Machine Learning

Hello, world!

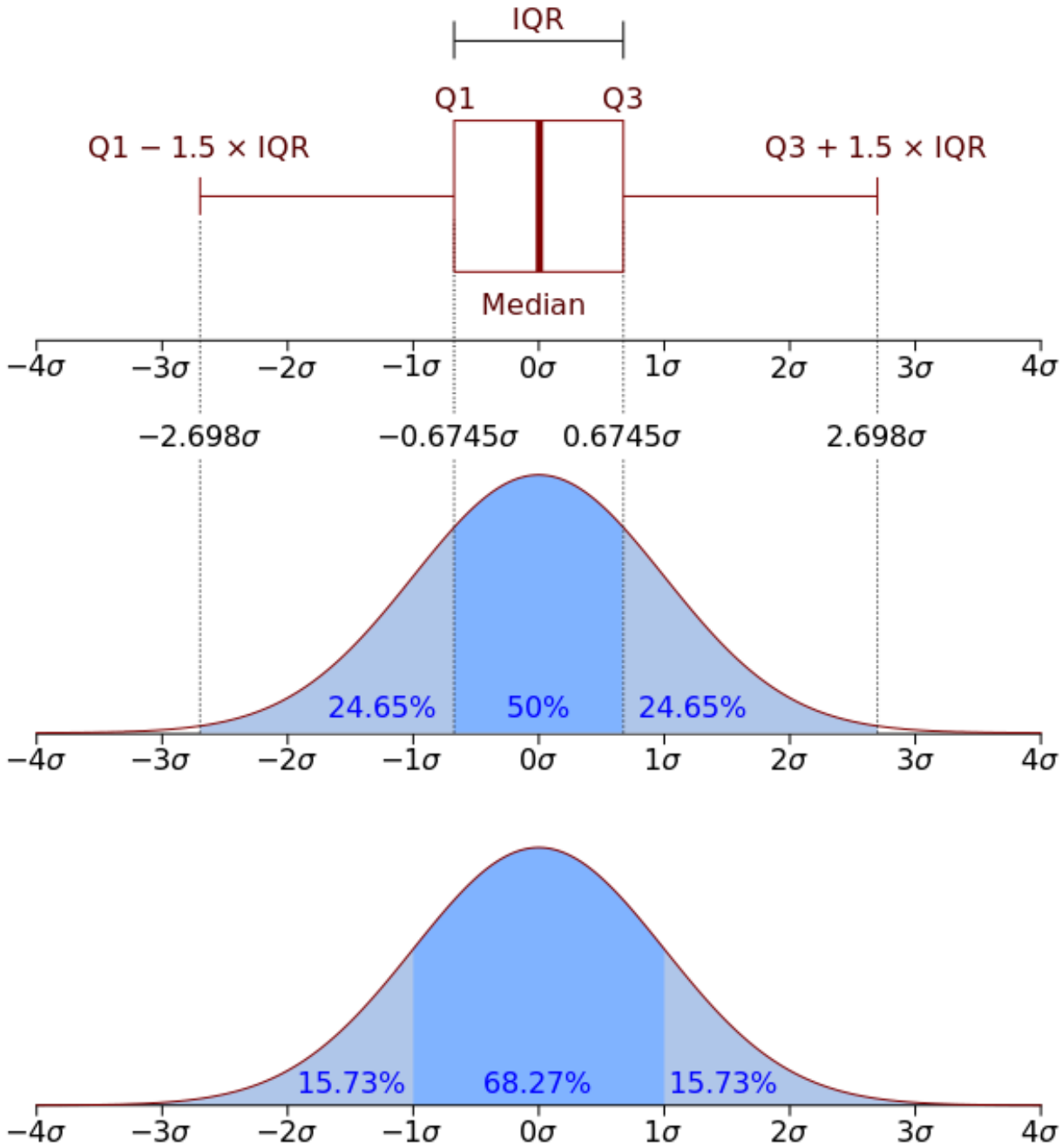
Jeff Abrahamson

2 March 2016

Machine Learning

Data Science

Statistics



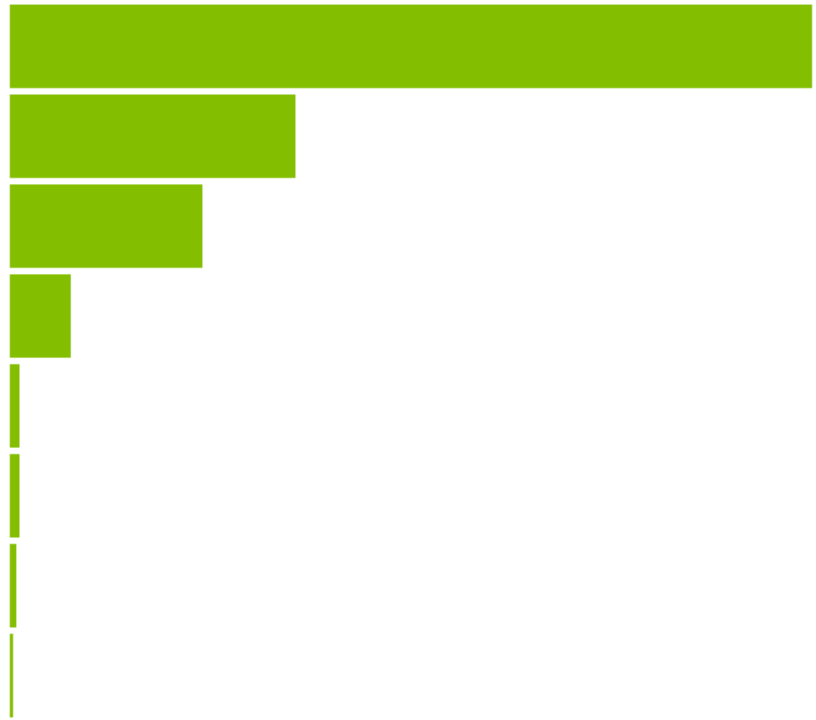
Maths

Vector Space

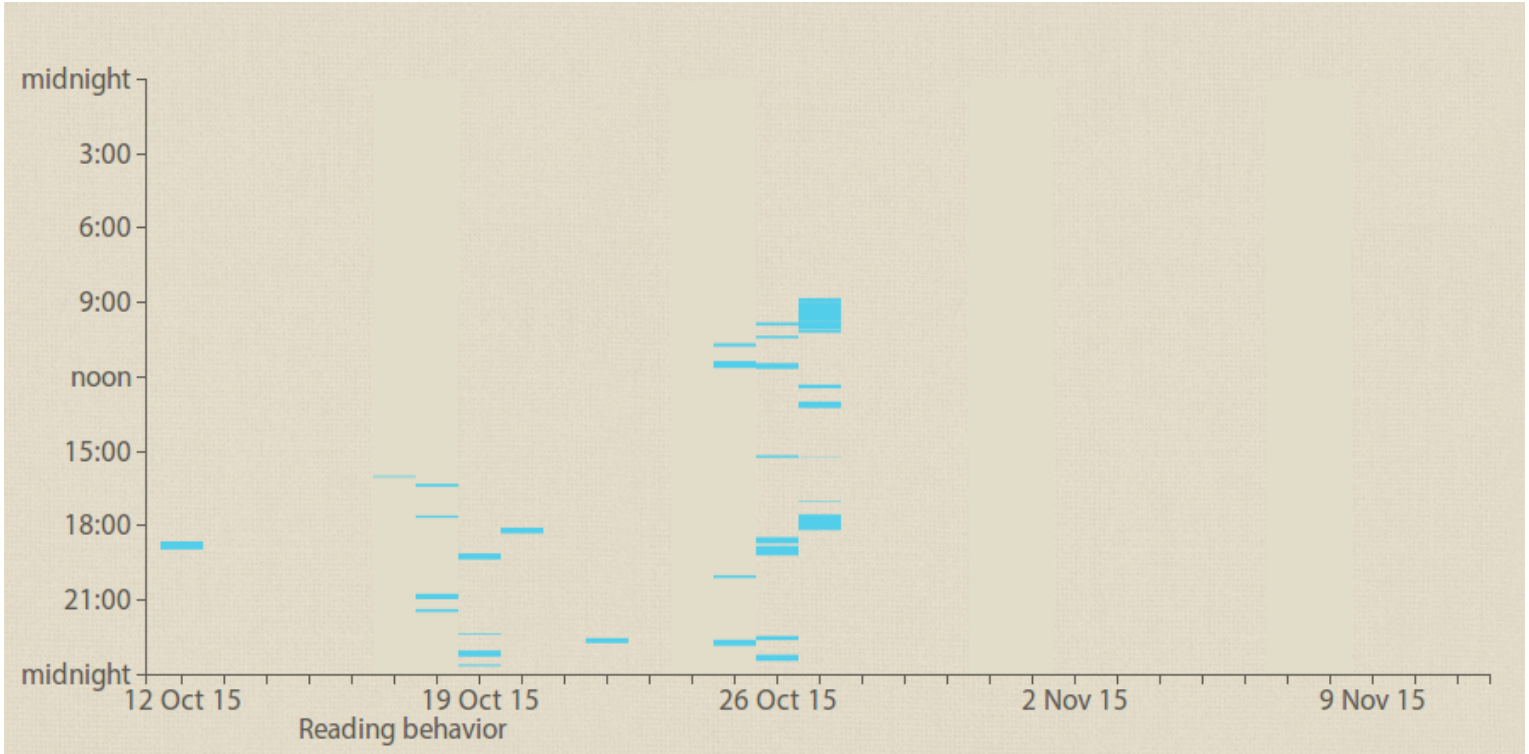
Features

