CLUSTERR
GROUPING SIMILAR NEWSPAPER ARTICLES

CHANJUN (CJ) PARK

RUOYI (ZOE) CAI

Project 72

DATA

“Transmittable and Storable Computer Information”

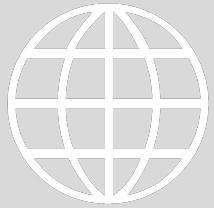
BIG DATA

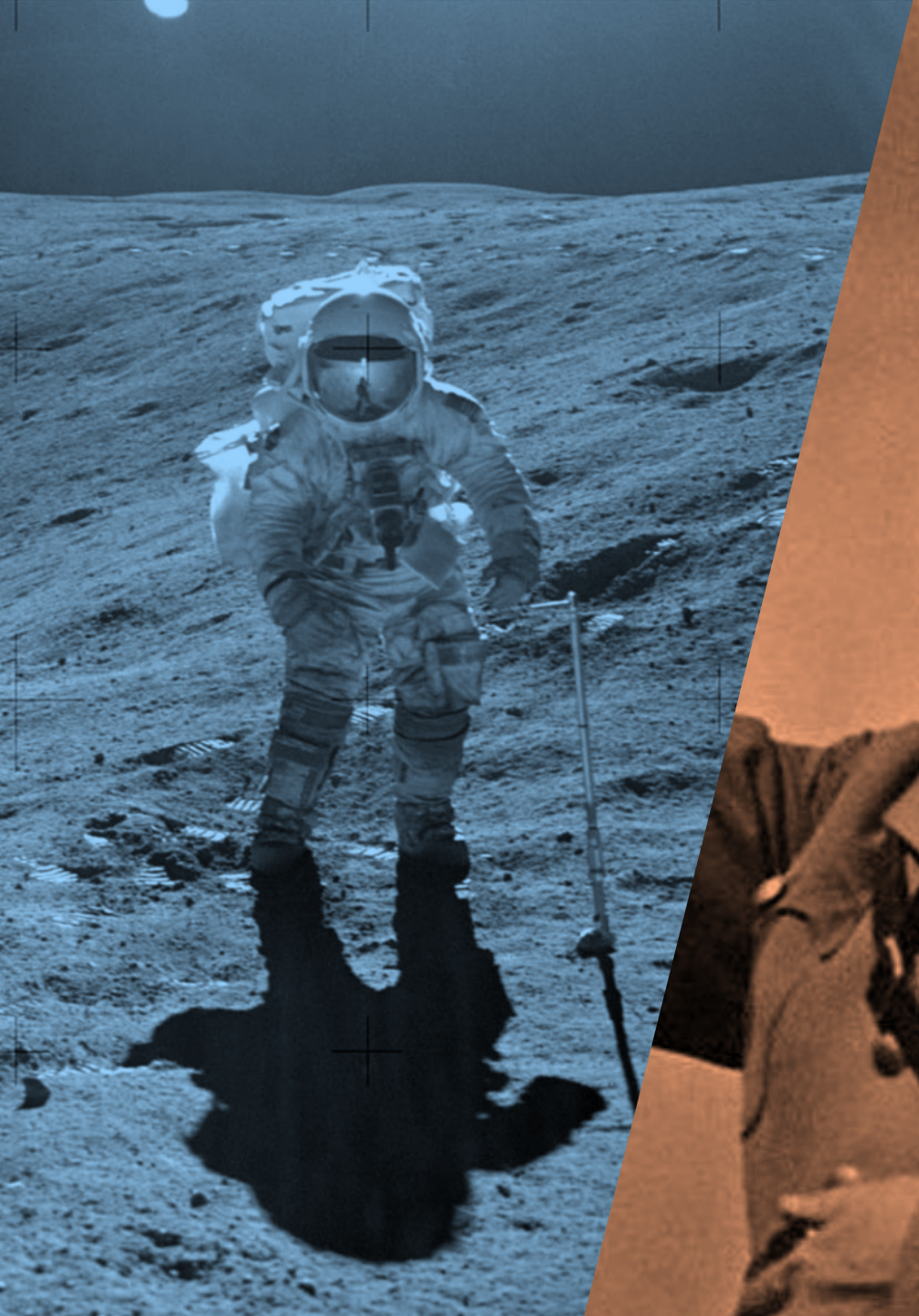
INTERNET OF THINGS

59BC

1605AD

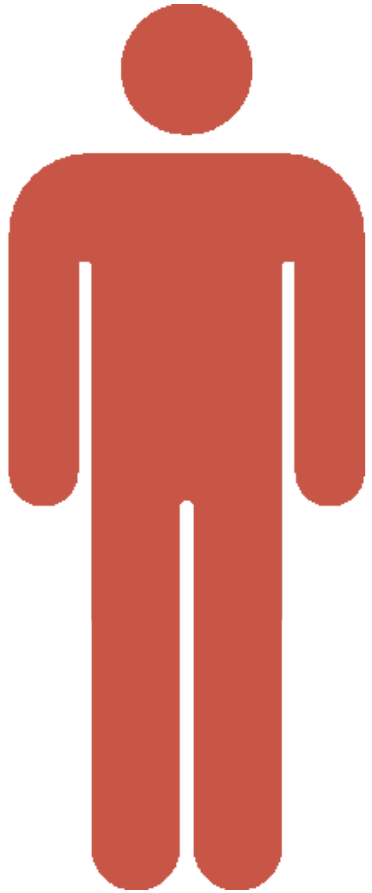
1991





So What is Missing?

END USER?



