



University
of Exeter

Data-driven Decisions

Week 09-BEM2031

Term2: 2023/24

Today:

- What is data driven decision making? What are the challenges?
- Understand how to estimate the expected value of different decisions
- Describe the different steps necessary to estimate the value of a decision
- Seminars: profit curves and recommendations

Assignment:

- I want to see that you understand what you have done, and what it is telling you about the usefulness of the model for making predictions, about the effect of varying the most useful (or one of the most useful) features etc.
- You need to make interpretations. Please do not just present me with some code and its outputs. You are not being assessed on your coding skills – hence I am happy to give you the code. You are being assessed on your interpretation.
- I would rather that you DID NOT use Python for this. If you do, please make sure you address all of the questions asked in the Rmd file.
- If you choose to leave out the PDP, that is fine, but say something about it in the context of interpreting the model.
- **Extra marks for:** great insight; improved/additional plots (and insight); more model exploration/more modelling (and insight). **No extra marks for:** a tonne more models/plots/code without explanation, justification, learning.

The focus of
your report
should
primarily be the
critique



Final project:

- Refer to the CRISP-DM document (WEEK 2)
- Python is fine, but would prefer NOT python for the assignment
- Business understanding – we don't know much
- Data understanding – how useful do you think the features are for providing understanding and solve the business problem?
- What motivated your choice of visual or model?
- Say what you did and how. What challenges did you face and how do you solve them?
- Say what you learned and what value it brings to the analysis. How do the results alter your perspective on the original business problem?

The focus of
your report
should
primarily be the
critique



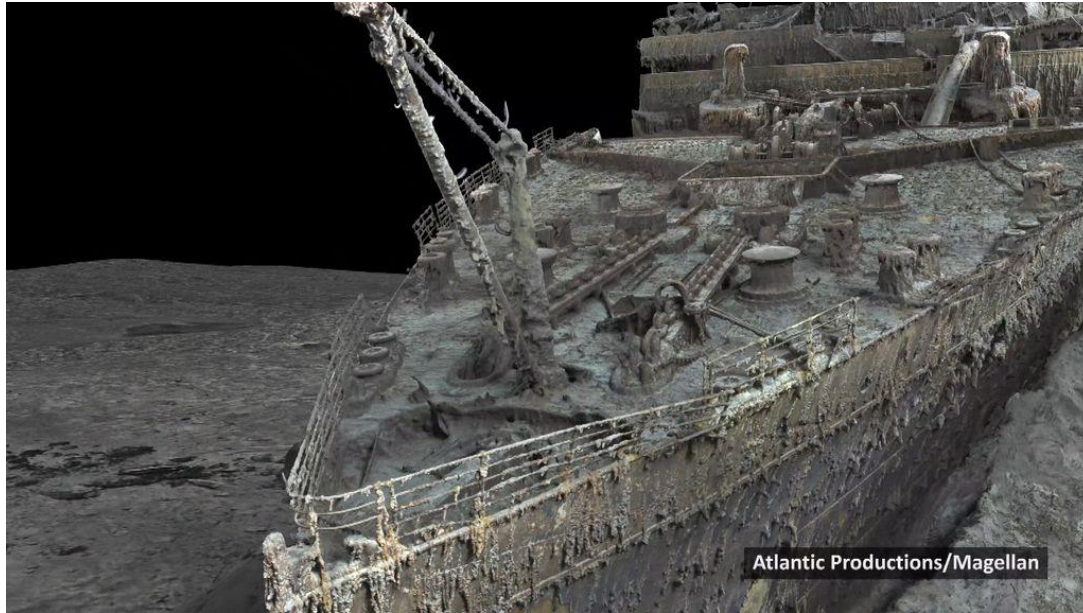
Final project:

- You don't need to try to critique the whole report – focus in on the areas that you will be improving on.
- Your critique should naturally lead to your new analysis. You will be making recommendations throughout the report, but your choice of model/visualisation will be the most useful way of demonstrating your critique.
- You can then critique your own visualisation/model and what you have learned from it. Do one of each really well, rather than a lot of stuff passably.
- Spend most of your wordcount where you have something tangible to say – I care about how you did the work, rather than did it improve the results.
- Visualisations can be provided alongside your model *in addition* to your own visualisation but don't go overboard. Remember you will need to provide a critique of both of your new elements, so it is better if your visualisation helps us to understand the problem a bit more (rather than being part of the model understanding).

