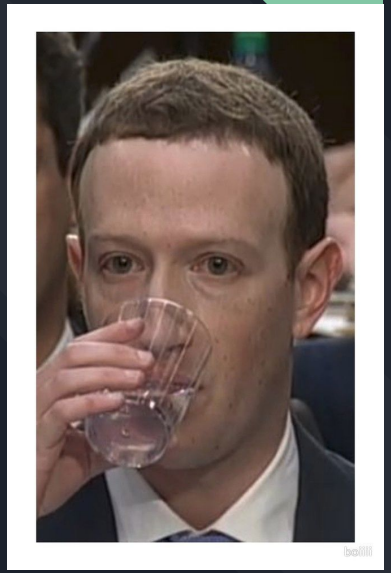
Data Mining 101

WISEST Summer Lunch Presentations 2019

Sarah Davis

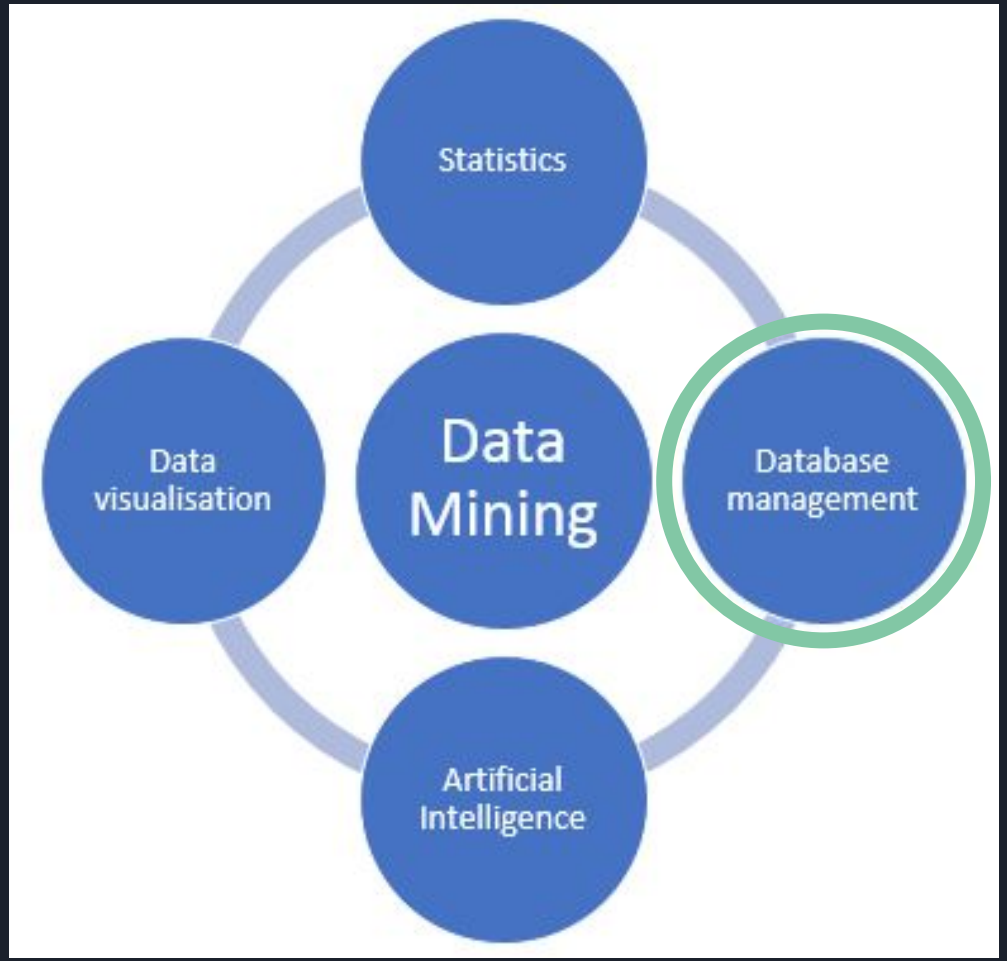


What is data mining anyway?

The use of mathematical, statistical, and computational techniques to extract trends and patterns from large amounts of data.

This data can be anything from credit card taps to Facebook statuses to hospital visits to brainwaves.

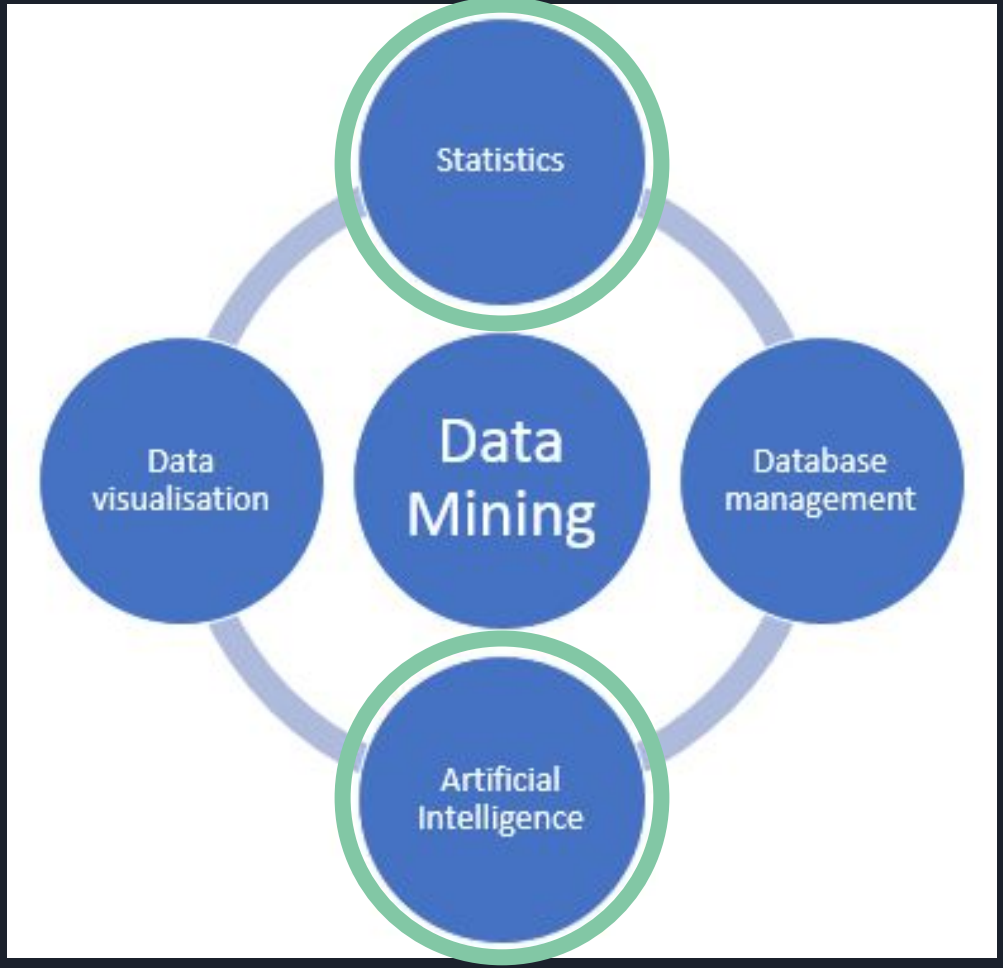
People’s habits, behaviors, and attributes en masse can reveal new and interesting information!


