CLUSTERING

2020-05-05
Mags Giust

April 30, 2020 Clustering

"Life comes in clusters, clusters of solitude, then a cluster when there is hardly time to breathe"

- May Sarton (Eleanor Marie Sarton)
Poet, Novelist, Memoirist
1912-1995

"over time, this growing tendency of like marrying like will only reinforce clustering and geographic sorting along class lines, giving the emerging map of social, economic, and cultural segregation even greater permanence."

- Richard Florida
Urban Studies - Social & Economic Theory
Prof. @ Univ. of Toronto
1957 - present

MACHINE LEARNING

Unsupervised Learning

DIMENSIONALITY REDUCTION

- PCA
- t-SNE
- Factor Analysis

CLUSTERING

AD, NLP may be solved with clustering.

Algorithms

- kMeans
- DBScan
- Hierarchical

Reinforcement Learning

AD, NLP, RS may be solved with reinforcement learning

- Collaborative Filtering
- Content Filtering
- Neural Networks (DL)

Supervised Learning

REGRESSION

- Continuous Target

TSA, AD may be regression probs.

- Linear Regression
- GLM
- Polynomial Reg.
- SVR
- DecisionTree Regr.
- Neural Networks

CLASSIFICATION

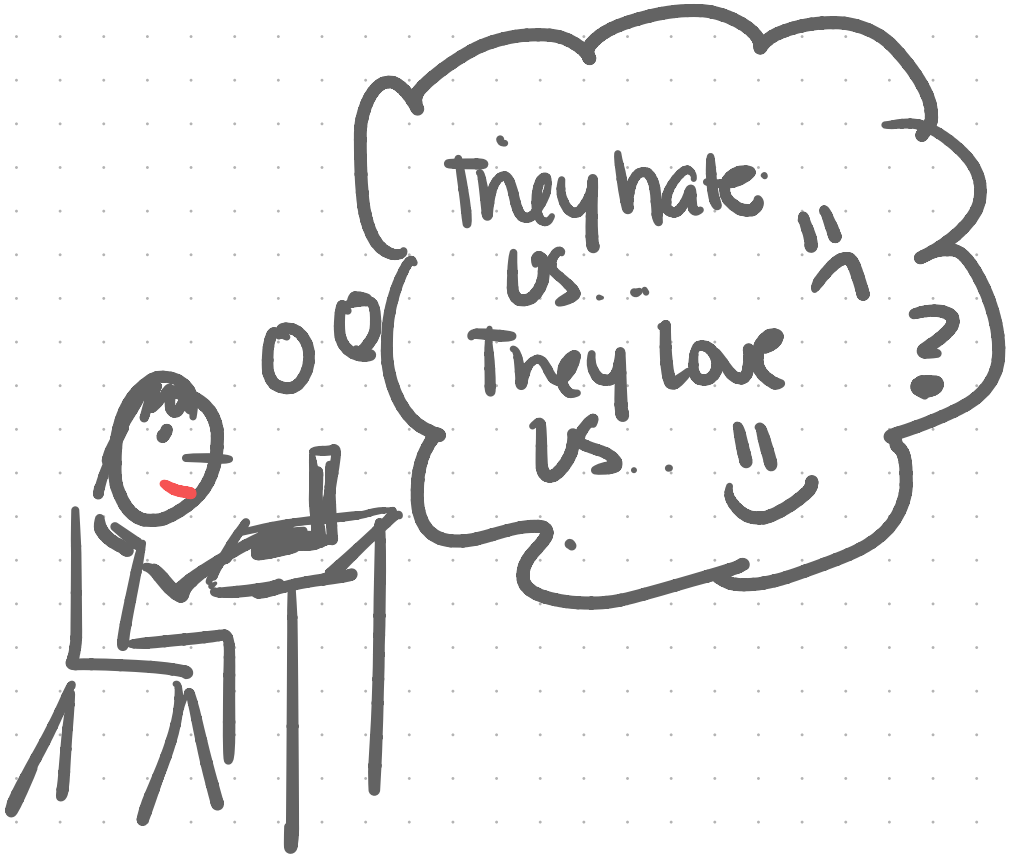
- Discrete, Categorical target

TSA, NLP, AD may be classification problems

- Logistic Regression
- Decision Tree
- Random Forest
- SVC
- KNN
- Neural Networks

SUB-METHODOLOGIES

- **TSA**: time series analysis
- **AD**: anomaly detection
- **NLP**: natural Language Processing
- **RS**: recommendation System



MAGGIE, customer intel analyst
RACKSPACE, C. 2010


I hear you can do Magic with numbers... SO... Who are our cloud Customers?
...They love...
What the f...
uh-oh!
LANHAM, CEO
RACKSPACE
MAGGIE,
aspiring data scientist

