Machine Learning in Practice: A Primer on Missing Data

James Allingham

Wolfram Research

15 April 2019

Missing Data?

π	0	1	?	1	NaN	0.1
e	0	0	?	?	∞	5
π	0	?	?	3	4	5
?	?	?	?	4	?	0.1
?	?	?	?	?	?	?
π	0	0	?	0.23	?	5
e	0	?	?	1.1	?	?

Missing Data?

Missing Data?

π	0	1	?	1	NaN	0.1
e	0	0	?	1.87	∞	5
π	0	0.33	?	3	4	5
2.97	0.	0.33	?	4	NaN	0.1
2.97	0.	0.33	?	1.87	NaN	3.04
π	0	0	?	0.23	NaN	5
e	0	0.33	?	1.1	NaN	3.04

Common Cases

- Faulty sensors

Common Cases

- Faulty sensors
- Unanswered survey questions

Common Cases

- Faulty sensors
- Unanswered survey questions
- Data corruption

Common Cases

- Faulty sensors
- Unanswered survey questions
- Data corruption
- Other examples?

Why do we care?

`np.mean(data)` $\longrightarrow$ `NaN`

Why do we care?

`np.mean(data)` $\longrightarrow$ `NaN`

Tip 1

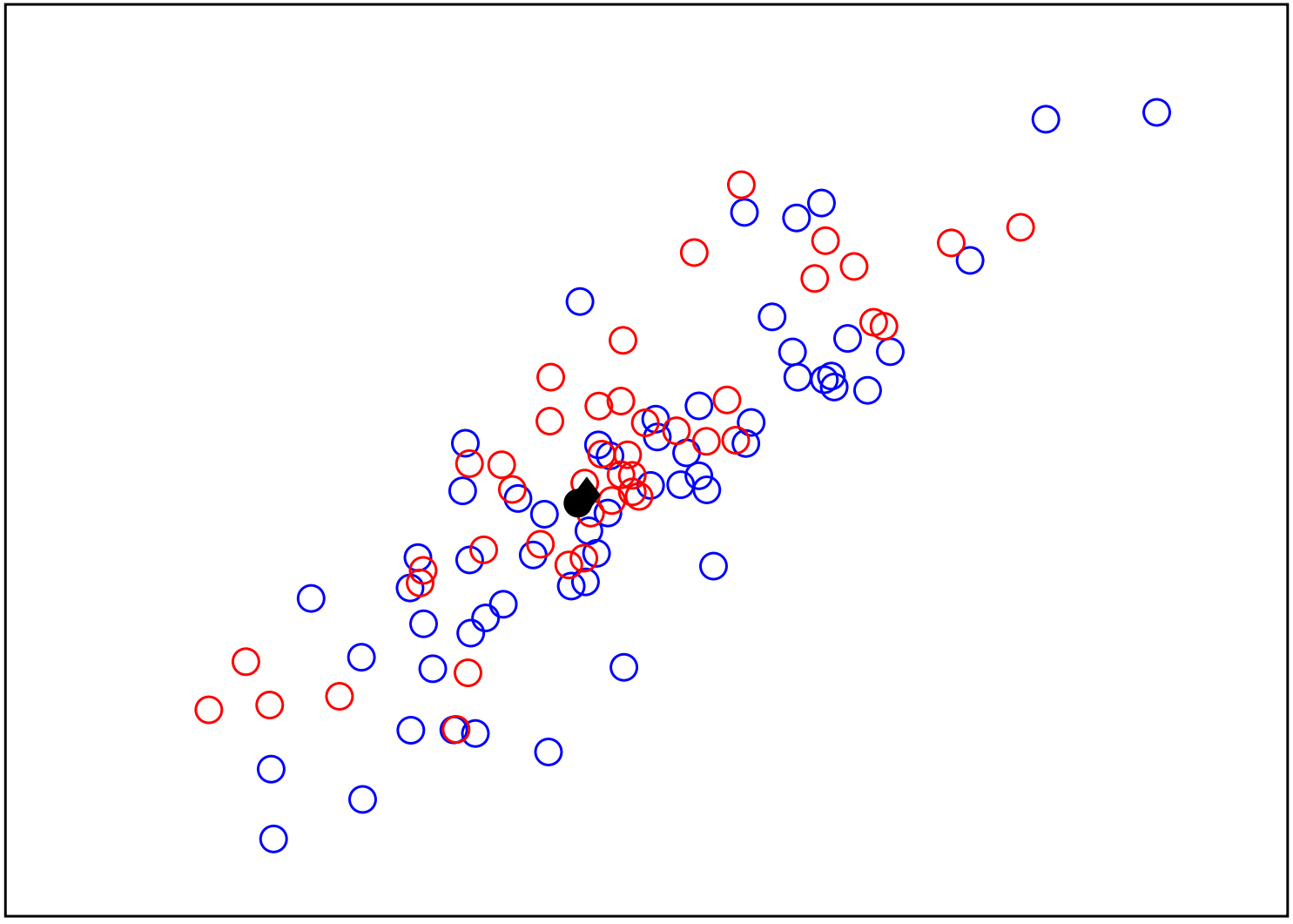
`numpy.ma` is very useful for representing missing data:

```
x = np.array([1, 2, 3, -1, 5])
mx = ma.masked_array(x, mask=[0, 0, 0, 1, 0])
mx.mean()
```