Feature engineering

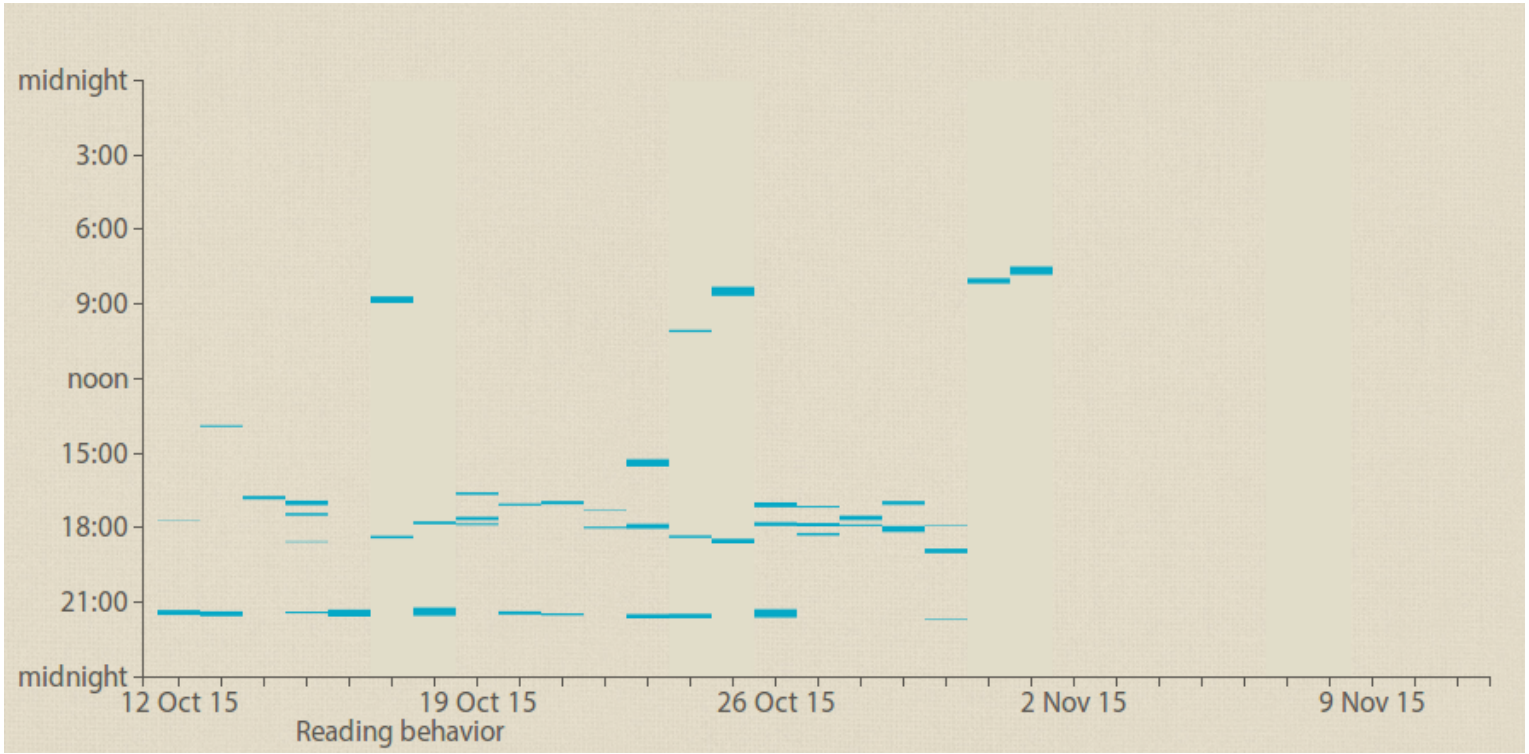
Value	Count	Percent
Mr.	517	58.025%
Miss.	185	20.763%
Mrs.	125	14.029%
Master.	40	4.489%
Dr.	7	0.786%
Rev.	6	0.673%
Sir.	5	0.561%
Col.	2	0.224%
Jonkheer.	1	0.112%
Lady.	1	0.112%
the Countess.	1	0.112%
Ms.	1	0.112%