Wrangle

Plan

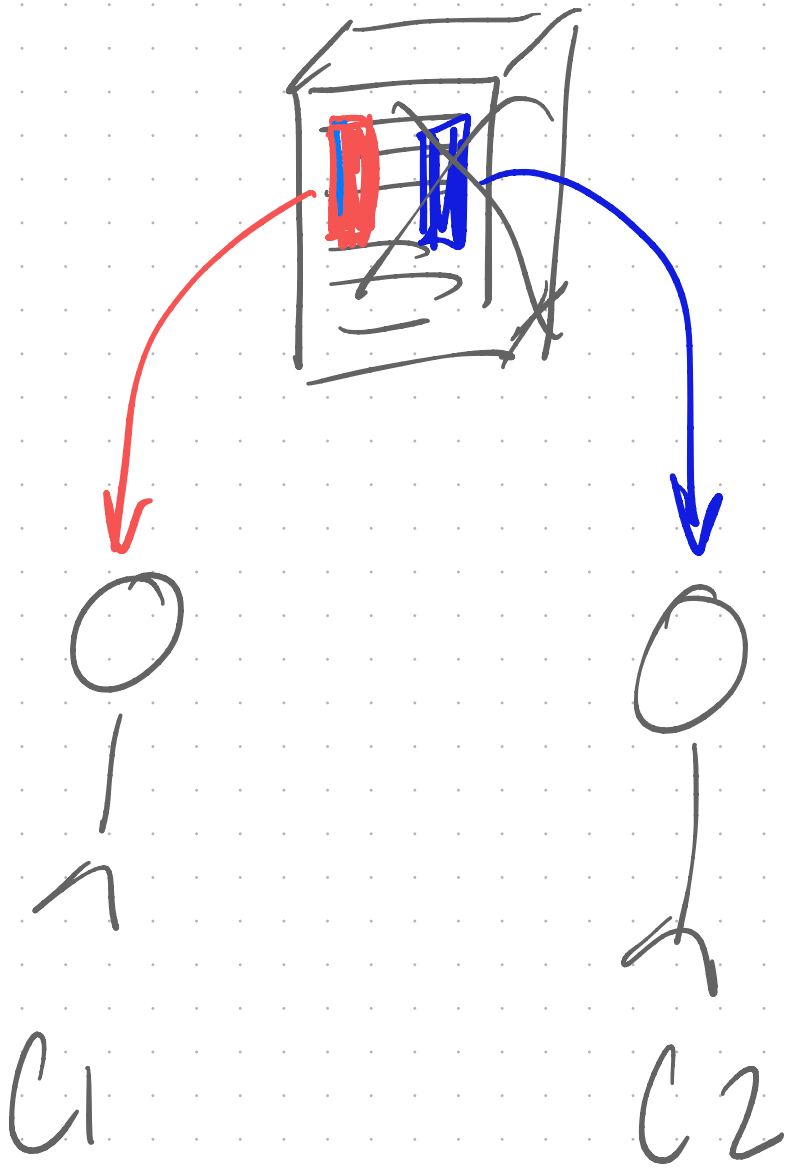
- define the ?, deliverables, goals - CLEARLY
- get necessary domain knowledge
- early hypotheses
- data definitions, locations, etc.
- timeline? how much do I have? how bad do I want this?

