Artificial Intelligence and Machine Learning in Chemistry

-how it learns, how it teaches

Aurora Clark

Examples of Artificial Intelligence

- **Learning from data to predict behavior**
  - large language models (LLMs)
    - predict new text that improves upon your own (ChatGPT)
  - decision making models
    - timing red and green lights on busy roads to optimize traffic flow
- **Learning from data to interpret observed behavior**
  - models that identify correlation of variables, even in high dimensional space
    - local vs. global weather patterns, how are pollutants transported in soil?
- **Learning from data to interpret AND predict behavior**
  - Models of personalized learning experiences
    - based on questions, identify what student's gaps are in understanding and predict what new information will help them

Examples of Artificial Intelligence

- **Learning from data to predict behavior**
  - large language models (LLMs)
    - predict new text that improves upon your own (Cha
  - decision making models
    - timing red and green lights on busy roads to optimi
- **Learning from data to interpret observed behavior**
  - models that identify correlation of variables, even in high dimensional space
    - local vs. global weather patterns, how are pollutants transported in soil?
- **Learning from data to interpret AND predict behavior**
  - Models of personalized learning experiences
    - based on questions, identify what student's gaps are in understanding and predict what new information will help them

Examples of Artificial Intelligence

- **Learning from data to predict behavior**
  - large language models (LLMs)
    - predict new text that improves upon your own (Cha
  - decision making models
    - timing red and green lights on busy roads to optimi
- **Learning from data to interpret observed behavior**
  - models that identify correlation of variables dimensional space
    - local vs. global weather patterns, how are pollutants
- **Learning from data to interpret AND predi**
  - Models of personalized learning experience
    - based on questions, identify what student's gaps ar predict what new information will help them

Examples of Artificial Intelligence

- **Learning from data to predict behavior**
  - large language models (LLMs)
    - predict new text that improves upon your own (ChatGPT)
  - decision making models
    - timing red and green lights on busy roads to optimize traffic flow
- **Learning from data to interpret observed behavior**
  - models that identify correlation of variables, even in high dimensional space
    - local vs. global weather patterns, how are pollutants transported in soil?
- **Learning from data to interpret AND predict behavior**
  - Models of personalized learning experiences
    - based on questions, identify what student's gaps are in understanding and predict what new information will help them

Examples of Artificial Intelligence

- **Learning from data to predict behavior**
  - large language models (LLMs)
    - predict new text that improves upon your own (ChatGPT)
  - decision making models
    - timing red and green lights on busy roads to optimize traffic flow
- **Learning from data to interpret observed behavior**
  - Models that identify correlation of variables, even in high dimensional space
    - how a misfolded protein causes disease
- **Learning from data to interpret AND predict behavior**
  - Models of personalized learning experiences
    - based on questions, identify what student's gaps are in understanding and predict what new information will help them

Examples of Artificial Intelligence

- **Learning from data to predict behavior**
  - large language models (LLMs)
    - predict new text that improves upon your own (ChatGPT)
  - decision making models
    - timing red and green lights on busy roads to optimize traffic flow
- **Learning from data to interpret observed behavior**
  - Models that identify correlation of variables, even in high dimensional space
    - how a misfolded protein causes disease
- **Learning from data to interpret AND predict behavior**
  - Models of personalized learning experiences
    - based on questions, identify what student's gaps are in understanding and predict what new information will help them

Examples of Artificial Intelligence

- **Learning from data to predict behavior**
  - large language models (LLMs)
    - predict new text that improves upon your own (ChatGPT)
  - decision making models
    - timing red and green lights on busy roads to optimize traffic flow
- **Learning from data to interpret observed behavior**
  - Models that identify correlation of variables, even in high dimensional space
    - how a misfolded protein causes disease
- **Learning from data to interpret AND predict behavior**
  - Models of personalized learning experiences
    - based on questions, identify what student's gaps are in understanding and predict what new information will help them

Examples of Artificial Intelligence

- **Learning from data to predict behavior**
  - large language models (LLMs)
    - predict new text that improves upon your own (ChatGPT)
  - decision making models
    - timing red and green lights on busy roads to optimize traffic flow

A model is a simplified representation of something in our real world

  - dimensional space
    - how a misfolded protein causes disease

- **Learning from data to interpret AND predict behavior**
  - Models of personalized learning experiences
    - based on questions, identify what student's gaps are in understanding and predict what new information will help them

Examples of Artificial Intelligence

- Learning from data to predict behavior
  - large language models (LLMs)
    - predict new text that improves upon your own (ChatGPT)
  - decision making models
    - timing red and green lights on busy roads to optimize traffic flow

A model is a simplified representation of something in our real world

- for an individual, our brains create models of our world based upon our sensory experiences and interactions

- Learning from data to interpret AND predict behavior
  - Models of personalized learning experiences
    - based on questions, identify what student's gaps are in understanding and predict what new information will help them

Examples of Artificial Intelligence

A model is a simplified representation of something in our real world

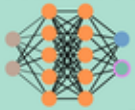
- for an individual, our brains create models of our world based upon our sensory experiences and interactions
- AI creates a model based upon information provided by humans

Artificial Intelligence

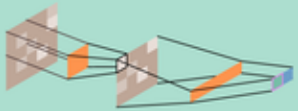
Machine Learning

Deep Learning

Deep Neural Network

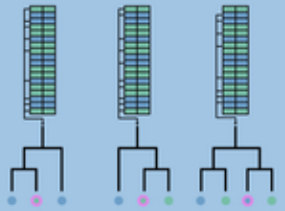


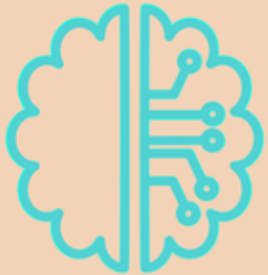
Convolutional Neural Network



Boosted Regression Tree

Random Forest





How can *we* use AI?

As scientists:

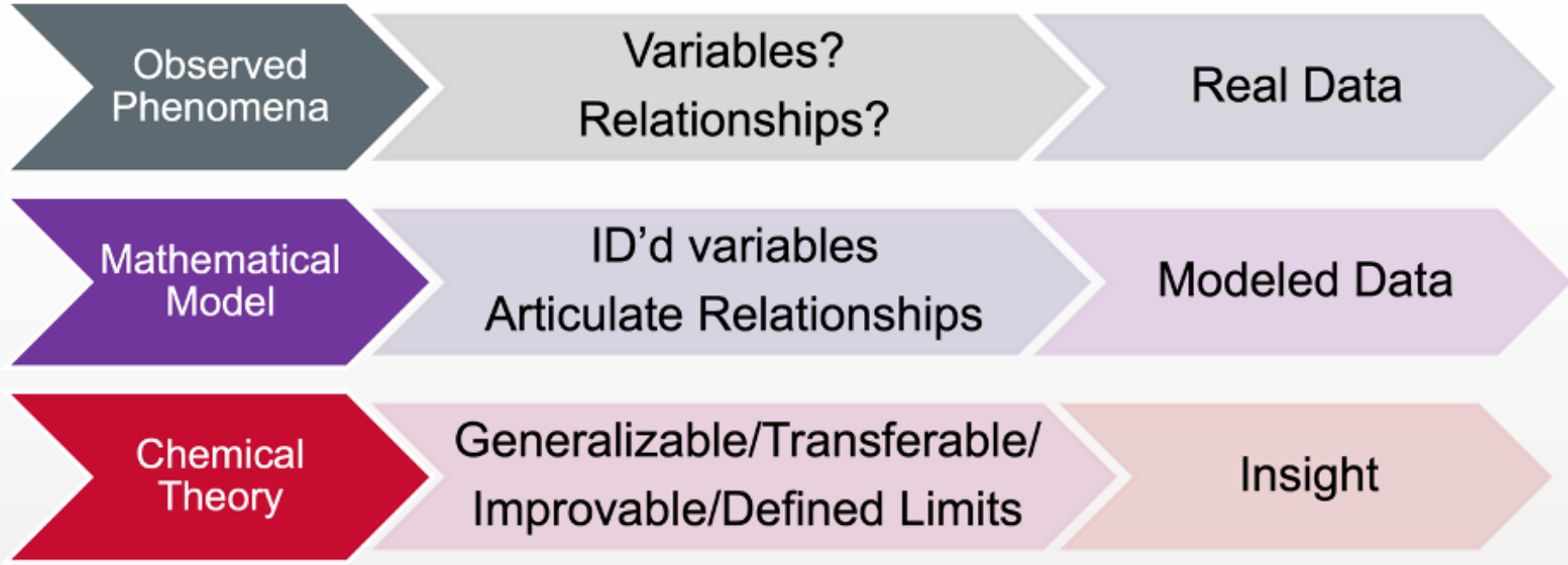
1) Gaining scientific insight

2) Predicting scientific phenomena

As educators:

3) Improving SLO's by learning what our students understand and how they can be supported

How do we gain scientific insight?

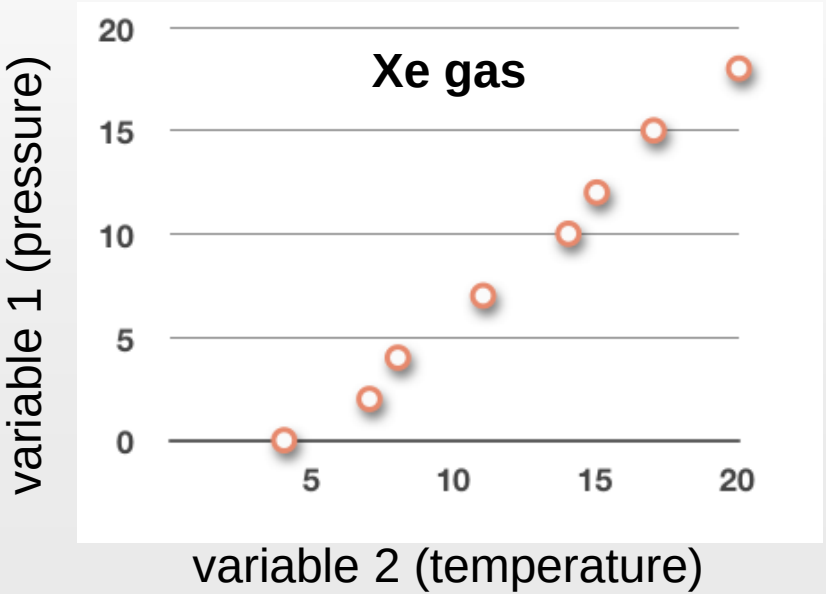


How do we gain scientific insight?

Observed Phenomena → Variables? Relationships? → Real Data

How does the pressure change as a function of temperature of a gas?

P, T



How do we gain scientific insight?

Mathematical Model → ID'd variables, Articulate Relationships → Modeled Data

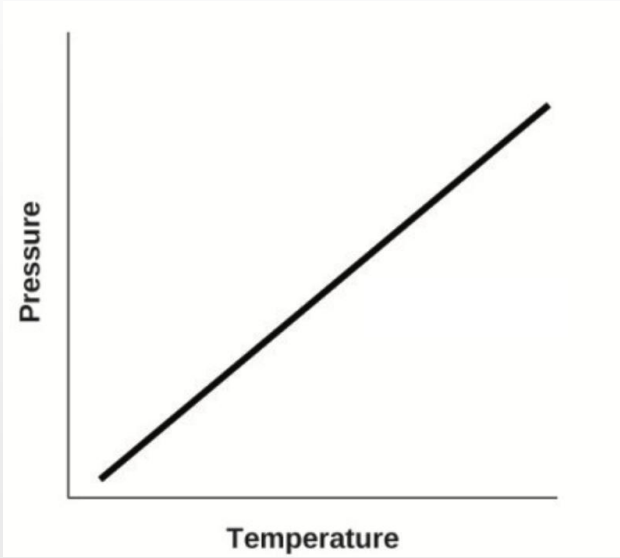
straight line

$$P \propto T$$

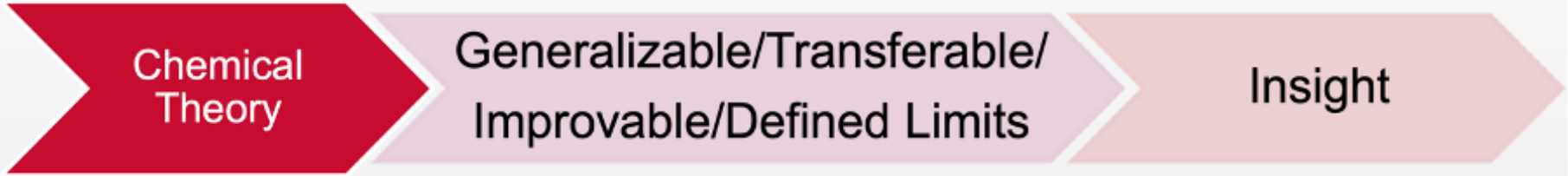
$$\frac{P}{T} = c$$

$$P = cT$$

what constant best fits the data?



How do we gain scientific insight?



What other variables are relevant for gases?

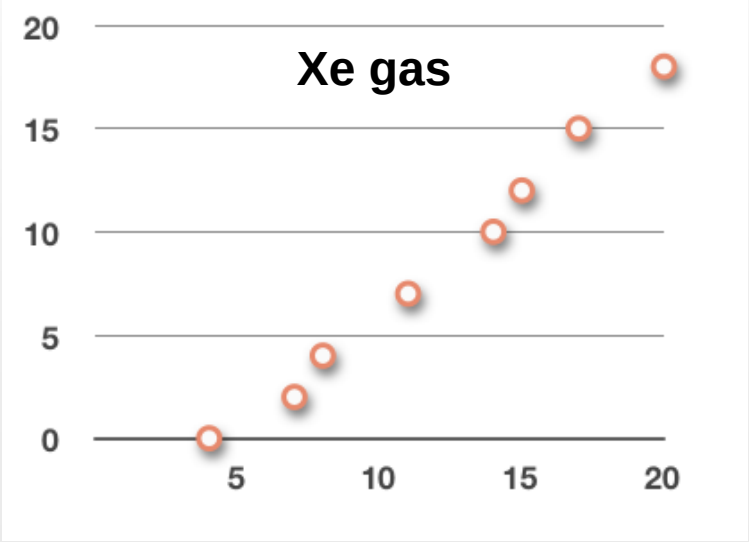
How do they behave?

How do we gain scientific insight?

Observed Phenomena → Variables? Relationships? → Real Data

What other variables are relevant for gases?

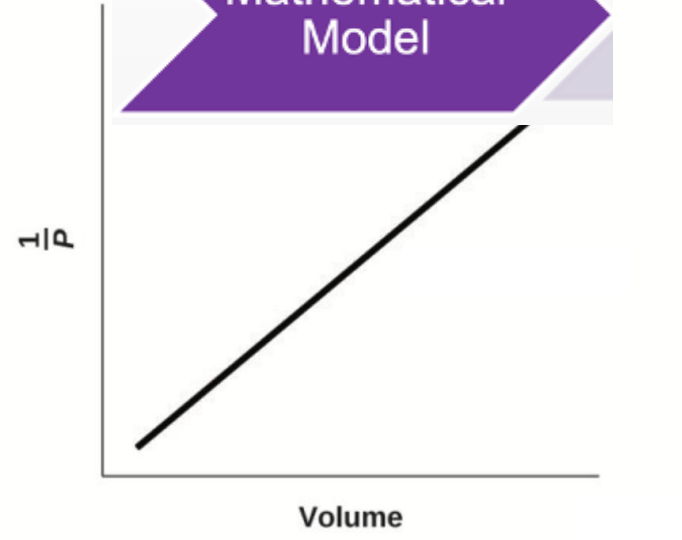
How do they behave?

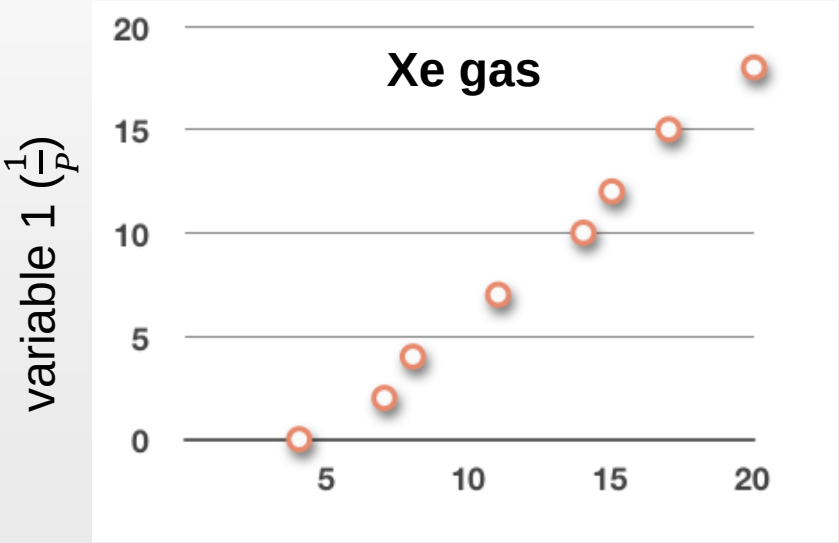


How do we gain scientific insight?

