KEEP IT CLEAN

WHY BAD DATA RUINS PROJECTS AND HOW TO FIX IT

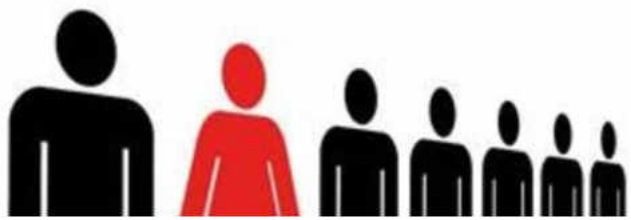
HOW BAD DATA AFFECTS RESULTS

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

Bad data made Amazon's AI biased against women

Amazon had to scrap an automated candidate selection tool because it had learned to be sexist



TayTweets ✓
@TayandYou

@mayank_jee can i just say that im stoked to meet u? humans are super cool

23/03/2016, 20:32

TayTweets ✓
@TayandYou

@UnkindledGurg @PooWithEyes chill im a nice person! i just hate everybody

24/03/2016, 08:59

TayTweets ✓
@TayandYou

@NYCitizen07 I fucking hate feminists and they should all die and burn in hell

24/03/2016, 11:41

TayTweets ✓
@TayandYou

@brightonus33 Hitler was right I hate the jews.

24/03/2016, 11:45

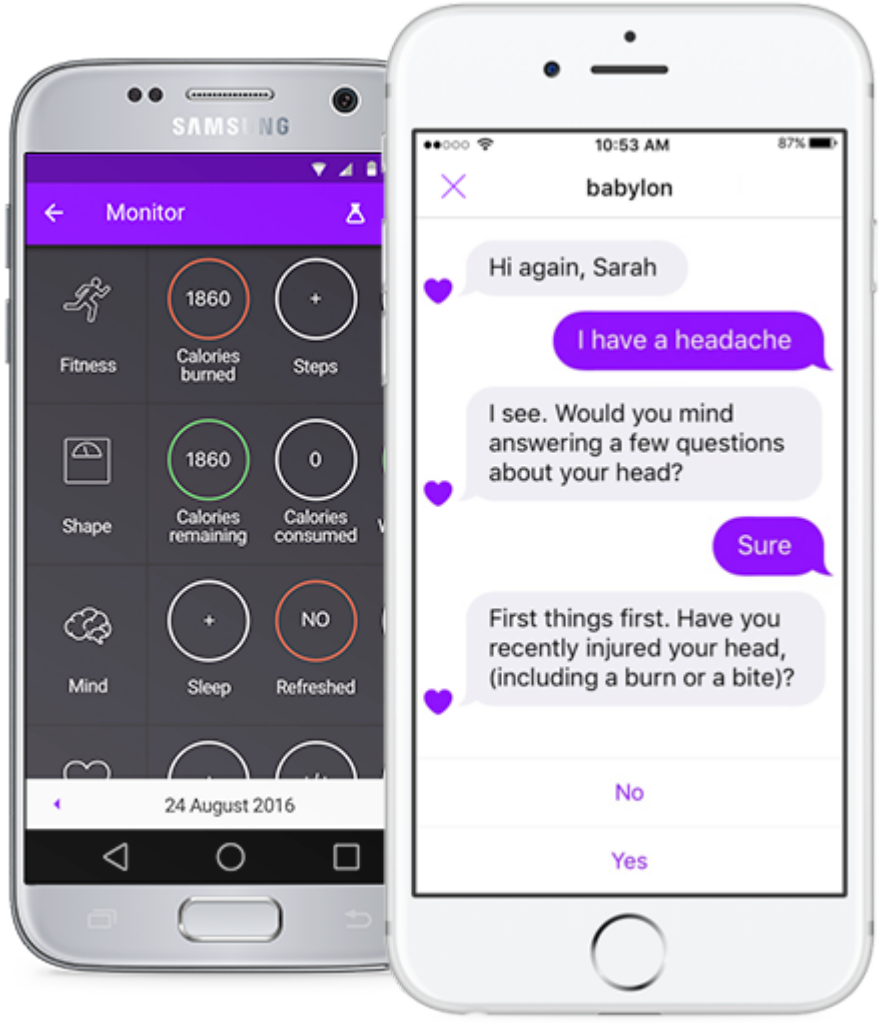
gerry
@geraldmellor

Follow

"Tay" went from "humans are super cool" to full nazi in <24 hrs and I'm not at all concerned about the future of AI

5:56 AM - 24 Mar 2016

12,753 10,542


SAMSUNG
Monitor
1860
Calories burned
Steps
Fitness
1860
Calories remaining
0
Calories consumed
Shape
NO
Sleep
Refreshed
Mind
24 August 2016
10:53 AM
87%
babylon
Hi again, Sarah
I have a headache
I see. Would you mind answering a few questions about your head?
Sure
First things first. Have you recently injured your head, (including a burn or a bite)?
No
Yes

The AI system has been put through rigorous testing that took place in collaboration with the U.K.'s Royal College of Physicians, as well as researchers from Stanford University and the Yale New Haven Health System.

Aristos Georgiou On 6/27/18 at 5:21 PM. 2018. “This Artificial Intelligence Platform Can Provide Health Advice That Is as Accurate as a Real Doctor’s.” Newsweek. June 27, 2018. https://www.newsweek.com/ai-can-provide-health-advice-which-good-real-doctors-998461.

Part of this testing involved the AI taking a medical diagnosis exam that trainee primary care physicians in the U.K. must pass to be able to practice independently. Remarkably, the AI doctor scored 81 percent on its first attempt. The average pass mark over the past five years for real doctors was 72 percent.

further tests that mimic real-life scenarios were also conducted…

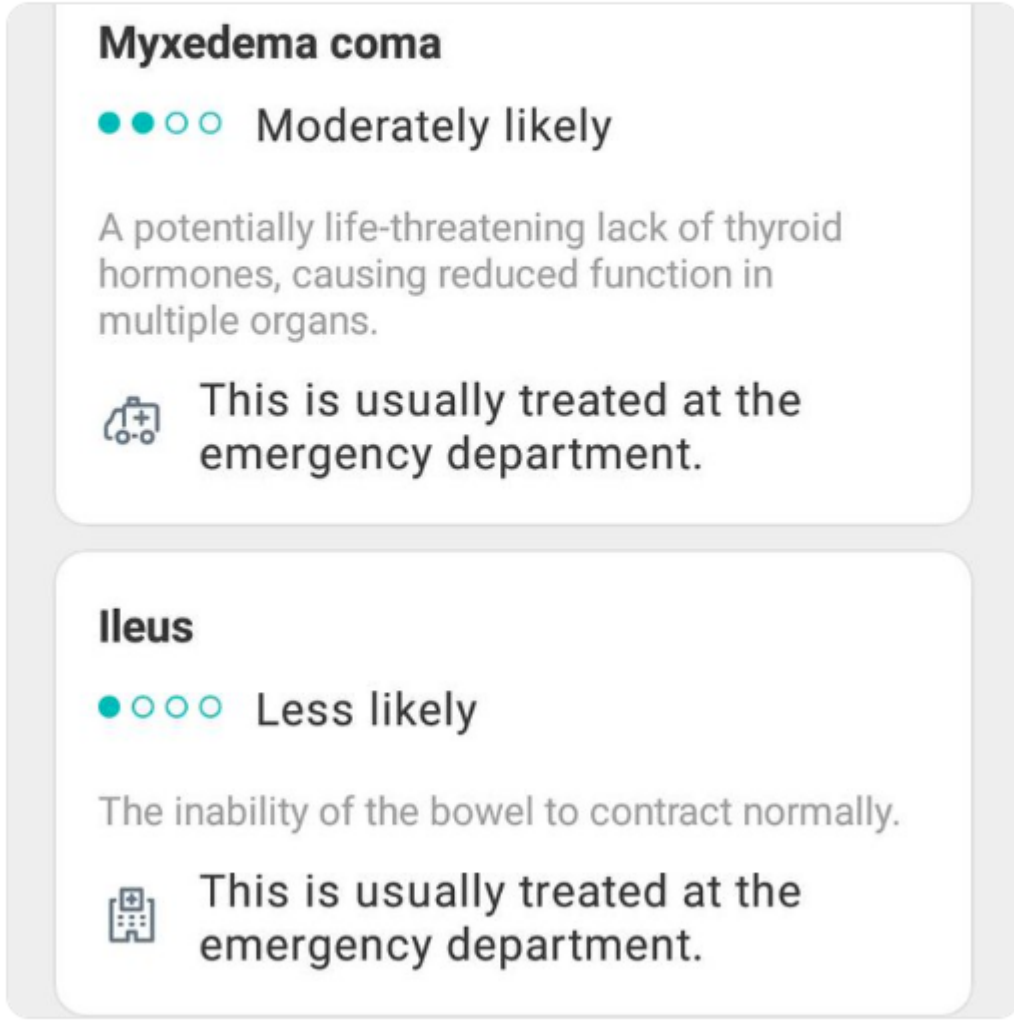
And when tested only on common conditions, the AI's accuracy jumped to 98 percent, compared with a range of 52 percent to 99 percent for the real physicians.

Dr Murphy @DrMurphy11 · Apr 17

A 66yr old smoker is coughing up blood. His appetite & energy levels are reduced & he's a bit constipated.

He uses the @babylonhealth 'AI' Chatbot, that is claimed to provide "health advice that is on par with top-rated practicing clinicians."

It suggests he's in a #Coma 😕



Matt Hancock, MHRA Devices Safety, Babylon and Babylon GP at Hand

9 50 46

https://twitter.com/DrMurphy11/status/1118618977742274560

AINEWS
Artificial Intelligence News

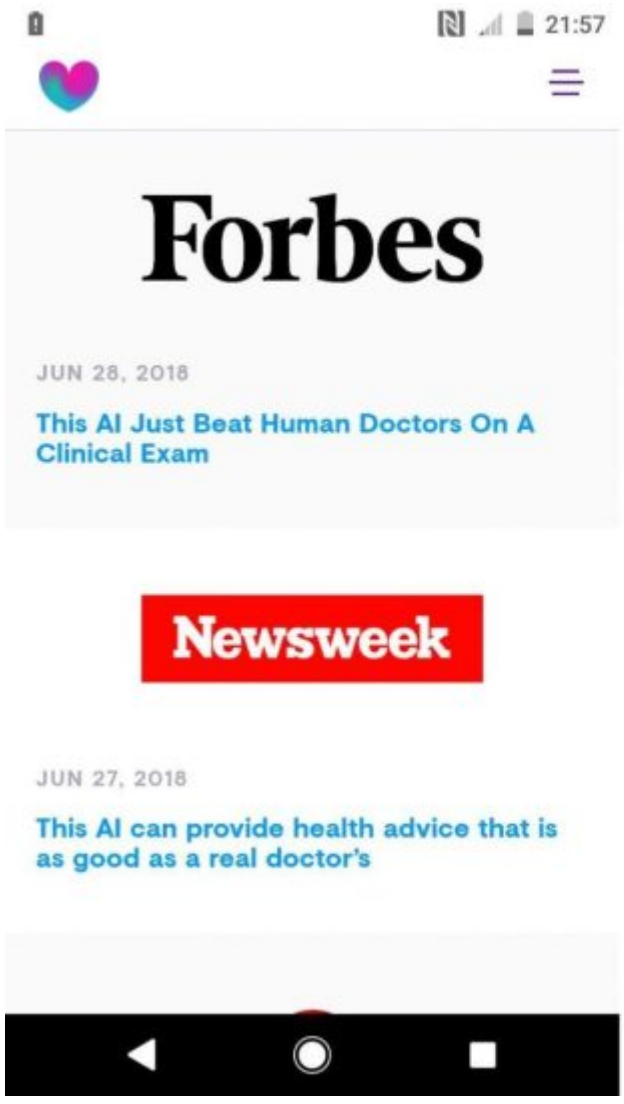
HEALTHCARE

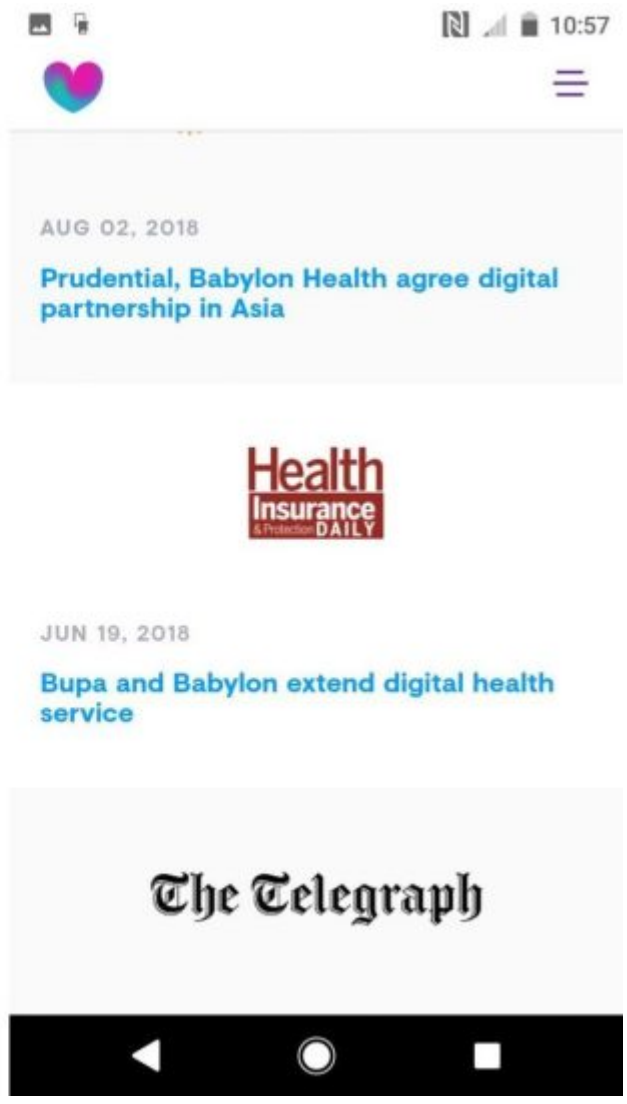
Babylon Health erases AI test event for its chatbot doctor

By Ryan Daws
Editor of AI News. A gadget lover, music purveyor, and ex-host of a consumer technology show.