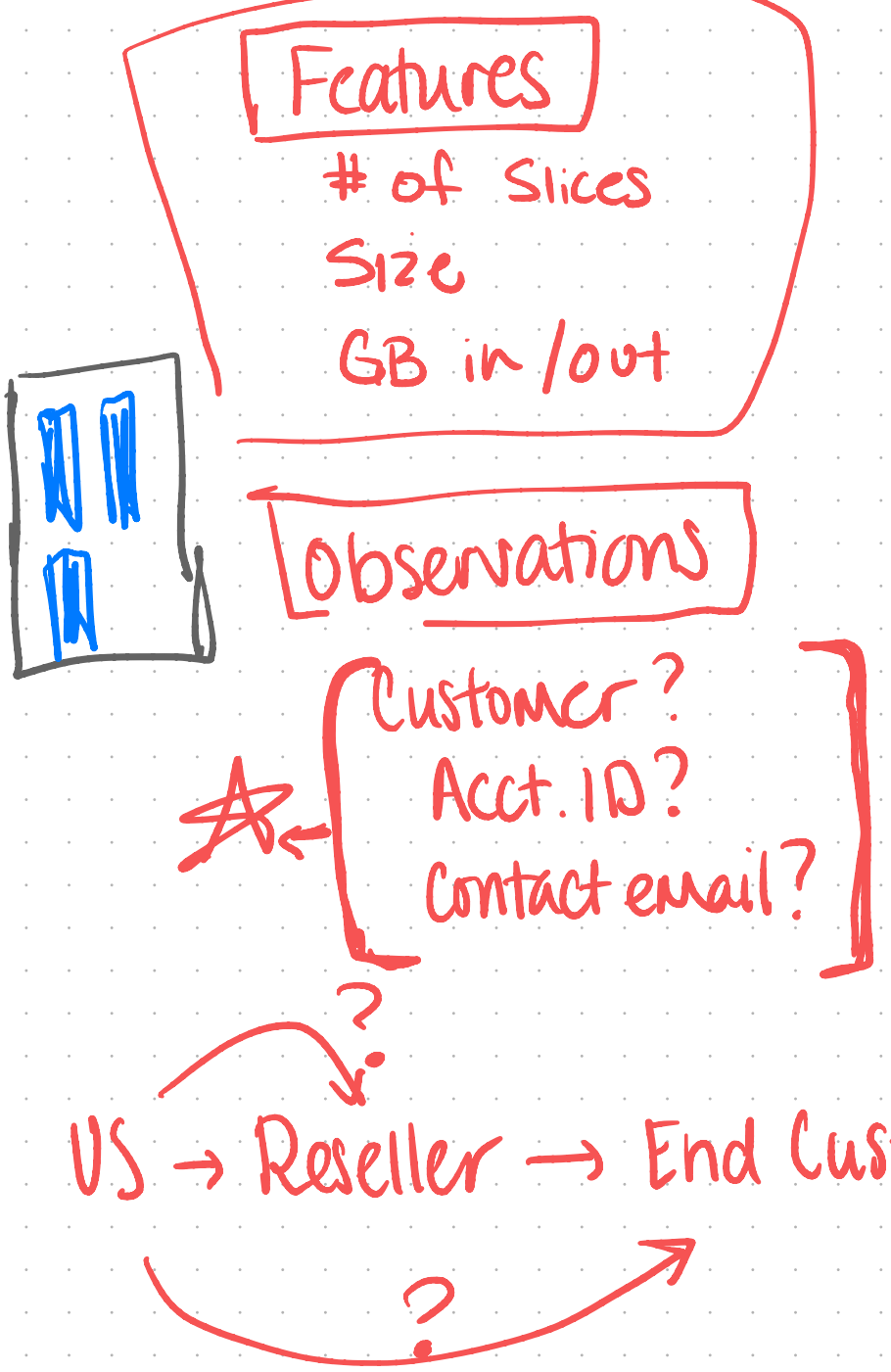
Explore/Analyze

Deliver

Model

* This Pipeline heuristic was not available/clear to much of anyone at this time.

