No Free Lunches in Machine Learning

Neill Campbell

February 2024

Department of Computer Science, University of Bath

Slide input credits: Carl Henrik Ek, Javier González, Simon Prince, Julian Faraway

AI Talks: AI & ML Research Group, Department of Computer Science

11 Oct 2023	Prof Simon Prince	Understanding Deep Learning: The Technology Behind Modern AI
15 Nov 2023	Prof Nello Cristianini	The Shortcut: How Machines Became Intelligent Without Thinking in a Human Way
13 Dec 2023	Prof Mike Tipping	The Irresistible Rise of Machine Learning
28 Feb 2024	Prof Neill Campbell	No Free Lunches in Machine Learning
20 Mar 2024	Prof Özgür Şimşek	Reinforcement Learning and the Pursuit of Artificial Intelligence
17 Apr 2024	Dr Harish Tayyar Madabushi	Emergent Abilities of Language Models: Do they pose an existential threat?
8 May 2024	Prof Darren Cosker	AI for Human Sensing: Research, Productisation and Ethics
TBD	Prof Mike Tipping	Bayesian Inference in Machine Learning: Indistinguishable from Magic?


THIS IS THE WAY
makeameme.org

Overview

Common Questions?

What questions do we have about ML?

- Can I use ML to solve x?

Common Questions?

What questions do we have about ML?

- Can I use ML to solve x?
- What does ML actually do?

Common Questions?

What questions do we have about ML?

- Can I use ML to solve x?
- What does ML actually do?
- Isn't ML just the same as y?

Common Questions?

What questions do we have about ML?

- Can I use ML to solve x?
- What does ML actually do?
- Isn't ML just the same as y?
- Can I replace myself/my research team with ML?

Common Questions?

What questions do we have about ML?

- Can I use ML to solve x?
- What does ML actually do?
- Isn't ML just the same as y?
- Can I replace myself/my research team with ML?
- How much data do I need?

Common Questions?

What questions do we have about ML?

- Can I use ML to solve x?
- What does ML actually do?
- Isn't ML just the same as y?
- Can I replace myself/my research team with ML?
- How much data do I need?
- Can I just use Deep Learning/Generative AI/ChatGPT?

Common Questions?

What questions do we have about ML?

- Can I use ML to solve x?
- What does ML actually do?
- Isn't ML just the same as y?
- Can I replace myself/my research team with ML?
- How much data do I need?
- Can I just use Deep Learning/Generative AI/ChatGPT?
- Surely Deep Learning/Generative AI/ChatGPT is all hype?

Common Questions?

What questions do we have about ML?

- Can I use ML to solve x?
- What does ML actually do?
- Isn't ML just the same as y?
- Can I replace myself/my research team with ML?
- How much data do I need?
- Can I just use Deep Learning/Generative AI/ChatGPT?
- Surely Deep Learning/Generative AI/ChatGPT is all hype?
- Can any of this be used for science/engineering?

Things to consider..

- Are all datasets equal or how to choose your data?

Things to consider..

- Are all datasets equal or how to choose your data?
- Gotchas: What Machine Learning can and can’t do for you.

Things to consider..

- Are all datasets equal or how to choose your data?
- Gotchas: What Machine Learning can and can't do for you.
- Average vs worst case Machine Learning

Things to consider..

- Are all datasets equal or how to choose your data?
- Gotchas: What Machine Learning can and can't do for you.
- Average vs worst case Machine Learning
- Machine Learning and Causality

Things to consider..

- Are all datasets equal or how to choose your data?
- Gotchas: What Machine Learning can and can't do for you.
- Average vs worst case Machine Learning
- Machine Learning and Causality
- Trade-offs in Machine Learning

Things to consider..

- Are all datasets equal or how to choose your data?
- Gotchas: What Machine Learning can and can't do for you.
- Average vs worst case Machine Learning
- Machine Learning and Causality
- Trade-offs in Machine Learning

Things to consider..

- Are all datasets equal or how to choose your data?
- Gotchas: What Machine Learning can and can't do for you.
- Average vs worst case Machine Learning
- Machine Learning and Causality
- Trade-offs in Machine Learning

Overview...

Overview

No Free Lunch

Uncertainty / Error Bars

Model Selection

Causality

Conclusions

No Free Lunch

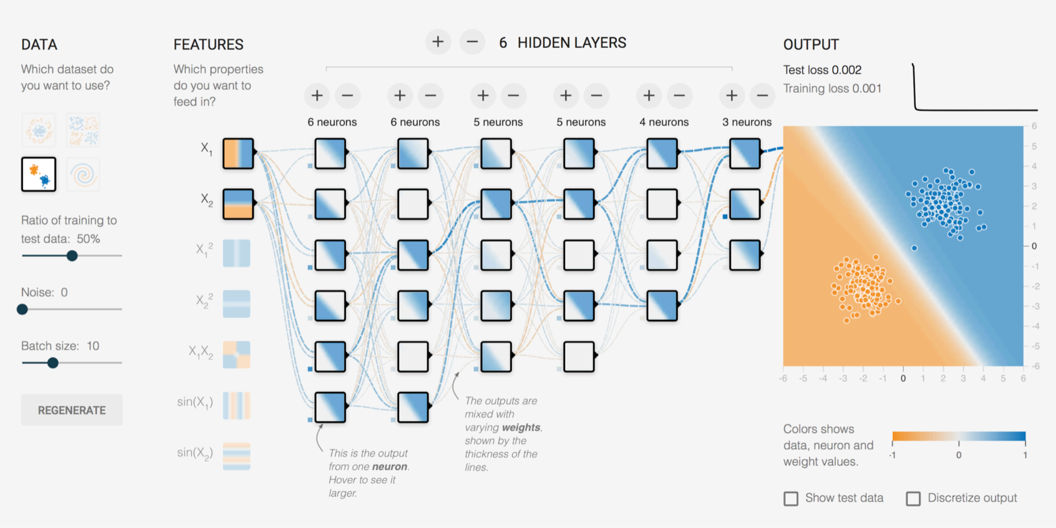
Overview...

Machine Learning illustration

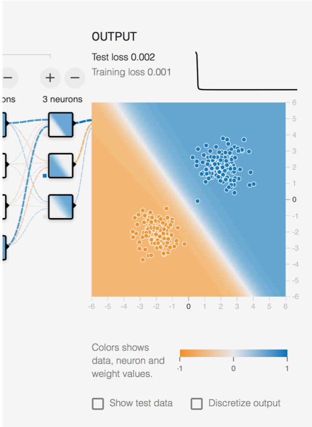
Neural Network Playground:

- https://playground.tensorflow.org/

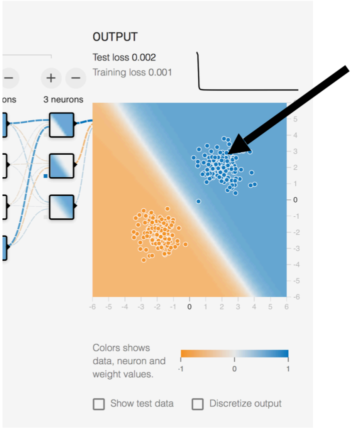
Machine Learning illustration



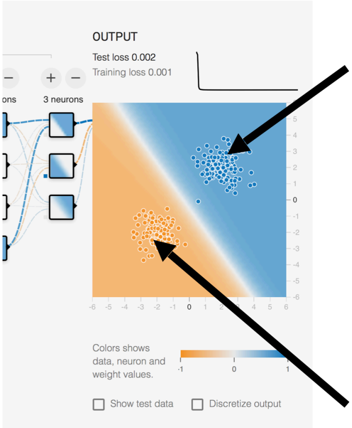
Machine Learning illustration

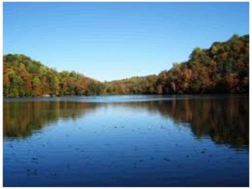


Machine Learning illustration

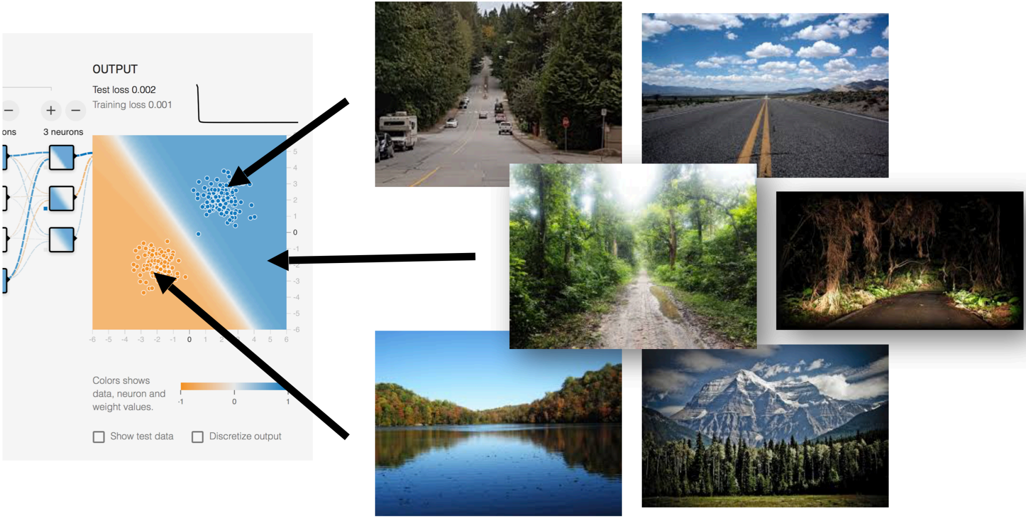


Machine Learning illustration





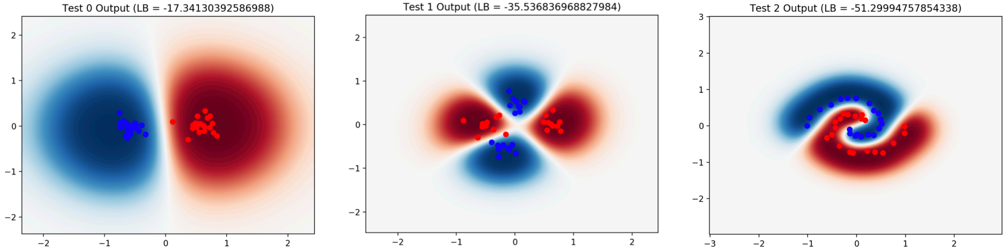
Machine Learning illustration



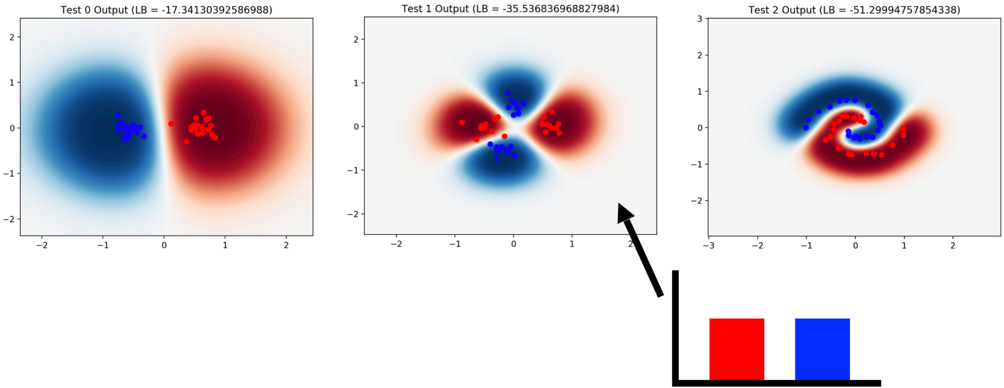
Machine Learning illustration

What if we use a probabilistic approach?

What if we use a probabilistic approach?