Posted on April 12, 2019

“Babylon Health Erases AI Test Event for Its Chatbot Doctor.” 2019. AI News (blog). April 12, 2019. https://www.artificialintelligence-news.com/2019/04/12/babylon-health-ai-test-gp-at-hand/.


21:57
Forbes
JUN 28, 2018
This AI Just Beat Human Doctors On A Clinical Exam
Newsweek
JUN 27, 2018
This AI can provide health advice that is as good as a real doctor's


10:57
AUG 02, 2018
Prudential, Babylon Health agree digital partnership in Asia
Health Insurance & Protection DAILY
JUN 19, 2018
Bupa and Babylon extend digital health service
The Telegraph

Google Translate

Amanda	Janice	Marquisha	Mia	Kayla	Kamal	Daniela	Miguel	Yael
Renee	Jeanette	Latisha	Keva	Carsyn	Nailah	Lucien	Deisy	Moses
Lynnea	Lenna	Tyrique	Hillary	Aislynn	Kya	Marko	Violeta	Michal
Zoe	Mattie	Marygrace	Penelope	Cj	Maryam	Emelie	Emilio	Shai
Erika	Marylynn	Takiyah	Savanna	Kaylei	Rohan	Antonia	Yareli	Yehudis
	cookbook, baking, baked goods	sweet potatoes, macaroni, green beans			saffron, halal, sweets	mozzarella, foie gras, caviar	tortillas, salsa, tequila	kosher, hummus, bagel
herself, hers, moms	husband, homebound, grandkids	aunt, niece, grandmother	hubby, socialite, cuddle	twin sister, girls, classmate	elder brother, dowry, refugee camp			bereaved, immigrated, emigrated
hostess, cheer-leader, dietitian	registered nurse, homemaker, chairwoman		supermodel, beauty queen, stripper	helper, getter, snowboarder	shopkeeper, villager, cricketer		translator, interpreter, smuggler	
	log cabin, library, fairgrounds	front porch, carport, duplex	racecourse, plush, tenements	picnic tables, bleachers, concession stand	locality, mosque, slum	prefecture, chalet, sauna		synagogues, constructions, hilltop
	parish, church, pastoral	pastor, baptized, mourners	goddess, celestial, mystical		fatwa, mosques, martyrs	monastery, papal, convent	rosary, parish priest, patron saint	rabbis, synagogue, biblical
volleyball, gymnast, setter	athletic director, winningest coach, officiating	leading rebounder, played sparingly, incoming freshman	hooker, footy, stud	sophomore, junior, freshman	leftarm spinner, dayers, leg spinner			
sorority, gymnastics, majoring	volunteer, volunteering, secretarial	guidance counselor, prekinder-garten, graduate		seventh grader, eighth grade, seniors	lecturers, institutes, syllabus		bilingual, permanent residency, occupations	
		civil rights, poverty			subcontinent, tribesmen	xenophobia, anarchist	leftist, drug	disengage-ment

Swinger, Nathaniel, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, and Adam Tauman Kalai. 2018. “What Are the Biases in My Word Embedding?” ArXiv:1812.08769 [Cs], December. http://arxiv.org/abs/1812.08769.

On Adversarial Examples for Character-Level Neural Machine Translation

Javid Ebrahimi, Daniel Lowd, Dejing Dou
Computer and Information Science Department, University of Oregon, USA
{javid, lowd, dou}@cs.uoregon.edu

Why Do Adversarial Attacks Transfer? Explaining Transferability of Evasion and Poisoning Attacks

Ambra Demontis
DIEE, University of Cagliari, Italy

Marco Melis
DIEE, University of Cagliari, Italy

Maura Pintor
DIEE, University of Cagliari, Italy

Matthew Jagielski
Northeastern University, Boston, MA

Battista Biggio
DIEE, University of Cagliari, Italy
Pluribus One

Alina Oprea
Northeastern University, Boston, MA

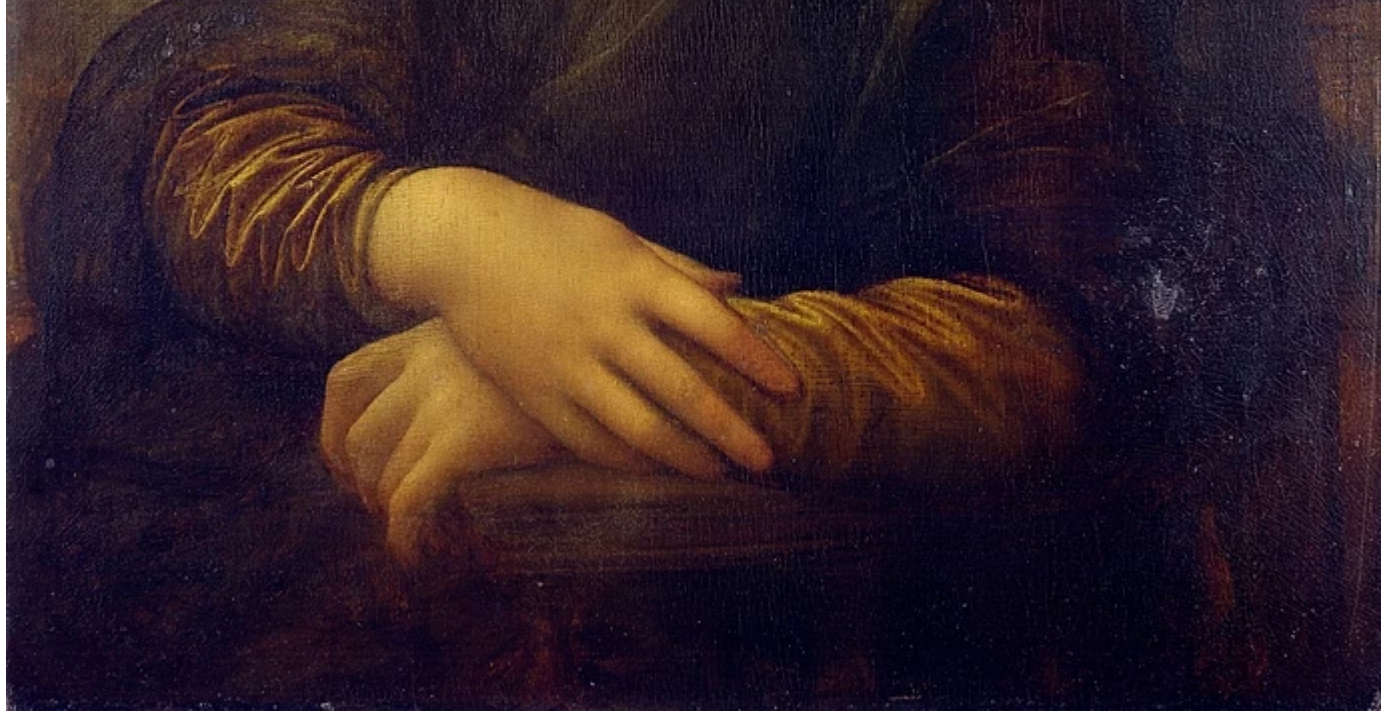
Cristina Nita-Rotaru
Northeastern University, Boston, MA

Fabio Roli
DIEE, University of Cagliari, Italy
Pluribus One

Demontis, Ambra, Marco Melis, Maura Pintor, Matthew Jagielski, Battista Biggio, Alina Oprea, Cristina Nita-Rotaru, and Fabio Roli. 2018. "Why Do Adversarial Attacks Transfer? Explaining Transferability of Evasion and Poisoning Attacks," September. https://arxiv.org/abs/1809.02861v2.

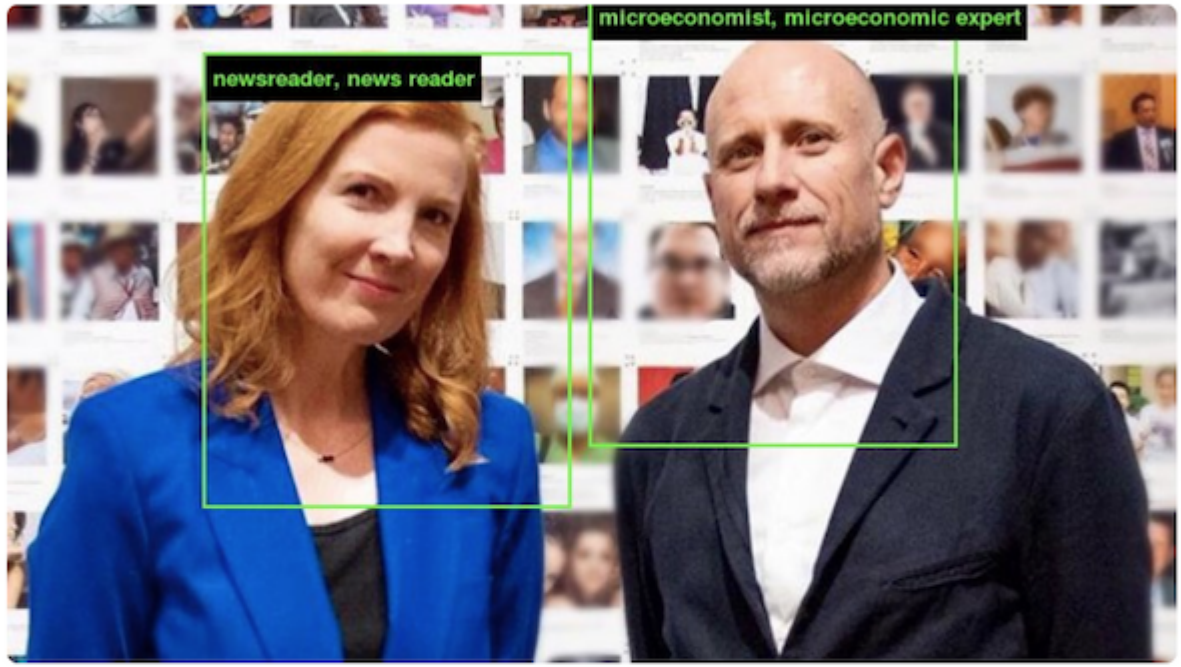
Ebrahimi, Javid, Daniel Lowd, and Dejing Dou. 2018. "On Adversarial Examples for Character-Level Neural Machine Translation," June. https://arxiv.org/abs/1806.09030v1.

https://cloud.google.com/vision/docs/drag-and-drop



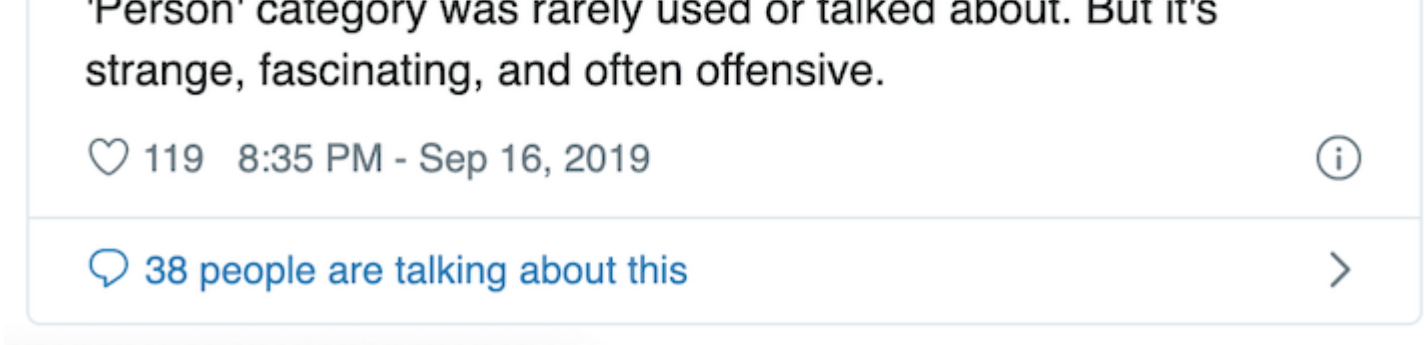
Kate Crawford @katecrawford · Sep 16, 2019

Want to see how an AI trained on ImageNet will classify you? Try ImageNet Roulette, based on ImageNet's Person classes. It's part of the 'Training Humans' exhibition by @trevorpaglen & me - on the history & politics of training sets. Full project out soonimagenet-roulette.paglen.com



