

Machine Learning for Molecular Graph Representations and Geometries

December 1st, 2021

Ichigaku Takigawa

<https://itakigawa.github.io/>

RIKEN Center for Advanced Intelligence Project (AIP)

Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University





TAKIGAWA Ichigaku
瀧川 一学

<https://itakigawa.github.io>

Hi, I am a machine-learning researcher

- 1995-2004 Hokkaido Univ (Grad School. Engineering)
2004 PhD Computer Science
- 2005-2011 Kyoto Univ (Inst. Chemical Research)
Bioinformatics Center, Assist. Prof.
Grad School Pharmaceutical Sciences, Assit. Prof.
- 2012-2018 Hokkaido Univ (Grad School. Info Sci & Tech)
Large-Scale Knowledge Processing Lab, Assoc. Prof.
2015-2018 JST PRESTO for Materials Informatics
- 2019- RIKEN Center for AI Project
2019- Hokkaido Univ
(Inst. Chemical Reaction Design & Discovery)
I belong to a joint research team based at Kyoto
with RIKEN AIP and Kyoto Univ CiRA,
working on stem cell biology.



TAKIGAWA Ichigaku
瀧川 一学

<https://itakigawa.github.io>

But also, I am a machine-learning user

- 1995-2004 Hokkaido Univ (Grad School. Engineering)
2004 PhD Computer Science
- 2005-2011 Kyoto Univ (Inst. Chemical Research)
Bioinformatics Center, Assist. Prof.
Grad School Pharmaceutical Sciences, Assit. Prof.
- 2012-2018 Hokkaido Univ (Grad School. Info Sci & Tech)
Large-Scale Knowledge Processing Lab, Assoc. Prof.
2015-2018 JST PRESTO for Materials Informatics
- 2019- RIKEN Center for AI Project
2019- Hokkaido Univ
(Inst. Chemical Reaction Design & Discovery)
I belong to a joint research team based at Kyoto
with RIKEN AIP and Kyoto Univ CiRA,
working on stem cell biology.

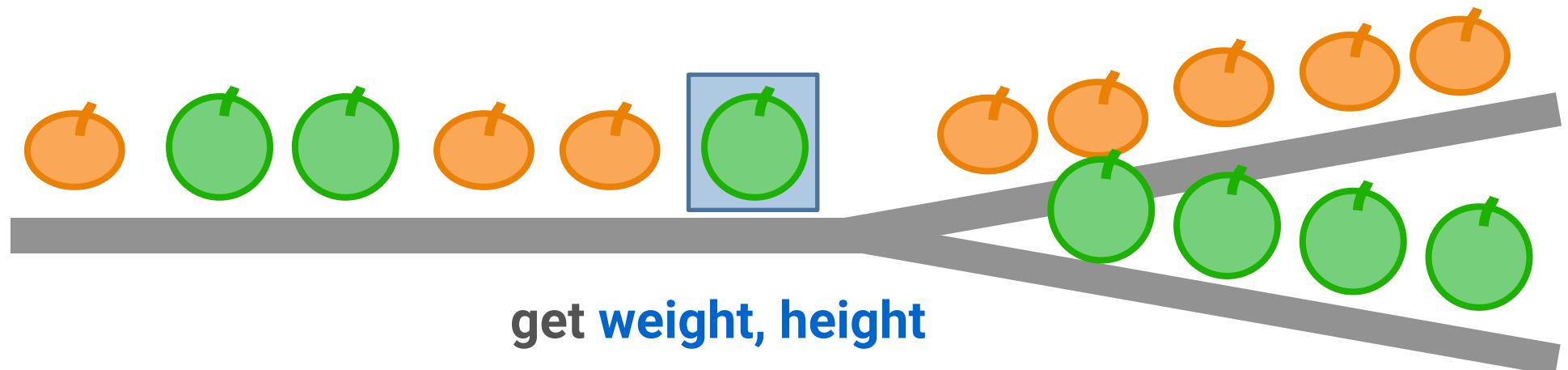
This talk

Machine Learning (ML) for Molecules

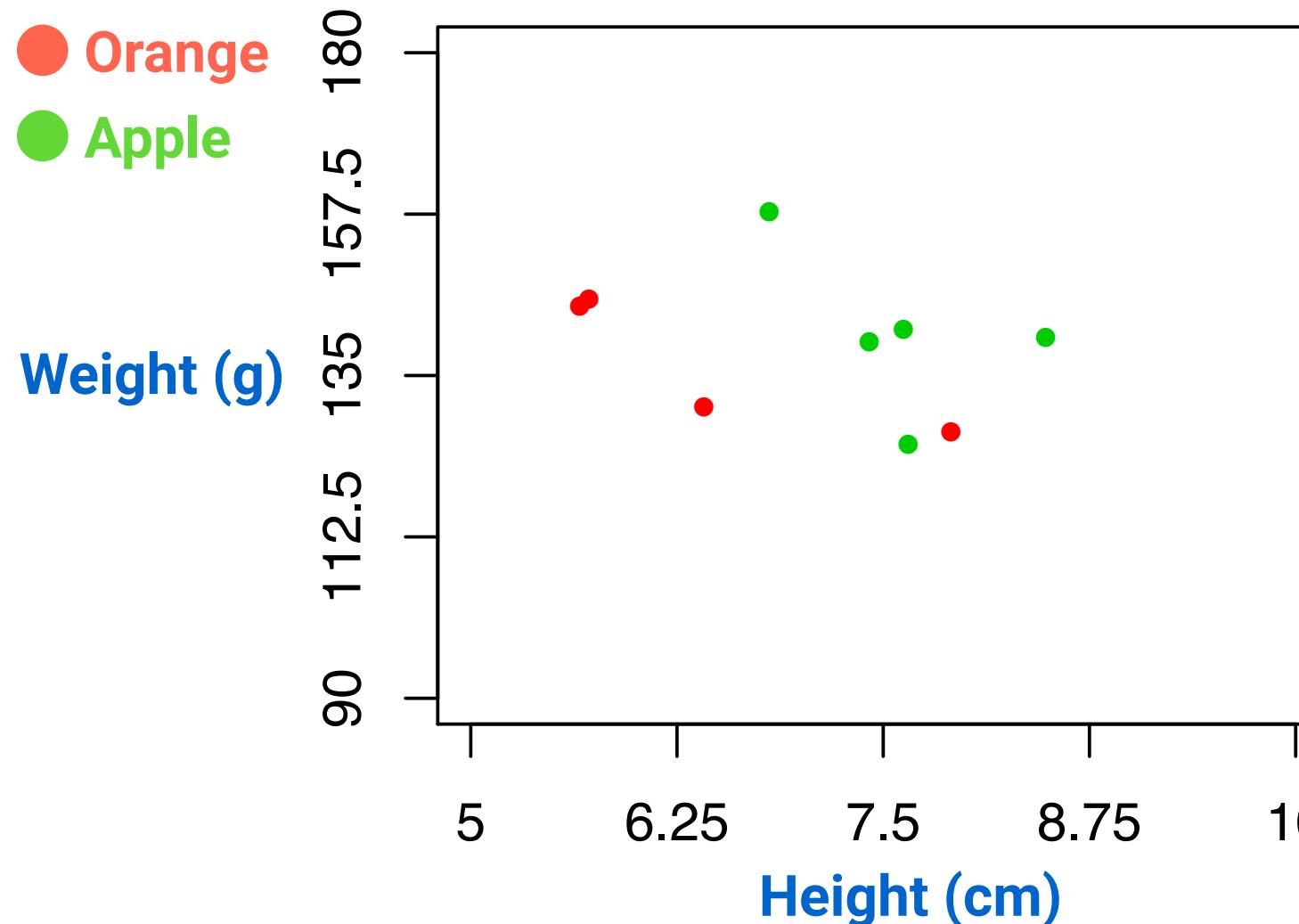
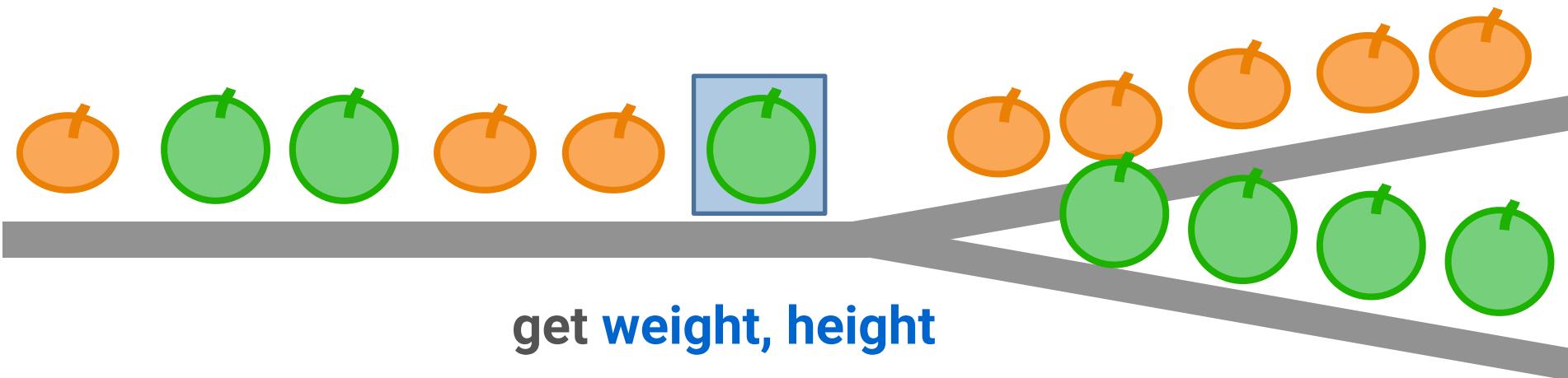
1. ML in a nutshell
2. The dark side: Modern aspects of ML
3. The light side: Deep learning for molecules
4. Challenges

May the ML Force be with you...

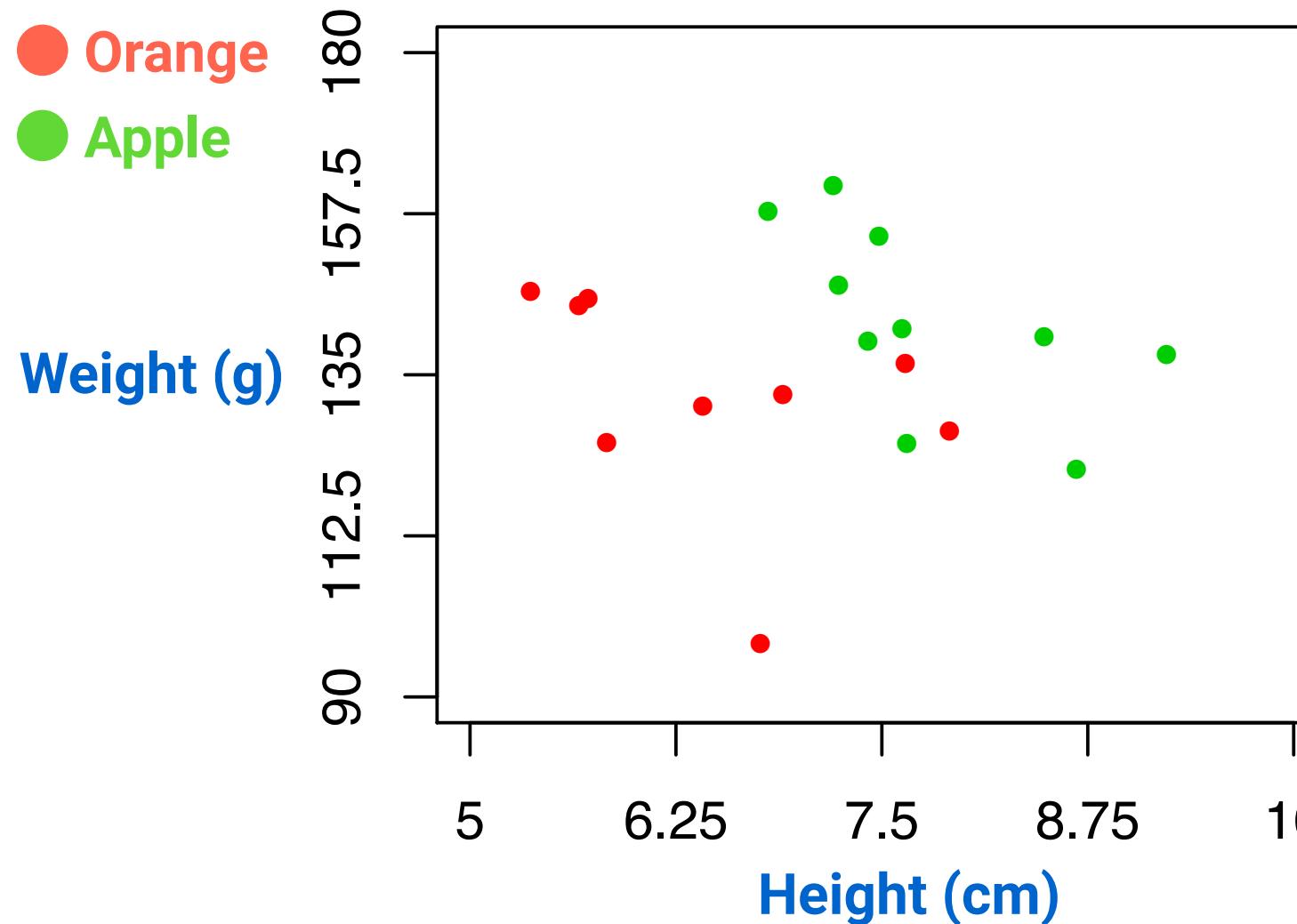
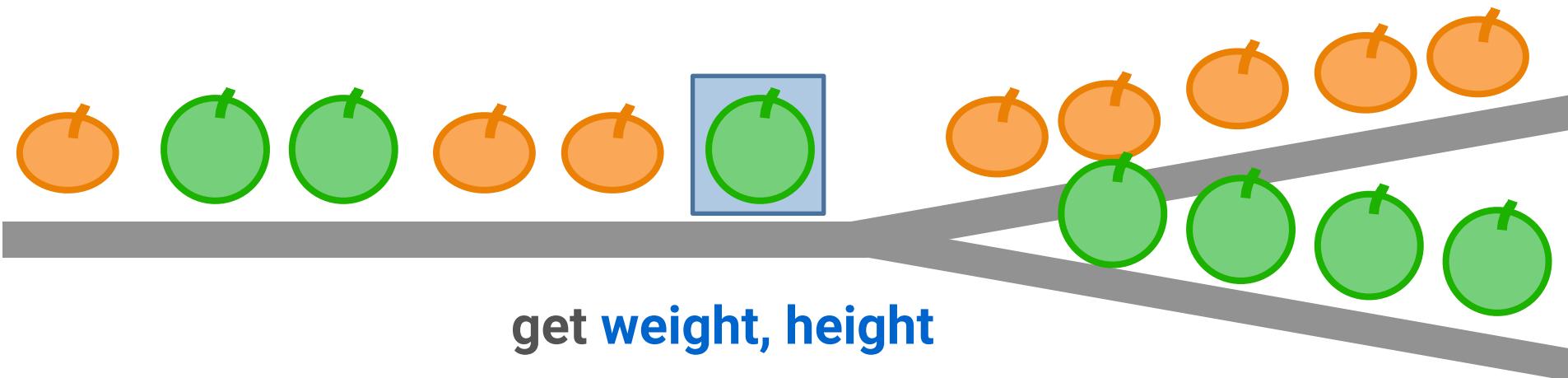
ML converts data into "prediction"



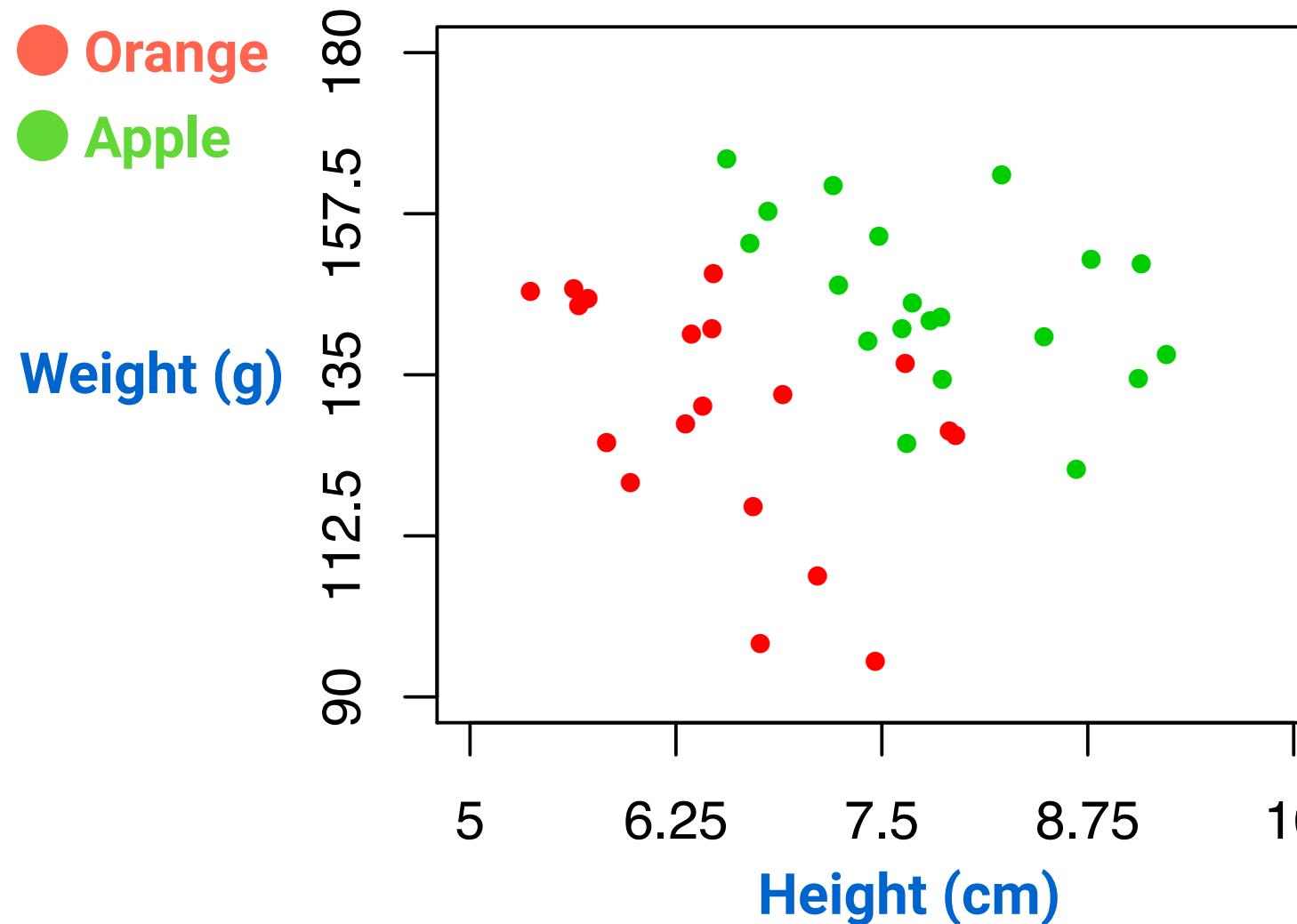
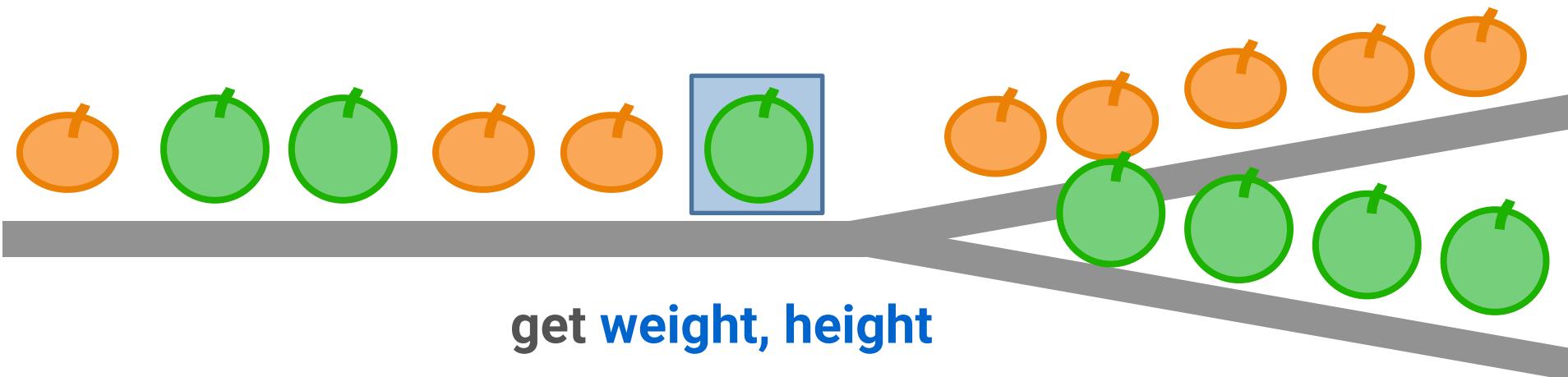
ML converts data into "prediction"



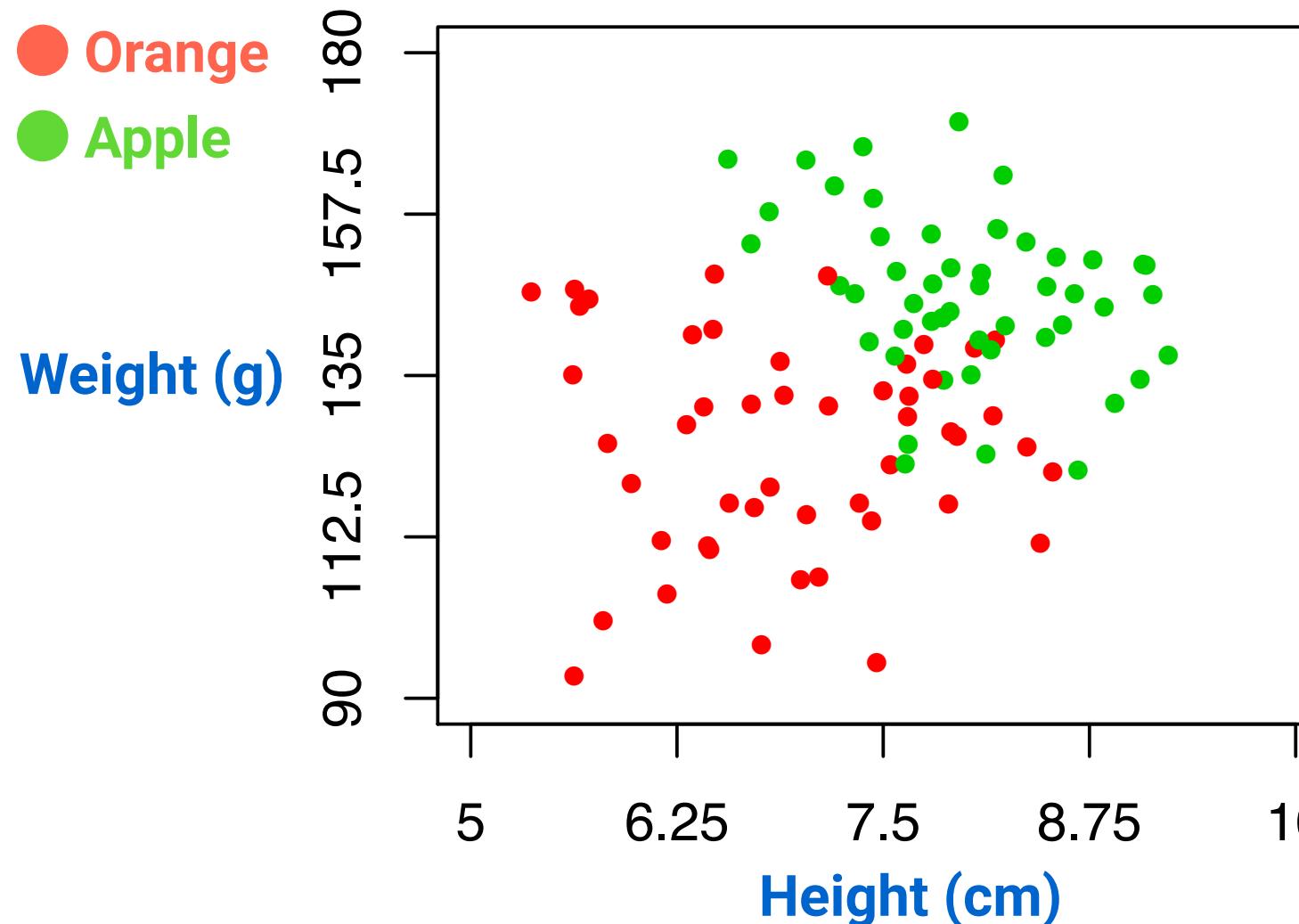
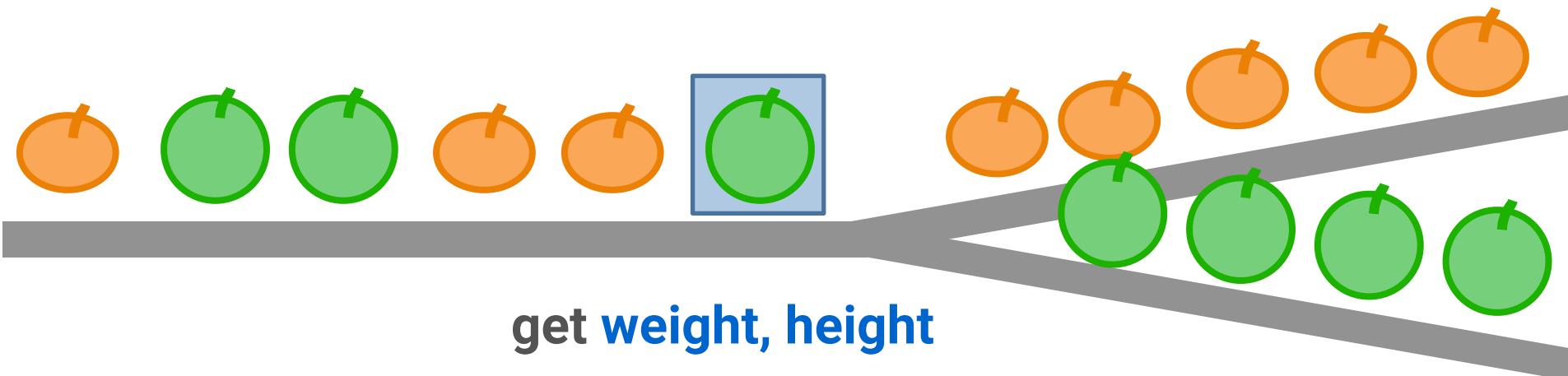
ML converts data into "prediction"



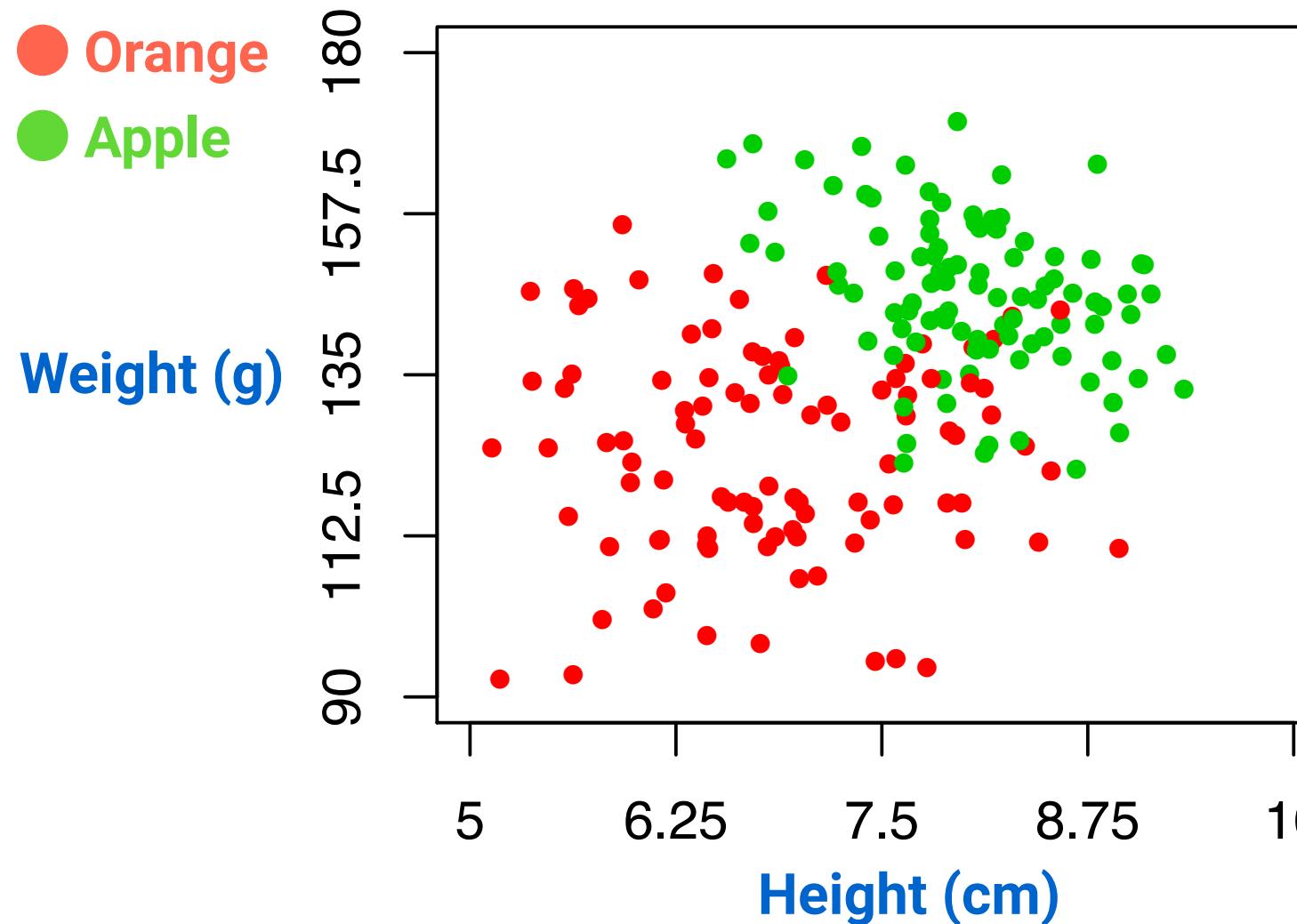
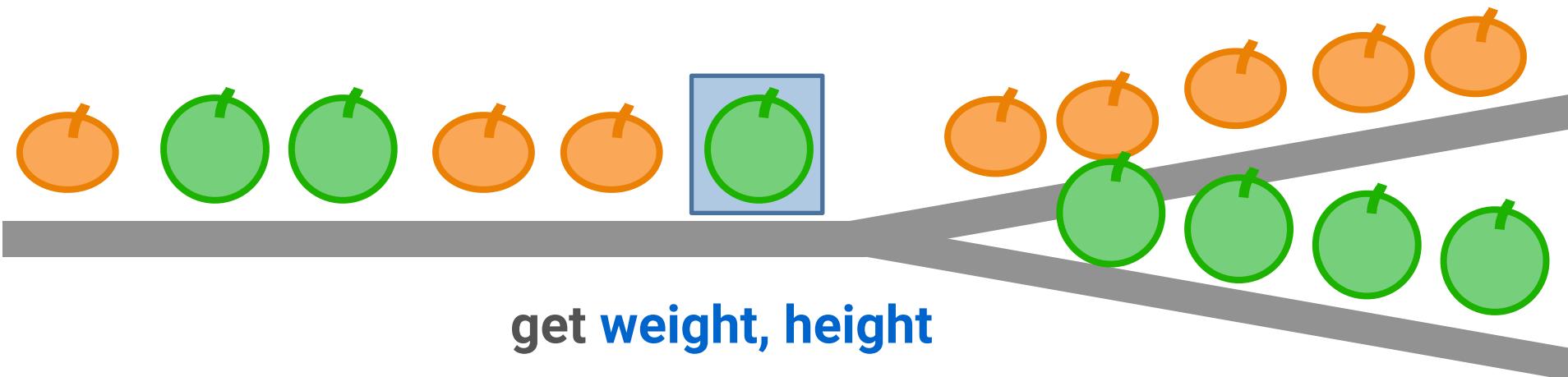
ML converts data into "prediction"