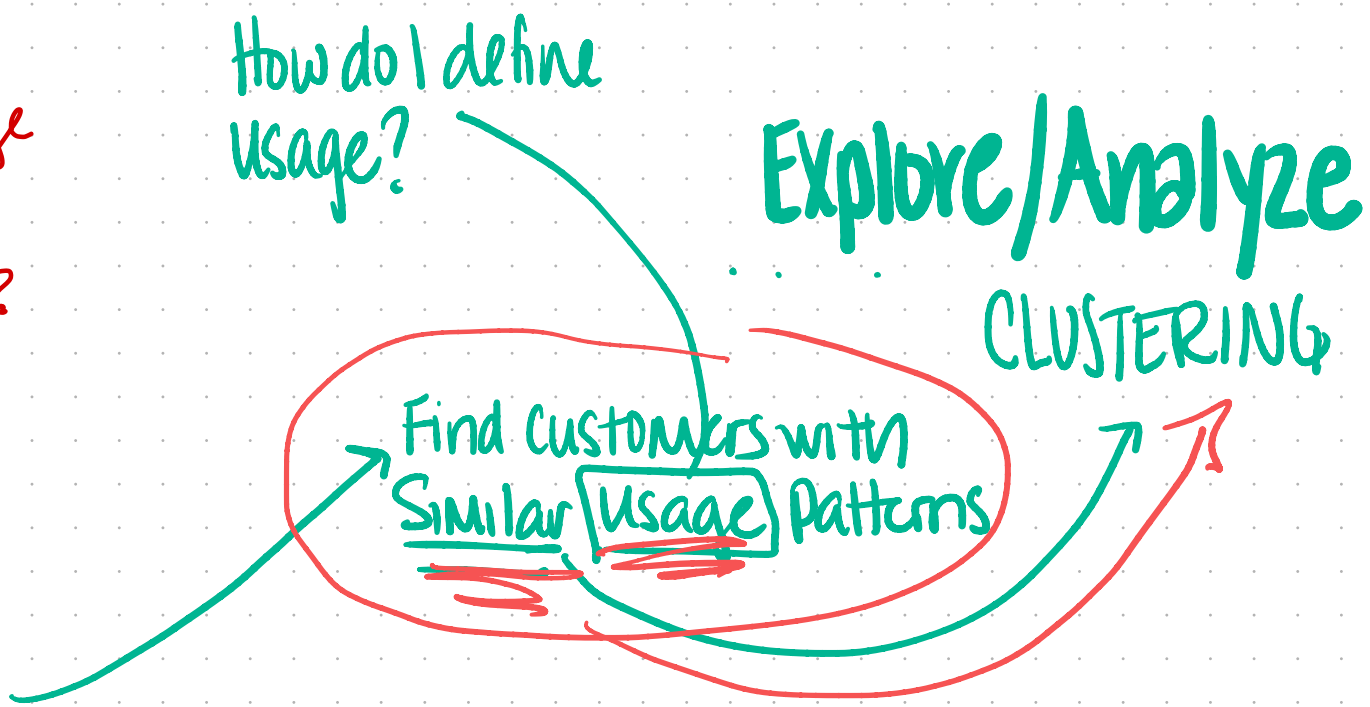

C1
C2


Features
# of Slices
Size
GB in/out
Observations
Customer?
Acct. ID?
Contact email?
?
US → Reseller → End Cust.
?

Wrangle

Plan

- define the ?, deliverables, goals - CLEARLY
- get necessary domain knowledge
- early hypotheses
- data definitions, locations, etc.
- timeline? how much do I have? how bad do I want this?



Deliver

- Common Usage Patters of Cloud Customers!
- NTH: Profile of each of those patterns - Who, how Many, $, ...

Plan

- define the?, deliverables, goals - CLEARLY
- get necessary domain knowledge
- early hypotheses
- data: definitions, locations, etc.
- timeline? how much do I have? how bad do I want this?

Wrangle

- hundreds of millions of records
- tens of tables
- handful of DB's
- NOT indexed well!
- overnormalized
- the "unique" ID vs. Acct vs. Customer issue

- how to aggregate
- how to sample
- how much history
- what to measure
- what is an "observation"

Scale data

Explore/Analyze

- lots of Viz.
- statistical analysis
- CLUSTERING

↓

★ Find customers whose usage patterns are similar

- define usage

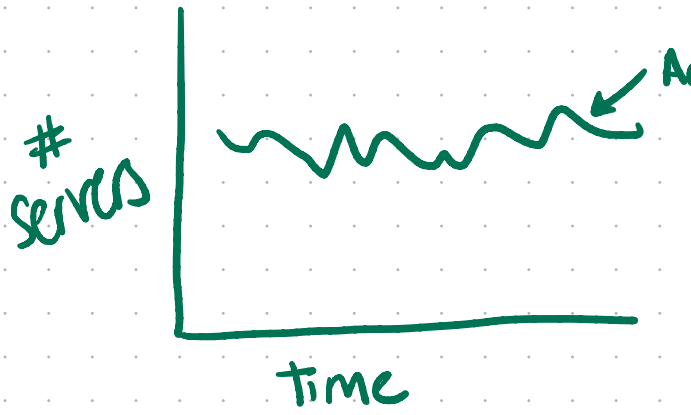
~~Model~~

Not yet!

Deliver

- Common Usage Patters of Cloud Customers!
  - NTH: Profile of each of those patterns - Who, how Many, $, ...

How do customers use Cloud Hosting? What are the common usage patterns?



# of servers active
observation = AccountID

Acct ID	Date	Server Ct

Reshape

Acct ID	S/1	S/2	S/3	...

