UNIVERSITY OF CAMBRIDGE

Machine Learning for Chemists

– Gaussian Processes –

---

19 February 2024

Outline

UNIVERSITY OF CAMBRIDGE

Introduction to Regression

Uncertainty of Predictions

The “big idea” behind Gaussian Processes

Examples of GPs used in the wild

Live Demo!

Regression

UNIVERSITY OF CAMBRIDGE

Answers a question that asks for a number.

Answers a question that asks for a number.

How much does this pet weigh?

Answers a question that asks for a number.

How much does this pet weigh?

This pet weighs **8 lbs**

Answers a question that asks for a number.

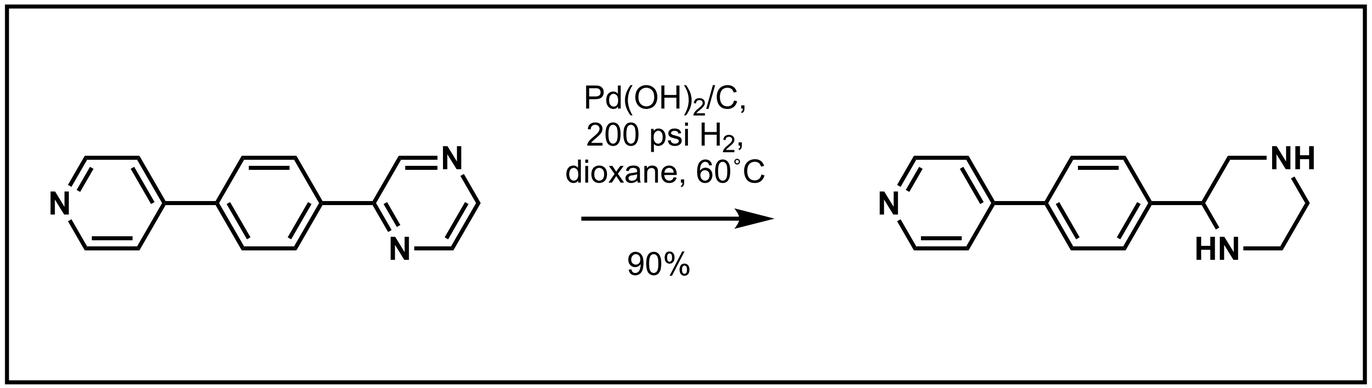
How much does this pet weigh?

This pet weighs **26 lbs**

UNIVERSITY OF CAMBRIDGE

Answers a question that asks for a number.

What is the yield for this reaction?

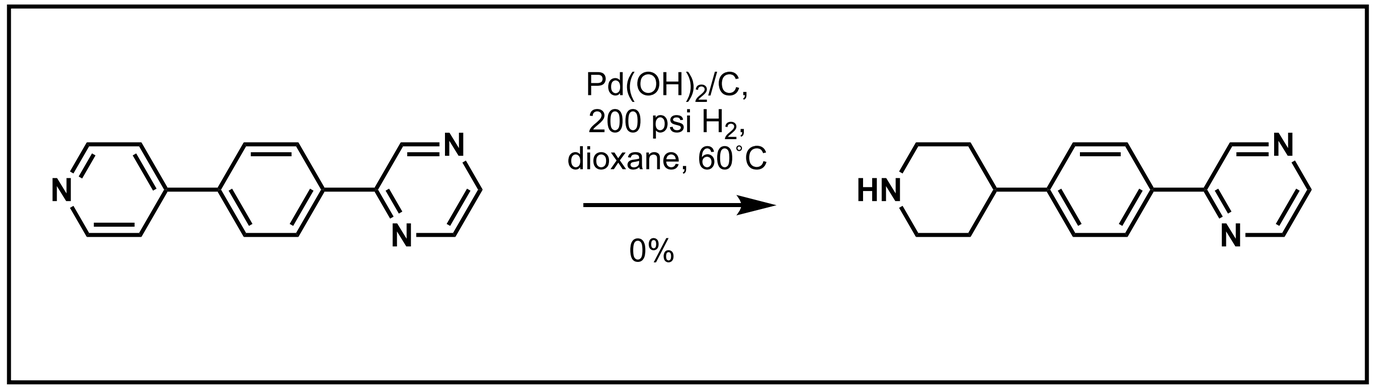


The yield of this reaction is **90%**

UNIVERSITY OF CAMBRIDGE

Answers a question that asks for a number.

What is the yield for this reaction?



The yield of this reaction is **0%**

UNIVERSITY OF CAMBRIDGE

Confusion Matrix

What we predicted (predictions) \ What actually happened (ground truth)	High Yielding	Not High Yielding
High Yielding	True Positive	False Positive
Not High Yielding	False Negative	True Negative

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad \text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall} = \frac{TP}{TP + FN} \qquad \text{F-Score} = \frac{2 \cdot TP}{TP + TN + FP}$$

Metrics quantify how far away from reality the predictions are.

UNIVERSITY OF CAMBRIDGE

Common Regression Metrics:

Mean Absolute Error (MAE)

UNIVERSITY OF CAMBRIDGE

Common Regression Metrics:

Mean Absolute Error (MAE)

Absolute Error = | What actually happened – What we predicted |

UNIVERSITY OF CAMBRIDGE

Common Regression Metrics:

Mean Absolute Error (MAE)

Absolute Error = | What actually happened – What we predicted |

	Exp 1	Exp 2	Exp 3
Ground truth	80%	25%	5%
Prediction	45%	30%	5%
Absolute Error			

UNIVERSITY OF CAMBRIDGE

Common Regression Metrics:

Mean Absolute Error (MAE)

Absolute Error = | What actually happened – What we predicted |

	Exp 1	Exp 2	Exp 3
Ground truth	**80%**	25%	5%
Prediction	**45%**	30%	5%
Absolute Error	**35%**		

Abs Error Exp 1 = | **80%** – **45%** | = **35%**

UNIVERSITY OF CAMBRIDGE

Common Regression Metrics:

Mean Absolute Error (MAE)

Absolute Error = | What actually happened – What we predicted |

	Exp 1	Exp 2	Exp 3
Ground truth	80%	**25%**	5%
Prediction	45%	**30%**	5%
Absolute Error	35%	**5%**	

Abs Error Exp 2 = | **25%** – **30%** | = **5%**

UNIVERSITY OF CAMBRIDGE

Common Regression Metrics:

Mean Absolute Error (MAE)

Absolute Error = | What actually happened – What we predicted |

	Exp 1	Exp 2	Exp 3
Ground truth	80%	25%	**5%**
Prediction	45%	30%	**5%**
Absolute Error	35%	5%	**0%**

Abs Error Exp 3 = | **5%** – **5%** | = **0%**

UNIVERSITY OF CAMBRIDGE

Common Regression Metrics:

Mean Absolute Error (MAE)

Absolute Error = | What actually happened – What we predicted |

	Exp 1	Exp 2	Exp 3
Ground truth	80%	25%	5%
Prediction	45%	30%	5%
Absolute Error	**35%**	**5%**	**0%**

$$\text{Mean Abs Error} = \frac{35\% + 5\% + 0\%}{3} = \textbf{13.3\% yield}$$

UNIVERSITY OF CAMBRIDGE

Common Regression Metrics:

Root Mean Squared Error (RMSE)

UNIVERSITY OF CAMBRIDGE

Common Regression Metrics:

Root Mean Squared Error (RMSE)

Squared Error = (What actually happened – What we predicted)^2

UNIVERSITY OF CAMBRIDGE

Common Regression Metrics:

Root Mean Squared Error (RMSE)

Squared Error = (What actually happened – What we predicted)^2

	Exp 1	Exp 2	Exp 3
Ground truth	**80%**	25%	5%
Prediction	**45%**	30%	5%

Squared Error **1225%**

Sqrd Error Exp 1 = (**80%** – **45%**)^2 = **1225%**

UNIVERSITY OF CAMBRIDGE

Common Regression Metrics:

Root Mean Squared Error (RMSE)

Squared Error = (What actually happened – What we predicted)^2

	Exp 1	Exp 2	Exp 3
Ground truth	80%	**25%**	5%
Prediction	45%	**30%**	5%
Squared Error	1225%	**25%**	

Sqrd Error Exp 2 = (**25%** – **30%**)^2 = **25%**

UNIVERSITY OF CAMBRIDGE

Common Regression Metrics:

Root Mean Squared Error (RMSE)

Squared Error = (What actually happened – What we predicted)^2

	Exp 1	Exp 2	Exp 3
Ground truth	80%	25%	**5%**
Prediction	45%	30%	**5%**
Squared Error	1225%	25%	**0%**

Sqrd Error Exp 3 = (**5%** – **5%**)^2 = **0%**

UNIVERSITY OF CAMBRIDGE

Common Regression Metrics:

Root Mean Squared Error (RMSE)

Squared Error = (What actually happened – What we predicted)^2

	Exp 1	Exp 2	Exp 3
Ground truth	80%	25%	5%
Prediction	45%	30%	5%
Squared Error	**1225%**	**25%**	**0%**

$$\text{Mean Squared Error} = \frac{1225\% + 25\% + 0\%}{3}$$

UNIVERSITY OF CAMBRIDGE

Common Regression Metrics:

Root Mean Squared Error (RMSE)

Squared Error = (What actually happened – What we predicted)^2

	Exp 1	Exp 2	Exp 3
Ground truth	80%	25%	5%
Prediction	45%	30%	5%
Squared Error	**1225%**	**25%**	**0%**

$$\text{Root Mean Squared Error} = \sqrt{\frac{1225\% + 25\% + 0\%}{3}} = \textbf{20.4\% yield}$$

UNIVERSITY OF CAMBRIDGE

Common Regression Metrics:

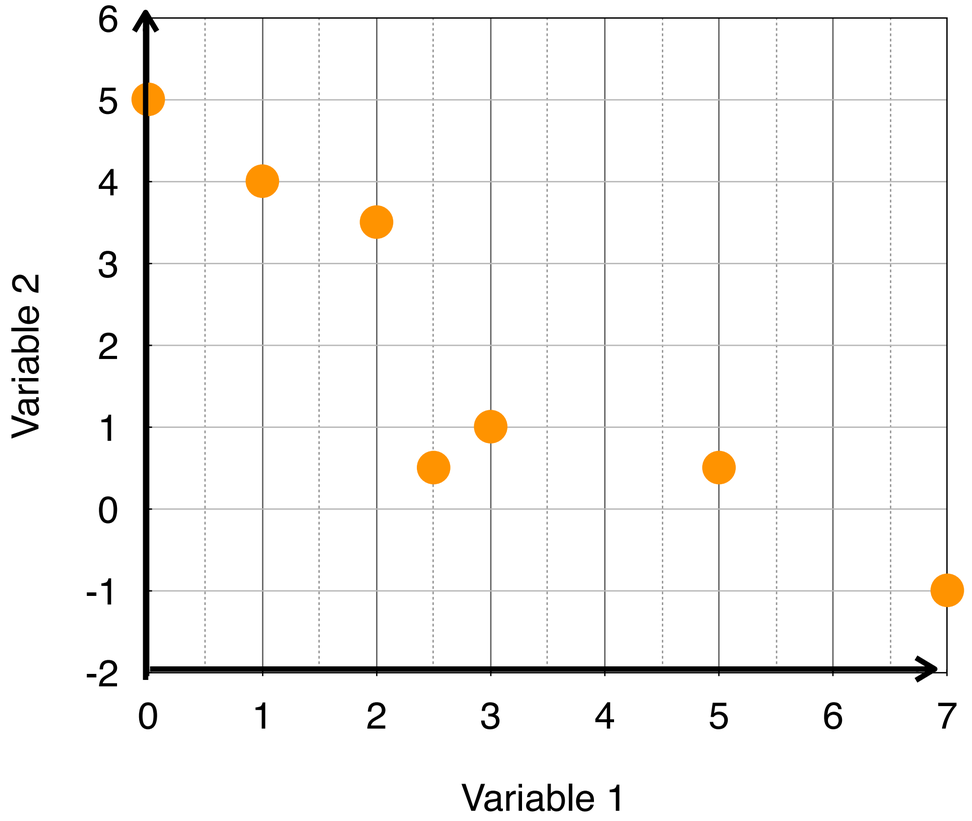
R^2 Error

UNIVERSITY OF CAMBRIDGE

Common Regression Metrics:

