



Machine Learning for Chemistry: Representing and Intervening

Ichigaku Takigawa

takigawa@icredd.hokudai.ac.jp

Joint Symposium of Engineering & Information Science & WPI-ICReDD

Apr 26, 2021 @ Hokkaido University

I am a graduate of School of Engineering and IST!

1995-2005 (10 years) Hokkaido Univ

School of Engineering
Grad School of Engineering
Grad School of Info Sci & Tech

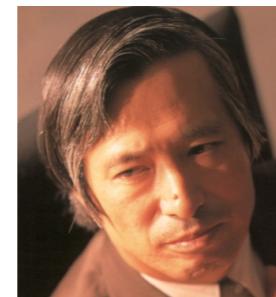
B.Eng (1999)
M.Eng (2001), PhD (2004)
Postdoc (2004-2005)



SHIMBO Masaru



KUDO Mineichi



TANAKA Yuzuru

2012-2019 (7 years) Hokkaido Univ

Grad School of Info Sci & Tech

Tenure Track (2012-2014)
Assoc Prof (2014-2019)



IMAI Hideyuki



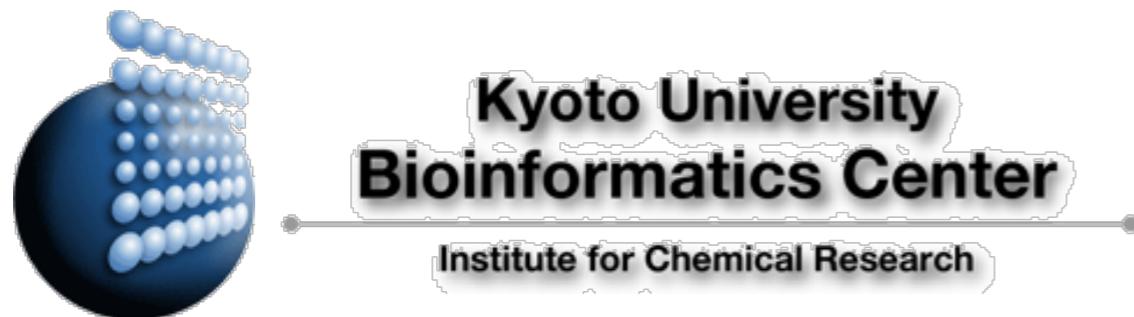
TANAKA Yuzuru



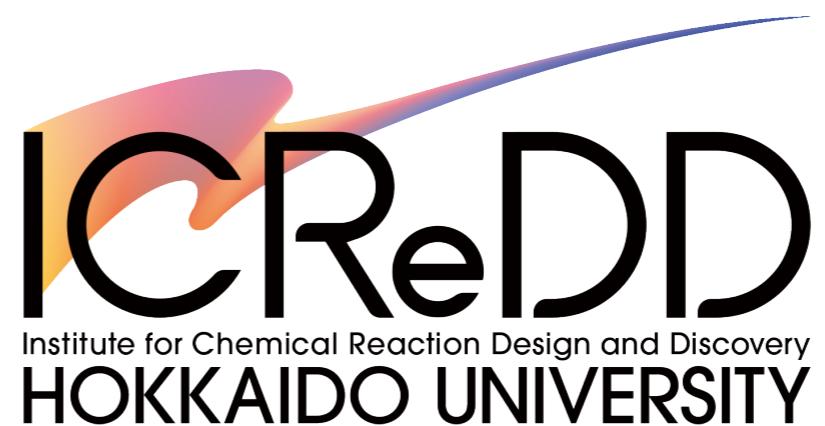
MINATO Shinichi

But when I stepped outside

2005-2011 (7 years) Kyoto Univ



2019-present (2 years) The “Cross-Appointment System”

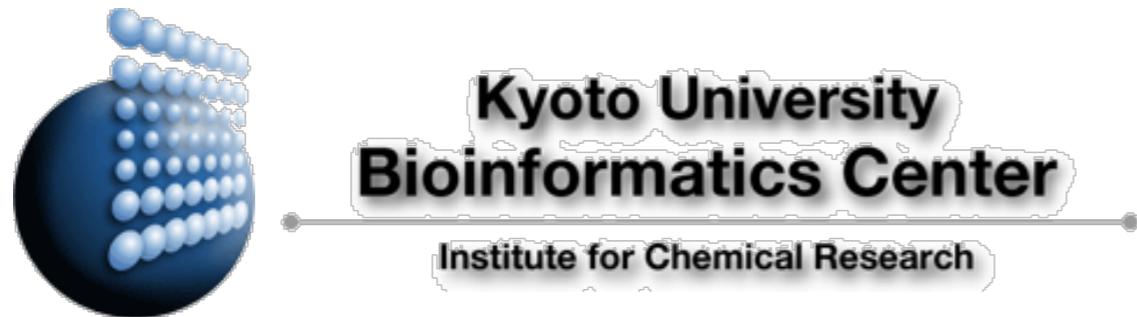


Physically I'm at Kyoto



Things go interdisciplinary...

2005-2011 (7 years) Kyoto Univ



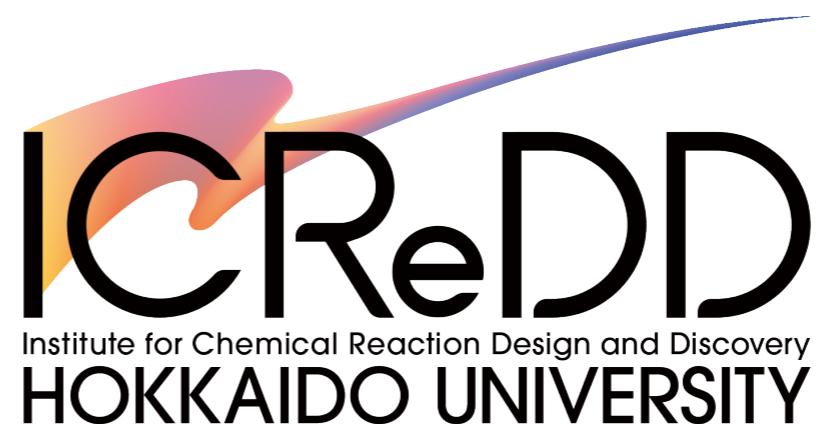
Assist Prof (2005-2011)

- Bioinformatics Center
Institute for Chemical Research
- Grad School of Pharmaceutical Sci

2019-present (2 years) The “Cross-Appointment System”



- Medical-risk Avoidance based on iPS Cells Team



- Institute for Chemical Reaction Design and Discovery



This talk

Machine Learning (ML) for Chemistry

- Why it is needed?
- What are exciting for computer scientists?

It's a hot topic in Chemistry

Science

REVIEW

Inverse molecular design using machine learning: Generative models for matter engineering

Benjamin Sanchez-Lengeling¹ and Alán Aspuru-Guzik^{2,3,4*}

Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning

Andrew F. Zahrt*, Jeremy J. Henle*, Brennan T. Rose, Yang Wang,
William T. Darrow, Scott E. Denmark†

nature reviews chemistry

REVIEWS

Synthetic organic chemistry driven by artificial intelligence

A. Filipa de Almeida¹, Rui Moreira¹ and Tiago Rodrigues^{2*}

PERSPECTIVES

Exploring chemical compound space with quantum-based machine learning

O. Anatole von Lilienfeld, Klaus-Robert Müller and Alexandre Tkatchenko¹

nature

REVIEW

Machine learning for molecular and materials science

Keith T. Butler¹, Daniel W. Davies², Hugh Cartwright³, Olexandr Isayev^{4*} & Aron Walsh^{5,6*}

Planning chemical syntheses with deep neural networks and symbolic AI

Marwin H. S. Segler^{1,2}, Mike Preuss³ & Mark P. Waller⁴

Holistic prediction of enantioselectivity in asymmetric catalysis

Jolene P. Reid¹ & Matthew S. Sigman^{1*}

Bayesian reaction optimization as a tool for chemical synthesis

Benjamin J. Shields¹, Jason Stevens², Jun Li², Marvin Parasram¹, Farhan Damani³, Jesus I. Martinez Alvarado¹, Jacob M. Janey², Ryan P. Adams³ & Abigail G. Doyle¹



But also in Machine Learning!

NeurIPS 2020

- *Self-Supervised Graph Transformer on Large-Scale Molecular Data*
- *RetroXpert: Decompose Retrosynthesis Prediction Like A Chemist*
- *Reinforced Molecular Optimization with Neighborhood-Controlled Grammars*
- *Autofocused Oracles for Model-based Design*
- *Barking Up the Right Tree: an Approach to Search over Molecule Synthesis DAGs*
- *On the Equivalence of Molecular Graph Convolution and Molecular Wave Function with Poor Basis Set*
- *CogMol: Target-Specific and Selective Drug Design for COVID-19 Using Deep Generative Models*

ICML 2020

- *A Graph to Graphs Framework for Retrosynthesis Prediction*
- *Hierarchical Generation of Molecular Graphs using Structural Motifs*
- *Learning to Navigate in Synthetically Accessible Chemical Space Using Reinforcement Learning*
- *Reinforcement Learning for Molecular Design Guided by Quantum Mechanics*
- *Multi-Objective Molecule Generation using Interpretable Substructures*
- *Improving Molecular Design by Stochastic Iterative Target Augmentation*
- *A Generative Model for Molecular Distance Geometry*

ICLR 2020

- *Directional Message Passing for Molecular Graphs*
- *GraphAF: a Flow-based Autoregressive Model for Molecular Graph Generation*
- *Augmenting Genetic Algorithms with Deep Neural Networks for Exploring the Chemical Space*
- *A Fair Comparison of Graph Neural Networks for Graph Classification*

Mixed feelings of curiosity, optimism, skepticism?

Computer Chemistry

How to cite: *Angew. Chem. Int. Ed.* **2020**, *59*, 18860–18865
International Edition: doi.org/10.1002/anie.202008366
German Edition: doi.org/10.1002/ange.202008366

Molecular Machine Learning: The Future of Synthetic Chemistry?

Philipp M. Pflüger and Frank Glorius*

Reaction Prediction

International Edition: DOI: 10.1002/anie.201803562
German Edition: DOI: 10.1002/ange.201803562

Machine Learning for Organic Synthesis: Are Robots Replacing Chemists?

Boris Maryasin, Philipp Marquetand, and Nuno Maulide*



AI-Assisted Synthesis Very Important Paper

International Edition: DOI: 10.1002/anie.201912083
German Edition: DOI: 10.1002/ange.201912083

Synergy Between Expert and Machine-Learning Approaches Allows for Improved Retrosynthetic Planning

Tomasz Badowski, Ewa P. Gajewska, Karol Molga, and Bartosz A. Grzybowski*

Computer-Aided Synthetic Planning

International Edition: DOI: 10.1002/anie.201506101
German Edition: DOI: 10.1002/ange.201506101

Computer-Assisted Synthetic Planning: The End of the Beginning

Sara Szymkuć, Ewa P. Gajewska, Tomasz Klucznik, Karol Molga, Piotr Dittwald, Michał Startek, Michał Bajczyk, and Bartosz A. Grzybowski*



Inseparably linked to automation

Automation is the use of technology to perform tasks with reduced human involvement or human labor.

“These illustrate how rapid advancements in hardware automation and machine learning continue to transform the nature of experimentation and modeling.”

Machine Learning

How to cite:

International Edition: doi.org/10.1002/anie.201909987

German Edition: doi.org/10.1002/ange.201909987

Autonomous Discovery in the Chemical Sciences Part I: Progress

Connor W. Coley, Natalie S. Eyke, and Klavs F. Jensen**

Computer Chemistry

How to cite:

International Edition: doi.org/10.1002/anie.201909989

German Edition: doi.org/10.1002/ange.201909989

Autonomous Discovery in the Chemical Sciences Part II: Outlook

Connor W. Coley, Natalie S. Eyke, and Klavs F. Jensen**



Towards machine autonomy in discovery

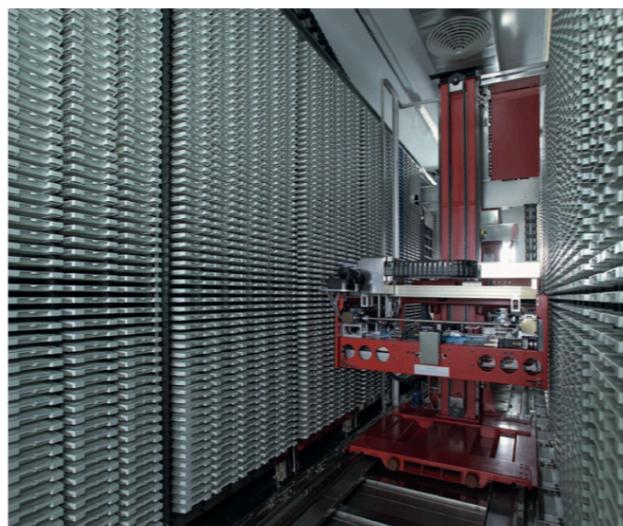
Automation has been impactfully changing our daily life, society, as well as scientific experiments and computations.



Organic synthesis in a modular robotic system. *Science* 363 (2019)



A mobile robotic chemist. *Nature* 583 (2020)



Automating drug discovery. *Nature Reviews Drug Discovery* 17 (2018)



This talk

Machine Learning (ML) for Chemistry

- Why it is needed?
- What are exciting for computer scientists?

