facebook

Using ML to Automate Dynamic Error Categorization

Antonio Davoli

Production Engineer, Servers Lifecycle Engineering

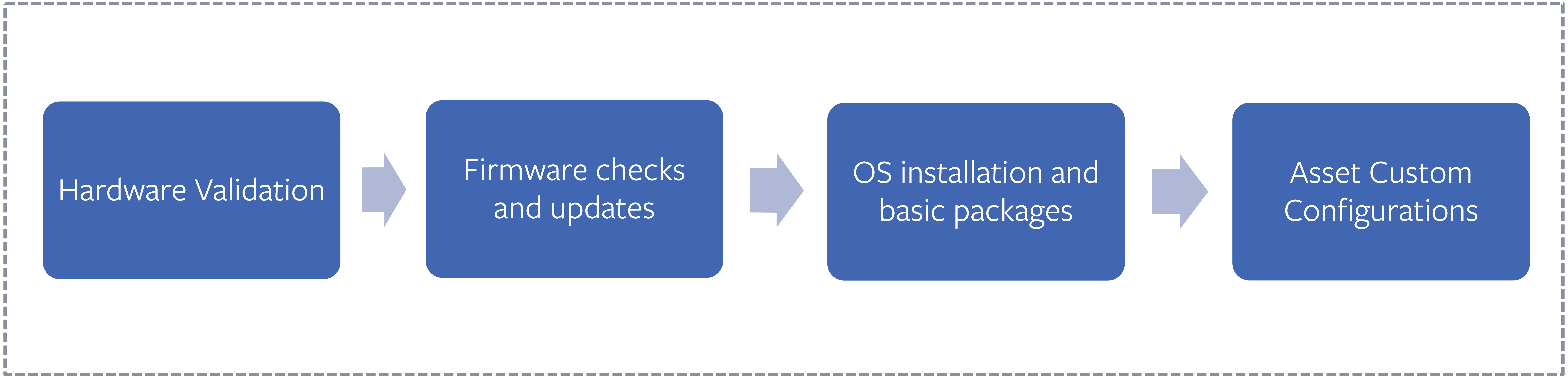
Agenda

- Servers Lifecycle
- Clustering
- SQClusters
- Results and future work

Servers Lifecycle

Servers Lifecycle

Distributed Jobs Orchestrator for handling server lifecycle stages (e.g. Provisioning)



Hardware Validation

Firmware checks and updates

OS installation and basic packages

Asset Custom Configurations

Suspended Jobs Queue be like:

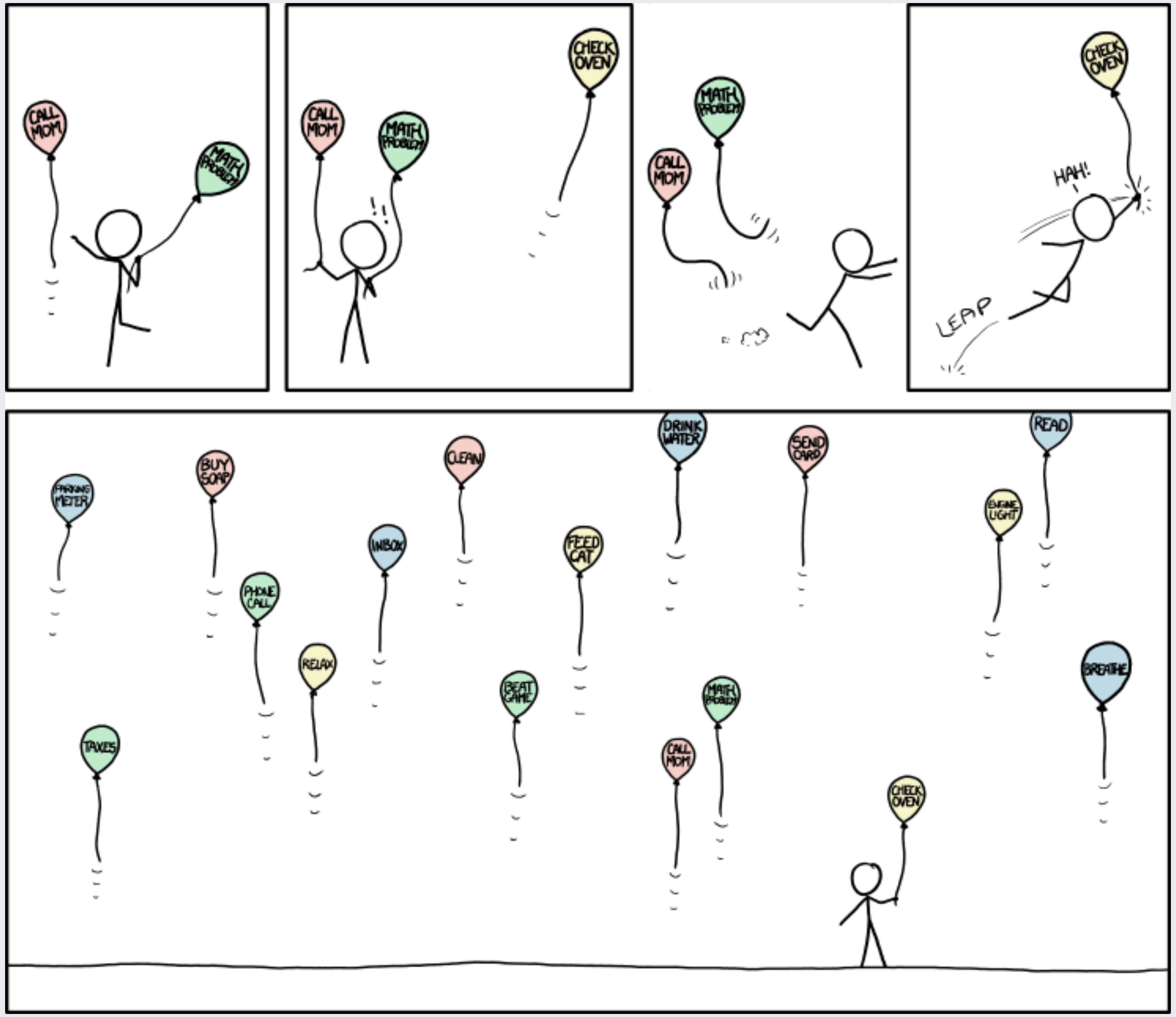
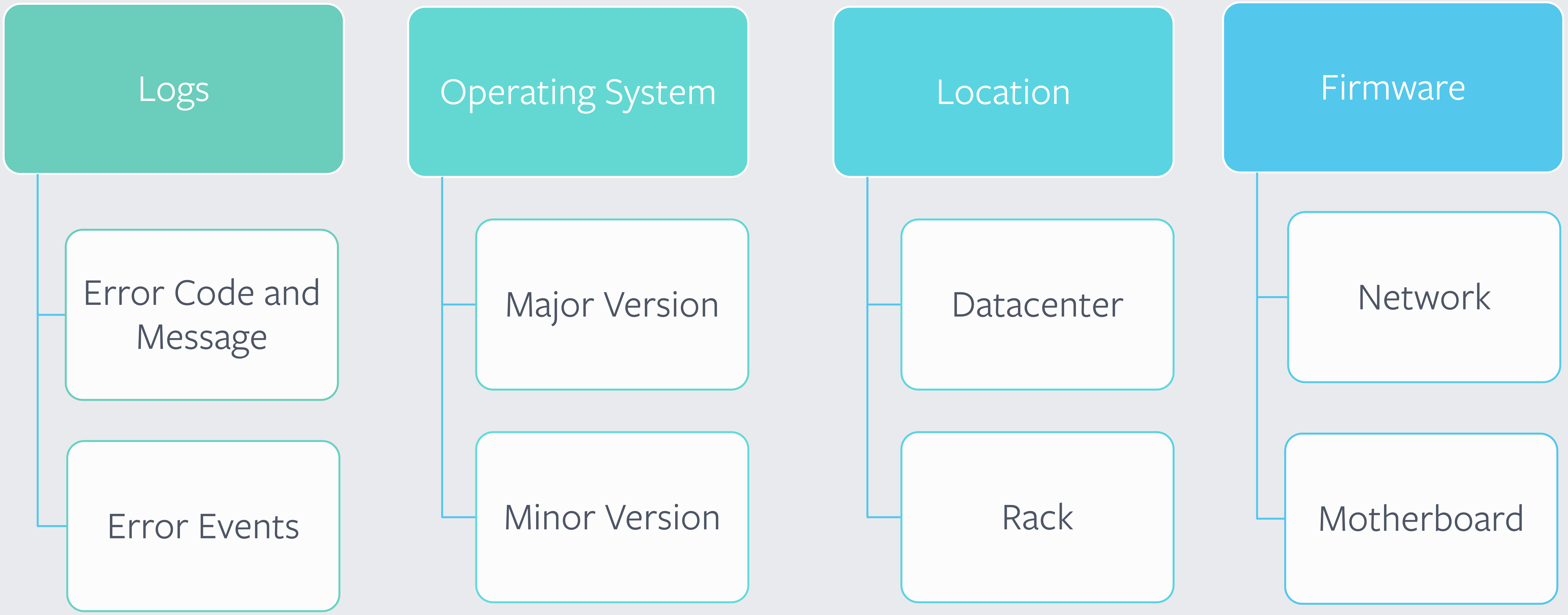


Image Credit: https://xkcd.com/1106

“if you torture the data long enough,
it will confess”

Ronald Coase, Economist

Moar data!



