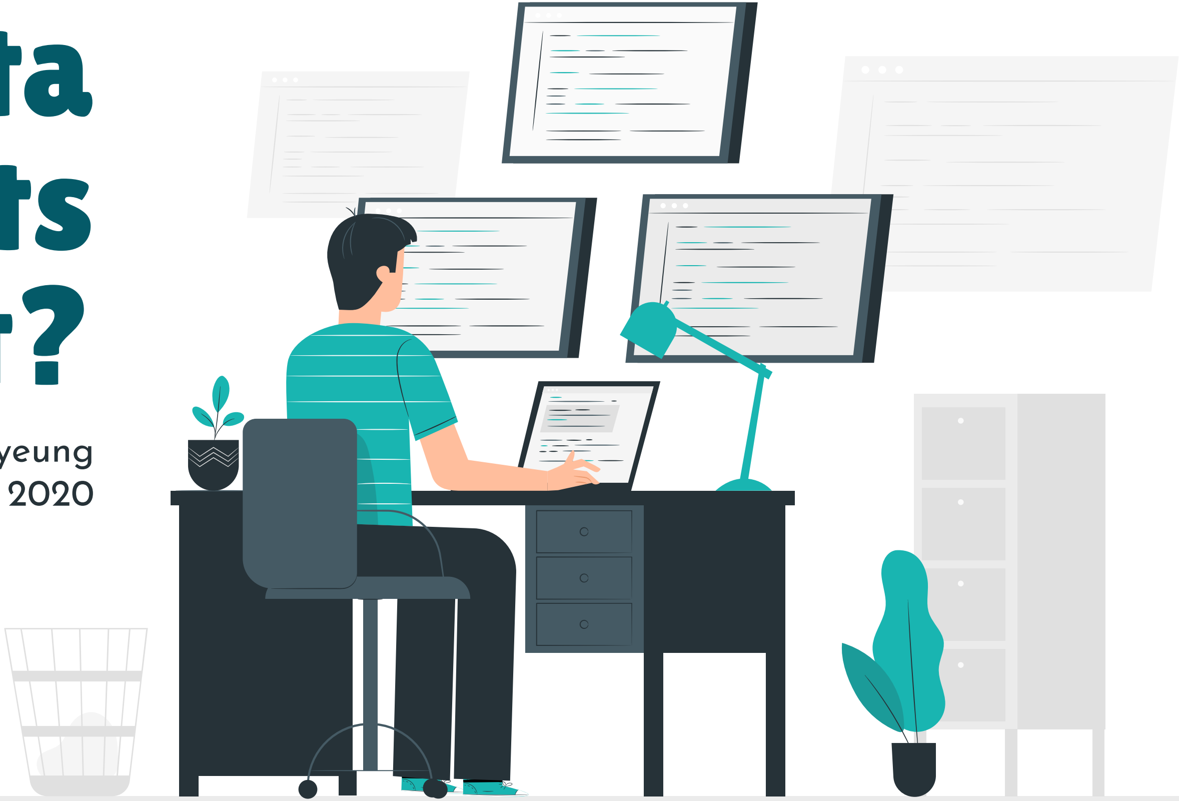
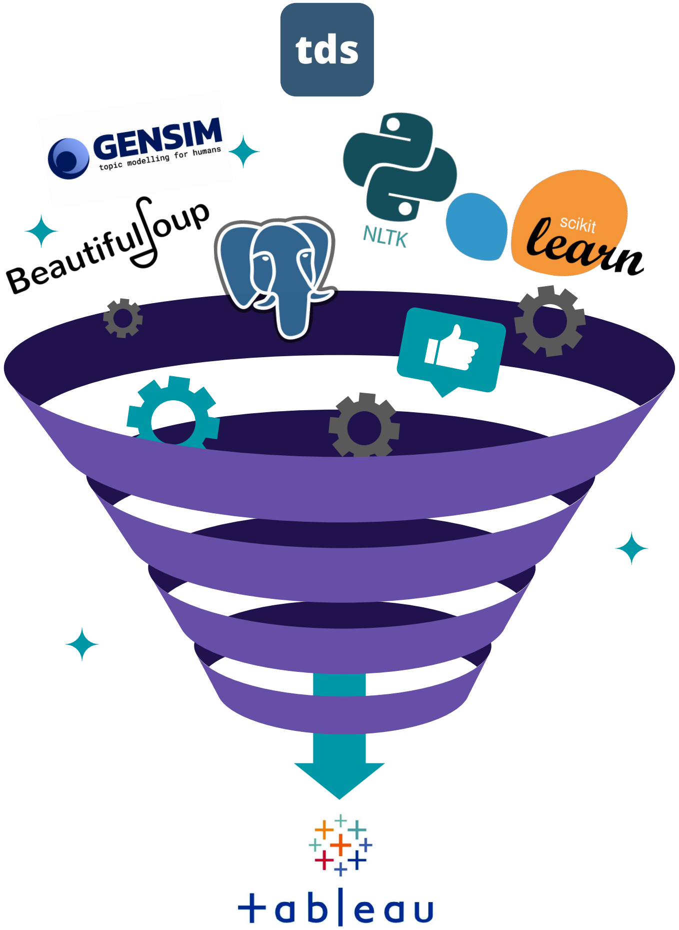
# What do Data Scientists talk about?

