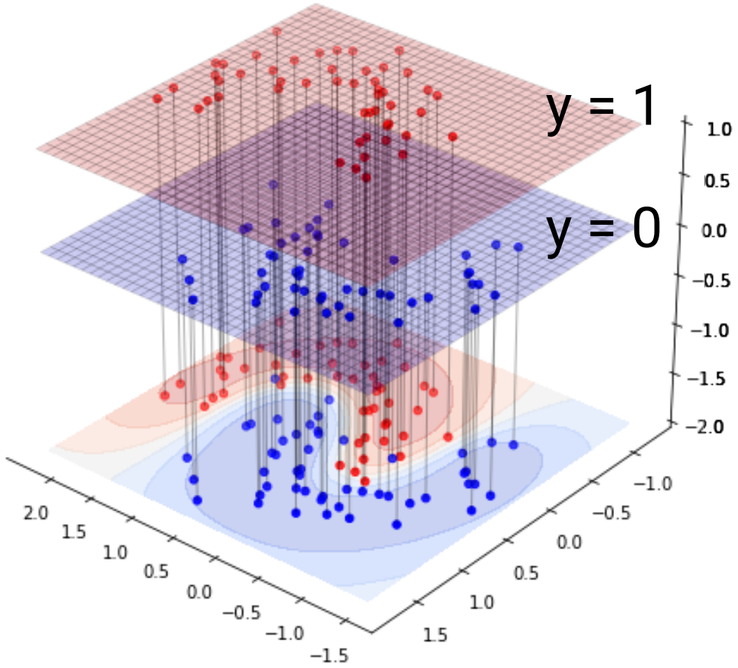
Every model just tries to fit a different type of functions to given data



Classification Setup

But all the inner workings are just function fitting to data

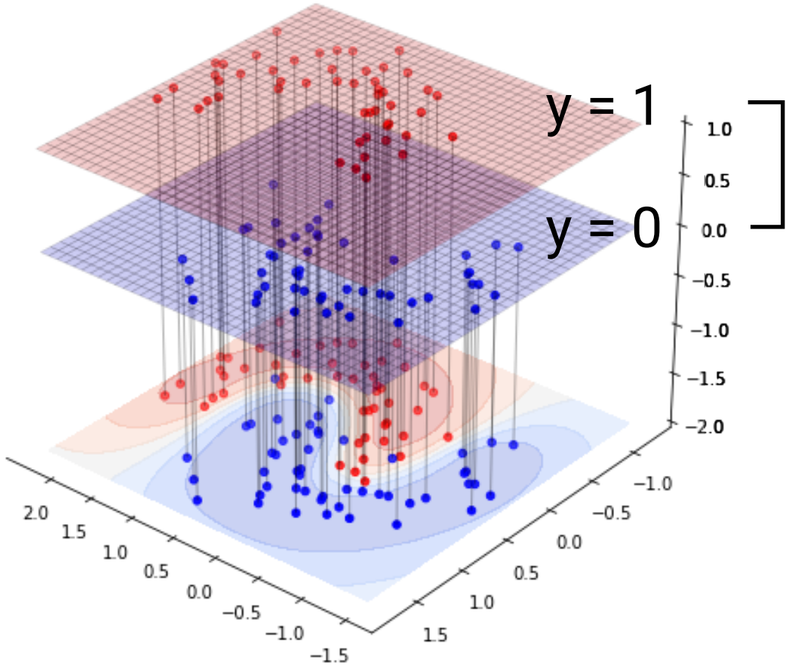
Every model just tries to fit a different type of functions to given data



Classification Setup

But all the inner workings are just function fitting to data

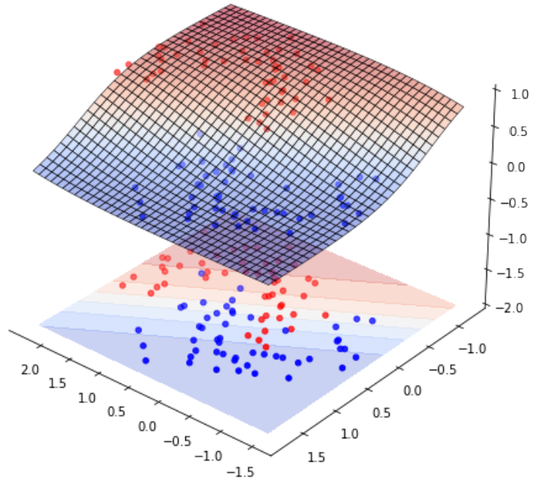
Every model just tries to fit a different type of functions to given data



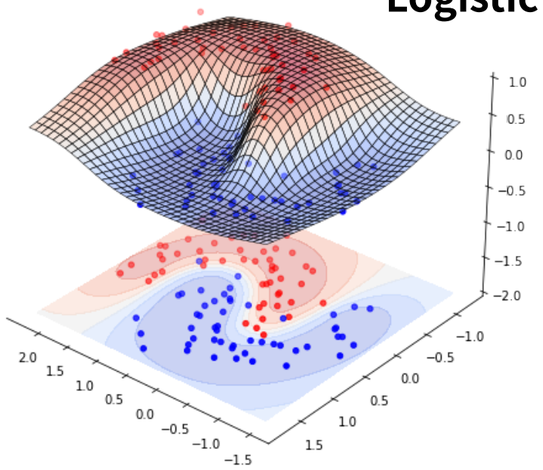
Classification Setup

Class probability

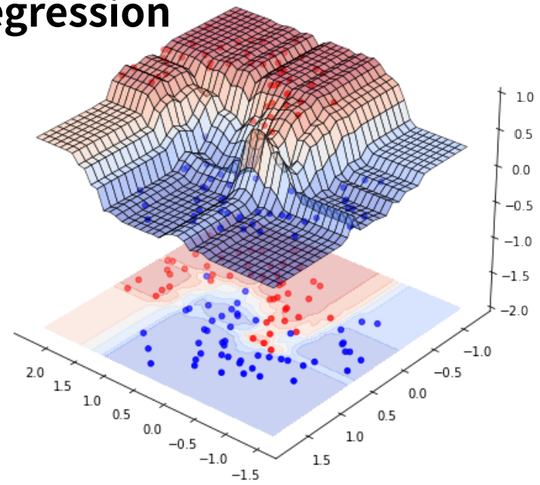
P(class=red)



Logistic Regression



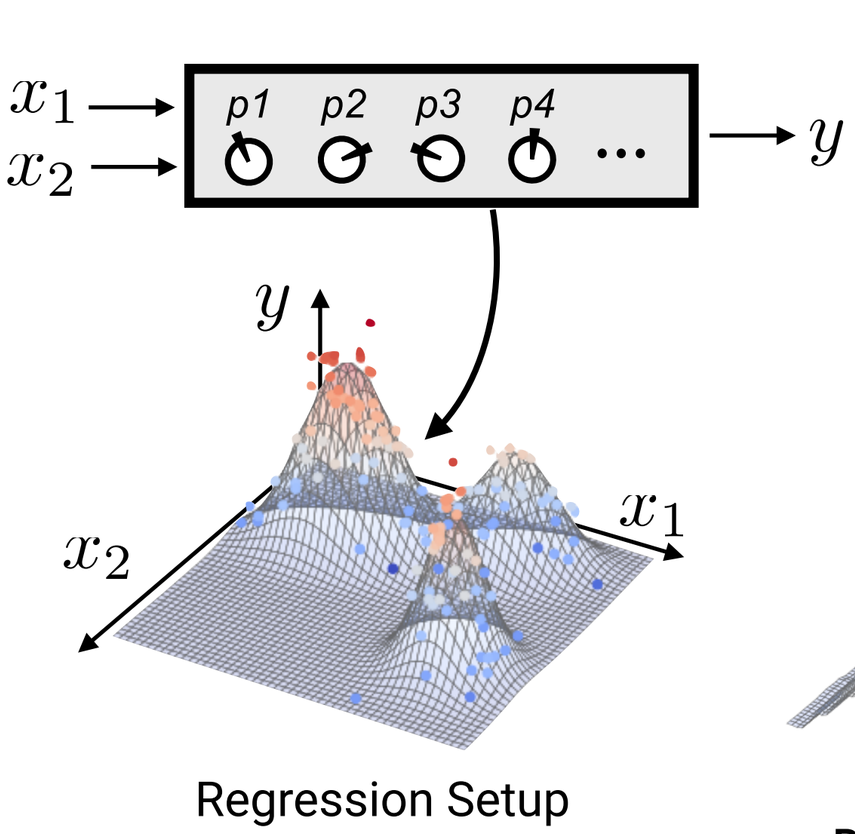
Gaussian Process



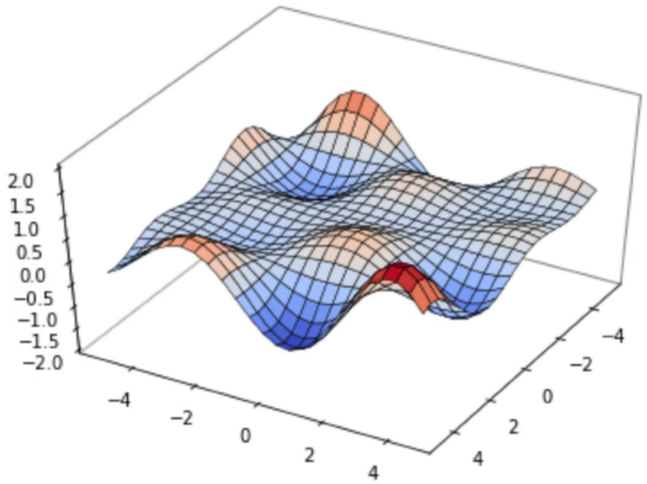
Random Forest

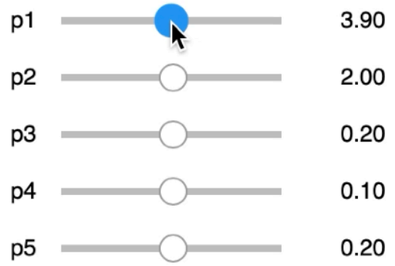
By just tweaking numeric values for model parameters

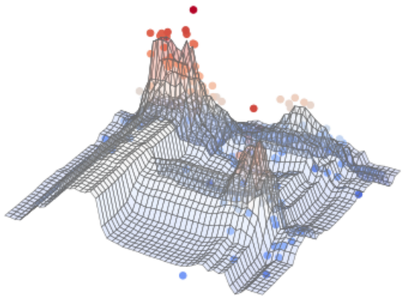
This fitting is done by optimally adjusting the model parameter values



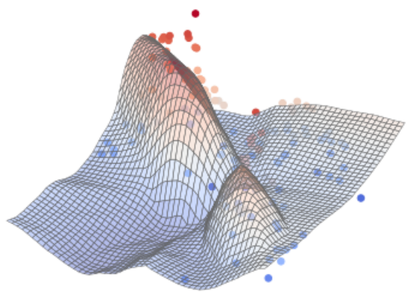
Regression Setup



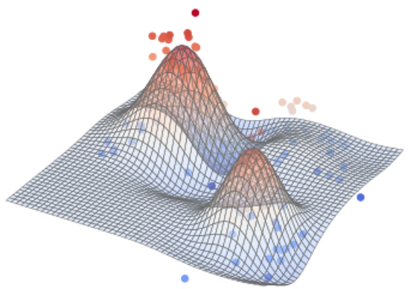




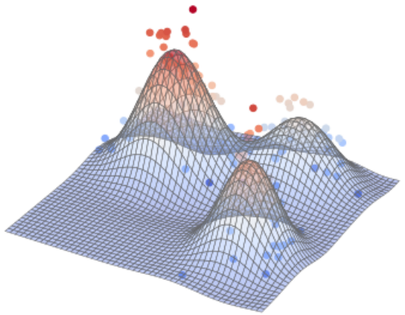
Random Forest



Neural Network



SVR



Kernel Ridge

Takeaways

Three lessons learned as I experienced **this illusion** being shattered…

1. The goals of ML and ‘materials/chemical science’ are **fundamentally different**. What we need here is not ML but a much harder problem of ‘machine discovery’
2. If we go for a hypothesis-free + off-the-shelf solution, exploration by **decision tree ensembles**, combined with UQ and abstracted (coarse grained) feature representations, will give a very strong baseline.
3. If we want more that that, we **can’t be hypothesis free**. Any strategies to narrow down the scope as well as domain expertise really matters.

The goals are fundamentally different.

To find "a material that is better than any existing materials today" or "a superior material that has never existed before.

The goals are fundamentally different.

To find "a material that is better than any existing materials today" or "a superior material that has never existed before.

To make a prediction for a given material on the basis of any similarities to the existing materials (i.e. the training data).