Observed Phenomena → Variables? Relationships? → Real Data

Mathematical Model





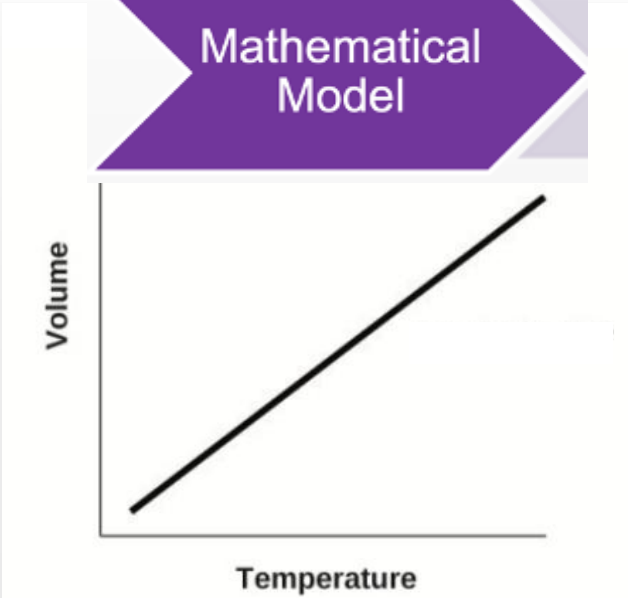
How do we gain scientific insight?

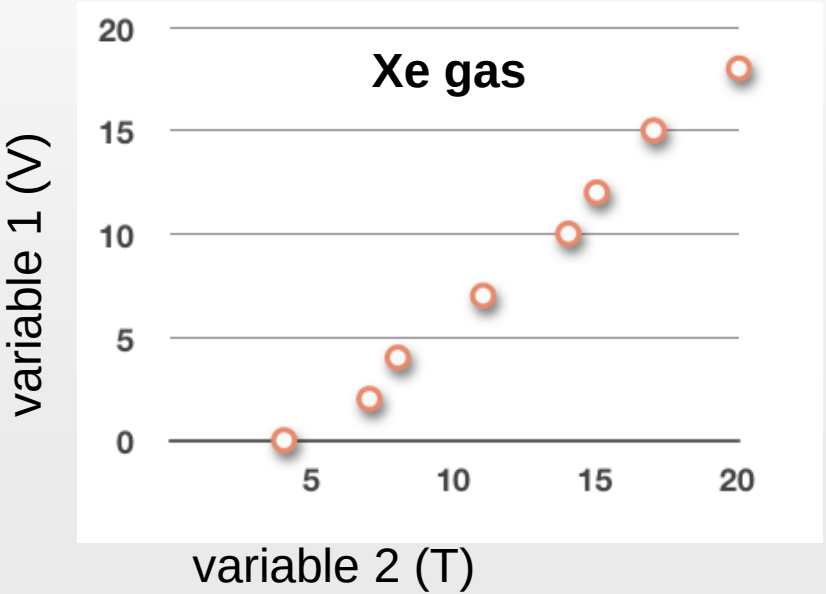
Observed Phenomena

Variables?
Relationships?

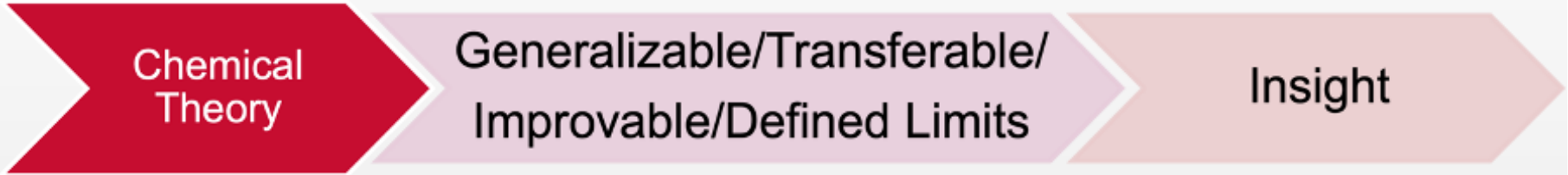
Real Data

Mathematical Model





How do we gain scientific insight?



Combine all observations mathematical models:

$$P = cT$$
$$V = cT$$
$$V = cn$$

$$PV = ncT$$

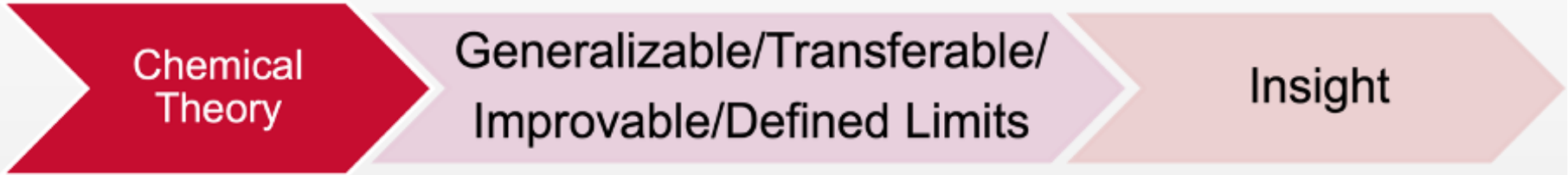
$$c = \frac{PV}{nT}$$

If we assume:

- that gases do not interact
- gases have no volume
- all collisions are elastic

c can be fit to lots of different data under different conditions, or measured directly

How do we gain scientific insight?



Combine all observations mathematical models:

$$P = cT$$
$$V = cT$$
$$V = cn$$
$$PV = nRT$$

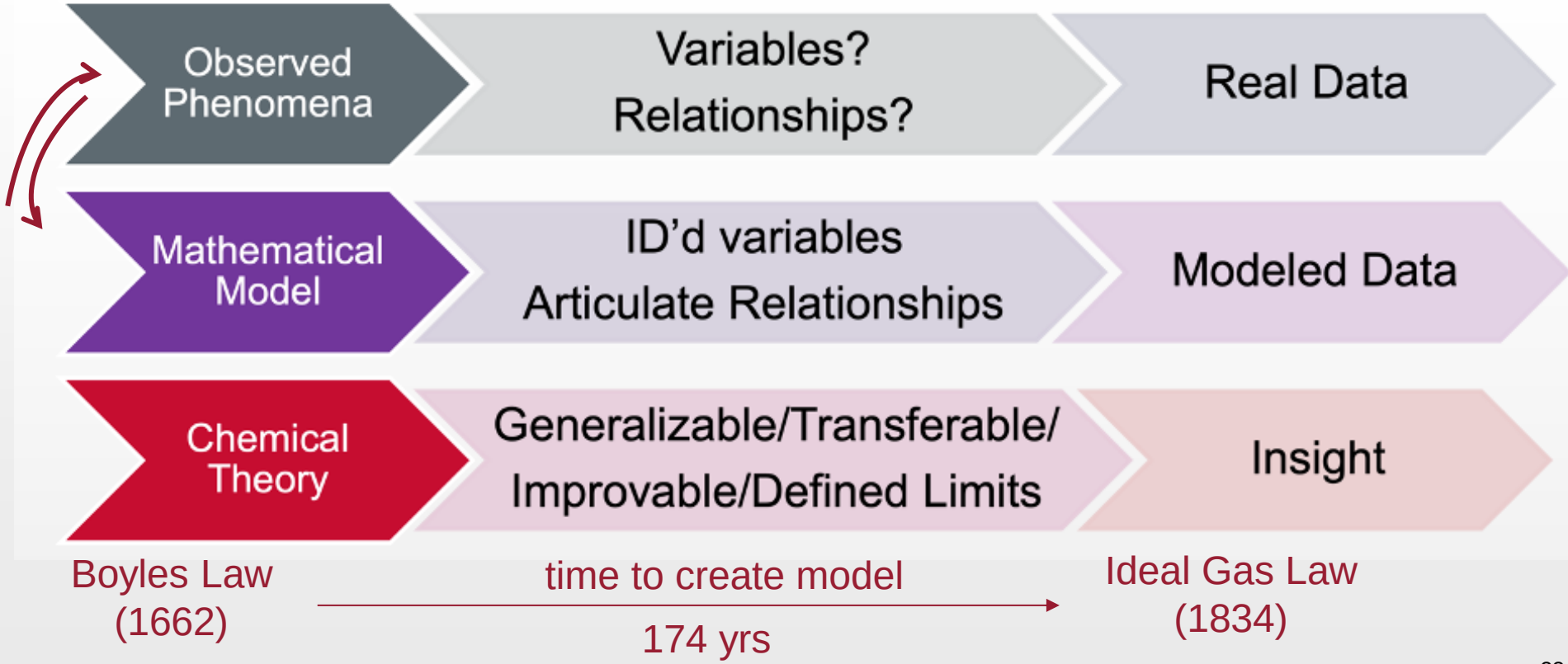
Ideal Gas Law
Equation of state for hypothetical gas

If we assume:

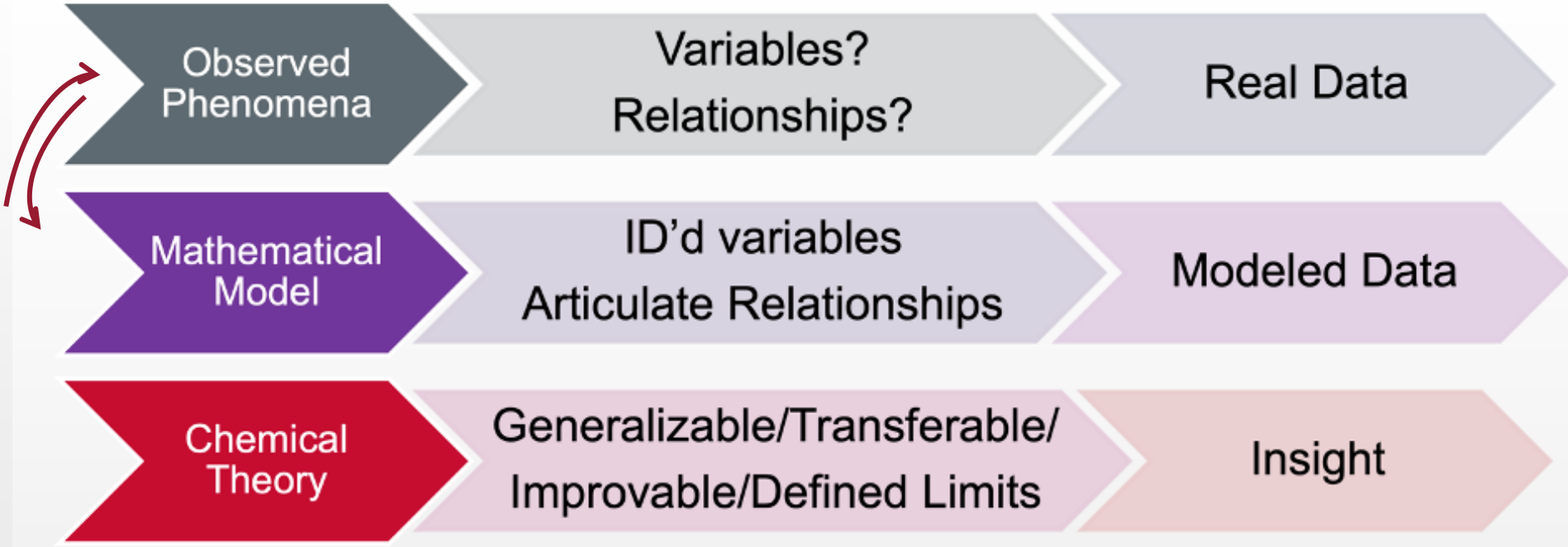
- that gases do not interact
- gases have no volume
- all collisions are elastic

- predict how gas behavior changes under different conditions
- why do hot air balloons float?
- how your lungs work during inhalation/exhalation

Gaining Insight with Human Intelligence

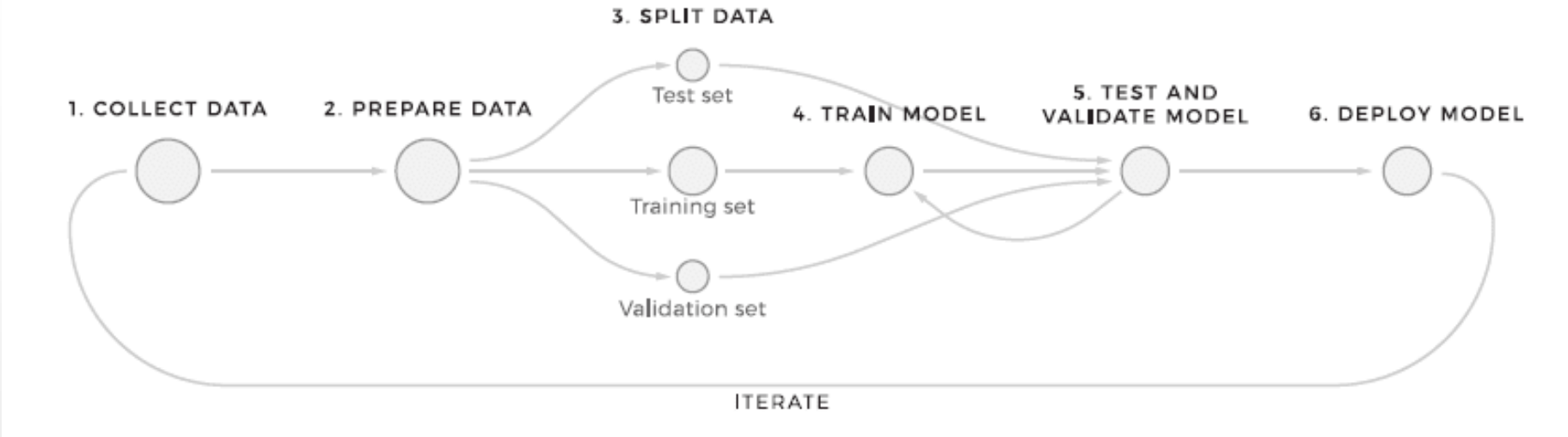


Gaining Insight with Artificial Intelligence



All observations — time to create model: ? → Ideal Gas Law

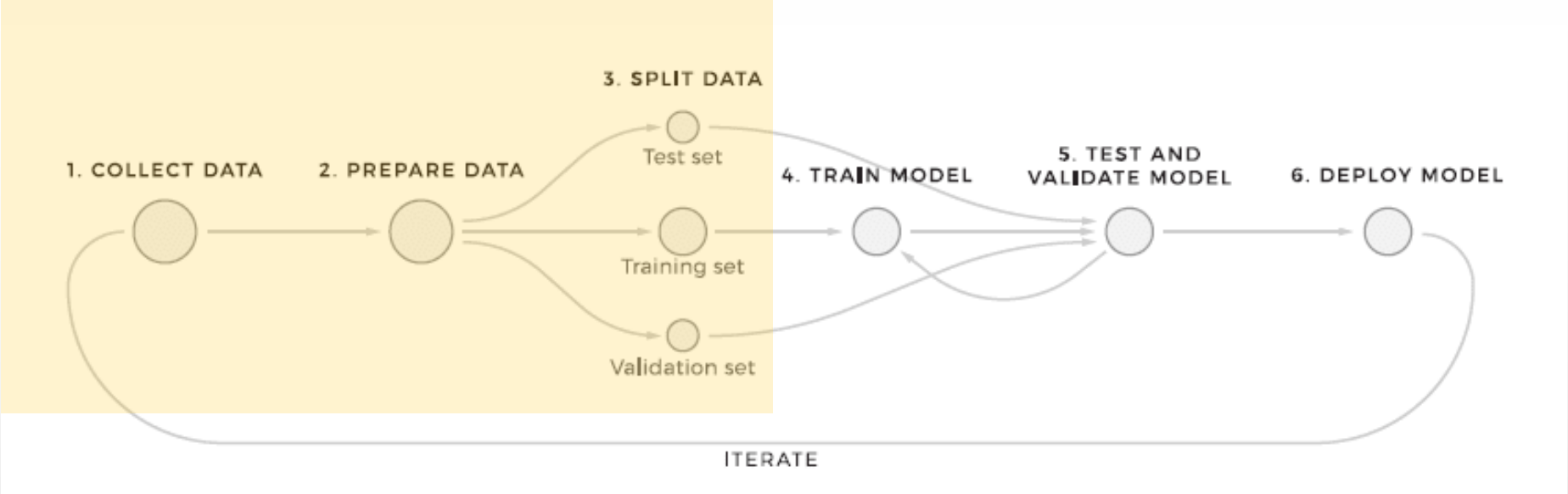
How We Create a Model with AI



Humans participate in each step (making choices), but machines and software do the work

-consider the 174 years of thought and observation that we are trying to replace in a single workflow that takes < 1 day

How We Create a Model with AI



The data you provide AI, is its experience of the world. If the data is flawed/limited, the model will be flawed/limited

-just like people – if a person has never experienced rock climbing, they don't know how to rock climb

How We Create a Model with AI

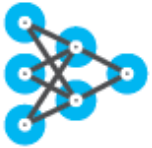
6 dimensions of data quality

Accuracy
The degree to which the data correctly represents the entity or attribute being described.

