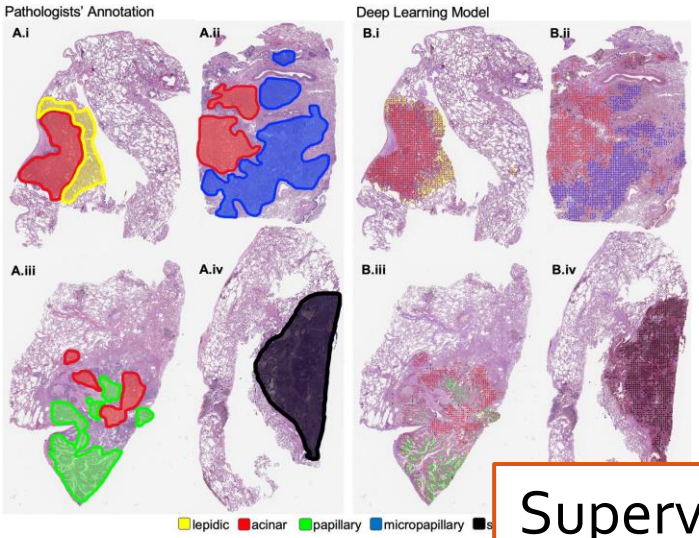
# REPRESENTATIONS: HOW TO PERFORM SUPERVISED LEARNING ON MATERIALS DATA

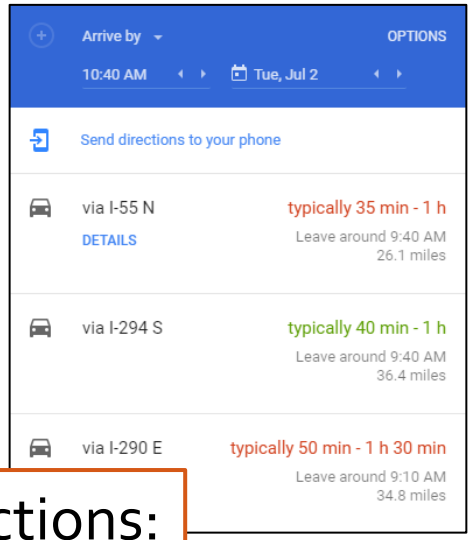
Logan Ward
Asst. Computational Scientist
Argonne National Laboratory

15 January 2022

# Supervised Learning: The ML you've definitely used

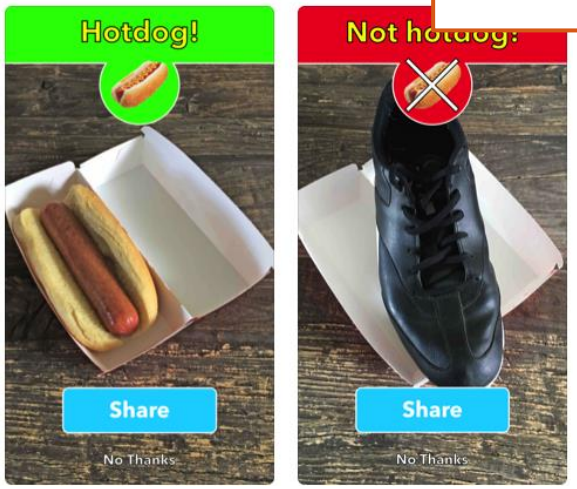



Wei *et al.* Sci Rep. (2019), 3358

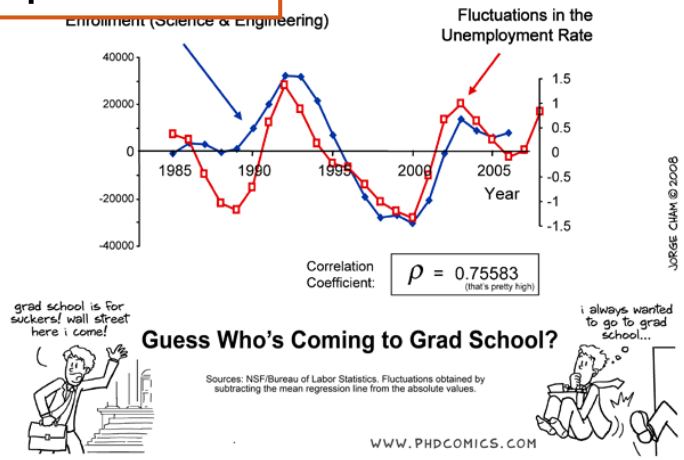



oogle Maps

Supervised Learning models functions:
"Given inputs, predict output"



