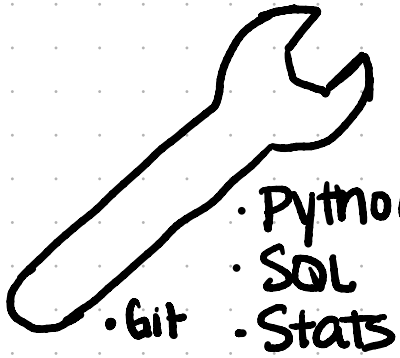
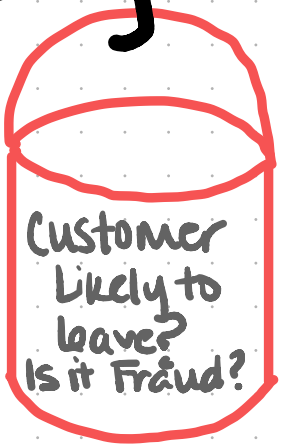
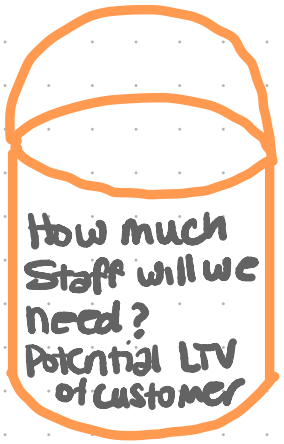
Learning Data Science



- Python
- SQL
- Stats
- Storytelling with Data
- Git
- Interview Skills
- Collaborating
- Presentation Skills