✓

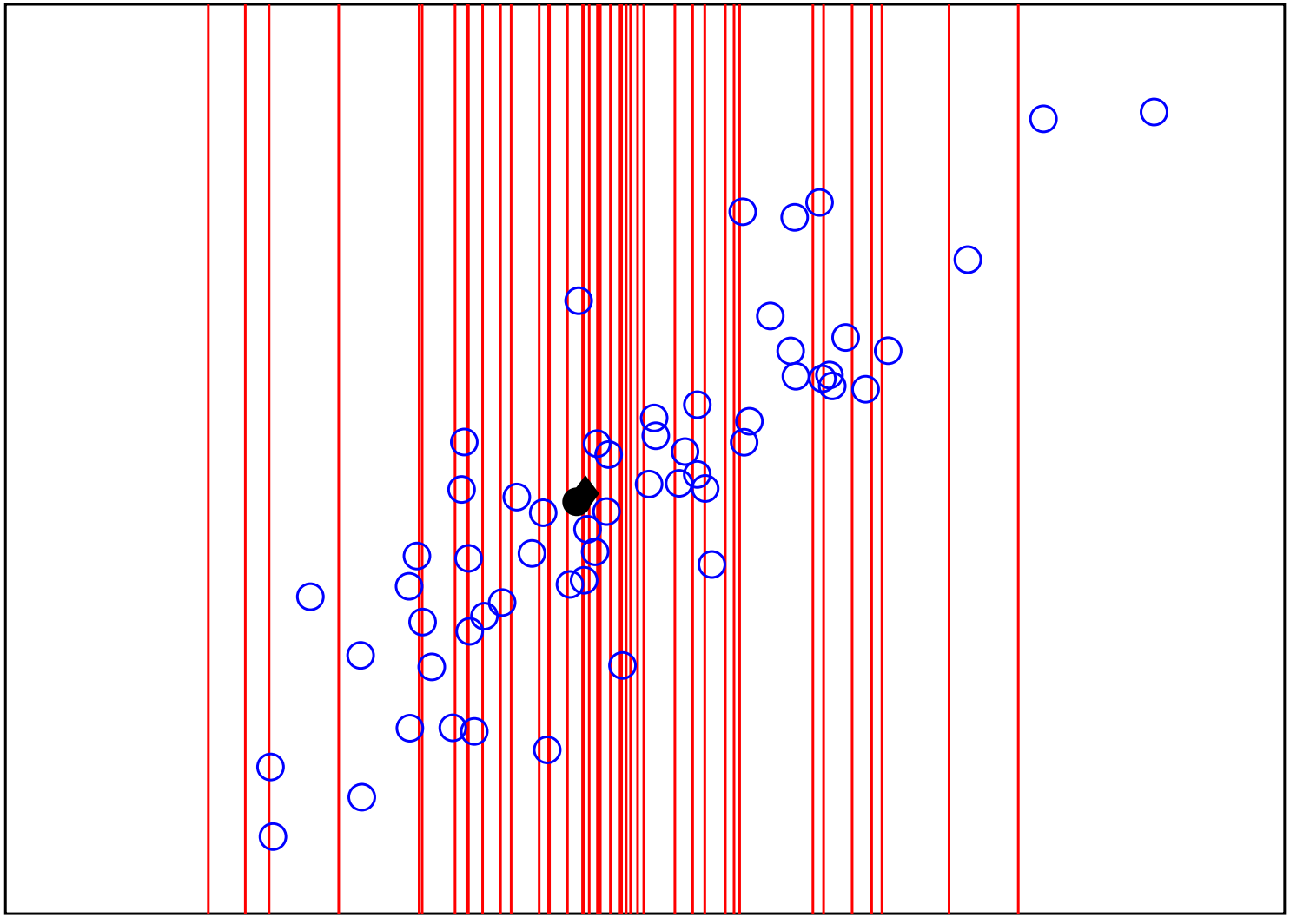
Types of Missing Data - MCAR

Missing Completely At Random



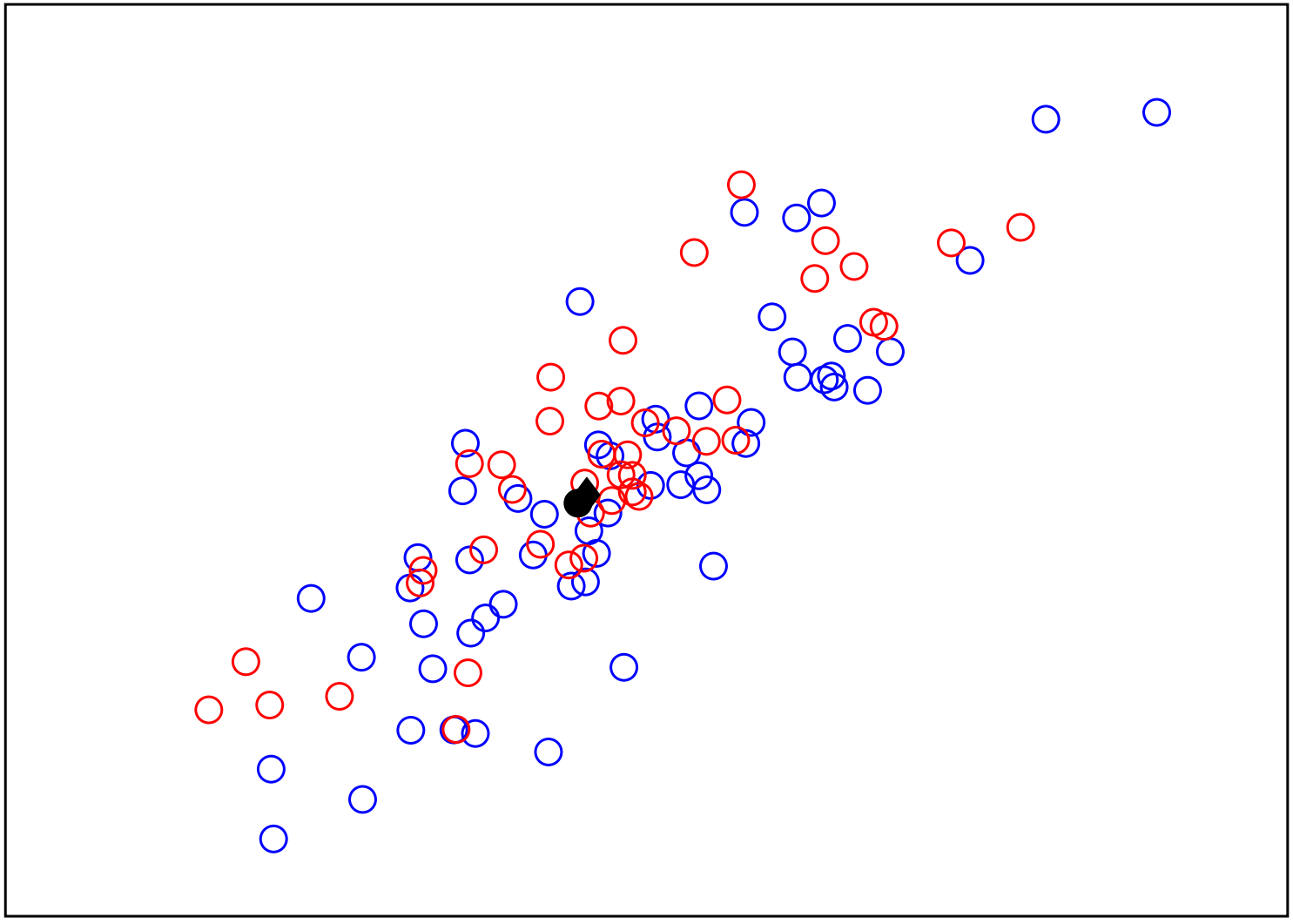
Types of Missing Data - MCAR

Missing Completely At Random



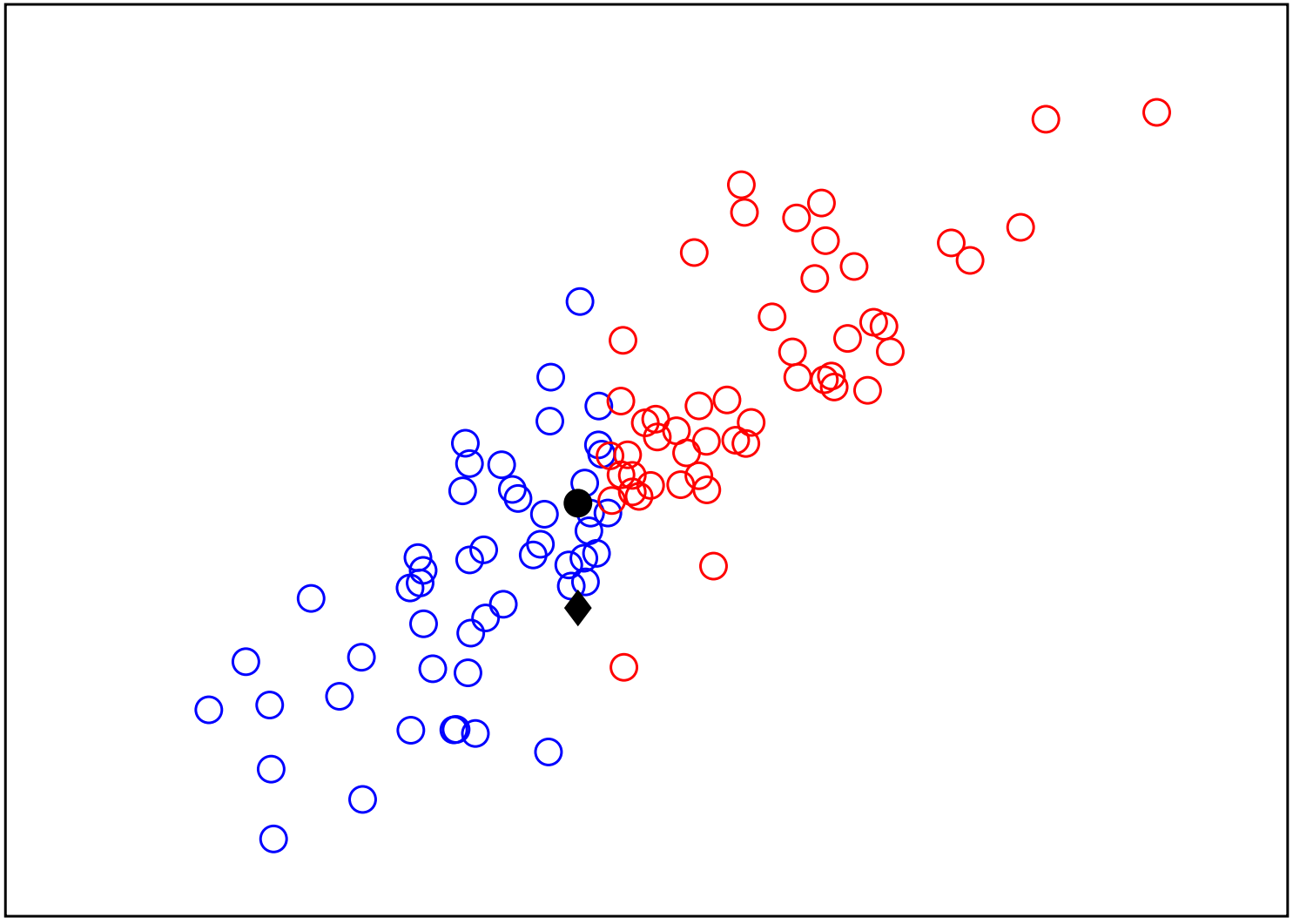