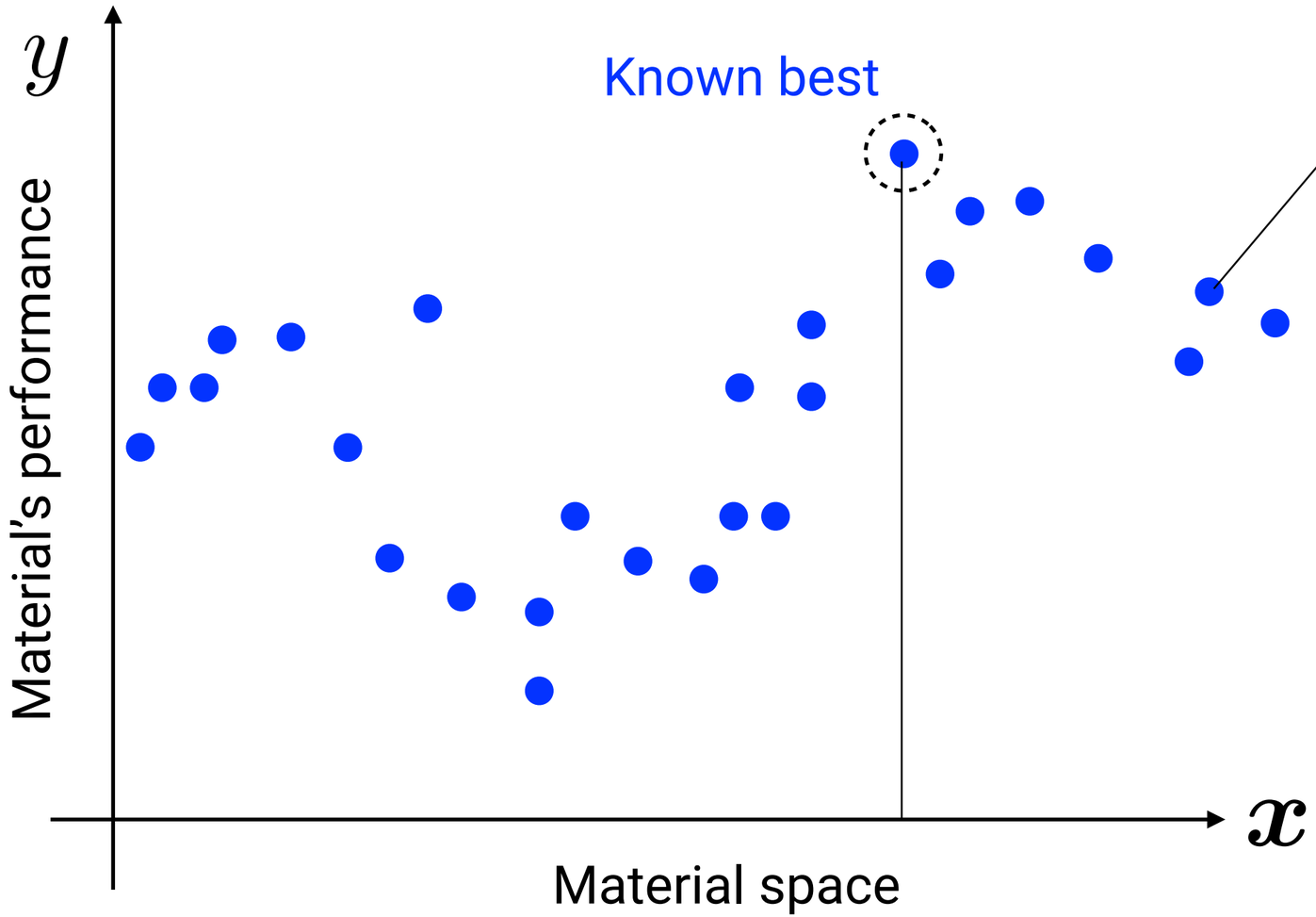
The goals are fundamentally different.

To find "a material that is better than any existing materials today" or "a superior material that has never existed before.

To make a prediction for a given material on the basis of any similarities to the existing materials (i.e. the training data).

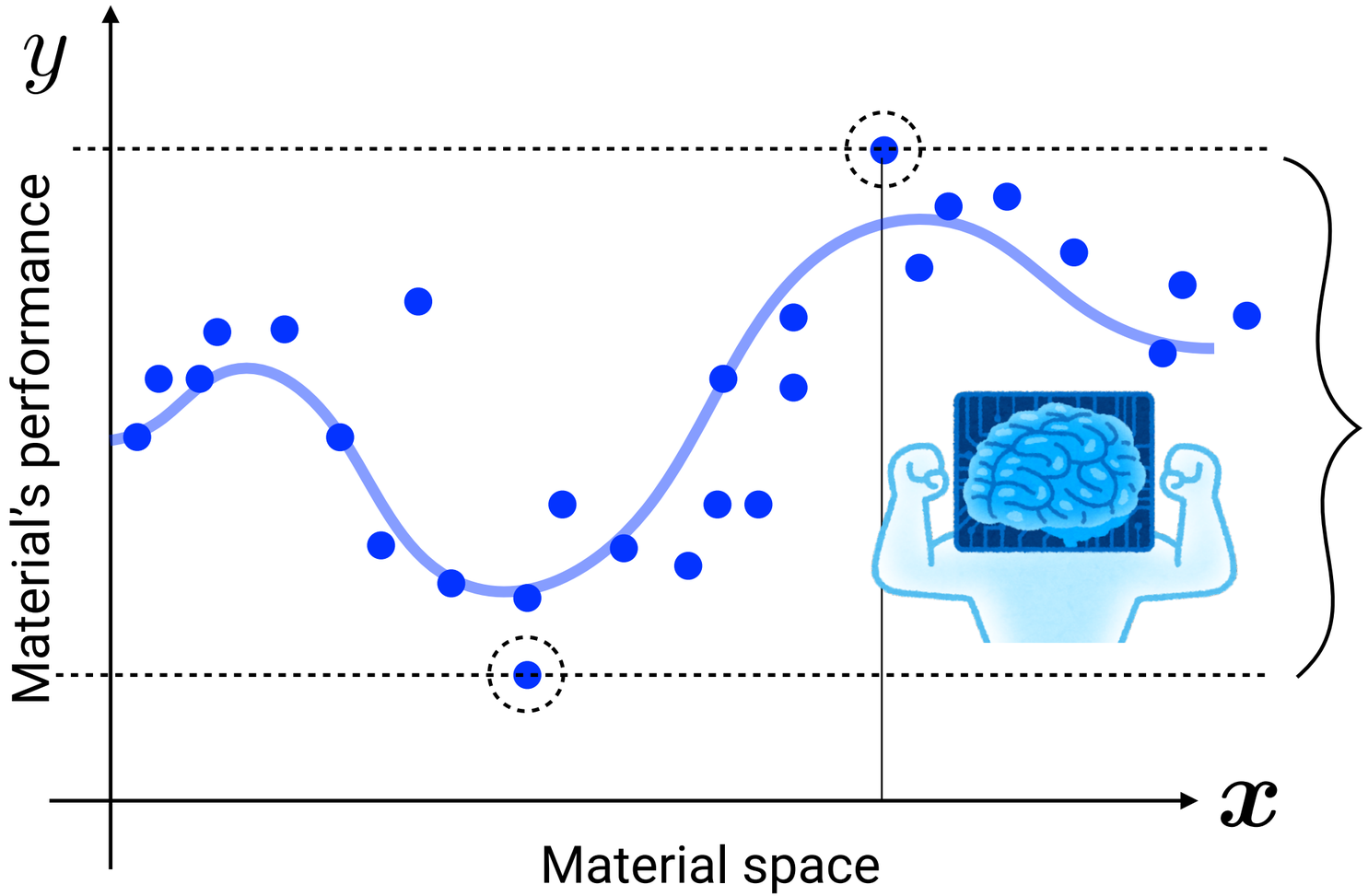
From a statistical point of view, this is the same as saying **"I want outliers (exceptions)."** The best known material is already a statistical outlier.

The setup is fundamentally different from ML's



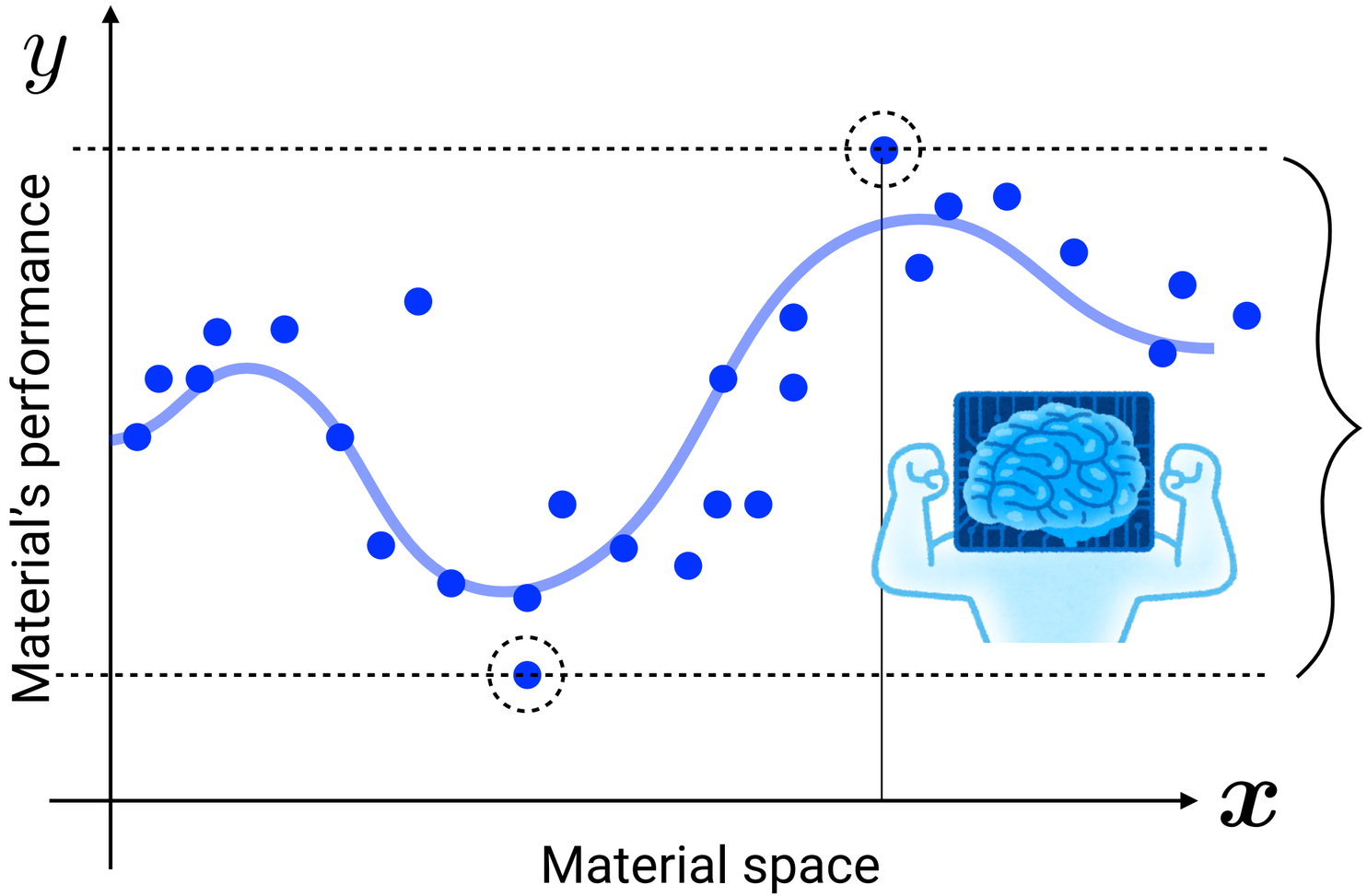
I want material x
with larger y anyway!

An inconvenient truth: ML is useless for this purpose



ML predicted values cut through **the middle** of the given training samples.

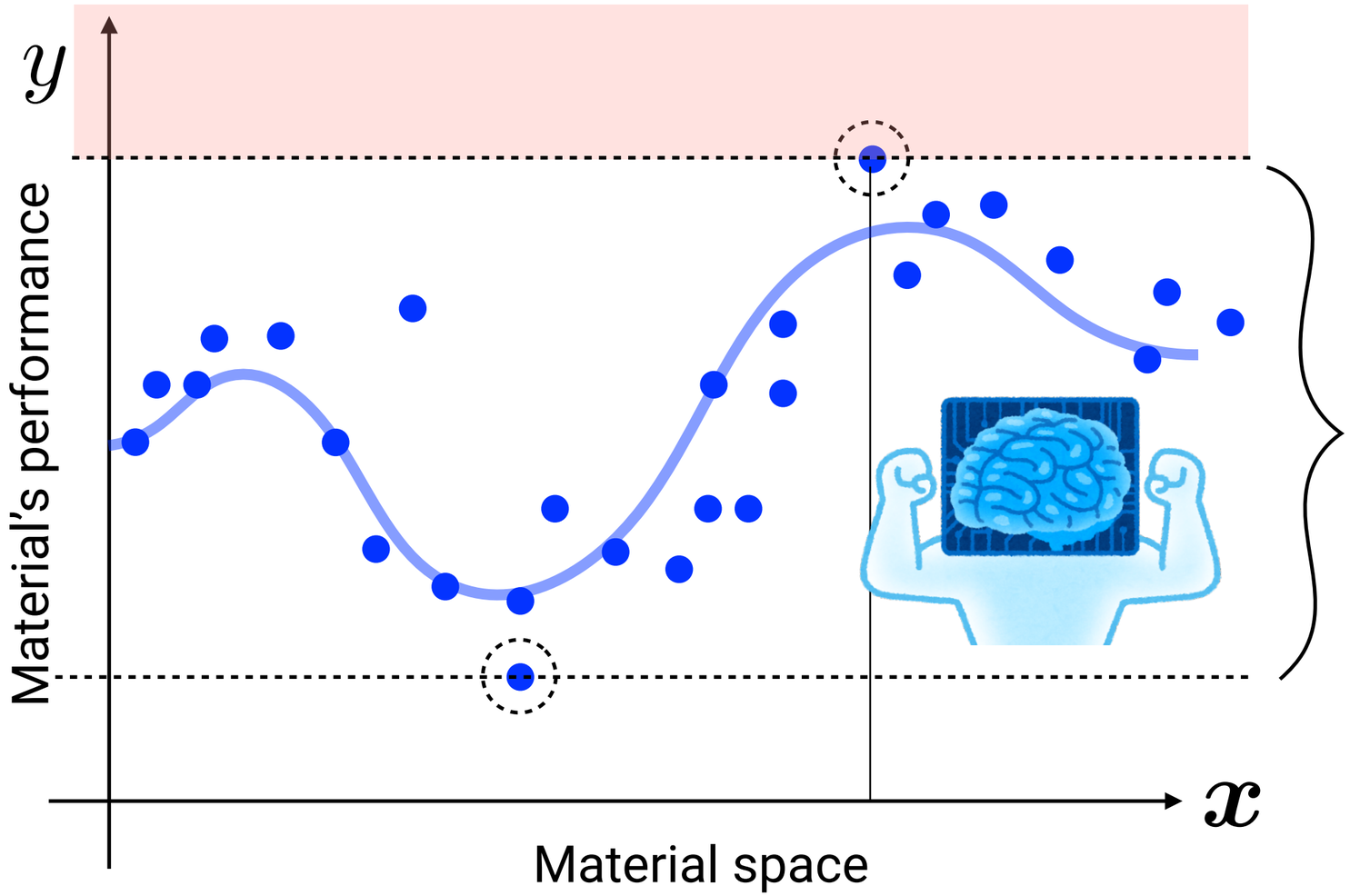
An inconvenient truth: ML is useless for this purpose



ML predicted values cut through **the middle** of the given training samples.

i.e. they take mediocre values **between the best and worst values** in the training data.