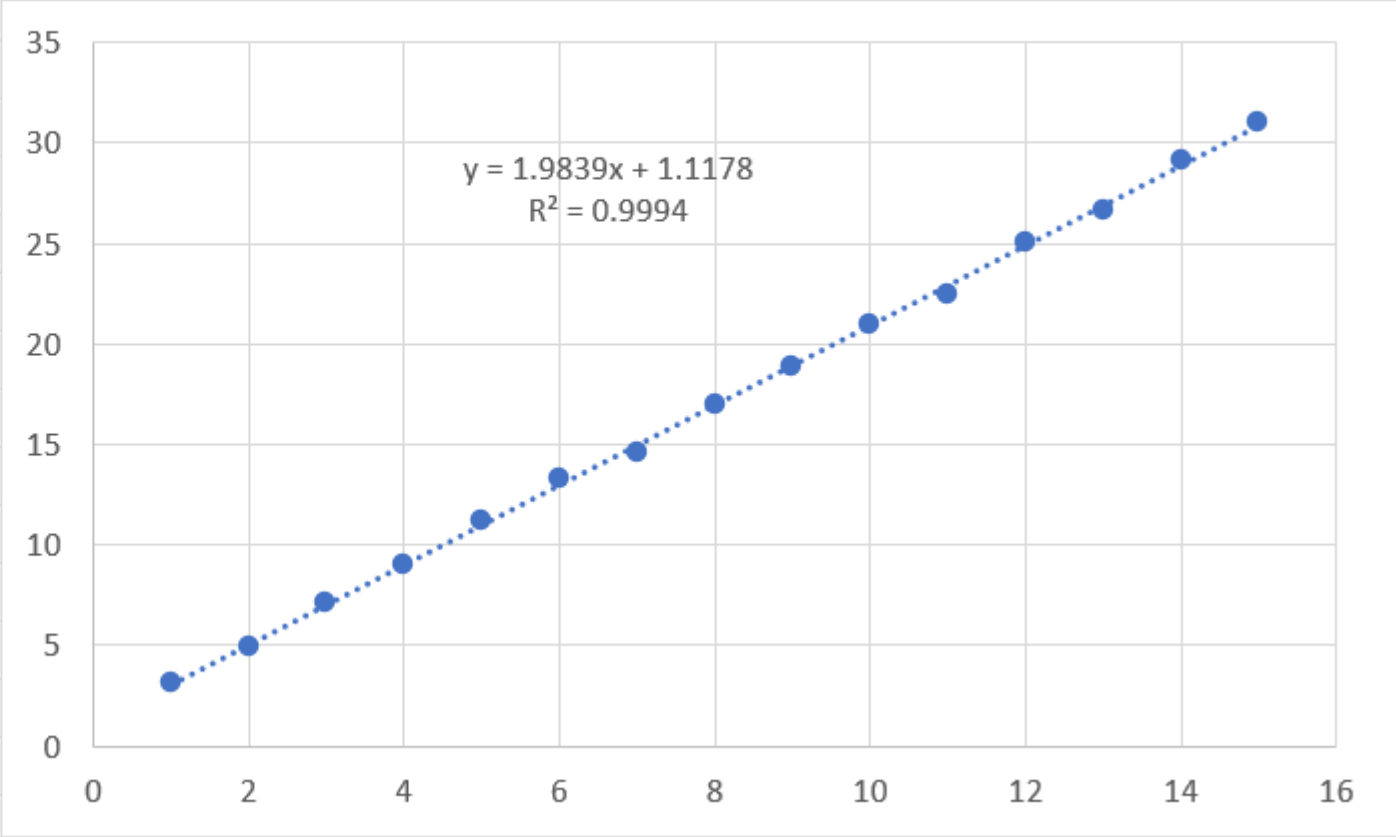

https://apps.apple.com/us/app/not-hotdog/id1212457521




http://phdcomics.com/comics/archive.php?comicid=1078

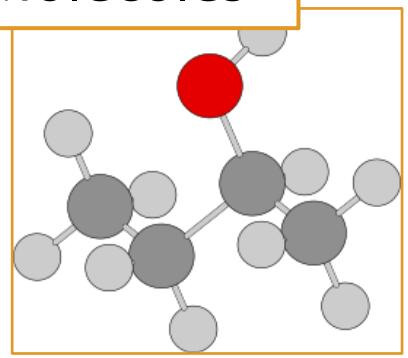
# You've already done machine learning

| | A | B |
|---|---|---|
| 1 | 1 | 3.18079 |
| 2 | 2 | 4.94926 |
| 3 | 3 | 7.1364 |
| 4 | 4 | 9.05278 |
| 5 | 5 | 11.2544 |
| 6 | 6 | 13.2715 |
| 7 | 7 | 14.5924 |
| 8 | 8 | 17.036 |
| 9 | 9 | 18.9159 |
| 10 | 10 | 21.0061 |
| 11 | 11 | 22.5182 |
| 12 | 12 | 25.046 |
| 13 | 13 | 26.6946 |
| 14 | 14 | 29.1493 |
| 15 | 15 | 31.0294 |
| 16 | | |