Types of Missing Data - MCAR

Missing Completely At Random



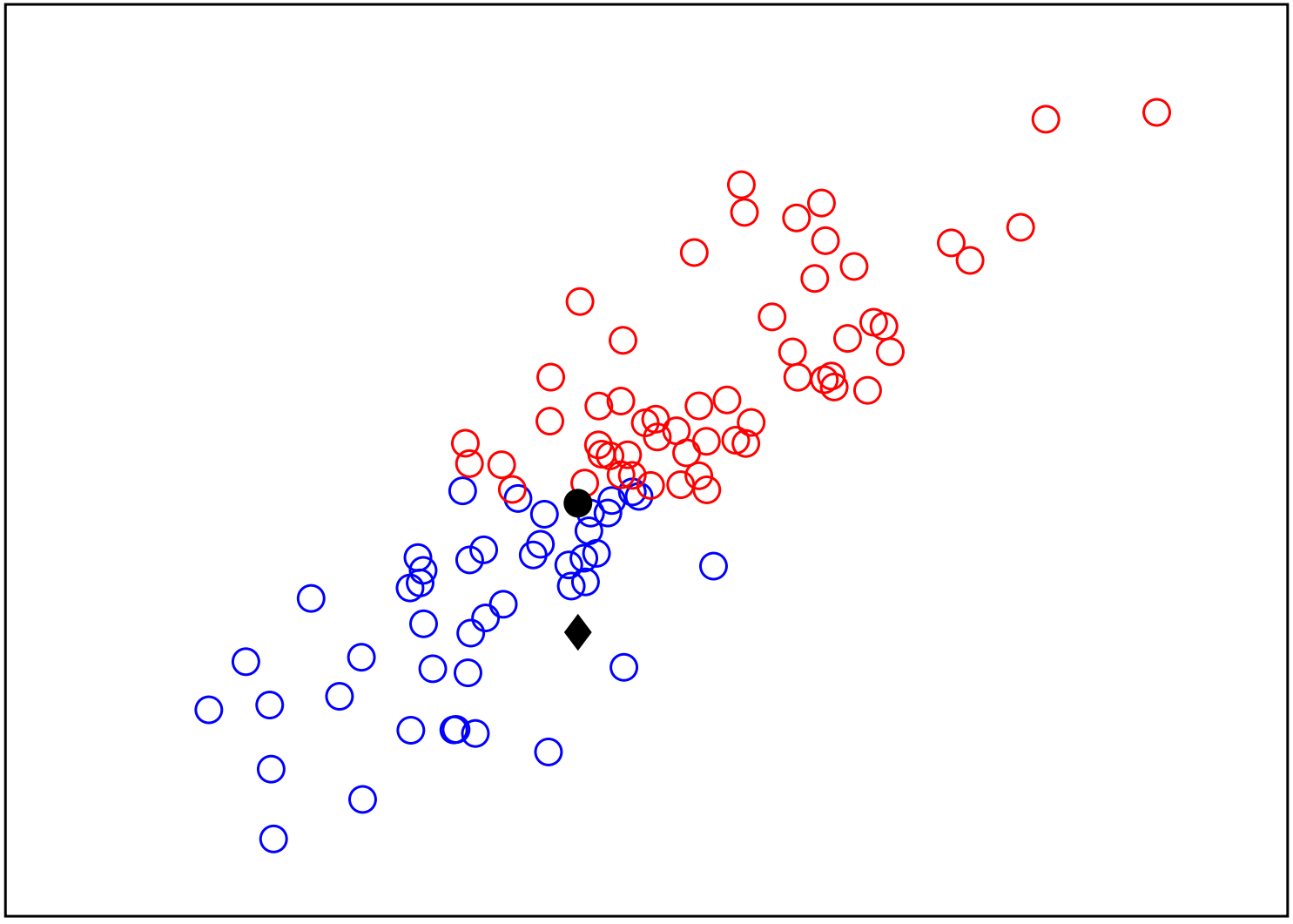
Types of Missing Data - MAR

Missing At Random



Types of Missing Data - NMAR

Not Missing At Random



A Few More Tips

Tip 2

If you have missing data, **domain knowledge is your friend**.

