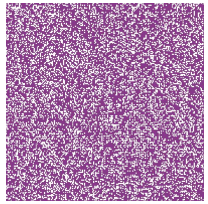
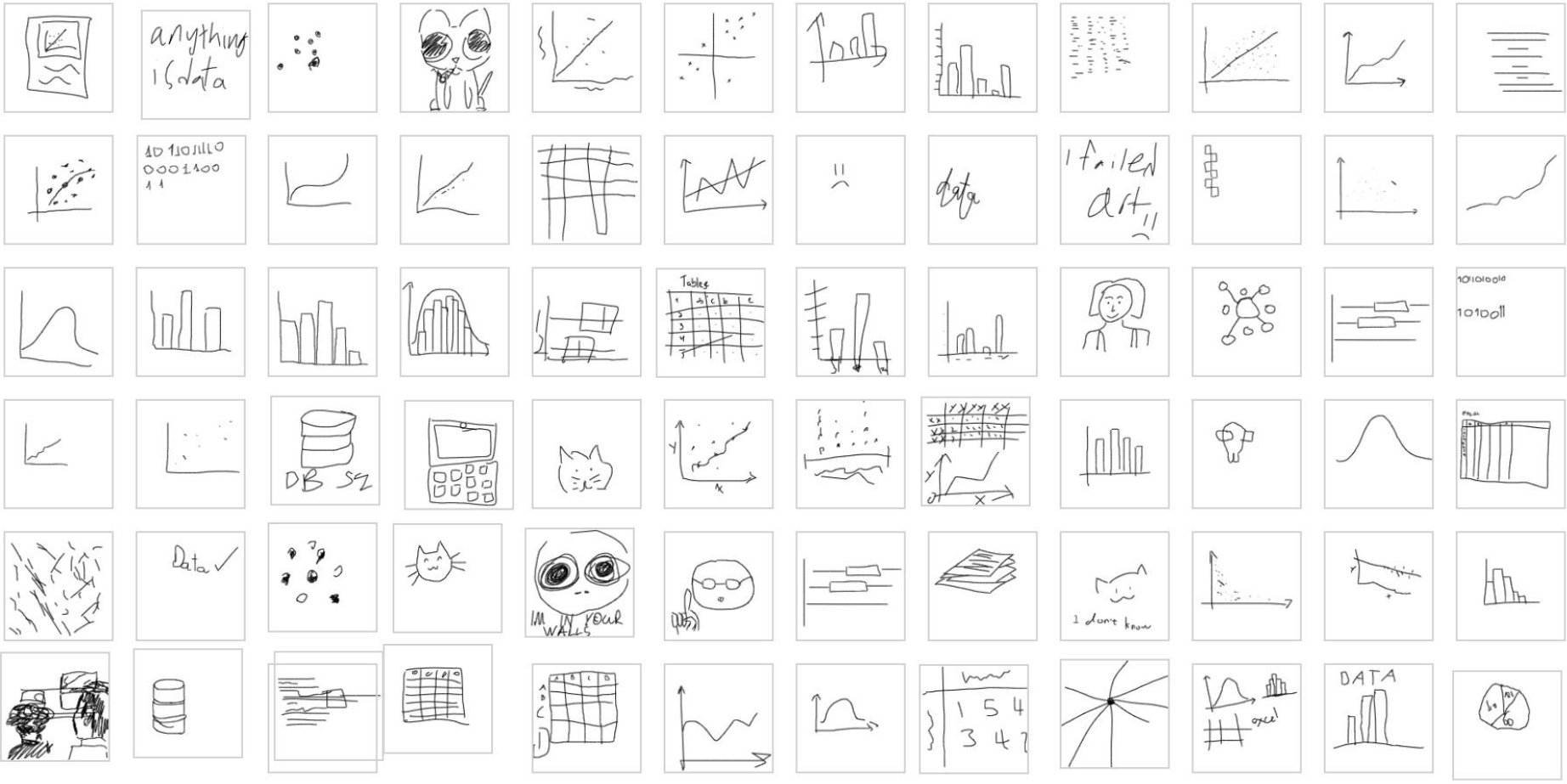
From *sketchy intuitions* to *imperfect rules*

Using digital image data from drawings to introduce informal classification models

Dr Anna Fergusson
Department of Statistics | Te Kura Tatauranga
University of Auckland | Waipapa Taumata Rau
Aotearoa New Zealand
a.fergusson@auckland.ac.nz | @annafergussonnz

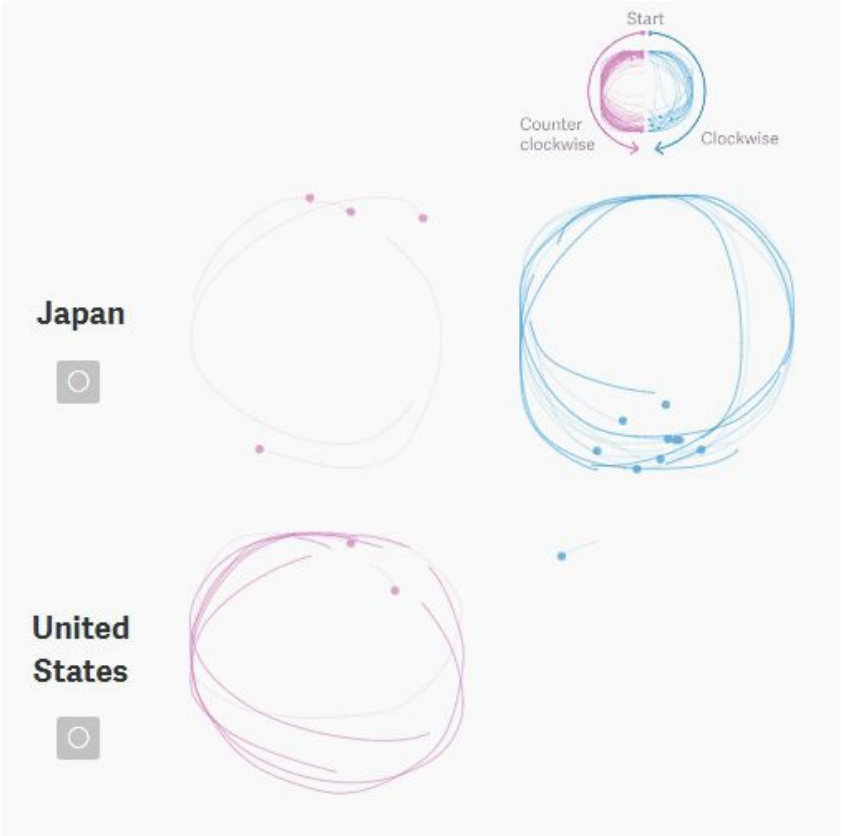


Drawings of “data” from my second stage/year data science students



What can we learn about humans based on how they sketch things?

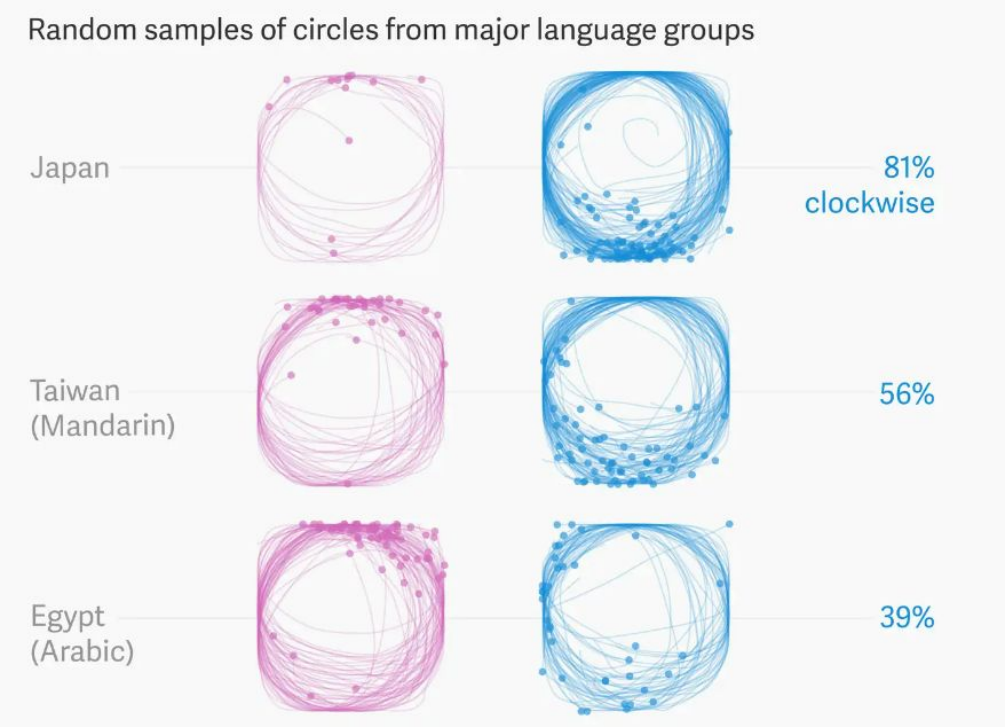
How you draw circles could be shaped by culture

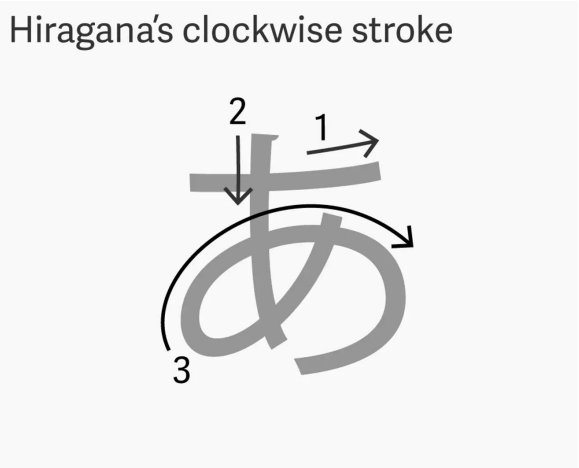


“Americans tend to draw circles counterclockwise. Of nearly 50,000 circles drawn in the US, 86% were drawn this way. People in Japan, on the other hand, tend to draw circles in the opposite direction. Of 800 circles drawn in Japan, 80% went clockwise.”