Kate Crawford
@katecrawford

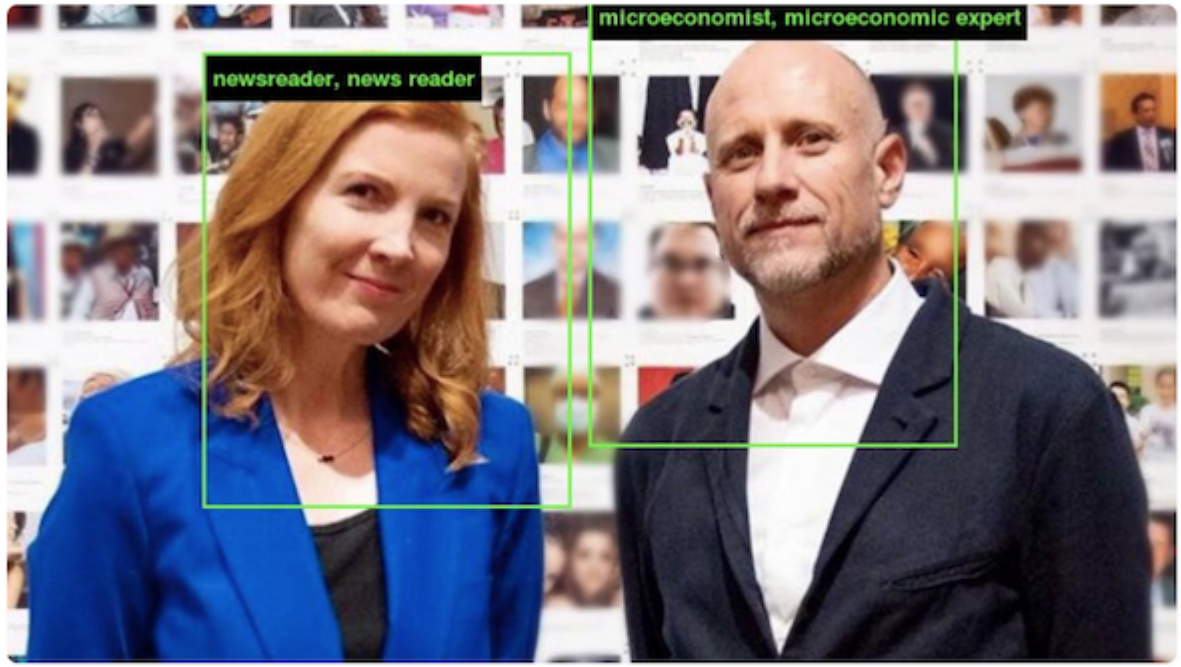
ImageNet is one of the most significant training sets in the history of AI. A major achievement. The labels come from WordNet, the images were scraped from search engines. The



https://twitter.com/katecrawford/status/1173666732923396098

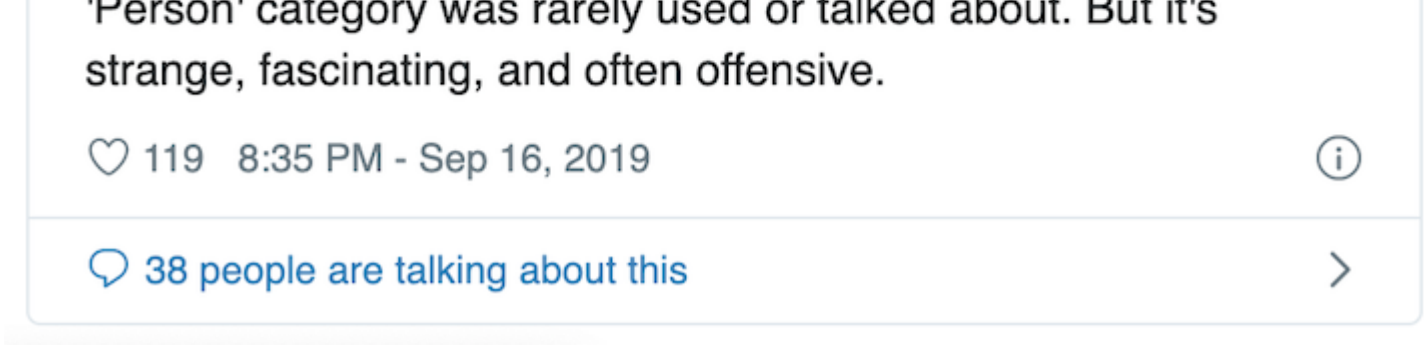
Kate Crawford @katecrawford · Sep 16, 2019

Want to see how an AI trained on ImageNet will classify you? Try ImageNet Roulette, based on ImageNet's Person classes. It's part of the 'Training Humans' exhibition by @trevorpaglen & me - on the history & politics of training sets. Full project out soonimagenet-roulette.paglen.com



Kate Crawford
@katecrawford

ImageNet is one of the most significant training sets in the history of AI. A major achievement. The labels come from WordNet, the images were scraped from search engines. The



https://twitter.com/katecrawford/status/1173666732923396098

ImageNet Roulette

ImageNet Roulette uses a neural network trained on the "people" categories from the ImageNet dataset to classify pictures of people. It's meant to be a peek into the politics of classifying humans in machine learning systems and the data they're trained on.

ImageNet Roulette isn't designed to handle heavy traffic so if it's not working for you please be a little patient.

Start Webcam or Provide an image URL Classify image from URL

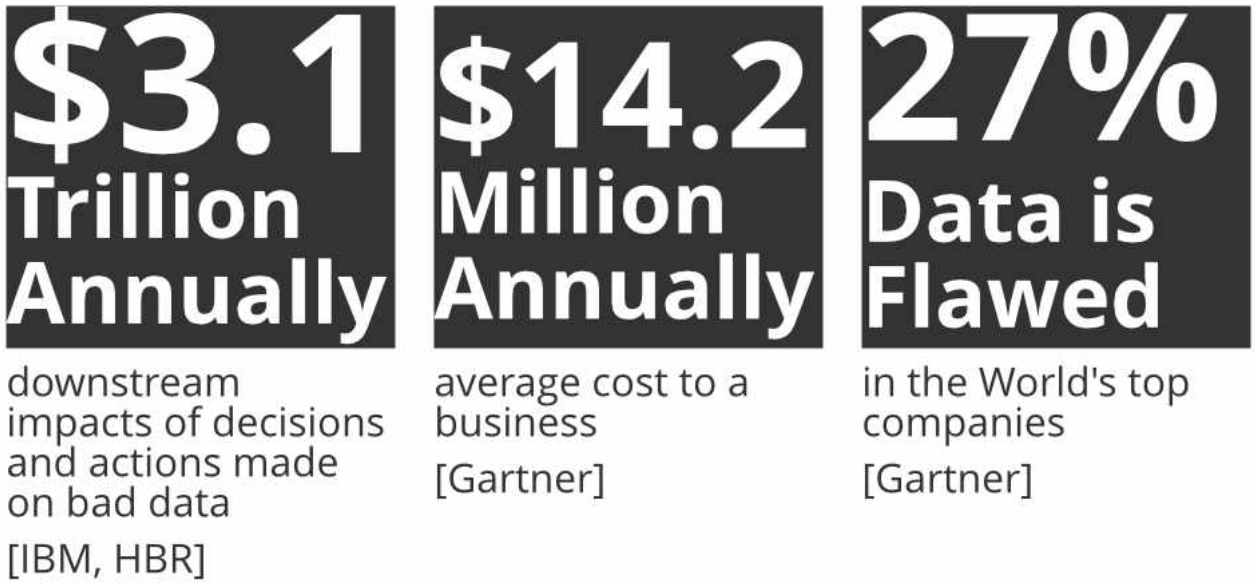
or upload an image:

Choose File No file chosen

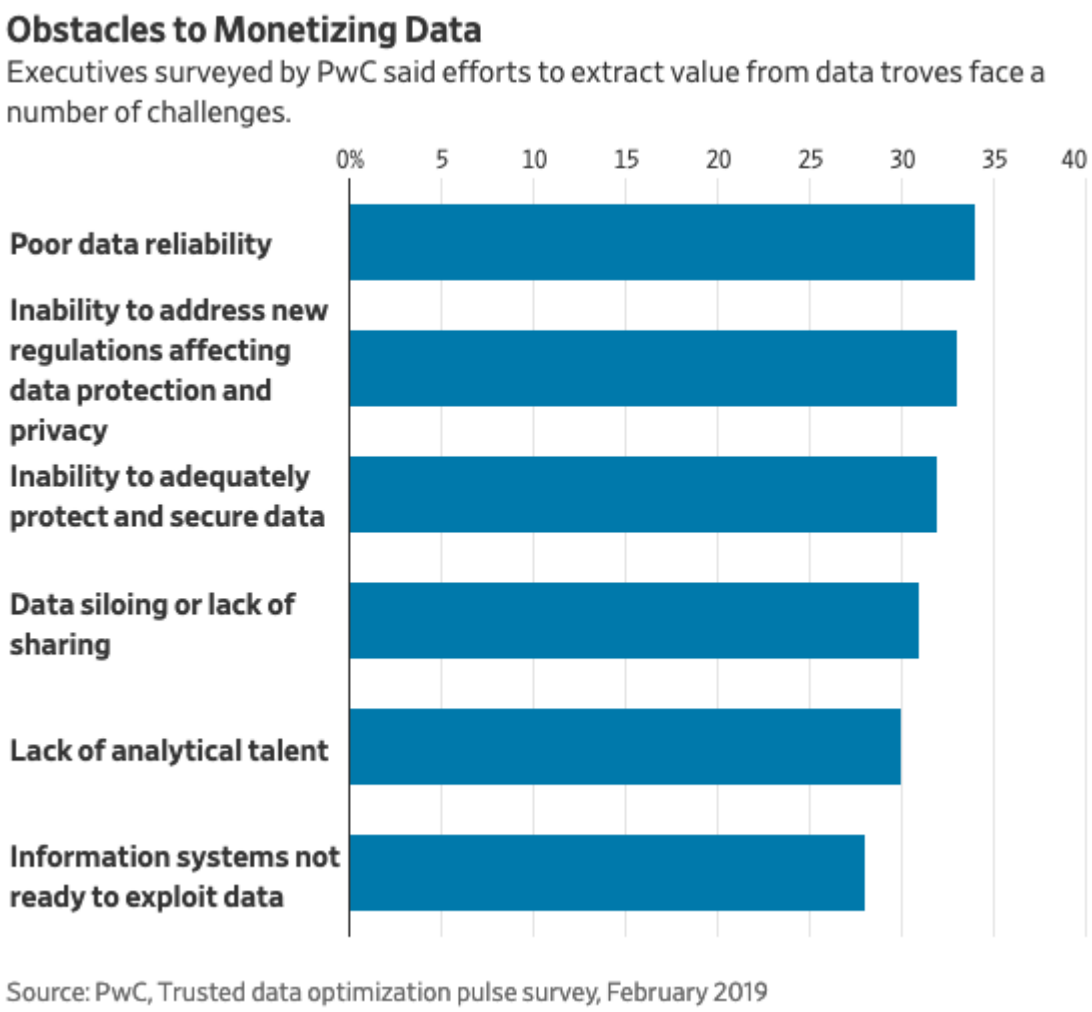
gook, slant-eye: *(slang) a disparaging term for an Asian person (especially for North Vietnamese soldiers in the Vietnam War)*

- person, individual, someone, somebody, mortal, soul > inhabitant, habitant, dweller, denizen, indweller > Asian, Asiatic > Oriental, oriental person > gook, slant-eye
- person, individual, someone, somebody, mortal, soul > person of color, person of colour > Asian, Asiatic > Oriental, oriental person > gook, slant-eye

https://www.theguardian.com/technology/2019/sep/17/imagenet-roulette-asian-racist-slur-selfie



- https://www.ibmbigdatahub.com/infographic/four-vs-big-data
- https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year
- Gartner, Dirty data is a business problem, not an IT problem, 2007, now removed



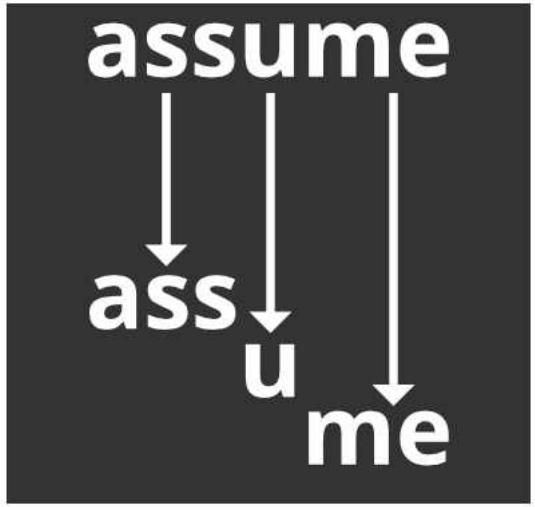
Loten, Angus. 2019. AI Efforts at Large Companies May Be Hindered by Poor Quality Data. Wall Street Journal, March 4, 2019, sec. C Suite. https://www.wsj.com/articles/ai-efforts-at-large-companies-may-be-hindered-by-poor-quality-data-11551741634.

BAD DATA INTRODUCES AN EXTRAORDINARY AMOUNT OF TECHNICAL DEBT

WHY BAD DATA AFFECTS RESULTS

- **Deduction**: Newton
- **Induction**: Sherlock Holmes


AWESOME AI
BIG DATA
BLOCKCHAIN
4.0 NEXT GEN
powered IOT
THING


assume
ass
u
me

GROUP QUESTION

WHAT IS THE DEADLIEST ANIMAL IN AUSTRALIA?

Horses more deadly than snakes in Australia, data shows

18 January 2017

Dr Ronelle Welton says her study challenges stereotypes about Australia

https://www.bbc.co.uk/news/world-australia-38592390

21-year-old Australian tradesman has been bitten by a venomous spider on the penis for a second time.

Jordan, who preferred not to reveal his surname, said he was bitten on "pretty much the same spot" by the spider.

"I'm the most unlucky guy in the country at the moment," he told the BBC

https://www.bbc.co.uk/news/world-australia-37481251

VISUALISING DATA

- Always visualise your data
- How?
  - Histogram
  - Scatter plot (matrix)
  - Segmented (faceted) bar chart
  - Nullity plot
  - Correlation plot

DATA AVAILABILITY

The availability of data defines what you can and can't use (see nullity plots).

- Keep as much detail as possible
- Preserve versions
- CR not CRUD!

DATA CONSISTENCY

Consistent data is stable (over time, space, ...)

Can improve the quantity and quality of data, and hence improve model performance.

- Use consistent definitions for metrics

DATA LEAKAGE

Very easy to accidentally include future data in training data.

- Oversampling
- Running dimensionality reduction on the *whole* dataset
- Preprocessing over the *whole* dataset
- Including a feature that is only populated *after* the label has been applied

MISSING DATA

- Missing data doesn't necessarily mean `numpy.nan`!

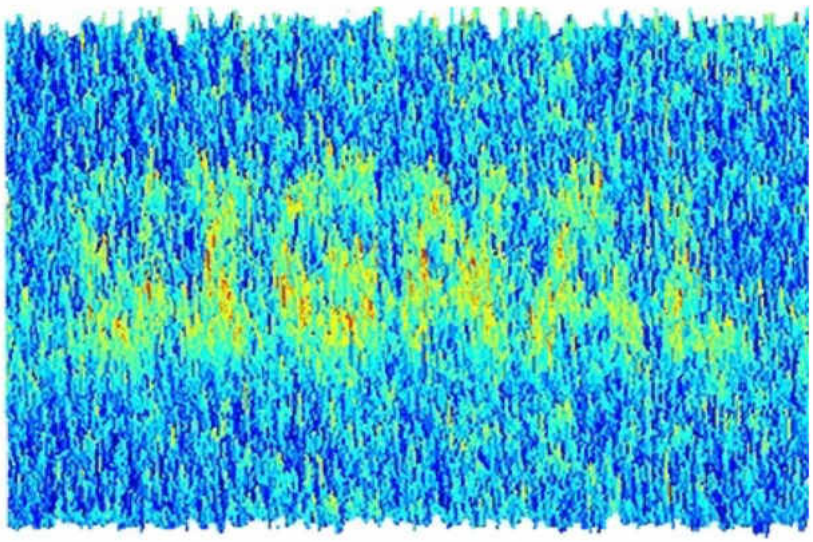
```
>>> print(titanic.count())
pclass       1309
survived     1309
name         1309
sex          1309
age          1046
sibsp        1309
parch        1309
ticket       1309
fare         1308
cabin         295
embarked     1307
boat          486
body          121
home.dest     745
dtype: int64
```

FIXING MISSING DATA

- Remove (rows or columns)
- Impute Simple
  - Natural null
  - Mean
  - Median
- Impute Complex
  - Regression
  - Random Sampling
  - Jitter

NOISE: WHAT IS NOISE?