clusterr

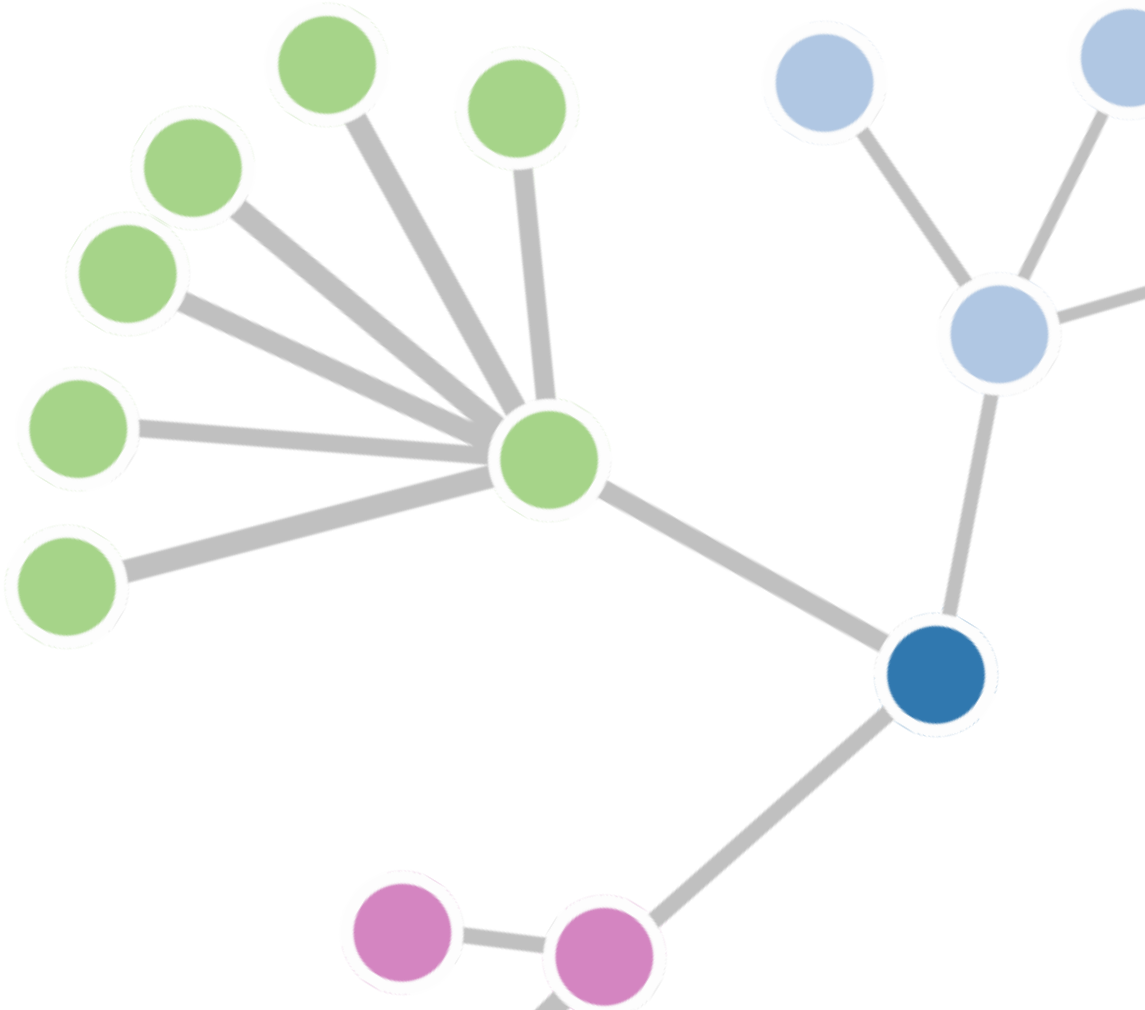

clusterr.

Related Work

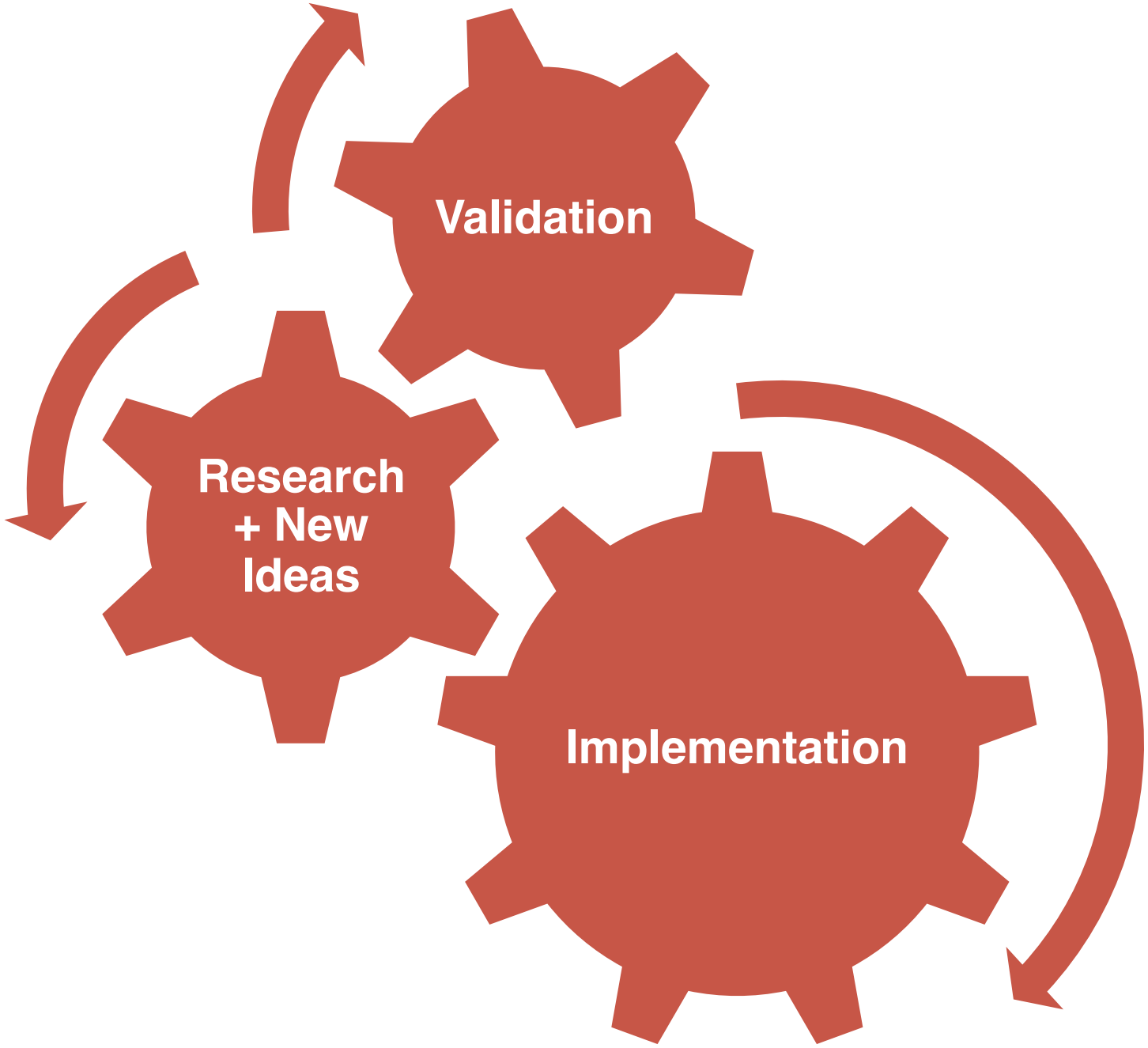
Google Trends

	Interactive User Interface	Aesthetic Design	Visualization of Clusters	Standardization of Input
carrot²	No	No	Somewhat	Yes
Google Trends	Yes	Yes	No	No
clusterr	Yes	Yes	Yes	Yes

However, clusterr is still a proof of concept.