R^2 Error

Recall from early graphing

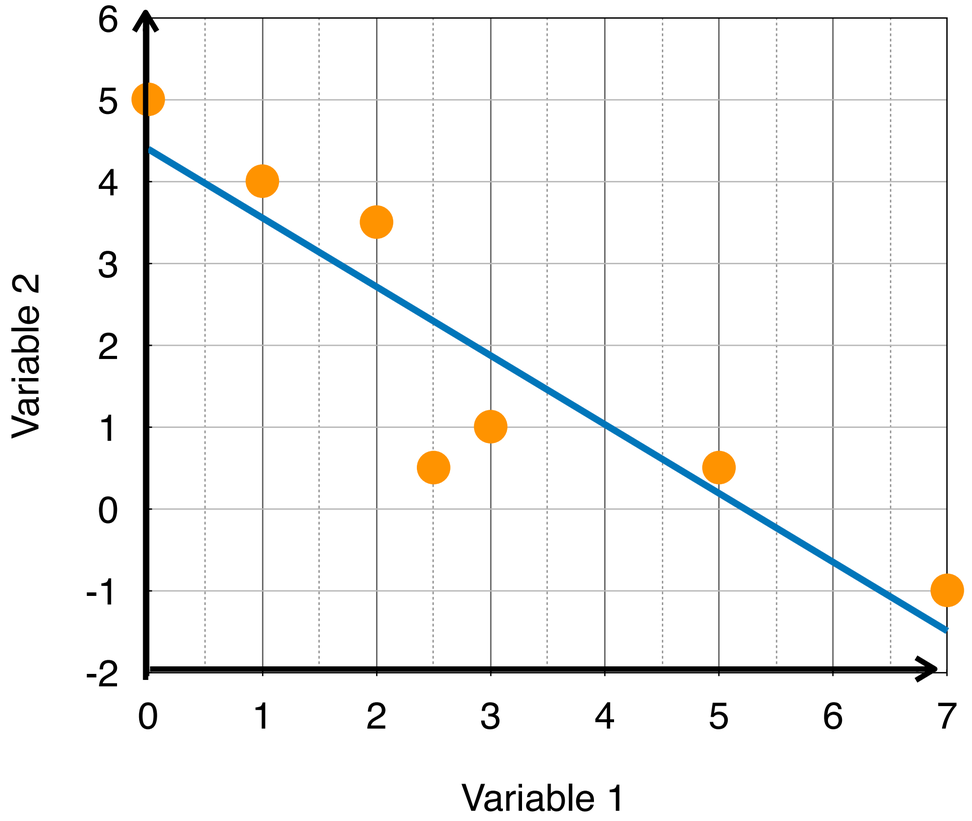


UNIVERSITY OF CAMBRIDGE

Common Regression Metrics:

R^2 Error

Recall from early graphing

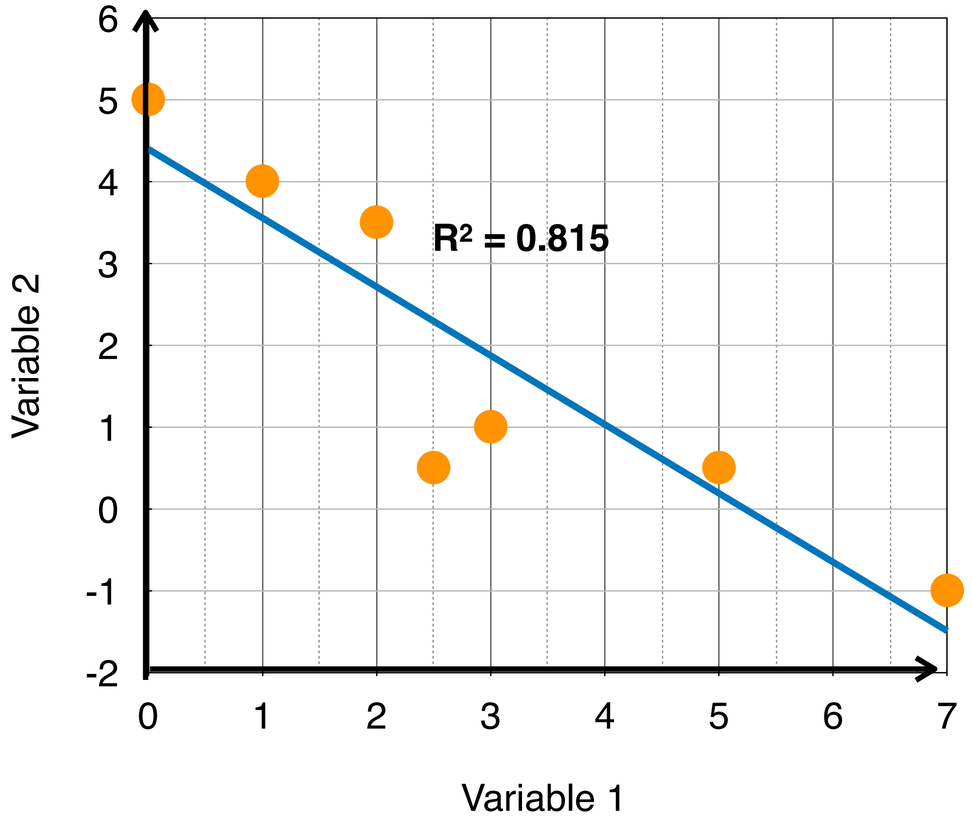


UNIVERSITY OF CAMBRIDGE

Common Regression Metrics:

R^2 Error

Recall from early graphing

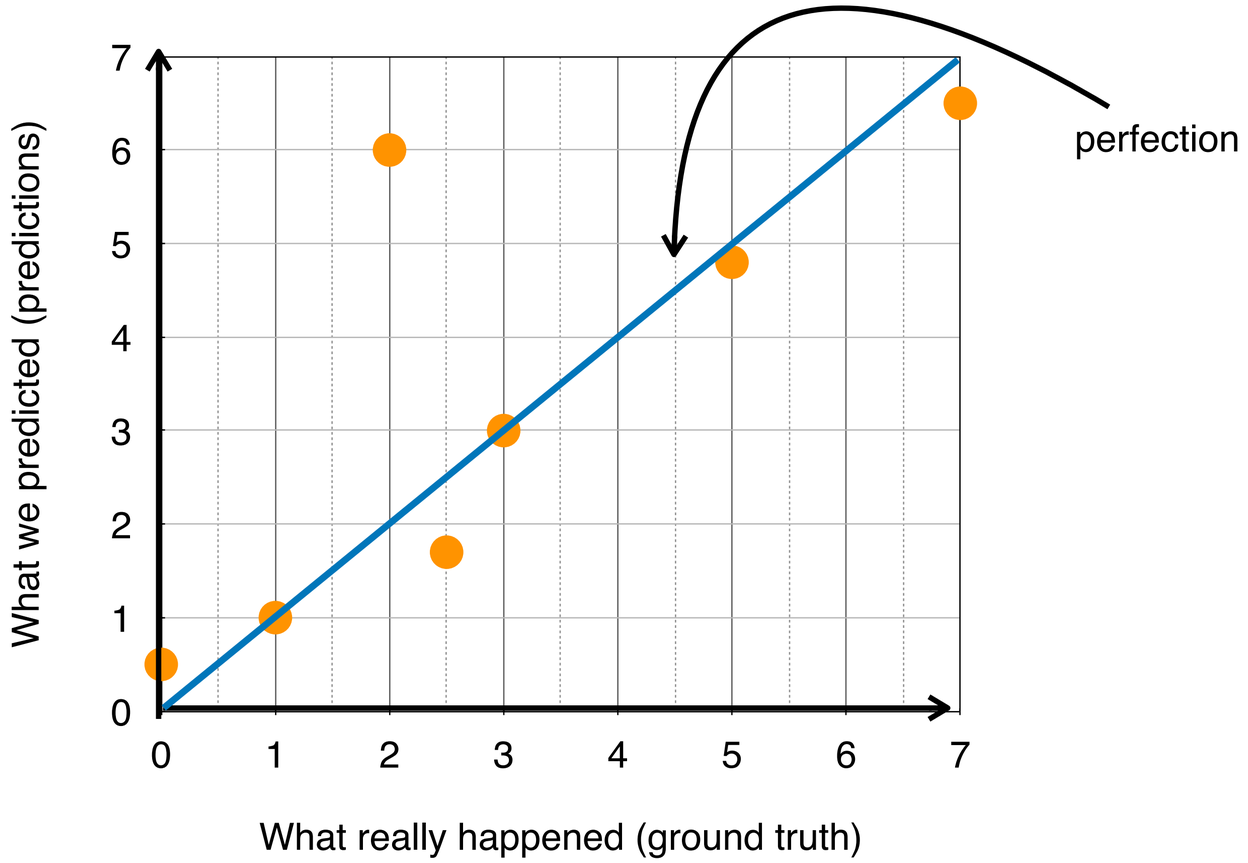


UNIVERSITY OF CAMBRIDGE

Common Regression Metrics:

R^2 Error

How well does the line y = x fit the data?



Answers a question that asks for a number AND answers with its confidence.

How much does this pet weigh?

How sure of your answer are you?

UNIVERSITY OF CAMBRIDGE

Answers a question that asks for a number AND answers with its confidence.

How much does this pet weigh?

How sure of your answer are you?

This pet weighs **8 lbs**

I’m not very sure (lots of fluff).

± 4 lbs

UNIVERSITY OF CAMBRIDGE

Answers a question that asks for a number AND answers with its confidence.

How much does this pet weigh?

How sure of your answer are you?

This pet weighs **26 lbs**

I'm very sure (I've seen lots of corgis).

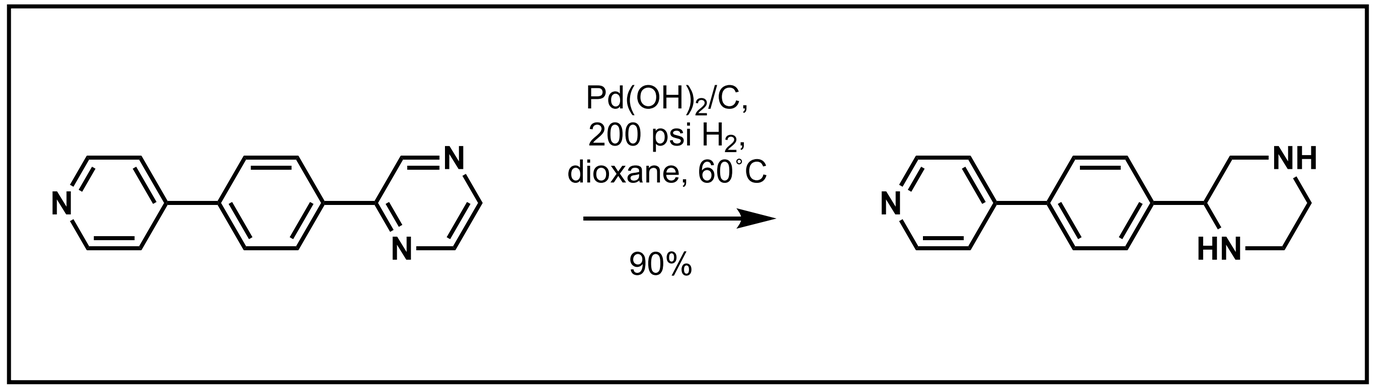
± 0.5 lbs

UNIVERSITY OF CAMBRIDGE

Answers a question that asks for a number AND answers with its confidence.

What is the yield of this reaction?

How sure of your answer are you?

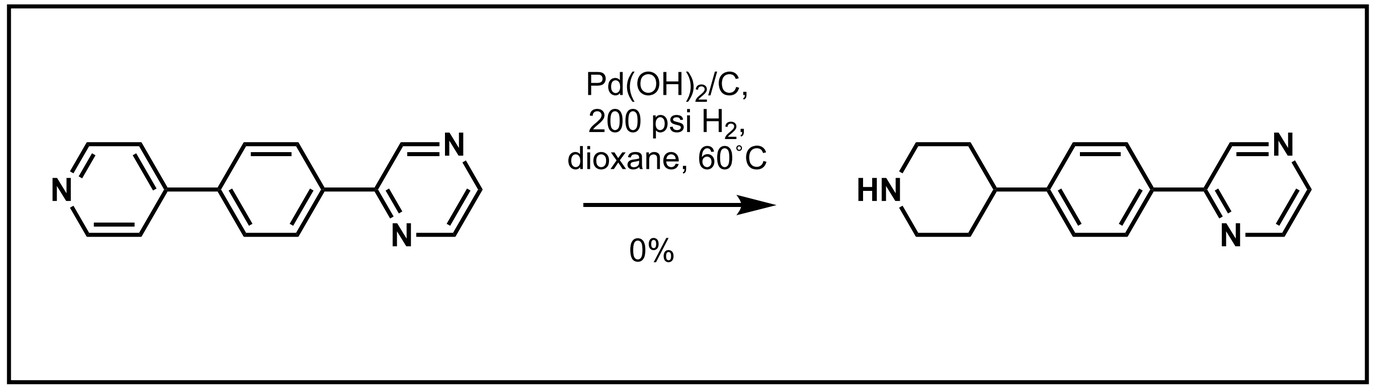


The yield of this reaction is **90% ± 10%**

Answers a question that asks for a number AND answers with its confidence.