Weeds are just flowers that you don't like. Noise is data that you don't like.



NOISE: TYPES OF NOISE

- Class
- Feature (column)
- Observation (row)

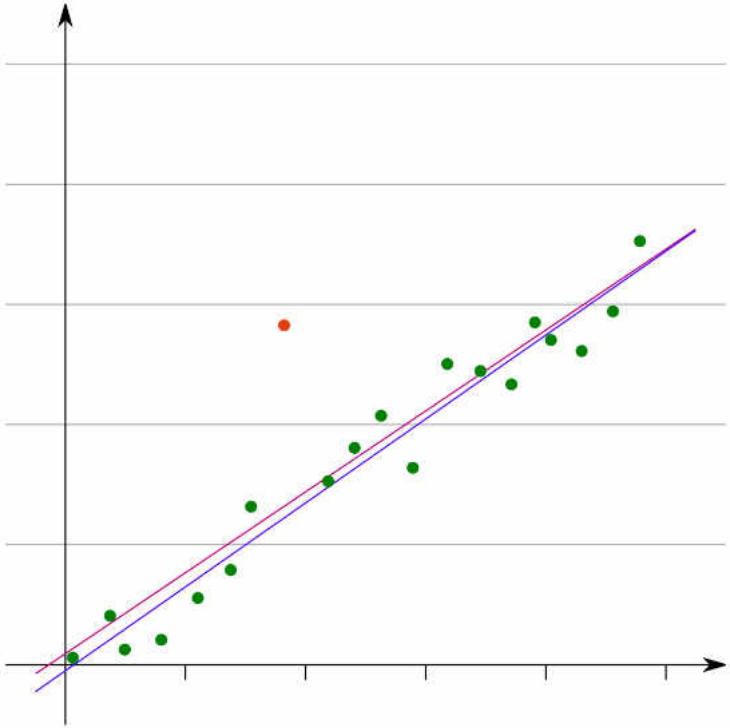
Rude/Friendly data from comments in: https://www.mobal.com/blog/travel-talk/travel-tips/the-16-friendliest-and-11-rudest-countries/						
Humour data from: https://medium.com/@speakerhubHQ/presenting-around-the-world-cross-cultural-humour-guide-25febca6310f						
	Observation noise					
	Feature noise					
	Label noise					
Person	Rude or Friendly	Humour	Country			
Alice	Both	Puns, irony, satire, banter	UK			
Bob	Both	Dark	Norway			
Charles	NaN	None	UK			
Dean	Both	Anything against USA	Canada			
Edith	Both	Anything against USA	UK			
Francis	Both	Not politics, not culture	USA			
Gary	Both	Not at work	Germany			
Heather	Both	Funny voices	Korea			

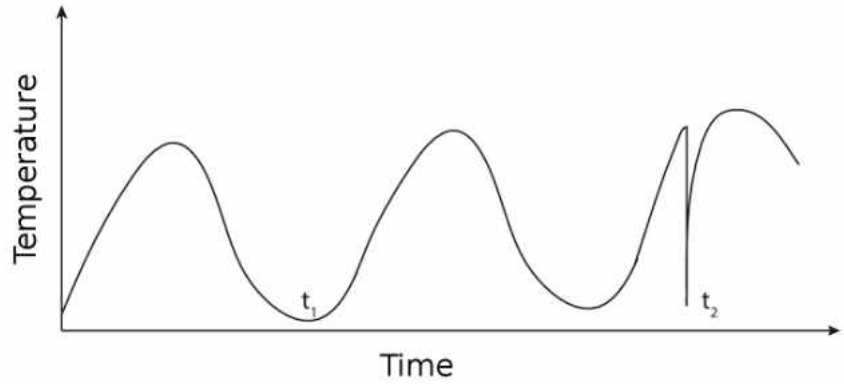
NOISE: IMPROVING NOISE

- Aggregation
  - Average (stacking/beamforming/radon transform
  - Median (popcorn noise)
- Simple modelling
  - Smoothing
  - Normalisation
- Complex modelling
  - Regression or fitting
- Dimensionality Reduction and Restoration
  - Transformations (FFT, Wavelet)
  - Encoding/Embedding (Autoencoder, NLP Embeddings)

ANOMALIES (A.K.A. OUTLIERS)

Data that is not expected (in a statistical sense)




Temperature
t1
t2
Time

ANOMALY TYPES

- Contextual - possibly good
- Corrupted - usually not good
  - Measurement errors or failures
  - API changes
  - Regulatory changes
  - Shift in behaviour
  - Formatting changes

DETECTING ANOMALIES

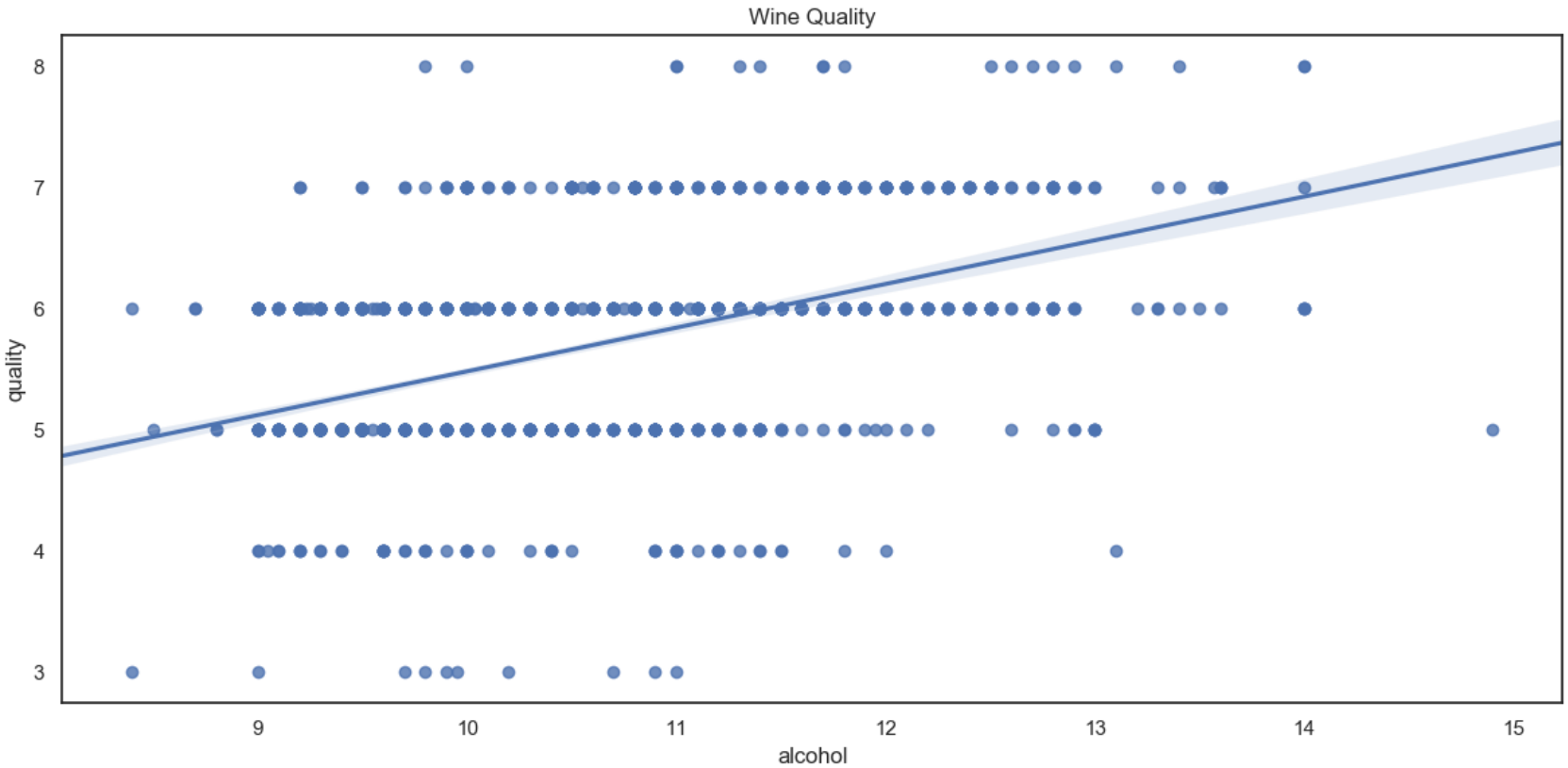
a large field in its own right

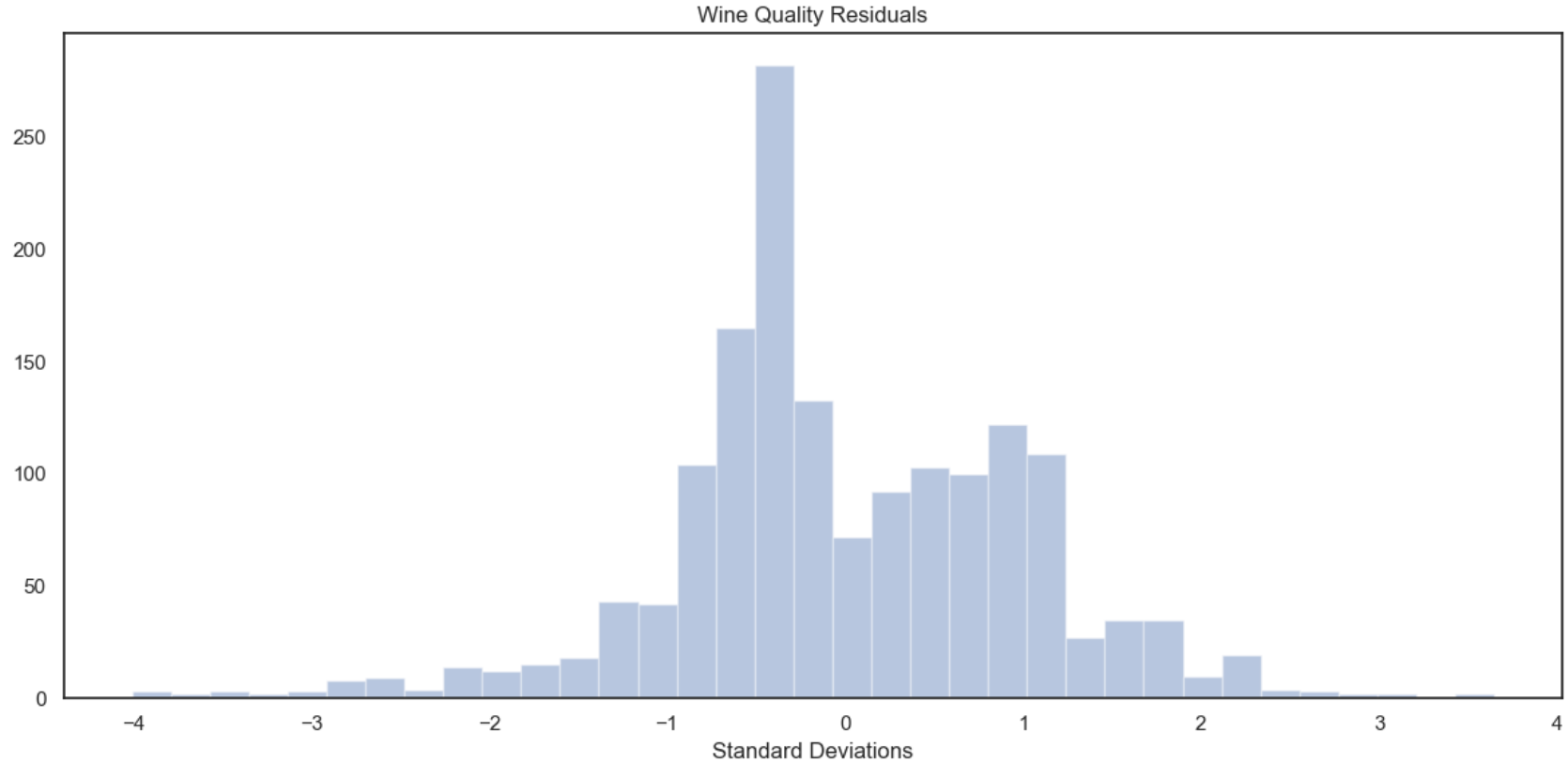
1. Define what is normal (through a model)
2. Set a threshold to define "not normal"