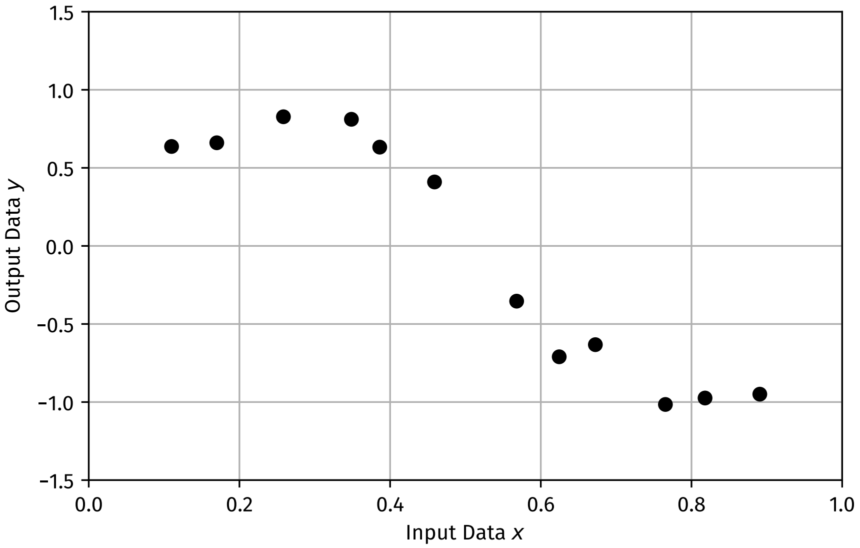
What if we use a probabilistic approach?



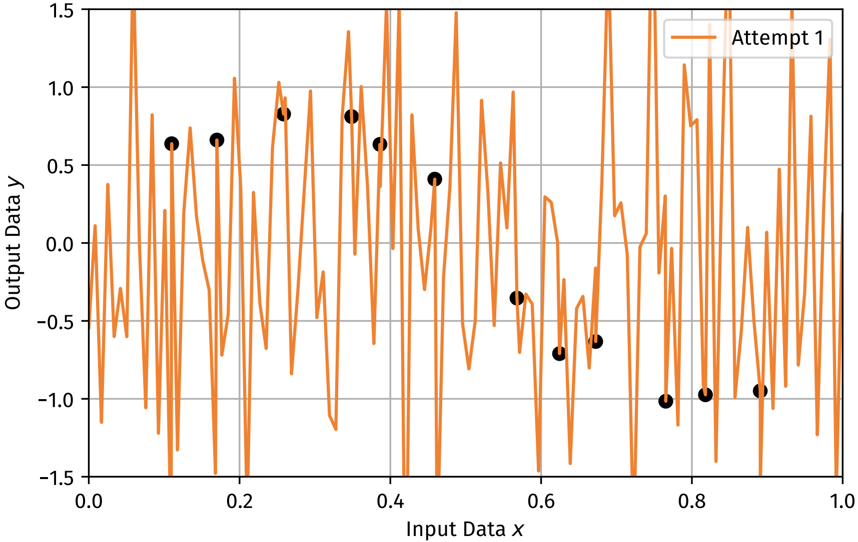
We need to consider properties of Machine Learning approaches

What happens between the dots?

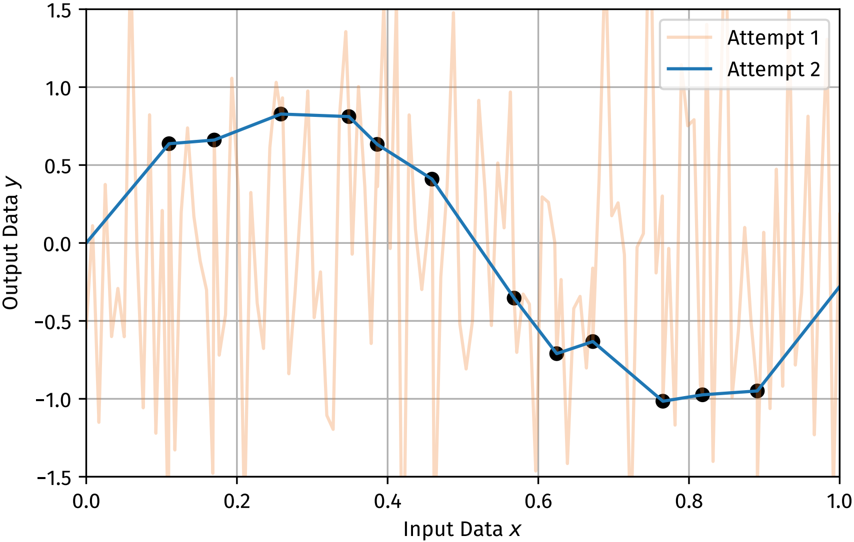
What happens between the dots?



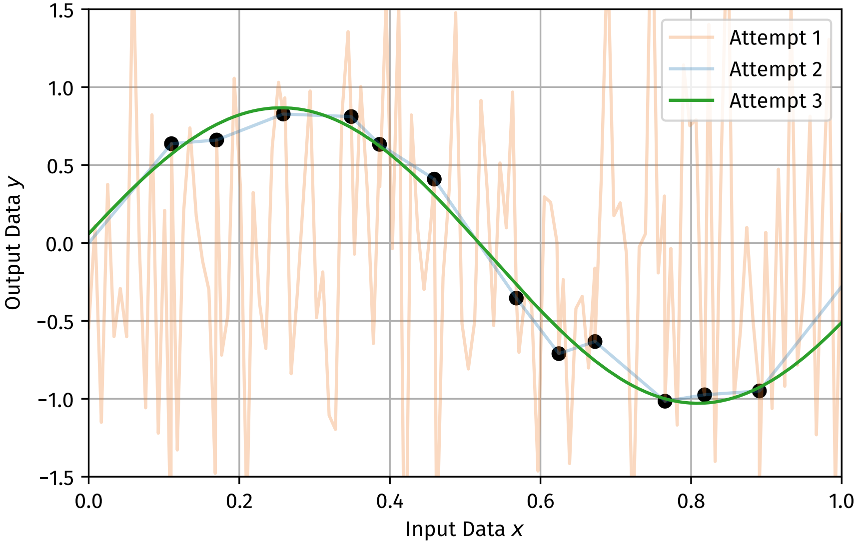
What happens between the dots?



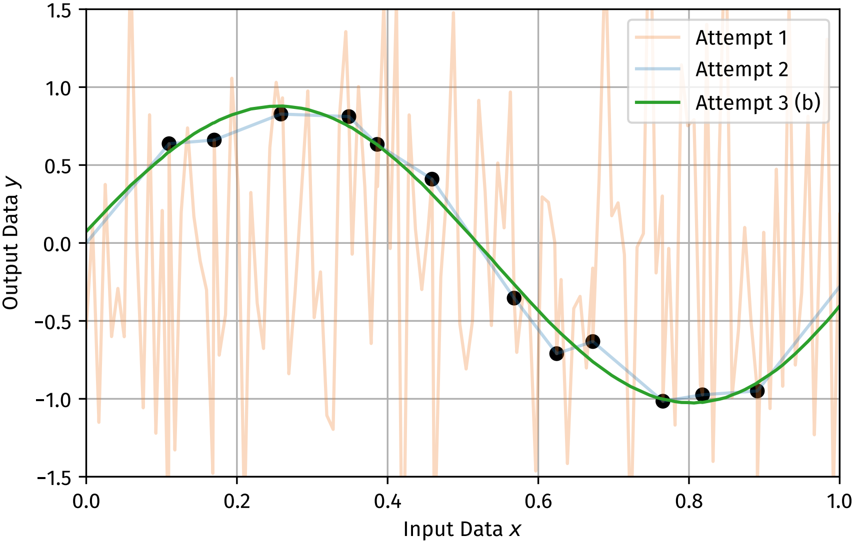
What happens between the dots?



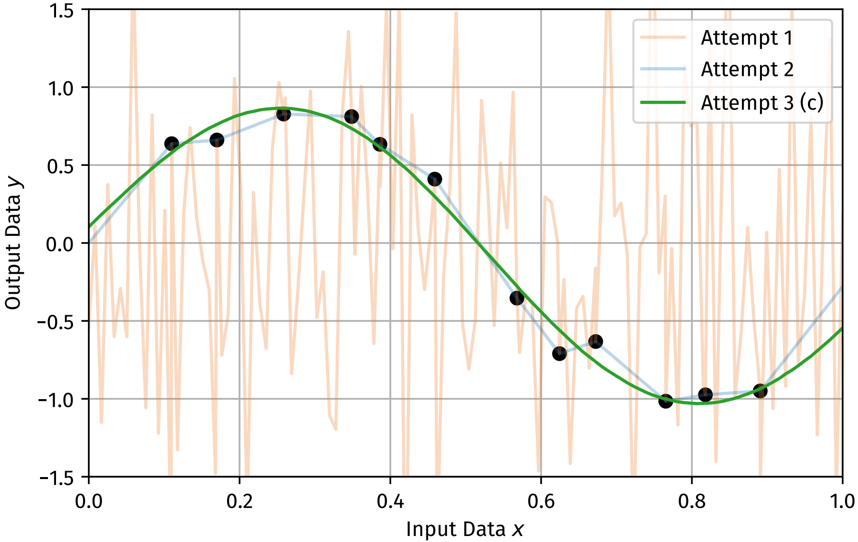
What happens between the dots?



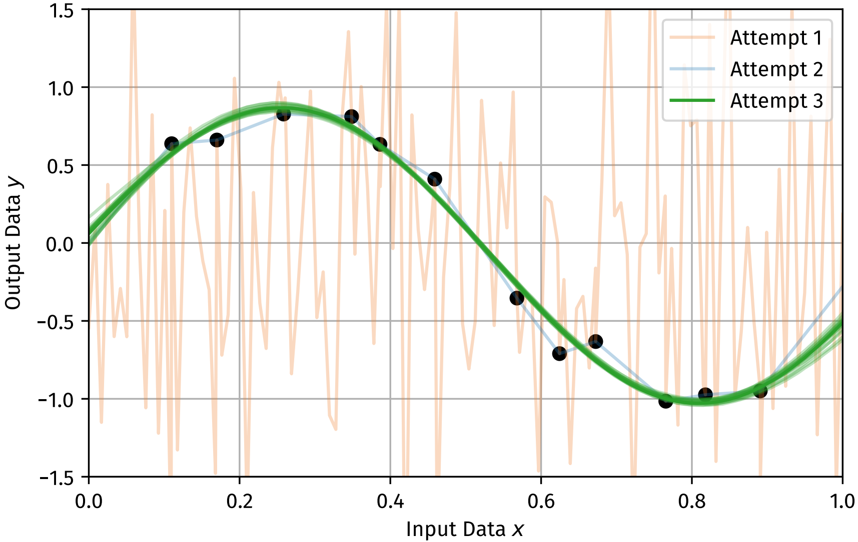
What happens between the dots?



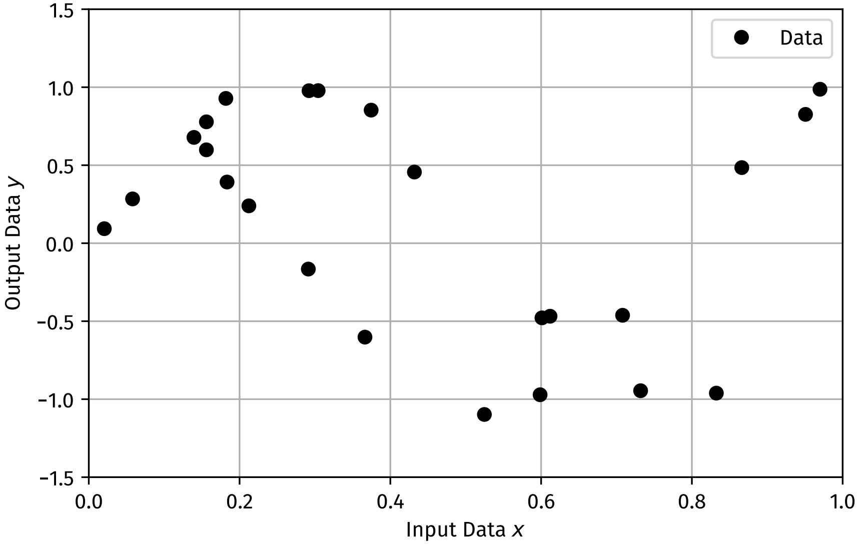
What happens between the dots?



What happens between the dots?