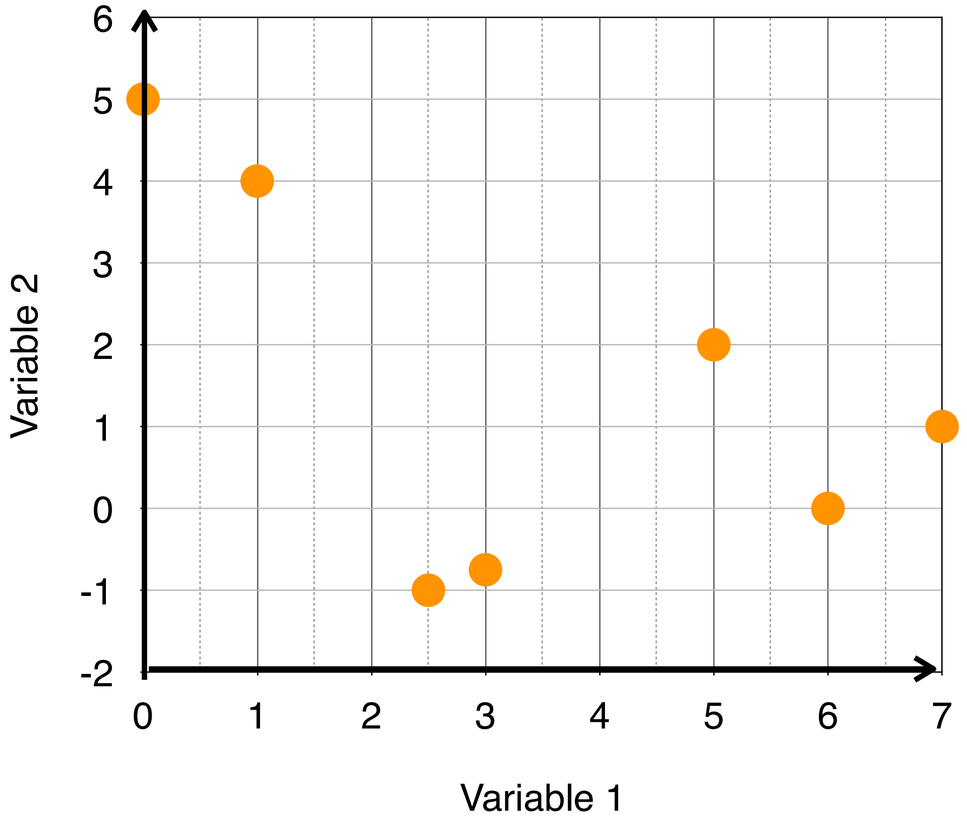
What is the yield of this reaction?

How sure of your answer are you?

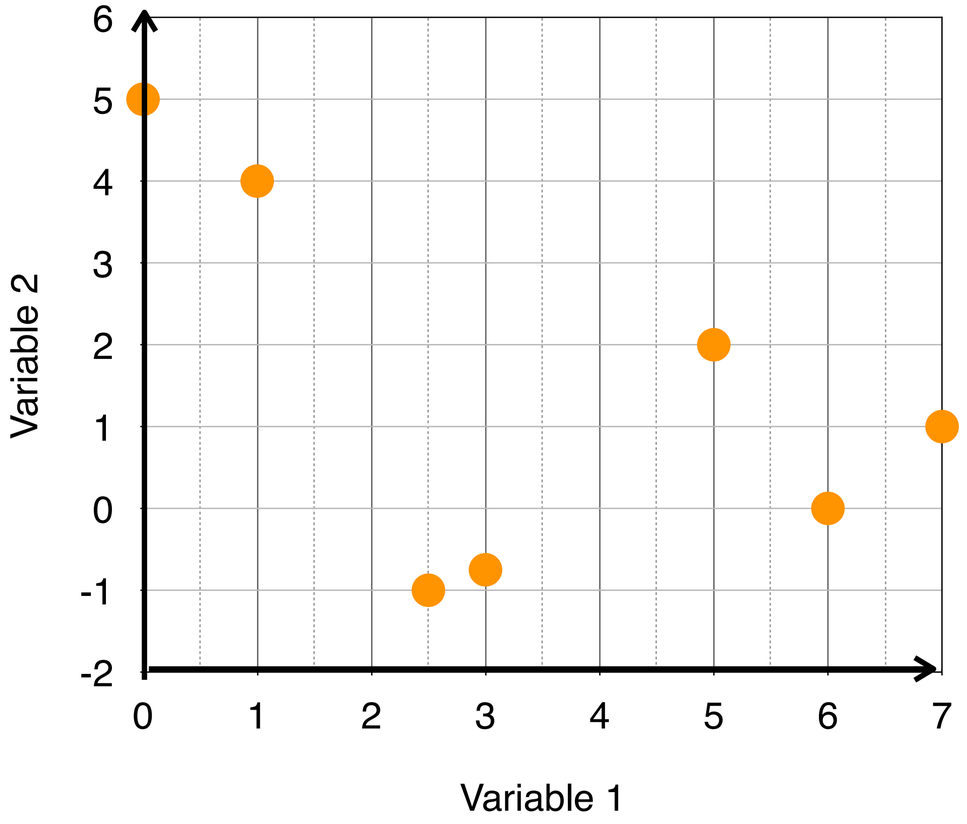


The yield of this reaction is **0% ± 50%**

Our Typical Workflow for Regression

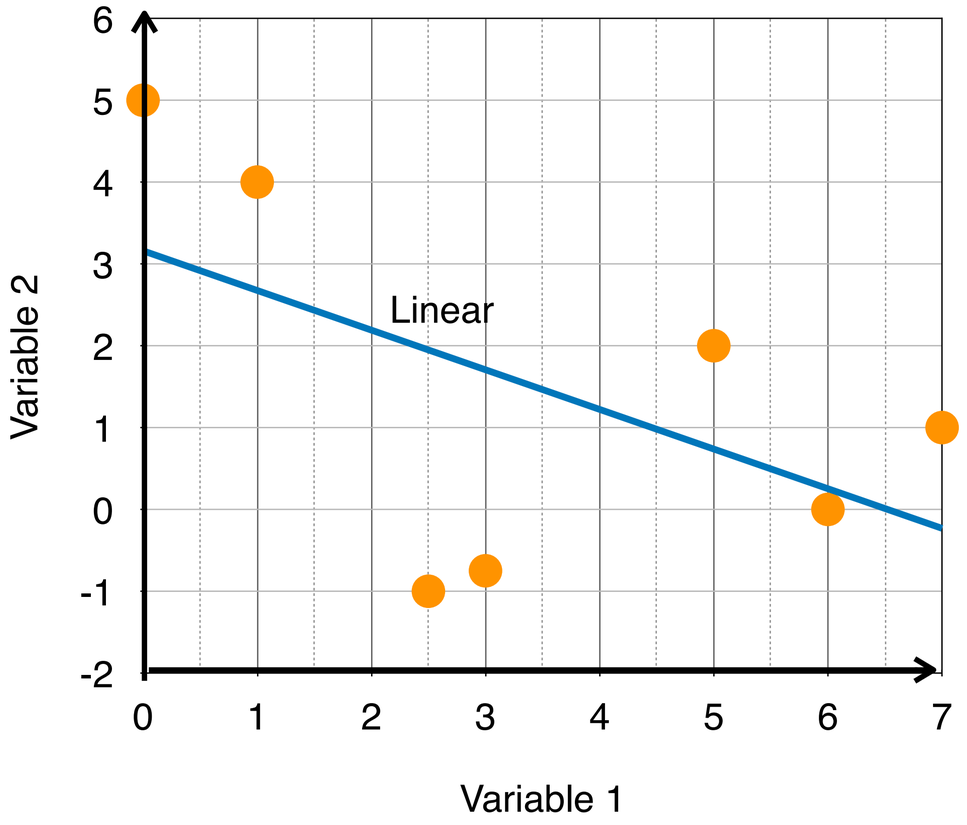


Our Typical Workflow for Regression



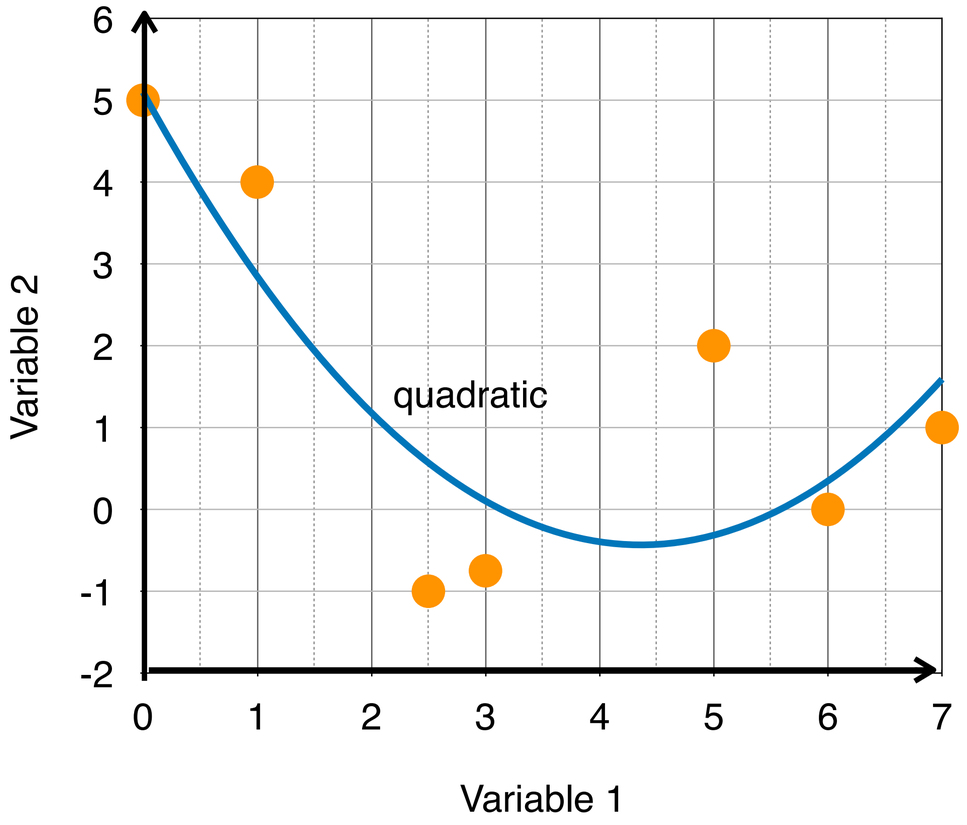
We have data → Single style of function to fit data