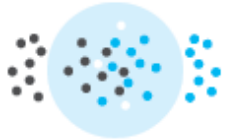
Completeness
The percentage of missing data from a given data set.

Consistency
The absence of difference or contradiction in data irrespective of the data's source.

Validity
Invalid data affects the accuracy and completeness of a data set.

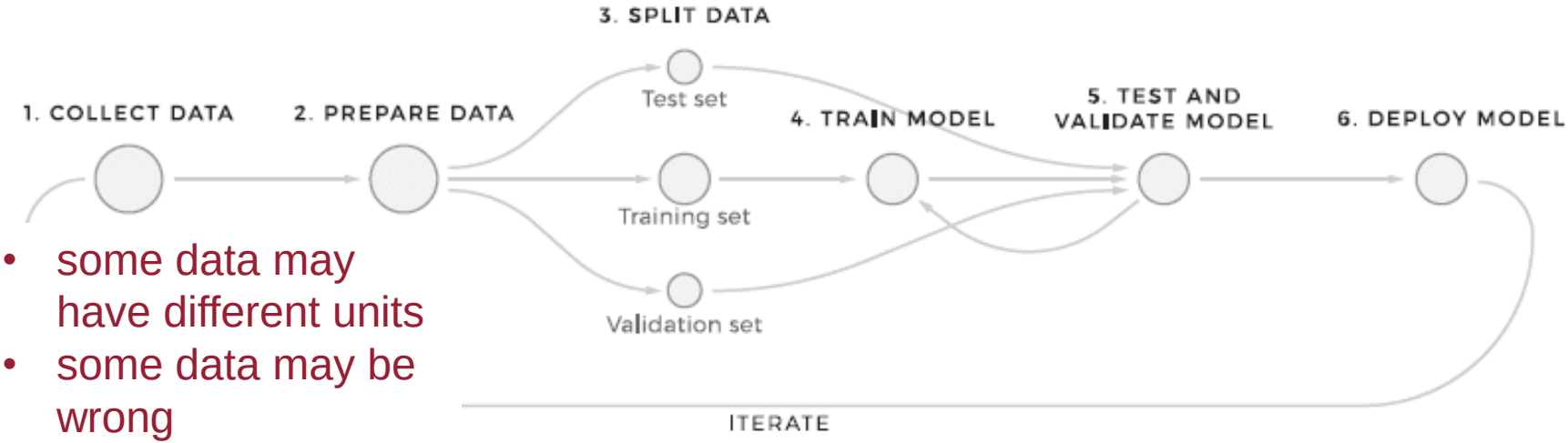


Integrity
The validity of relationships across various data entities.



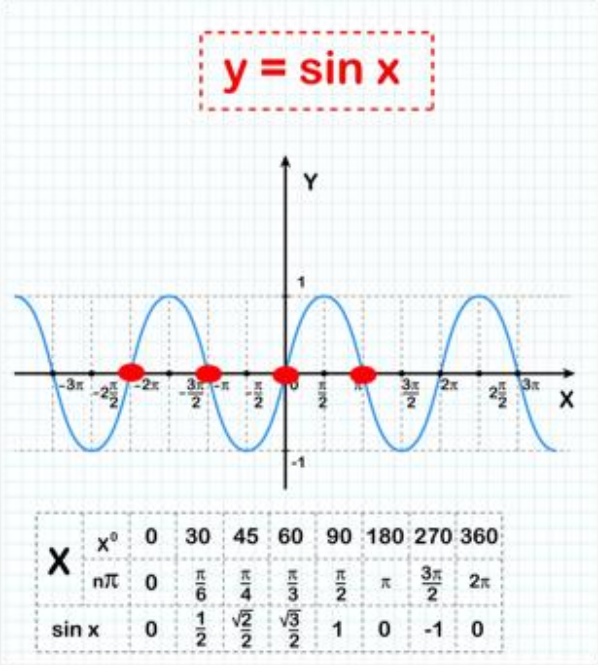
Uniqueness
Ensures duplicate or overlapping data is identified and marked.

How We Create a Model with AI

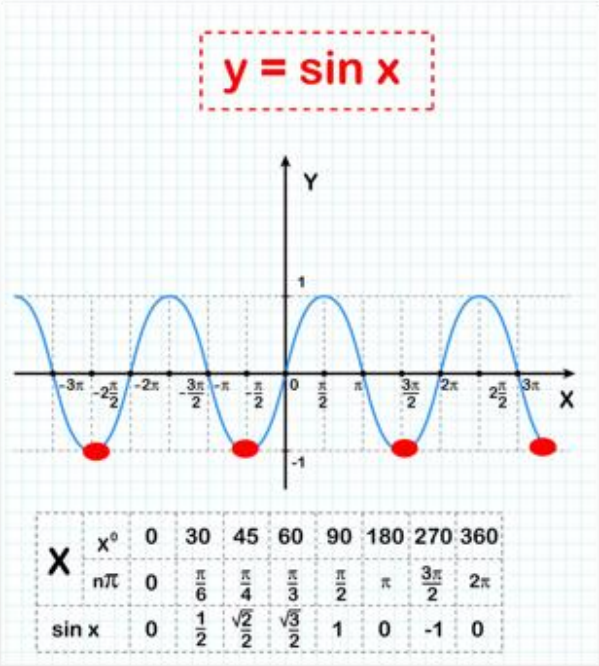


- some data may have different units
- some data may be wrong
- **are we sampling reality "enough"**

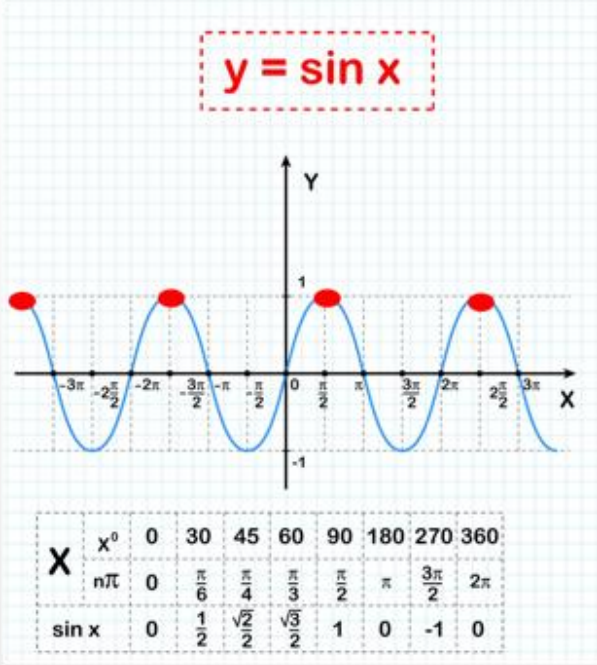
Why Sampling Matters



X	x°	0	30	45	60	90	180	270	360
	nπ	0	π/6	π/4	π/3	π/2	π	3π/2	2π
sin x		0	1/2	√2/2	√3/2	1	0	-1	0



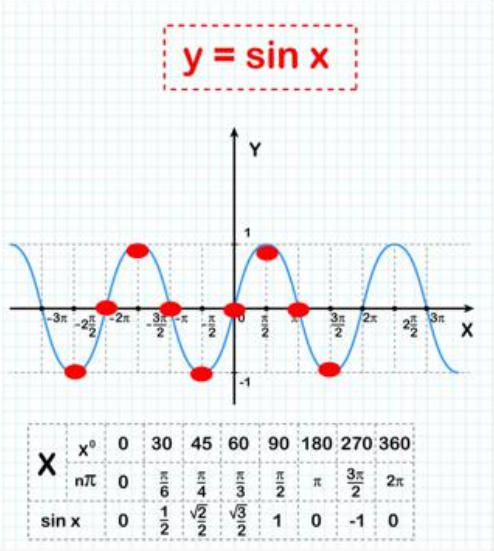
X	x°	0	30	45	60	90	180	270	360
	nπ	0	π/6	π/4	π/3	π/2	π	3π/2	2π
sin x		0	1/2	√2/2	√3/2	1	0	-1	0



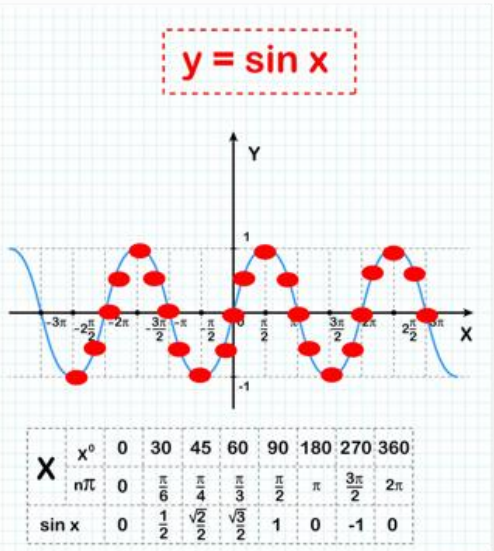
X	x°	0	30	45	60	90	180	270	360
	nπ	0	π/6	π/4	π/3	π/2	π	3π/2	2π
sin x		0	1/2	√2/2	√3/2	1	0	-1	0

Why Sampling Matters

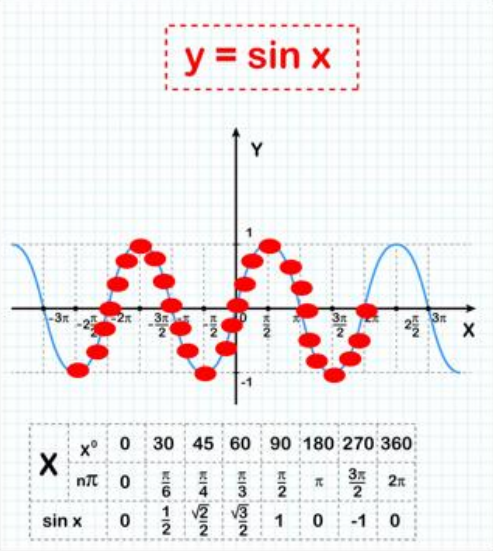
Systematic sampling



X	x°	0	30	45	60	90	180	270	360
	nπ	0	π/6	π/4	π/3	π/2	π	3π/2	2π
	sin x	0	1/2	√2/2	√3/2	1	0	-1	0



X	x°	0	30	45	60	90	180	270	360
	nπ	0	π/6	π/4	π/3	π/2	π	3π/2	2π
	sin x	0	1/2	√2/2	√3/2	1	0	-1	0

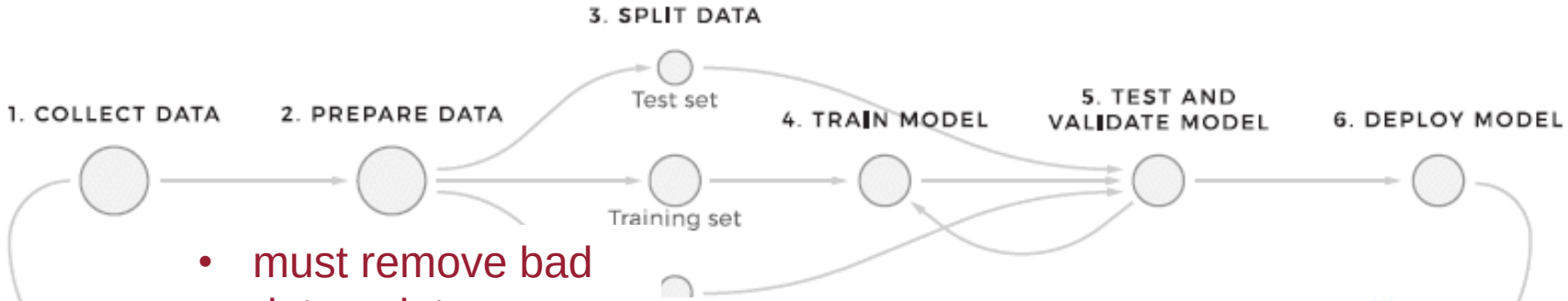


X	x°	0	30	45	60	90	180	270	360
	nπ	0	π/6	π/4	π/3	π/2	π	3π/2	2π
	sin x	0	1/2	√2/2	√3/2	1	0	-1	0

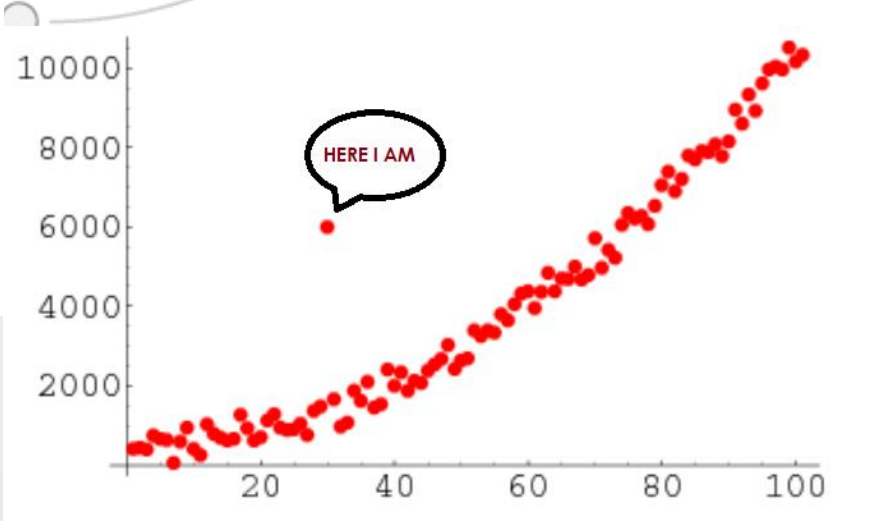
Sparse
(discrete)

Dense
(approaching continuity)

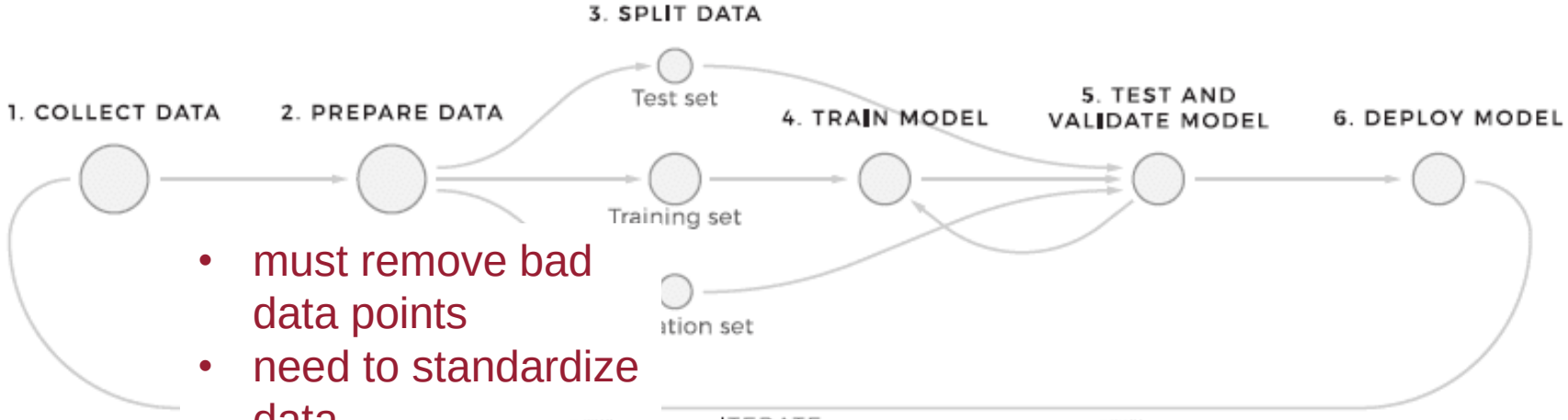
How We Create a Model with AI



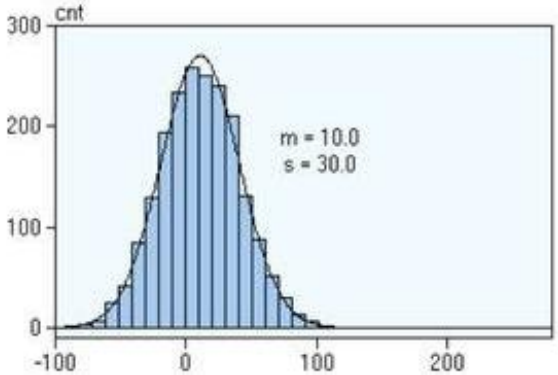
- must remove bad data points
  - how do you do this? find statistical outliers