An inconvenient truth: ML is useless for this purpose



ML predicted values cut through **the middle** of the given training samples.

i.e. they take mediocre values **between the best and worst values** in the training data.

In conclusion,
ML can't predict better material than ones in the training data

It’s not a bug, it’s a feature

- This is **not a bug, it’s a feature!**

It’s not a bug, it’s a feature

- This is **not a bug, it’s a feature!**
- Furthermore, “Let’s try ML” situations usually imply the **paucity** of data.

  If we already have sufficient data, experts would already identify promising materials and there is **no need to use ML predictions**.

It’s not a bug, it’s a feature

- This is **not a bug, it’s a feature!**
- Furthermore, “Let’s try ML” situations usually imply the **paucity** of data.

  If we already have sufficient data, experts would already identify promising materials and there is **no need to use ML predictions**.

- In such a situation, it is **extremely difficult** to accurately evaluate the ML predictions since it means that we don’t have enough data **for testing** either.

It’s not a bug, it’s a feature

- This is **not a bug, it’s a feature!**
- Furthermore, “Let’s try ML” situations usually imply the **paucity** of data.

  If we already have sufficient data, experts would already identify promising materials and there is **no need to use ML predictions**.

- In such a situation, it is **extremely difficult** to accurately evaluate the ML predictions since it means that we don’t have enough data **for testing** either.

  It is a matter of course for ML to be able to predict the training examples. So we need to ensure if ML can predict **other examples than the training ones.**

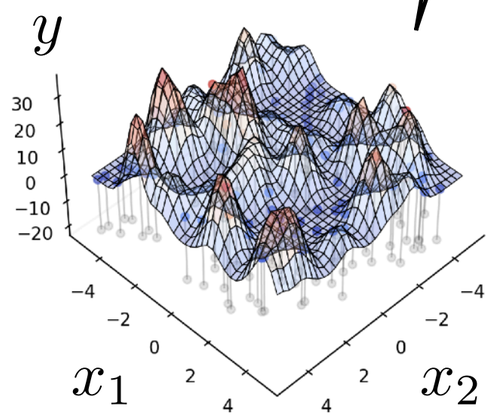
  However, test data means **everything but the training examples**…

The training and test data also fundamentally differ

For discovery, accurate prediction **for the entire input space** is expected because we are interested in **any possible materials**! (no probability things here)

Materials/Chemical Sciences

Test data

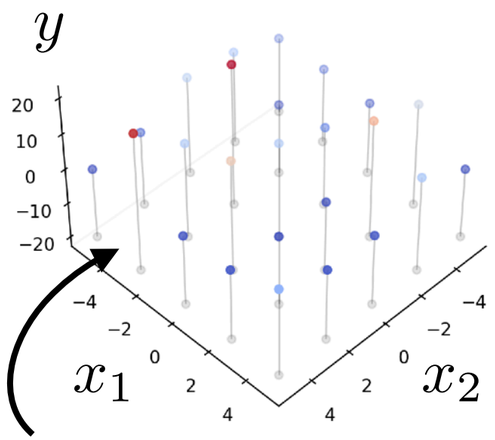


The training and test data also fundamentally differ

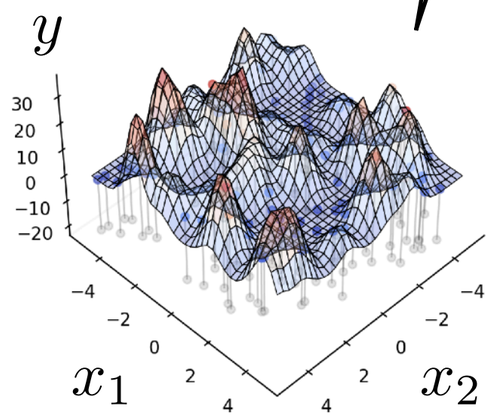
For discovery, accurate prediction **for the entire input space** is expected because we are interested in **any possible materials**! (no probability things here)

Materials/Chemical Sciences

Training data



Test data



Training samples should **cover the entire input space.**

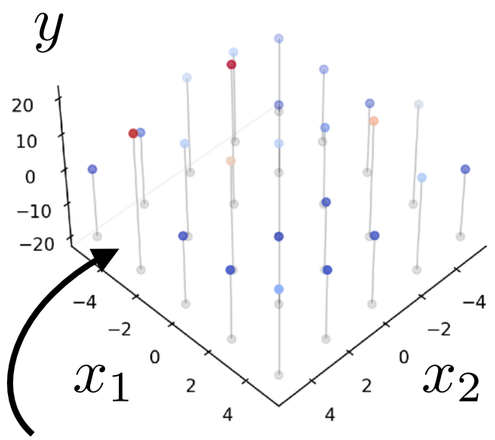
With considering Fisher's three principles for DoE.
Replication, Randomization, Local Control (Blocking)

The training and test data also fundamentally differ

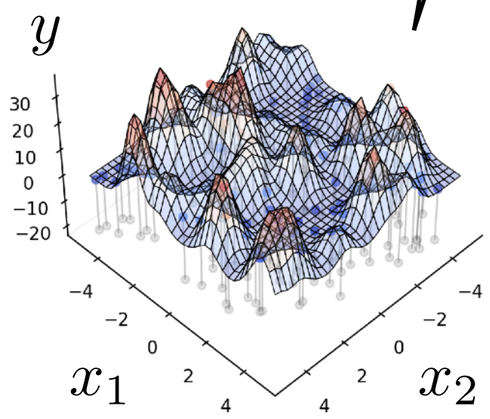
For discovery, accurate prediction **for the entire input space** is expected because we are interested in **any possible materials**! (no probability things here)

Materials/Chemical Sciences

Training data



Test data

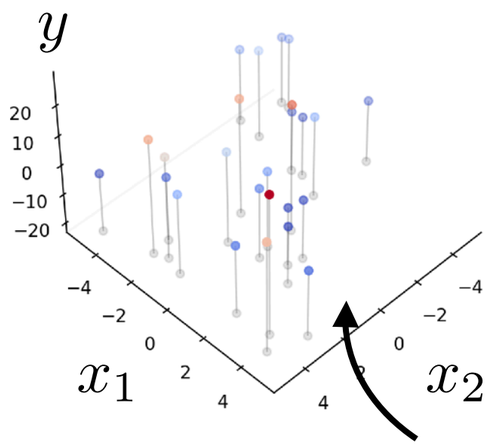


Training samples should **cover the entire input space.**

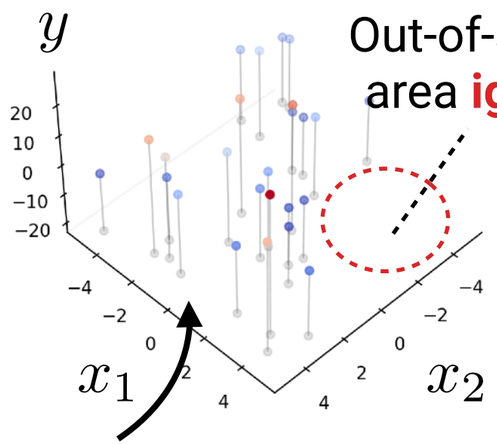
With considering Fisher's three principles for DoE.
Replication, Randomization, Local Control (Blocking)

Machine Learning

Training data



Test data



Out-of-sample area **ignored**

Both samples follow **the same** distribution

“Machine Discovery” Problem

We should recognize this problem as **a quite different problem** from standard ML!

“Machine Discovery” Problem

We should recognize this problem as **a quite different problem** from standard ML!

It is **way harder than ML**, and requires systematic study on whether any ‘scientific discovery’ can be rationalized by using “hard” sciences as a compelling testbed. Indeed now is the best time to revisit this theme **with modern methods and data**.

Herbert A. Simon

Setsuo Arikawa

- Simon, Machine Discovery. (1997)
- Langley, Simon, Bradshaw, Zytkow, Scientific Discovery: Computational Explorations of the Creative Process (1987).
- Arikawa, Our Studies on Machine Learning and Machine Discovery. (1996)
- Arikawa et al, The Discovery Science Project (2000).

Won Nobel Prize & Turing Award

Human and machine discovery are **gradual problem-solving processes** of searching large problem spaces for incompletely defined goal objects. (Simon)

Takeaways

Three lessons learned as I experienced **this illusion** being shattered…

1. The goals of ML and ‘materials/chemical science’ are **fundamentally different**. What we need here is not ML but a much harder problem of ‘machine discovery’

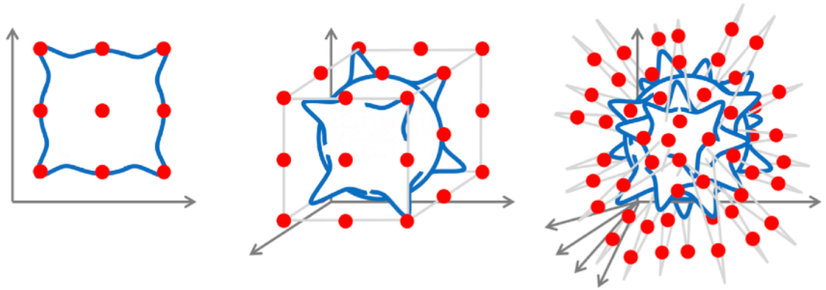
2. If we go for a hypothesis-free + off-the-shelf solution, exploration by **decision tree ensembles**, combined with UQ and abstracted (coarse grained) feature representations, will give a very strong baseline.

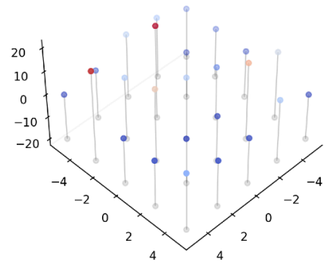
3. If we want more that that, we **can’t be hypothesis free**. Any strategies to narrow down the scope as well as domain expertise really matters.

Approx for the entire input space is practically impossible

Inconvenient mathematical truths (Curse of dimensionality)

1. The number of samples required to ensure the accurate prediction for the entire input space (uniform approximation) is necessarily exponential in the dimension.



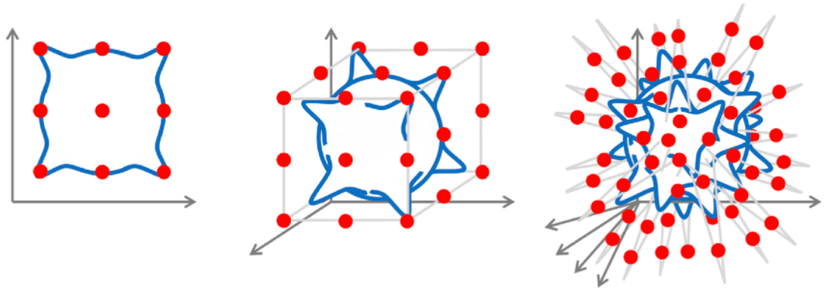


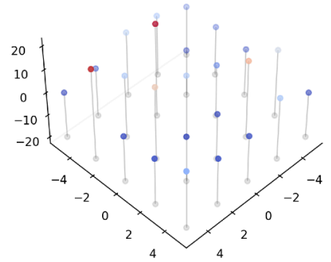
If we take 5 levels for each variable,
we need $5^2 = 25$ for 2 variables;
we need $5^{10} \approx 10$ millions for just 10 variables.

Approx for the entire input space is practically impossible

Inconvenient mathematical truths (Curse of dimensionality)

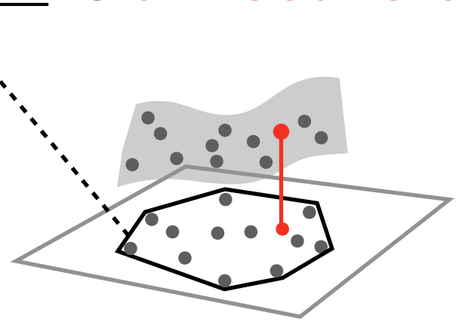
1. The number of samples required to ensure the accurate prediction for the entire input space (uniform approximation) is necessarily exponential in the dimension.





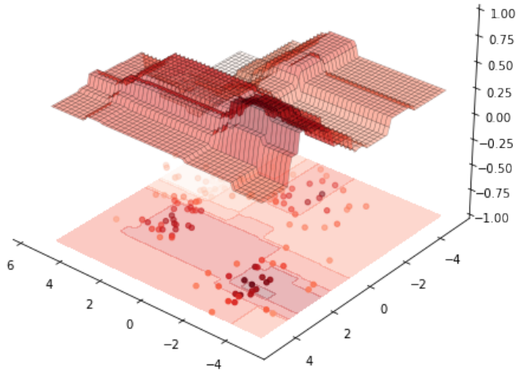
If we take 5 levels for each variable,
we need $5^2 = 25$ for 2 variables;
we need $5^{10} \approx 10$ millions for just 10 variables.

2. The probability that a new sample falls in training set's convex hull is almost zero for a high-dimensional (>100) space.