Our Typical Workflow for Regression



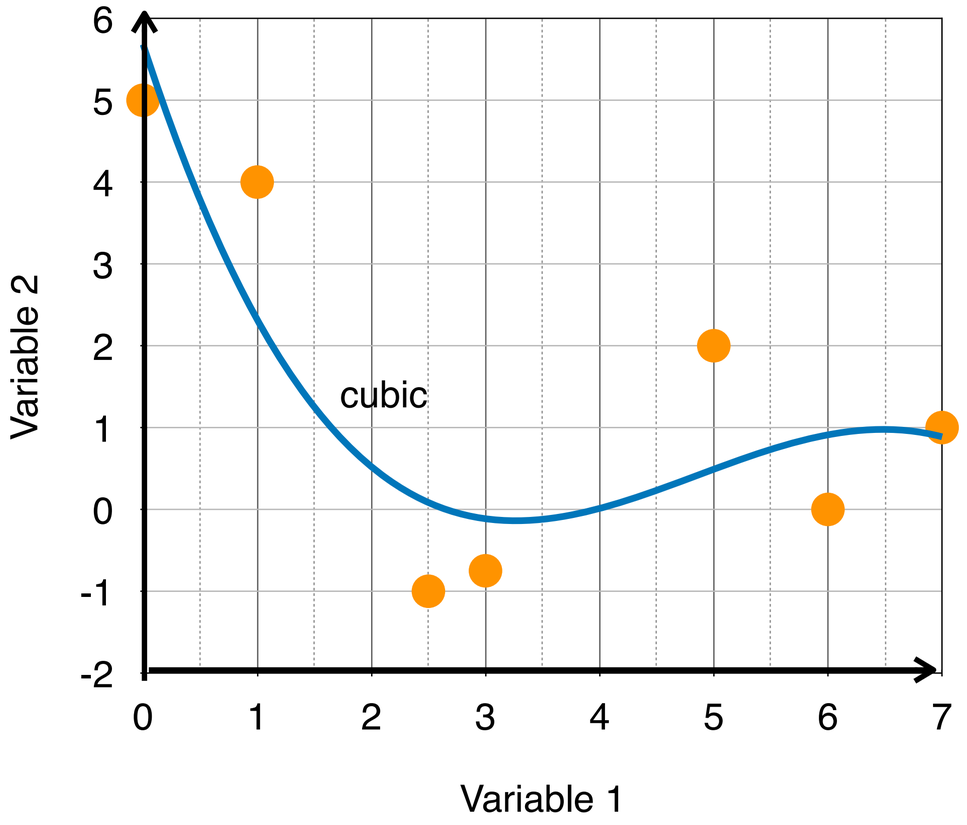
We have data → Single style of function to fit data

Our Typical Workflow for Regression



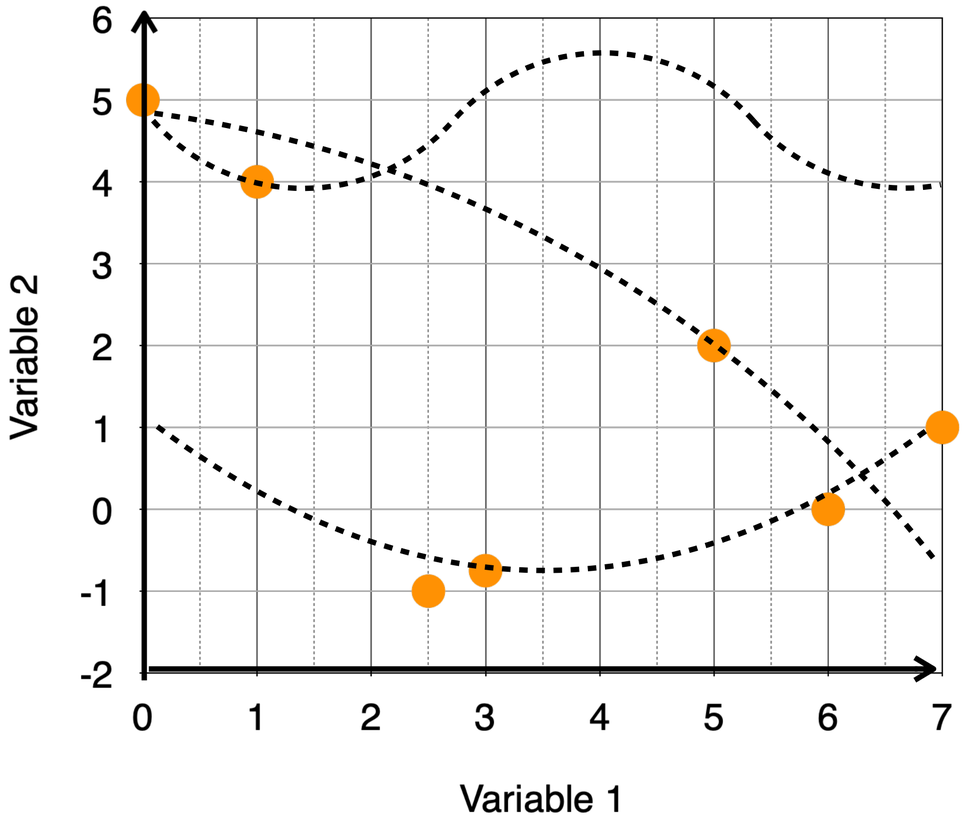
We have data → Single style of function to fit data

Our Typical Workflow for Regression



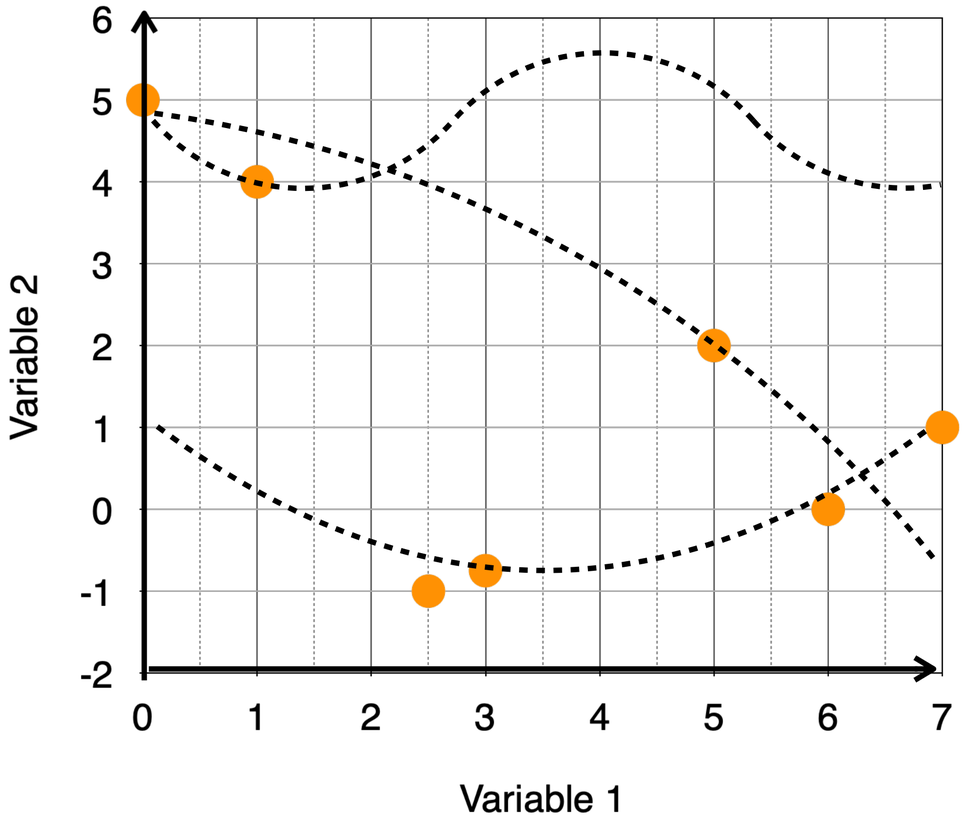
We have data → Single style of function to fit data

The Big Idea



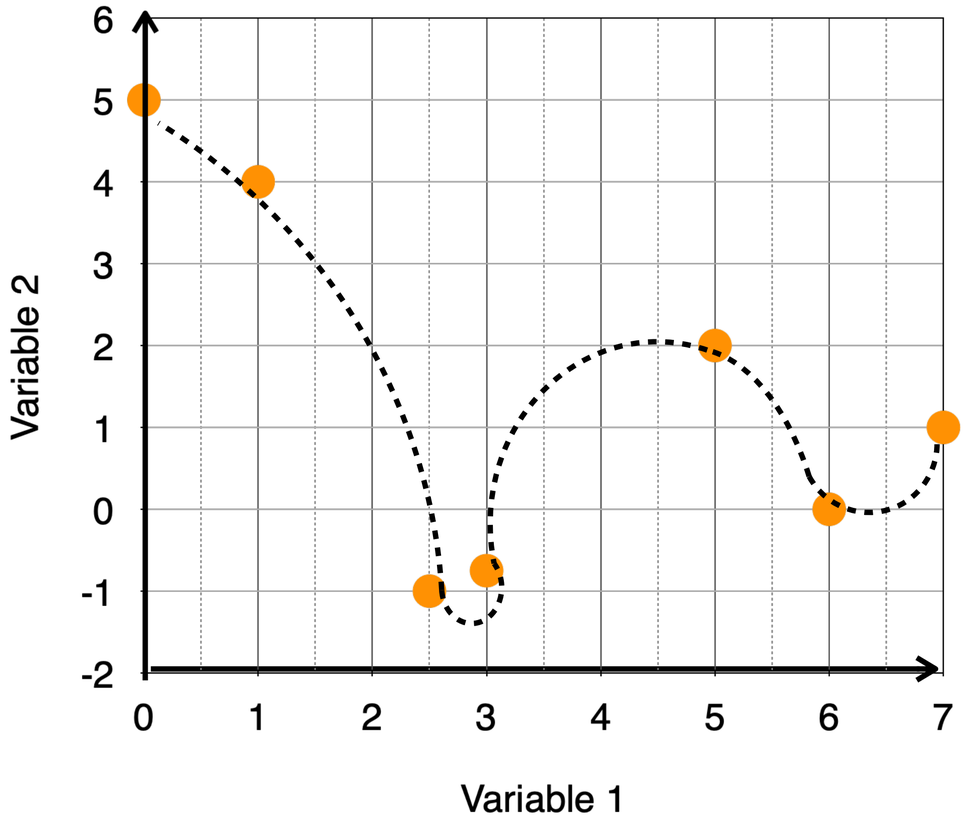
We have data → All possible functions that fit the data

The Big Idea



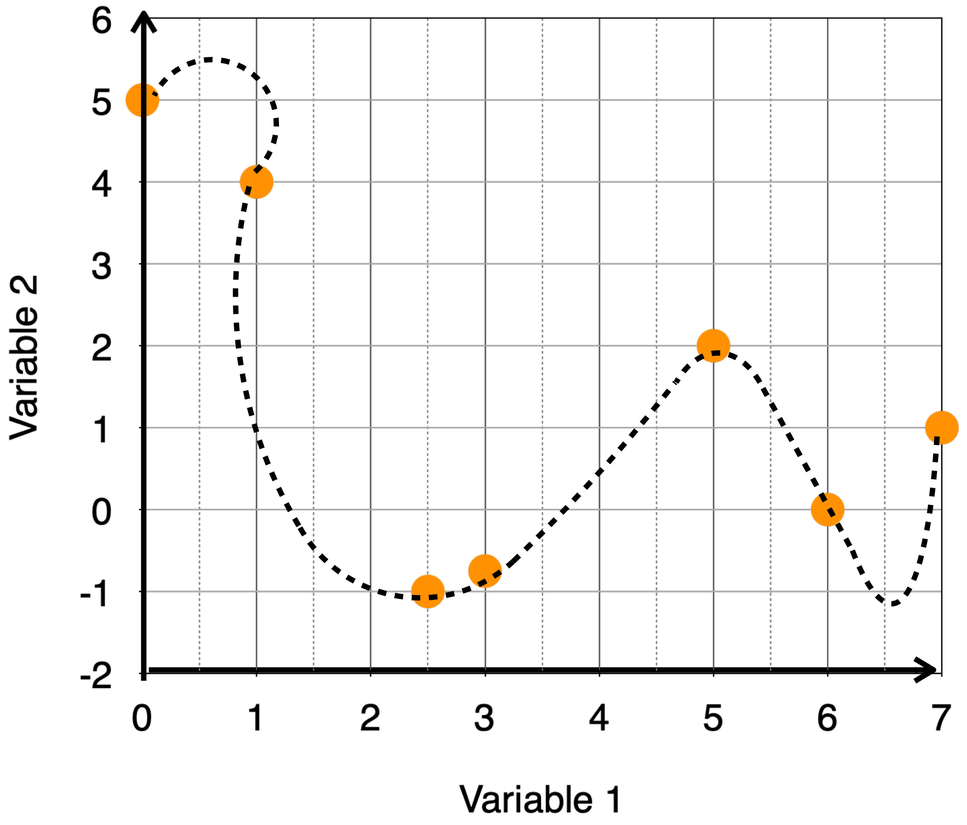
We have data → All possible functions that **go through ALL our points**