The focus of
your report
should
primarily be the
critique



Who was the luckiest person on the Titanic?

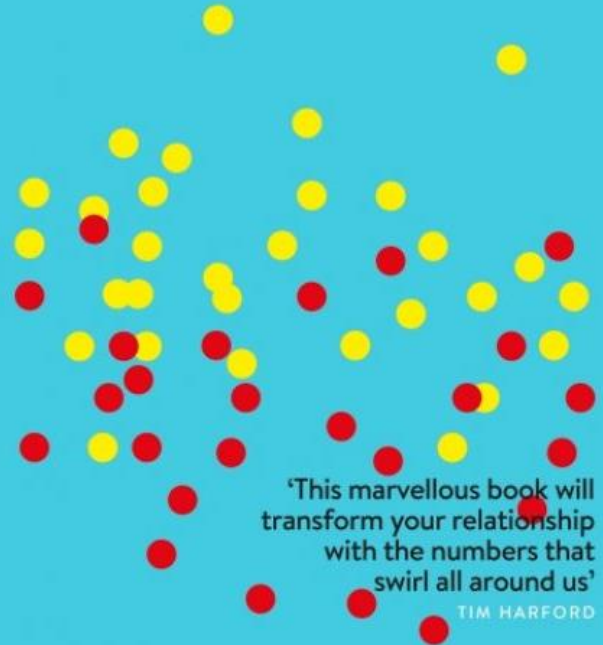


Atlantic Productions/Magellan



A PELICAN
BOOK

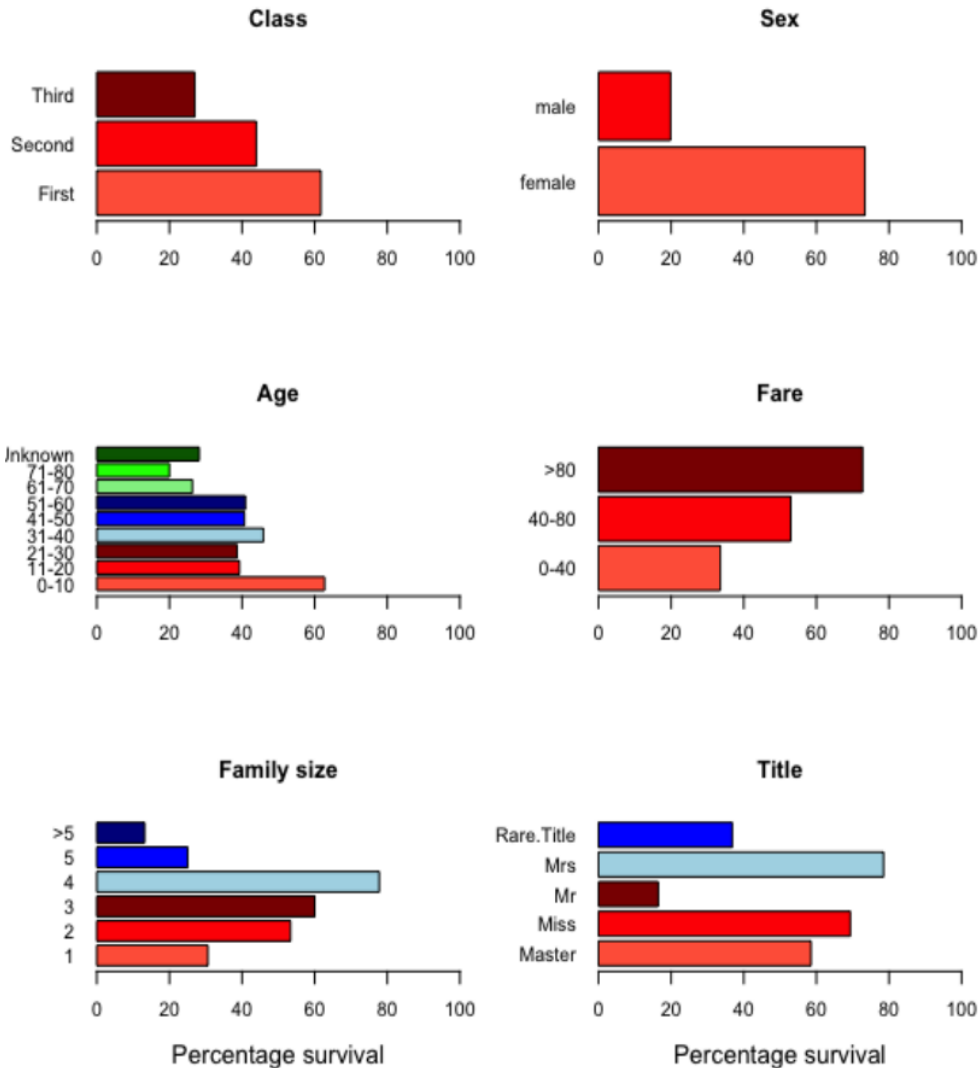
The Art of Statistics Learning from Data David Spiegelhalter