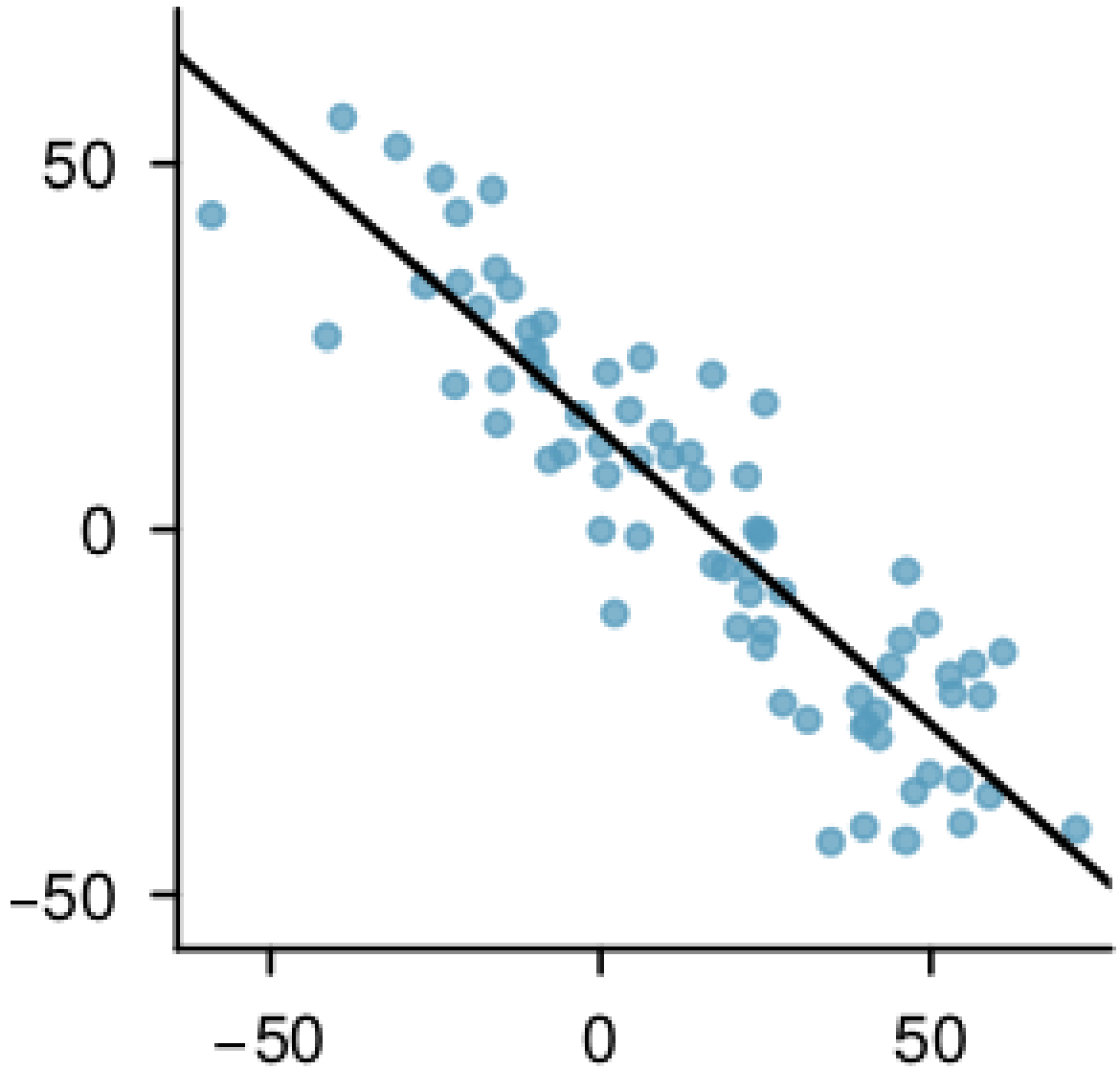
Kaggle

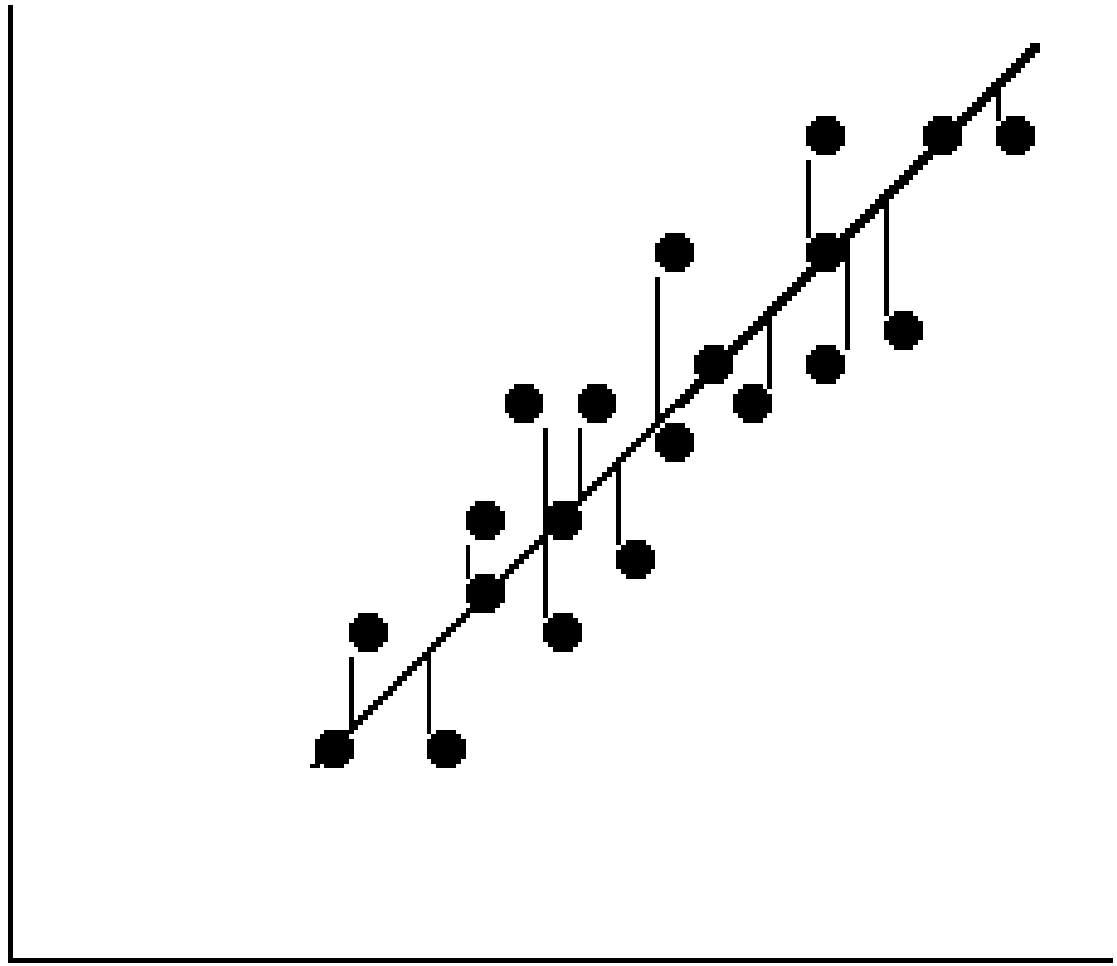


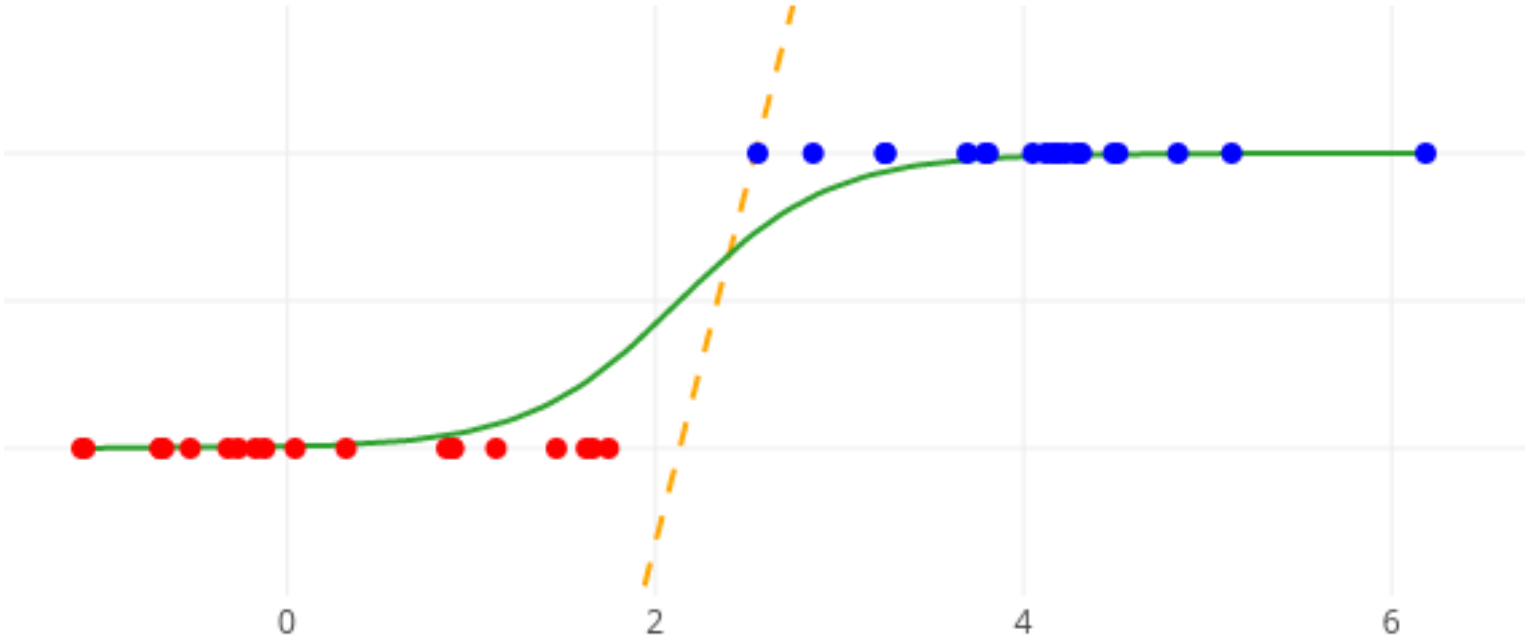
Jellybooks

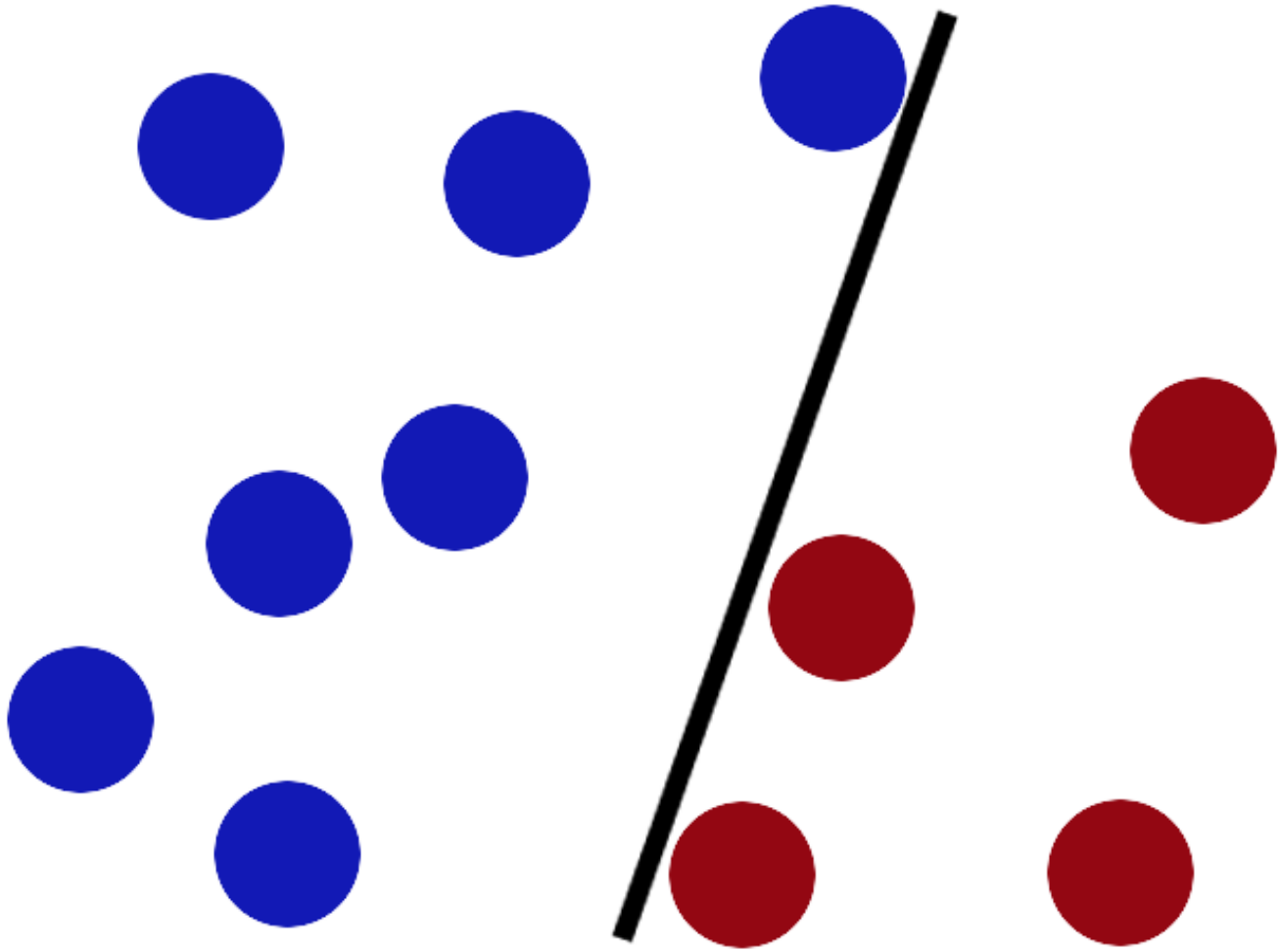


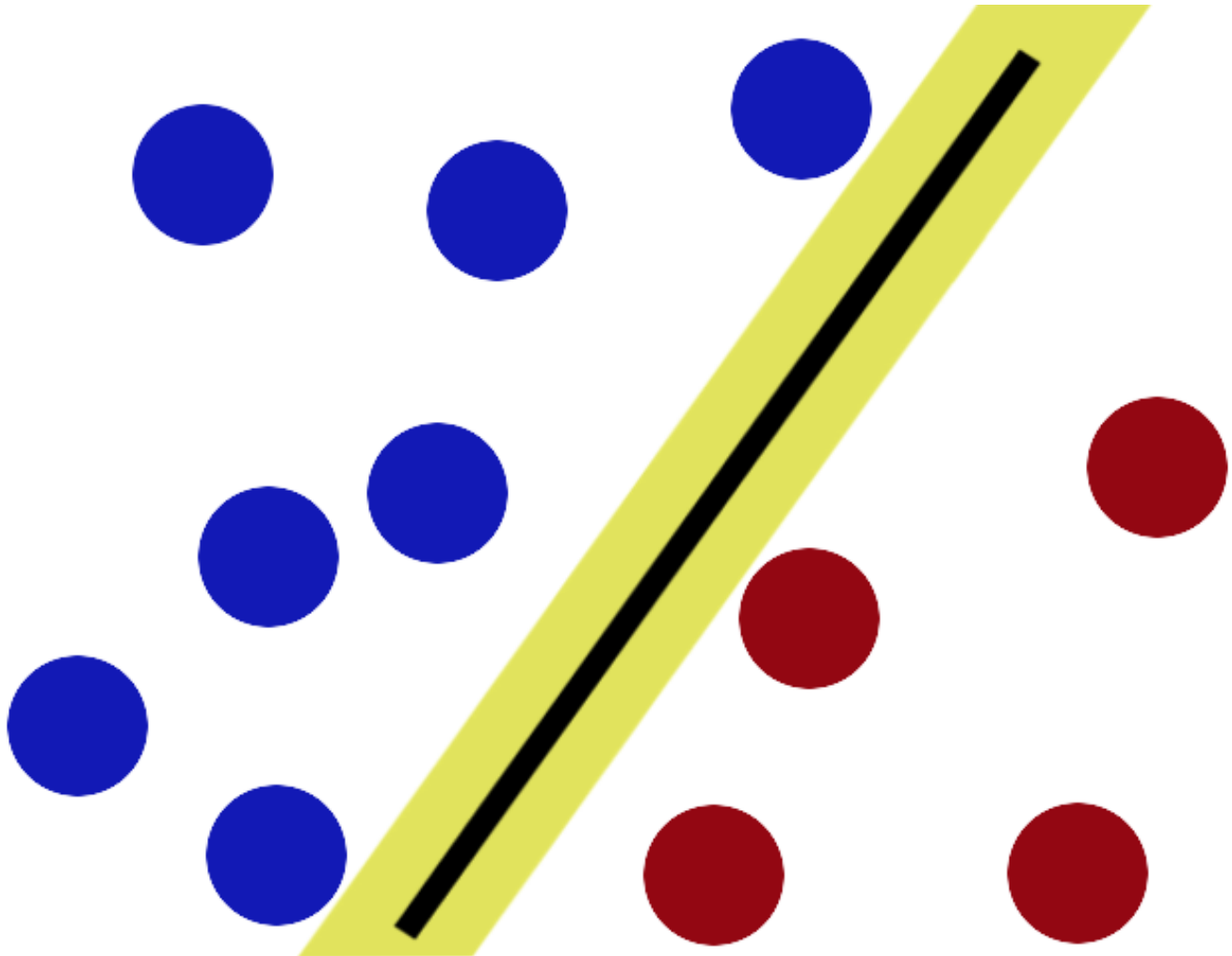
Jellybooks

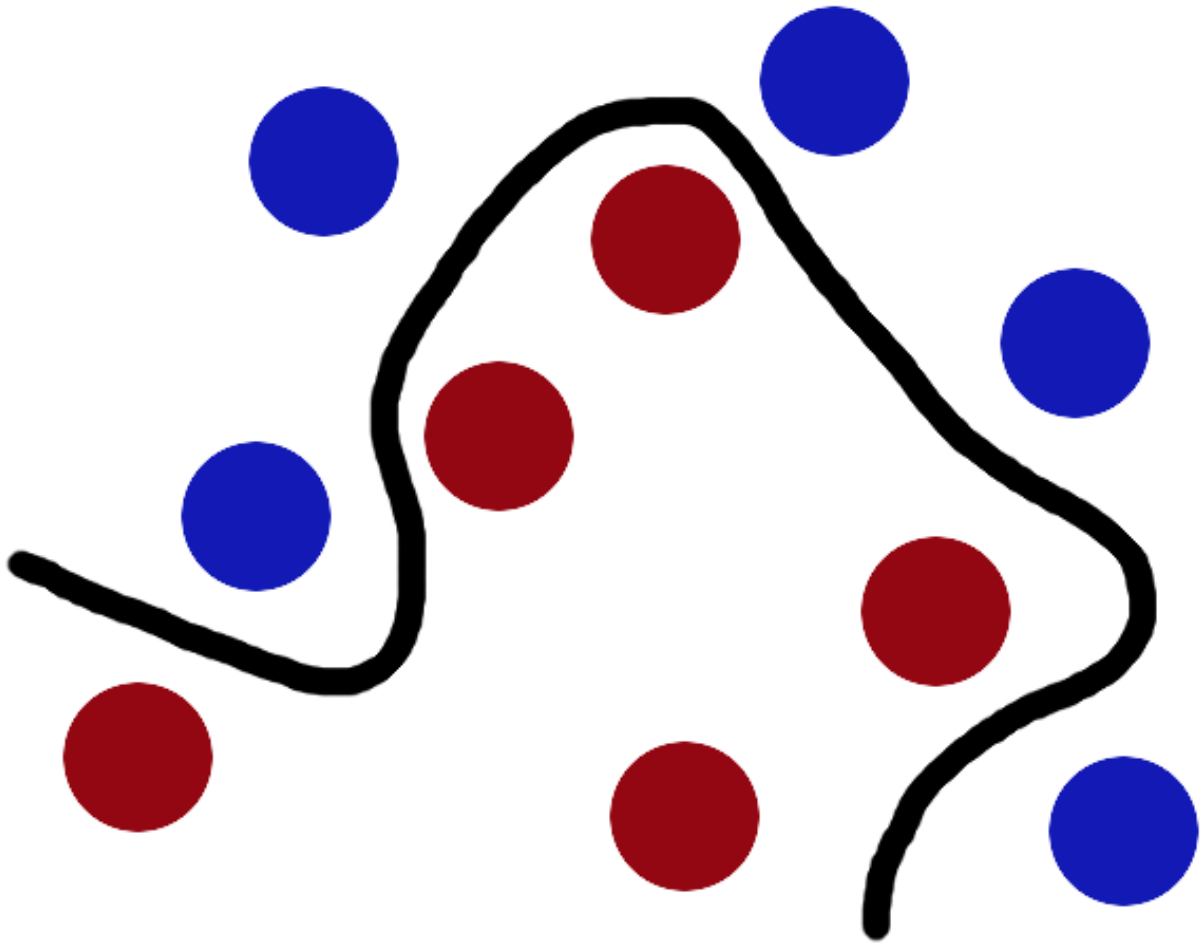