The Big Idea



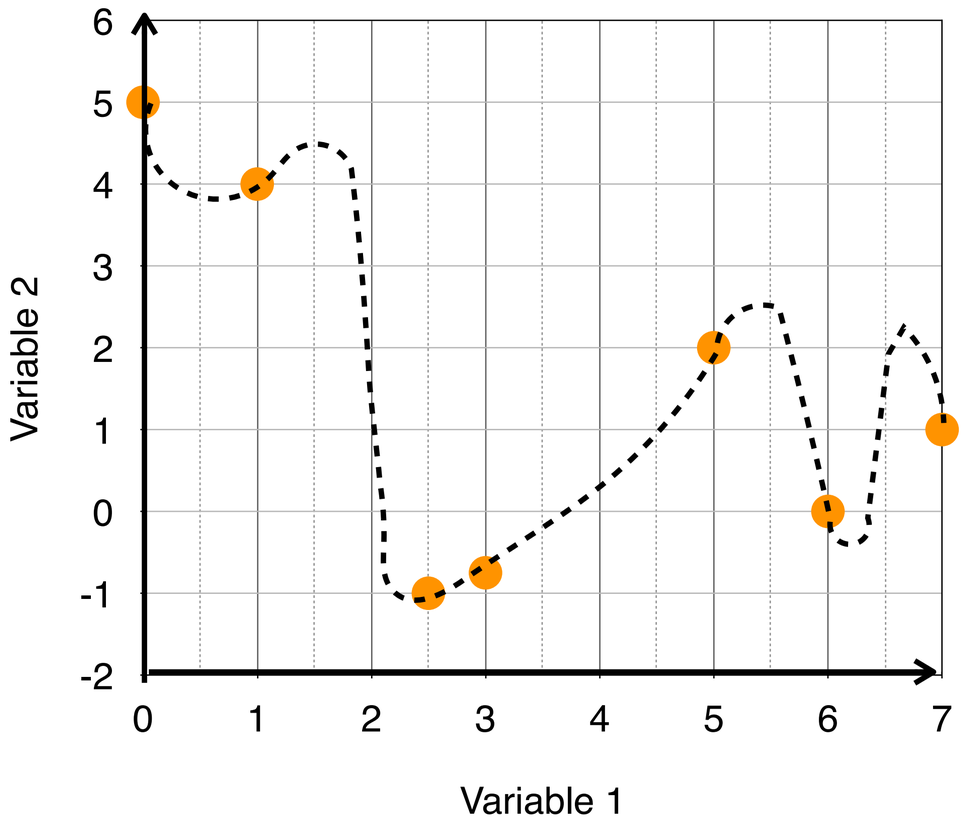
We have data → All possible functions that **go through ALL our points**

The Big Idea



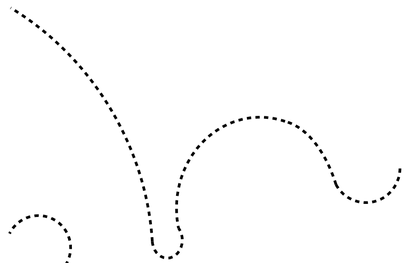
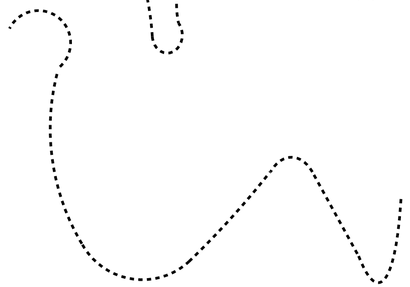
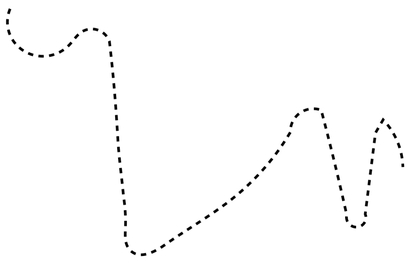
We have data → All possible functions that **go through ALL our points**

The Big Idea

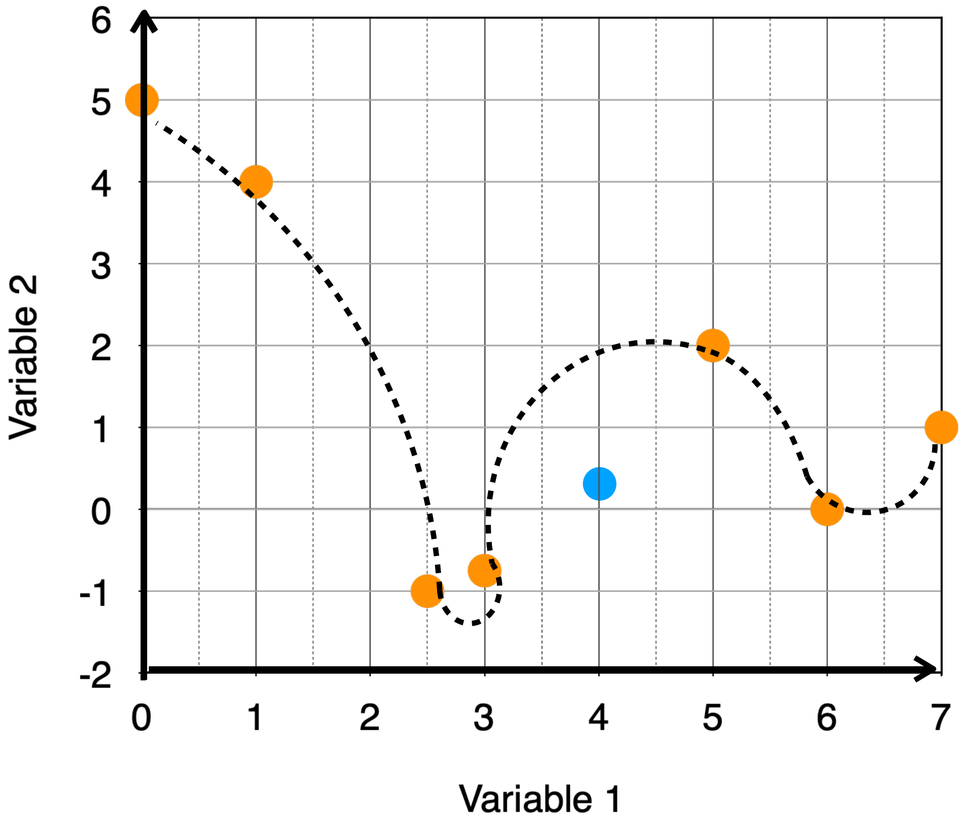


We have data → All possible functions that **go through ALL our points**

The Big Idea

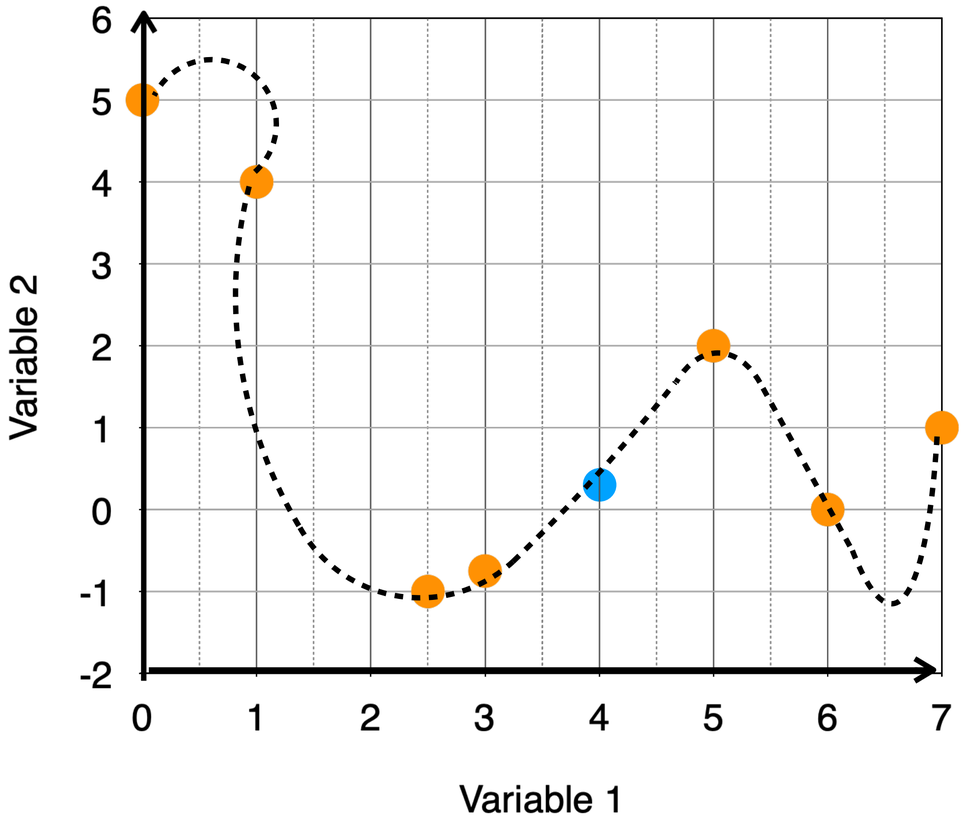
Possible Function	Probability
	5%
	3%
	4%
⋮	

The Big Idea



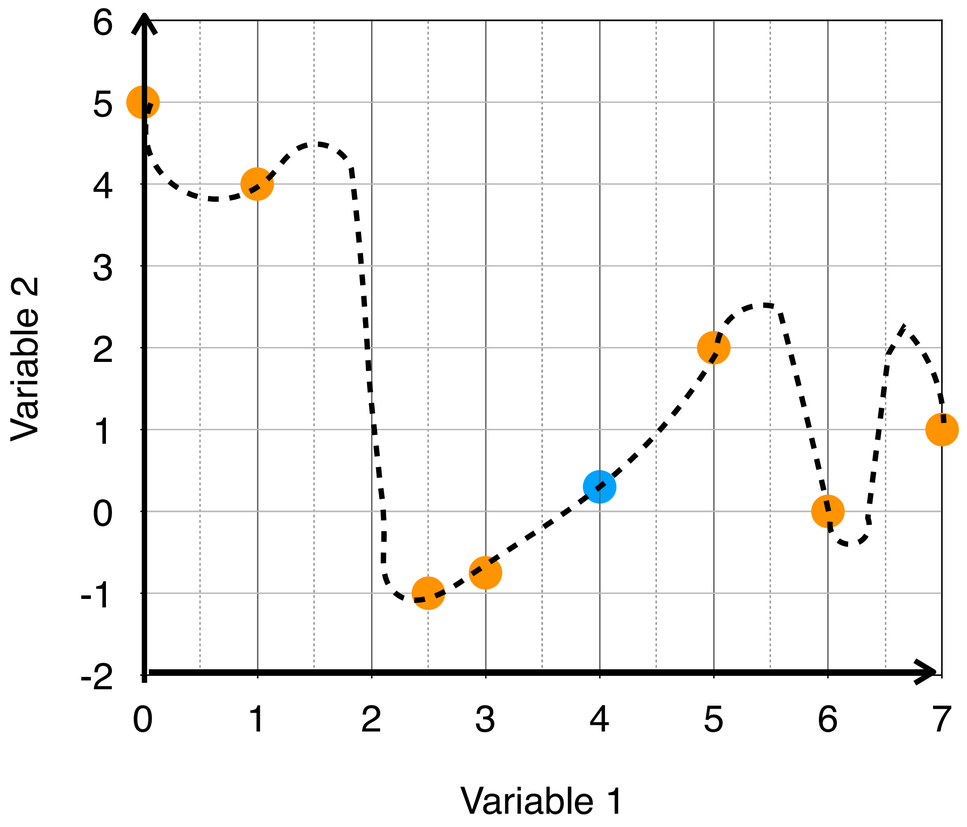
Add new data point → Update the probabilities

The Big Idea



Add new data point → Update the probabilities

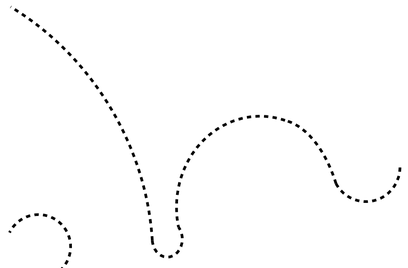
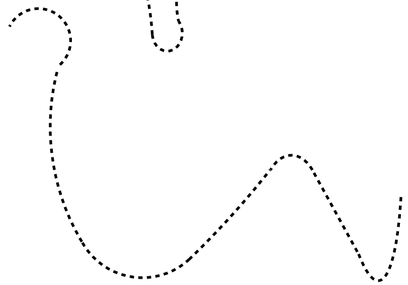
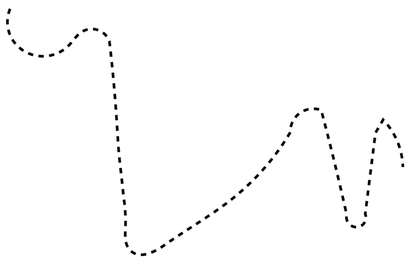
The Big Idea