I'll briefly cover these from two aspects:

1. Representation

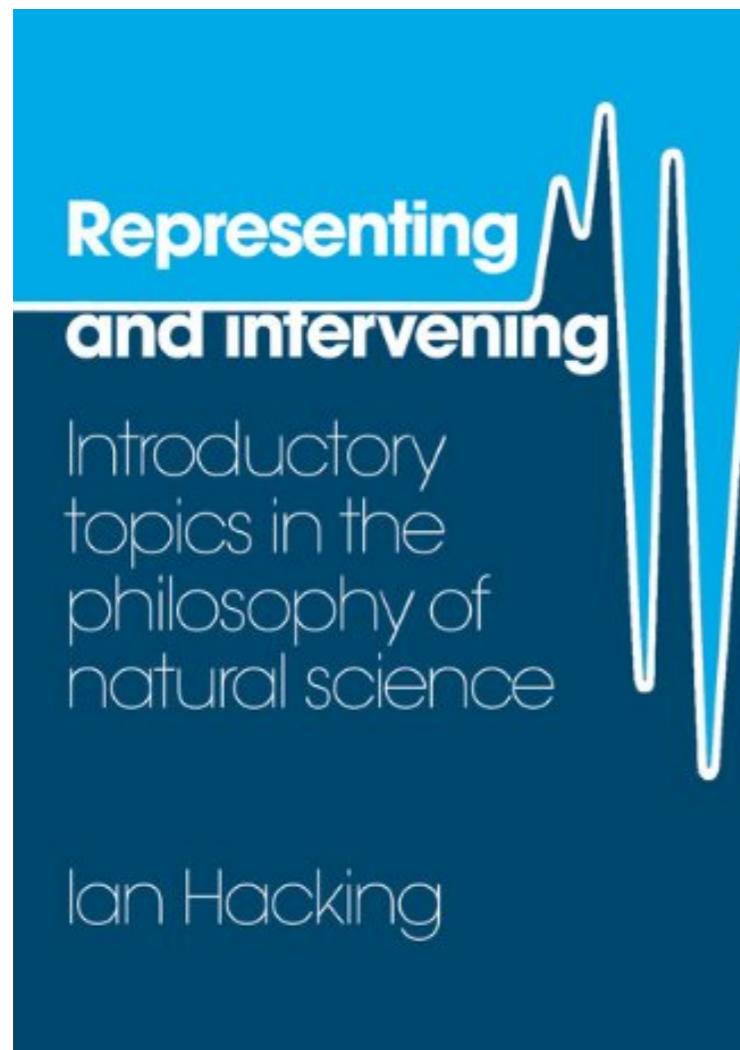
- What are good ML-readable representations for chemistry?
- What information should be recorded and given to ML?

2. (Experimental) Intervention

- What are essential to make real chemical discoveries?
- Any principled ways for data acquisition and experimental design?

Two pillars for scientific discovery?

In essence, ML for chemistry is *metascience (the science on how to do science)* unexpectedly hitting age-old unsolved questions in the philosophy of natural science.



Machine Learning (ML)



Feb 19, 2020, 08:00am EST | 2,187 views

In Praise Of Boring AI (A.K.A. Machine Learning)

JC Schutterle Forbes Councils Member
Forbes Technology Council COUNCIL POST | Paid Program

Matt Velloso, a technical advisor to Microsoft's CEO, got 24,000 likes on this [tweet](#) posted in November 2018: “Difference between machine learning and AI: If it is written in Python, it's probably machine learning. If it is written in PowerPoint, it's probably AI.”

:

Whether or not machine learning is paving the way for a sci-fi movie type of AI in the distant future is a pointless question. The benefits of a data-driven approach to automating nitty-gritty processes and transforming organizations as a whole are far from being exhausted. Machine learning offers enough value potential for the new decade. It's time to stop staring at boring PowerPoint decks and start coding in Python. It's time for boring AI.

<https://www.forbes.com/sites/forbestechcouncil/2020/02/19/in-praise-of-boring-ai-a-k-a-machine-learning/>

The AI frenzy: hope & hype
“Let's face it:
So far, the artificial intelligence plastered all over PowerPoint slides hasn't lived up to its hype.”

Machine Learning (ML)



Feb 19, 2020, 08:00am EST | 2,187 views

In Praise Of Boring AI (A.K.A. Machine Learning)

JC Schutterle Forbes Councils Member
Forbes Technology Council COUNCIL POST | Paid Program

Matt Velloso, a technical advisor to Microsoft's CEO, got 24,000 likes on this [tweet](#) posted in November 2018: "Difference between machine learning and AI: If it is written in Python, it's probably machine learning. If it is written in PowerPoint, it's probably AI."

:

Whether or not machine learning is paving the way for a sci-fi movie type of AI in the distant future is a pointless question. The benefits of a data-driven approach to automating nitty-gritty processes and transforming organizations as a whole are far from being exhausted. Machine learning offers enough value potential for the new decade. It's time to stop staring at boring PowerPoint decks and start coding in Python. It's time for boring AI.

<https://www.forbes.com/sites/forbestechcouncil/2020/02/19/in-praise-of-boring-ai-a-k-a-machine-learning/>

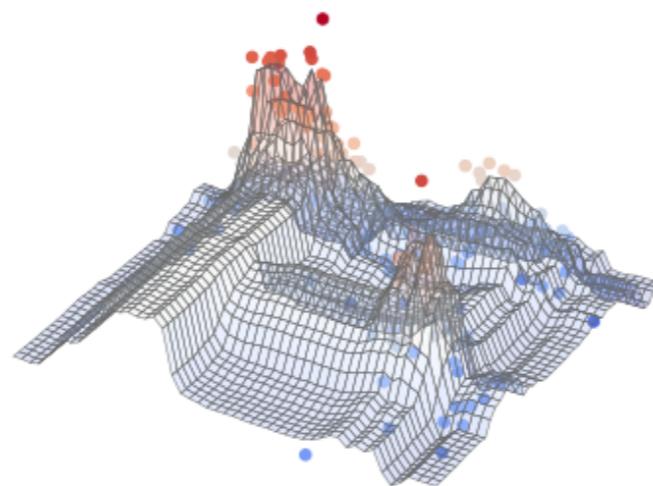
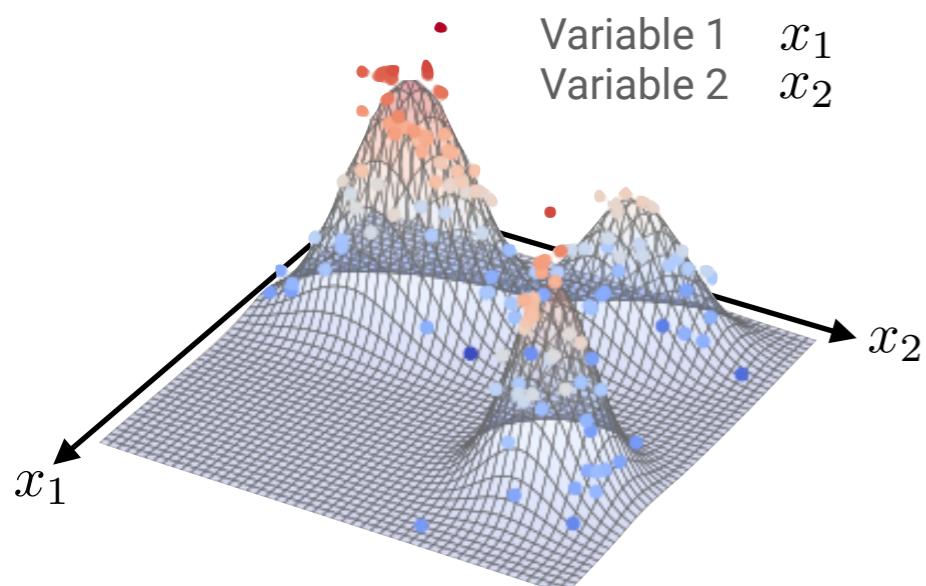
The AI frenzy: hope & hype
“Let’s face it:
So far, the artificial
intelligence plastered all
over PowerPoint slides
hasn’t lived up to its hype.”



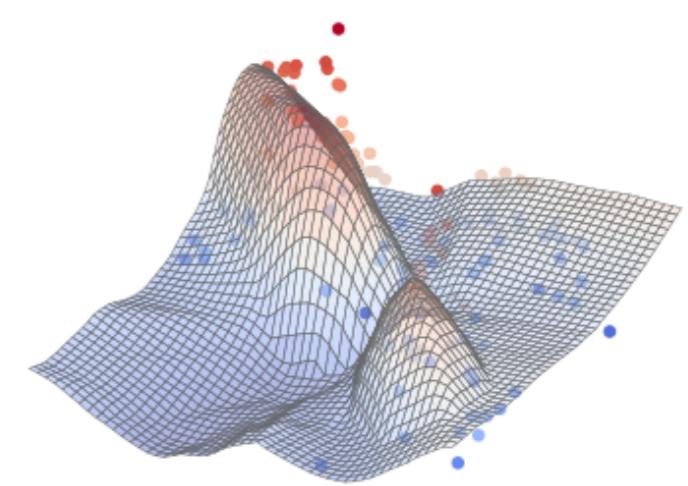
From AAAI-20 Oxford-Style Debate

Machine Learning (ML)

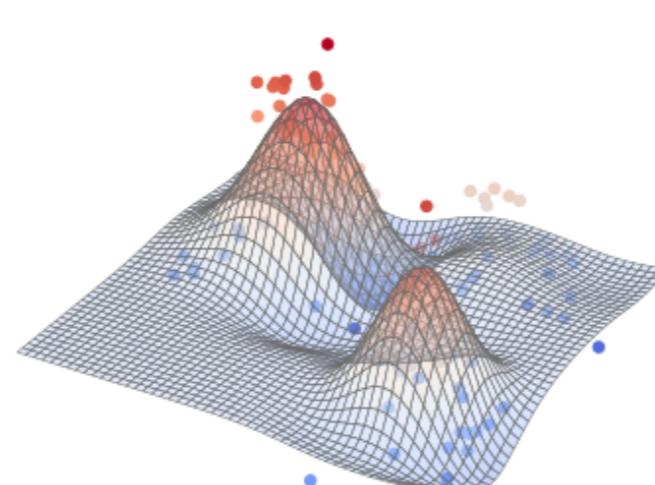
All about statistical and algorithmic techniques for surface-model fitting to data points by adjusting model parameters.



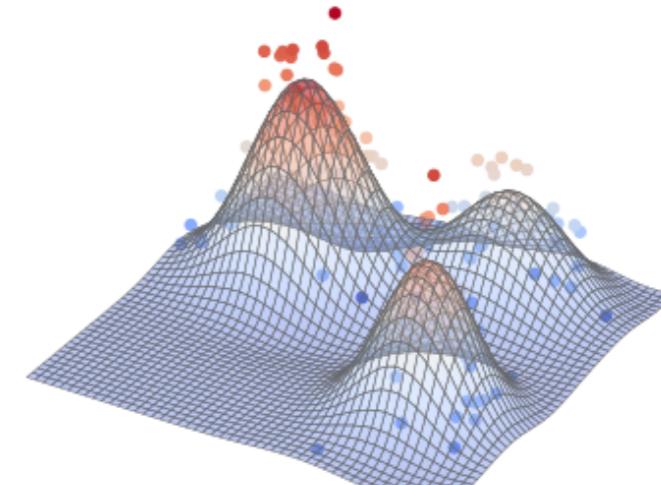
Random Forest



Neural Networks



SVR



Kernel Ridge

"Predictive Modeling"

Fitted surface used for making predictions on unseen data points

Modern aspects of ML

1. High dimensionality: Data can have many input variables.

a 100x100 pixel grayscale image = 10000 input variables (a 10000-dimensional array)

Modern aspects of ML

1. High dimensionality: Data can have many input variables.

a 100x100 pixel grayscale image = 10000 input variables (a 10000-dimensional array)

2. Multiformity and multimodality: Data take many forms + modes

Numerical values, discrete structures, networks, variable-length sequences, etc.

Images, volumes, videos, audios, texts, point clouds, geometries, sensor signals, etc.

Modern aspects of ML

1. High dimensionality: Data can have many input variables.

a 100x100 pixel grayscale image = 10000 input variables (a 10000-dimensional array)

2. Multiformity and multimodality: Data take many forms + modes

Numerical values, discrete structures, networks, variable-length sequences, etc.

Images, volumes, videos, audios, texts, point clouds, geometries, sensor signals, etc.

3. Overrepresentation: ML models can have many parameters.

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

Modern aspects of ML

1. High dimensionality: Data can have many input variables.

a 100x100 pixel grayscale image = 10000 input variables (a 10000-dimensional array)

2. Multiformity and multimodality: Data take many forms + modes

Numerical values, discrete structures, networks, variable-length sequences, etc.

Images, volumes, videos, audios, texts, point clouds, geometries, sensor signals, etc.

