

Machine Learning for Chemists

– *Gaussian Processes* –

19 February 2024



Introduction to Regression

Uncertainty of Predictions

The “big idea” behind Gaussian Processes

Examples of GPs used in the wild

Live Demo!

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?

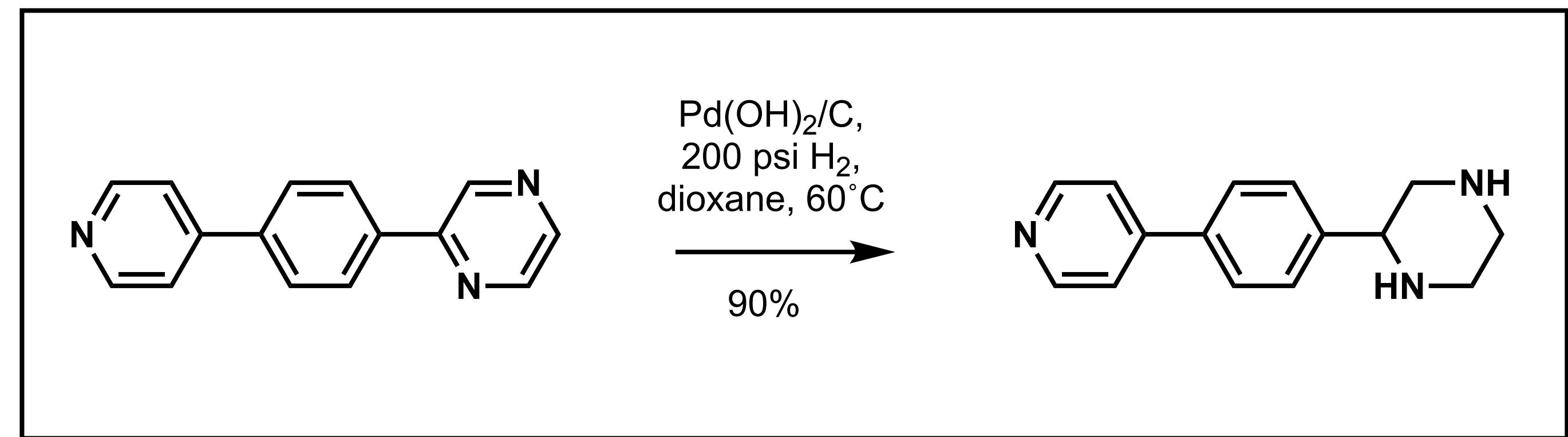


This pet weighs **26 lbs**



Answers a question that asks for a number.

What is the yield for this reaction?

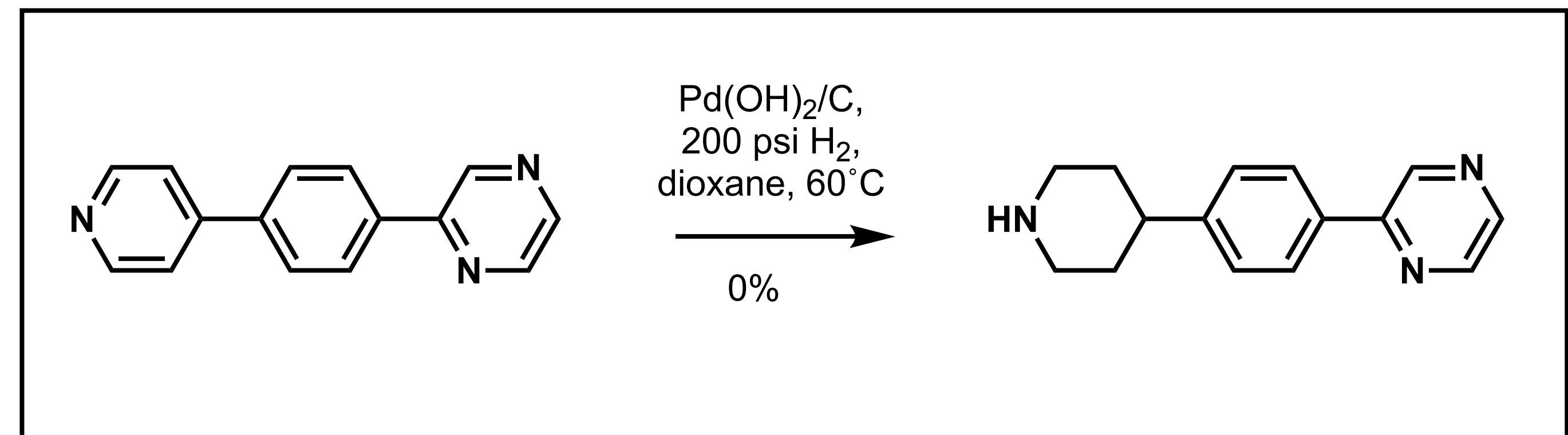


The yield of this reaction is **90%**



Answers a question that asks for a number.

What is the yield for this reaction?



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



Confusion Matrix

What actually happened (ground truth)

		What we predicted (predictions)	
		High Yielding	Not High Yielding
What we predicted (predictions)	High Yielding	True Positive	False Positive
	Not High Yielding	False Negative	True Negative

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F-Score} = \frac{2 \cdot \text{TP}}{\text{TP} + \text{TN} + \text{FP}}$$

Metrics quantify how far away from reality the predictions are.

Common Regression Metrics:

Mean Absolute Error (MAE)

Common Regression Metrics:

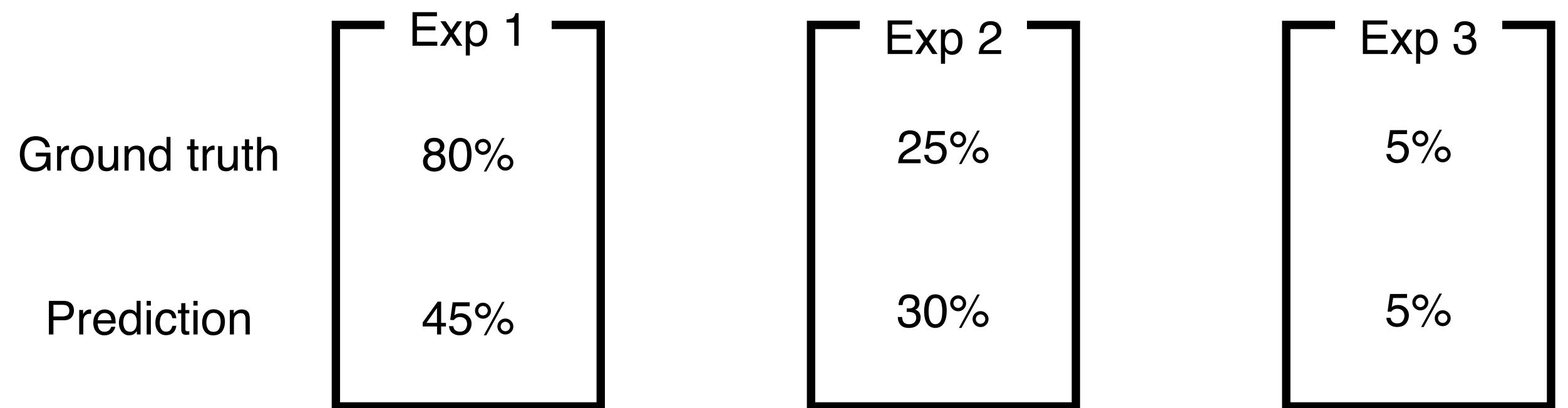
Mean Absolute Error (MAE)

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

Common Regression Metrics:

Mean Absolute Error (MAE)

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



Absolute Error

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%		

$$\text{Abs Error Exp 1} = | 80\% - 45\% | = 35\%$$

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%	

$$\text{Abs Error Exp 2} = | 25\% - 30\% | = 5\%$$

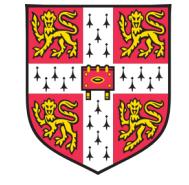
Common Regression Metrics:

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%

$$\text{Abs Error Exp 3} = | 5\% - 5\% | = 0\%$$



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%

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

Common Regression Metrics:

Root Mean Squared Error (RMSE)

Common Regression Metrics:

Root Mean Squared Error (RMSE)

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

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%**

$$\text{Sqr Err Exp 1} = (\text{80\%} - \text{45\%})^2 = \text{1225\%}$$

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%	

$$\text{Sqr Error Exp 2} = (\text{25\%} - \text{30\%})^2 = \text{25\%}$$

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%

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%

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



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%

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

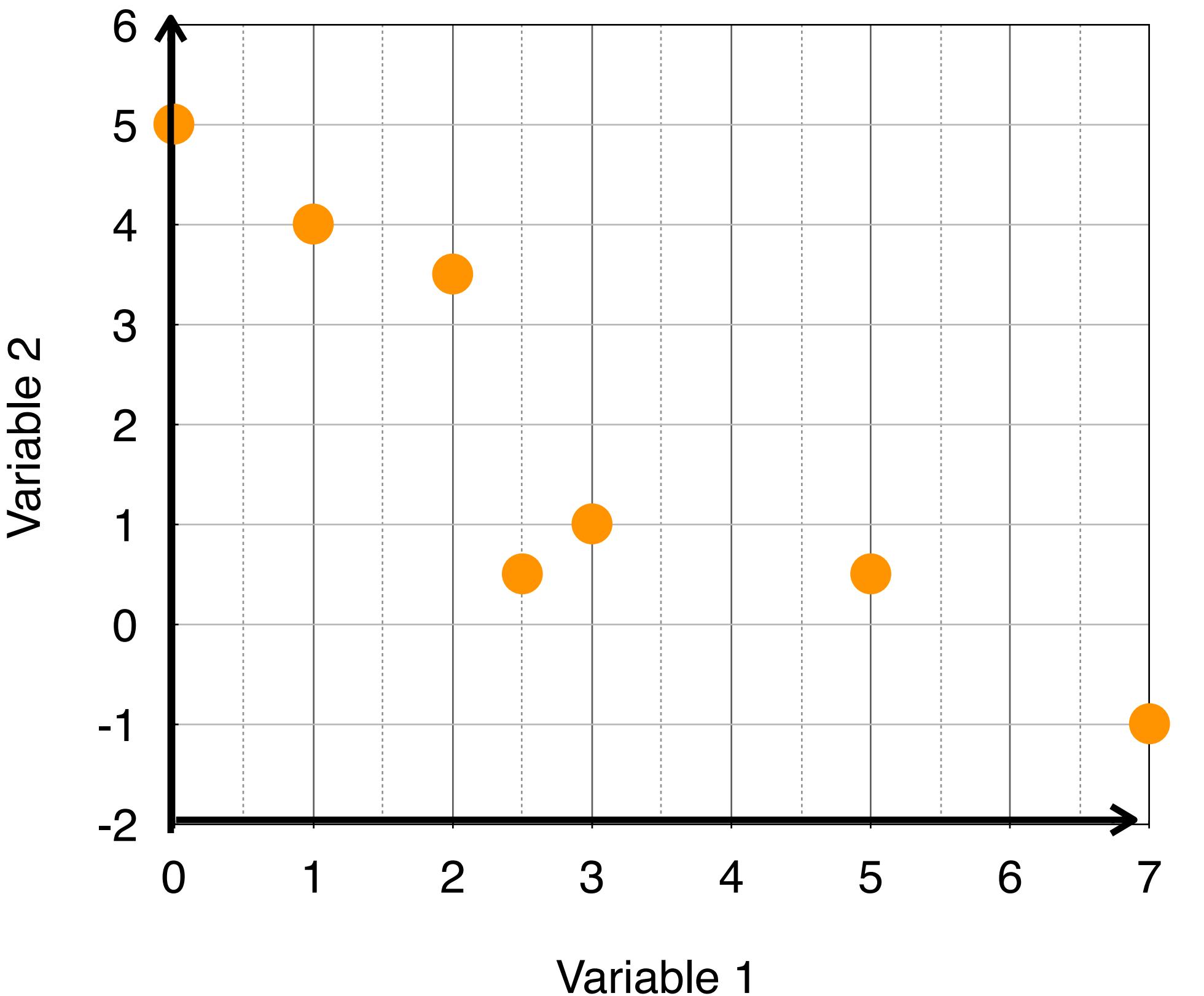
Common Regression Metrics:

R^2 Error

Common Regression Metrics:

R^2 Error

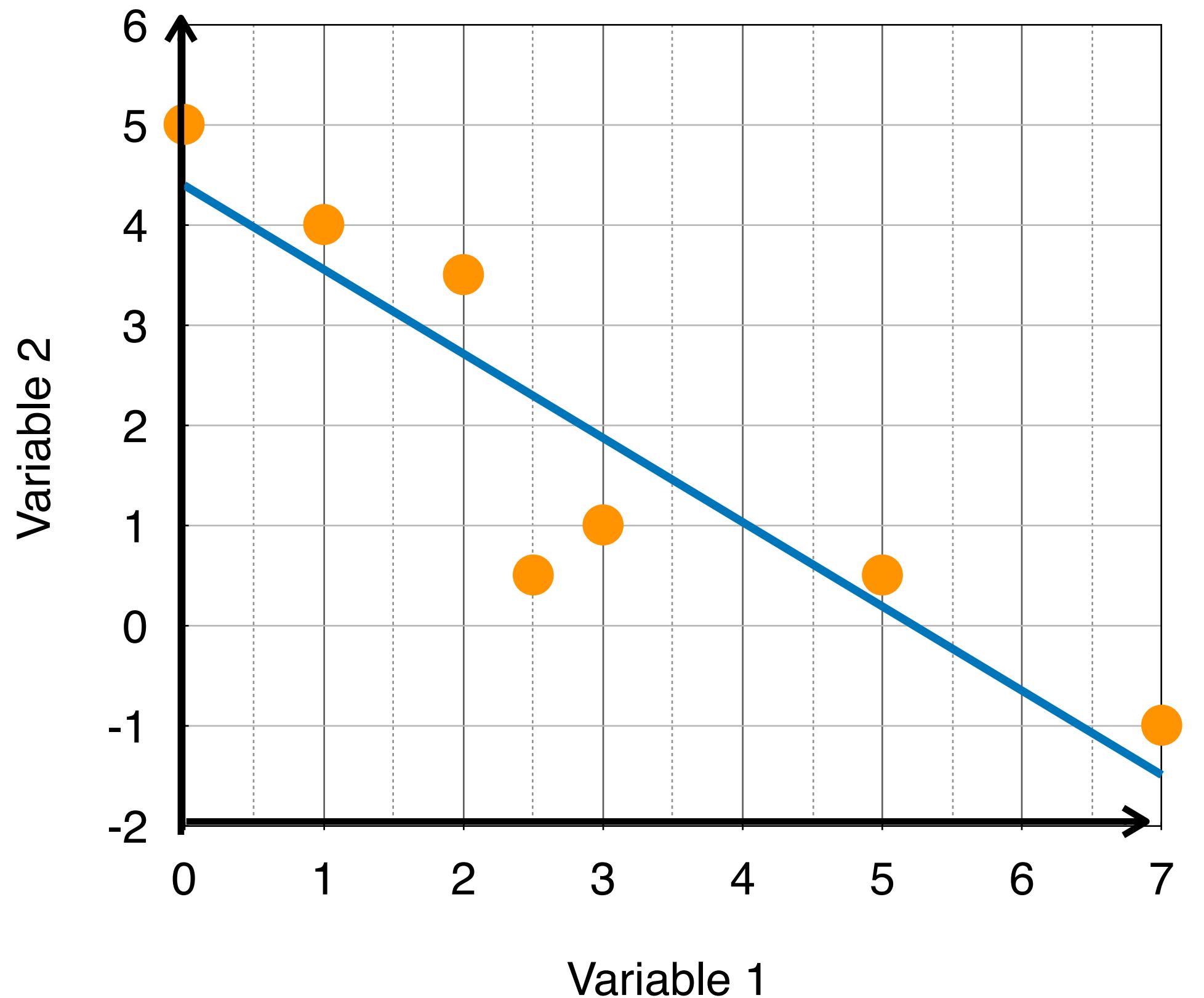
Recall from early graphing



Common Regression Metrics:

R^2 Error

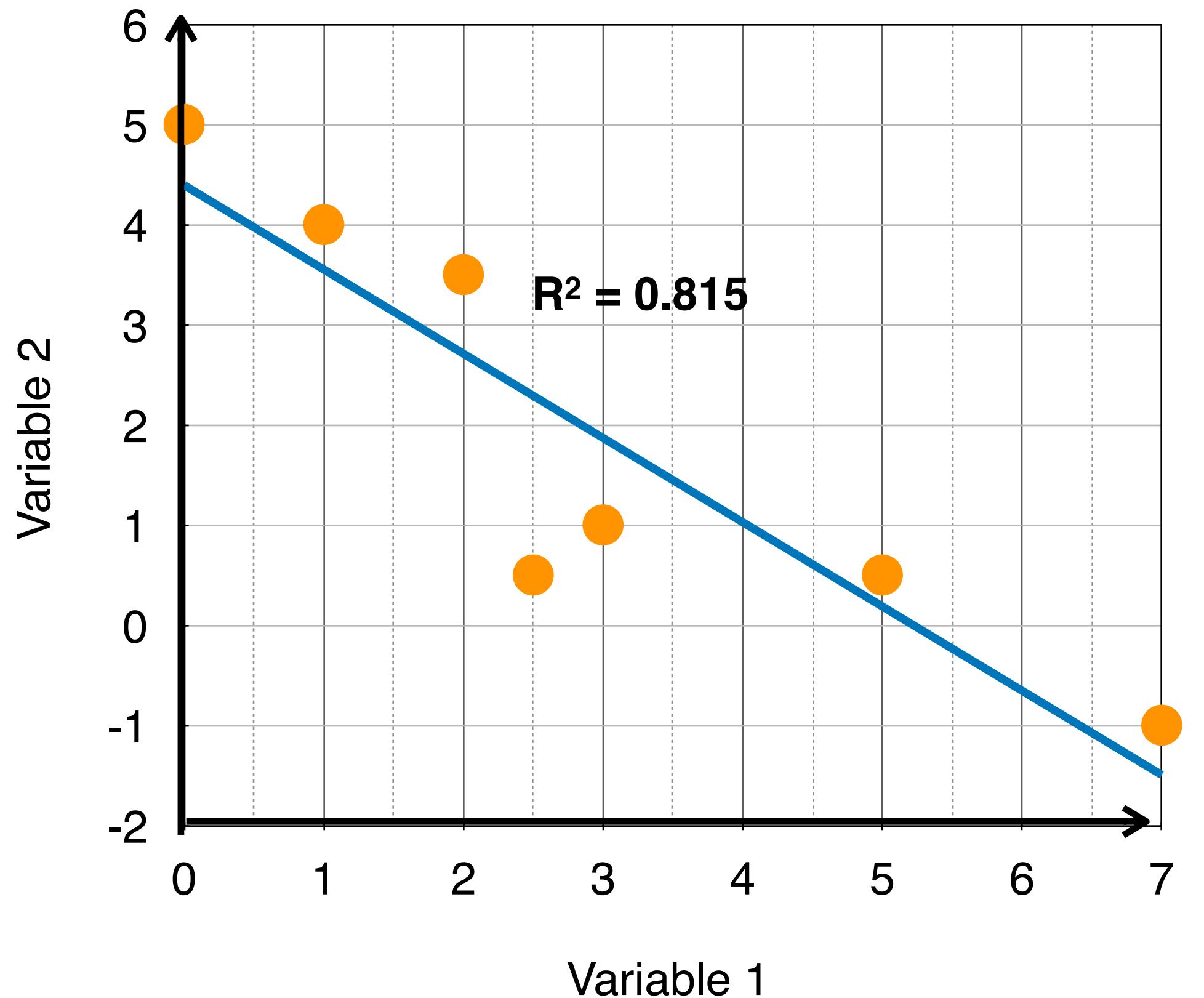
Recall from early graphing



Common Regression Metrics:

R^2 Error

Recall from early graphing



Common Regression Metrics:**R² Error**How well does the line $y = x$ fit the data?

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

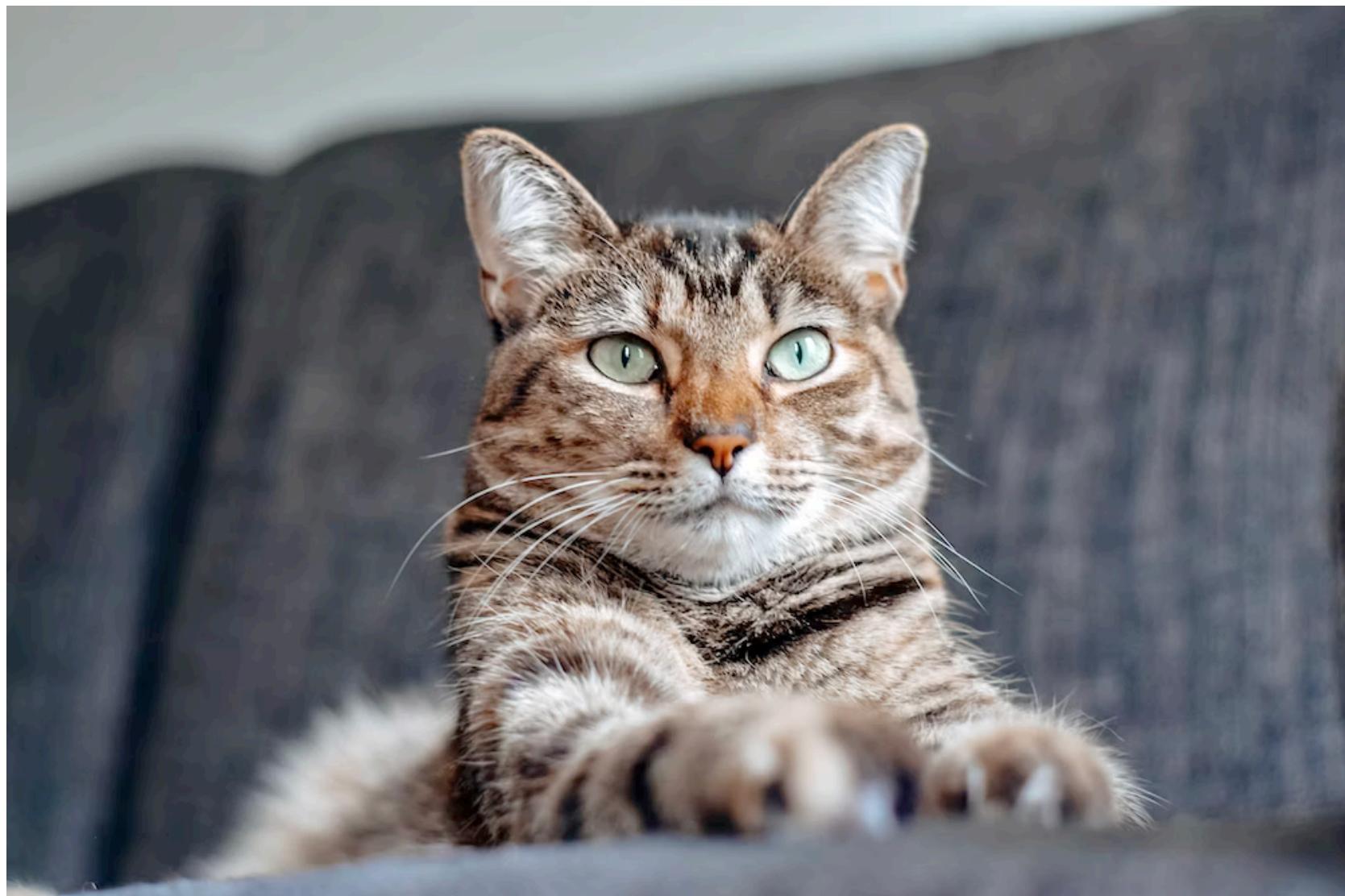
How much does this pet weigh?

How sure of your answer are you?

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

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

How much does this pet weigh?

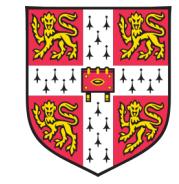
How sure of your answer are you?



This pet weighs **26 lbs**

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

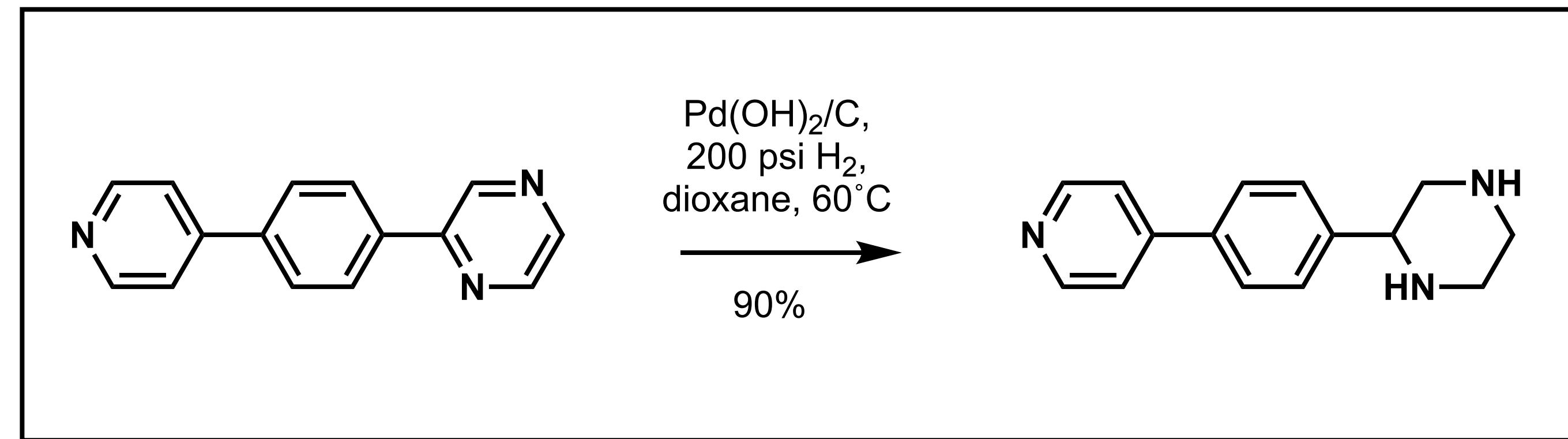
± 0.5 lbs



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?



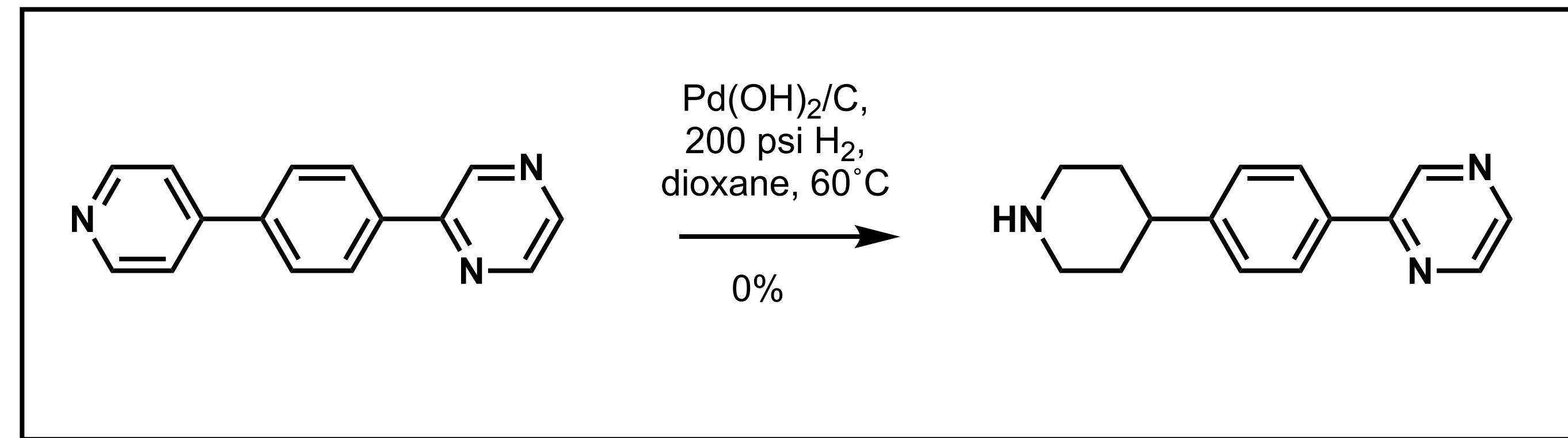
The yield of this reaction is **90% \pm 10%**



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

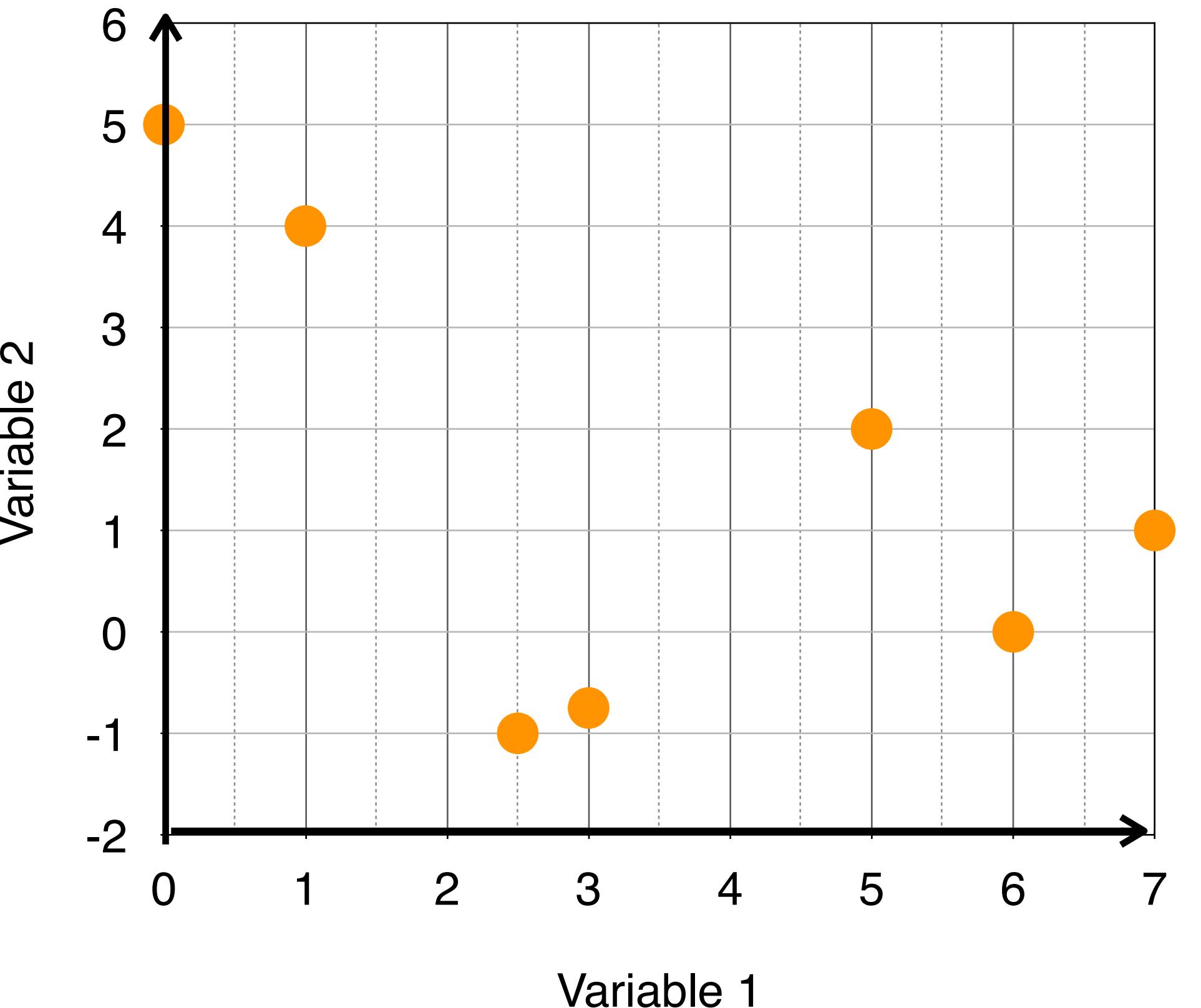
What is the yield of this reaction?

How sure of your answer are you?