Statistics
Data visualisation
Data Mining
Database management
Artificial Intelligence

Database Management

- All about protecting the “health” of the data!

- Adding/removing data points
  - Make sure all data being added is good quality in the proper format
- Storage and capacity planning
  - How do we store data of multiple forms? Minimize storage cost?
- Backing up and recovering data
  - Data loss can be detrimental


Statistics
Data visualisation
Data Mining
Database management
Artificial Intelligence

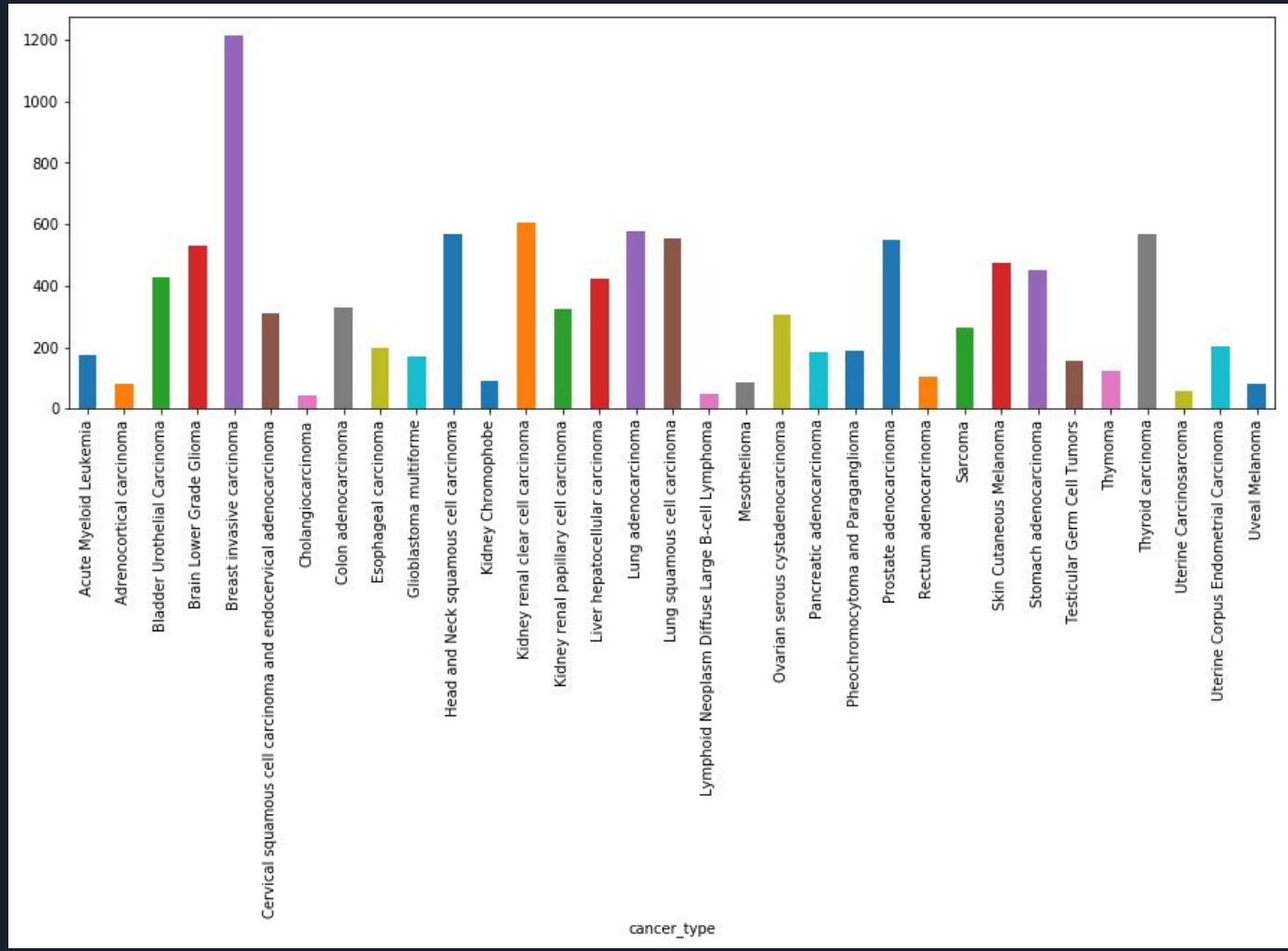
What are we trying to do?