- Logs
  - Error Code and Message
  - Error Events
- Operating System
  - Major Version
  - Minor Version
- Location
  - Datacenter
  - Rack
- Firmware
  - Network
  - Motherboard

Inferring Similarities

Considered all the various data sources we can pull data from, why don’t we try to ***infer more similarities*** that we can exploit to ***fix the highest number of servers*** *in the shorter possible time*?

Clustering

Clustering is the task of **grouping a set of objects** in such a way that objects in the same group are *more similar to each other* than to those in other groups.

— Wikipedia

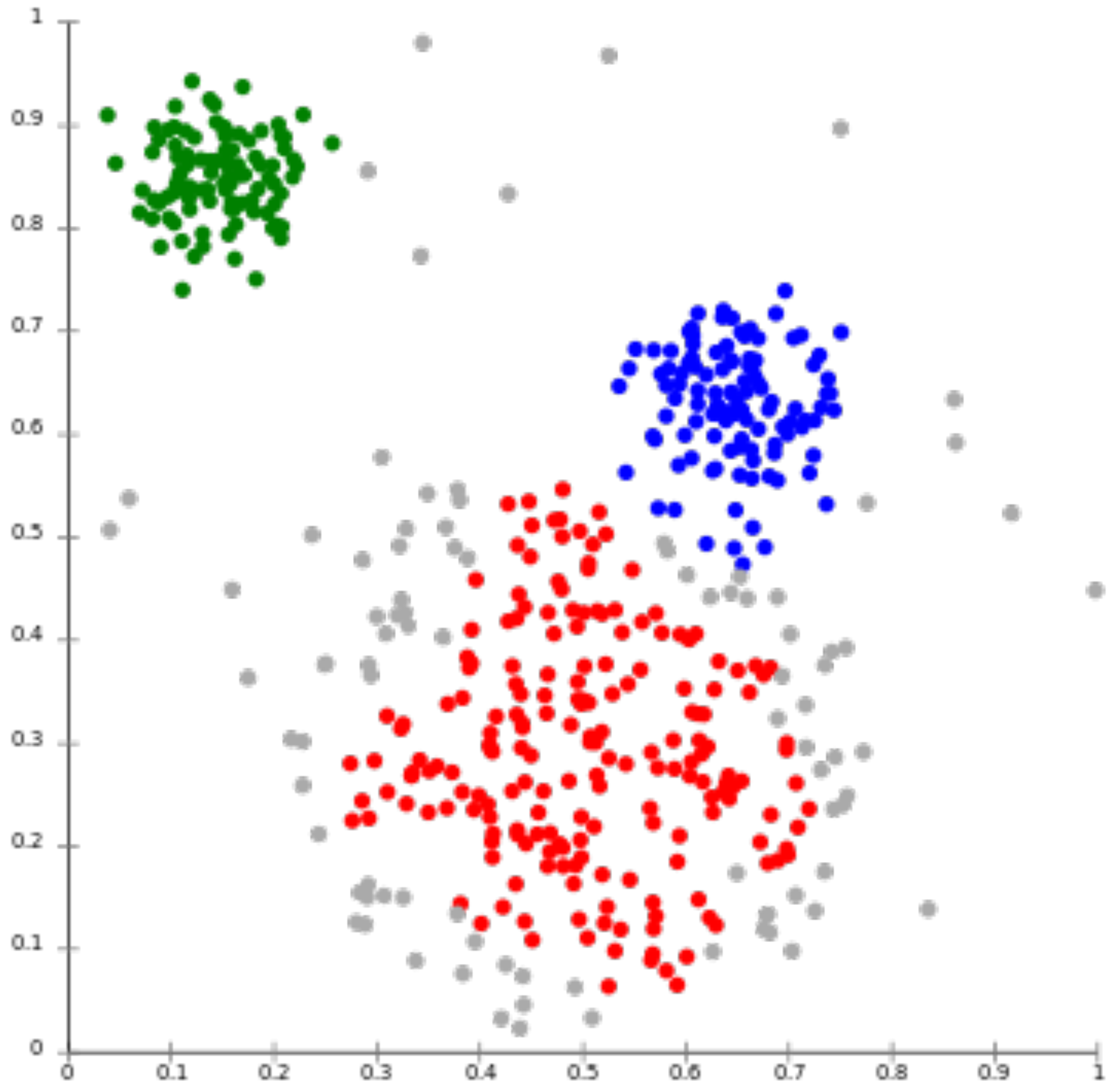
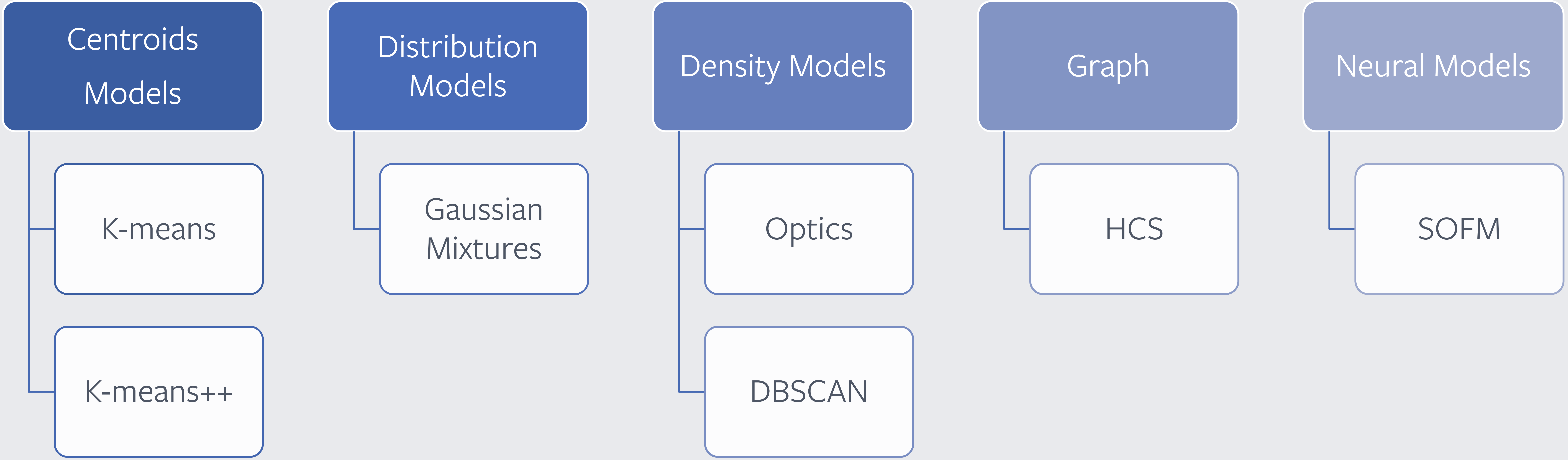