3. Overrepresentation: ML models can have many parameters.

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

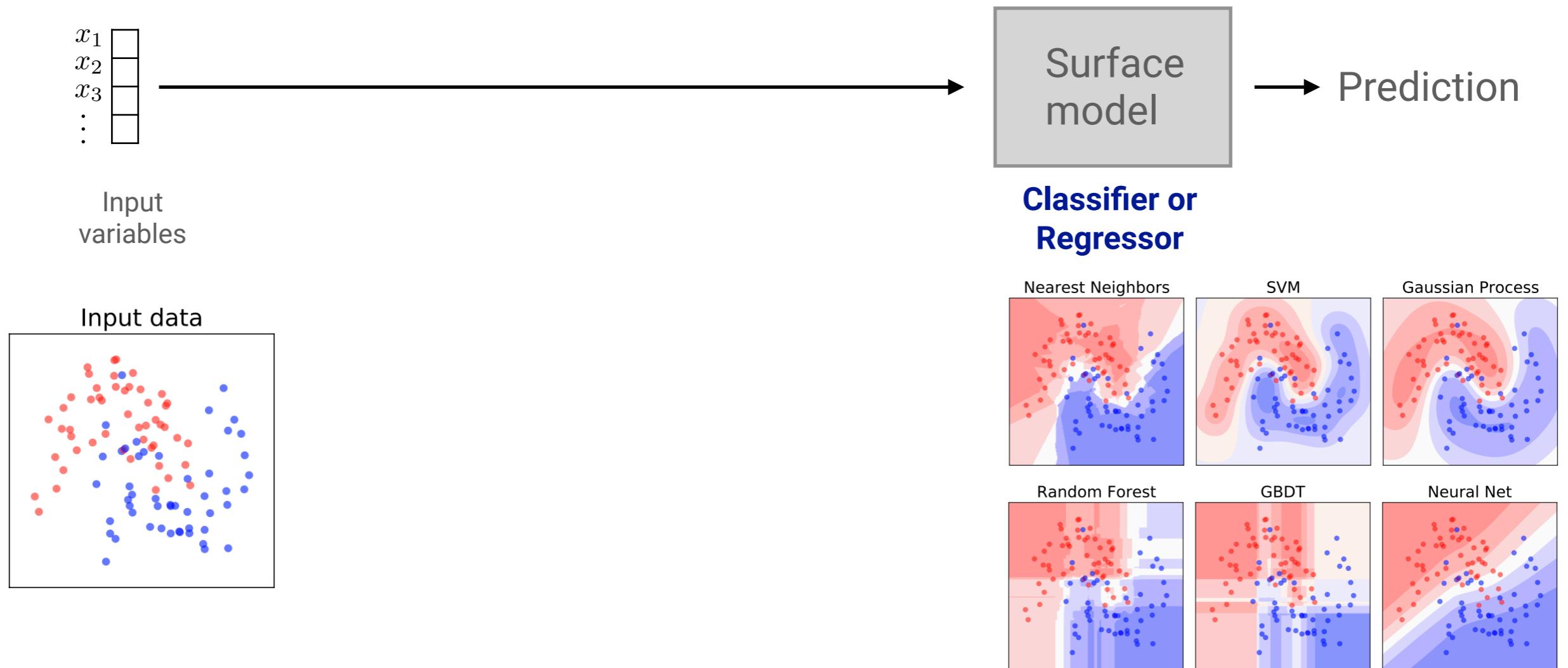
GPT-2 XL: **1558 million** params

GPT-3: **175 billion** params

Can you imagine what would happen if we try to fit a surface model having **175 billion** parameters to **100 million** data points in **10 thousand** dimension??

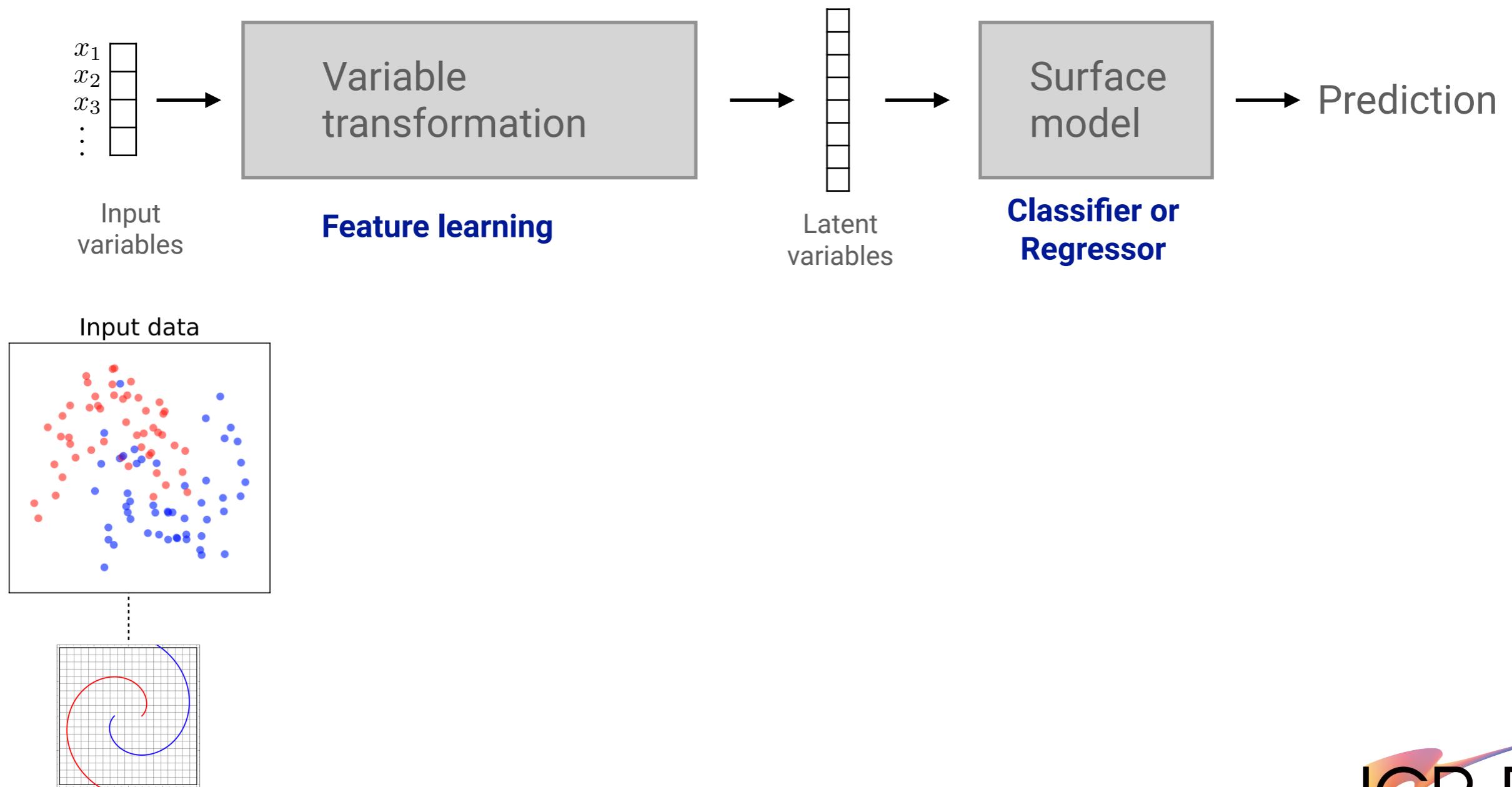
Modern aspects of ML

4. Representation learning: Models can have “feature learning” blocks, and they can be “pre-trained” by different large datasets.



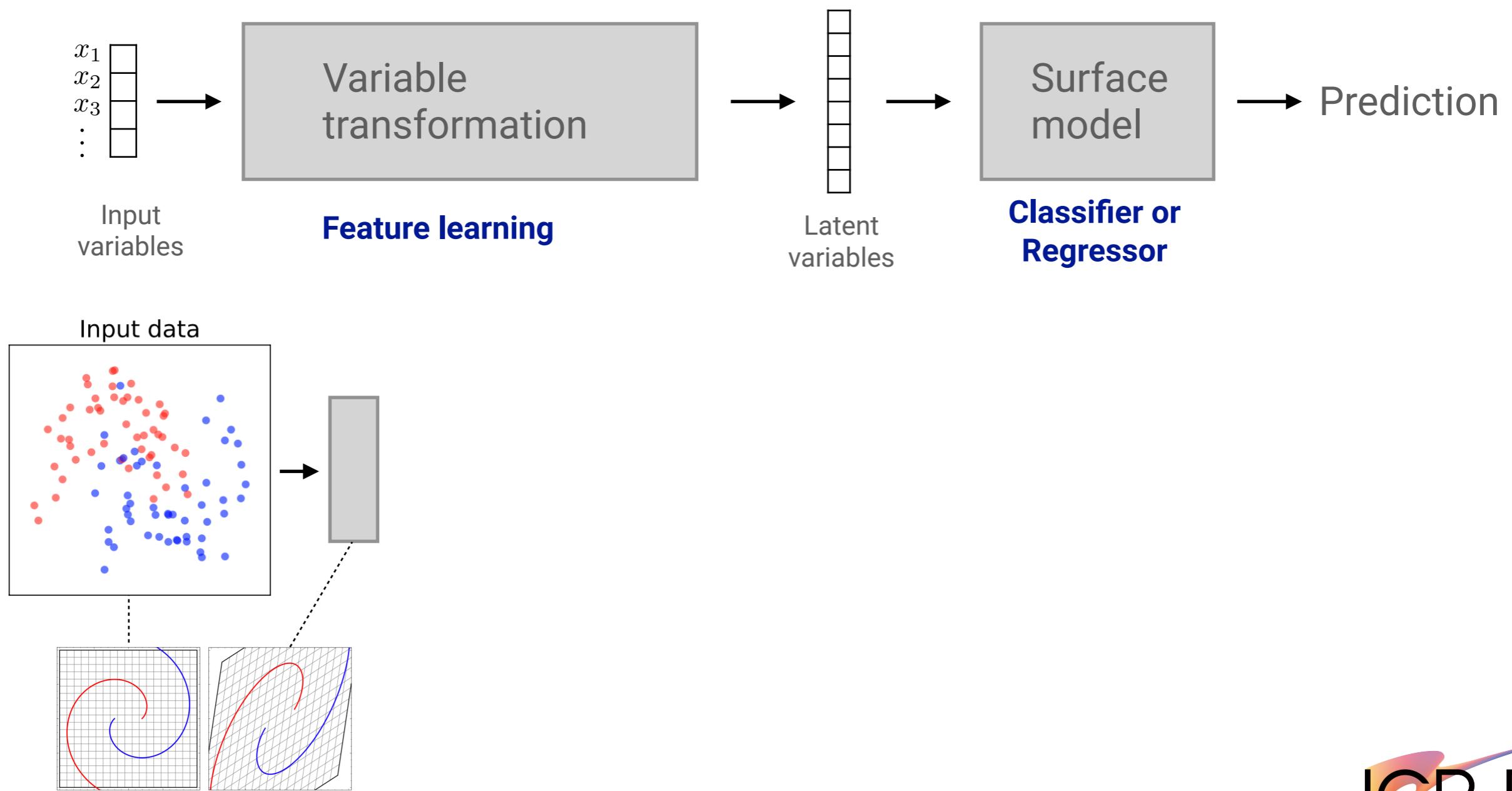
Modern aspects of ML

4. **Representation learning:** Models can have “feature learning” blocks, and they can be “pre-trained” by different large datasets.



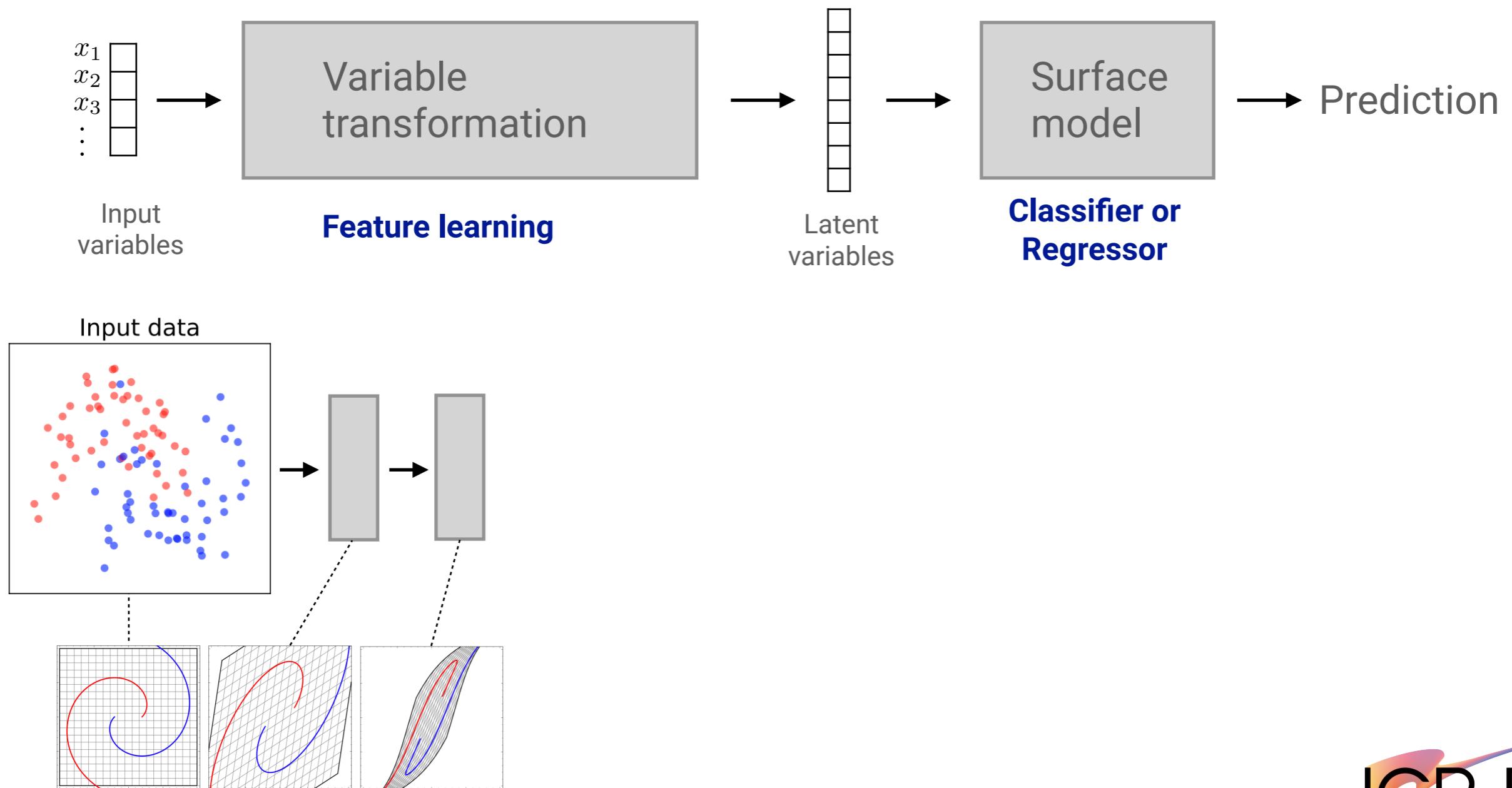
Modern aspects of ML

4. **Representation learning:** Models can have “feature learning” blocks, and they can be “pre-trained” by different large datasets.