Example based on analysis and visualisations by Thu-Huong Ha & Nikhil Sonnad:
https://qz.com/994486/the-way-you-draw-circles-says-a-lot-about-you

More comparisons by country (and assumed language)





"Both Japanese and Chinese scripts follow a strict stroke order. On the whole, characters are drawn from top left in the direction of the bottom right."

Example based on analysis and visualisations by Thu-Huong Ha & Nikhil Sonnad: https://qz.com/994486/the-way-you-draw-circles-says-a-lot-about-you

Other cool things we can learn from sketches/drawings!

Draw a scientist

Source: time.com/5201175/draw-a-scientist-studies/

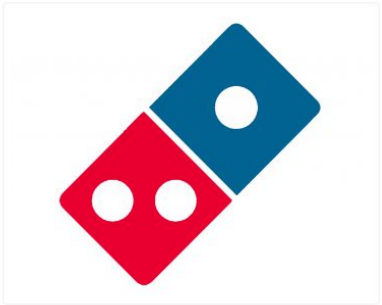
Draw a computing student

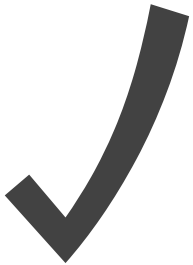
Source: https://dl.acm.org/doi/10.1145/3545945.3569795

A study by University of Auckland researchers (Varoy, Lee, Luxton-Reilly & Giacman) explored drawings made by teachers to gain insight into the characteristics that teachers look for in computing students, and to prompt discussion with the teacher about how these characteristics or related teachings are promoted in their classrooms.

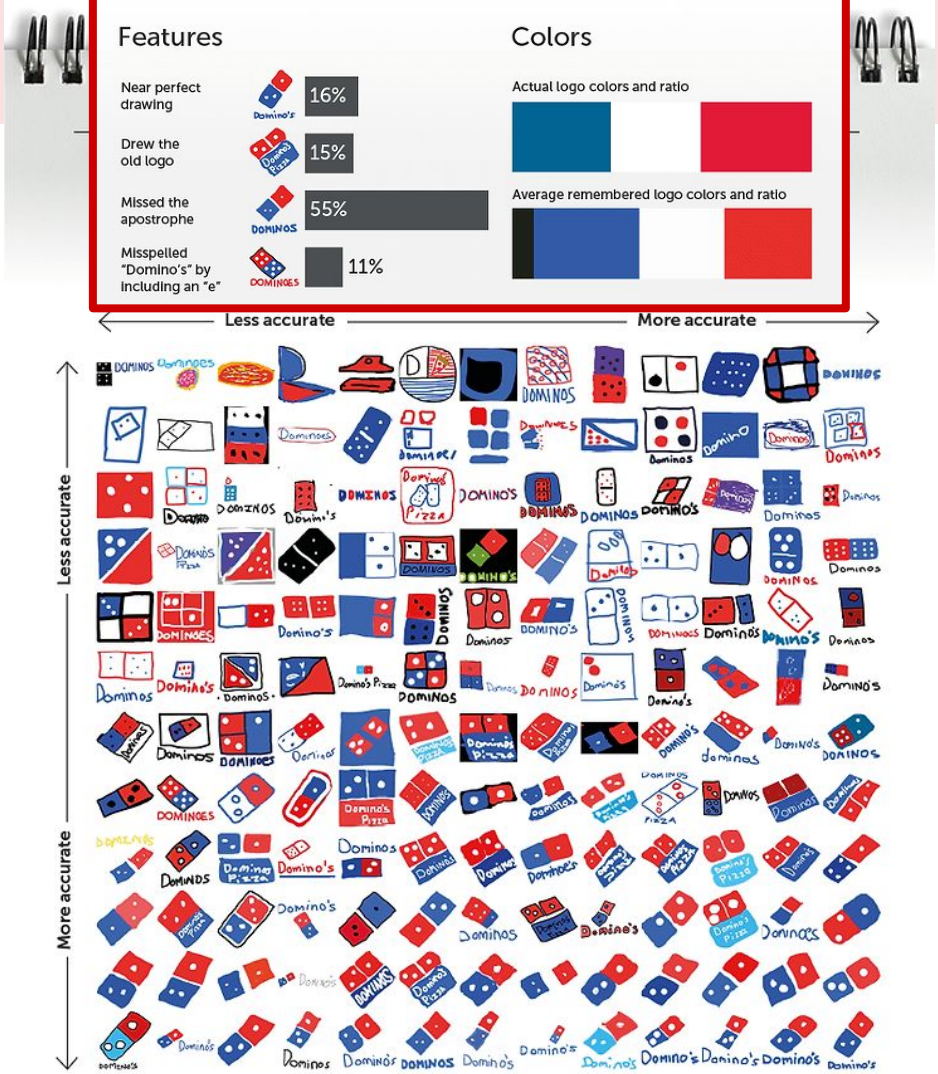
Brand or logo memory (recognition)

Which one is the correct logo for dominoes?

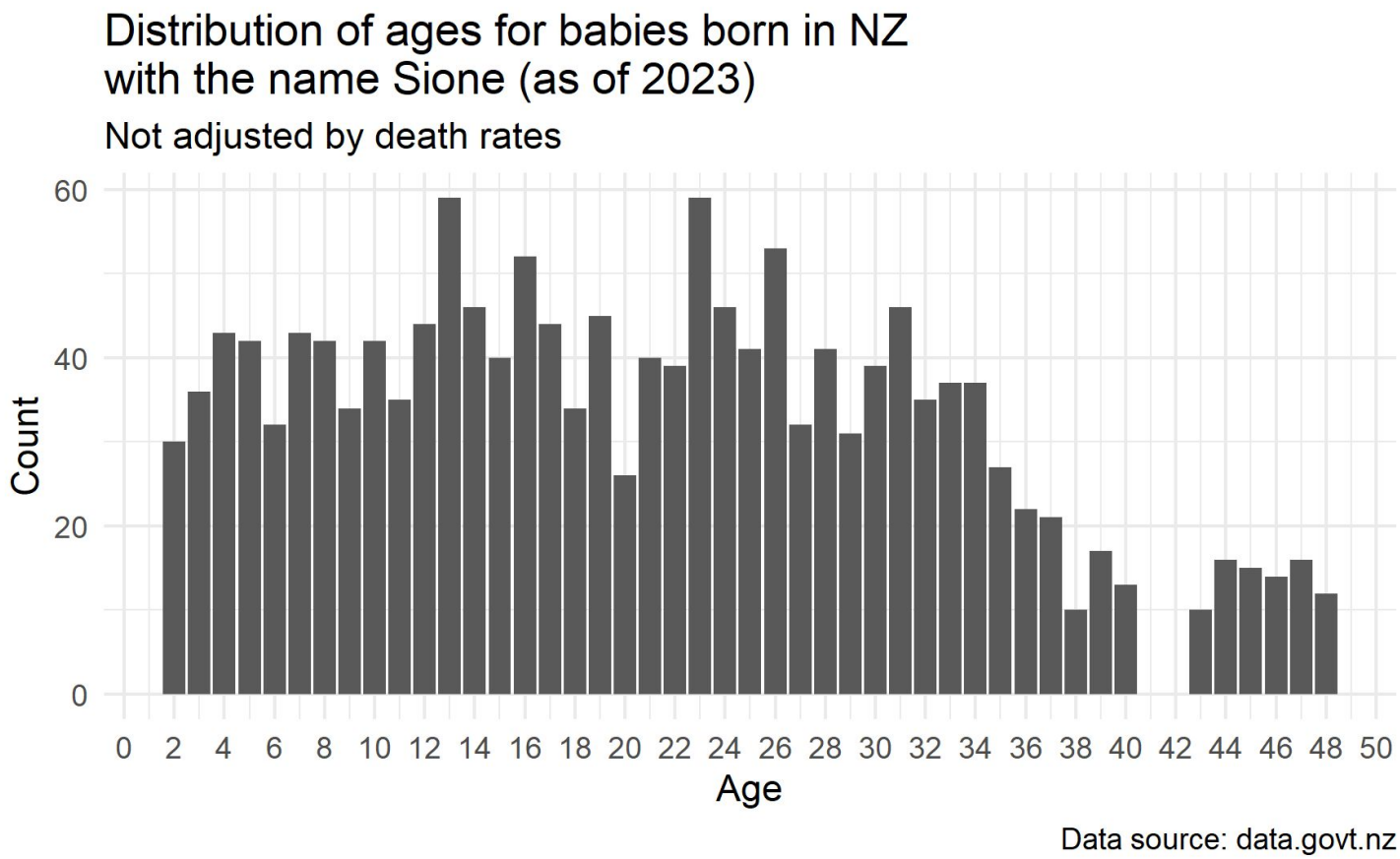




Source:
https://www.signs.com/branded-in-memory/



Using drawings or sketches to help teach statistics and data science



Do we all see the same shape?