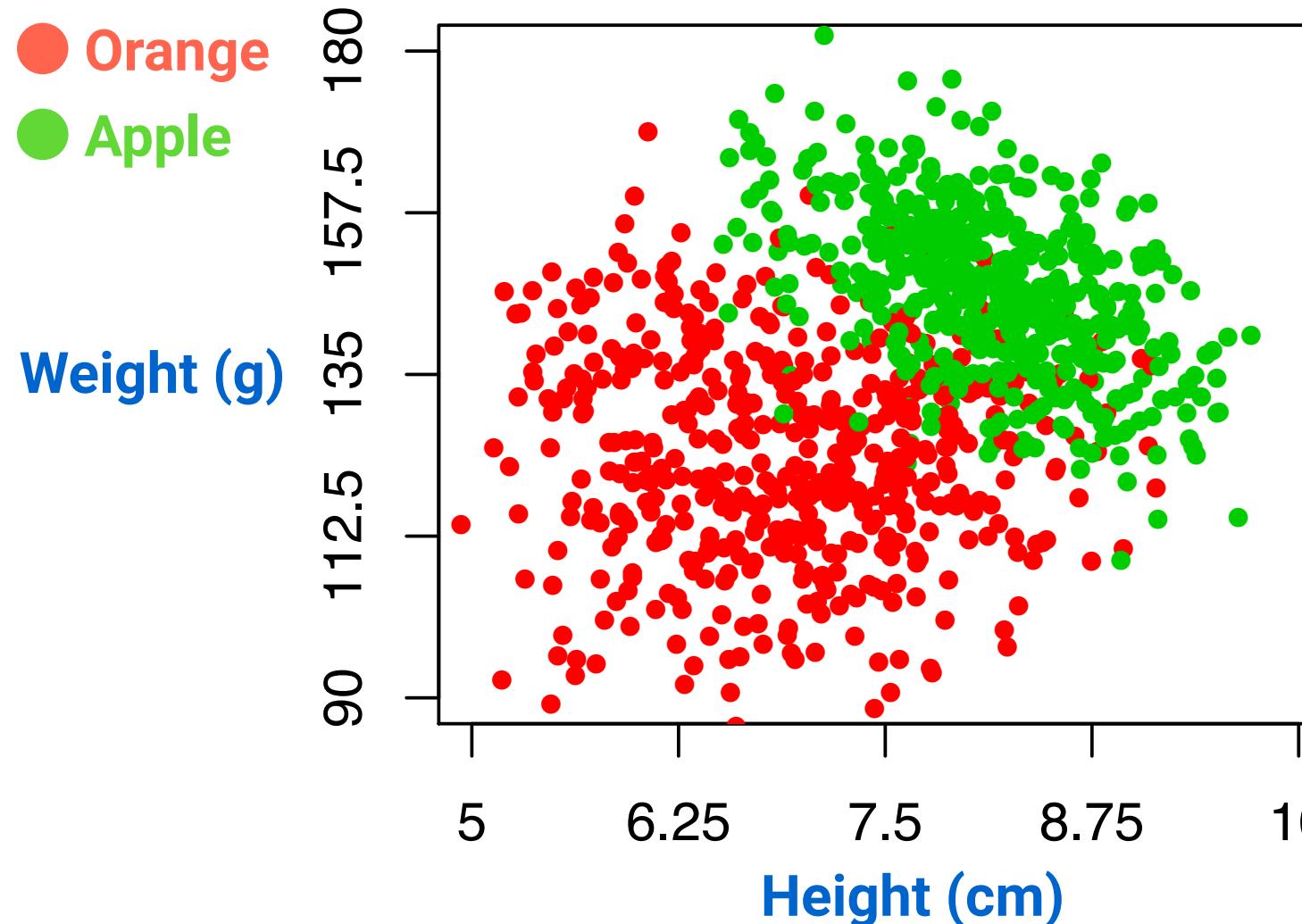
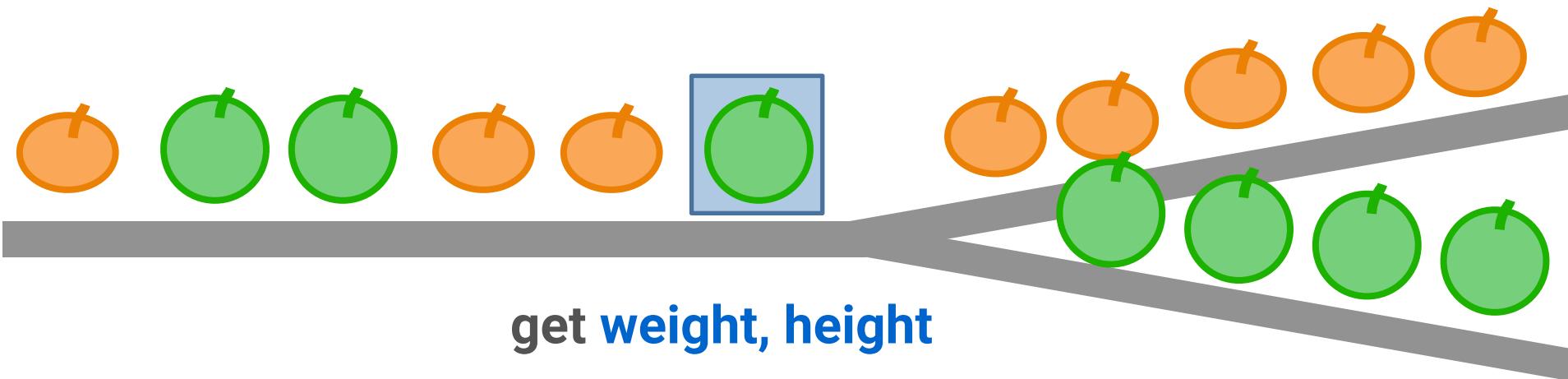
ML converts data into "prediction"



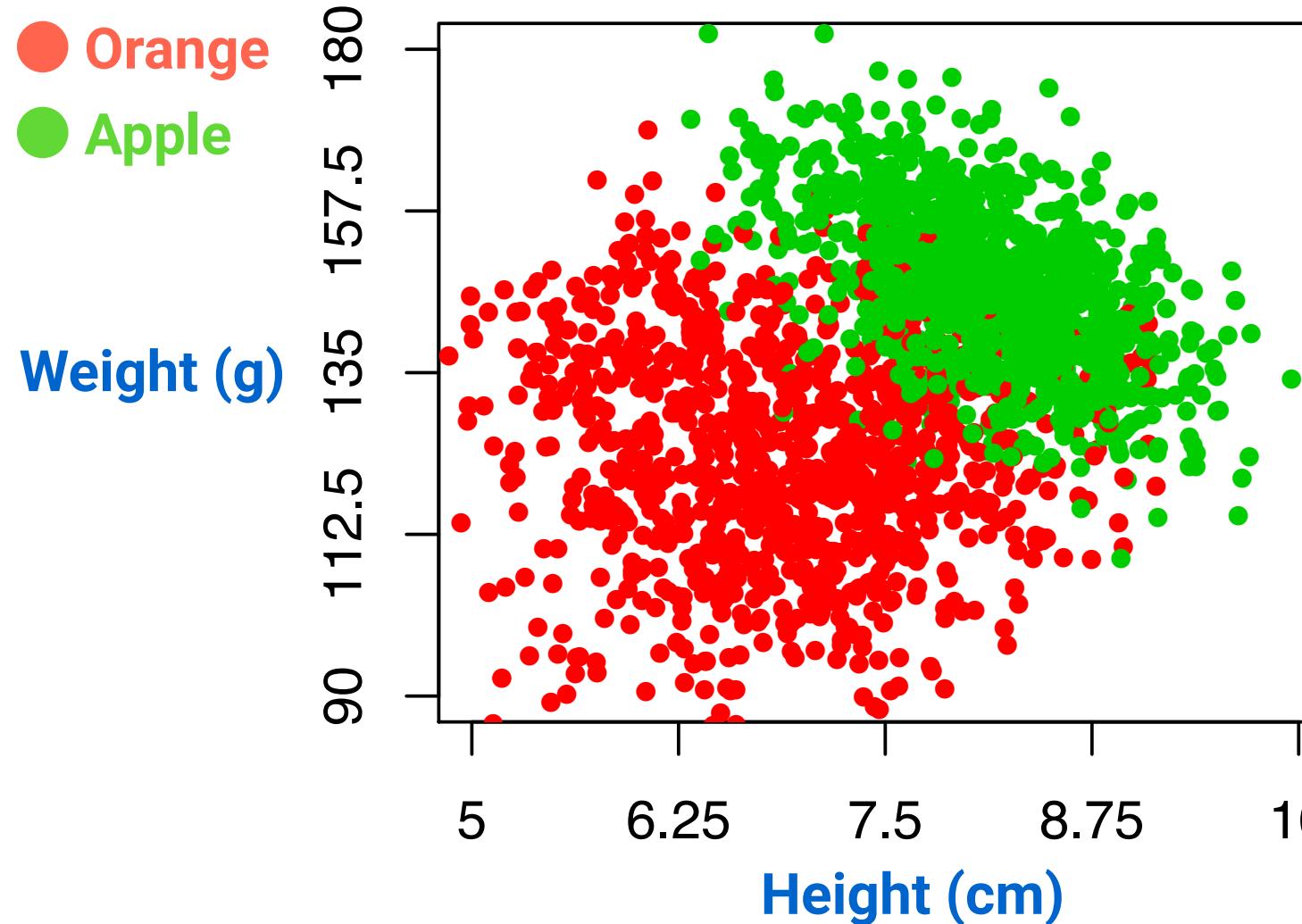
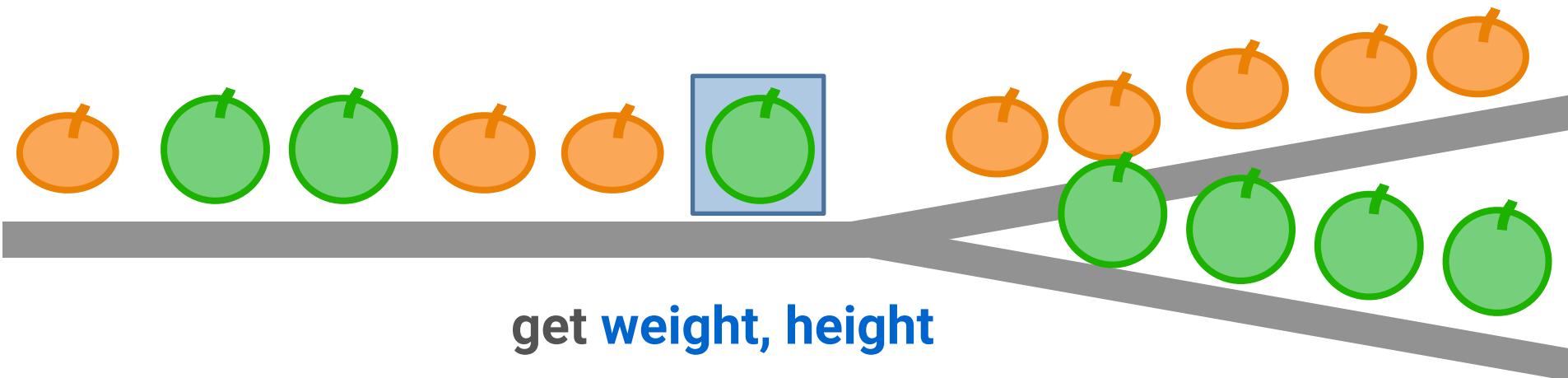
ML converts data into "prediction"



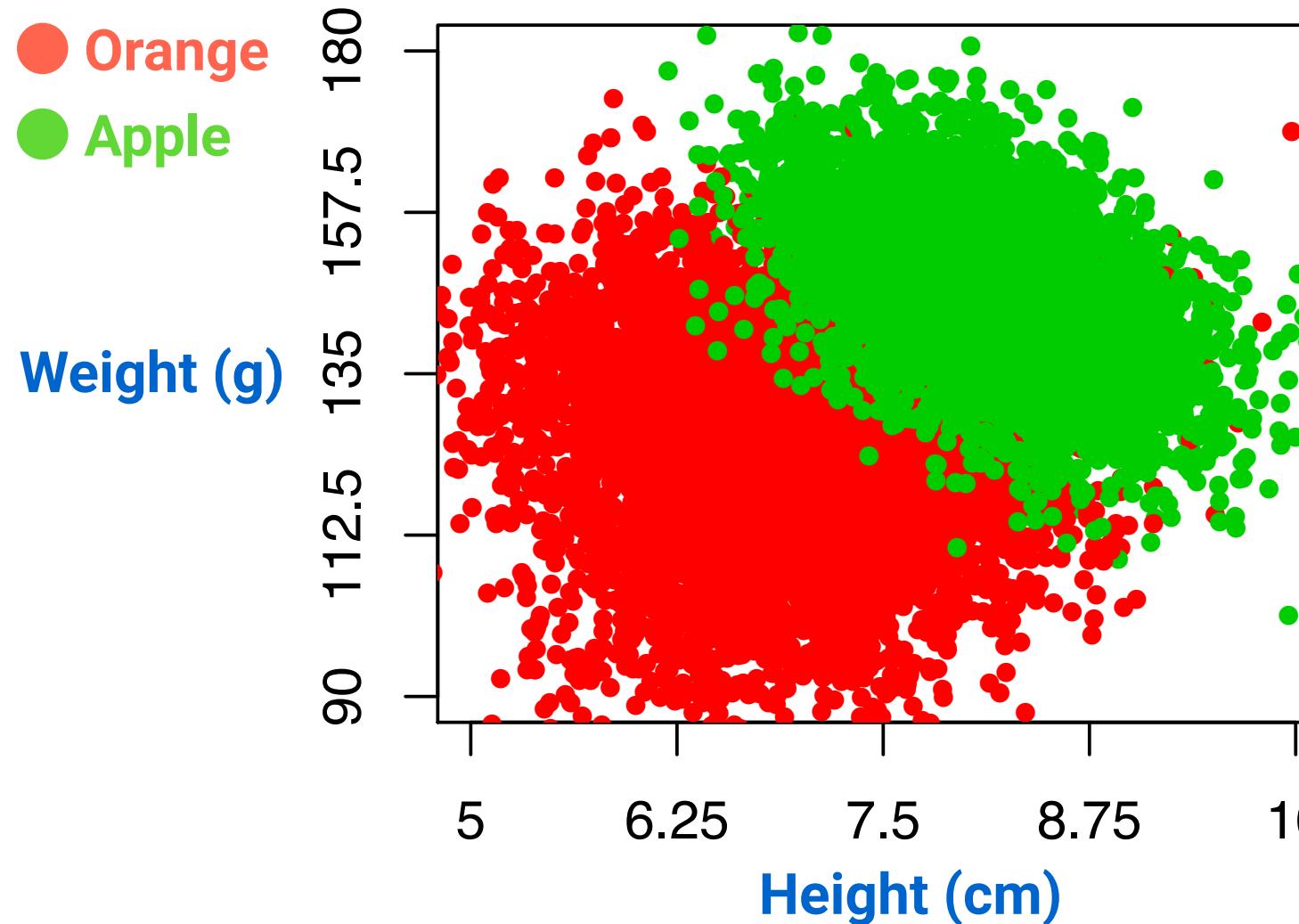
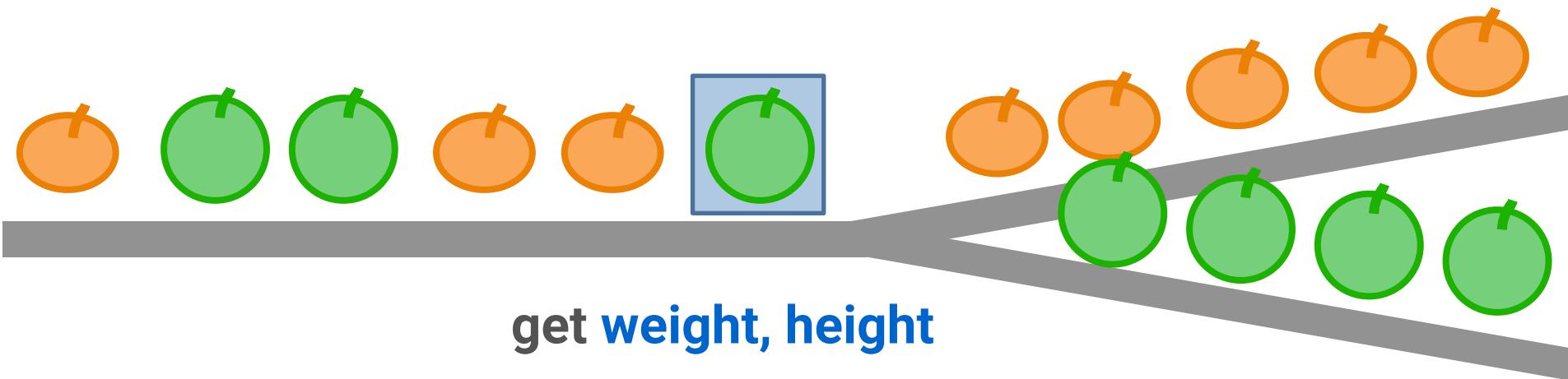
ML converts data into "prediction"



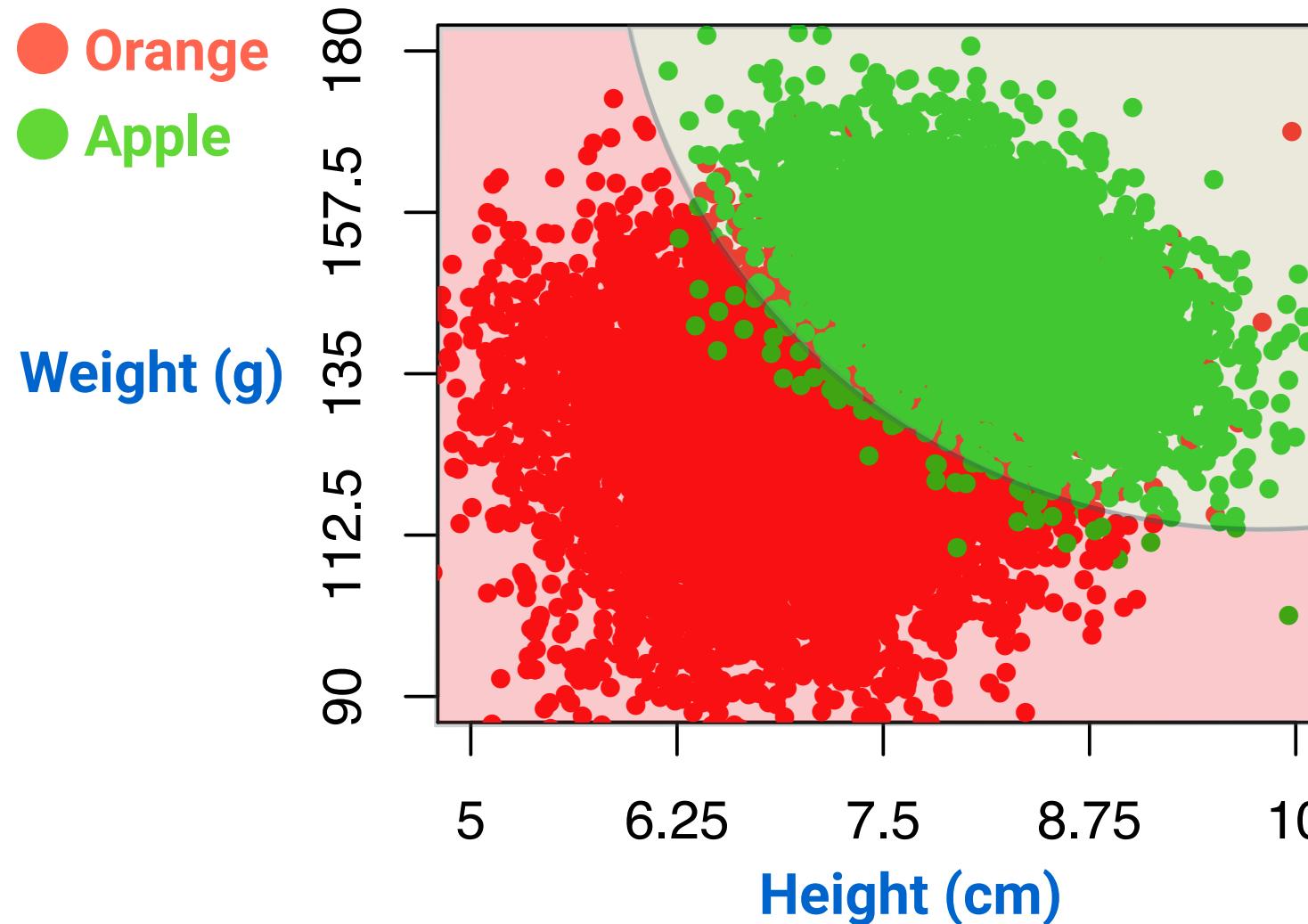
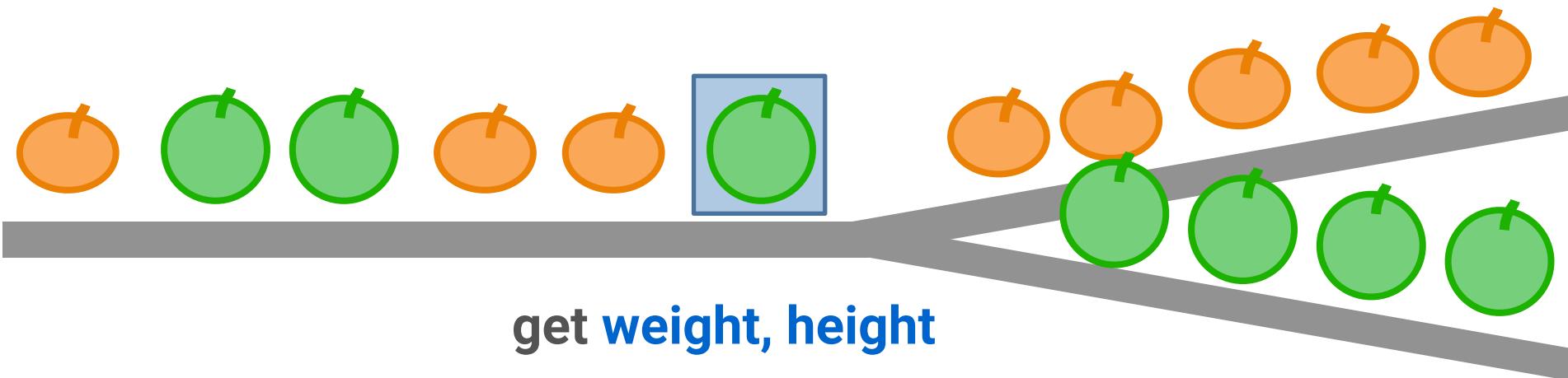
ML converts data into "prediction"



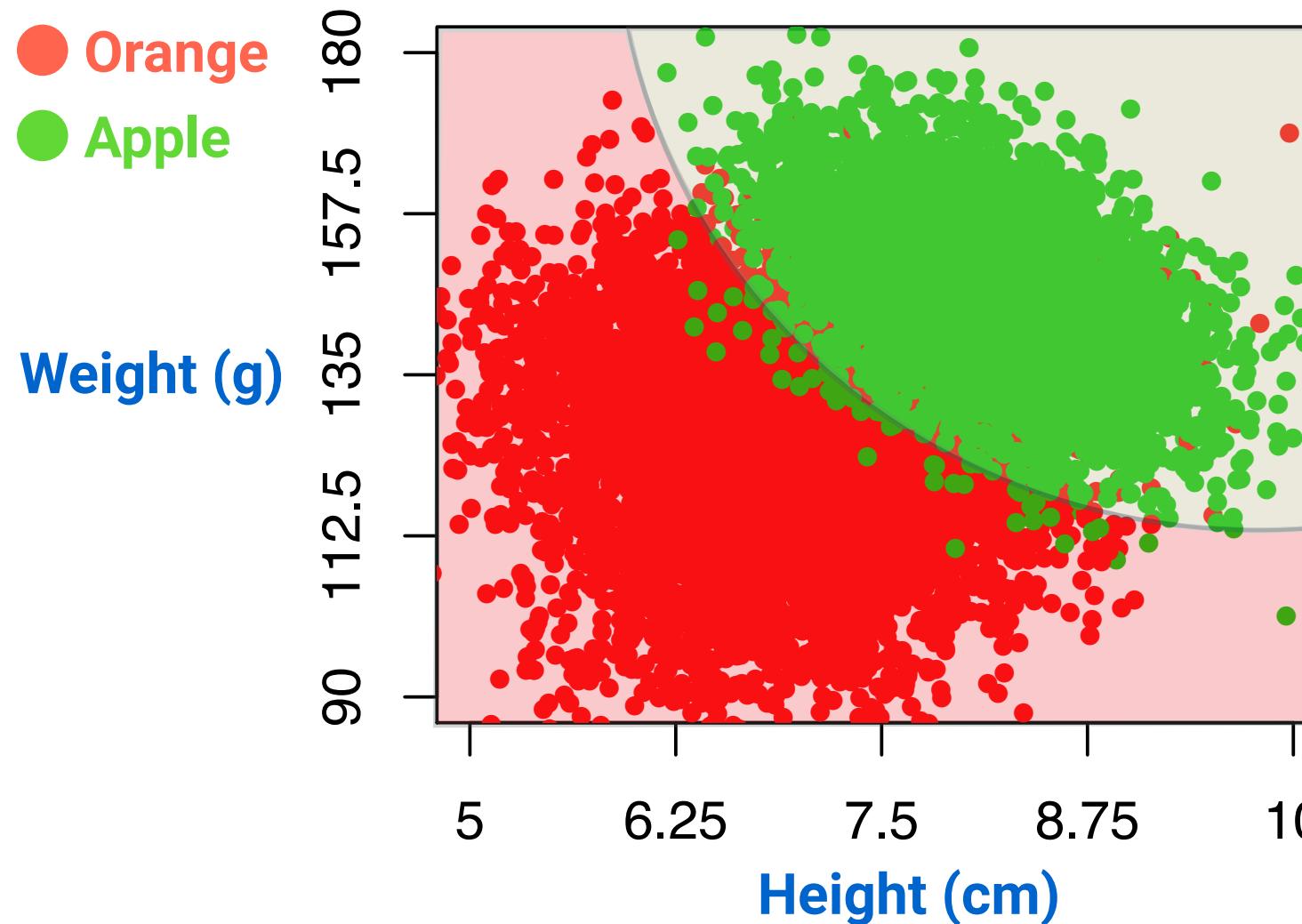
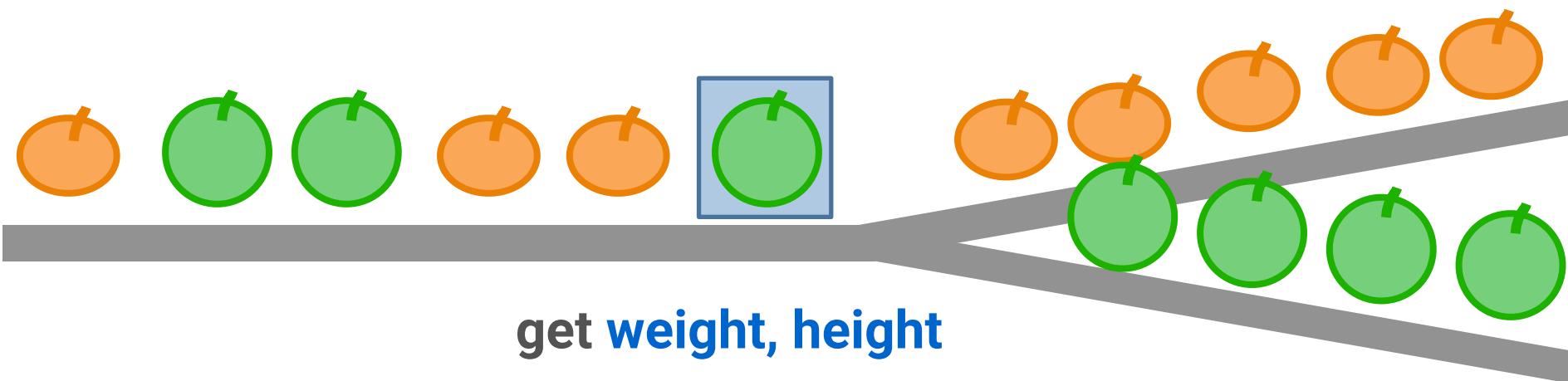
ML converts data into "prediction"



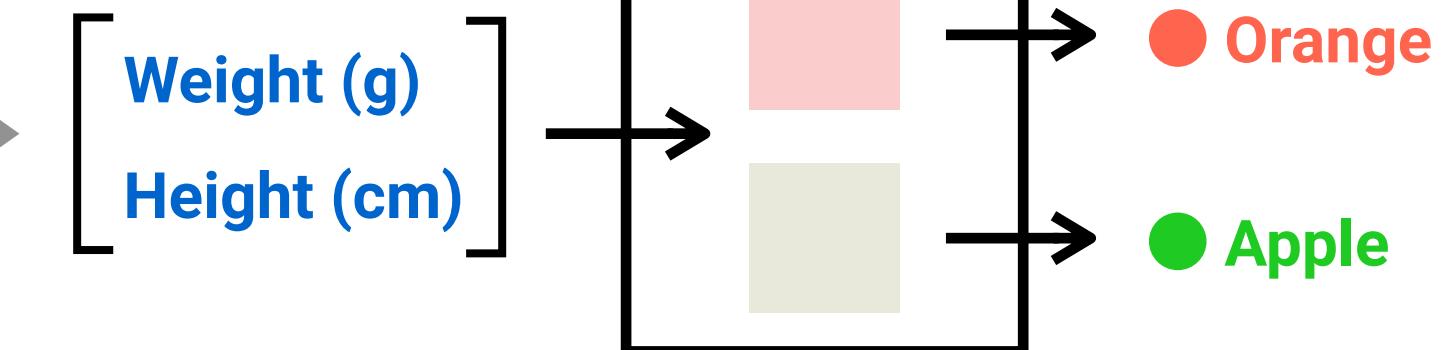
ML converts data into "prediction"



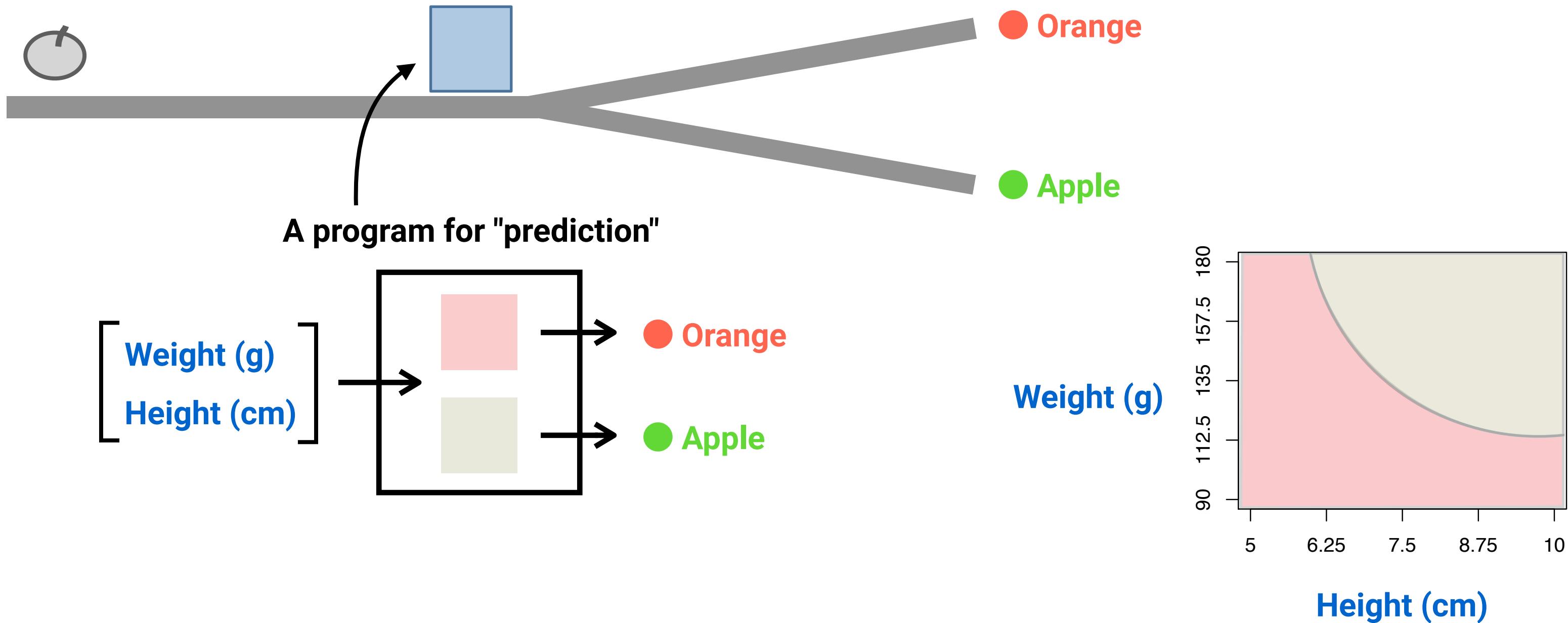
ML converts data into "prediction"



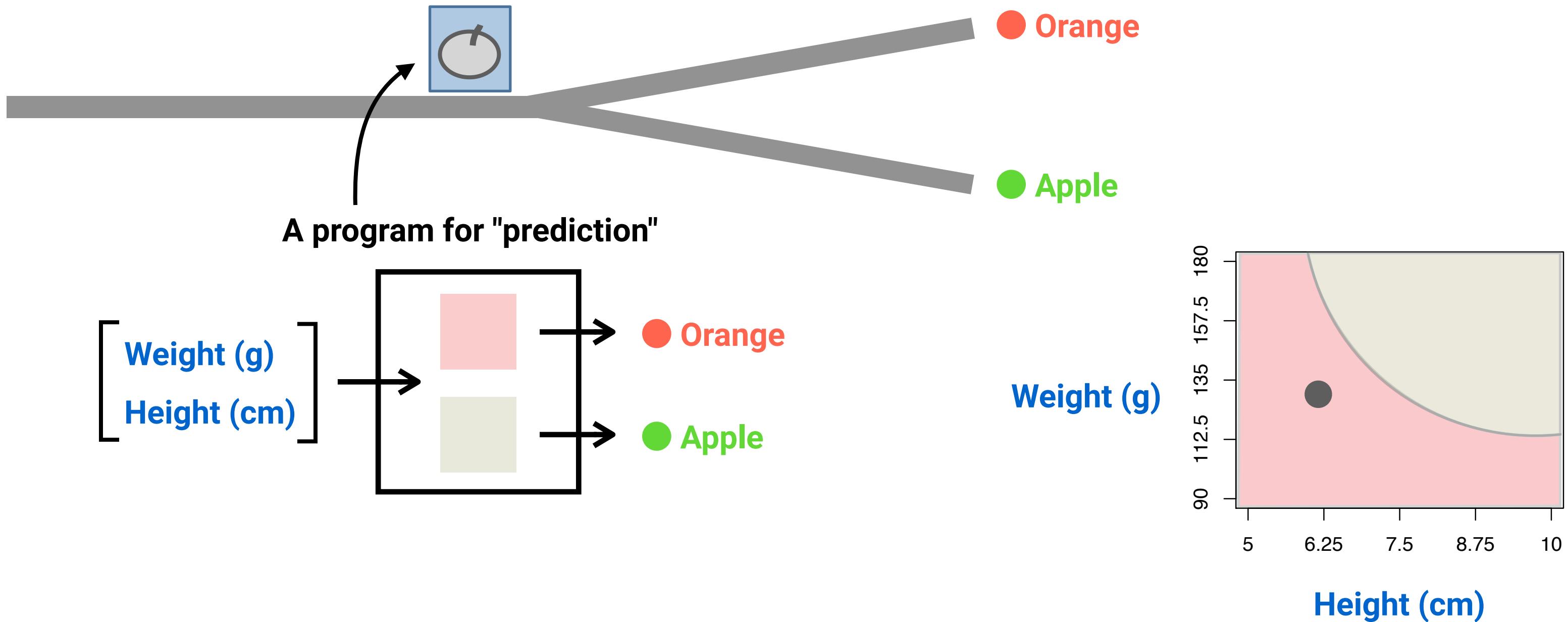
A program for "prediction"



