UNIVERSITY OF CAMBRIDGE

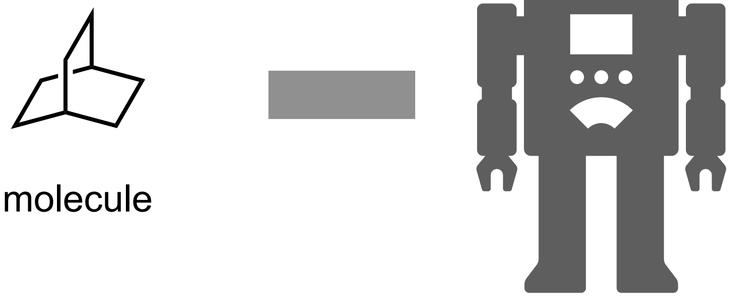
Machine Learning for Chemists

– Principal Component Analysis & Support Vector Machines –

---

12 February 2024

Recall: What is Machine Learning, Really?



molecule

Computer readable format

Made up of numbers

molecule property

Computed Variable

HOMO-LUMO gap

Vibrational Energies

Stability

Observed Variable

Yield

Solubility

Toxicity

Outline

UNIVERSITY OF CAMBRIDGE

The “big idea” behind PCA

An example of PCA used in the wild

Introduction to Classifications

The “big idea” behind SVMs

An example of SVMs used in the wild

Live Demo!

Name	Height
Max	50”
Judy	55”
Charles	60”
Hanna	48”

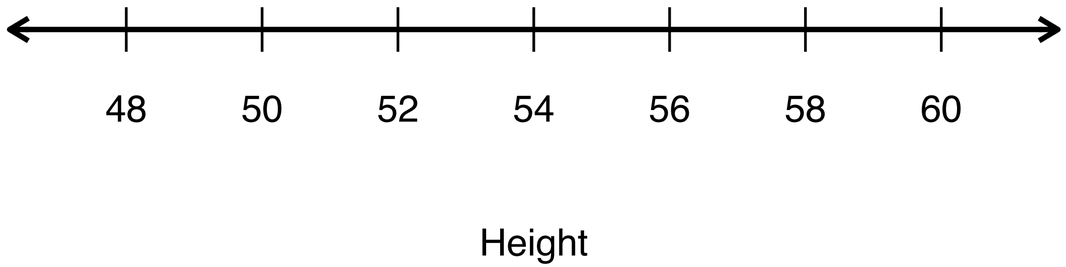
1 Variable

Name	Height
Max	50”
Judy	55”
Charles	60”
Hanna	48”

A Brief Introduction to Dimensions

1 Variable → 1 dimension

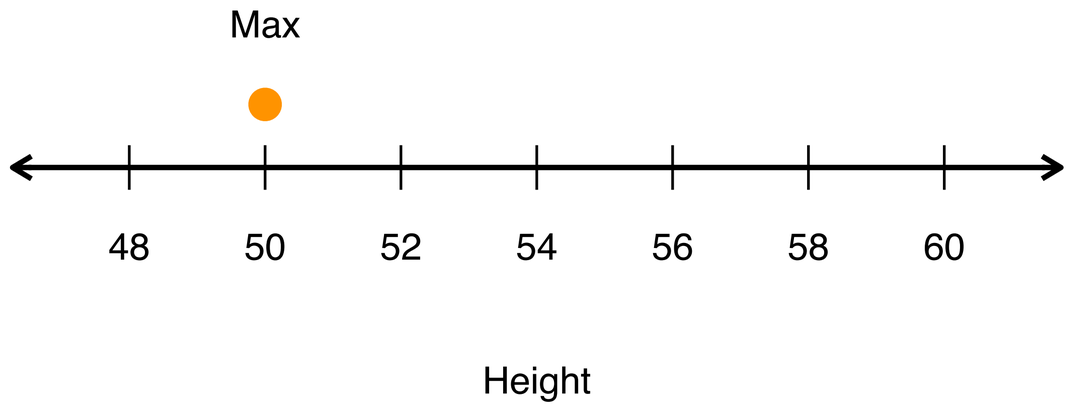
Name	Height
Max	50”
Judy	55”
Charles	60”
Hanna	48”



A Brief Introduction to Dimensions

1 Variable → 1 dimension

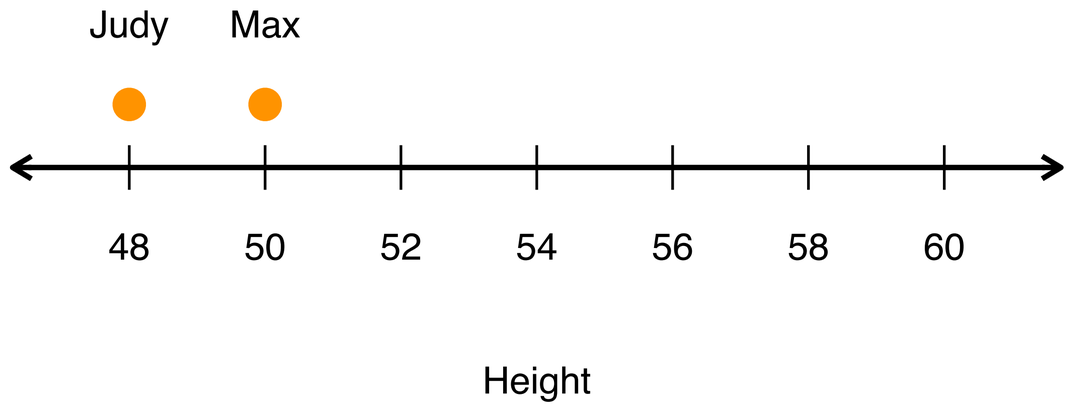
Name	Height
Max	50”
Judy	55”
Charles	60”
Hanna	48”



A Brief Introduction to Dimensions

1 Variable → 1 dimension

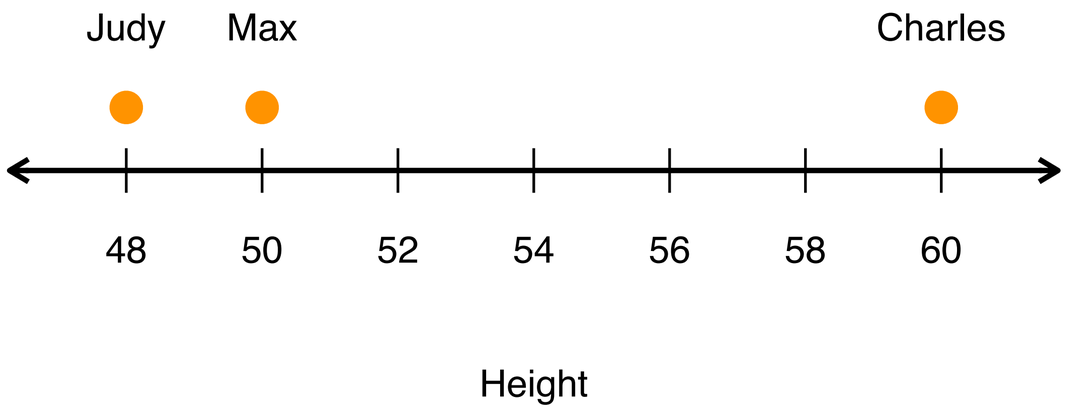
Name	Height
Max	50”
Judy	55”
Charles	60”
Hanna	48”



UNIVERSITY OF CAMBRIDGE

1 Variable → 1 dimension

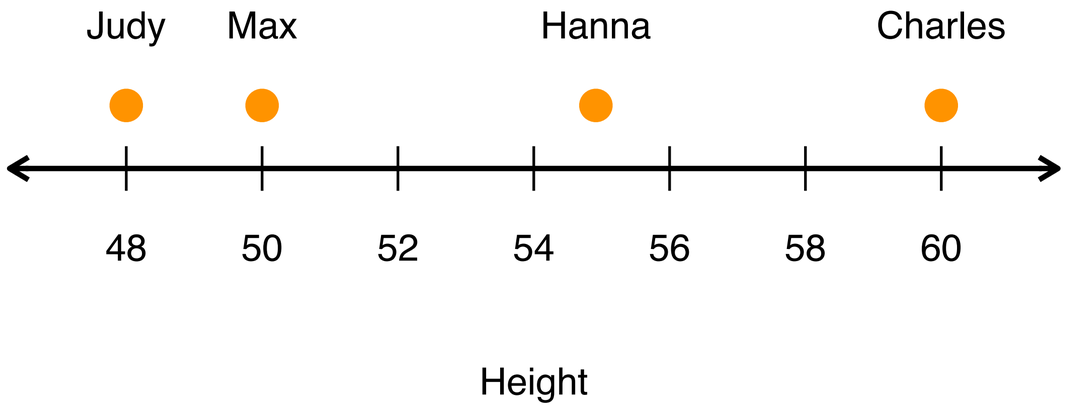
Name	Height
Max	50”
Judy	55”
Charles	60”
Hanna	48”



UNIVERSITY OF CAMBRIDGE

1 Variable → 1 dimension

Name	Height
Max	50”
Judy	55”
Charles	60”
Hanna	48”



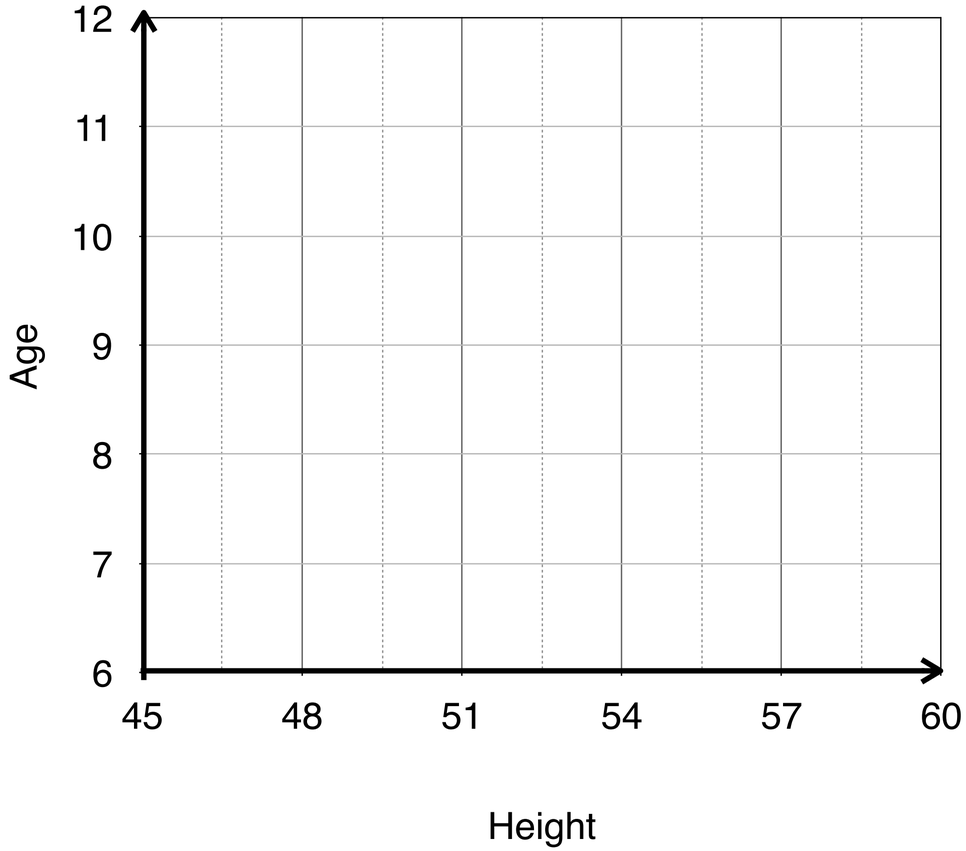
2 Variables

Name	Height	Age
Max	50”	10
Judy	55”	8
Charles	60”	11
Hanna	48”	7

A Brief Introduction to Dimensions

2 Variables → 2 dimensions

Name	Height	Age
Max	50”	10
Judy	55”	8
Charles	60”	11
Hanna	48”	7

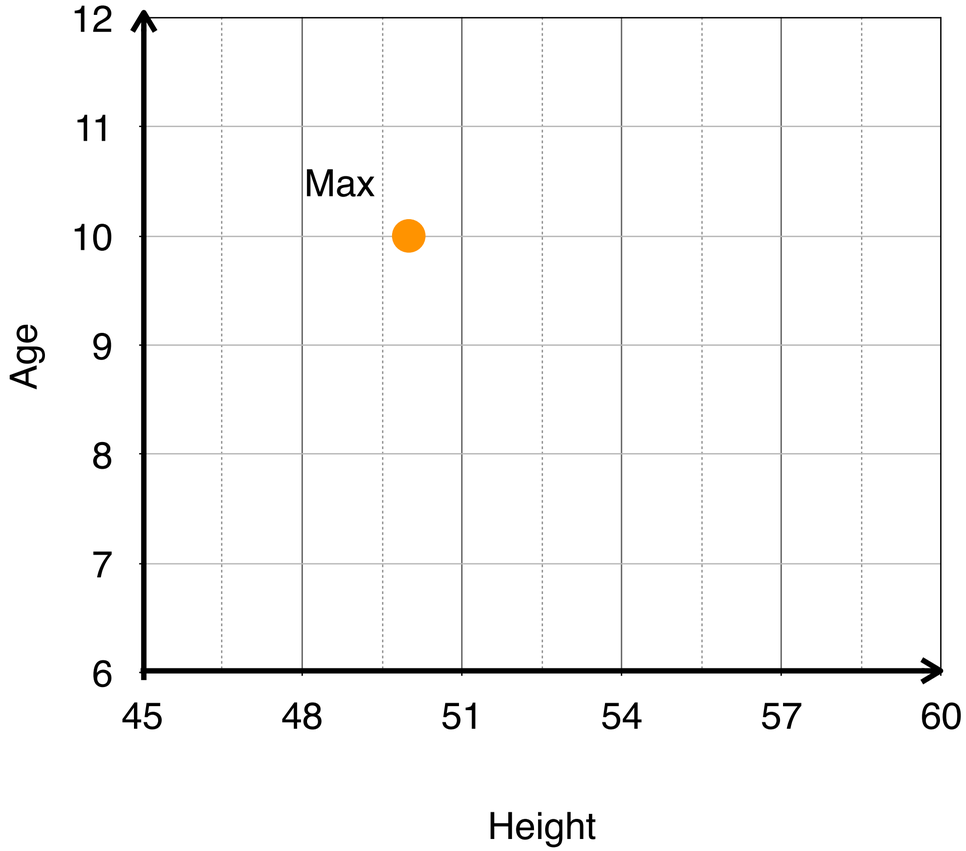


A Brief Introduction to Dimensions

UNIVERSITY OF CAMBRIDGE

2 Variables → 2 dimensions

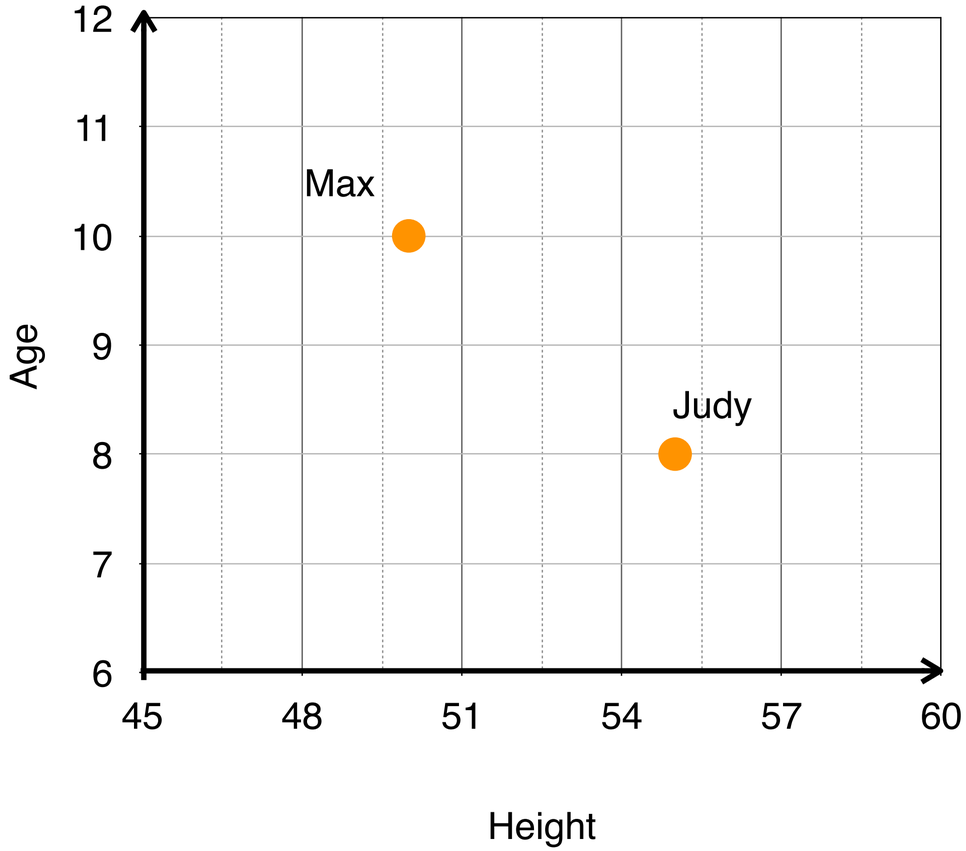
Name	Height	Age
Max	50”	10
Judy	55”	8
Charles	60”	11
Hanna	48”	7



A Brief Introduction to Dimensions

2 Variables → 2 dimensions

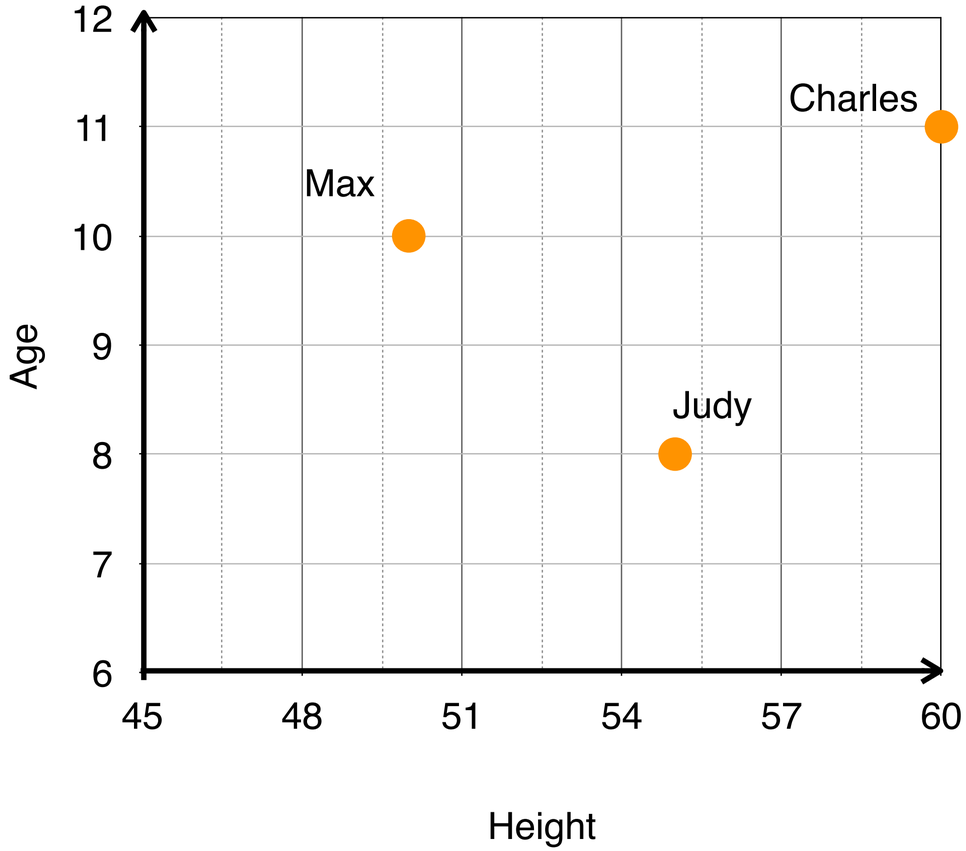
Name	Height	Age
Max	50”	10
Judy	55”	8
Charles	60”	11
Hanna	48”	7



A Brief Introduction to Dimensions

2 Variables → 2 dimensions

Name	Height	Age
Max	50”	10
Judy	55”	8
Charles	60”	11
Hanna	48”	7

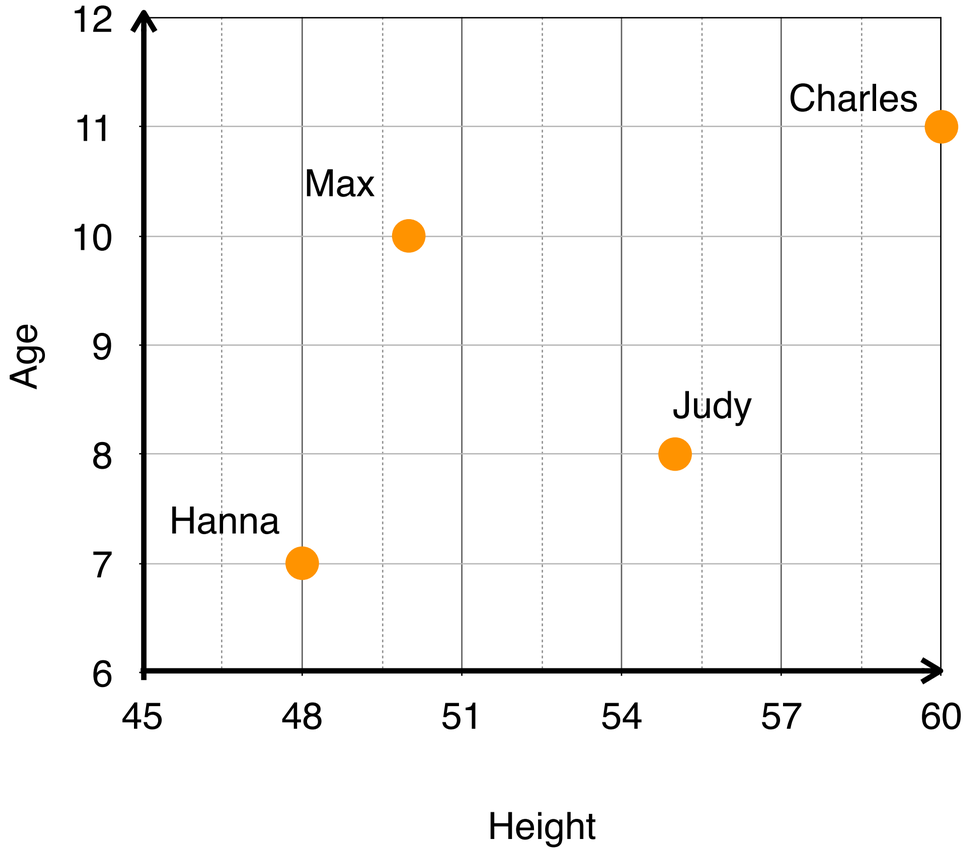


A Brief Introduction to Dimensions

UNIVERSITY OF CAMBRIDGE

2 Variables → 2 dimensions

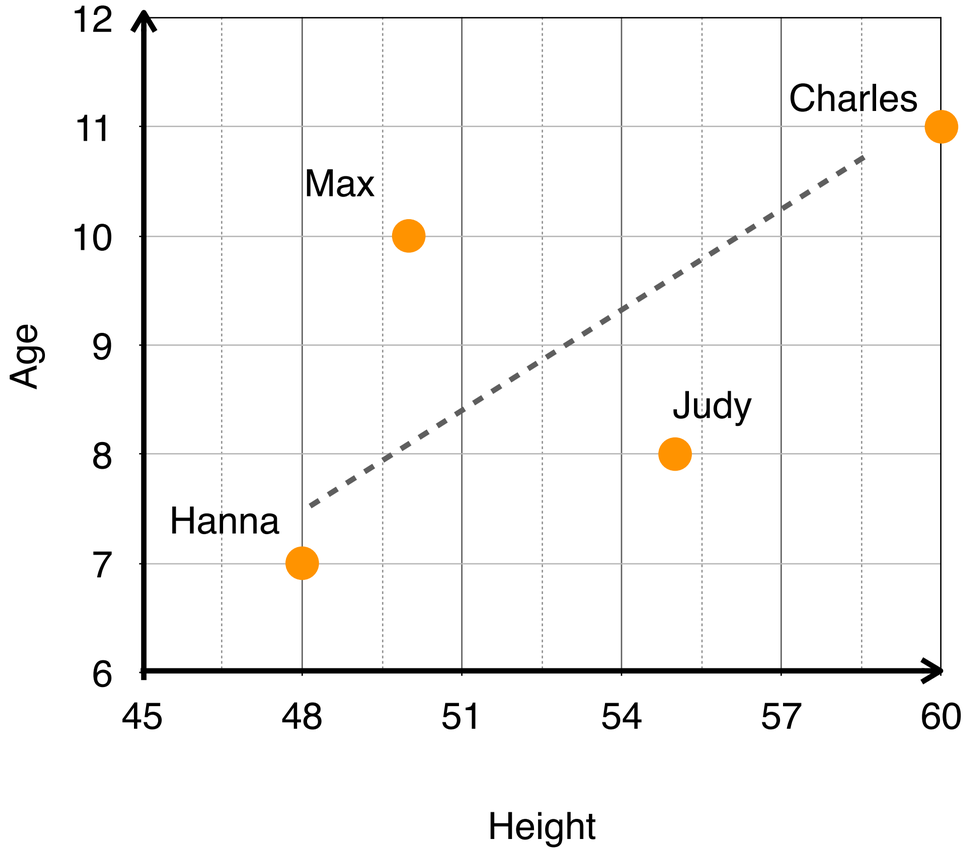
Name	Height	Age
Max	50”	10
Judy	55”	8
Charles	60”	11
Hanna	48”	7