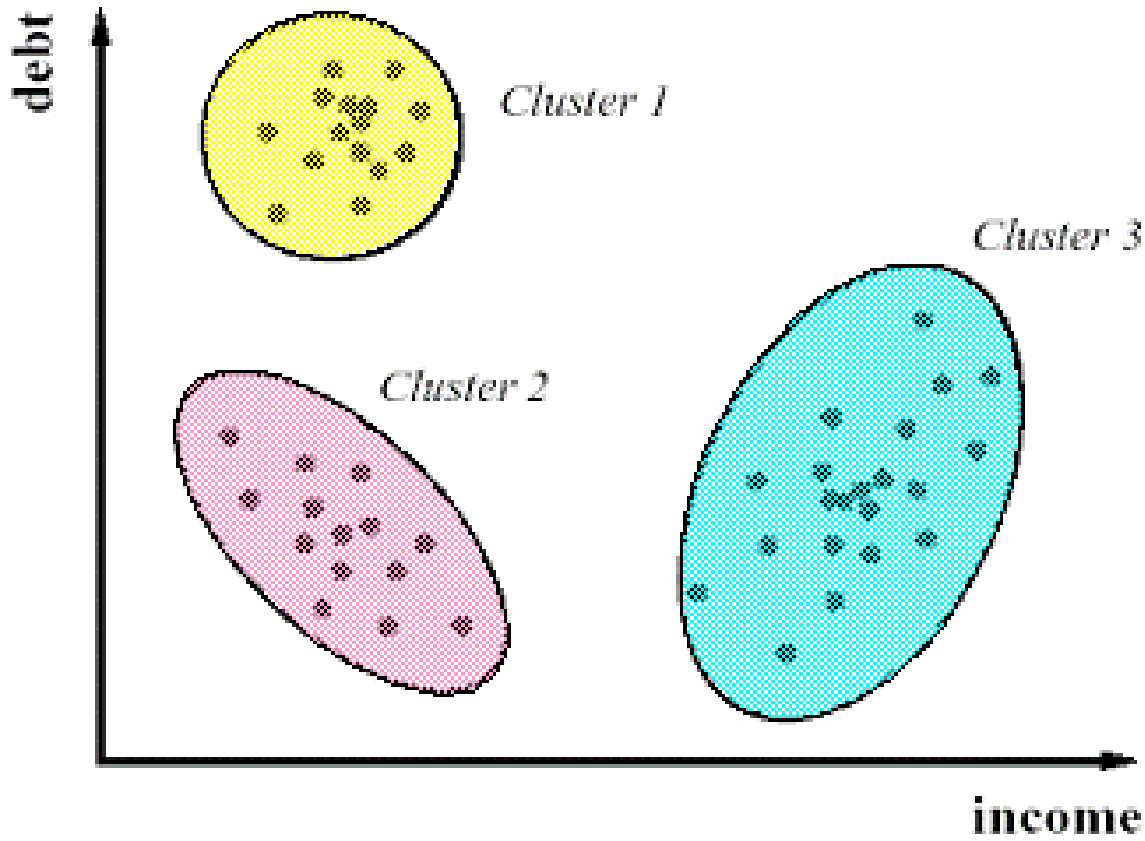


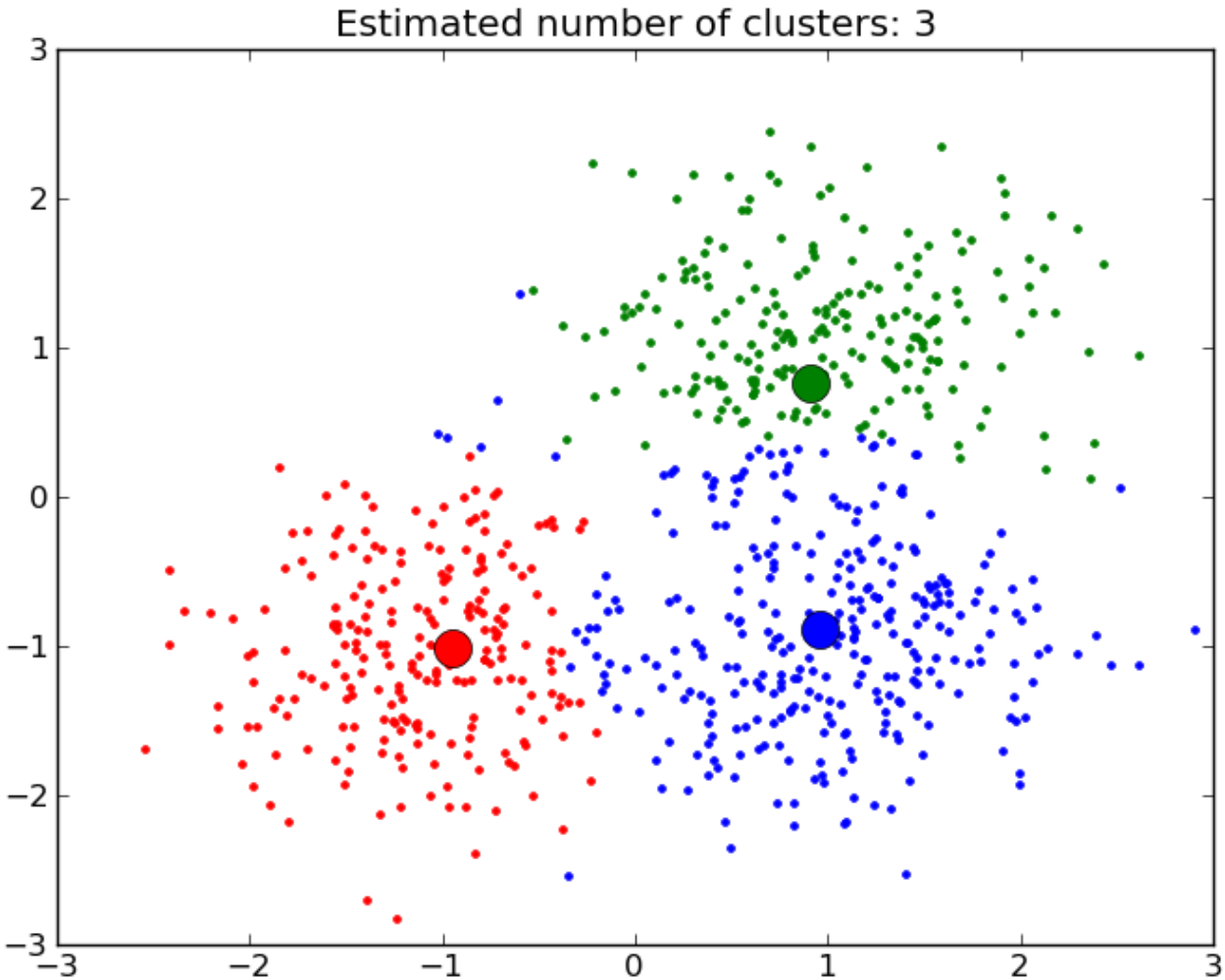


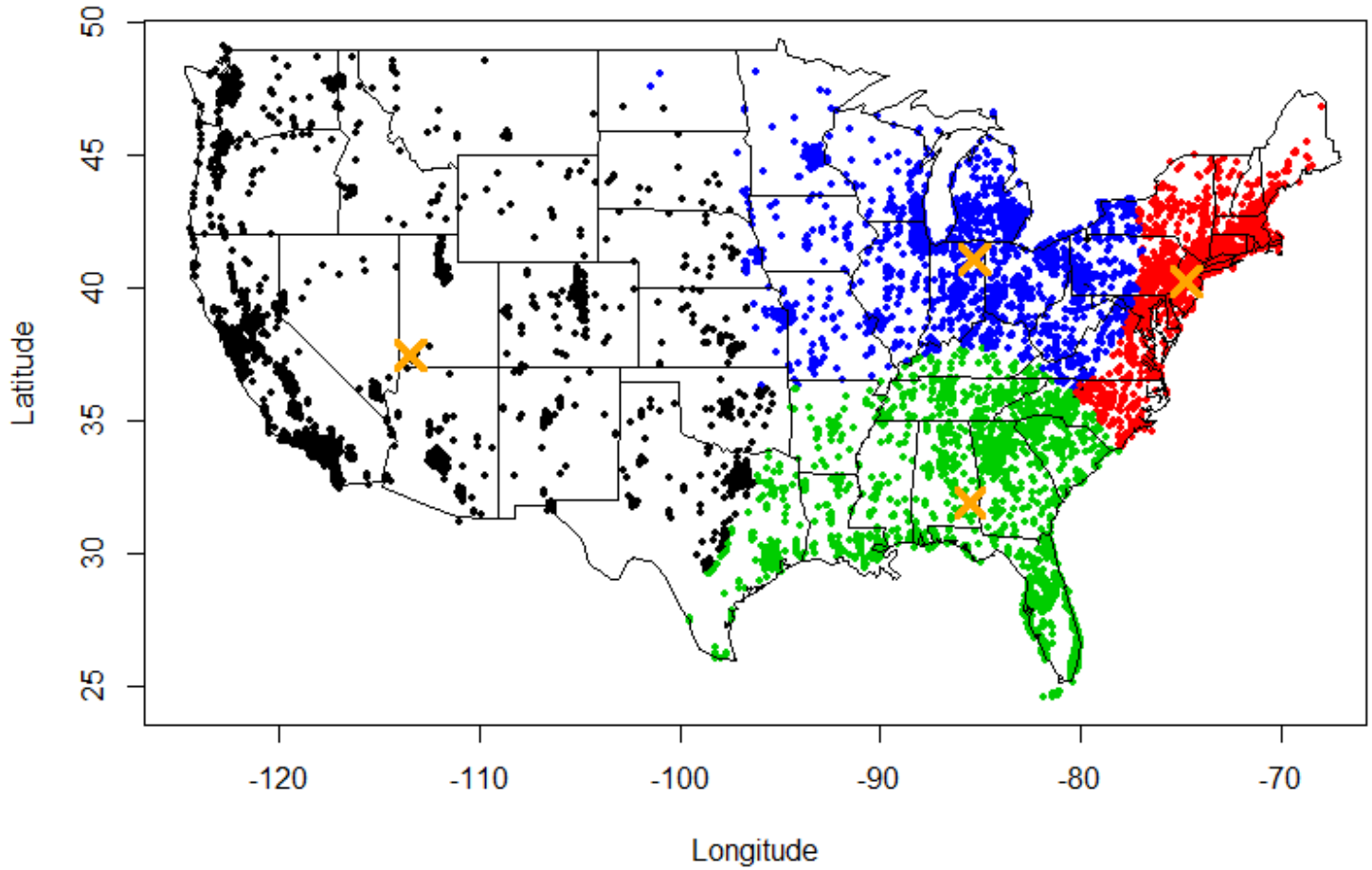


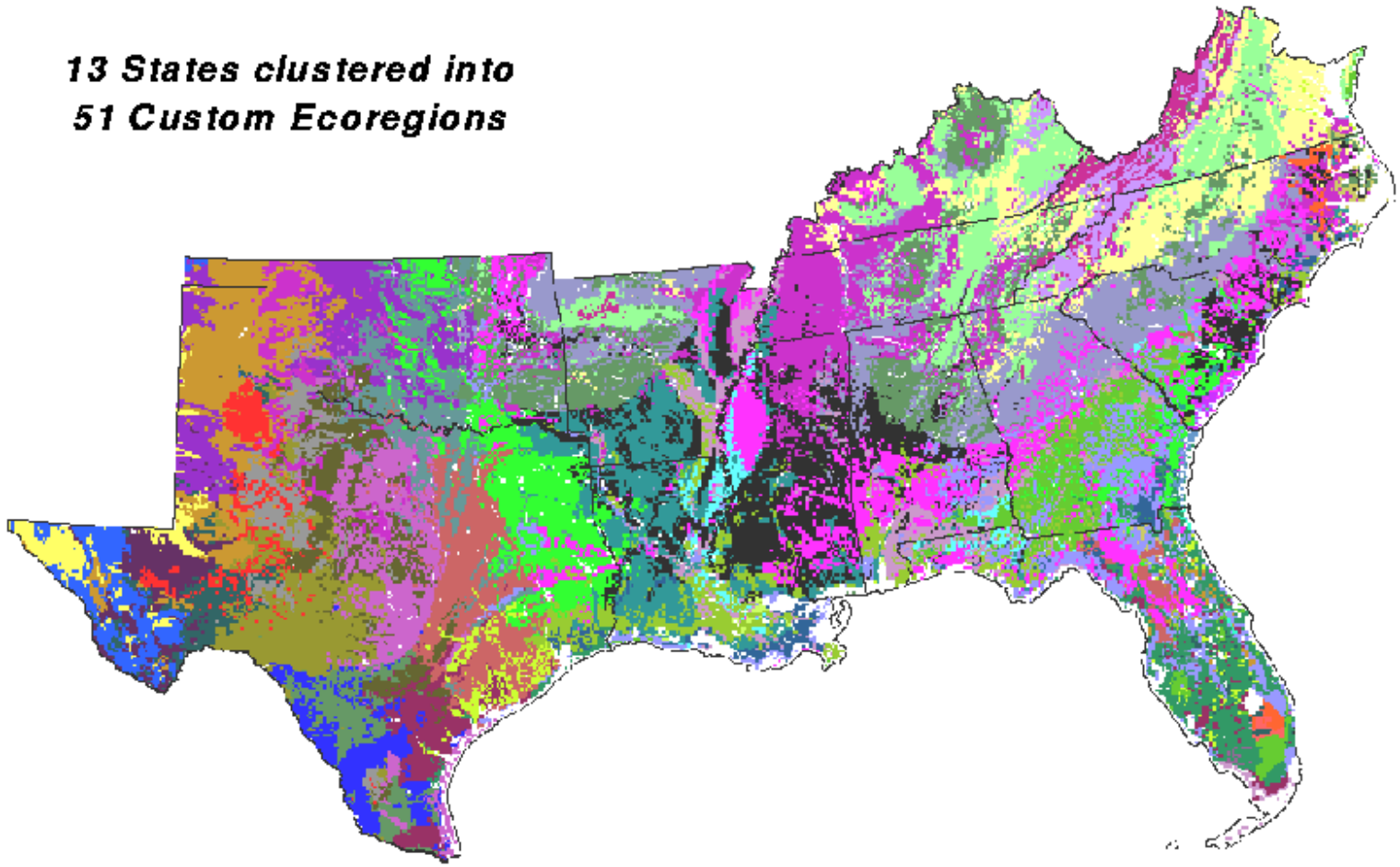


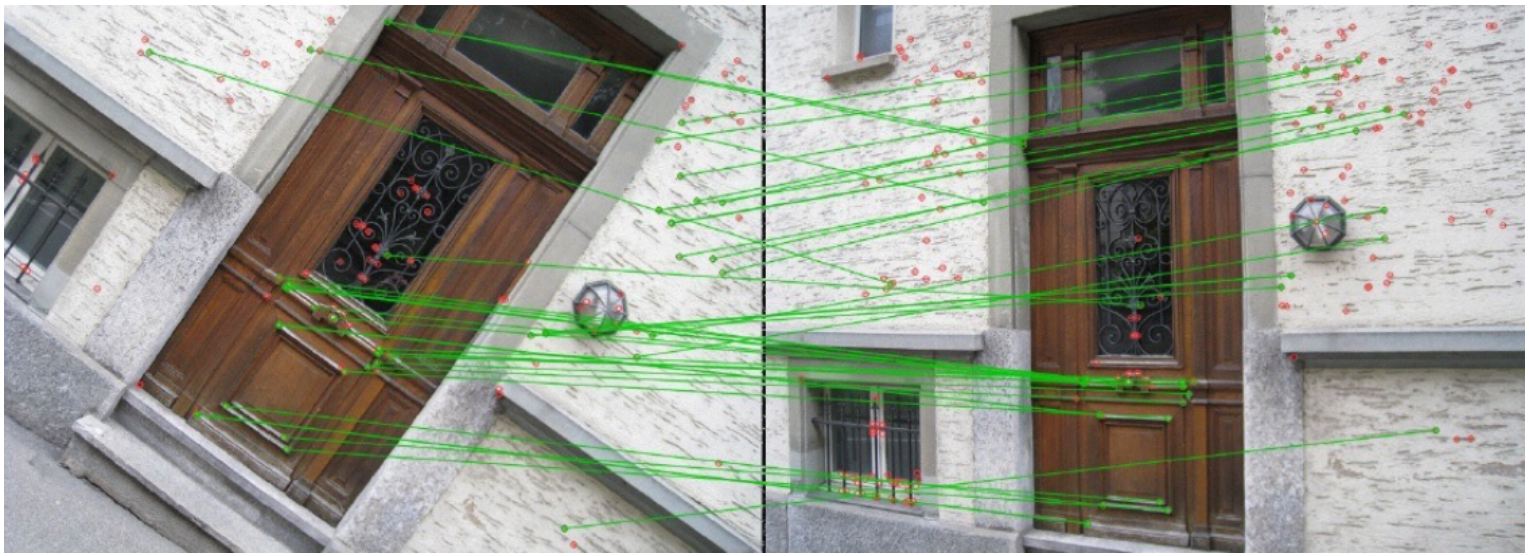












Eddie Bell @ Lyst