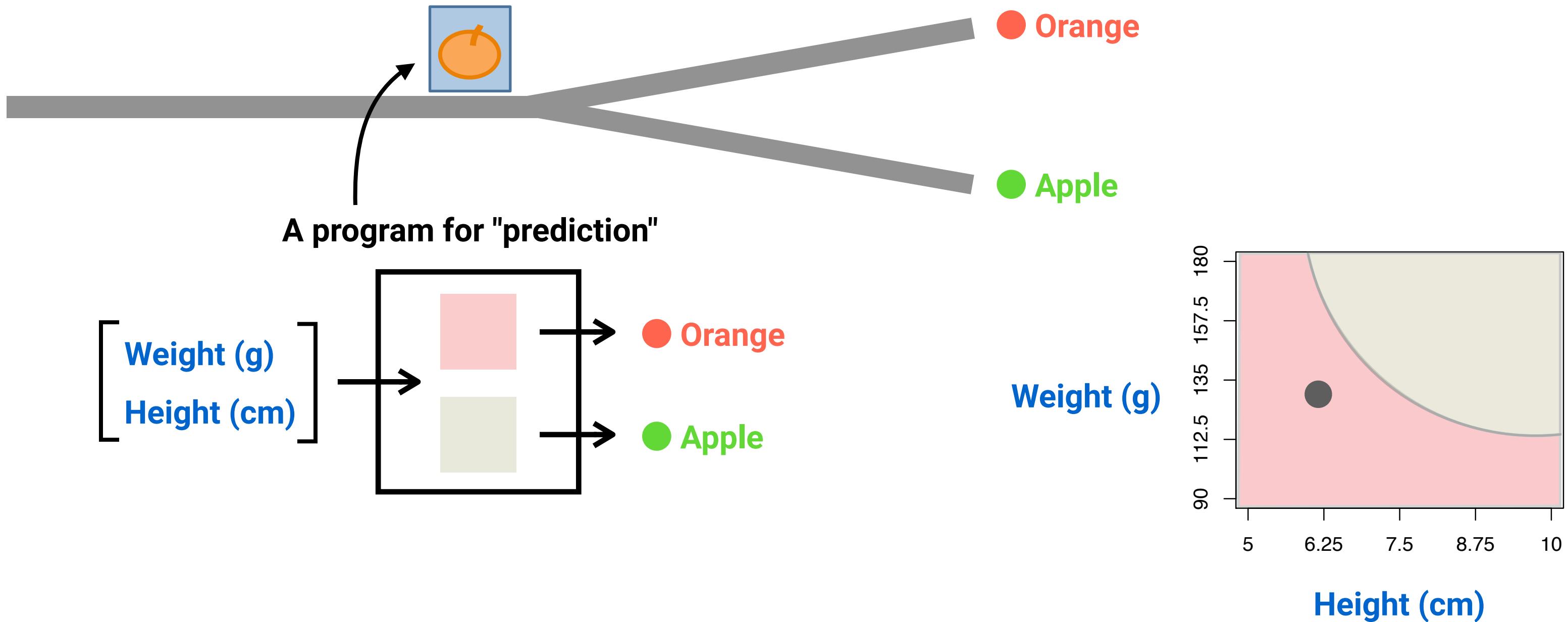
ML converts data into "prediction"



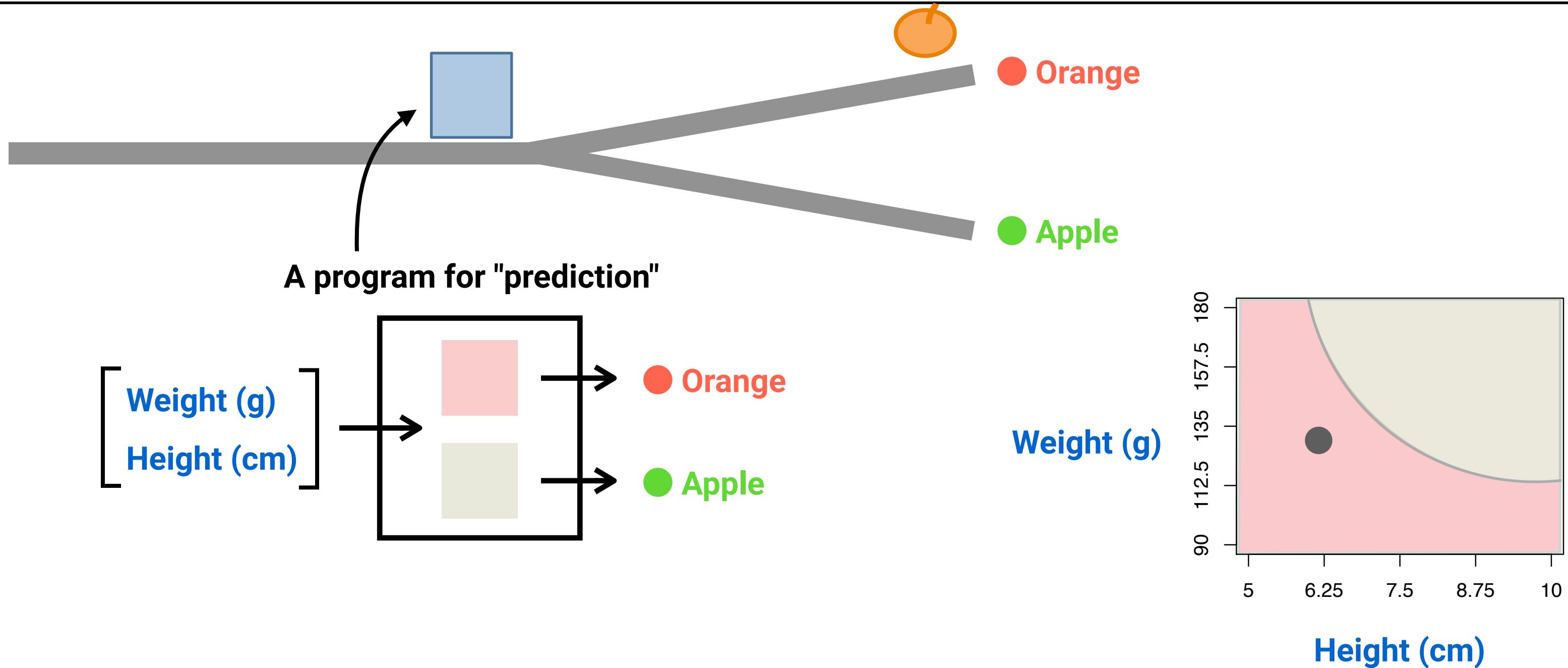
ML converts data into "prediction"



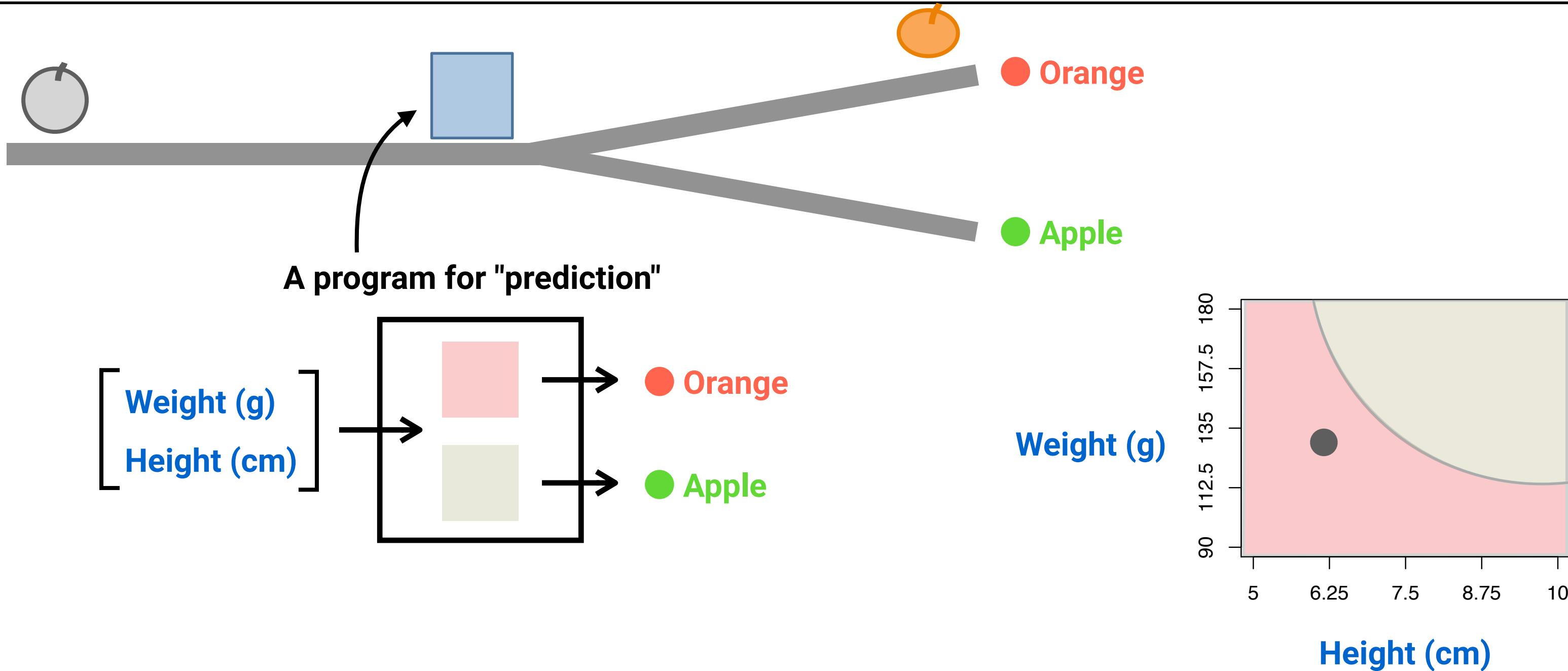
ML converts data into "prediction"



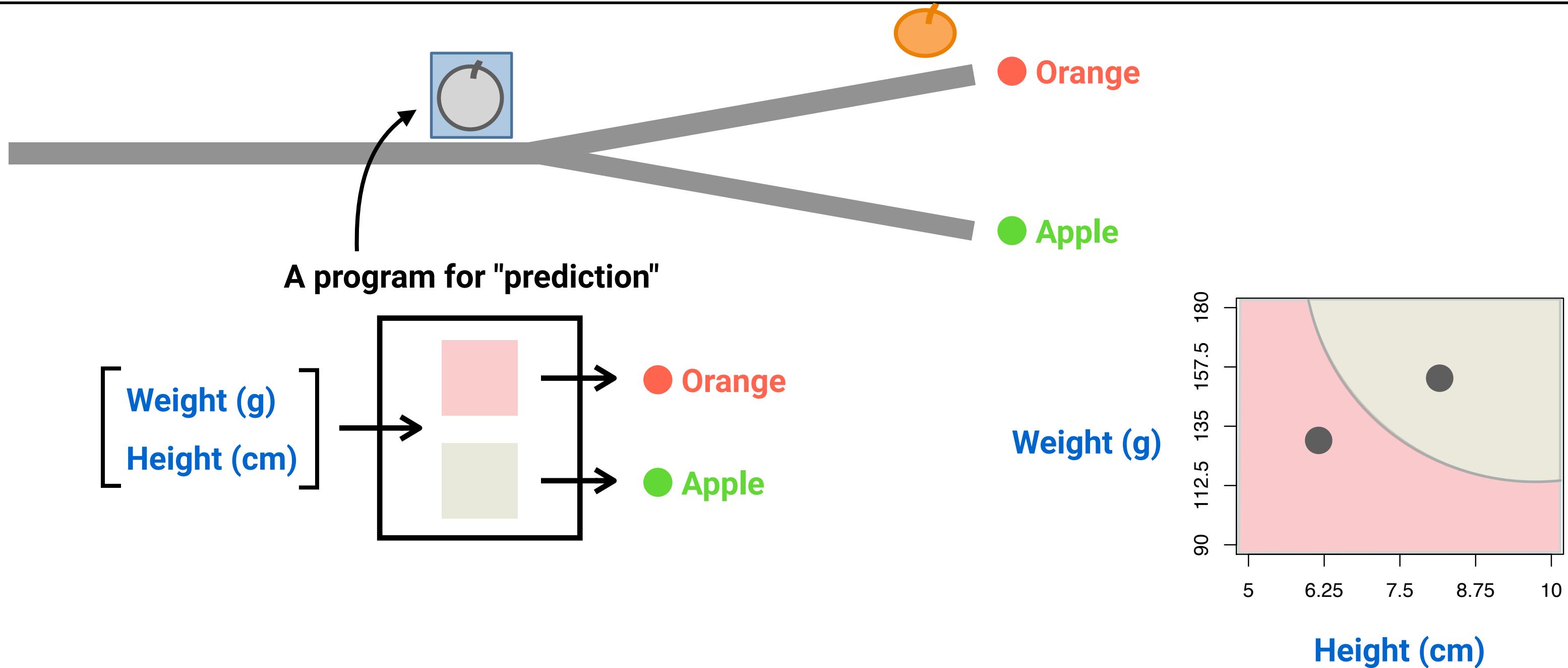
ML converts data into "prediction"



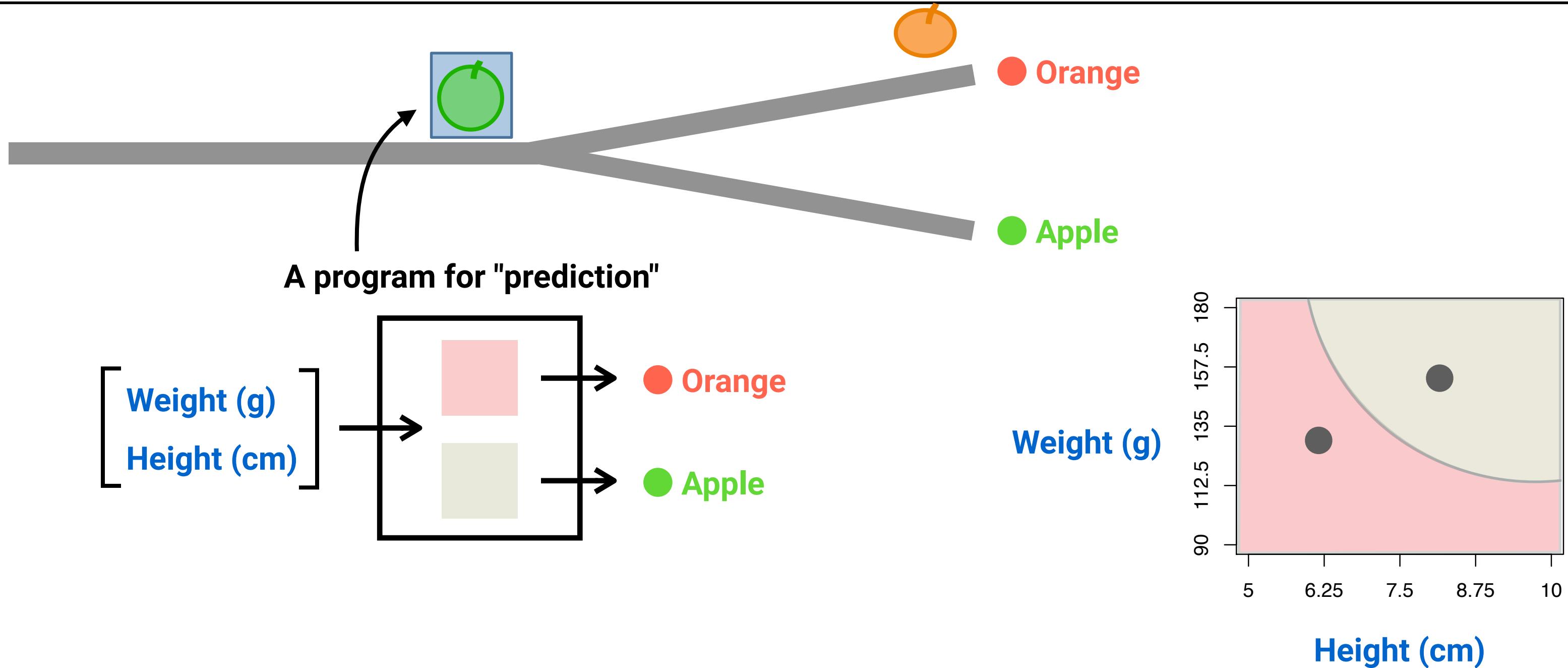
ML converts data into "prediction"



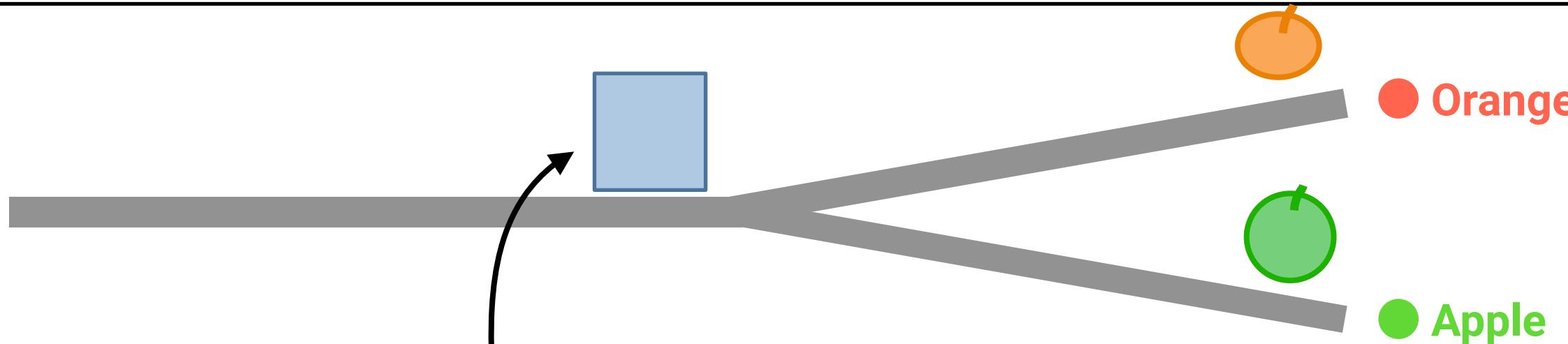
ML converts data into "prediction"



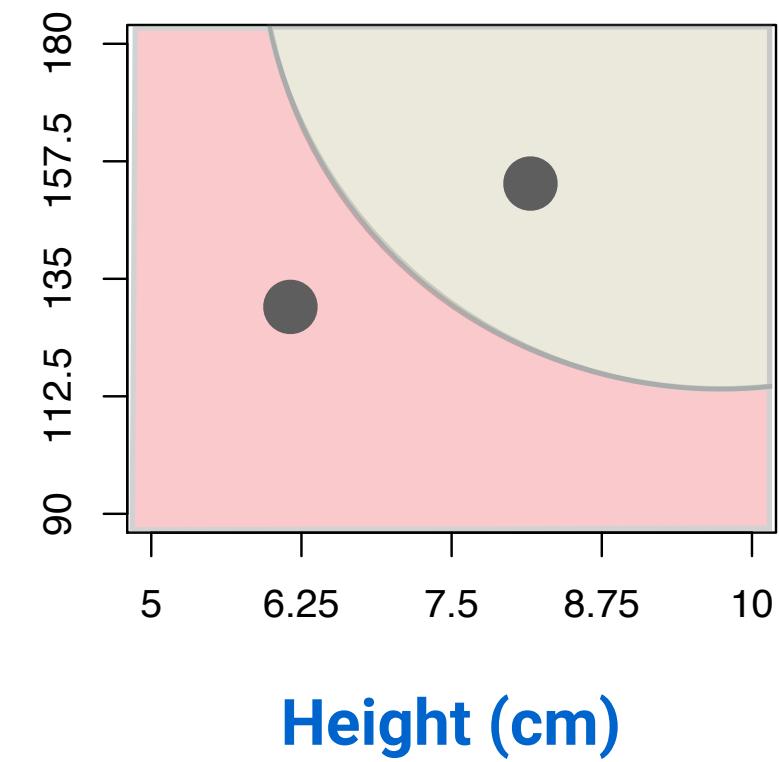
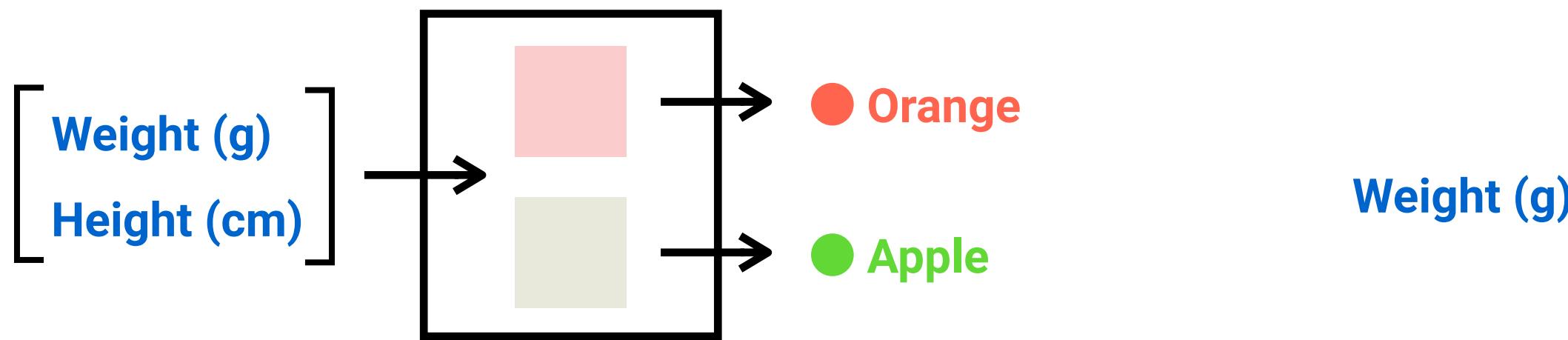
ML converts data into "prediction"



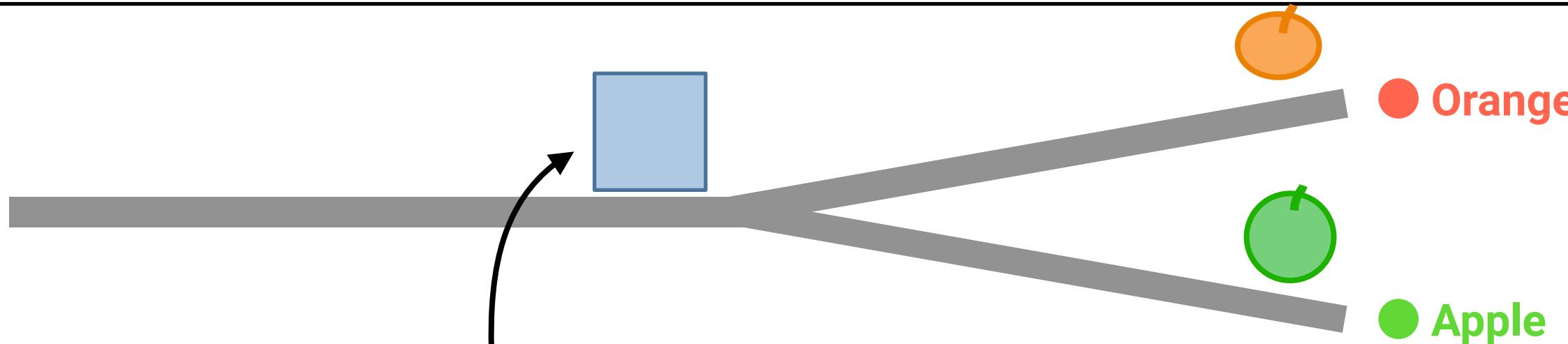
ML converts data into "prediction"



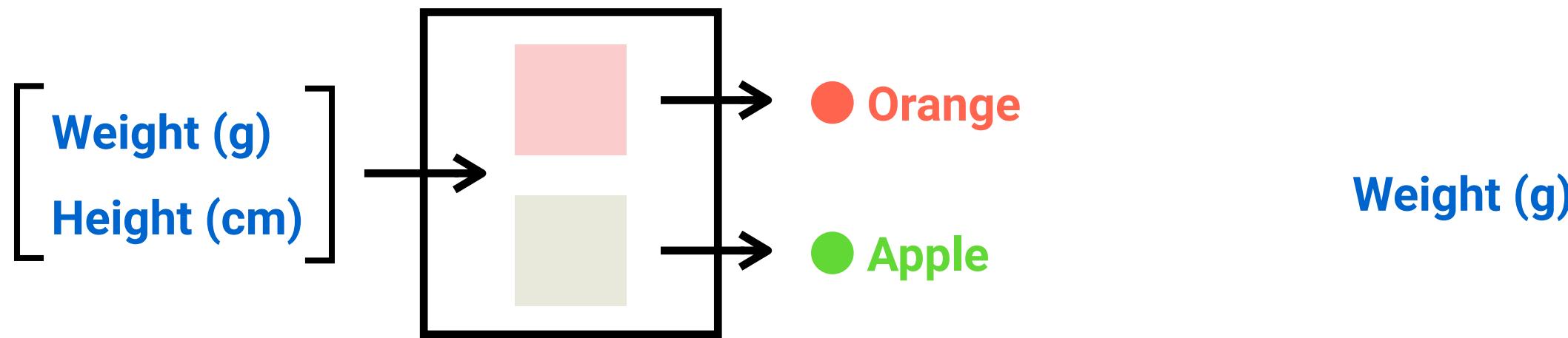
A program for "prediction"



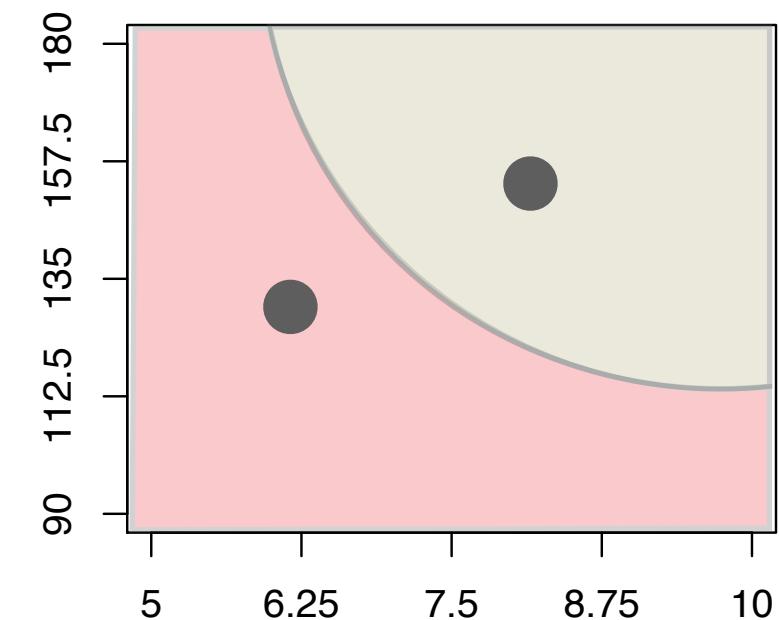
ML converts data into "prediction"



A program for "prediction"

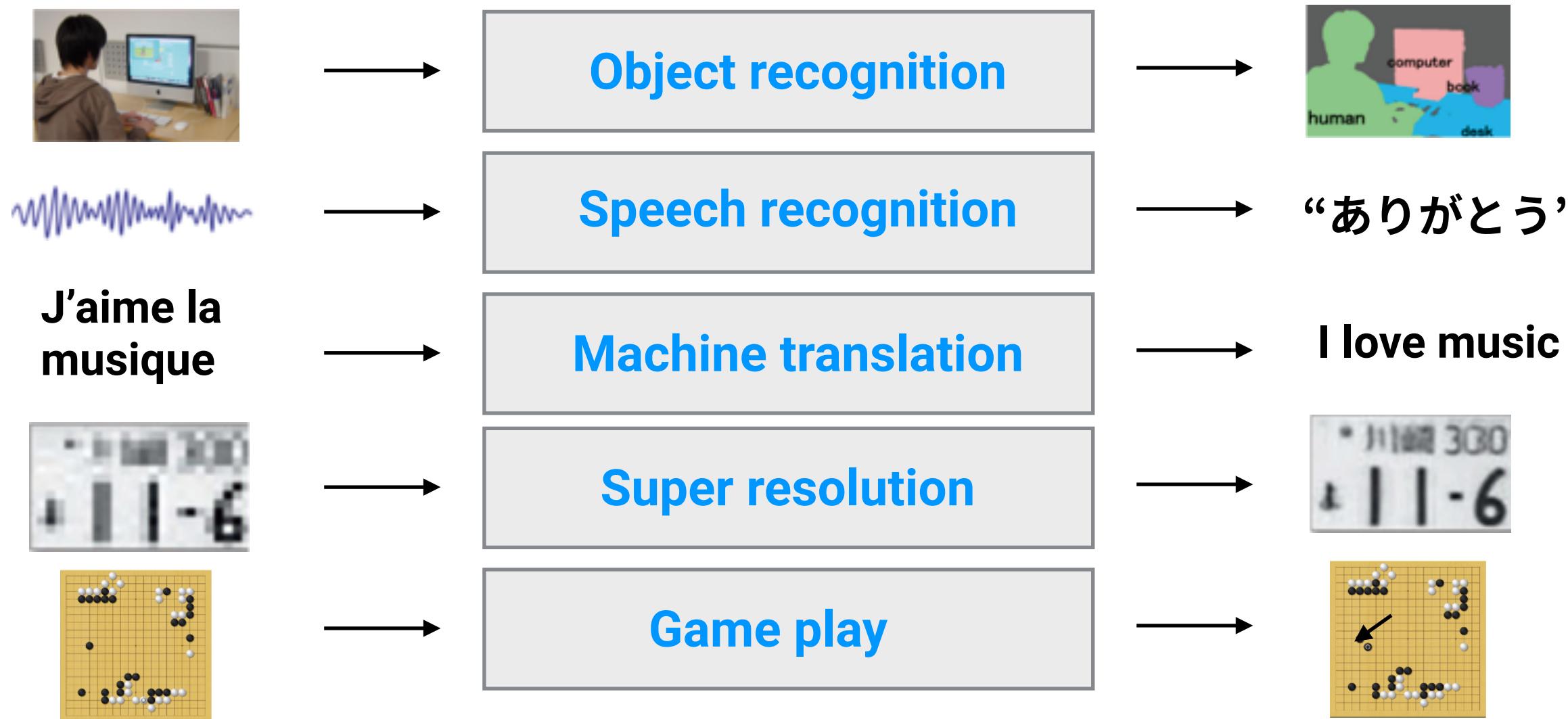


Now we got a computer program to predict "orange or apple" for any **unseen** ones **directly from collected data**



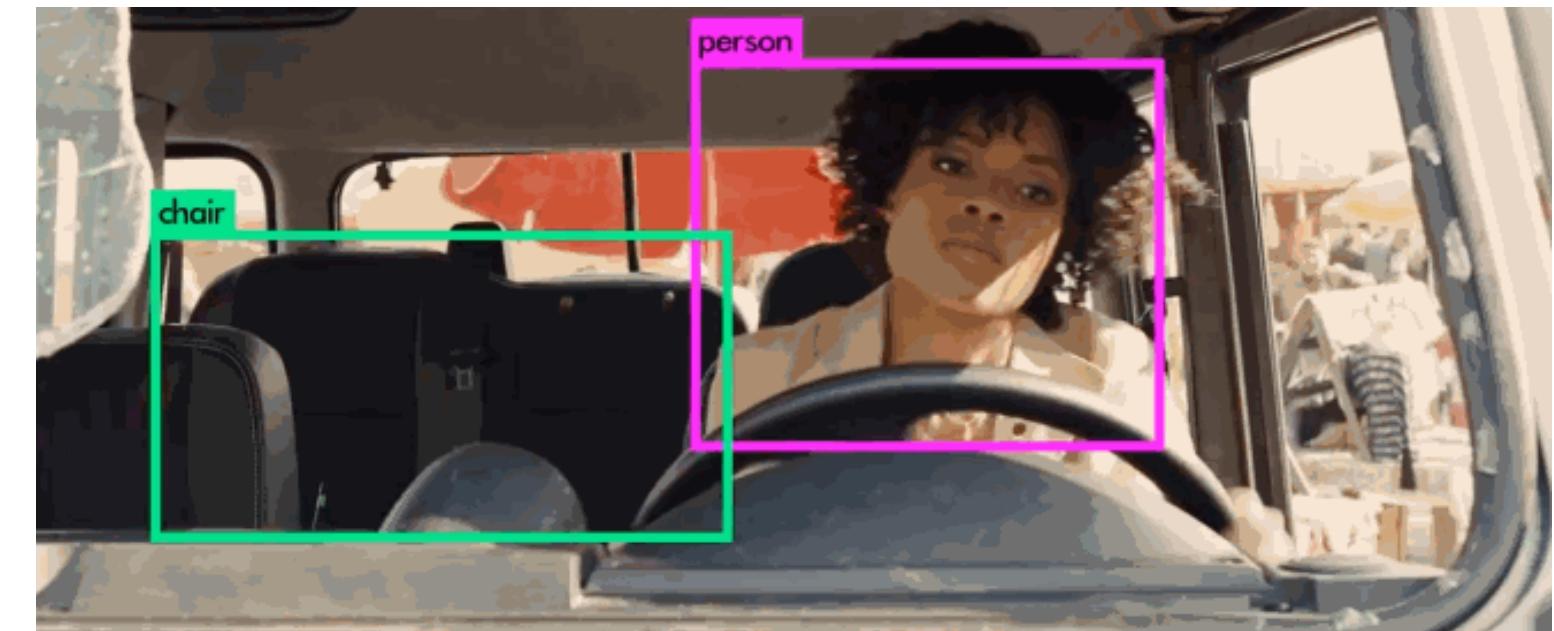
ML is a new (lazy) way of programming

ML generates a computer program just **by giving many input-output examples** even when we **don't know** the underlying mechanism between inputs and outputs.