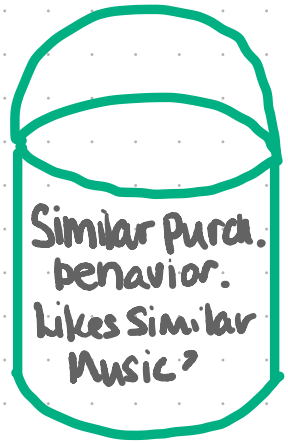


CLASSIFICATION

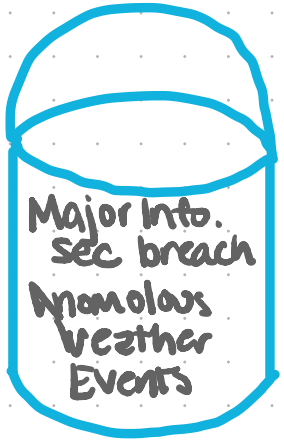


REGRESSION

TIME SERIES ANALYSIS

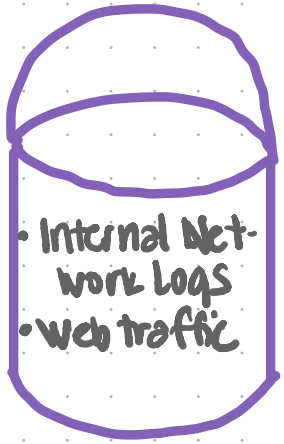


CLUSTERING

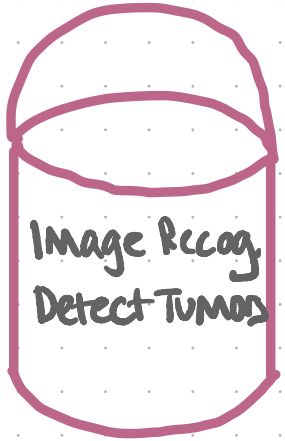


ANOMALY DETECTION

NATURAL LANGUAGE PROCESSING



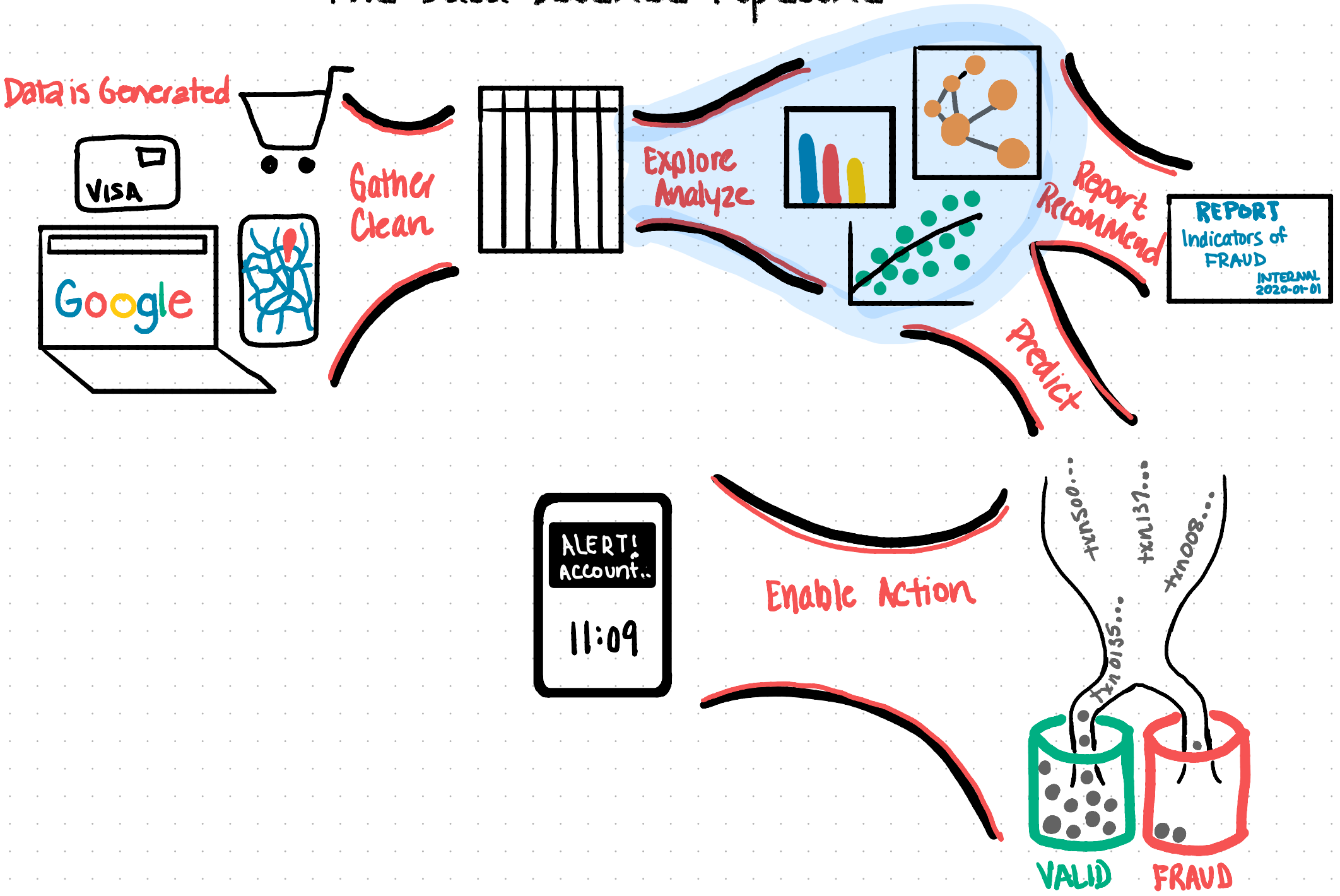
BIG DATA SPARK

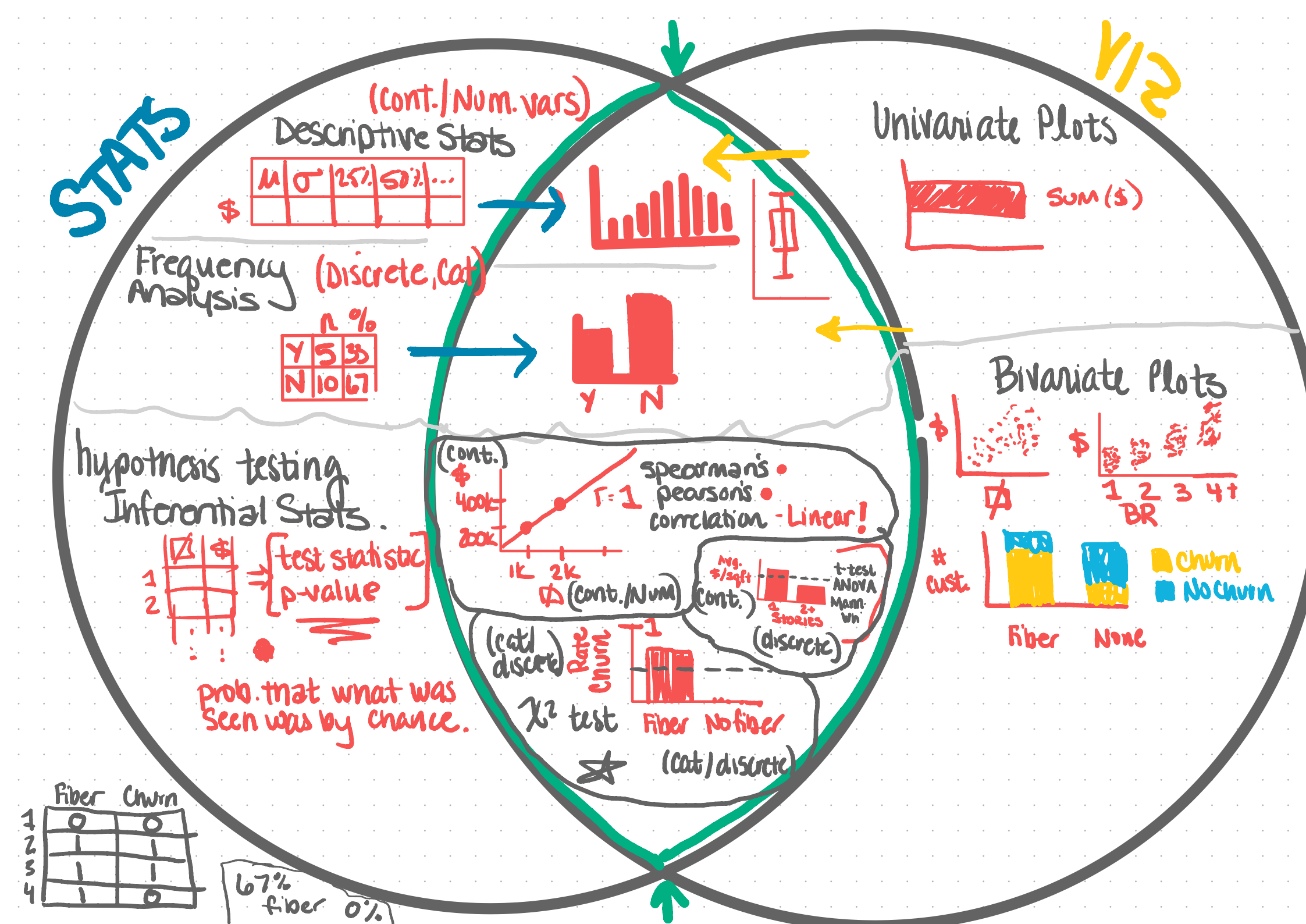


NEURAL NETWORKS DEEP LEARNING

end with the... FINAL CAPSTONE

The Data Science Pipeline

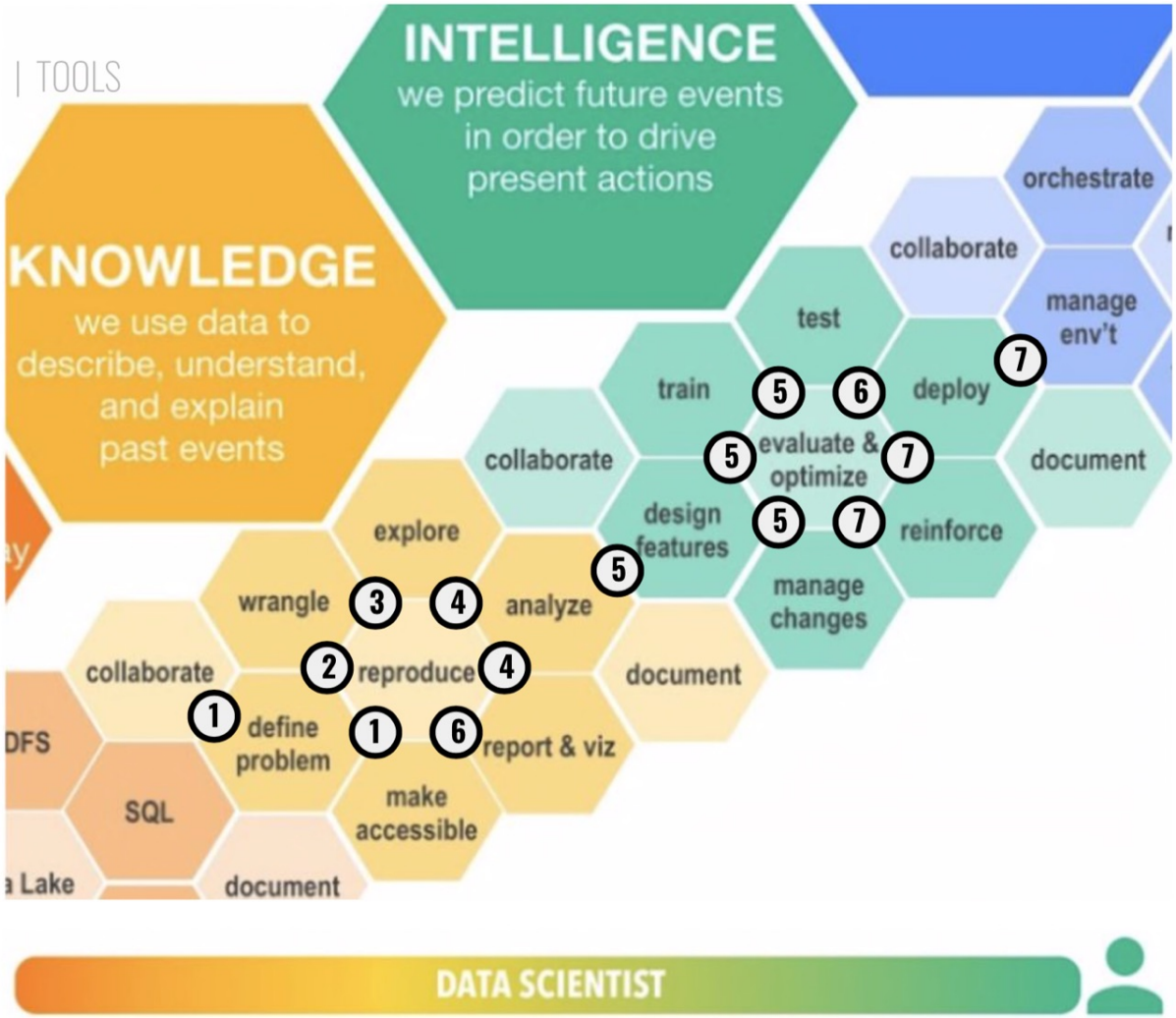



STATS
VIZ
(Cont./Num. vars)
Descriptive Stats
M
σ
25%
50%
...
$
Frequency Analysis
(Discrete, Cat)
n %
Y 5 33
N 10 67
Univariate Plots
SUM($)
Bivariate Plots
$
Churn
No Churn
Fiber
None
BR
1 2 3 4+
cust.
hypothesis testing
Inferential Stats.
test statistic
p-value
prob. that what was seen was by chance.