servers

getfeatures

Acct ID	σ	μ	%.b	cv	m	Δ

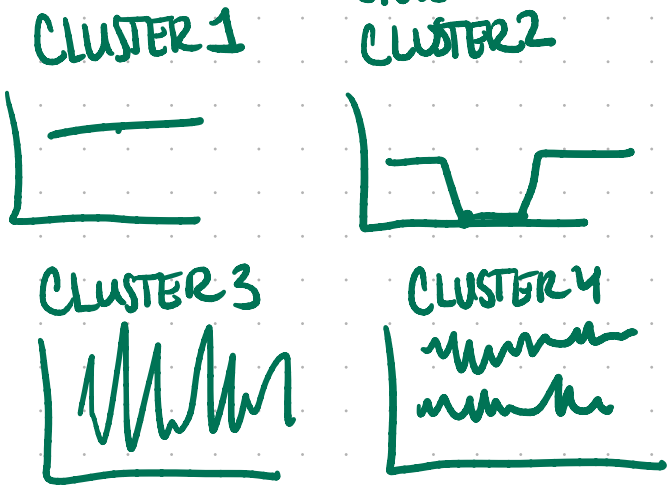
features

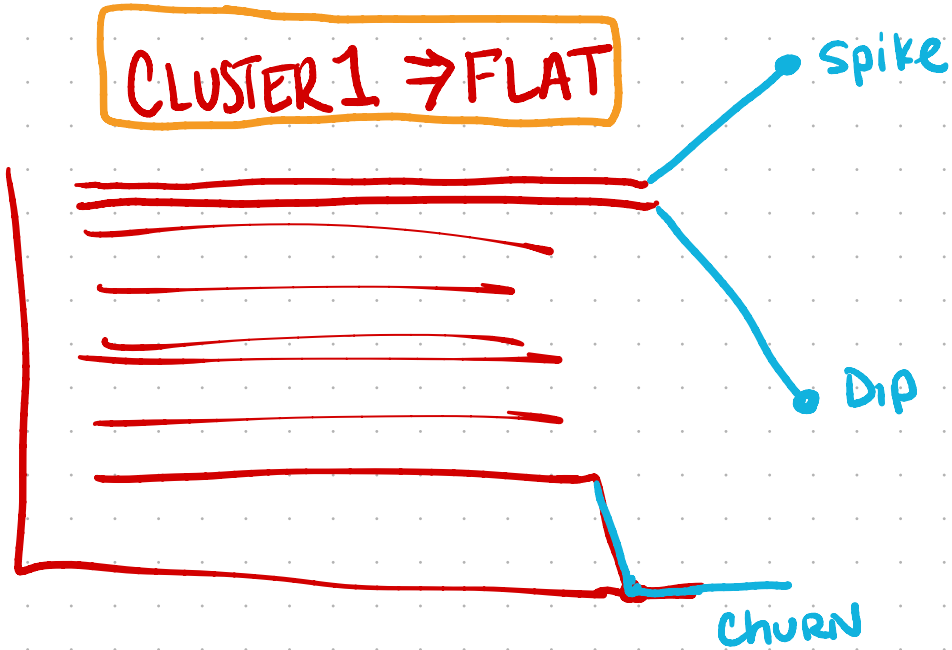
Run Clusterim Algorithm here

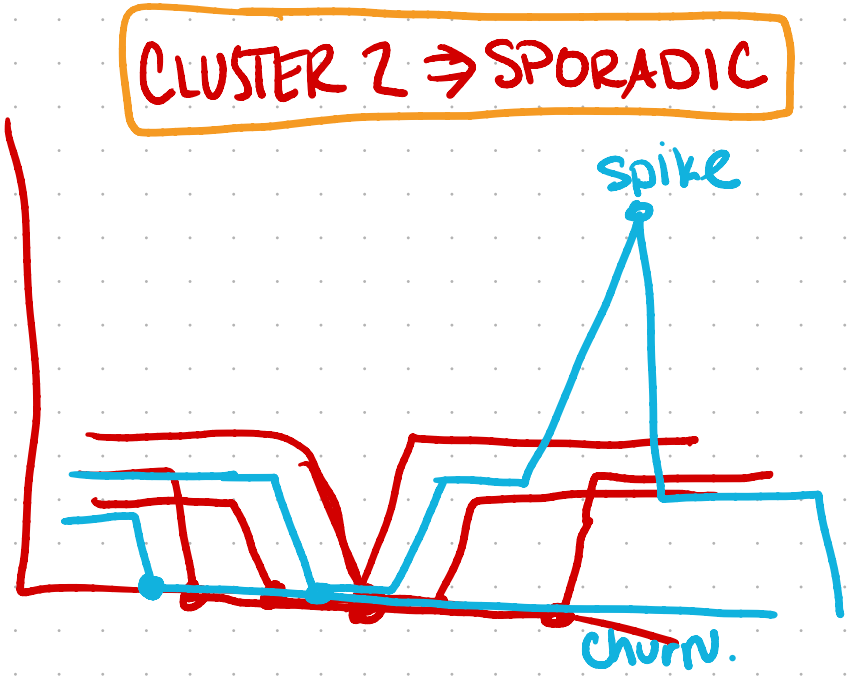