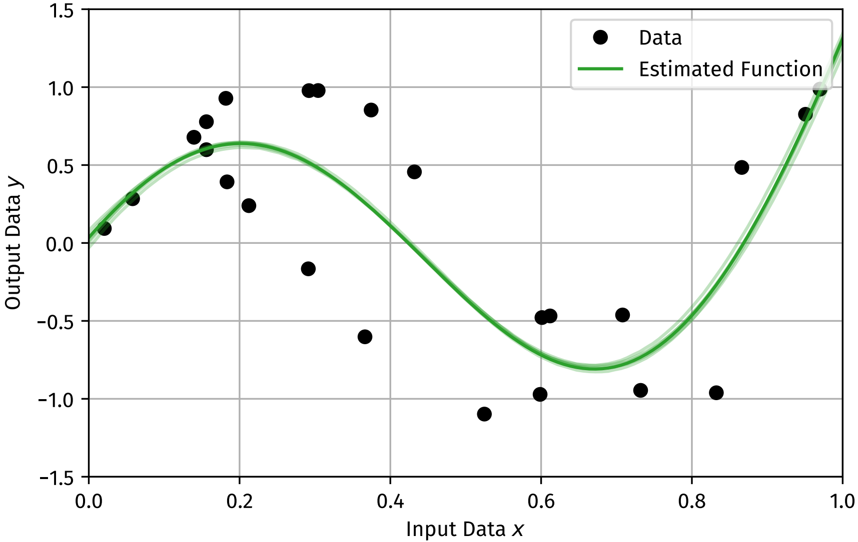


Ambiguity..



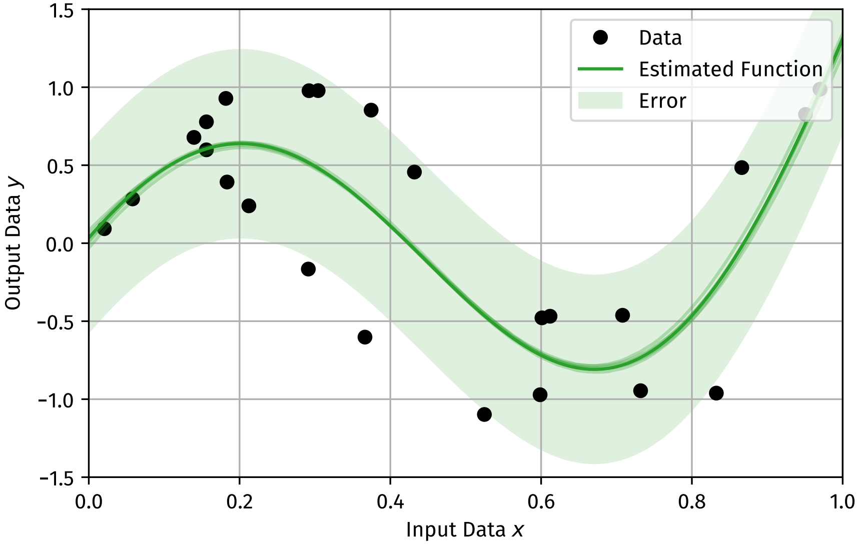
Data
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Output Data y
0.0
0.2
0.4
0.6
0.8
1.0
Input Data x

Ambiguity..



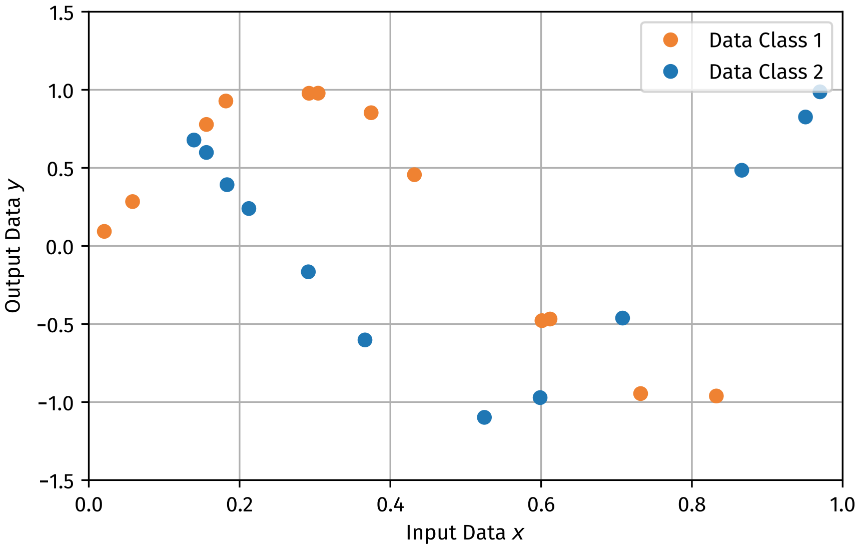
Data
Estimated Function
Output Data y
Input Data x
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0.0
0.2
0.4
0.6
0.8
1.0

Ambiguity..

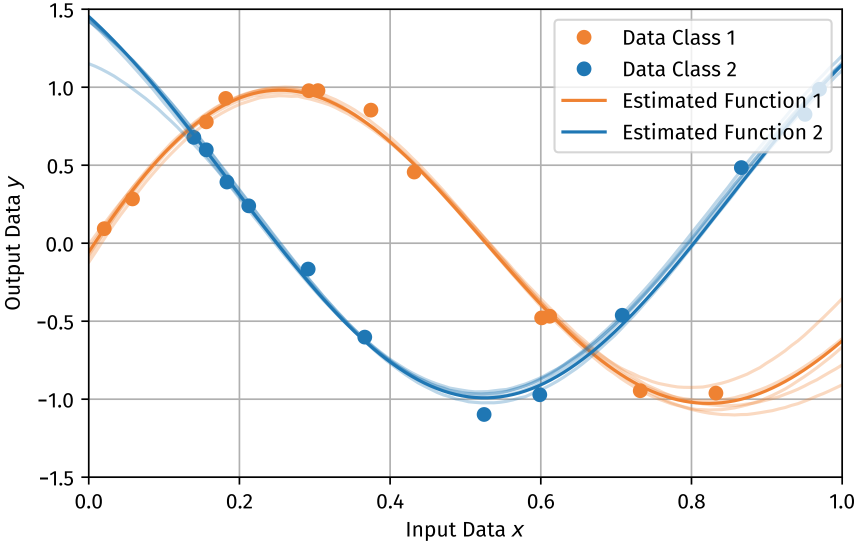


Data
Estimated Function
Error
Output Data y
Input Data x
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0.0
0.2
0.4
0.6
0.8
1.0

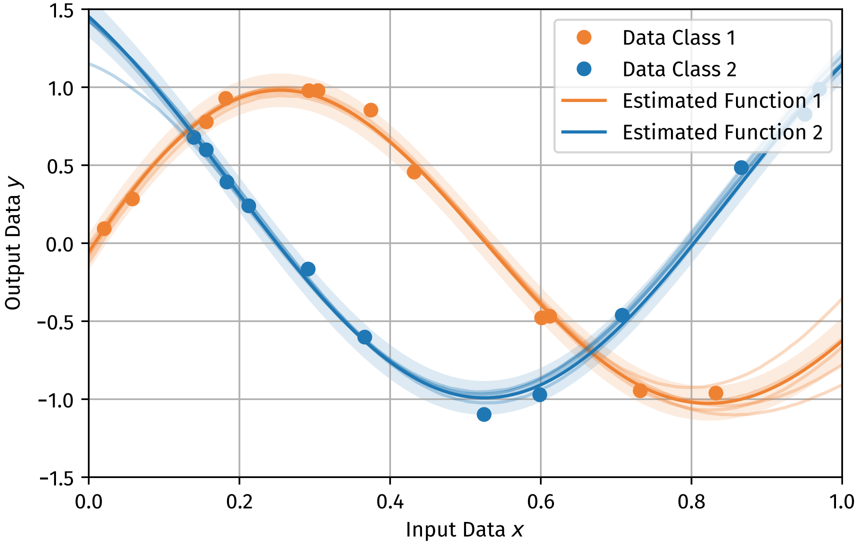
Ambiguity..


Data Class 1
Data Class 2
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0.0
0.2
0.4
0.6
0.8
1.0
Output Data y
Input Data x

Ambiguity..

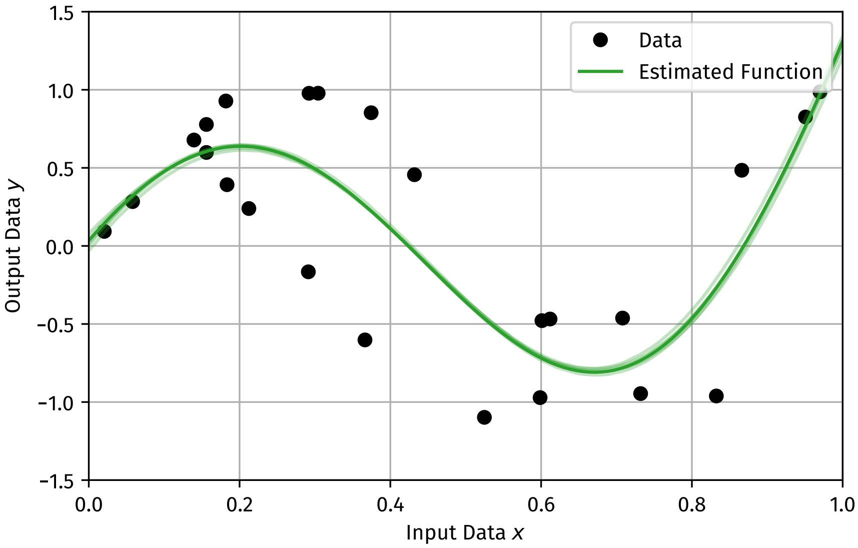

Data Class 1
Data Class 2
Estimated Function 1
Estimated Function 2
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Output Data y
0.0
0.2
0.4
0.6
0.8
1.0
Input Data x

Ambiguity..

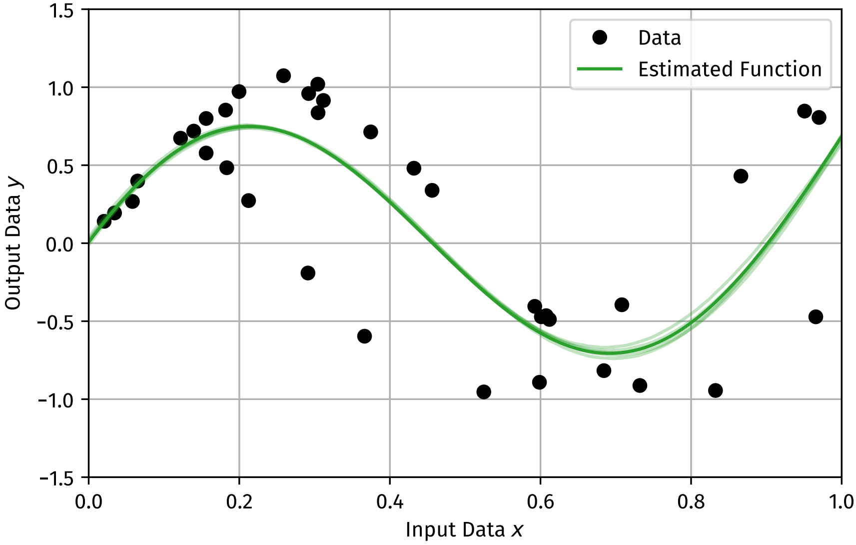

Data Class 1
Data Class 2
Estimated Function 1
Estimated Function 2
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0.0
0.2
0.4
0.6
0.8
1.0
Output Data y
Input Data x

Average vs Worst Case..

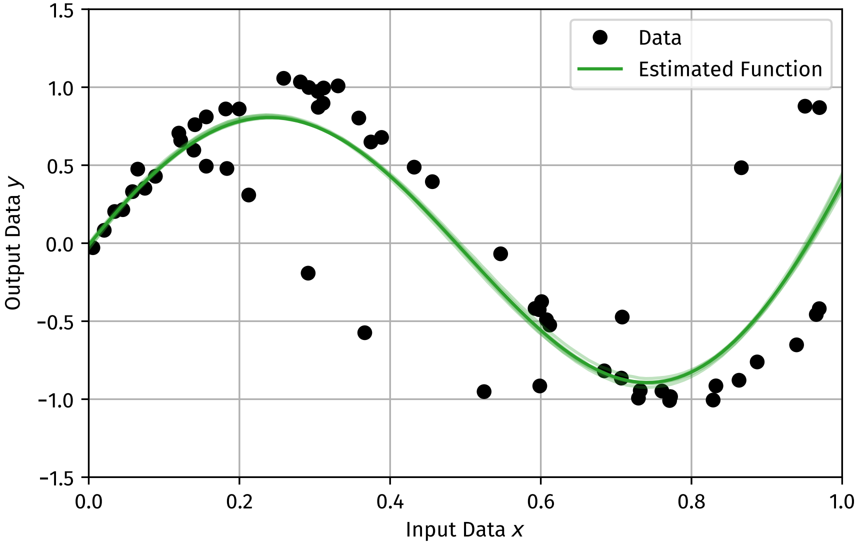
Average vs Worst Case..