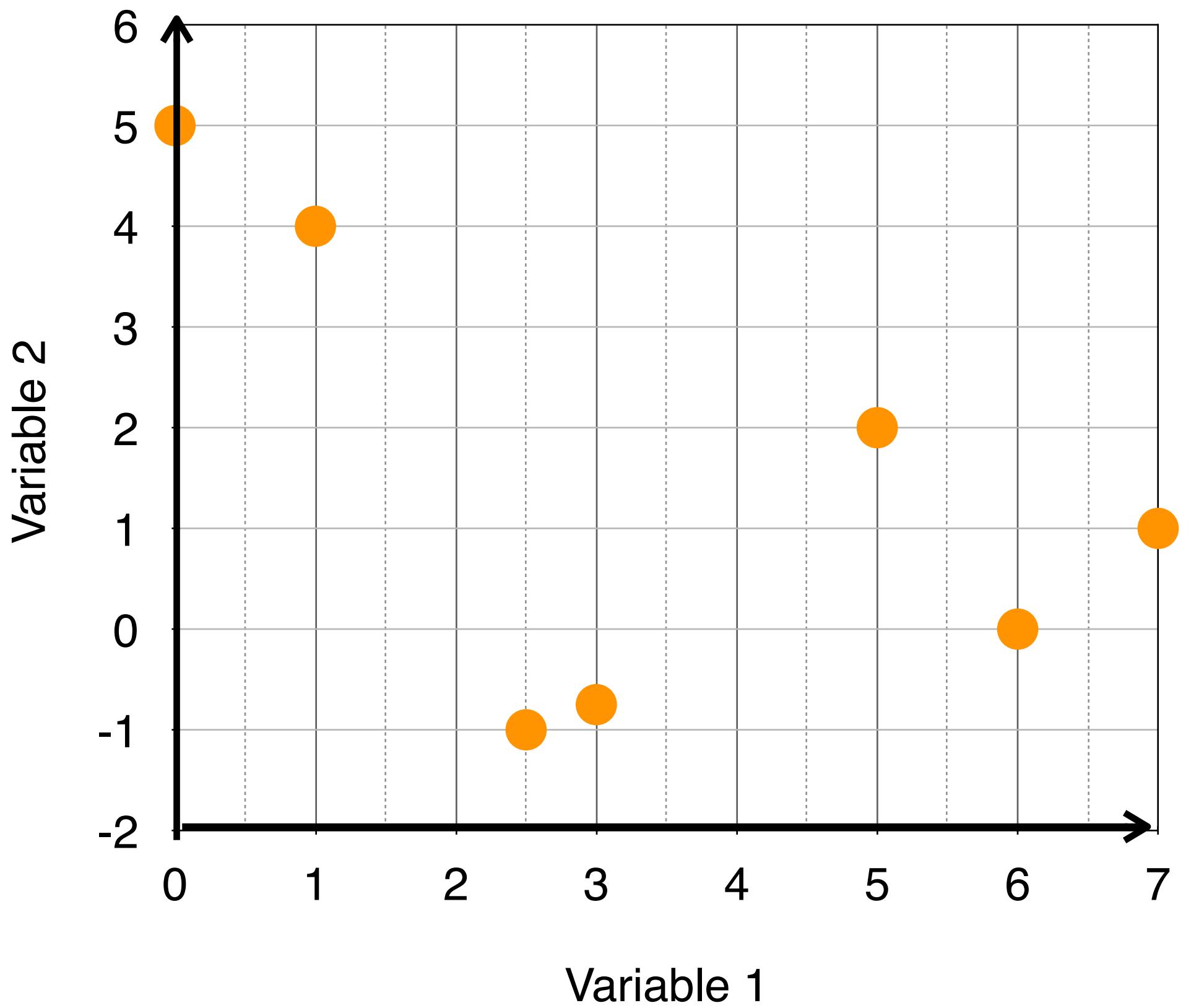
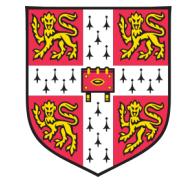