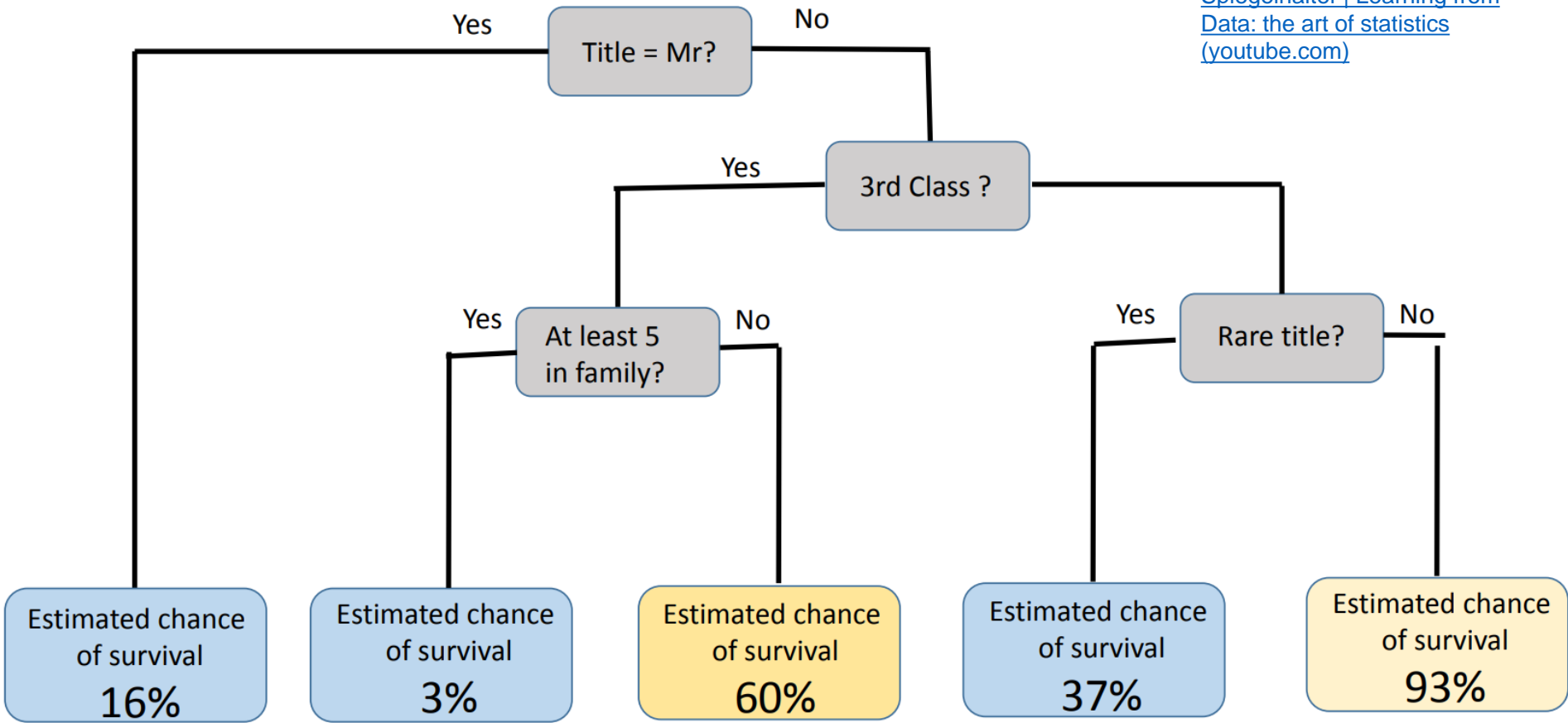
1309 passengers (39% survive)

pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body
3	0	Somerton, Mr. Francis William	male	30	0	0	A.5. 18509	8.0500		S		
3	0	Spector, Mr. Woolf	male		0	0	A.5. 3236	8.0500		S		
3	0	Spinner, Mr. Henry John	male	32	0	0	STON/OQ. 369943	8.0500		S		
3	0	Staneff, Mr. Ivan	male		0	0	349208	7.8958		S		
3	0	Stankovic, Mr. Ivan	male	33	0	0	349239	8.6625		C		
3	1	Stanley, Miss. Amy Zillah Elsie	female	23	0	0	CA. 2314	7.5500		S	C	
3	0	Stanley, Mr. Edward Roland	male	21	0	0	A/4 45380	8.0500		S		

Unsurprising factors predict survival



A simple classification tree



- Prediction is not saying what is going to happen – it is giving the probability of what is going to happen.
- You can score those with a Brier Score (mean squared error):
If probability p is given to event $X(0,1)$, then the Brier score is $(X-p)^2$

Method	Accuracy (high is good)	Brier score (low is good)
Everyone has a 39% chance of surviving	0.639	0.232
All females survive, all males do not	0.786	0.214
Simple classification tree	0.806	0.139
Classification tree (over-fitted)	0.806	0.150
Logistic regression	0.789	0.146
Random forest	0.799	0.148
Support Vector Machine (SVM)	0.782	0.153
Neural network	0.794	0.146
Averaged neural network	0.794	0.142
K-nearest-neighbour	0.774	0.180

So who was the luckiest person on the titanic?

Of all survivors, who got the lowest predicted chance of survival, across all the algorithms?



Karl Dahl, a 45-year old Norwegian/Australian joiner travelling on his own in 3rd class, had the highest average Brier score among the survivors – a very surprising survivor.

He dived into the freezing water. He tried to get onto a lifeboat, despite some trying to push him off.

Interpretable Machine Learning

If a machine learning model performs well, **why do we not just trust the model** and ignore **why** it made a certain decision? “The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks.” (Doshi-Velez and Kim 2017)

The process of integrating machines and algorithms into our daily lives requires interpretability to increase **social acceptance**. People attribute beliefs, desires, intentions and so on to objects.

It's not just about building an accurate model, but also about being able to explain how the model arrived at its decisions in a way that is comprehensible to stakeholders.