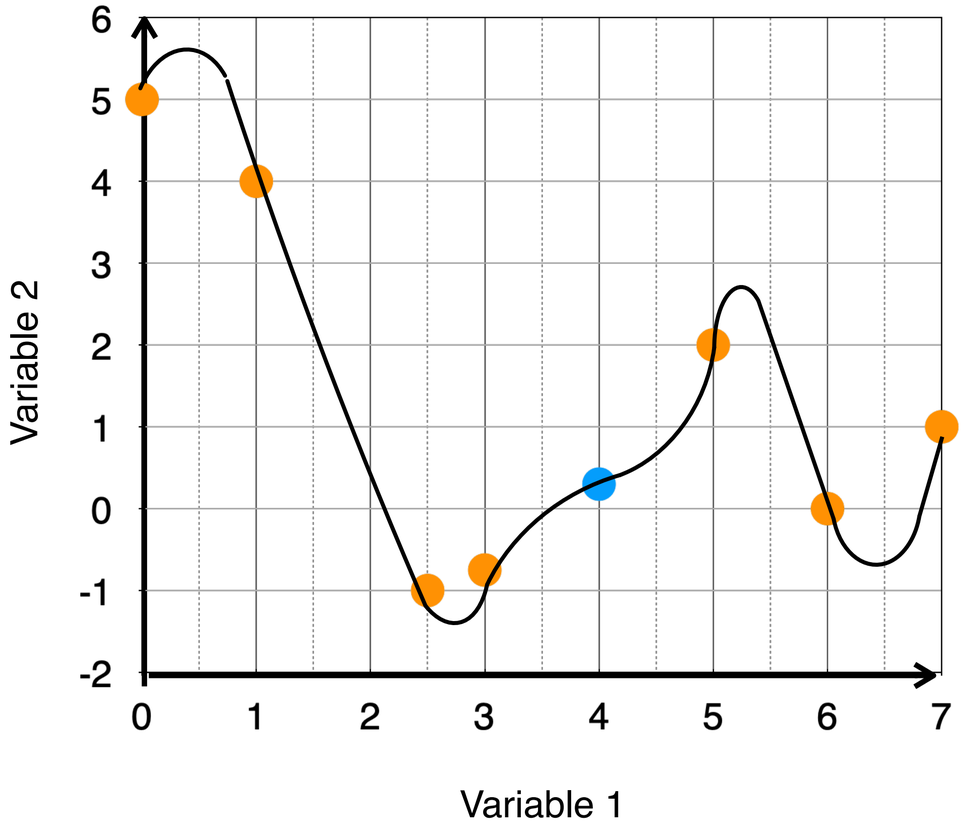
Add new data point → Update the probabilities

The Big Idea

UNIVERSITY OF CAMBRIDGE

Possible Function	Probability
	**0%**
	3%
	7%
⋮	

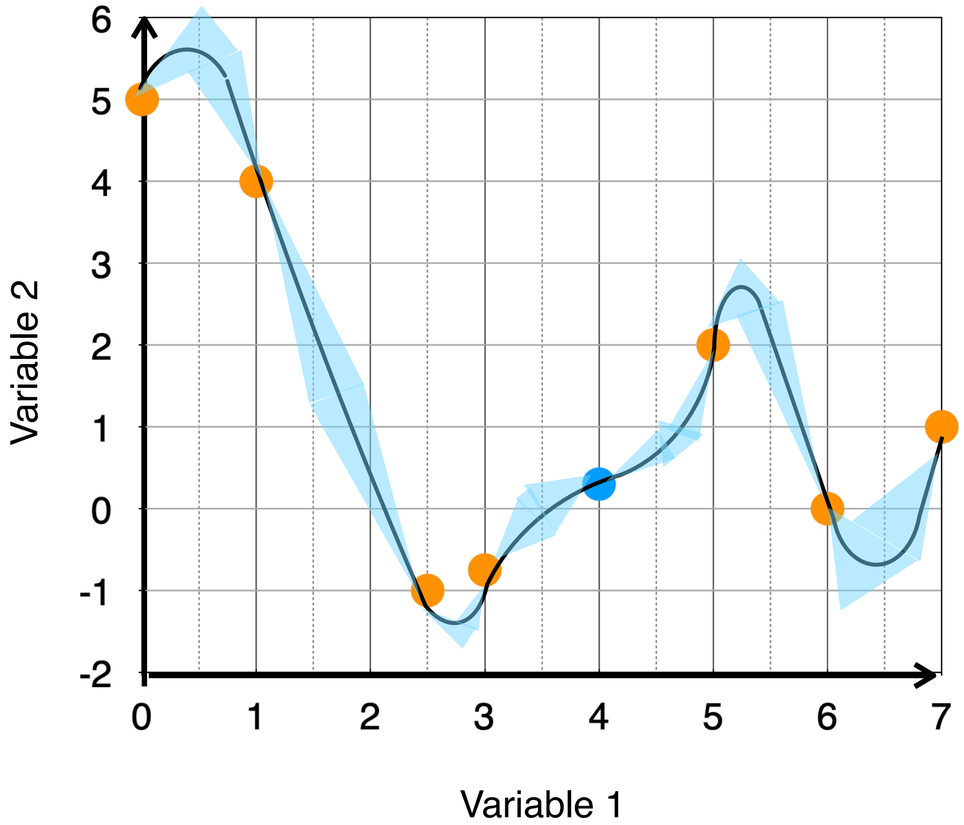
The Big Idea



Weighted average of the possible functions → Mean = most probable value

https://distill.pub/2019/visual-exploration-gaussian-processes/

The Big Idea



Weighted average of the possible functions

Mean = most probable value

Standard Deviation = uncertainty of each prediction
(based on values of other probable functions)

https://distill.pub/2019/visual-exploration-gaussian-processes/

A Challenge for Gaussian Processes

UNIVERSITY OF CAMBRIDGE

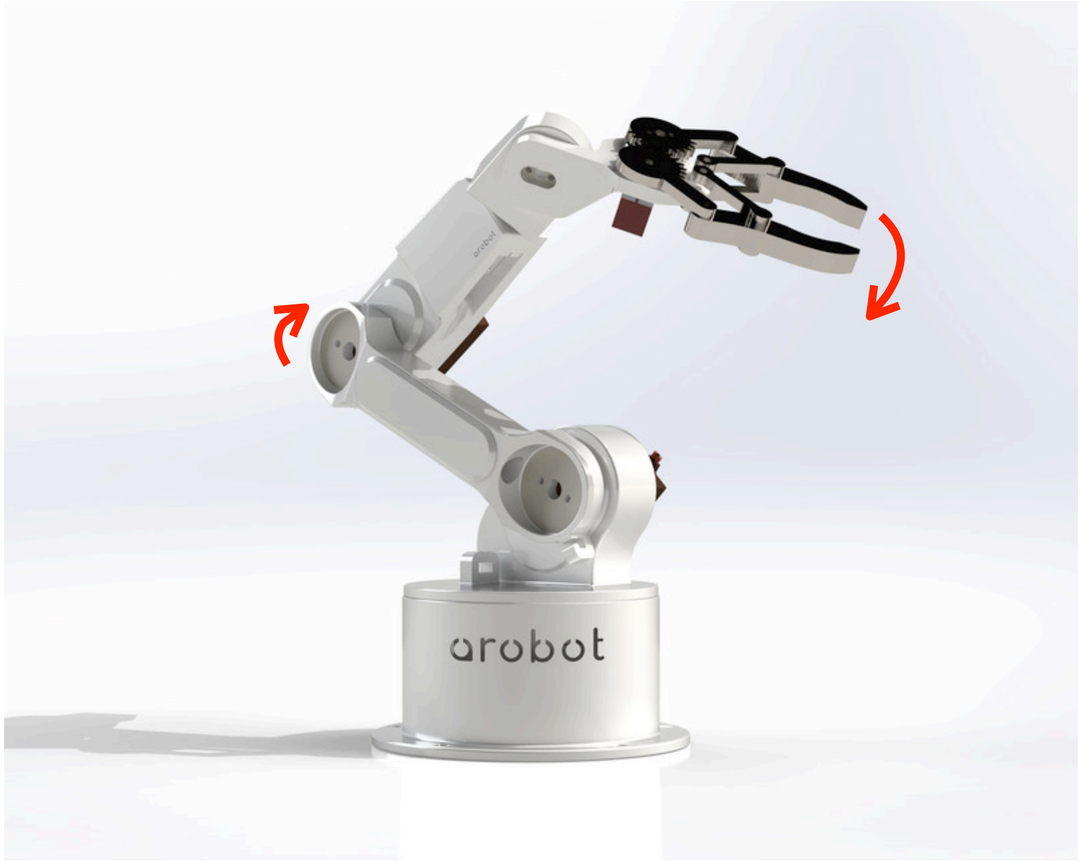
Gaussian Processes work best for continuous processes.

https://2020blogfor.github.io/posts/2020/04/kf_gp/

A Challenge for Gaussian Processes

UNIVERSITY OF CAMBRIDGE

Gaussian Processes work best for continuous processes.

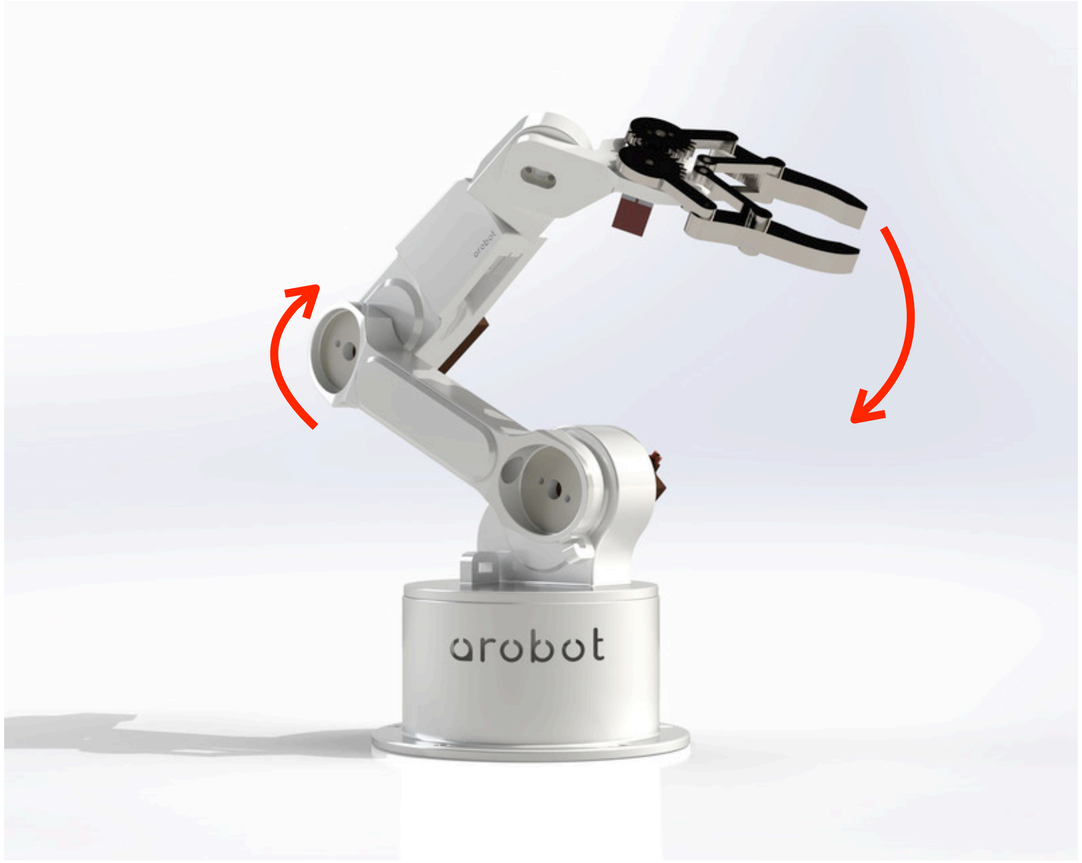


https://2020blogfor.github.io/posts/2020/04/kf_gp/

A Challenge for Gaussian Processes

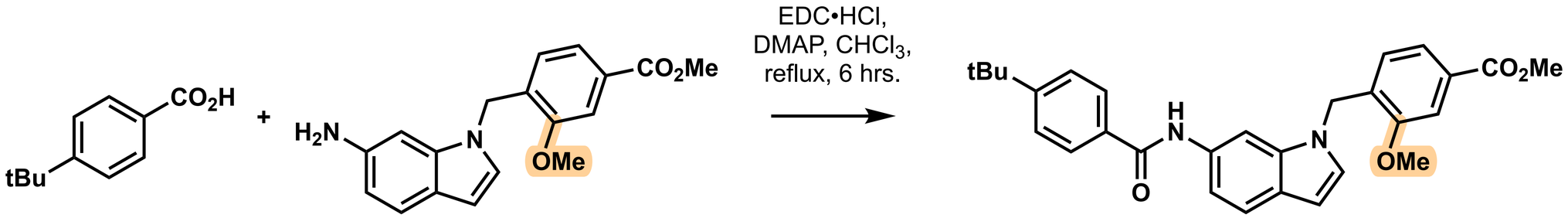
UNIVERSITY OF CAMBRIDGE

Gaussian Processes work best for continuous processes.