The yield of this reaction is **0% ± 50%**

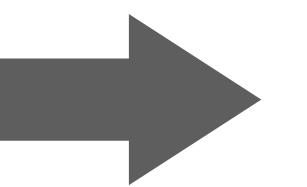
Our Typical Workflow for Regression



Our Typical Workflow for Regression

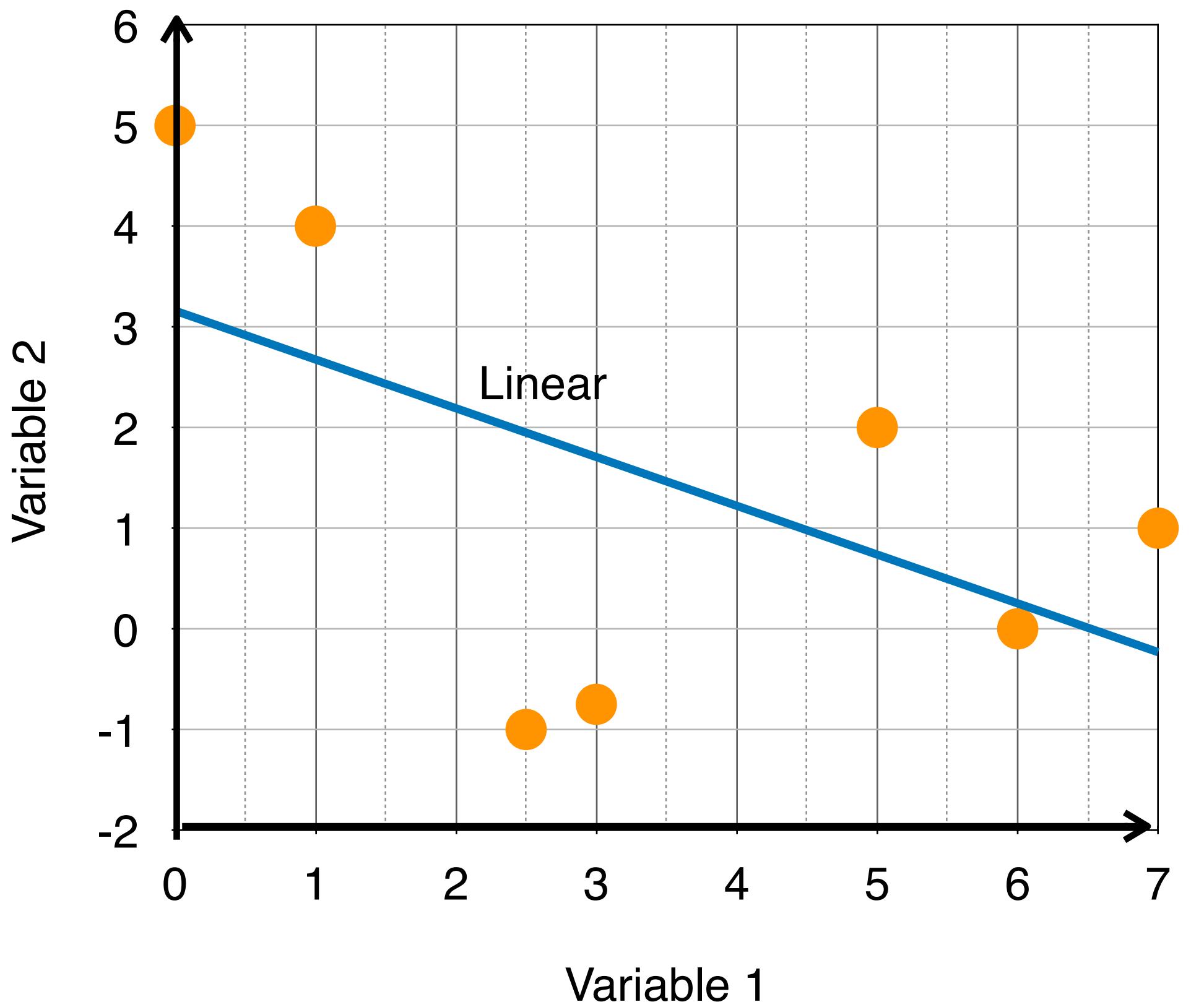


We have data

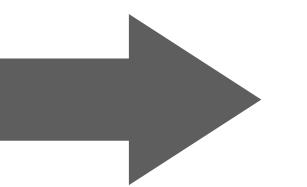


Single style of function
to fit data

Our Typical Workflow for Regression

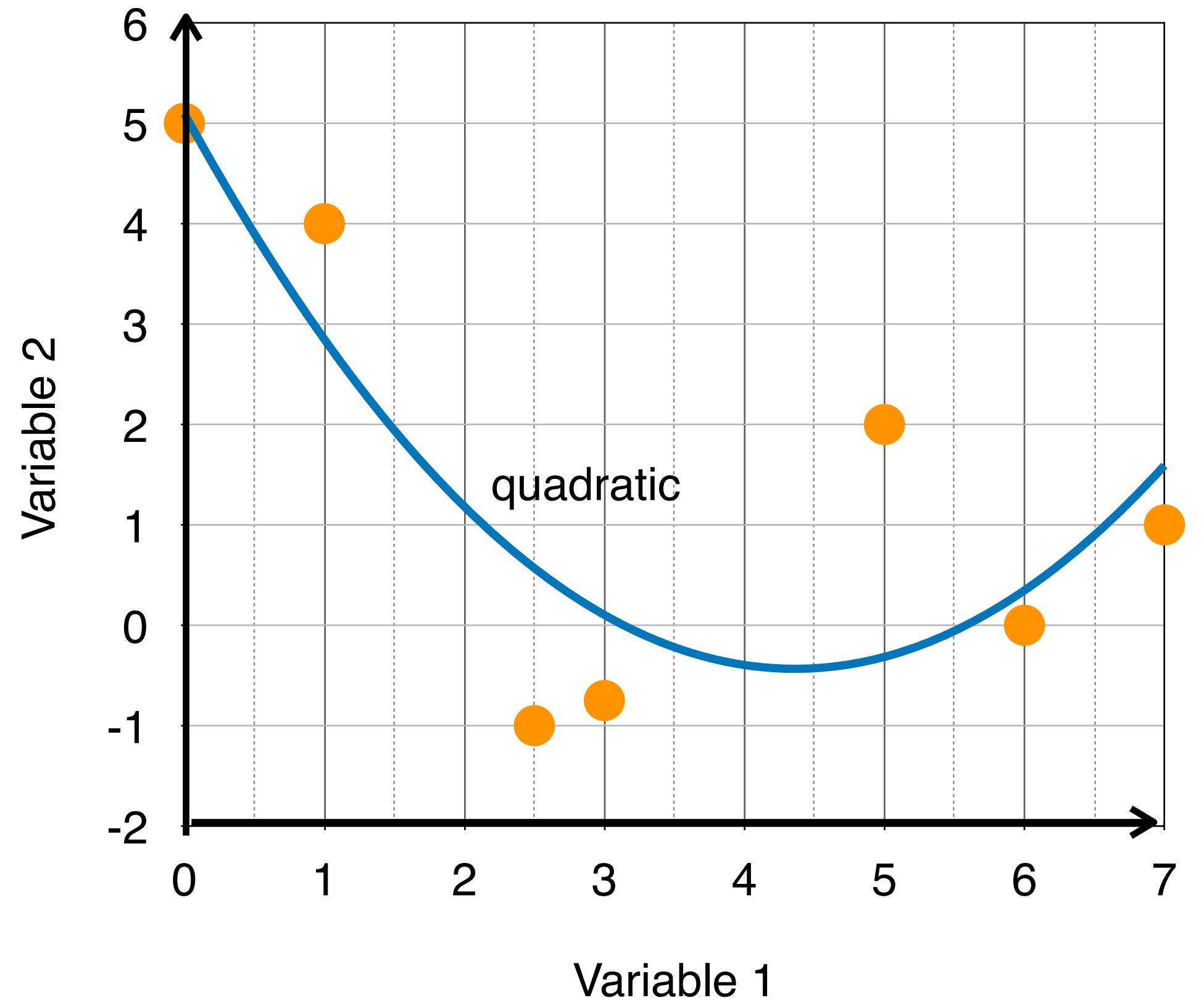


We have data

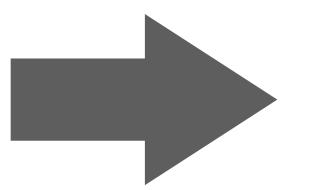


Single style of function
to fit data

Our Typical Workflow for Regression

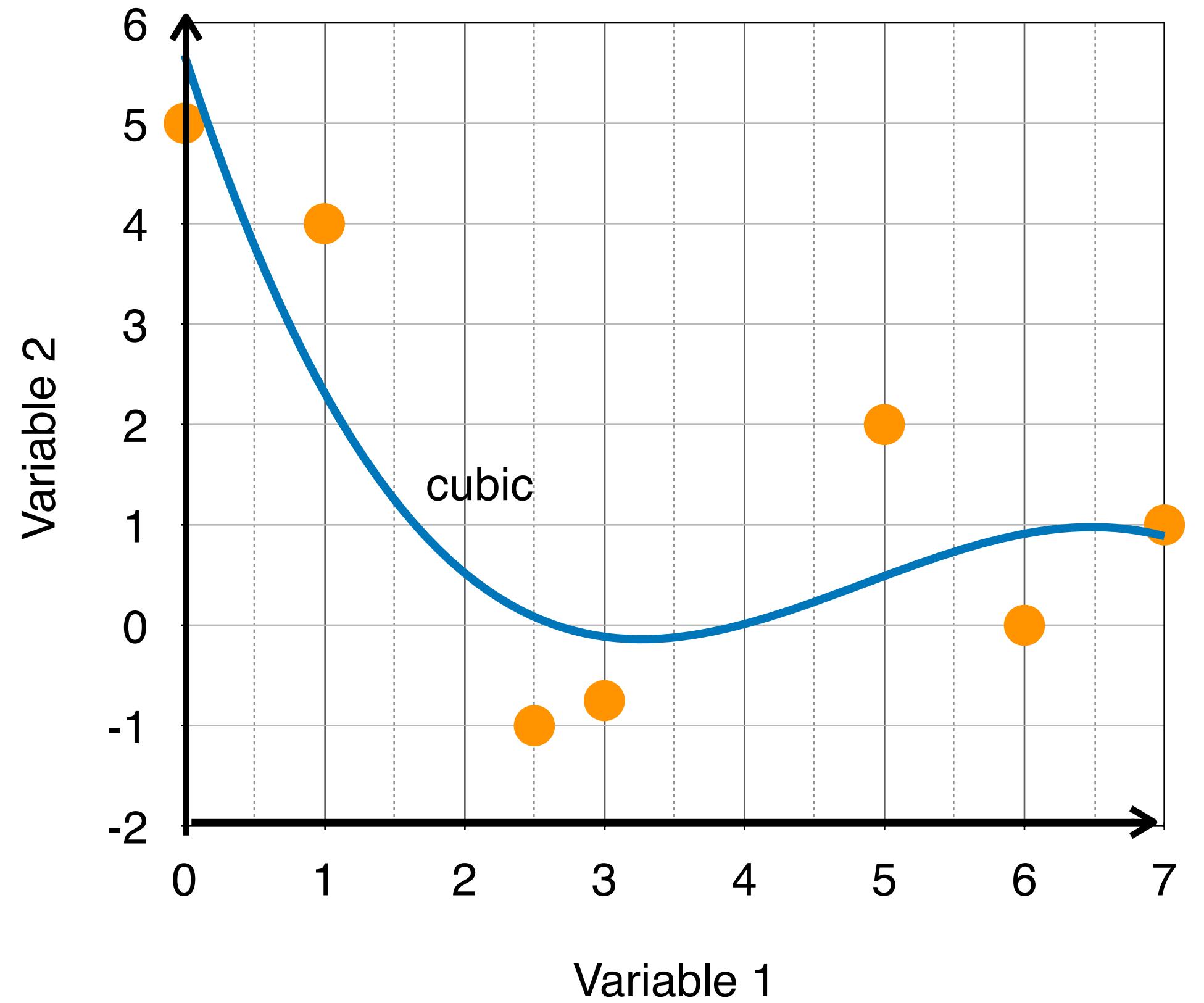


We have data

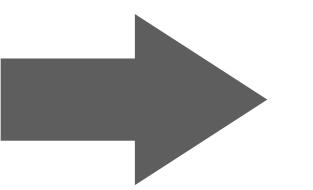


Single style of function
to fit data

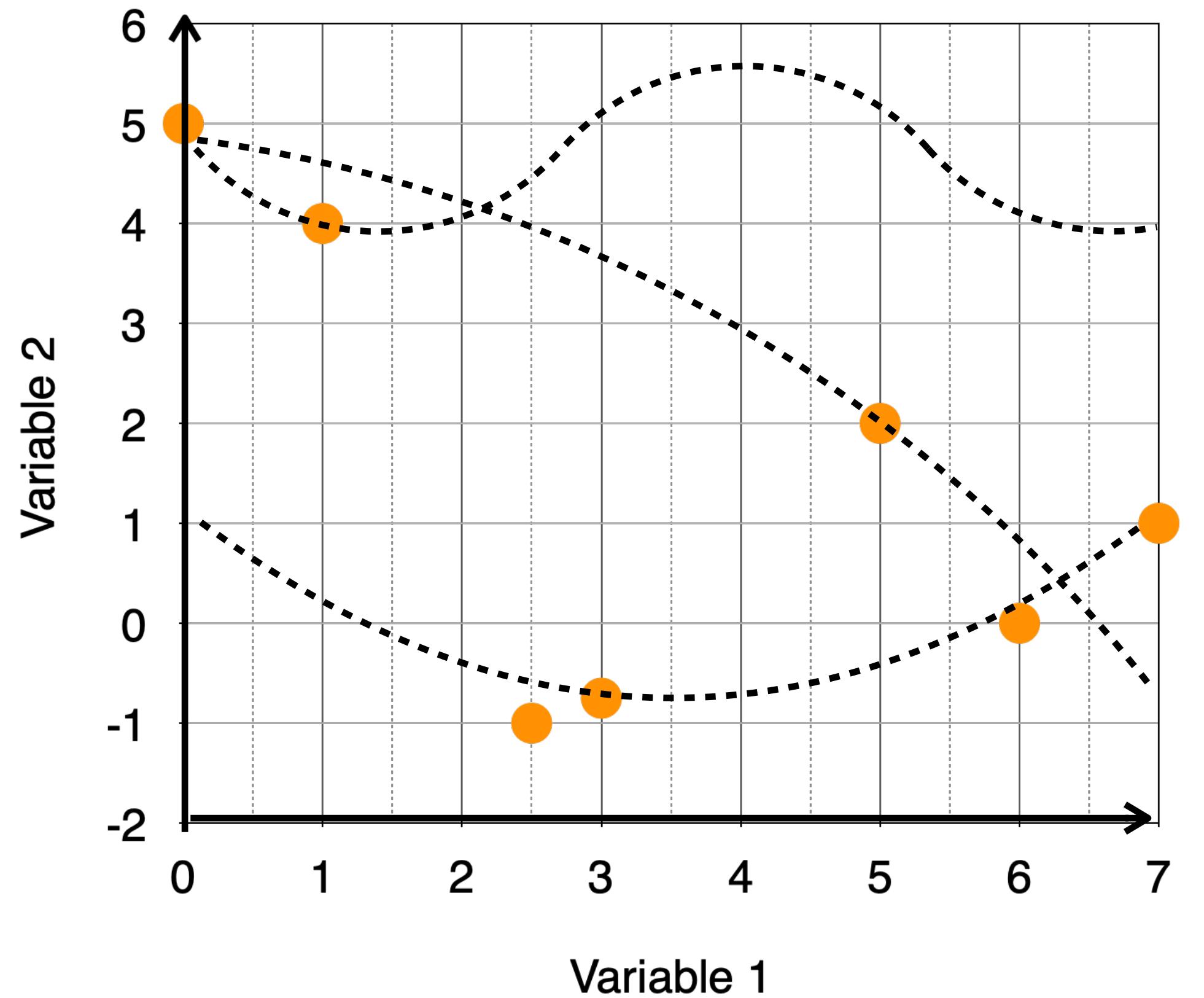
Our Typical Workflow for Regression



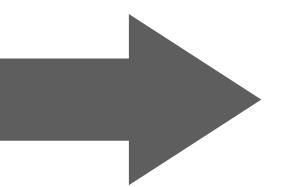
We have data



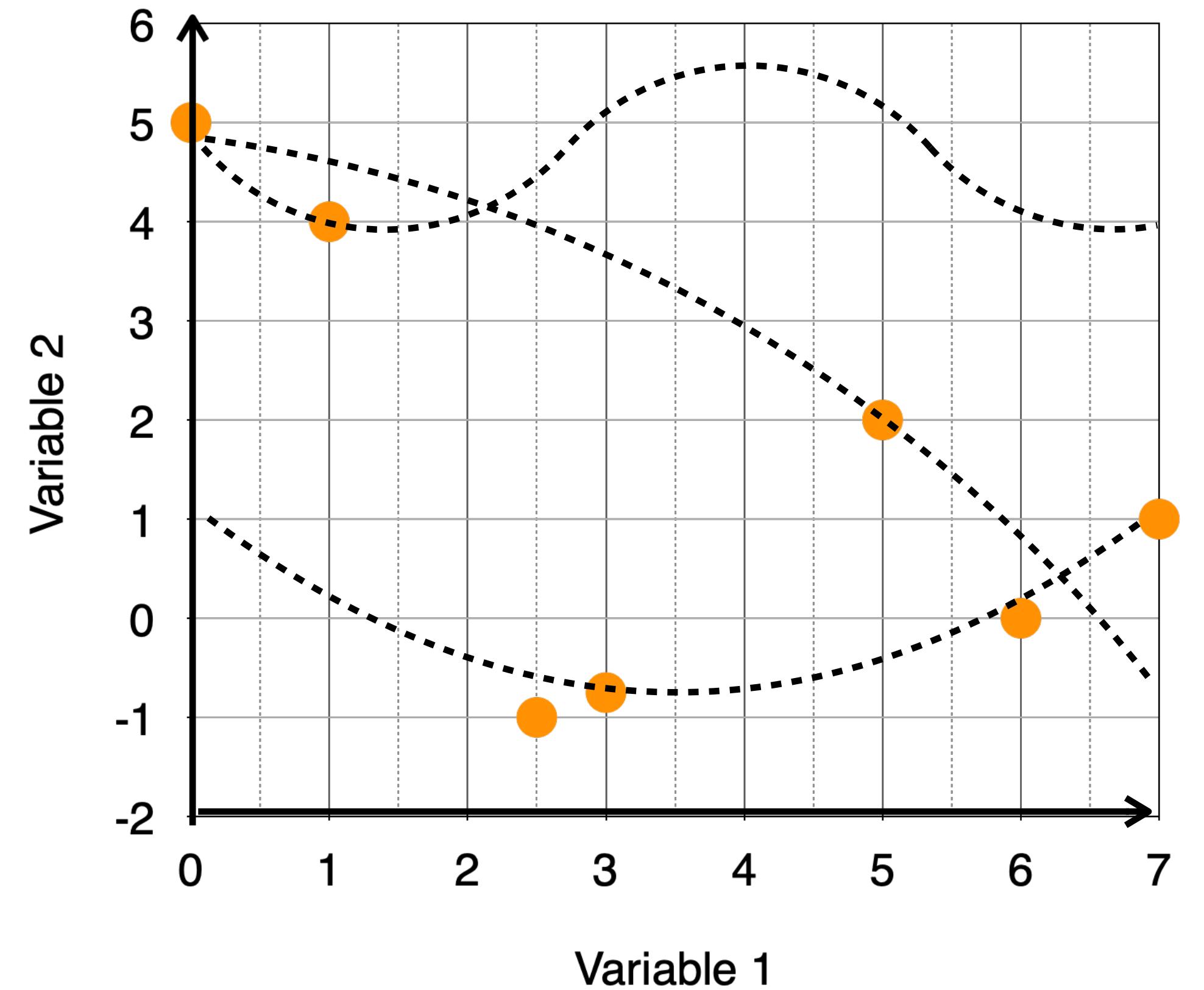
Single style of function
to fit data



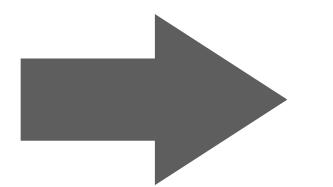
We have data



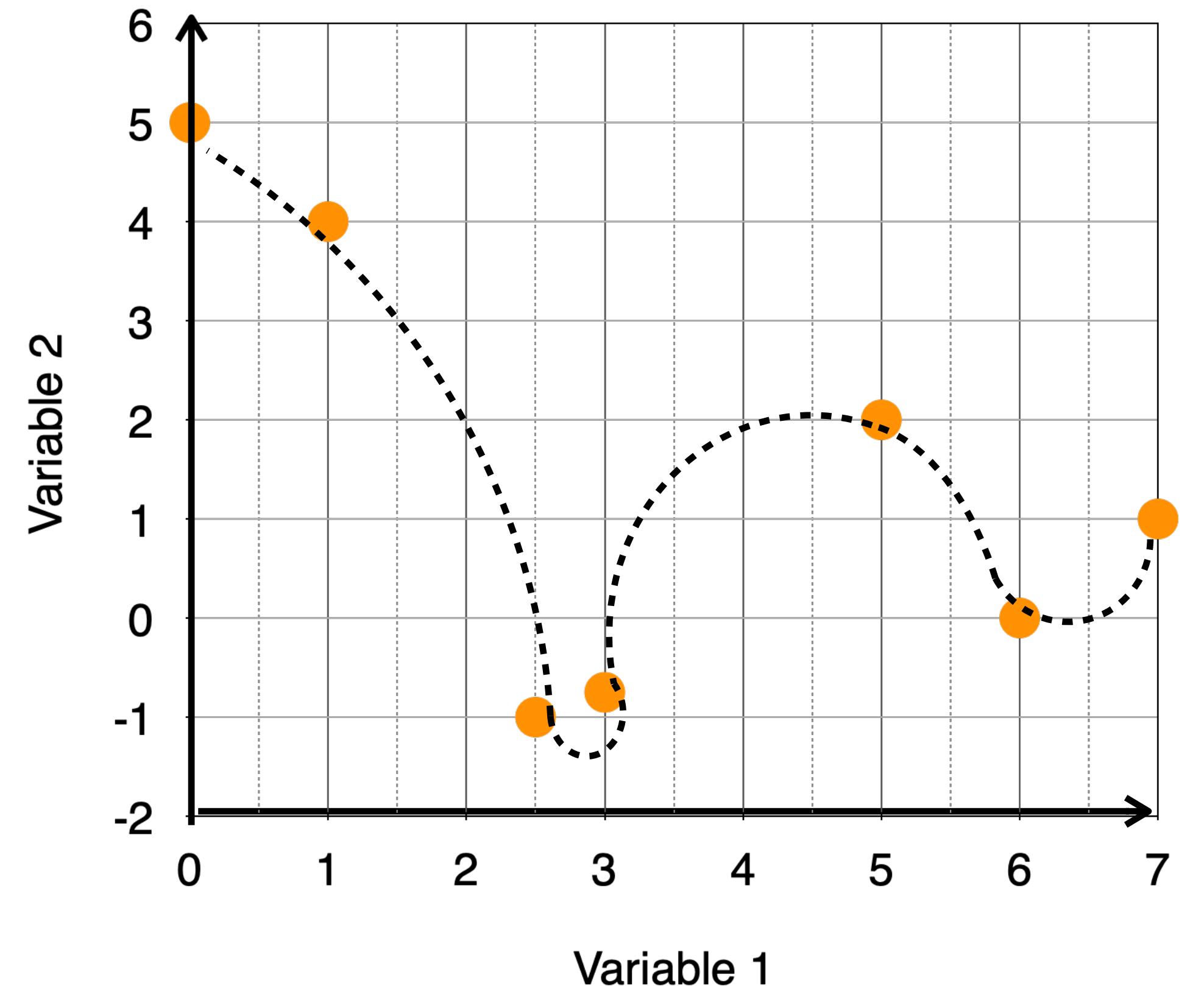
All possible functions
that fit the data



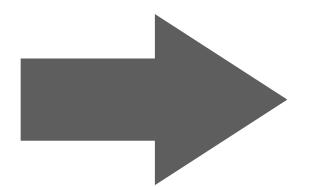
We have data



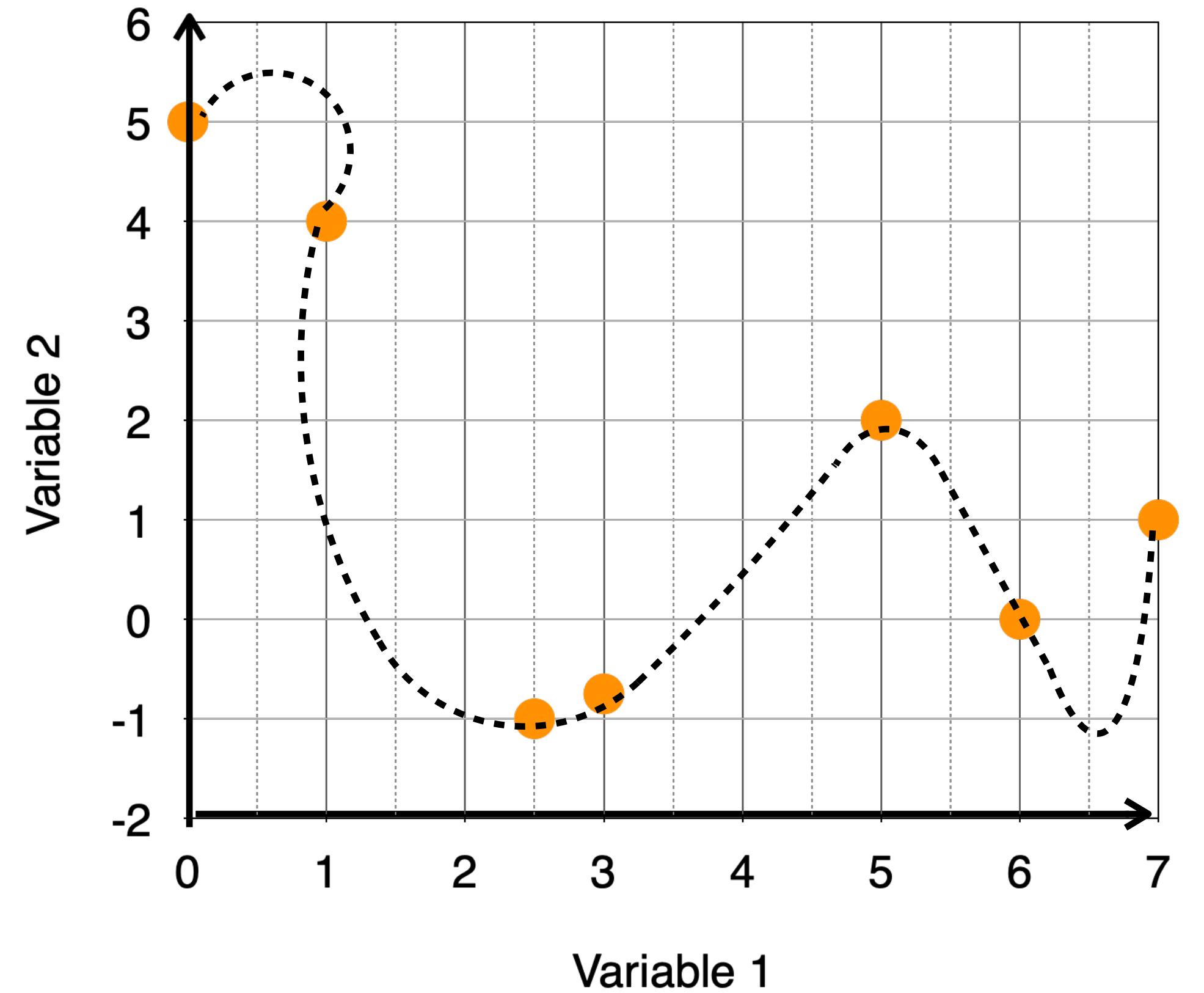
All possible functions that
go through ALL our points