This simple idea is more powerful than you may think

Remarkably powerful when we have relevant input-output examples (**it's useless if we don't**)



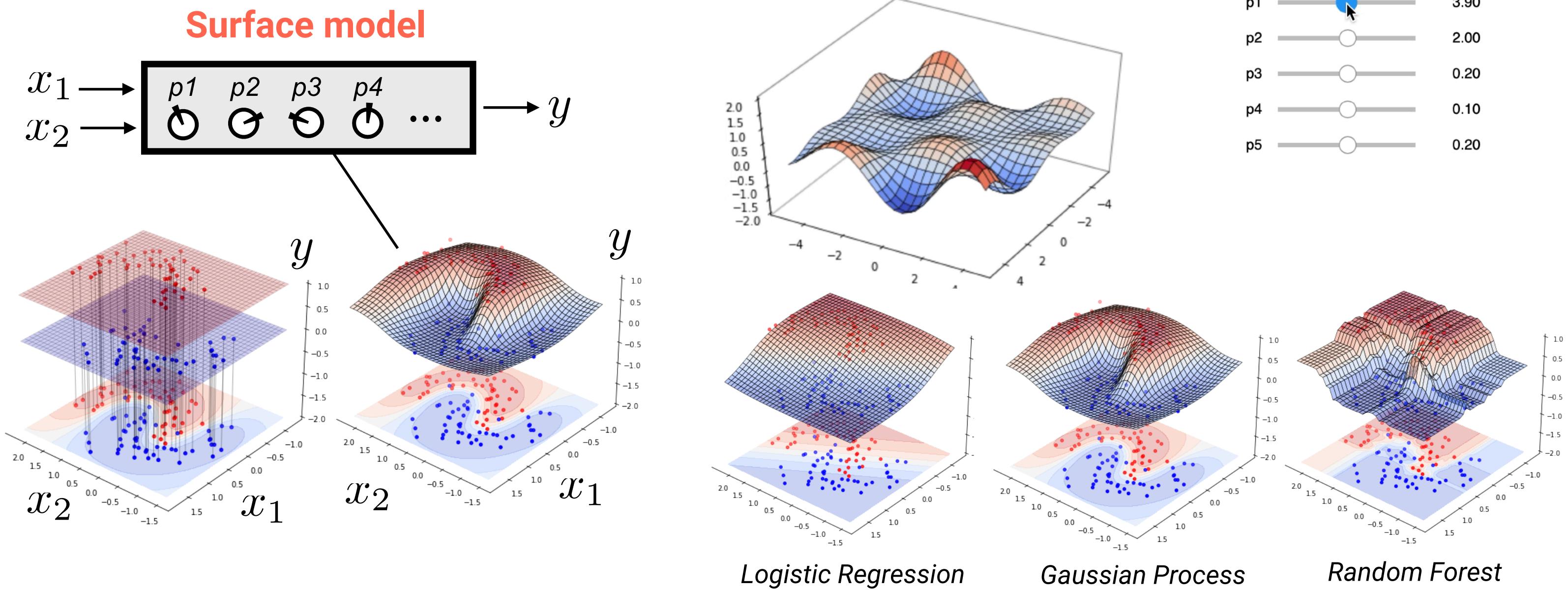
Many ways to mathematically represent the boundary

This is why you see too many algorithms when you start to learn ML...



But anyway, we're just tweaking parameters for a good fit

Internally, we're just **fitting a surface** to given points by adjusting its parameter values.



Pitfall: The lure of wishful wordings

The current ML is stunningly powerful but it's very different from our sci-fi image of "AI".
Be careful about these **"wishful"** wordings that needlessly distract and mislead us!

*"Artificial intelligence" doesn't mean that we have something artificial also intelligent like us.
"Machine learning" doesn't mean that machines actually learn things like us.*



The image shows the October 2021 issue of IEEE Spectrum magazine. The cover features a large, bold title 'IEEE Spectrum' at the top. Below the title, there is a photograph of several white, cylindrical AI hardware components, possibly neural network modules. A red banner across the middle of the cover contains the text 'Why Is AI So Dumb?' in white, followed by 'A SPECIAL REPORT' in smaller white text. At the bottom left, there is a small circular logo with the letters 'S' and 'P' inside. The right side of the cover has a vertical column of text with article titles and their page numbers: 'What's Next for Deep Learning' (p. 26), 'Inside DeepMind's Robot Lab' (p. 34), 'The 7 Biggest Weaknesses of Neural Nets' (p. 42), and 'FOR THE TECHNOLOGY INSIDER OCTOBER 2021'.

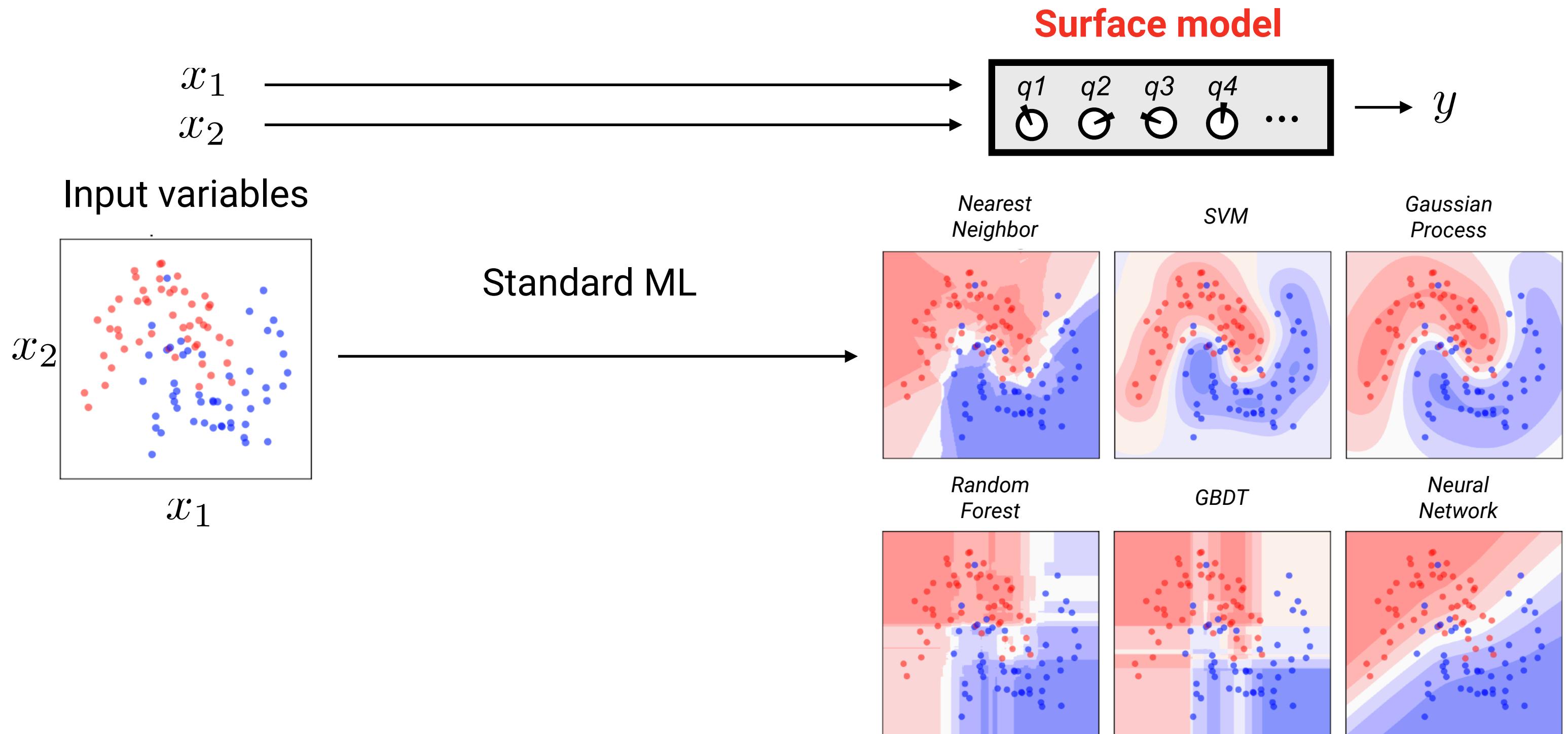
SIGART Newsletter No. 57 April 1976

ARTIFICIAL INTELLIGENCE MEETS NATURAL STUPIDITY

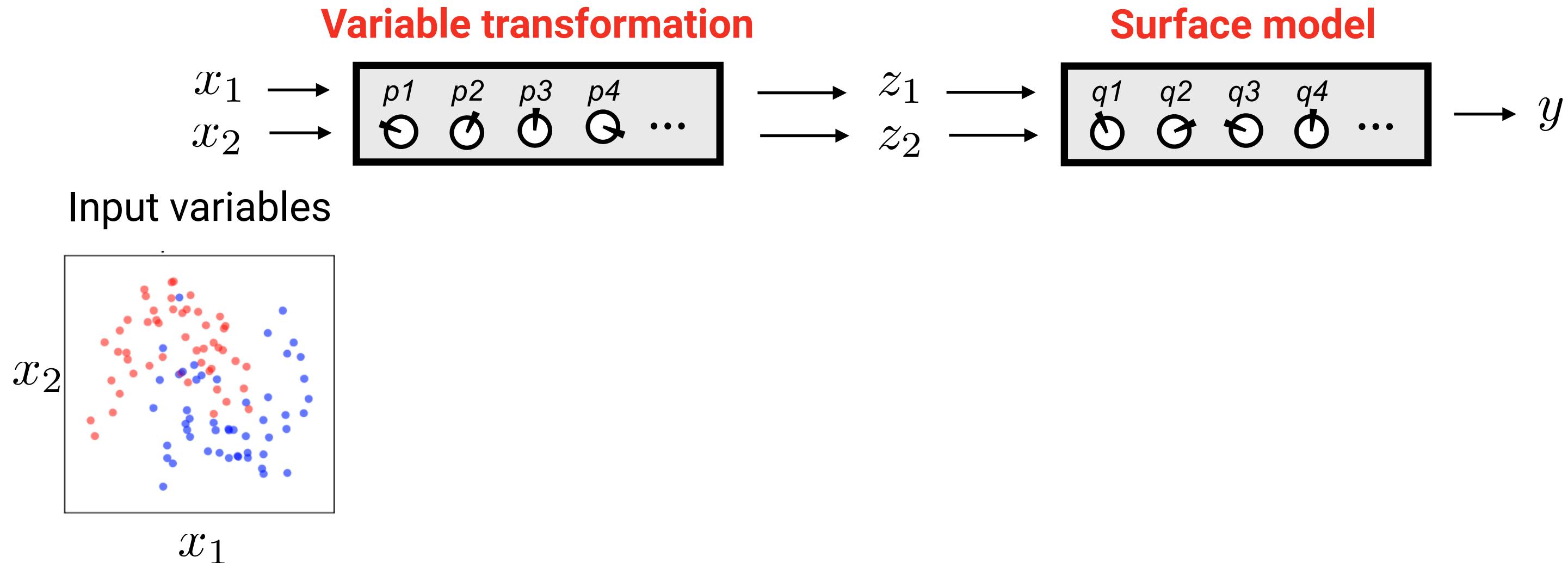
Drew McDermott
MIT AI Lab Cambridge, Mass 02139

As a field, artificial intelligence has always been on the border of respectability, and therefore on the border of crackpottery. Many critics <Dreyfus, 1972>, <Lighthill, 1973> have urged that we are over the border. We have been very defensive toward this charge, drawing ourselves up with dignity when it is made and folding the cloak of Science about us. On the other hand, in private, we have been justifiably proud of our willingness to explore weird ideas, because pursuing them is the only way to make progress.

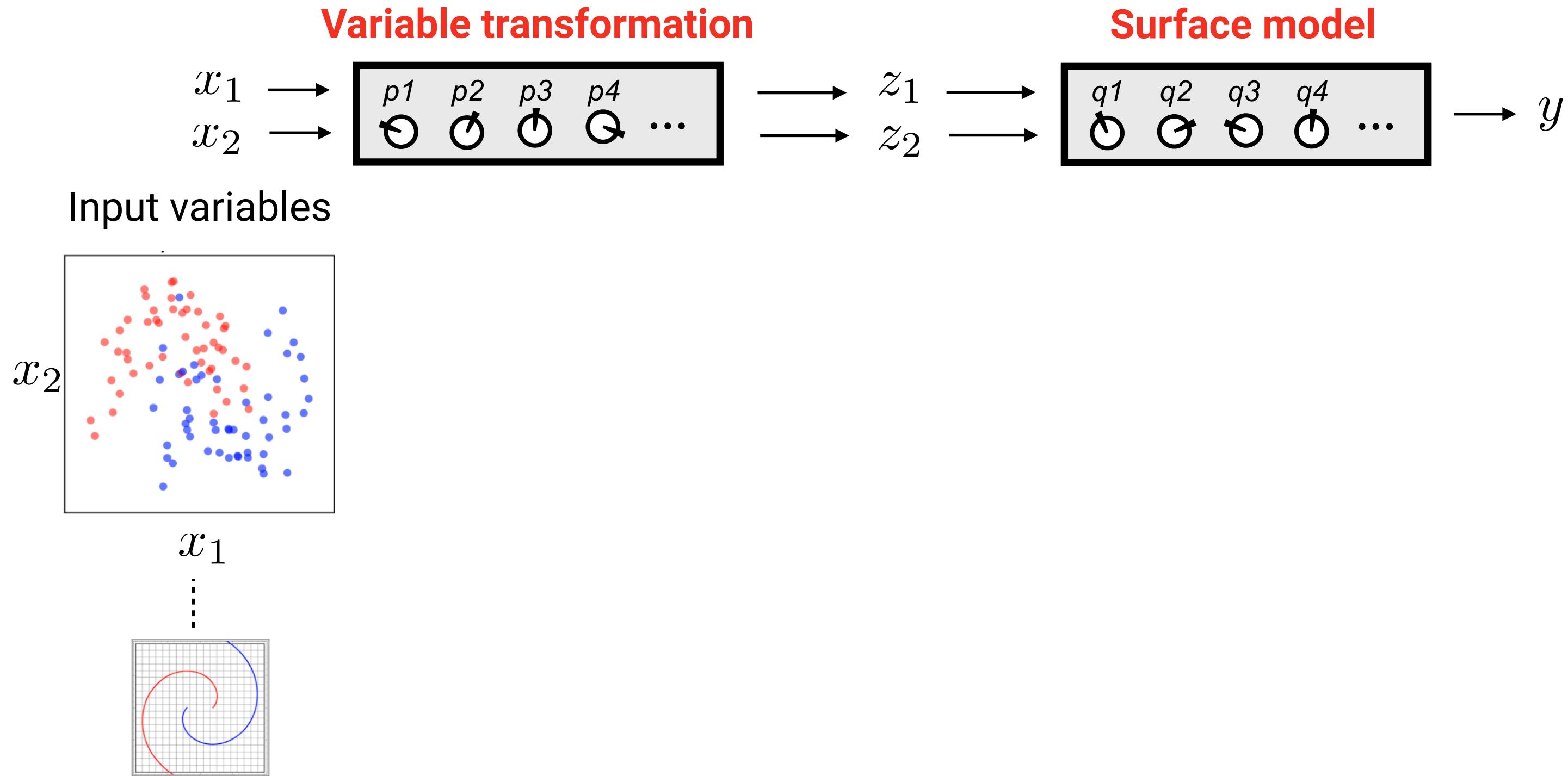
Deep Learning (Representation Learning)



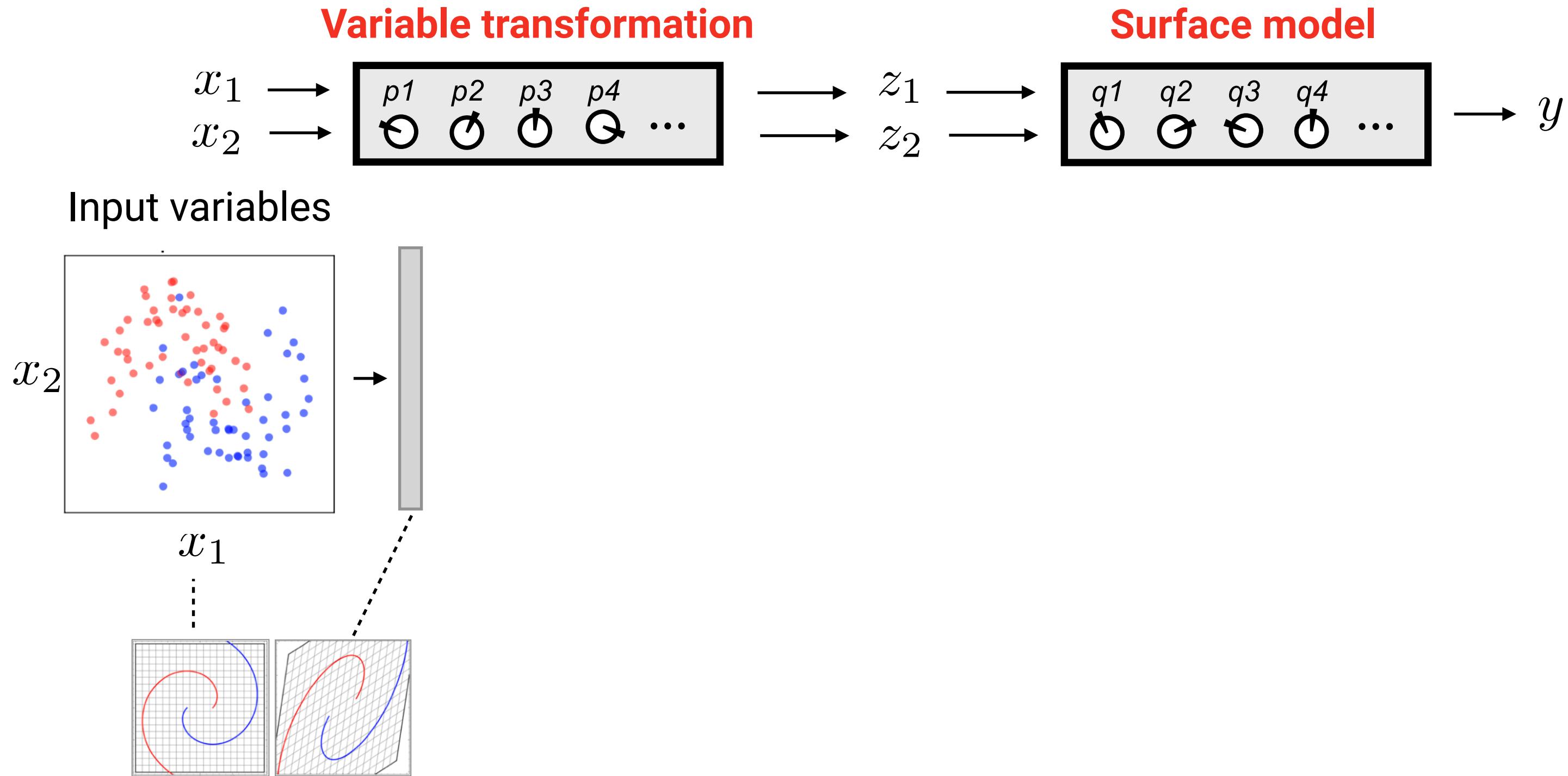
Deep Learning (Representation Learning)



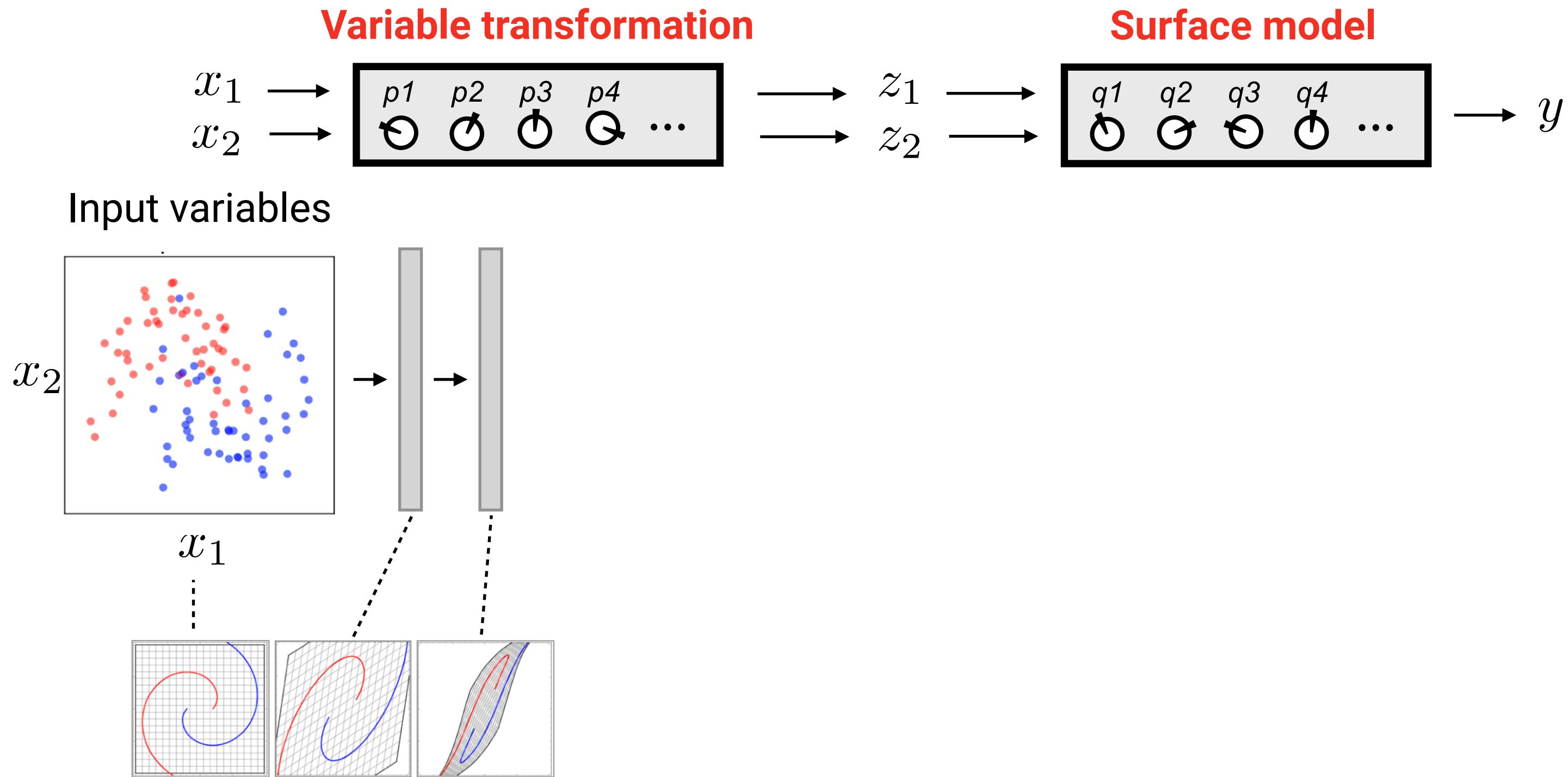
Deep Learning (Representation Learning)



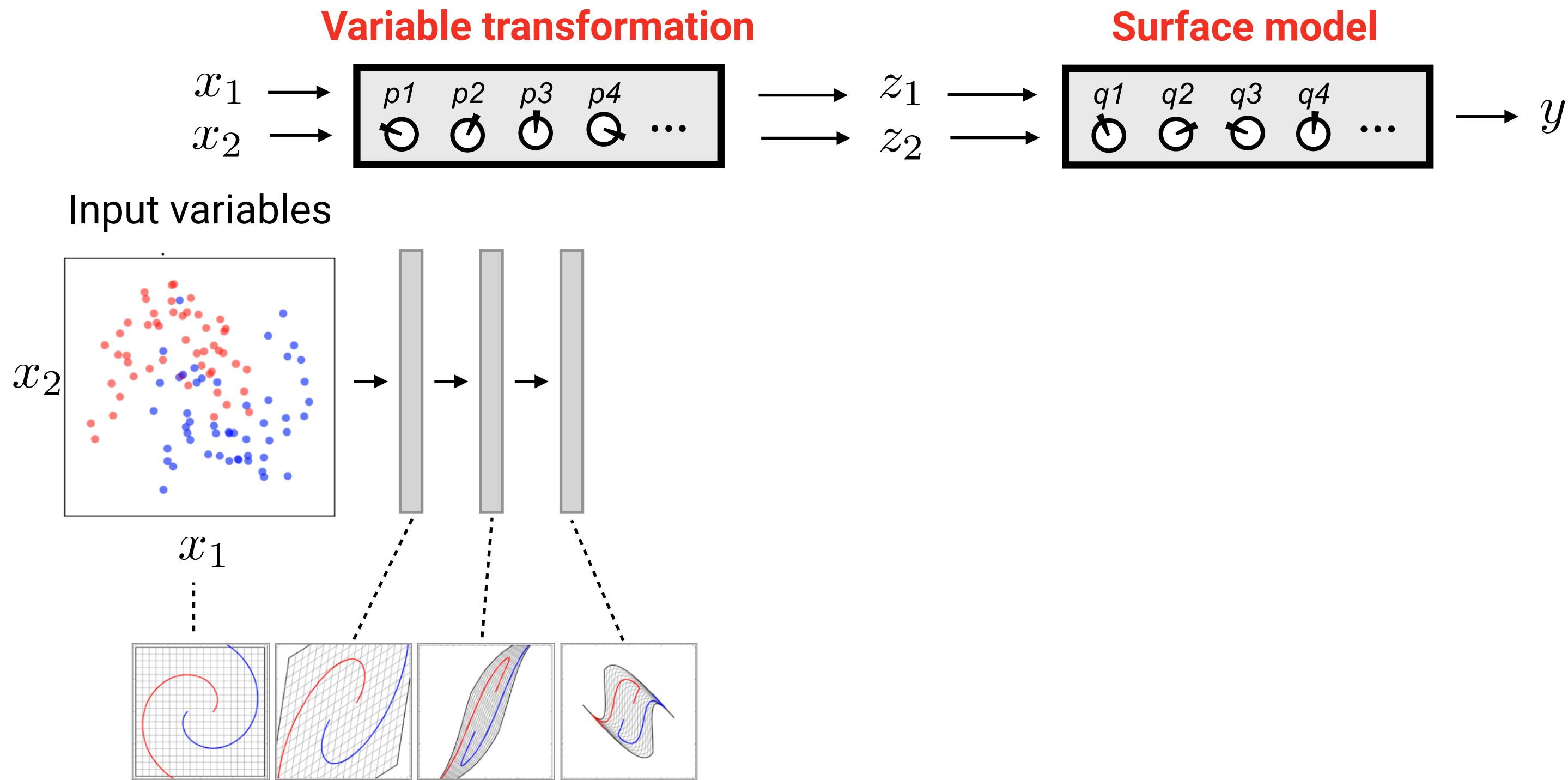
Deep Learning (Representation Learning)



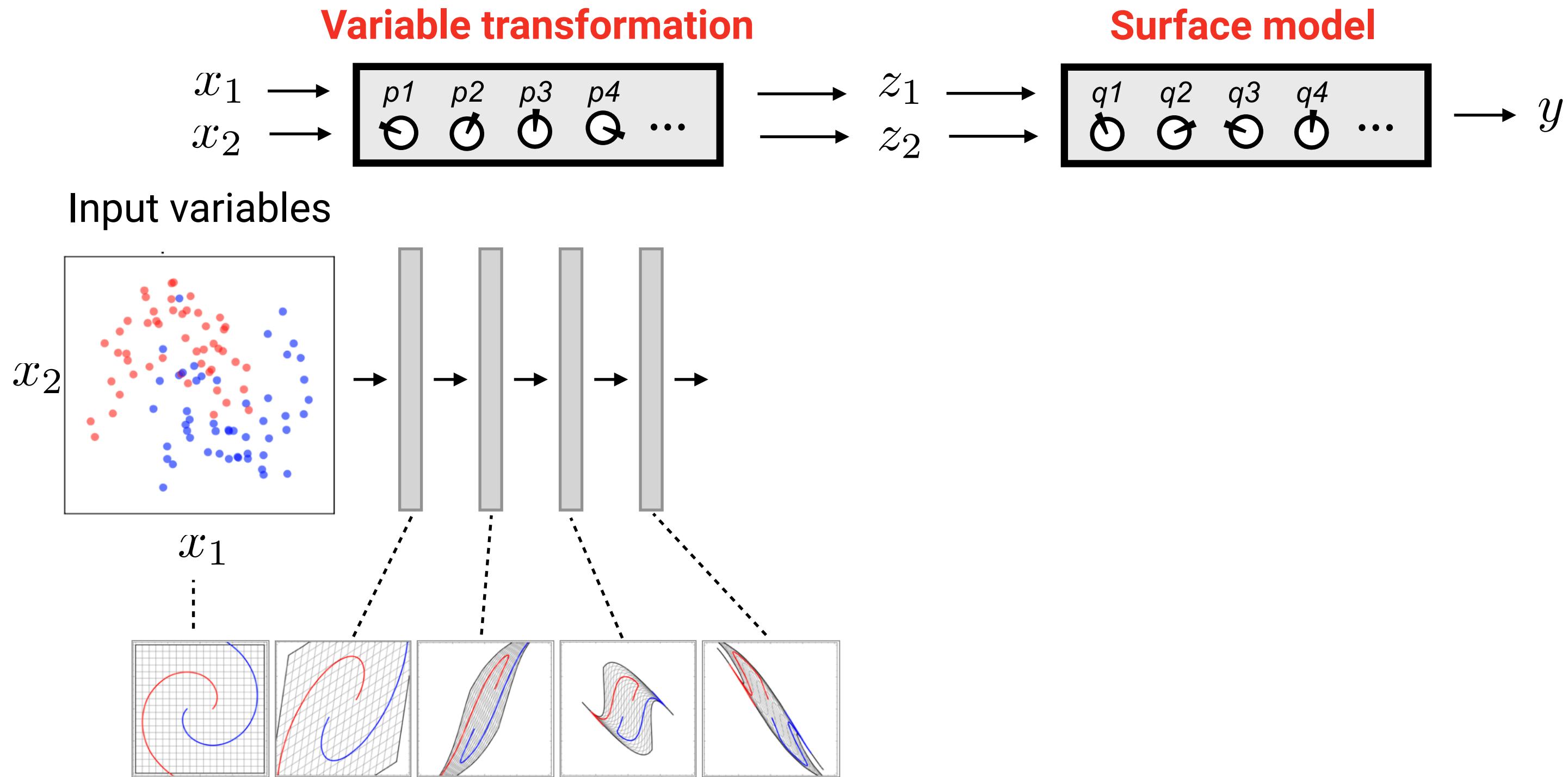
Deep Learning (Representation Learning)