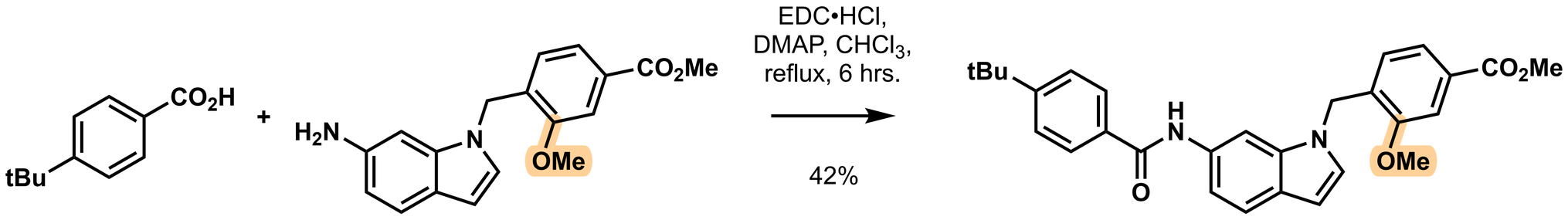
https://2020blogfor.github.io/posts/2020/04/kf_gp/

Discontinuous processes much harder to model



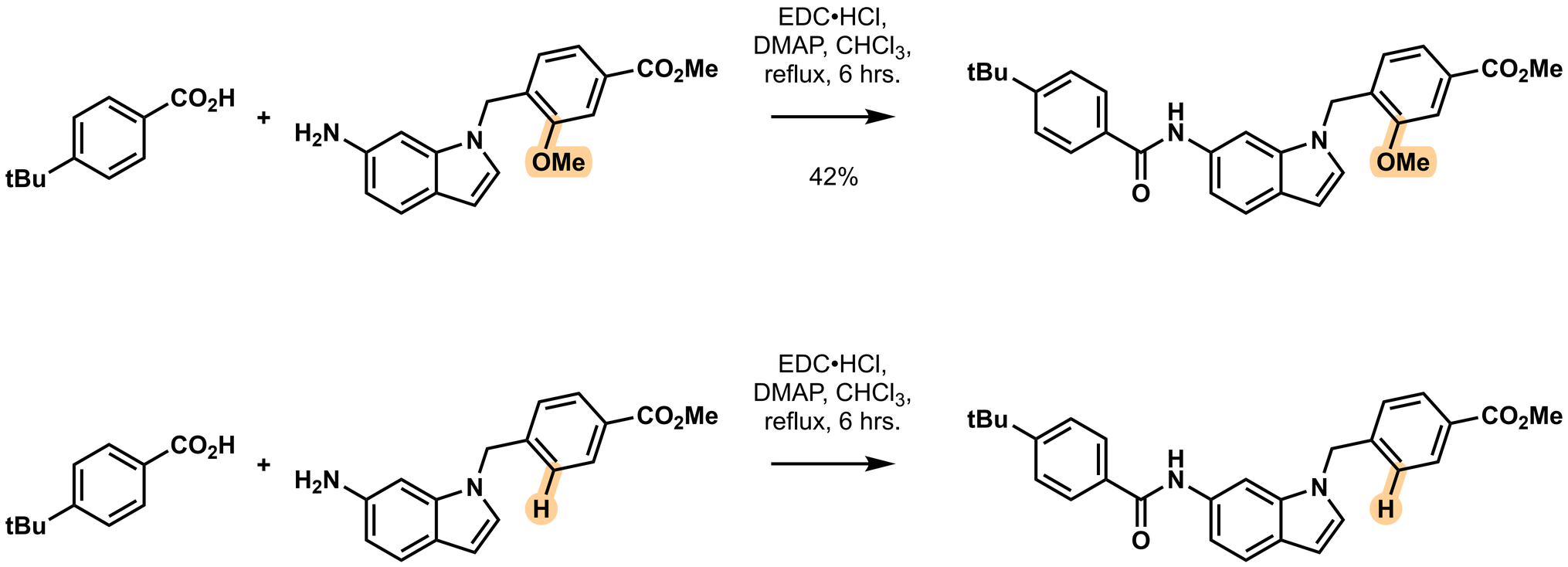
Liu *et al. Chem. Sci.* **2023**, *14*, 10835.

Discontinuous processes much harder to model

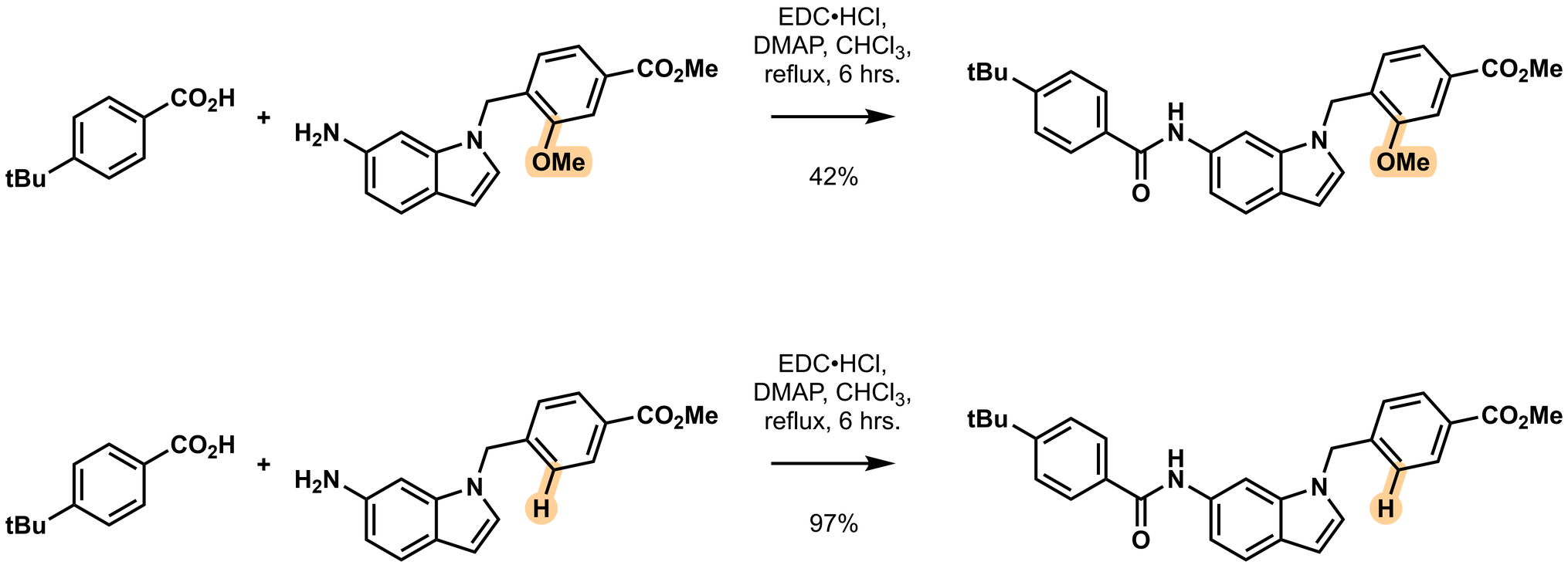


Liu *et al. Chem. Sci.* **2023**, *14*, 10835.

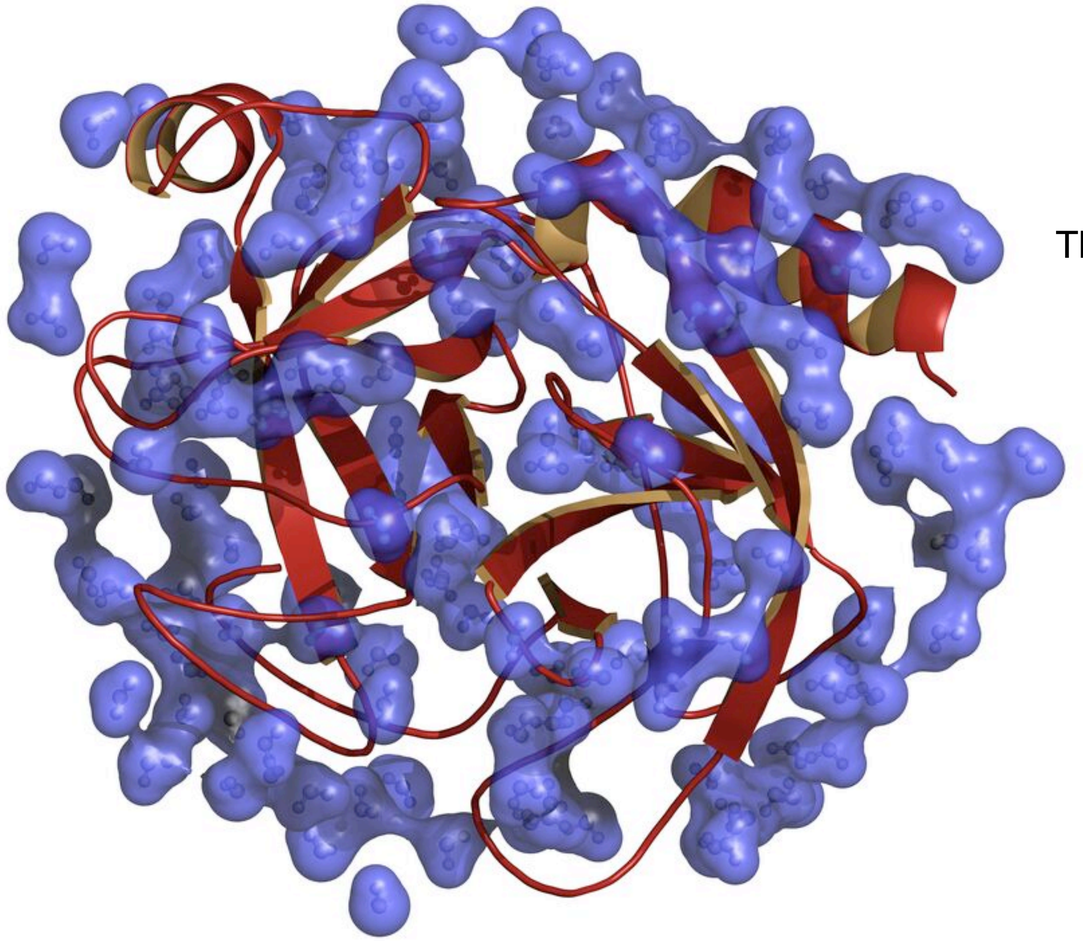
Discontinuous processes much harder to model



Discontinuous processes much harder to model



Enzymes!



There are a lot of enzymes in the world.

There are a lot of enzymes with unknown function.

A lot of metabolites made from unknown enzymes.

Can we predict unknown reactivity?

Some Real World Examples of Gaussian Processes

Does this enzyme catalyze this reaction?

Mellor *et al. ACS Synth. Biol.* **2016**, *5*, 518.

Some Real World Examples of Gaussian Processes

Is there affinity between the enzyme and the starting materials?

Mellor *et al. ACS Synth. Biol.* **2016**, *5*, 518.

Some Real World Examples of Gaussian Processes

Is there affinity between the enzyme and the starting materials?

predicting K_M

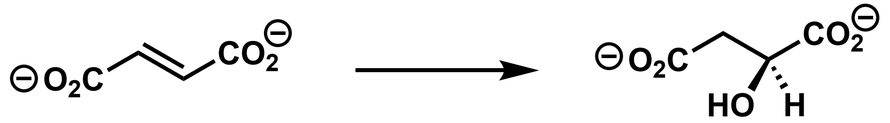
Mellor *et al. ACS Synth. Biol.* **2016**, *5*, 518.