Image Credit: Wikipedia, https://en.wikipedia.prg/DBSCAN

Clustering Algorithms



Centroids Models

- K-means
- K-means++

Distribution Models

- Gaussian Mixtures

Density Models

- Optics
- DBSCAN

Graph

- HCS

Neural Models

- SOFM

SQClusters

SQClusters

Applying DBSCAN to the Orchestrator Suspend Queue

DBSCAN is a density-based clustering algorithm.

Given a set of points in some space, it **groups points that are closely packed together**, marking as outliers points that lie alone in low-density regions.

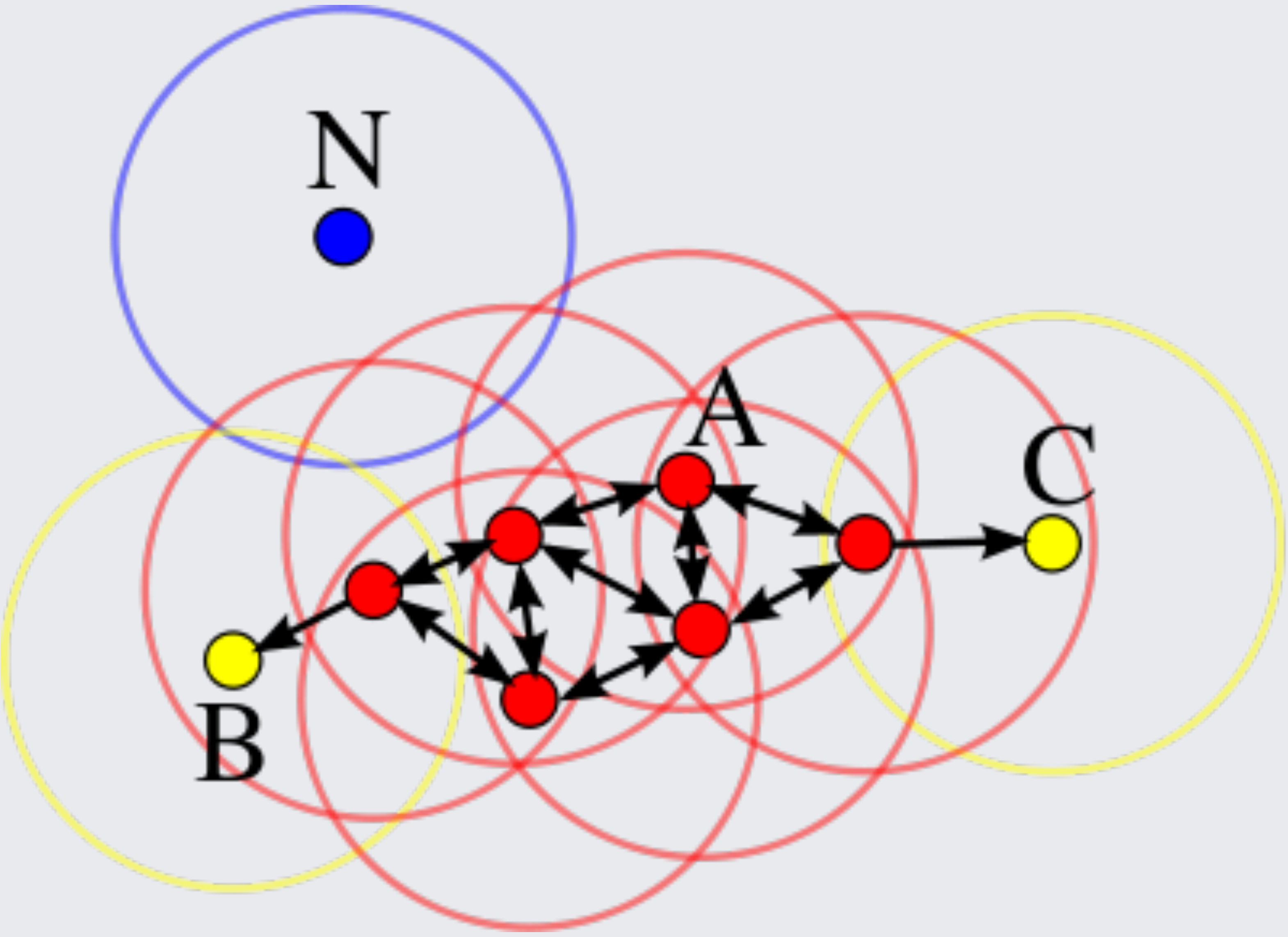