Methodology
&
Technical Concepts


Validation
Research
+ New
Ideas
Implementation

Breaking the problem into small steps

Dataset

20 Newsgroups

Preprocessing

Stop Words
Lemmatisation

TF-IDF

“Blue”
tf = 1/5
idf = log(2/2) = 0

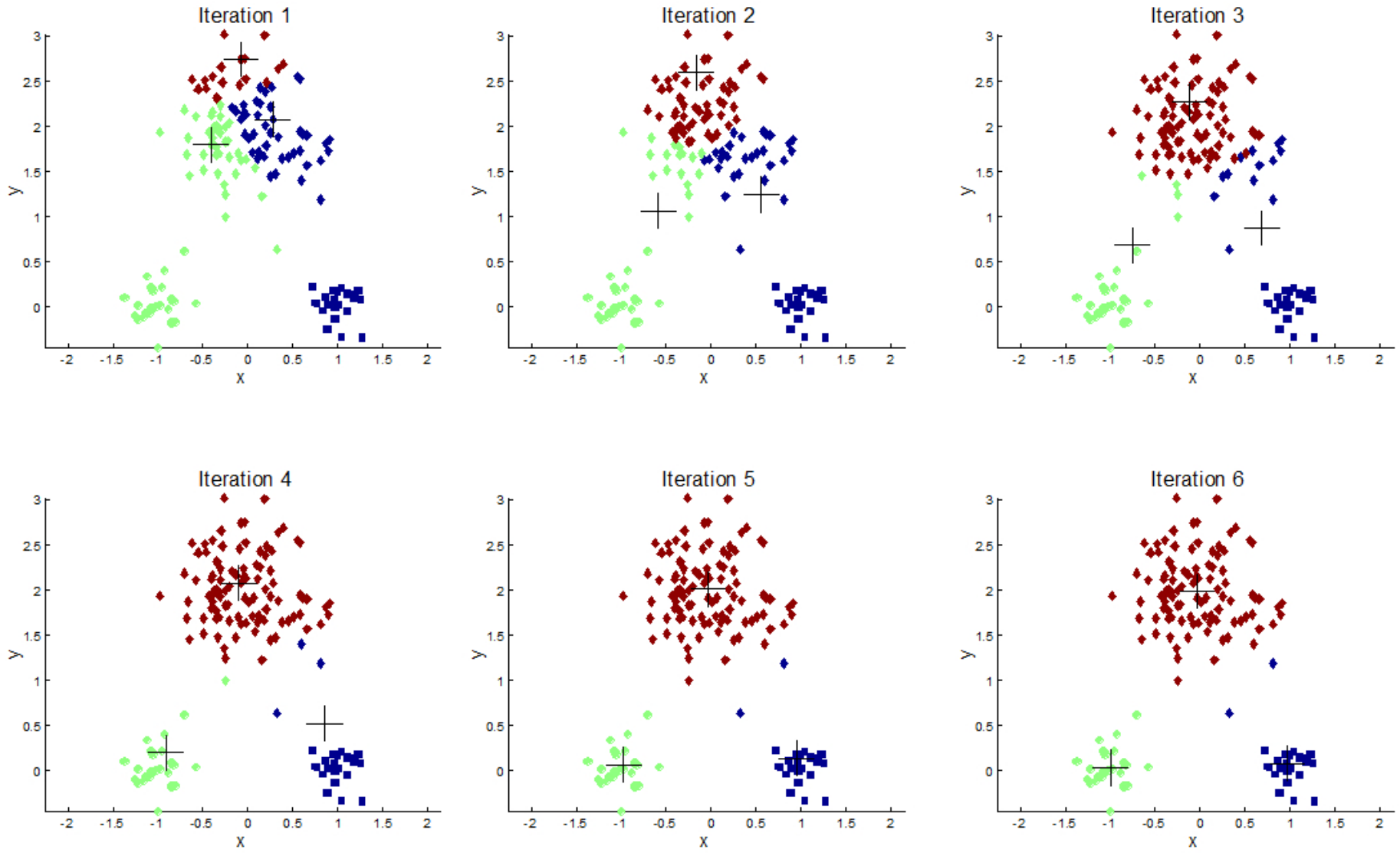
“Blue”
tf = 1/7