Andrew Auyeung
Metis Fall 2020





# Project Pipeline

35000 Articles

- BeautifulSoup Web scraping
- PostgreSQL Database

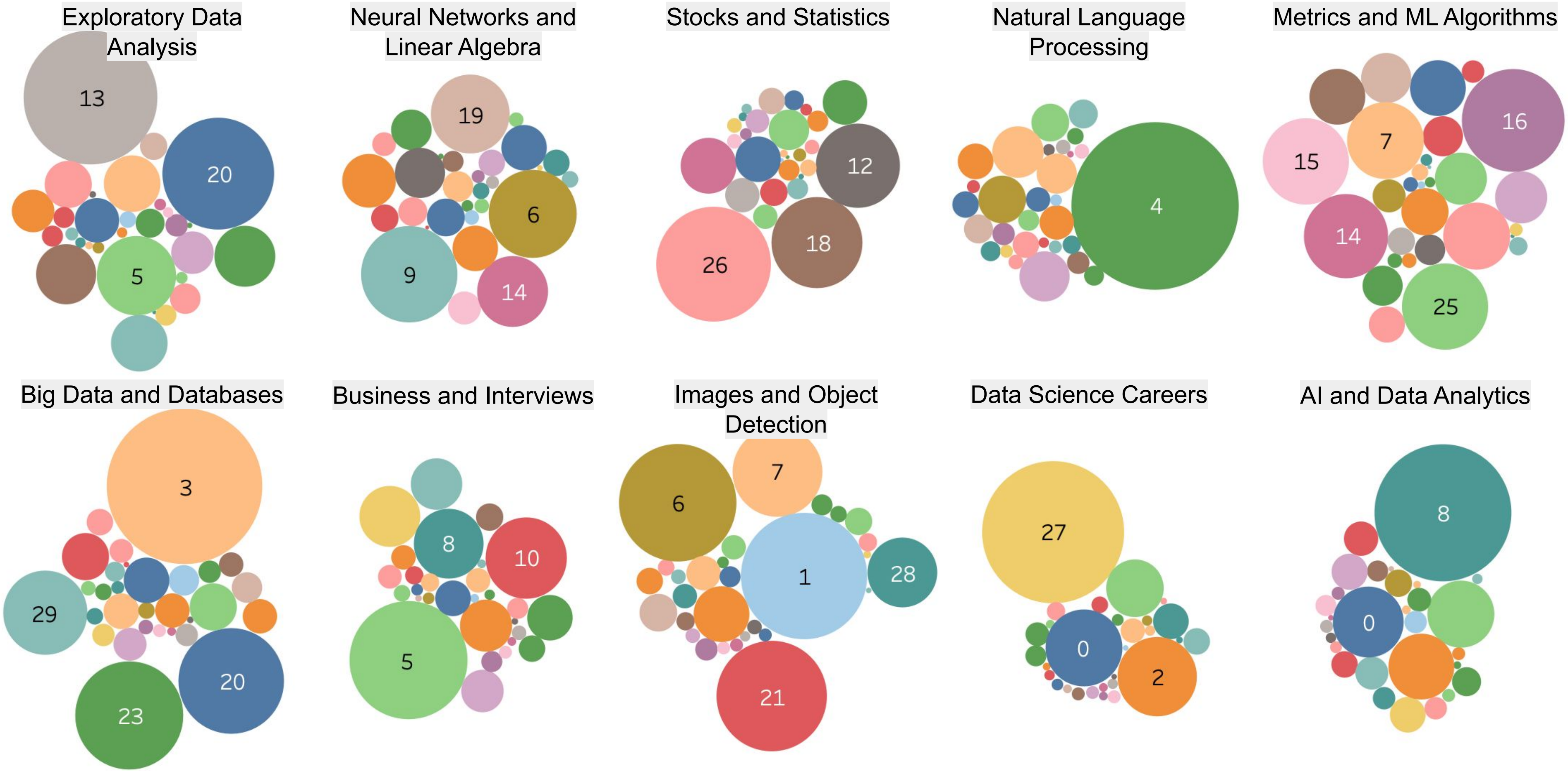
30 Unique Topics

- Non-negative Matrix Factorization
- Groups similar words into topics

10 Clusters

- Gensim Doc2Vec
- KMeans Clustering

# Topic Distribution Per Cluster



Exploratory Data Analysis

Neural Networks and Linear Algebra

Stocks and Statistics

Natural Language Processing

Metrics and ML Algorithms

Big Data and Databases