A Brief Introduction to Dimensions

2 Variables → 2 dimensions

Name	Height	Age
Max	50”	10
Judy	55”	8
Charles	60”	11
Hanna	48”	7

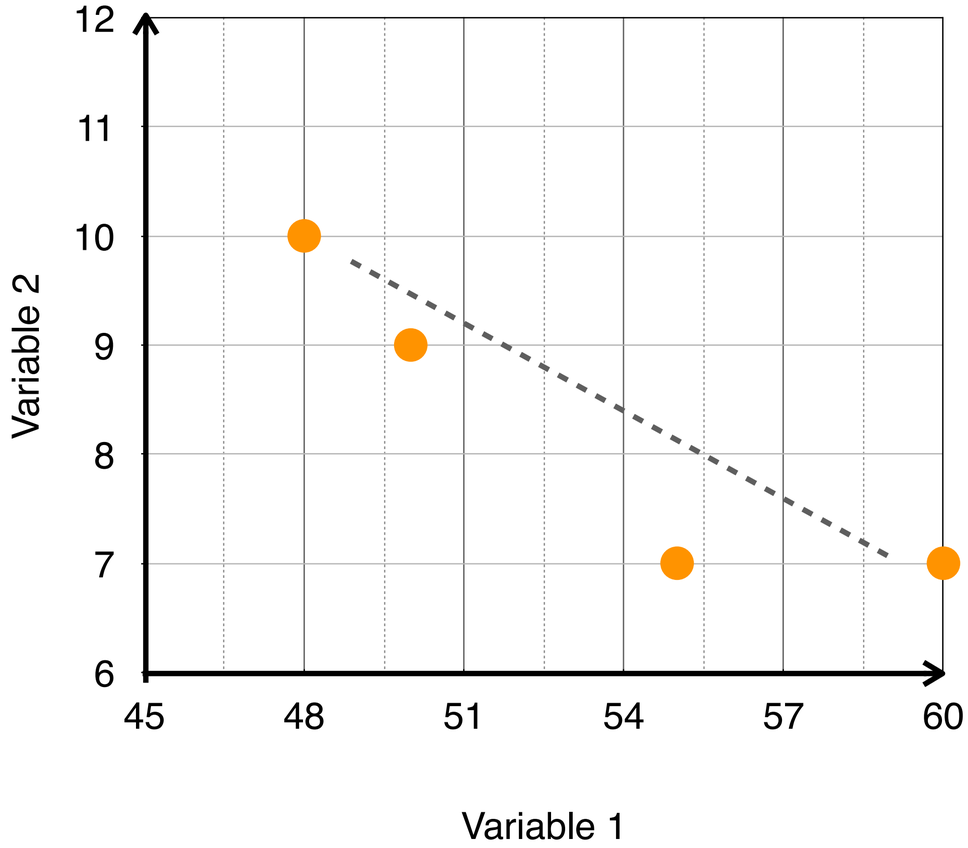


positive correlation
between height and age

A Brief Introduction to Dimensions

2 Variables → 2 dimensions

Name	Height	Age
Max	50”	10
Judy	55”	8
Charles	60”	11
Hanna	48”	7



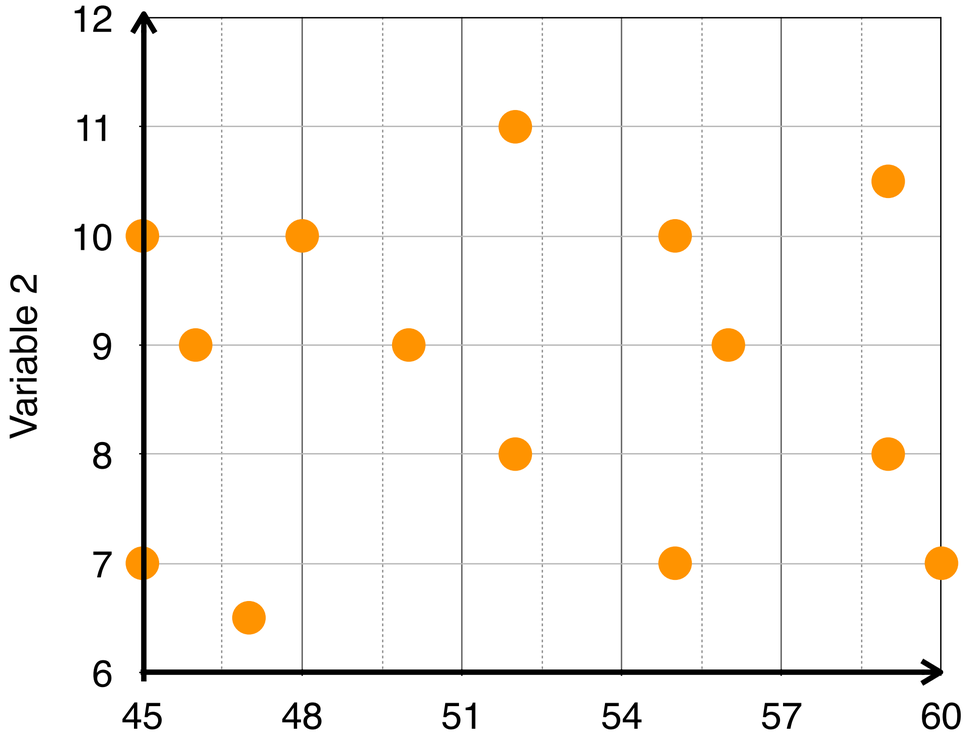
negative correlation between variable 1 and variable 2

A Brief Introduction to Dimensions

UNIVERSITY OF CAMBRIDGE

2 Variables → 2 dimensions

Name	Height	Age
Max	50”	10
Judy	55”	8
Charles	60”	11
Hanna	48”	7



no correlation between variable 1 and variable 2

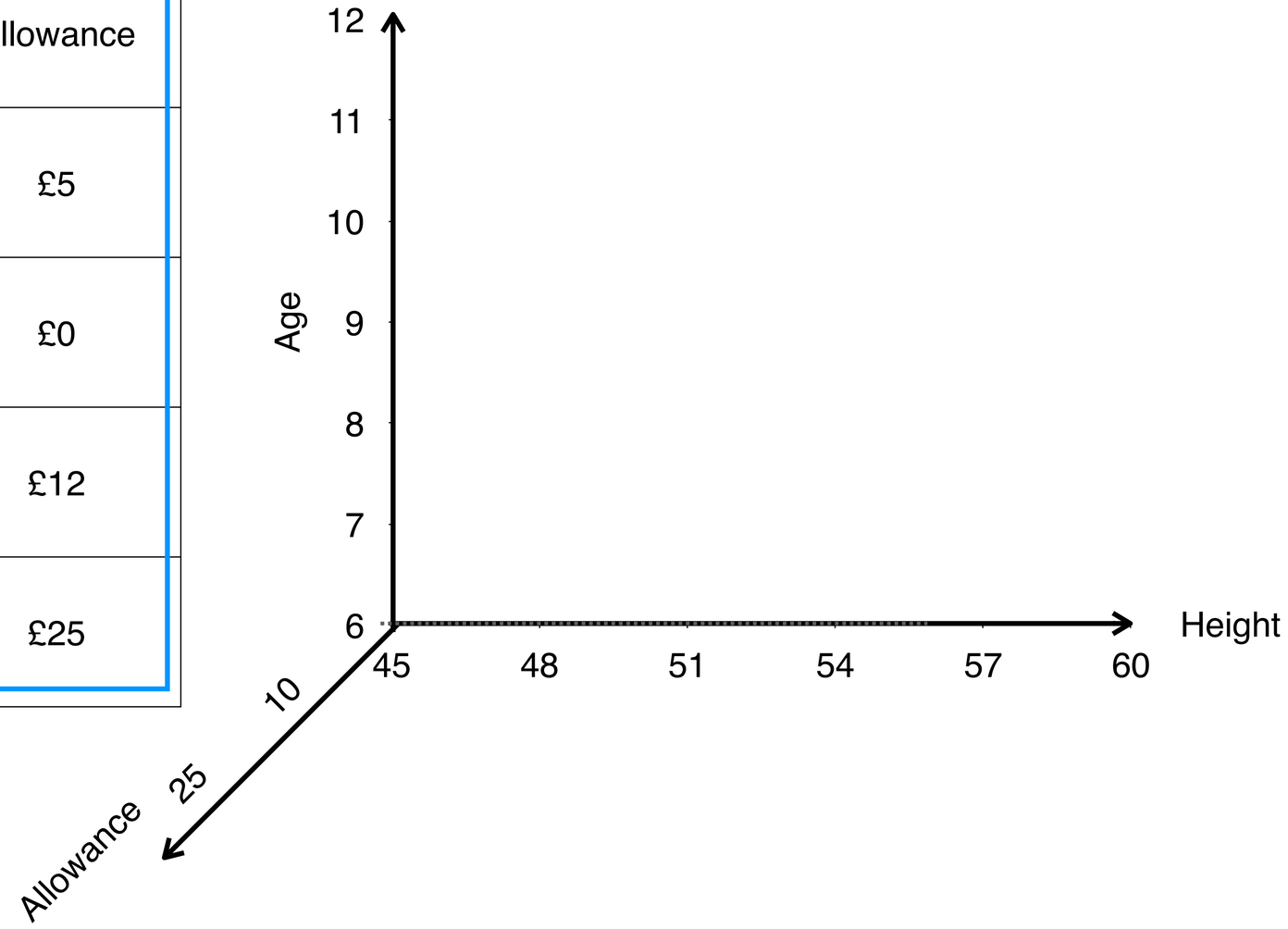
3 Variables

Name	Height	Age	Allowance
Max	50”	10	£5
Judy	55”	8	£0
Charles	60”	11	£12
Hanna	48”	7	£25

A Brief Introduction to Dimensions

3 Variables → 3 dimensions

Name	Height	Age	Allowance
Max	50”	10	£5
Judy	55”	8	£0
Charles	60”	11	£12
Hanna	48”	7	£25

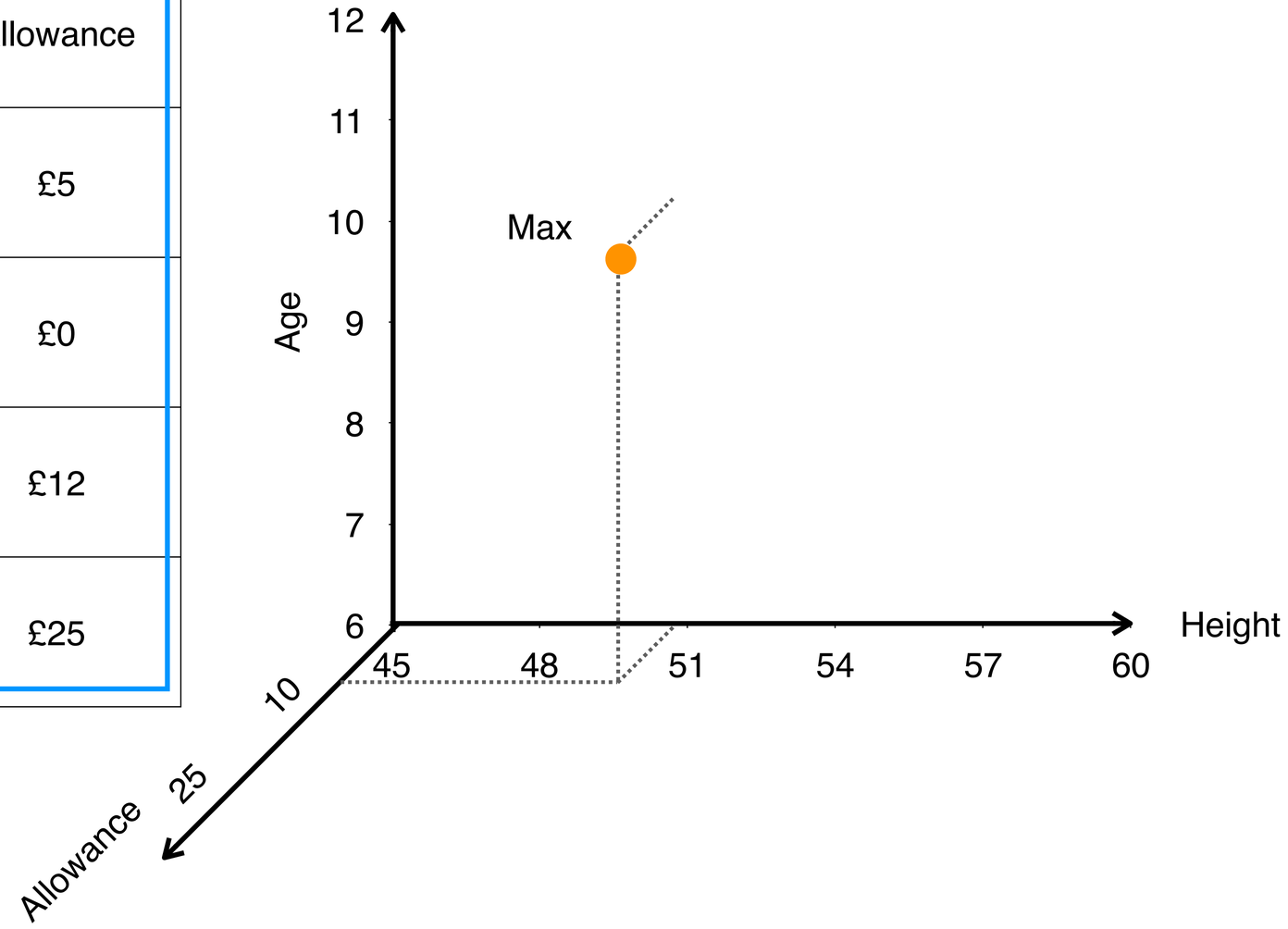


A Brief Introduction to Dimensions

UNIVERSITY OF CAMBRIDGE

3 Variables → 3 dimensions

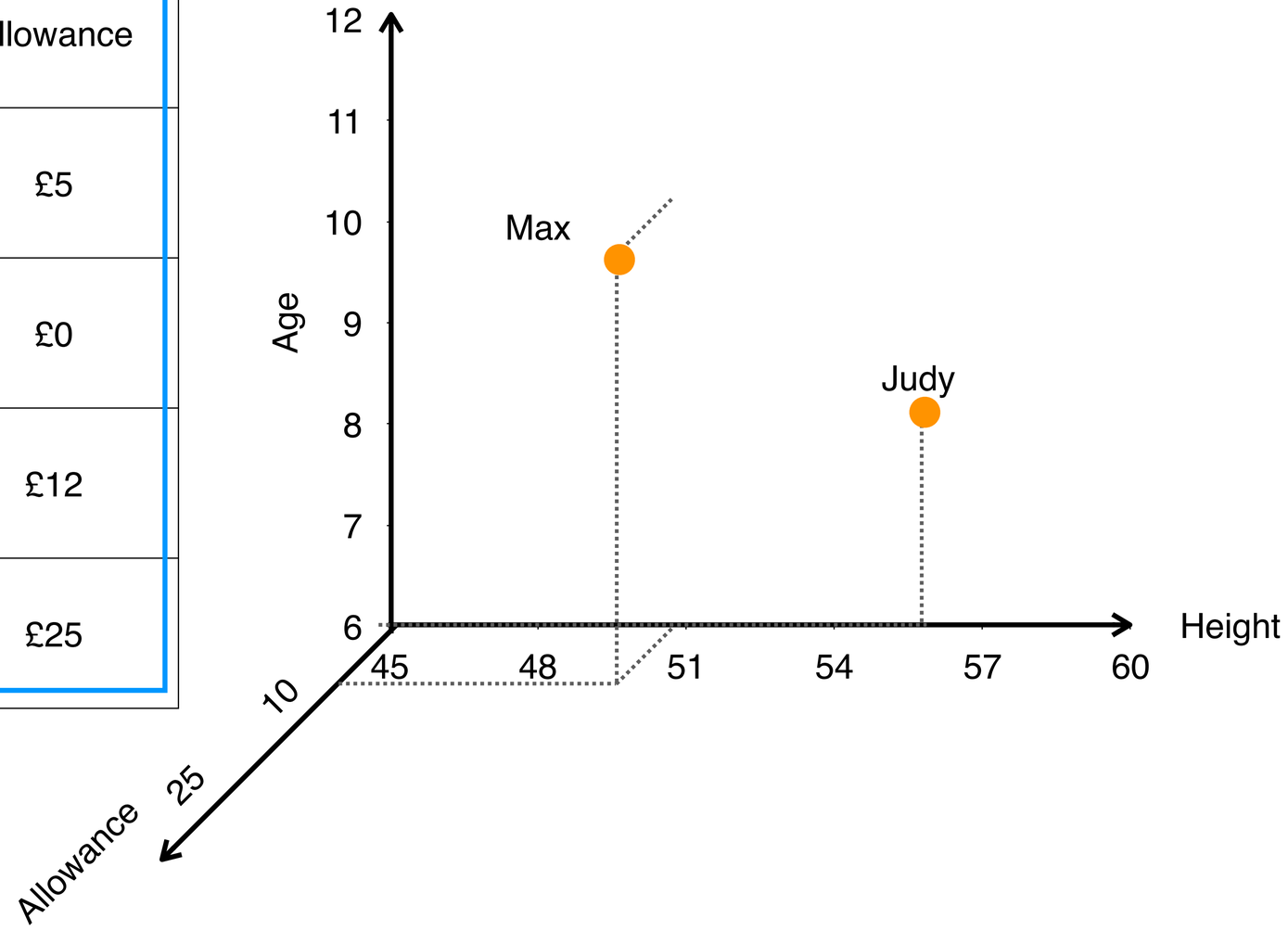
Name	Height	Age	Allowance
Max	50”	10	£5
Judy	55”	8	£0
Charles	60”	11	£12
Hanna	48”	7	£25



A Brief Introduction to Dimensions

3 Variables → 3 dimensions

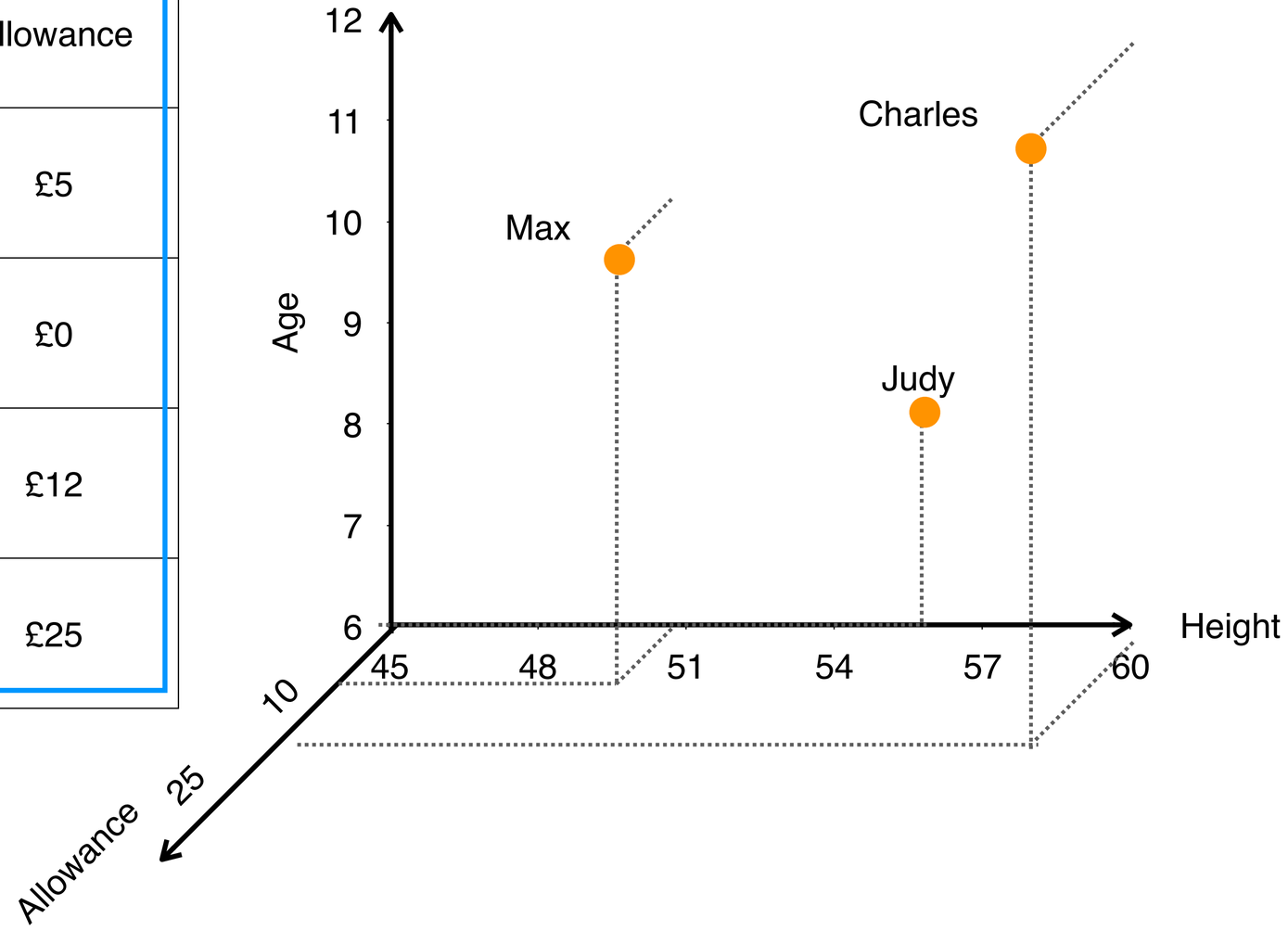
Name	Height	Age	Allowance
Max	50”	10	£5
Judy	55”	8	£0
Charles	60”	11	£12
Hanna	48”	7	£25



A Brief Introduction to Dimensions

3 Variables → 3 dimensions

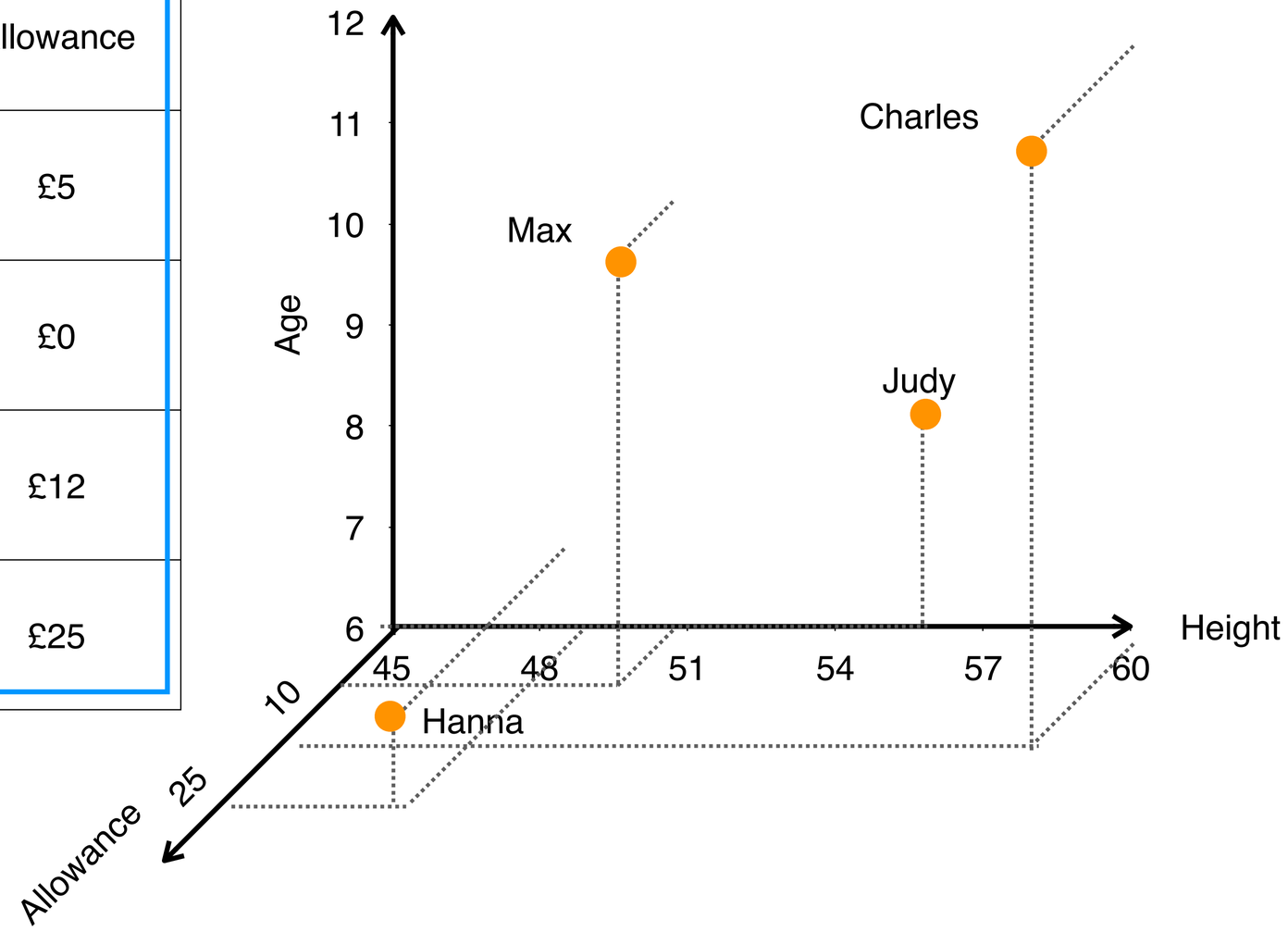
Name	Height	Age	Allowance
Max	50”	10	£5
Judy	55”	8	£0
Charles	60”	11	£12
Hanna	48”	7	£25



A Brief Introduction to Dimensions

3 Variables → 3 dimensions

Name	Height	Age	Allowance
Max	50”	10	£5
Judy	55”	8	£0
Charles	60”	11	£12
Hanna	48”	7	£25



Why Would You Ever Want to Reduce Your Dimensions?

UNIVERSITY OF CAMBRIDGE

N Variables → N dimensions

Name	Height	Age	Allowance	Fav. #	…	…	…
Max	50”	10	£5	2	…	…	…
Judy	55”	8	£0	7	…	…	…
Charles	60”	11	£12	12	…	…	…
Hanna	48”	7	£25	55	…	…	…

Why Would You Ever Want to Reduce Your Dimensions?

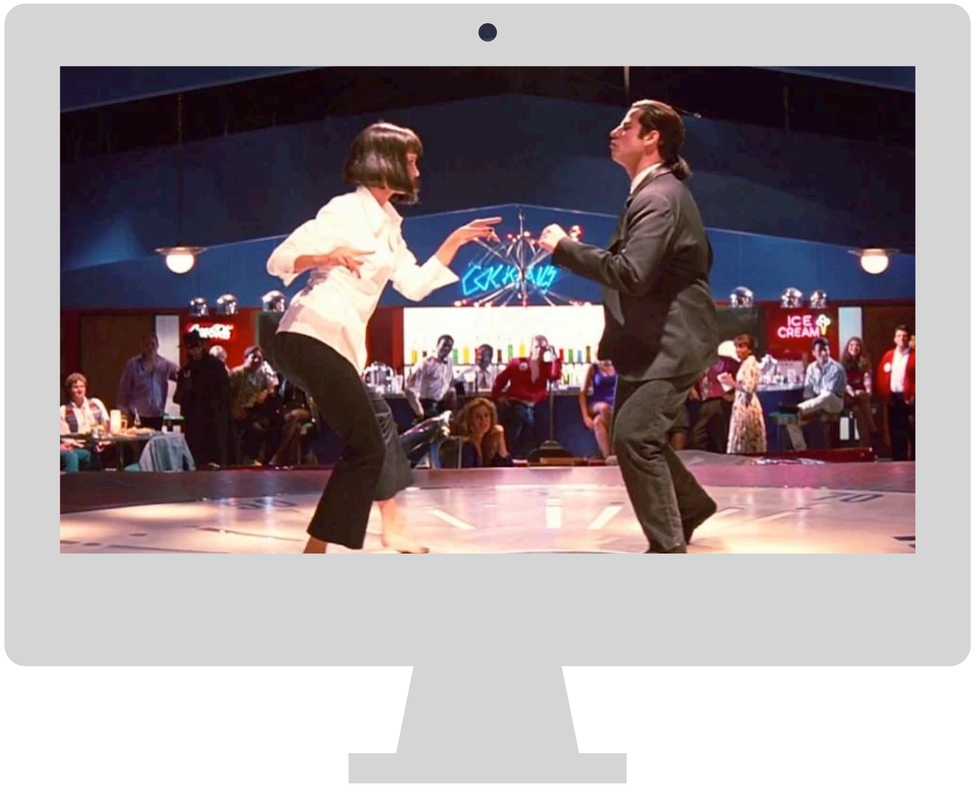
UNIVERSITY OF CAMBRIDGE

N Variables → N dimensions

Name	Height	Age	Allowance	Fav. #	…	…	…
Max	50”	10	£5	2	…	…	…
Judy	55”	8	£0	7	…	…	…
Charles	60”	11	£12	12	…	…	…
Hanna	48”	7	£25	55	…	…	…

You cannot draw N > 3 dimensions

Are All These Dimensions Really That Important?