- Describe concepts
  - Contrast data with other data
- Classify and predict
  - Build models that can tell the difference between groups
  - Predict some unknown value
- Cluster
  - Let the algorithm decide how to categorize data
- Analyze outliers
  - Noise or exception?

Example (from my own research!)

- We want to know what kind of cancer each patient has based on their gene expression (mRNA counts for each gene, measured from a tumor biopsy)

- To do this, we use the statistical technique *logistic regression* implemented in Python


1200
1000
800
600
400
200
0
Acute Myeloid Leukemia
Adrenocortical carcinoma
Bladder Urothelial Carcinoma
Brain Lower Grade Glioma
Breast invasive carcinoma
Cervical squamous cell carcinoma and endocervical adenocarcinoma
Cholangiocarcinoma
Colon adenocarcinoma
Esophageal carcinoma
Glioblastoma multiforme
Head and Neck squamous cell carcinoma
Kidney Chromophobe
Kidney renal clear cell carcinoma
Kidney renal papillary cell carcinoma
Liver hepatocellular carcinoma
Lung adenocarcinoma
Lung squamous cell carcinoma
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
Mesothelioma
Ovarian serous cystadenocarcinoma
Pancreatic adenocarcinoma
Pheochromocytoma and Paraganglioma
Prostate adenocarcinoma
Rectum adenocarcinoma
Sarcoma
Skin Cutaneous Melanoma
Stomach adenocarcinoma
Testicular Germ Cell Tumors
Thymoma
Thyroid carcinoma
Uterine Carcinosarcoma
Uterine Corpus Endometrial Carcinoma
Uveal Melanoma
cancer_type

	A1BG|1	A1CF|29974	A2BP1|54715	A2LD1|87769	A2ML1|144568	A2M|2	A4GALT|53947	A4GNT|51146	AAA1|404744	AAAS|8086
TCGA-CV-6436-01A-11R-1686-07	28.0426	0.000	0.0000	78.7044	14049.9159	1595.5524	1809.3102	0.0	0.0000	1290.5216
TCGA-KB-A93H-01A-11R-A39E-31	18.9607	123.061	0.5171	82.3164	0.5171	2909.5553	41.8821	0.0	0.0000	1237.3320
TCGA-DD-AAEB-01A-11R-A41C-07	94409.7087	2496.150	0.0000	550.7466	0.0000	11215.3063	124.5397	0.0	1.3391	1375.2929
TCGA-OR-A5LD-01A-11R-A29S-07	36.6350	0.000	0.0000	39.1859	1.3569	5242.1574	104.4776	0.0	0.0000	3084.1248
TCGA-OR-A5L9-01A-11R-A29S-07	1091.3978	0.000	16.1290	17.2581	75.2688	7851.8817	655.9140	0.0	0.0000	2881.7204
TCGA-BS-A0V7-01A-21R-A118-07	1235.5140	0.000	13.0841	81.2897	0.0000	13809.1589	960.7477	0.0	0.0000	2357.0093

← What we have

10k rows (each one is a patient with some form of cancer)

20k columns (each one is a gene)