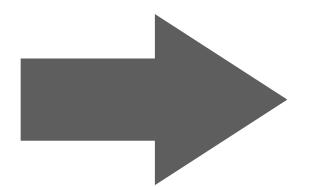
We have data



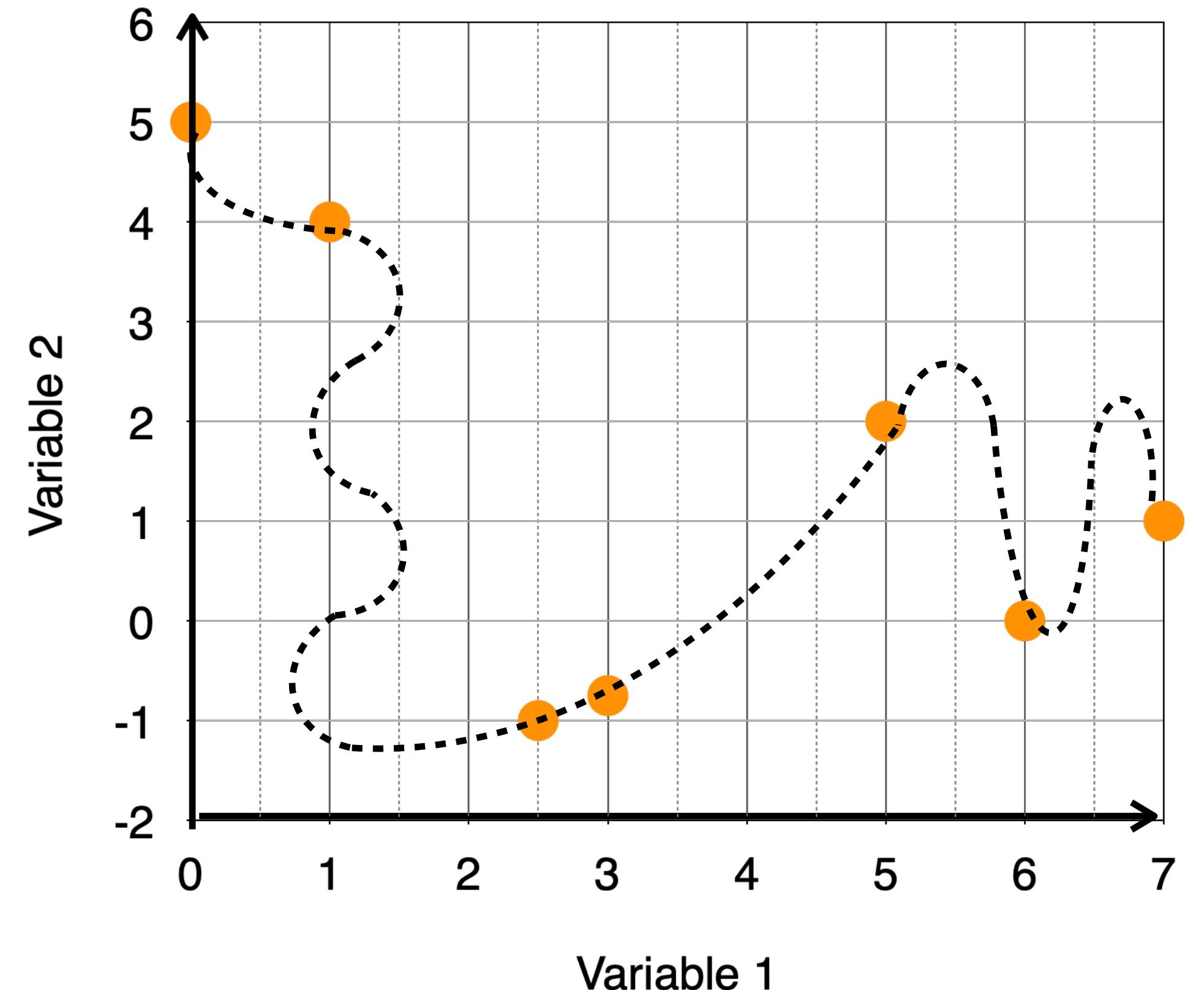
All possible functions that
go through ALL our points



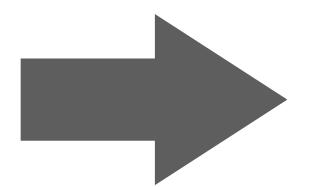
We have data



All possible functions that
go through ALL our points

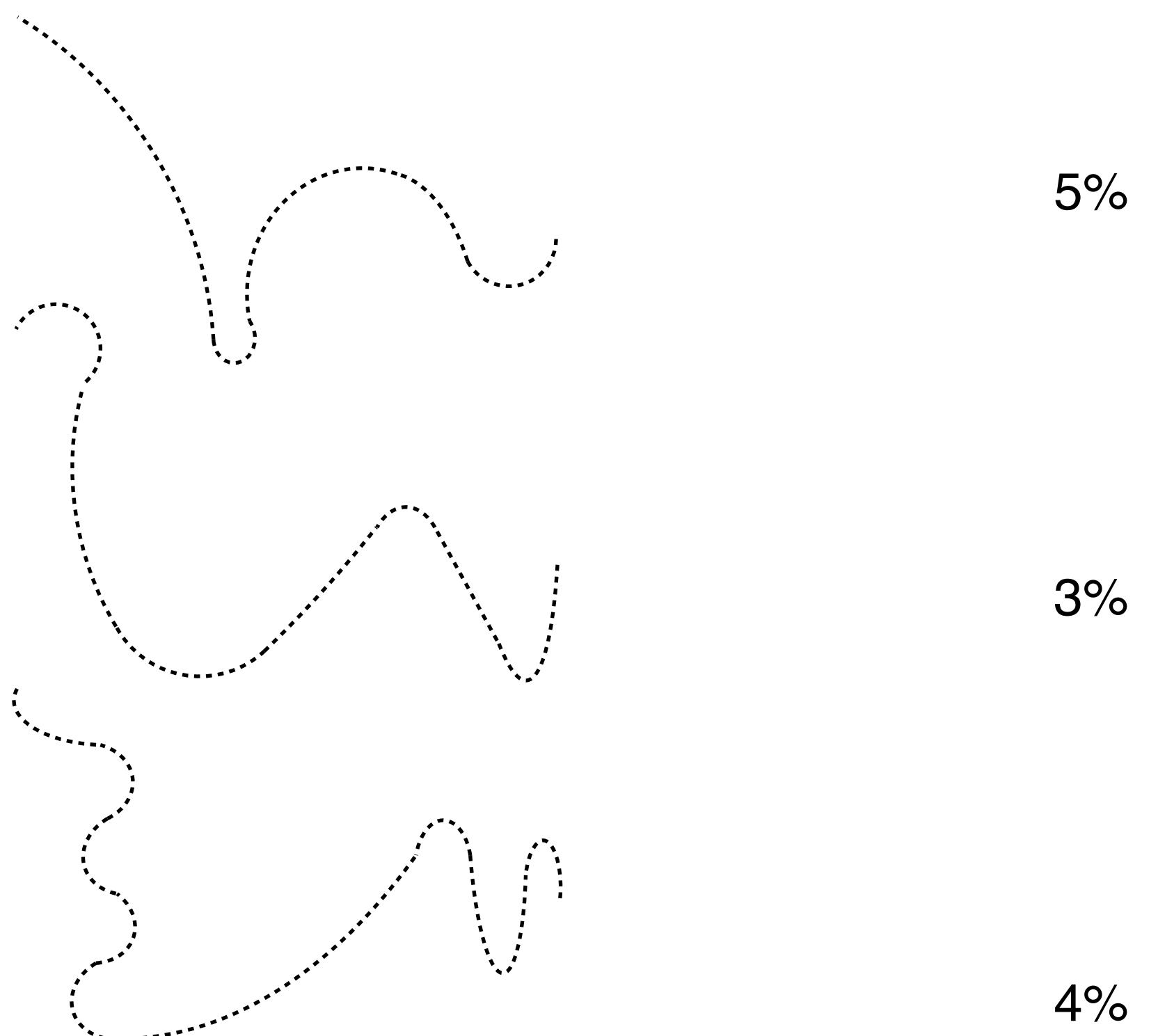


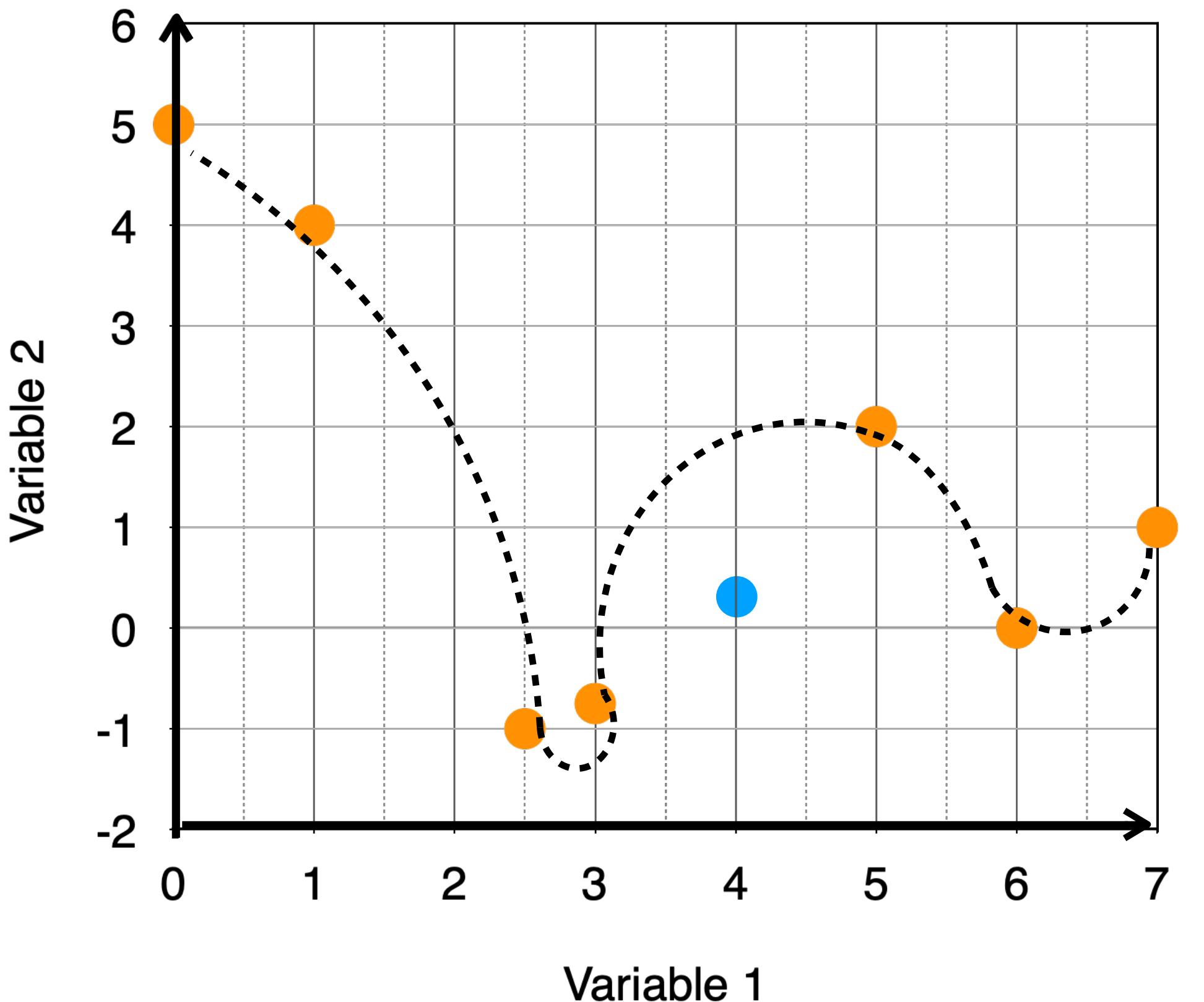
We have data



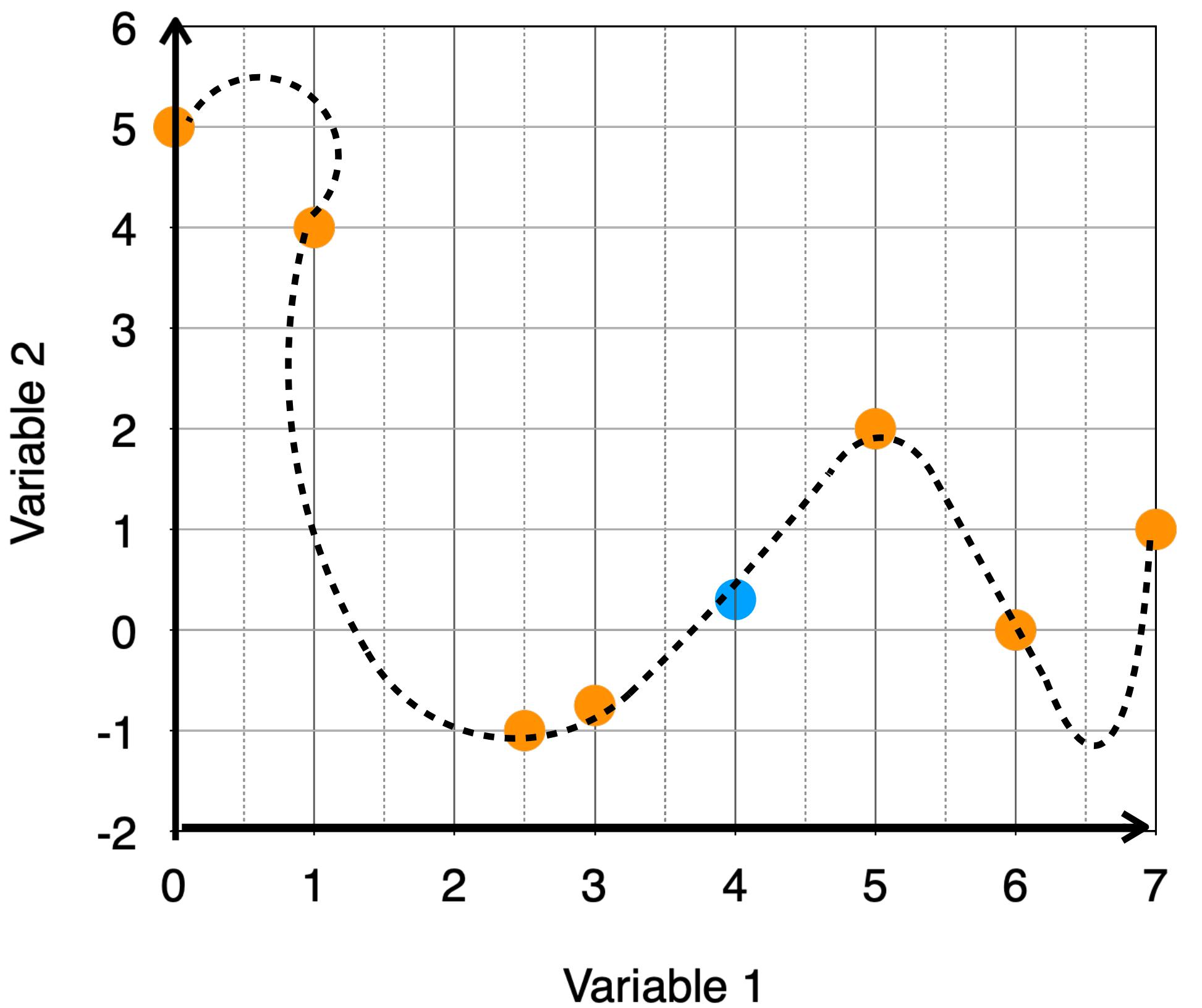
All possible functions that
go through ALL our points