Business and Interviews

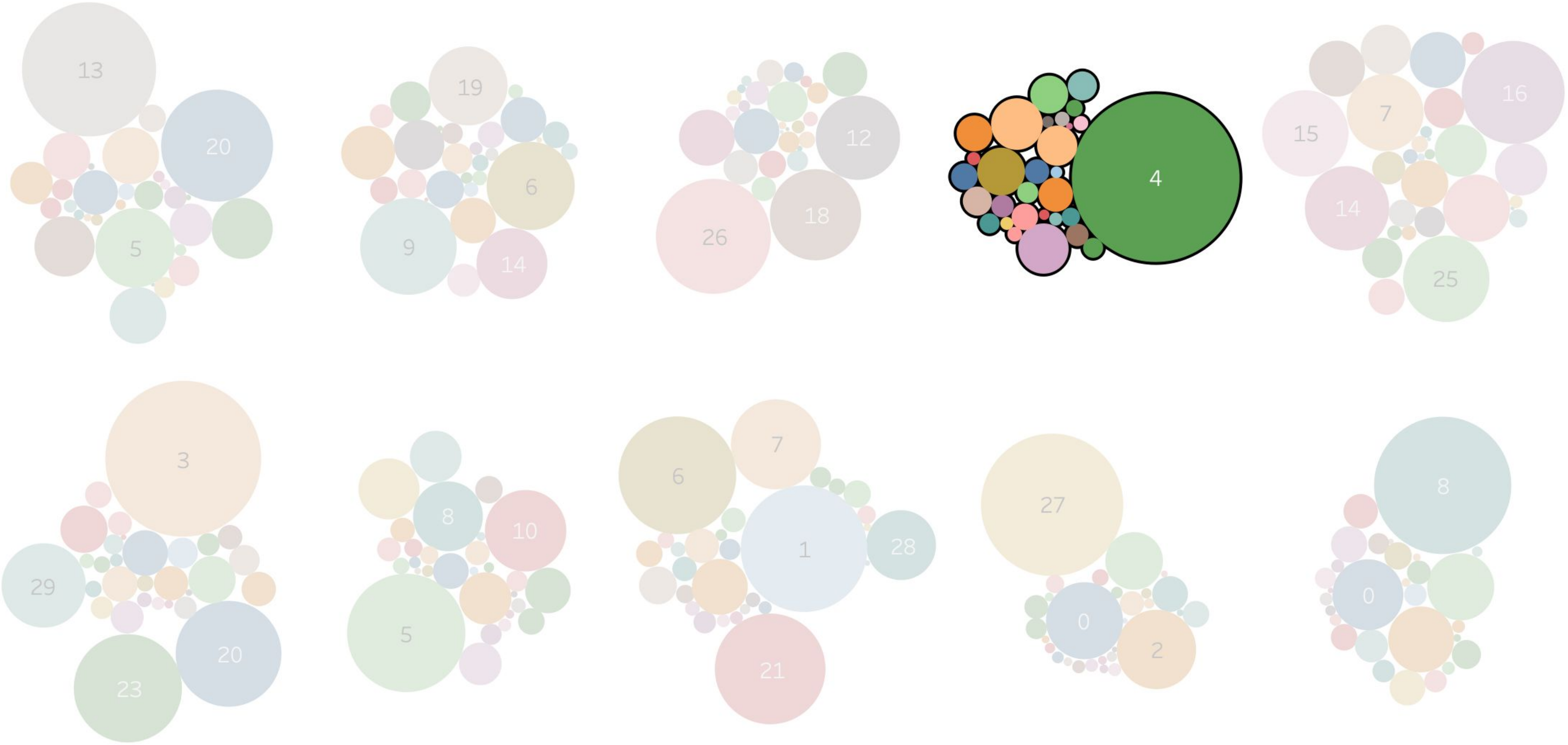
Images and Object Detection

Data Science Careers

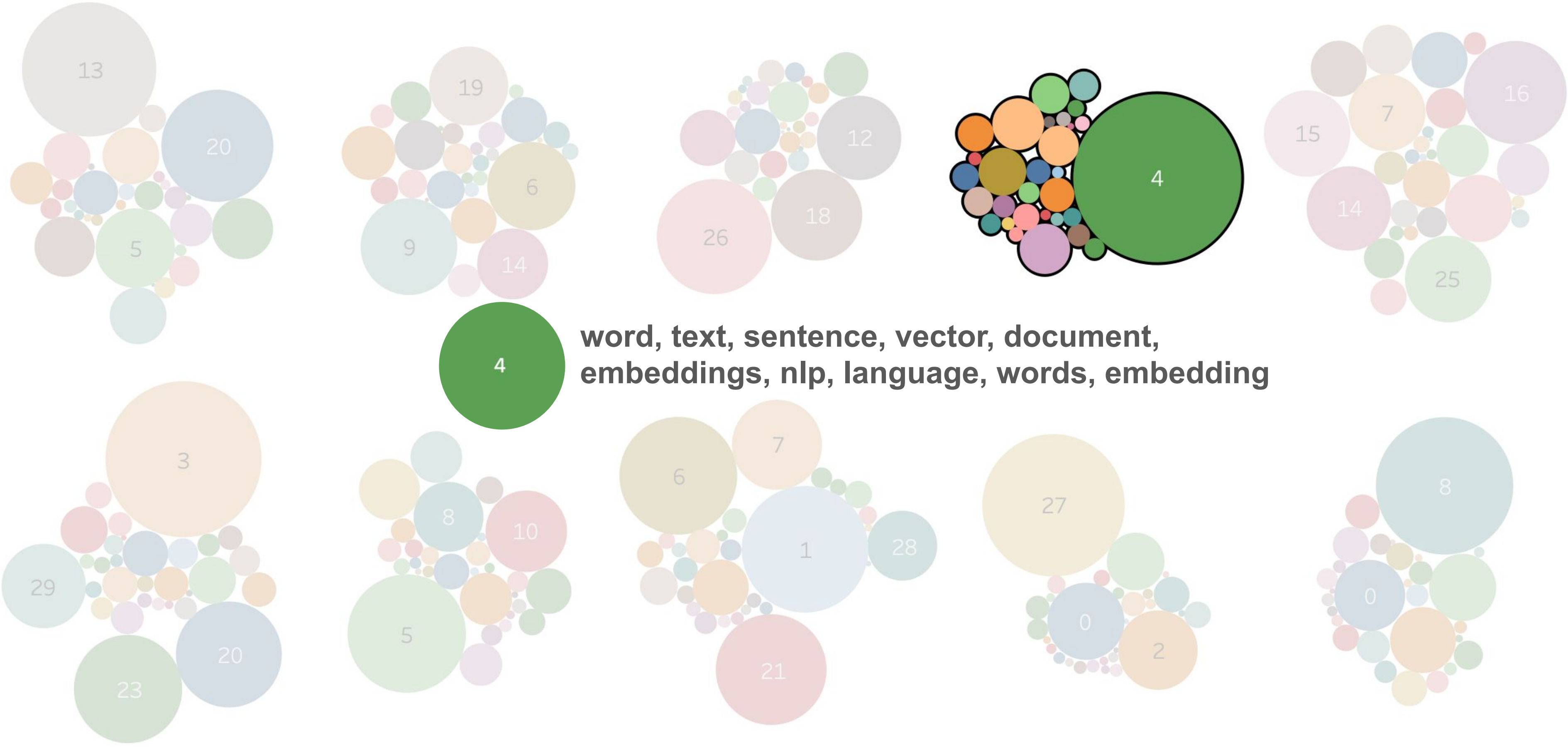
AI and Data Analytics

* Words associated with Topics can be found in the appendix of this presentation

# Topic Distribution Per Cluster

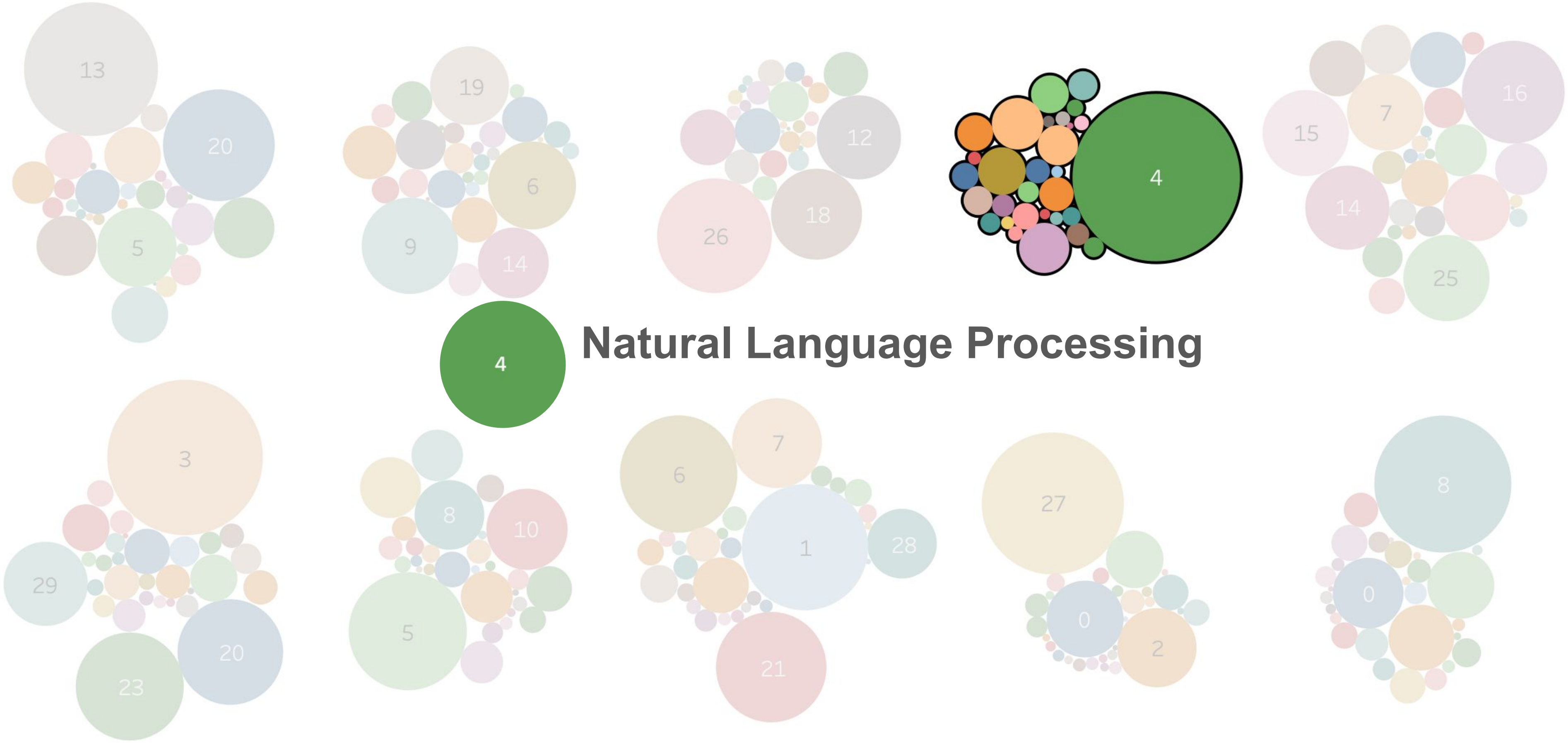


# Topic Distribution Per Cluster



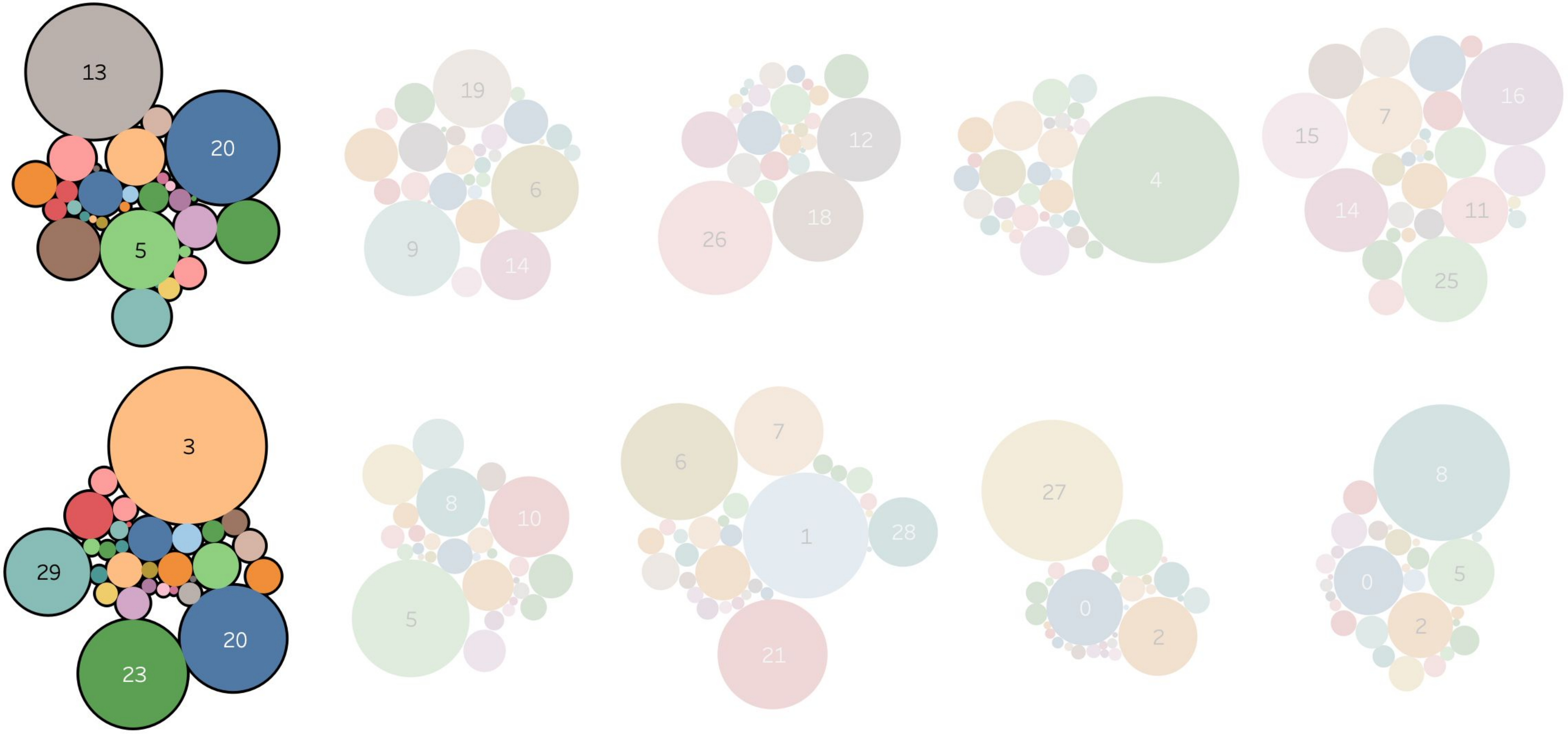
word, text, sentence, vector, document, embeddings, nlp, language, words, embedding