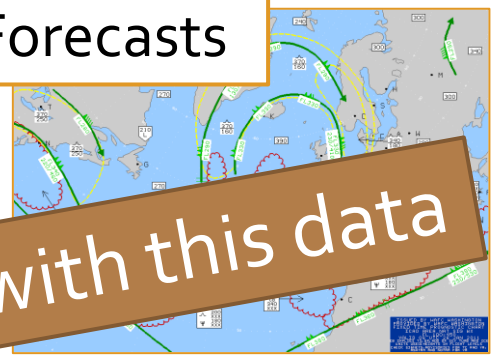
# Problem: Not all data are tensors

Molecules



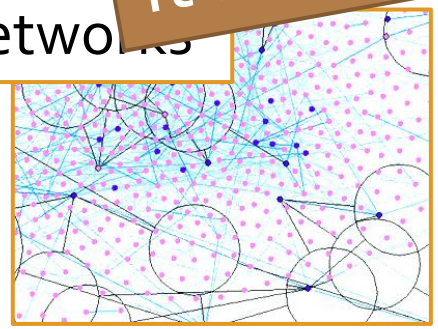
Images

Weather Forecasts

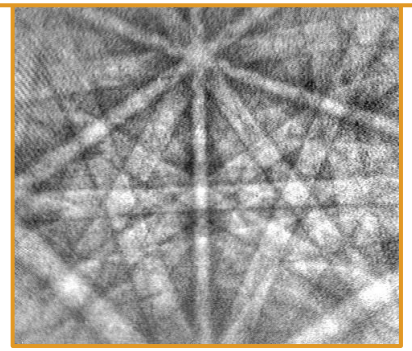


It's still possible to use machine learning with this data

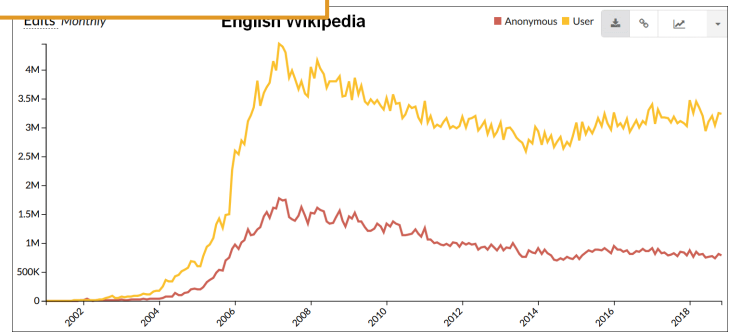
Social Networks



EBSD Patterns



Timeseries



# What do NaCl crystals look like?

High Electronegativity Difference

Large Band Gap

Minimal Interaction with Visible Light

**Clear**

**NaCl**

Representations make chemical rules accessible to learning algorithms

Similar Atomic Radii

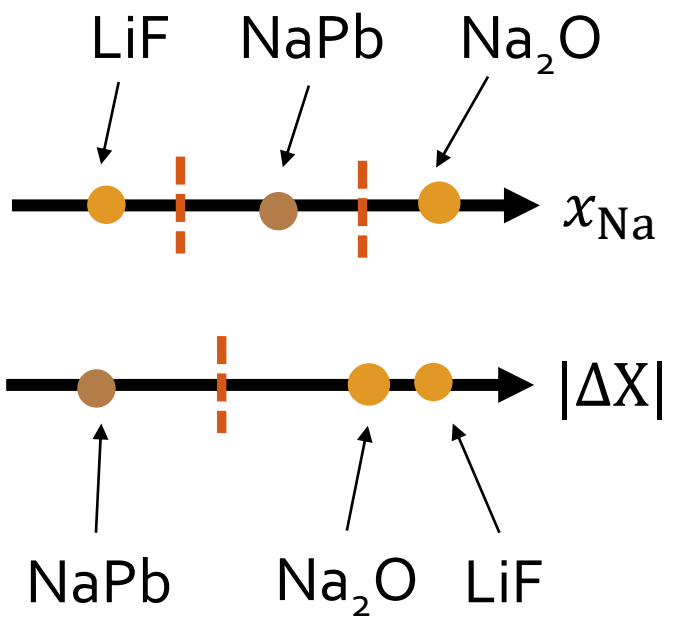
Coordination Number of 6

Cubic Symmetry