Moneyball

The situation

0:05:10

How it normally works

0:08:25

Finding Peter

0:13:30

How it works now

0:31:30

Adapt or die

0:46:30

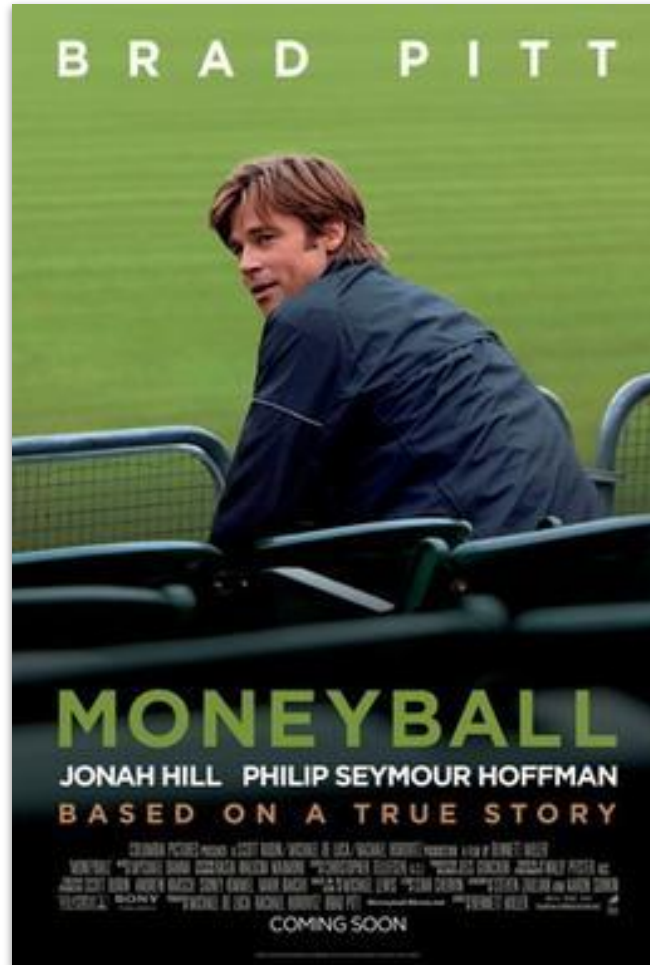
Implementing

1:21:00

Meeting

1:55:00

<https://www.youtube.com/watch?v=odfK3GiBb6k>



Moneyball

Data-Driven Decisions: The movie focuses on the use of sports analytics, to evaluate and recruit players. The manager teams up with an economics graduate who uses statistical data analysis to identify undervalued players.

Impact on Baseball: The analytics approach challenges the traditional methods of scouting and evaluating players, which were based more on the scouts' experience and intuition rather than empirical data.

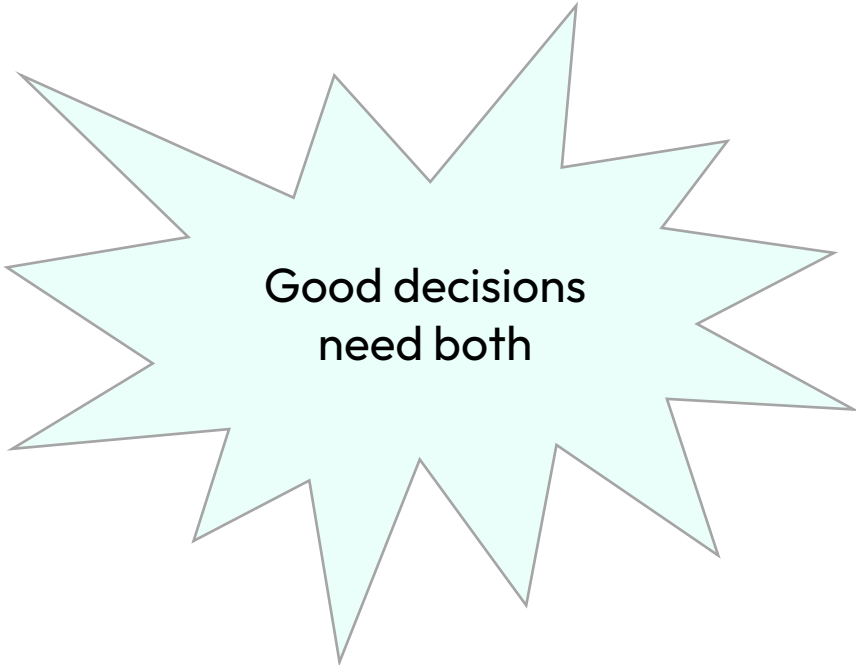
Underlying Themes: The story is not just about baseball but also about how innovative thinking and data analytics can disrupt traditional practices and lead to success in any field.

Data-Driven Decisions:

The process of making organisational decision based on actual data rather than intuition, experience, or observation alone.

“In several studies I’ve conducted over the past eight years looking at high-stakes decisions, such as surgeons making life-or-death emergency room decisions, or early-stage investors deciding how to allocate millions of dollars in startup capital, I found that the role of gut feel is often to inspire a leader to make a call, particularly when the decision is risky.”

—Laura Huang, associate professor of business administration at Harvard Business School



Good decisions
need both

Solving Business Problems

Data science in business should be solving business problems.

They are often treated as classification or regression problems, but we use these tools to solve **business problems**.

Charity Mailing

In the assignment you created a model to predict who will respond to a request for a donation.

If people donate £1 on average, and each attempt costs £1 on average, we have accomplished nothing.

We create predictive models because we want to *use them*.