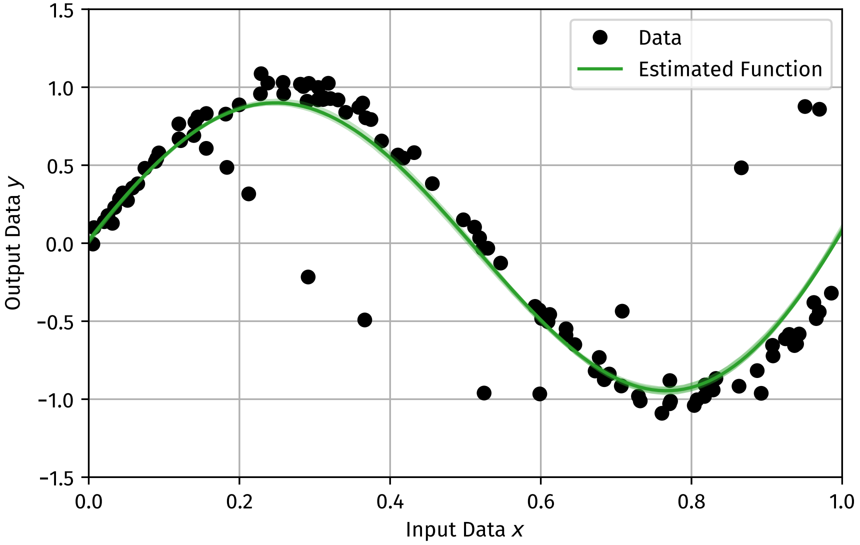
Average vs Worst Case..



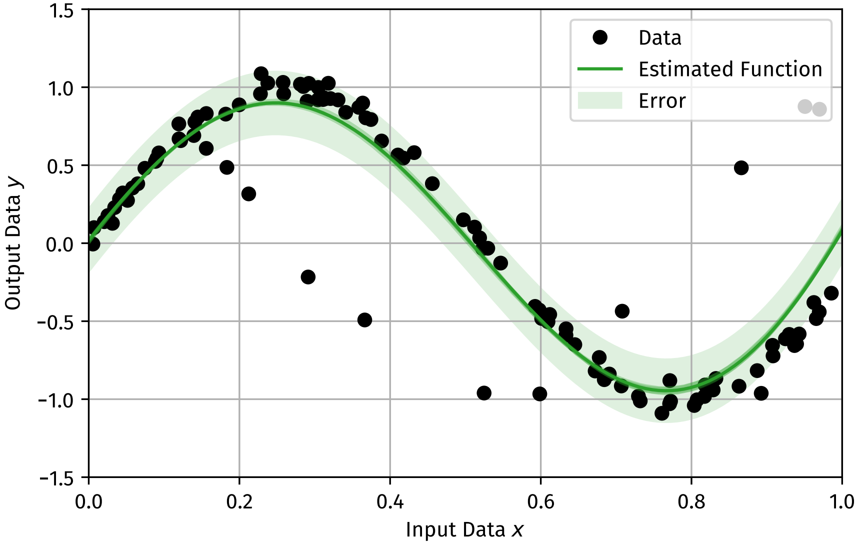
Average vs Worst Case..



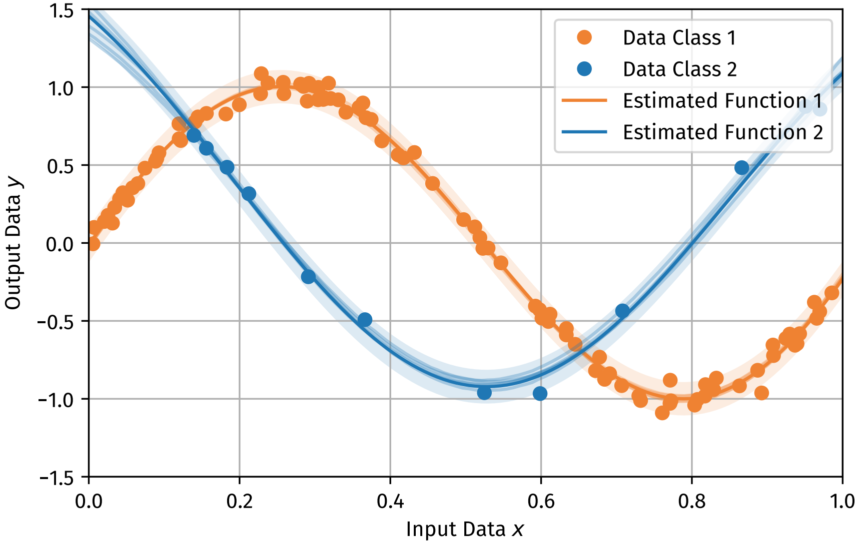
Average vs Worst Case..



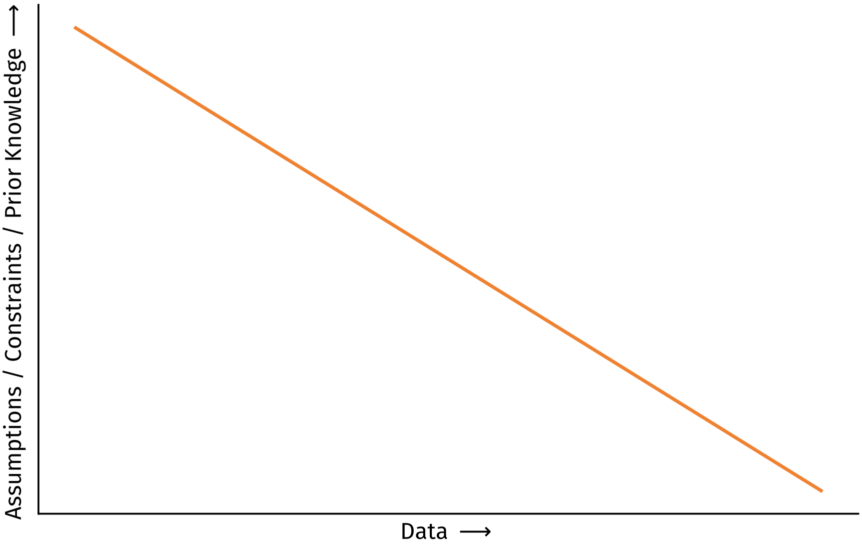
Average vs Worst Case: Failure to model..



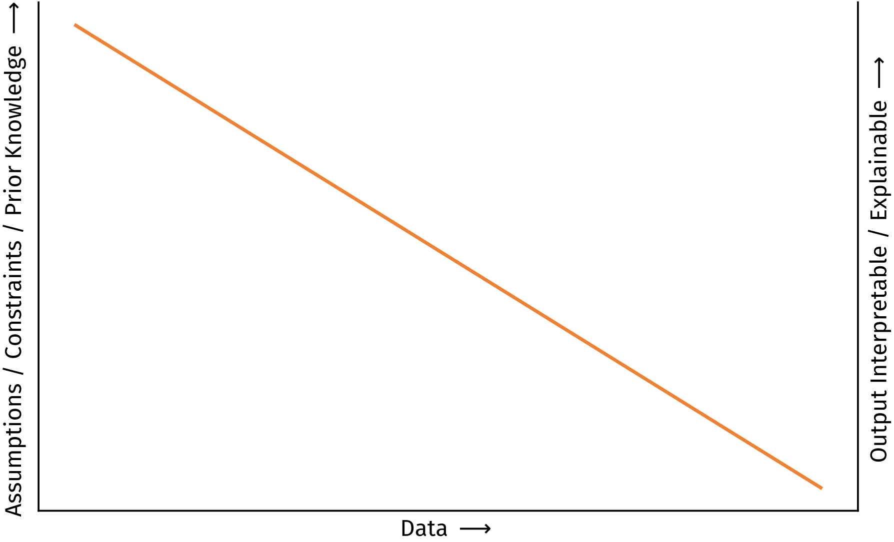
Average vs Worst Case: Explicitly accounting for imbalance..



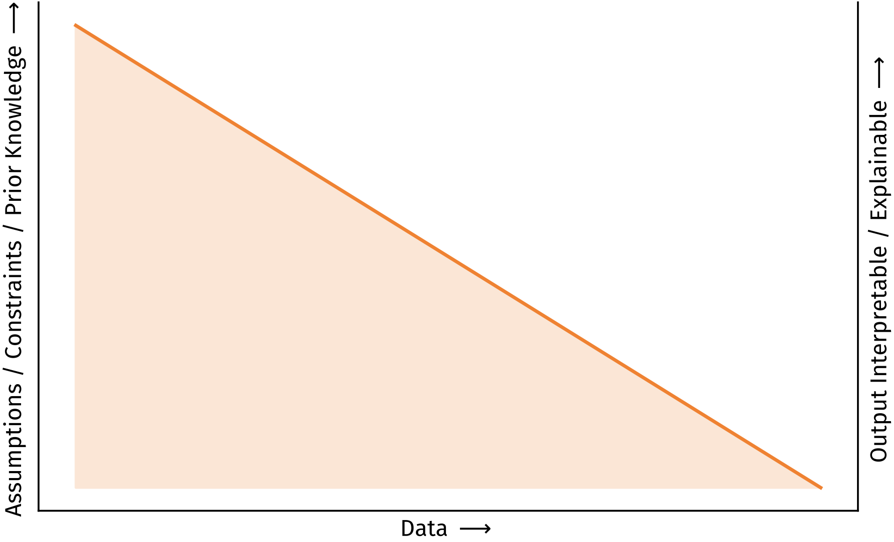
No free lunch



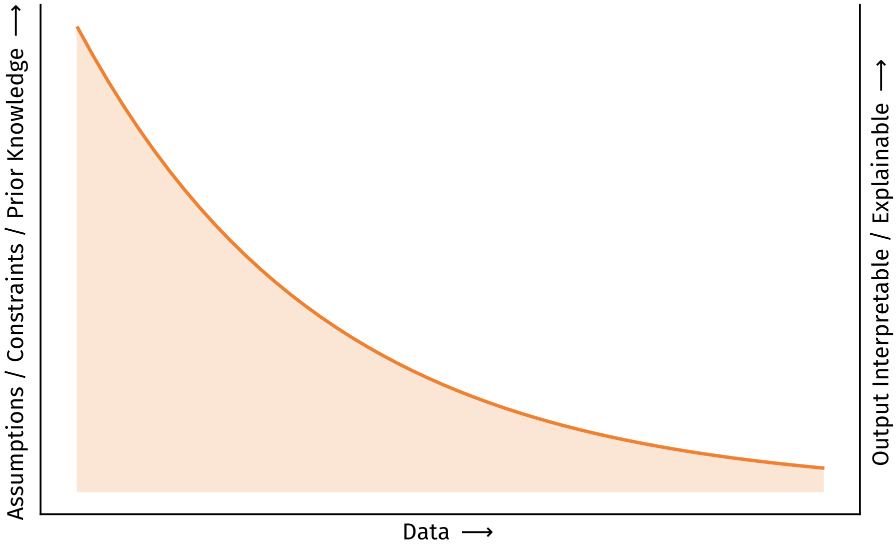
No free lunch



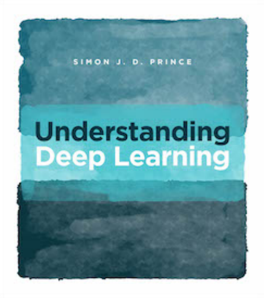
No free lunch



No free lunch (more realistic)



Understanding Deep Learning



Excellent new text book from Simon Prince (visiting Prof in Bath for semester 1):

Understanding Deep Learning, Simon J.D. Prince, MIT Press

Final draft available on the website: `https://udlbook.github.io/udlbook/`

Uncertainty / Error Bars

Overview...