**Cubic Crystals**

# How to translate chemistry to a computer?

**Representation:** *Set of quantitative attributes that describe a material, molecule, …*




**Why do we need them?**

- Machine learning tools take <u>tensors</u>

**What makes a good one?**

- Easy to learn generalizable rules

# Rules for representations: Faber et al. (2015)

**Complete:** Different materials should have different representations

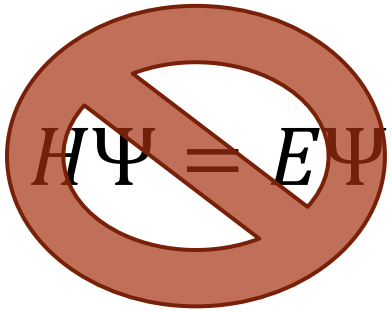
A != B

$x_A \neq x_B$

**Compact:** Representations should not include redundant features

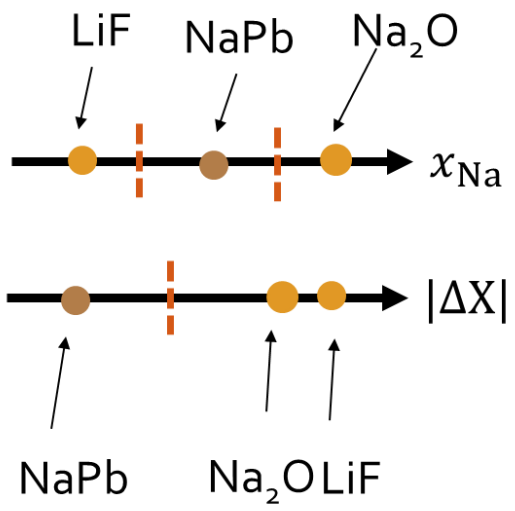
**Descriptive:** Similar materials should have similar representations

**Simple:** Representations should be fast to complete




Ref: Faber et al. Int. J. Quantum Chem. (2015) [And other excellent papers by Faber, von Lilienfeld, and team]