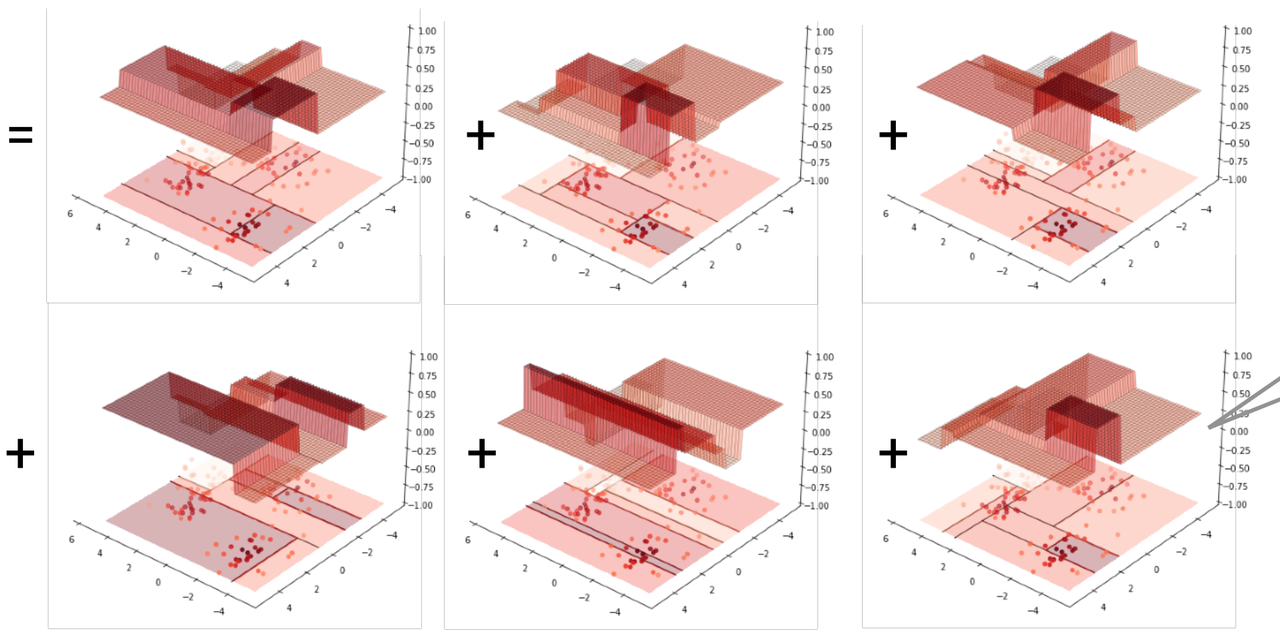


Interpolation almost surely never happens, and "Learning in high dimension always amounts to extrapolation".
(Balestriero, Pesenti, LeCun, 2021; arXiv:2110.09485)

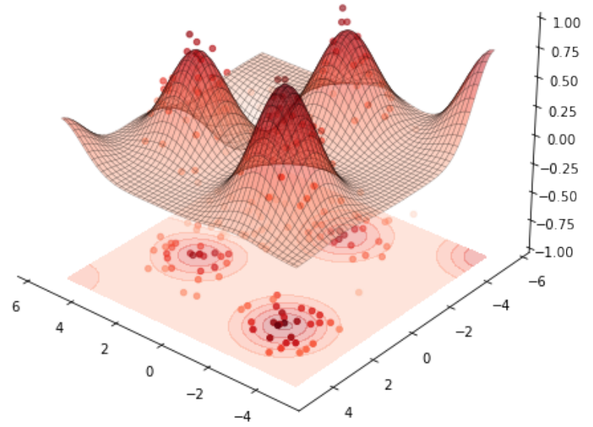
Decision tree ensembles: Local-averaging estimators

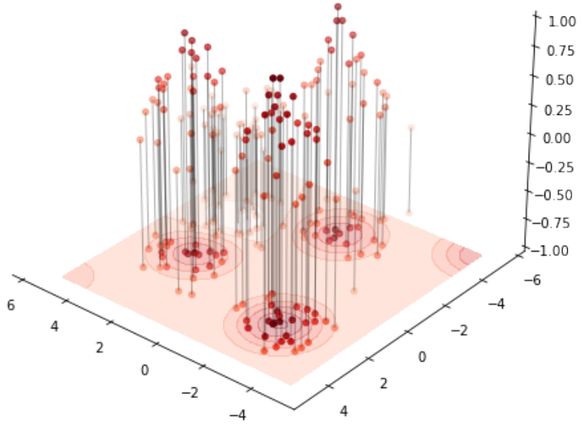


(piecewise constant)



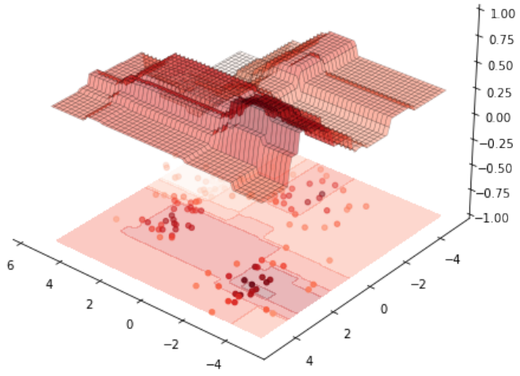
i.e. Histogram rules on data-dependent partitions (piecewise constant)





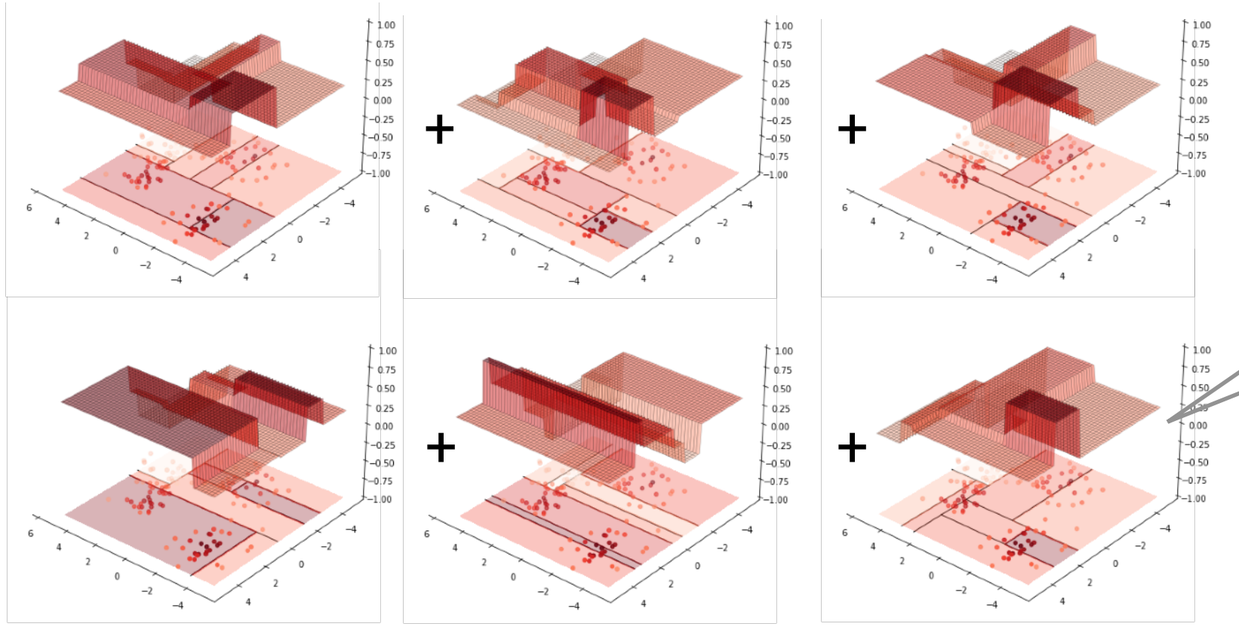
- Make prediction by a histogram rule, i.e. the average of subset of training samples **even for the out-of-sample area**
- It's a histogram and unintentional interpolation just by ungrounded inductive biases never happens **even in a high-dimensional space.**

Decision tree ensembles: Local-averaging estimators

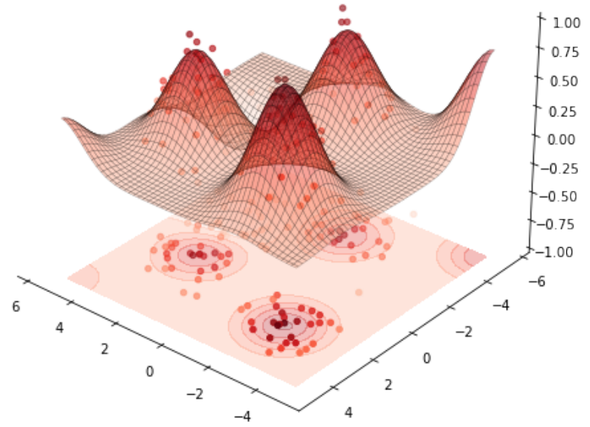


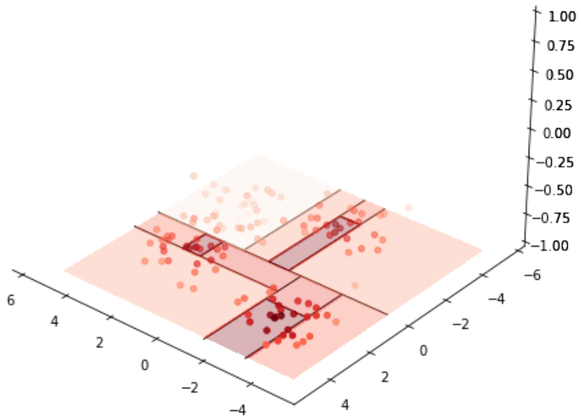
(piecewise constant)

=



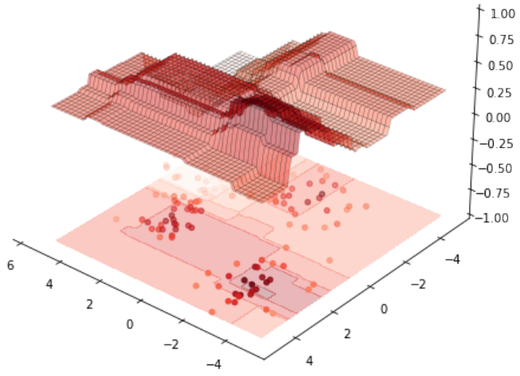
i.e. Histogram rules on data-dependent partitions (piecewise constant)



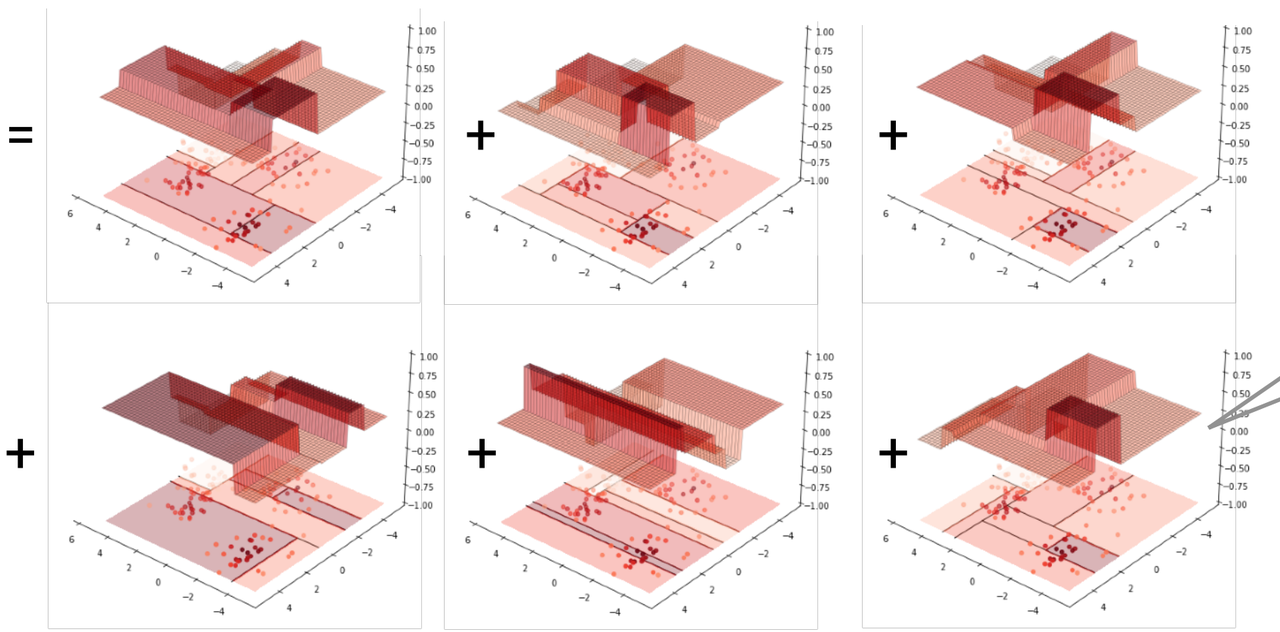


- Make prediction by a histogram rule, i.e. the average of subset of training samples **even for the out-of-sample area**
- It's a histogram and unintentional interpolation just by ungrounded inductive biases never happens **even in a high-dimensional space.**

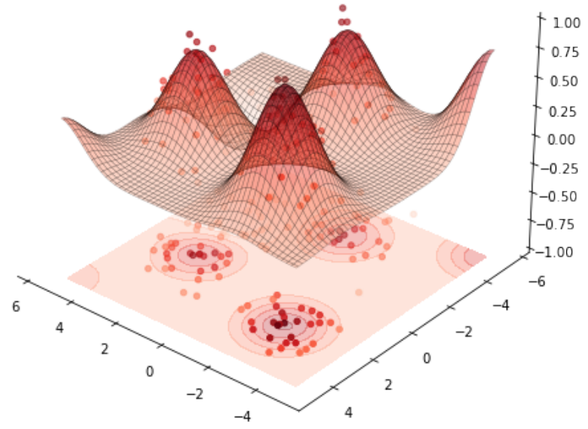
Decision tree ensembles: Local-averaging estimators