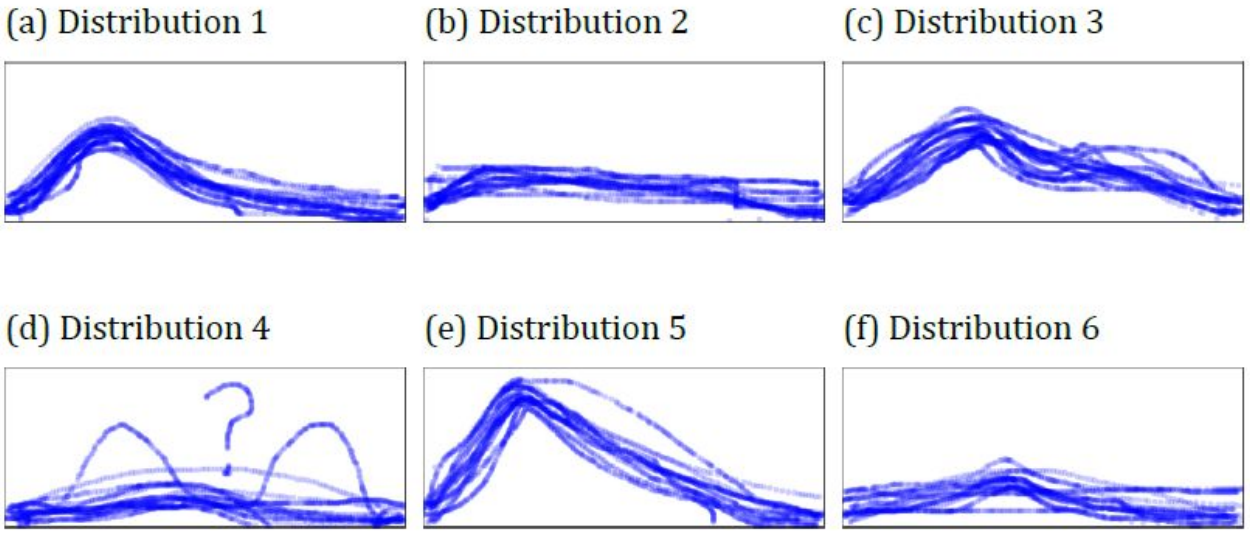
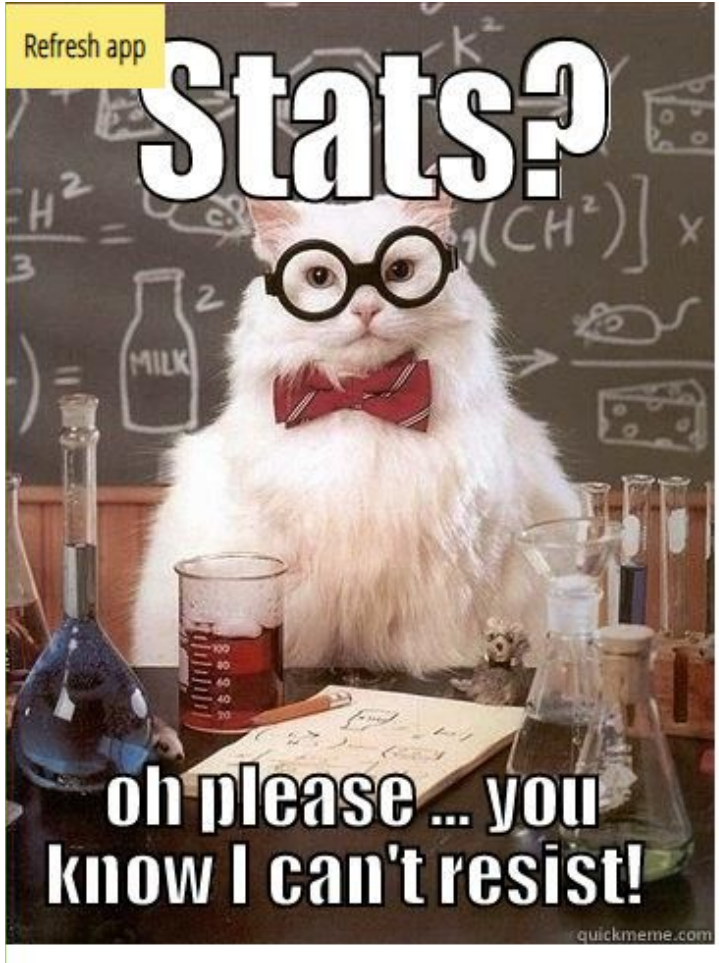


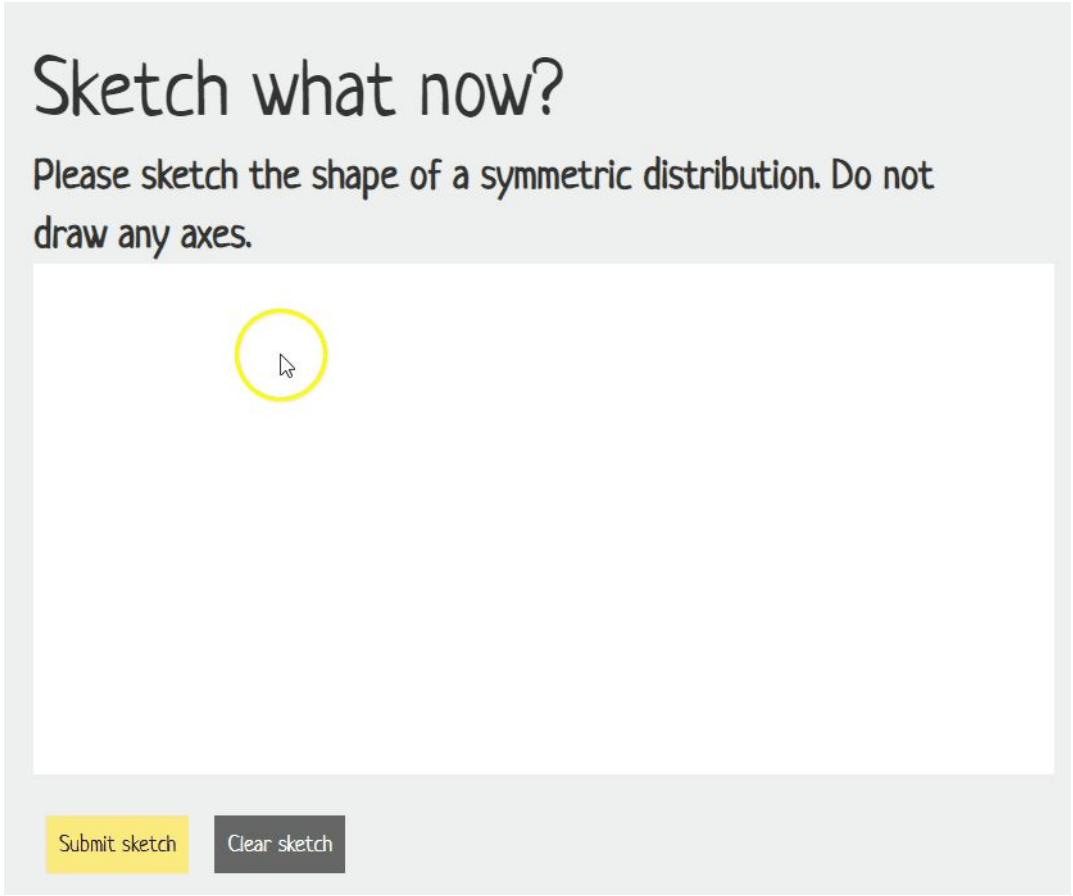
Figure 25: Task C shape sketches for each distribution produced by combining the sketches from all 12 teachers and overlaid on the same plot without the reference distribution

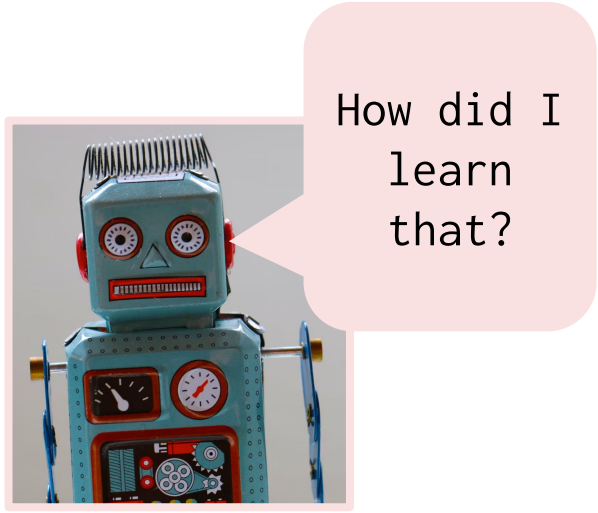
Source: researchspace.auckland.ac.nz/handle/2292/36909



bit.ly/sketchymodels

Using AI to help students learn how to sketch shapes of distributions



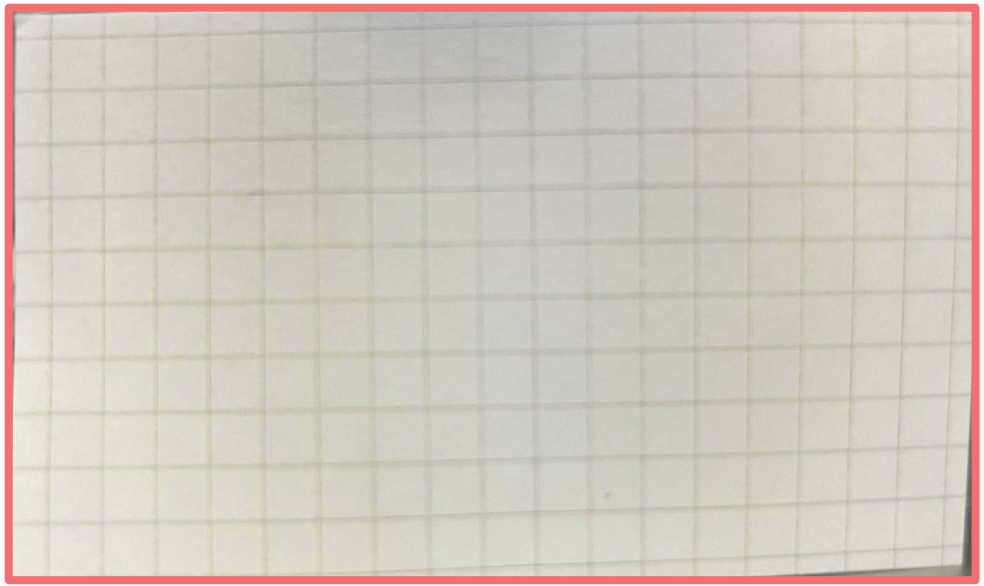


https://annafergusson.online/sketchy/

What else do your drawings reveal?

