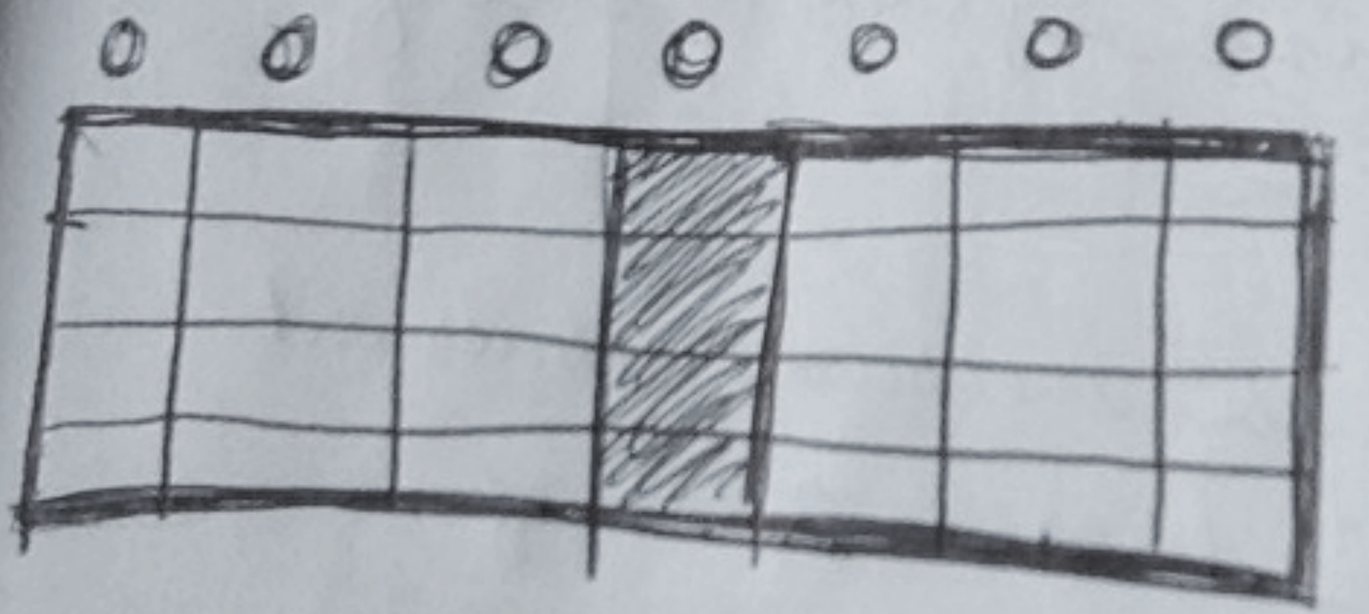
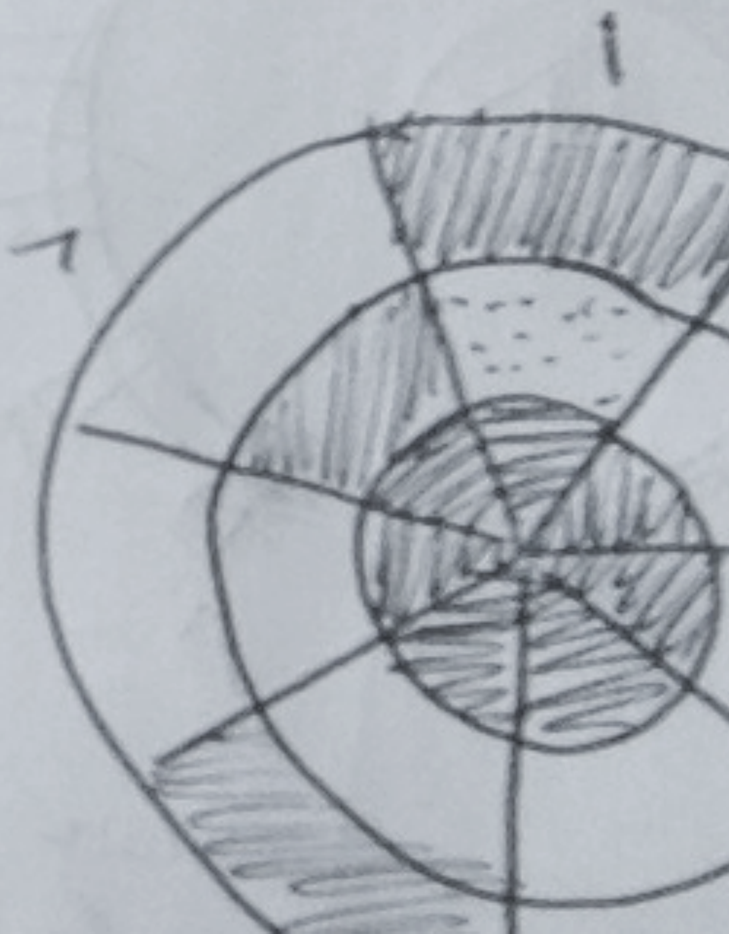