UNIVERSITY OF CAMBRIDGE

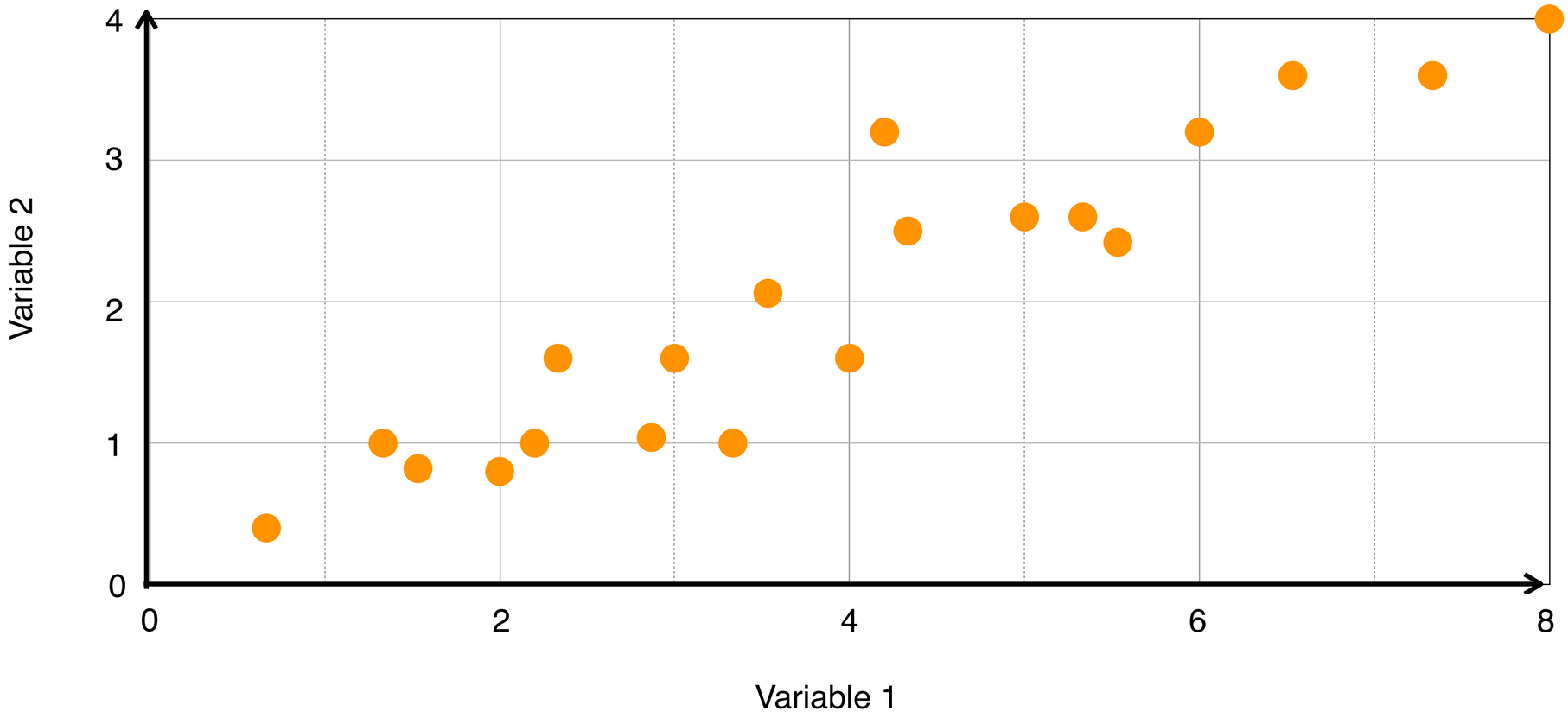
Principal Component Analysis (PCA)

Finds a way to squeeze multidimensional data to more manageable dimensional data

Minimizes information lost in “squeezing” process

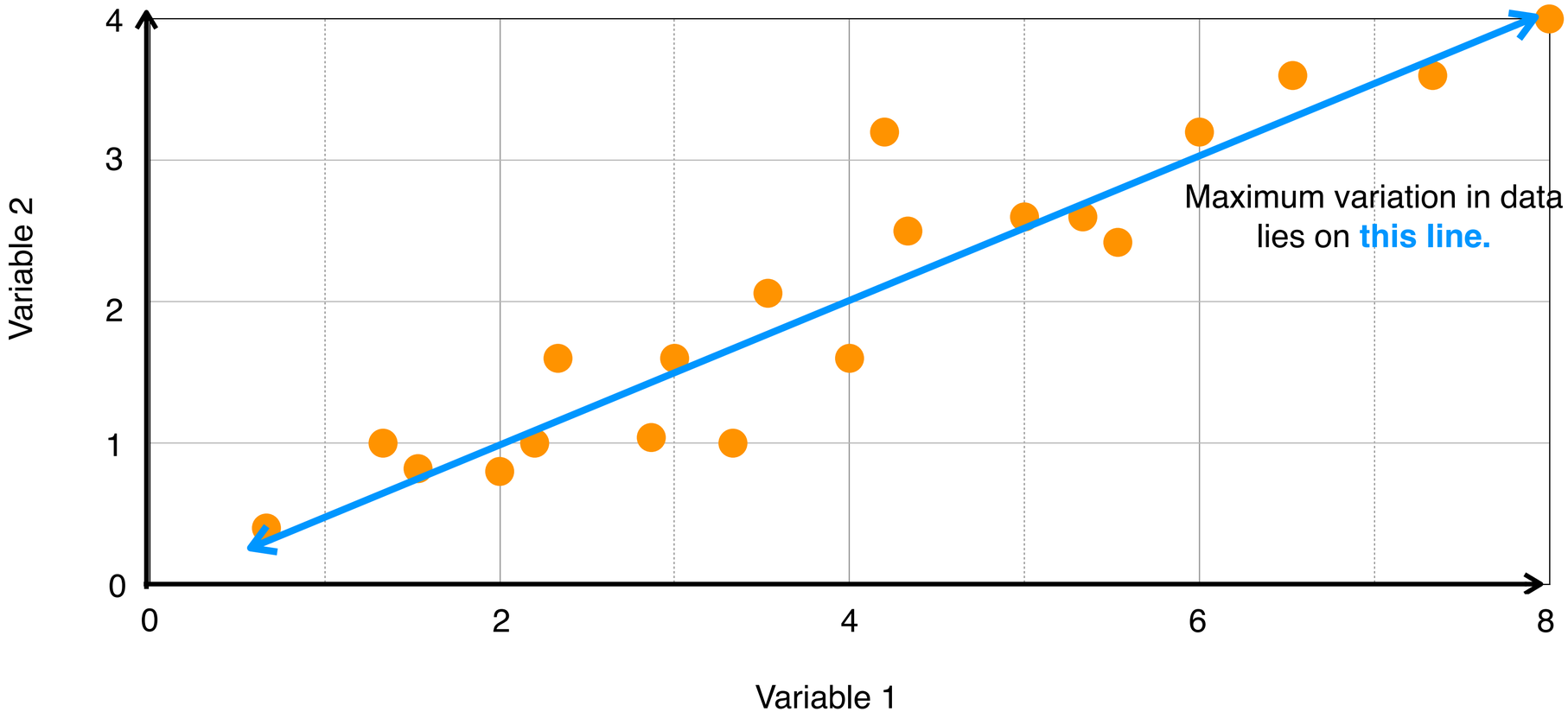
Allows us to visualize our data

The Big Idea: How Does PCA Work?



We want to find 2 perpendicular axes that encapsulate the most information (variation) possible.

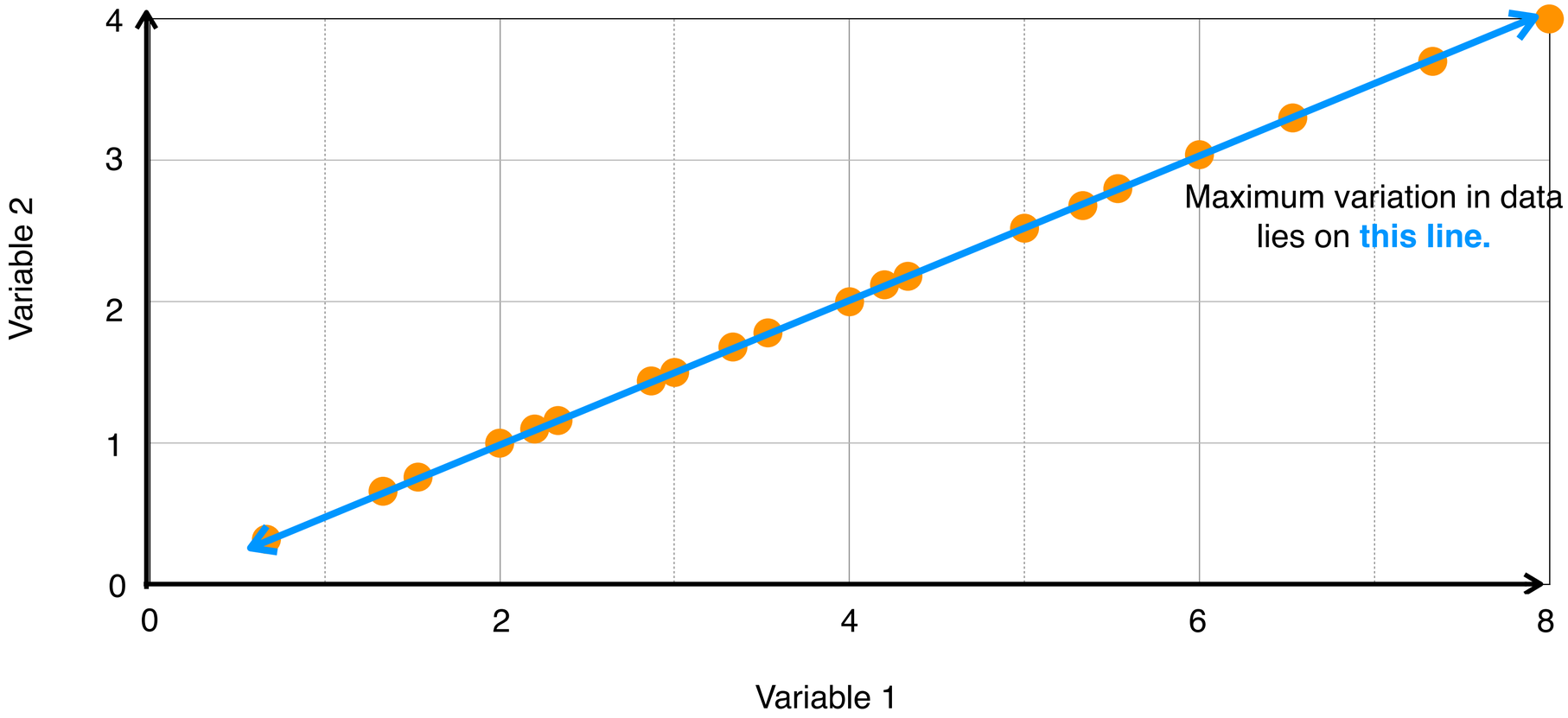
The Big Idea: How Does PCA Work?



We want to find 2 perpendicular axes that encapsulate the most information (variation) possible.

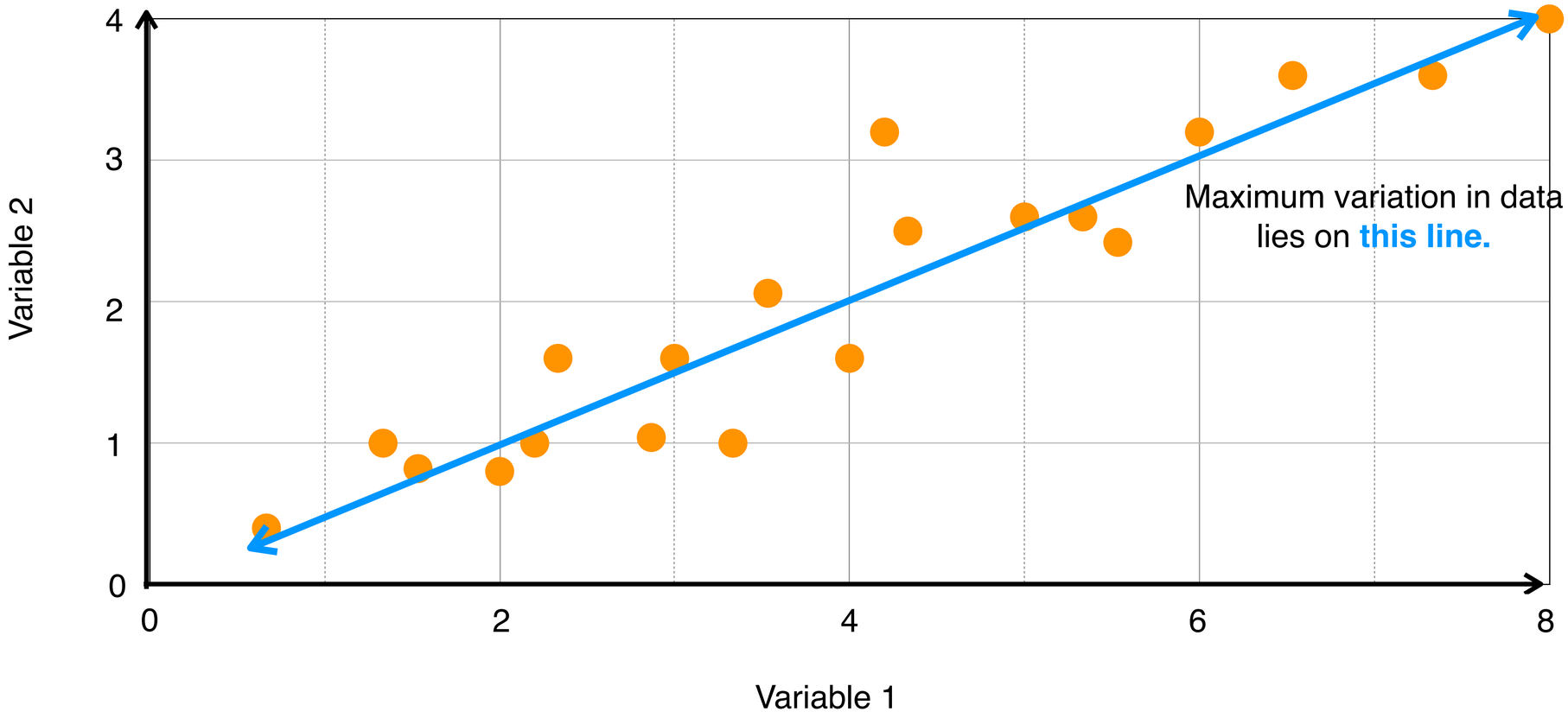
The Big Idea: How Does PCA Work?

UNIVERSITY OF CAMBRIDGE



We want to find 2 perpendicular axes that encapsulate the most information (variation) possible.

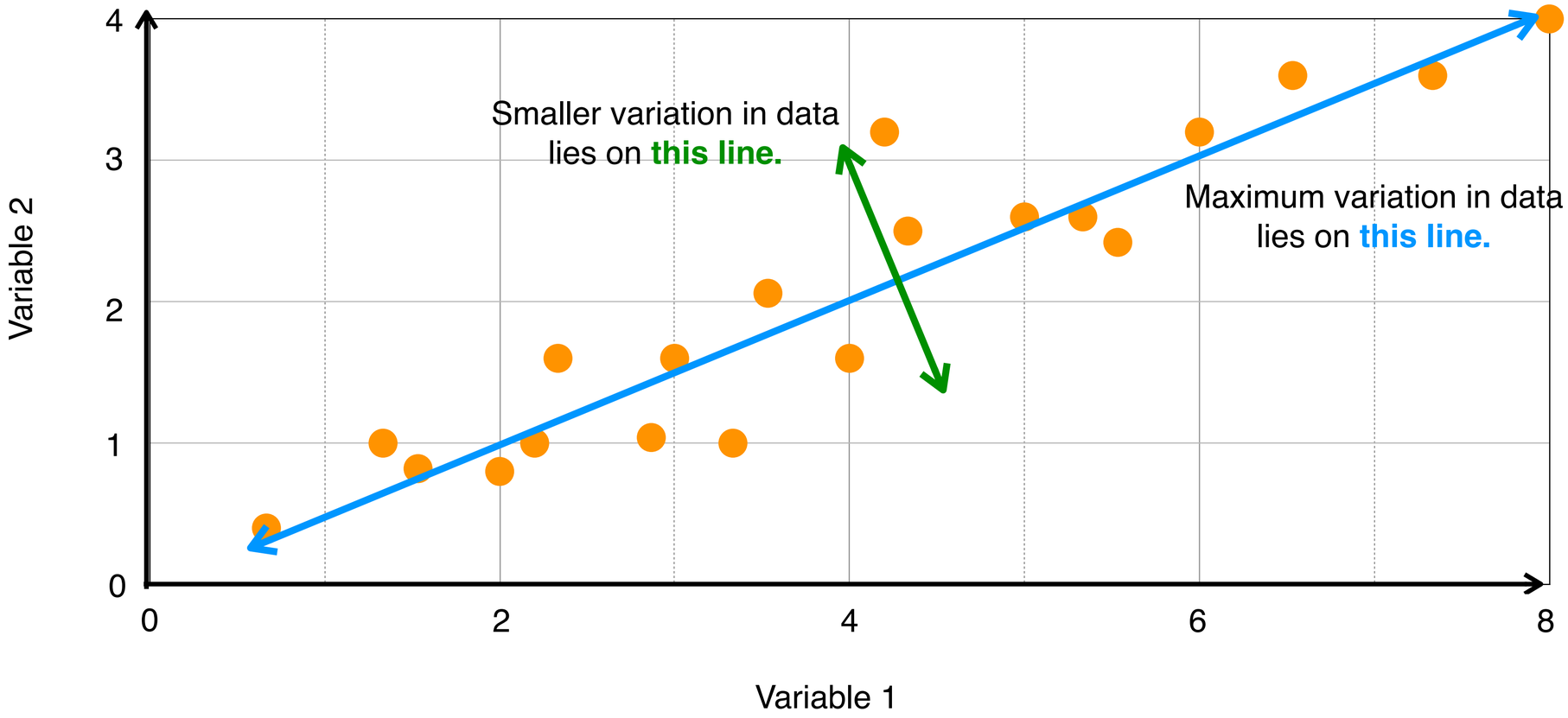
The Big Idea: How Does PCA Work?



We want to find 2 perpendicular axes that encapsulate the most information (variation) possible.

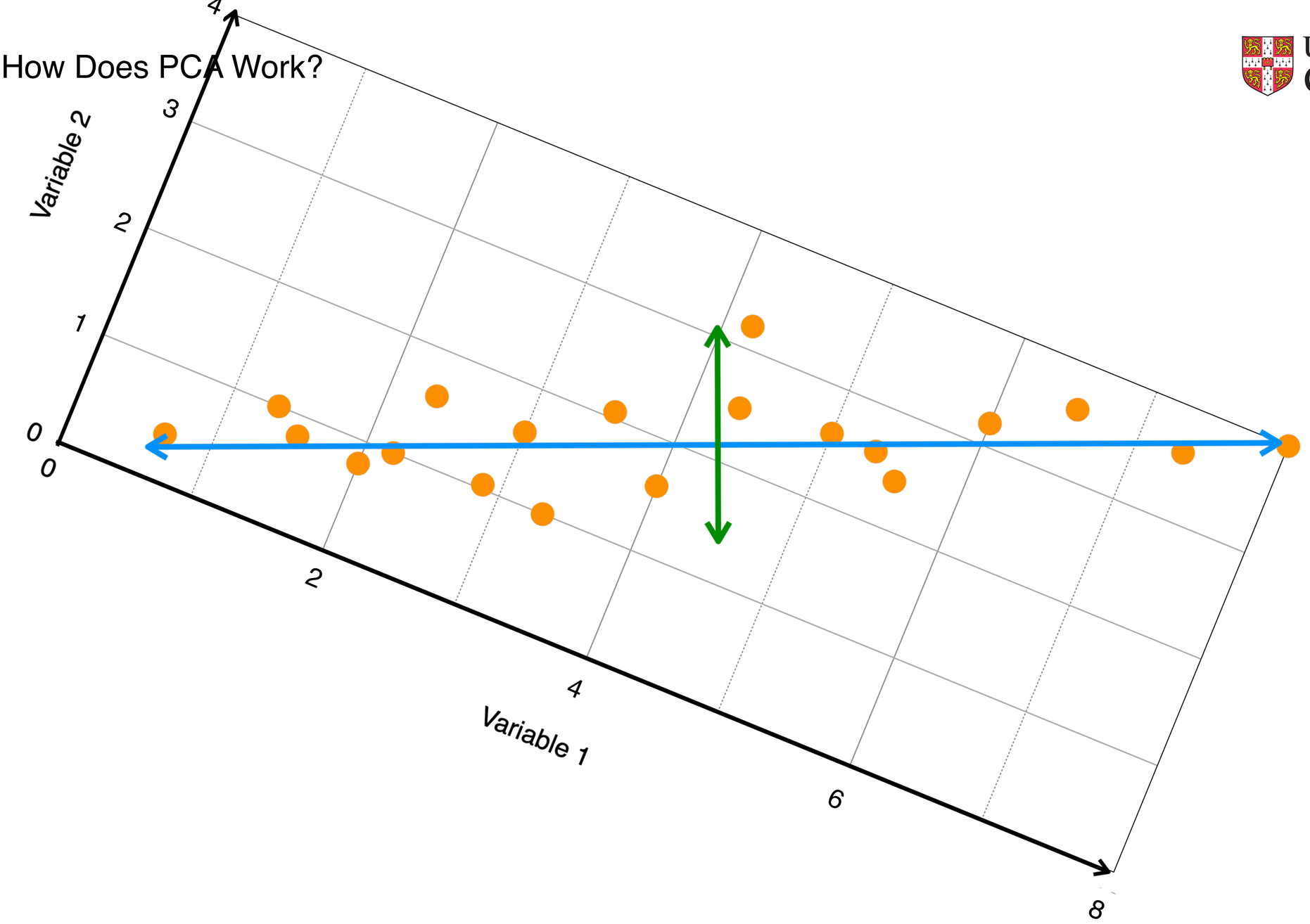
The Big Idea: How Does PCA Work?

UNIVERSITY OF CAMBRIDGE



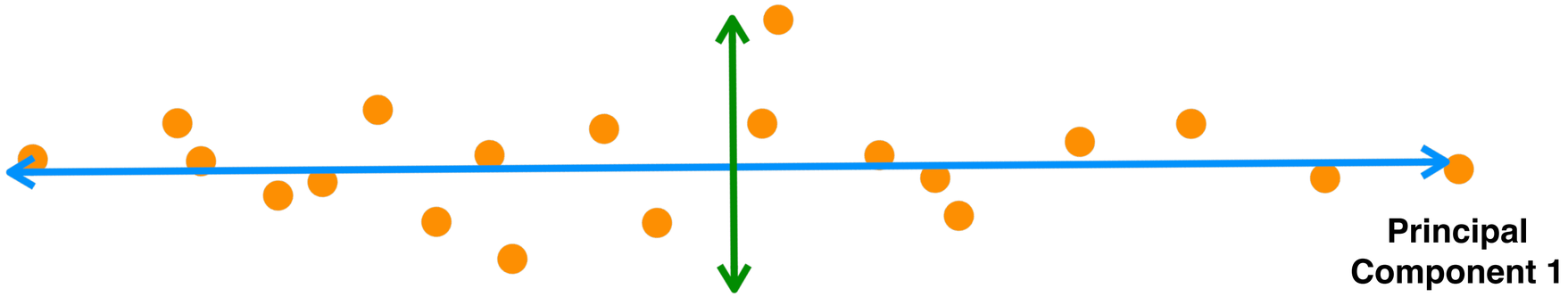
We want to find 2 perpendicular axes that encapsulate the most information (variation) possible.

The Big Idea: How Does PCA Work?



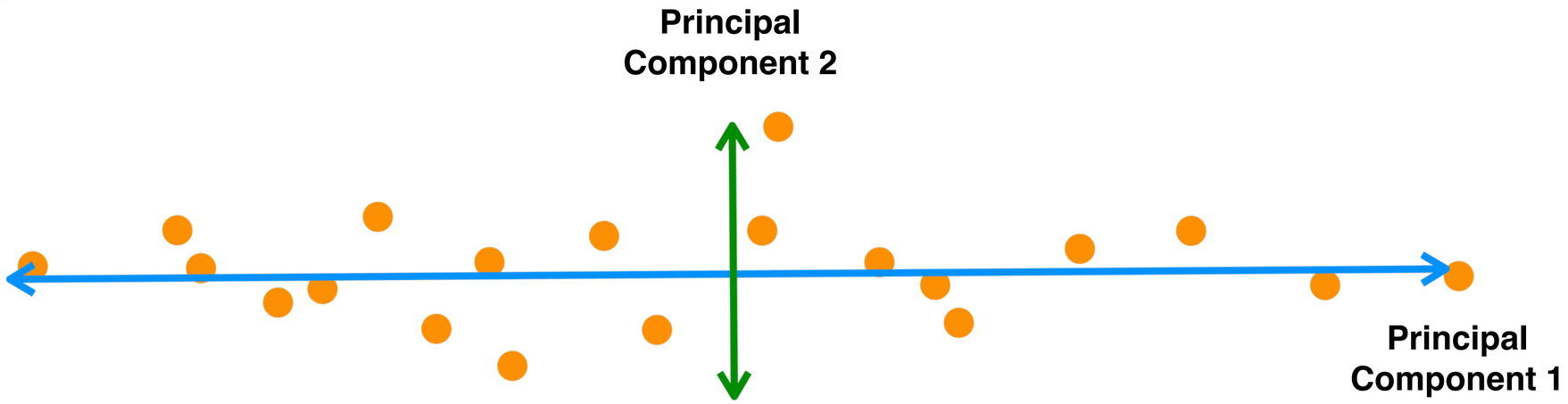
The Big Idea: How Does PCA Work?

UNIVERSITY OF CAMBRIDGE



Principal Component 1 (PC1) captures the direction where most of the data variation occurs.

The Big Idea: How Does PCA Work?