# Topic Distribution Per Cluster

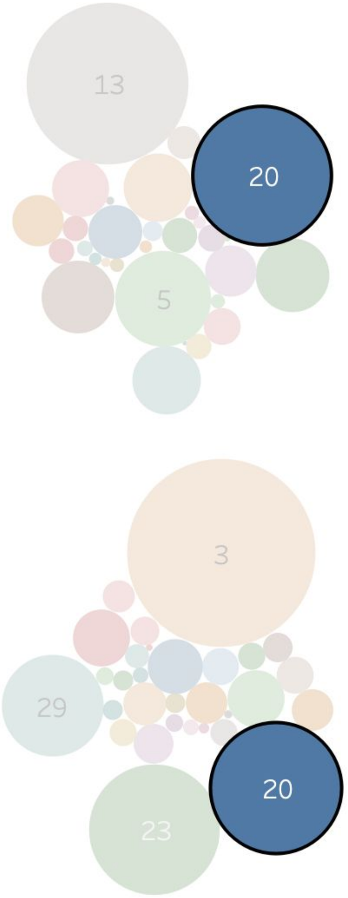


Natural Language Processing

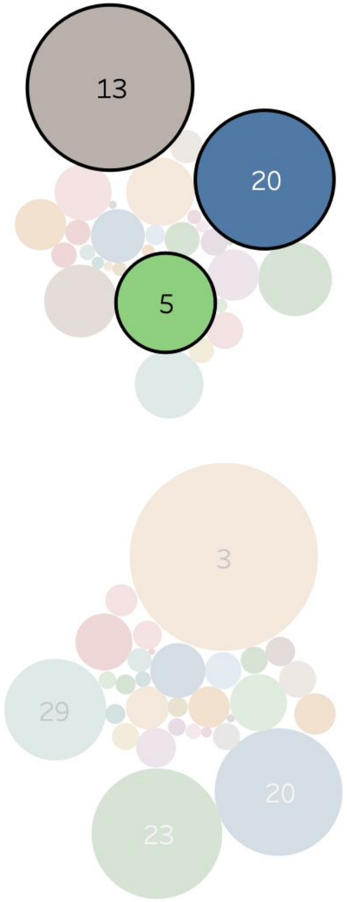
# Topic Distribution Per Cluster



# Topic Distribution Per Cluster

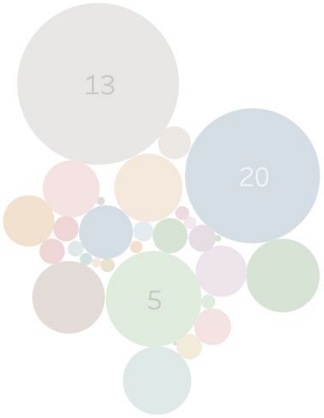


## Topic Distribution Per Cluster

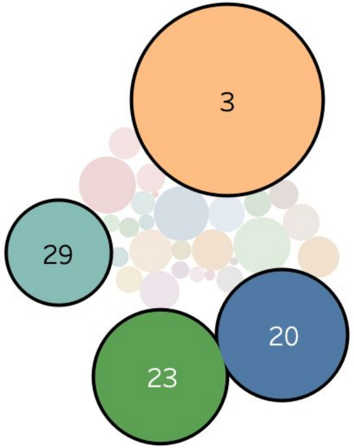


# Exploratory Data Analysis

# Topic Distribution Per Cluster

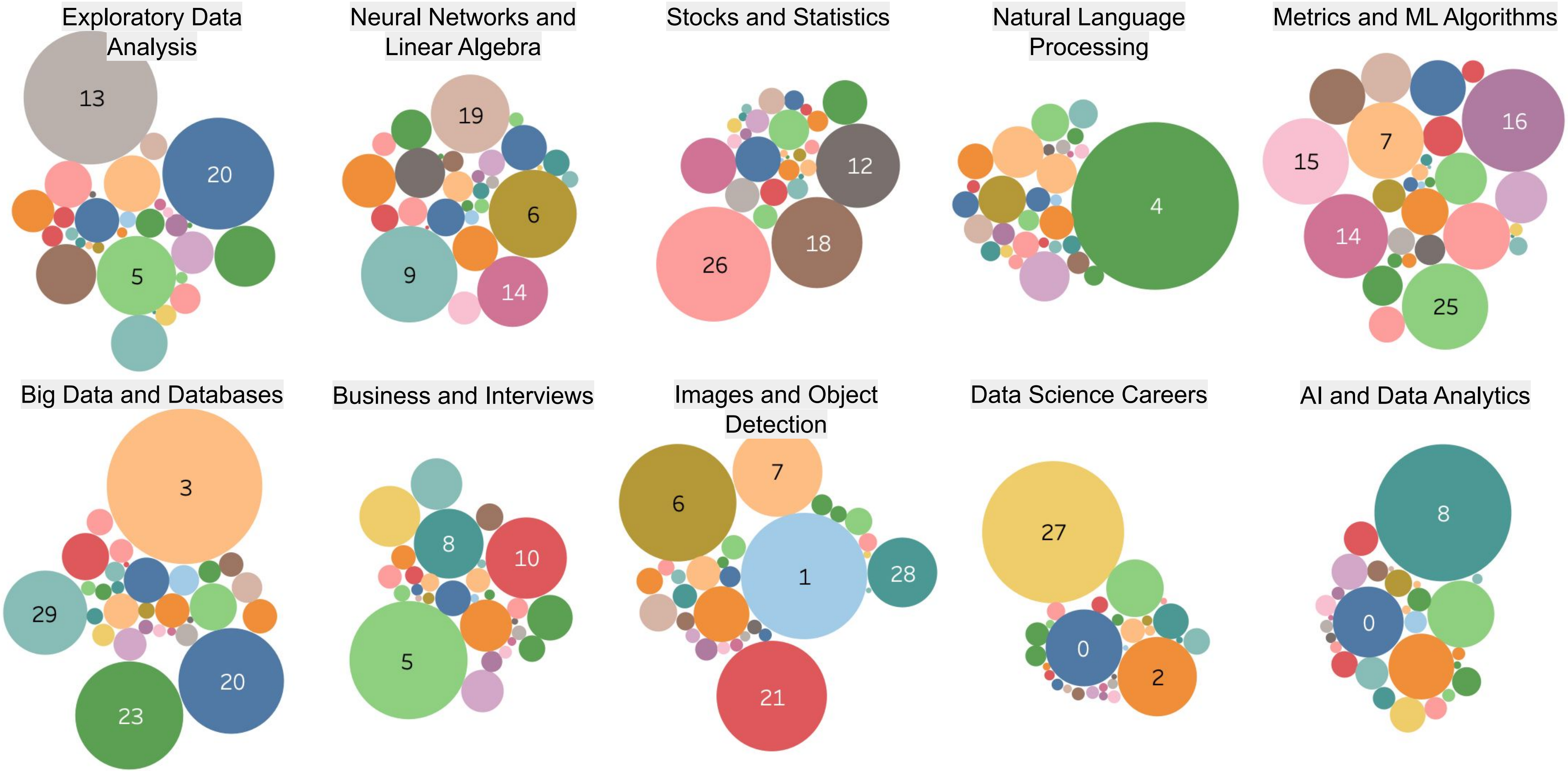


Exploratory Data Analysis



Big Data and Databases