Image Credit: Wikipedia, https://en.wikipedia.prg/DBSCAN

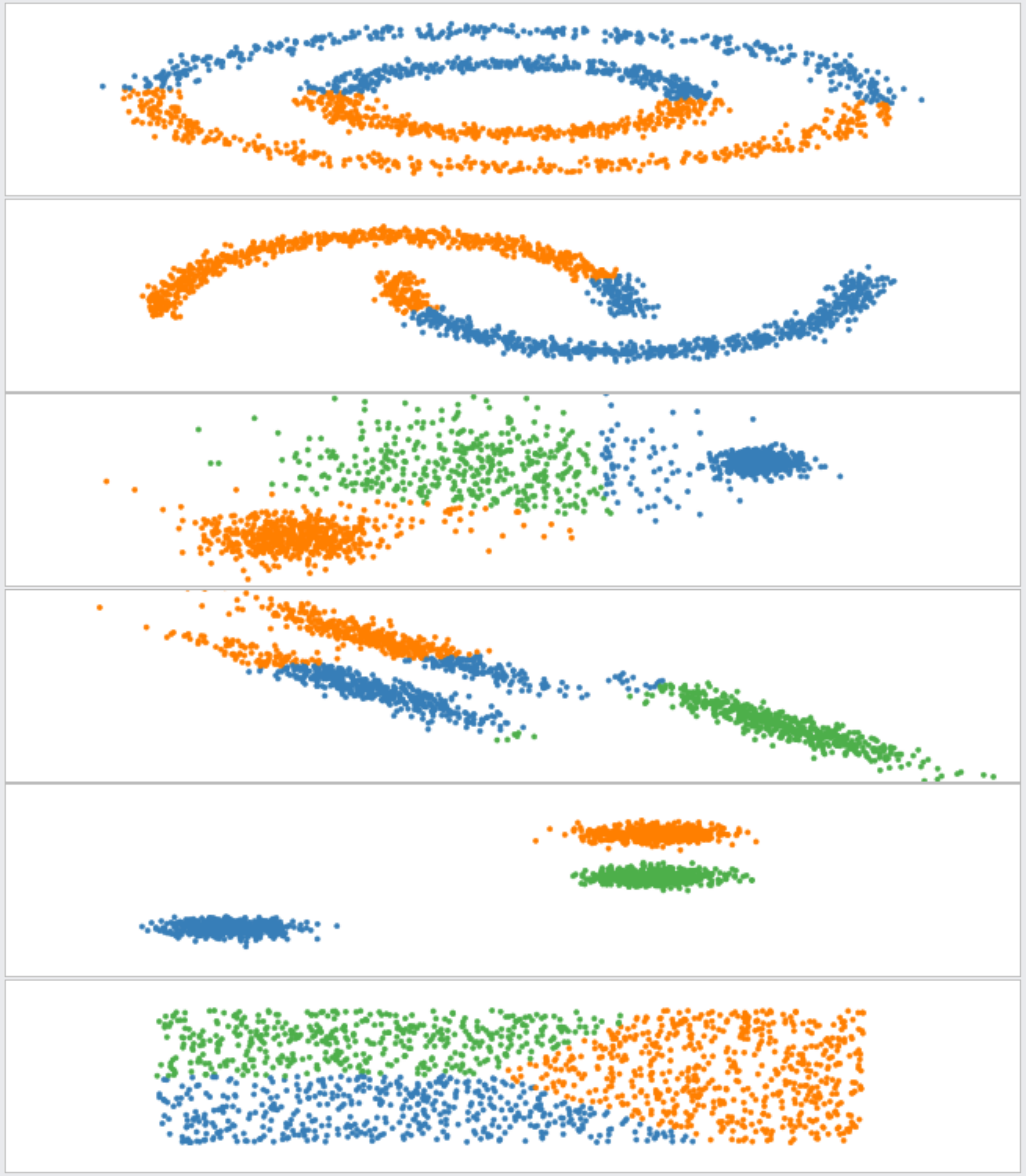
DBSCAN

Algorithm Internals

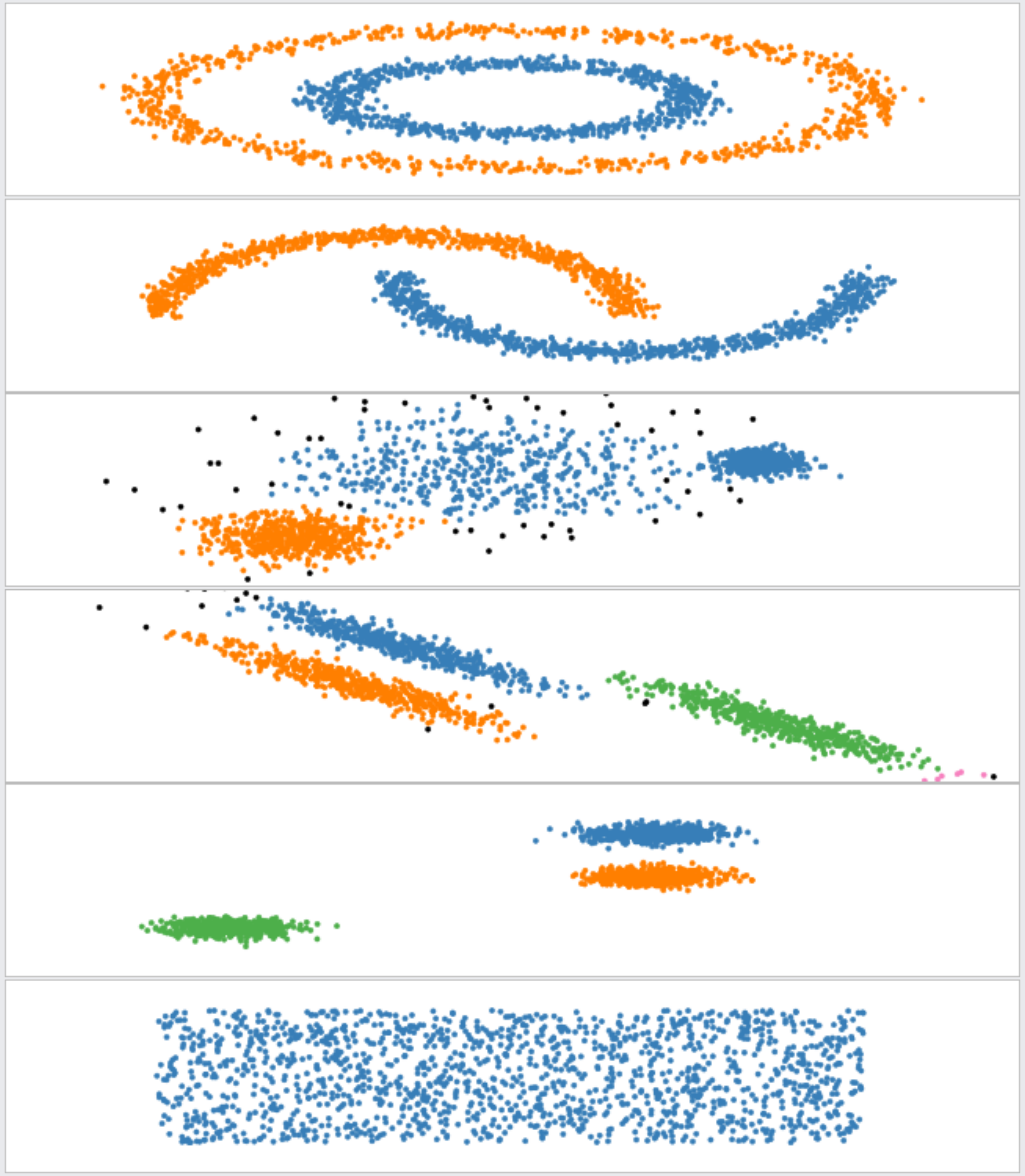
Doesn’t require to specify the number of clusters, it does have a notion of noise which makes it robust to outliers.