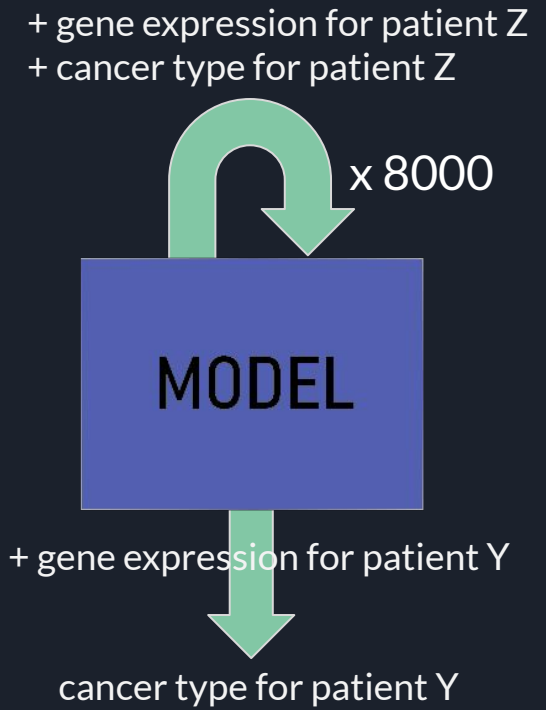
What we want to predict →

Each patient's cancer type

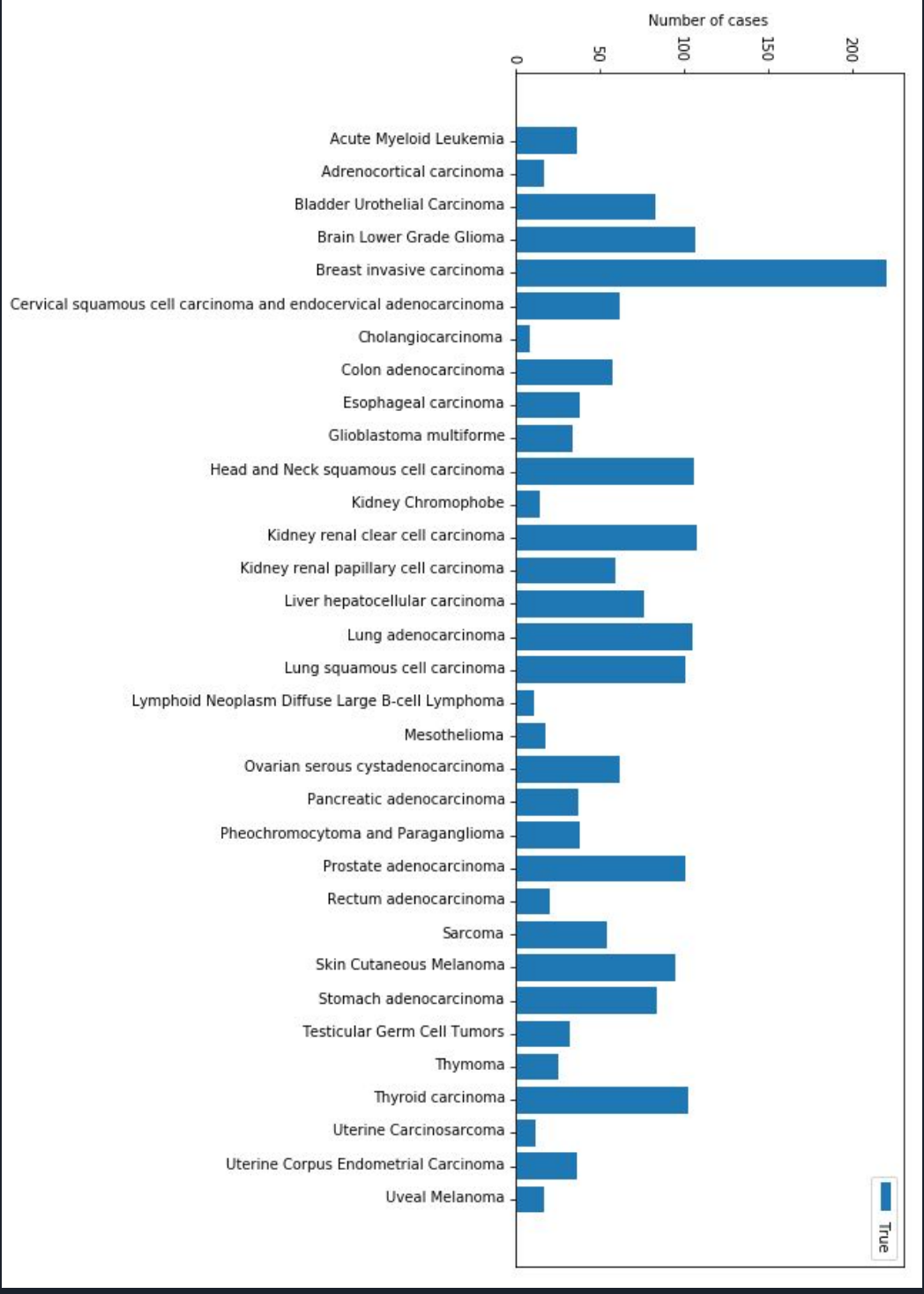
	cancer_type
TCGA-CV-6436-01A-11R-1686-07	Head and Neck squamous cell carcinoma
TCGA-KB-A93H-01A-11R-A39E-31	Stomach adenocarcinoma
TCGA-DD-AAEB-01A-11R-A41C-07	Liver hepatocellular carcinoma
TCGA-OR-A5LD-01A-11R-A29S-07	Adrenocortical carcinoma
TCGA-OR-A5L9-01A-11R-A29S-07	Adrenocortical carcinoma
TCGA-BS-A0V7-01A-21R-A118-07	Uterine Corpus Endometrial Carcinoma

Process

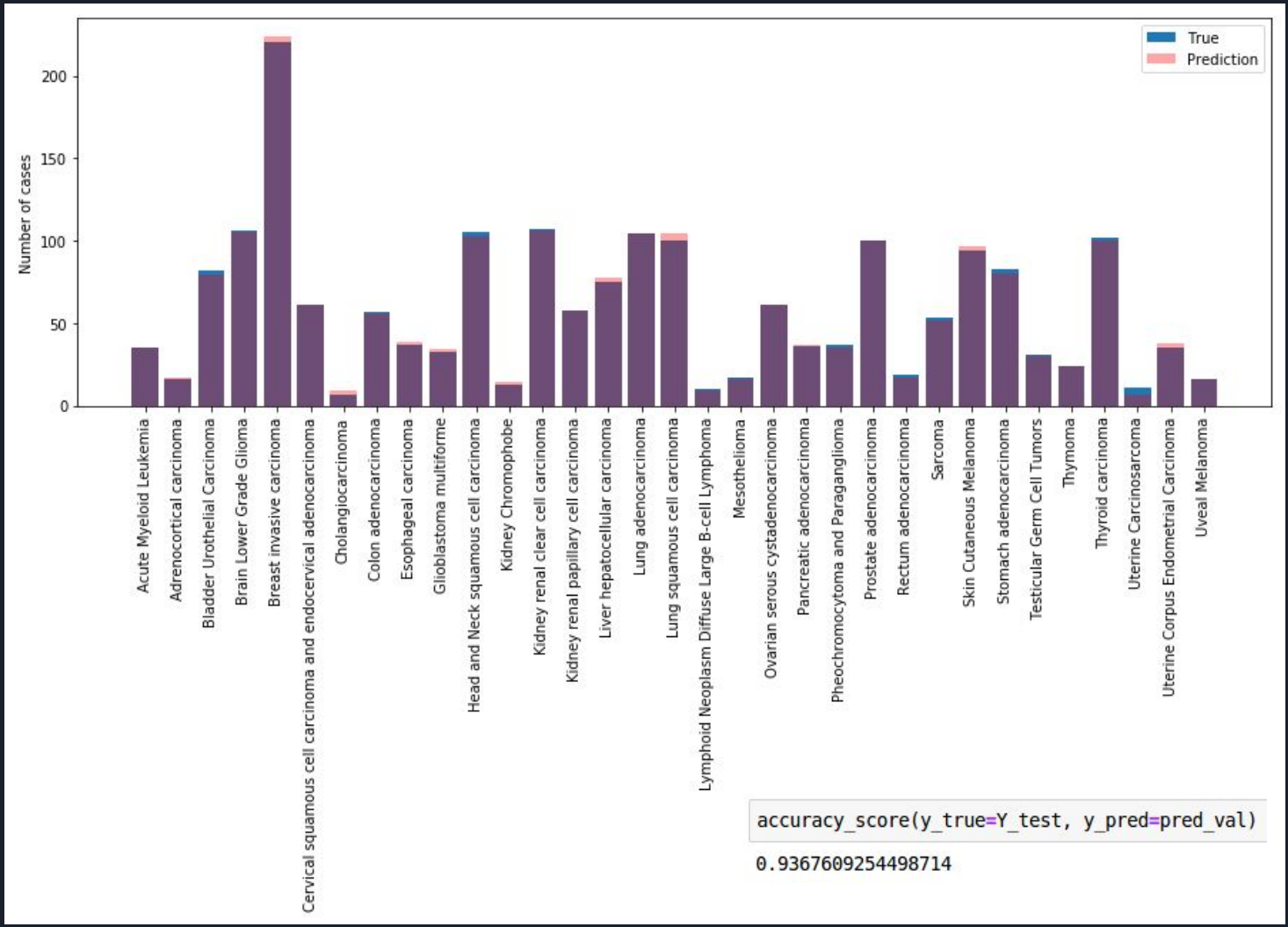
- The model starts at a generic baseline
- During training, the model is exposed to many correct pairings of gene expression and cancer type
- The formula slowly changes to better and better predict the correct answer
- Apply model to patients the model has never seen before
- Measure the accuracy!