A Few More Tips

Tip 3

If you don't know what kind of missing data you have, assume that it is MAR.

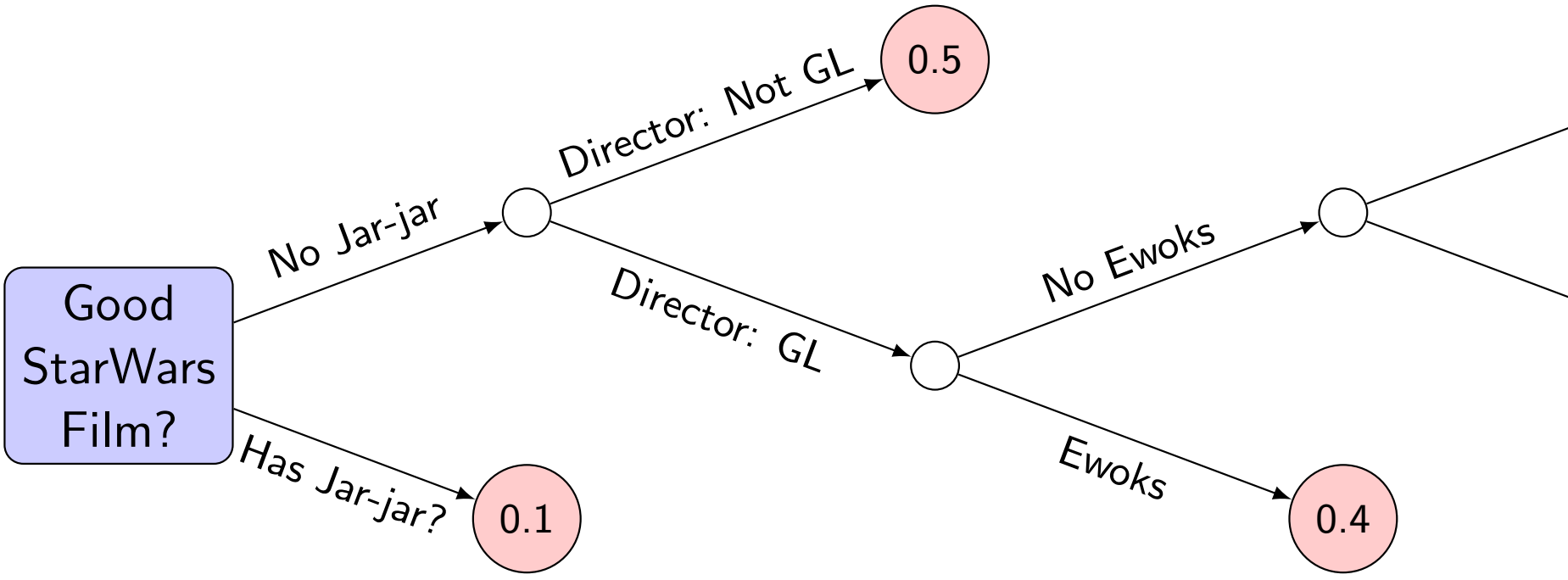
A Few More Tips

Tip 4

If the presence of missing data is important, don't throw that information away!