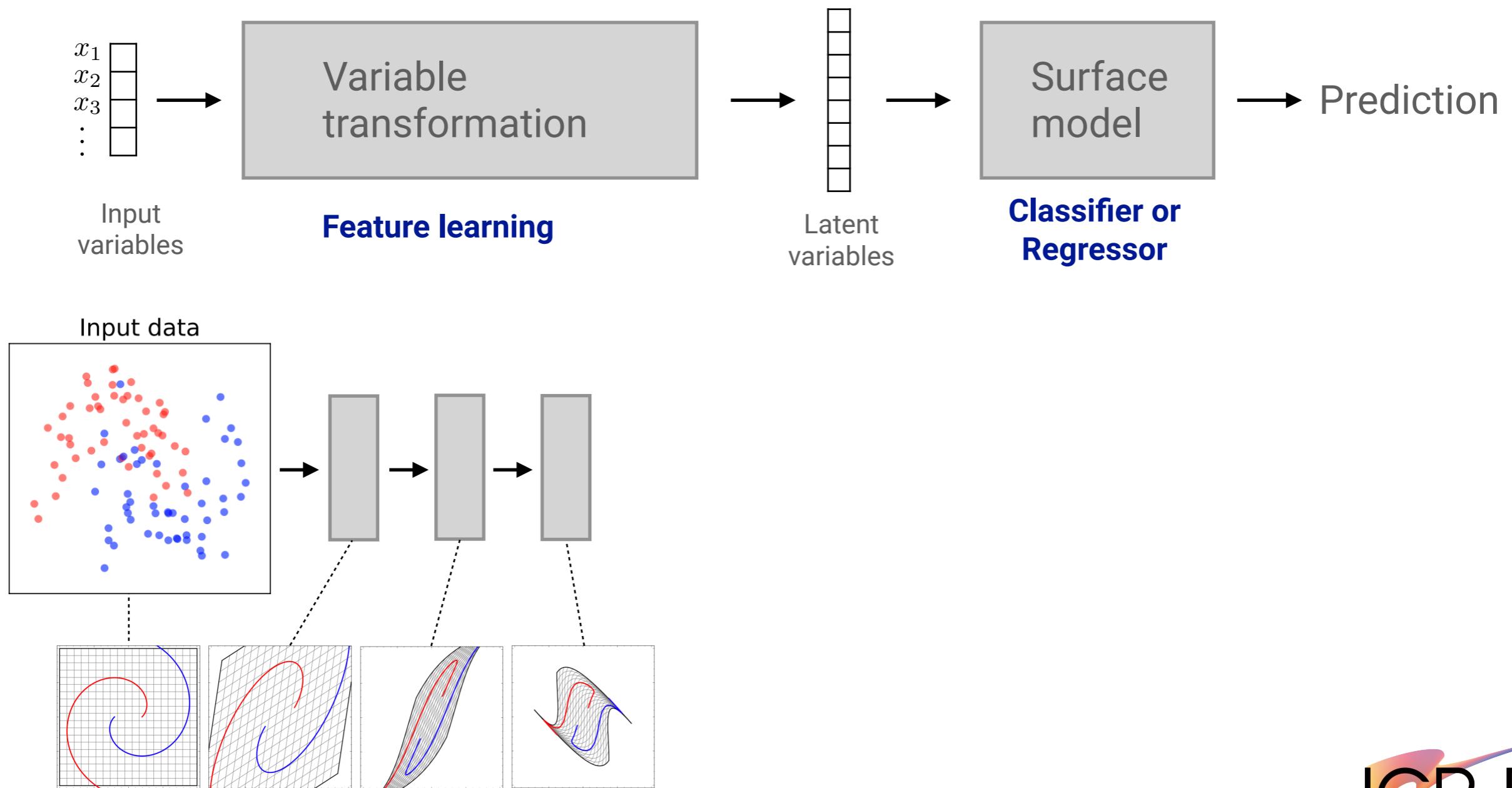
Modern aspects of ML

4. **Representation learning:** Models can have “feature learning” blocks, and they can be “pre-trained” by different large datasets.



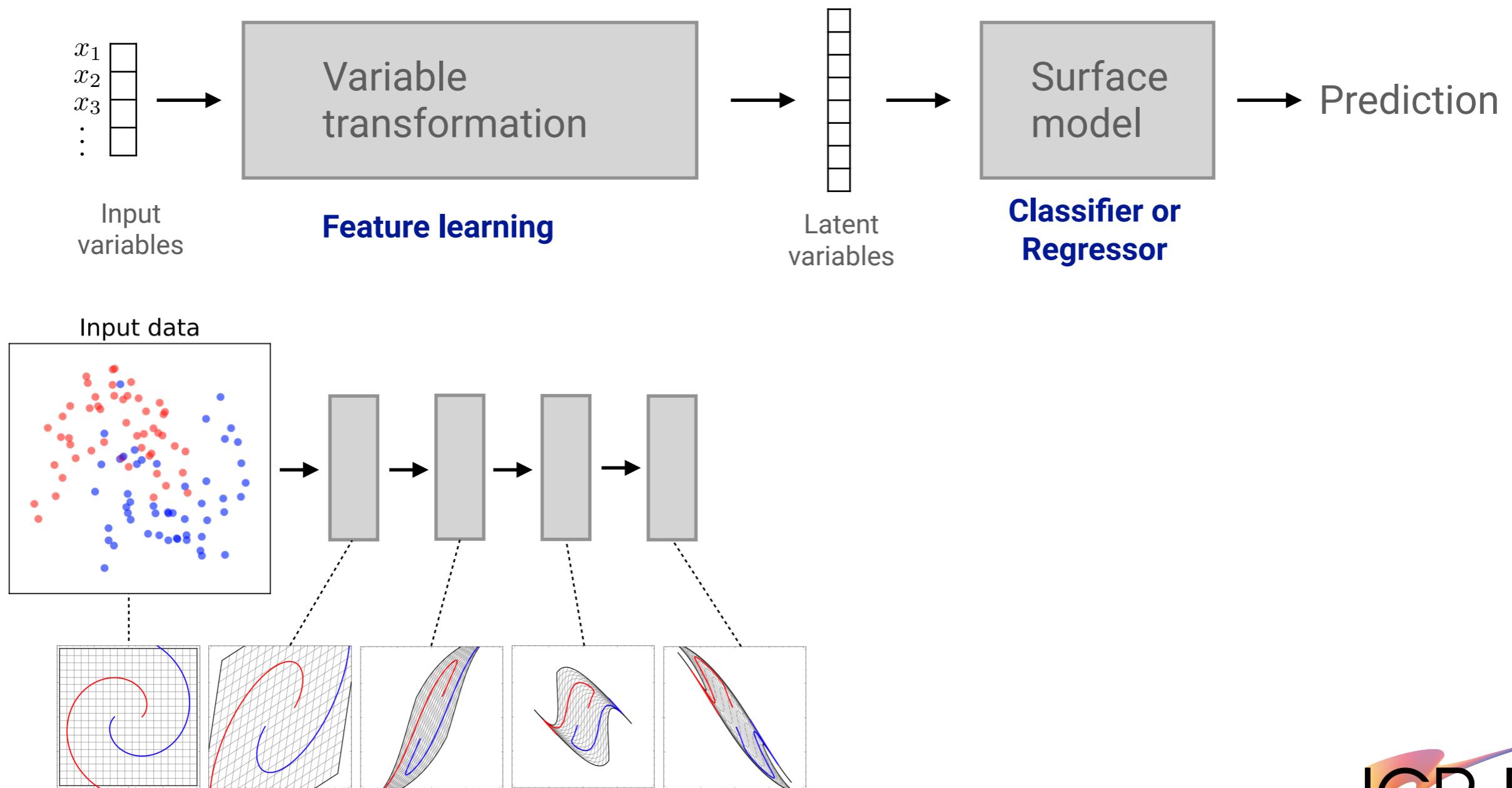
Modern aspects of ML

4. Representation learning: Models can have “feature learning” blocks, and they can be “pre-trained” by different large datasets.



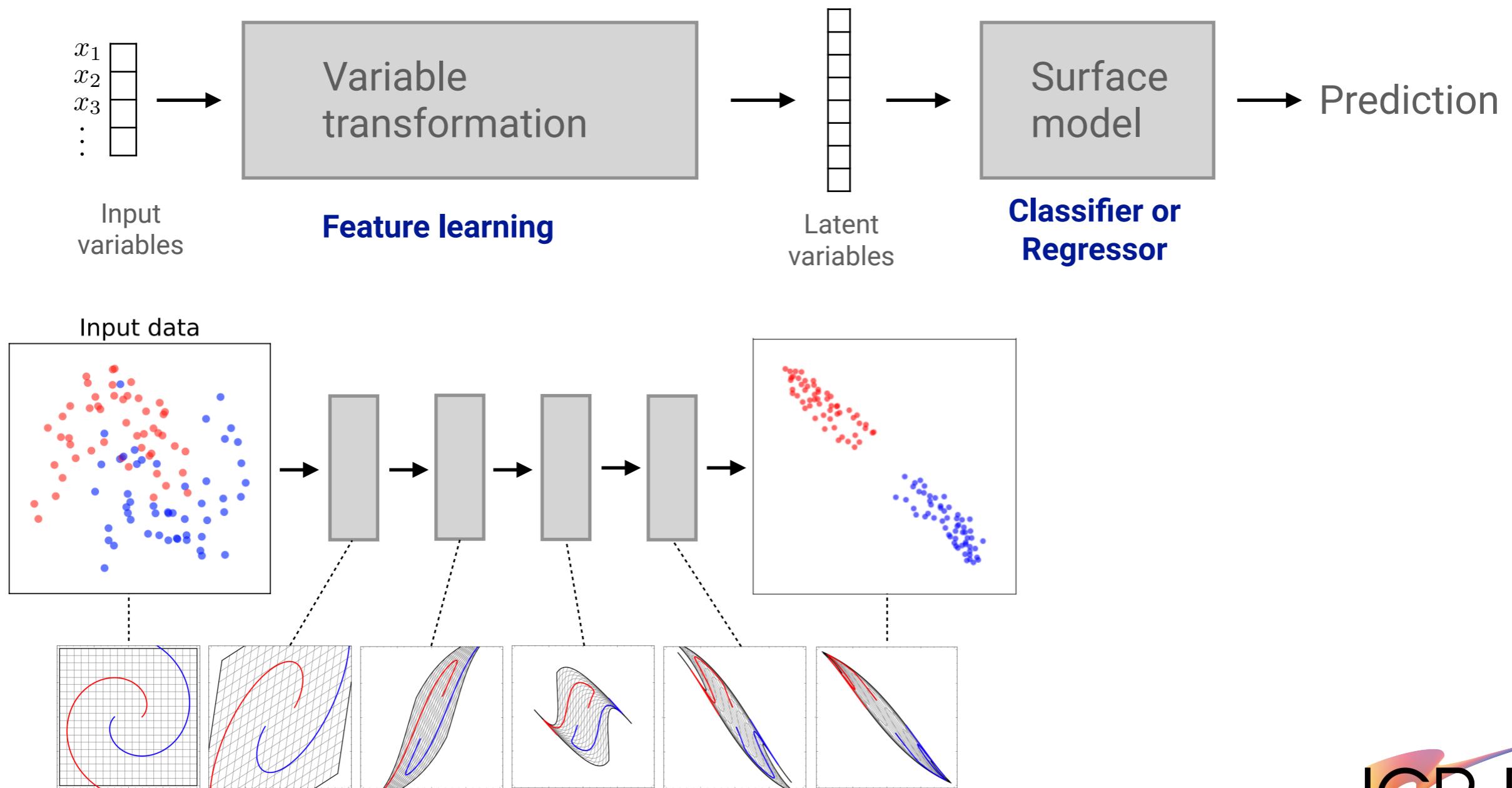
Modern aspects of ML

4. Representation learning: Models can have “feature learning” blocks, and they can be “pre-trained” by different large datasets.