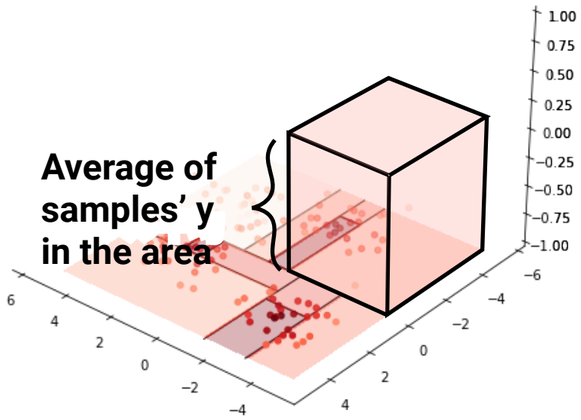


(piecewise constant)



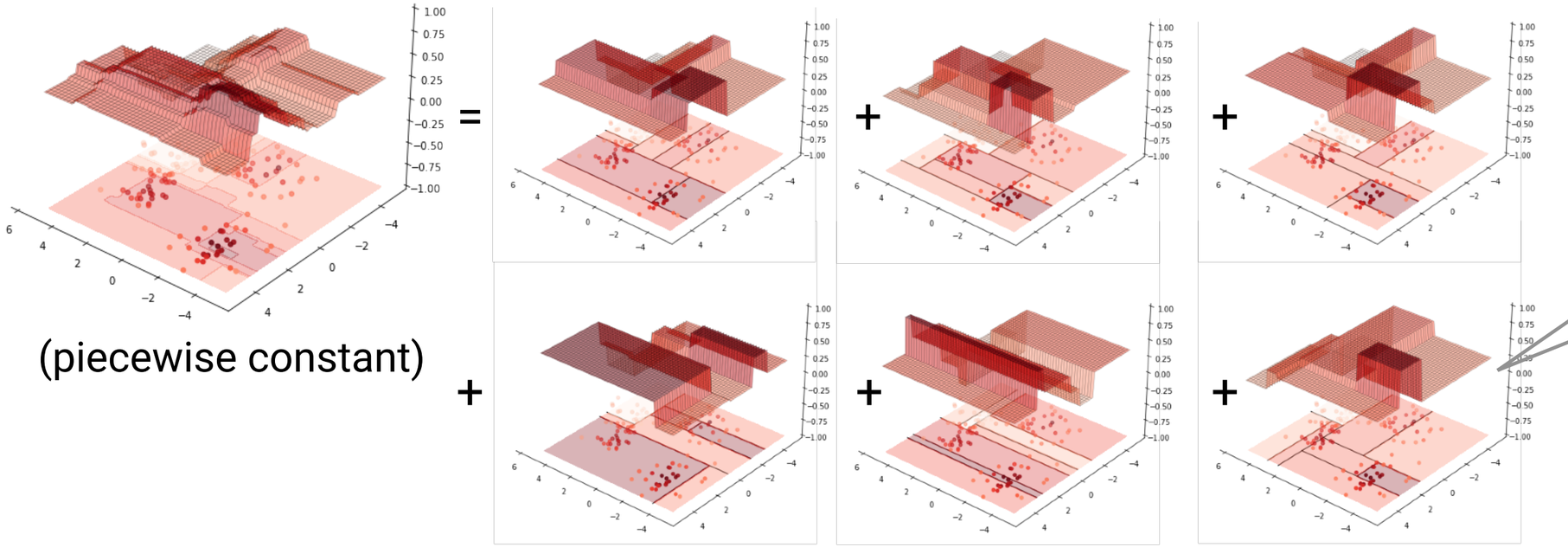
i.e. Histogram rules on data-dependent partitions (piecewise constant)





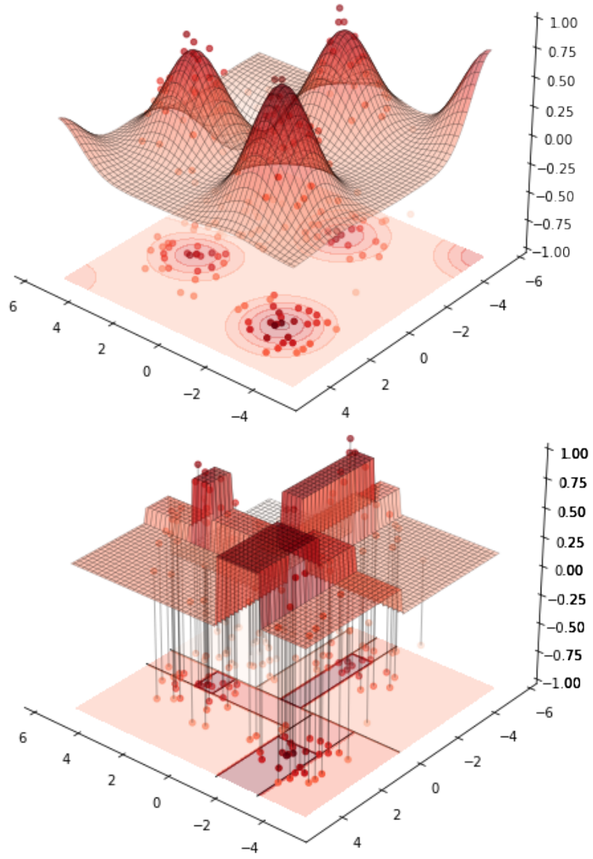
- Make prediction by a histogram rule, i.e. the average of subset of training samples **even for the out-of-sample area**
- It's a histogram and unintentional interpolation just by ungrounded inductive biases never happens **even in a high-dimensional space.**

Decision tree ensembles: Local-averaging estimators



(piecewise constant)

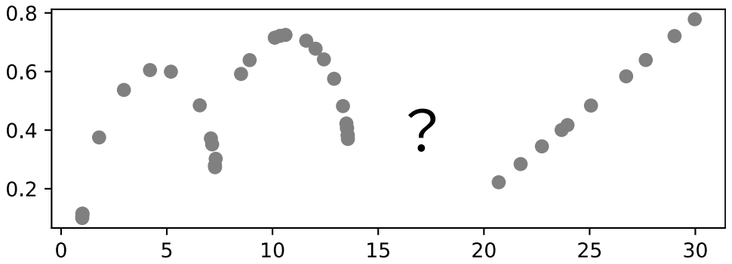
i.e. Histogram rules on data-dependent partitions (piecewise constant)



- Make prediction by a histogram rule, i.e. the average of subset of training samples **even for the out-of-sample area**
- It's a histogram and unintentional interpolation just by ungrounded inductive biases never happens **even in a high-dimensional space.**

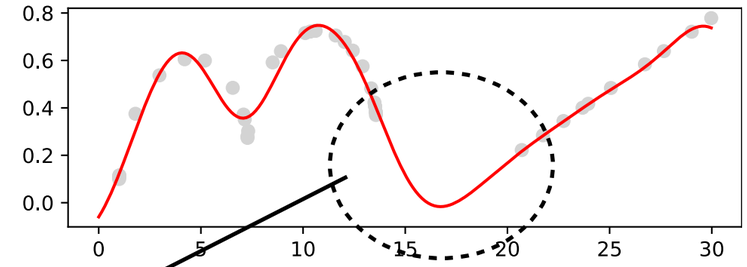
Evidence-based behavior for out-of-sample area

For out-of-sample area, we cannot say anything confident without any assumptions

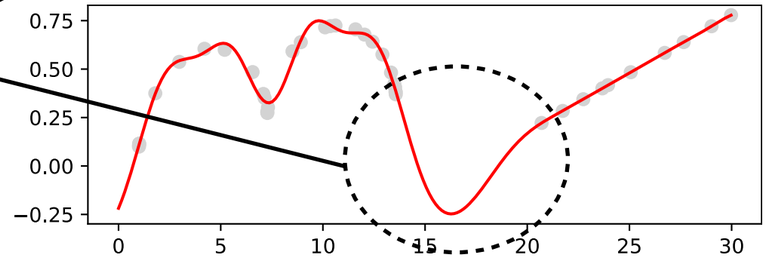


by inductive biases (continuity)

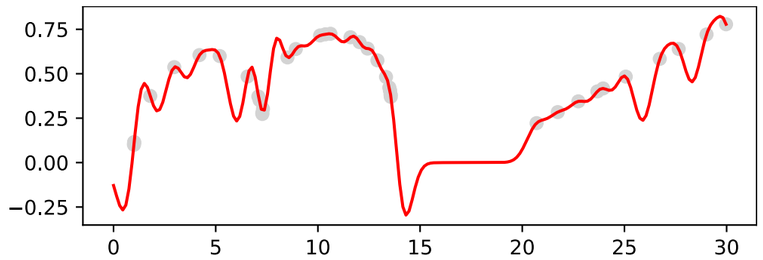
```
KernelRidge(kernel='rbf', alpha=0.05, gamma=0.1)
```




```
KernelRidge(kernel='rbf', alpha=1e-4, gamma=0.1)
```

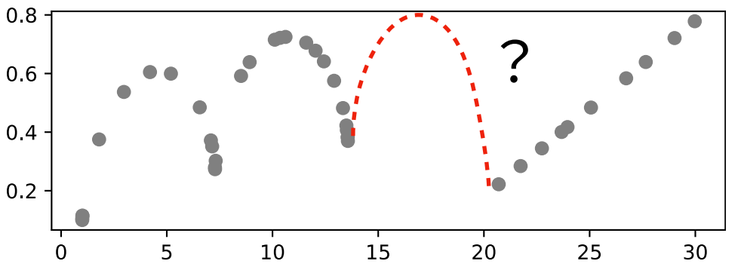



```
KernelRidge(kernel='rbf', alpha=1e-4, gamma=2.0)
```



Evidence-based behavior for out-of-sample area

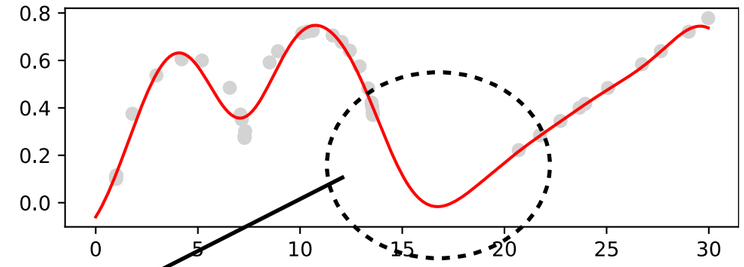
For out-of-sample area, we cannot say anything confident without any assumptions



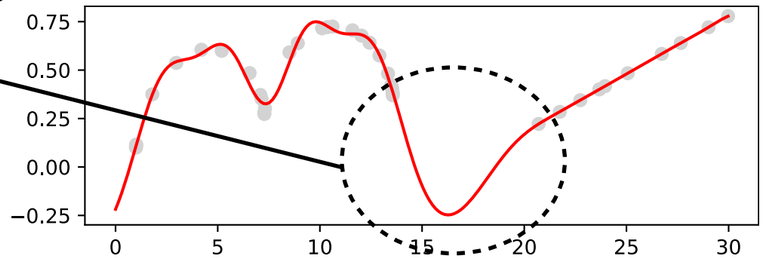
by inductive biases
(continuity)

But this can be **not necessarily continuous**
(selectivity cliffs,
activity cliffs, etc)

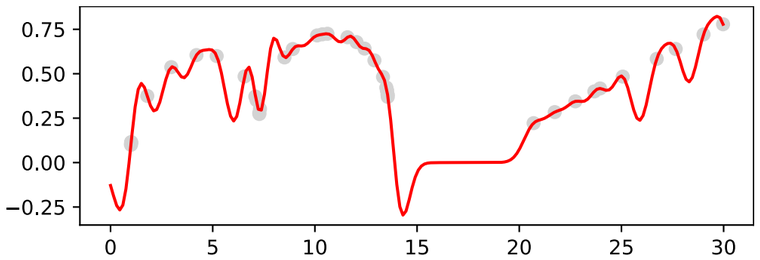
```
KernelRidge(kernel='rbf', alpha=0.05, gamma=0.1)
```



```
KernelRidge(kernel='rbf', alpha=1e-4, gamma=0.1)
```

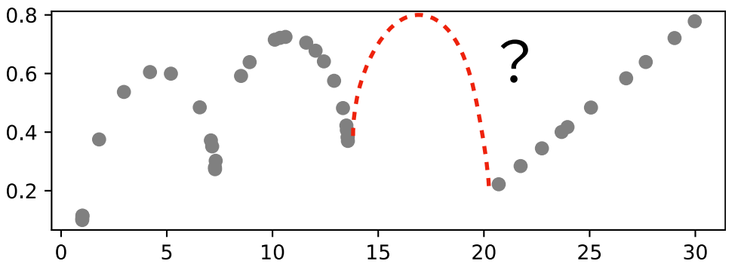


```
KernelRidge(kernel='rbf', alpha=1e-4, gamma=2.0)
```



Evidence-based behavior for out-of-sample area

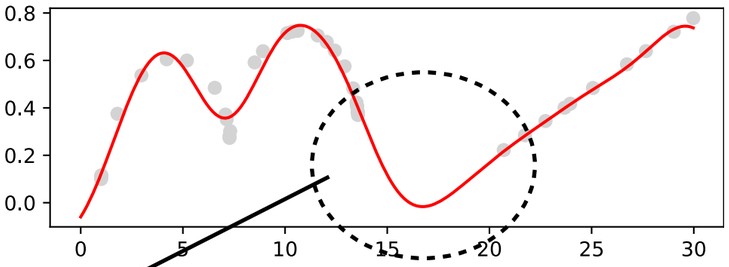
For out-of-sample area, we cannot say anything confident without any assumptions



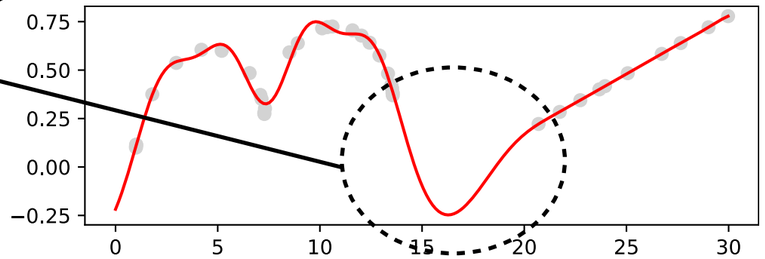
by inductive biases
(continuity)