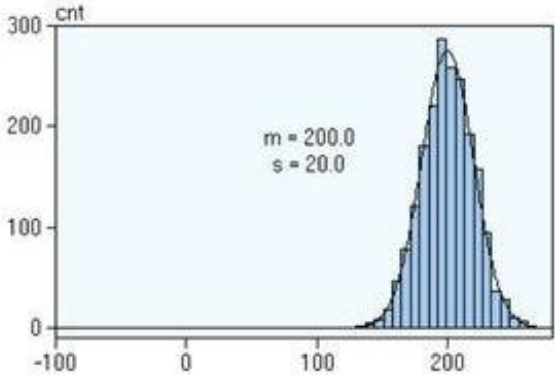


How We Create a Model with AI

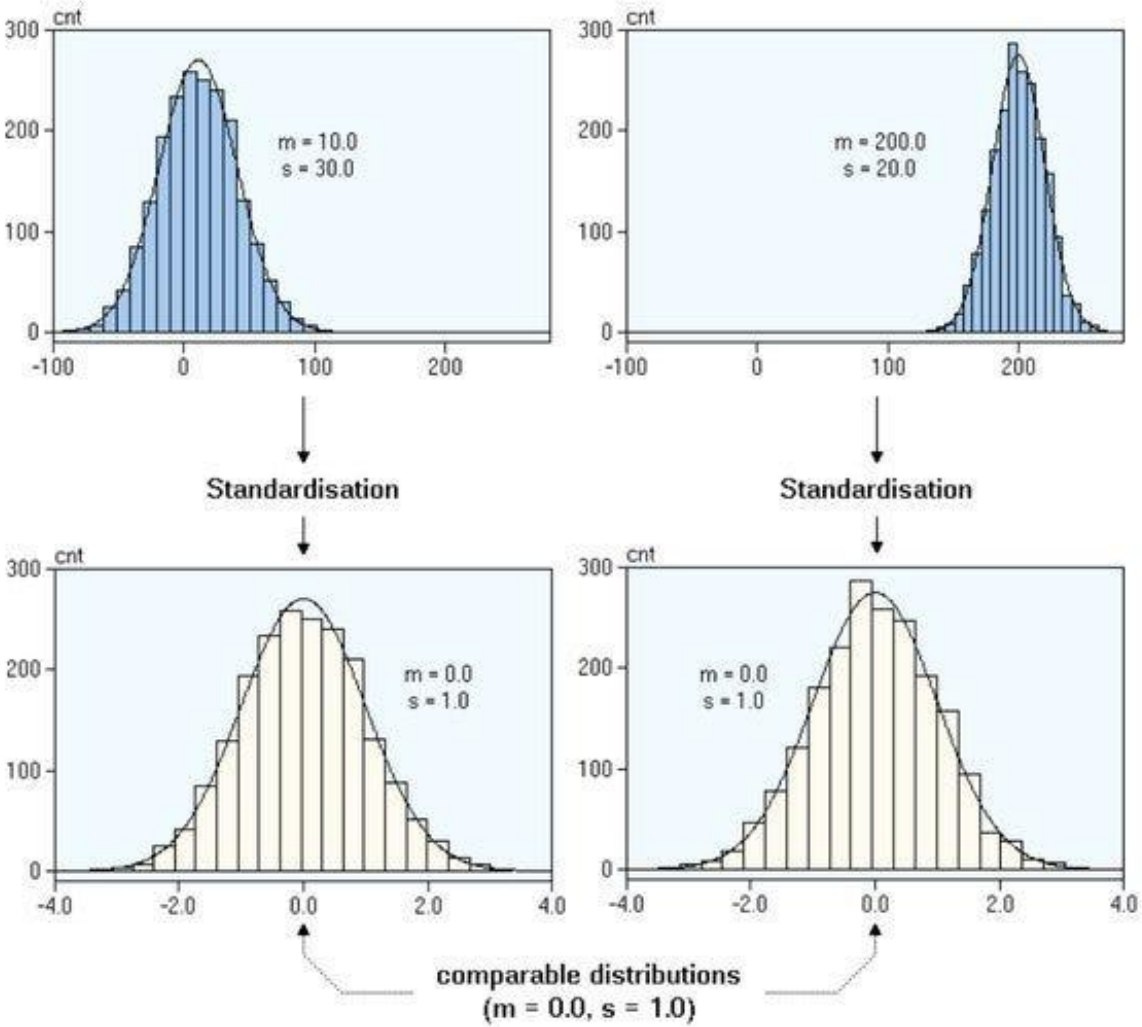


- must remove bad data points
- need to standardize data



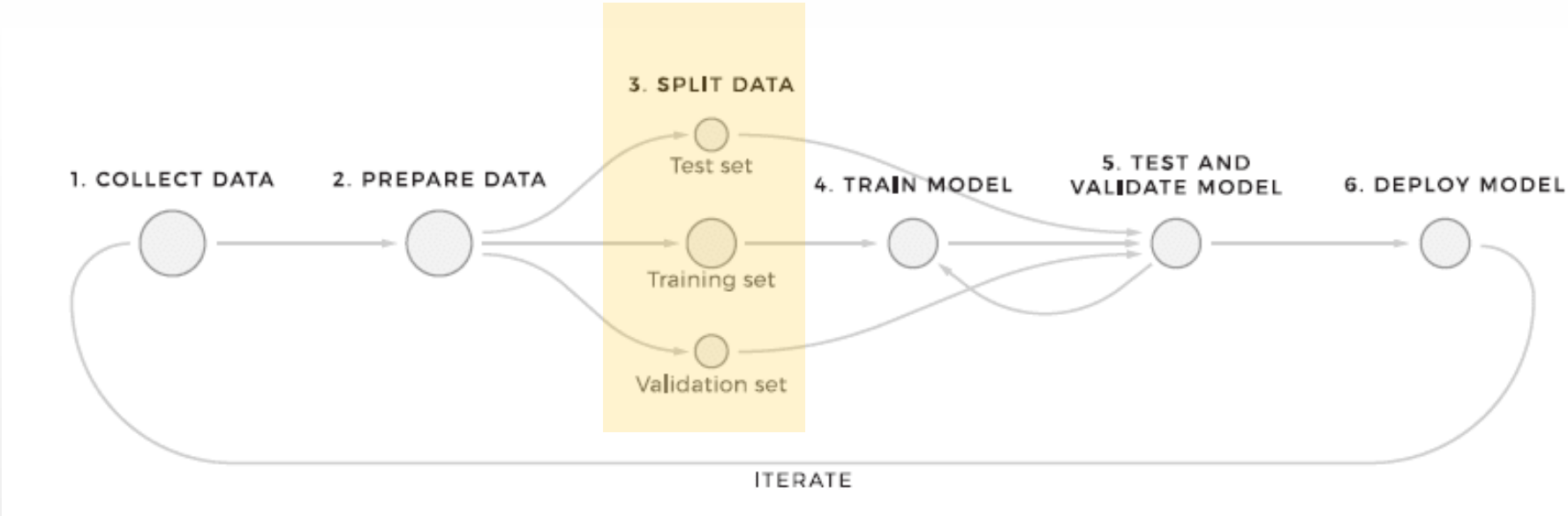


1. COLLECT DAT



5. DEPLOY MODEL

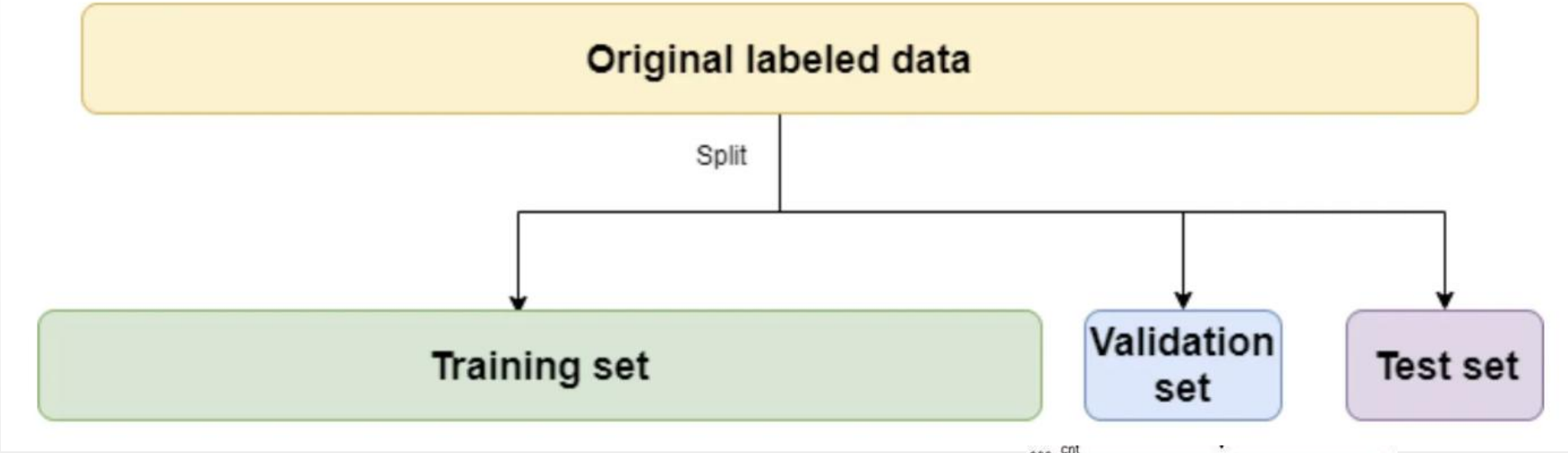
How We Create a Model with AI



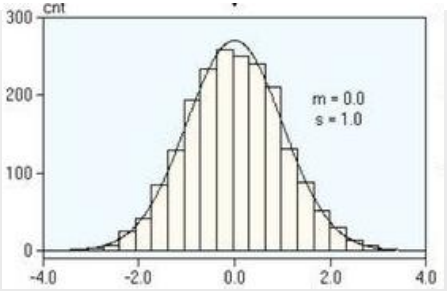
We split the data because we want the model to be able to predict observations from data it was not trained on

- a model is useless if it has only “memorized” what the answer is, it isn’t “learning”
- just like teaching a student: materials we use to teach and materials we use to assess

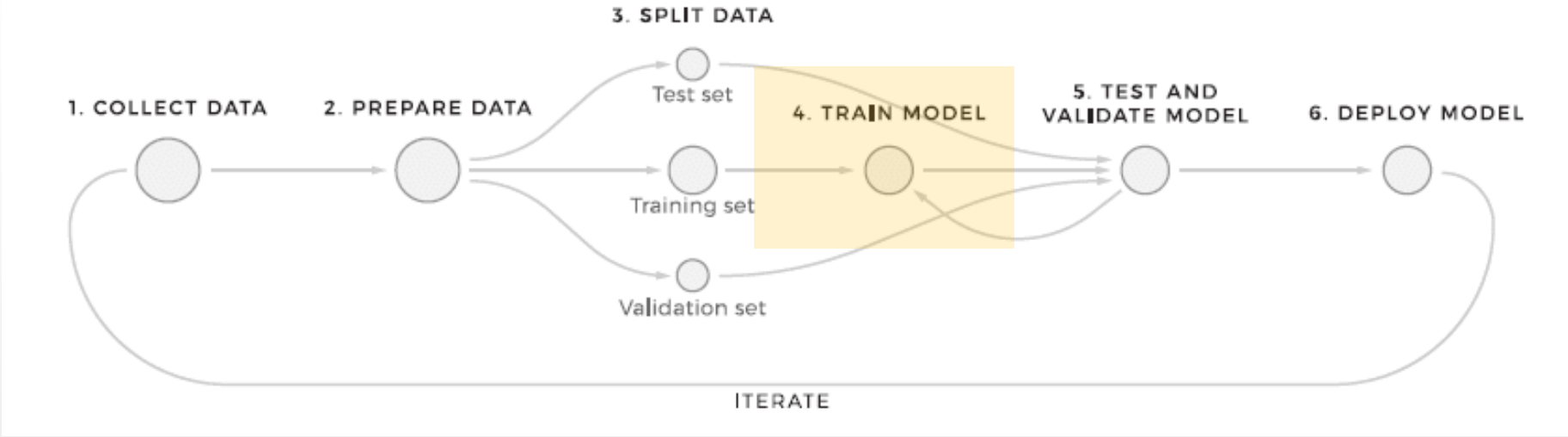
How We Create a Model with AI



- must not be biased toward specific observations or parts of the distribution of the observations



How We Create a Model with AI



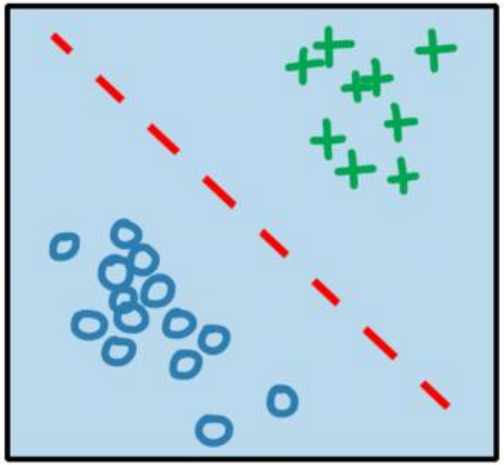
This is 2 parts:

1) a human chooses the model (the function, set of functions that *we* think will describe or help us interpret the observations)

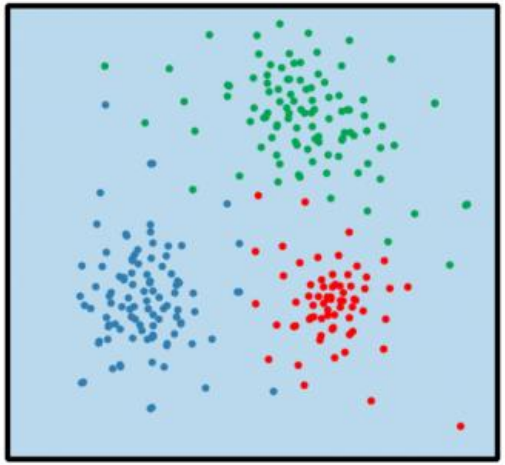
2) we train the model (often fitting, but not always)