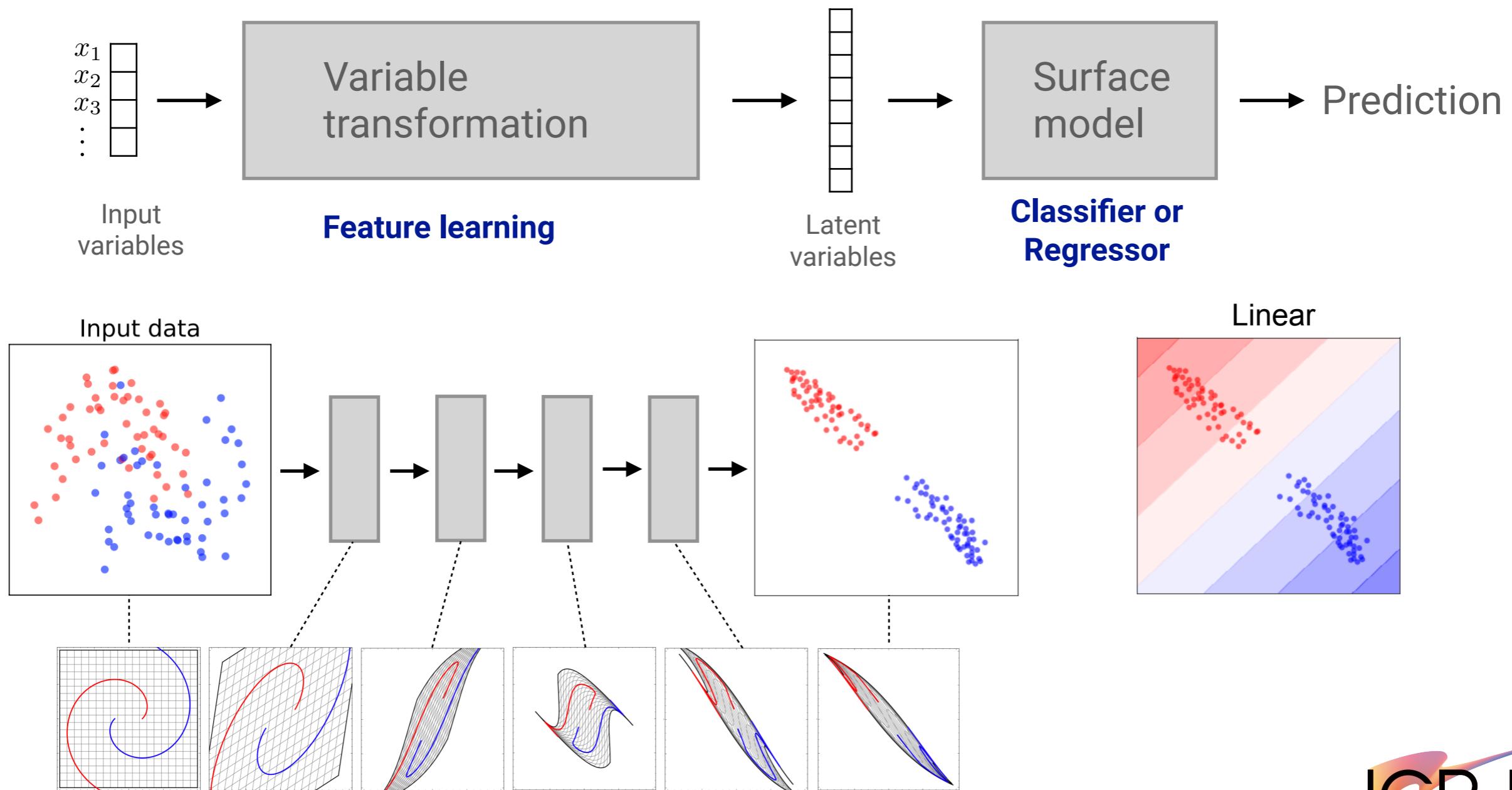
Modern aspects of ML

4. Representation learning: Models can have “feature learning” blocks, and they can be “pre-trained” by different large datasets.



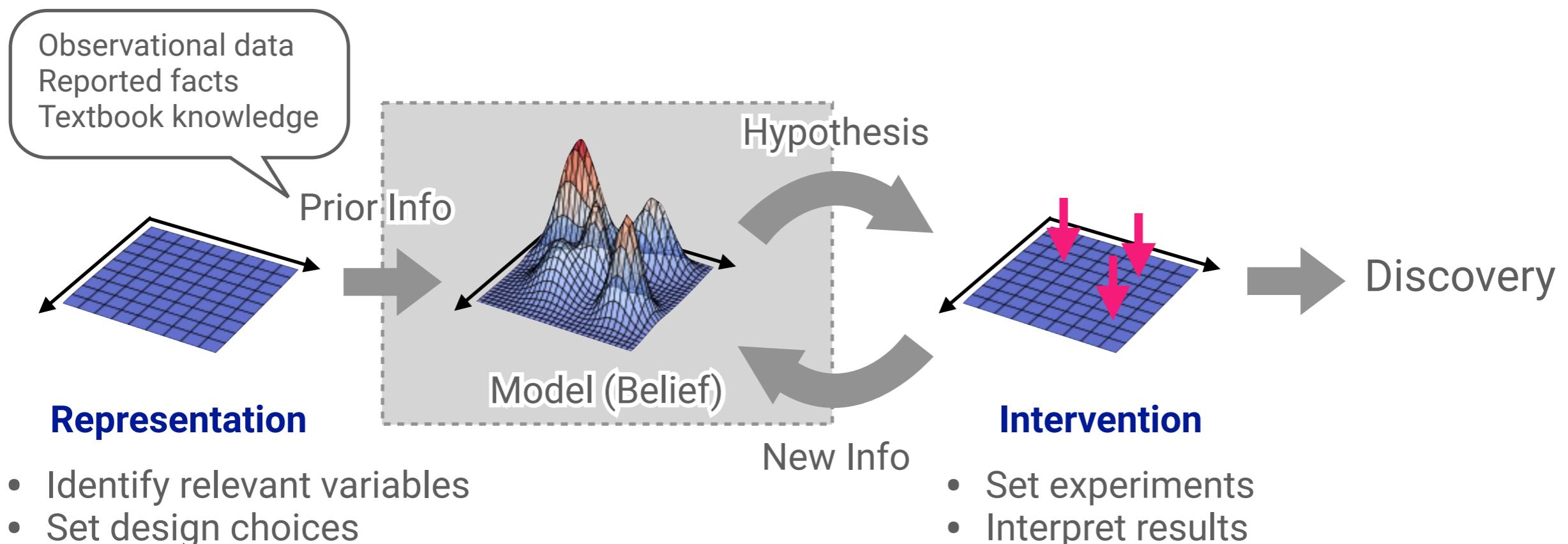
Modern aspects of ML

4. Representation learning: Models can have “feature learning” blocks, and they can be “pre-trained” by different large datasets.



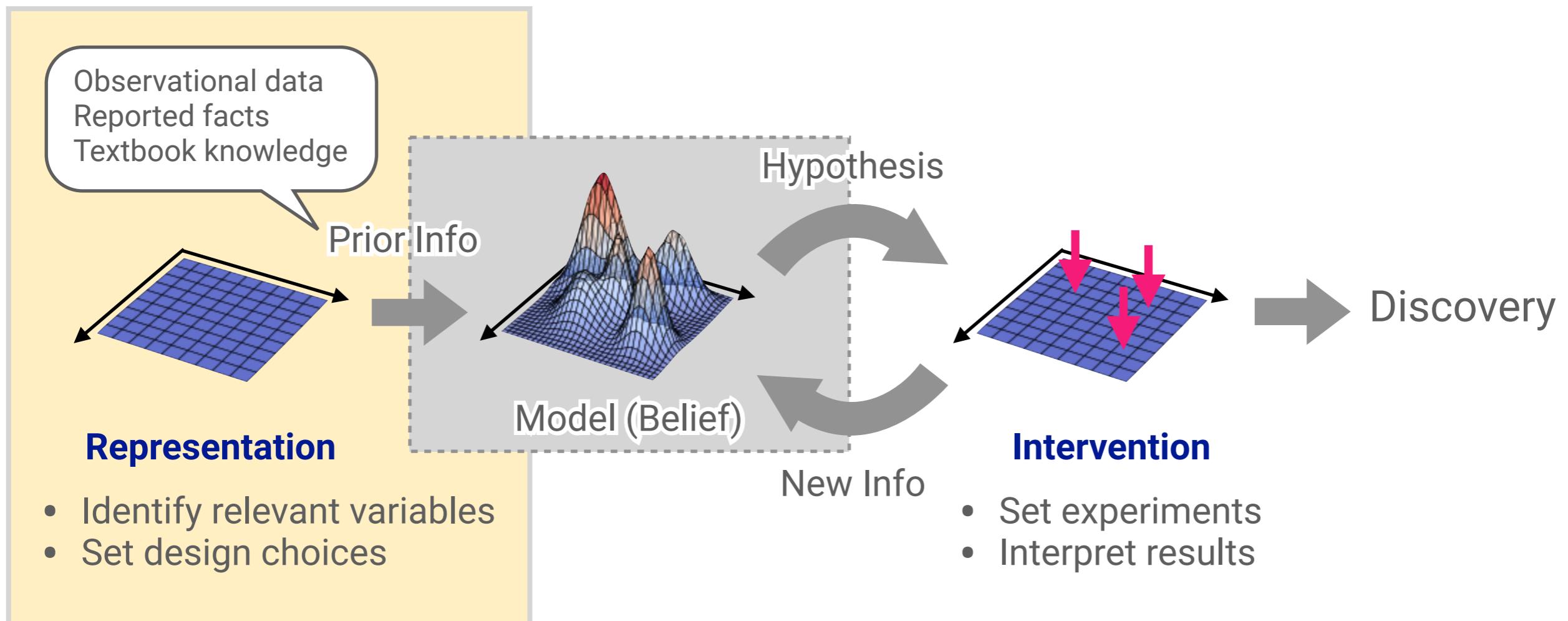
Needs and excitement around ML for Chemistry

Can we somehow externalize “experience and intuition” of experienced chemists to rationalize and accelerate discoveries?



Needs and excitement around ML for Chemistry

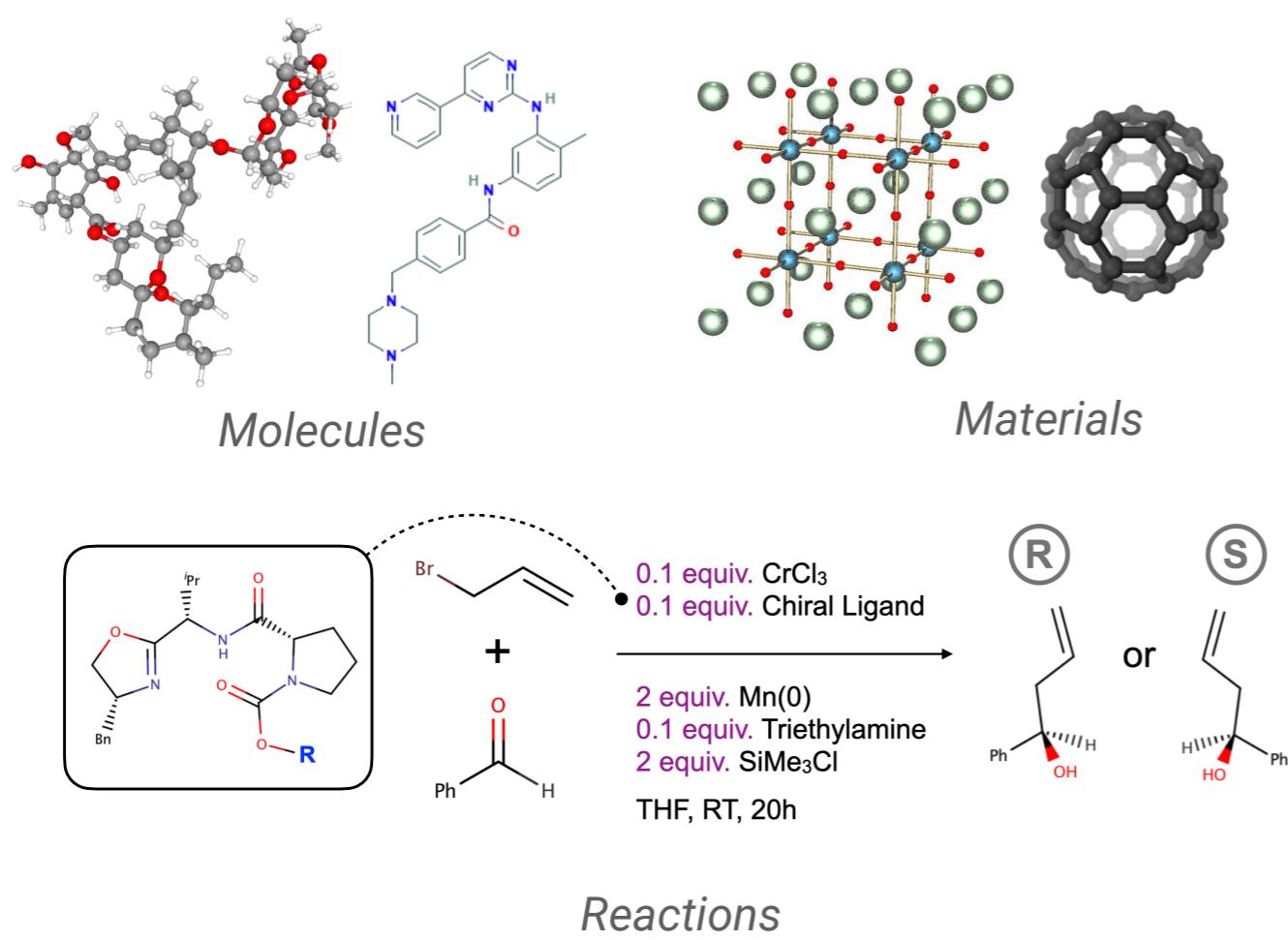
Can we somehow externalize “experience and intuition” of experienced chemists to rationalize and accelerate discoveries?