But this can be **not necessarily continuous**
(selectivity cliffs,
activity cliffs, etc)

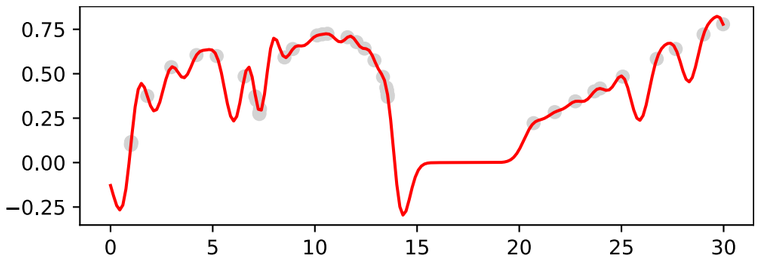
`KernelRidge(kernel='rbf', alpha=0.05, gamma=0.1)`



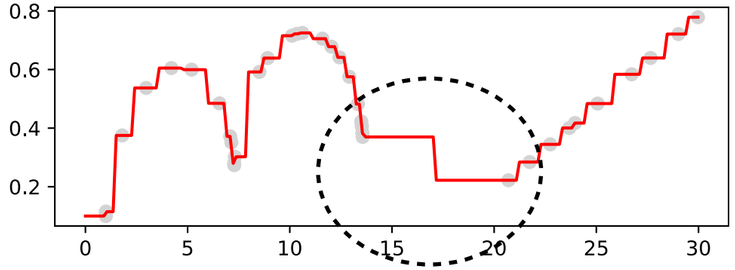
`KernelRidge(kernel='rbf', alpha=1e-4, gamma=0.1)`



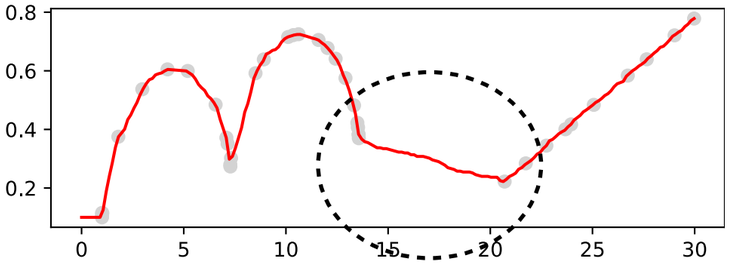
`KernelRidge(kernel='rbf', alpha=1e-4, gamma=2.0)`



`DecisionTreeRegressor()`



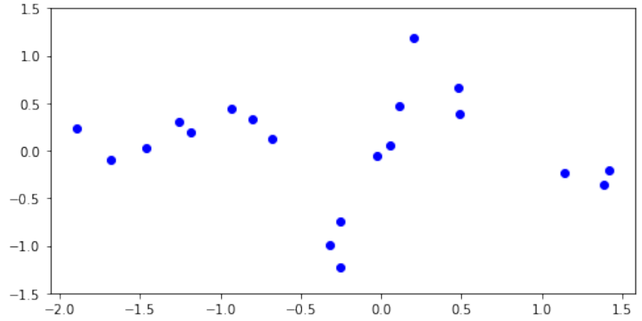
`ExtraTreesRegressor(n_estimators=50)`

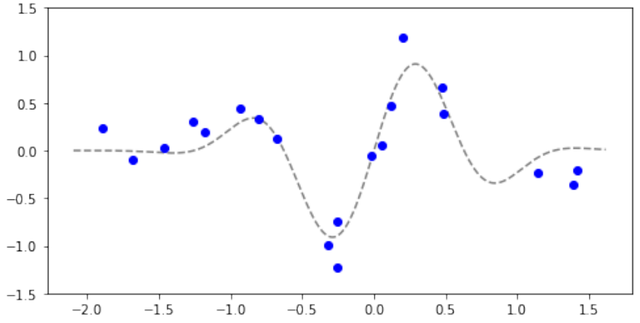


conservative and safer prediction

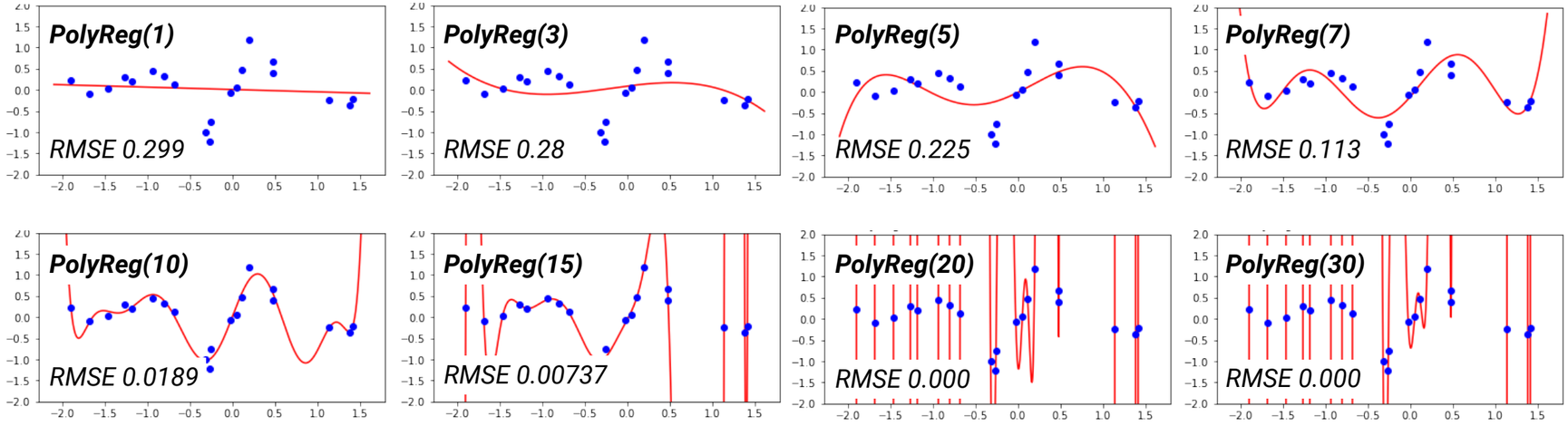
at least, grounded by some given data

Adaptability for non-smooth changes (Benign overfitting)

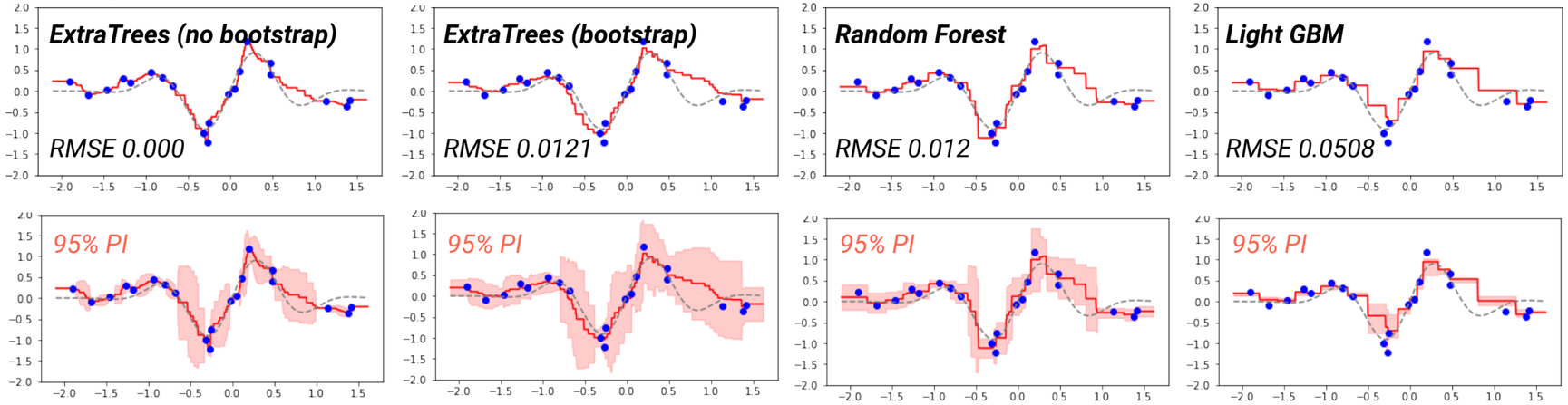




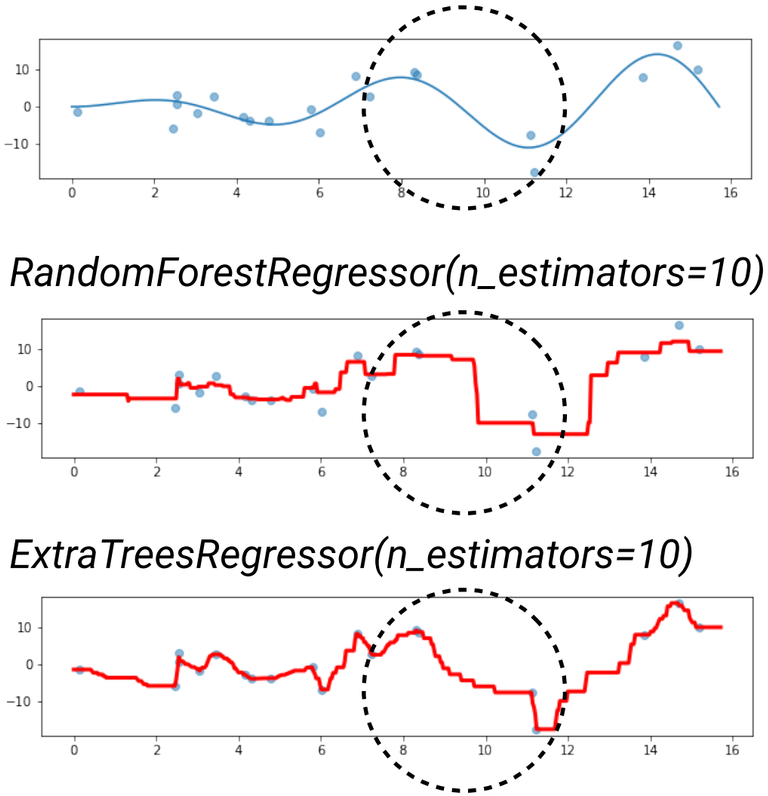
Problematic overfitting by polynomial regression of order k



Clearly overfitted but harmless (still informative)



Pseudo-continuous interpolation of ExtraTrees



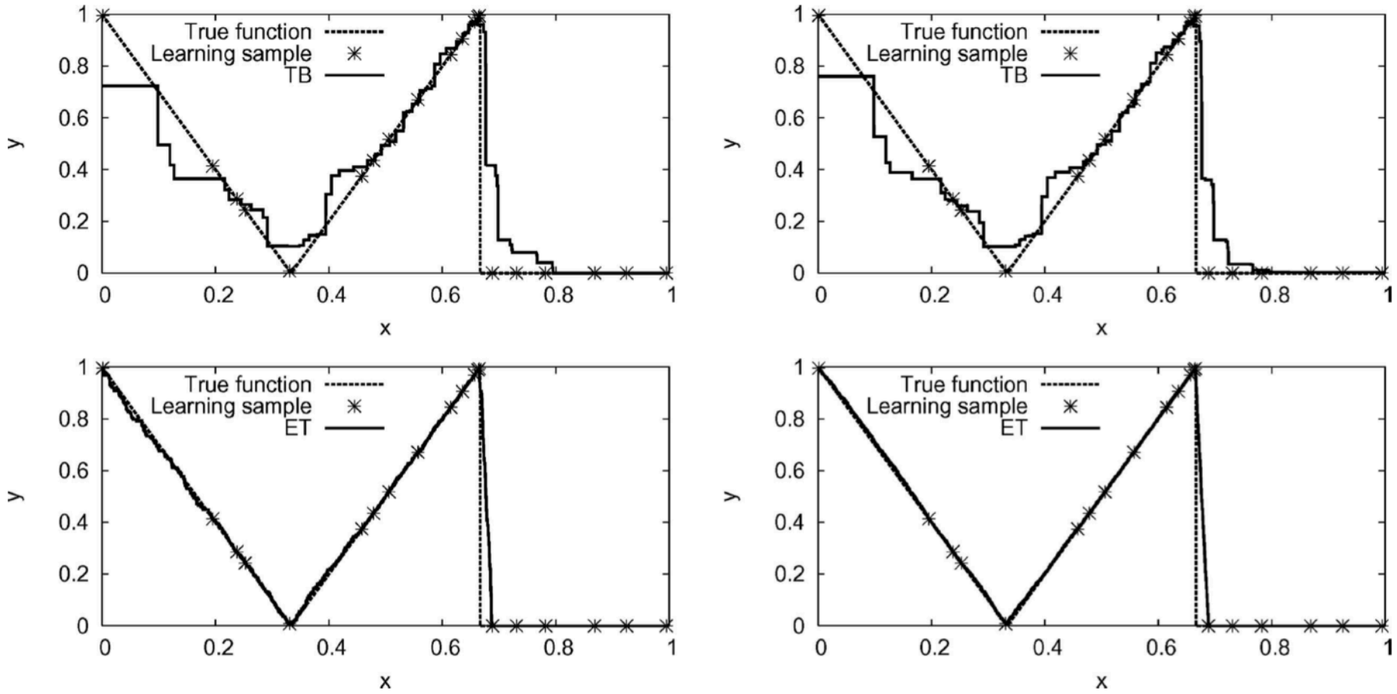


Fig. 10 Tree Bagging, and fully developed Extra-Trees ($n_{\min} = 2$) on a one-dimensional piecewise linear problem ($N = 20$). Left with $M = 100$ trees, right with $M = 1000$ trees.