Overview

No Free Lunch

Uncertainty / Error Bars

Model Selection

Causality

Conclusions

- Bayes' Rule

$$\text{Posterior Probability (after)} = \frac{\text{Likelihood (of event)} \times \text{Prior Probability (before)}}{\text{Evidence}}$$

Example of Bayes' Rule..

- Consider a legal trial..

$$\underbrace{p(\text{guilt} \mid \text{observations})}_{\text{Posterior}} = \frac{\overbrace{p(\text{observations} \mid \text{guilt})}^{\text{Likelihood}} \times \overbrace{p(\text{guilt})}^{\text{Prior}}}{\underbrace{p(\text{observations})}_{\text{Evidence}}}$$

$p(A \mid B)$ means "probability of A being the case given that B occurs"

Example of Bayes' Rule..

- Consider a legal trial..

$$\underbrace{p(\text{guilt} \mid \text{observations})}_{\text{Posterior}} = \underbrace{\frac{\overbrace{p(\text{observations} \mid \text{guilt})}^{\text{Likelihood}} \times \overbrace{p(\text{guilt})}^{\text{Prior}}}{\sum_{\text{guilt}} p(\text{observations} \mid \text{guilt})\, p(\text{guilt})}}_{\text{Evidence}}$$

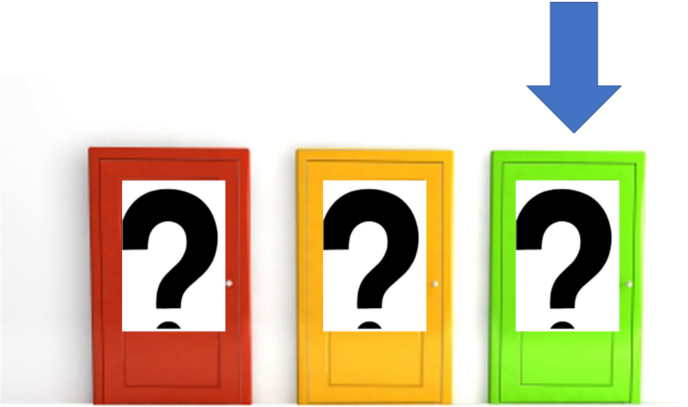
$p(A \mid B)$ means "probability of A being the case given that B occurs"

Monty Hall..

Monty Hall..

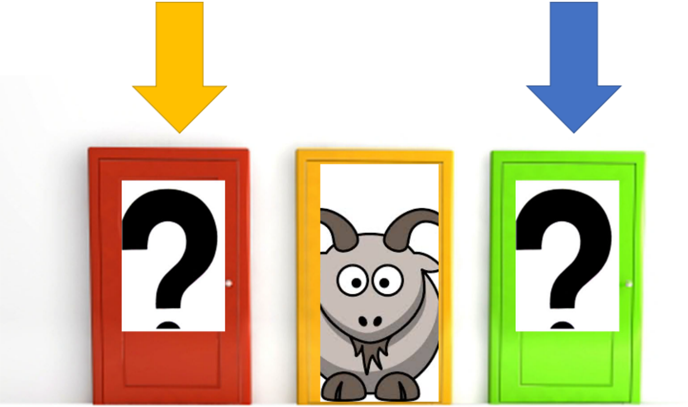



Monty Hall..



Monty Hall..

Monty Hall..



Monty Hall: How would we generate data (or simulate)?

Monty Hall: How would we generate data (or simulate)?

```
door_with_car = pick_random({1, 2, 3})
door_with_goat = {1, 2, 3} - door_with_car

door_picked = pick_random({1, 2, 3})

if door_picked == door_with_car:
    door_to_open = pick_random(door_with_goat)
else:
    door_to_open = door_with_goat - door_picked
```

Monty Hall: How would we generate data (or simulate)?

```
door_with_car = pick_random({1, 2, 3})          # 1/3 equal chance
door_with_goat = {1, 2, 3} - door_with_car

door_picked = pick_random({1, 2, 3})            # 1/3 equal chance

if door_picked == door_with_car:
    door_to_open = pick_random(door_with_goat)  # 1 times in 3
else:
    door_to_open = door_with_goat - door_picked # 2 times in 3
```

Consider Modelling and ML as a

Generative Process

Bayes' Rule with models and functions..

$$\underbrace{p(\text{functions} \mid \text{observed data})}_{\text{Posterior}} = \frac{\overbrace{p(\text{observed data} \mid \text{functions})}^{\text{Likelihood}} \times \overbrace{p(\text{functions})}^{\text{Prior}}}{\underbrace{p(\text{observed data})}_{\text{Evidence}}}$$

Bayes' Rule with models and functions..

$$\underbrace{p(\text{functions} \mid \text{observed data})}_{\text{Posterior}} = \frac{\overbrace{p(\text{observed data} \mid \text{functions})}^{\text{Likelihood}} \times \overbrace{p(\text{functions})}^{\text{Prior}}}{\underbrace{p(\text{observed data})}_{\text{Evidence}}}$$

$$\underbrace{p(f \mid \mathcal{D})}_{\text{Posterior}} = \frac{\overbrace{p(\mathcal{D} \mid f)}^{\text{Likelihood}} \times \overbrace{p(f)}^{\text{Prior}}}{\underbrace{p(\mathcal{D})}_{\text{Evidence}}} \qquad p(\mathcal{D}) = \sum_f p(\mathcal{D} \mid f)\, p(f)$$

Data $\mathcal{D} = \{X, Y\}$, pairs of inputs $\{x_n\}$ and outputs $\{y_n\}$, and functions f

Bayes' Rule with models and functions..

$$\underbrace{p(\text{functions} \mid \text{observed data})}_{\text{Posterior}} = \frac{\overbrace{p(\text{observed data} \mid \text{functions})}^{\text{Likelihood}} \times \overbrace{p(\text{functions})}^{\text{Prior}}}{\underbrace{p(\text{observed data})}_{\text{Evidence}}}$$

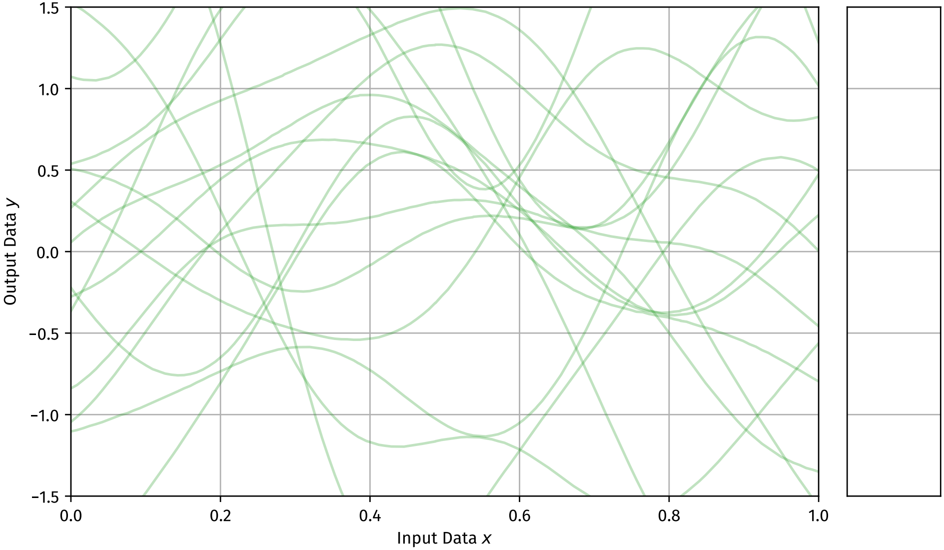
$$\underbrace{p(f \mid \mathcal{D})}_{\text{Posterior}} = \frac{\overbrace{p(\mathcal{D} \mid f)}^{\text{Likelihood}} \times \overbrace{p(f)}^{\text{Prior}}}{\underbrace{p(\mathcal{D})}_{\text{Evidence}}} \qquad p(\mathcal{D}) = \sum_f p(\mathcal{D} \mid f)\, p(f)$$

Data $\mathcal{D} = \{X, Y\}$, pairs of inputs $\{x_n\}$ and outputs $\{y_n\}$, and functions f

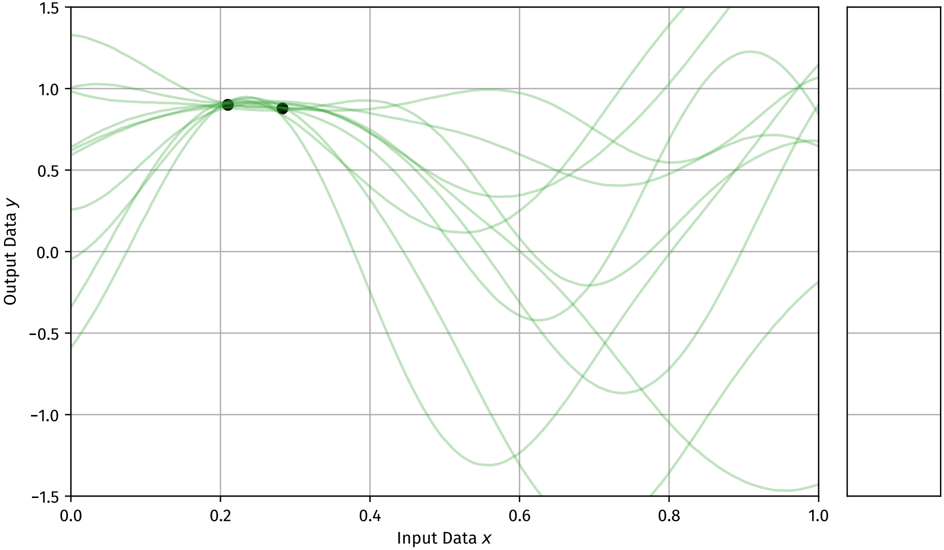
Average over functions to predict unknown output y^* for a **new input** x^*:

$$p(y^* \mid x^*, \mathcal{D}) = \sum_f p(y^* \mid x^*, f)\, p(f \mid \mathcal{D})$$

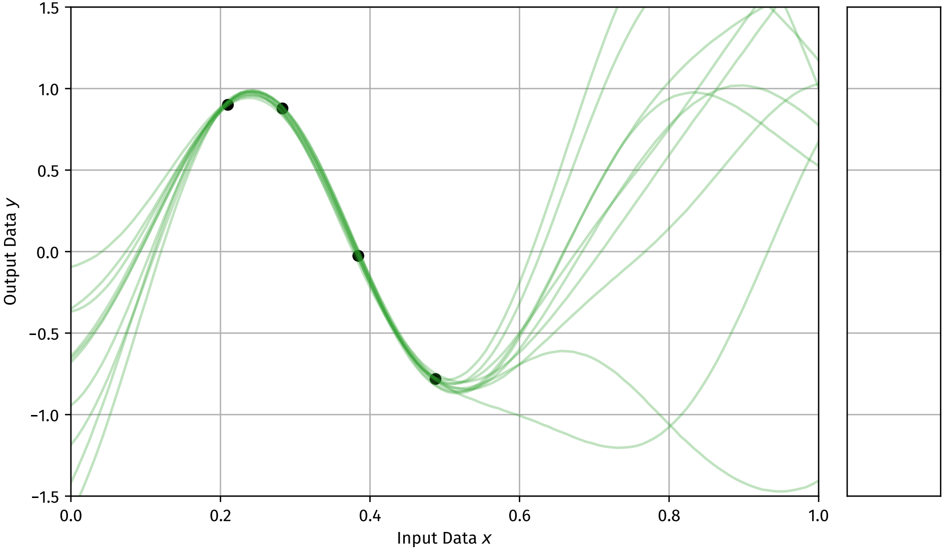
Prior over functions...