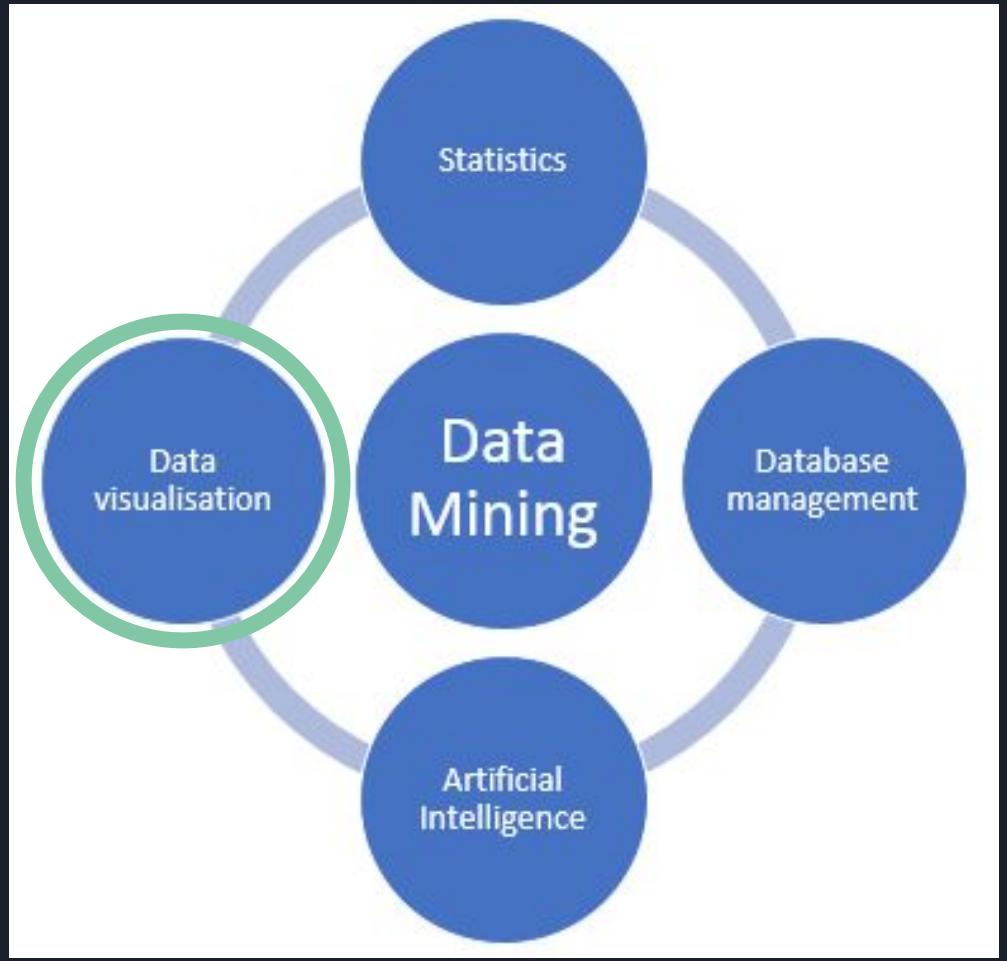


Results


Number of cases
0
50
100
150
200
Acute Myeloid Leukemia
Adrenocortical carcinoma
Bladder Urothelial Carcinoma
Brain Lower Grade Glioma
Breast invasive carcinoma
Cervical squamous cell carcinoma and endocervical adenocarcinoma
Cholangiocarcinoma
Colon adenocarcinoma
Esophageal carcinoma
Glioblastoma multiforme
Head and Neck squamous cell carcinoma
Kidney Chromophobe
Kidney renal clear cell carcinoma
Kidney renal papillary cell carcinoma
Liver hepatocellular carcinoma
Lung adenocarcinoma
Lung squamous cell carcinoma
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
Mesothelioma
Ovarian serous cystadenocarcinoma
Pancreatic adenocarcinoma
Pheochromocytoma and Paraganglioma
Prostate adenocarcinoma
Rectum adenocarcinoma
Sarcoma
Skin Cutaneous Melanoma
Stomach adenocarcinoma
Testicular Germ Cell Tumors
Thymoma
Thyroid carcinoma
Uterine Carcinosarcoma
Uterine Corpus Endometrial Carcinoma
Uveal Melanoma
True