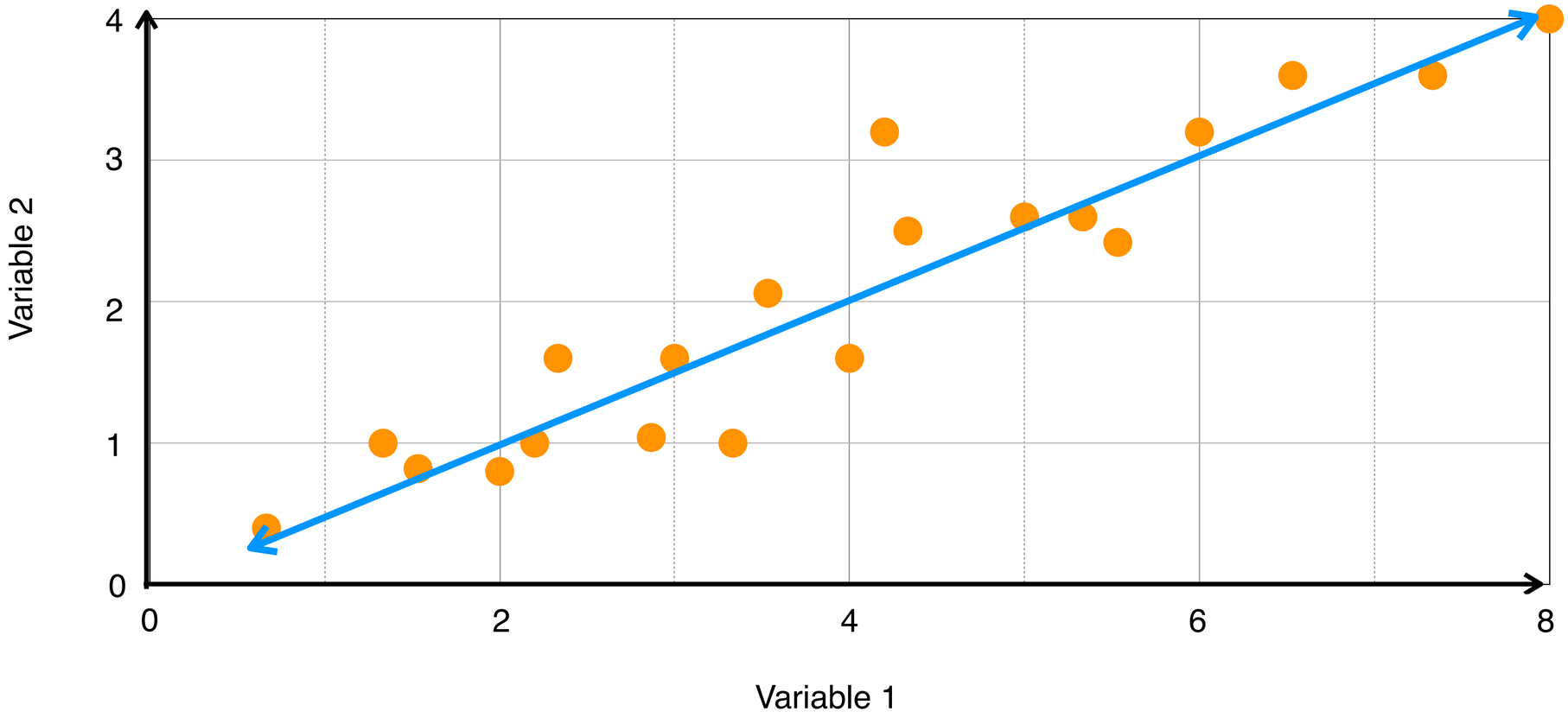


Principal Component 1 (PC1) captures the direction where most of the data variation occurs.

Principal Component 2 (PC2) captures the direction where the 2nd most data variation occurs.

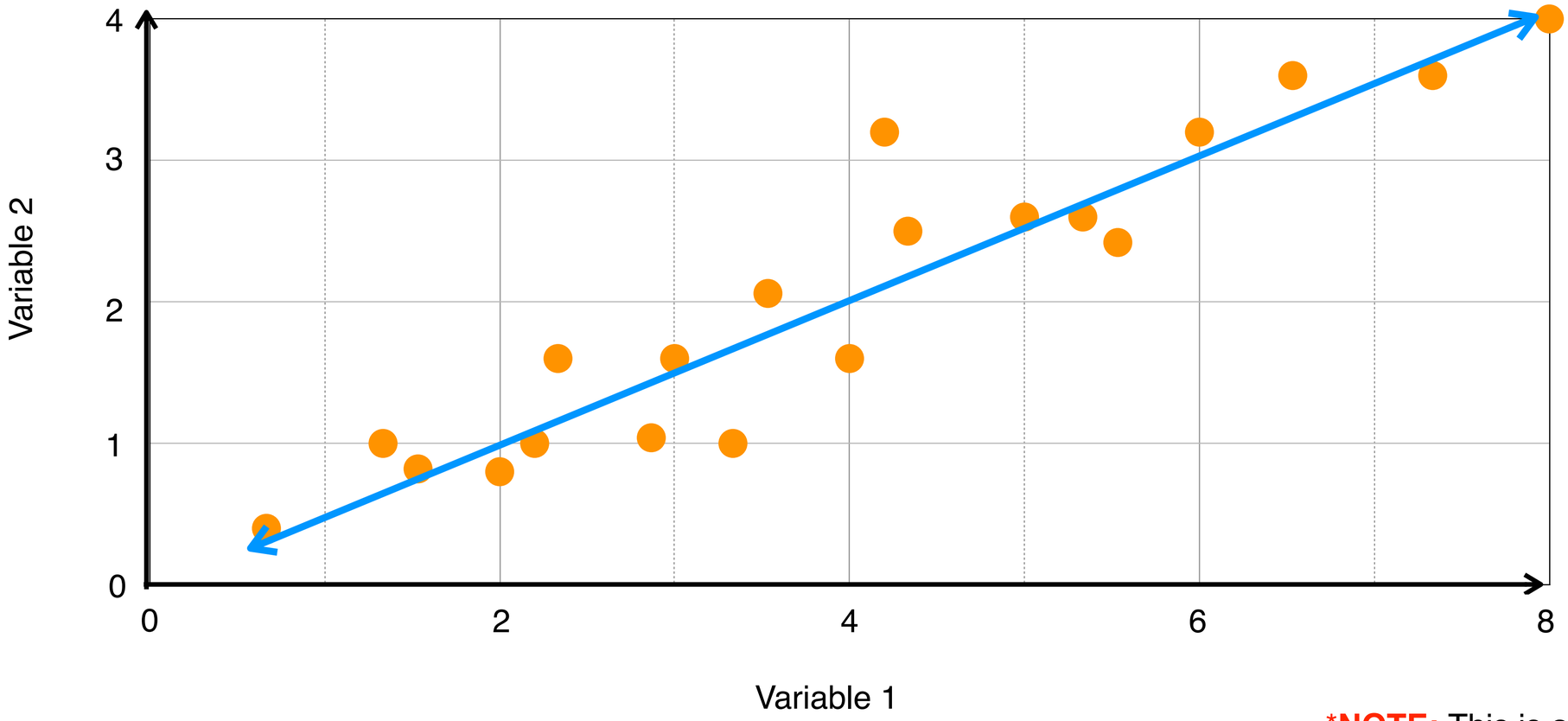
Beware! Principal Components Aren't "Real World" Variables!

UNIVERSITY OF CAMBRIDGE



Principal Component 1 = (2 • Variable 1) + (1 • Variable 2)

Beware! Principal Components Aren't "Real World" Variables!



Principal Component 1 ≈ (2 • Variable 1) + (1 • Variable 2)*

*__NOTE:__ This is a slight simplification! In reality PC1 would be expressed as a linear combination of variable 1 and 2 that has this 2:1 relationship, but has been scaled up or down.

UNIVERSITY OF CAMBRIDGE

Guiding library design

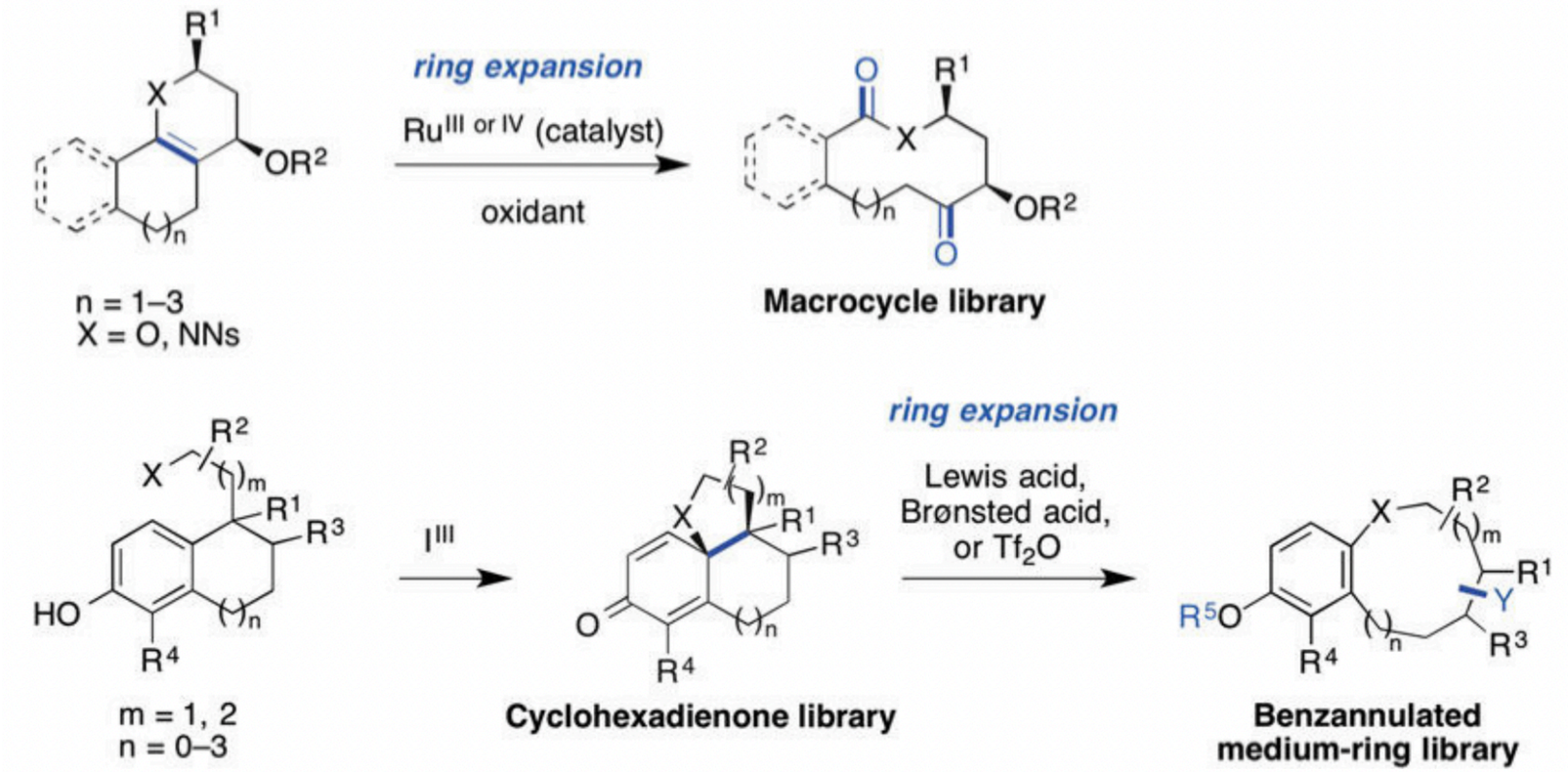


Fig. 1.
Recently developed routes to natural product-like macrocycle and medium-ring libraries using ring expansion strategies

PCA: A Real World Example

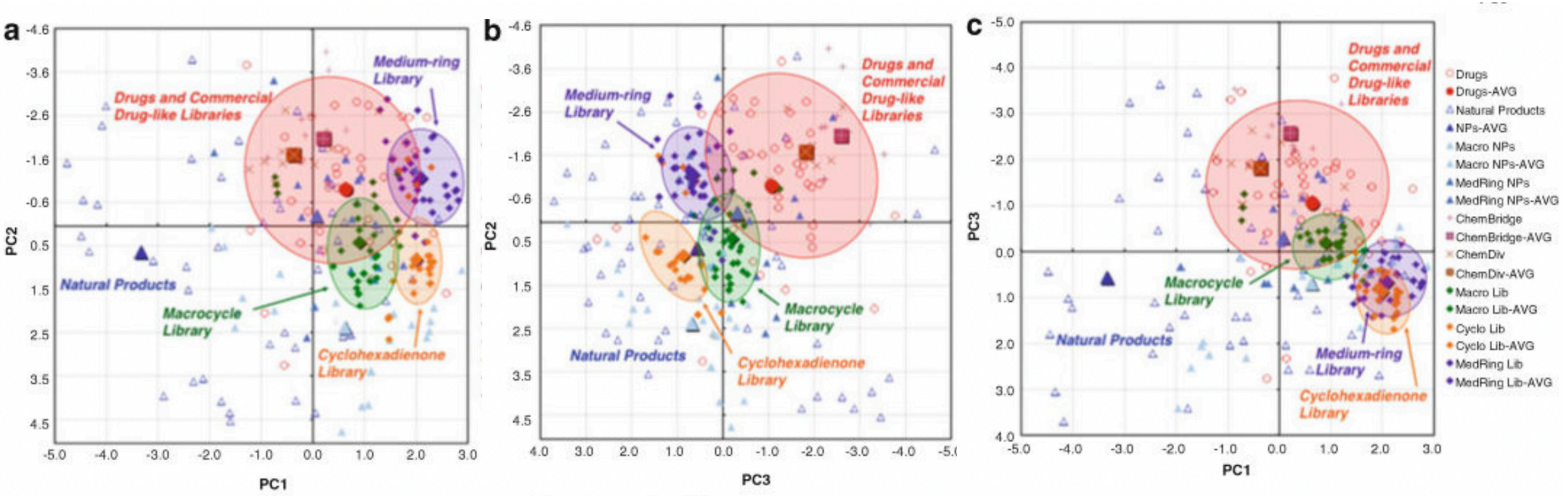
UNIVERSITY OF CAMBRIDGE

PCA to take 20-dimensional molecular featurization down to 3D.

	PC1	PC2	PC3
MW	-0.3207	-0.1482	0.1008
N	-0.2080	-0.1055	-0.3331
O	-0.3180	0.0179	0.1159
HBD	-0.3087	0.0677	-0.1180
HBA	-0.3346	-0.0057	-0.0256
RotB	-0.2587	-0.1209	-0.0540
tPSA	-0.3336	0.0109	-0.0818
nStereo	-0.2789	0.0936	0.2882
nStMW	-0.1612	0.2880	0.2964
Rings	-0.1669	-0.2384	0.0698
RngAr	-0.0475	-0.3660	-0.3227
RngSys	-0.1761	-0.2315	-0.1506
RngLg	-0.1459	-0.0128	0.2521
RRSys	-0.0015	-0.0362	0.1920
ALOGPs	0.0859	-0.4127	0.2677
ALOGpS	0.0378	0.4319	-0.2060
Fsp3	-0.0910	0.3464	0.3224
LogD	0.1850	-0.2569	0.3114
relPSA	-0.1797	0.2262	-0.3112
VWSA	-0.3122	-0.1253	0.1626

Methods Mol. Bio. **2015**, *1263*, 225.

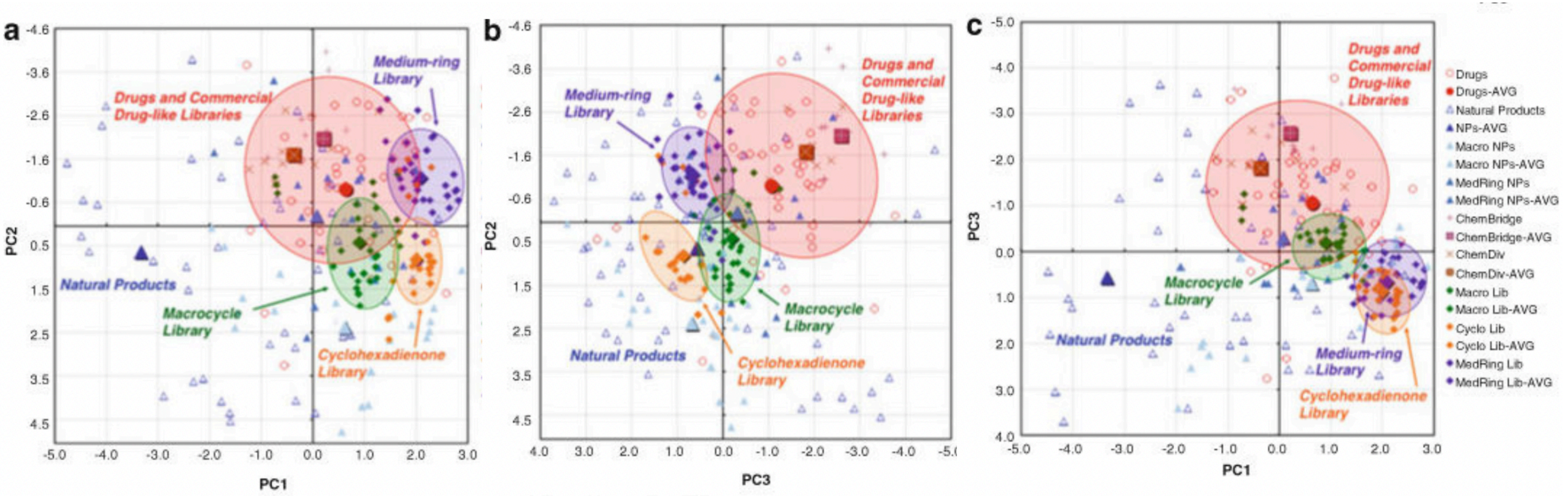
PCA: A Real World Example



Modify library molecules so that they fall into natural product space using principal component factors.

Wenderski *et al. Methods Mol. Bio.* **2015**, *1263*, 225.

PCA: A Real World Example



Modify library molecules so that they fall into natural product space using principal component factors.

Move PC1 left and PC2 down

PCA: A Real World Example

Important (negatively correlated) features for PC1 are molecular weight, number of oxygen atoms, and number of hydrogen bond donors and acceptors.

	PC1	PC2	PC3
MW	-0.3207	-0.1482	0.1008
N	-0.2080	-0.1055	-0.3331
O	-0.3180	0.0179	0.1159
HBD	-0.3087	0.0677	-0.1180
HBA	-0.3346	-0.0057	-0.0256
RotB	-0.2587	-0.1209	-0.0540
tPSA	-0.3336	0.0109	-0.0818
nStereo	-0.2789	0.0936	0.2882
nStMW	-0.1612	0.2880	0.2964
Rings	-0.1669	-0.2384	0.0698
RngAr	-0.0475	-0.3660	-0.3227
RngSys	-0.1761	-0.2315	-0.1506
RngLg	-0.1459	-0.0128	0.2521
RRSys	-0.0015	-0.0362	0.1920
ALOGPs	0.0859	-0.4127	0.2677
ALOGpS	0.0378	0.4319	-0.2060
Fsp3	-0.0910	0.3464	0.3224
LogD	0.1850	-0.2569	0.3114
relPSA	-0.1797	0.2262	-0.3112
VWSA	-0.3122	-0.1253	0.1626

Wenderski *et al. Methods Mol. Bio.* **2015**, *1263*, 225.

PCA: A Real World Example

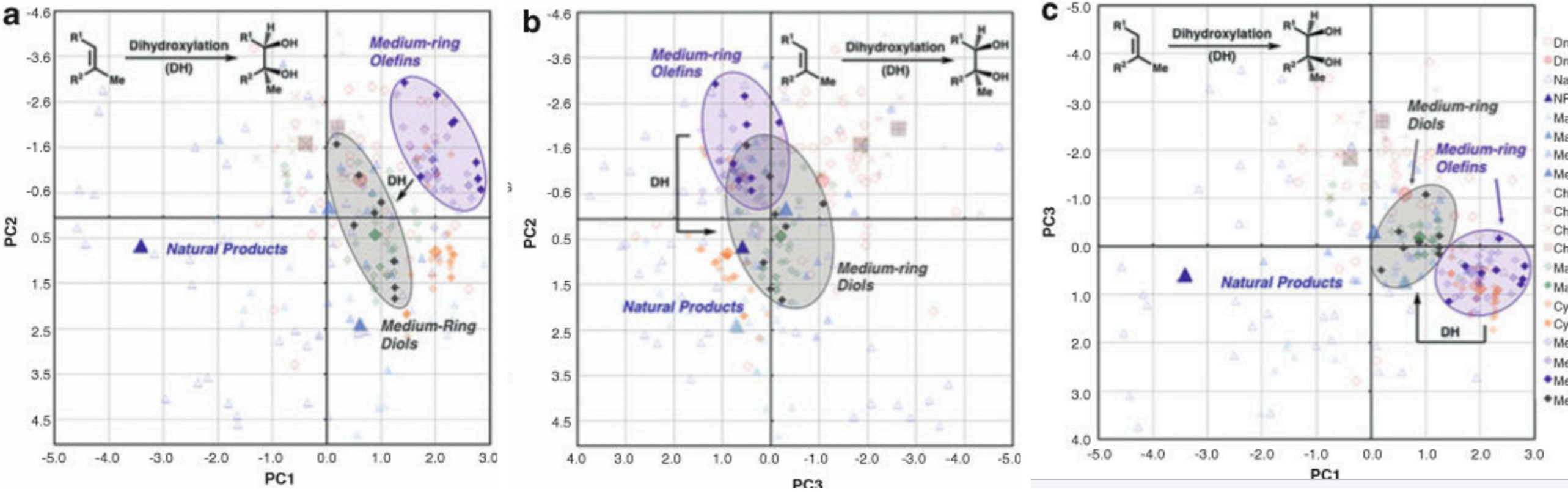
UNIVERSITY OF CAMBRIDGE

Important features for PC2 are density of stereocenters and the number of sp^3 bonds.

	PC1	PC2	PC3
MW	-0.3207	-0.1482	0.1008
N	-0.2080	-0.1055	-0.3331
O	-0.3180	0.0179	0.1159
HBD	-0.3087	0.0677	-0.1180
HBA	-0.3346	-0.0057	-0.0256
RotB	-0.2587	-0.1209	-0.0540
tPSA	-0.3336	0.0109	-0.0818
nStereo	-0.2789	0.0936	0.2882
nStMW	-0.1612	0.2880	0.2964
Rings	-0.1669	-0.2384	0.0698
RngAr	-0.0475	-0.3660	-0.3227
RngSys	-0.1761	-0.2315	-0.1506
RngLg	-0.1459	-0.0128	0.2521
RRSys	-0.0015	-0.0362	0.1920
ALOGPs	0.0859	-0.4127	0.2677
ALOGpS	0.0378	0.4319	-0.2060
Fsp3	-0.0910	0.3464	0.3224
LogD	0.1850	-0.2569	0.3114
relPSA	-0.1797	0.2262	-0.3112
VWSA	-0.3122	-0.1253	0.1626

Wenderski *et al. Methods Mol. Bio.* **2015**, *1263*, 225.

PCA: A Real World Example



Answers typically a “yes/no” question.

Is this a cat?

Answers typically a “yes/no” question.

Is this a cat?

Yes, this is **a cat**

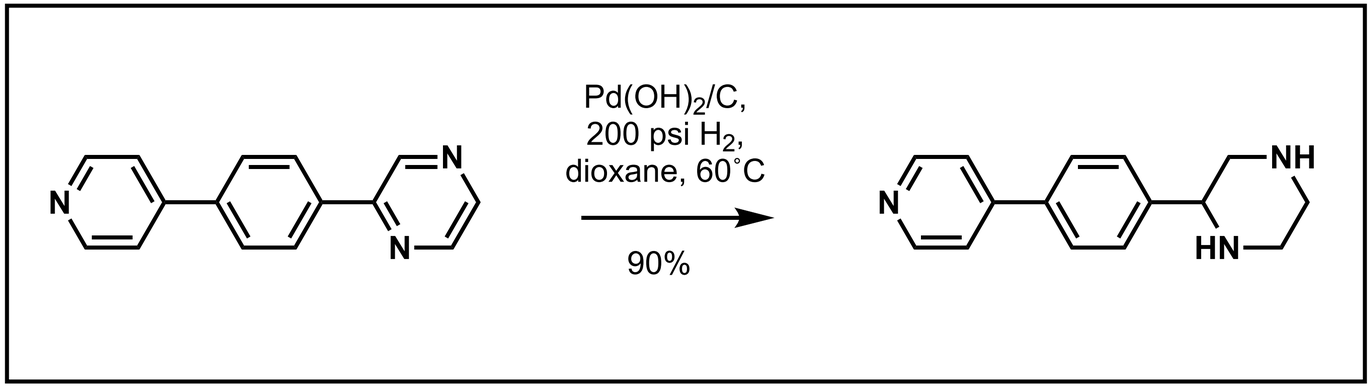
Answers typically a “yes/no” question.

Is this a cat?

No, this is a **not a cat**

Answers typically a “yes/no” question.

Is this reaction high yielding?

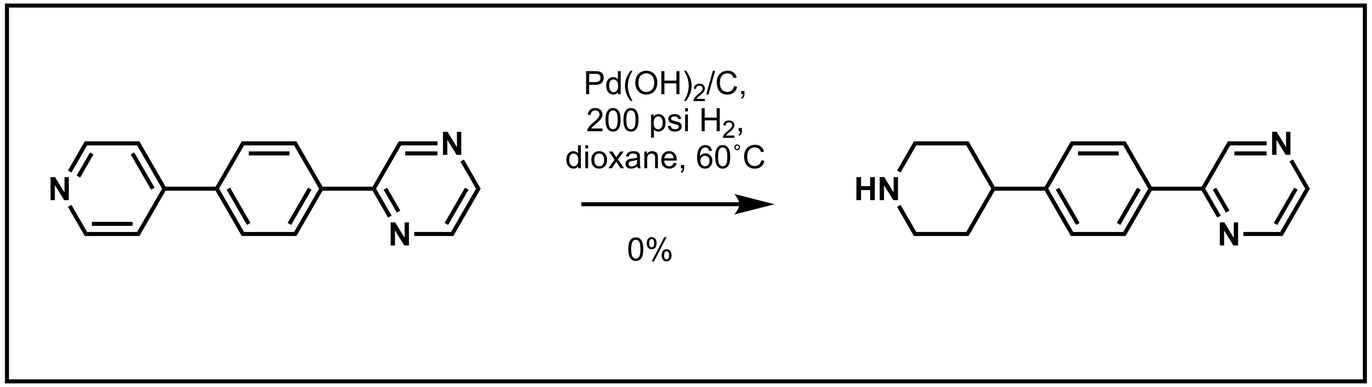


Yes, is a **high yielding** rxn

Lyons *et al. J. Org. Chem.* **2024**, *89*, 1438.

Answers typically a “yes/no” question.

Is this reaction high yielding?



No, this is **not a high yielding** rxn

Lyons *et al. J. Org. Chem.* **2024**, *89*, 1438.

Can also assign labels from a predetermined list

What is this? A cat? A dog? An owl?

Classification

Can also assign labels from a predetermined list

What is this? A cat? A dog? An owl?

This is **a cat**

Can also assign labels from a predetermined list

What is this? A cat? A dog? An owl?