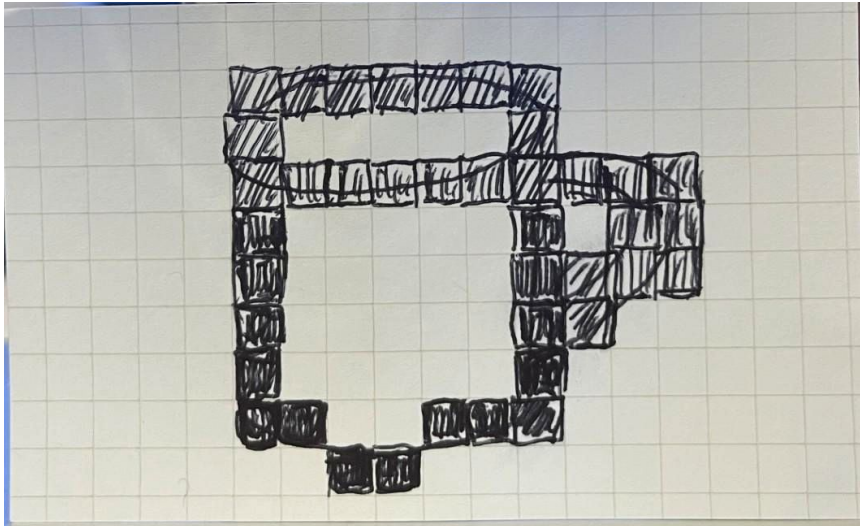
Draw what now?

Mark off a rectangle 18 squares wide and 11 squares high
The rectangle should be landscape as shown below

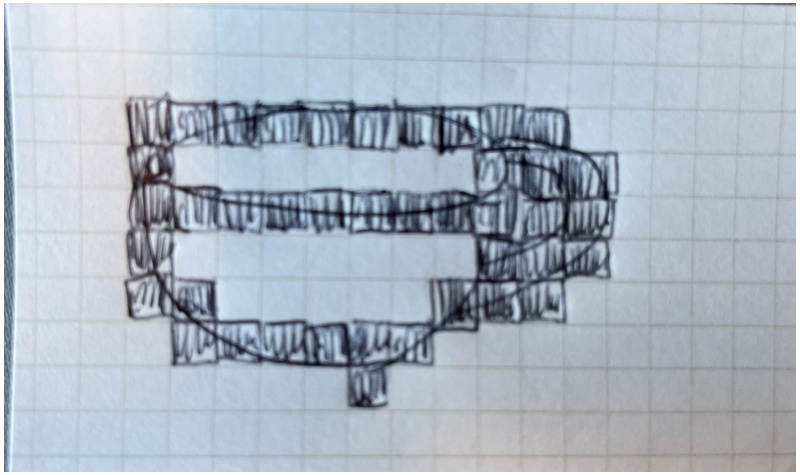


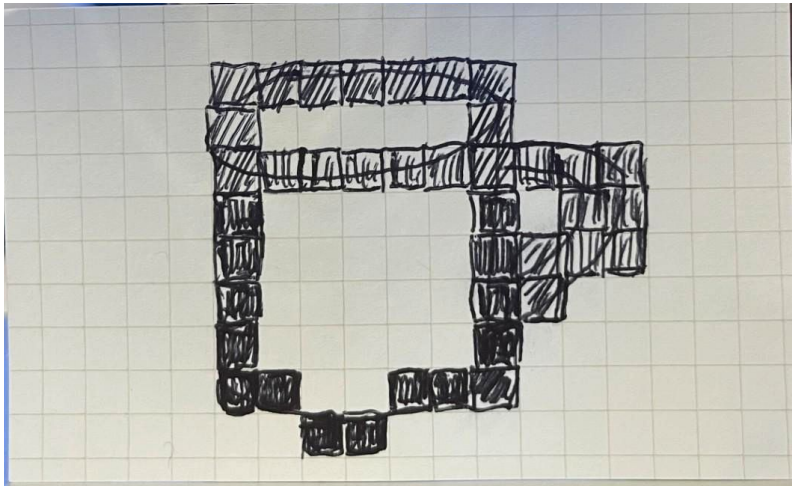
Draw a mug e.g. something you would drink something from!

From lines to pixels



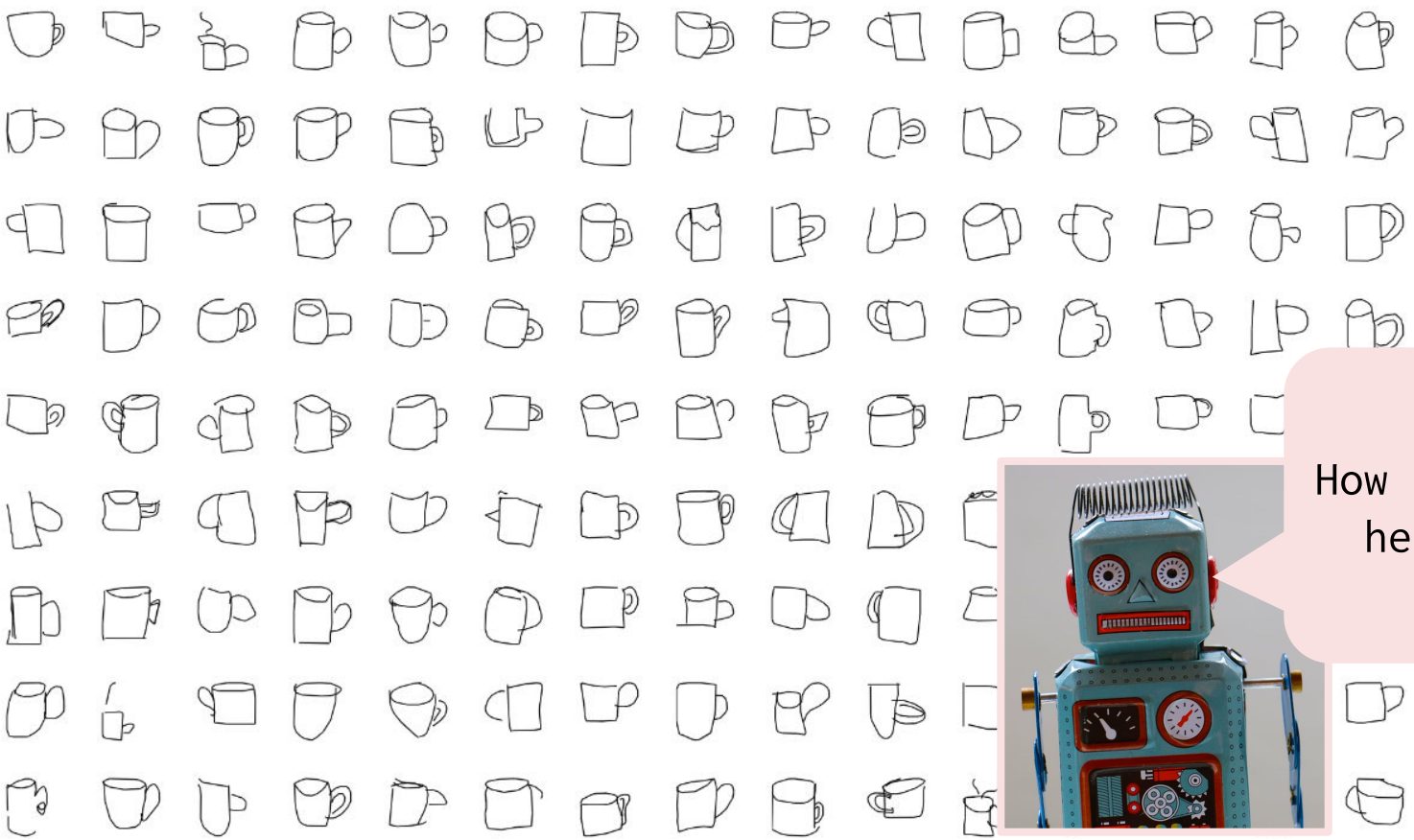
Shade in any squares that your lines in your drawings cross





What is similar about your drawings? What is different?

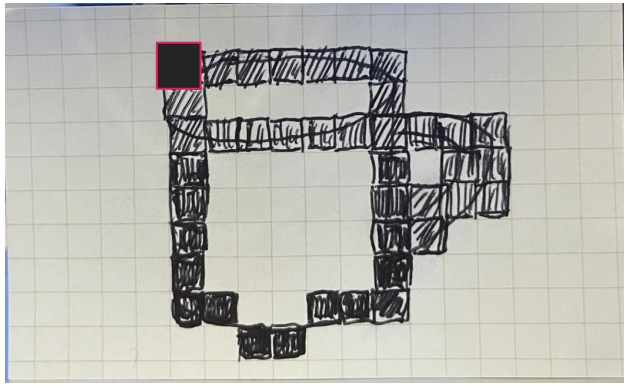
You are looking at 140,649 mug drawings made by real people... on the internet.

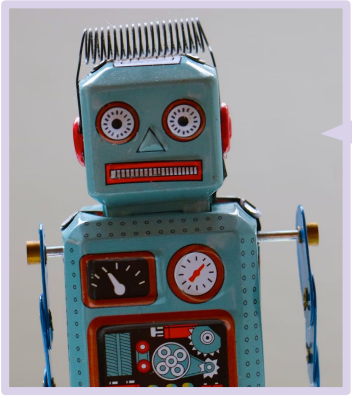


How can I help?