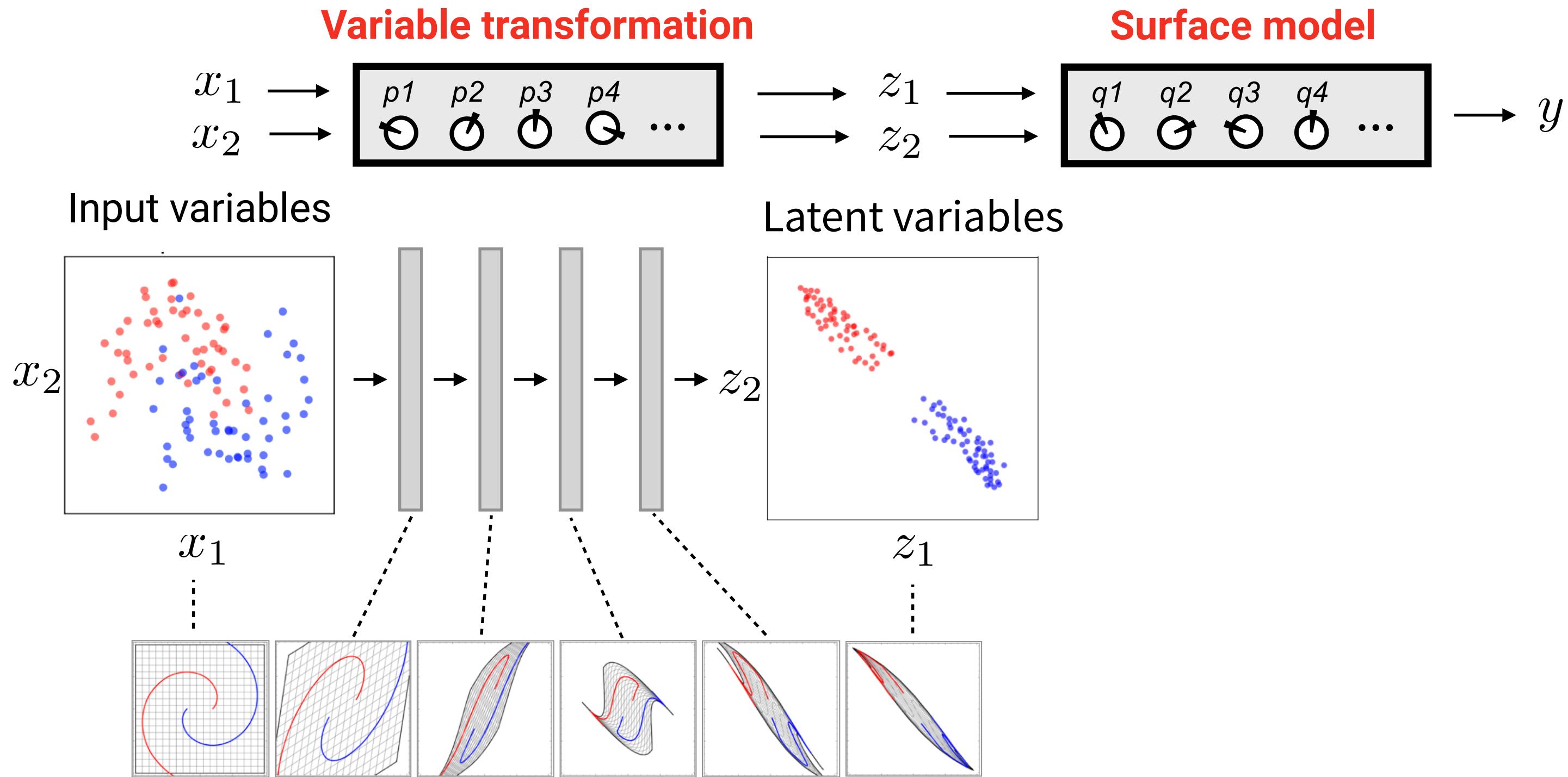
Deep Learning (Representation Learning)



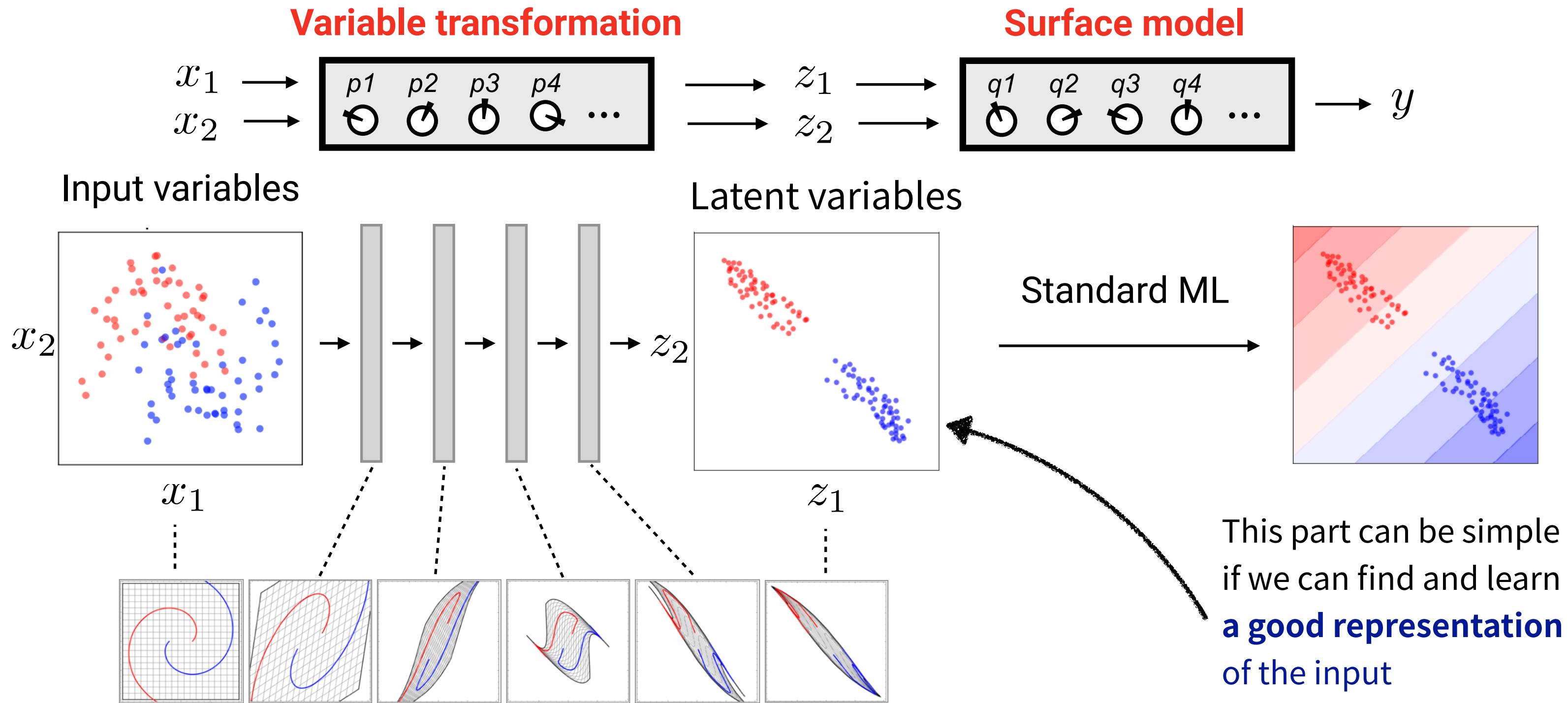
Deep Learning (Representation Learning)



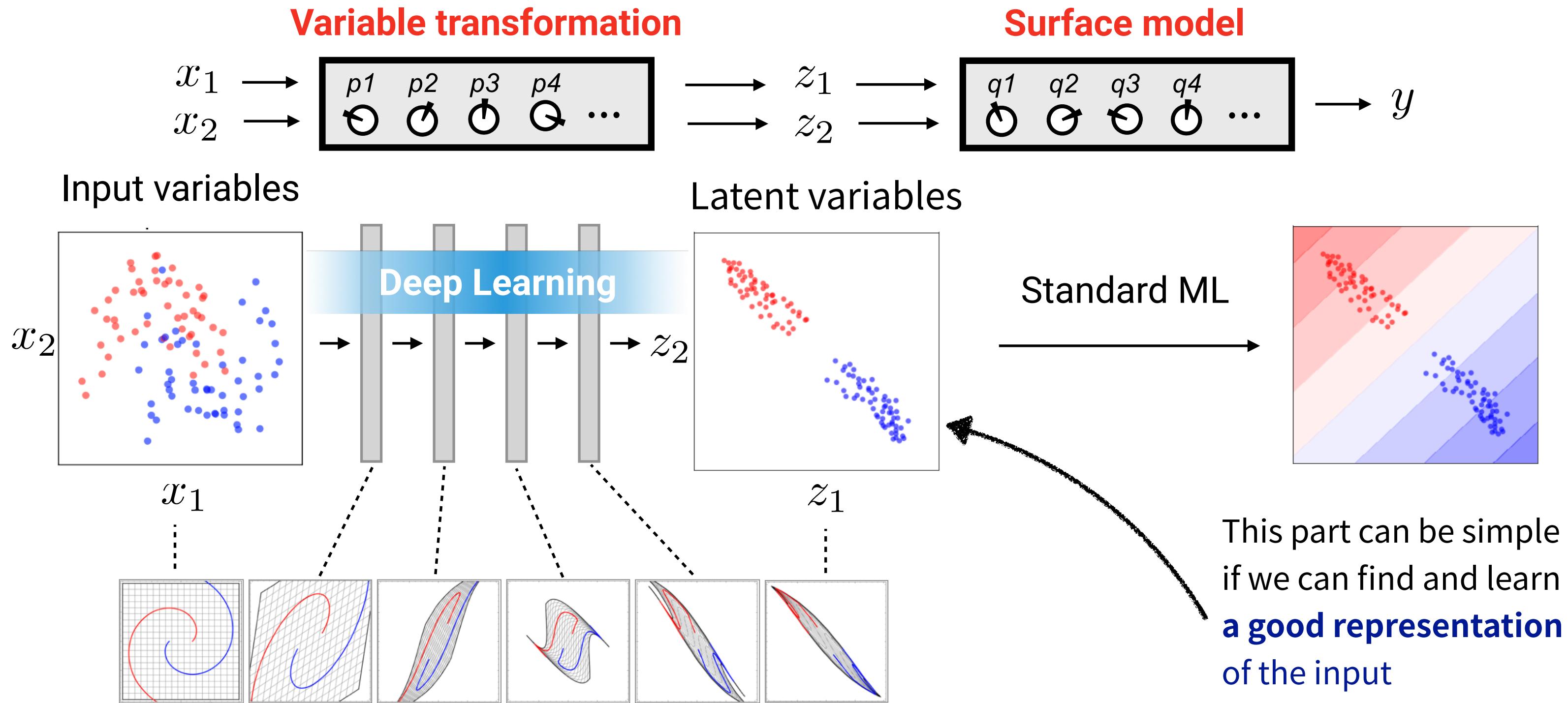
Deep Learning (Representation Learning)



Deep Learning (Representation Learning)



Deep Learning (Representation Learning)



The Dark Side: Modern Aspects of ML

- **High dimensionality:** Too many input variables

We tend to use **many input variables** because ML is completely unaware of any information **not** in the input variables. Missing relevant factors results in spurious correlation.

e.g.) 100 x 100 RGB image = **30 thousand** variables

1000 x 1000 RGB image = **3 million** variables

The Dark Side: Modern Aspects of ML

- **High dimensionality:** Too many input variables

We tend to use **many input variables** because ML is completely unaware of any information **not** in the input variables. Missing relevant factors results in spurious correlation.

e.g.) 100 x 100 RGB image = **30 thousand** variables

1000 x 1000 RGB image = **3 million** variables

- **Overrepresentation:** Too many parameters

Remember that we're fitting a surface with *hundreds million* parameters in a *several million* dimensional space!

e.g.) ResNet50: **26 million** params

ResNet101: **45 million** params

EfficientNet-B7: **66 million** params

VGG19: **144 million** params

12-layer, 12-heads BERT: **110 million** params

24-layer, 16-heads BERT: **336 million** params

GPT-2 XL: **1558 million** params

GPT-3: **175 billion** params

The Dark Side: Modern Aspects of ML

- **Data hungeriness:** Big data is big for human, but can be too small for ML models...
As a result, it **requires huge data** to make current ML models work.

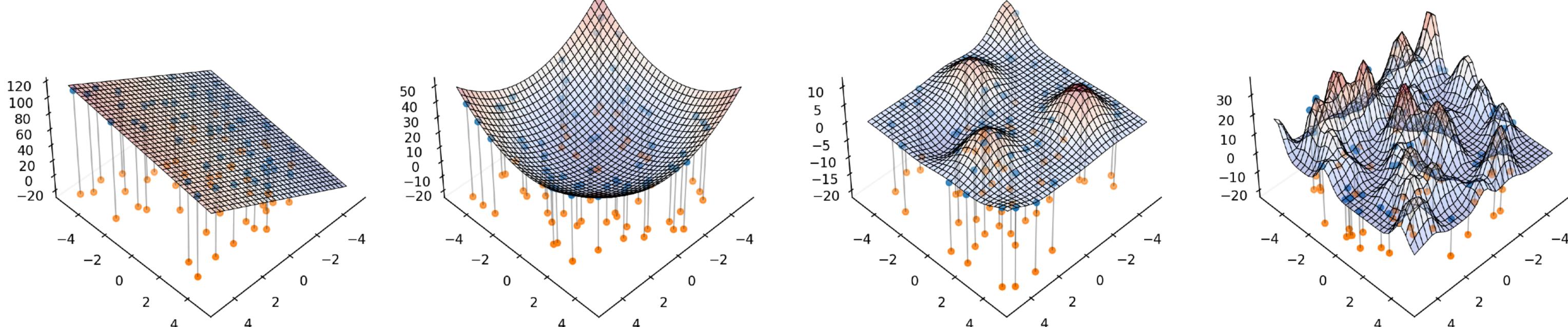
The Dark Side: Modern Aspects of ML

- **Data hungeriness:** Big data is big for human, but can be too small for ML models...

As a result, it **requires huge data** to make current ML models work.

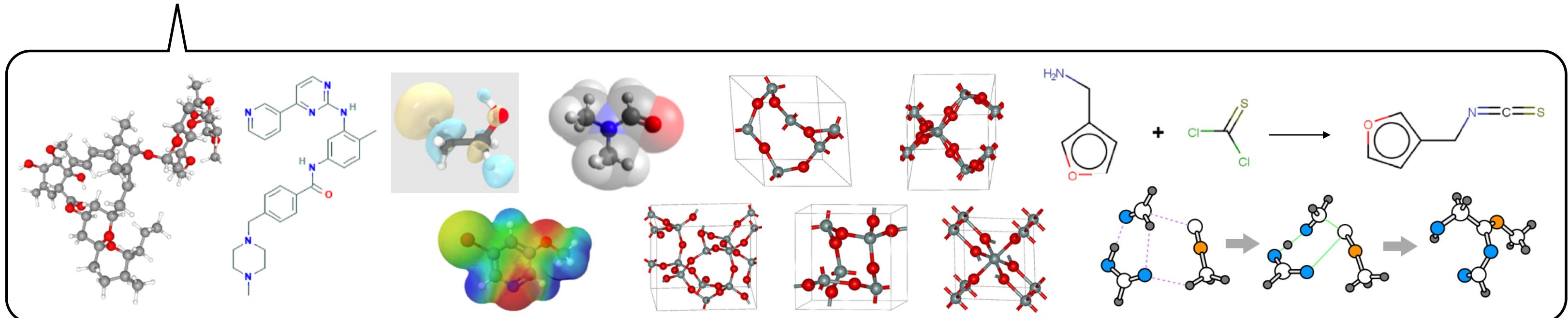
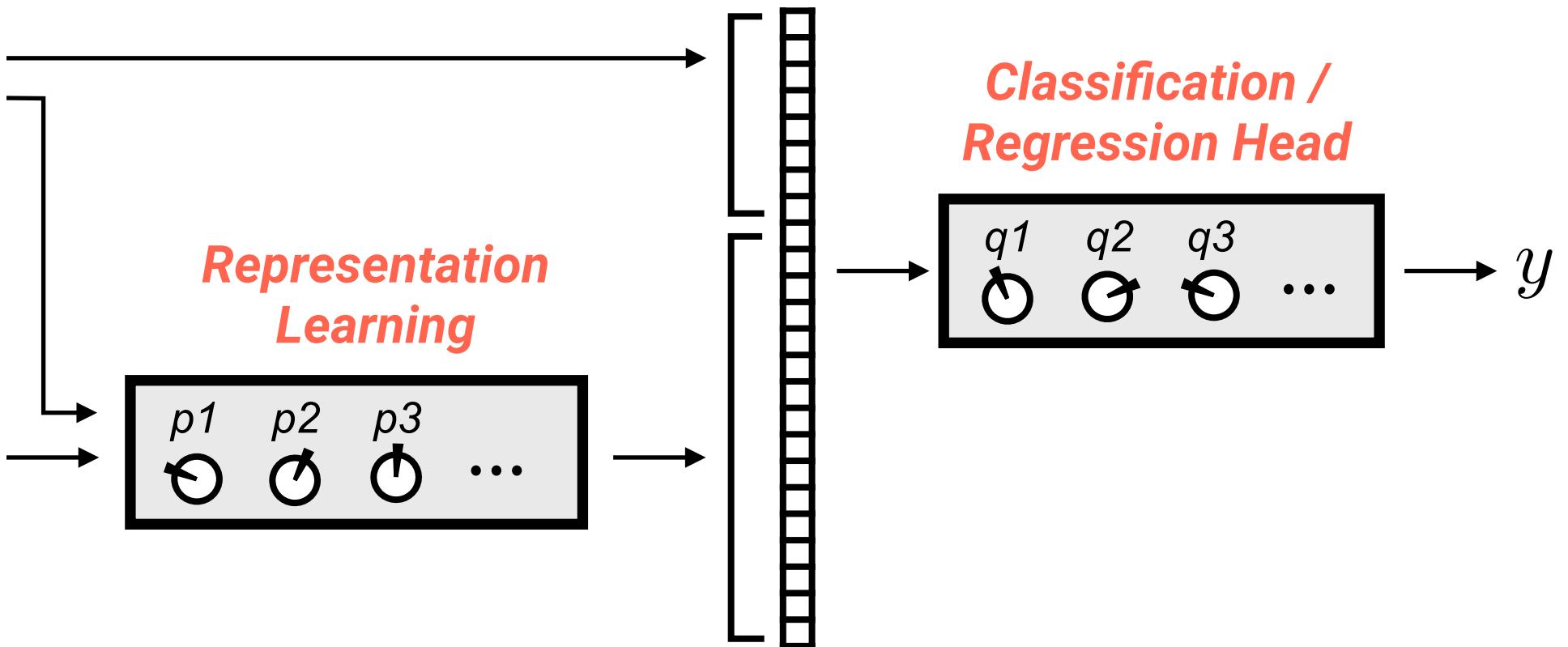
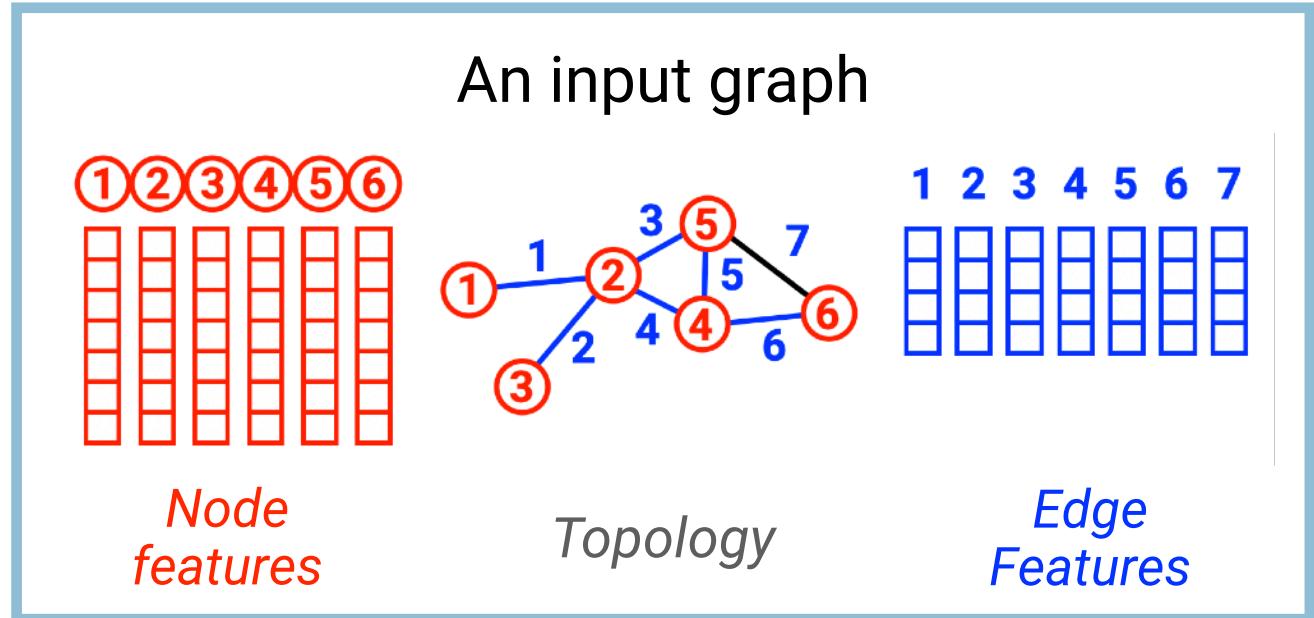
Think twice about how complex the input-output relationship you are trying to find will be.

How many samples will be ***statistically sufficient*** to estimate **2-variable** functions like these?
What if you're fitting a **100-variable** function?



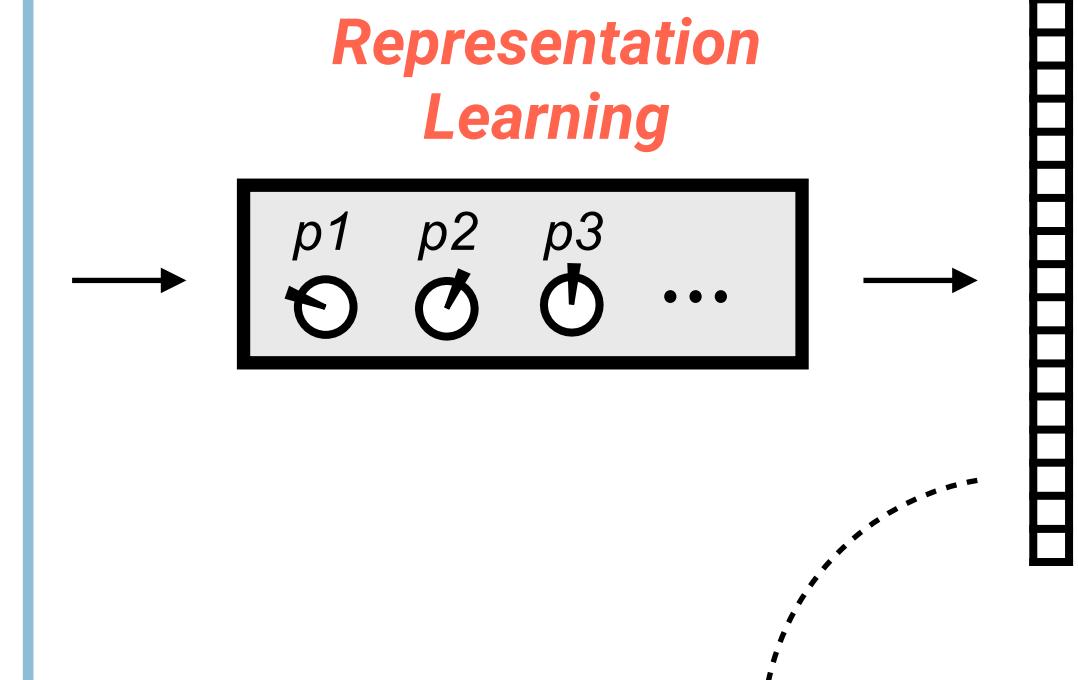
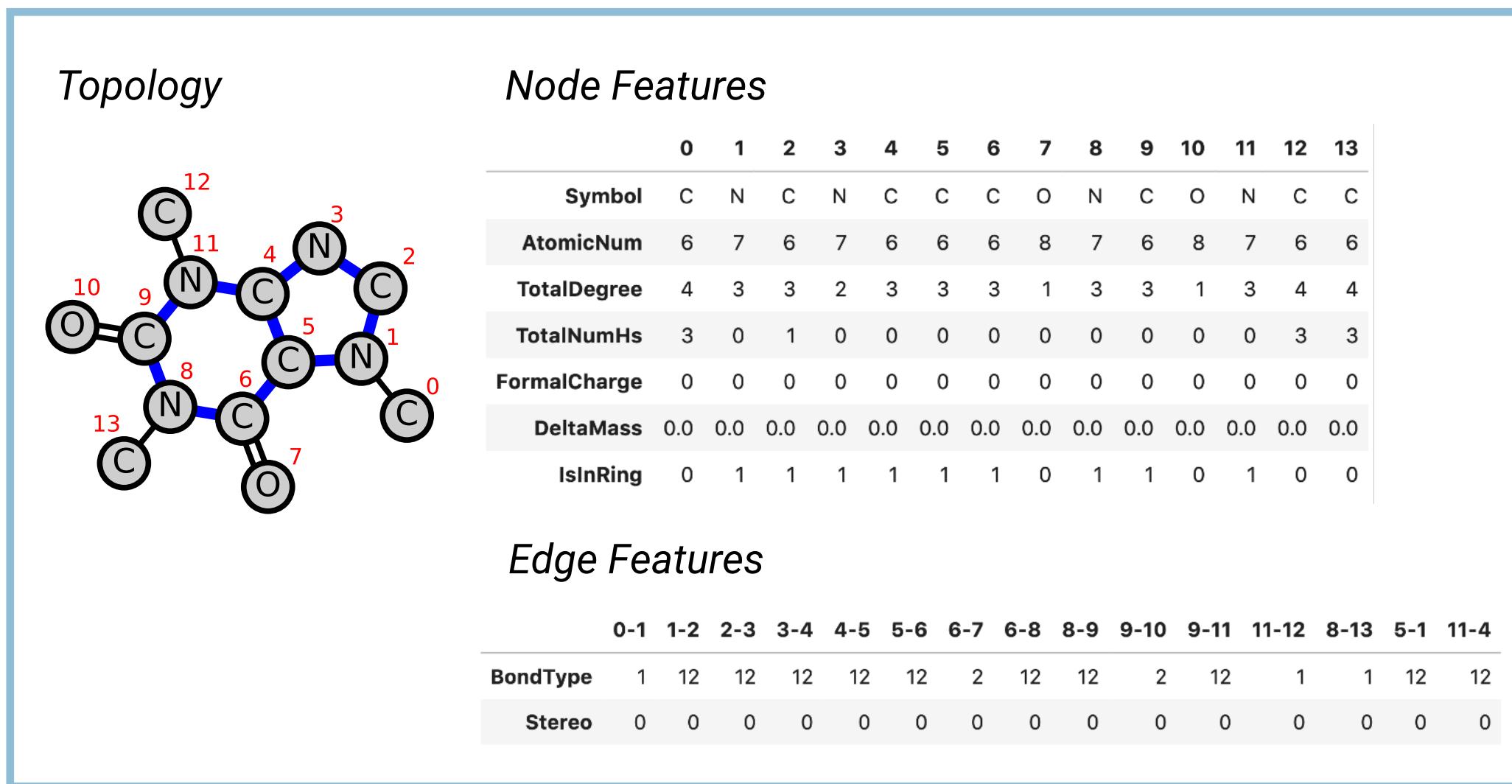
Graph Neural Networks (aka Geometric Deep Learning)

Other Info (Conditions, Environment, ...)



Graph Representation Learning

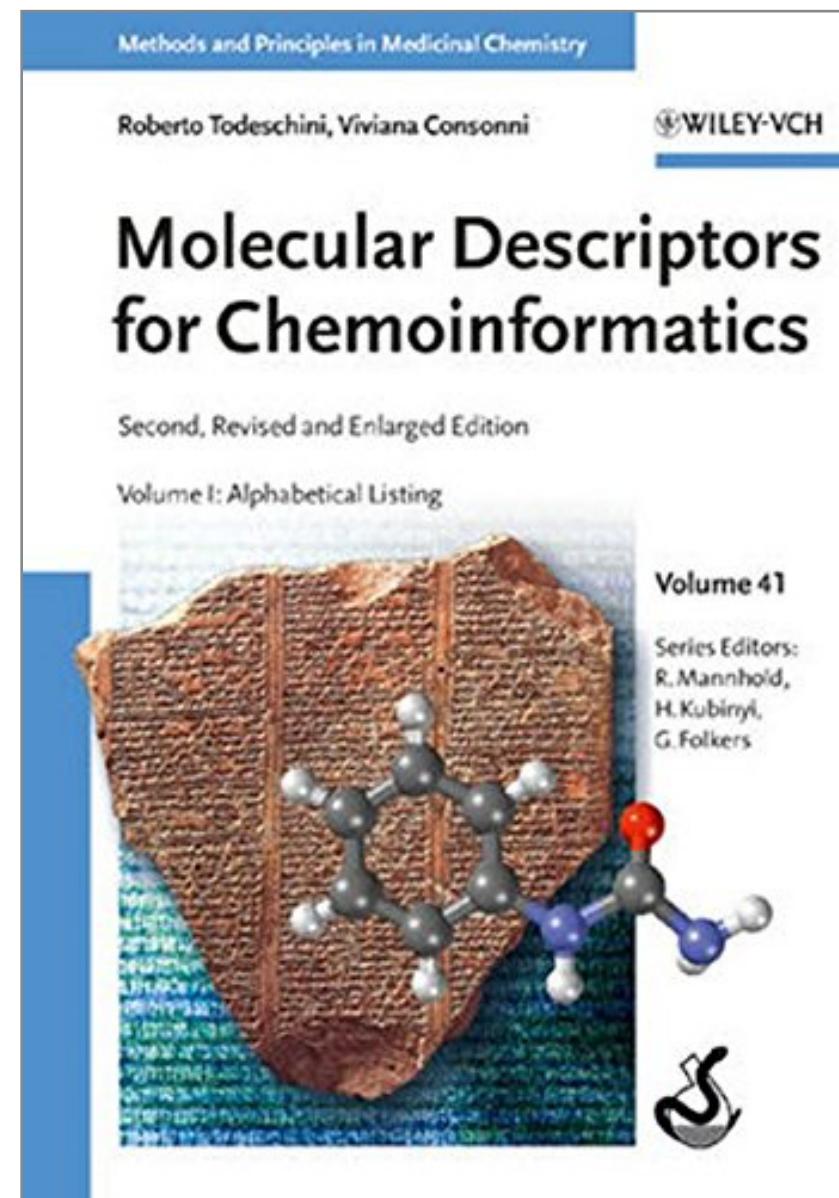
We can seek for **a good representation vector** that can be computed from a molecular graph, which is expected to be superior to any man-made descriptors!



Beyond Man-Made Descriptors

- **0-Dimensional Descriptors**
 - Constitutional descriptors
 - Count descriptors
- **1-Dimensional Descriptors**
 - List of structural fragments
 - Fingerprints
- **2-Dimensional Descriptors**
 - Graph invariants
- **3-Dimensional Descriptors**
 - 3D MoRSE, WHIM, GETAWAY, ...
 - Quantum-chemical descriptors
 - Size, steric, surface, volume, ...
- **4-Dimensional Descriptors**
 - GRID, CoMFA, Volsurf, ...