Some Real World Examples of Gaussian Processes

Is there affinity between the enzyme and the starting materials?

predicting K_M

MYRALRLLARSRPLVRAPAAALASAPGLGGAAVPSFWPPNAARMASQNSFRIEYDTFGEL
KVPNDKYYGAQTVRSTMNFKIGGVTERMPTPVIKAFGILKRAAAEVNQDYGLDPKIANAI
MKAADEVAEGKLNDHFPLVVWQTGSGTQTNMNVNEVISNRAIEMLGGELGSKIPVHPNDH
VNKSQSSNDTFPTAMHIAAAIEVHEVLLPGLQKLHDALDAKSKEFAQIIKIGRTHTQDAV PLTL
TGLPFVTAPNKFEALAAHDALVELSGAMNTTACSLMKIANDIRFLGSGPRSGLGELILPE
NEPGSSIMPGKVNPTQCEAMTMVAAQVMGNHVAVTVGGSNGHFELNVFKPMMIKNVLHSA
RLLGDASVSFTENCVVGIQANTERINKLMNESLMLVTALNPHIGYDKAAKIAKTAHKNGS
TLKETAIELGYLTAEQFDEWVKPKDMLGPK



fumarate (*S*)-malate

Mellor *et al. ACS Synth. Biol.* **2016**, *5*, 518.

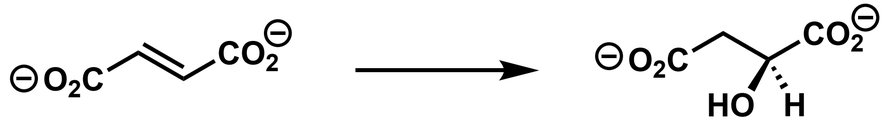
Some Real World Examples of Gaussian Processes

Is there affinity between the enzyme and the starting materials?

predicting K_M

MYRALRLLARSRPLVRAPAAALASAPGLGGAAVPSFWPPNAARMASQNSFRIEYDTFGEL
KVPNDKYYGAQTVRSTMNFKIGGVTERMPTPVIKAFGILKRAAAEVNQDYGLDPKIANAI
MKAADEVAEGKLNDHFPLVVWQTGSGTQTNMNVNEVISNRAIEMLGGELGSKIPVHPNDH
VNKSQSSNDTFPTAMHIAAAIEVHEVLLPGLQKLHDALDAKSKEFAQIIKIGRTHTQDAV PLTL
TGLPFVTAPNKFEALAAHDALVELSGAMNTTACSLMKIANDIRFLGSGPRSGLGELILPE
NEPGSSIMPGKVNPTQCEAMTMVAAQVMGNHVAVTVGGSNGHFELNVFKPMMIKNVLHSA
RLLGDASVSFTENCVVGIQANTERINKLMNESLMLVTALNPHIGYDKAAKIAKTAHKNGS
TLKETAIELGYLTAEQFDEWVKPKDMLGPK

fumarate hydratase sequence

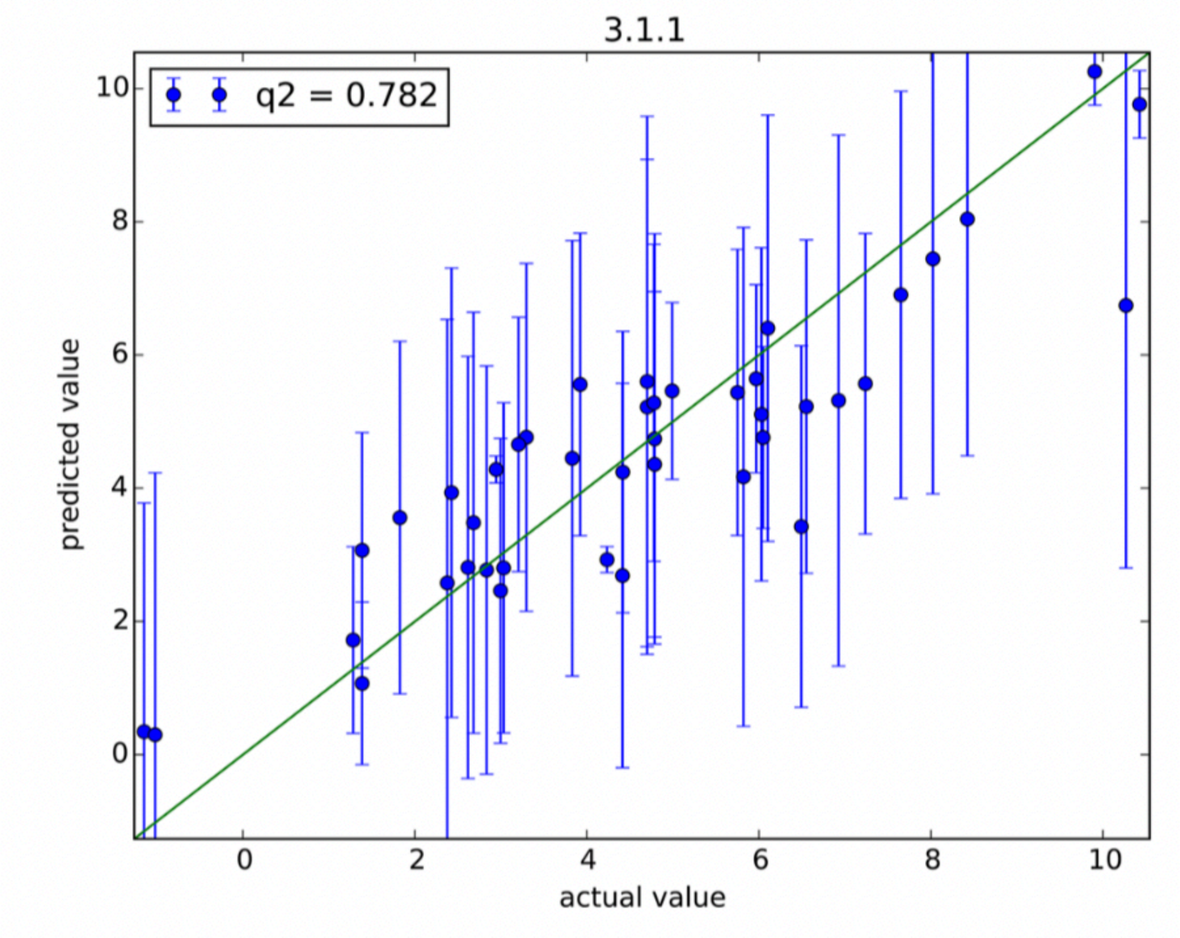


fumarate (*S*)-malate

K_M = 5 x 10^{-6} M

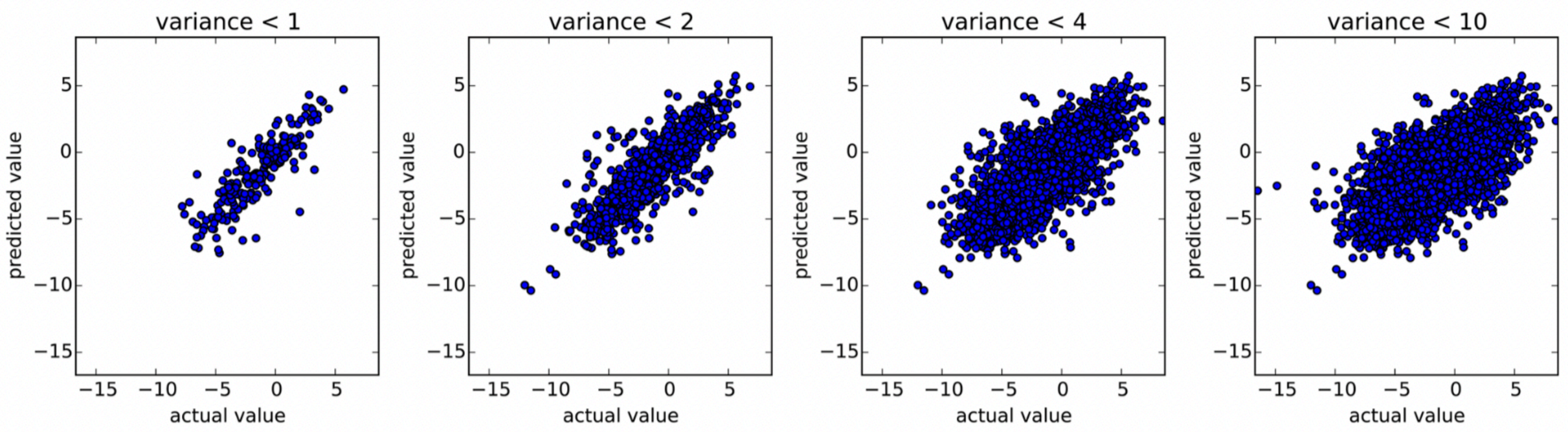
Mellor *et al. ACS Synth. Biol.* **2016**, *5*, 518.

Some Real World Examples of Gaussian Processes



Mellor *et al. ACS Synth. Biol.* **2016**, *5*, 518.

Some Real World Examples of Gaussian Processes



Mellor *et al. ACS Synth. Biol.* **2016**, *5*, 518.

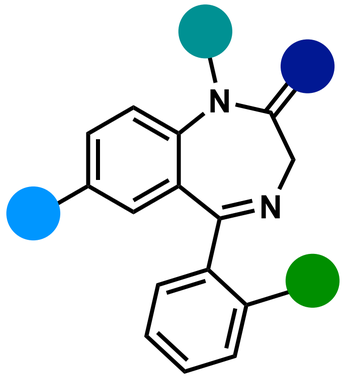
UNIVERSITY OF CAMBRIDGE

ADME Property optimization

Absorption, Distribution, Metabolism, Excretion

Obrezanova *et al. J. Chem. Inf. Model.* **2007**, *47*, 2007.

Some Real World Examples of Gaussian Processes



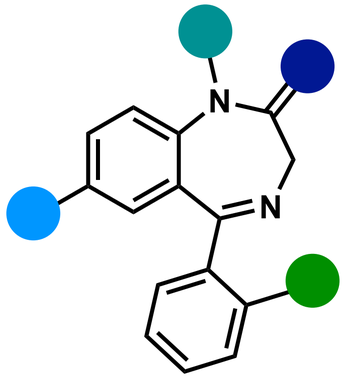
245 benzodiazepines

predicting in vitro
binding affinities (pIC_{50})

Table 1. BZD Set with Molecular Indices Descriptors

		Our Results[a]					
		training set			test set		
method	desc.	RMSE	R^2	r_{corr}^2	RMSE	R^2	r_{corr}^2
PLS	3[b]	0.63	0.32	0.32	0.52	0.51	0.53
GP-Basic	38	0.53	0.51	0.52	0.52	0.51	0.53
GP-FVS	15	0.53	0.51	0.52	0.52	0.52	0.54
GP-Opt	9	0.47	0.62	0.62	0.55	0.47	0.51
GP-Nest	38	0.44 (0.14)[c]	0.67	0.68	0.46 (0.15)[c]	0.63	0.65

Obrezanova *et al. J. Chem. Inf. Model.* **2007**, *47*, 2007.



245 benzodiazepines

predicting in vitro binding affinities (pIC_{50})

Table 1. BZD Set with Molecular Indices Descriptors

Our Results[a]

		training set			test set		
method	desc.	RMSE	R^2	r_{corr}^2	RMSE	R^2	r_{corr}^2
PLS	3[b]	0.63	0.32	0.32	0.52	0.51	0.53
GP-Basic	38	0.53	0.51	0.52	0.52	0.51	0.53
GP-FVS	15	0.53	0.51	0.52	0.52	0.52	0.54
GP-Opt	9	0.47	0.62	0.62	0.55	0.47	0.51
GP-Nest	38	0.44 (0.14)[c]	0.67	0.68	0.46 (0.15)[c]	0.63	0.65

Burden Results[5]

		training set			test set		
method	desc.	SEF[d]	R^2	r_{corr}^2	SEP[e]	R^2	r_{corr}^2
MLR	39	0.18		0.47	0.20		0.32
ANN	22	0.13		0.73	0.14		0.66
BRANN	39	0.12		0.75	0.12		0.71
GPmodel	39	0.12		0.76	0.14		0.66
GPlinear	39	0.12		0.78	0.13		0.71

[a] Training set 208 compounds, test set 37 compounds. [b] Number of PLS components. [c] In brackets: RMSE for unit interval scaled Y values. [d] SEF = Standard error of fit. [e] SEP = Standard error of prediction.

The actual data split into training and test sets used to model BZD by Burden[5] was not available to us.

Obrezanova *et al. J. Chem. Inf. Model.* **2007**, *47*, 2007.