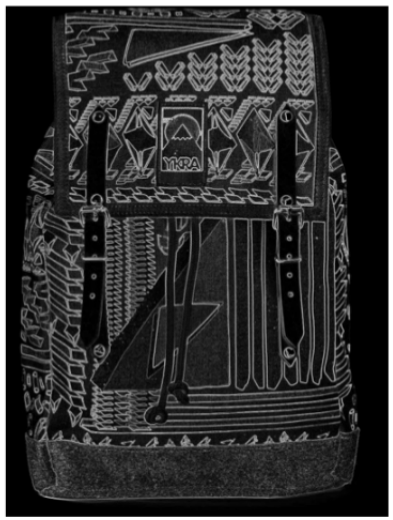
Eddie Bell @ Lyst

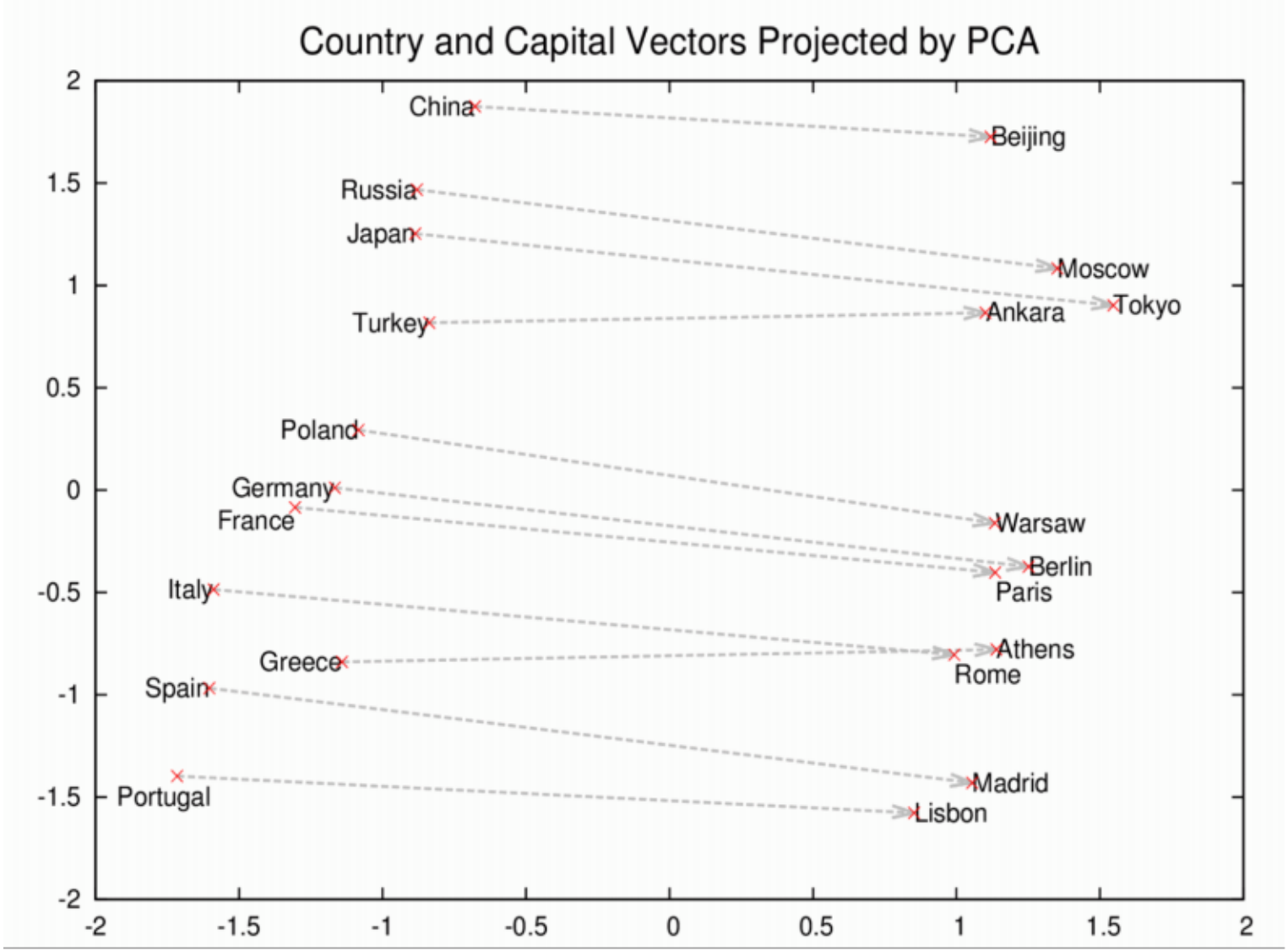
Eddie Bell @ Lyst

Eddie Bell @ Lyst

-1	0	+1
-2	0	+2
-1	0	+1

-1	-2	-1
0	0	0
+1	+2	+1





	Term	Similarity
	"shift"	0.933104
	"gown"	0.887743
	"skirt"	0.881672
	"bandage"	0.880162