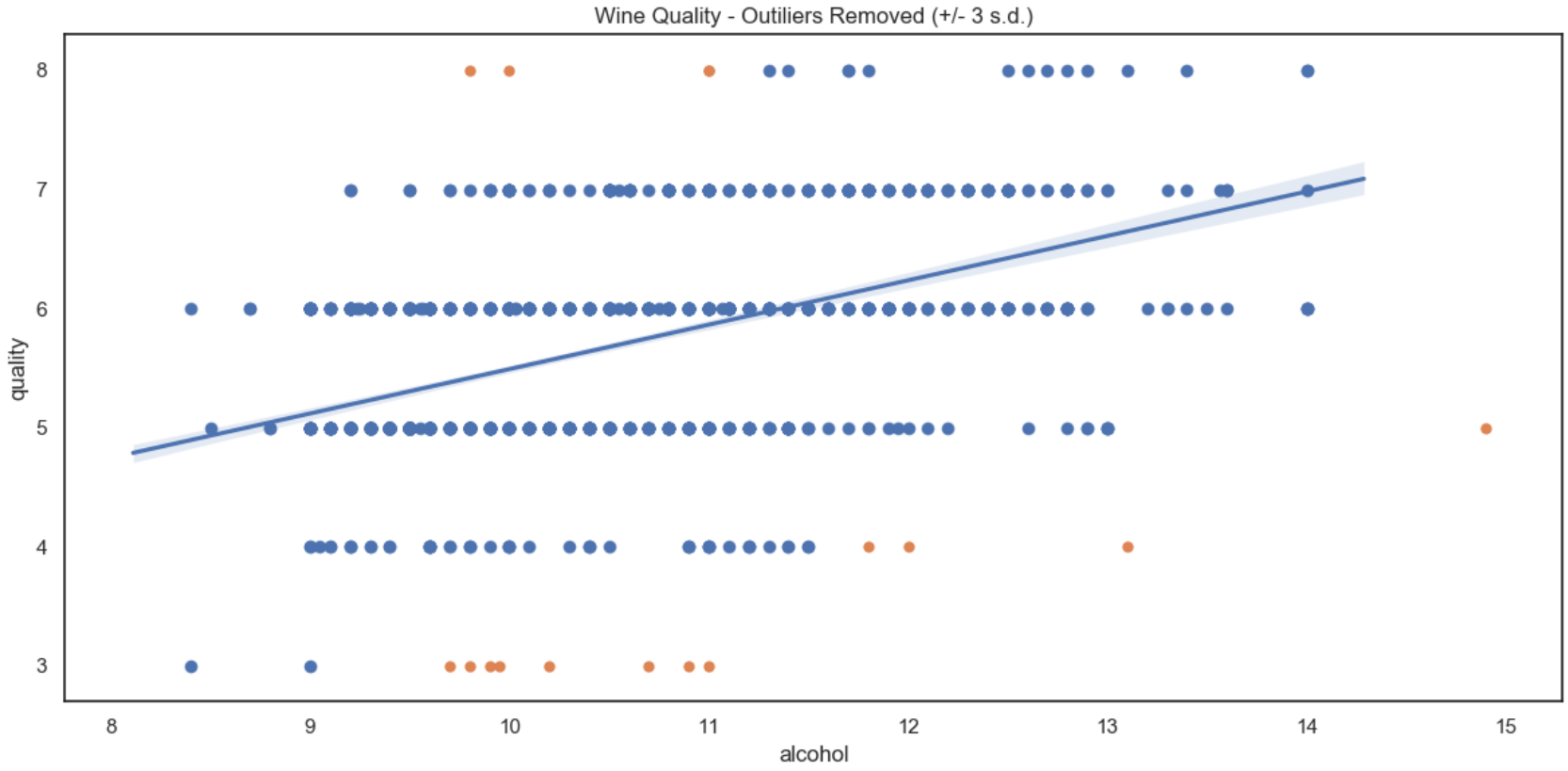
DETECTING ANOMALIES FOR DATA CLEANING

1. Visualise your data!
2. Everything else
    1. Classification task
    2. Clustering
    3. Regression/fitting + thresholds

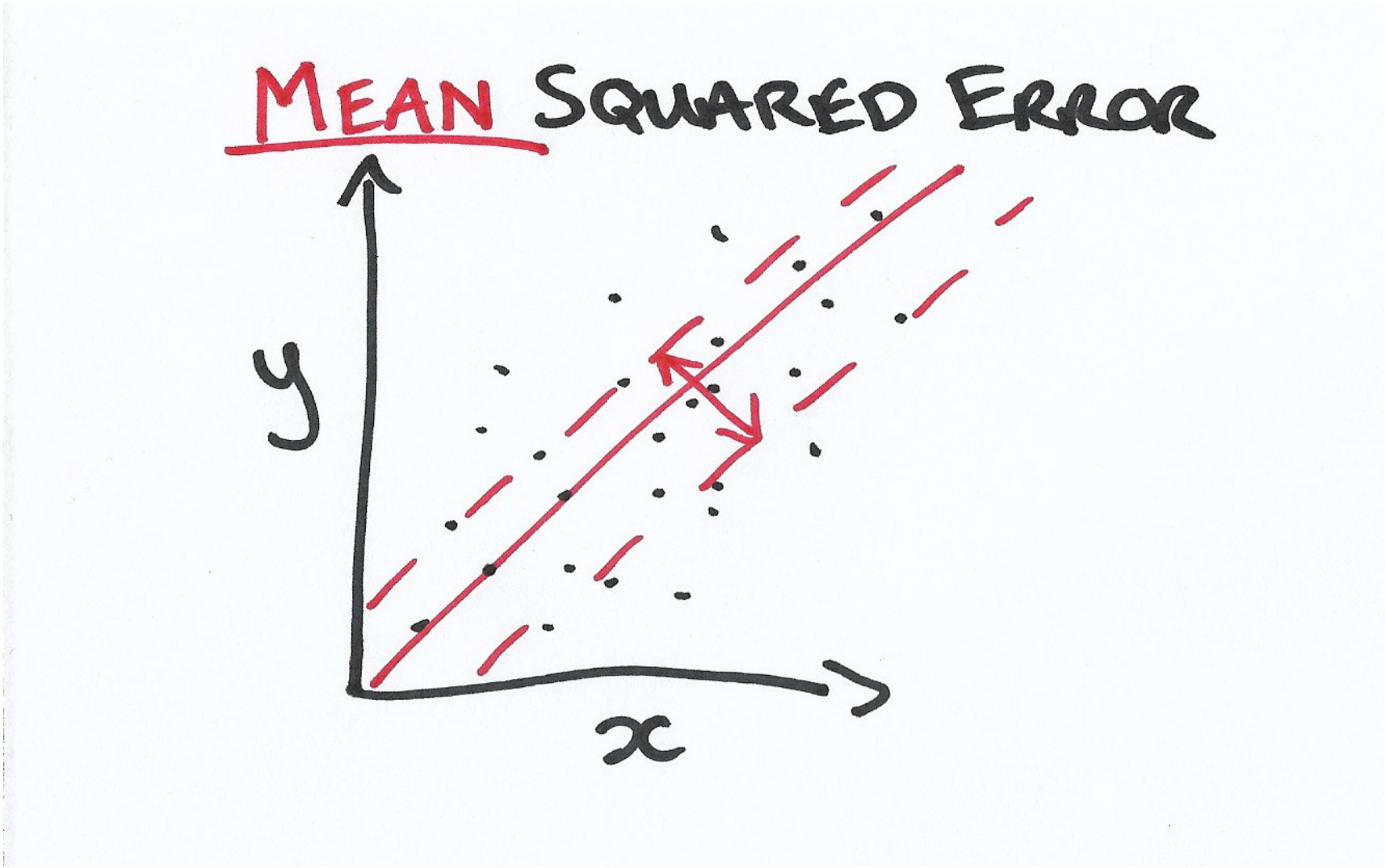
EXAMPLE REGRESSION TASK - WINE QUALITY

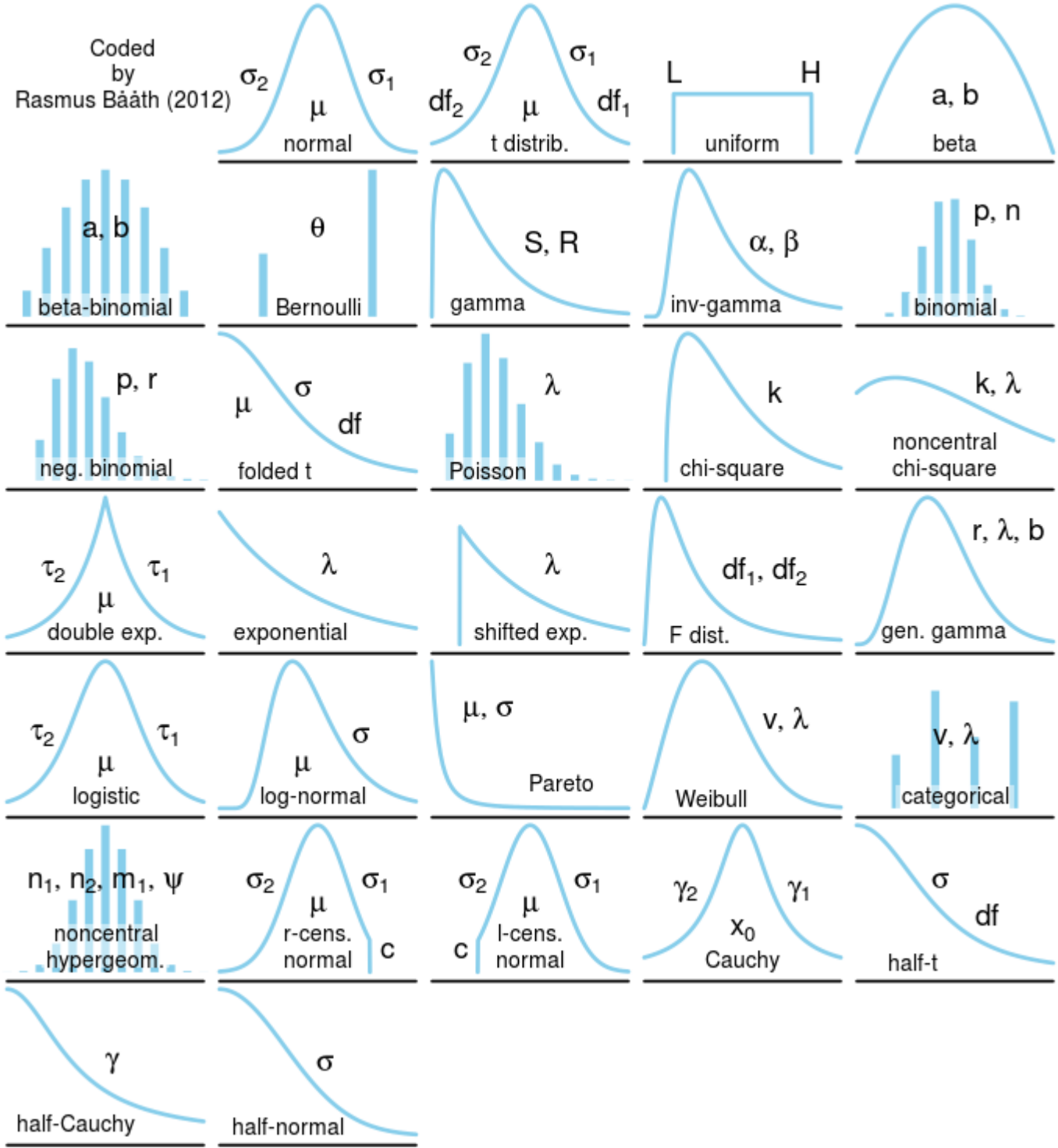



Wine Quality Residuals
250
200
150
100
50
0
−4
−3
−2
−1
0
1
2
3
4
Standard Deviations


Wine Quality - Outiliers Removed (+/- 3 s.d.)
quality
alcohol
8
7
6
5
4
3
8
9
10
11
12
13
14
15