This is **a dog**

Can also assign labels from a predetermined list

What is this? A cat? A dog? An owl?

This is **an owl**

True Positive (TP) = Correctly identified as **X**

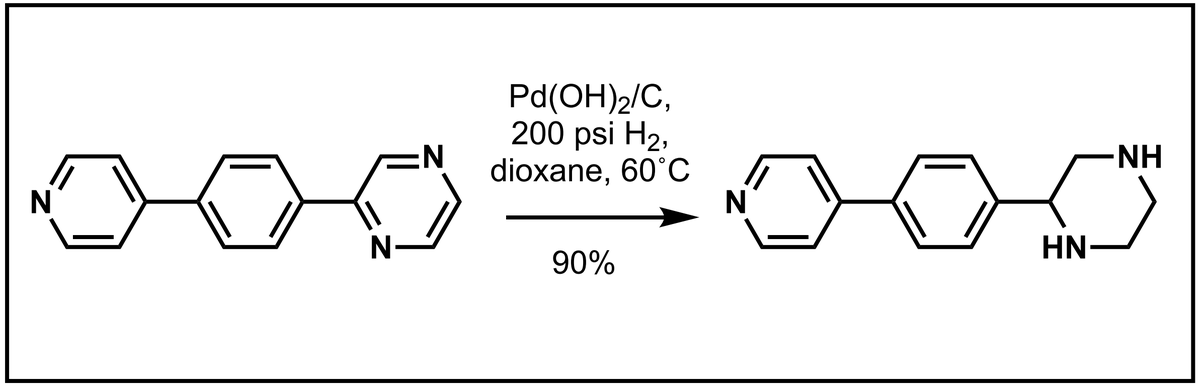
True Negative (TN) = Correctly identified as **not X**

False Positive (FP) = Incorrectly identified as **X**

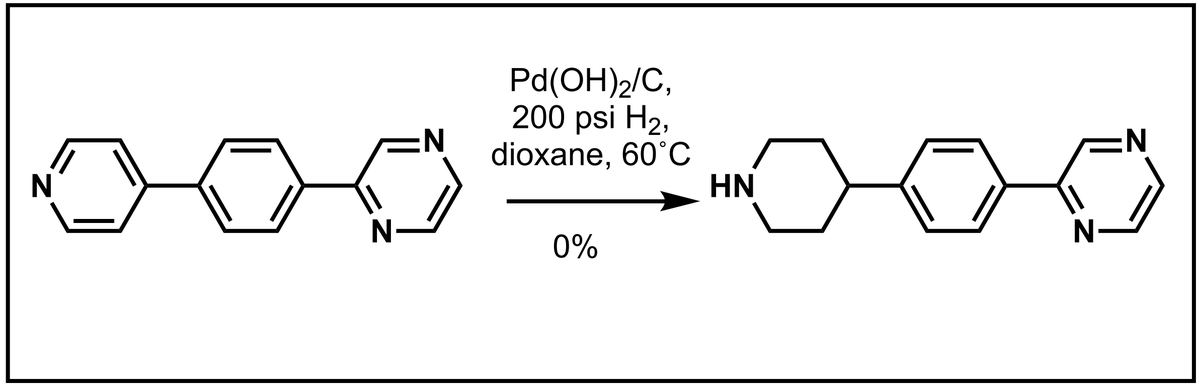
False Negative (FN) = Incorrectly identified as **not X**

Classification Metrics

Is this reaction high yielding?



Rxn A

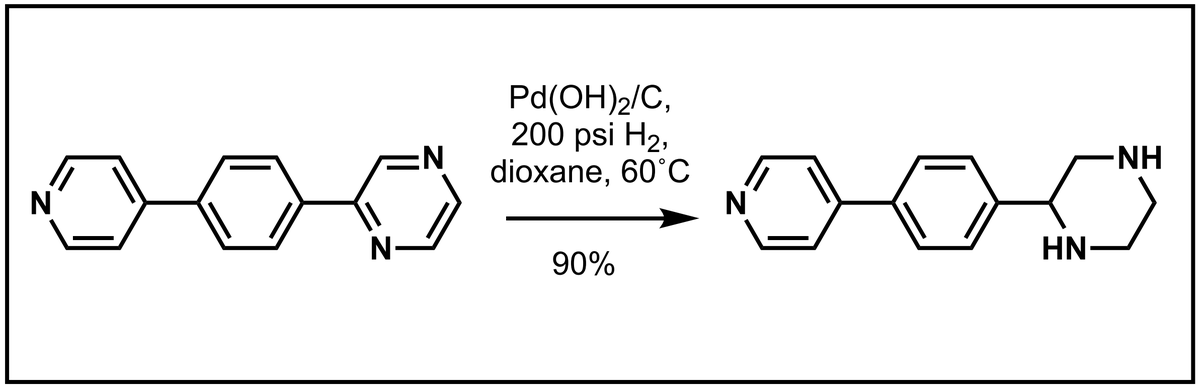


Rxn B

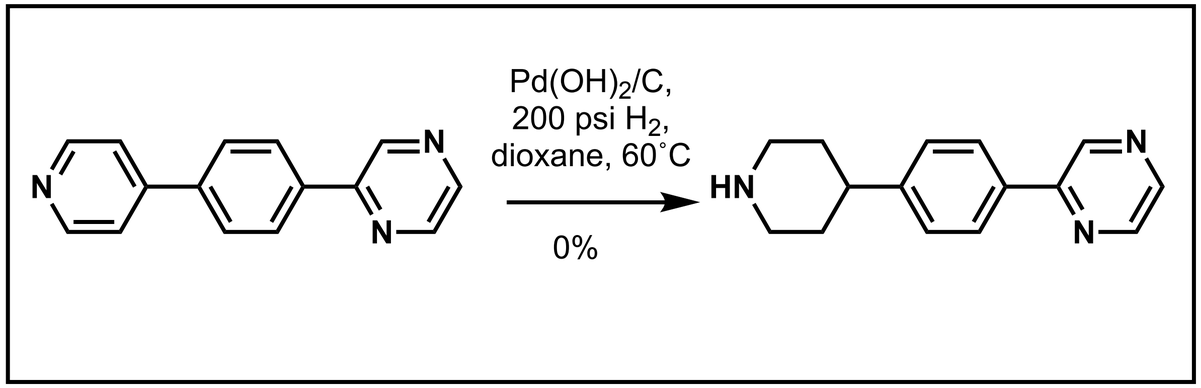
Classification	Truth	Result

Classification Metrics

Is this reaction high yielding?



Rxn A



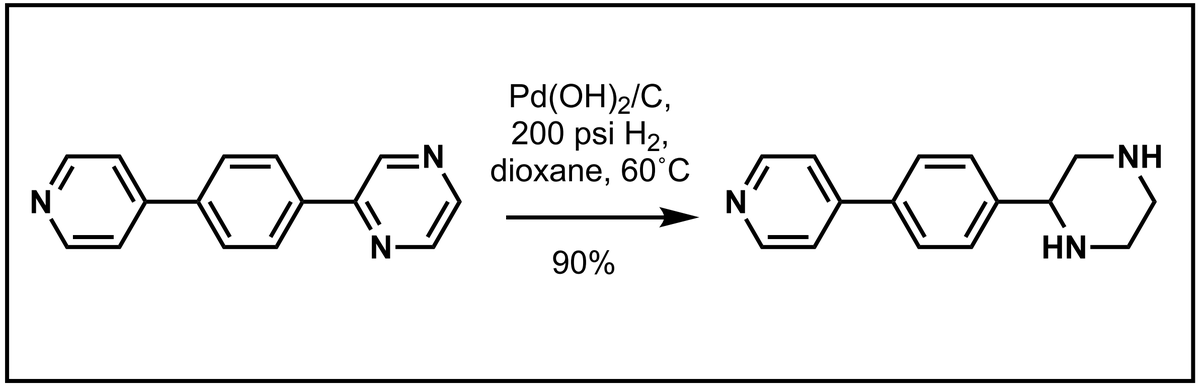
Rxn B

Classification	Truth	Result
	High Yielding	
	High Yielding	

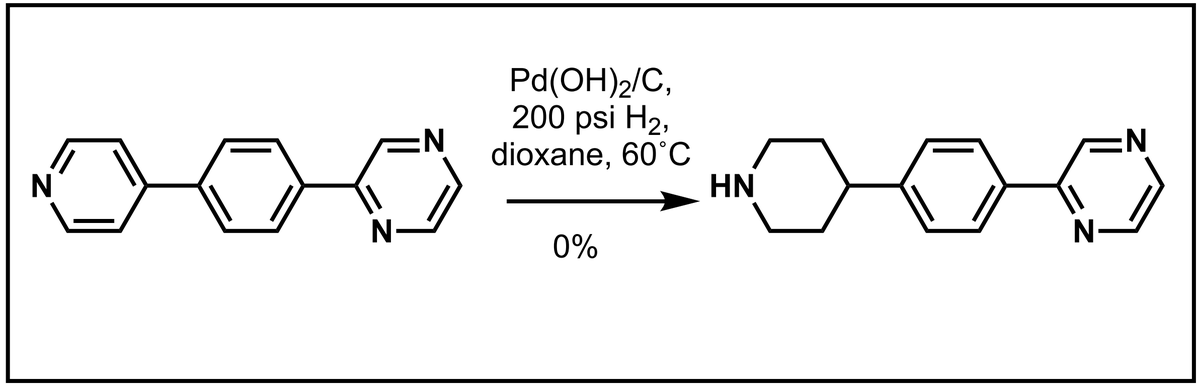
Lyons *et al. J. Org. Chem.* **2024**, *89*, 1438.

Classification Metrics

Is this reaction high yielding?



Rxn A



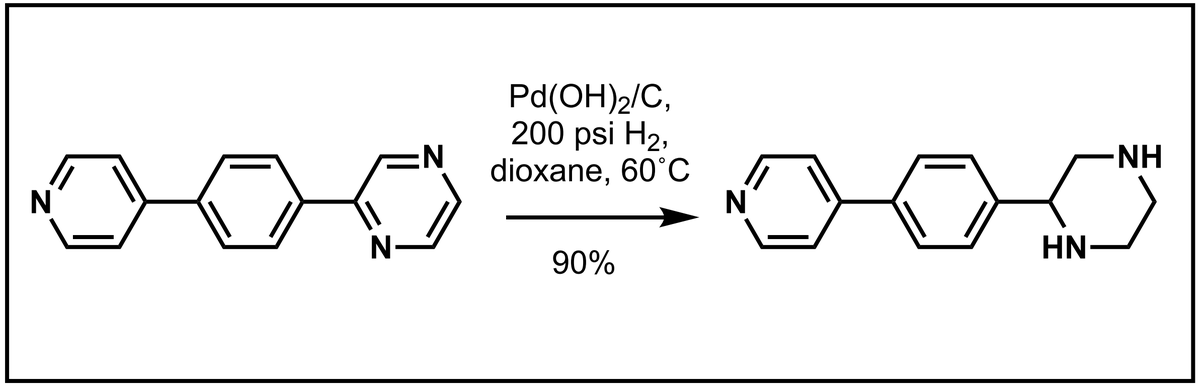
Rxn B

Classification	Truth	Result
High Yielding	High Yielding	True Positive
	High Yielding	

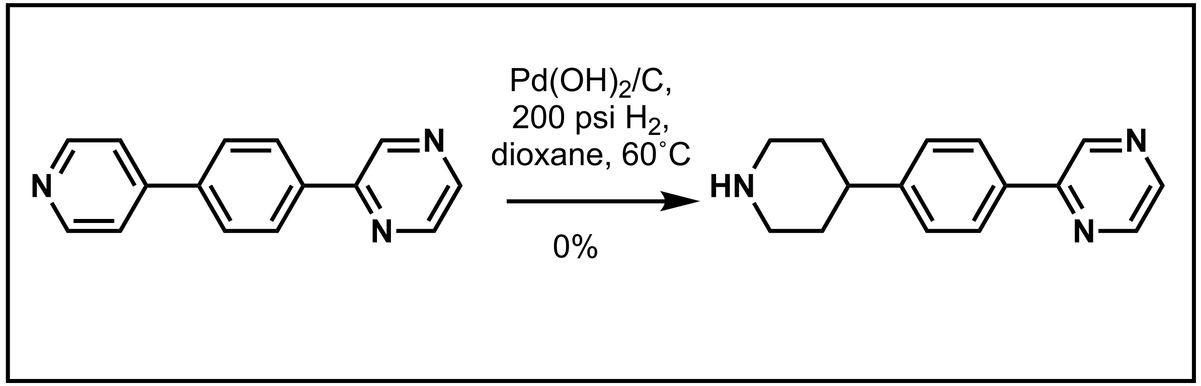
Lyons *et al. J. Org. Chem.* **2024**, *89*, 1438.

Classification Metrics

Is this reaction high yielding?



Rxn A

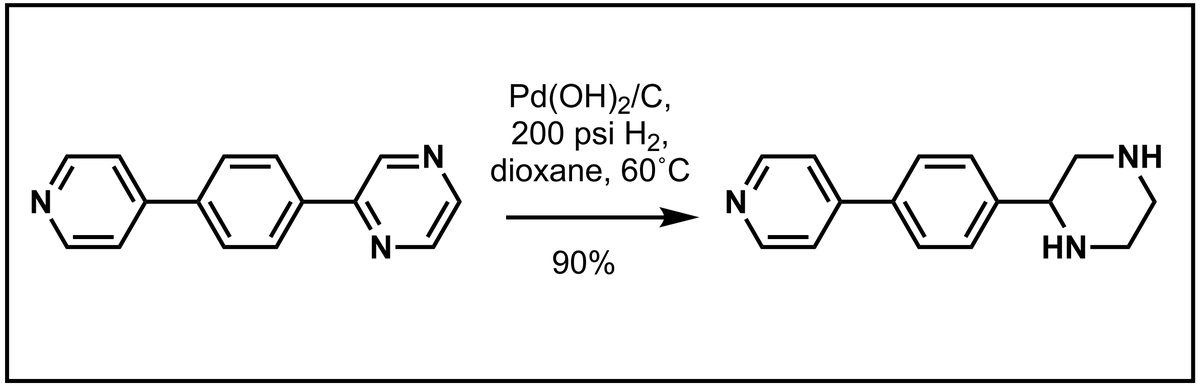


Rxn B

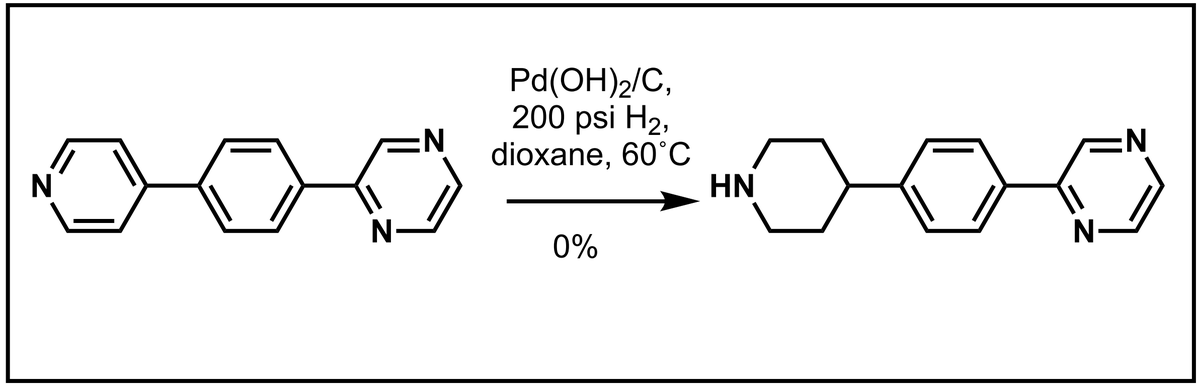
Classification	Truth	Result
High Yielding	High Yielding	True Positive
Low Yielding	High Yielding	False Negative

Classification Metrics

Is this reaction high yielding?



Rxn A



Rxn B

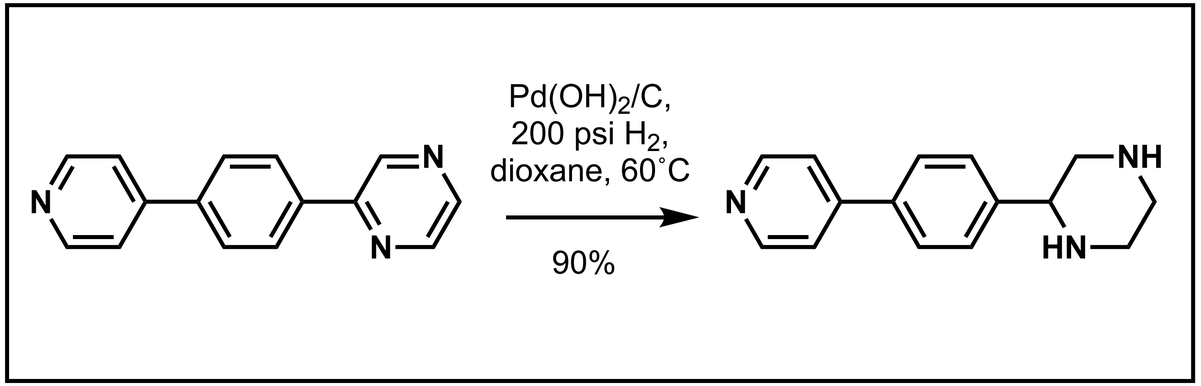
Classification	Truth	Result
High Yielding	High Yielding	True Positive
Low Yielding	High Yielding	False Negative

Classification	Truth	Result
	Low Yielding	
	Low Yielding	

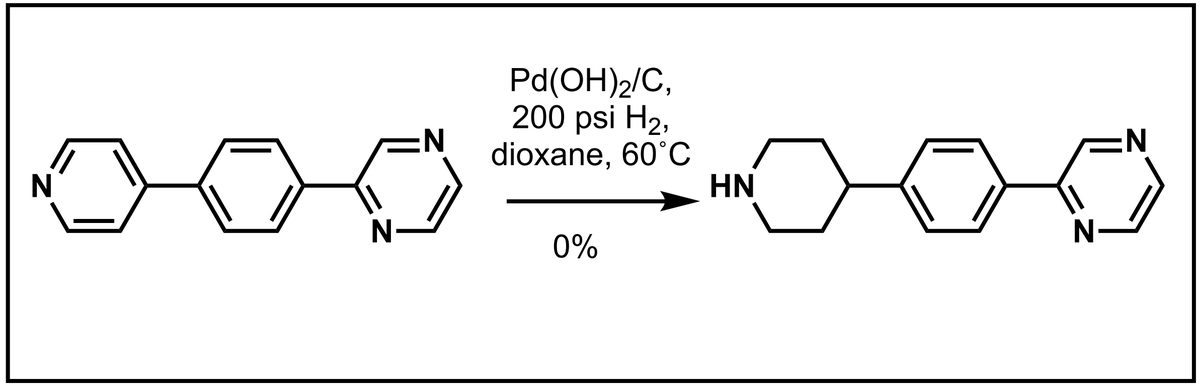
Lyons *et al. J. Org. Chem.* **2024**, *89*, 1438.

Classification Metrics

Is this reaction high yielding?



Rxn A



Rxn B

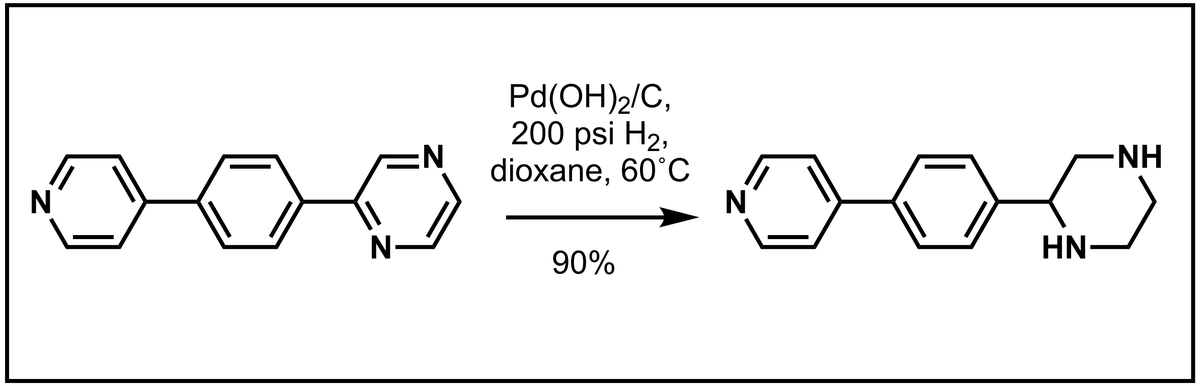
Classification	Truth	Result
High Yielding	High Yielding	True Positive
Low Yielding	High Yielding	False Negative

Classification	Truth	Result
High Yielding	Low Yielding	False Positive
	Low Yielding	

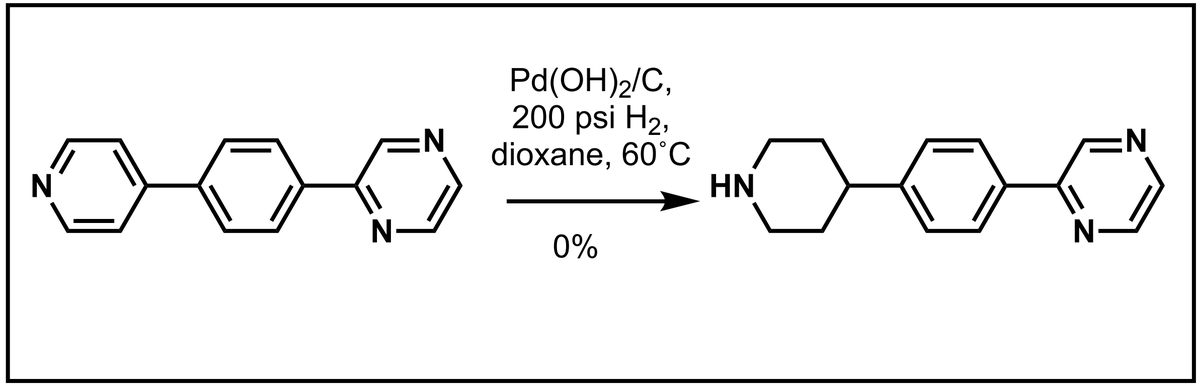
Lyons *et al. J. Org. Chem.* **2024**, *89*, 1438.

Classification Metrics

Is this reaction high yielding?



Rxn A



Rxn B

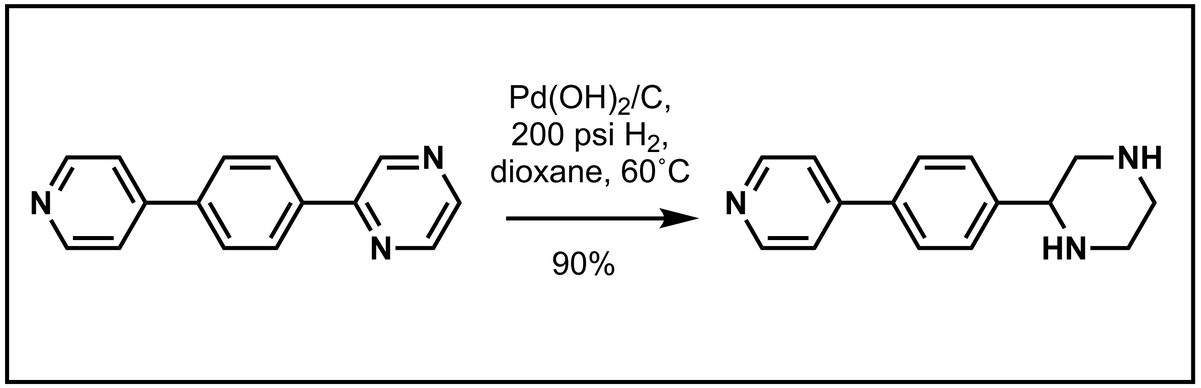
Classification	Truth	Result
High Yielding	High Yielding	True Positive
Low Yielding	High Yielding	False Negative

Classification	Truth	Result
High Yielding	Low Yielding	False Positive
Low Yielding	Low Yielding	True Negative

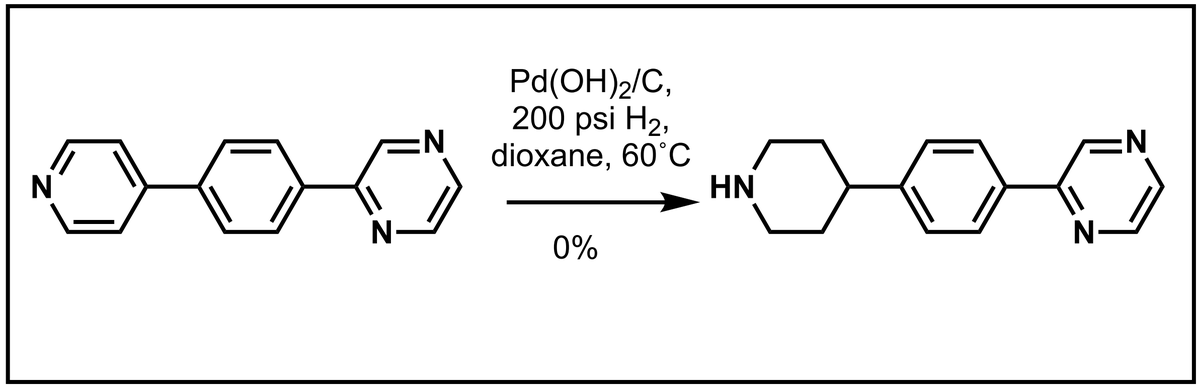
Lyons *et al. J. Org. Chem.* **2024**, *89*, 1438.

Classification Metrics

Is this reaction high yielding?