idf = log(2/2) = 0

Document 1		**Document 2**	
Term	**Term Count**	**Term**	**Term Count**
Blue	1	Blue	1
shirt	1	skirt	1
elephant	2	another	2
sample	1	example	3
Total	**5**	**Total**	**7**

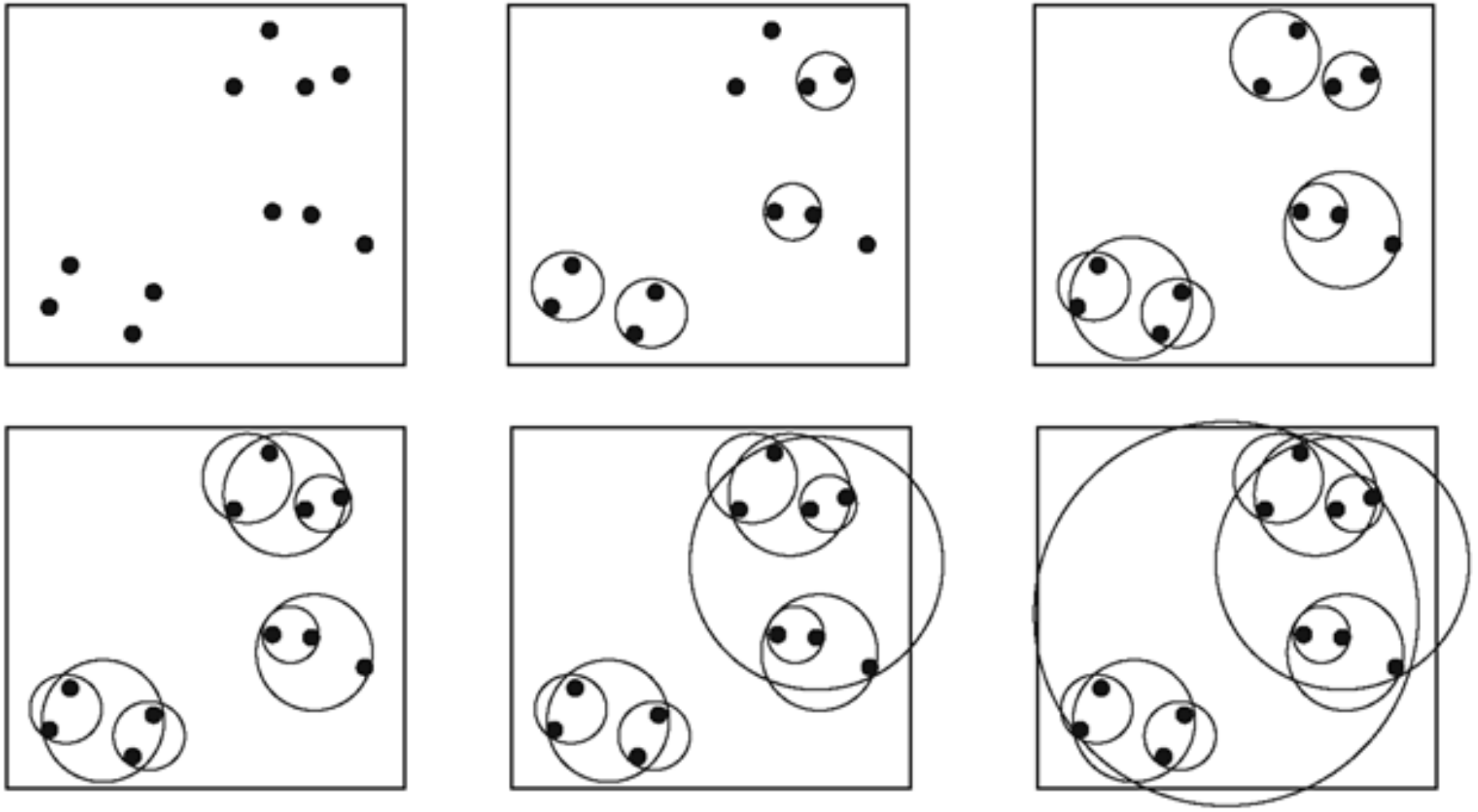
“example”
tf = 3/7
idf = log(2/1) = 0.301
tf-idf = 0.129