"Vol 1 contains an alphabetical listing of **more than 3,300 descriptors**"

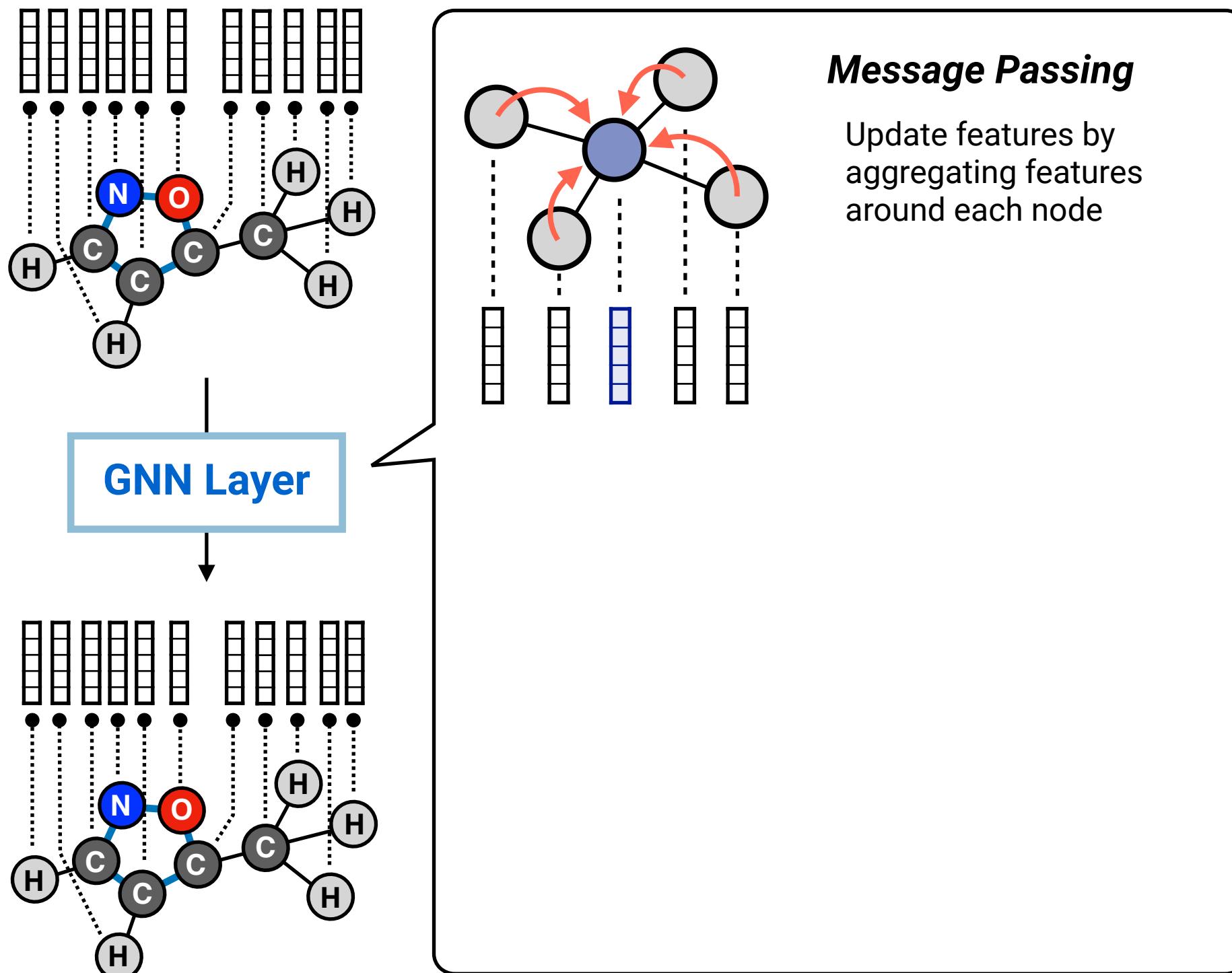


"Dragon calculates **5,270** molecular descriptors"

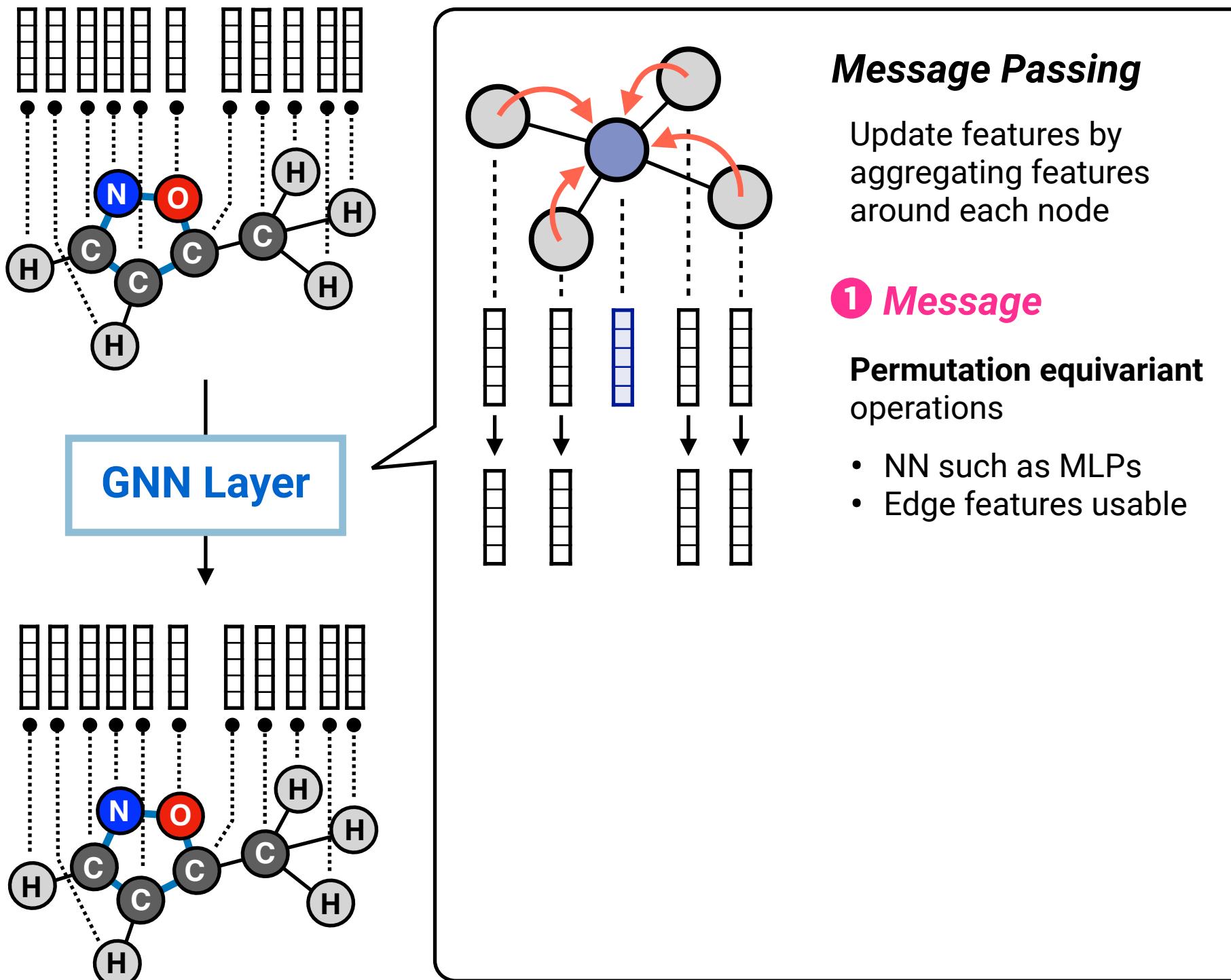
The screenshot displays the DRAGON 7.0 software interface. At the top, the Kode chemoinformatics logo and the DRAGON 7.0 version are shown. Below is a navigation bar with tabs for DESCRIPTORS, FINGERPRINTS, and PROJECT AND SCIENTIFIC WORKS. The DESCRIPTORS tab is selected. A text block states: "Dragon 7.0 calculates 5,270 molecular descriptors, organized in different logical blocks as in the previous versions. Blocks are further divided into sub-blocks to make management, selection, and analysis of descriptors easier. Following, the summary of molecular descriptors blocks calculated by Dragon 7.0 is reported." A table below lists the blocks, their names, and the number of descriptors they contain.

BLOCK NO	BLOCK NAME	DESCRIPTORS
1	Constitutional	47
2	Ring descriptors	32
3	Topological indices	75
4	Walk and path counts	46
5	Connectivity indices	37
6	Information indices	50
7	2D matrix-based descriptors	607
8	2D autocorrelations	213
9	Burden eigenvalues	96
10	P-VSA-like descriptors	55

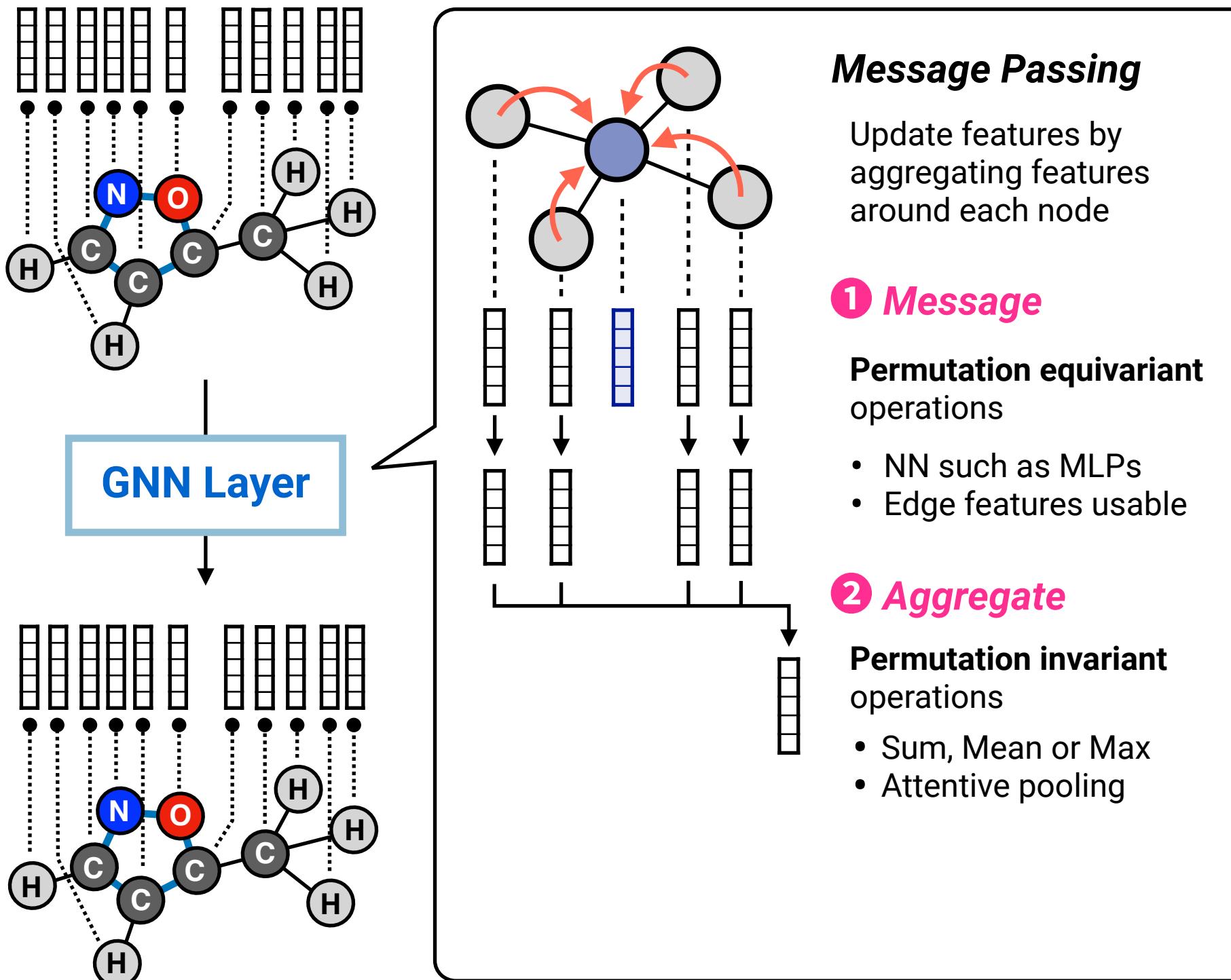
Message Passing: The Inner Workings of GNNs



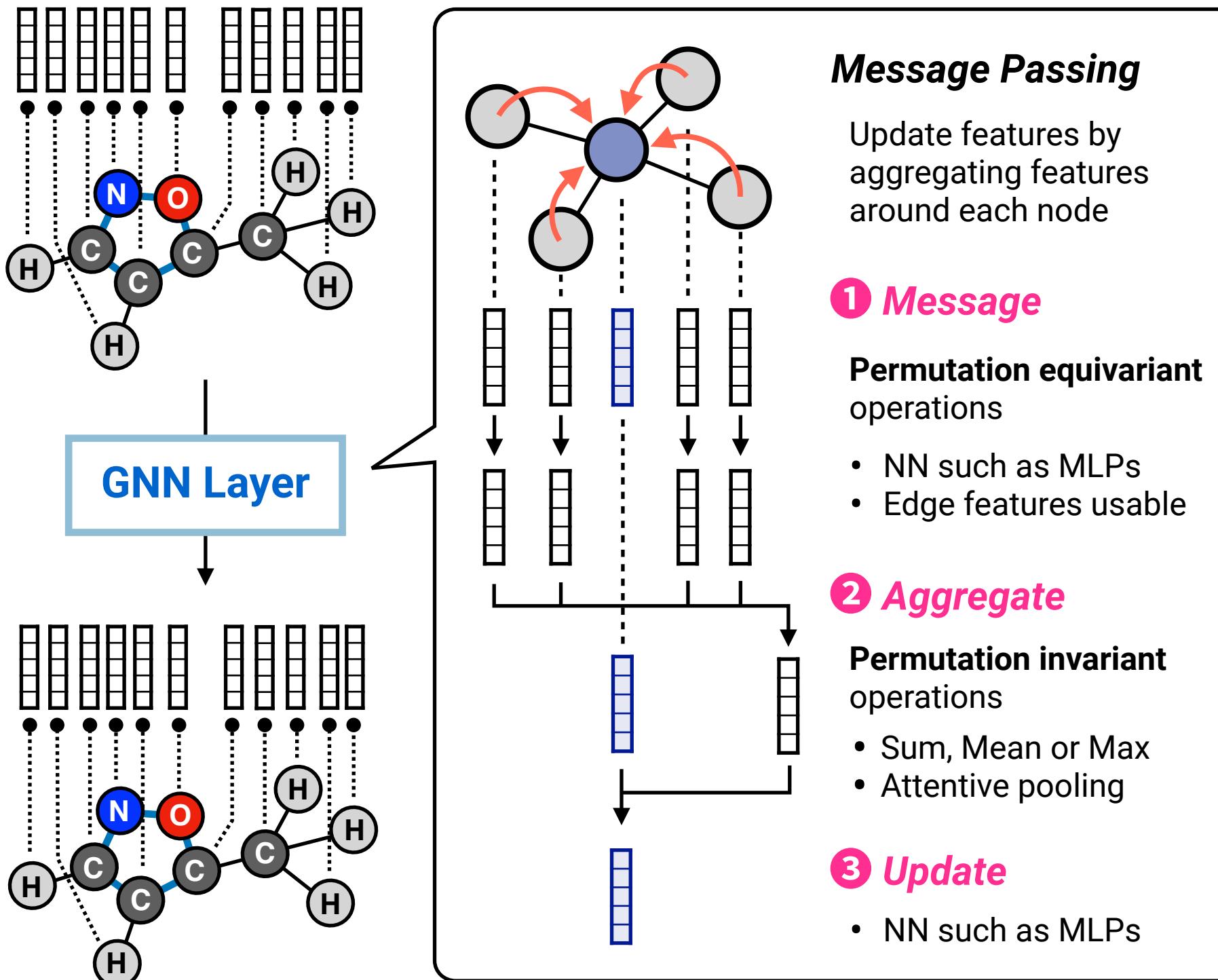
Message Passing: The Inner Workings of GNNs



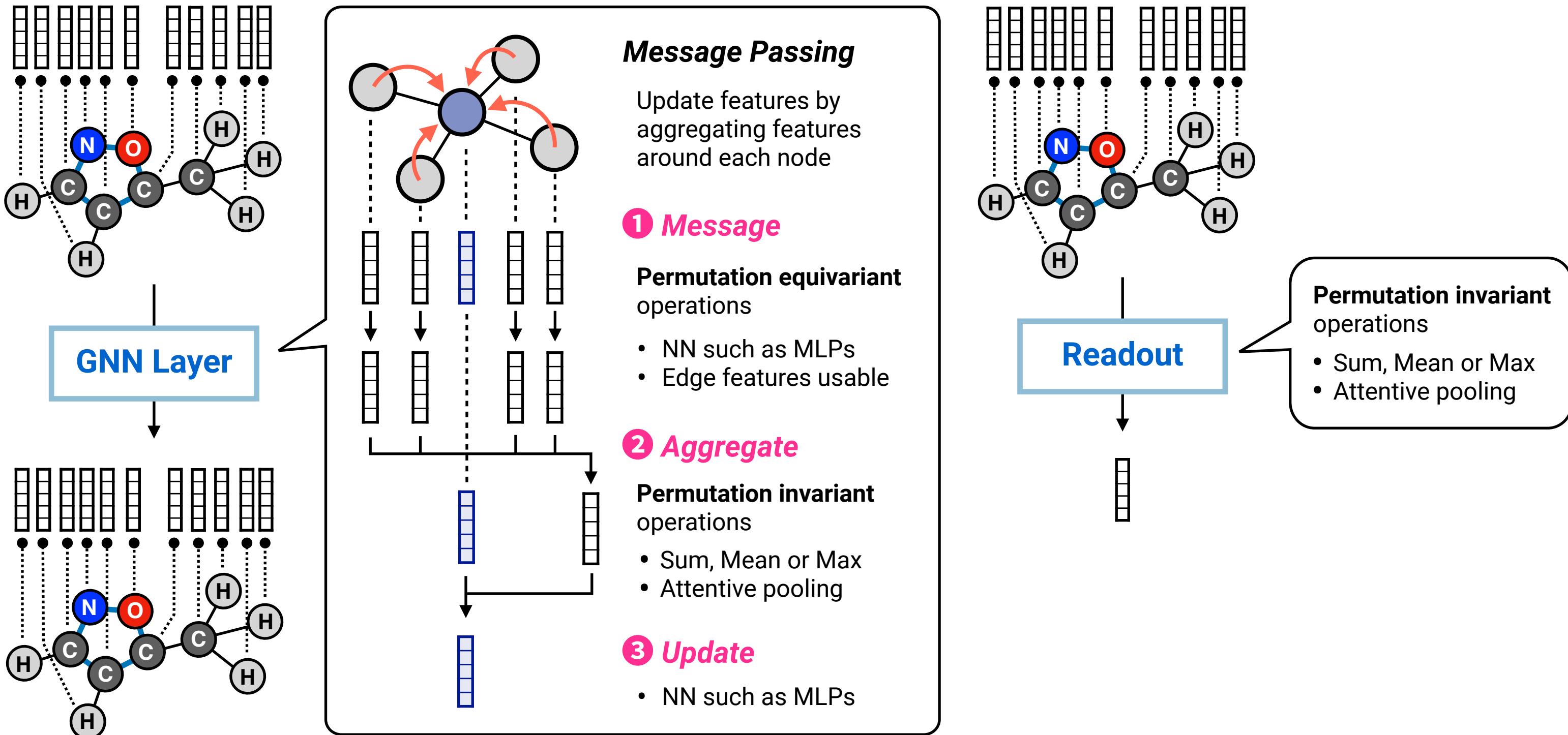
Message Passing: The Inner Workings of GNNs



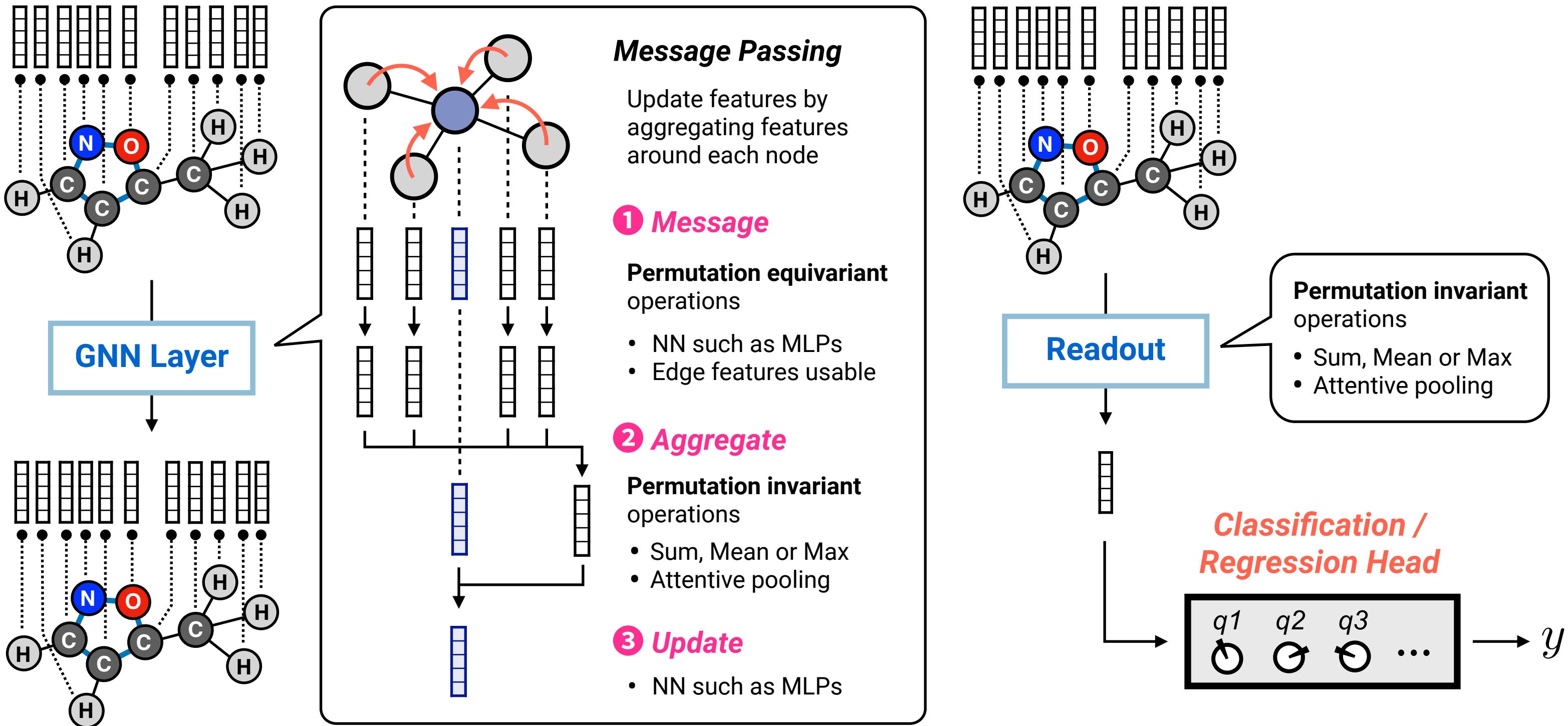
Message Passing: The Inner Workings of GNNs



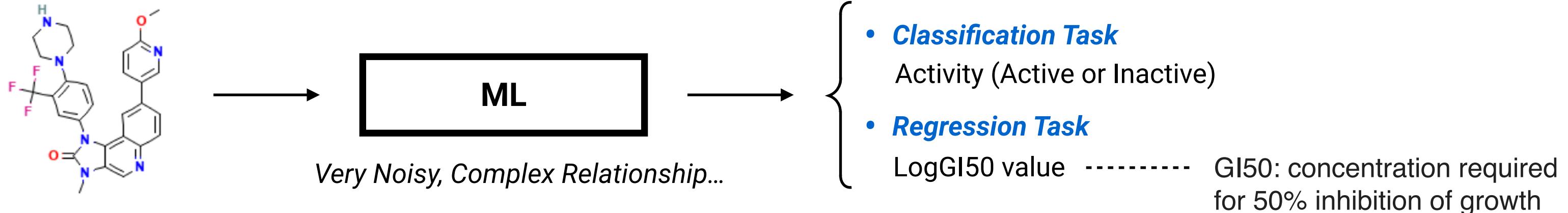
Message Passing: The Inner Workings of GNNs



Message Passing: The Inner Workings of GNNs



Use Case 1: Virtual Screening (QSAR/QSPR)



NCI Human Tumor Cell Line Growth Inhibition Assay (PubChem AID 1)

Active (2,814)			Activity	Score	LogGI50_M ⓘ
Structure	CID	SID			
	5298	121832	Active	67	-8
	363173	493713	Active	43	-6.5871
	399631	530868	Active	51	-7.0678
Inactive (48,922)			Activity	Score	LogGI50_M ⓘ
Structure	CID	SID			
	390324	521601	Inactive	0	-4
	390311	521588	Inactive	0	-4
	390312	521589	Inactive	4	-4.214

Use Case 1: Virtual Screening (QSAR/QSPR)

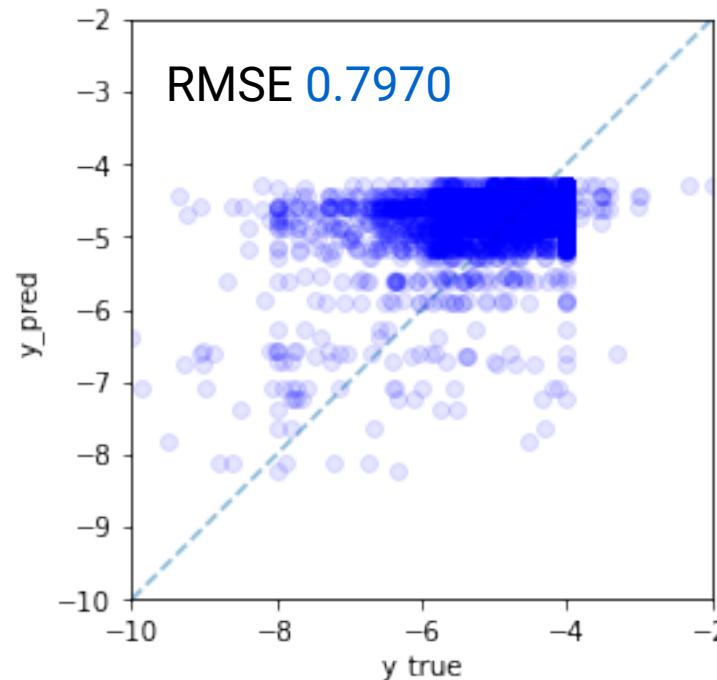
Standard ML

ExtraTrees
w/ ECFP6(1024)

- **Classification Task** Activity (Active or Inactive)

95.079%

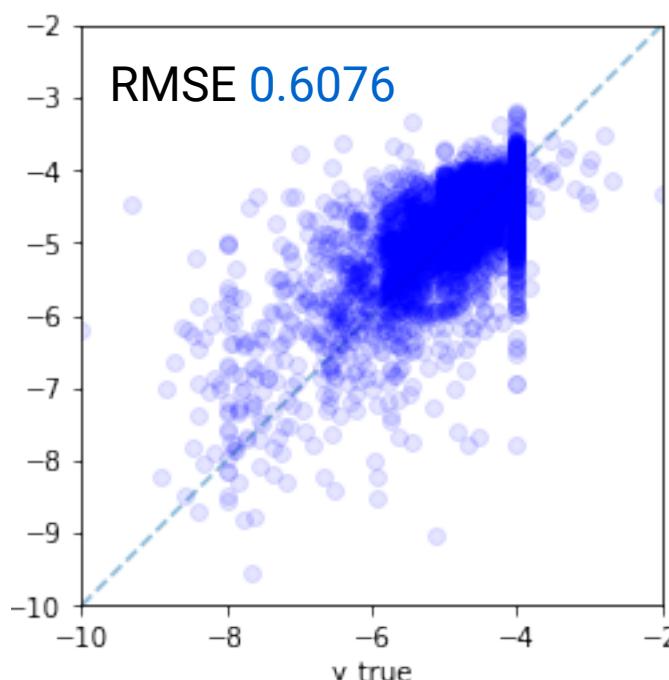
- **Regression Task** LogGI50 value



GNN

ChemProp
(Directed MPNN)

95.604%



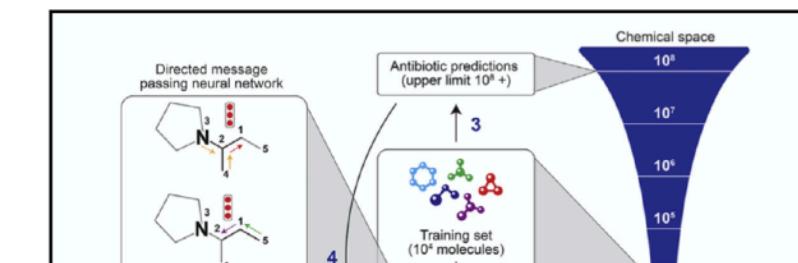
ChemProp (Yang et al, 2019)

from MIT MLPDS (Machine Learning for Pharmaceutical Discovery and Synthesis) Consortium

Cell

A Deep Learning Approach to Antibiotic Discovery

Graphical Abstract



Authors

Jonathan M. Stokes, Kevin Yang,
Kyle Swanson, ..., Tommi S. Jaakkola,
Regina Barzilay, James J. Collins

Correspondence

regina@csail.mit.edu (R.B.),
jimjc@mit.edu (J.J.C.)

nature

NEWS | 20 February 2020

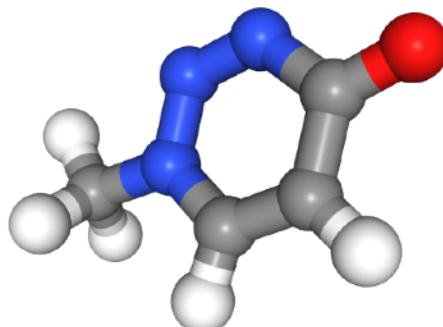
Powerful antibiotics discovered using AI

Machine learning spots molecules that work even against 'untreatable' strains of bacteria.

Jo Marchant

Use Case 2: Fast Approximation for QM Calculations

input



gdb_21014

	x	y	z
O	0.314096	-0.129589	-0.389150
C	0.111219	2.102676	-0.051749
C	2.331344	3.941075	0.212303
O	4.667017	2.677399	0.437948
C	6.152491	3.062553	-1.780599
C	4.732264	5.009654	-3.282819
C	2.562527	5.549427	-2.143825
H	-1.771427	3.048695	0.071772
H	1.977918	5.086871	1.919865
H	8.050245	3.696867	-1.222422
H	6.372399	1.276980	-2.825015
H	5.428656	5.805758	-5.033531
H	1.118529	6.857080	-2.763050

~ 1000 sec

Quantum mechanical calculations

by solving a one-electron Schrödinger equation (Kohn–Sham equation)

$$\hat{H}\Psi = E\Psi$$

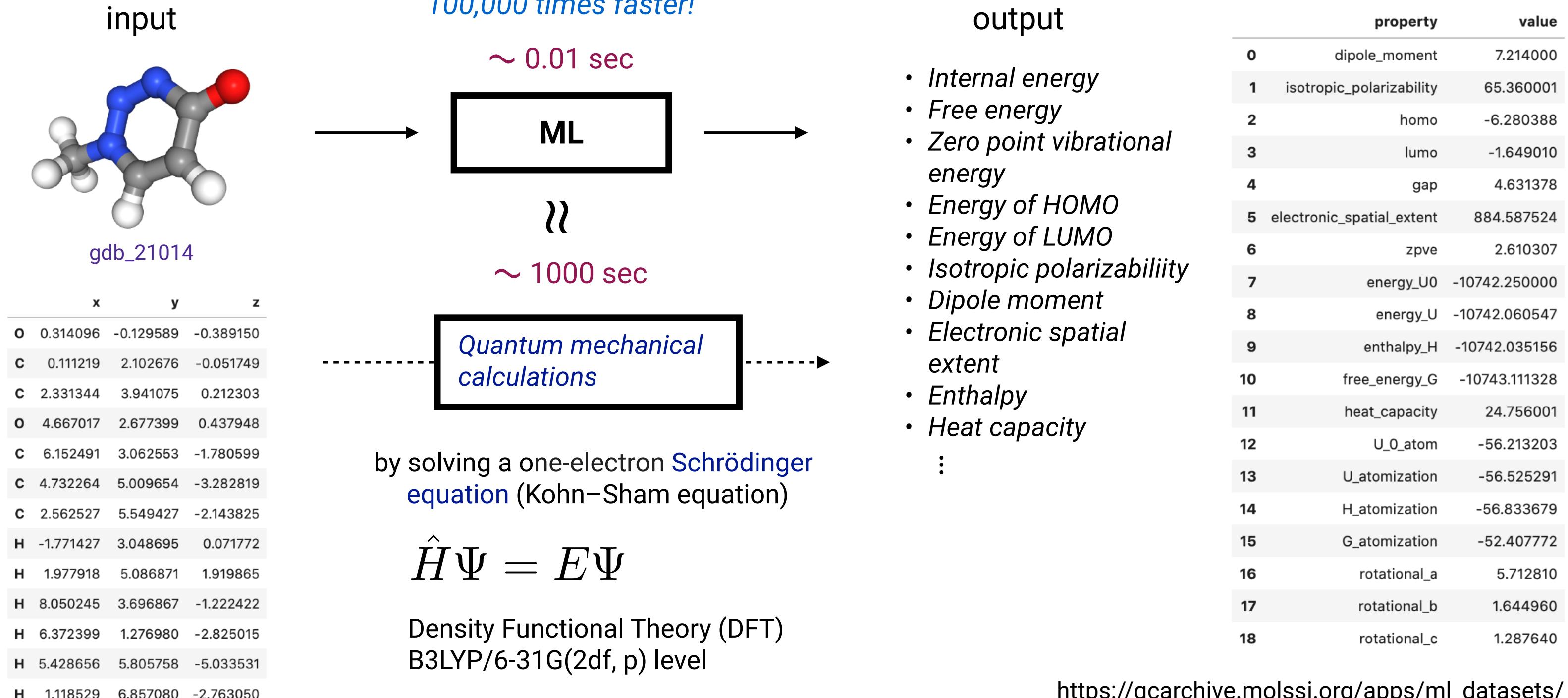
Density Functional Theory (DFT)
B3LYP/6-31G(2df, p) level

output

- Internal energy
- Free energy
- Zero point vibrational energy
- Energy of HOMO
- Energy of LUMO
- Isotropic polarizability
- Dipole moment
- Electronic spatial extent
- Enthalpy
- Heat capacity
- ⋮

	property	value
0	dipole_moment	7.214000
1	isotropic_polarizability	65.360001
2	homo	-6.280388
3	lumo	-1.649010
4	gap	4.631378
5	electronic_spatial_extent	884.587524
6	zpve	2.610307
7	energy_U0	-10742.250000
8	energy_U	-10742.060547
9	enthalpy_H	-10742.035156
10	free_energy_G	-10743.111328
11	heat_capacity	24.756001
12	U_0_atom	-56.213203
13	U_atomization	-56.525291
14	H_atomization	-56.833679
15	G_atomization	-52.407772
16	rotational_a	5.712810
17	rotational_b	1.644960
18	rotational_c	1.287640

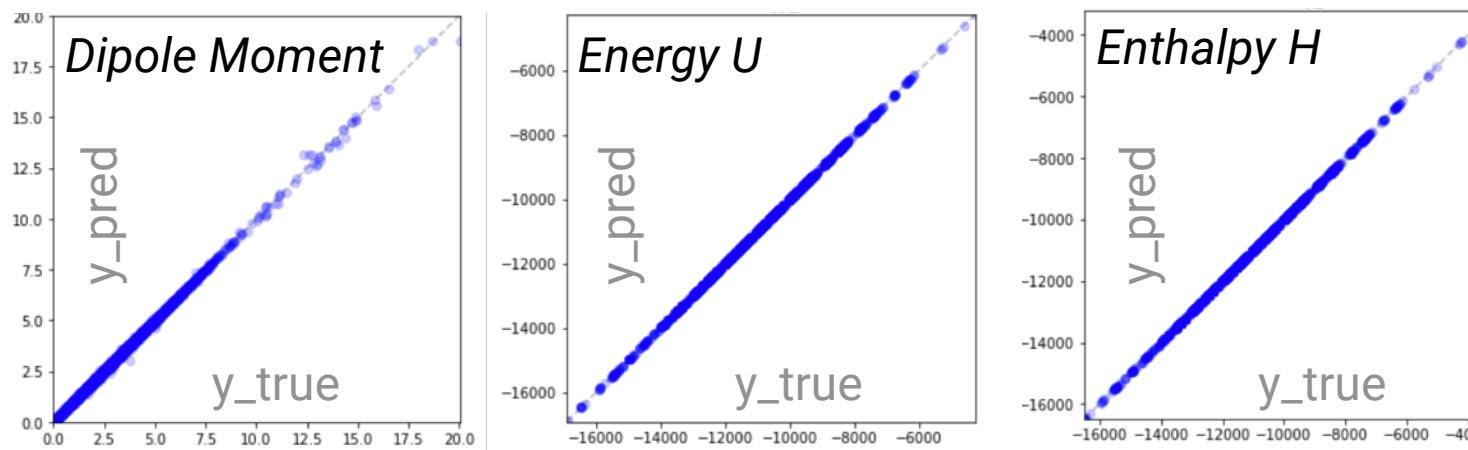
Use Case 2: Fast Approximation for QM Calculations