Cluster

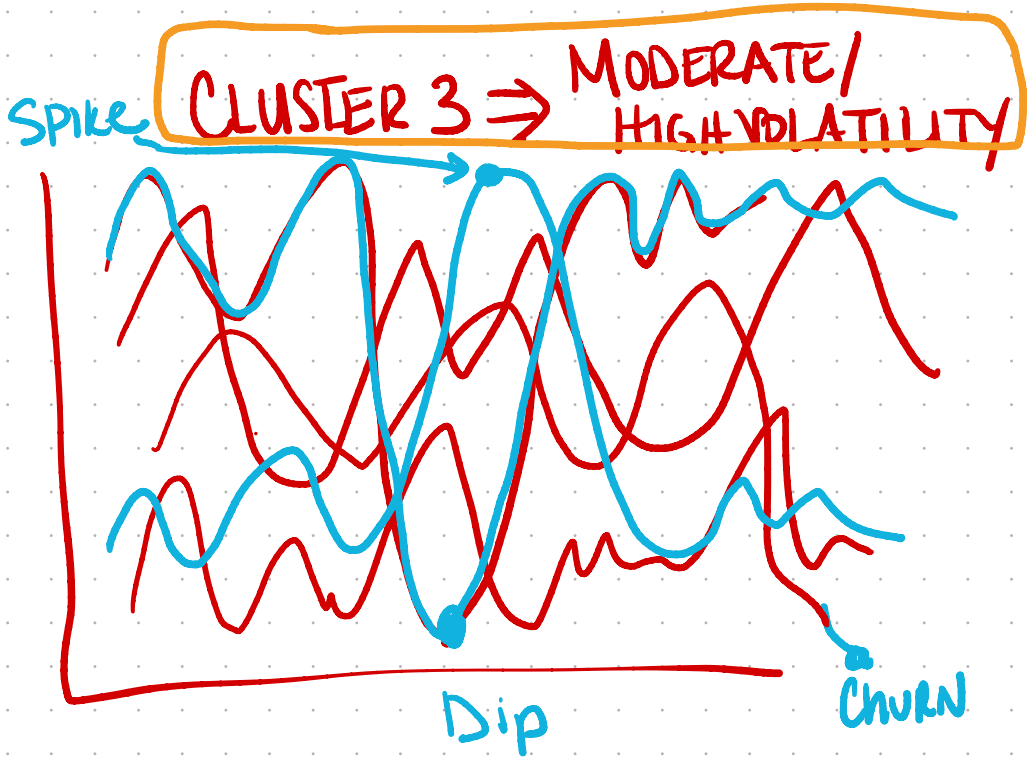
Acct ID	Cluster ID

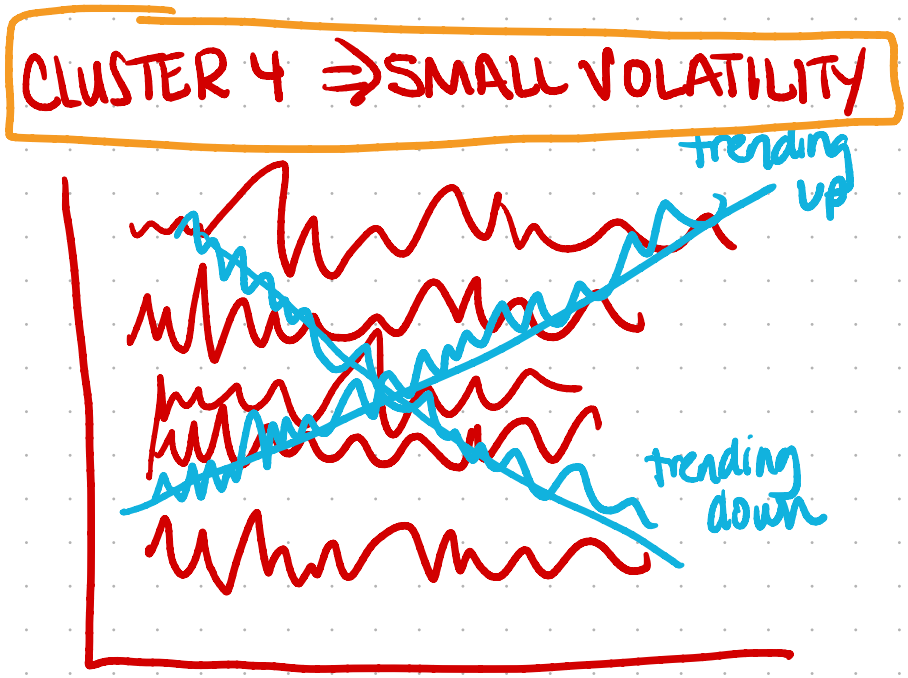
EXPLORE each Cluster




CLUSTER 1 ⇒ FLAT
Spike
Dip
CHURN


CLUSTER 2 ⇒ SPORADIC
Spike
CHURN.