Types of Training

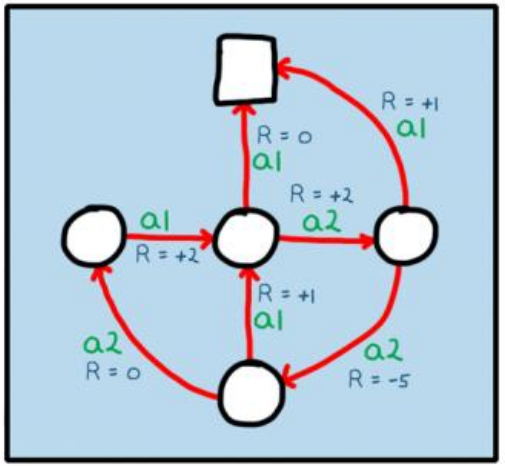
supervised learning



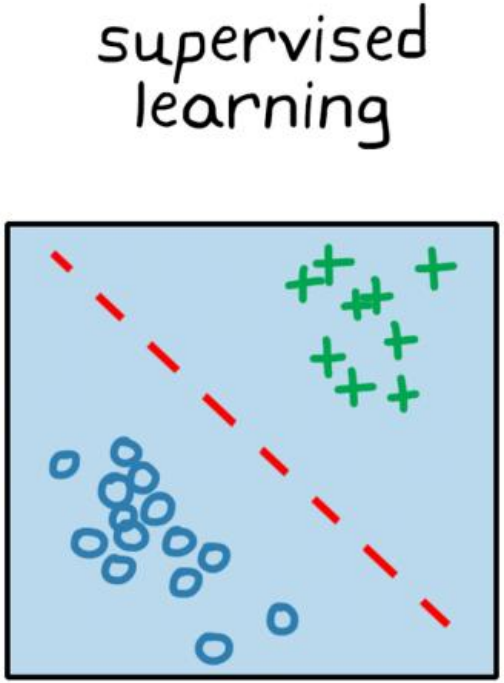
unsupervised learning



reinforcement learning

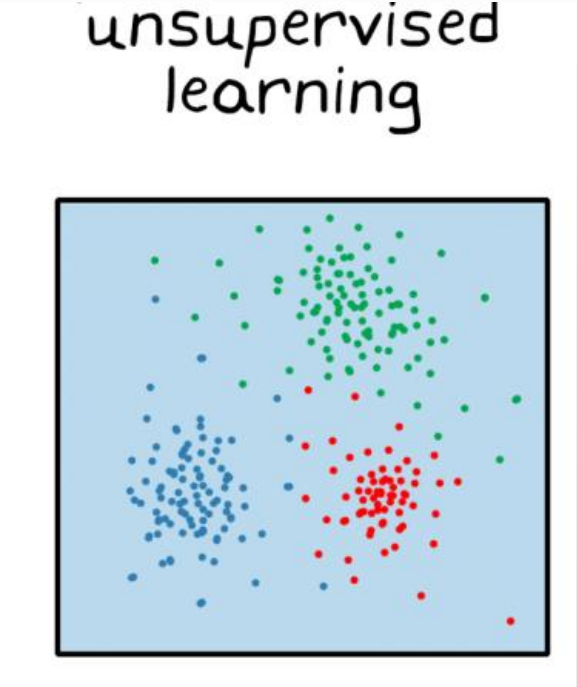


Types of Training

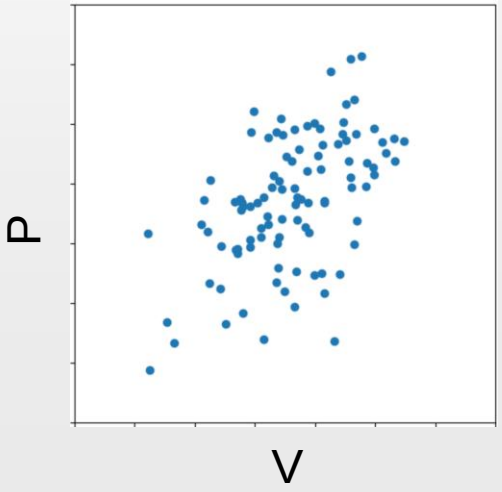


- We select function/functions that we think describes the underlying physics that causes the observation
  - Ex: PV=nRT
    - we use some type of regression or optimization to find R
  - Ex: I think that the data on left can be split into 2 groups
    - optimize variables in a function that best splits them

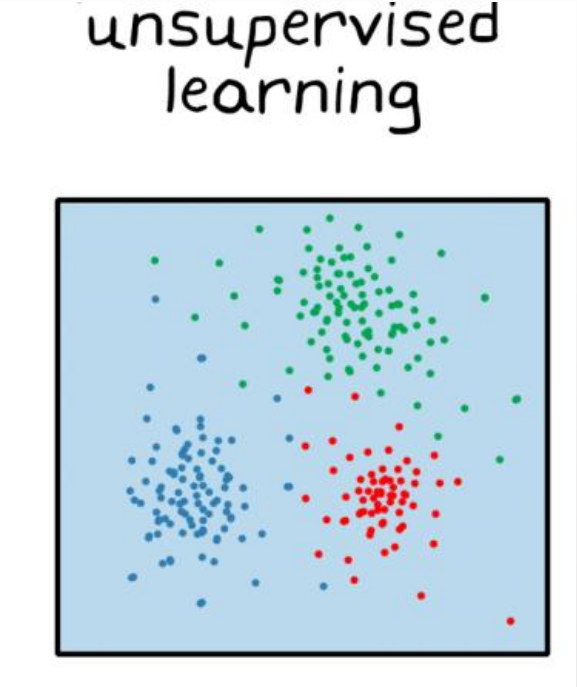
Types of Training



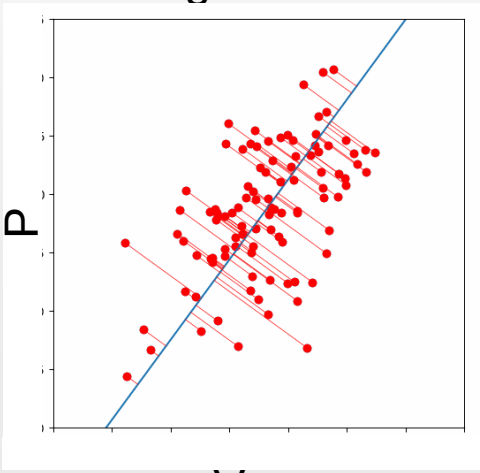
- I hypothesize there are relationships between certain variables in my dataset, lets let math find out what they are
  - Ex: P is related to V, how is it related?
    - use linear algebra to find out



Types of Training

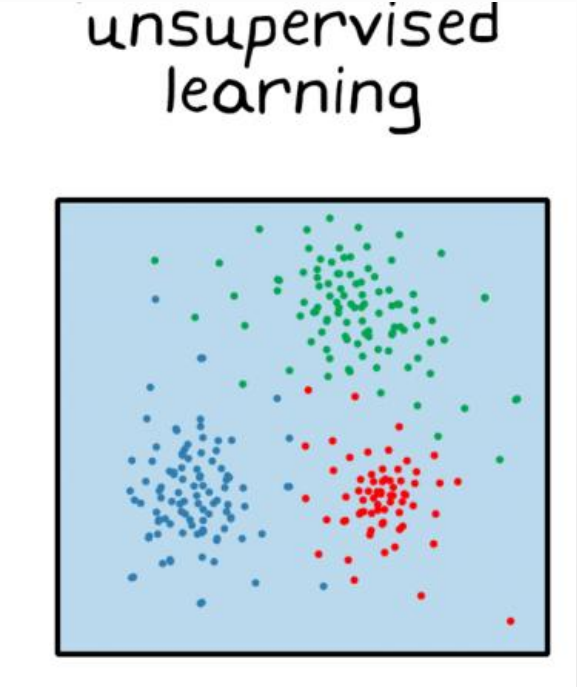


- I hypothesize there are relationships between certain variables in my dataset, lets let math find out what they are
  - Ex: P is related to V, how is it related?
    - use linear algebra to find out

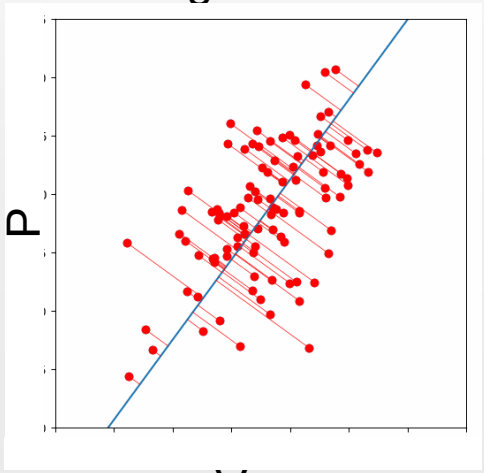


Principal Component Analysis (PCA)

Types of Training

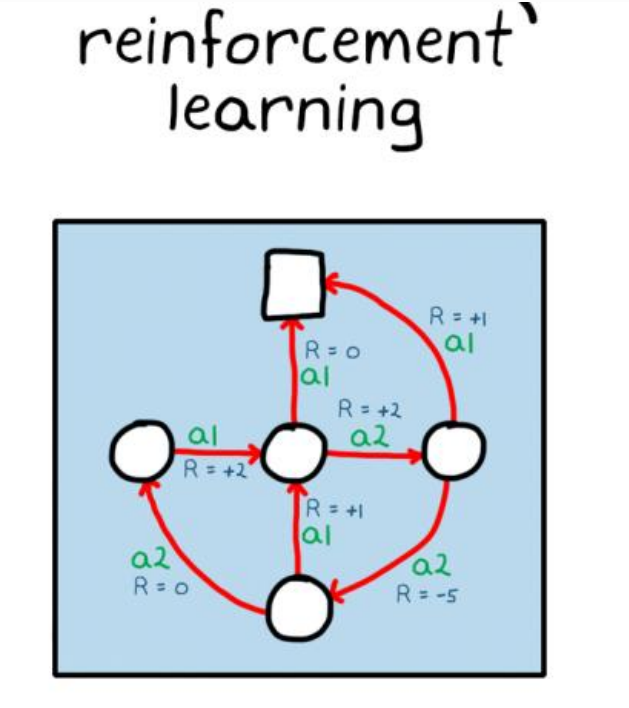


- I hypothesize there are relationships between certain variables in my dataset, lets let math find out what they are
  - Ex: P is related to V, how is it related?
    - use linear algebra to find out



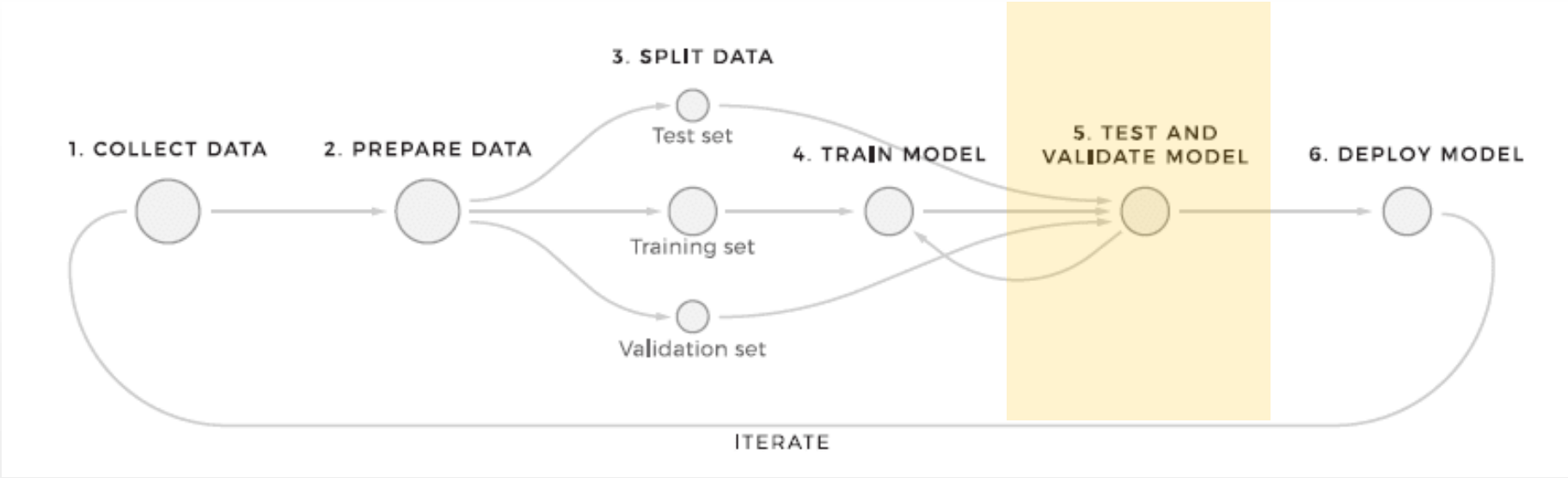
Principal Component Analysis (PCA)

Types of Training



- We select function/functions that we think describes the underlying physics that causes the observation
  - Ex: PV=nRT
    - use software to tell us what what observations agree or disagree with the function that we chose

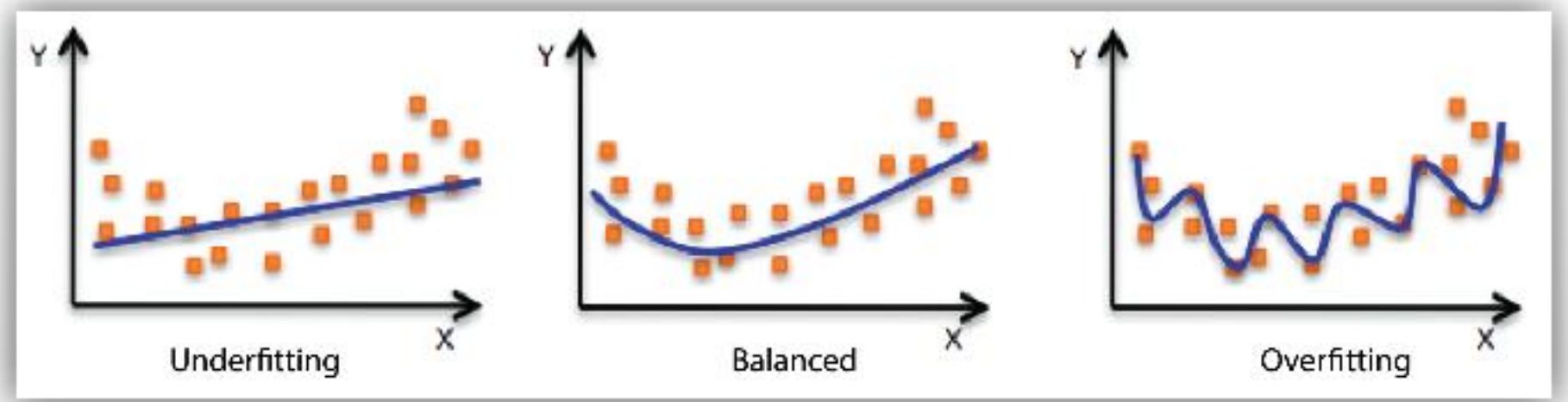
How We Create a Model with AI



Here, we do assessment:

- do we need to tune any parameters in our model (uses validation set)?
  - does our model overfit the data?
- can we predict observations in the test set?

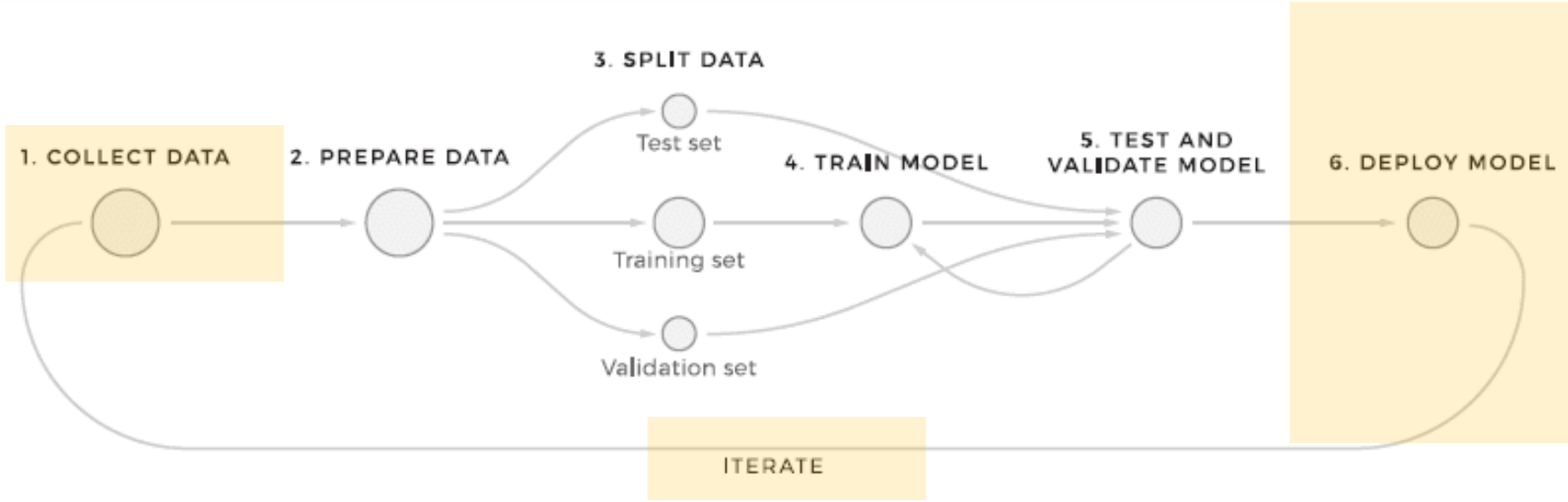
How We Create a Model with AI



Has our model learned the data?
NO

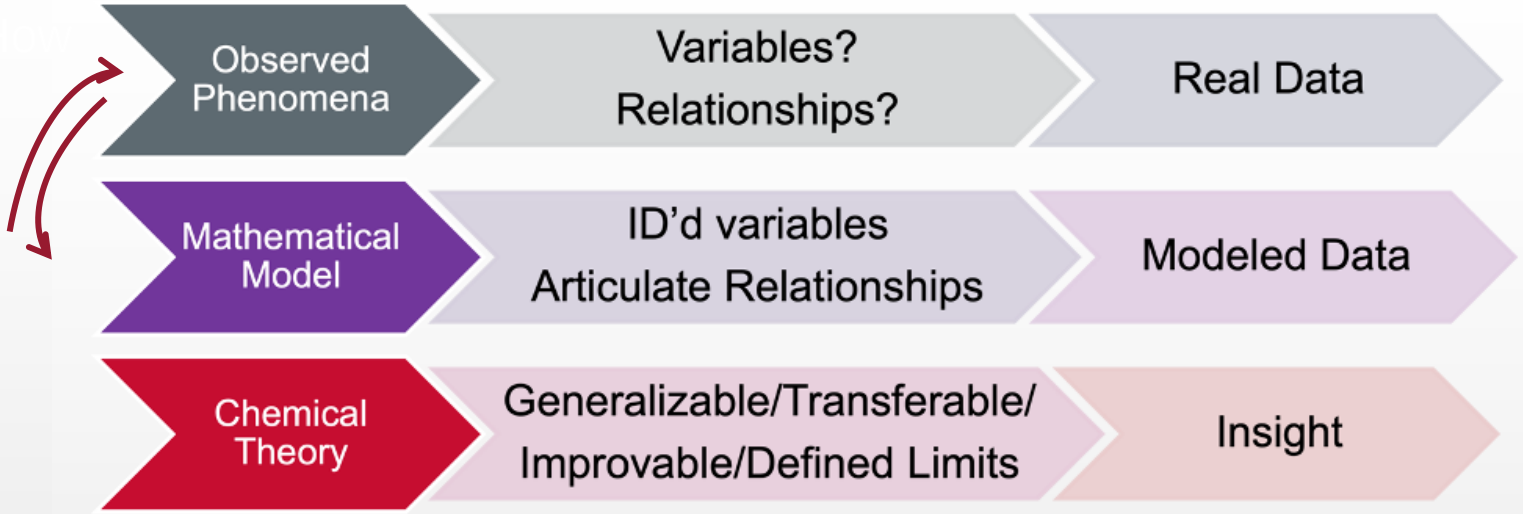
Has our model memorized the data?
YES

How We Create a Model with AI



A model should be constantly updated with new information

Gaining Insight with Artificial Intelligence



All observations ⟶ Ideal Gas Law

Once data is collected – model created in < 1 day (down from 174 yrs)