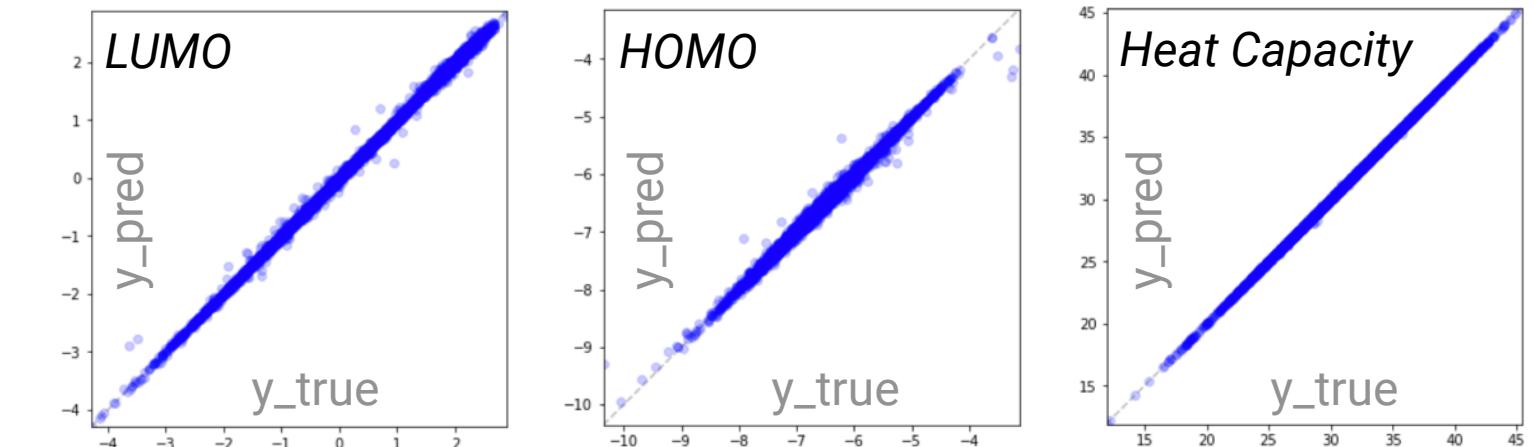
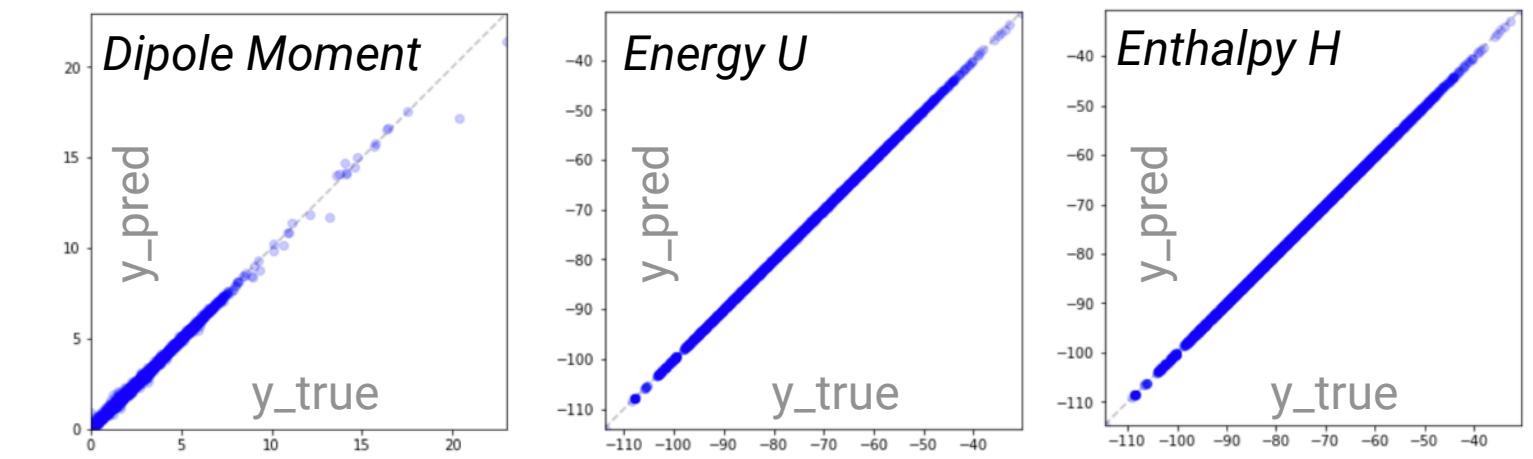
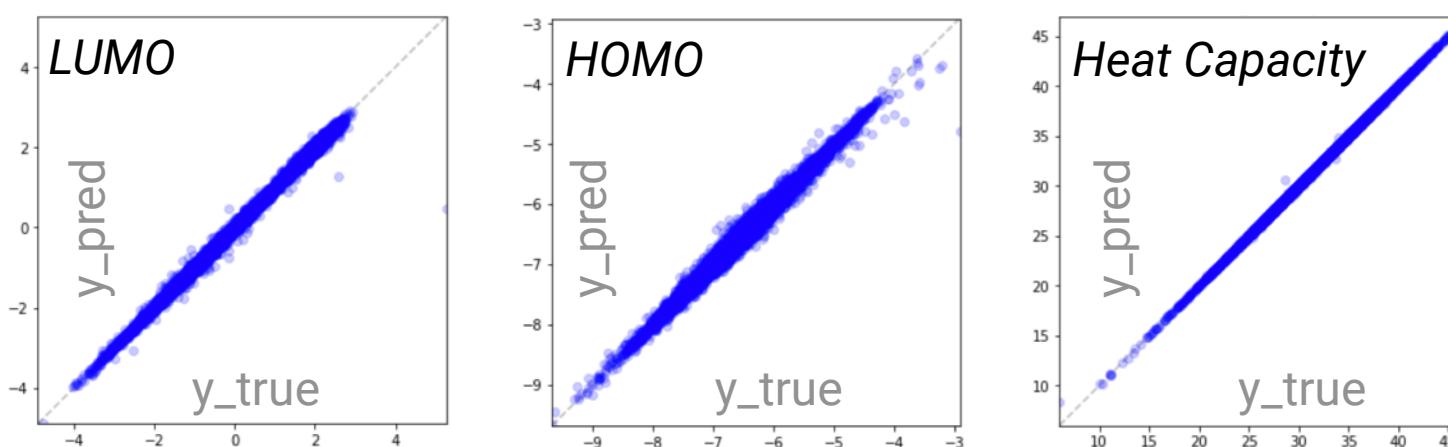
Use Case 2: Fast Approximation for QM Calculations

GNN predictions are **strikingly accurate**, in particular, for predicting **energies** of a molecule of a conformation or **forces at each atom** to transition towards a more stable conformation!

Predictions for Test Data by **SchNet** (Schütt et al, 2017)



Predictions for Test Data by **DimeNet** (Klicpera et al, 2020)



Big Challenge: Rashomon Effect and Underspecification

Rashomon Effect: The multiplicity of good ML models

In general, we can have **many good but very different ML models** that give equally accurate predictions for the given data.

Big Challenge: Rashomon Effect and Underspecification

Rashomon Effect: The multiplicity of good ML models

In general, we can have **many good but very different ML models** that give equally accurate predictions for the given data.

- Many explanations can exist for a single set of **finite** observations in general .
(whether they are given by ML or by human experts.)

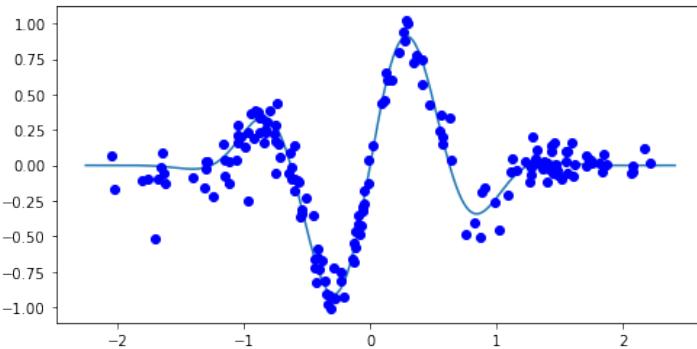
Big Challenge: Rashomon Effect and Underspecification

Rashomon Effect: The multiplicity of good ML models

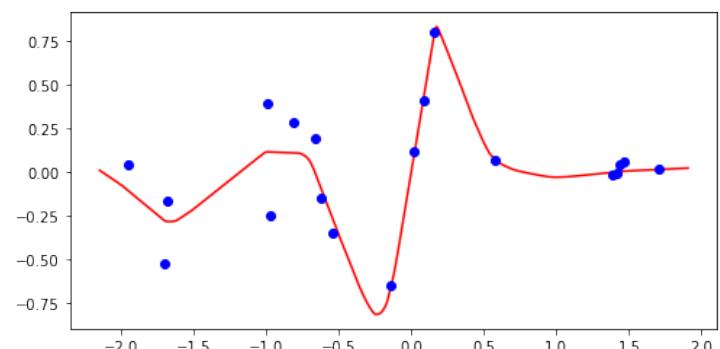
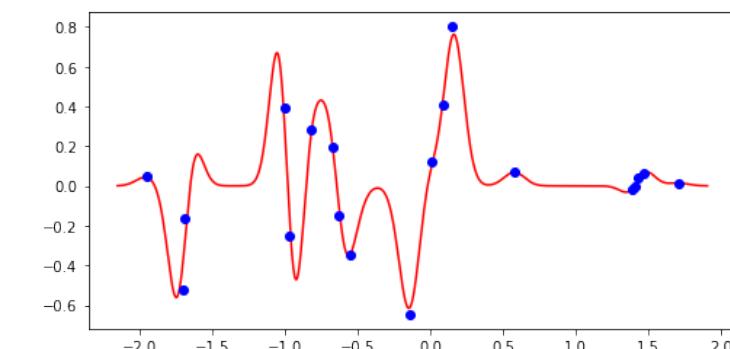
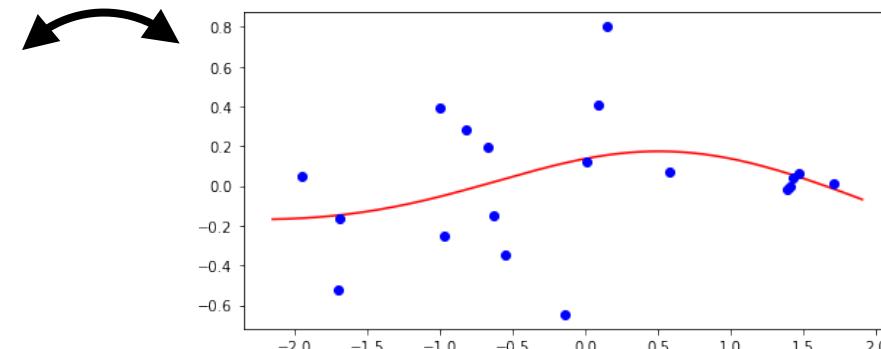
In general, we can have **many good but very different ML models** that give equally accurate predictions for the given data.

- Many explanations can exist for a single set of **finite** observations in general . (whether they are given by ML or by human experts.)
- They can **largely disagree in a underspecified situation** where data is statistically insufficient.

Any ML model will work



Different models can give very different predictions for out-of-sample cases



Designing Relevant Inductive Bias for Molecules

In reality, almost all cases might be *statistically insufficient* to fit modern ML models.

Designing Relevant Inductive Bias for Molecules

In reality, almost all cases might be *statistically insufficient* to fit modern ML models.

- Even the number of data is big for us, it can be small given the number of input variables and model parameters.

Designing Relevant Inductive Bias for Molecules

In reality, almost all cases might be *statistically insufficient* to fit modern ML models.

- Even the number of data is big for us, it can be small given the number of input variables and model parameters.
- Most experiments are planned by human scientists and therefore are subject to a variety of biases from human cognitive biases, heuristics and social influences.

Designing Relevant Inductive Bias for Molecules

In reality, almost all cases might be *statistically insufficient* to fit modern ML models.

- Even the number of data is big for us, *it can be small* given the number of input variables and model parameters.
- Most experiments are planned by human scientists and therefore are *subject to a variety of biases* from human cognitive biases, heuristics and social influences.
- Robotization/automation of experiments is promising in terms of reproducibility, but ... the chemical space is *astronomical (10^{60} or so)*. We might be able to have 10^6 robots, but 10^{60} would be physically hopeless...

Designing Relevant Inductive Bias for Molecules

In reality, almost all cases might be *statistically insufficient* to fit modern ML models.

- Even the number of data is big for us, *it can be small* given the number of input variables and model parameters.
- Most experiments are planned by human scientists and therefore are *subject to a variety of biases* from human cognitive biases, heuristics and social influences.
- Robotization/automation of experiments is promising in terms of reproducibility, but ... the chemical space is *astronomical (10^{60} or so)*. We might be able to have 10^6 robots, but 10^{60} would be physically hopeless...

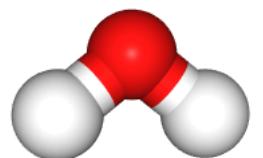
One remedy: Fusing rationalism (theory/simulation) and empiricism (ML/data-driven)

Encoding what we already know for ML *not to needlessly explore chemically invalid forms*

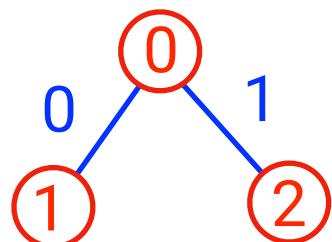
→ In other words, restrict ML models not to represent irrelevant input-output mappings.

Designing Relevant Inductive Bias for Molecules

Water molecule H_2O



A molecular graph (RDKit)



① ① ②

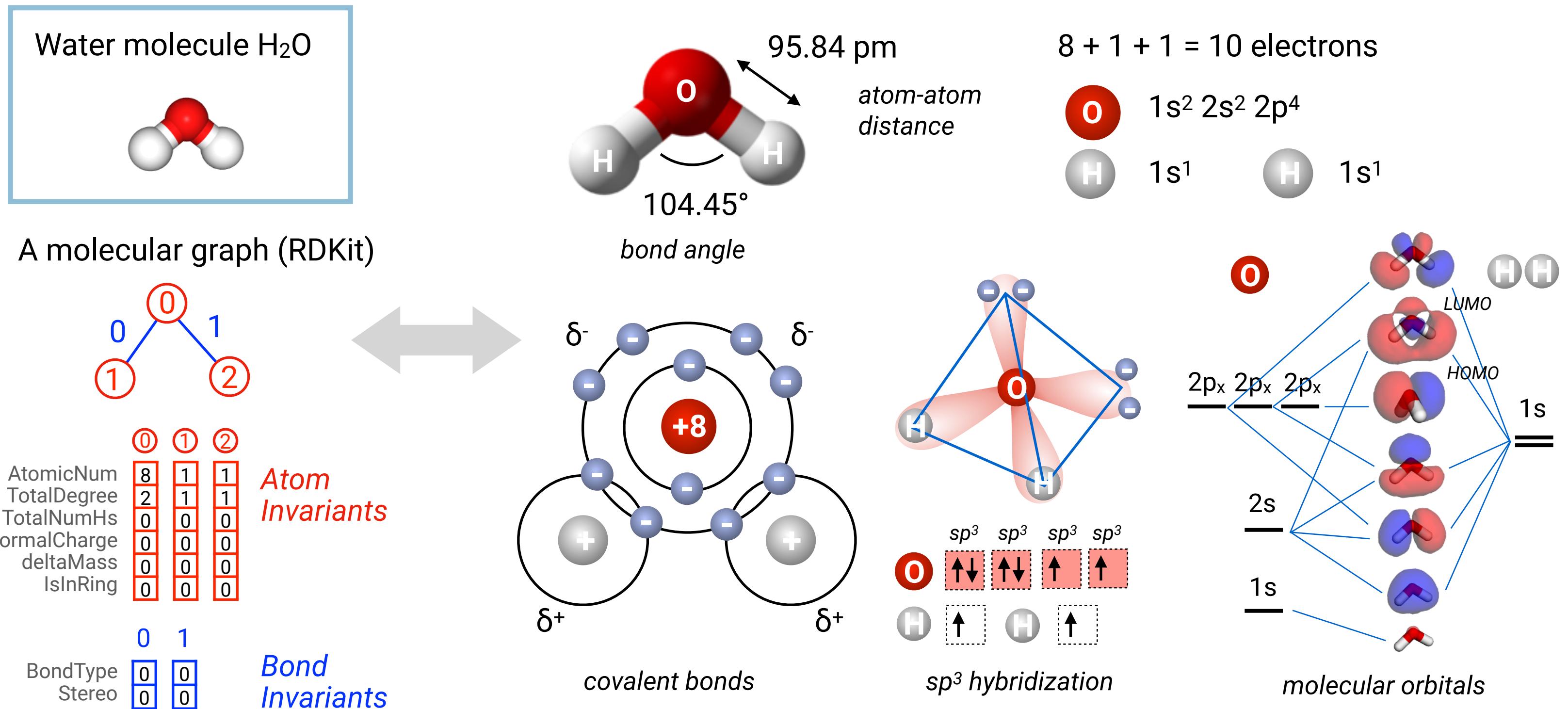
	0	1	2
AtomicNum	8	1	1
TotalDegree	2	1	1
TotalNumHs	0	0	0
FormalCharge	0	0	0
deltaMass	0	0	0
IsInRing	0	0	0

*Atom
Invariants*

BondType	0	1
Stereo	0	0

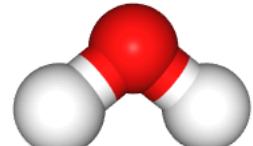
*Bond
Invariants*

Designing Relevant Inductive Bias for Molecules



SchNet (Schütt et al, 2017): Standard Geometric GNN

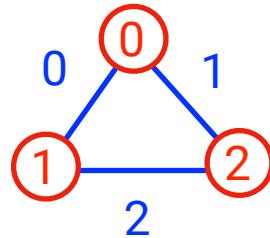
H₂O



gdb_3

	x	y	z	AN
O	-0.0344	0.9775	0.0076	8
H	0.0648	0.0206	0.0015	1
H	0.8718	1.3008	0.0007	1

Graph

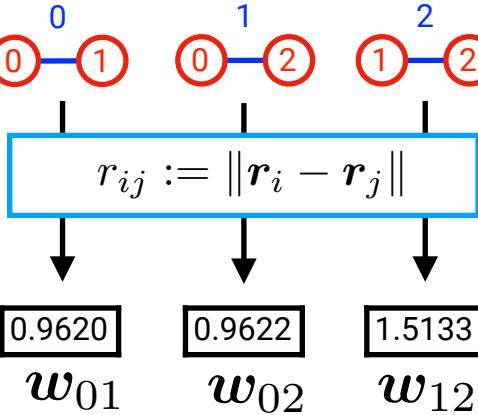


edges w/
cutoff (10Å)

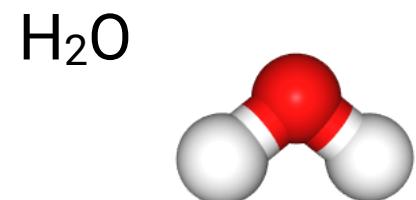
atom features

0	8	x_0
1	1	x_1
2	1	x_2

bond features

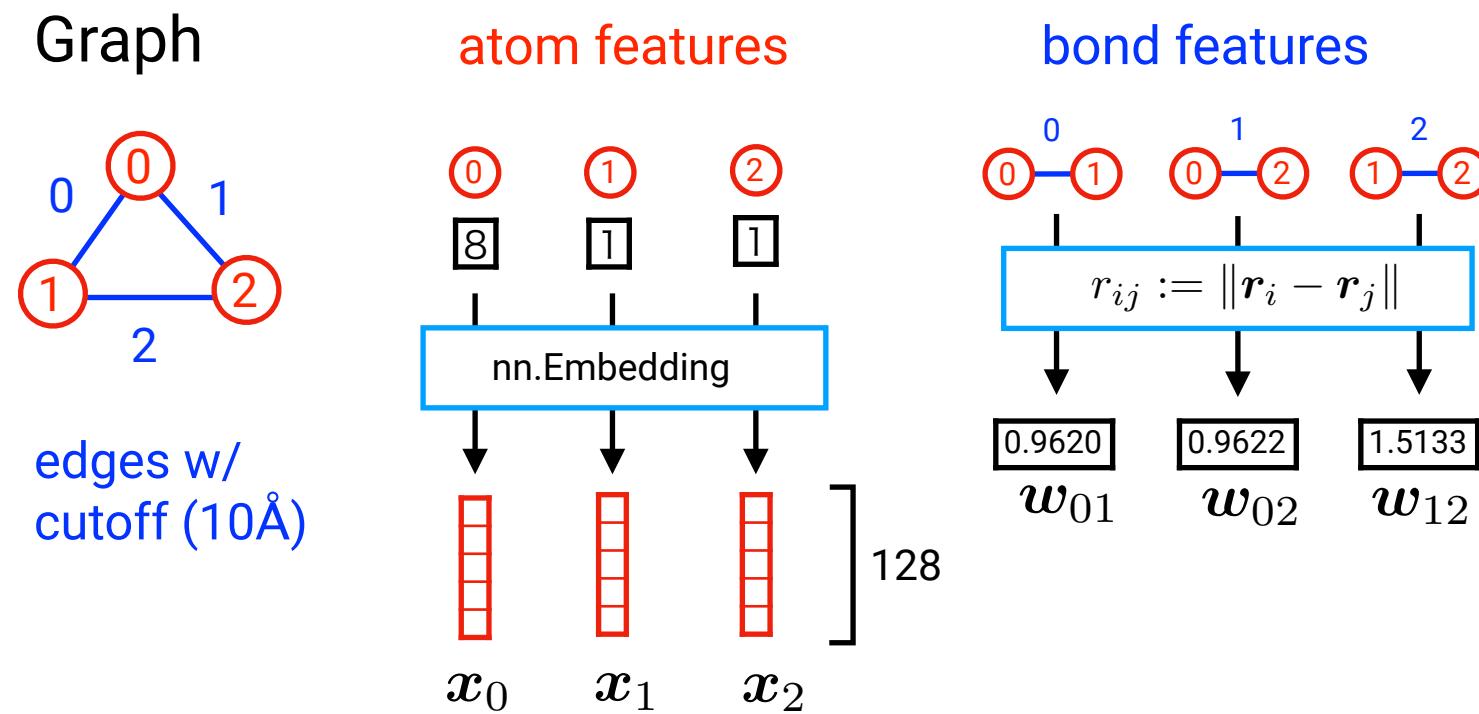


SchNet (Schütt et al, 2017): Standard Geometric GNN

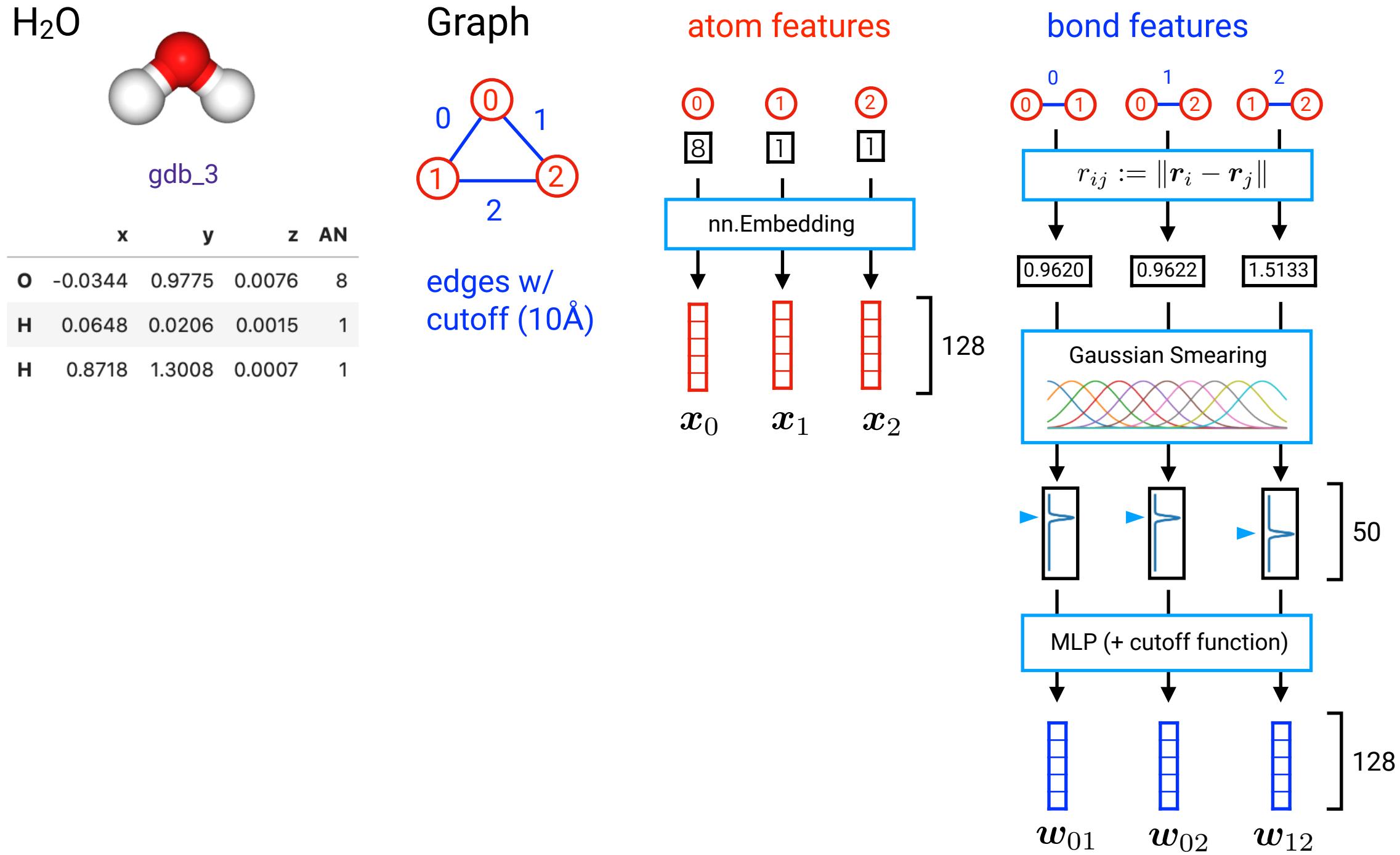


gdb_3

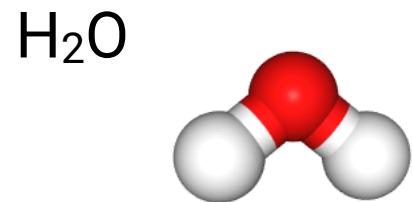
	x	y	z	AN
O	-0.0344	0.9775	0.0076	8
H	0.0648	0.0206	0.0015	1
H	0.8718	1.3008	0.0007	1



SchNet (Schütt et al, 2017): Standard Geometric GNN



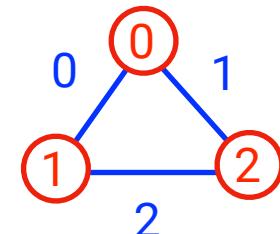
SchNet (Schütt et al, 2017): Standard Geometric GNN



gdb_3

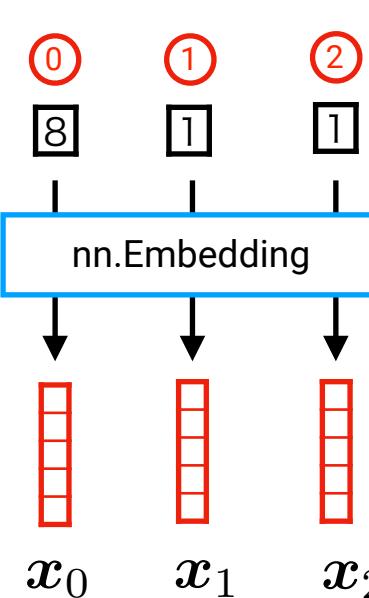
	x	y	z	AN
O	-0.0344	0.9775	0.0076	8
H	0.0648	0.0206	0.0015	1
H	0.8718	1.3008	0.0007	1

Graph

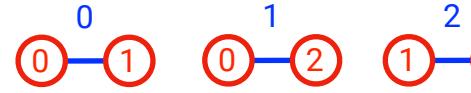


edges w/
cutoff (10Å)

atom features



bond features



$$r_{ij} := \|\mathbf{r}_i - \mathbf{r}_j\|$$

0.9620 0.9622 1.5133

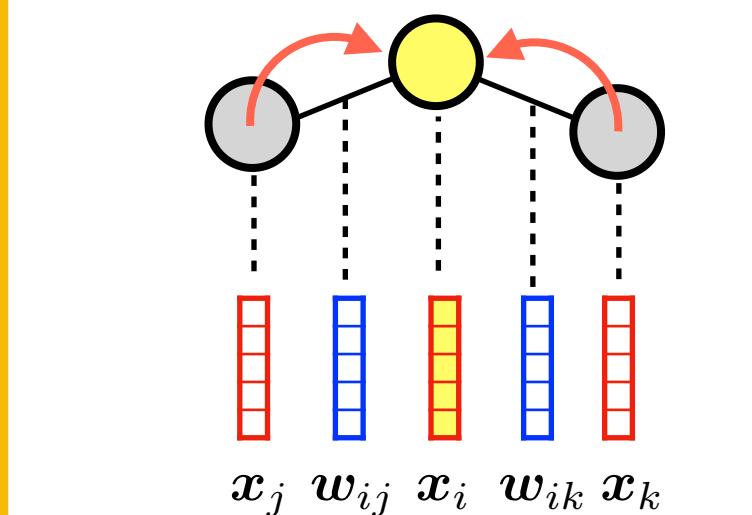
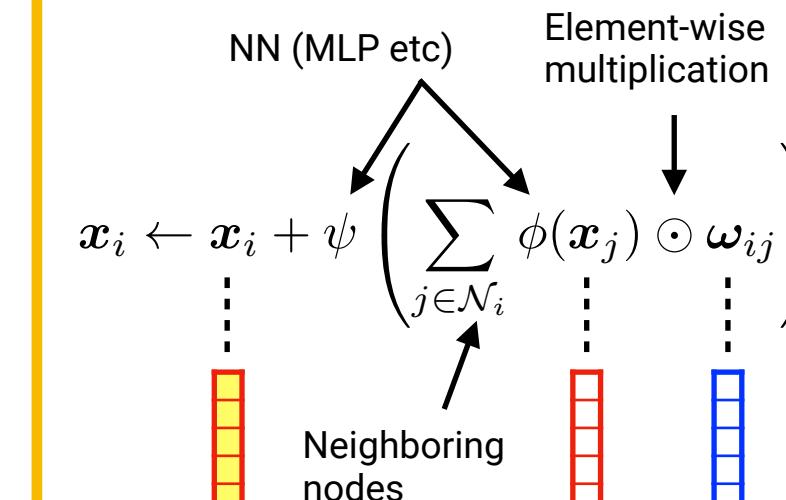


MLP (+ cutoff function)

128

w_{01} w_{02} w_{12}

Message Passing with
residual connections

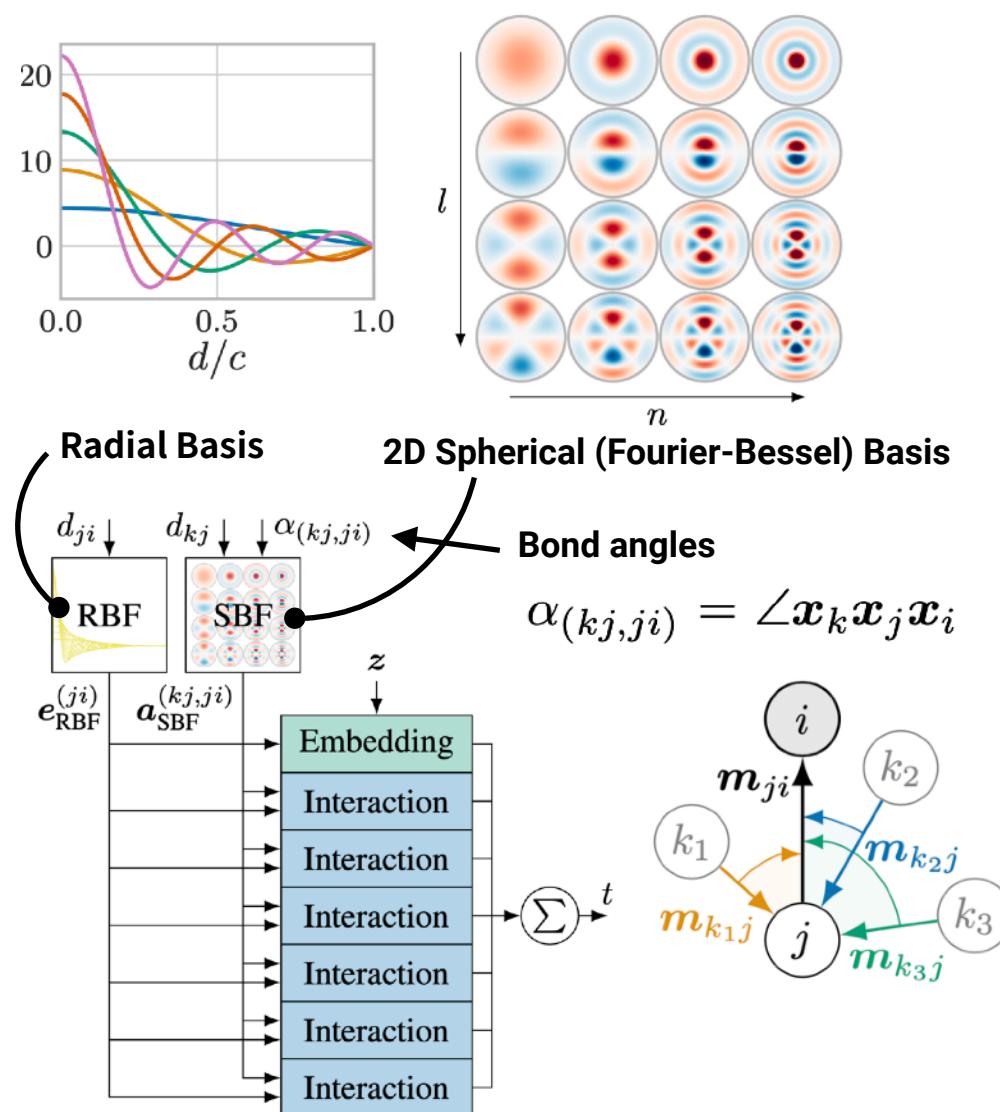


x_j w_{ij} x_i w_{ik} x_k

QSAR/QSPR, QM Approximation, Molecule Generations, ...