Rxn A



Rxn B

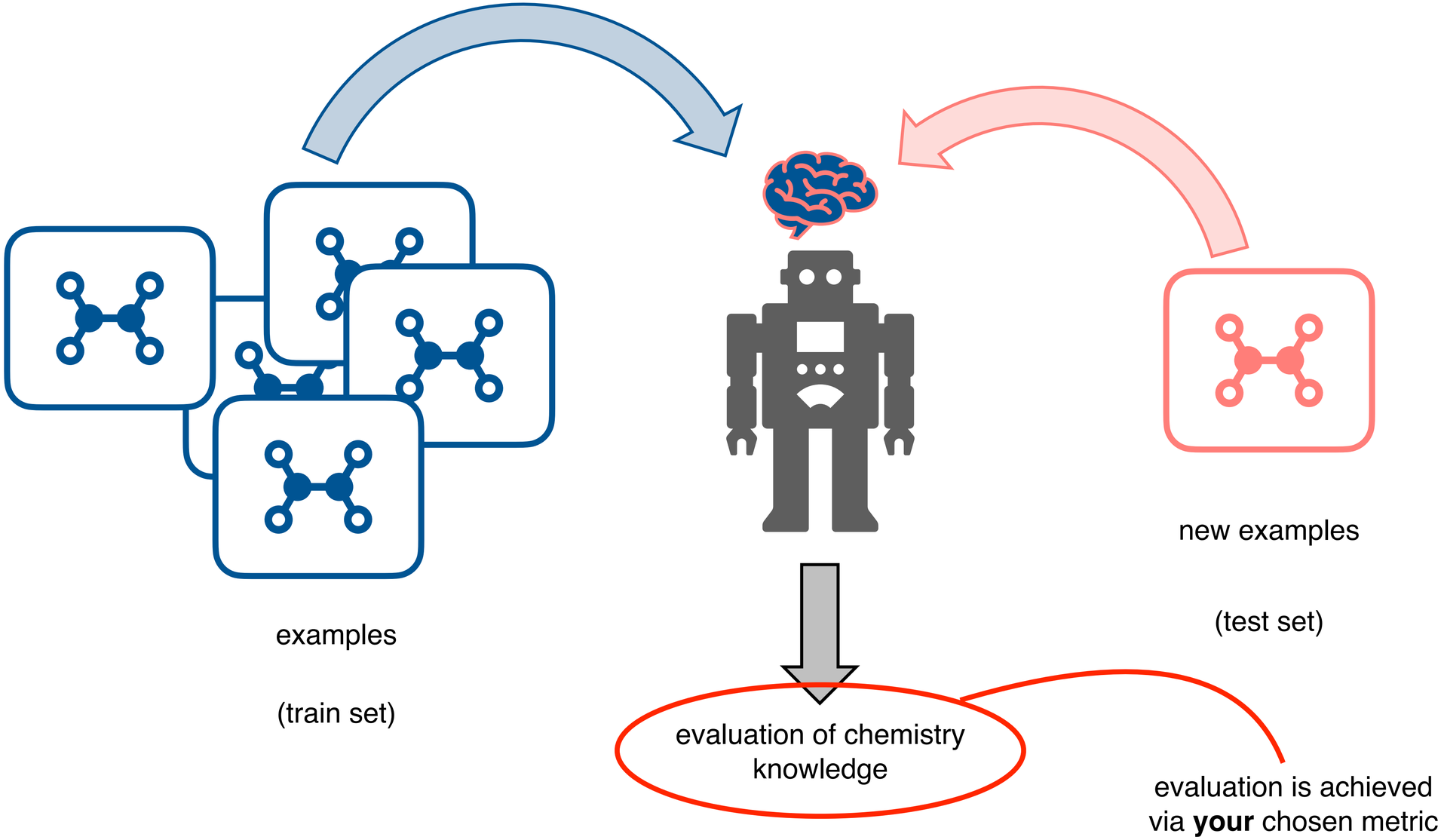
Confusion Matrix

What actually happened (ground truth)

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

Lyons *et al. J. Org. Chem.* **2024**, *89*, 1438.

Why Are Metrics **SO** Important? Recall Our Robot



UNIVERSITY OF CAMBRIDGE

Confusion Matrix

What actually happened (ground truth)

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

High accuracy model = often gets it right

UNIVERSITY OF CAMBRIDGE

Confusion Matrix

What actually happened (ground truth)

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

High accuracy model = often gets it right

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

UNIVERSITY OF CAMBRIDGE

Confusion Matrix

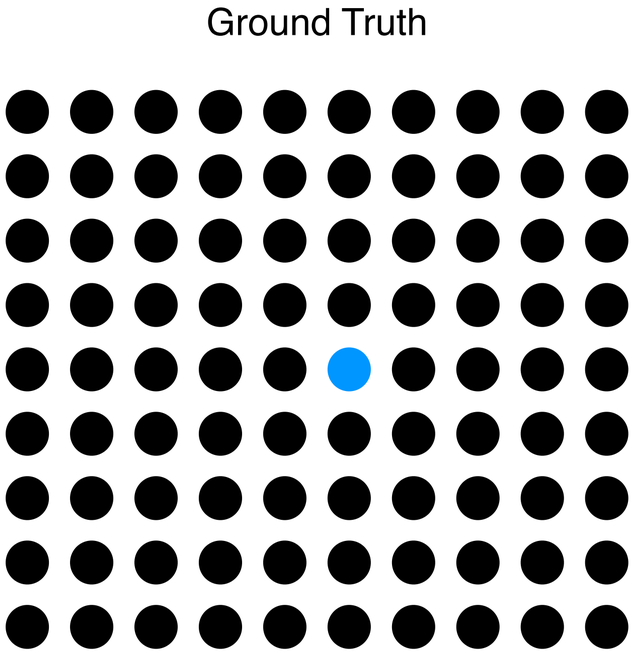
What we predicted (predictions) \ What actually happened (ground truth)	High Yielding	Not High Yielding
High Yielding	True Positive	False Positive
Not High Yielding	False Negative	True Negative

High accuracy model = often gets it right

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

Not always a good metric in unbalanced data

Classification Metrics

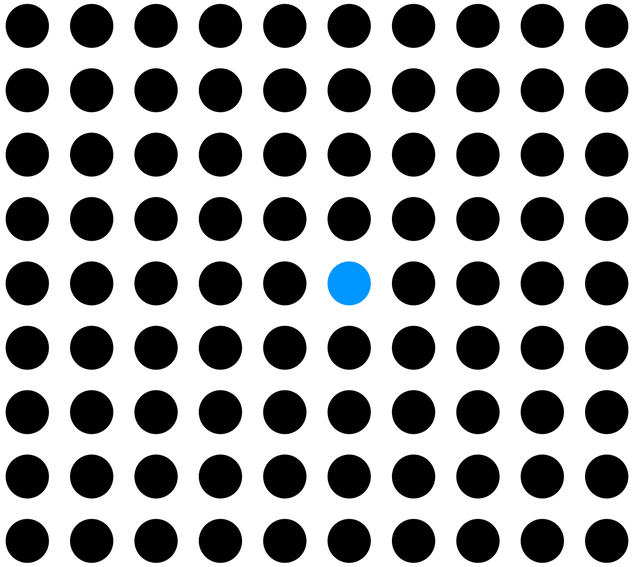


$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

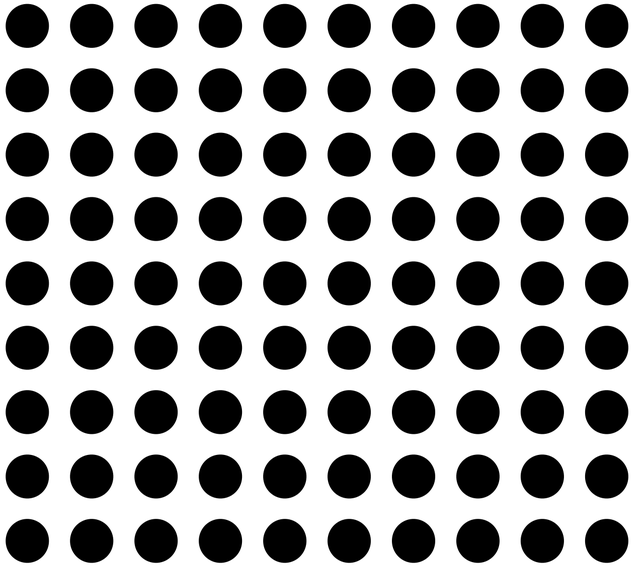
Classification Metrics

UNIVERSITY OF CAMBRIDGE

Ground Truth



Predictions



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

Accuracy = 99%

UNIVERSITY OF CAMBRIDGE

Confusion Matrix

What actually happened (ground truth)

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

High precision model = when it does predict a TP, it gets it right.

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

High precision model = when it does predict a TP, it gets it right.

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

UNIVERSITY OF CAMBRIDGE

Confusion Matrix

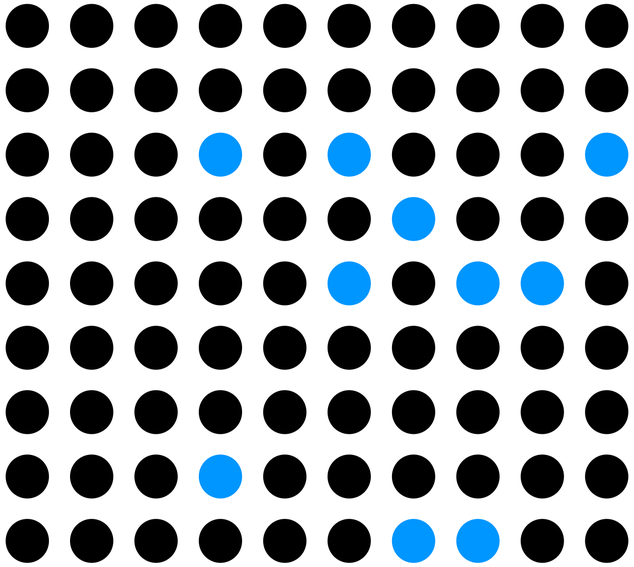
What we predicted (predictions) \ What actually happened (ground truth)	High Yielding	Not High Yielding
High Yielding	True Positive	False Positive
Not High Yielding	False Negative	True Negative

High precision model = when it does predict a TP, it gets it right.

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

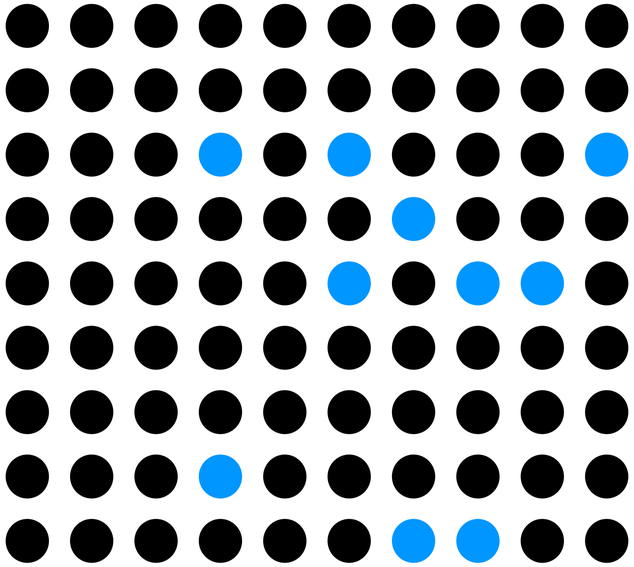
Precision says nothing about how good the model is with finding all the positives.

Ground Truth

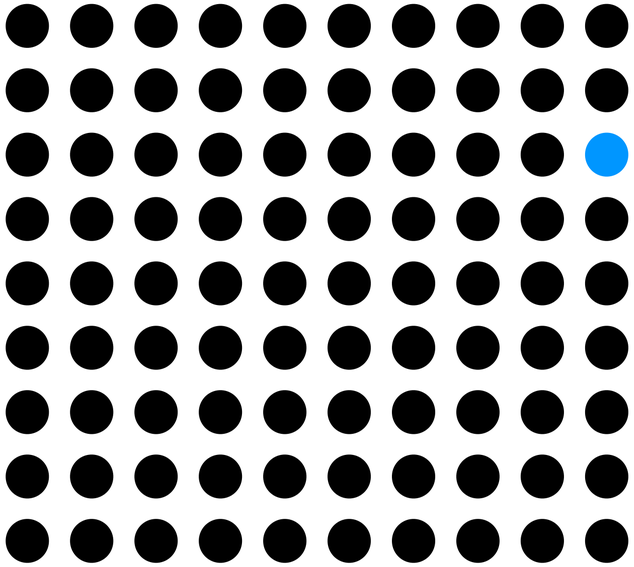


$$\text{Precision} = \frac{TP}{TP + FP}$$

Ground Truth



Predictions



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

Precision = 100%

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

High recall model = finds most of the TP.

UNIVERSITY OF CAMBRIDGE

Confusion Matrix

What actually happened (ground truth)

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

High recall model = finds most of the TP.

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

UNIVERSITY OF CAMBRIDGE

Confusion Matrix

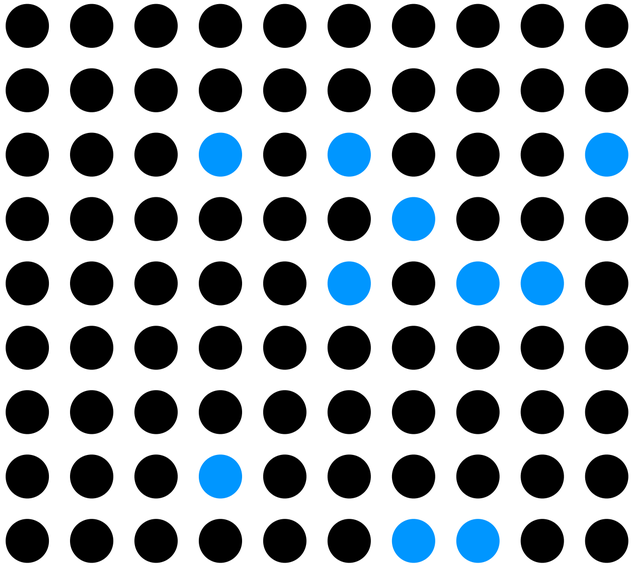
What we predicted (predictions) \ What actually happened (ground truth)	High Yielding	Not High Yielding
High Yielding	True Positive	False Positive
Not High Yielding	False Negative	True Negative

High recall model = finds most of the TP.

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

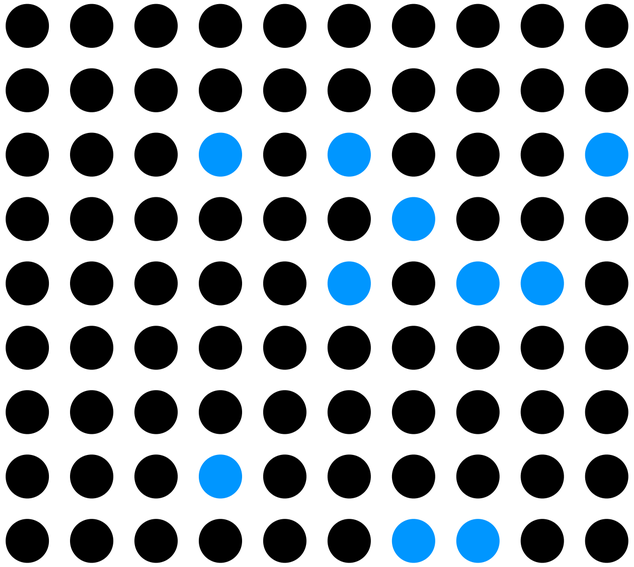
Recall says nothing about about how good the model is at distinguishing between true positives and false positives.

Ground Truth

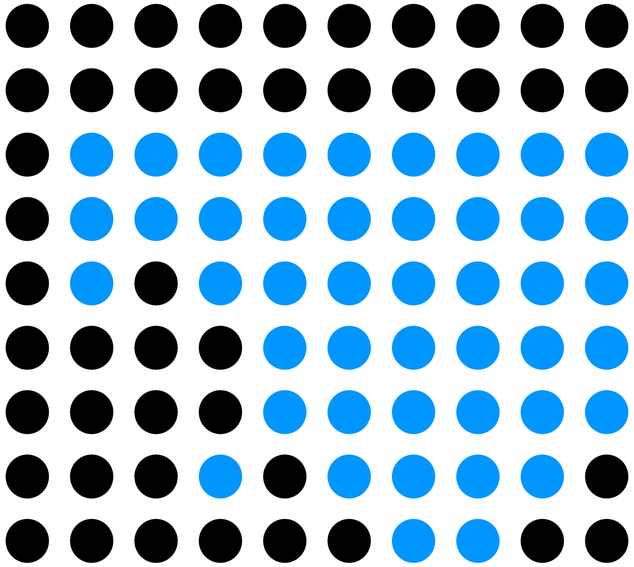


$$\text{Recall} = \frac{TP}{TP + FN}$$

Ground Truth



Predictions



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

Recall = 100%

UNIVERSITY OF CAMBRIDGE

Confusion Matrix

What actually happened (ground truth)

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

It is up to **you** if you'd like to prioritize high precision or high recall.

UNIVERSITY OF CAMBRIDGE

Confusion Matrix

What actually happened (ground truth)

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

It is up to **you** if you'd like to prioritize high precision or high recall.

Oftentimes, we'd like both.

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

It is up to **you** if you'd like to prioritize high precision or high recall.

Oftentimes, we'd like both.

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

SVMs: Separating Data Points

UNIVERSITY OF CAMBRIDGE

Less about visualization, more about prediction

Height	Job?
50”	No
60"	Yes
52”	No
64”	Yes
49”	No
51.5”	No
48”	No
63”	Yes
51”	No
65”	Yes

SVMs: Separating Data Points

UNIVERSITY OF CAMBRIDGE

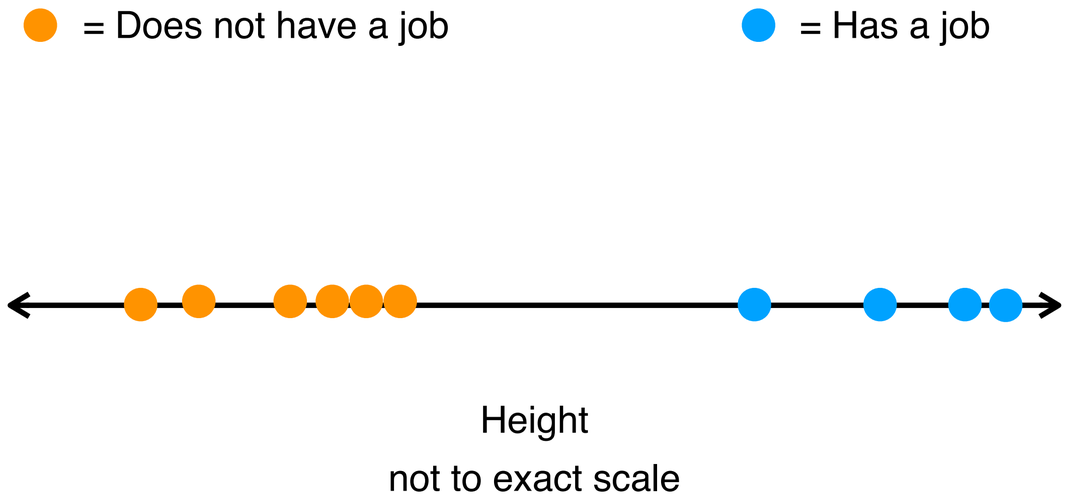
Less about visualization, more about prediction

Height	Job?
50”	No
60"	Yes
52”	No
64”	Yes
49”	No
51.5”	No
48”	No
63”	Yes
51”	No
65”	Yes

SVMs: Separating Data Points

Less about visualization, more about prediction

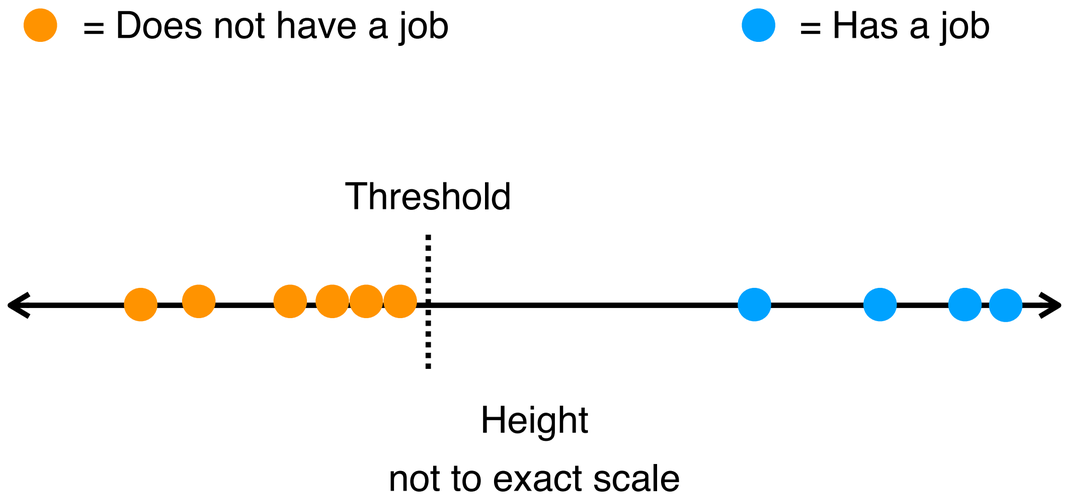
Height	Job?
50”	No
60"	Yes
52”	No
64”	Yes
49”	No
51.5”	No
48”	No
63”	Yes
51”	No
65”	Yes



SVMs: Separating Data Points

Less about visualization, more about prediction

Height	Job?
50”	No
60"	Yes
52”	No
64”	Yes
49”	No
51.5”	No
48”	No
63”	Yes
51”	No
65”	Yes



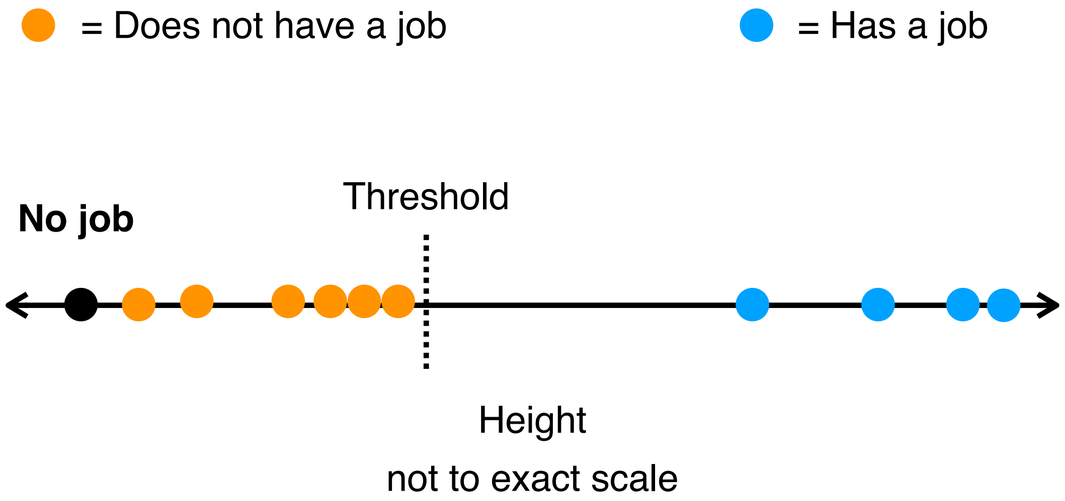
Above threshold = predicted to have a job

Below threshold = predicted to not have a job

SVMs: Separating Data Points

Less about visualization, more about prediction

Height	Job?
50”	No
60"	Yes
52”	No
64”	Yes
49”	No
51.5”	No
48”	No
63”	Yes
51”	No
65”	Yes



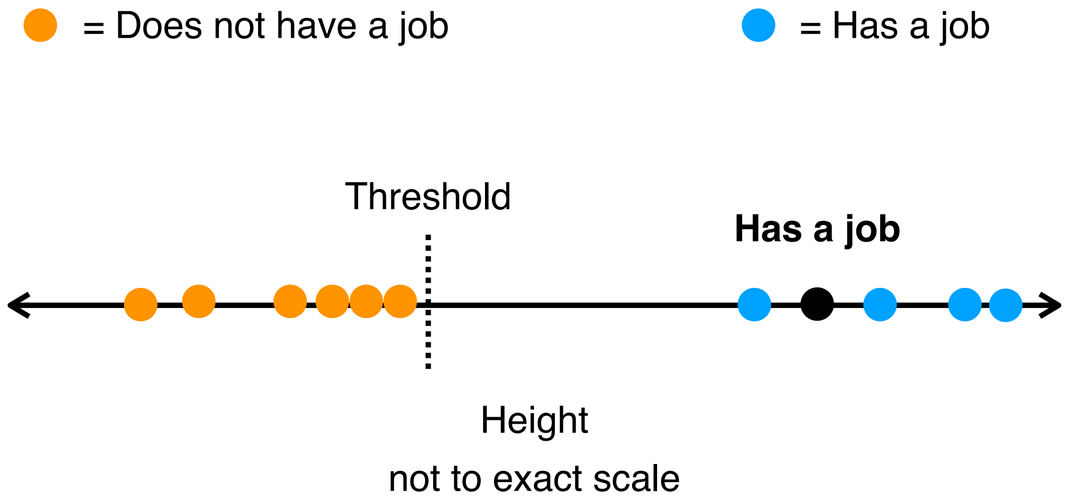
Above threshold = predicted to have a job

Below threshold = predicted to not have a job

SVMs: Separating Data Points

Less about visualization, more about prediction

Height	Job?
50”	No
60"	Yes
52”	No
64”	Yes
49”	No
51.5”	No
48”	No
63”	Yes
51”	No
65”	Yes



Above threshold = predicted to have a job

Below threshold = predicted to not have a job

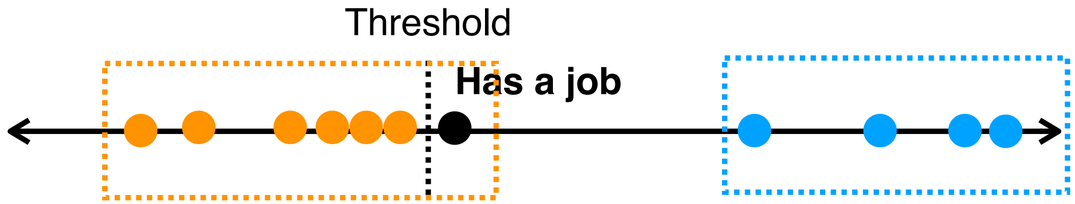
SVMs: Separating Data Points

UNIVERSITY OF CAMBRIDGE

Less about visualization, more about prediction

Height	Job?
50”	No
60"	Yes
52”	No
64”	Yes
49”	No
51.5”	No
48”	No
63”	Yes
51”	No
65”	Yes

● = Does not have a job ● = Has a job



Height

not to exact scale

Above threshold = predicted to have a job

Below threshold = predicted to not have a job

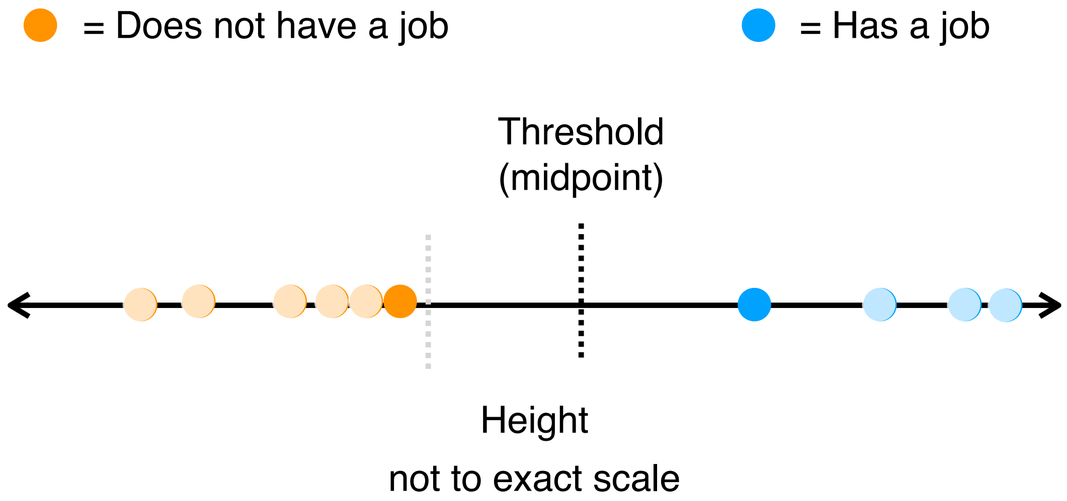
Even though ● is classified as having a job, it’s closer to ● than ●.