FREAKS AND GEEKS


FREQSANDGEEKS

WHAT IS OUR PLAN!?

CIRCULA



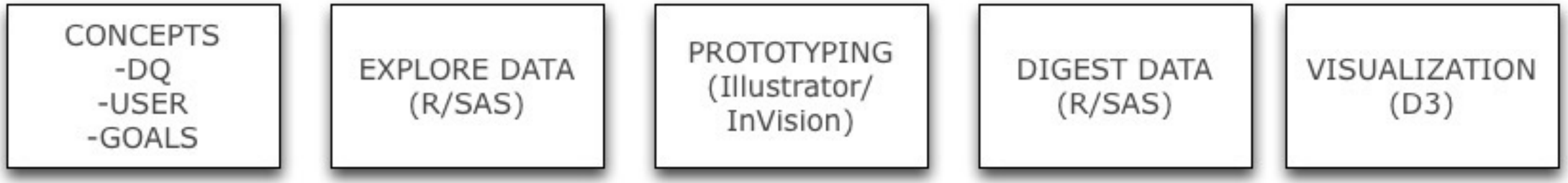


OMOPP V5

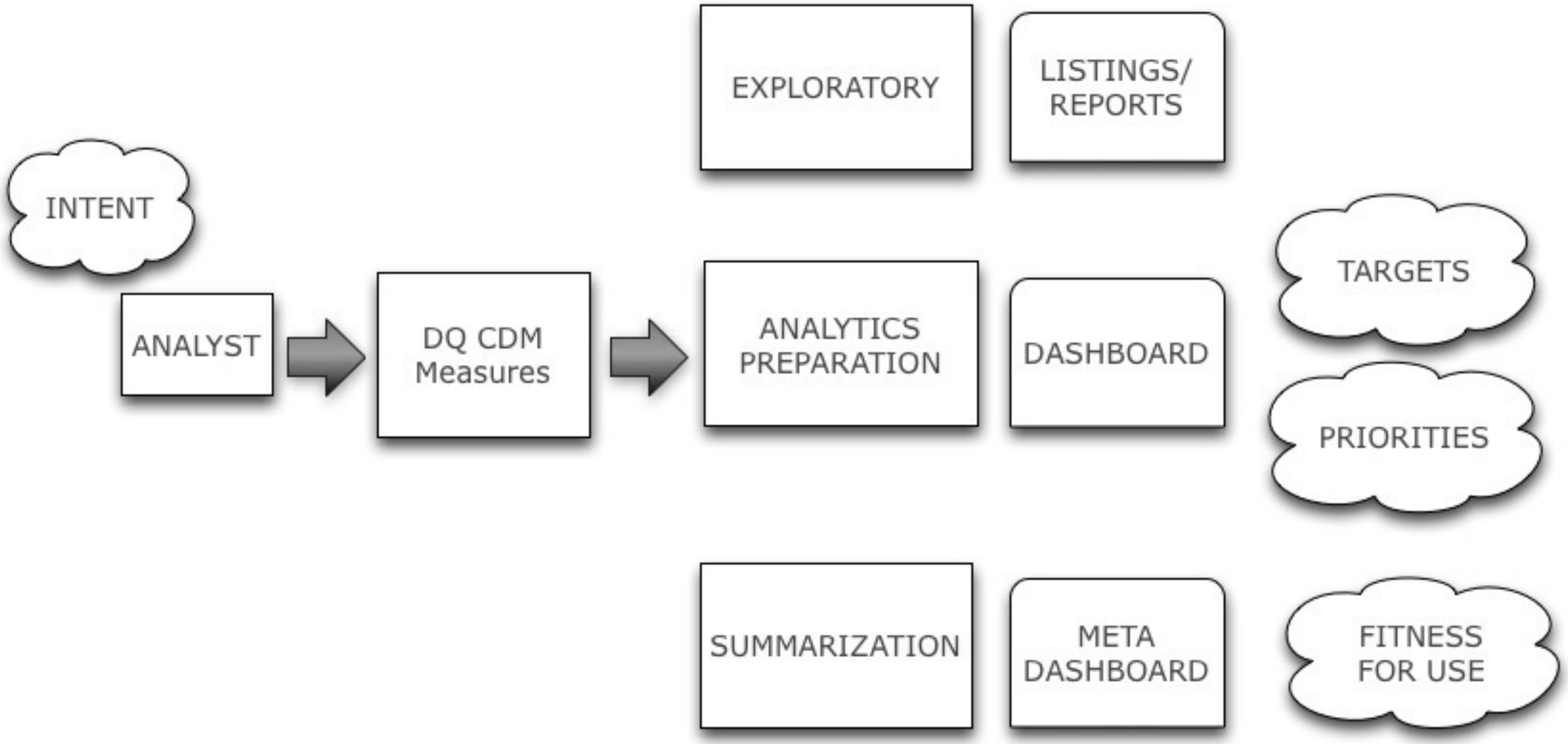
What are these data about?