Representation

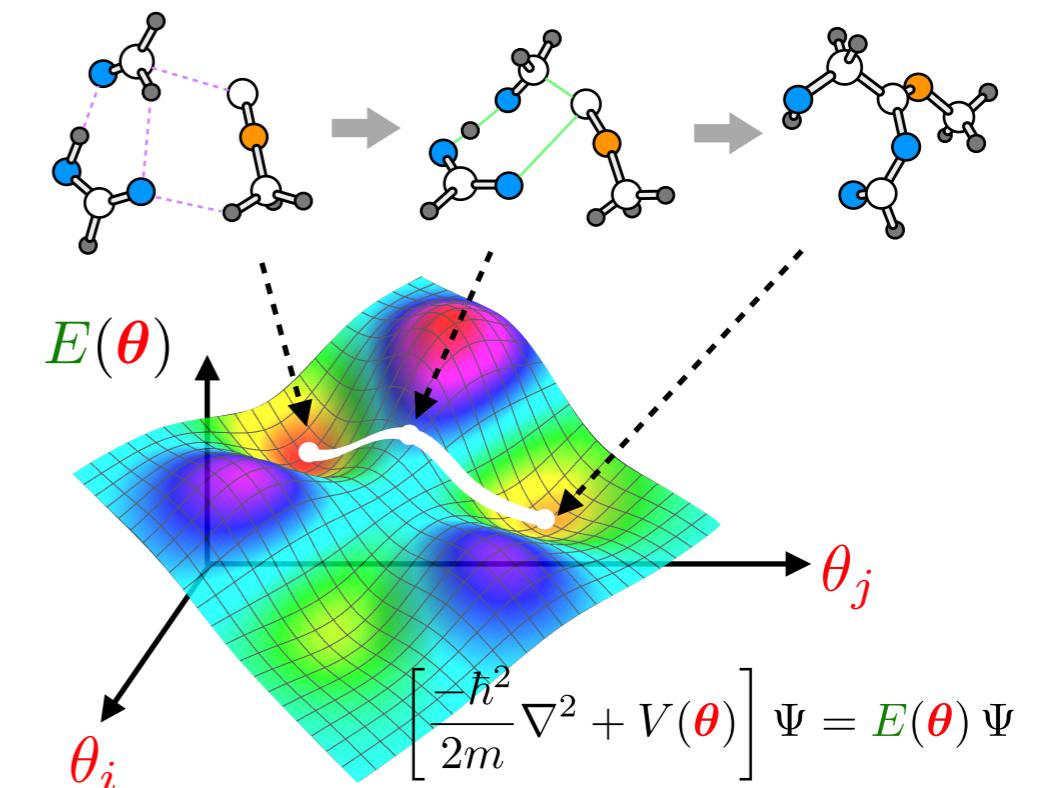
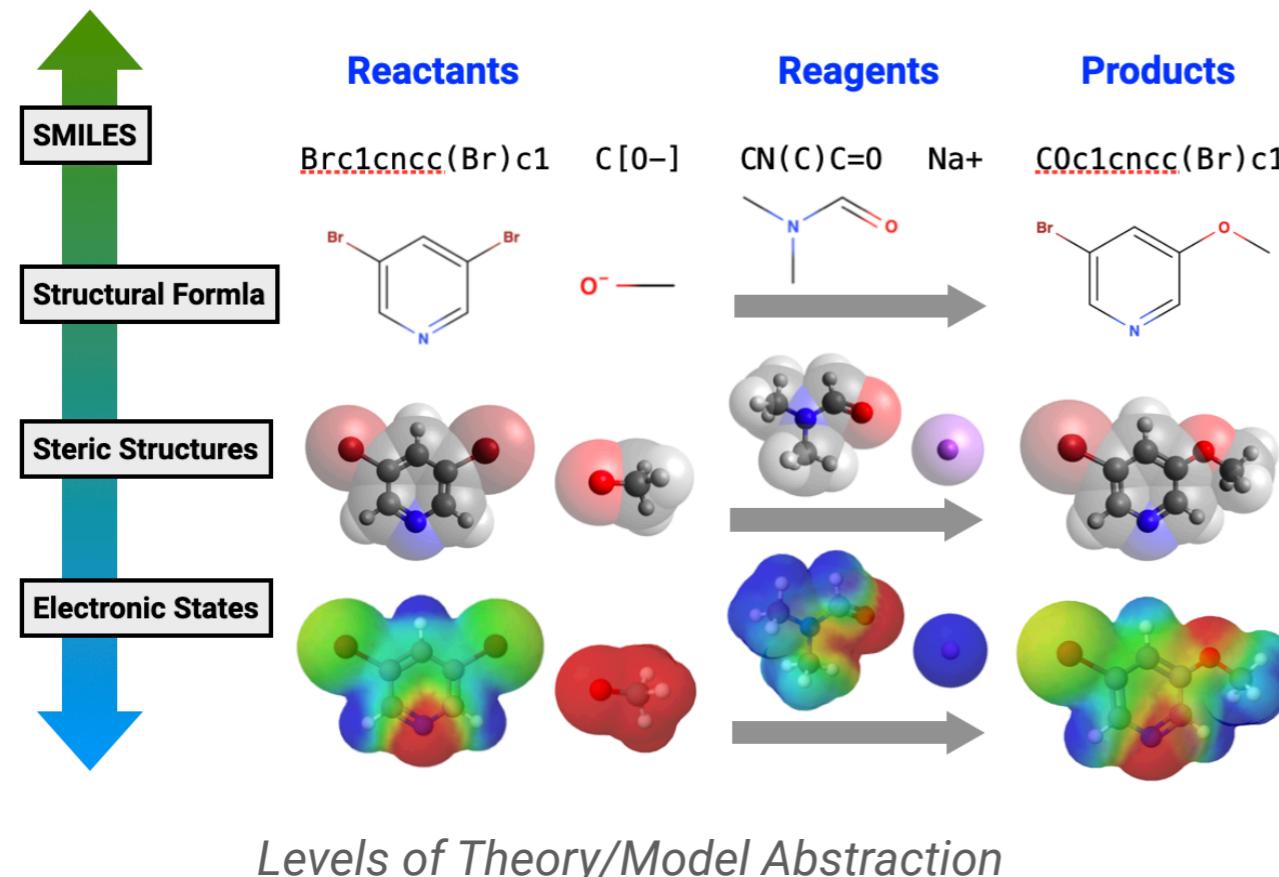
Any rationalized “real” discovery only comes from understanding and discovery of the causal relations between relevant factors.

Identifying relevant factors and establishing any necessary and sufficient computer-readable representations are inevitable preconditions, but this is far from trivial and quite paradoxical since we haven’t understood the target.

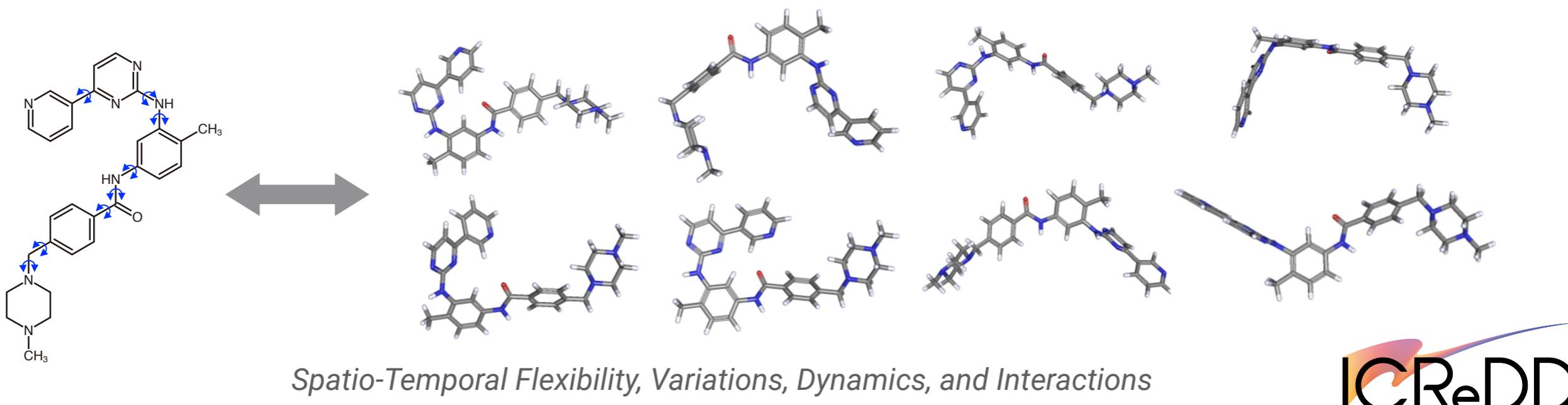


- Observational data
- Reported facts
- Textbook knowledge

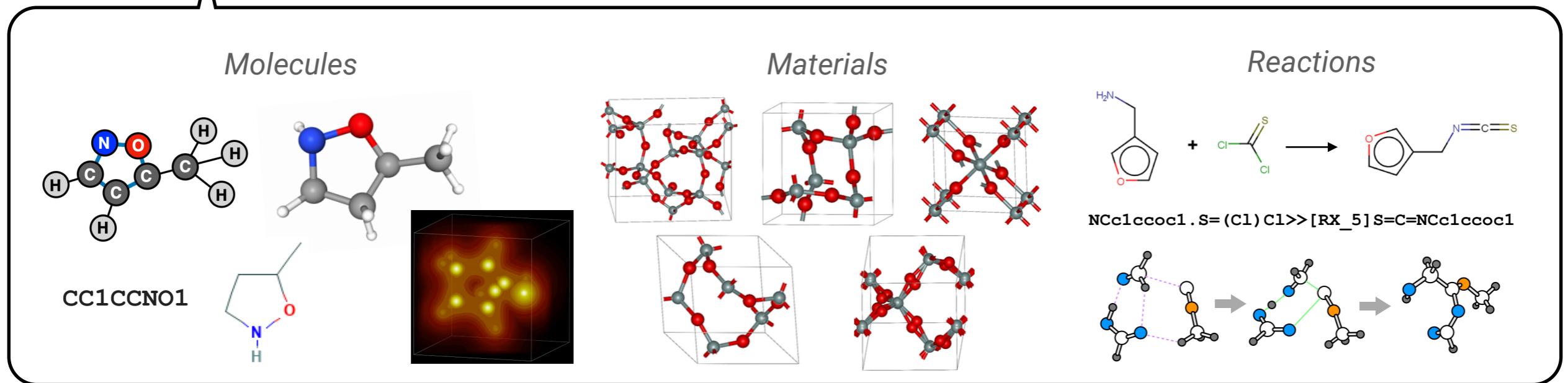
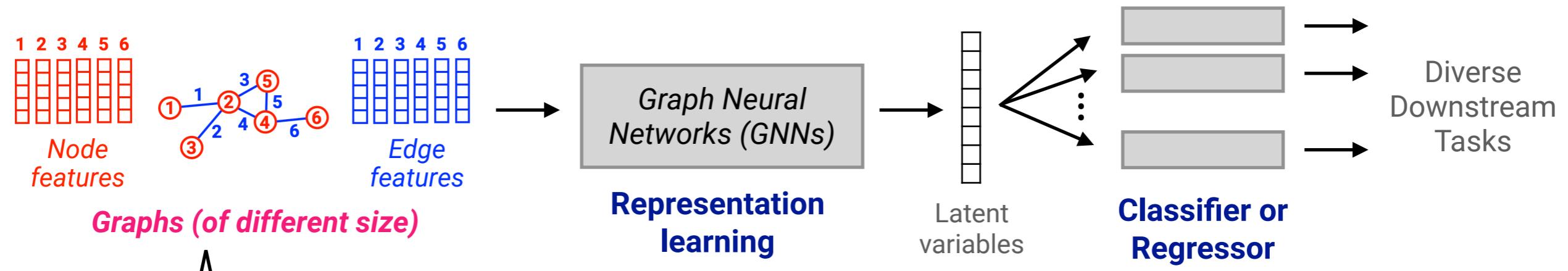
Representation



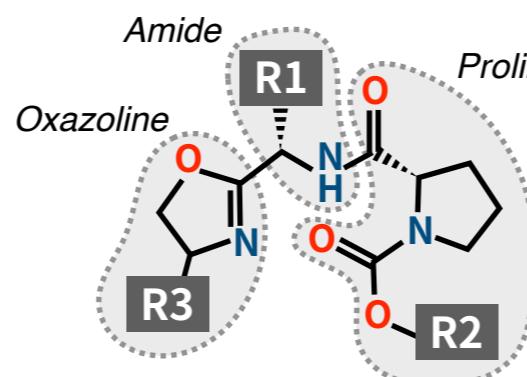
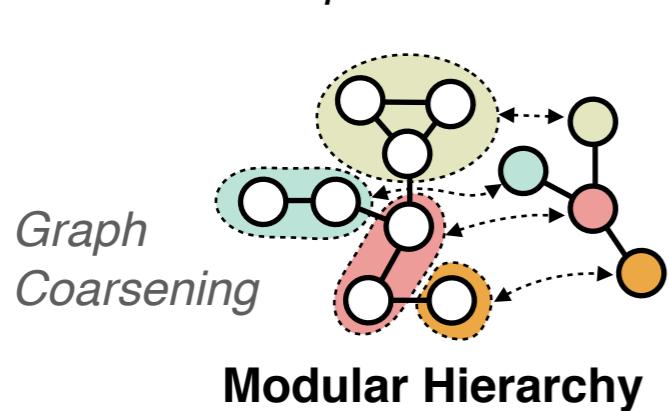
First Principle and Simulation (Quantum Chemistry)