- any mistake in the different steps in ML will create a bad model
  - ML quick to create models, but we have to scrutinize their behavior

Artificial Intelligence and Machine Learning in Chemistry

-how it learns, how it teaches

- AI reflects how we learn, but at a much-accelerated pace
- All of the pitfalls of learning occur in AI
- As scientists, we have to be more critical, more judicious and have higher standards than ever when using AI

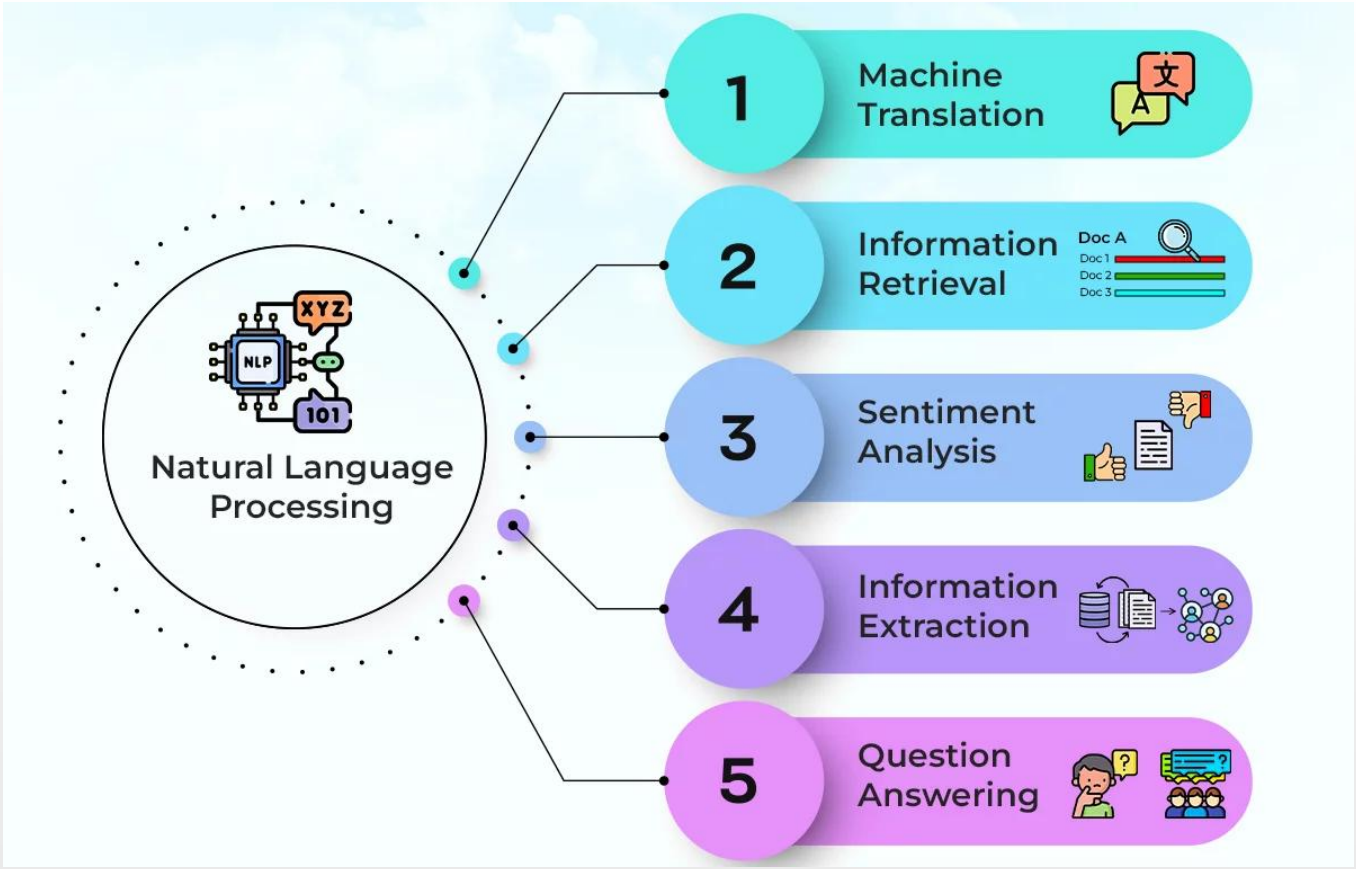
Artificial Intelligence and Machine Learning in Chemistry

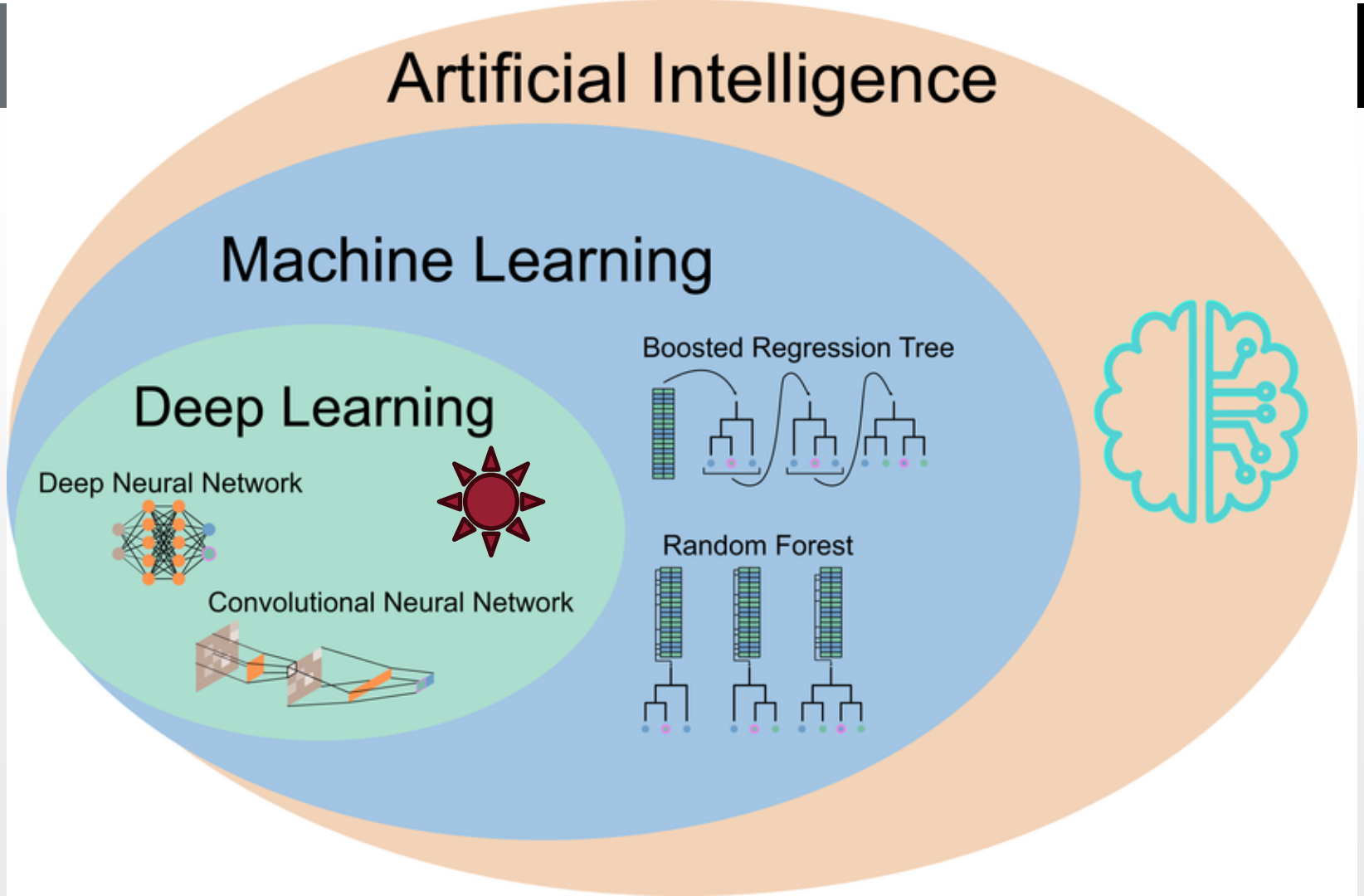
-how it learns, how it teaches

Examples of Artificial Intelligence

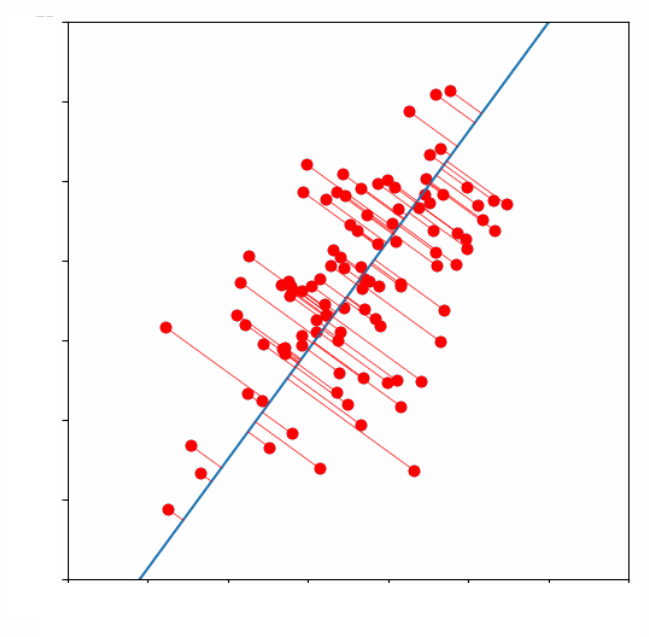
- **Learning from data to predict behavior**
  - large language models (LLMs)
    - predict new text that improves upon your own (ChatGPT)
  - decision making models
    - timing red and green lights on busy roads to optimize traffic flow
- **Learning from data to interpret observed behavior**
  - models that identify correlation of variables, even in high dimensional space
    - local vs. global weather patterns, how are pollutants transported in soil?
- **Learning from data to interpret AND predict behavior**
  - Models of personalized learning experiences
    - based on questions, identify what student's gaps are in understanding and predict what new information will help them

Natural Language Processing Models




Artificial Intelligence
Machine Learning
Deep Learning
Deep Neural Network
Convolutional Neural Network
Boosted Regression Tree
Random Forest

How a Neural Network Works



How we would normally treat data:

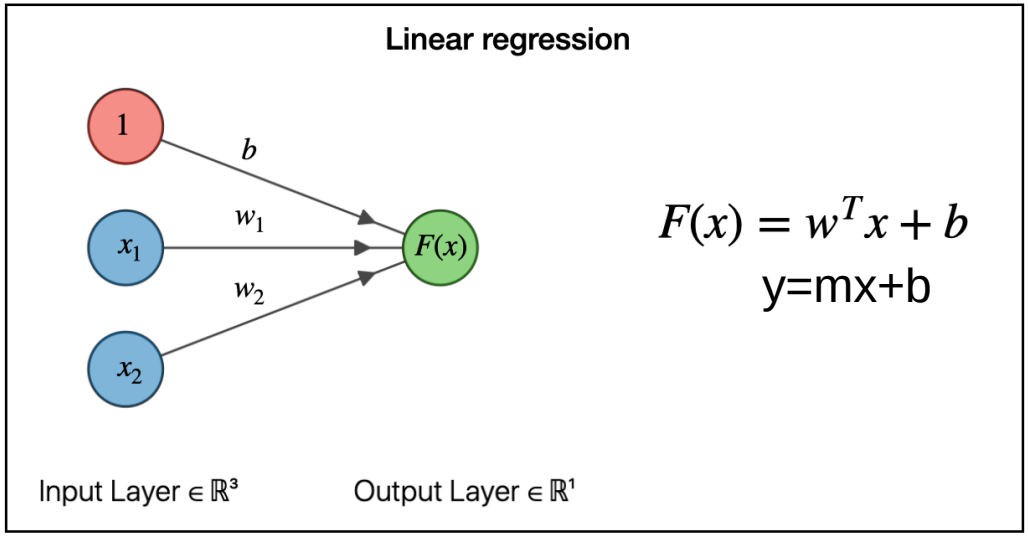
- I have data, it looks like a line
- I write a function for a line
- I tune the parameters of the function to best fit the data (best fit slope, linear regression)

How a neural networks treats data:

- I have data, it looks like a line
- I write the function for a line
- How do I weight or bias the data so that the data sits on the line?