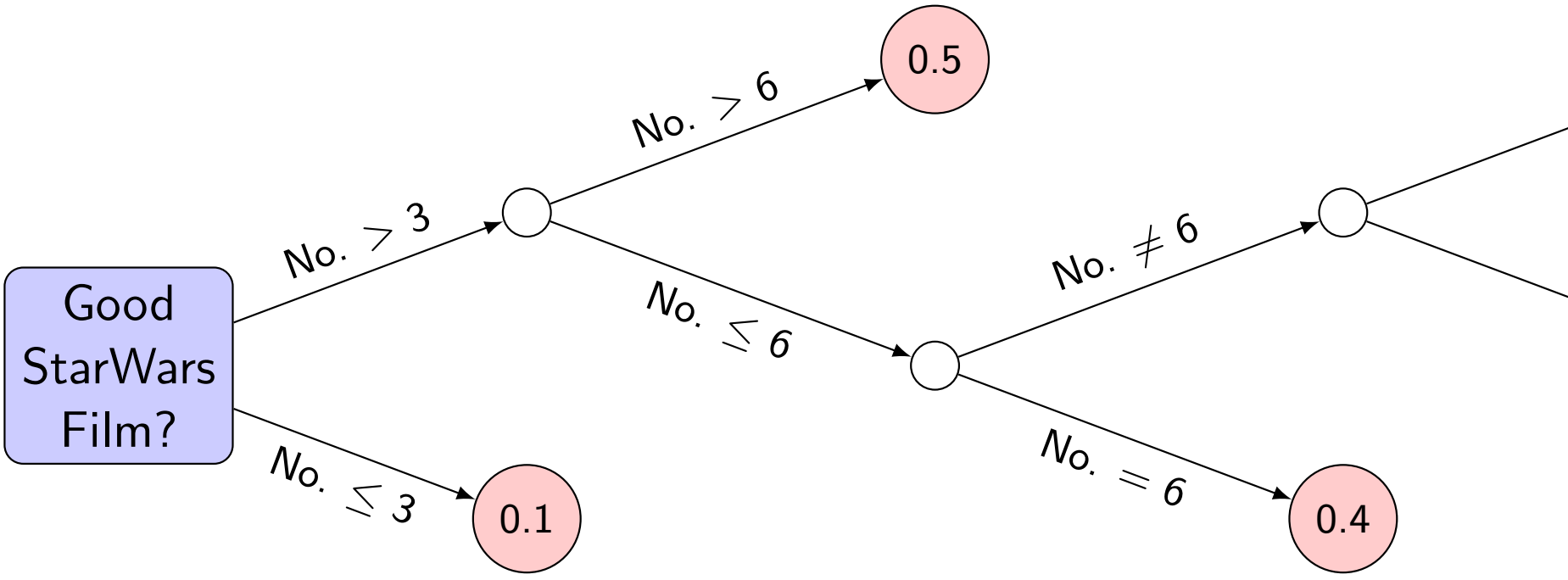
MissForest

R Package (Stekhoven and Bühlmann, 2012)



MissForest

R Package (Stekhoven and Bühlmann, 2012)



Data Wig

Python Library

Product Type	Description	Picture	Size	Colour
SD Card	Best SD Card ever ...		128MB	Black
Half Life 3	Coming Soon™		30GB	NA
Dress	Fantastic summer dress ...		M	?
Shoe	... in size 11 ...		?	Blue

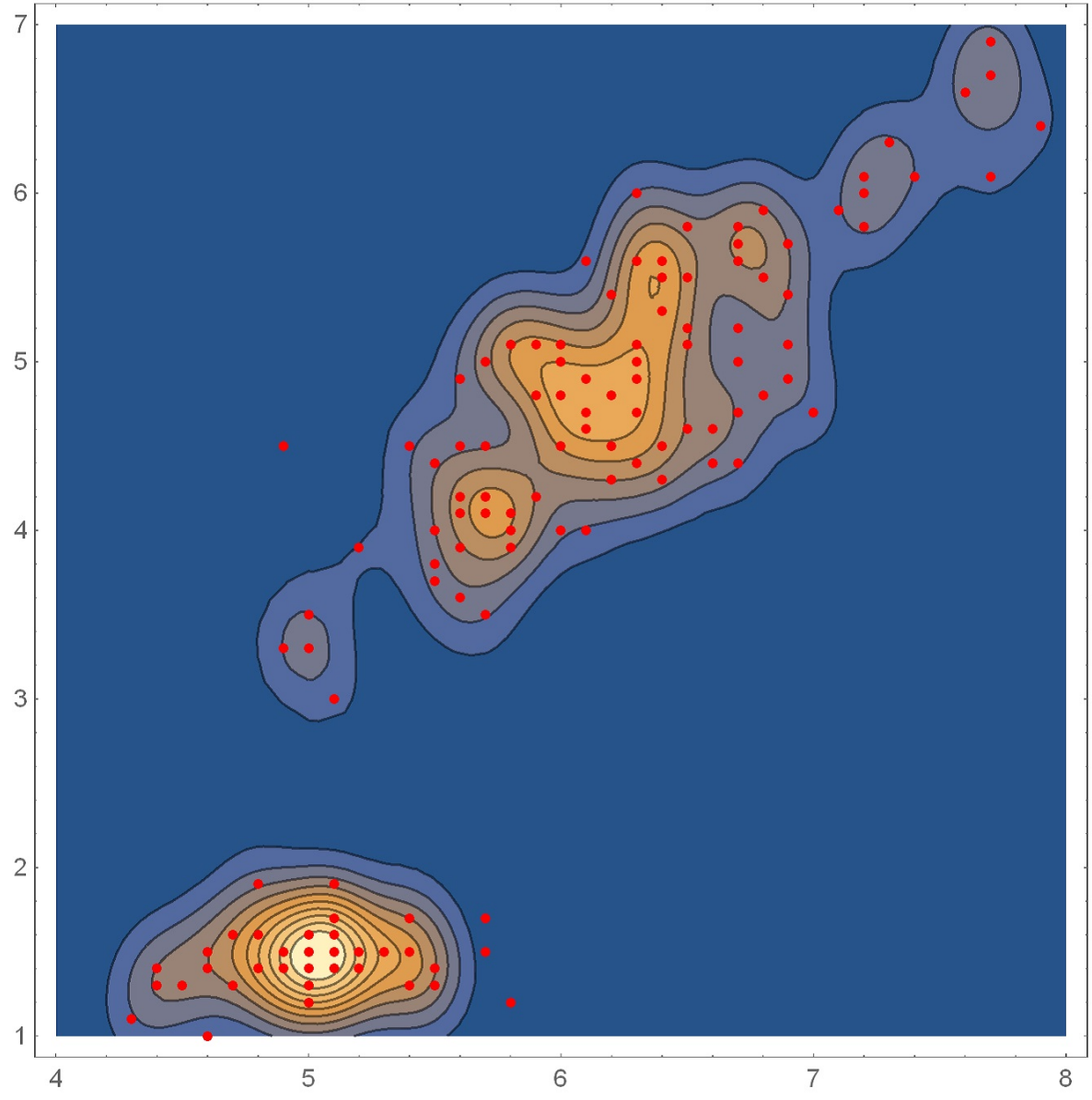
Mathematica

LearnDistribution[] + SynthesizeMissingValues[]

```
iris = ExampleData[{"MachineLearning", "FisherIris"},
"Data"][[All,1, {1, 3}]];
ld = LearnDistribution[iris];
```