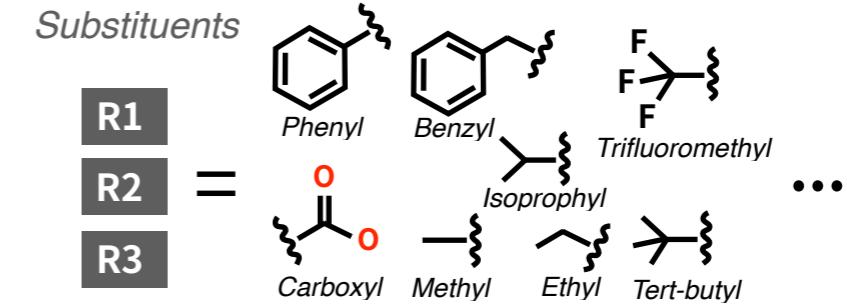
Representation



Combinatorial aspects



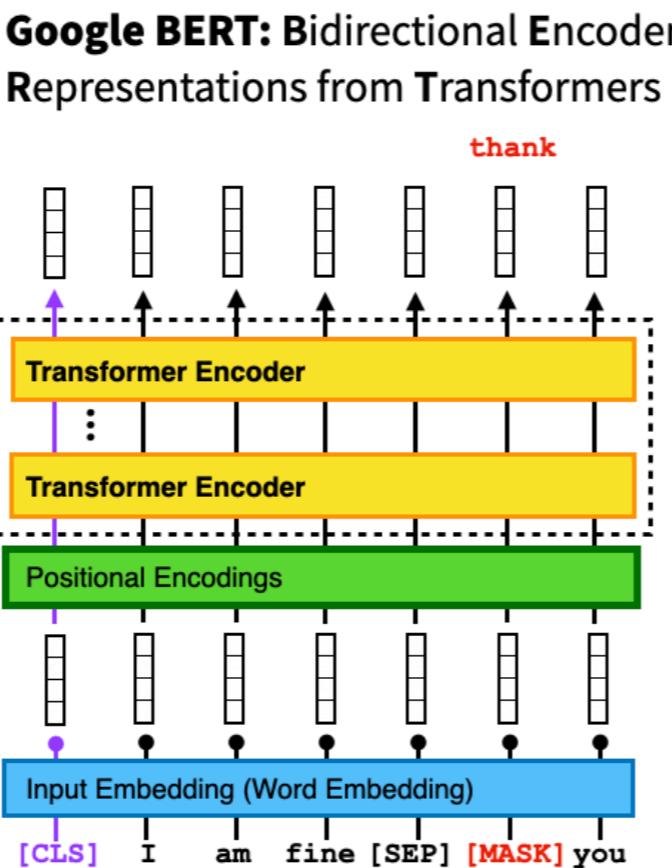
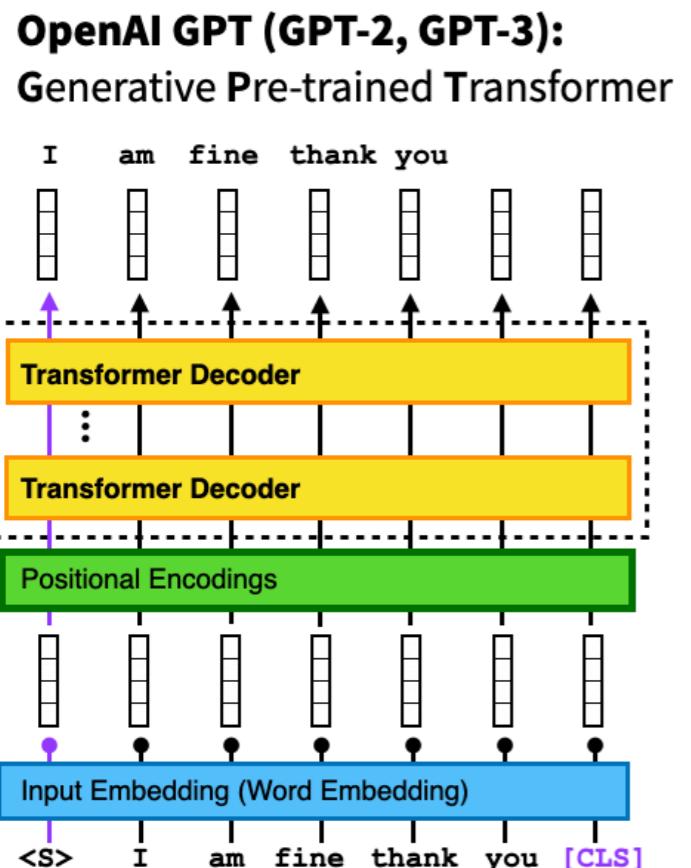
Compositionality



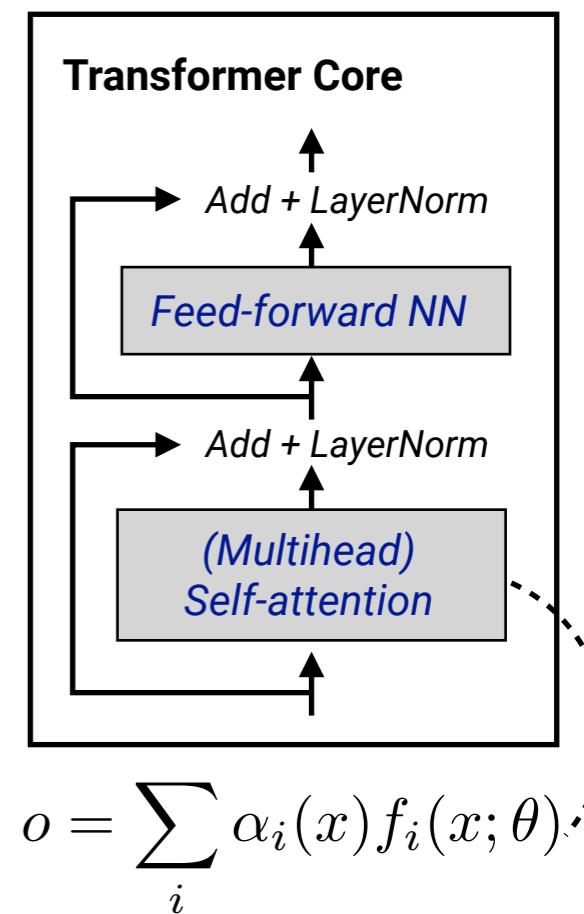
Representation

Effective pretraining is a crucial open problem because in practice, we can only access to limited data for each specific problem.

Pretraining with self-supervised pretext tasks have transformed NLP



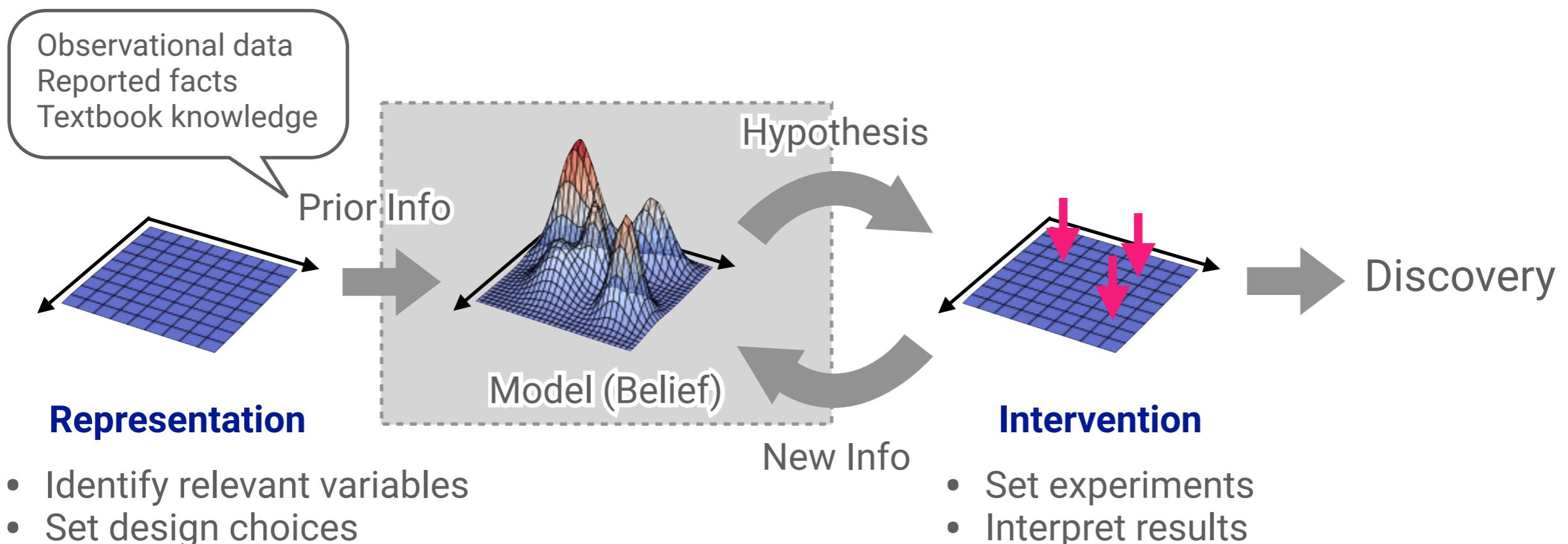
- Self-supervised pretext tasks**
- GPT-2, GPT-3**
 - Left-to-Right Language Models (LM)
 - BERT**
 - Masked Language Models (MLM)
 - Next Sentence Prediction (NSP)
 - ALBERT**
 - Sentence-Order Prediction (SOP)
 - ELECTRA**
 - Replaced Token Detection (RTD)



NB: Transformers can be considered as a special case of GNNs, and many Transformer-type GNNs are also developed.

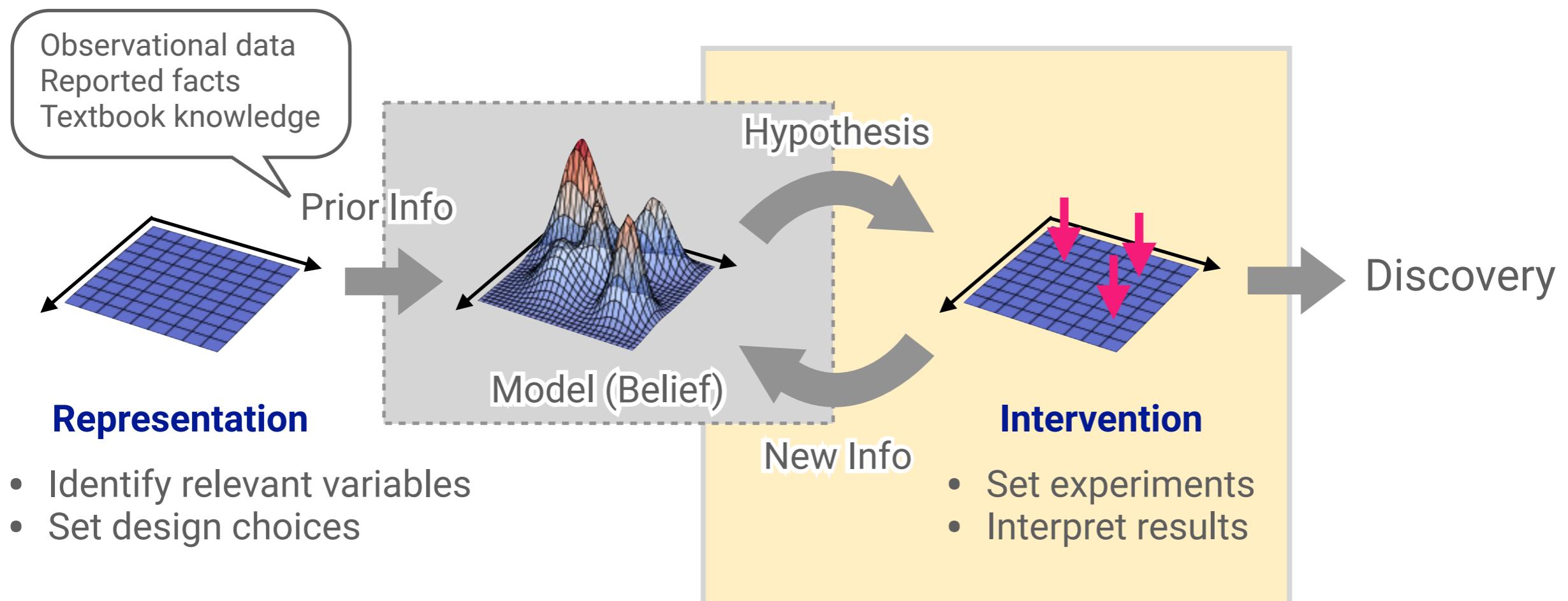
Needs and excitement around ML for Chemistry

Can we somehow externalize “experience and intuition” of experienced chemists to rationalize and accelerate discoveries?