CLUSTER 3 ⇒ MODERATE/ HIGH VOLATILITY
Spike
Dip
CHURN


CLUSTER 4 ⇒ SMALL VOLATILITY
trending up
trending down

What do you do with clusters?

- explore your new clusters, compare & contrast, analyze
- create new features for a future supervised model
- split observations into cluster to build a better model for each cluster - makes patterns easier to detect / extract!
- use clusters as labels for supervised learning - (classification, usually)

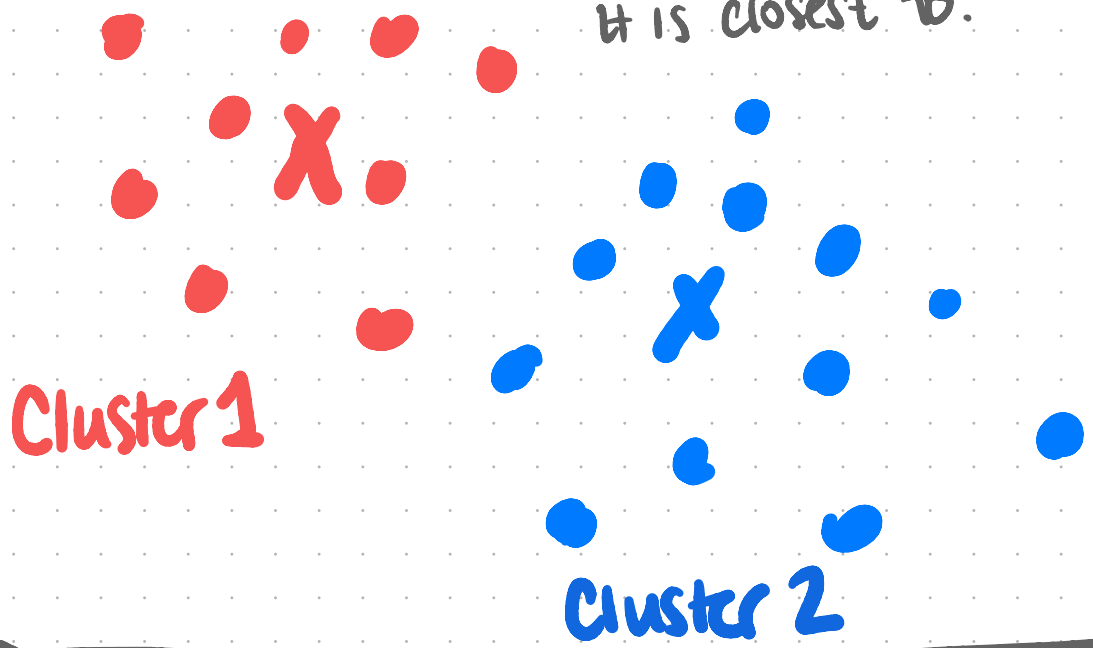
When is clustering useful?

- targeted marketing - cust. segmentation, market research
- anomaly detection - account take over, fraud, data breach
- crime zones, housing prices
- text - document classification
- image processing

CLUSTERING

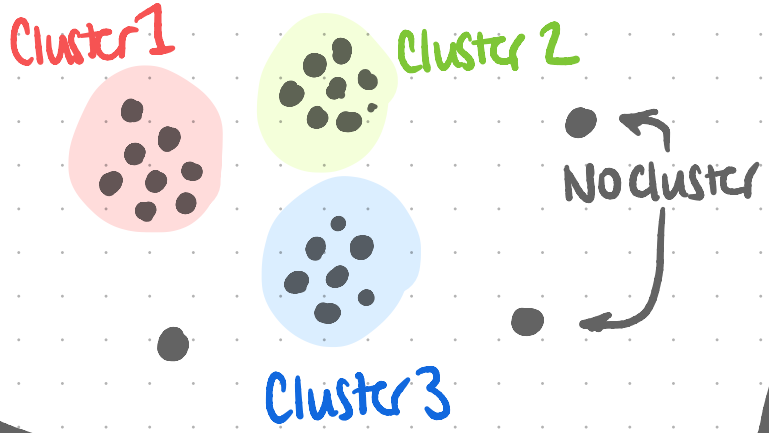
K-Means:

K = # of cluster
a point is put in the cluster whose centroid (X) it is closest to.