Mathematica

`LearnDistribution[] + SynthesizeMissingValues[]`



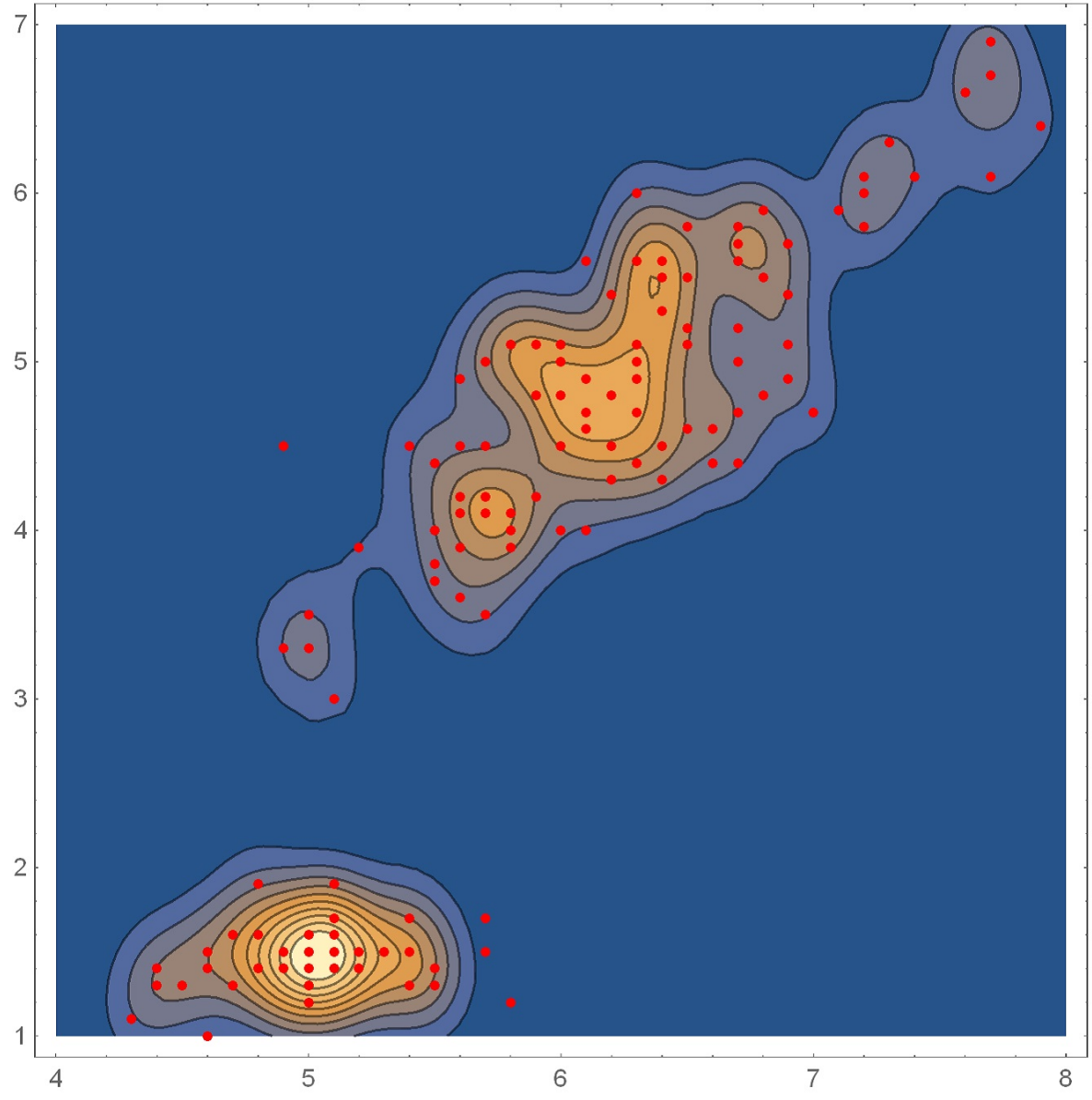
Mathematica

LearnDistribution[] + SynthesizeMissingValues[]

```
iris = ExampleData[{"MachineLearning", "FisherIris"},
"Data"][[All,1, {1, 3}]];
ld = LearnDistribution[iris];
SynthesizeMissingValues[ld, {5.5, Missing[]}] ⟶
{5.5, 1.83949}
```