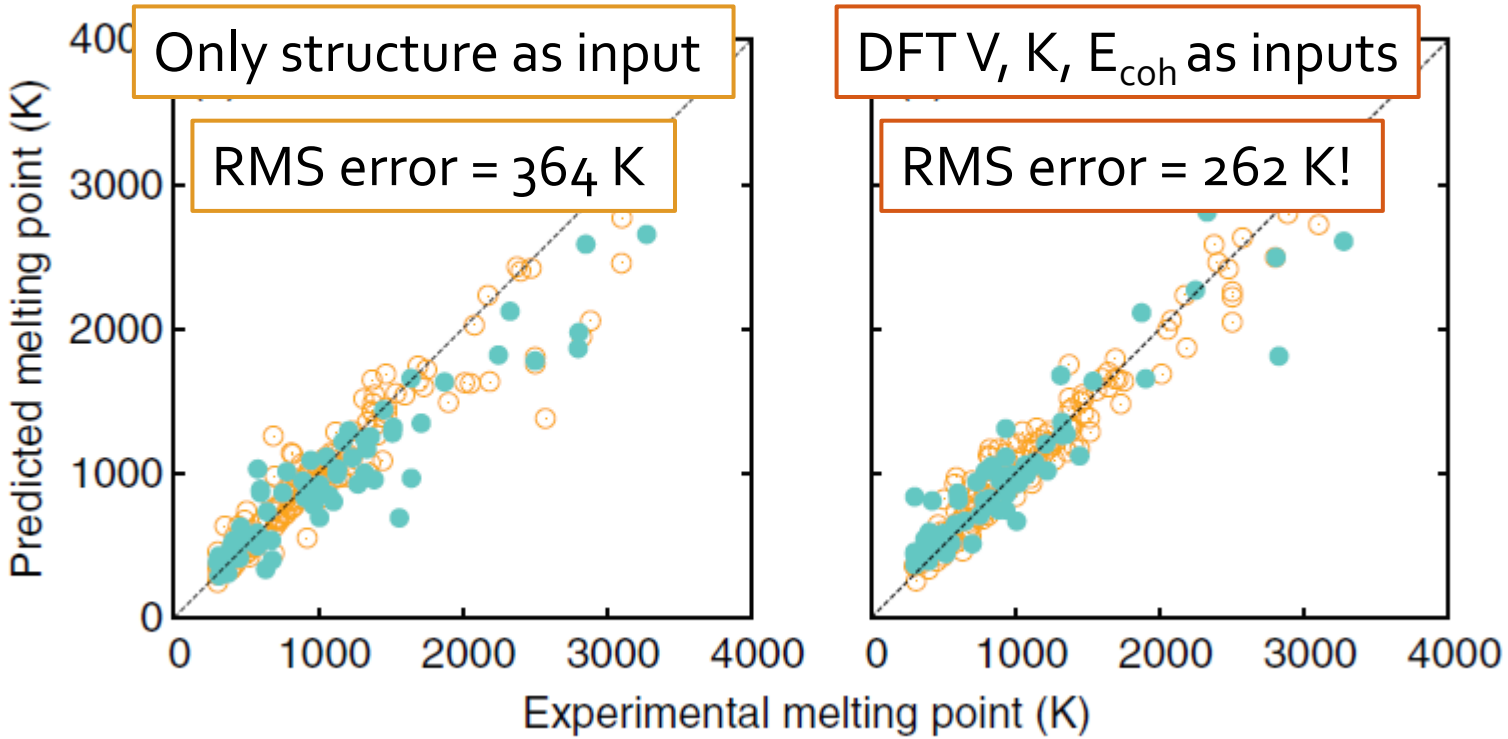
# Example: Back to "Is it Ionic?"




| | Fraction of Na | Difference in EN |
|---|---|---|
| **Complete** | ✓ Yes, no overlapping points | ✓ Yes, no overlapping points |
| **Compact** | ✓ Yes, single feature! | ✓ Yes, single feature! |
| **Descriptive** | ✗ No, rules are complicated | ✓ Yes, matches textbook |
| **Simple** | ✓ Yes, no compute | ✓ Yes, minimal compute |

# Representations standards are not clear cut

Seko et al. used DFT-computed properties as inputs to an ML model




**Ok for $10^2$ compounds.** Not OK for $10^5$ compounds

**Key Concept:** There is not and will not be a "one representation for all uses."

**Ref:** Seko *et al.*, PRB. (2014), 054303

# Types of representations: Discriminative vs Descriptive

| Discriminative | | Descriptive |
|---|---|---|
| Make features that capture intuition | **Concept** | Make features that distinguish examples |
| Element properties ($\lvert\Delta X\rvert$)<br>Interocular distances | **Examples** | Atomic fractions ($x_{Na}$)<br>Pixel values in images |
| Learn from little data<br>(potentially) Interpretable models | **Advantages** | Maximum expressivity |
| Introduces biases into the model<br>Need to know physics to solve problem | **Challenges** | Requires more data<br>(learn model *and features*) |