Not the ideal threshold.

SVMs: Separating Data Points

Less about visualization, more about prediction

Height	Job?
50”	No
60"	Yes
52”	No
64”	Yes
49”	No
51.5”	No
48”	No
63”	Yes
51”	No
65”	Yes



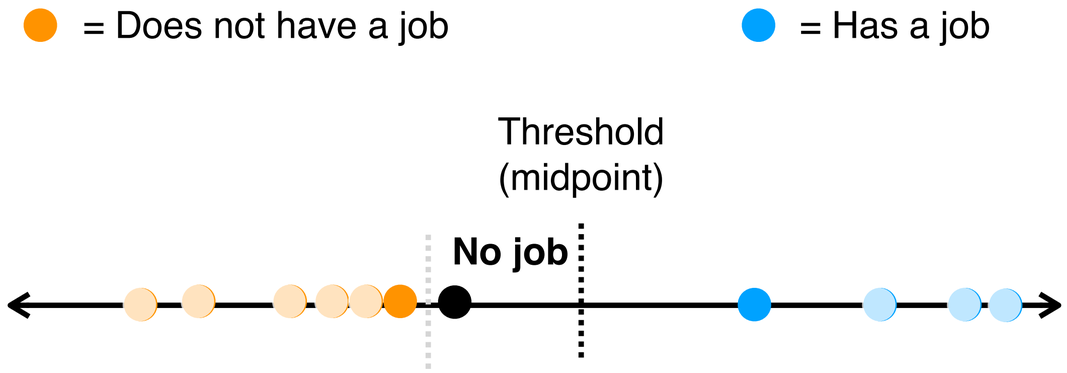
Above threshold = predicted to have a job

Below threshold = predicted to not have a job

SVMs: Separating Data Points

Less about visualization, more about prediction

Height	Job?
50”	No
60"	Yes
52”	No
64”	Yes
49”	No
51.5”	No
48”	No
63”	Yes
51”	No
65”	Yes



Above threshold = predicted to have a job

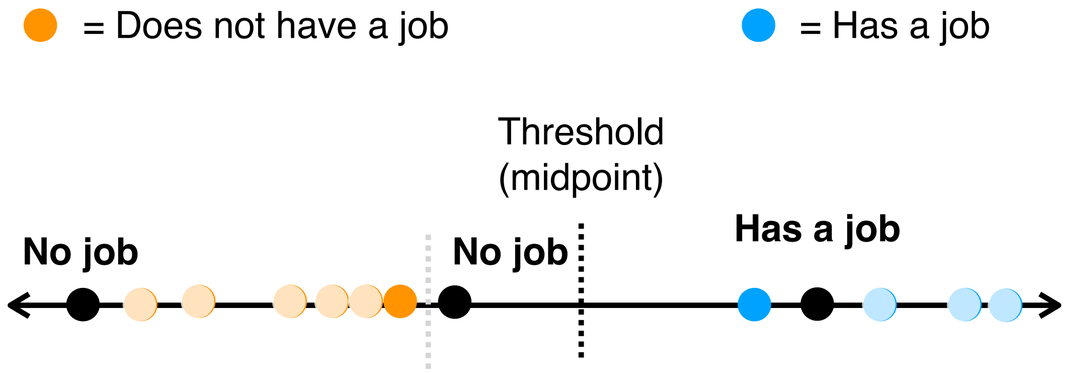
Below threshold = predicted to not have a job

Now ● classified to the class it’s closest to (●).

SVMs: Separating Data Points

Less about visualization, more about prediction

Height	Job?
50”	No
60"	Yes
52”	No
64”	Yes
49”	No
51.5”	No
48”	No
63”	Yes
51”	No
65”	Yes



* not to exact scale

Above threshold = predicted to have a job

Below threshold = predicted to not have a job

Now ● classified to the class it’s closest to (●).

Other points still fall within their clusters.

SVMs: Separating Data Points

UNIVERSITY OF CAMBRIDGE

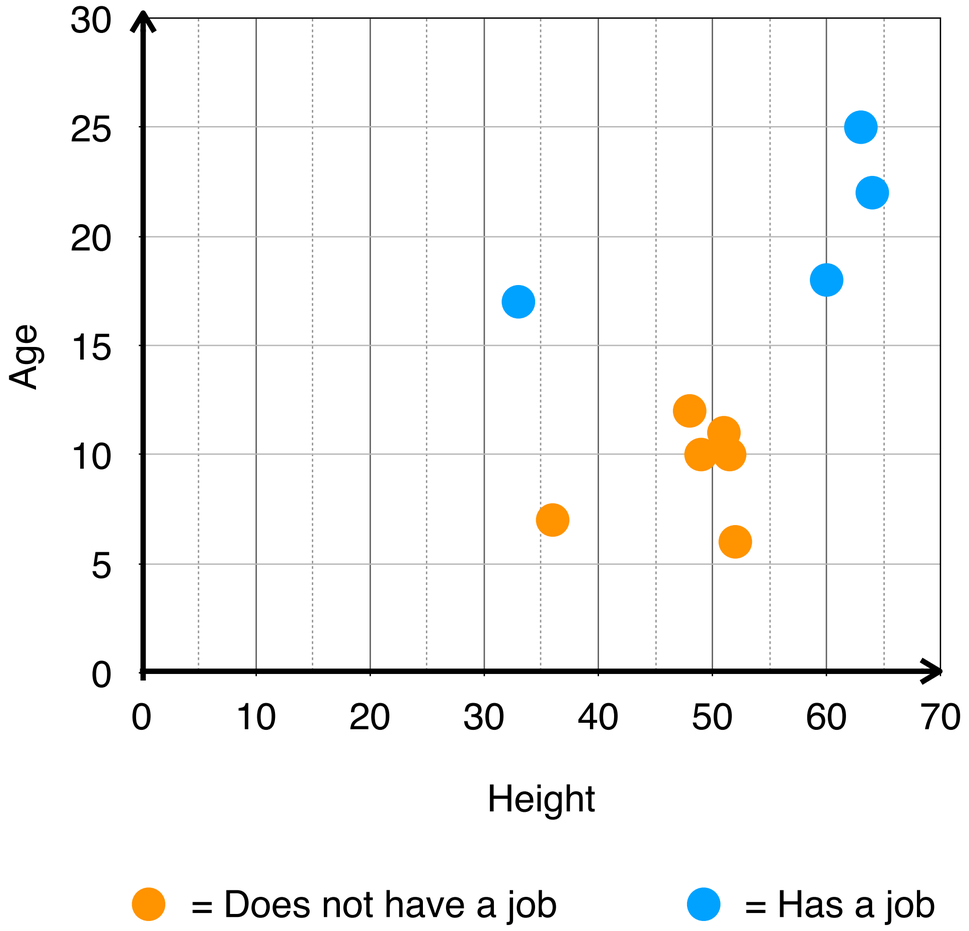
In 2D, we no longer have a threshold point, but a threshold **line**.

Height	Age	Job?
36”	7	No
60"	18	Yes
52”	6	No
64”	22	Yes
49”	10	No
51.5”	10	No
48”	12	No
63”	25	Yes
51”	11	No
33”	17	Yes

SVMs: Separating Data Points

In 2D, we no longer have a threshold point, but a threshold **line**.

Height	Age	Job?
36”	7	No
60"	18	Yes
52”	6	No
64”	22	Yes
49”	10	No
51.5”	10	No
48”	12	No
63”	25	Yes
51”	11	No
33”	17	Yes

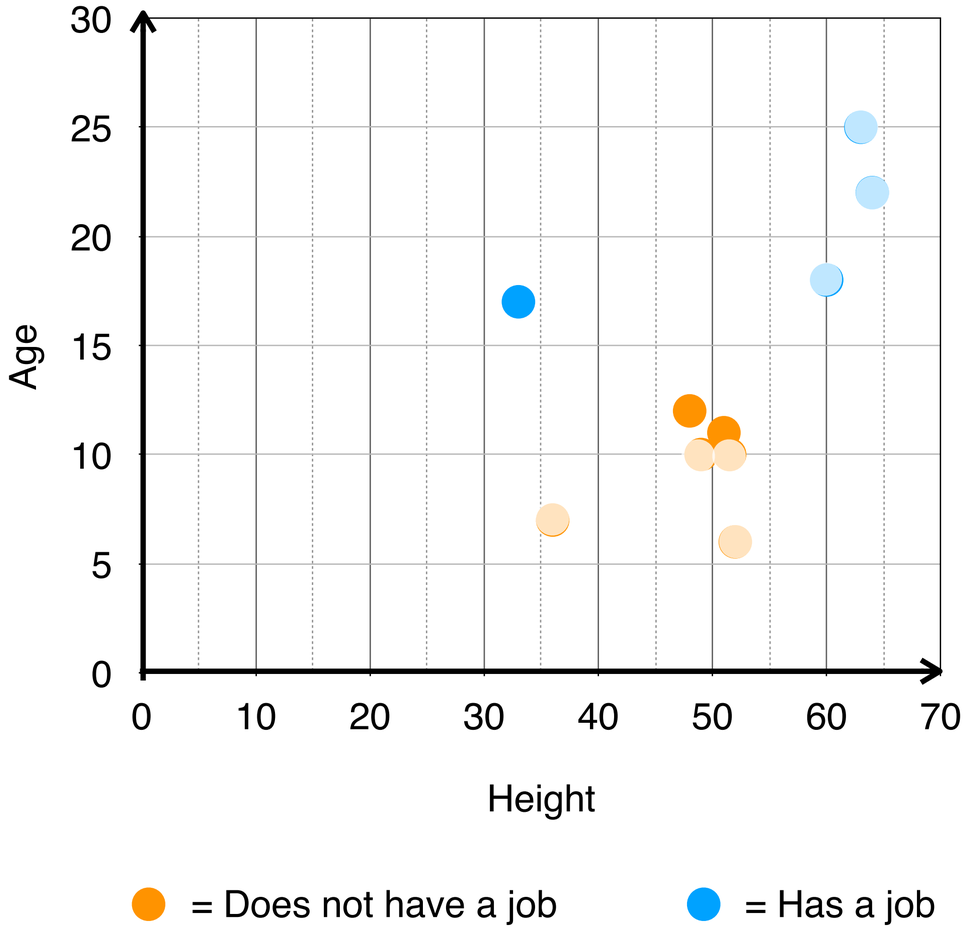


SVMs: Separating Data Points

UNIVERSITY OF CAMBRIDGE

In 2D, we no longer have a threshold point, but a threshold **line**.

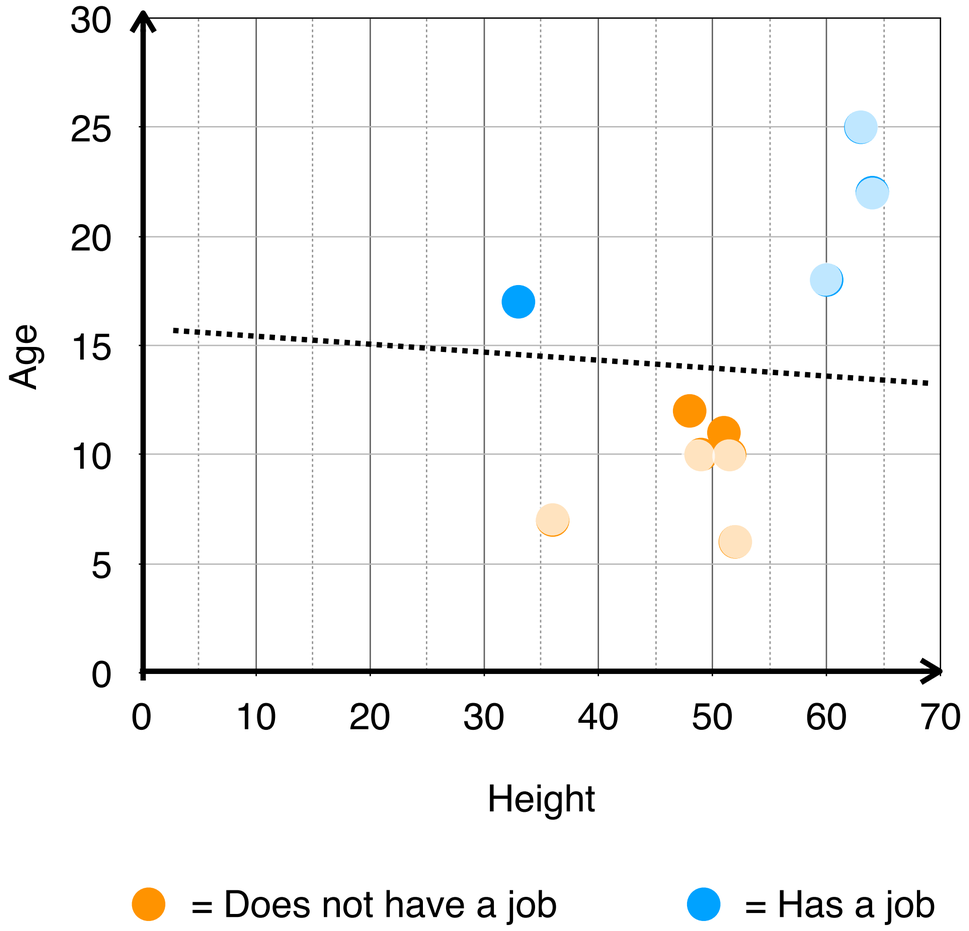
Height	Age	Job?
36”	7	No
60"	18	Yes
52”	6	No
64”	22	Yes
49”	10	No
51.5”	10	No
48”	12	No
63”	25	Yes
51”	11	No
33”	17	Yes



SVMs: Separating Data Points

In 2D, we no longer have a threshold point, but a threshold **line**.

Height	Age	Job?
36”	7	No
60"	18	Yes
52”	6	No
64”	22	Yes
49”	10	No
51.5”	10	No
48”	12	No
63”	25	Yes
51”	11	No
33”	17	Yes



Threshold

maximizes distance between line and highlighted points

SVMs: Separating Data Points

UNIVERSITY OF CAMBRIDGE

In 3D, we no longer have a threshold line, but a threshold **plane**.

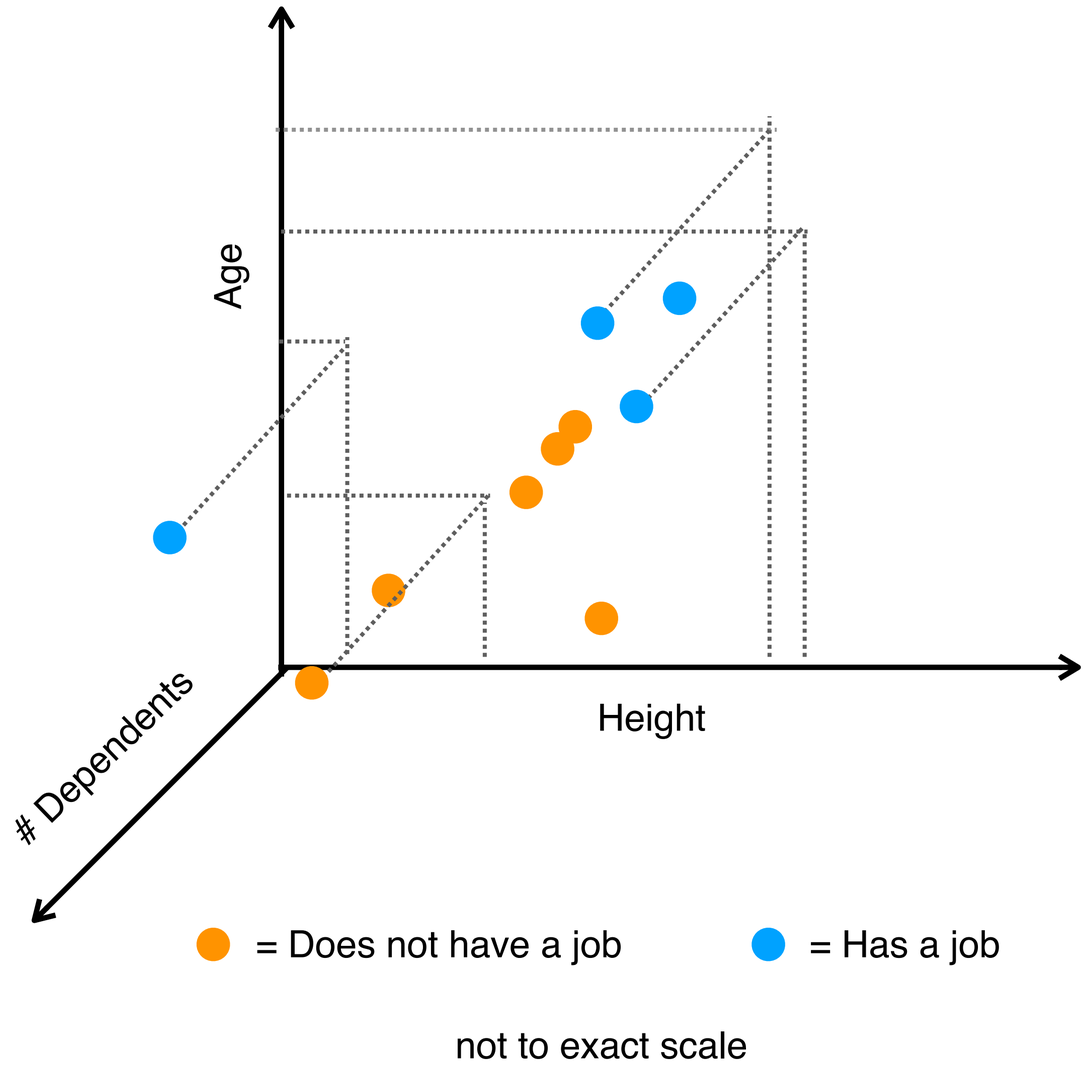
Height	Age	Number of Dependents	Job?
36”	7	0	No
60"	18	0	Yes
52”	6	0	No
64”	22	1	Yes
49”	10	0	No
51.5”	10	0	No
48”	12	1	No
63”	25	1	Yes
51”	11	0	No
33”	17	1	Yes

SVMs: Separating Data Points

UNIVERSITY OF CAMBRIDGE

In 3D, we no longer have a threshold line, but a threshold **plane**.

Height	Age	Number of Dependents	Job?
36”	7	0	No
60"	18	0	Yes
52”	6	0	No
64”	22	1	Yes
49”	10	0	No
51.5”	10	0	No
48”	12	1	No
63”	25	1	Yes
51”	11	0	No
33”	17	1	Yes

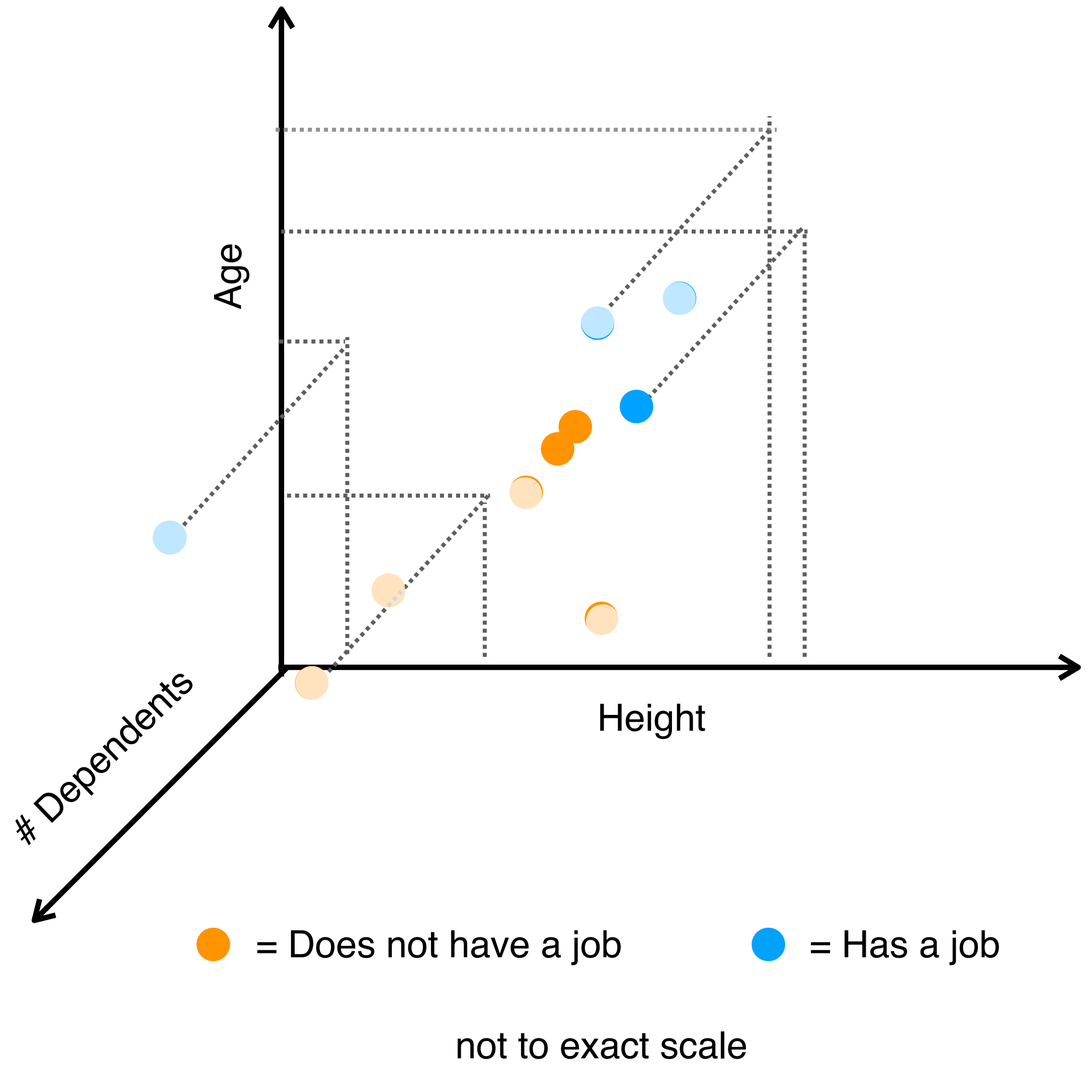


SVMs: Separating Data Points

UNIVERSITY OF CAMBRIDGE

In 3D, we no longer have a threshold line, but a threshold **plane**.

Height	Age	Number of Dependents	Job?
36”	7	0	No
60"	18	0	Yes
52”	6	0	No
64”	22	1	Yes
49”	10	0	No
51.5”	10	0	No
48”	12	1	No
63”	25	1	Yes
51”	11	0	No
33”	17	1	Yes

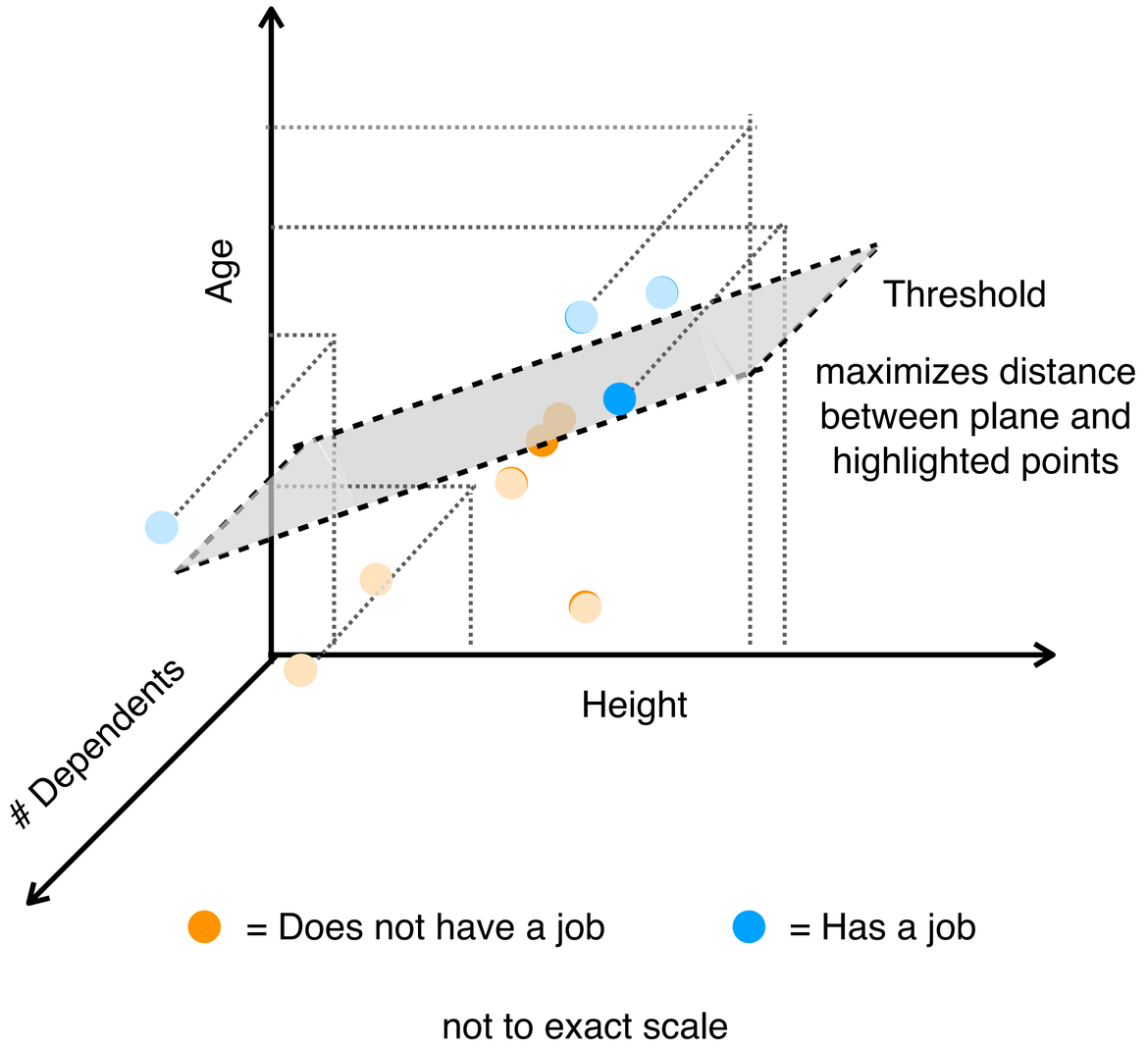


SVMs: Separating Data Points

UNIVERSITY OF CAMBRIDGE

In 3D, we no longer have a threshold line, but a threshold **plane**.