Possible Function Probability

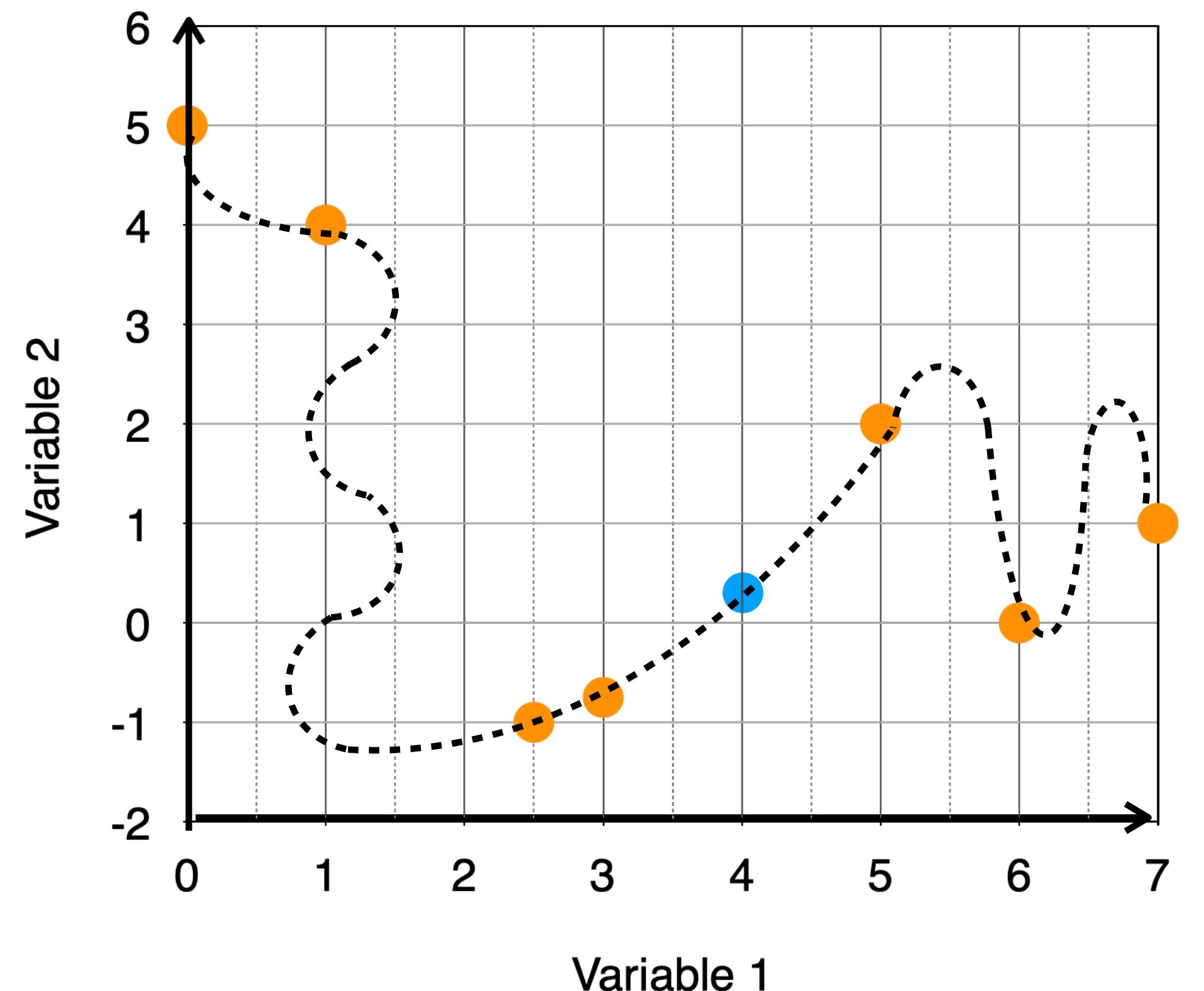




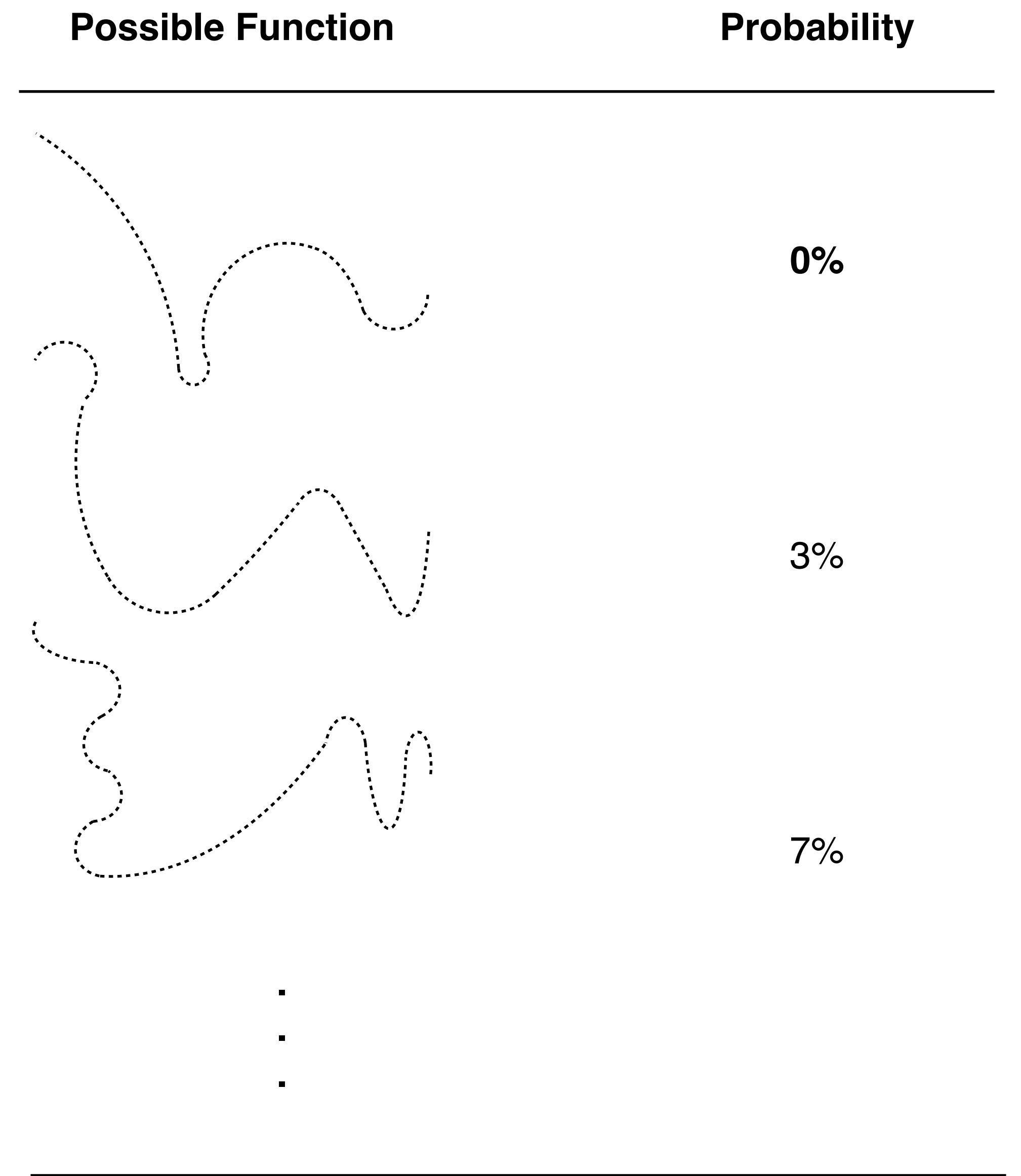
Add new data point → Update the probabilities

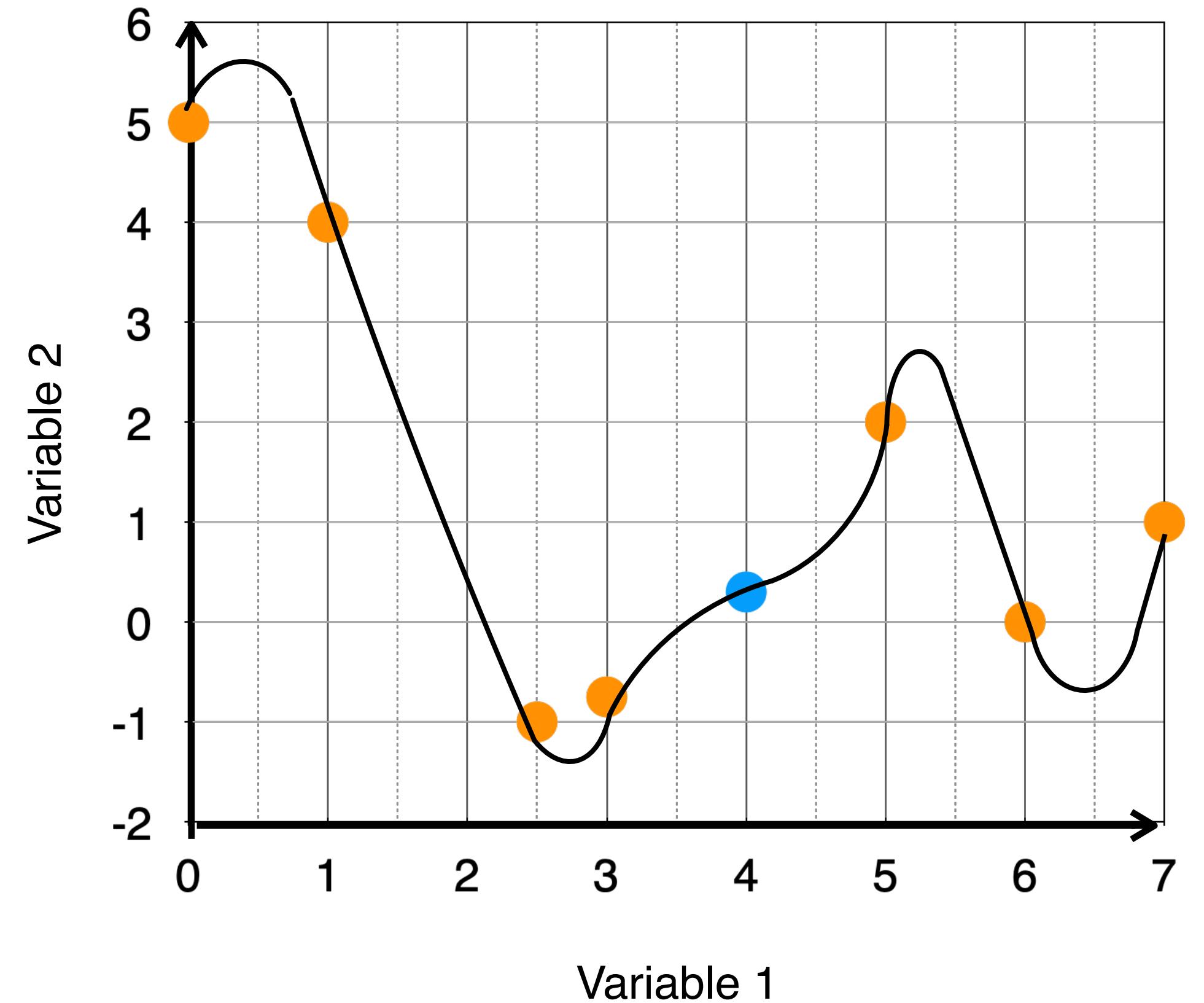


Add new data point → Update the probabilities

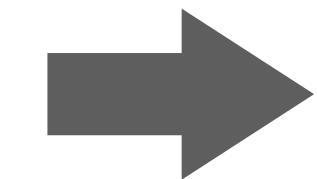


Add new data point → Update the probabilities

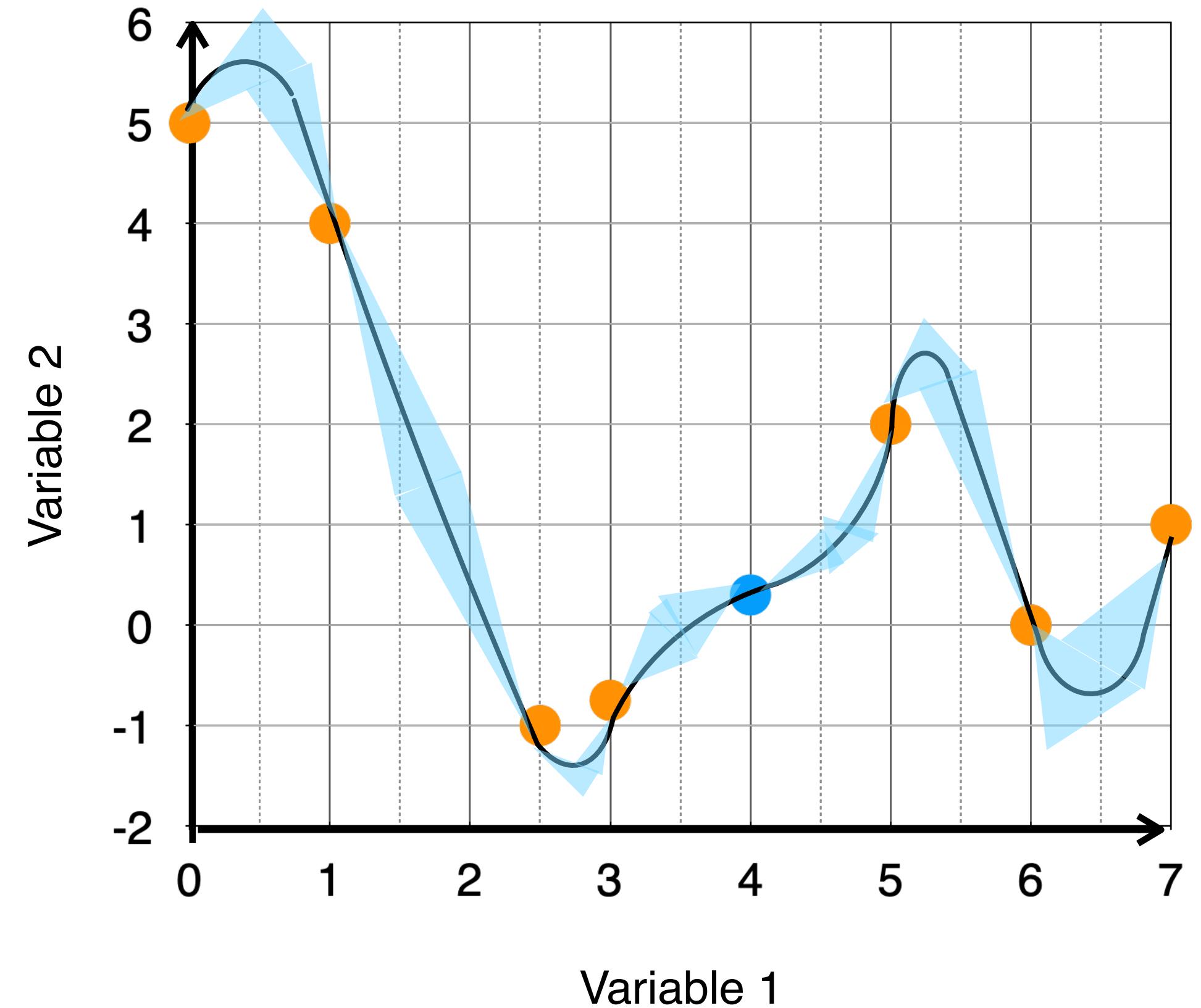




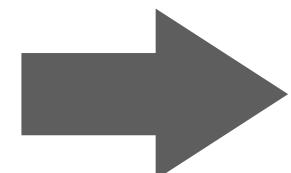
Weighted average of the possible functions