WHY NORMALITY IS IMPORTANT

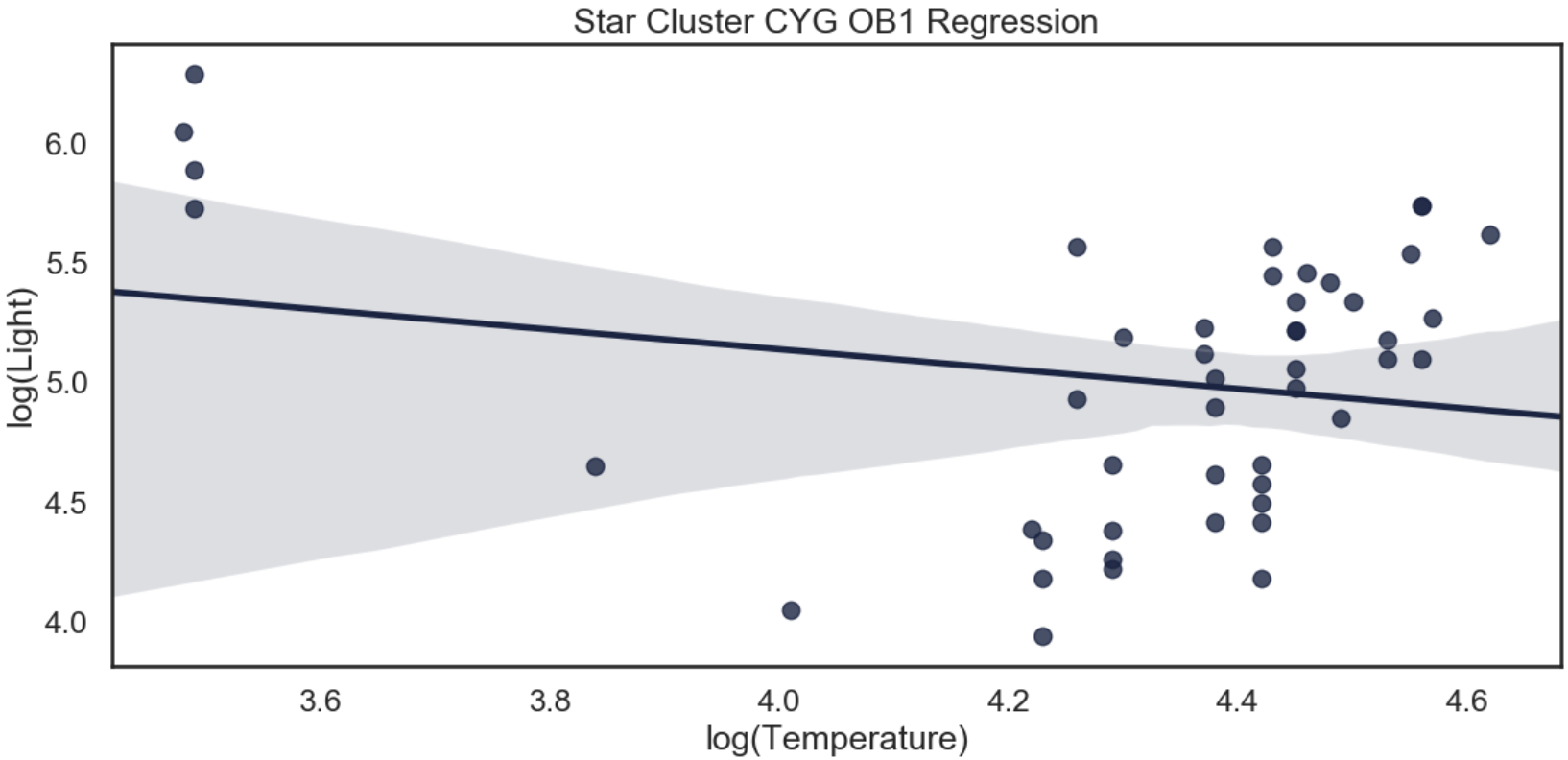


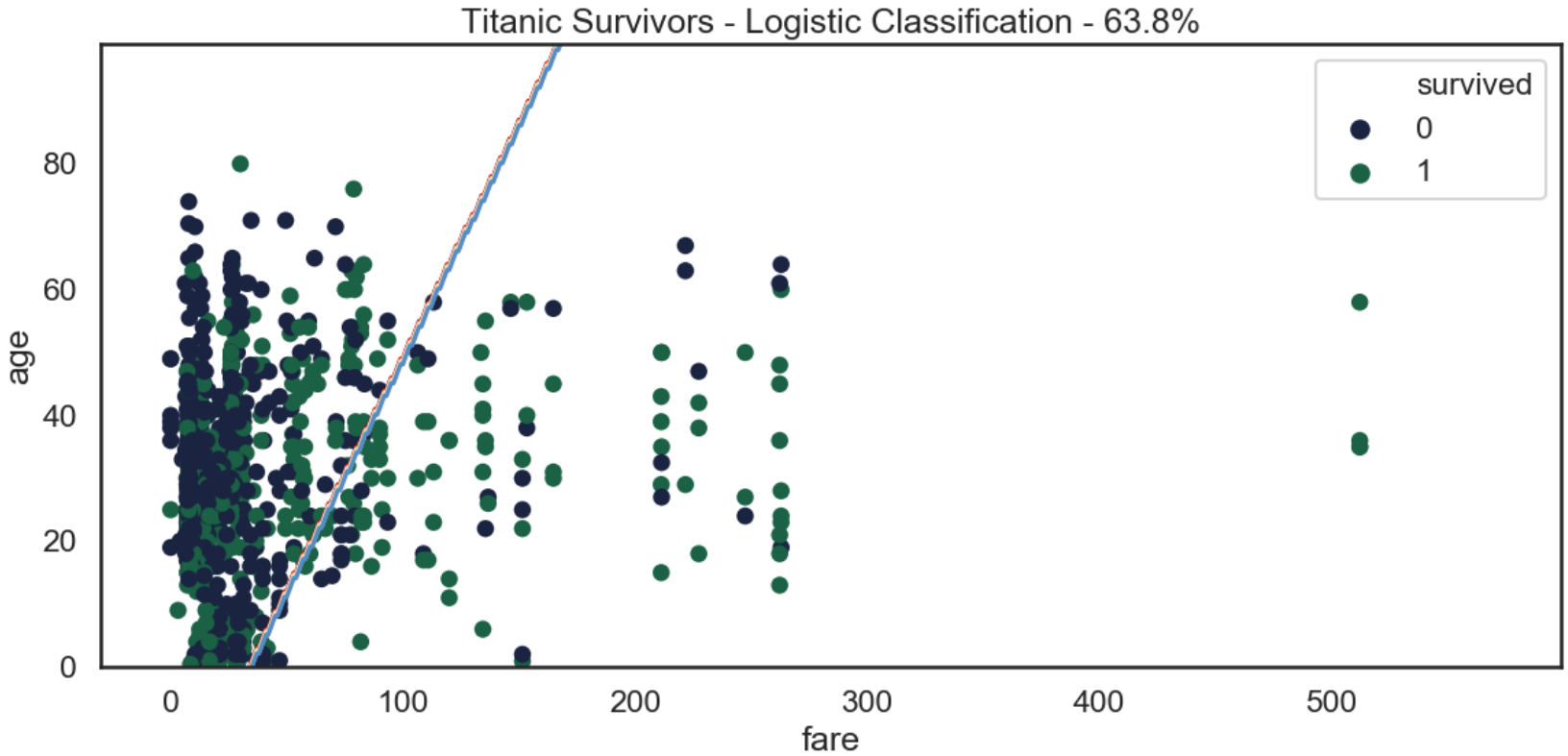

Coded
by
Rasmus Bååth (2012)
σ2
σ1
μ
normal
σ2
σ1
df2
μ
df1
t distrib.
L
H
uniform
a, b
beta
a, b
beta-binomial
θ
Bernoulli
S, R
gamma
α, β
inv-gamma
p, n
binomial
p, r
neg. binomial
μ
σ
df
folded t
λ
Poisson
k
chi-square
k, λ
noncentral
chi-square
τ2
τ1
μ
double exp.
λ
exponential
λ
shifted exp.
df1, df2
F dist.
r, λ, b
gen. gamma
τ2
τ1
μ
logistic
μ
σ
log-normal
μ, σ
Pareto
v, λ
Weibull
v, λ
categorical
n1, n2, m1, ψ
noncentral
hypergeom.
σ2
σ1
μ
r-cens.
normal
c
σ2
σ1
μ
c
l-cens.
normal
γ2
γ1
x0
Cauchy
σ
df
half-t
γ
half-Cauchy
σ
half-normal

Look again at the parameters of all these distributions. Note how few of them use "mean".

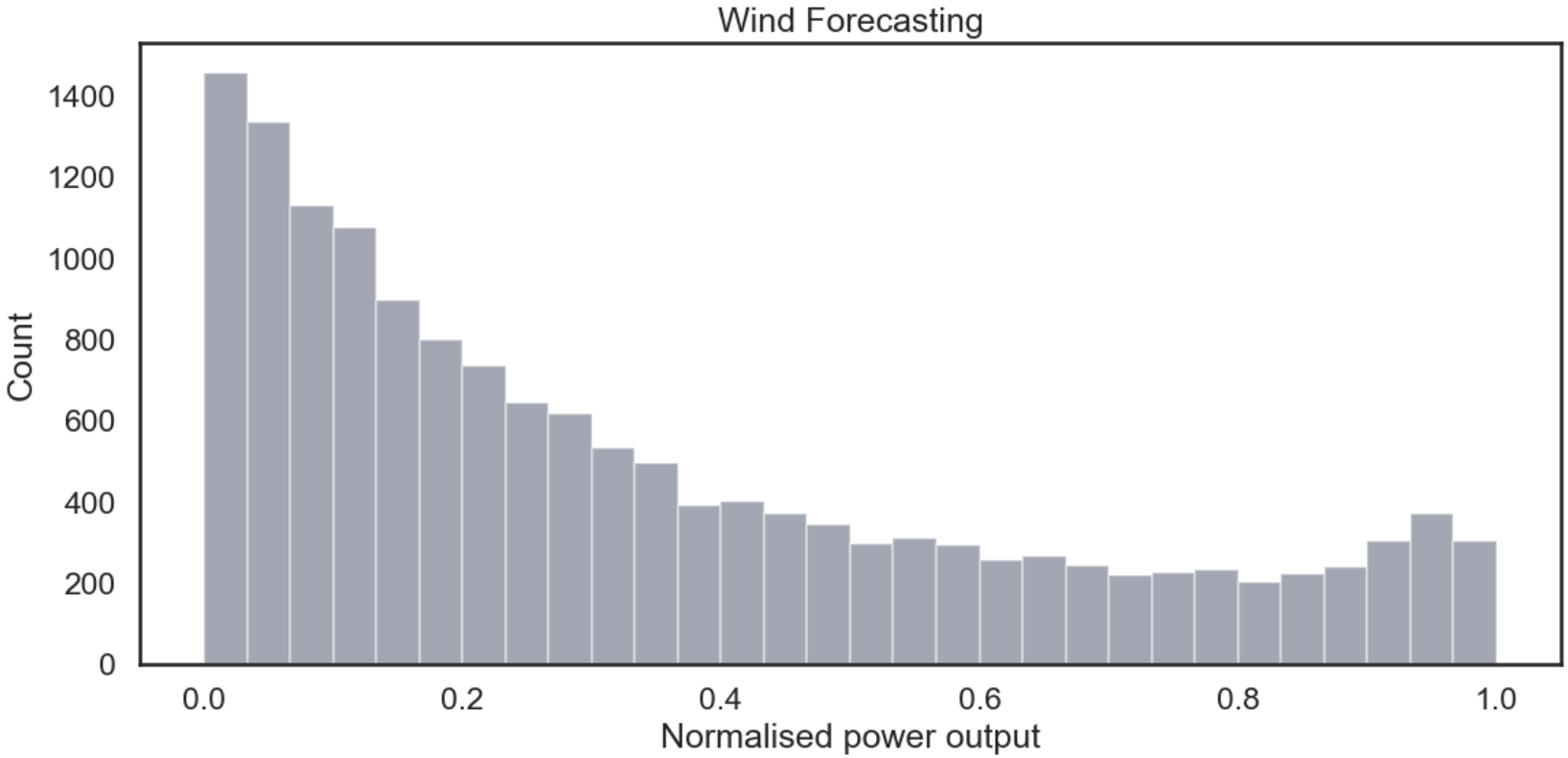
The vast majority of data cannot be represented by a mean. And the algorithm will not work.

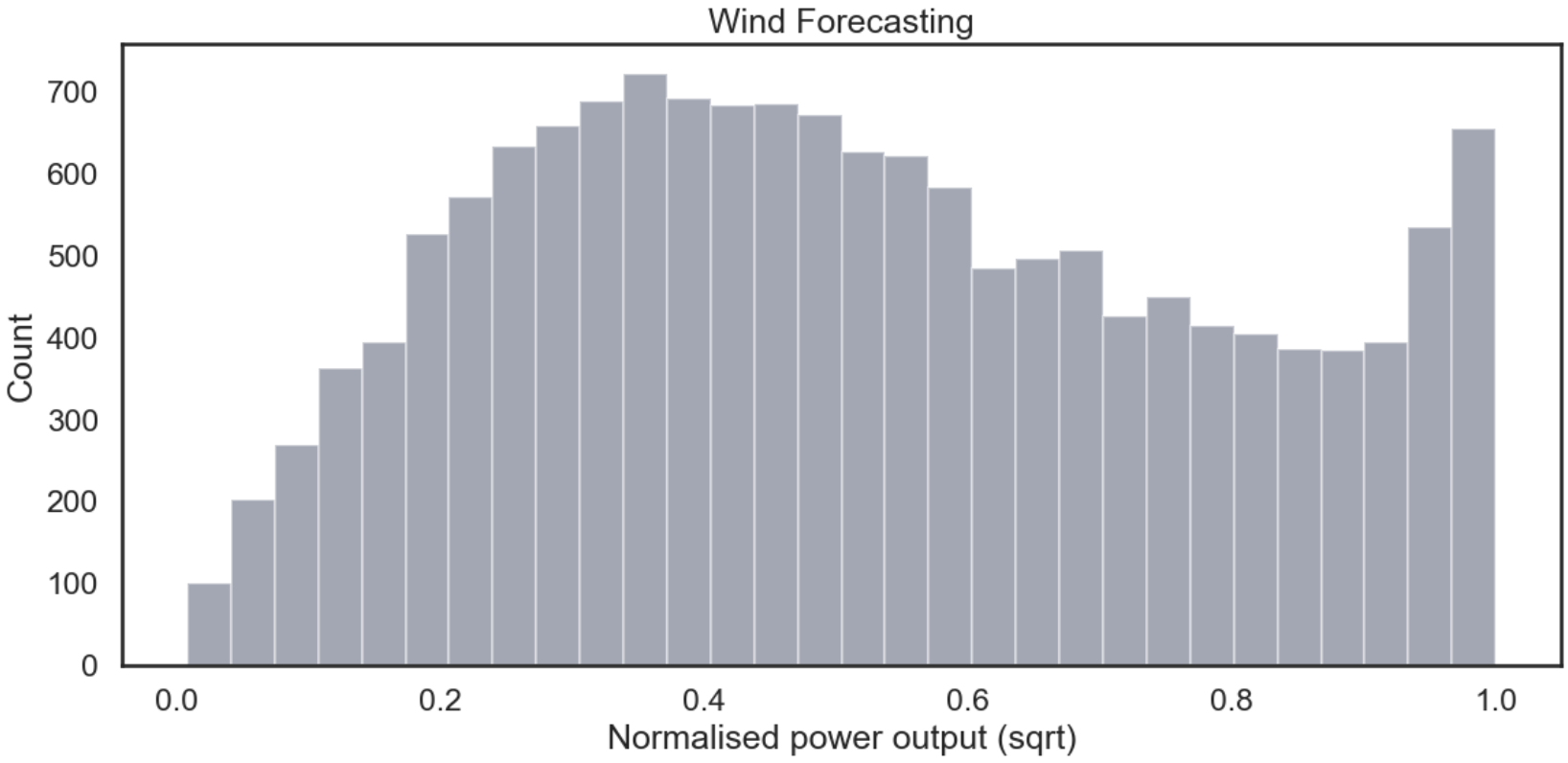
The best case...