University
of Exeter

Expected Benefit

$$EB = \sum (P_i \times U_i)$$

Where:

- P_i represents the probability of each possible outcome i
- U_i represents the utility of each outcome i

‘Utility’ refers to a measure of the desirability or value of a particular outcome (0-1).

For example, consider a treatment with two possible outcomes: recovery or no improvement. If the treatment has a 70% chance of recovery (with a utility value of 1) and a 30% chance of no improvement (with a utility value of 0), the expected benefit of the treatment would be:

$$EB = (0.7 \times 1) + (0.3 \times 0) = 0.7$$

Probability of responding (R)
given some attributes of the
person (\mathbf{x})

This is what your model
predicts based on the data
you gathered.

Value you get from a
response. How much do they
donate?

This comes from another
model, a regression.

$$\text{Expected benefit of targeting} = p(R \mid \mathbf{x}) \cdot v_R(\mathbf{x}) + [1 - p(R \mid \mathbf{x})] \cdot v_{NR}(\mathbf{x})$$

Probability of not
responding.

Value of not responding.

Charity Mailing

If people donate £18 on average, and each attempt costs £1 on average, we could gain £17.

		Observed	
		Ignore	Donate
Predicted	Ignore	TN £0	FN £0
	Donate	FP -£1	TP £17



Charity Mailing

Cutoff = 50%



	Reference	
Prediction	0	1
0	182063	9716
1	0	0



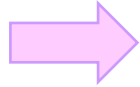
£0 – so we decide
its not worth it!