Who are we designing for?

30 HOUR PROCESS



FRAMEWORK


INTENT
ANALYST
DQ CDM Measures
EXPLORATORY
ANALYTICS PREPARATION
SUMMARIZATION
LISTINGS/ REPORTS
DASHBOARD
META DASHBOARD
TARGETS
PRIORITIES
FITNESS FOR USE

GOALS

1. ANALYST DASHBOARD

3 DIFF MEASURE TYPES

1. MEASURE 1 ☐
2. MEASURE 2 ☐
3. MEASURE 3 ☐

TARGETS ☐

PRIORITIES

2. ANALYST/RESEARCHER PI META DASHBOARD

1. SUMMARY OF MEASURES ☐

META DQ SUMMARY ☐

DEM

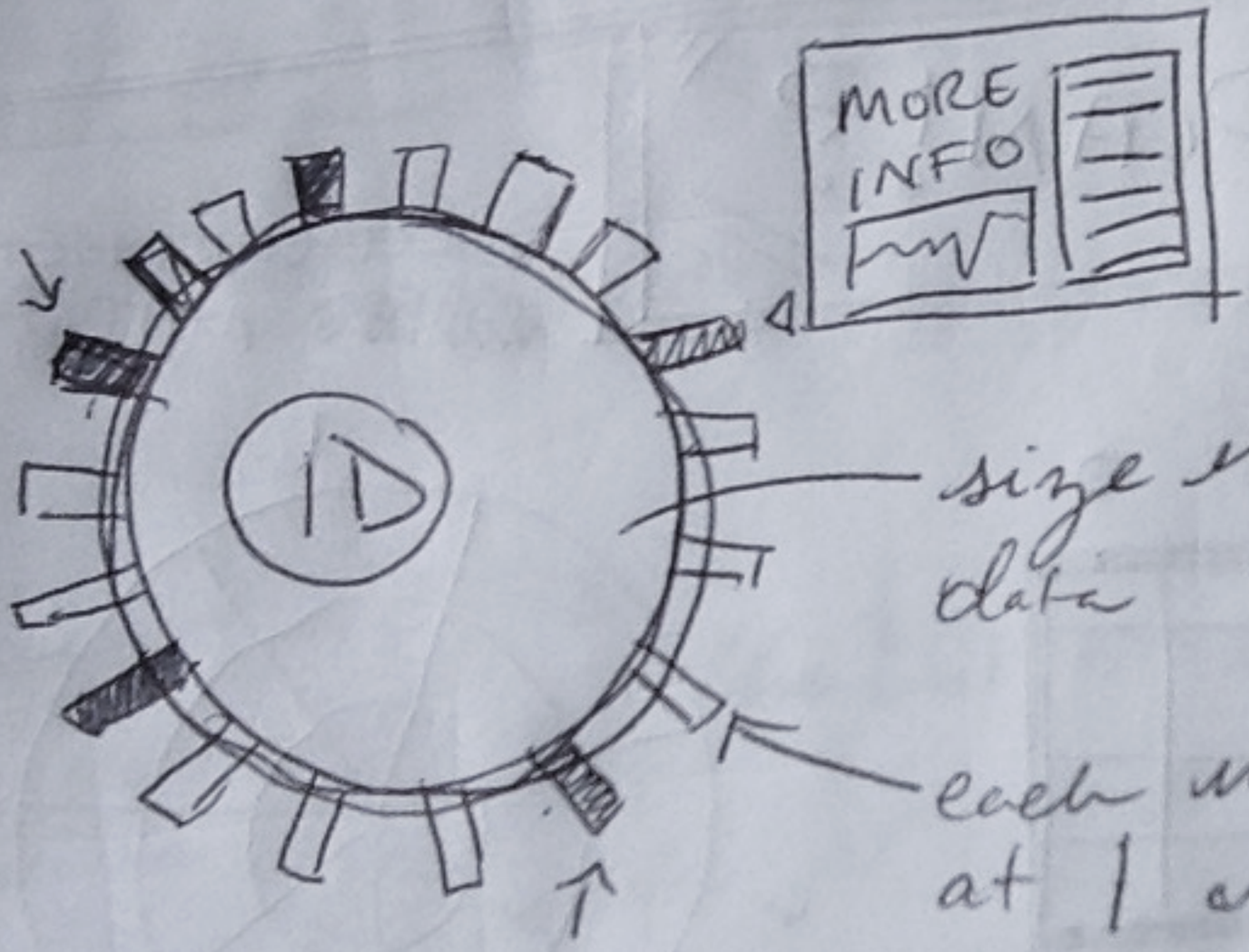
MS DATA → {1 COOK BOOK}

L1 = Completeness, Content { Missing, Content/formats }

L2 = More complex between vars between tables { between vars/tables }