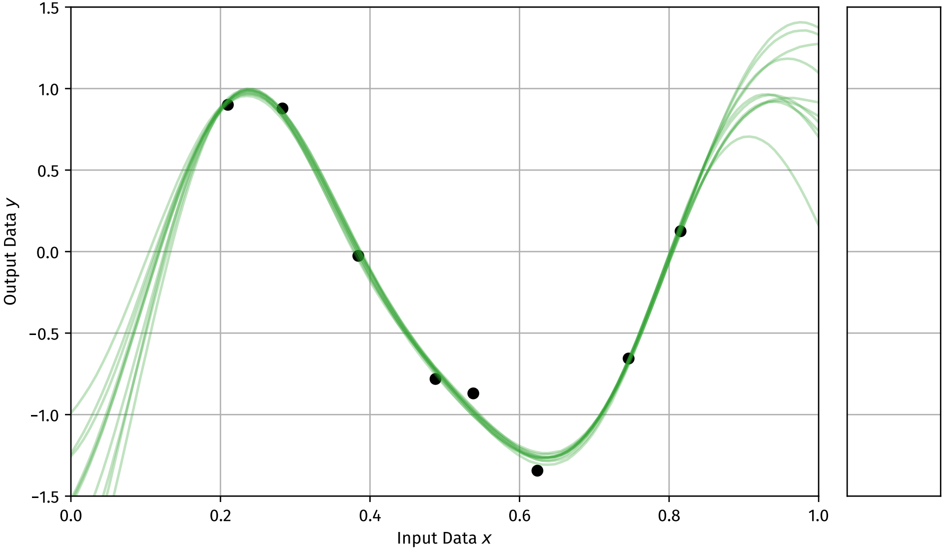
Combine prior with data...



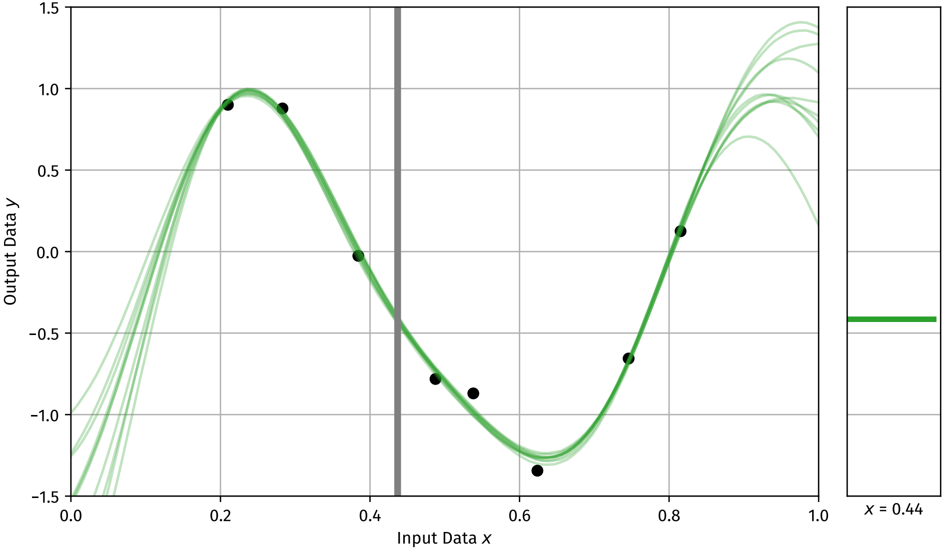
Combine prior with data...



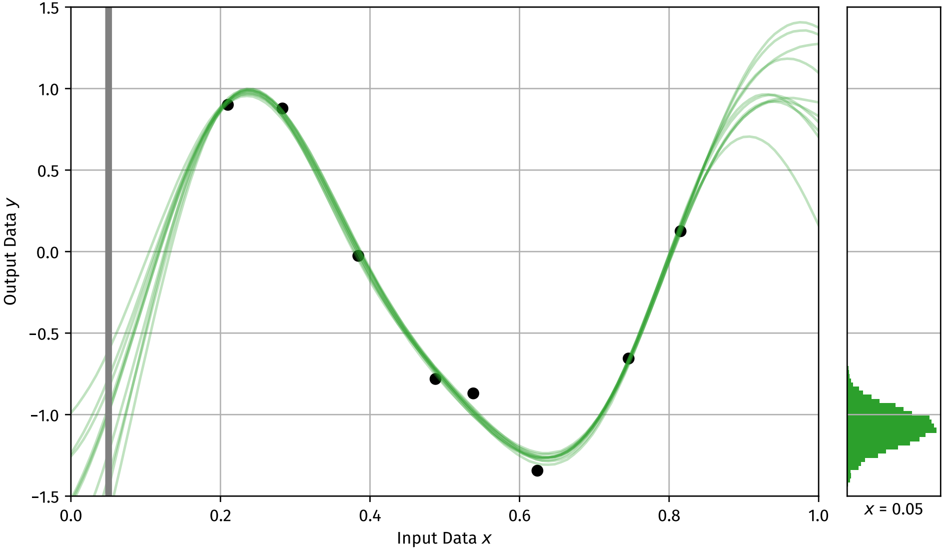
Combine prior with data...



Average over functions to predict...



Averaging over functions gives us (Epistemic) Uncertainty!



Model Selection

- Overview
- No Free Lunch
- Uncertainty / Error Bars
- Model Selection
- Causality
- Conclusions

Model selection

- How much data do we need?

Model selection

- How much data do we need?
- Might not be the right question..

Model selection

- How much data do we need?
- Might not be the right question..
- What can we actually say?

Model selection

- How much data do we need?
- Might not be the right question..
- What can we actually say? The odds!

Model selection

- How much data do we need?
- Might not be the right question..
- What can we actually say? The odds!

Science (and Machine Learning) cannot prove things to be true via data

Science (and Machine Learning) cannot prove things to be true via data

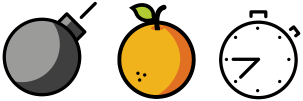
we can only demonstrate that things are inconsistent with data

Model selection illustration: Gravity!

Stable Diffusion: *“Drop cannonball and orange off the leaning tower of Pisa.”*

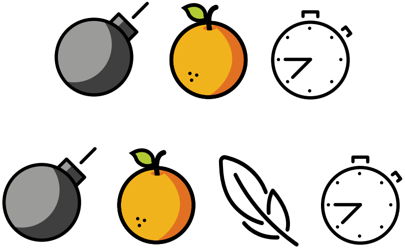
Model selection illustration: Gravity!

Stable Diffusion: *“Drop cannonball and orange off the leaning tower of Pisa.”*



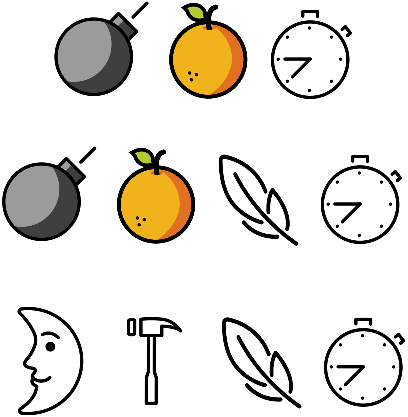
Model selection illustration: Gravity!

Stable Diffusion: *"Drop cannonball and orange off the leaning tower of Pisa."*



Model selection illustration: Gravity!

Stable Diffusion: *“Drop cannonball and orange off the leaning tower of Pisa.”*



Model selection illustration: Gravity!

Stable Diffusion: *“Drop cannonball and orange off the leaning tower of Pisa.”*

Apollo 15 Hammer-Feather Drop

NASASolarSystem
14.9K subscribers

Subscribe

4.8K

Share

576K views 8 years ago
At the end of the last Apollo 15 moon walk, Commander David Scott (pictured above) performed a live demonstration for the television cameras. He held out a geologic hammer and a feather and dropped them at the same time. Because they were essentially in a vacuum, there w ...more

Bayes' Rule for model selection..

$$\underbrace{p(w \mid \mathcal{D})}_{\text{Posterior}} = \frac{\overbrace{p(\mathcal{D} \mid \mathrm{w})}^{\text{Likelihood}} \times \overbrace{p(\mathrm{w})}^{\text{Prior}}}{\underbrace{p(\mathcal{D})}_{\text{Evidence}}}$$

Data $\mathcal{D} = \{X, Y\}$, input/output pairs, and parameters w

Bayes' Rule for model selection..

$$\underbrace{p(w \mid \mathcal{D}, \mathcal{M} = m)}_{\text{Posterior under model}} = \frac{\overbrace{p(\mathcal{D} \mid \mathrm{w}, \mathcal{M} = m)}^{\text{Likelihood under model}} \times \overbrace{p(\mathrm{w}, \mathcal{M} = m)}^{\text{Prior}}}{\underbrace{p(\mathcal{D} \mid \mathcal{M} = m)}_{\text{Evidence for model}}}$$

Data $\mathcal{D} = \{X, Y\}$, input/output pairs, and parameters w for Model $\mathcal{M} = m$

Bayes' Rule for model selection..

$$\underbrace{p(w \mid \mathcal{D}, \mathcal{M} = m)}_{\text{Posterior under model}} = \frac{\overbrace{p(\mathcal{D} \mid \mathrm{w}, \mathcal{M} = m)}^{\text{Likelihood under model}} \times \overbrace{p(\mathrm{w}, \mathcal{M} = m)}^{\text{Prior}}}{\underbrace{p(\mathcal{D} \mid \mathcal{M} = m)}_{\text{Evidence for model}}}$$

Data $\mathcal{D} = \{X, Y\}$, input/output pairs, and parameters w for Model $\mathcal{M} = m$

$$\underbrace{p(\mathcal{M} = m \mid \mathcal{D})}_{\text{Posterior for model}} = \frac{\overbrace{p(\mathcal{D} \mid \mathcal{M} = m)}^{\text{Evidence for model}} \times \overbrace{p(\mathcal{M} = m)}^{\text{Prior for model}}}{\underbrace{p(\mathcal{D})}_{\text{Data}}}$$

Bayes' Rule for model selection..

$$\underbrace{p(w \mid \mathcal{D}, \mathcal{M} = m)}_{\text{Posterior under model}} = \frac{\overbrace{p(\mathcal{D} \mid \mathrm{w}, \mathcal{M} = m)}^{\text{Likelihood under model}} \times \overbrace{p(\mathrm{w}, \mathcal{M} = m)}^{\text{Prior}}}{\underbrace{p(\mathcal{D} \mid \mathcal{M} = m)}_{\text{Evidence for model}}}$$

Data $\mathcal{D} = \{X, Y\}$, input/output pairs, and parameters w for Model $\mathcal{M} = m$

$$\underbrace{p(\mathcal{M} = m \mid \mathcal{D})}_{\text{Posterior for model}} = \frac{\overbrace{p(\mathcal{D} \mid \mathcal{M} = m)}^{\text{Evidence for model}} \times \overbrace{p(\mathcal{M} = m)}^{\text{Prior for model}}}{\underbrace{p(\mathcal{D})}_{\text{Data}}}$$

If prior over models is equal, we compare via the Evidence for the Model: $p(\mathcal{D} \mid \mathcal{M} = m)$

Model selection example

Fitting polynomial models to data under Gaussian noise, $\varepsilon_n \sim \mathcal{N}(0, \sigma^2)$:

$$\text{Model 1} : y_n = a_0 + a_1 x_n + \varepsilon_n$$

$$\text{Model 2} : y_n = a_0 + a_1 x_n + a_2 x^2 + \varepsilon_n$$

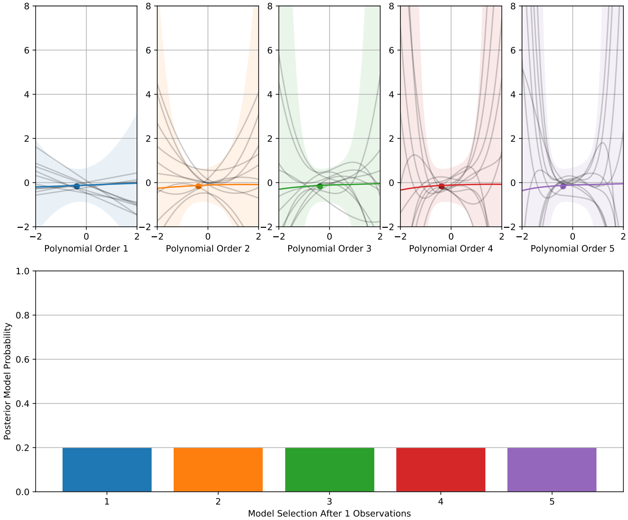
$$\text{Model 3} : y_n = a_0 + a_1 x_n + a_2 x^2 + a_3 x^3 + \varepsilon_n$$

$$\text{Model 4} : y_n = a_0 + a_1 x_n + a_2 x^2 + a_3 x^3 + a_4 x^4 + \varepsilon_n$$

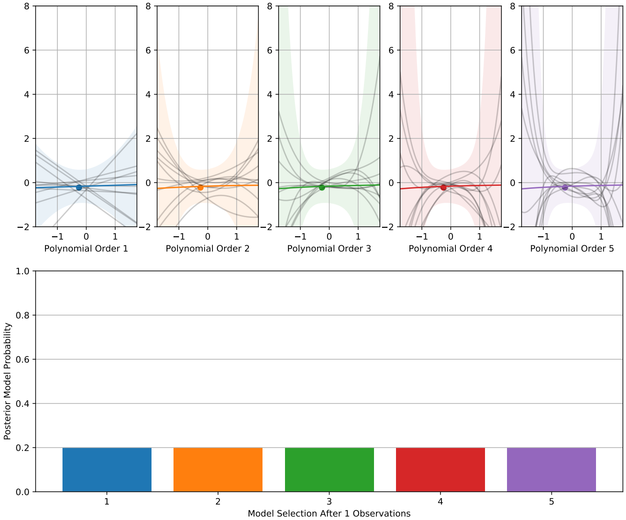
$$\text{Model 5} : y_n = a_0 + a_1 x_n + a_2 x^2 + a_3 x^3 + a_4 x^4 + a_5 x^5 + \varepsilon_n$$

Parameters $w_m = [a_0, \ldots, a_m]$ for model m, where $m \in [1, \ldots, 5]$.