- **ε** (eps): minimum distance between points in space,
- **min_points**: minimum number of points required to form a dense region

K-means



dbscan



Code for synthetic data: https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html#sphx-glr-auto-examples-cluster-plot-cluster-comparison-py

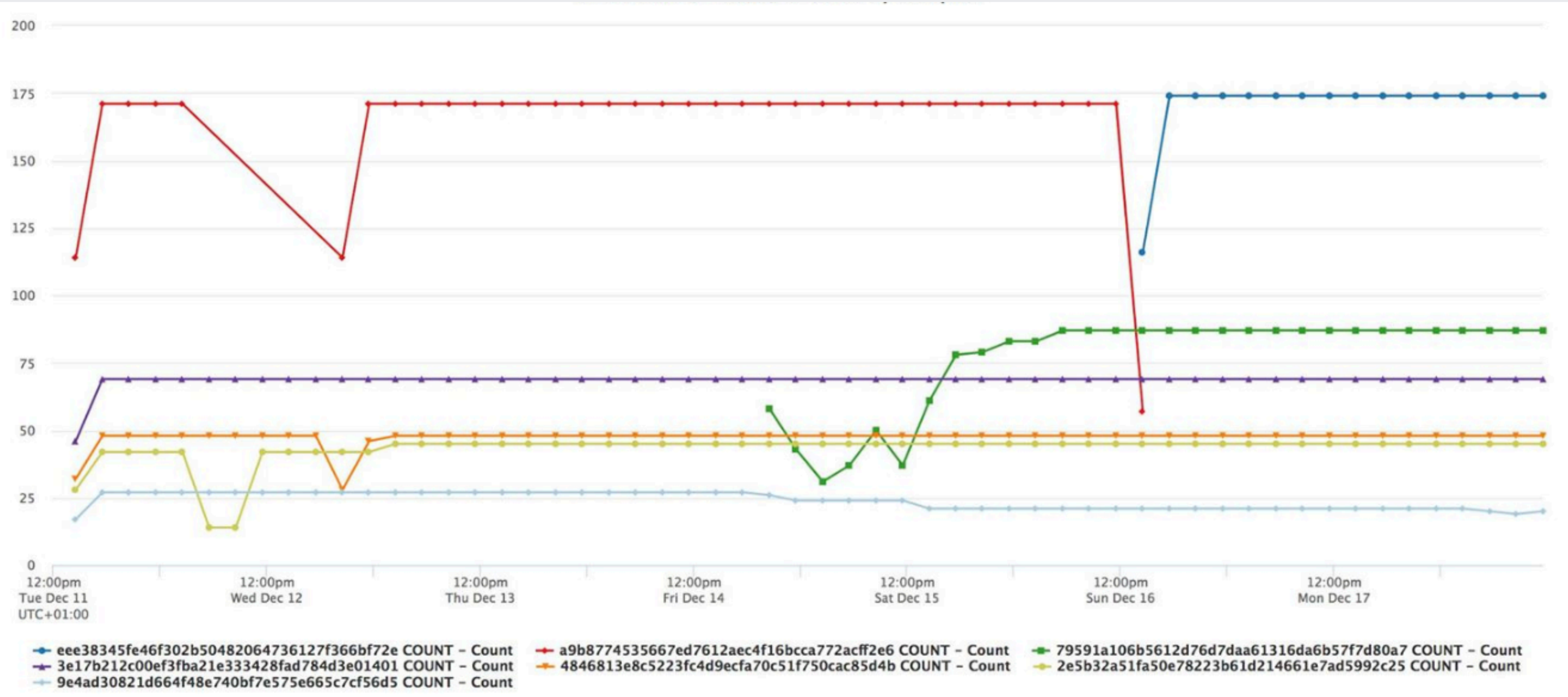
One-Hot Encoding for Categorical Features