Results


True
Prediction
Number of cases
0
50
100
150
200
Acute Myeloid Leukemia
Adrenocortical carcinoma
Bladder Urothelial Carcinoma
Brain Lower Grade Glioma
Breast invasive carcinoma
Cervical squamous cell carcinoma and endocervical adenocarcinoma
Cholangiocarcinoma
Colon adenocarcinoma
Esophageal carcinoma
Glioblastoma multiforme
Head and Neck squamous cell carcinoma
Kidney Chromophobe
Kidney renal clear cell carcinoma
Kidney renal papillary cell carcinoma
Liver hepatocellular carcinoma
Lung adenocarcinoma
Lung squamous cell carcinoma
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
Mesothelioma
Ovarian serous cystadenocarcinoma
Pancreatic adenocarcinoma
Pheochromocytoma and Paraganglioma
Prostate adenocarcinoma
Rectum adenocarcinoma
Sarcoma
Skin Cutaneous Melanoma
Stomach adenocarcinoma
Testicular Germ Cell Tumors
Thymoma
Thyroid carcinoma
Uterine Carcinosarcoma
Uterine Corpus Endometrial Carcinoma
Uveal Melanoma
accuracy_score(y_true=Y_test, y_pred=pred_val)
0.9367609254498714

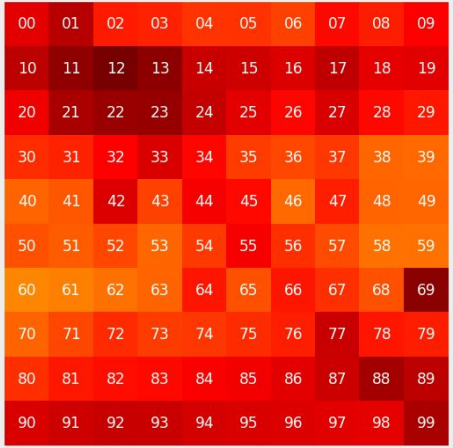

Statistics
Data visualisation
Data Mining
Database management
Artificial Intelligence