(cont.)
400k
200k
1k
2k
r=1
Spearman's
pearson's
correlation - Linear!
(cont./Num)
(cont.)
Avg. $/sqft
t-test
ANOVA
Mann Wh
1 Stories
2+
(discrete)
(cat/discrete)
Rate Churn
χ² test
Fiber
No fiber
(cat/discrete)
Fiber
Churn
1
2
3
4
67% fiber
0%

PIPELINE

PIPELINE | METHODOLOGIES | TOOLS

1. Planning
2. Acquisition
3. Preparation
4. Exploration
5. Modeling
6. Delivery
7. Maintenance



are cust. who churn charged a significantly different amount?
are younger customers more likely to churn?
are people w/ dependents more likely to stay longer (tenure)
are customers w/ more phone lines more likely to stay longer? (tenure)

Is there a relationship between those who
Churn & # of phone lines?
# of dependents
fiber or not
SR citizen

All the discrete, categorical vars.

χ^2

Do cust. who churn have a sign. diff.
- Monthly charges
- Age
- total charges
- tenure

compare mean
t-test,
ANOVA
Mann-Whitner

Is there a sign. diff. in monthly charges for those who churn vs. those who do not?

1) Plot.

Add color -

2) Set hypothesis

H_0: There is NO sign. diff. in $ b/w those who churn & those who don't.

$\mu_{churn}(\text{monthly charges}) == \mu_{No\,churn}(\text{monthly charges})$

3) Set alpha $\alpha = .05$

4) Verify Assumptions

H_a: There IS sign diff.

$\mu_{churn}(\text{charges}) != \mu_{no\,churn}(\text{charges})$

5) Run the test → test statistic (t-stat)
p-value.

6) Conclude ← test statistic (t-stat)
p-value.

If p-value $< \alpha$, we reject our H_0. - Type I error - Actually is NO diff

If p-value $> \alpha$, we fail to reject our H_0. - Type II if there is a diff.

LOTS of data

Too little data