Categorical features are substituted by their integer representation.

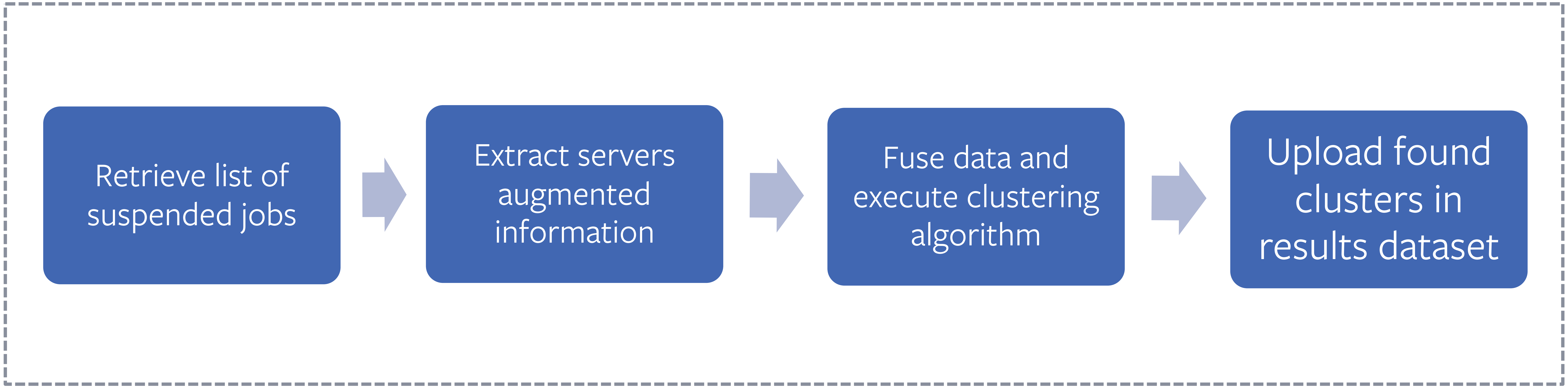
Server	Datacenter
1	Singapore
2	Sweden
3	Ireland

Server	Datacenter_Singapore	Datacenter_Sweden	Datacenter_Ireland
1	1	0	0
2	0	1	0
3	0	0	1

Hash values for clusters identifiers



SQClusters Pipeline



Retrieve list of suspended jobs

Extract servers augmented information

Fuse data and execute clustering algorithm

Upload found clusters in results dataset

Real example of clustering results

Cluster	Size	Error Message	Hostname Scheme	Model	Datacenter
abc	231	chef_error_msg	hadoop	Model #1	SGP, SWE
xyz	91	dhcp_error_msg, pxe_boot_error_msg	cache	Model #2	IRL

Lessons learned

- Structured logging helps (use it, it'll pay back!),
- Spend all the time you need in cleaning your data,
- When you do this sort of exploratory work, listen to your data and make them “confess”,
- Using ML tooling is extremely easy to use: `dbscan.fit(X)`

What next?

- Experiments with more clustering algorithms, especially hierarchical approach based on density,
- Improve hashing techniques,
- Extract data on trends analysis and seasonality

facebook

Questions?

facebook

Thank You!

Backup: DBSCAN Internals

- NearestNeighbors based (Pair-wise or KD-Tree)
- Depth-first search, very similar to the classic algorithm for computing connected

Backup: k-means Internals

- Iterative approach (Expectation–Maximization), continues to compute centroids continuously
  - The ″cluster center″ is the arithmetic mean of all the points belonging to the cluster.