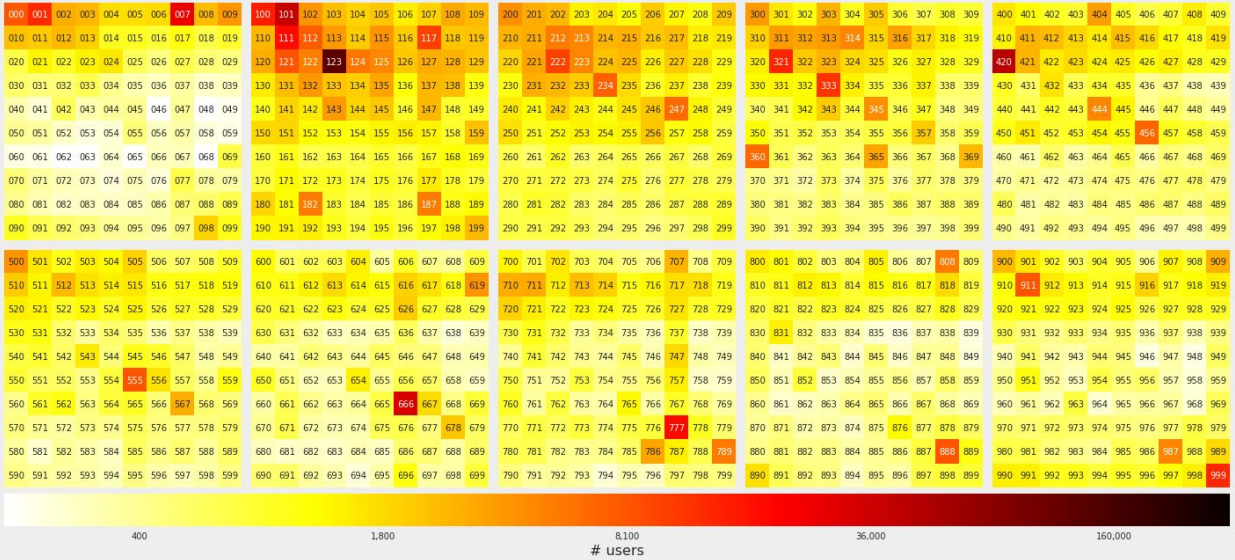
Numbers in Reddit Usernames

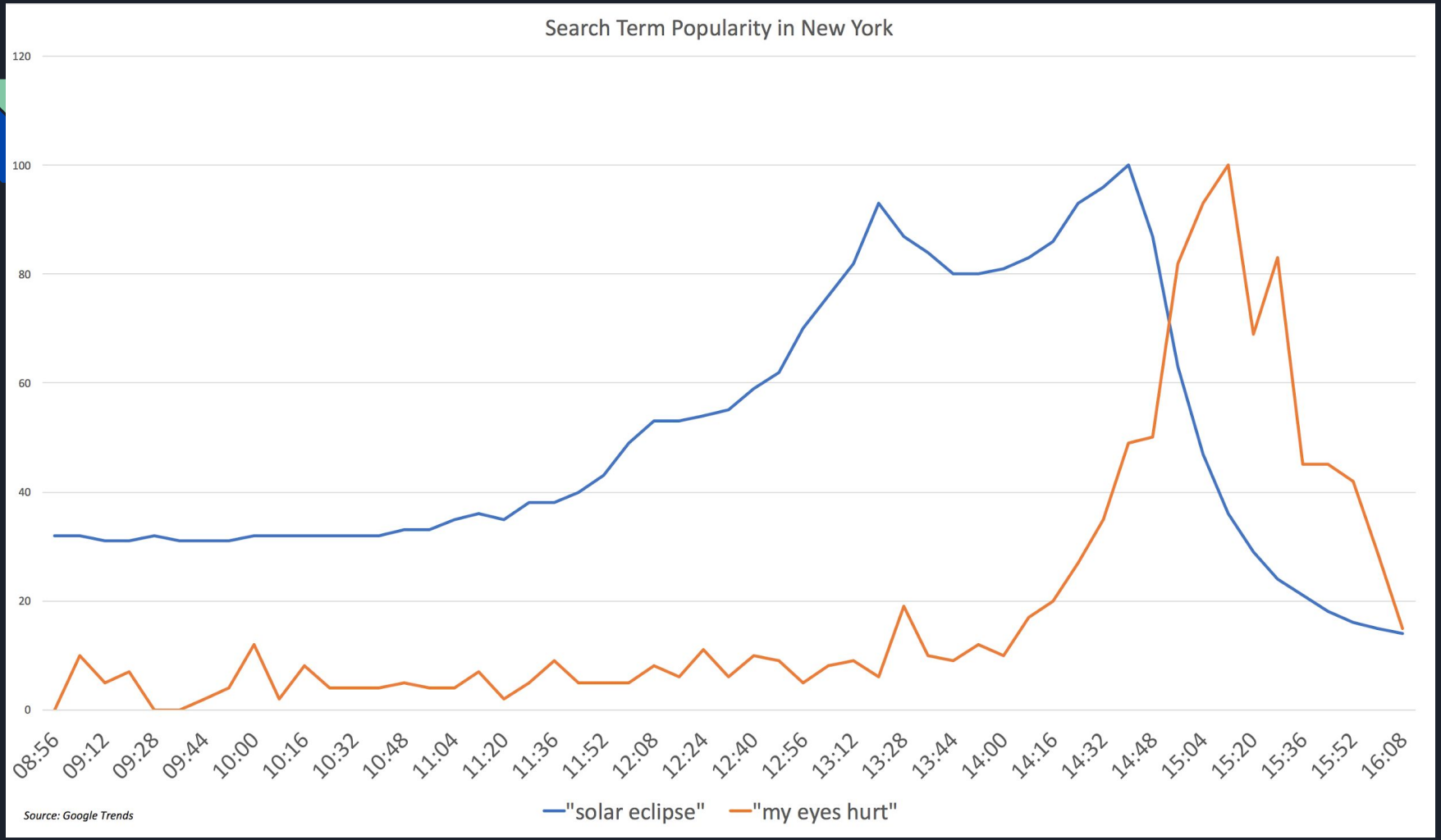
What numbers do Reddit users put on the end of their names? e.g.:

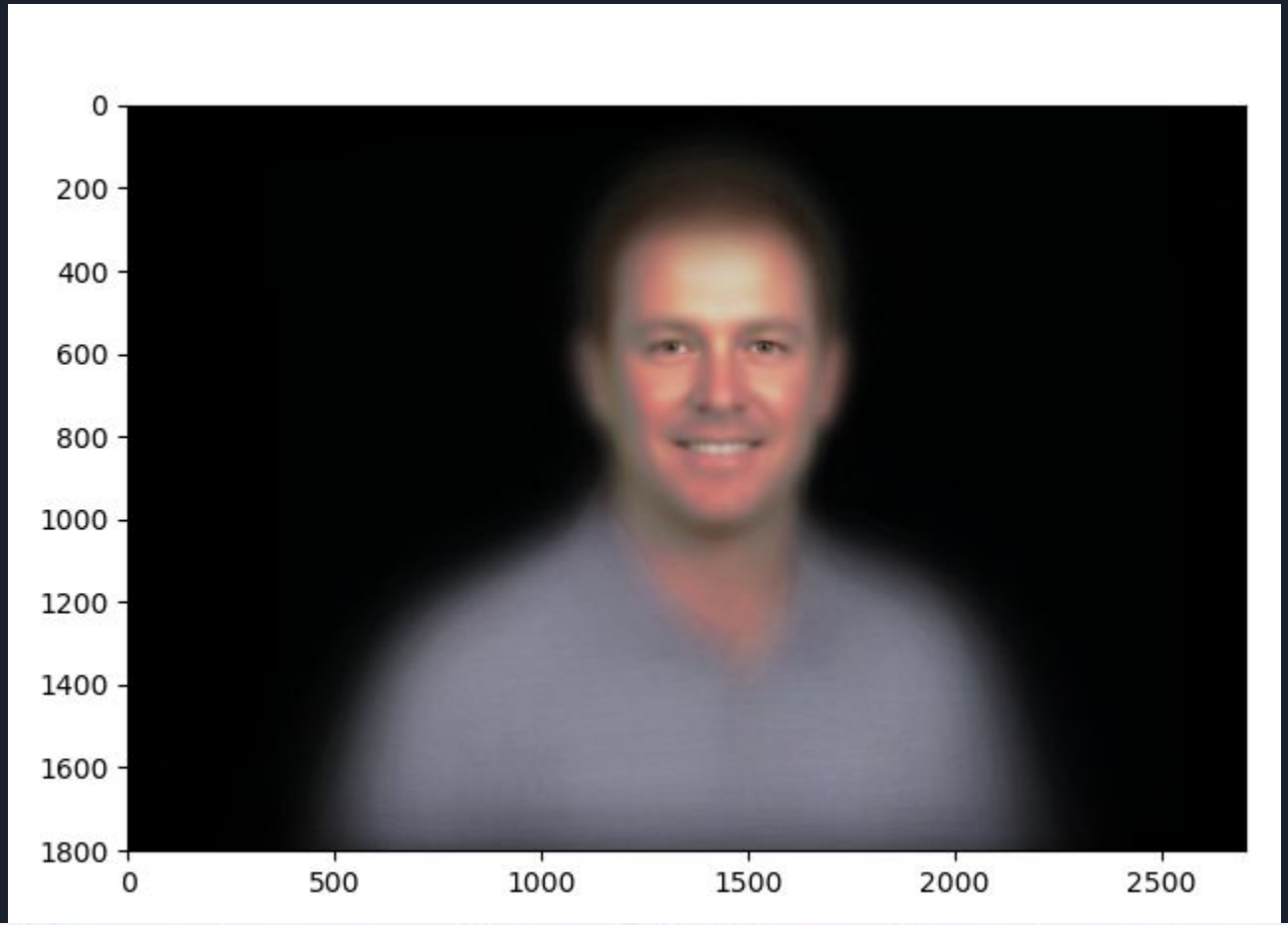
PM_ME_YOUR_DISTRO_5 Sephiroth117 princessunicorn51 xxkillslayer4457 aznhacker52
patchythepirate2 Bandersnatch13 Rainbow_Jetski69 ThePokemonMaster123