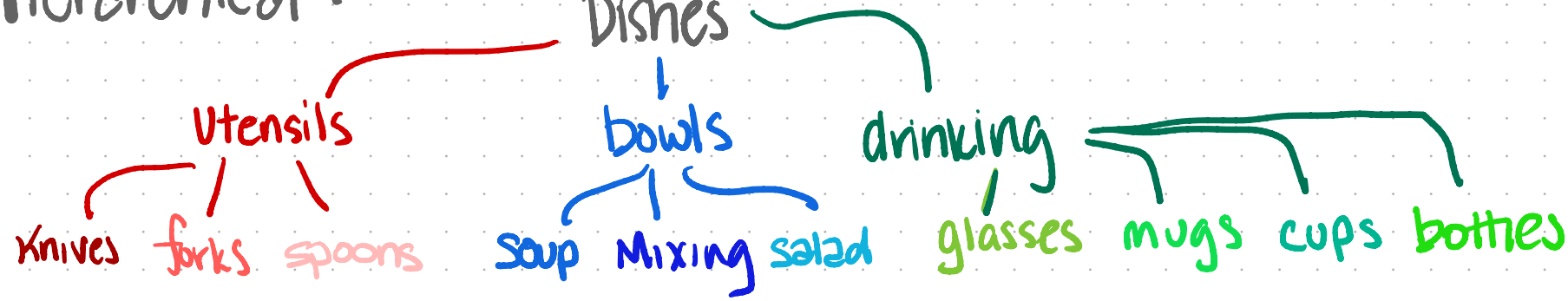


DBScan

A cluster is formed when it has a min. # of pts. within a max distance.

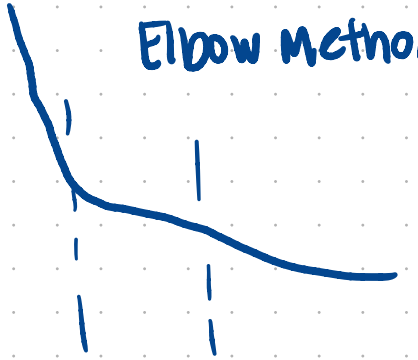


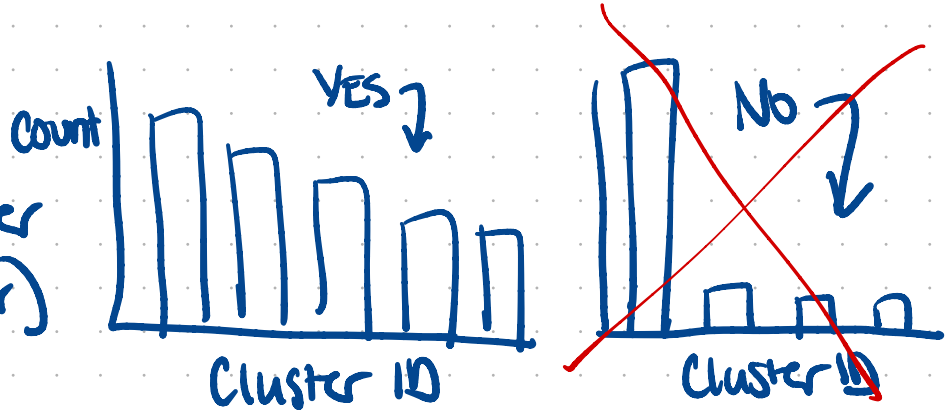
Hierarchical:



① KMeans: ⟹ SCALED DATA! Which vars?

② Elbow Method





③ Range of k ⇒ Get # of observations/cluster
select k (There is NOT ONE right answer)

④ Run KMeans(k, random_state)
fit_predict(train_scaled[cluster vars])
predict(test_scaled[cluster vars]) ⇒ 2 arrays 1 length of train
1 length of test

Add to DF

⑤ Join cluster ID's to Df's (x)
var 1 | var 2 | .. | var n | cluster_id

⑥ get centroids of each cluster: | cluster ID | centroid_var1 | centroid_var2 .. |

⑦ join back to DFs (x) var 1 | .. | var n | cluster_ID | centroid_var1 | centroid_var2 |

① SQL - filter !!!

② NULLS

③ DUMMY Vars of County, & City of LA

④ New features - reduce noise & dependencies → AGE, Tax Rate, Acres, $/sqft, $/acre, Bed:Bath Ratio + Create binned Vars

⑤ Remove Outliers for now - Need to Reduce Noise to find Patterns!

SAMPLE Data as we try out things - Most diff. will be LA, so use LA County

⑥ Split into train/test

⑦ Scale

⑧ Cluster Area - LAT, LON, AGE

⑨ Cluster Size - Sqft, ACRES

} see cluster steps

⑩ once clusters are in df's, group by the 2 cluster ID fields & get summary stats.

⑪ join Median + Stdev back to df's

⑫ USE stdev as feature

⑬ use Median + $/sqft to compute new feature $/sqft − Median

What have we done?

- Captured area in a useable format (numeric)
- separated dependencies on Y
- extracted a feature that measures variance in assessed value
- decreased dimensions.