Needs and excitement around ML for Chemistry

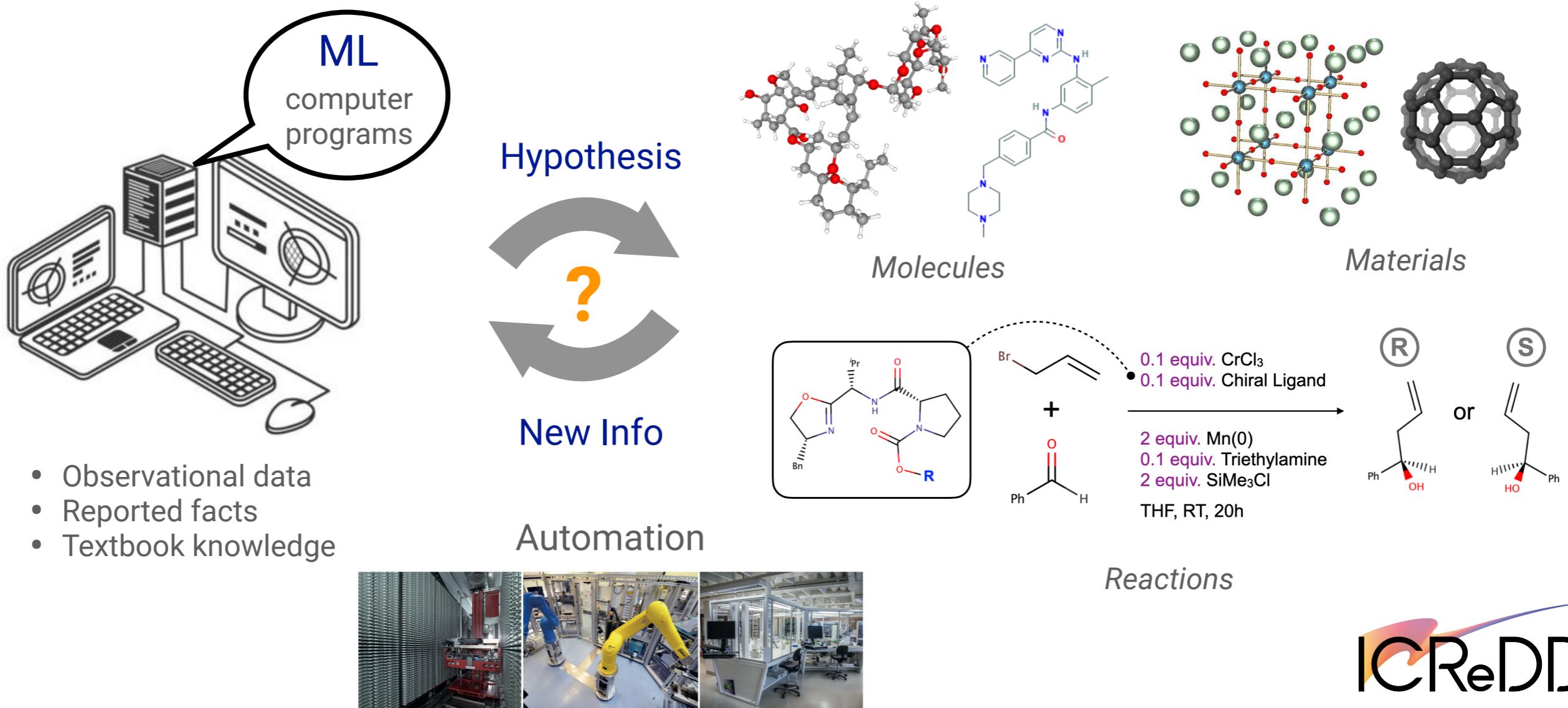
Can we somehow externalize “experience and intuition” of experienced chemists to rationalize and accelerate discoveries?



(Experimental) Intervention

Any rationalized “real” discovery only comes from understanding and discovery of the causal relations between relevant factors.

Information about causal relations can be acquired by passive observation and active intervention. Correlation does not imply causation.



(Experimental) Intervention

We need to carefully rethink how an experiment should be performed to be informative about causal structure of targets.

(Experimental) Intervention

We need to carefully rethink how an experiment should be performed to be informative about causal structure of targets.

- ***Correlation vs Causation***

ML models trained over passive observational data can be trapped by **spurious correlations between variables**, being totally ignorant of the underlying causality.

(Experimental) Intervention

We need to carefully rethink how an experiment should be performed to be informative about causal structure of targets.

- ***Correlation vs Causation***

ML models trained over passive observational data can be trapped by **spurious correlations between variables**, being totally ignorant of the underlying causality.

- ***Garbage In, Garbage Out (GIGO)***

ML models are **just representative** of the given data. If it has any bias, ML predictions can be miserably misleading.

(Experimental) Intervention

We need to carefully rethink how an experiment should be performed to be informative about causal structure of targets.

- ***Correlation vs Causation***

ML models trained over passive observational data can be trapped by **spurious correlations between variables**, being totally ignorant of the underlying causality.

- ***Garbage In, Garbage Out (GIGO)***

ML models are **just representative** of the given data. If it has any bias, ML predictions can be miserably misleading.

- ***Unavoidable Human-Caused Biases***

Always remember that “most chemical experiments are planned by human scientists and therefore are **subject to a variety of human cognitive biases, heuristics and social influences.**”

* Jia, X., Lynch, A., Huang, Y. et al. Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. *Nature* **573**, 251–255 (2019).

(Experimental) Intervention

REPORT

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

 Derek T. Ahneman¹, Jesús G. Estrada¹,  Shishi Lin²,  Spencer D. Dreher^{2,*},  Abigail G. Doyle^{1,*}

 See all authors and affiliations

Science 13 Apr 2018:
Vol. 360, Issue 6385, pp. 186-190
DOI: 10.1126/science.aar5169

Science

- Main paper <https://doi.org/10.1126/science.aar5169>
- Erratum <https://doi.org/10.1126/science.aat7648>
- Negative comment paper <https://doi.org/10.1126/science.aat8603>
- Author's response <https://doi.org/10.1126/science.aat8763>



<https://www.chemistryworld.com/news/dispute-over-reaction-prediction-puts-machine-learnings-pitfalls-in-spotlight/3009912.article>

(Experimental) Intervention

Keys: fusing modern ML with first-principles, simulations, domain knowledge, and collaboratively working with experimental experts.

Current ML is **too data-hungry and vulnerable to any data bias**, but acquisition of clean representative data is often quite impractical.

- Deep learning techniques thus far have proven to be **data hungry, shallow, brittle, and limited in their ability to generalize** (Marcus, 2018)
- Current machine learning techniques are **data-hungry and brittle**—they can only make sense of patterns they've seen before. (Chollet, 2020)
- A growing body of evidence shows that state-of-the-art models **learn to exploit spurious statistical patterns in datasets**... instead of learning meaning in the flexible and generalizable way that humans do. (Nie et al., 2019)
- Current machine learning methods seem **weak when they are required to generalize beyond the training distribution**, which is what is often needed in practice. (Bengio et al., 2019)

(Experimental) Intervention

This has reignited the old war between induction and deduction, and we're re-encountering the long-standing problems in AI.

- **Knowledge acquisition / Principled data acquisition**
Experimental design, Model-based optimization, Evolutionary computation
- **Reconciliation between inductive and deductive ML**
Hybrid models of causal/logical/algorithmic ML and deep learning
- **Balancing exploitation and exploration**
Model-based reinforcement learning or search in a combinatorial space

AlphaGo
(Nature, 2016)



AlphaGo Zero
(Nature, 2017)

ARTICLE

Mastering the game of Go without human knowledge

David Silver^{1*}, Julian Schrittwieser^{1*}, Karen Simonyan¹, Ioannis Antonoglou¹, Aja Huang¹, Arthur Guez¹, Thomas Hubert¹, Lucas Baker¹, Matthew Lai¹, Adrian Bolton¹, Yutian Chen¹, Timothy Lillicrap¹, Fan Hui¹, Laurent Sifre¹, George van den Driessche¹, Thore Graepel¹ & Demis Hassabis¹

A long-standing goal of artificial intelligence is an algorithm that learns, *tabula rasa*, superhuman proficiency in challenging domains. Recently, AlphaGo became the first program to defeat a world champion in the game of Go. The tree search in AlphaGo evaluated positions and selected moves using deep neural networks. These neural networks were trained by supervised learning from human expert moves, and by reinforcement learning from self-play. Here we introduce an algorithm that solely relies on reinforcement learning, without human data, gathering knowledge beyond game rules. AlphaZero becomes its own teacher: a neural network is trained to predict AlphaGo's own moves. AlphaZero also becomes the winner of AlphaGo's games. This neural network improves the strength of the tree search, resulting in higher quality move selection and stronger self-play in the next iteration. Starting *tabula rasa*, our new program AlphaGo Zero achieved superhuman performance, winning 100–0 against the previously published champion-defeating AlphaGo.

Much progress towards artificial intelligence has been made using supervised learning systems that are trained to replace the decisions of human experts^{1–4}. However, expert data sets are often expensive, unreliable or simply unavailable. Even when reliable data sets are available, they may impose a ceiling on the performance of systems trained in this way. By contrast, AlphaZero uses a simple tree search that relies upon no such external knowledge. Instead, it uses a single neural network that is trained from its own experience, in principle allowing them to exceed human capabilities, and to operate in domains where humans expertise is sparse. AlphaZero uses a combination of planning, reinforcement learning, and precise and stable learning. Further technical differences in the search algorithm, training procedure and network architecture are described in Methods.

AlphaZero
(Science, 2018)

Silver *et al.*, *Science* **362**, 1140–1144 (2018) 7 December 2018

COMPUTER SCIENCE

A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play

David Silver^{1,2,*†}, Thomas Hubert^{1*}, Julian Schrittwieser^{1*}, Ioannis Antonoglou¹, Matthew Lai¹, Arthur Guez¹, Marc Lanctot¹, Laurent Sifre¹, Dharshan Kumaran¹, Thore Graepel¹, Timothy Lillicrap¹, Karen Simonyan¹, Demis Hassabis¹

The game of chess is the longest-studied domain in the history of artificial intelligence. The strongest programs are based on a combination of sophisticated search techniques, domain-specific adaptations, and handcrafted evaluation functions that have been refined by human experts over several decades. By contrast, the AlphaGo Zero program recently achieved superhuman performance in the game of Go by reinforcement learning from self-play. In this paper, we generalize this approach into a single AlphaZero algorithm that can achieve superhuman performance in many challenging games. Starting from random play and given no domain knowledge except the game rules, AlphaZero convincingly defeated a world champion program in the games of chess and shogi (Japanese chess), as well as Go.

MuZero
(Nature, 2020)

Article

Mastering Atari, Go, chess and shogi by planning with a learned model

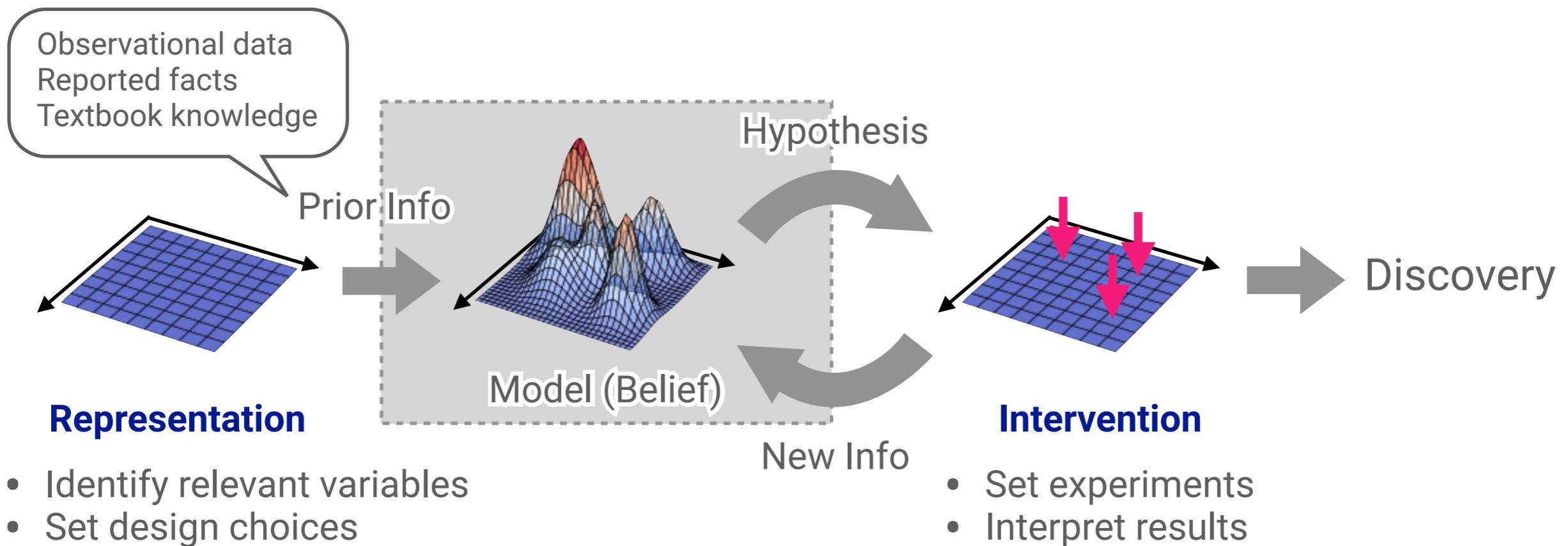
<https://doi.org/10.1038/s41586-020-03051-4>
Received: 3 April 2020
Accepted: 7 October 2020
Published online: 23 December 2020
Check for updates

Julian Schrittwieser^{1,2}, Ioannis Antonoglou^{1,2}, Thomas Hubert^{1,2}, Karen Simonyan¹, Laurent Sifre¹, Simon Schmitt¹, Arthur Guez¹, Edward Lockhart¹, Demis Hassabis¹, Thore Graepel^{1,2}, Timothy Lillicrap¹ & David Silver^{1,2,3,4,5}

Constructing agents with planning capabilities has long been one of the main challenges in the pursuit of artificial intelligence. Tree-based planning methods have enjoyed huge success in challenging domains, such as chess¹ and Go², where a perfect simulator is available. However, in real-world problems, the dynamics governing the environment are often complex and unknown. Here we present the MuZero algorithm, which, by combining a tree-based search with a learned model, achieves superhuman performance in a range of challenging and visually complex domains, without any knowledge of their underlying dynamics. The MuZero algorithm learns an iterable model that produces predictions relevant to planning: the action-selection policy, the value function and the reward. When evaluated on 57 different Atari games³—the canonical video game environment for testing artificial intelligence techniques, in which model-based planning approaches have historically struggled^{4–6}—the MuZero algorithm achieved state-of-the-art performance. When evaluated on Go, chess and shogi—canonical environments for high-performance planning—the MuZero algorithm matched, without any knowledge of the game dynamics, the superhuman performance of the AlphaZero algorithm⁷ that was supplied with the rules of the game.

ML for Chemistry to me (a ML researcher)

An exciting “real” test bench for the long-standing unsolved but attractive fundamental problems in “AI for automating discovery”, involving many fascinating technical topics of modern ML.



Summary

Machine Learning (ML) for Chemistry

- Why it is needed?
- What are exciting for computer scientists?

Two aspects:

1. Representation

- What are good ML-readable representations for chemistry?
- What information should be recorded and given to ML?

2. (Experimental) Intervention

- What are essential to make real chemical discoveries?
- Any principled ways for data acquisition and experimental design?