Source: https://quickdraw.withgoogle.com/data/mug

Computational approaches with pixels





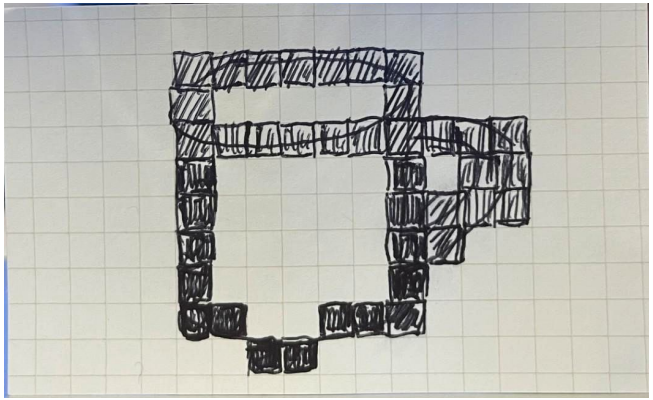
I know the cards are 18 pixels by 11 pixels (ish)

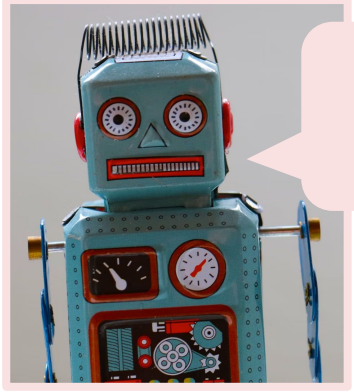
I know which pixels are black and where they are on the card

For example, the highlighted black pixel is on the third row and in the sixth column

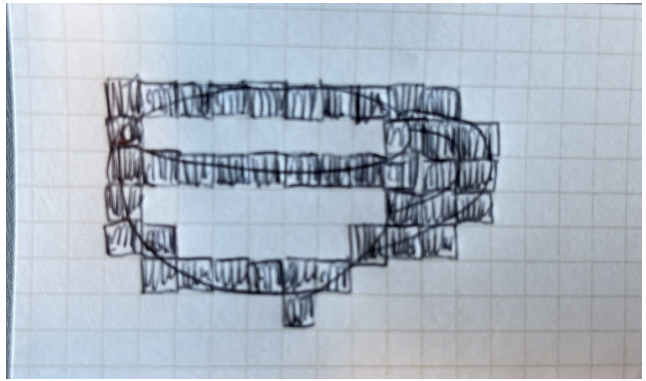
What would you like me to calculate using the black pixels?

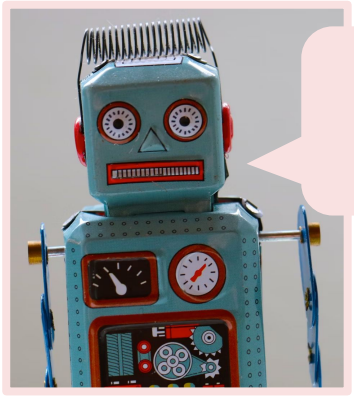
From pixels to variables





40





41

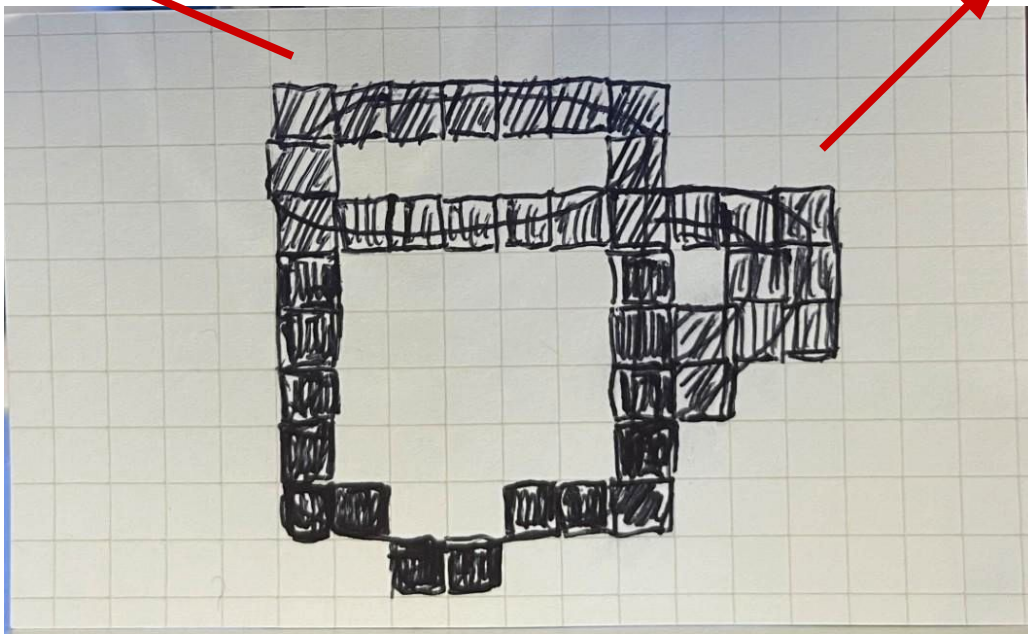
`total_pixels`

Count how many of the pixels are black

What other variables could we create using pixels?

An example of thinking computationally

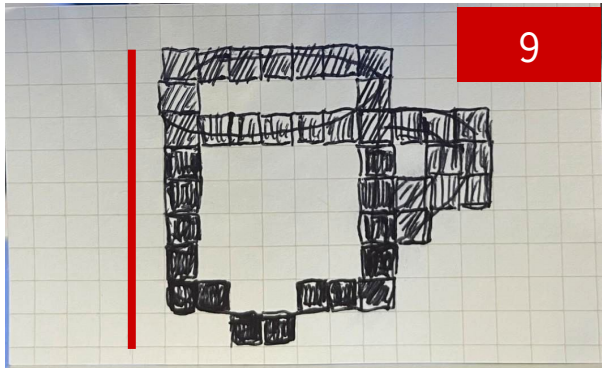
The leftmost pixel is in column 6

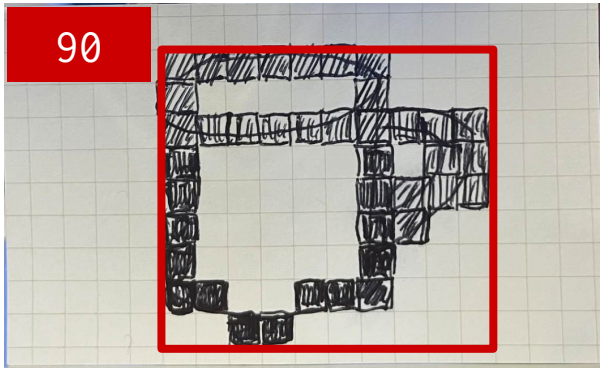


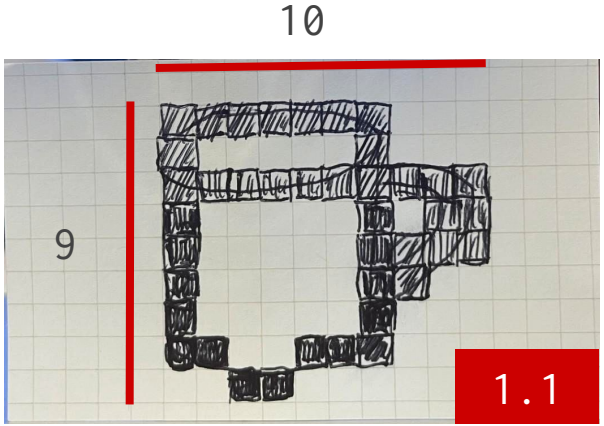
The rightmost pixel is in column 15

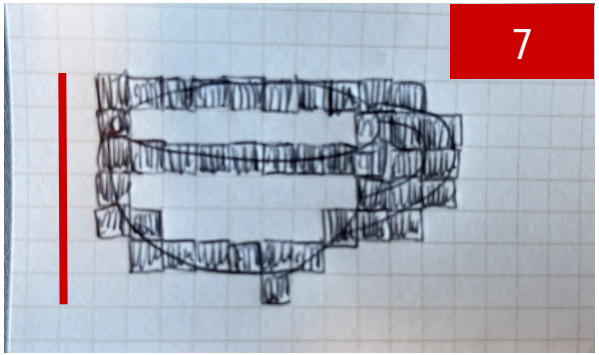
The width of the drawing, in pixels, is 10

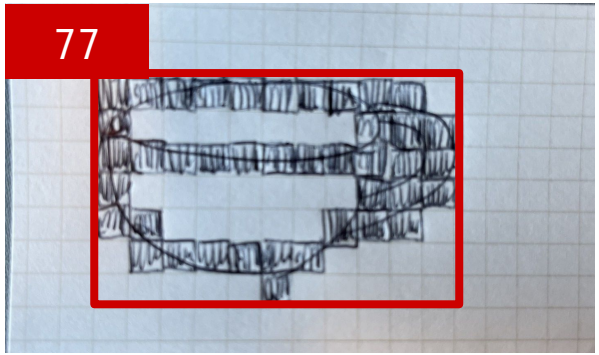
What could these variables be? How were the values calculated?

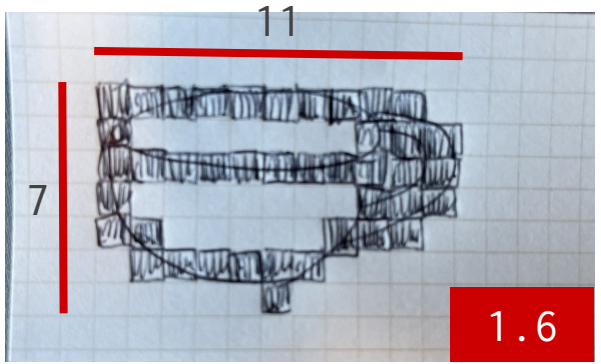












AI note: What is feature engineering?

Feature engineering for classification models involves creatively combining contextual and statistical knowledge to identify which data transformations and input variables (features) are predictive of the target variable, **so that the model performance can be improved**.