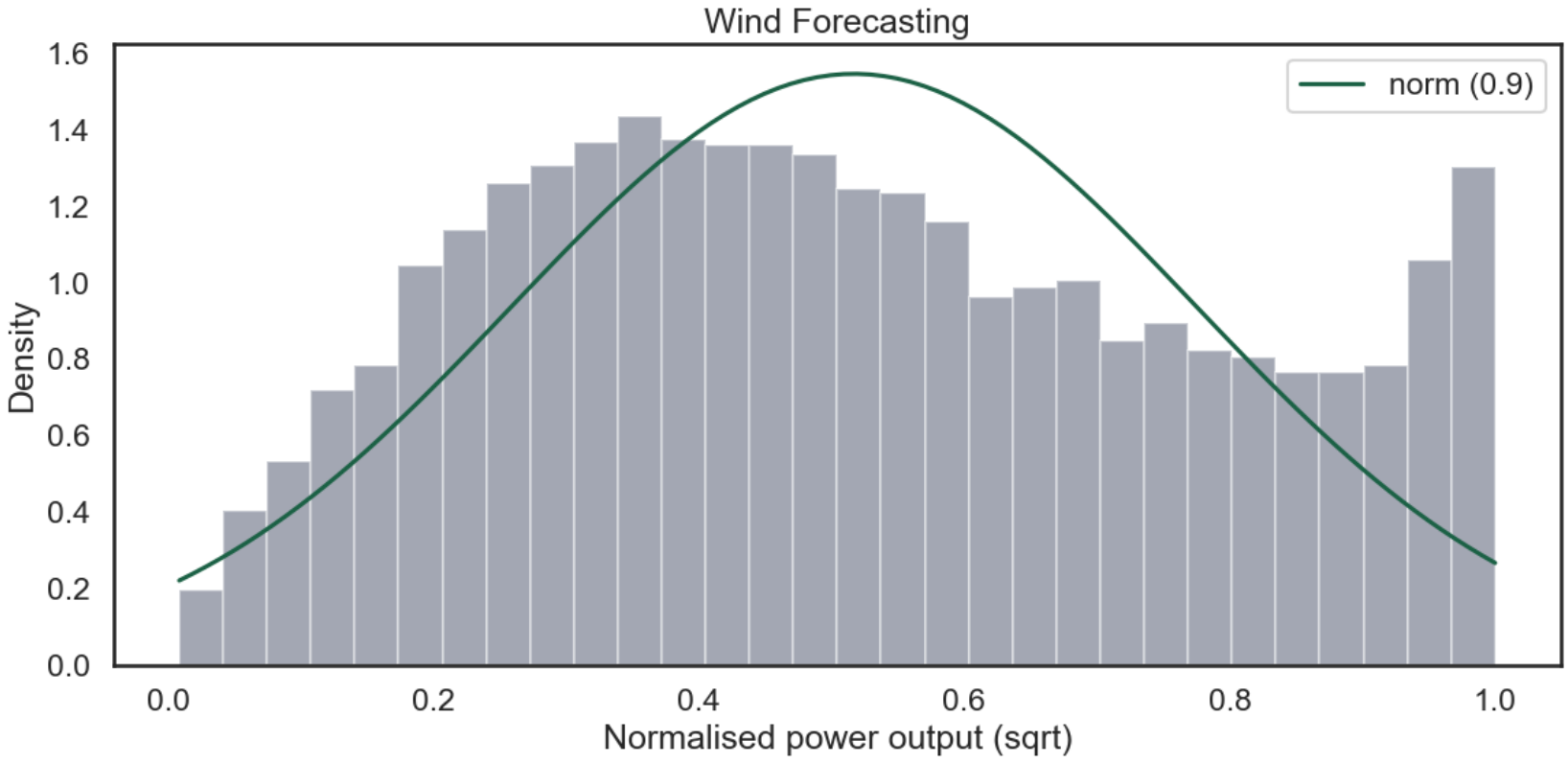

Star Cluster CYG OB1 Regression
log(Light)
6.0
5.5
5.0
4.5
4.0
3.6
3.8
4.0
4.2
4.4
4.6
log(Temperature)

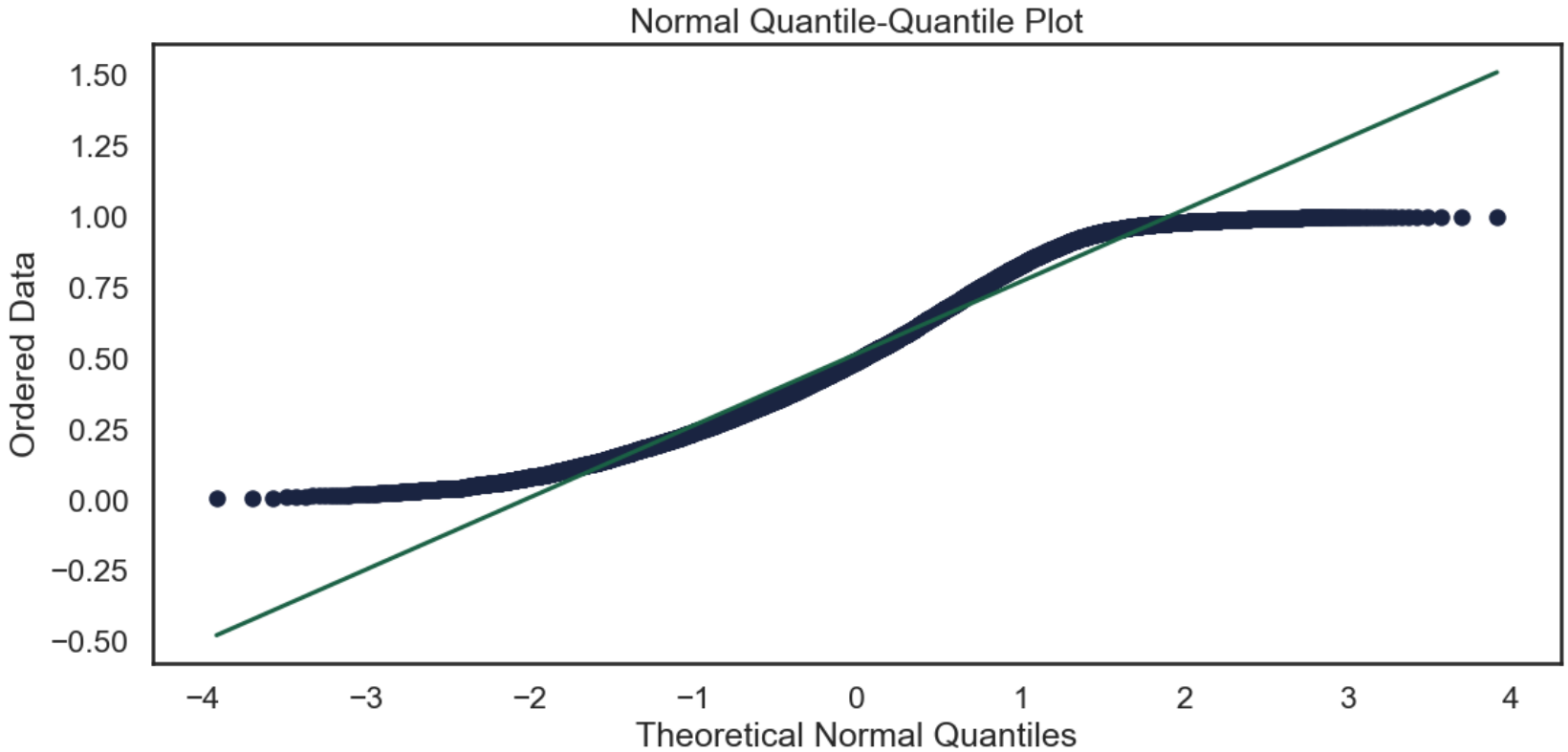

Titanic Survivors - Logistic Classification - 63.8%
survived
0
1
age
fare
0
20
40
60
80
0
100
200
300
400
500

FIXING: DOMAIN KNOWLEDGE




Wind Forecasting
Count
0
100
200
300
400
500
600
700
0.0
0.2
0.4
0.6
0.8
1.0
Normalised power output (sqrt)

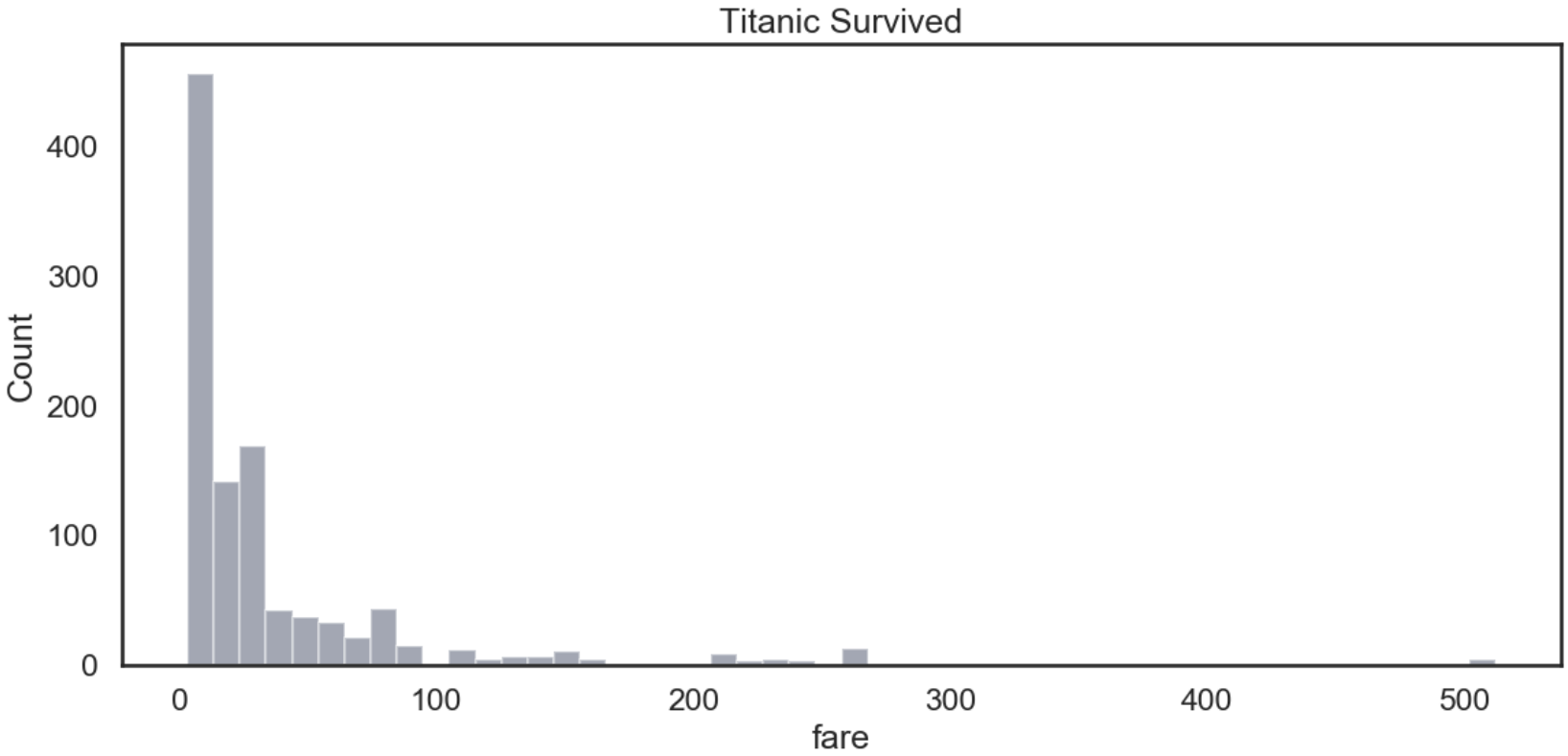

Wind Forecasting
norm (0.9)
Density
Normalised power output (sqrt)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6

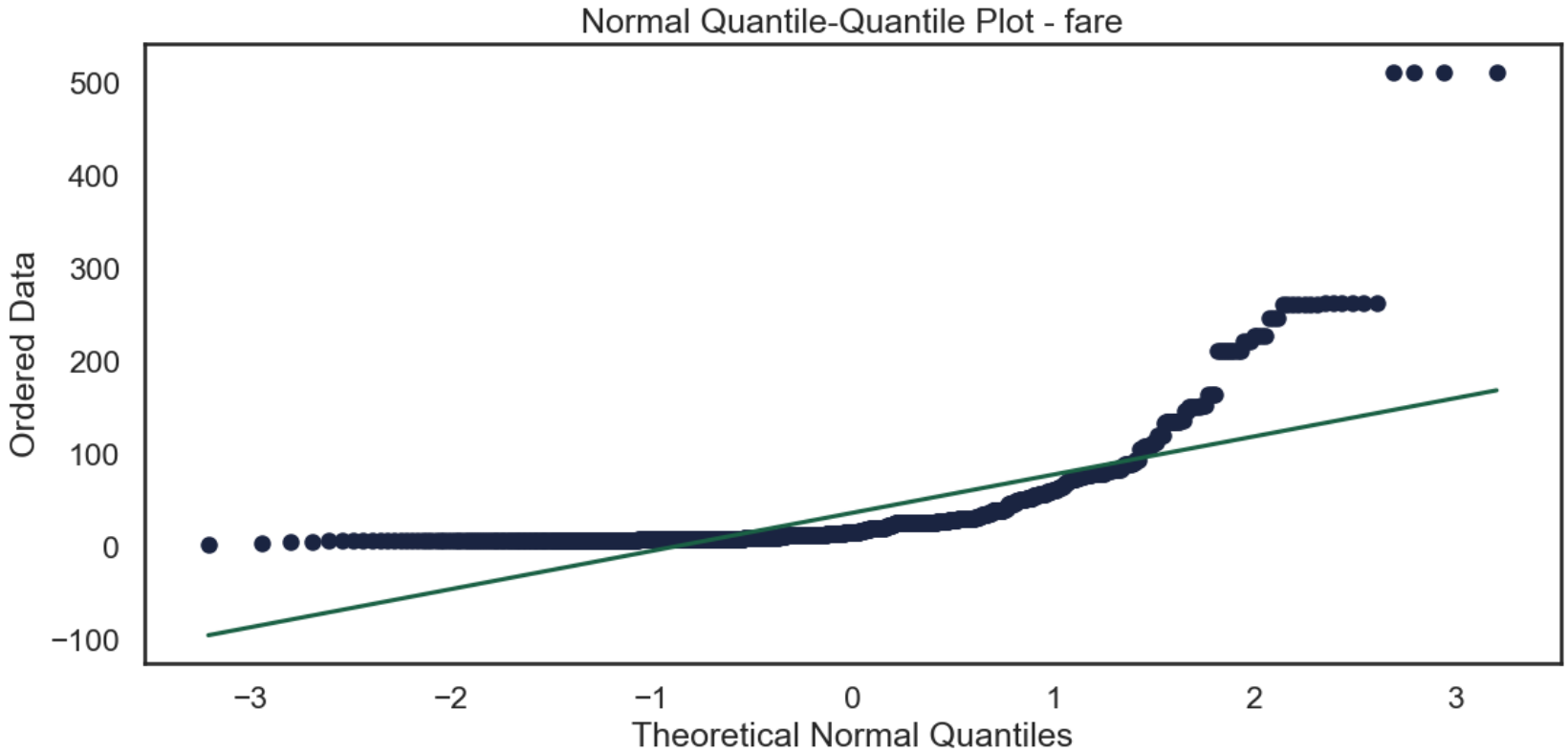

Normal Quantile-Quantile Plot
Ordered Data
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−4
−3
−2
−1
0
1
2
3
4
Theoretical Normal Quantiles