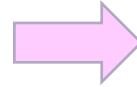
University
of Exeter

Charity Mailing

Cutoff = 5%



	Reference	
Prediction	0	1
0	109162	4010
1	72901	5706

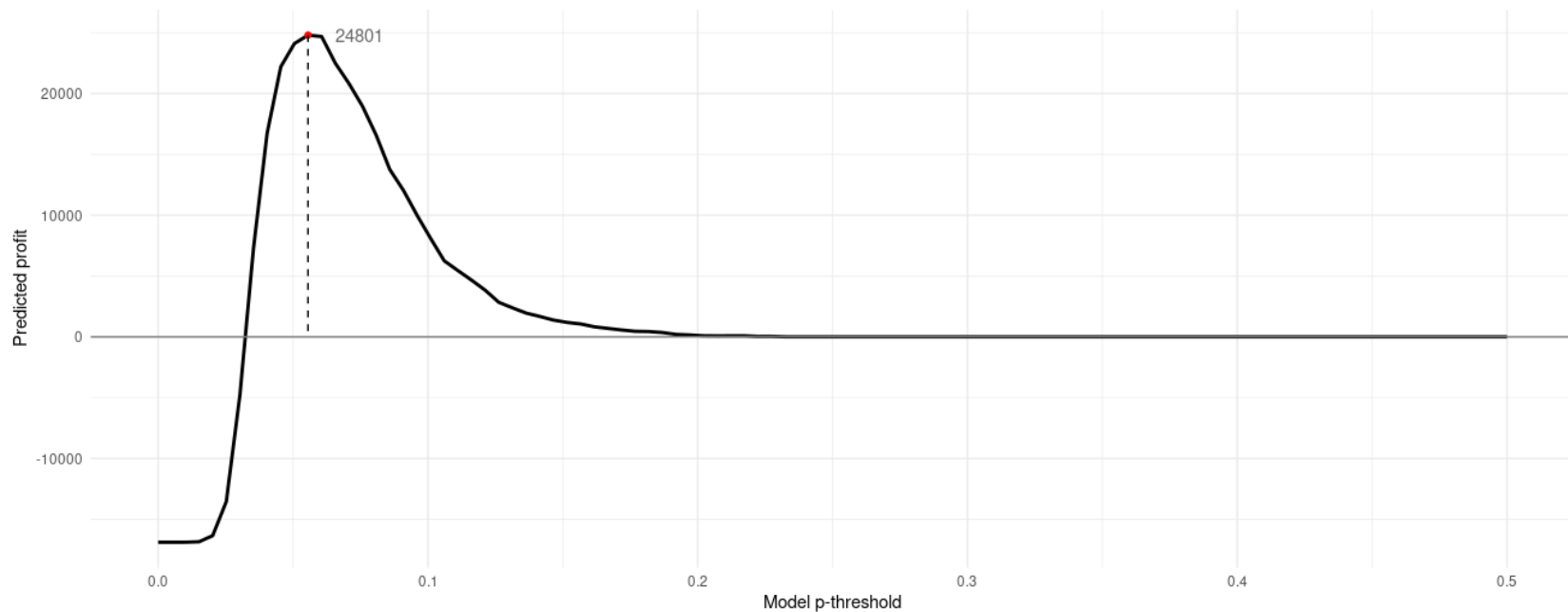


£0.13 profit per person



University
of Exeter

Charity Mailing: Profit Curve



There is a profitable strategy though, where the non-profit gains £24,801

Selection Bias

Our models are based on data from the past.

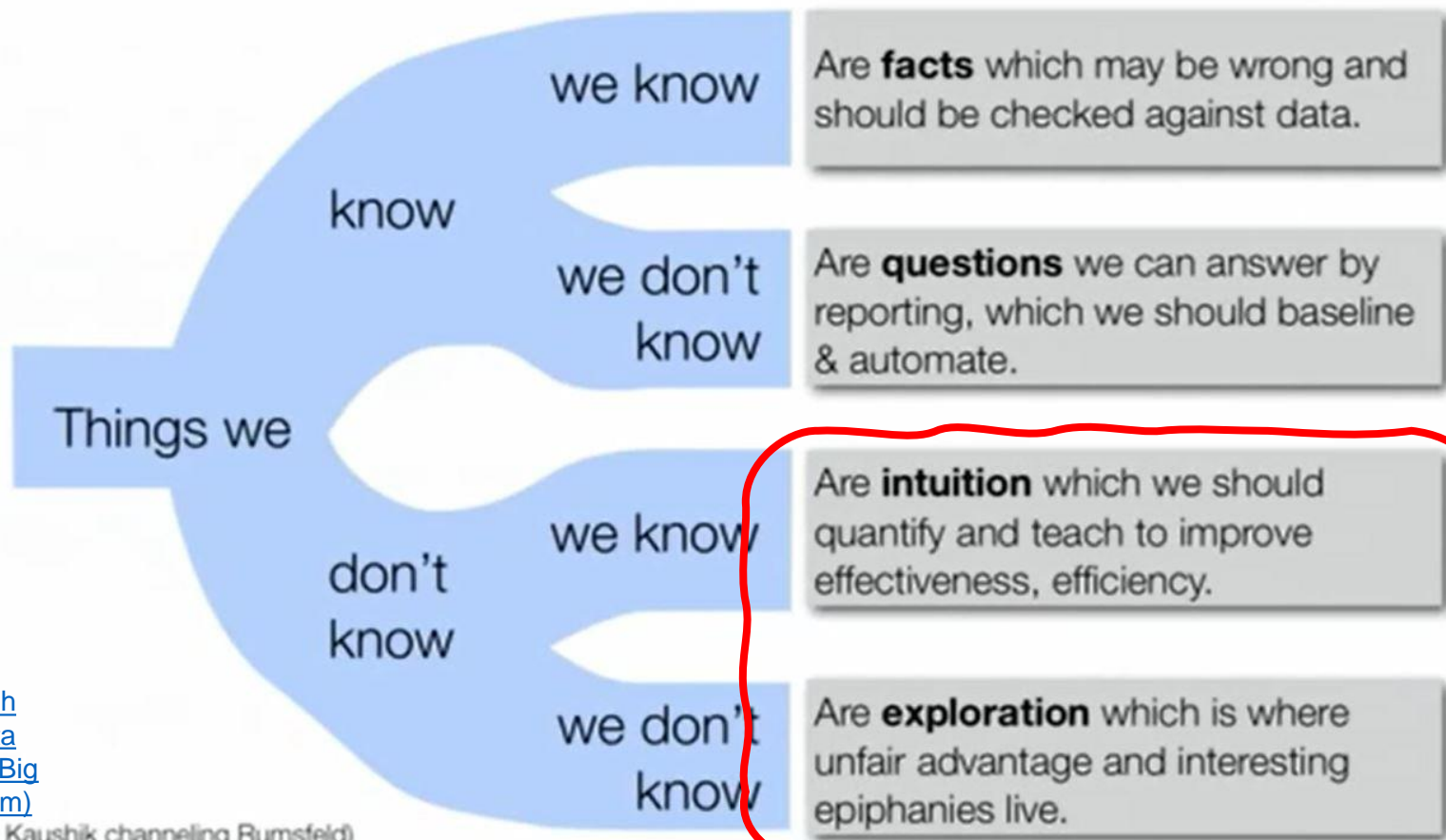
The people in our charity mailing list are people who likely gave already in the past.