Height	Age	Number of Dependents	Job?
36”	7	0	No
60"	18	0	Yes
52”	6	0	No
64”	22	1	Yes
49”	10	0	No
51.5”	10	0	No
48”	12	1	No
63”	25	1	Yes
51”	11	0	No
33”	17	1	Yes



A Problem Case

UNIVERSITY OF CAMBRIDGE

= beneficial effect

= no beneficial effect

(no effect / toxic)

Drug Dosage

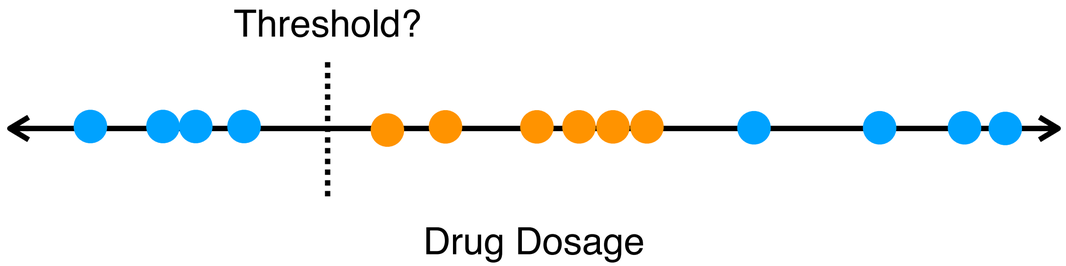
A Problem Case

UNIVERSITY OF CAMBRIDGE

= beneficial effect

= no beneficial effect

(no effect / toxic)



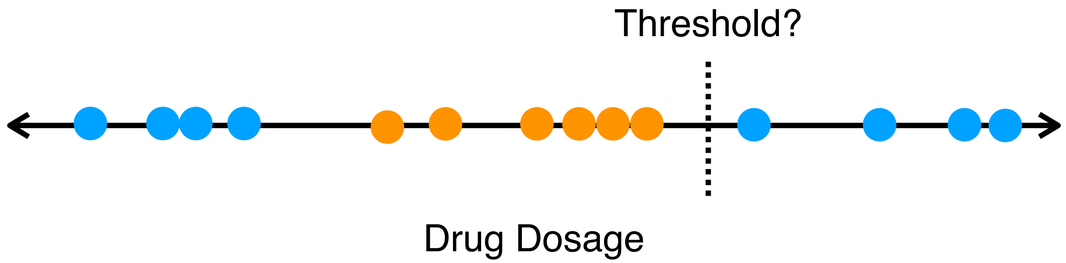
A Problem Case

UNIVERSITY OF CAMBRIDGE

= beneficial effect

= no beneficial effect

(no effect / toxic)



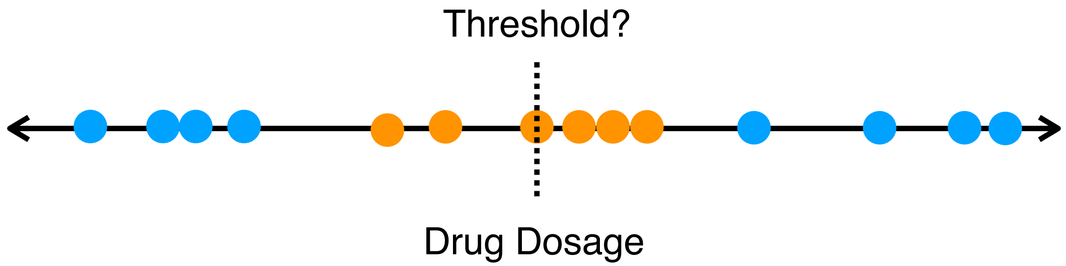
A Problem Case

UNIVERSITY OF CAMBRIDGE

= beneficial effect

= no beneficial effect

(no effect / toxic)



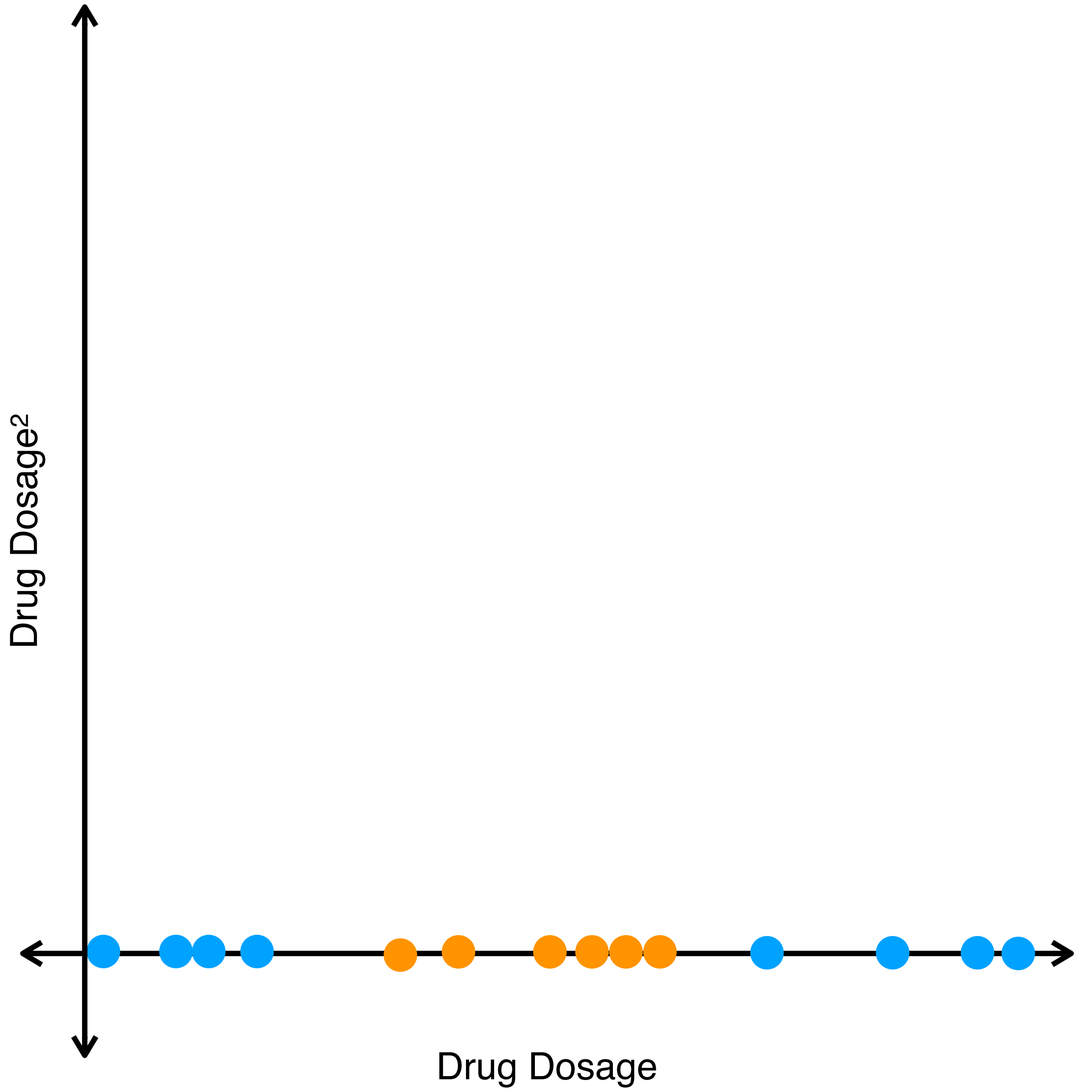
No matter where we place the threshold, we get poor classification!

A Problem Case

= beneficial effect

= no beneficial effect

(no effect / toxic)

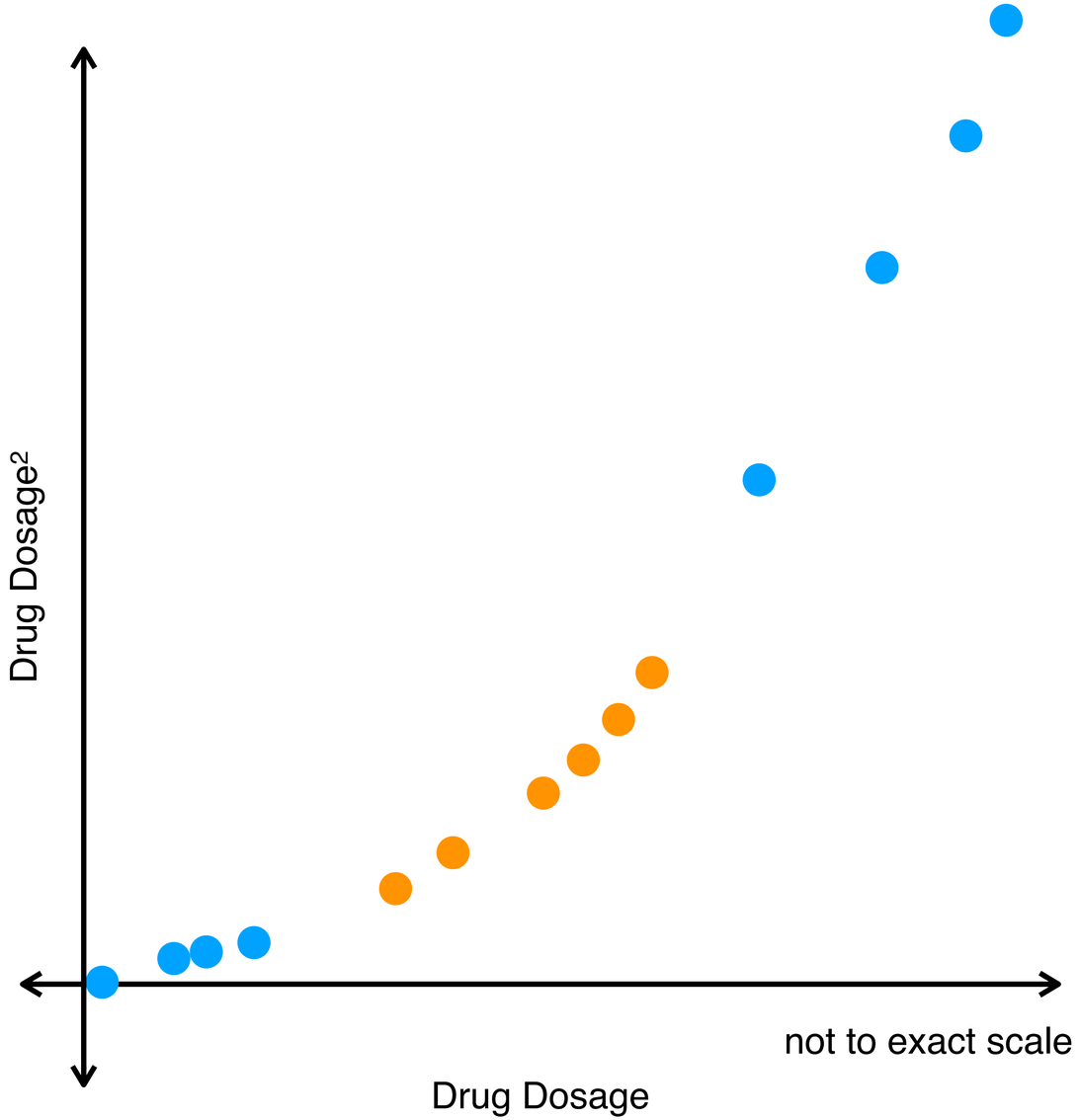


A Problem Case

= beneficial effect

= no beneficial effect

(no effect / toxic)

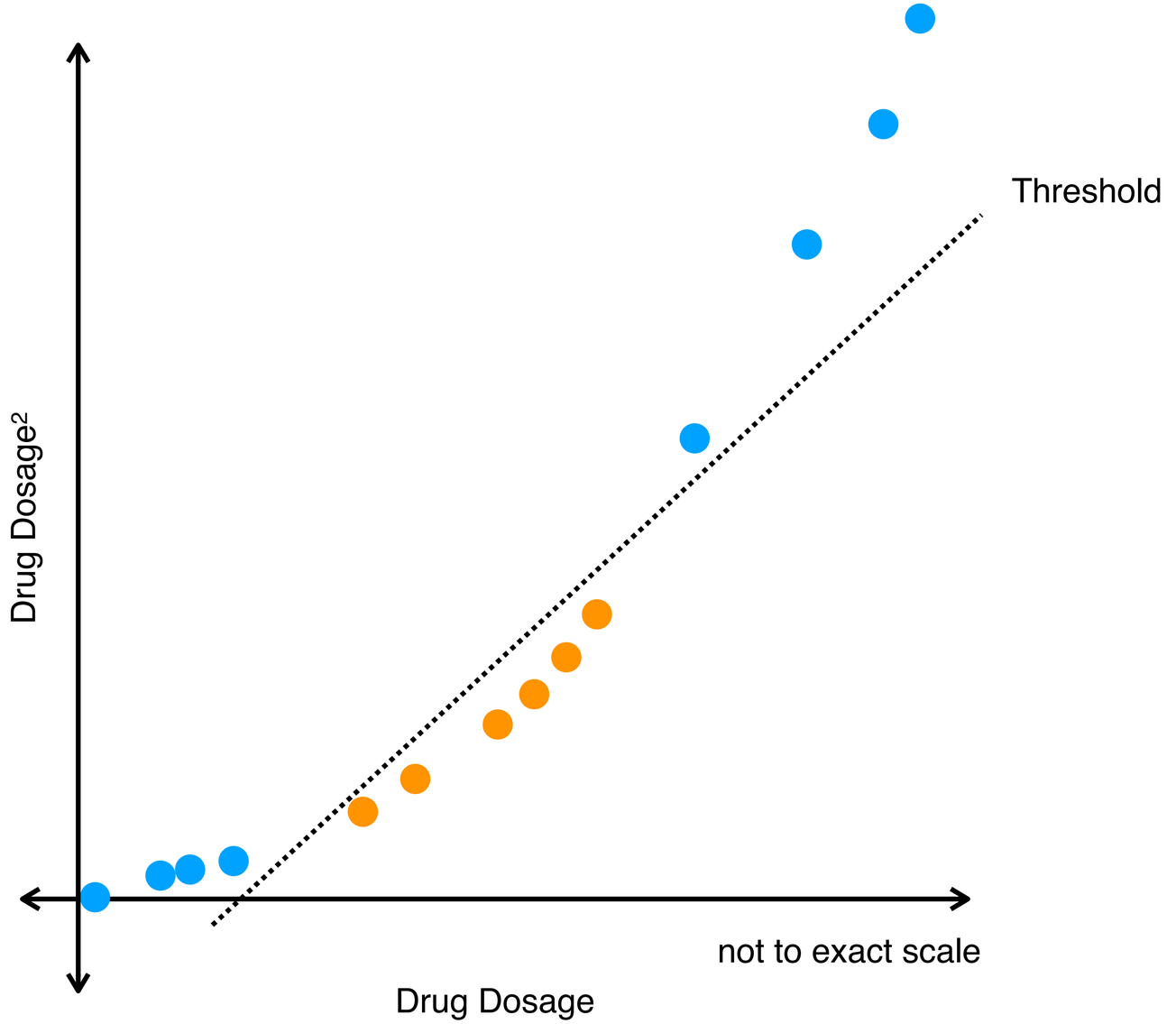


A Problem Case

= beneficial effect

= no beneficial effect

(no effect / toxic)



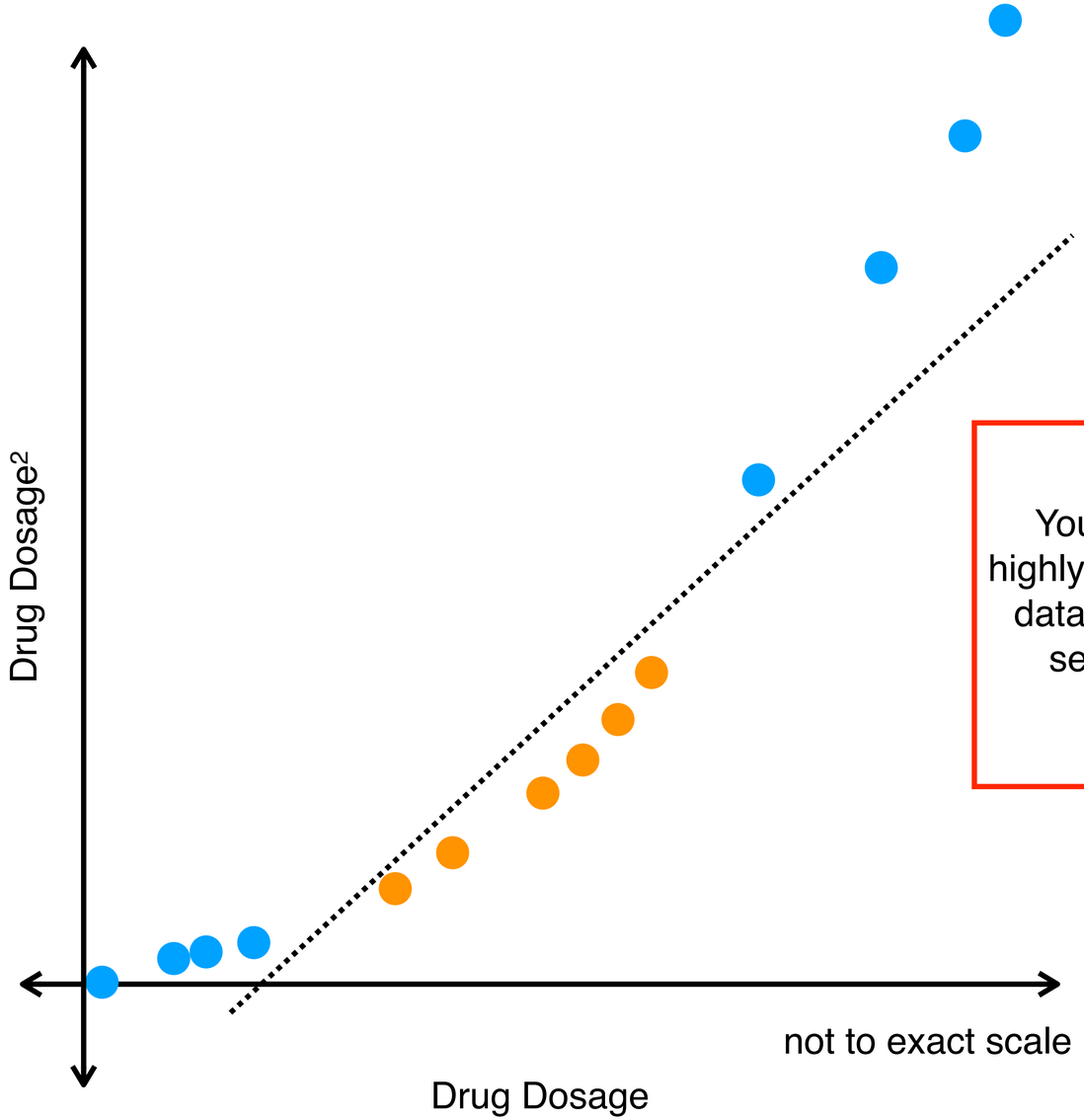
We have increased the dimensionality of the data to find a threshold!

A Problem Case

= beneficial effect

= no beneficial effect

(no effect / toxic)



Practical Tip

Your SVM's ability to solve problems is highly dependent on the function that moves data into higher dimension. Be sure to try several options before settling on the optimal SVM

We have increased the dimensionality of the data to find a threshold!

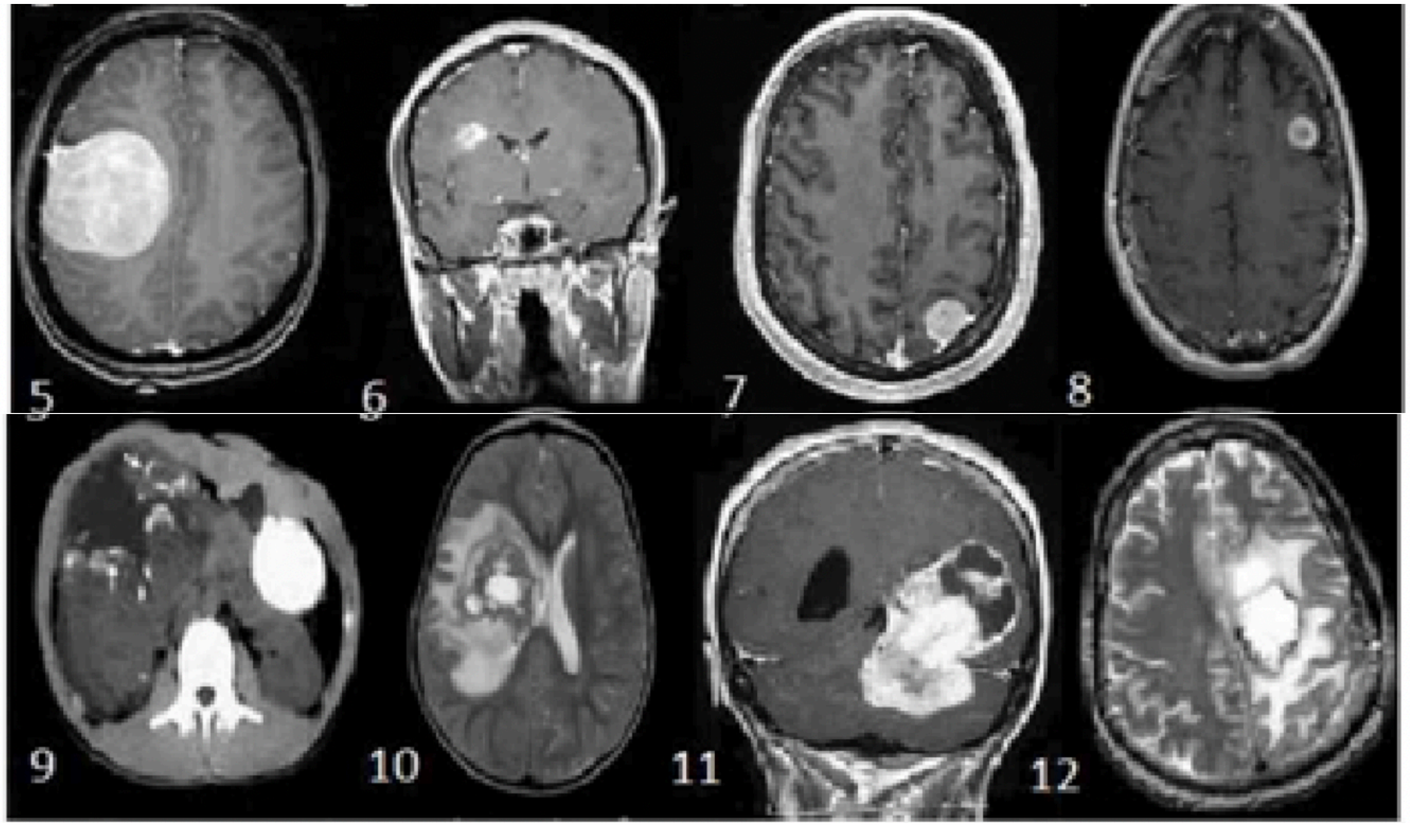
The "reverse" of PCA (reducing dimensions to simplify finding correlations).

SVM: A Real World Example

UNIVERSITY OF CAMBRIDGE

Which brain tumor is malignant and which is benign?

253 MRI images dataset



Accuracy ≈ 80%

Alam *et al. IEEE Asia-Pacific Conference on Computer Science and Data Engineering* **2019**, 1.

Malarvizhi *et al. J. Physics* **2022**, *2318*, 012042.

SVM: A Real World Example

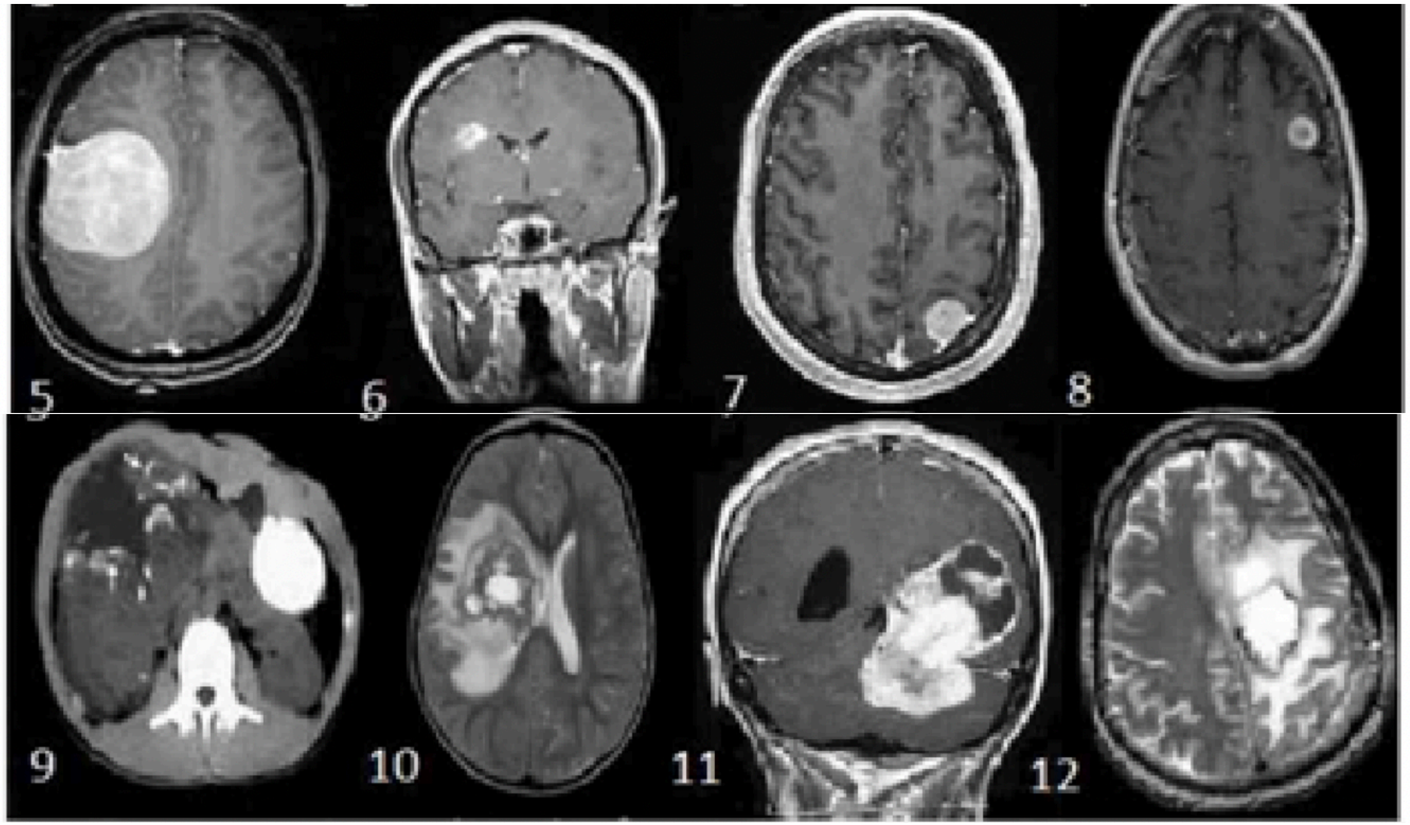
UNIVERSITY OF CAMBRIDGE

Which brain tumor is malignant and which is benign?

253 MRI images dataset

These are benign

These are not benign (malignant)



Accuracy ≈ 80%

Alam *et al. IEEE Asia-Pacific Conference on Computer Science and Data Engineering* **2019**, 1.

Malarvizhi *et al. J. Physics* **2022**, *2318*, 012042.

Playing with Dimensions

UNIVERSITY OF CAMBRIDGE

SVM		PCA
	Data Visualization	
	Classification	
	Regression	
Increases dimensions for predictions	What do dimensions have to do with it?	Reduces dimensions for human visualization