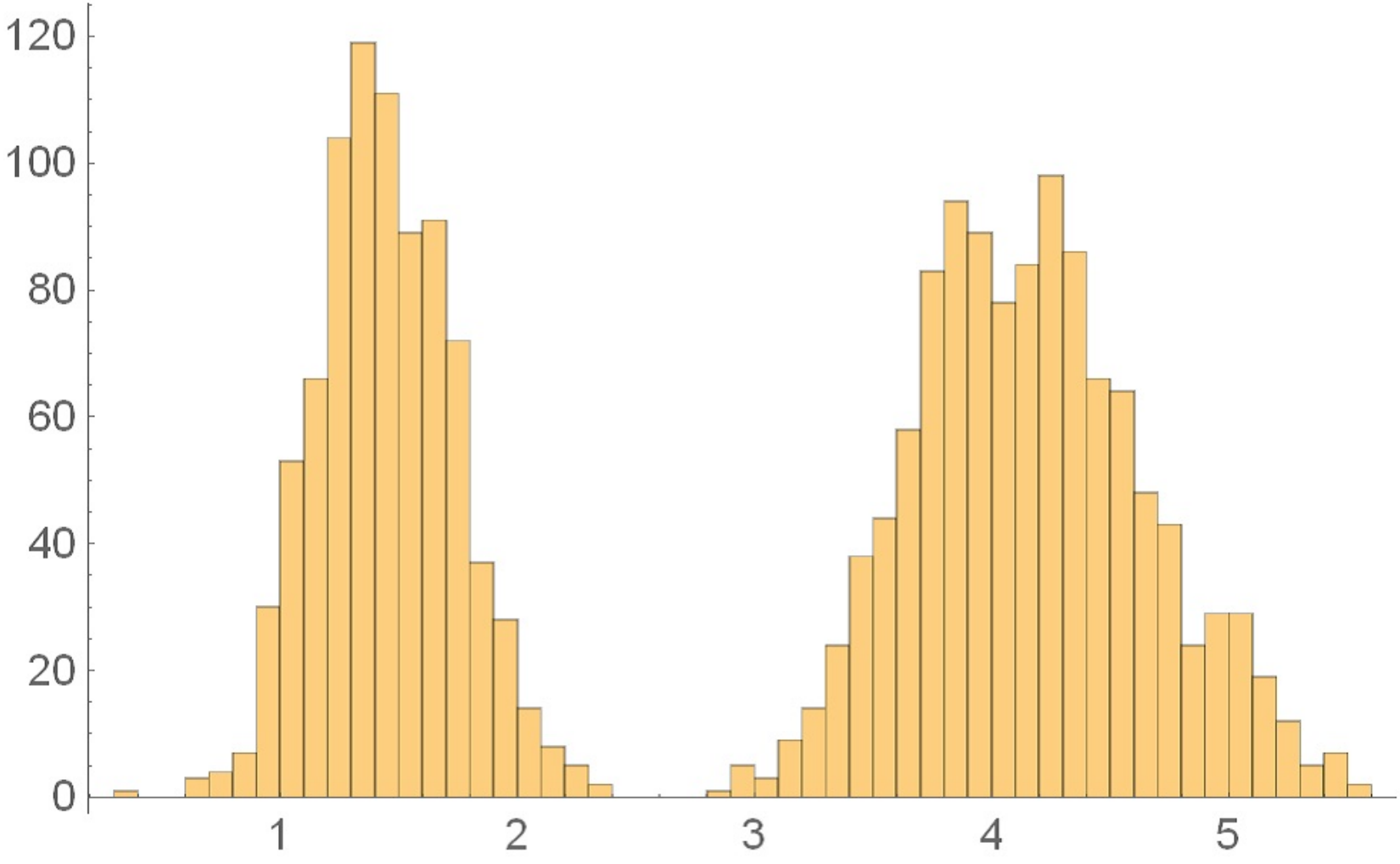
Mathematica

`LearnDistribution[] + SynthesizeMissingValues[]`



Mathematica

`LearnDistribution[] + SynthesizeMissingValues[]`

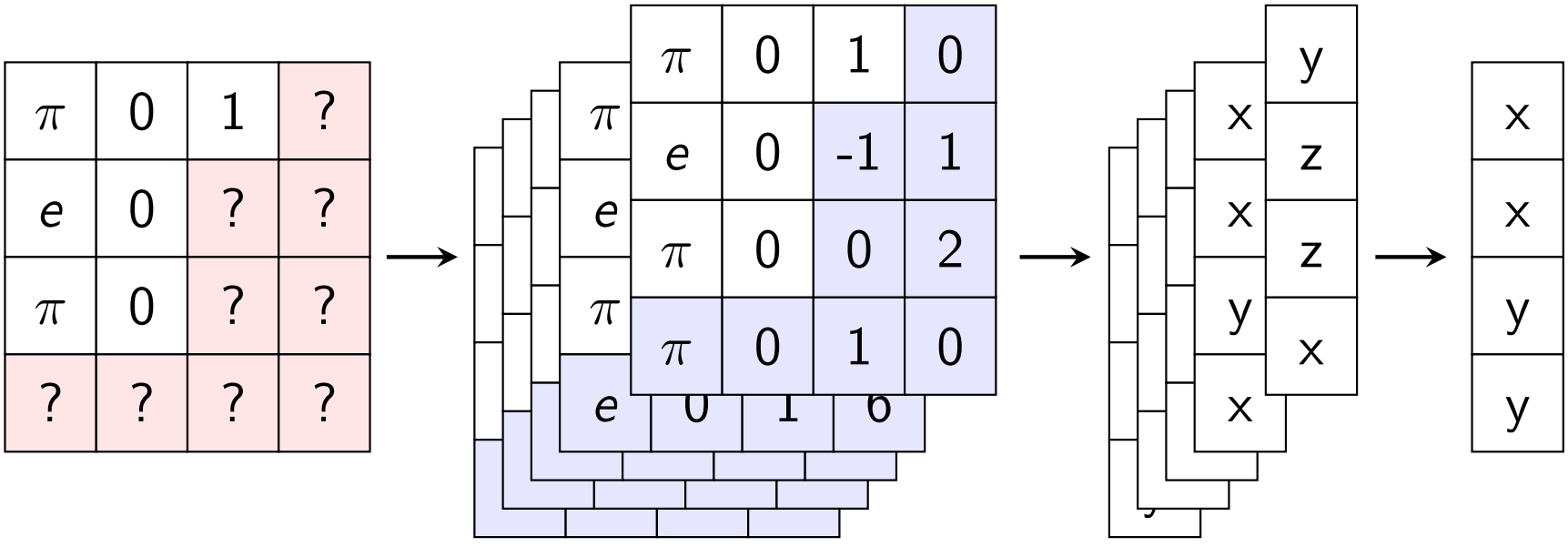


Multiple Imputation

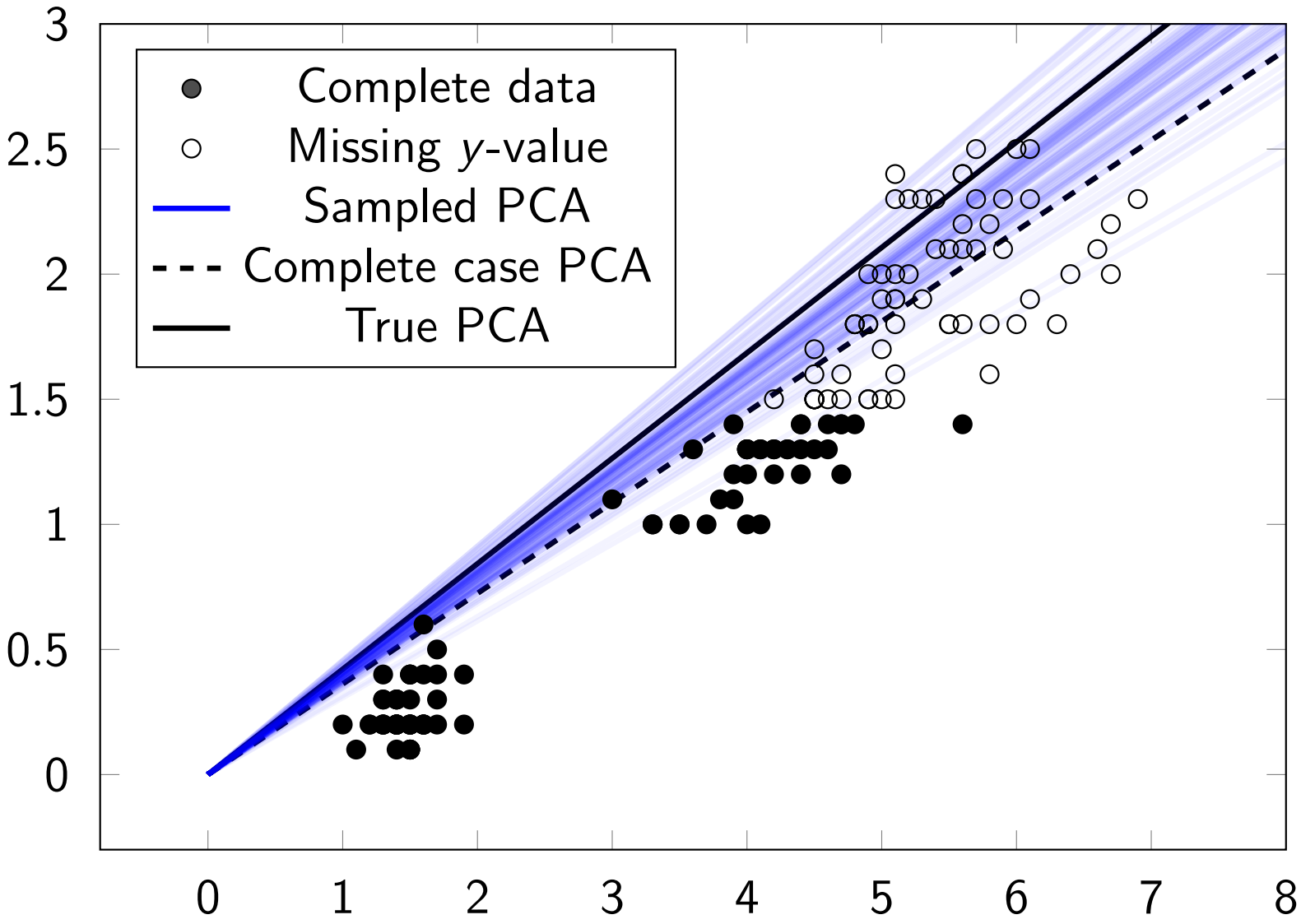
Tip 5

If you need some idea of the uncertainty in your imputation do **Multiple Imputation**.

Multiple Imputation



Multiple Imputation



Automatic Imputation

My research

π	0	1	?	1	NaN	0.1
e	0	0	?	?	∞	5
π	0	?	?	3	4	5
?	?	?	?	4	?	0.1
?	?	?	?	?	?	?
π	0	0	?	0.23	?	5
e	0	?	?	1.1	?	?

Work supervised by Prof. Zoubin Ghahramani and Dr. Christian Steinruecken.

What I didn't get to talk about

- Imputation for time-series. (I'd use GPs.)

What I didn't get to talk about

- Imputation for time-series. (I'd use GPs.)
- Specifics of density imputation using EM (Ghahramani and Jordan, 1996)

What I didn't get to talk about

- Imputation for time-series. (I'd use GPs.)
- Specifics of density imputation using EM (Ghahramani and Jordan, 1996)
- Specifics of regression imputation, especially MissForest (Stekhoven and Bühlmann, 2012)

What I didn't get to talk about

- Imputation for time-series. (I'd use GPs.)
- Specifics of density imputation using EM (Ghahramani and Jordan, 1996)
- Specifics of regression imputation, especially MissForest (Stekhoven and Bühlmann, 2012)
- Specifics of automatic imputation

What I didn't get to talk about

- Imputation for time-series. (I'd use GPs.)
- Specifics of density imputation using EM (Ghahramani and Jordan, 1996)
- Specifics of regression imputation, especially MissForest (Stekhoven and Bühlmann, 2012)