L3 = "higher level"
Tables of Freqs / x-freqs
Aggregate stats
variability

L4 - Further....
Freq of Preg codes

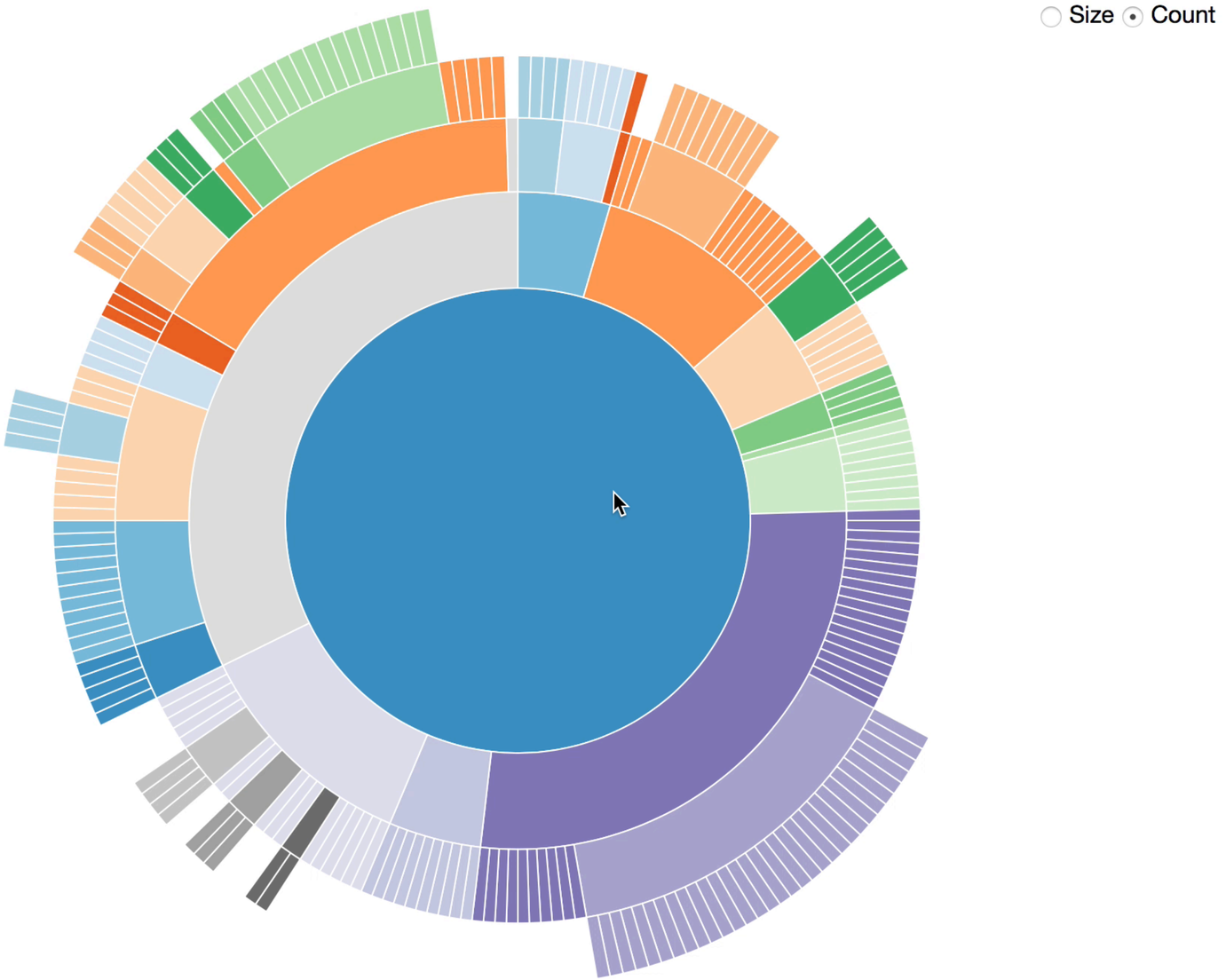


size responds to size of data

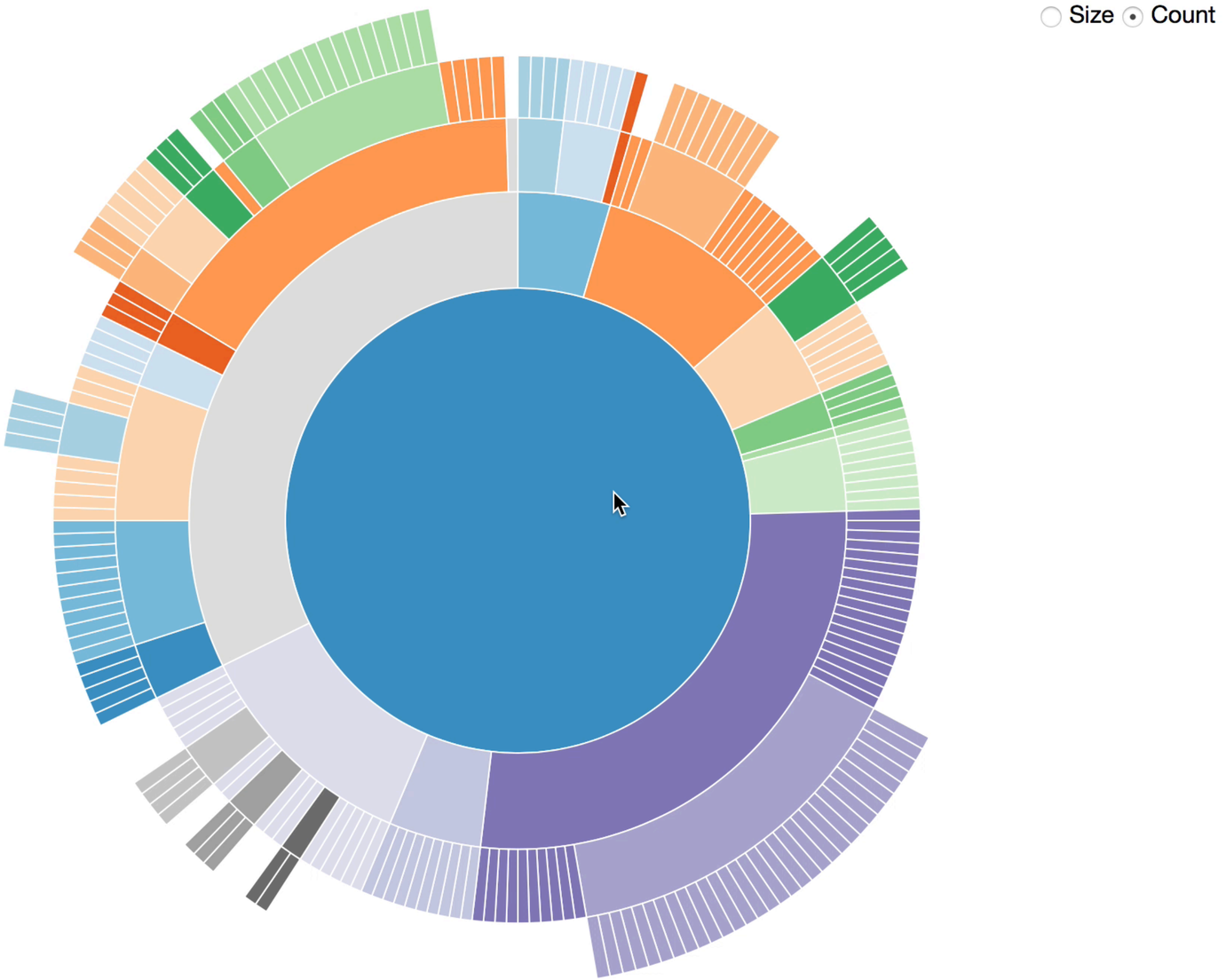
each measure comes out at 1 on 2 points of circumference