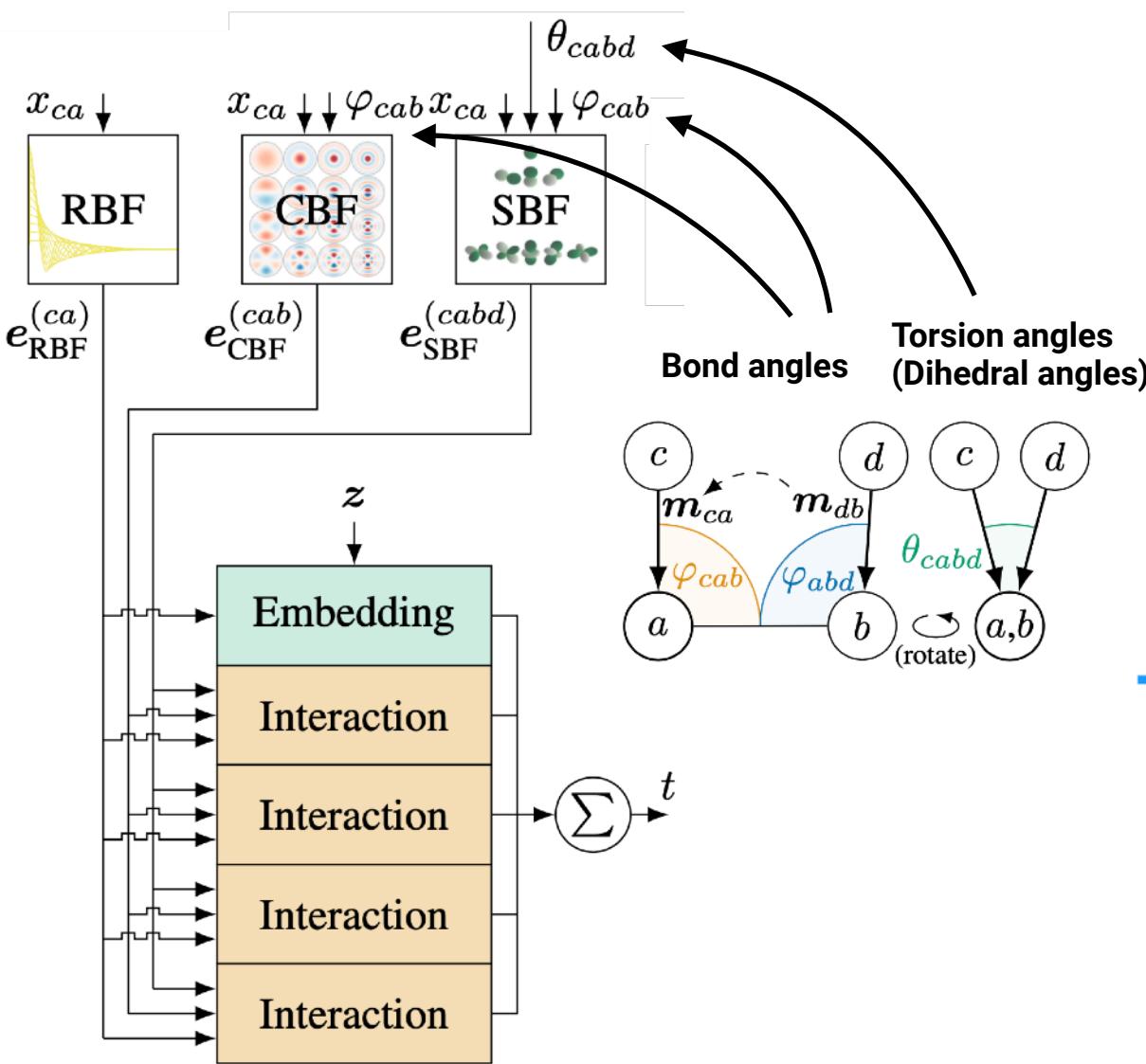
DimeNet++

Klicpera et al (NeurIPS WS2022)
<https://arxiv.org/abs/2011.14115>



GemNet

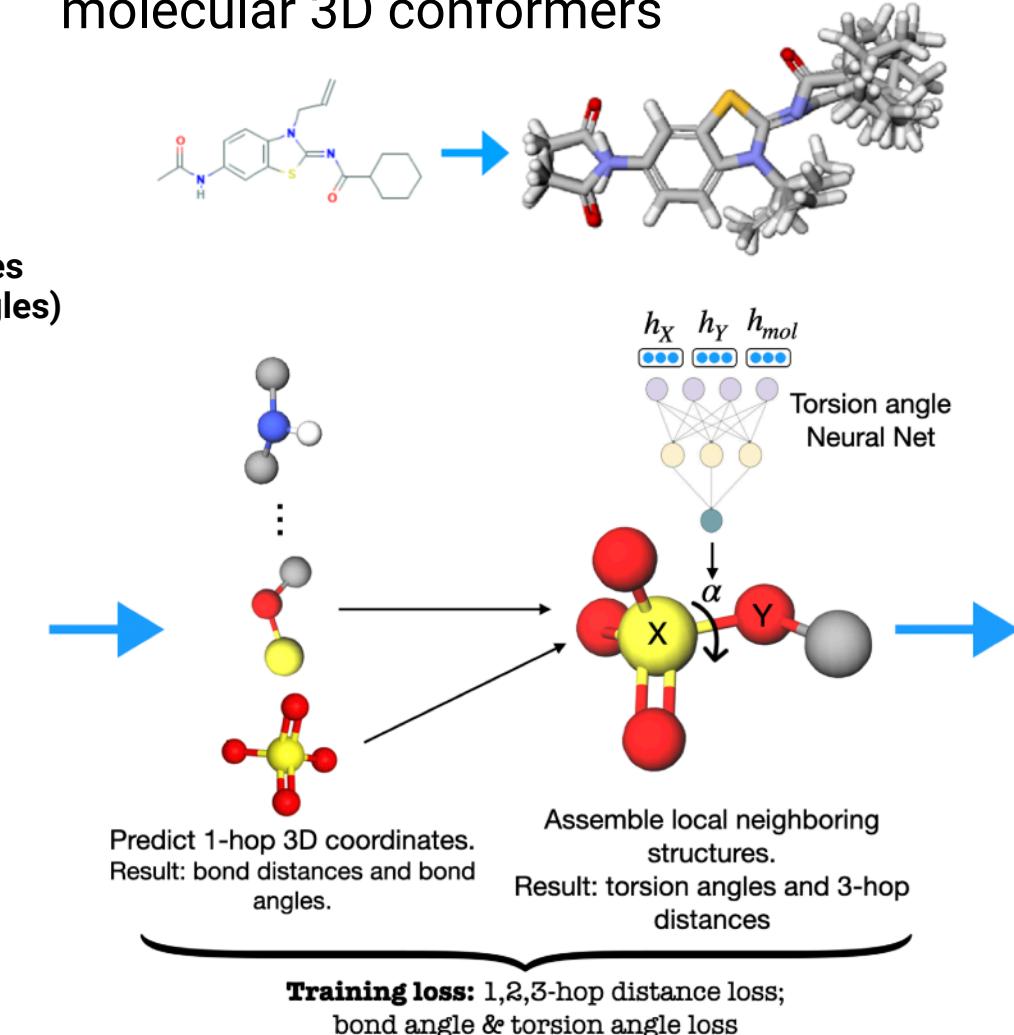
Klicpera et al (NeurIPS2021)
<https://arxiv.org/abs/2106.08903>



GeoMol

Ganea et al (NeurIPS2021)
<https://arxiv.org/abs/2106.07802>

generates distributions of low-energy molecular 3D conformers



"Learn to Simulate"

DeepMind > Research > Learning to Simulate Complex Physics with Graph Networks

PUBLICATIONS

SHARE

PUBLICATION LINKS

DOWNLOAD

VIEW PUBLICATION

DATASETS & CODE

VIDEO SITE

→ VIEW OPEN SOURCE

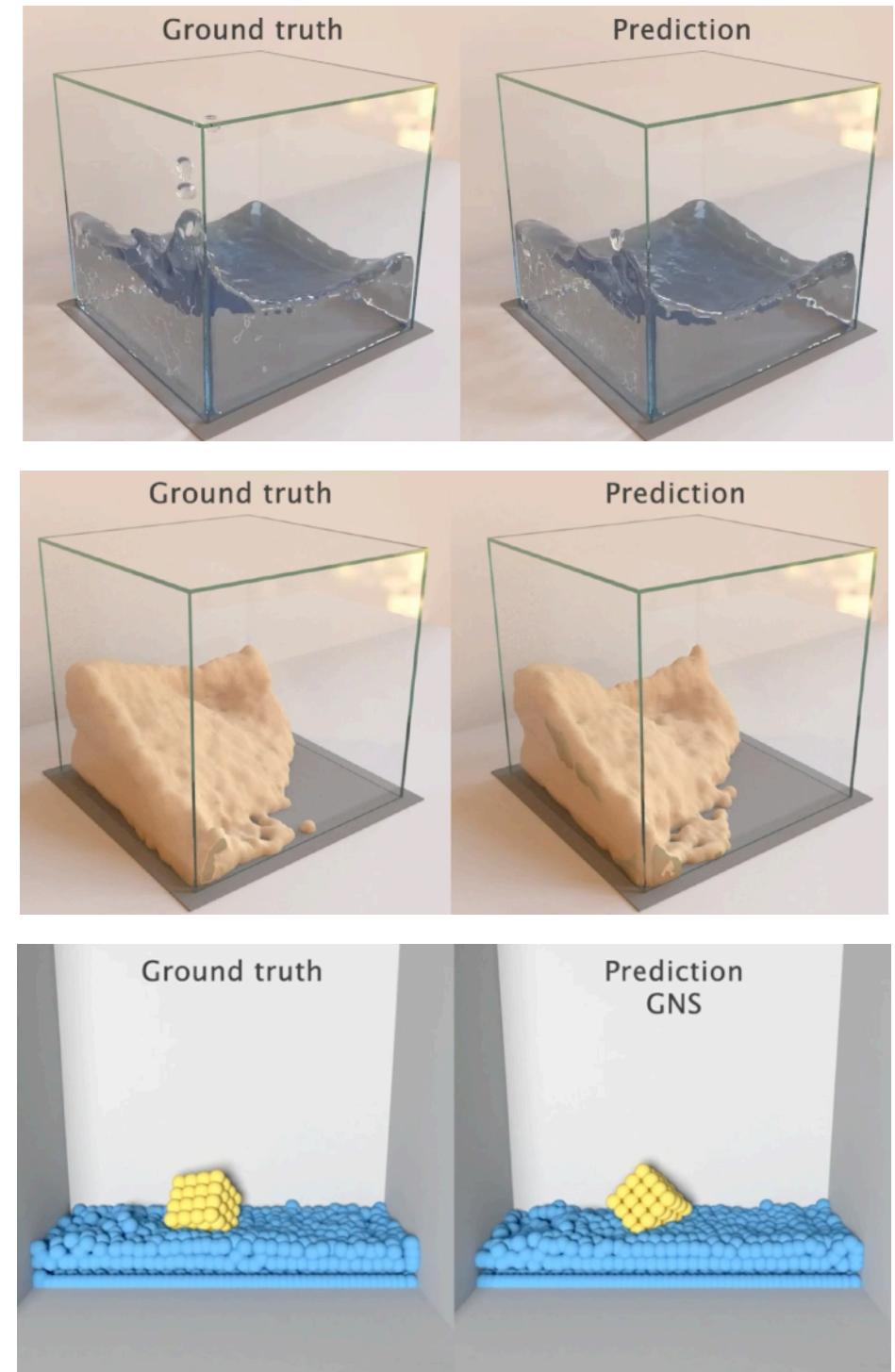
PUBLICATION
ICML

Learning to Simulate Complex Physics with Graph Networks

Abstract

Here we present a machine learning framework and model implementation that can learn to simulate a wide variety of challenging physical domains, involving fluids, rigid solids, and deformable materials interacting with one another. Our framework—which we term “Graph Network-based Simulators” (GNS)—represents the state of a physical system with particles, expressed as nodes in a graph, and computes dynamics via learned message-passing. Our results show that our model can generalize from single-timestep predictions with thousands of particles during training, to different initial conditions, thousands of timesteps, and at least an order of magnitude more particles at test time. Our model was robust to hyperparameter choices across various evaluation metrics: the main determinants of long-term performance were the number of message-passing steps, and mitigating the accumulation of error by corrupting the training data with noise. Our GNS framework advances the state-of-the-art in learned physical simulation, and holds promise for solving a wide range of complex forward and inverse problems.

Datasets and example model and training code available.



"Learn to Simulate"

DeepMind > Research > Learning to Simulate Complex Physics with Graph Networks

PUBLICATIONS

SHARE

PUBLICATION LINKS

DOWNLOAD

VIEW PUBLICATION

DATASETS & CODE

VIDEO SITE

→ VIEW OPEN SOURCE

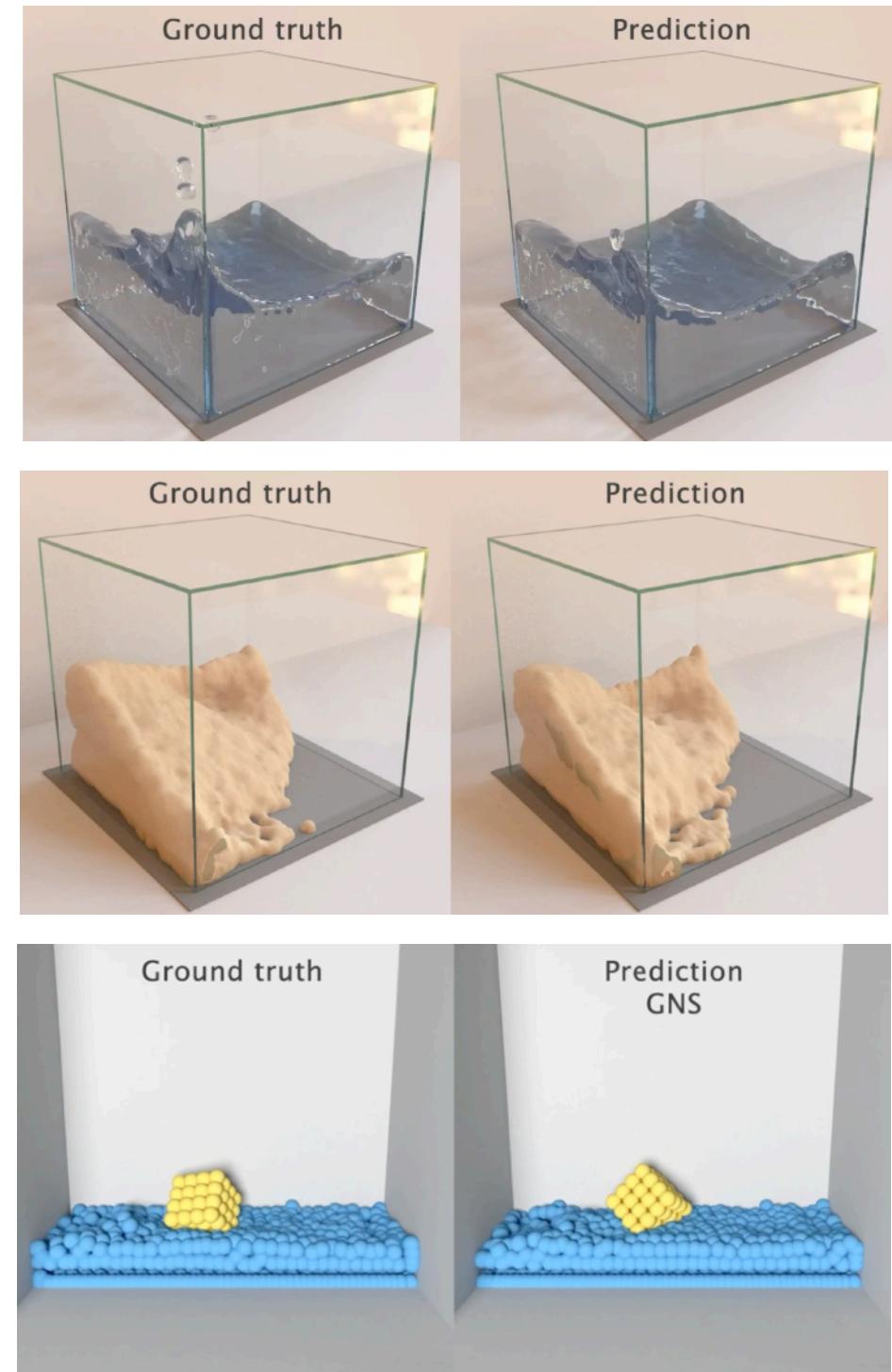
PUBLICATION
ICML

Learning to Simulate Complex Physics with Graph Networks

Abstract

Here we present a machine learning framework and model implementation that can learn to simulate a wide variety of challenging physical domains, involving fluids, rigid solids, and deformable materials interacting with one another. Our framework—which we term “Graph Network-based Simulators” (GNS)—represents the state of a physical system with particles, expressed as nodes in a graph, and computes dynamics via learned message-passing. Our results show that our model can generalize from single-timestep predictions with thousands of particles during training, to different initial conditions, thousands of timesteps, and at least an order of magnitude more particles at test time. Our model was robust to hyperparameter choices across various evaluation metrics: the main determinants of long-term performance were the number of message-passing steps, and mitigating the accumulation of error by corrupting the training data with noise. Our GNS framework advances the state-of-the-art in learned physical simulation, and holds promise for solving a wide range of complex forward and inverse problems.

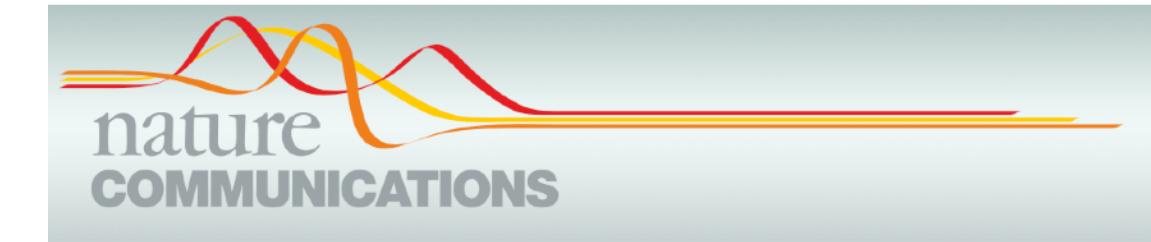
Datasets and example model and training code available.



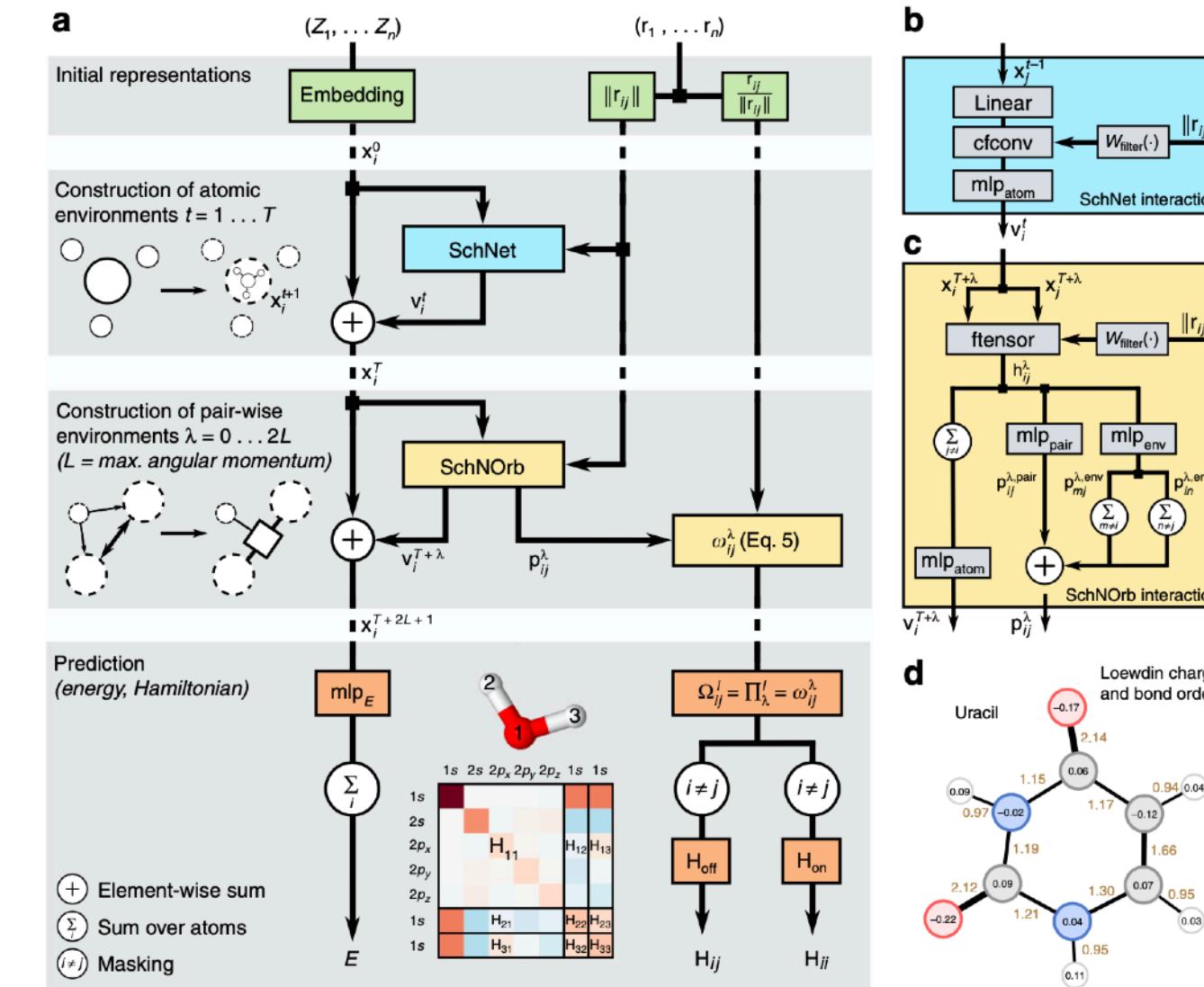
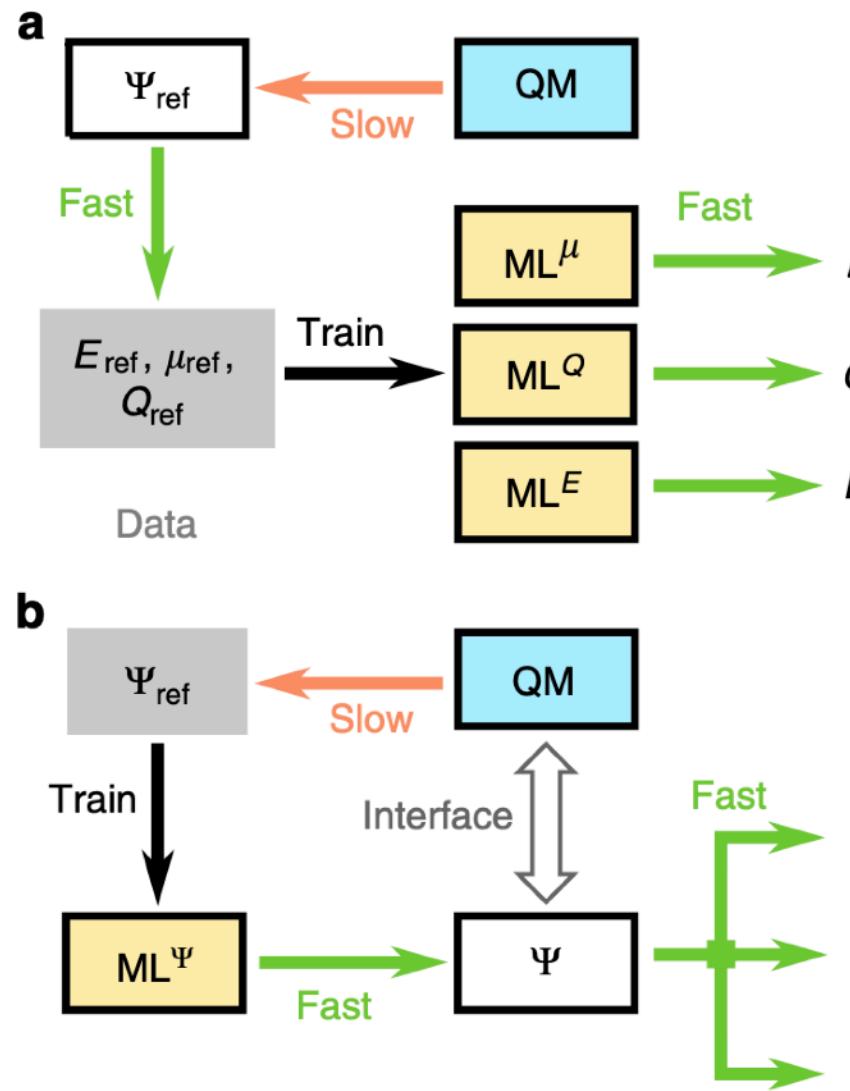
Fusing ML + Quantum Chemistry

**Unifying machine learning and quantum chemistry
with a deep neural network for molecular
wavefunctions**

K. T. Schütt, M. Gastegger, A. Tkatchenko✉, K.-R. Müller✉ & R. J. Maurer✉



Nature Communications 10, Article number: 5024 (2019)



Fusing ML + Quantum Chemistry

Machine Learning at the Atomic Scale (Chem. Rev.)

<https://pubs.acs.org/toc/chreay/121/16>

CHEMICAL REVIEWS

pubs.acs.org/CR

Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems

John A. Keith,* Valentin Vassilev-Galindo, Bingqing Cheng, Stefan Chmiela, Michael Gastegger, Klaus-Robert Müller,* and Alexandre Tkatchenko*

 Cite This: <https://doi.org/10.1021/acs.chemrev.1c00107>

 Read Online



Review

Data Science Meets Chemistry (Acc. Chem. Res.)

<https://pubs.acs.org/page/achre4/data-science-meets-chemistry>

CHEMICAL REVIEWS

pubs.acs.org/CR

Physics-Inspired Structural Representations for Molecules and Materials

Felix Musil, Andrea Grisafi, Albert P. Bartók, Christoph Ortner, Gábor Csányi, and Michele Ceriotti*

 Cite This: *Chem. Rev.* 2021, 121, 9759–9815

 Read Online



Review

CHEMICAL REVIEWS

pubs.acs.org/CR

Ab Initio Machine Learning in Chemical Compound Space

Bing Huang and O. Anatole von Lilienfeld*

 Cite This: *Chem. Rev.* 2021, 121, 10001–10036

 Read Online



Review

ACCOUNTS of chemical research

pubs.acs.org/accounts

Article

Learning to Approximate Density Functionals

Published as part of the Accounts of Chemical Research special issue “Data Science Meets Chemistry”.

Bhupalee Kalita, Li Li, Ryan J. McCarty, and Kieron Burke*

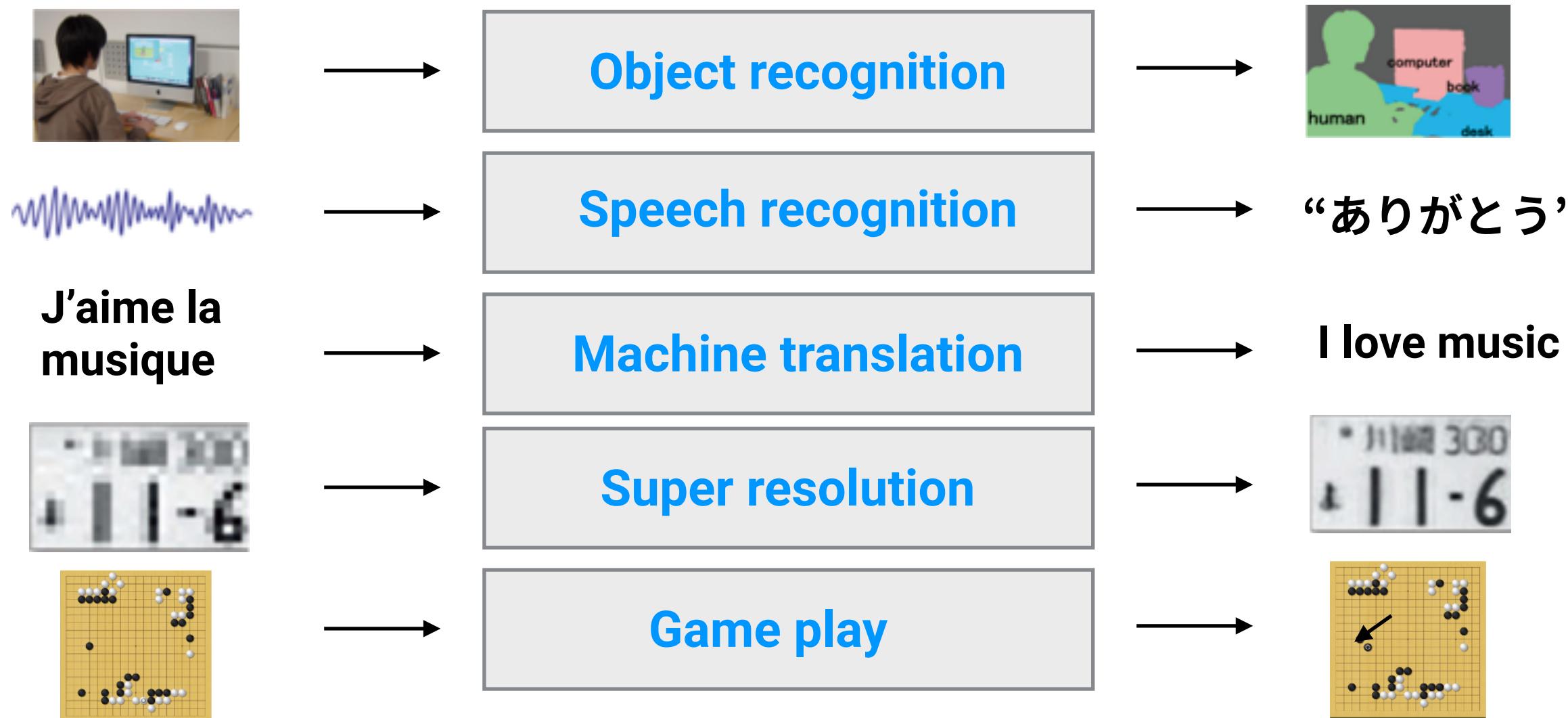
 Cite This: *Acc. Chem. Res.* 2021, 54, 818–826

 Read Online



One Final Remark: "The True Dark Side"

"We're able to predict" does not directly mean "We're able to understand" nor "We're able to discover". We still need further considerations to impact natural sciences...



Summary

This slide is available at

<https://itakigawa.github.io/news.html>

Machine Learning (ML) for Molecules

1. ML in a nutshell

- *ML converts data into "prediction"*
- *ML is a new (lazy) way of programming*

2. The dark side:

Modern aspects of ML

- *High dimensionality: Too many input variables*
- *Overrepresentation: Too many parameters*
- *Data hungriness: Big data is big for human, but can be too small for ML models...*

3. The light side :

Deep learning for molecules

- *Graph Neural Networks (GNNs)*
- *Case 1: Virtual Screening (QSAR/QSPR)*
- *Case 2: Fast Approximation for QM Calculations*

4. Challenges

- *Rashomon Effect and Underspecification*
- *Designing Relevant Inductive Bias for Chemistry*
- *Prediction does not directly mean Understanding or Discovery*

May the ML Force be with you...