# “Every feature” vs “Hand Selected”

| Use Every Features You Can Think Of | | Selecting by Hand |
|---|---|---|
| • Use when physics poorly understood<br>• Can find unexpected linkages | **Advantages** | • Avoid fitting to bias in data<br>• Little need for feature selection |
| • Danger of overfitting<br>  • Find patterns not due to physics<br>  • More complex models<br>• Requires algorithms with <u>feature selection</u> | **Disadvantages** | • Requires understanding physics<br>• Limited to rules you expect to find |

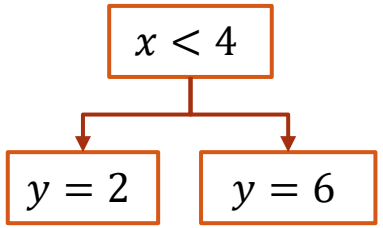
# Different representations for different algorithms

## Kernel Ridge Regression

$$f(x) = K(x, x'_i)$$

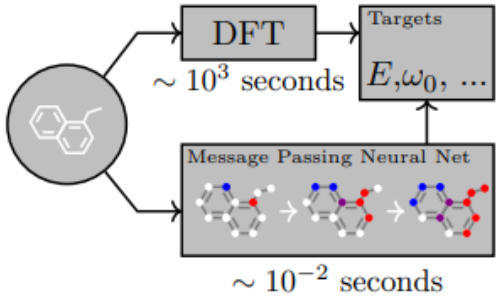
- Model requires a "kernel" and not "features"
- Only needs a single feature
- Kernels must be relevant to target property

## Random Forest Regression




- Can use with $10^2$-$10^3$ features
- Features work best if they compose simple rules
- Automatic feature selection
- Each feature must have same number of features

## Neural Networks




- Automatically learn representations from data
- Variable number of features, input shapes, …

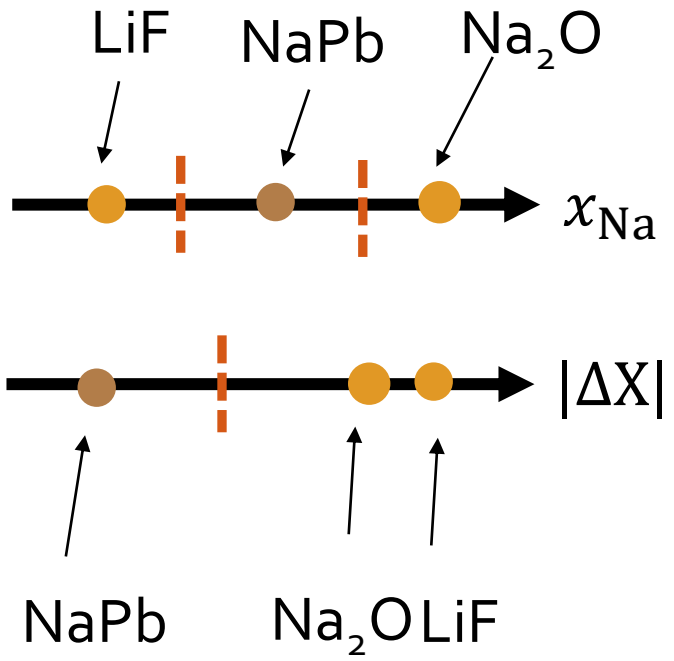
if you can design the architecture

# Take-away Points

- Representations are the key for encoding chemistry/physics into machine learning

- Representations <u>should</u> be:
  - Complete
  - Compact
  - Descriptive
  - Simple

- There is no best representation for "all problems"



# Learning goals for the following lectures

- Explain why representations for kernel ridge regression and neural networks are different
- Identify appropriate representations for inorganic and organic materials
- Design studies for testing different approaches that expose problems in models