Geurts, Ernst, Wehenkel, Extremely randomized trees. *Mach Learn* 63, 3–42 (2006).
https://doi.org/10.1007/s10994-006-6226-1

Our recent research: Method

Very Conservative Prediction

- Evidence-based behavior for out-of-sample area
- Adaptability for non-smooth changes

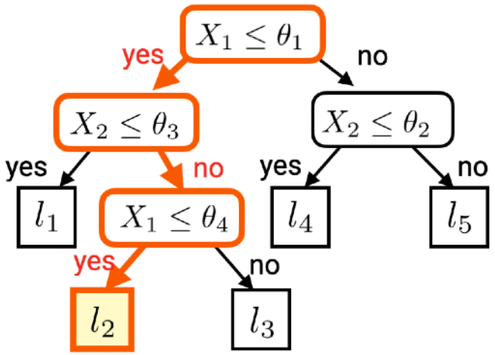
Very Radical Representation

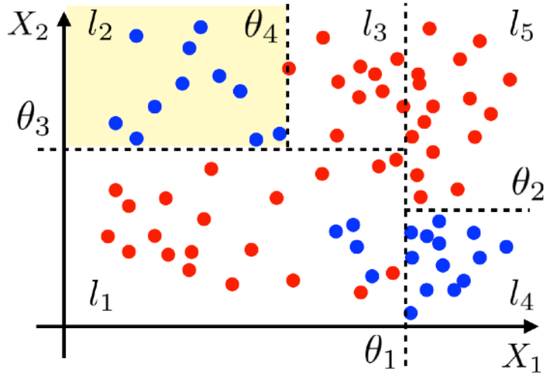
- avoid fragmented memorization
- compensate for elemental sparsity and data paucity

Decision tree ensembles (with UQ)
i.e. histogram on data-dependent partitions

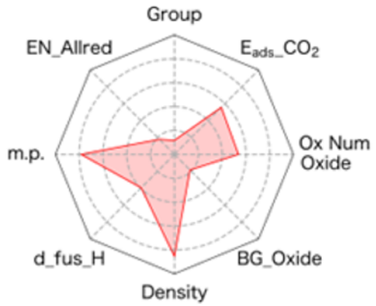
+

Abstracted (coarse grained) featurization of chemical compositions

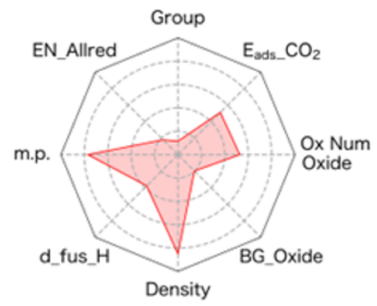




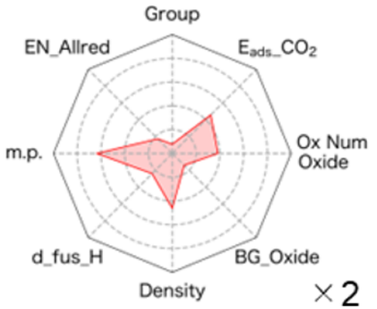
Pt(3)/**Ba(2)-Mo(1)-Tm(1)-Eu(0.5)-Dy(0.5)**/TiO_2



Pt(3)/**Mo(1)-Ba(1)-Tb(1)-Ho(1)-Cs(0.5)**/TiO_2



Pt(3)/**Rb(1)-Ba(1)-Mo(0.6)-Nb(0.2)**/TiO_2



×2

Takeaways

Three lessons learned as I experienced **this illusion** being shattered…

1. The goals of ML and ‘materials/chemical science’ are **fundamentally different**. What we need here is not ML but a much harder problem of ‘machine discovery’
2. If we go for a hypothesis-free + off-the-shelf solution, exploration by **decision tree ensembles**, combined with UQ and abstracted (coarse grained) feature representations, will give a very strong baseline.
3. If we want more that that, we **can’t be hypothesis free**. Any strategies to narrow down the scope as well as domain expertise really matters.

Can ML contribute to scientific discovery/understanding?

I assume that ML-based exploration like ours is **used, calibrated, and carefully monitored by human experts**. I am skeptical so far about whether scientific discovery can be **fully automated by AI.**

What kinds of elemental features are used…?

What level of coarse graining is effective…?

:

×

Can ML contribute to scientific discovery/understanding?

I assume that ML-based exploration like ours is **used, calibrated, and carefully monitored by human experts**. I am skeptical so far about whether scientific discovery can be **fully automated by AI.**

What kinds of elemental features are used…?

What level of coarse graining is effective…?

:

×

- In the first place, the majority of scientific research, particularly experimental science, is still largely empirical, and much is irrationally left to **luck and inertia**.

Can ML contribute to scientific discovery/understanding?

I assume that ML-based exploration like ours is **used, calibrated, and carefully monitored by human experts**. I am skeptical so far about whether scientific discovery can be **fully automated by AI.**

What kinds of elemental features are used…?

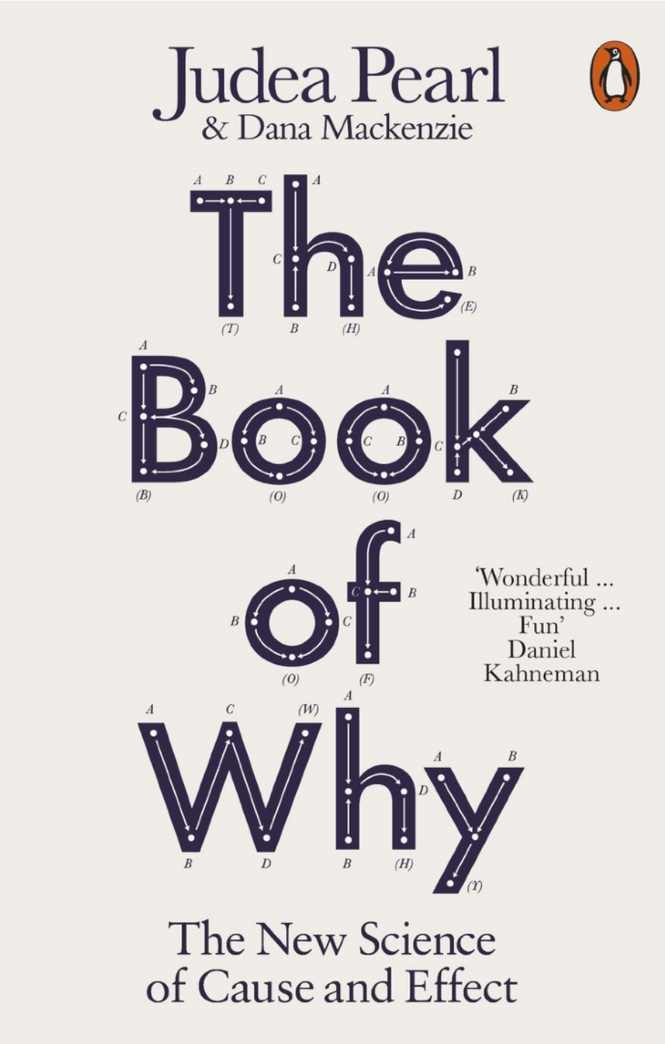
What level of coarse graining is effective…?

:

×

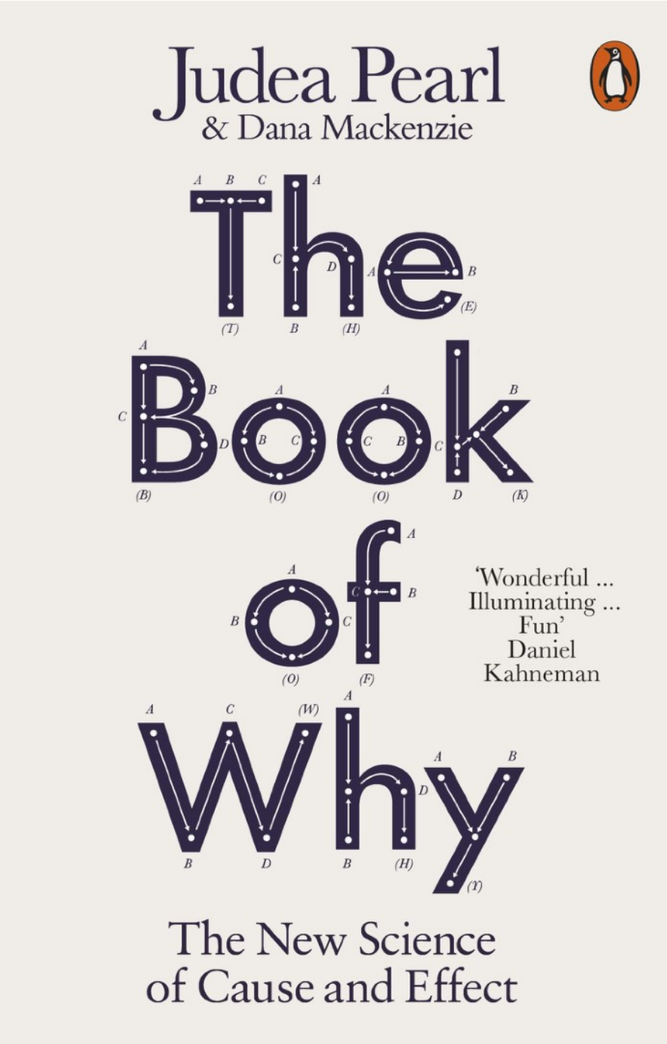
- In the first place, the majority of scientific research, particularly experimental science, is still largely empirical, and much is irrationally left to **luck and inertia**.
- ML-based exploration is just a glorified version of empirical exploration, and exhibits different types of **"bounded rationality** (Herb Simon again!)**"** as we are bounded by our own "cognitive limits."

Science requires causal understanding



We cannot be hypothesis free when we want **causality**.

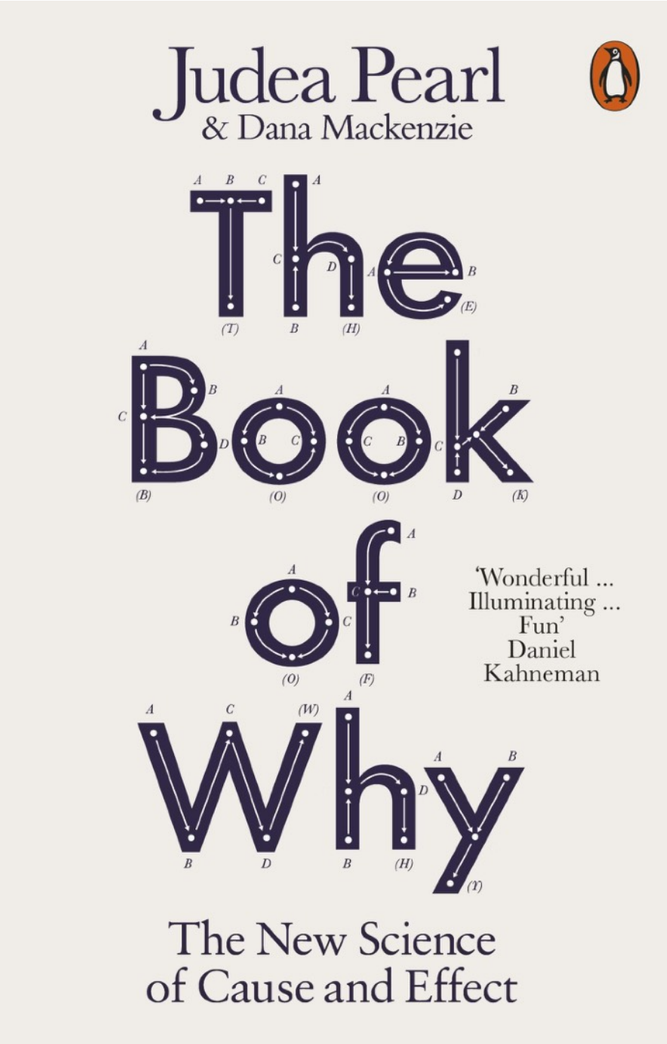
Science requires causal understanding



We cannot be hypothesis free when we want **causality**.

- *"Causal analysis is emphatically **not just about data**; in causal analysis we must incorporate some understanding of the process that produces the data, and then we get something that was not in the data to begin with."*

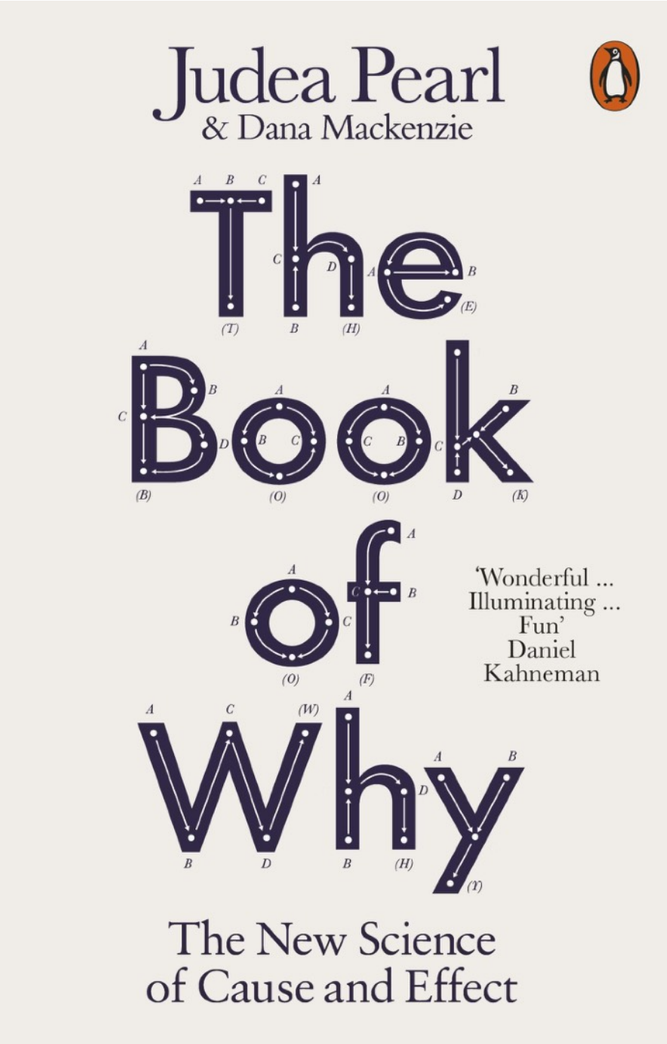
Science requires causal understanding



We cannot be hypothesis free when we want **causality**.