Mean = most probable value



Weighted average of the possible functions



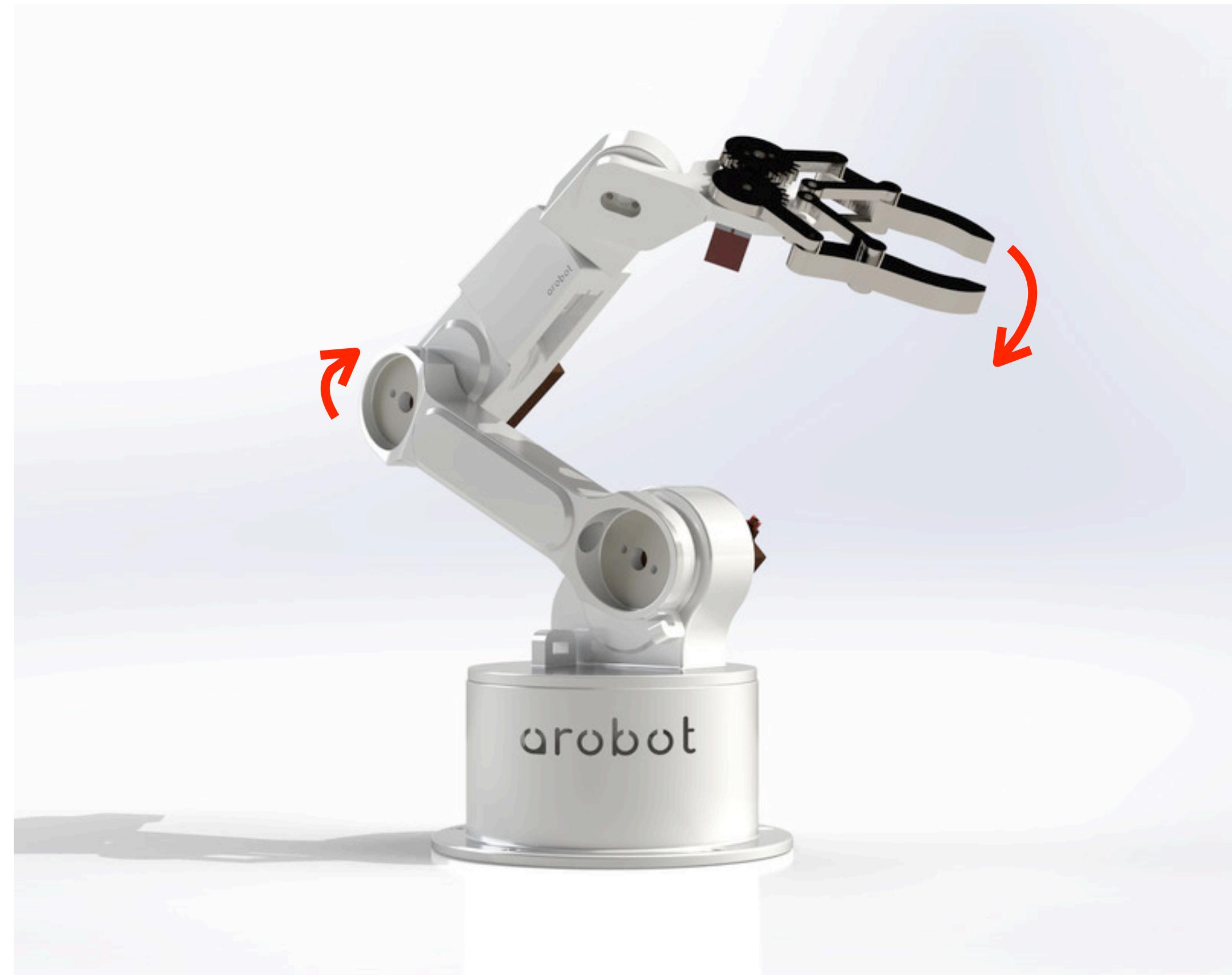
Mean = most probable value

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

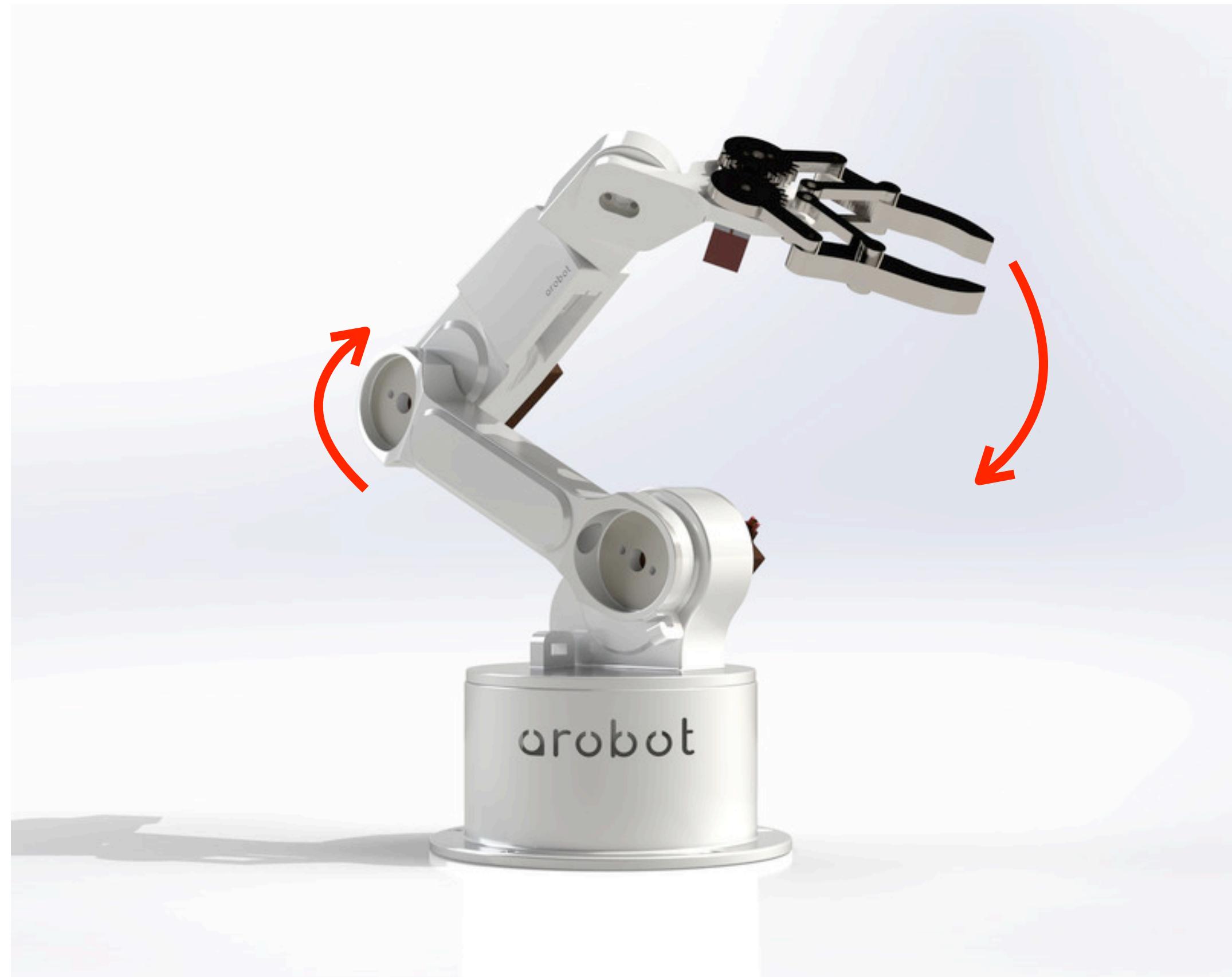
Gaussian Processes work best for continuous processes.



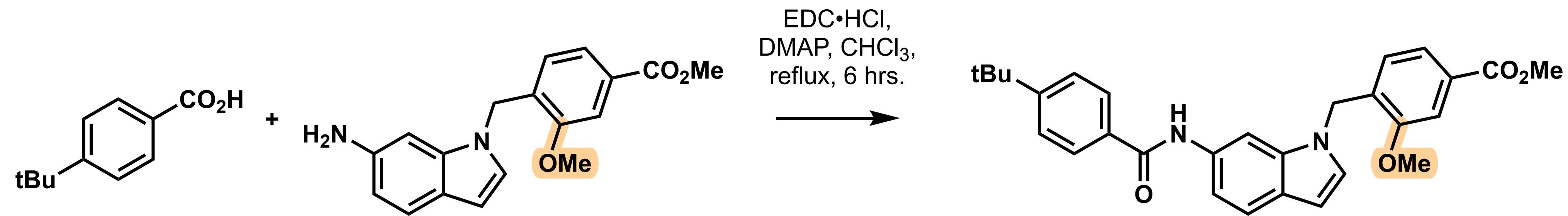
Gaussian Processes work best for continuous processes.



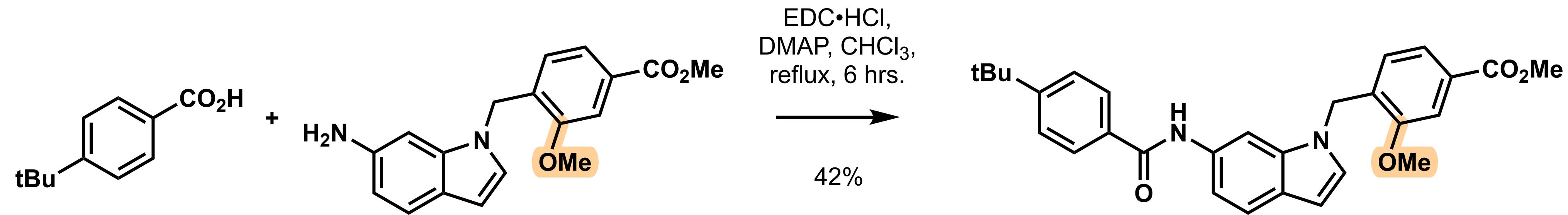
Gaussian Processes work best for continuous processes.



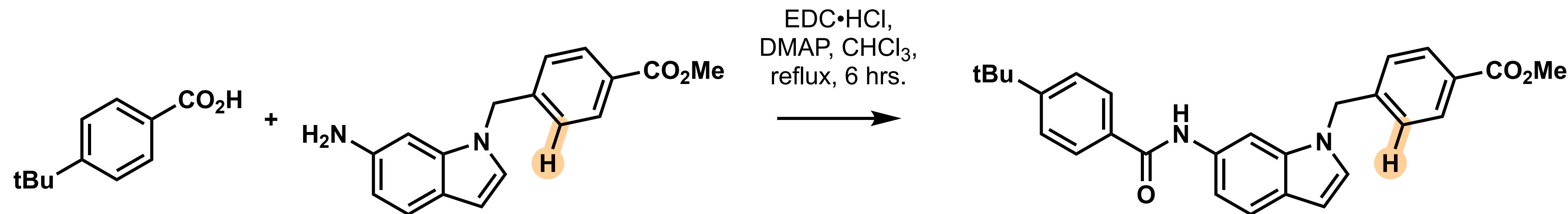
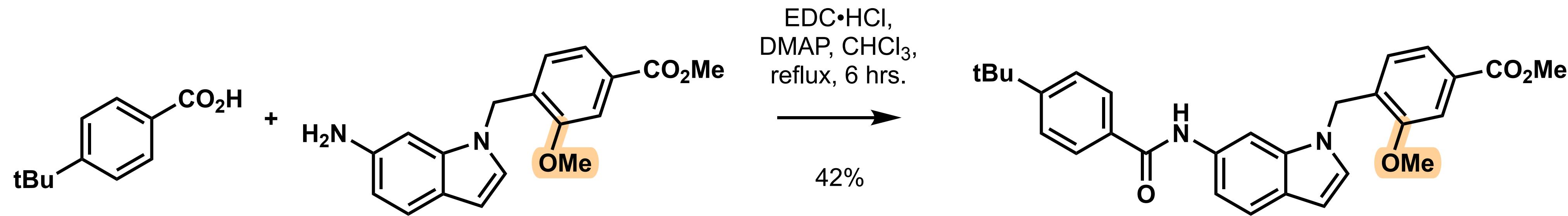
Discontinuous processes much harder to model



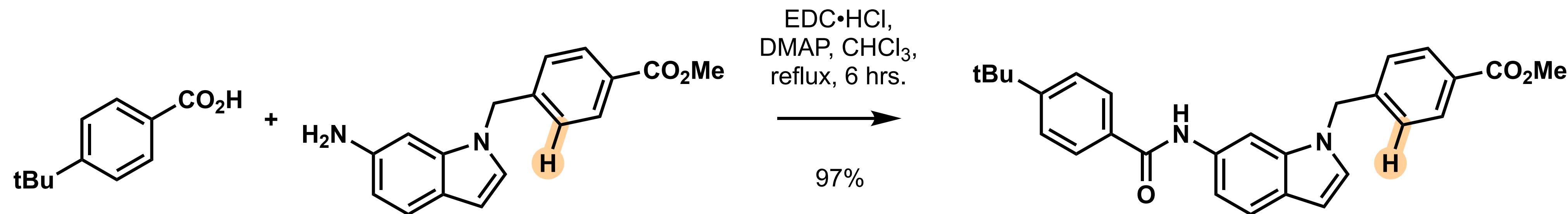
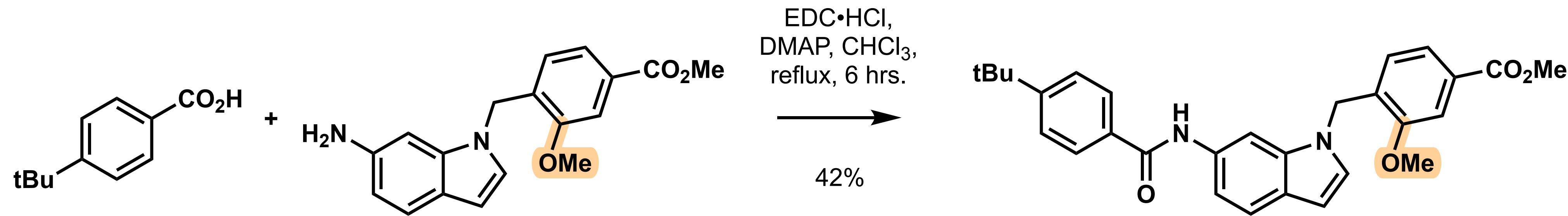
Discontinuous processes much harder to model



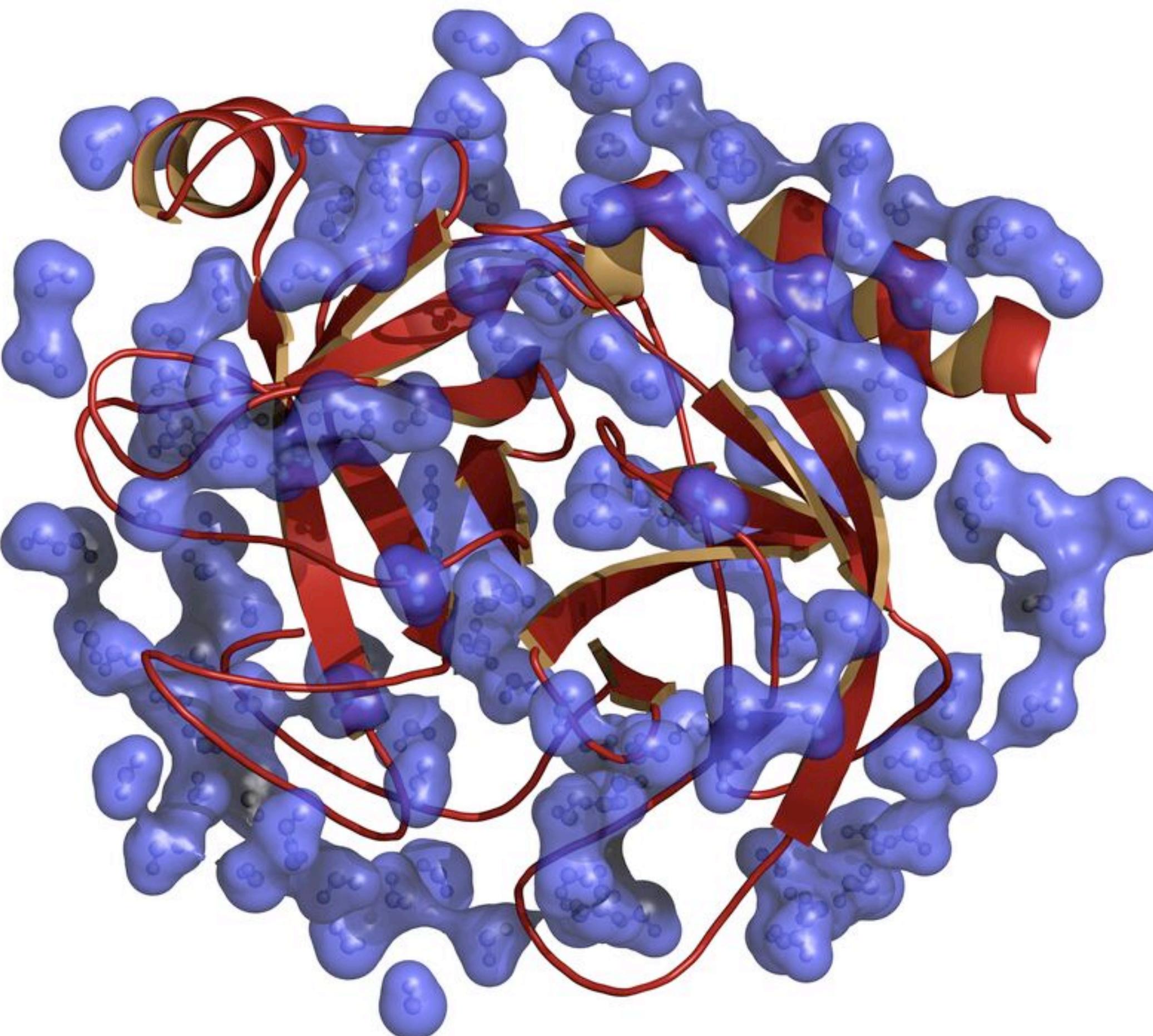
Discontinuous processes much harder to model



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

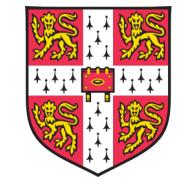
Does this enzyme catalyze this reaction?

Is there affinity between the enzyme and the starting materials?

Some Real World Examples of Gaussian Processes

Is there affinity between the enzyme and the starting materials?