How a Neural Network Works

Simplest neural network is a linear fit



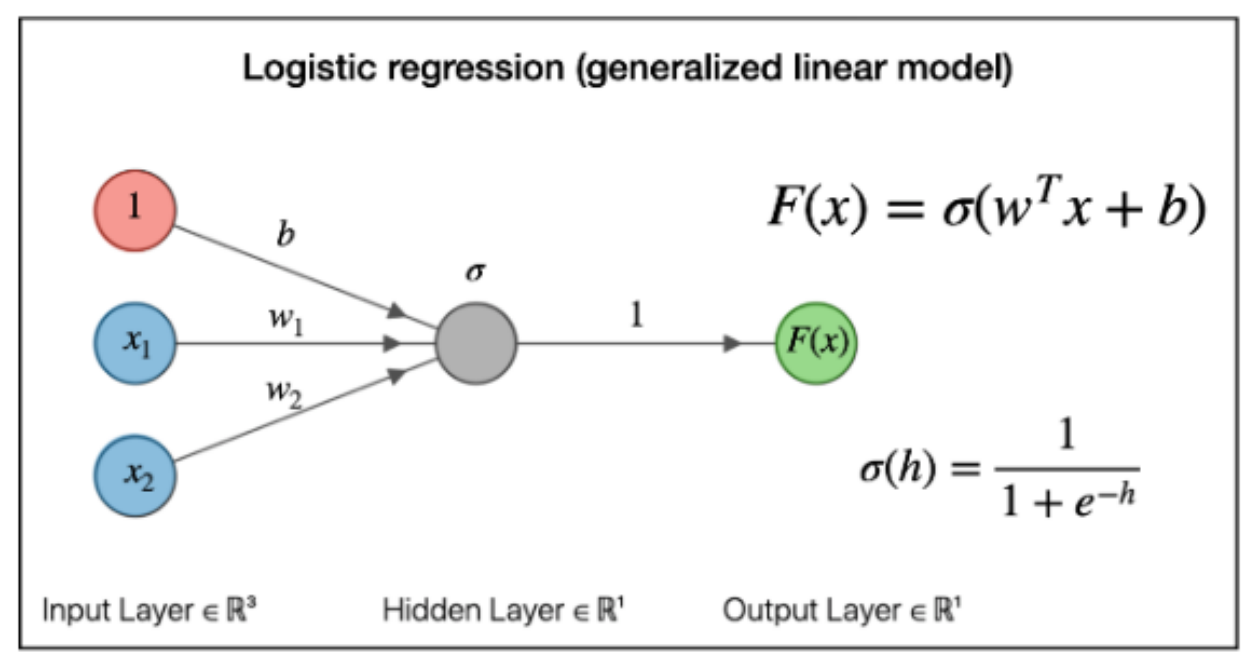
$$F(x) = w^T x + b$$

y=mx+b

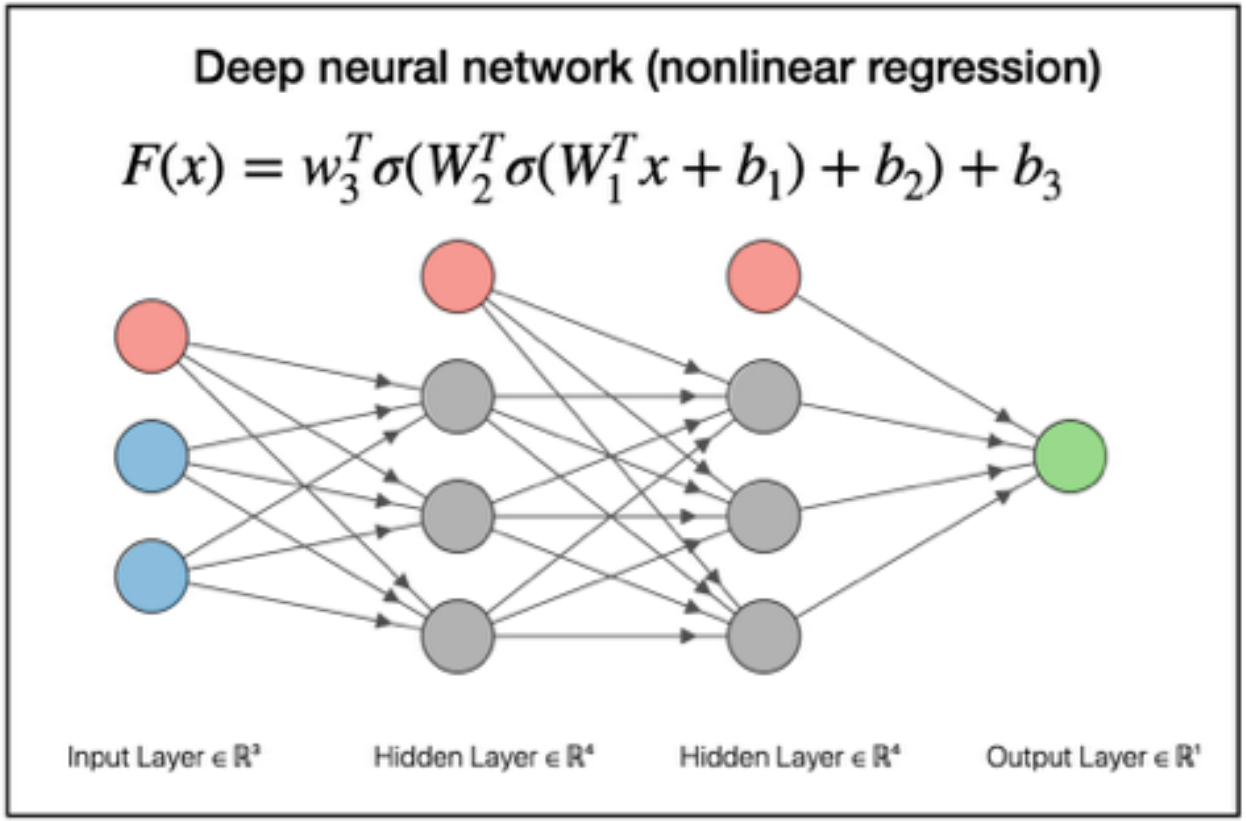
$$F(x) = w_1 \cdot x_1 + w_2 \cdot x_2 + 1 \cdot b$$

How a Neural Network Works

-add a layer of complexity – pass the linear function you fit into another function → a generalized linear model (logistic regression)

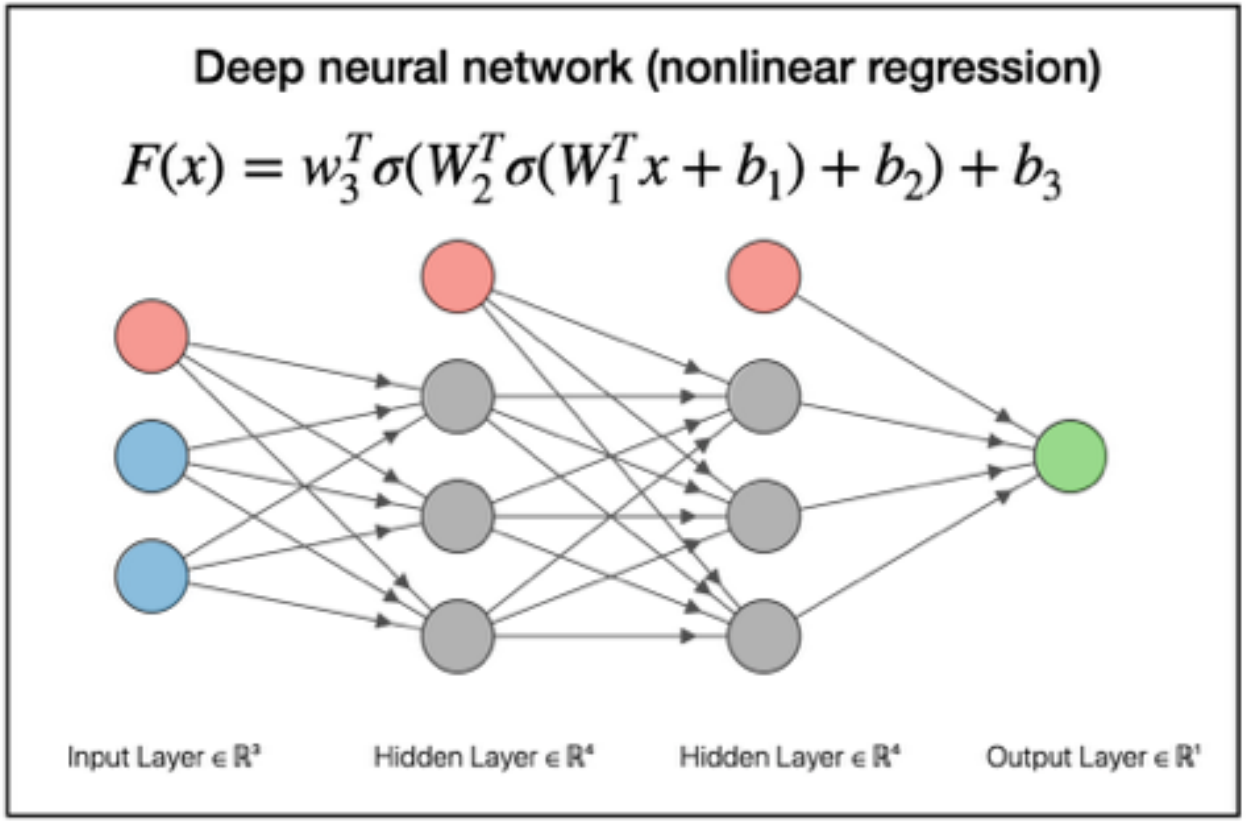


How a Neural Network Works



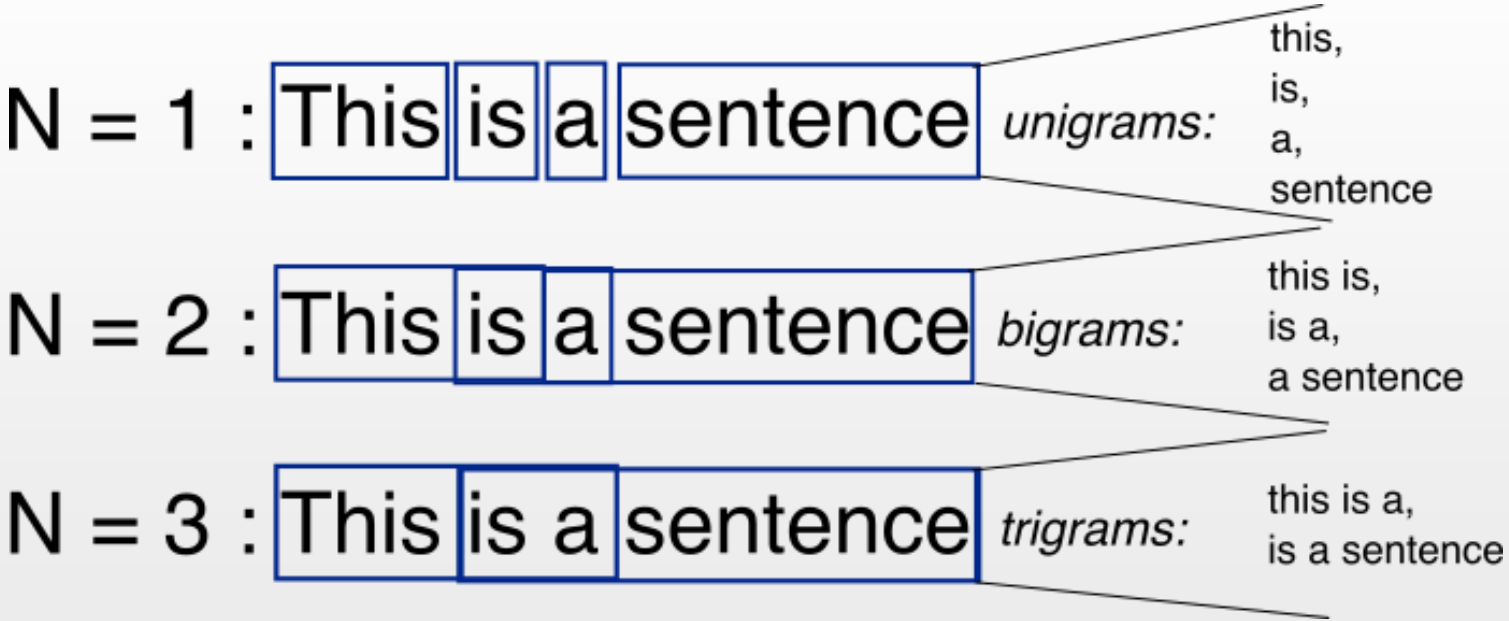
Every time we pass one function into another function, the model is able to fit the data more completely, including nonlinear trends

How a Neural Network Works

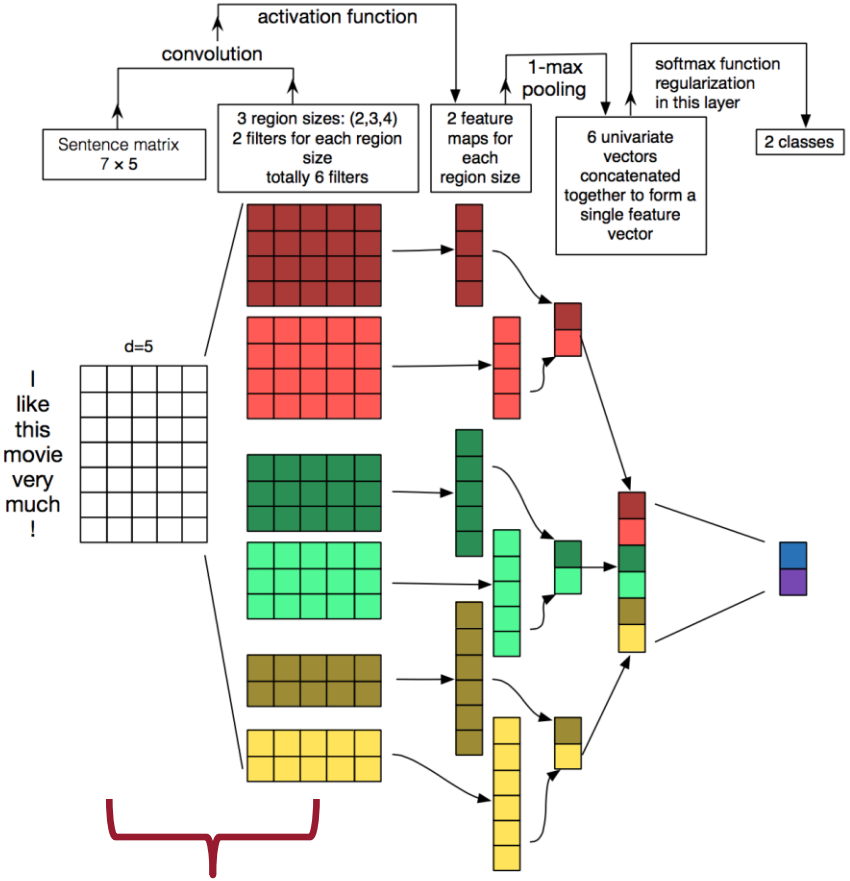


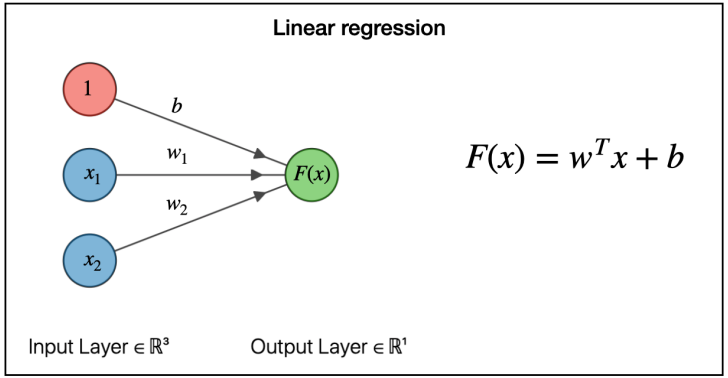
Every time we pass one function into another function, the model is able to fit the data more completely, including nonlinear trends