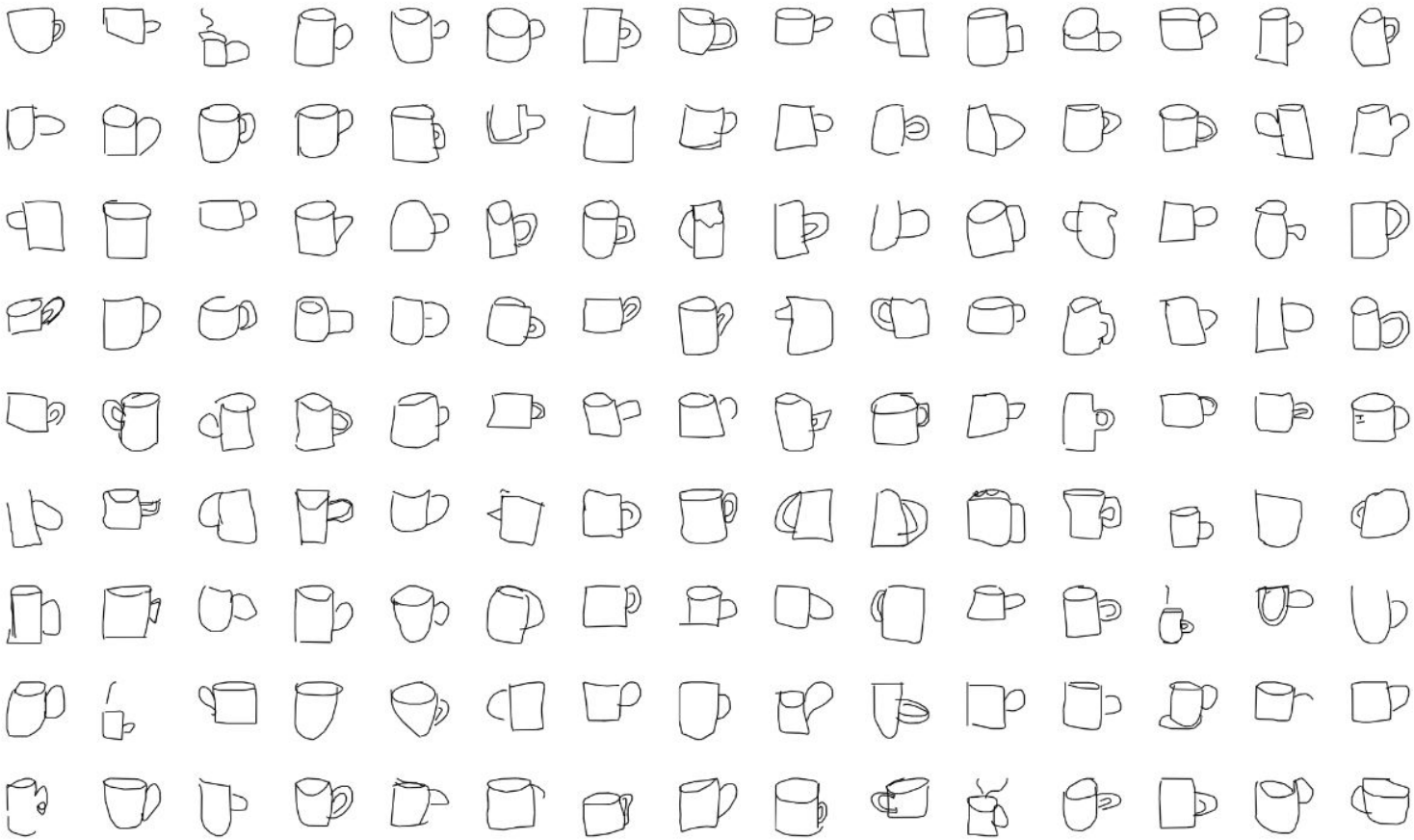
To know what how to transform the data and/or create new input variables from the pixelated drawings, we need to know the context for the model: **WHAT are we trying to predict - what is our goal?**

Left right out?

Do our line drawings reveal the existence of “right hand bias”?

If we had access to a very large number of digital images of drawings of mugs, could we develop an algorithm to sort the drawings into “right hand” or “not right hand”?

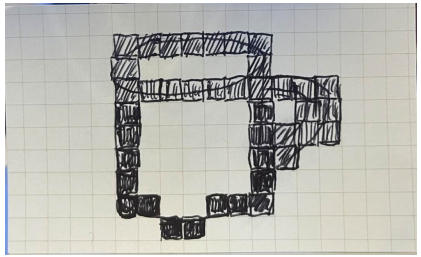
You are looking at 140,649 mug drawings made by real people... on the internet.



Source: https://quickdraw.withgoogle.com/data/mug

Let’s use our drawings as training data to try to develop an algorithm

drawing → input variable → decision rule



?

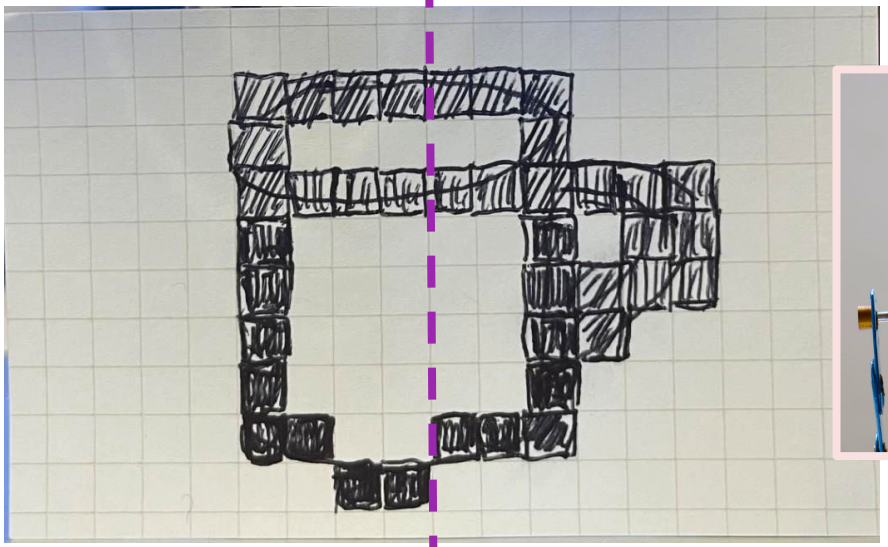
```
if _____
then _____
else _____
```

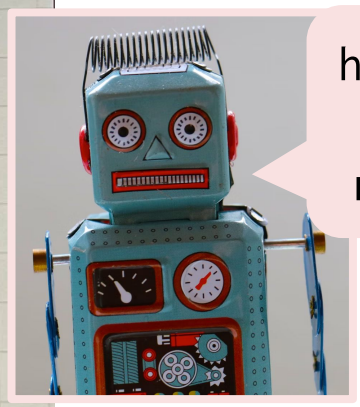
Are there any features of the drawings that make it difficult to develop a good classification model?

One approach: counting left vs right pixels

left_pixels: 17

right_pixels: 23



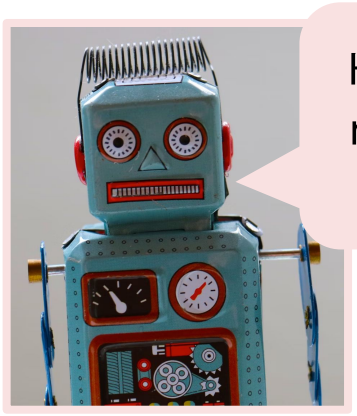


Are there any features of the drawings that interfere with using this calculation to identify “right” handles?

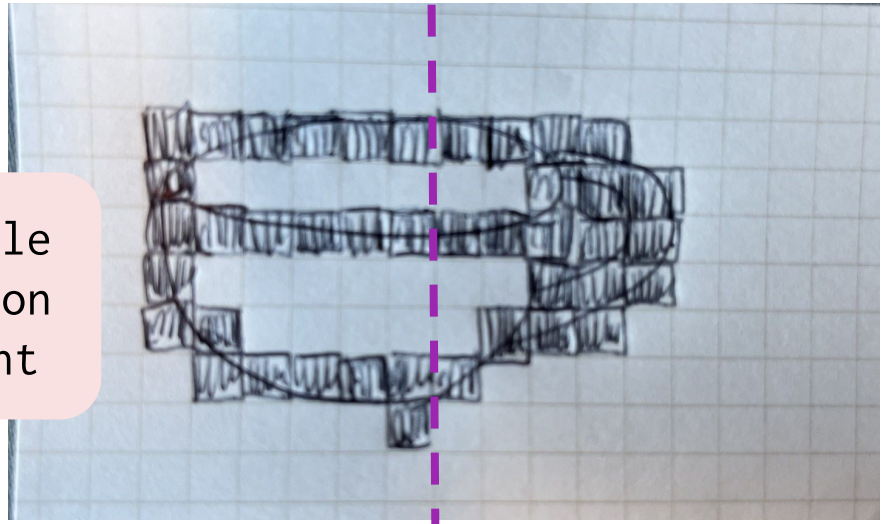
Not right!