Model selection example



Model selection example (more noise)



Causality

Overview...

Overview

No Free Lunch

Uncertainty / Error Bars

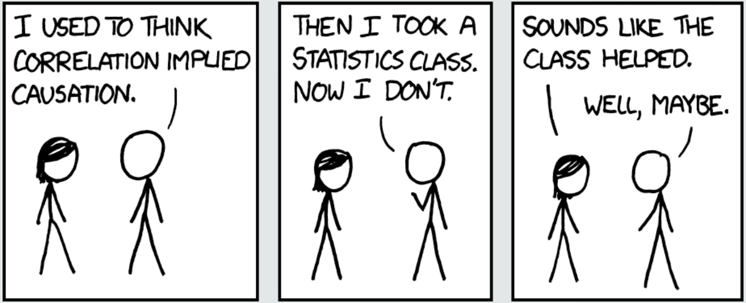
Model Selection

Causality

Conclusions

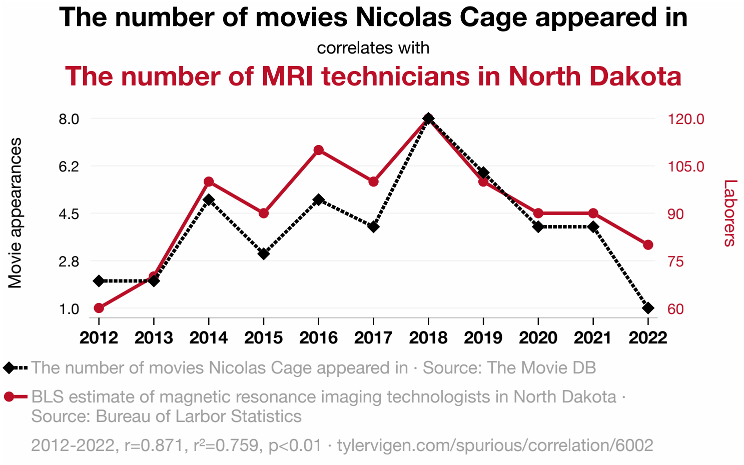
Correlation is not Causation

Correlation is not Causation



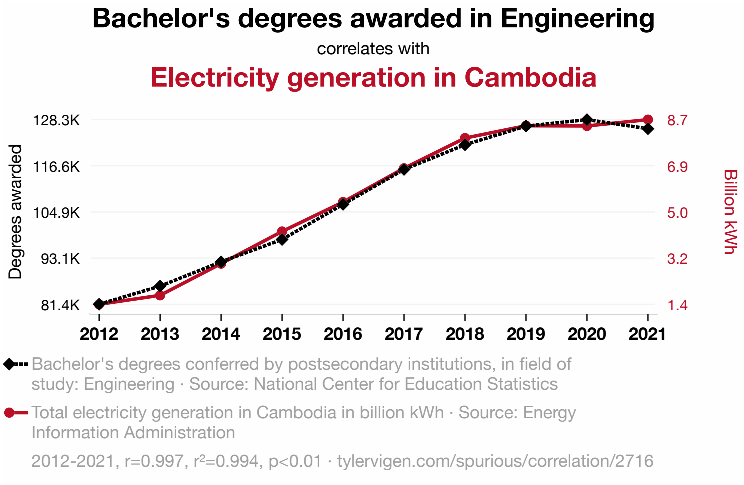
Danger Batman..

Danger Batman..



[https://tylervigen.com/spurious-correlations]

Danger Batman..



[https://tylervigen.com/spurious-correlations]

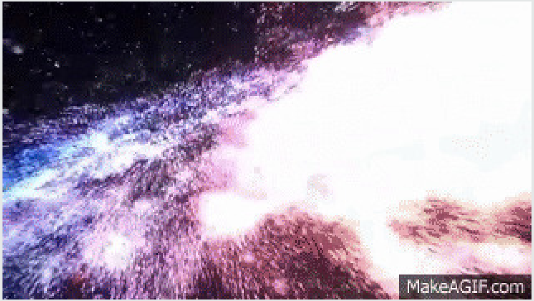
“Correlation is not Causation”

- Do we need causation?
- Is science not just correlation?

“Correlation is not Causation”

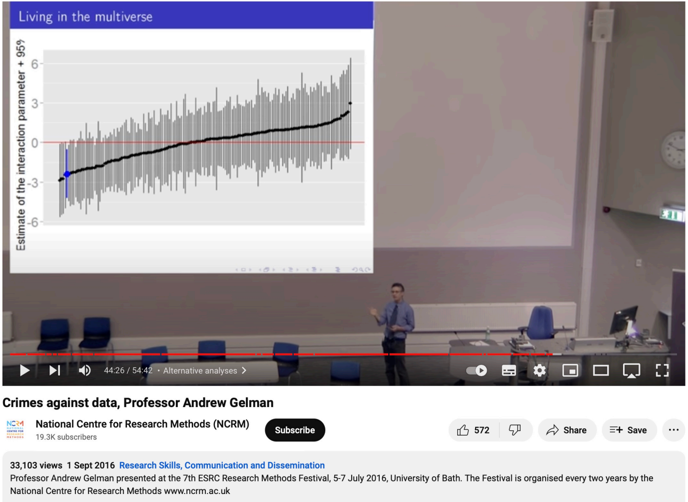
- Do we need causation?
- Is science not just correlation?

Importance simultaneously undervalued and overestimated?



Objectivity..

Objectivity..



[Andrew Gelman: “Crimes against Data”]

Correlation vs Causation: What’s the difference?

Correlation vs Causation: What's the difference?

- Well we all know what the difference is..
  - e.g. Atmospheric pressure and barometer needle reading

Correlation vs Causation: What's the difference?

- Well we all know what the difference is..
    - e.g. Atmospheric pressure and barometer needle reading

Formal definitions tricky but:

An object is the cause of another …

"if the first object had not been, the second never had existed"

[David Hume, Enquiry Concerning Human Understanding, 1748]

Correlation vs Causation: What's the difference?

- Well we all know what the difference is..
  - e.g. Atmospheric pressure and barometer needle reading

Formal definitions tricky but:

An object is the cause of another ...

"if the first object had not been, the second never had existed"

[David Hume, Enquiry Concerning Human Understanding, 1748]

- Introduces the idea of a counterfactual

Counterfactuals

- Real world:

$$\textsc{Action} \quad \longrightarrow \quad \textsc{Outcome}$$

Counterfactuals

- Real world:

$$\text{ACTION} \longrightarrow \text{OUTCOME}$$

- Hypothetical world:

$$\text{COUNTERFACUTAL ACTION} \longrightarrow \text{COUNTERFACTUAL OUTCOME}$$

Counterfactuals

- Real world:

$$\textsc{Action} \quad \longrightarrow \quad \textsc{Outcome}$$

- Hypothetical world:

$$\textsc{Counterfacutal Action} \quad \longrightarrow \quad \textsc{Counterfactual Outcome}$$

- Difference in outcome $=$ effect of the action!

Counterfactuals

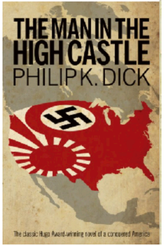
- Real world:

ACTION $\longrightarrow$ OUTCOME

- Hypothetical world:

COUNTERFACUTAL ACTION $\longrightarrow$ COUNTERFACTUAL OUTCOME

- Difference in outcome = effect of the action!



So we are all done?

So we are all done?

Problems with counterfactuals..

So we are all done?

Problems with counterfactuals..

- We never get to observe the counterfactual :-(

So we are all done?

Problems with counterfactuals..

- We never get to observe the counterfactual :-(
- Could the counterfactual possibly occur?
  - All the time inside our heads!

So we are all done?

Problems with counterfactuals..

- We never get to observe the counterfactual :-(
- Could the counterfactual possibily occur?
  - All the time inside our heads!
  - *What if I'd bought some tasty chocolates for Neill?*

So we are all done?

Problems with counterfactuals..

- We never get to observe the counterfactual :-(
- Could the counterfactual possibly occur?
    - All the time inside our heads!
    - *What if I'd bought some tasty chocolates for Neill?*
- Philosophical difficulties/objections..

So we are all done?

Problems with counterfactuals..

- We never get to observe the counterfactual :-(
- Could the counterfactual possibily occur?
  - All the time inside our heads!
  - *What if I'd bought some tasty chocolates for Neill?*
- Philosophical difficulties/objections..
- Can we approximate the counterfactual?

So we are all done?

Problems with counterfactuals..

- We never get to observe the counterfactual :-(
- Could the counterfactual possibily occur?
  - All the time inside our heads!
  - *What if I'd bought some tasty chocolates for Neill?*
- Philosophical difficulties/objections..
- Can we approximate the counterfactual?
  - Lots of the time in science $\longrightarrow$ the Randomised Control Trial (RCT)!

So we are all done?

Problems with counterfactuals..

- We never get to observe the counterfactual :-(
- Could the counterfactual possibly occur?
  - All the time inside our heads!
  - *What if I'd bought some tasty chocolates for Neill?*
- Philosophical difficulties/objections..
- Can we approximate the counterfactual?
  - Lots of the time in science $\longrightarrow$ the Randomised Control Trial (RCT)!
- Exciting question: what if we can't do RCT?

So we are all done?

Problems with counterfactuals..

- We never get to observe the counterfactual :-(
- Could the counterfactual possibly occur?
    - All the time inside our heads!
    - *What if I'd bought some tasty chocolates for Neill?*
- Philosophical difficulties/objections..
- Can we approximate the counterfactual?
    - Lots of the time in science $\longrightarrow$ the Randomised Control Trial (RCT)!
- Exciting question: what if we can't do RCT?
    - Can we use ML to estimate the counterfactuals?

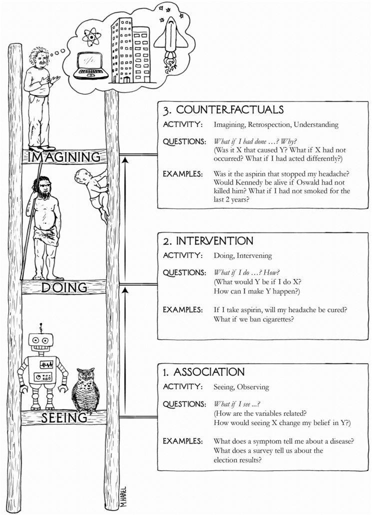
So we are all done?

Problems with counterfactuals..