Interpreting Sentences



Interpreting Sentences

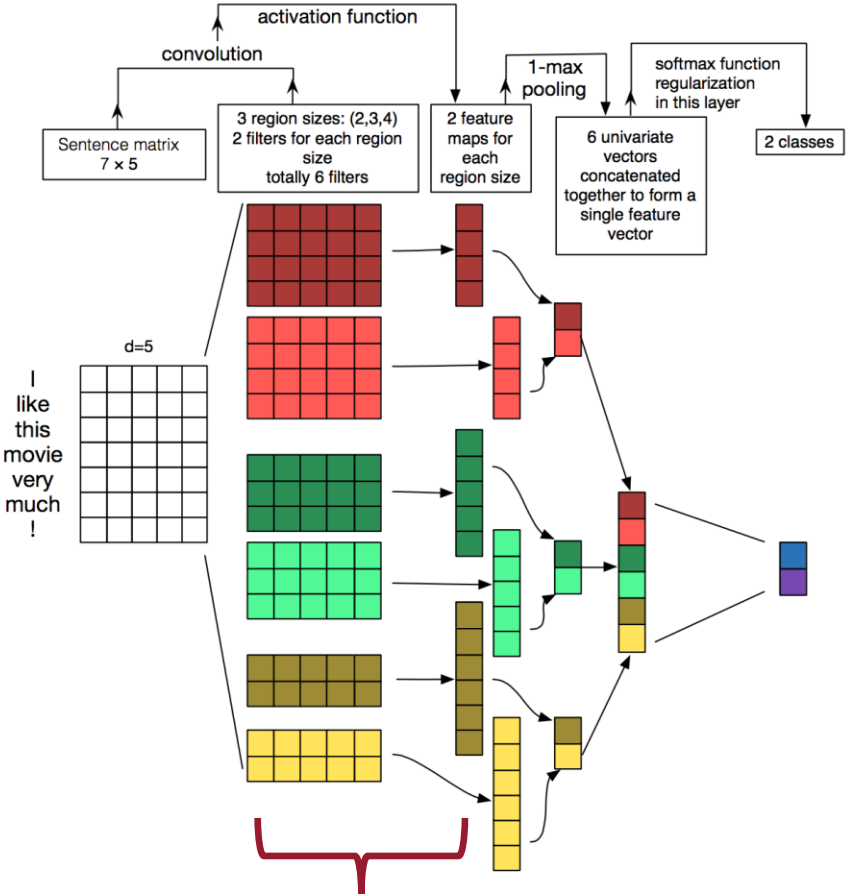


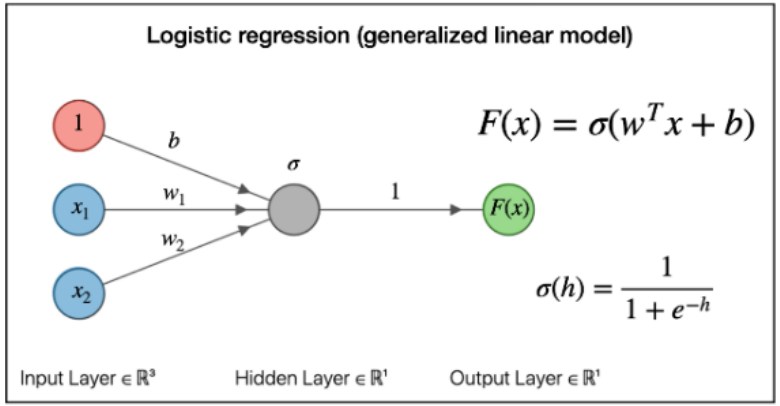


$$F(x) = w^T x + b$$

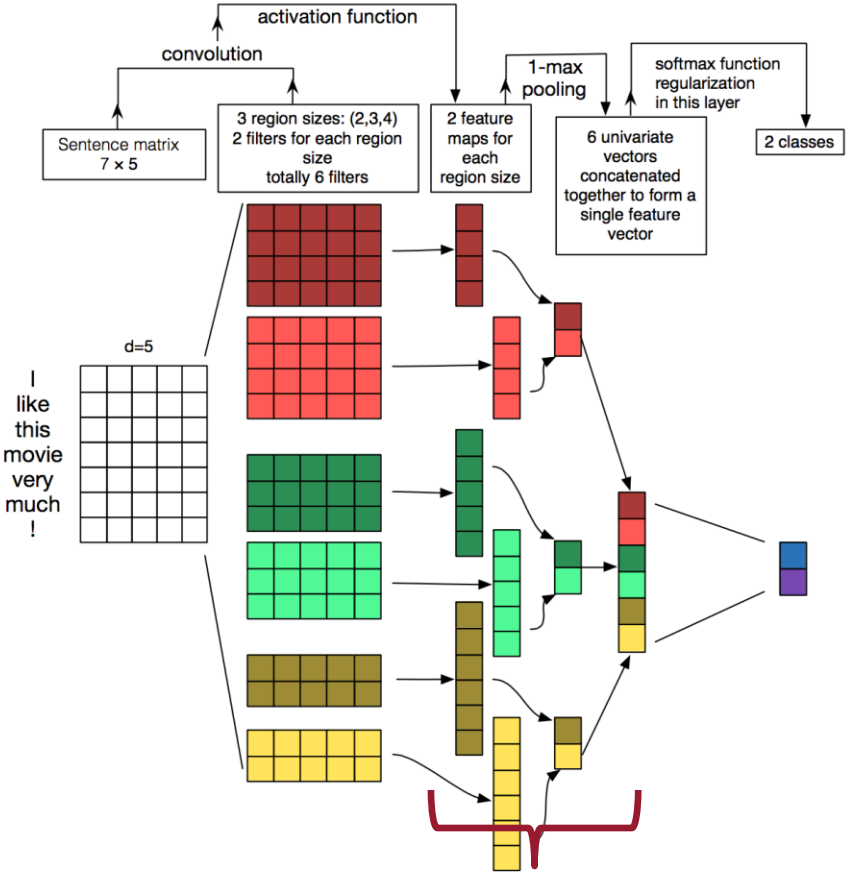
$$F(x) = w_1 \cdot x_1 + w_2 \cdot x_2 + 1 \cdot b$$

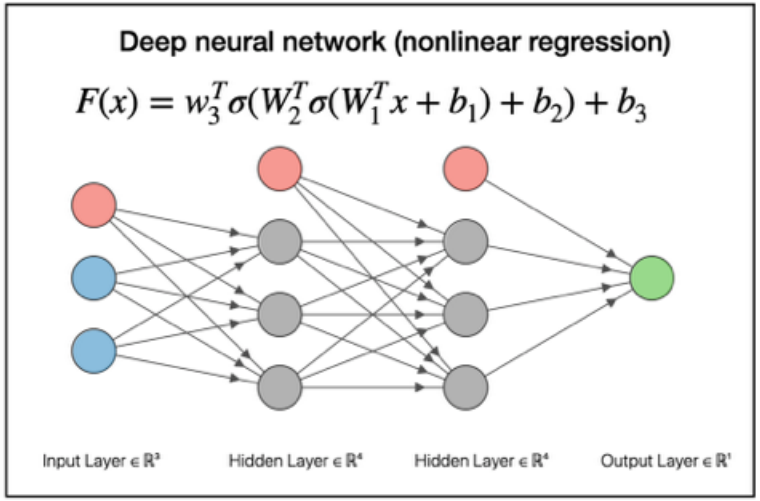
Interpreting Sentences





Interpreting Sentences

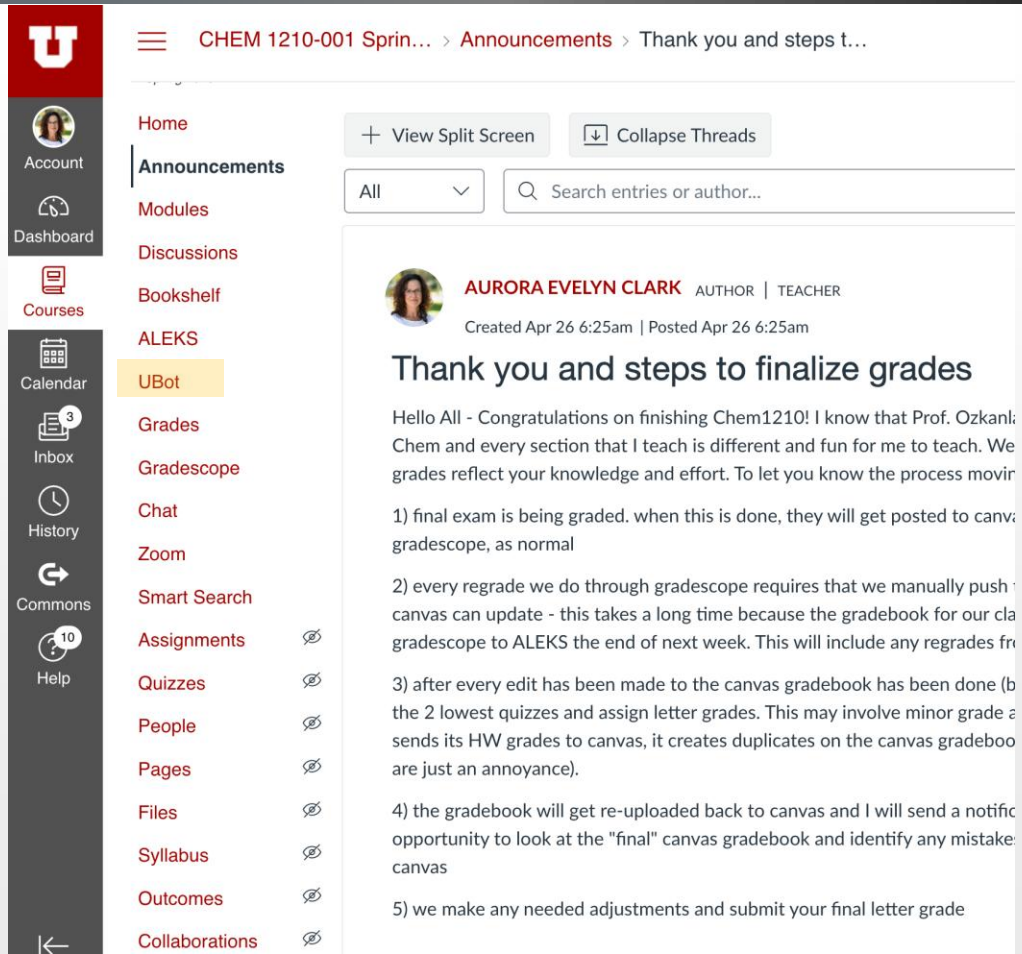




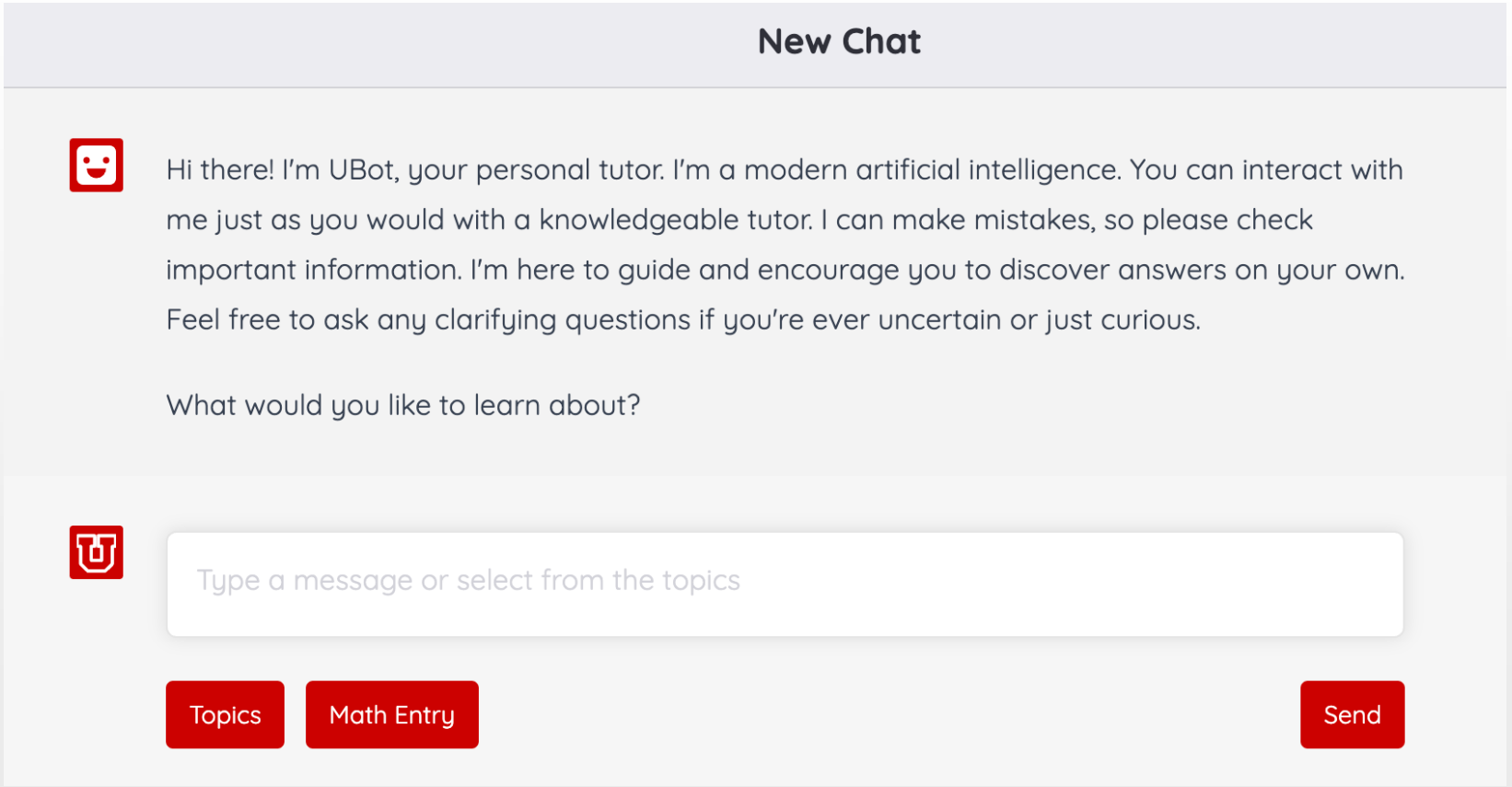
ChatGPT vs. Tools to Help Students Learn

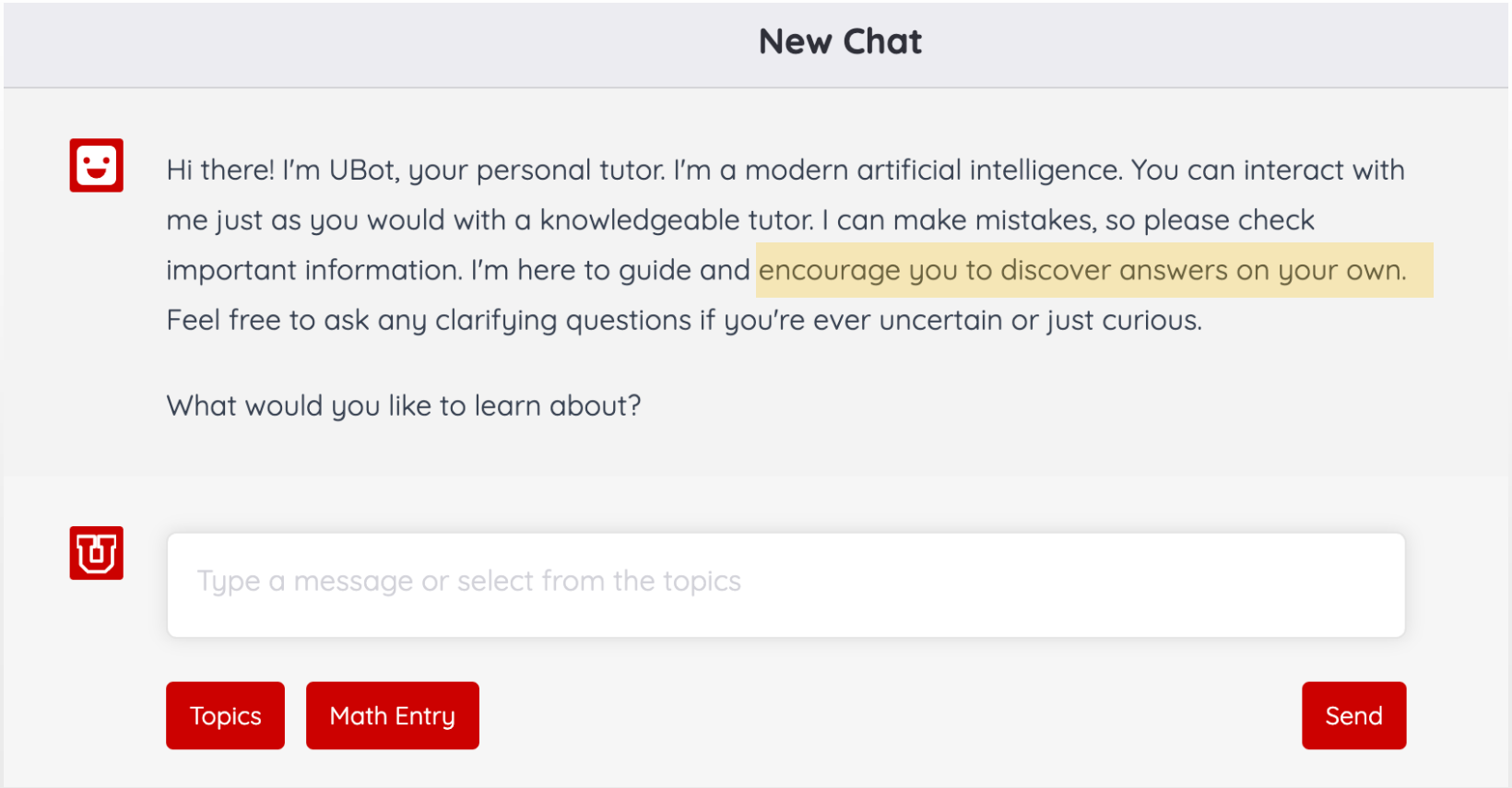
- ChatGPT: Complex nonlinear functions that associate words and word patterns –
  - train the model on lots and lots of text
  - when you input text and ask it to rewrite – the ChatGPT model rewrites it based upon its observed patterns
- Student Learning Tools: complex nonlinear functions that relate concepts, text and student performance
  - trained on text, associates text with science concepts
  - trained on student questions about concepts and student performance data
- Example - UBot

UBot – a tool for Chem1210 (Gen Chem)



UBot – a tool for Chem1210 (Gen Chem)



UBot – a tool for Chem1210 (Gen Chem)

Meant to be a tutor – not give answers!