left_pixels: 22

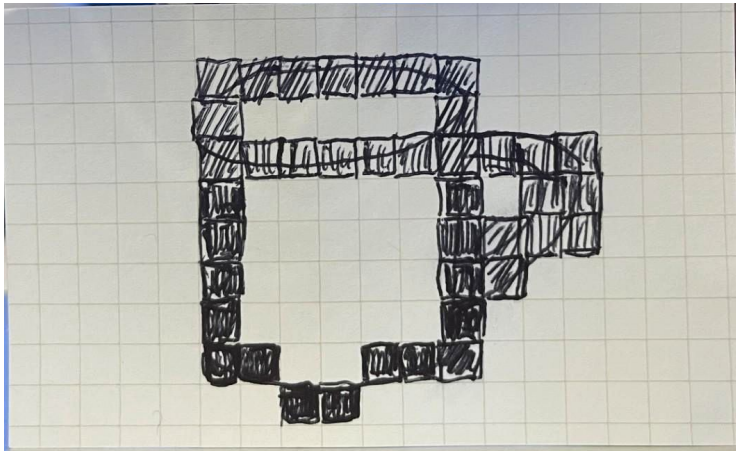
right_pixels: 19

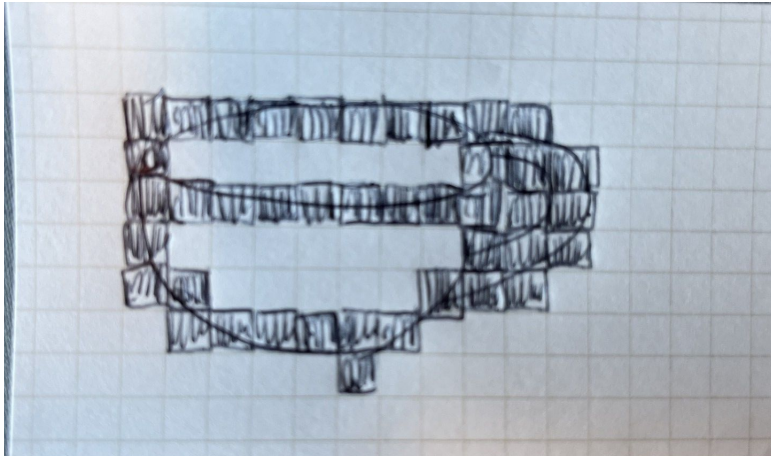


Handle not on right



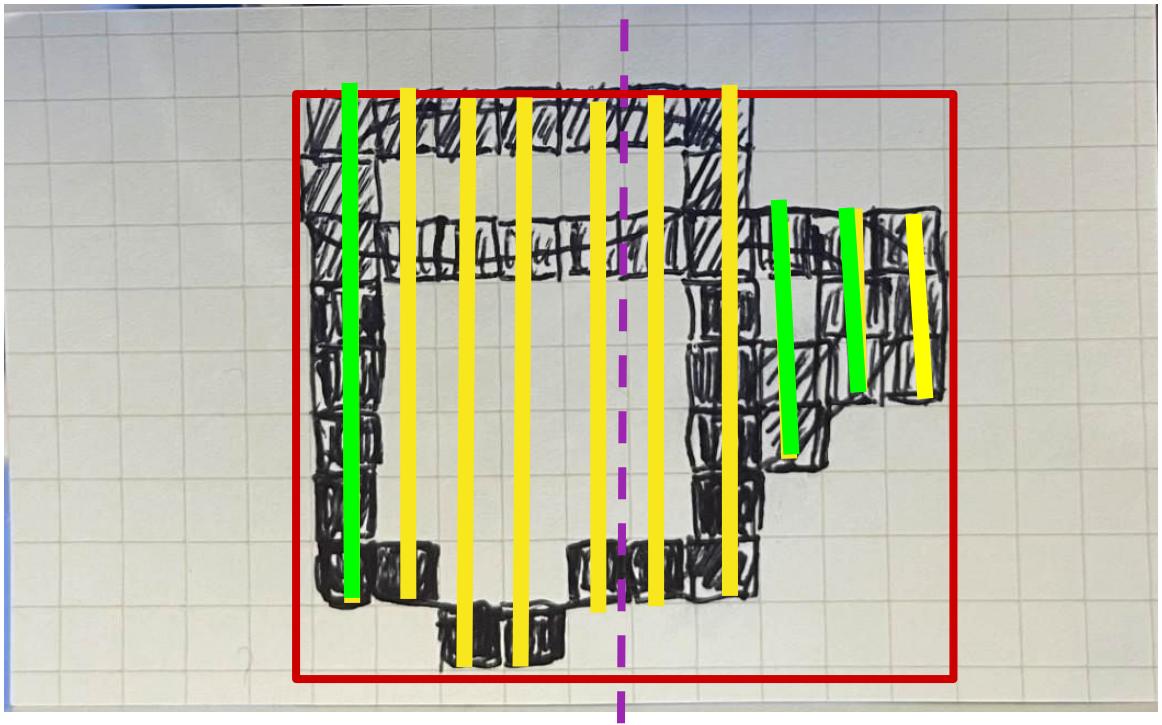
Challenge: Can you devise a better approach to identify right handles?

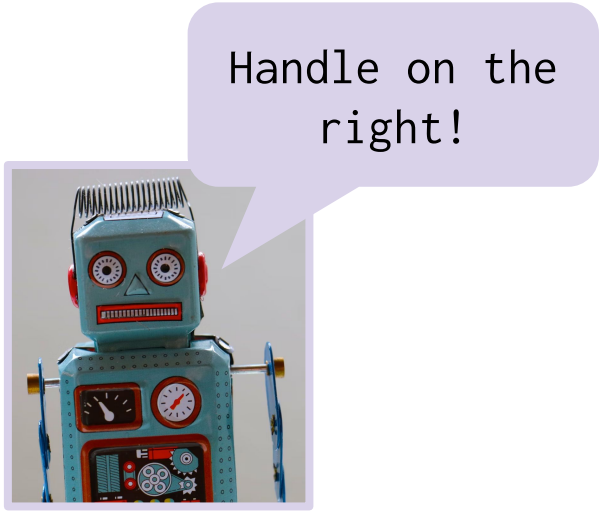




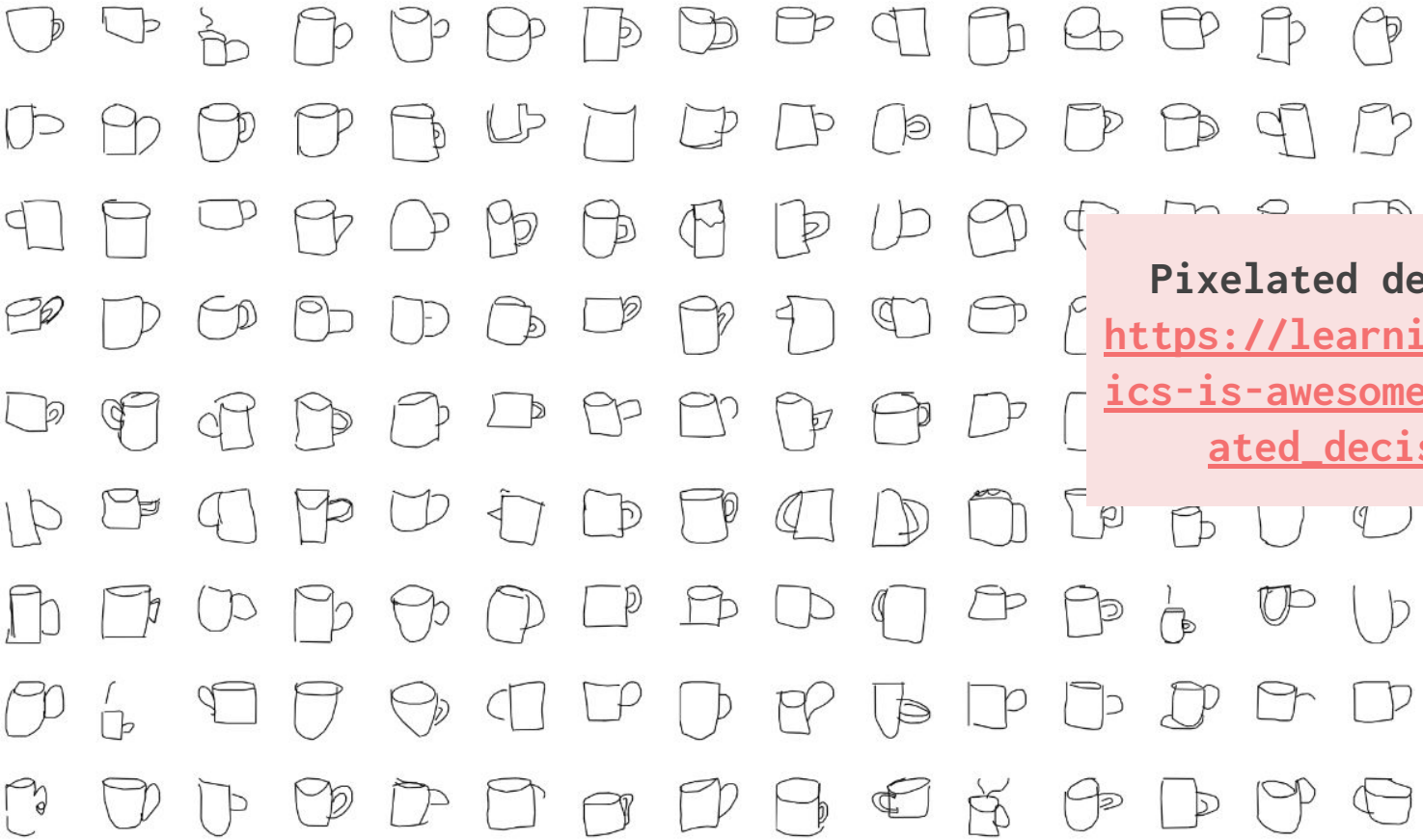
What are the key pixel-based calculations, decisions, and outputs for your approach?

What could this approach involve computationally?



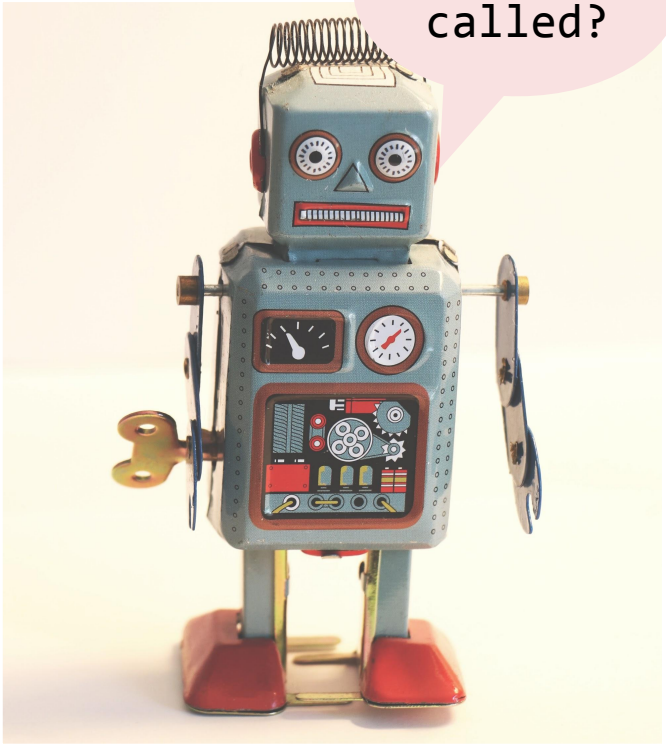


How many times would that approach not work? And why?