# Topic Distribution Per Cluster



Exploratory Data Analysis

Neural Networks and Linear Algebra

Stocks and Statistics

Natural Language Processing

Metrics and ML Algorithms

Big Data and Databases

Business and Interviews

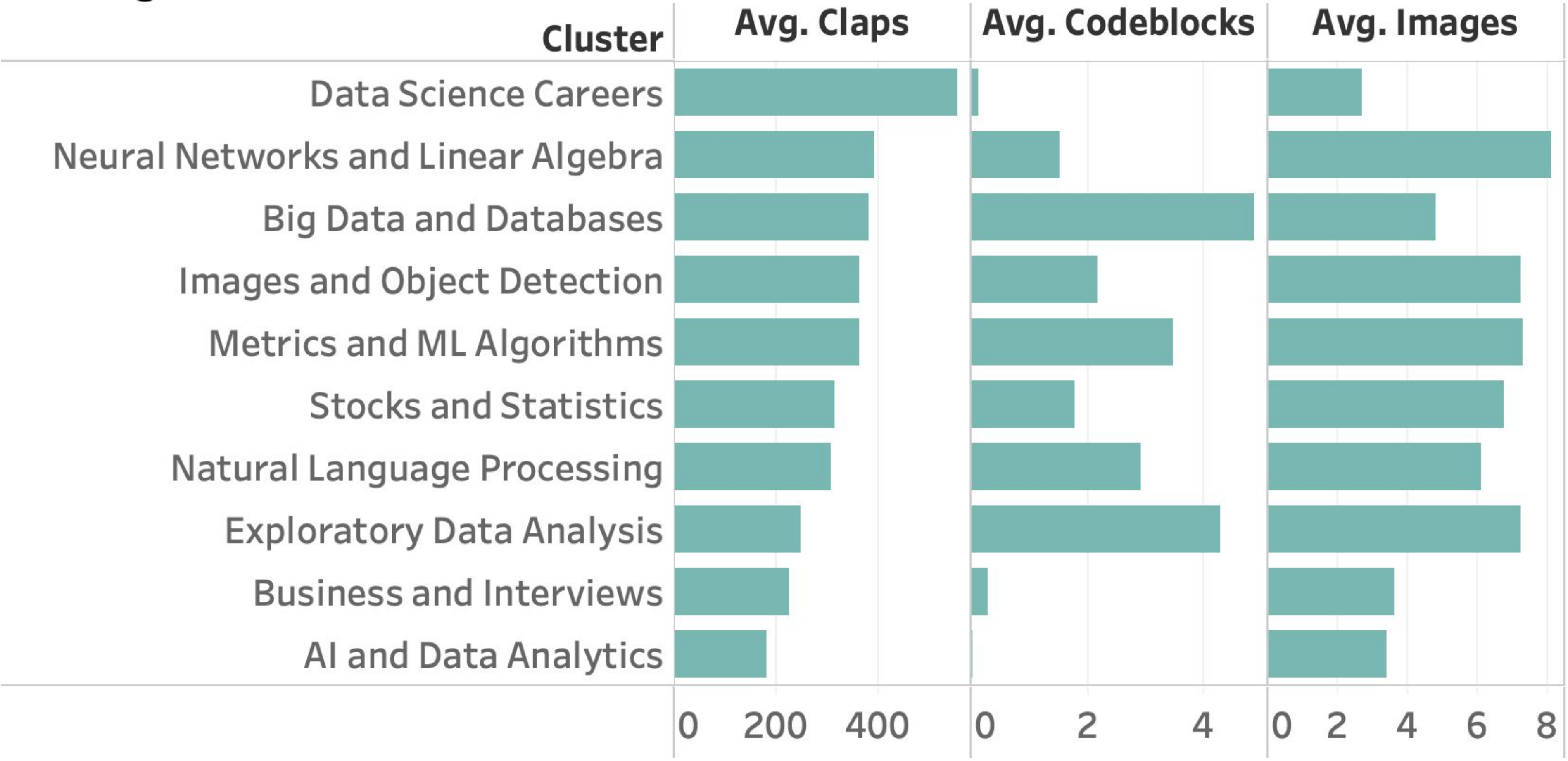
Images and Object Detection

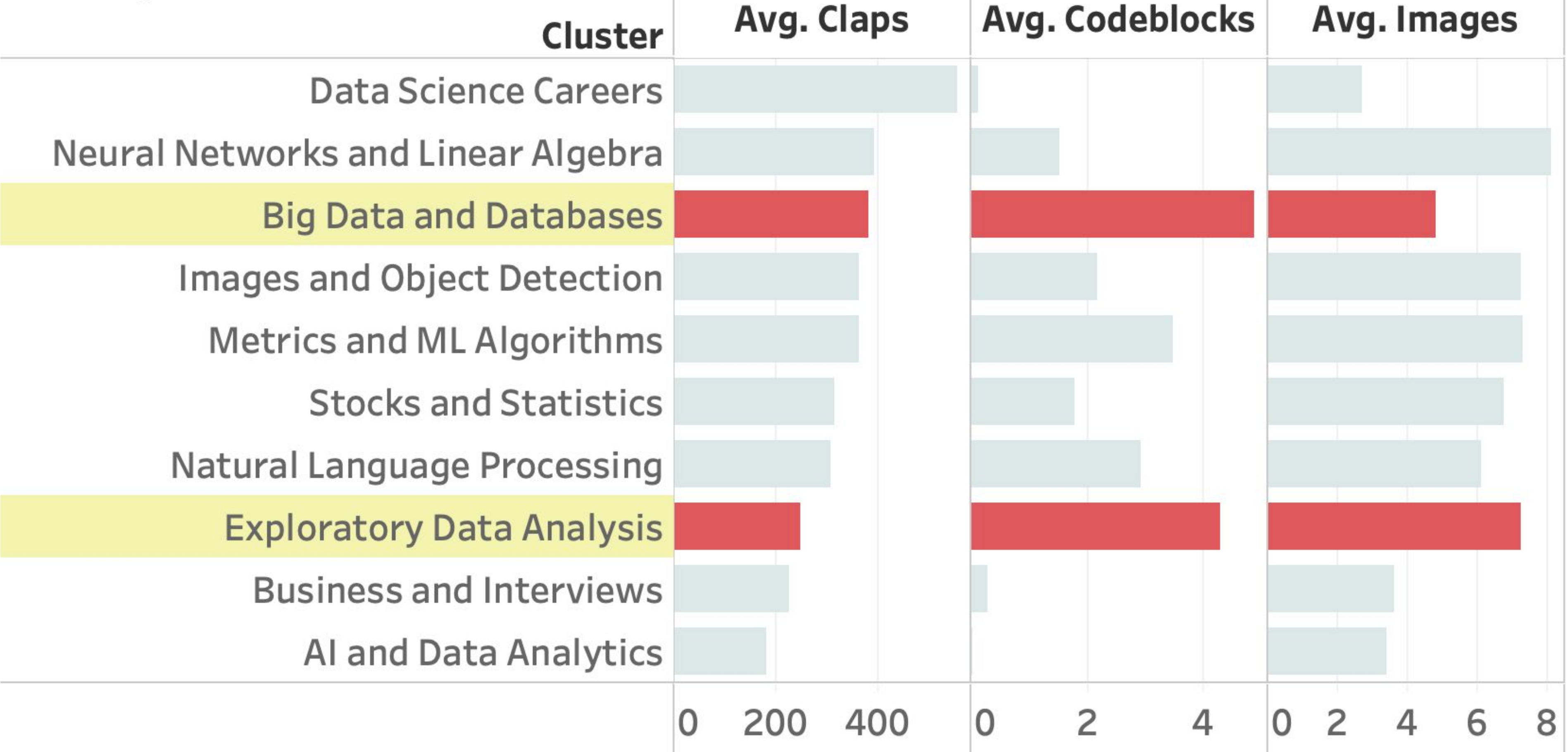
Data Science Careers

AI and Data Analytics

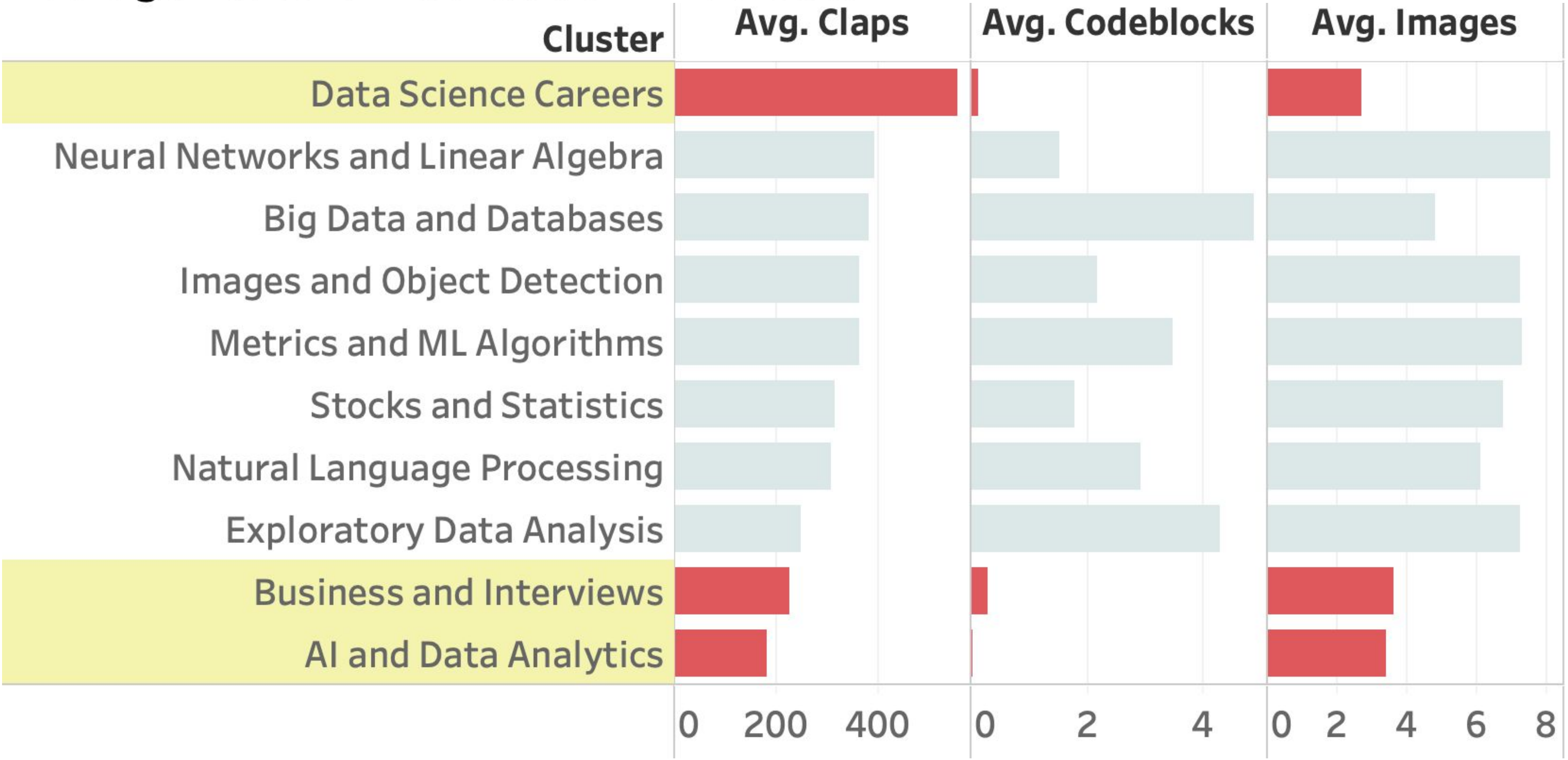
* Words associated with Topics can be found in the appendix of this presentation