	"midi"	0.869786

Similar to ‘dress’

Eddie Bell @ Lyst

a group of young girls standing next to each other on the beach

A clock tower with a clock on top of it

Google?

A bunch of bananas hanging from a tree

Google?

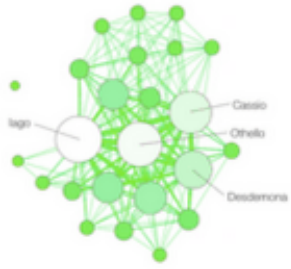
TITUS ANDRONICUS
Number of characters 36 | 50% Network density

ROMEO AND JULIET
Number of characters 41 | 37% Network density

JULIUS CAESAR
Number of characters 46 | 34% Network density

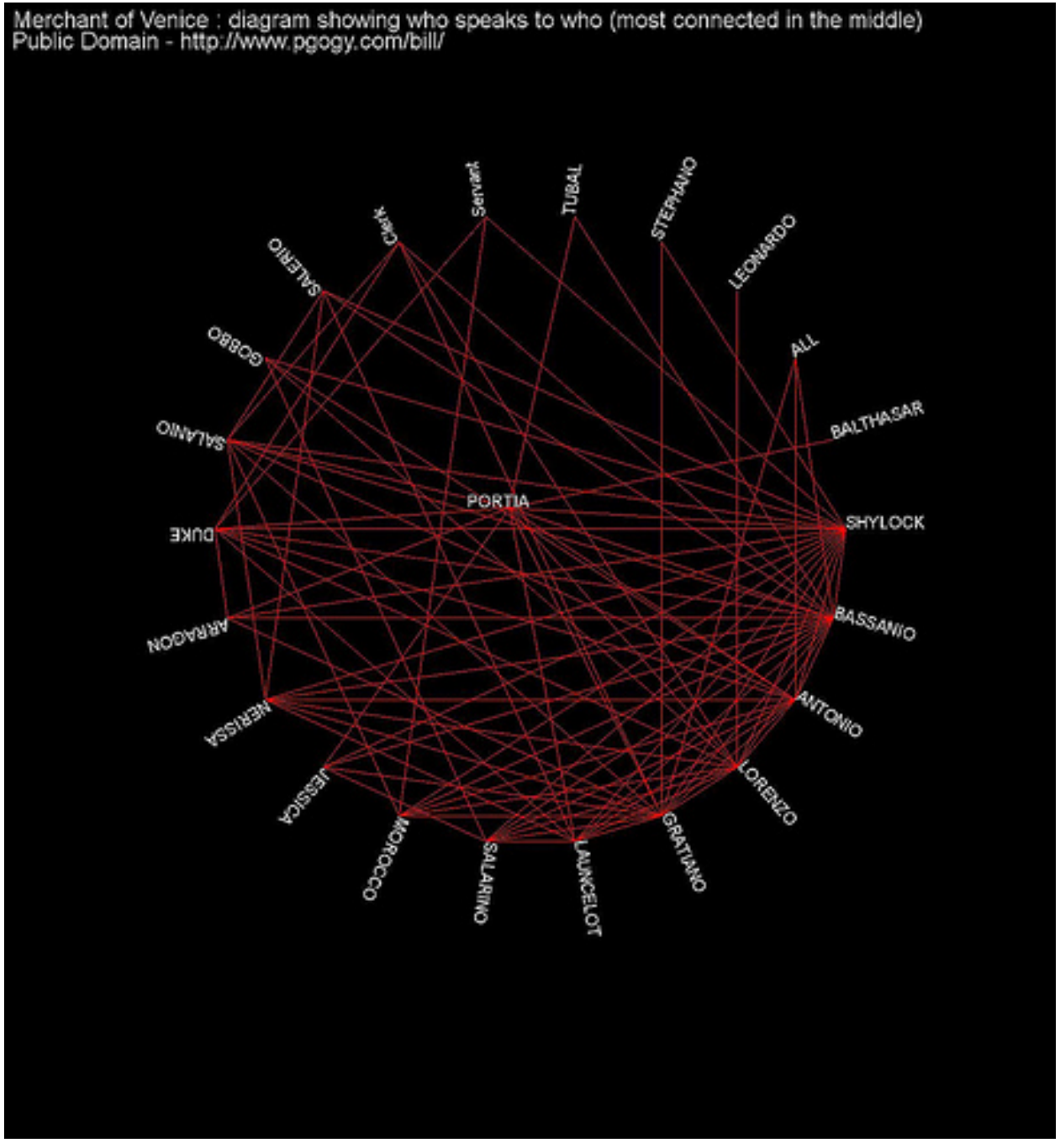
HAMLET
Number of characters 37 | 39% Network density