Clustering

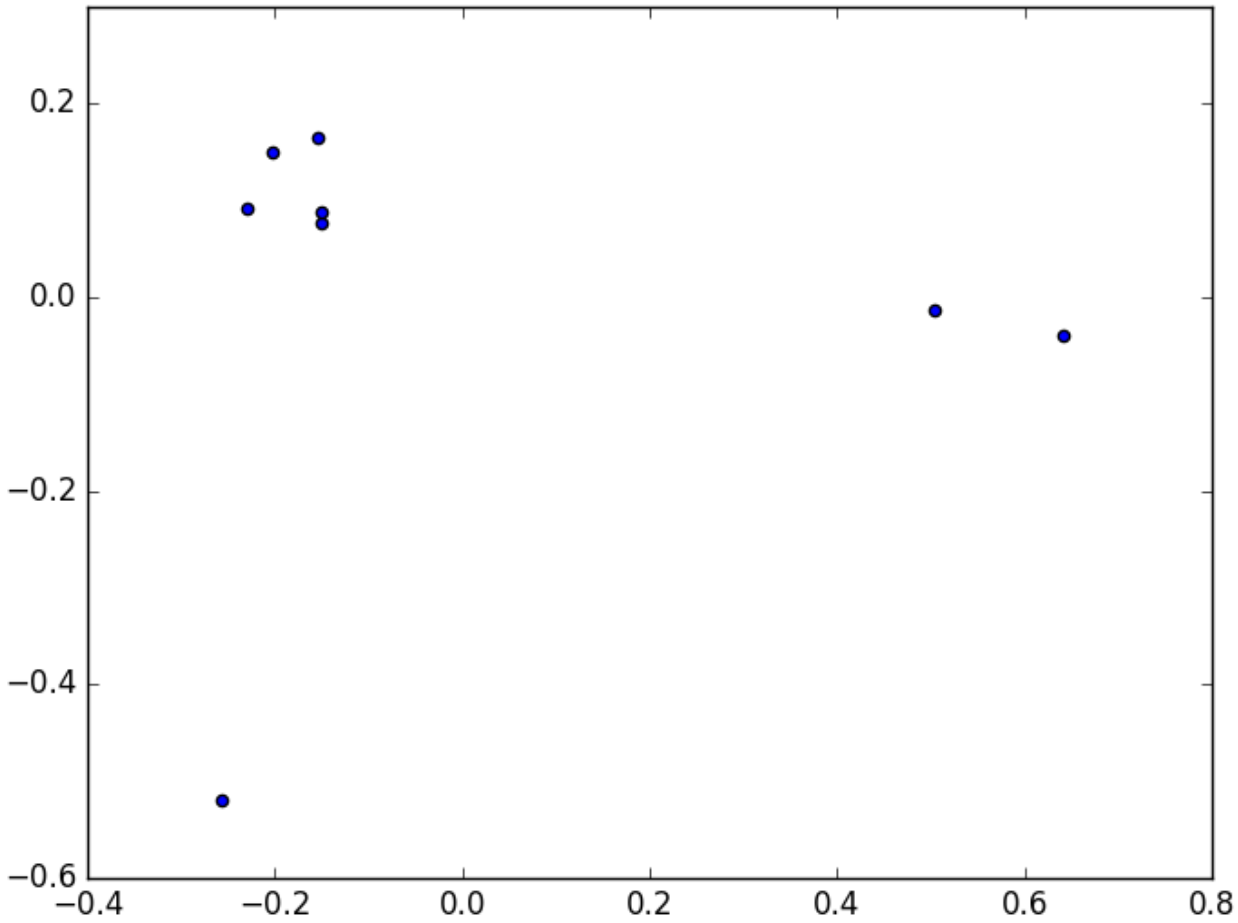
CLUSTERING – K-MEANS



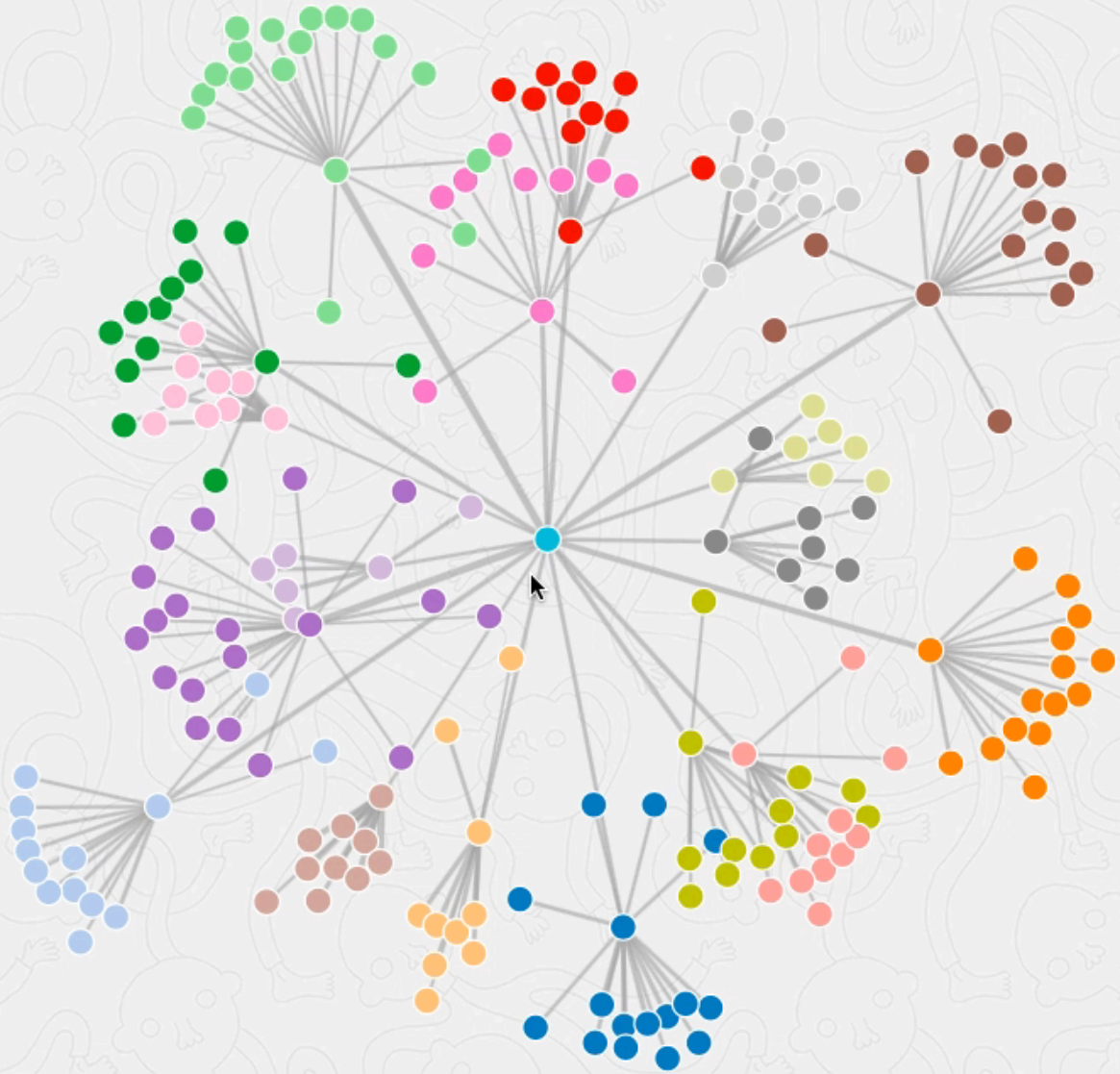
CLUSTERING – AGGLOMERATIVE



VISUALISATION



Primitive &
Not User
Friendly



FUTURE WORK

Use semantic meaning of the words
Look into other clustering algorithms
Add other visualization and graphs

LEARNED

Machine Learning
Full stack web development

Newspaper articles are quite hard to cluster!

CHALLENGES

Incorrect mappings bug
Making a web app with no prior experience

Not perfect, but it’s better than humans!

Try it yourself!

clusterr.zoecai.com

(be gentle though, it’s a free server and has processor limitations)