- We never get to observe the counterfactual :-(
- Could the counterfactual possibily occur?
  - All the time inside our heads!
  - *What if I'd bought some tasty chocolates for Neill?*
- Philosophical difficulties/objections..
- Can we approximate the counterfactual?
  - Lots of the time in science $\longrightarrow$ the Randomised Control Trial (RCT)!
- Exciting question: what if we can't do RCT?
  - Can we use ML to estimate the counterfactuals? **Possibly!**

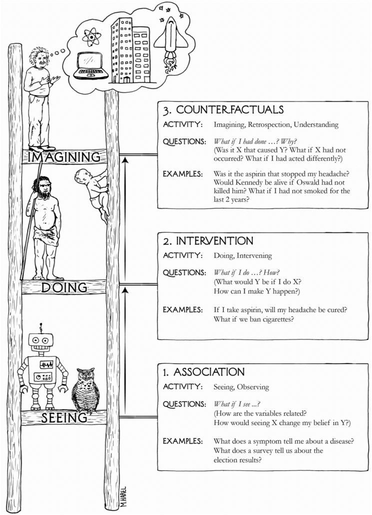
ML for Causation: Pearl's "Ladder of Causation"..

ML for Causation: Pearl's "Ladder of Causation"..



[Pearl and Mackensie 2017]

ML for Causation: Pearl's "Ladder of Causation"..



[Pearl and Mackensie 2017]

Causal reasoning cannot be answered by data alone we will need a model as well!

Illustration: Disease treatments

- Two disease treatments (surgical/non-surgival for kidney stones)

	Positive Outcome
Treatment A	273/350 = 78%
Treatment B	289/350 = **83%**

Illustration: Disease treatments

- Two disease treatments (surgical/non-surgival for kidney stones)

	Positive Outcome	Small Stones	Large Stones
Treatment A	273/350 = 78%	81/87 = **93%**	192/263 = **73%**
Treatment B	289/350 = **83%**	234/270 = 87%	55/80 = 69%

- What's going on?

Illustration: Disease treatments

- Two disease treatments (surgical/non-surgival for kidney stones)

	Positive Outcome	Small Stones	Large Stones
Treatment A	273/350 = 78%	81/87 = **93%**	192/263 = **73%**
Treatment B	289/350 = **83%**	234/270 = 87%	55/80 = 69%

- What's going on?
- Not a fair **RCT**: uneven allocation of patients

Illustration: Disease treatments

- Two disease treatments (surgical/non-surgival for kidney stones)

	Positive Outcome	Small Stones	Large Stones
Treatment A	273/350 = 78%	81/87 = **93%**	192/263 = **73%**
Treatment B	289/350 = **83%**	234/270 = 87%	55/80 = 69%

- What's going on?
- Not a fair **RCT**: uneven allocation of patients

The stone size S is a confounder:

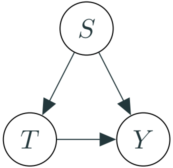


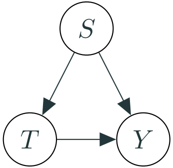
Illustration: Disease treatments

- Two disease treatments (surgical/non-surgival for kidney stones)

	Positive Outcome	Small Stones	Large Stones
Treatment A	273/350 = 78%	81/87 = **93%**	192/263 = **73%**
Treatment B	289/350 = **83%**	234/270 = 87%	55/80 = 69%

- What's going on?
- Not a fair **RCT**: uneven allocation of patients

The stone size S is a confounder:



RCT would remove the confounder:

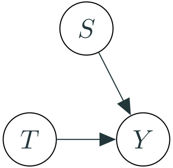
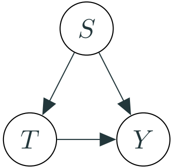


Illustration: Disease treatments

	Positive Outcome	Small Stones	Large Stones
Treatment A	273/350 = 78%	81/87 = **93%**	192/263 = **73%**
Treatment B	289/350 = **83%**	234/270 = 87%	55/80 = 69%

The stone size S is a confounder:



RCT would remove the confounder:

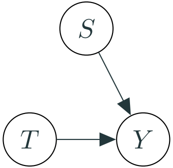
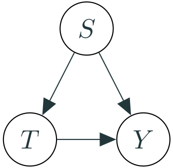


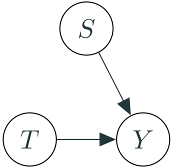
Illustration: Disease treatments

	Positive Outcome	Small Stones	Large Stones
Treatment A	273/350 = 78%	81/87 = **93%**	192/263 = **73%**
Treatment B	289/350 = **83%**	234/270 = 87%	55/80 = 69%

The stone size S is a confounder:



RCT would remove the confounder:



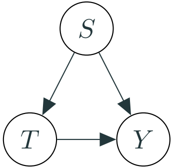
$p(A \mid B)$ means "probability of A being the case given that B occurs" **by observation alone**

$$p(Y \mid T) = \sum_s p(Y \mid S, T)\, p(T \mid S)\, p(S) \,/\, p(T)$$

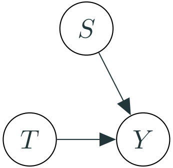
Illustration: Disease treatments

	Positive Outcome	Small Stones	Large Stones
Treatment A	273/350 = 78%	81/87 = **93%**	192/263 = **73%**
Treatment B	289/350 = **83%**	234/270 = 87%	55/80 = 69%

The stone size S is a confounder:



RCT would remove the confounder:



$p(A \mid B)$ means "probability of A being the case given that B occurs" by observation alone

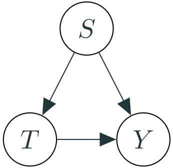
$$p(Y \mid T) = \sum_s p(Y \mid S, T)\, p(T \mid S)\, p(S) \,/\, p(T)$$

Probability of outcome Y given *intervening with treatment* T is $p(Y \mid \mathrm{do}(T))$

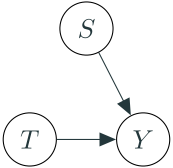
Illustration: Disease treatments

	Positive Outcome	Small Stones	Large Stones
Treatment A	273/350 = 78%	81/87 = **93%**	192/263 = **73%**
Treatment B	289/350 = **83%**	234/270 = 87%	55/80 = 69%

The stone size S is a confounder:



RCT would remove the confounder:



$p(A \mid B)$ means "probability of A being the case given that B occurs" **by observation alone**

$$p(Y \mid T) = \sum_s p(Y \mid S, T)\, p(T \mid S)\, p(S) \,/\, p(T)$$

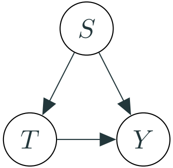
Probability of outcome Y given *intervening with treatment* T is $p(Y \mid \mathrm{do}(T))$

$$p(Y \mid \mathrm{do}(T)) = \sum_s p(Y, S \mid \mathrm{do}(T)) = \sum_s p(Y \mid S, \mathrm{do}(T))\, p(S)$$

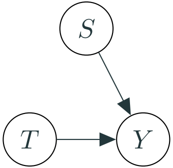
Illustration: Disease treatments

	Positive Outcome	Small Stones	Large Stones
Treatment A	273/350 = 78%	81/87 = **93%**	192/263 = **73%**
Treatment B	289/350 = **83%**	234/270 = 87%	55/80 = 69%

The stone size S is a confounder:



RCT would remove the confounder:

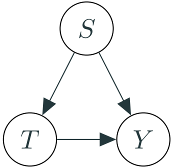


$$p(Y \mid \mathrm{do}(T)) = \sum_s p(Y, S \mid \mathrm{do}(T)) = \sum_s p(Y \mid S, \mathrm{do}(T))\, p(S)$$

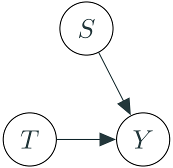
Illustration: Disease treatments

	Positive Outcome	Small Stones	Large Stones
Treatment A	273/350 = 78%	81/87 = **93%**	192/263 = **73%**
Treatment B	289/350 = **83%**	234/270 = 87%	55/80 = 69%

The stone size S is a confounder:



RCT would remove the confounder:



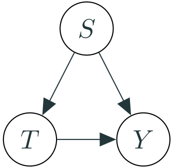
$$p(Y \mid \mathrm{do}(T)) = \sum_s p(Y, S \mid \mathrm{do}(T)) = \sum_s p(Y \mid S, \mathrm{do}(T))\, p(S)$$

$$p(Y \mid \mathrm{do}(T = a)) = \sum_s p(Y \mid S, \mathrm{do}(T = a))\, p(S) = \frac{81}{87}\frac{357}{700} + \frac{192}{263}\frac{343}{700} = 83\%$$

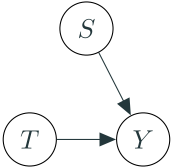
Illustration: Disease treatments

	Positive Outcome	Small Stones	Large Stones
Treatment A	273/350 = 78%	81/87 = **93%**	192/263 = **73%**
Treatment B	289/350 = **83%**	234/270 = 87%	55/80 = 69%

The stone size S is a confounder:



RCT would remove the confounder:



$$p(Y \mid \mathsf{do}(T)) = \sum_s p(Y, S \mid \mathsf{do}(T)) = \sum_s p(Y \mid S, \mathsf{do}(T))\, p(S)$$

$$p(Y \mid \mathsf{do}(T{=}a)) = \sum_s p(Y \mid S, \mathsf{do}(T{=}a))\, p(S) = \frac{81}{87}\frac{357}{700} + \frac{192}{263}\frac{343}{700} = 83\%$$

$$p(Y \mid \mathsf{do}(T{=}b)) = \sum_s p(Y \mid S, \mathsf{do}(T{=}b))\, p(S) = 78\%$$

Causality

- Statistical/Probabilistic reasoning alone cannot support causal inference
- Determining the joint probability distribution of variables says nothing about causation
- Causal Inference: promises to determine the necessary set of (non-data) assumptions sufficient to make a causal conclusion

[Thanks to Julian Faraway for Causal Illustrations]

Conclusions

Conclusions

Did we answer any of the questions?

- Can I use ML to solve x?
- What does ML actually do?
- Isn't ML just the same as y?
- Can I replace myself/my research team with ML?
- How much data do I need?
- Can I just use Deep Learning/Generative AI/ChatGPT?
- Surely Deep Learning/Generative AI/ChatGPT is all hype?
- Can any of this be used for science/engineering?

Conclusions

Need to think about what we really want..

- Computationally efficient look-up table
  - e.g. have loads of data that spans the space
  - Could use deep learning

Conclusions

Need to think about what we really want..

- Computationally efficient look-up table
    - e.g. have loads of data that spans the space
    - Could use deep learning
- Need data efficiency / care about uncertainty
    - e.g. clinical/safety applications
    - Need a Bayesian method

Conclusions

Need to think about what we really want..

- Computationally efficient look-up table
  - e.g. have loads of data that spans the space
  - Could use deep learning
- Need data efficiency / care about uncertainty
  - e.g. clinical/safety applications
  - Need a Bayesian method
- Want to analyse scientific results
  - e.g. does my new model explain dark matter
  - Need causal inference

Conclusions

Loads of gotchas..

- Availability (using the data you have not the data you need)
- Evaluation measure (is a human baseline sensible?)
- Ignore uncertainty/error bars
- Sample / dataset bias
- Bias / variance trade-off
- Haven't spoken about Decision Theory
- Lots to talk about regarding Causality
- "All models are wrong but some models are useful"
- ...

That’s all folks..

AI Talks: AI & ML Research Group, Department of Computer Science

11 Oct 2023	Prof Simon Prince
	Understanding Deep Learning: The Technology Behind Modern AI
15 Nov 2023	Prof Nello Cristianini
	The Shortcut: How Machines Became Intelligent Without Thinking in a Human Way
13 Dec 2023	Prof Mike Tipping
	The Irresistible Rise of Machine Learning
28 Feb 2024	Prof Neill Campbell
	No Free Lunches in Machine Learning
20 Mar 2024	Prof Özgür Şimşek
	Reinforcement Learning and the Pursuit of Artificial Intelligence
17 Apr 2024	Dr Harish Tayyar Madabushi
	Emergent Abilities of Language Models: Do they pose an existential threat?
8 May 2024	Prof Darren Cosker
	AI for Human Sensing: Research, Productisation and Ethics
TBD	Prof Mike Tipping
	Bayesian Inference in Machine Learning: Indistinguishable from Magic?