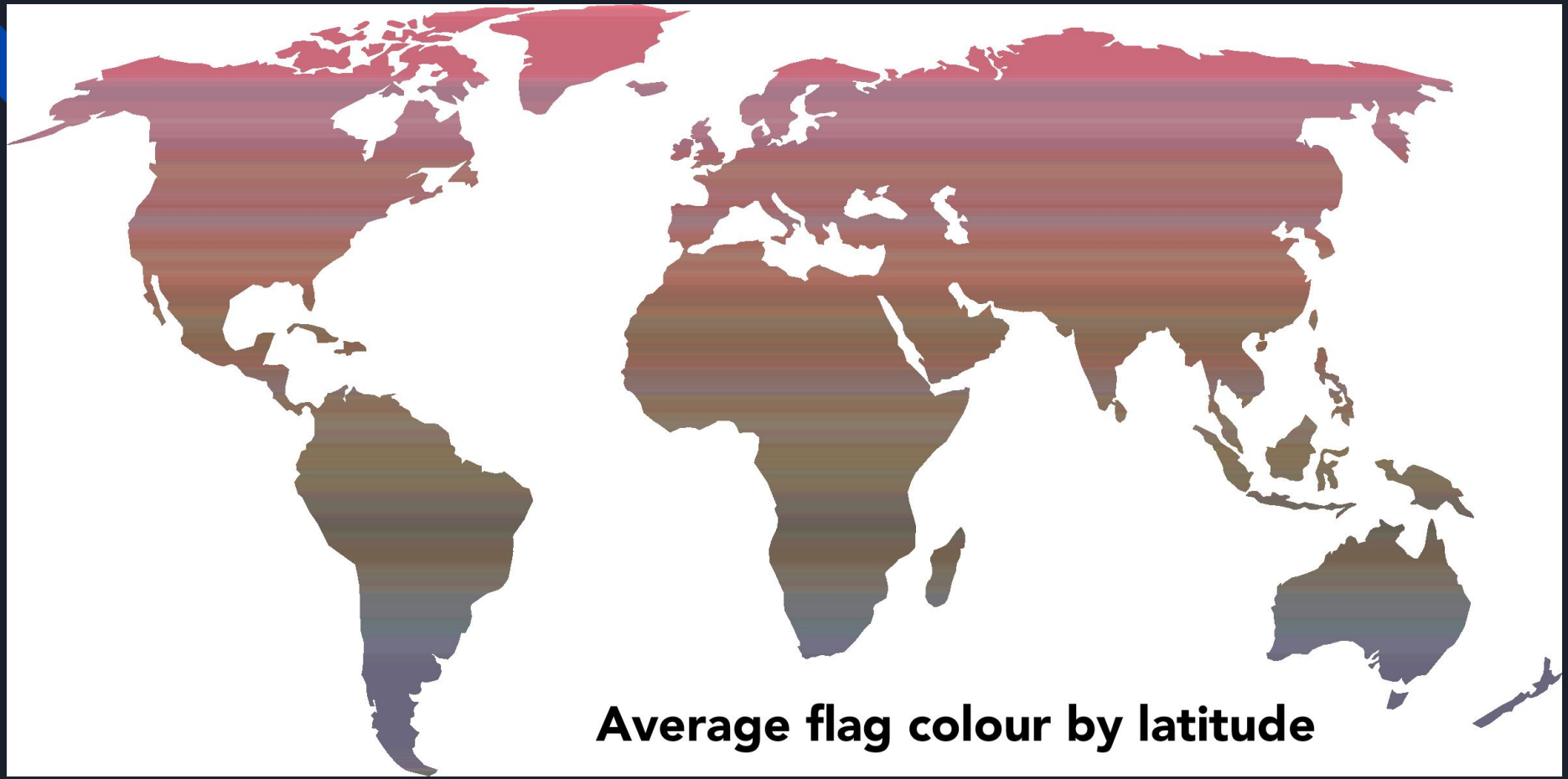





Search Term Popularity in New York
120
100
80
60
40
20
0
08:56
09:12
09:28
09:44
10:00
10:16
10:32
10:48
11:04
11:20
11:36
11:52
12:08
12:24
12:40
12:56
13:12
13:28
13:44
14:00
14:16
14:32
14:48
15:04
15:20
15:36
15:52
16:08
"solar eclipse"
"my eyes hurt"
Source: Google Trends



Combined faces of top 500 professional golfers


Average flag colour by latitude

Interesting Technical Problems of Data Mining

- Data is often noisy and imperfect
- Data collections become so large that it’s hard to manage
- Data comes in many different forms
  - text, multimedia, time series, etc.
- Unavailability of data or difficulty accessing data
- Keeping data secure (data leaks!)

Interesting Ethical Problems of Data Mining

- Who owns the data? The person it was gathered from? The institution that gathered it? The government?
  - Big problem in health informatics…
- When does aggressive personalized advertising become predatory?
- Where is the line drawn between ethical data acquisition and unethical data acquisition?

Thank you!



Questions?