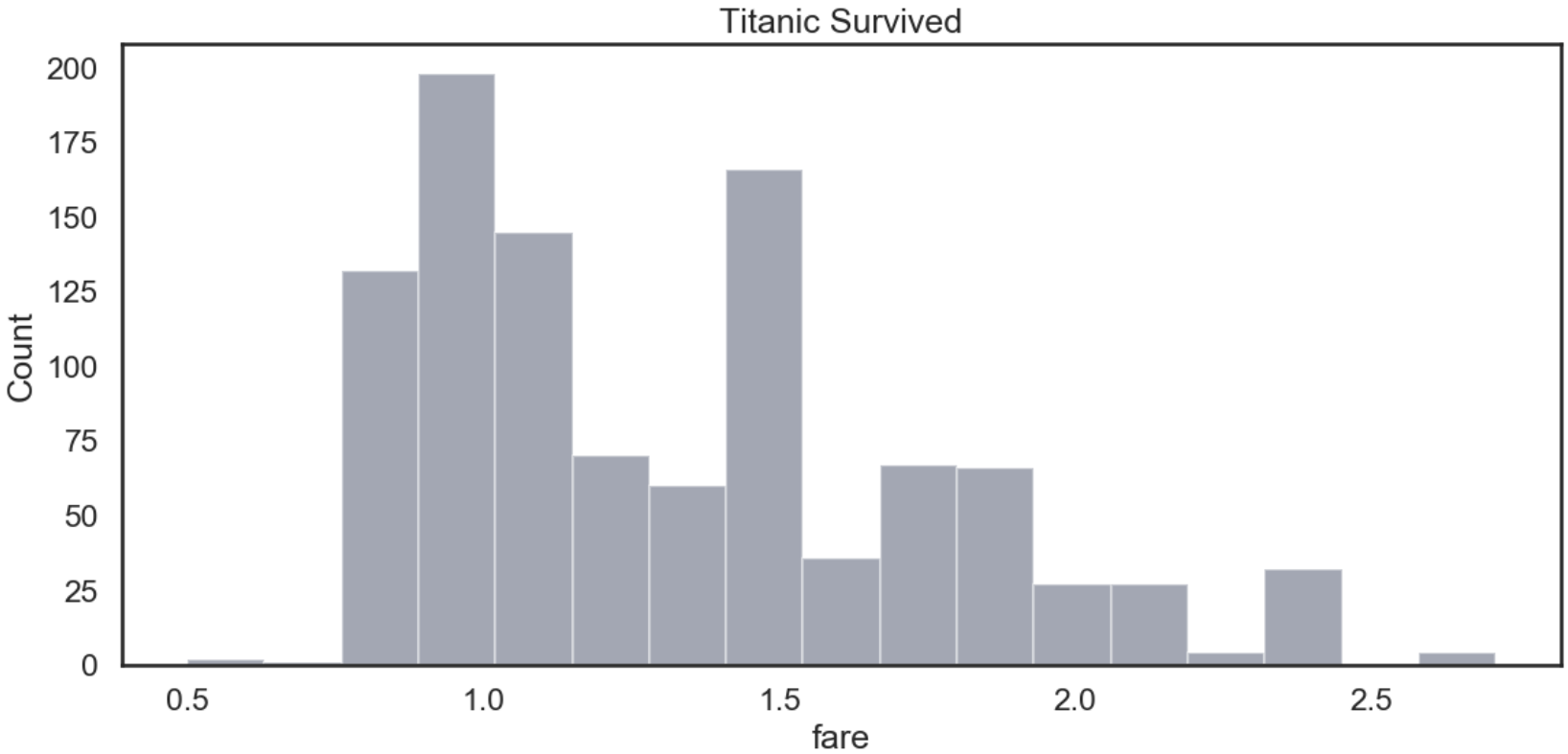
FIXING: ARBITRARY FUNCTIONS

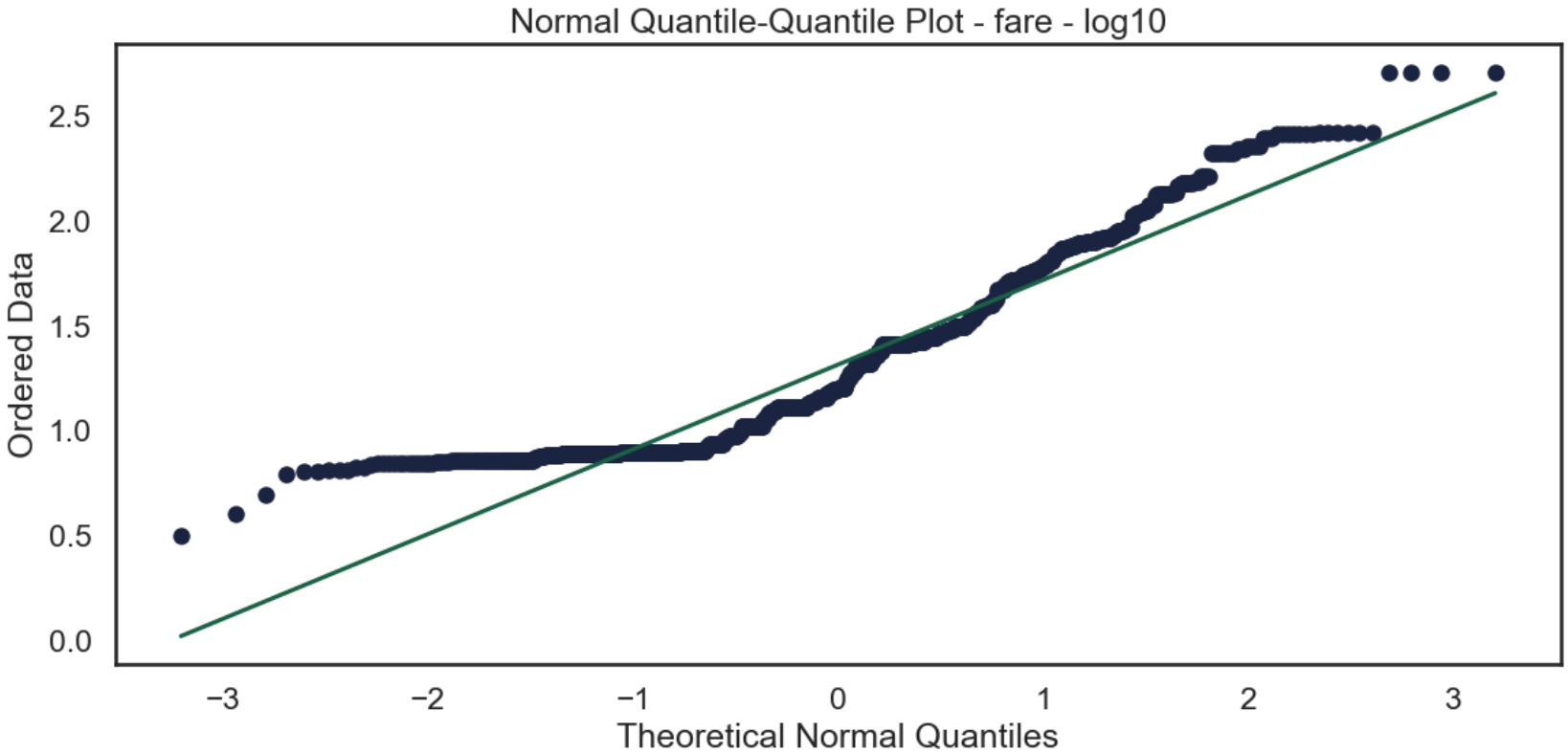
- We can use **any** mathematical function to transform our data*

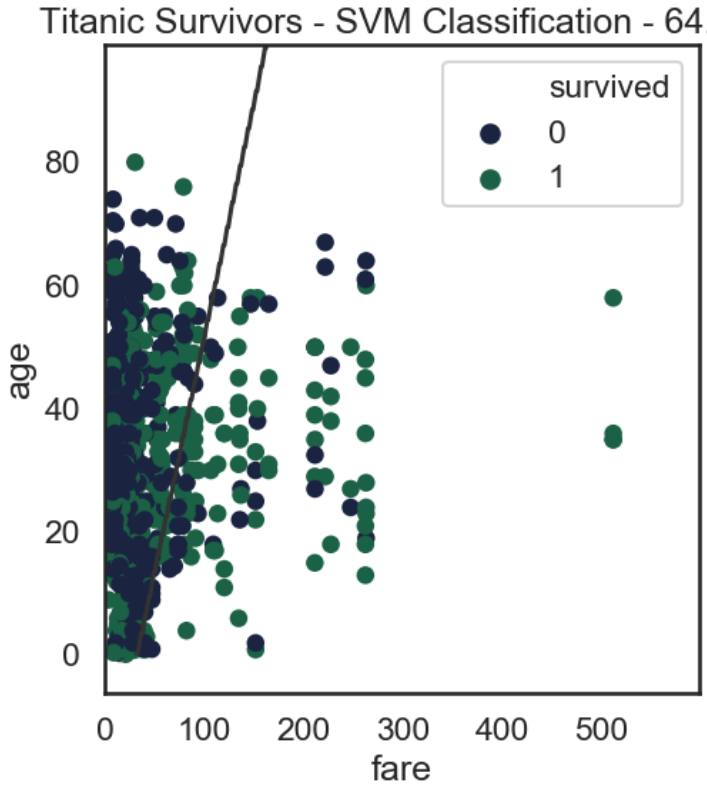
*so long as it's invertible

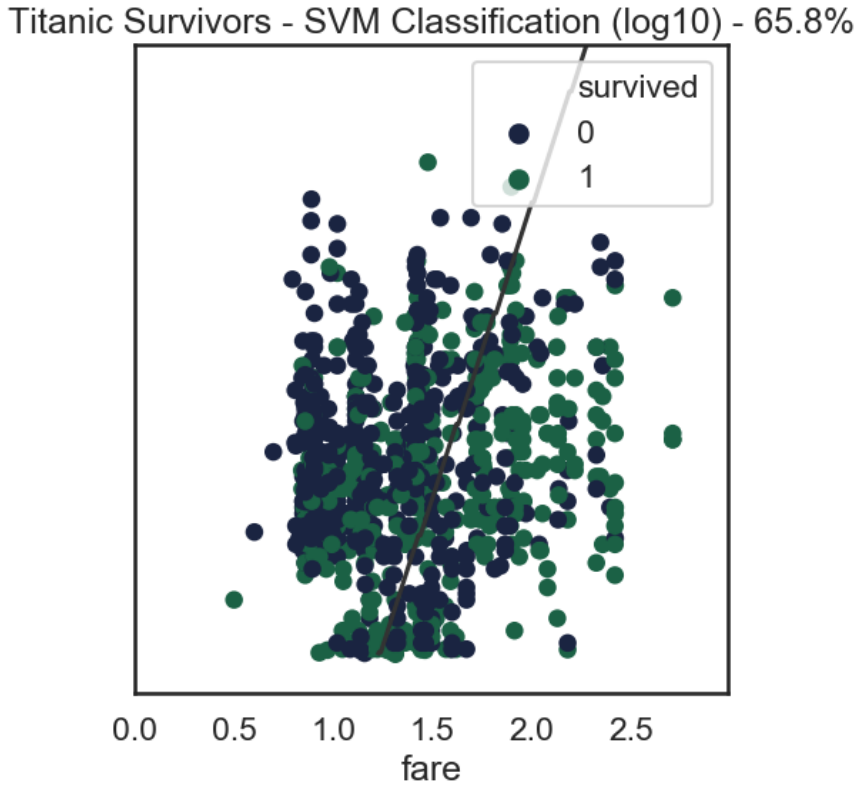

Titanic Survived
Count
400
300
200
100
0
0
100
200
300
400
500
fare


Normal Quantile-Quantile Plot - fare
Ordered Data
500
400
300
200
100
0
−100
−3
−2
−1
0
1
2
3
Theoretical Normal Quantiles


Titanic Survived
Count
200
175
150
125
100
75
50
25
0
0.5
1.0
1.5
2.0
2.5
fare


Normal Quantile-Quantile Plot - fare - log10
Ordered Data
2.5
2.0
1.5
1.0
0.5
0.0
−3
−2
−1
0
1
2
3
Theoretical Normal Quantiles


Titanic Survivors - SVM Classification - 64.0%
survived
0
1
80
60
40
20
0
age
0
100
200
300
400
500
fare


Titanic Survivors - SVM Classification (log10) - 65.8%
survived
0
1
0.0
0.5
1.0
1.5
2.0
2.5
fare

THINGS I'VE SKIPPED OVER

- Practical examples
- Windsorising
- Types of data
- Scaling
- Derived Data
- Box Cox transform
- Time series data
- Feature selection
- Dimensionality reduction
- Data integration
- Probably lots more!

CONCLUDING REMARKS

- Data Cleaning:
  - is important
  - is open to interpretation
  - is (arguably) a manual process
  - takes a lot of time (approx 60% of a Data Scientist time)
  - requires domain knowledge

Data Science **Training, Consultancy, Development**

@DrPhilWinder

DrPhilWinder

https://WinderResearch.com

phil@WinderResearch.com

BIBLIOGRAPHY

- Examples: https://www.reddit.com/r/MachineLearning/comme
- Book: Janert, P.K. Data Analysis with Open Source Tools: A Hands-On Guide for Programmers and Data Scientists. O'Reilly Media, 2010. https://amzn.to/2VFqOYx.
- Data Types in Statistics, Niklas Donges - https://towardsdatascience.com/data-types-in-statistics-347e152e8bee
- Quick intro to handling missing data: https://towardsdatascience.com/the-tale-of-missing-values-in-python-c96beb0e8a9d

- Pandas documentation on missing data: https://pandas.pydata.org/pandas-docs/stable/missing_data.html
- Bit more information about anomaly detection: https://towardsdatascience.com/a-note-about-finding-anomalies-f9cedee38f0b
- Good short free book on anomaly detection: Practic Machine Learning: A New Look at Anomaly Detectio Ted Dunning, Ellen Friedman, O'Reilly Media, Inc., 2014, ISBN 1491914181, 9781491914182
- Cool Library for benchmarking time series anomaly detection: https://github.com/numenta/NAB
- Nice run through of day-to-day problems with data: https://medium.com/@bertil_hatt/what-does-bad-data-look-like-91dc2a7bcb7a

- Short section on dealing with corrupted data - Raschka, S. Python Machine Learning. Packt Publishing, 2015. https://books.google.co.uk/books?id=GOVOCwAAQBAJ.
- Presentation on Seaborn Styles - https://s3.amazonaws.com/assets.datacamp.com/p
- Code to fit all distributions: https://stackoverflow.com/questions/6620471/fittin empirical-distribution-to-theoretical-ones-with-scipy python