TROILUS AND CRESSIDA
Number of characters 35 | 40% Network density

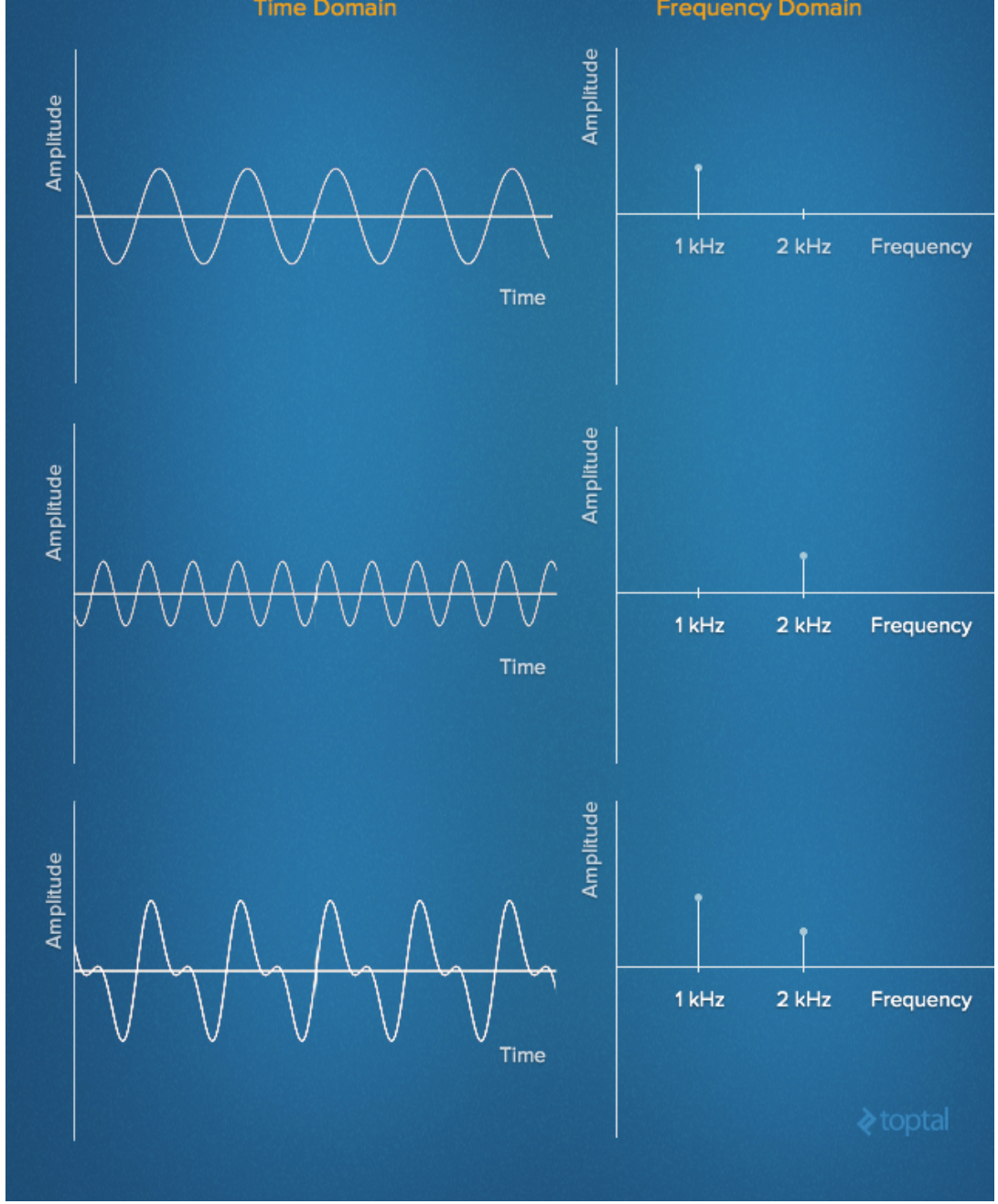


OTHELLO
Number of characters 24 | 55% Network density

https://flowingdata.com/2015/12/30/shakespeare-tragedies-as-network-graphs/



http://www.nand.io/visualisation/understanding-shakespeare



http://www.toptal.com/algorithms/

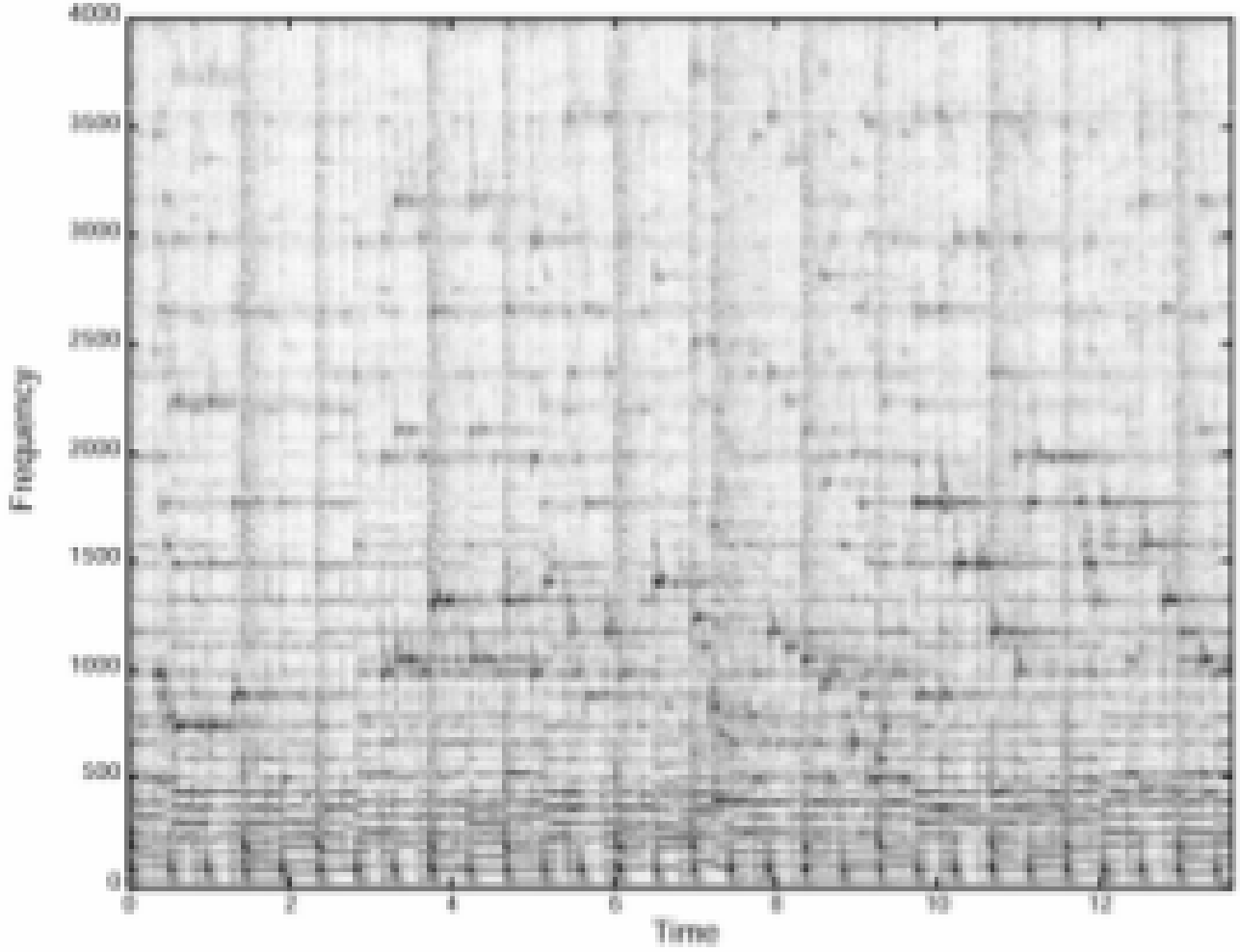


Fig. 1A - Spectrogram

https://www.ee.columbia.edu/~dpwe/papers/Wang03-shazam.pdf

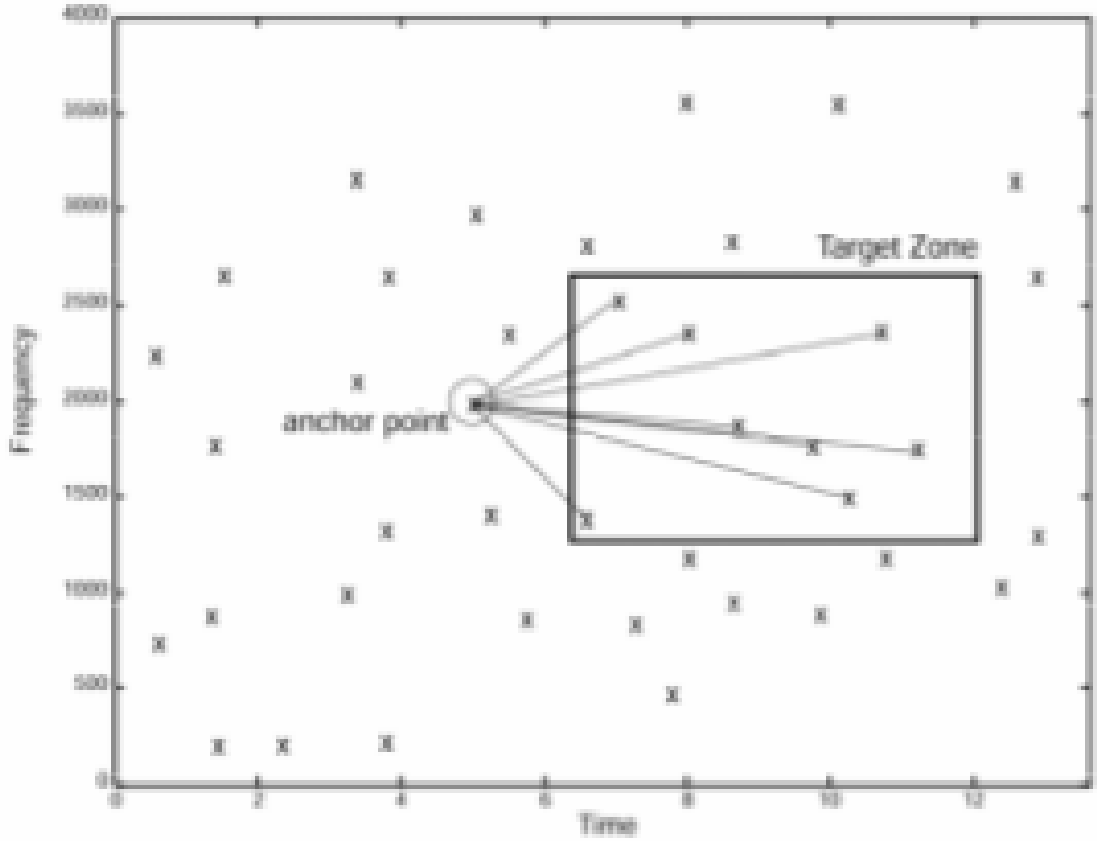


Fig. 1C - Combinatorial Hash Generation

https://www.ee.columbia.edu/~dpwe/papers/Wang03-shazam.pdf

Google

Big data?

http://aadrake.com/command-line-tools-can-be-235x-faster-than-your-hadoop-cluste
html

Some maths. . .

- 100 features
- int *(4 bytes per feature)*
- 16 GB RAM
- $O(n \log n)$ in RAM

$$400N + 400N \log(400N) < \frac{16 \cdot 10^9}{400}$$

Some maths...

- 100 features
- int *(4 bytes per feature)*
- 16 GB RAM
- $O(n \log n)$ in RAM

$$400N + 400N \log(400N) < \frac{16 \cdot 10^9}{400}$$

$$N \approx 2 \cdot 10^6$$

Questions?