- *"Causal analysis is emphatically **not just about data**; in causal analysis we must incorporate some understanding of the process that produces the data, and then we get something that was not in the data to begin with."*
- "Unlike correlation and most of the other tools of mainstream statistics, causal analysis requires the user to make a **subjective commitment**."

Science requires causal understanding



We cannot be hypothesis free when we want **causality**.

- *"Causal analysis is emphatically **not just about data**; in causal analysis we must incorporate some understanding of the process that produces the data, and then we get something that was not in the data to begin with."*
- "Unlike correlation and most of the other tools of mainstream statistics, causal analysis requires the user to make a **subjective commitment**."

For causal understanding, data is not everything. We need **something else that doesn't come from the data themselves**.

ML gives prediction; We want discovery/understanding

Science is built up with facts, as a house is with stones.
But **a collection of facts is no more a science** than a heap of stones is a house.

Henri Poincaré *"Science and Hypothesis"*

ML gives prediction; We want discovery/understanding

Science is built up with facts, as a house is with stones.
But **a collection of facts is no more a science** than a heap of stones is a house. Henri Poincaré *"Science and Hypothesis"*

- *"Theory-driven models can be wrong. But data-driven models cannot be wrong or right. Data-driven are not trying to describe an underlying reality."* (David Hand)

ML gives prediction; We want discovery/understanding

Science is built up with facts, as a house is with stones.
But **a collection of facts is no more a science** than a heap of stones is a house. Henri Poincaré *"Science and Hypothesis"*

- *"Theory-driven models can be wrong. But data-driven models cannot be wrong or right. Data-driven are not trying to describe an underlying reality."* (David Hand)
- *"The goal of finding models that are predictively accurate differs from the goal of finding models that are true."* Statistical Learning from a regression perspective.

ML gives prediction; We want discovery/understanding

Science is built up with facts, as a house is with stones.
But **a collection of facts is no more a science** than a heap of stones is a house. Henri Poincaré *"Science and Hypothesis"*

- *"Theory-driven models can be wrong. But data-driven models cannot be wrong or right. Data-driven are not trying to describe an underlying reality."* (David Hand)
- *"The goal of finding models that are predictively accurate differs from the goal of finding models that are true."* Statistical Learning from a regression perspective.

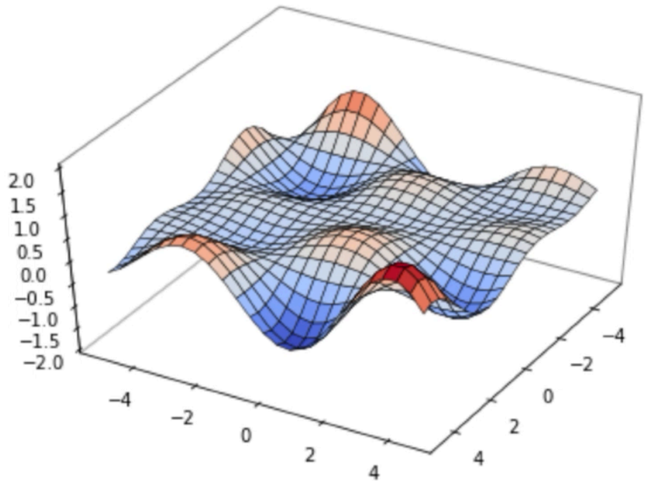
If we seek not prediction but (scientific) understanding, we basically **cannot remain hypothesis-free** because "understanding" is the problem of human recognition.

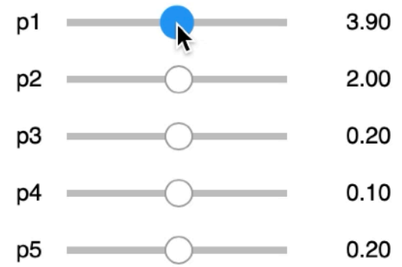
Giving Up on ML's Versatility

Modern ML models have the virtue of being able to **represent** **any function** just by changing parameter values.

e.g.
The *universal approximation theorem* says that neural networks can **approximate any function**.

"Blackbox" vs. "Hypothesis-free"



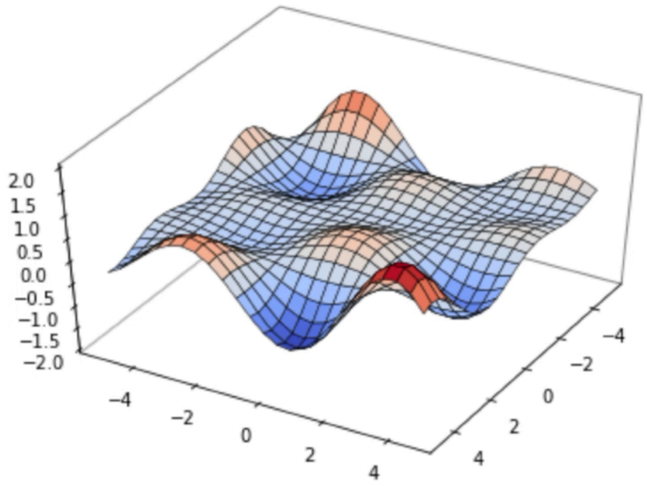


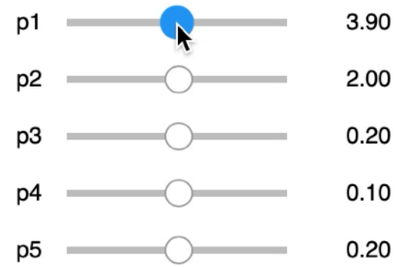
Giving Up on ML's Versatility

Modern ML models have the virtue of being able to **represent** **any function** just by changing parameter values.

e.g.
The *universal approximation theorem* says that neural networks can **approximate any function**.

"Blackbox" vs. "Hypothesis-free"



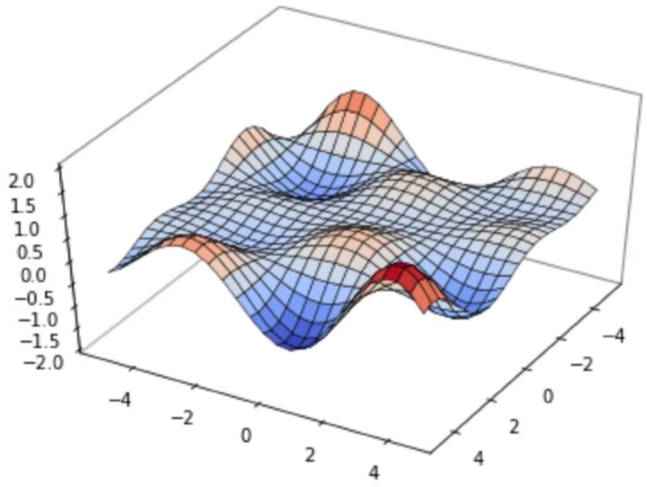


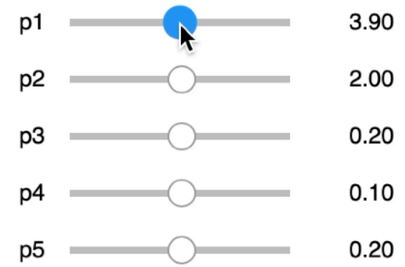
Giving Up on ML's Versatility

Modern ML models have the virtue of being able to **represent** **any function** just by changing parameter values.

e.g.
The *universal approximation theorem* says that neural networks can **approximate any function**.

"Blackbox" vs. "Hypothesis-free"





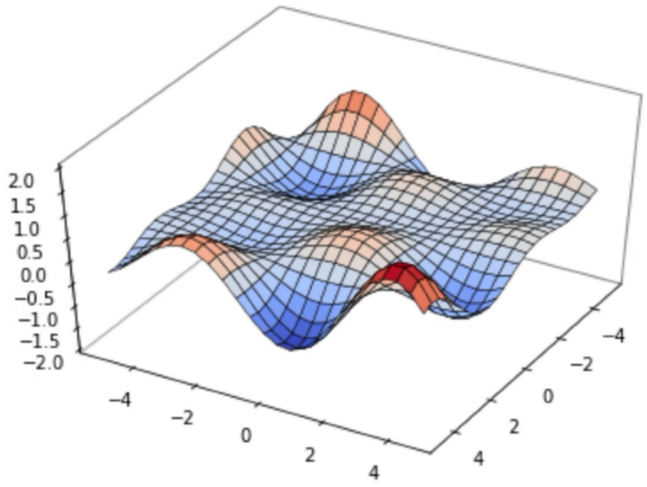
- However, when used in the natural sciences, this virtue leads to scientifically invalid predictions just by "spurious correlations" in the given finite data…

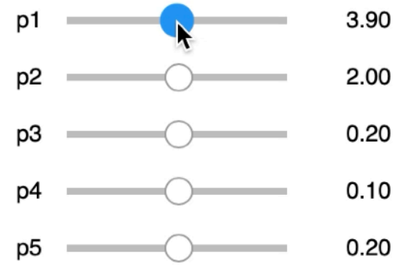
Giving Up on ML's Versatility

Modern ML models have the virtue of being able to **represent** **any function** just by changing parameter values.

e.g.
The *universal approximation theorem* says that neural networks can **approximate any function**.

"Blackbox" vs. "Hypothesis-free"





- However, when used in the natural sciences, this virtue leads to scientifically invalid predictions just by "spurious correlations" in the given finite data…
- It is **not good to be able to "represent any function,"** but it is **better to restrict the model** so that "it **cannot** represent scientifically invalid functions by design."

Path to Machine Discovery: 1st step is physics-informed?

https://doi.org/10.1038/s42254-021-00314-5

NATURE REVIEWS | PHYSICS
422 | JUNE 2021 | VOLUME 3

REVIEWS

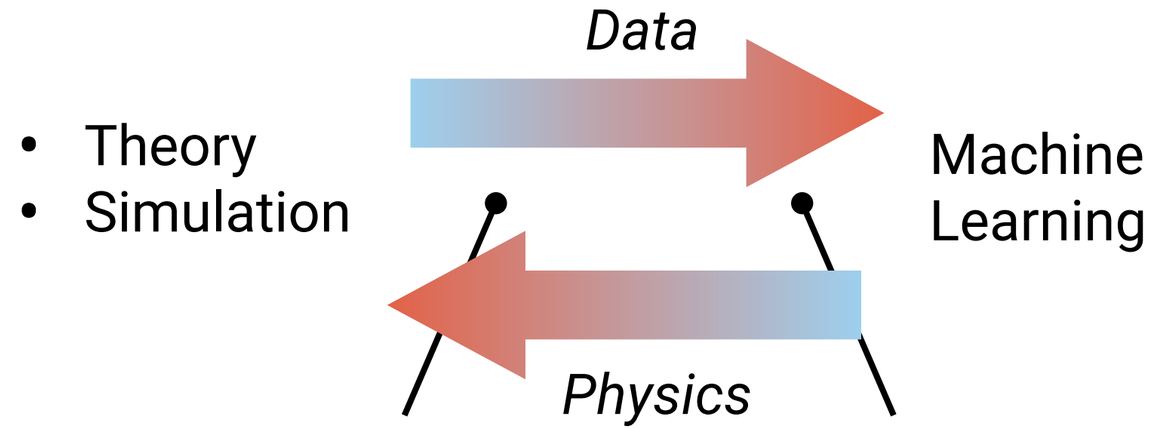
Physics-informed machine learning

George Em Karniadakis[1,2]✉, Ioannis G. Kevrekidis[3,4], Lu Lu[5], Paris Perdikaris[6], Sifan Wang[7] and Liu Yang[1]

Abstract | Despite great progress in simulating multiphysics problems using the numerical discretization of partial differential equations (PDEs), one still cannot seamlessly incorporate noisy data into existing algorithms, mesh generation remains complex, and high-dimensional problems governed by parameterized PDEs cannot be tackled. Moreover, solving inverse problems with hidden physics is often prohibitively expensive and requires different formulations and elaborate computer codes. Machine learning has emerged as a promising alternative, but training deep neural networks requires big data, not always available for scientific problems. Instead, such networks can be trained from additional information obtained by enforcing the physical laws (for example, at random points in the continuous space-time domain). Such physics-informed learning integrates (noisy) data and mathematical models, and implements them through neural networks or other kernel-based regression networks. Moreover, it may be possible to design specialized network architectures that automatically satisfy some of the physical invariants for better accuracy, faster training and improved generalization. Here, we review some of the prevailing trends in embedding physics into machine learning, present some of the current capabilities and limitations and discuss diverse applications of physics-informed learning both for forward and inverse problems, including discovering hidden physics and tackling high-dimensional problems.

Fusion between rationalism & empiricism (deduction & induction)

- **ML × Simulation**
- **ML × Theoretical Chemistry/Physics**
- **ML × Logic & Symbol Manipulations**



Data Assimilation
Simulation with Prediction

Sim2Real
Geometric ML

Summary

Three lessons learned as I experienced **this illusion** being shattered…

1. The goals of ML and ‘materials/chemical science’ are **fundamentally different**. What we need here is not ML but a much harder problem of ‘machine discovery’
2. If we go for a hypothesis-free + off-the-shelf solution, exploration by **decision tree ensembles**, combined with UQ and abstracted (coarse grained) feature representations, will give a very strong baseline.
3. If we want more that that, we **can’t be hypothesis free**. Any strategies to narrow down the scope as well as domain expertise really matters.

PDF of this slide: https://itakigawa.page.link/acs2023spring