Average Feature Distribution in a Cluster



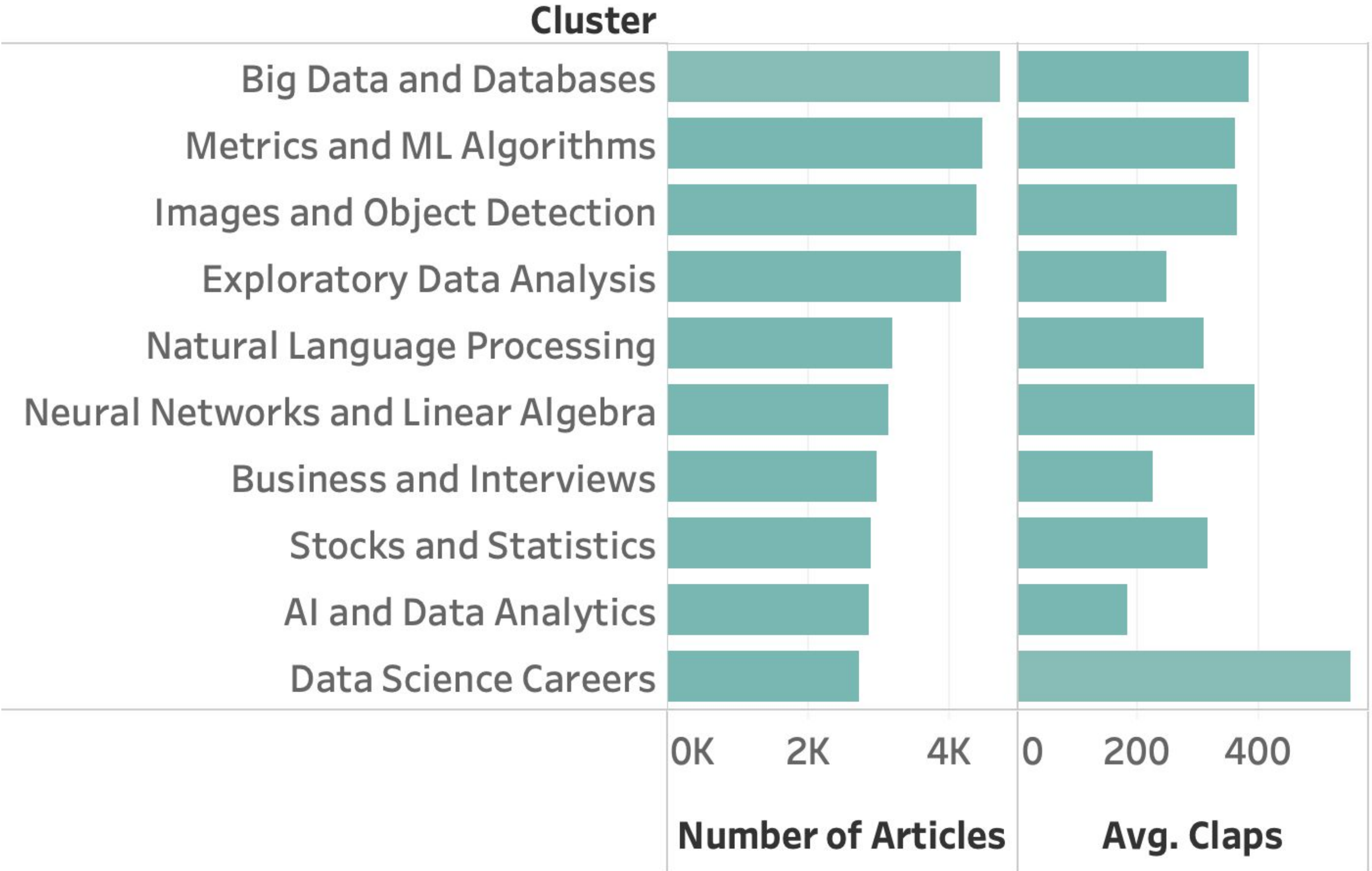


Average Feature Distribution in a Cluster

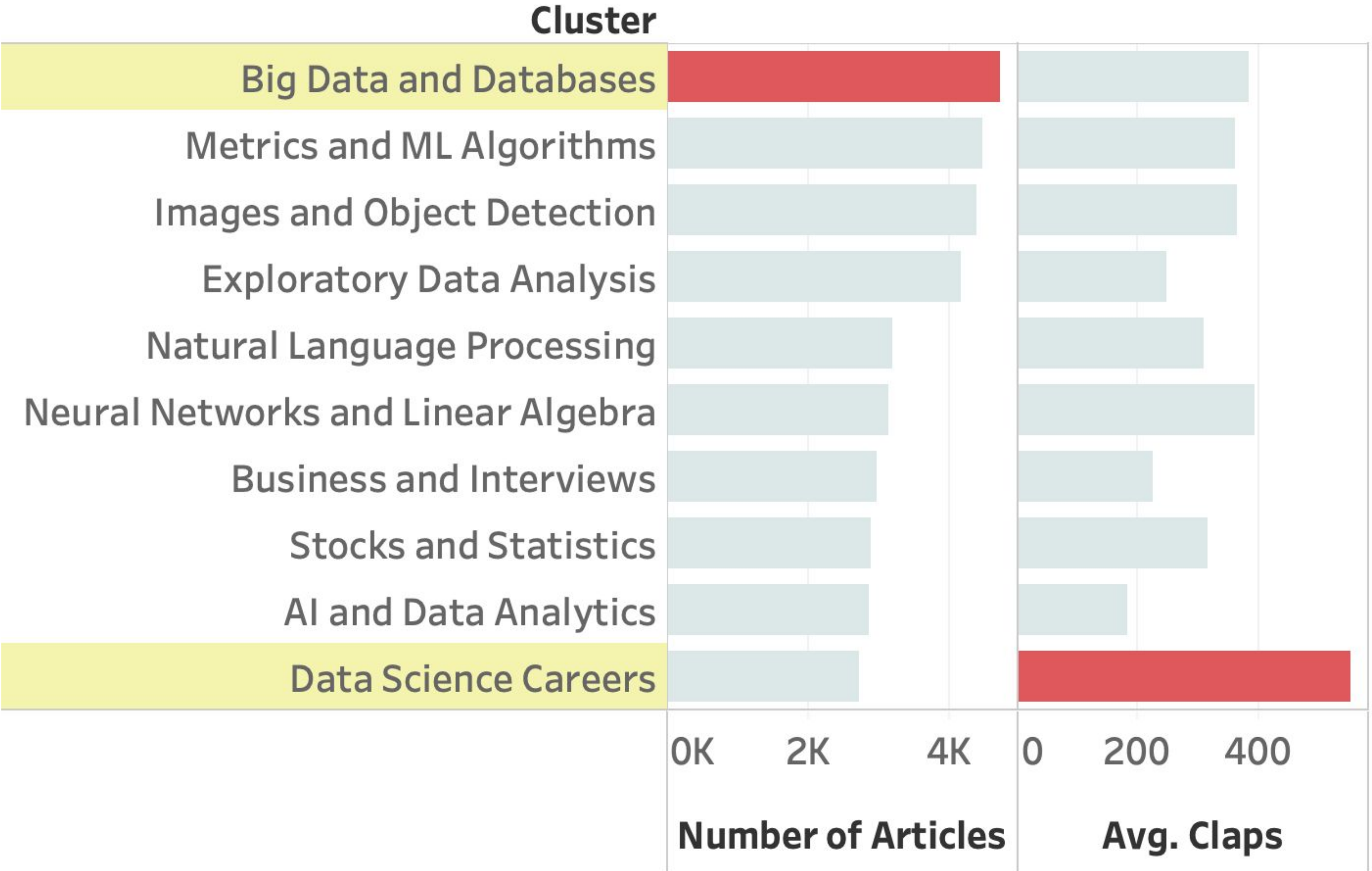


Average Feature Distribution in a Cluster

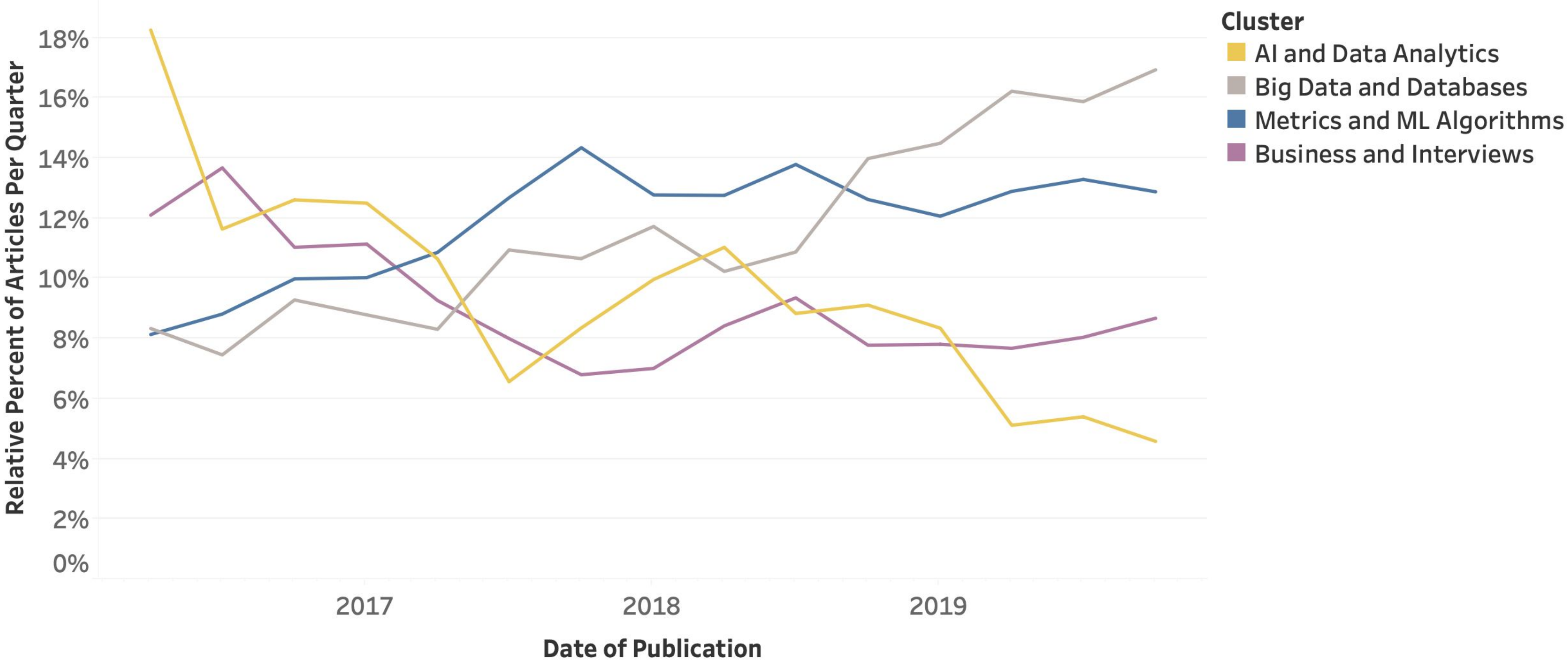
# Cluster Popularity

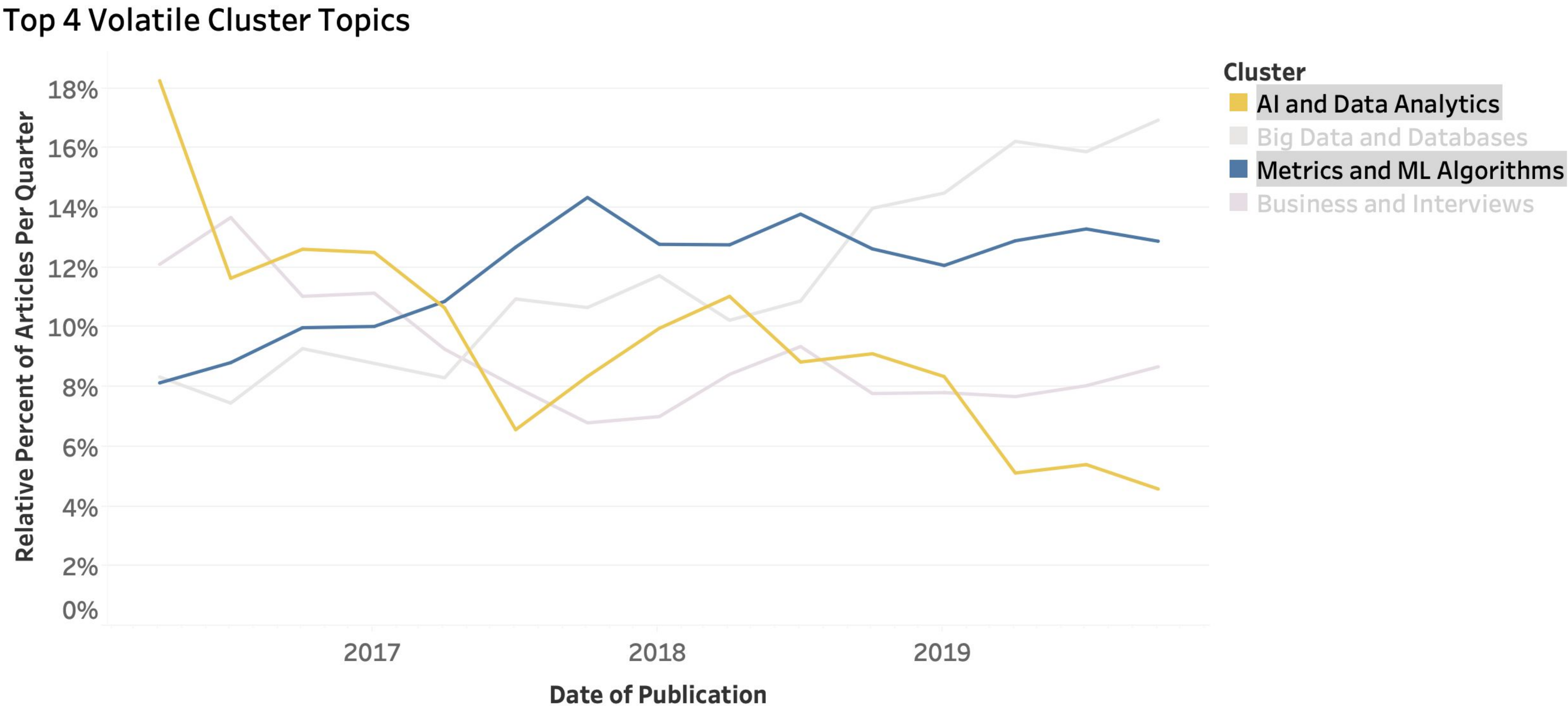


# Cluster Popularity

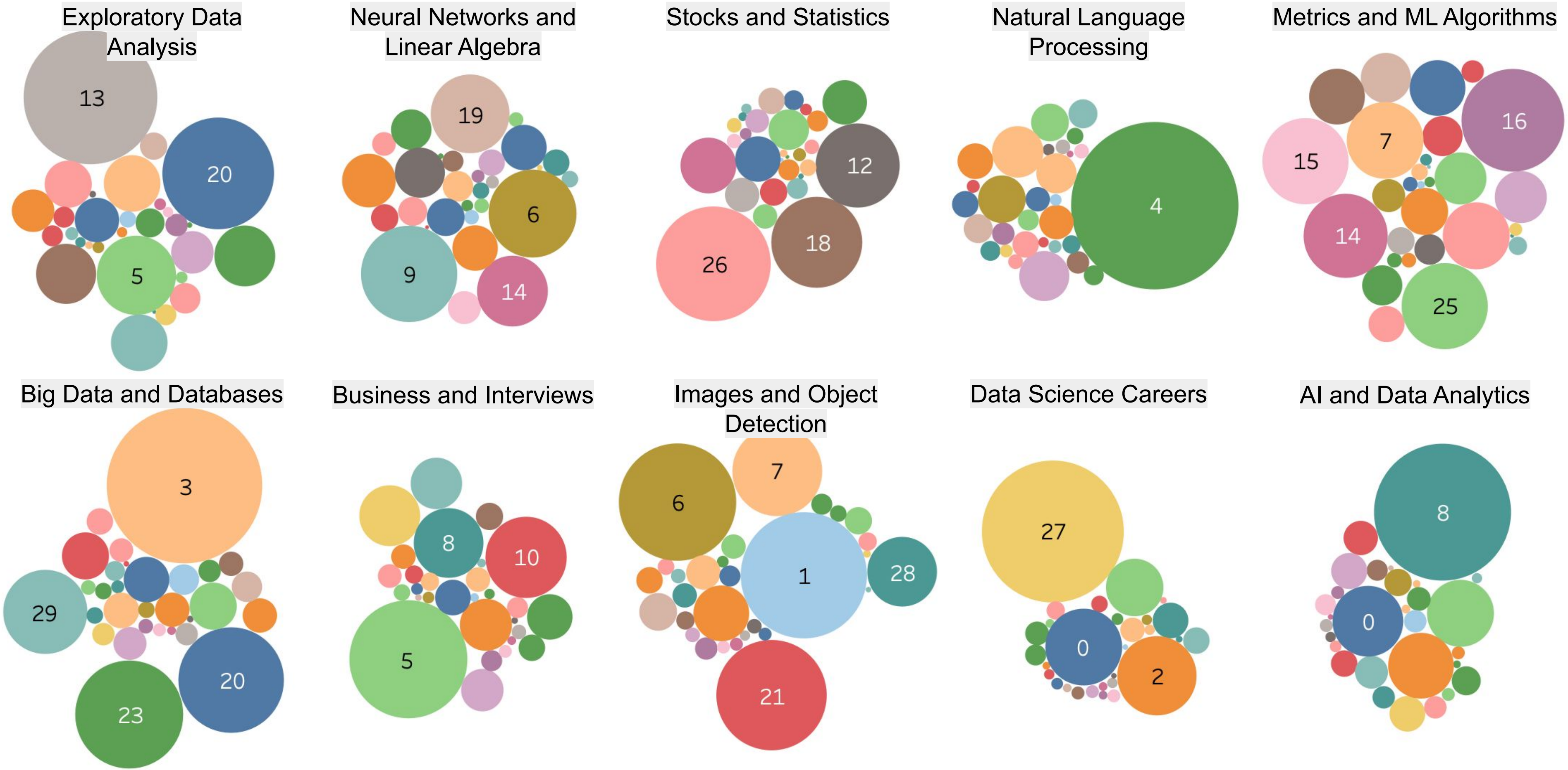


## Top 4 Volatile Cluster Topics




Top 4 Volatile Cluster Topics
Relative Percent of Articles Per Quarter
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
2017
2018
2019
Date of Publication
Cluster
AI and Data Analytics
Big Data and Databases
Metrics and ML Algorithms
Business and Interviews

# Topic Distribution Per Cluster

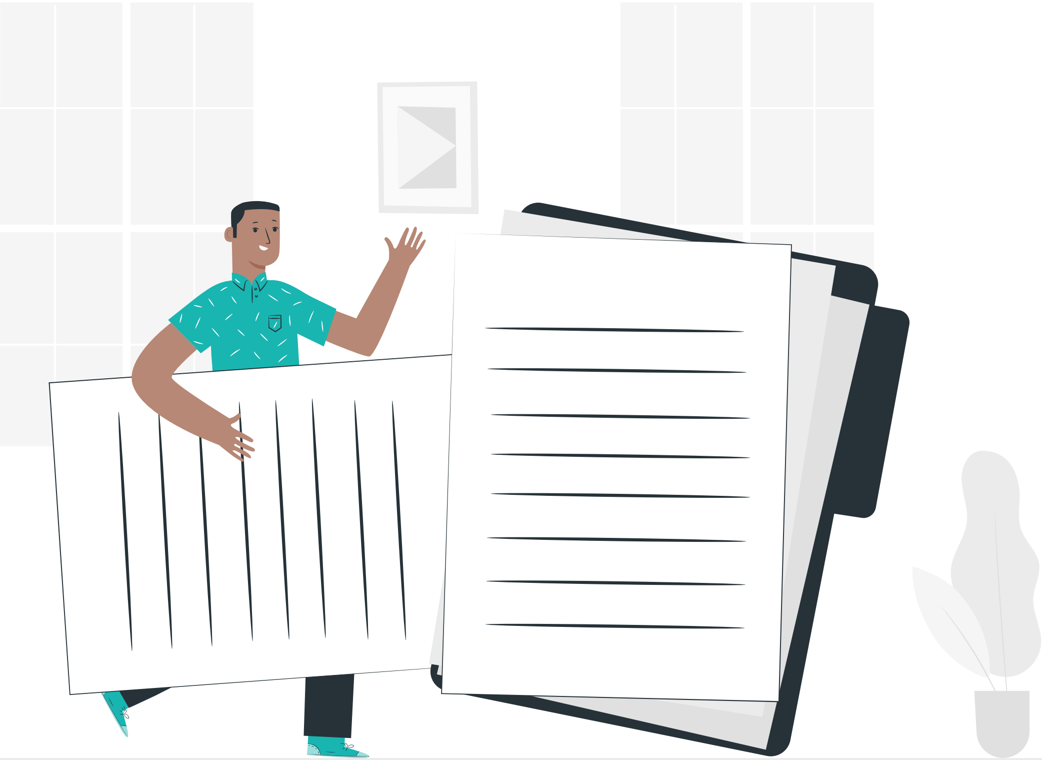


* Words associated with Topics can be found in the appendix of this presentation

# THANKS

Do you have any questions?

andrew.k.auyeung@gmail.com
732 666 4020



CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik

Please keep this slide for attribution.

# Sub-Topic Distribution

Topic 0
time, one, would, like, thing, get, work, could, way, know

Topic 1
image, images, pixel, convolution, color, convolutional, style, cnn, filter, augmentation

Topic 2
learning, machine, deep, algorithm, ml, learn, course, neural, supervised, book