Pixelated decisions:
https://learning.statistics-is-awesome.org/pixelated_decisions/

The robots are coming!



Exploring pixel-based data

In the game Quick, draw!, the drawings are much larger, and so there are many pixels to consider.

How can we use information about the pixels (digital image data) to create numeric variables about each sketch?

Link to app

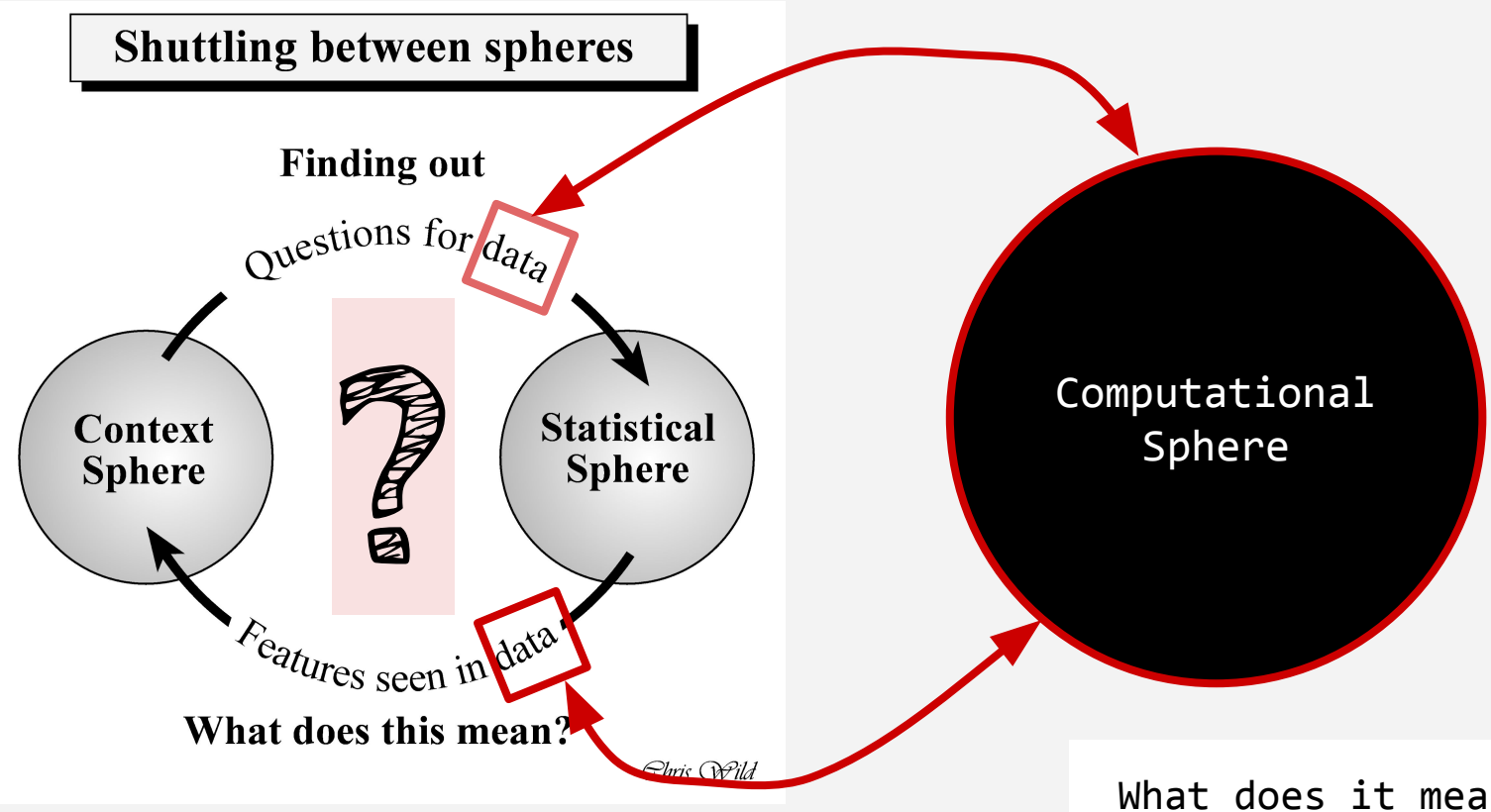
Can you figure out what each of the variables is measuring?

Do any of these computational measures seem to make sense for sorting sketches of cats?

Do the Quick, Draw! sketches reveal about how we draw cats?

1. Generate a random sample of 40 cat drawings to use as your training data
2. Use the pixel power app to sort the drawings
3. Generate links and use the labelled data to select the numeric variable to use for your decision rule AND the cutoff value
4. Use the sketch sorter app to test out your classification model with 500 sketches
5. Review the result and click on any sketches that were misclassified

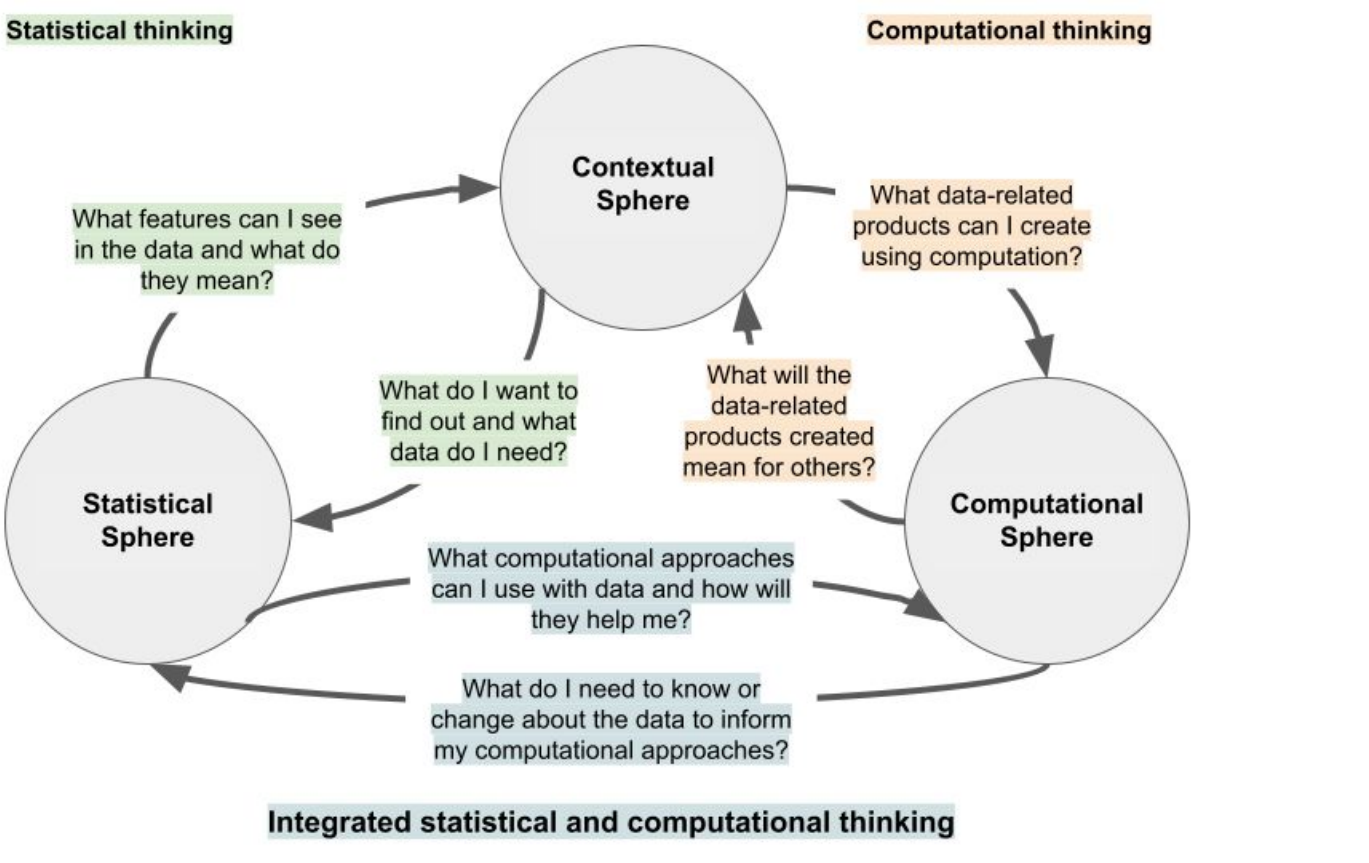
Why explore sketches?



Reprinted with permission from Wild & Pfannkuch (1999)

What does it mean to learn from data when shuttling between all three spheres?

My framework for integrating statistical & computational thinking
AKA connecting spheres



Three phase approach to learning data science

gradual learning “on ramp”

Awareness

Gain a greater awareness of what is possible with models

Participation

Participate in modelling activities

Production

Produce and test a model of your own

We can't escape uncertainty when modelling!

What proportion of mug drawings have handles on the right?

Current statistical approach

- take a random sample of papers
- manually as humans classify them as either handle on the right or not
- construct a confidence interval for the proportion of ALL mug drawings that have handles on the right

Uncertainty comes from using a sample (i.e. sampling variation), as well as subjectiveness of human labelling

New "data science" approach

- take two small random distinct samples of papers
  - use one as training data set
  - use one as testing data set
- manually as humans classify all papers in both data sets as either teaching experience/conceptual OR empirical data-based research
- using the training data set
  - develop a decision rule for classifying the paper that can be automated, using variables extracted from the words
- using the testing data set
  - apply the decision rule and evaluate the classification model using classification rates (e.g., PCC percent correct classifications)
- [repeat from the beginning until develop "good" model]
- apply decision rule to all papers to assign labels

Uncertainty comes from using a model (i.e. misclassification), as well as subjectiveness of human labelling