They are not a random selection of the population. This is a form of *selection bias*.

The data we have on our donors assumes they have donated in the past.

To make the best decisions we need better / more data.

Rumsfeld on Analytics



[Strata 2012: Avinash Kaushik, "A Big Data Imperative: Driving Big Action" \(youtube.com\)](#)

(Or rather, Avinash Kaushik channeling Rumsfeld)



University
of Exeter

Challenges

Data
availability

Resistance to
change

Lack of
skilled staff

Cost

Data quality

Managing
expectations

Legacy
systems

Data
interpretation

Legacy
cultures

Data silos

Data privacy
and ethics

Legacy
skillsets

Complexity
of tools

Lack of
infrastructure

Analysis
paralysis

Change
management

Privacy and
security
concerns

We need more than data

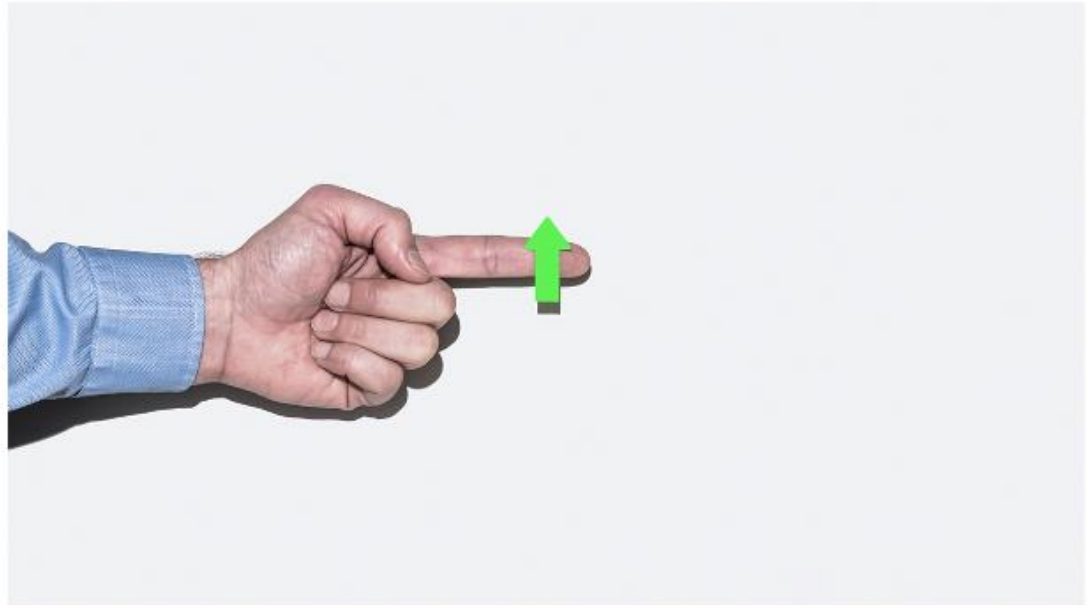
- Focus data initiative and identify high impact cases
 - Reconsider how organisations handle data
 - A transformation to being data driven is a long-term process
-
- Being data driven is for everyone. This is key to cultural change
 - Improve data fluency of all staff

Why Is It So Hard to Become a Data-Driven Company?

by Randy Bean

February 05, 2021

[Why Is It So Hard to Become a Data-Driven Company? \(hbr.org\)](https://hbr.org)



Jorg Greuel/Getty Images

Next Week: Week 10

Textbook Ch. 12, 13

Watch: [Introduction to Ethical AI](#)

Listen: Talking Machines - AI for Good and The Real World

Watch: [Getting Specific about Algorithmic Bias](#)

Watch: [7 minutes to understand AI](#) – A set of UNESCO videos

For more detail:

- Watch: [Deep Learning State of the Art \(2020\) | MIT Deep Learning Series](#)



Any questions?

?