Topic 3
app, api, web, command, notebook, project, install, page, create, package

Topic 4
word, text, sentence, vector, document, embeddings, nlp, language, words, embedding

Topic 5
data, analysis, science, big, analytics, business, scientist, tool, set, information

Topic 6
layer, network, neural, input, output, activation, neuron, weight, function, hidden

Topic 7
training, validation, dataset, train, test, set, accuracy, data, trained, tensorflow

Topic 8
ai, intelligence, human, artificial, technology, system, company, ml, world, business

Topic 9
agent, action, reward, state, policy, reinforcement, environment, rl, value, game

Topic 10
customer, business, product, company, analytics, marketing, churn, customers, team, service

Topic 11
cluster, clustering, means, distance, clusters, algorithm, point, centroid, number, group

Topic 12
distribution, probability, bayesian, random, normal, likelihood, posterior, event, prior, sample

Topic 13
plot, chart, visualization, bar, plotly, matplotlib, color, map, axis, country

Topic 14
regression, linear, function, gradient, variable, value, descent, logistic, error, equation

# Sub-Topic Distribution

Topic 15
tree, decision, forest, random, trees, node, split, algorithm, boosting, gini

Topic 16
feature, features, variable, categorical, dataset, selection, missing, importance, correlation, target

Topic 17
user, item, movie, rating, recommendation, recommender, system, filtering, collaborative, users

Topic 18
price, stock, series, time, market, forecast, forecasting, trading, arima, day

Topic 19
matrix, vector, array, pca, numpy, dimension, dimensional, space, covariance, eigenvectors

Topic 20
column, function, dataframe, pandas, value, row, list, index, method, string

Topic 21
object, detection, box, bounding, yolo, class, cnn, segmentation, video, car

Topic 22
node, graph, edge, nodes, network, vertex, graphs, centrality, neo, connected

Topic 23
docker, container, spark, aws, cloud, run, service, kubernetes, ml, command

Topic 24
player, game, team, season, players, win, nba, league, play, games

Topic 25
class, positive, recall, precision, classification, false, score, negative, classifier, roc

Topic 26
hypothesis, test, sample, null, population, value, mean, testing, statistical, significance

Topic 27
science, scientist, job, skill, data, project, company, interview, team, role

Topic 28
discriminator, generator, gan, gans, fake, loss, adversarial, generative, network, training

Topic 29
sql, query, table, database, bigquery, join, mysql, relational, tables, select