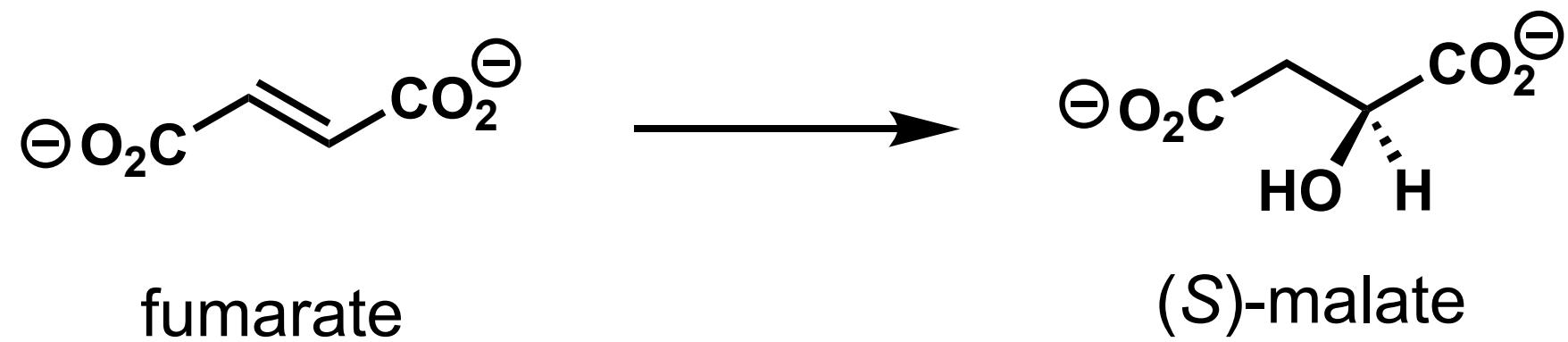
predicting K_M



Is there affinity between the enzyme and the starting materials?

predicting K_M

MYRALRLLARSRPLVRAPAAALASAPGLGGAAVPSFWPPNAARMASQNSFRIEYDTFHEL
KVPNDKYYGAQTVRSTMNFKIGGVTERMPTPVIAFGILKRAAAEVNQDYGLDPKIANAI
MKAADAEVAEGKLNDHFPLVVWQTSGTQTNMVNEVISNRAIEMLGGEGLGSIPVHPNDH
VNKSQSSNDTFPTAMHIAAAIEVHEVLLPGLQKLHDALDAKSKEFAQIIKIGRTHTQDAV PLTL
TGLPFVTAPNKFEALAAHDALVELSGAMNTTACSLMKIANDIRFLGSGPRSGLGEILPE
NEPGSSIMPGKVNPQTQCEAMTMVAAQVMGNHVAVTVGGSNGHFELNVFKPMMIKNVLHSA
RLLGDASVSFTENCVVGIQANTERINKLMNESMLVTALNPHIGYDKAAKIAKTAHKNGS
TLKETAIELGYLTAEQFDEWVKPKDMLGPK



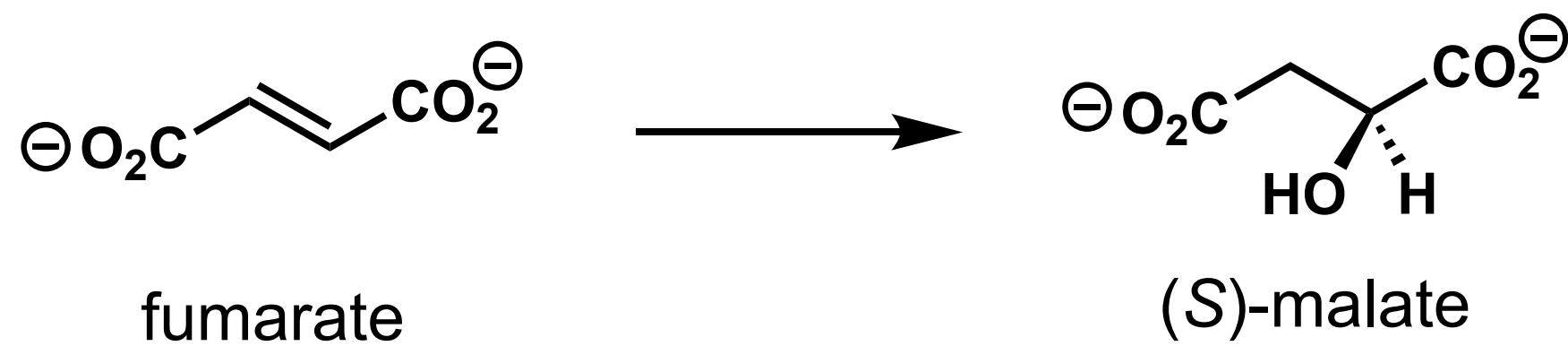


Is there affinity between the enzyme and the starting materials?

predicting K_M

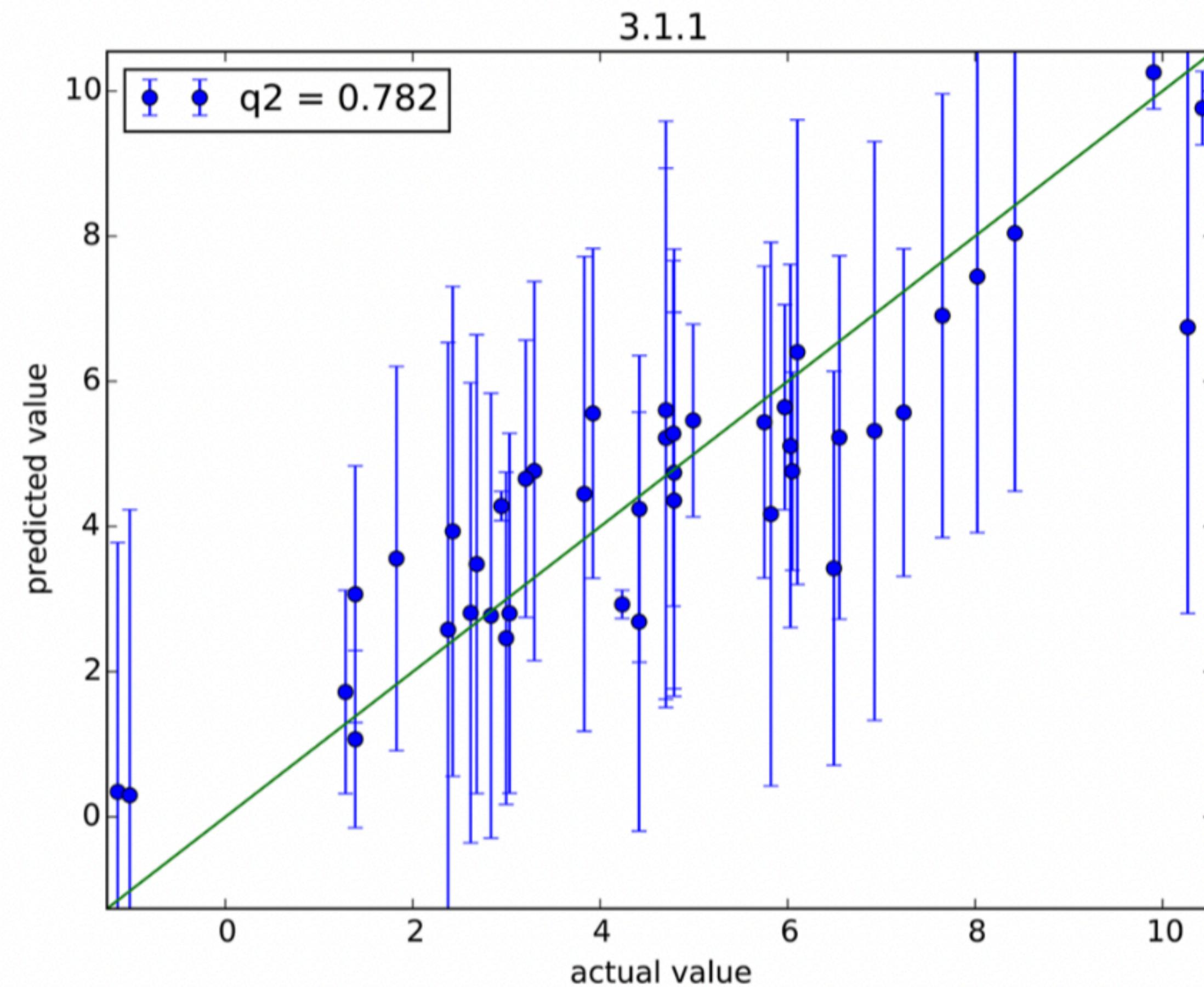
MYRALRLLARSRPLVRAPAAALASAPGLGGAAVPSFWPPNAARMASQNSFRIEYDTFHEL
KVPNDKYYGAQTVRSTMNFKIGGVTERMPTPVIAFGILKRAAAEVNQDYGLDPKIANAI
MKAADAEVAEGKLNDHFPLVVWQTSGTQTNMVNEVISNRAIEMLGGEGLGSIPVHPNDH
VNKSQSSNDTFPTAMHIAAAIEVHEVLLPGLQKLHDALDAKSKEFAQIIKIGRTHTQDAV PLTL
TGLPFVTAPNKFEALAAHDALVELSGAMNTTACSLMKIANDIRFLGSGPRSGLGEILPE
NEPGSSIMPGKVNPQTQCEAMTMVAAQVMGNHVAVTVGGSNGHFELNVFKPMMIKNVLHSA
RLLGDASVSFTENCVVGIQANTERINKLMNESMLVTALNPHIGYDKAAKIAKTAHKNGS
TLKETAIELGYLTAEQFDEWVKPKDMLGPK

fumarate hydratase sequence

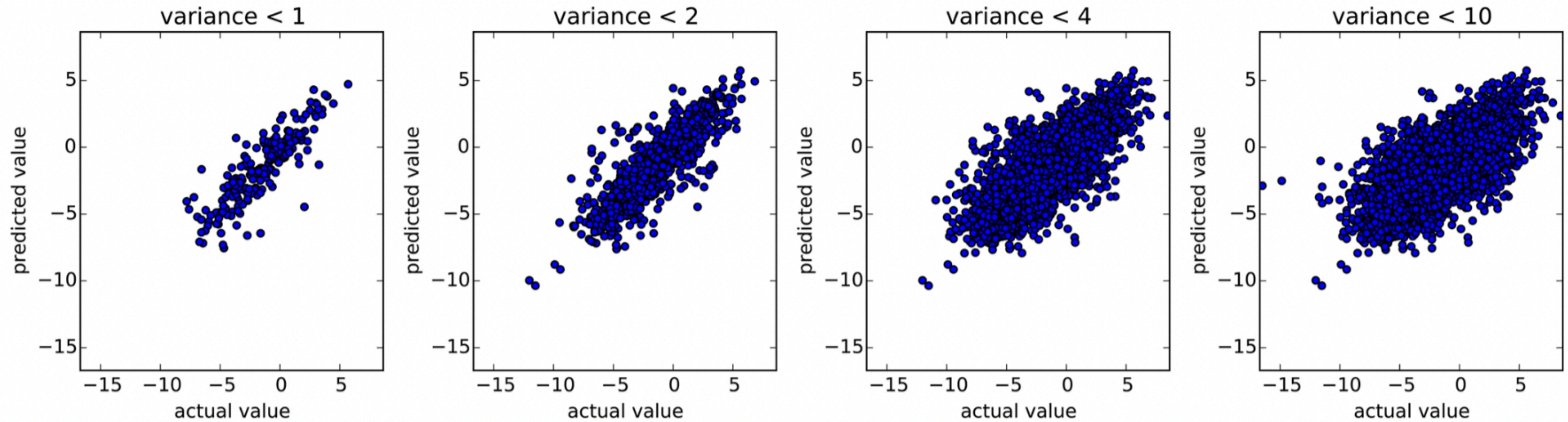


$$K_M = 5 \times 10^{-6} \text{ M}$$

Some Real World Examples of Gaussian Processes



Some Real World Examples of Gaussian Processes



ADME Property optimization



Absorption, Distribution, Metabolism, Excretion

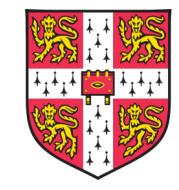
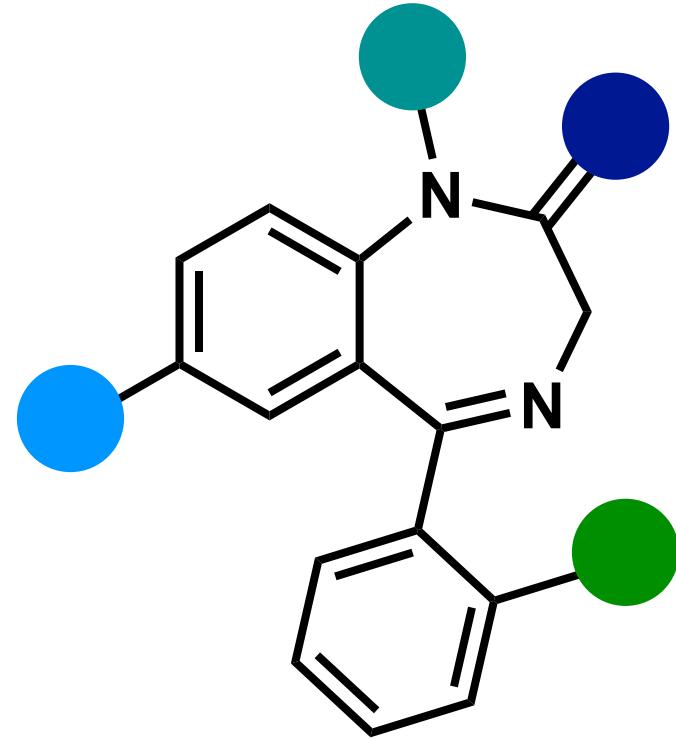


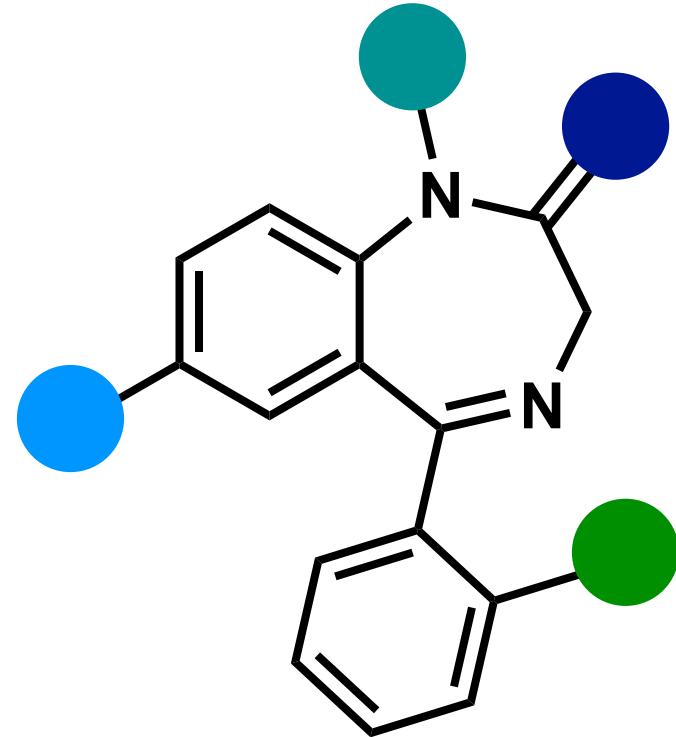
Table 1. BZD Set with Molecular Indices Descriptors

Our Results ^a								
method	desc.	training set			test set			r_{corr}^2
		RMSE	R^2	r_{corr}^2	RMSE	R^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	



245 benzodiazepines

predicting in vitro
binding affinities (pIC_{50})

Table 1. BZD Set with Molecular Indices Descriptors

245 benzodiazepines

predicting in vitro
binding affinities (pIC_{50})

Our Results ^a								
method	desc.	training set			test set			r_{corr}^2
		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								
method	desc.	training set			test set			r_{corr}^2
		SEF ^d	R^2	r_{corr}^2	SEP ^e	R^2	r_{corr}^2	
MLR	39	0.18	0.47	0.20	0.20	0.32		
ANN	22	0.13	0.73	0.14	0.14	0.66		
BRANN	39	0.12	0.75	0.12	0.12	0.71		
GPmodel	39	0.12	0.76	0.14	0.14	0.66		
GPlinear	39	0.12	0.78	0.13	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 Burden⁵ was not available to us.