- Specifics of automatic imputation
- Cutting edge techniques using deep learning – GAIN. (Yoon et al., 2018)

What I didn't get to talk about

- Imputation for time-series. (I'd use GPs.)
- Specifics of density imputation using EM (Ghahramani and Jordan, 1996)
- Specifics of regression imputation, especially MissForest (Stekhoven and Bühlmann, 2012)
- Specifics of automatic imputation
- Cutting edge techniques using deep learning – GAIN. (Yoon et al., 2018)
- Probabilistic Deep Learning (Gal and Ghahramani, 2016)

Further Reading

Gal, Yarin and Ghahramani, Zoubin (2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. URL https://arxiv.org/pdf/1506.02142.pdf.

Ghahramani, Zoubin and Jordan, Michael I. (1996). Supervised learning from incomplete data via an EM approach. URL http://papers.nips.cc/paper/767-supervised-learning-from-incomplete-data-via-an-em-approach.pdf.

Little, Roderick J. A. and Rubin, Donald B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., 2nd edition. ISBN 978-0-471-18386-0.

Stekhoven, Daniel J. and Bühlmann, Peter (2012). MissForest – non-parametric missing value imputation for mixed-type data. *Bioinformatics*, **28**(1) 112–118. URL http://dx.doi.org/10.1093/bioinformatics/btr597.

Yoon, Jinsung, Jordon, James and Van Der Schaar, Mihaela (2018). GAIN: Missing Data Imputation using Generative Adversarial Nets. Technical report. URL http://medianetlab.ee.ucla.edu/papers/ICML{_}GAIN.pdf.

My dissertation:
https://jamesallingham.co.za/pdf/james_allingham_dissertation.pdf
My poster:
https://jamesallingham.co.za/pdf/missing_data_imputation_poster.pdf

Types of Missing Data