Color coded

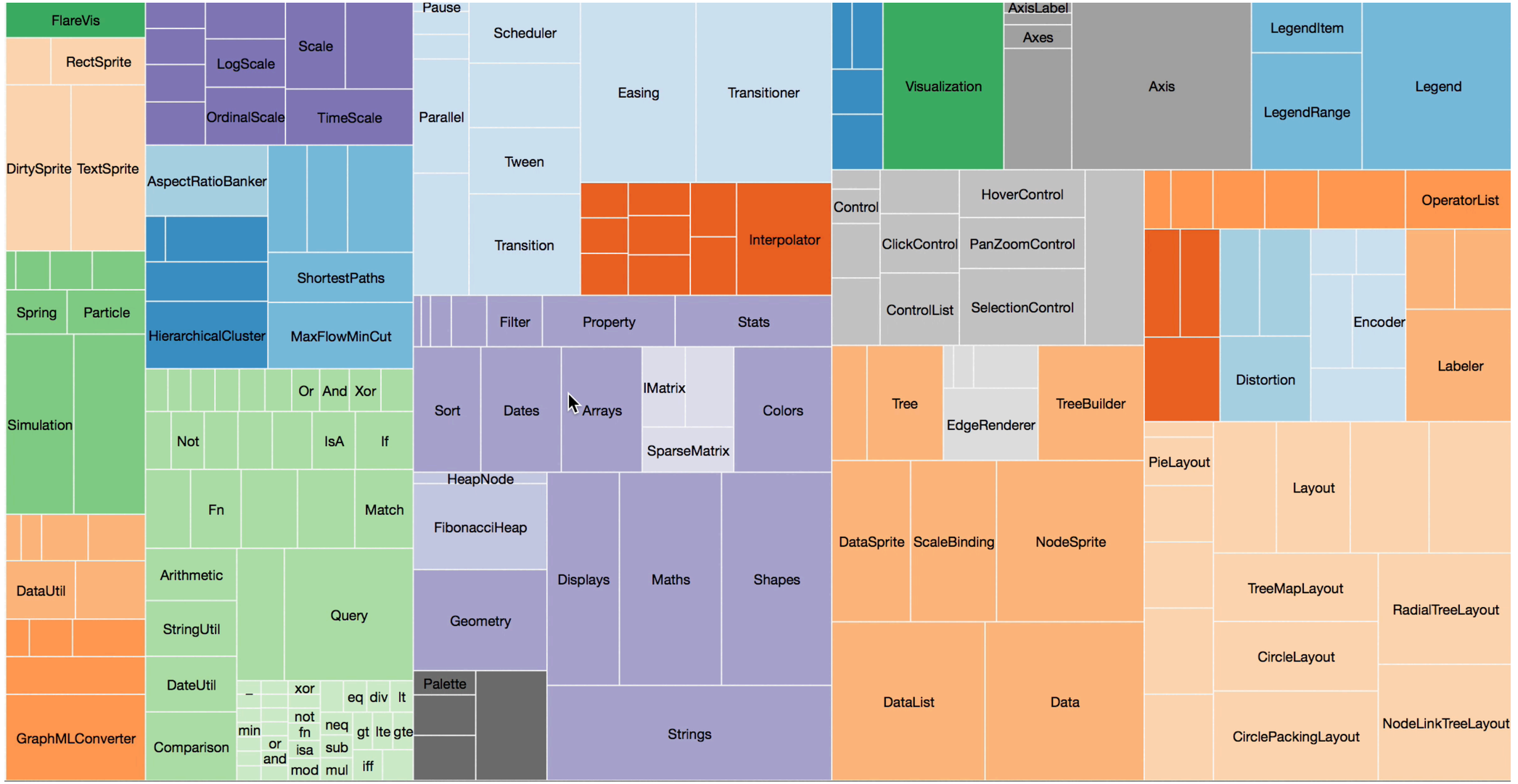
First, we wanted to make something like this.



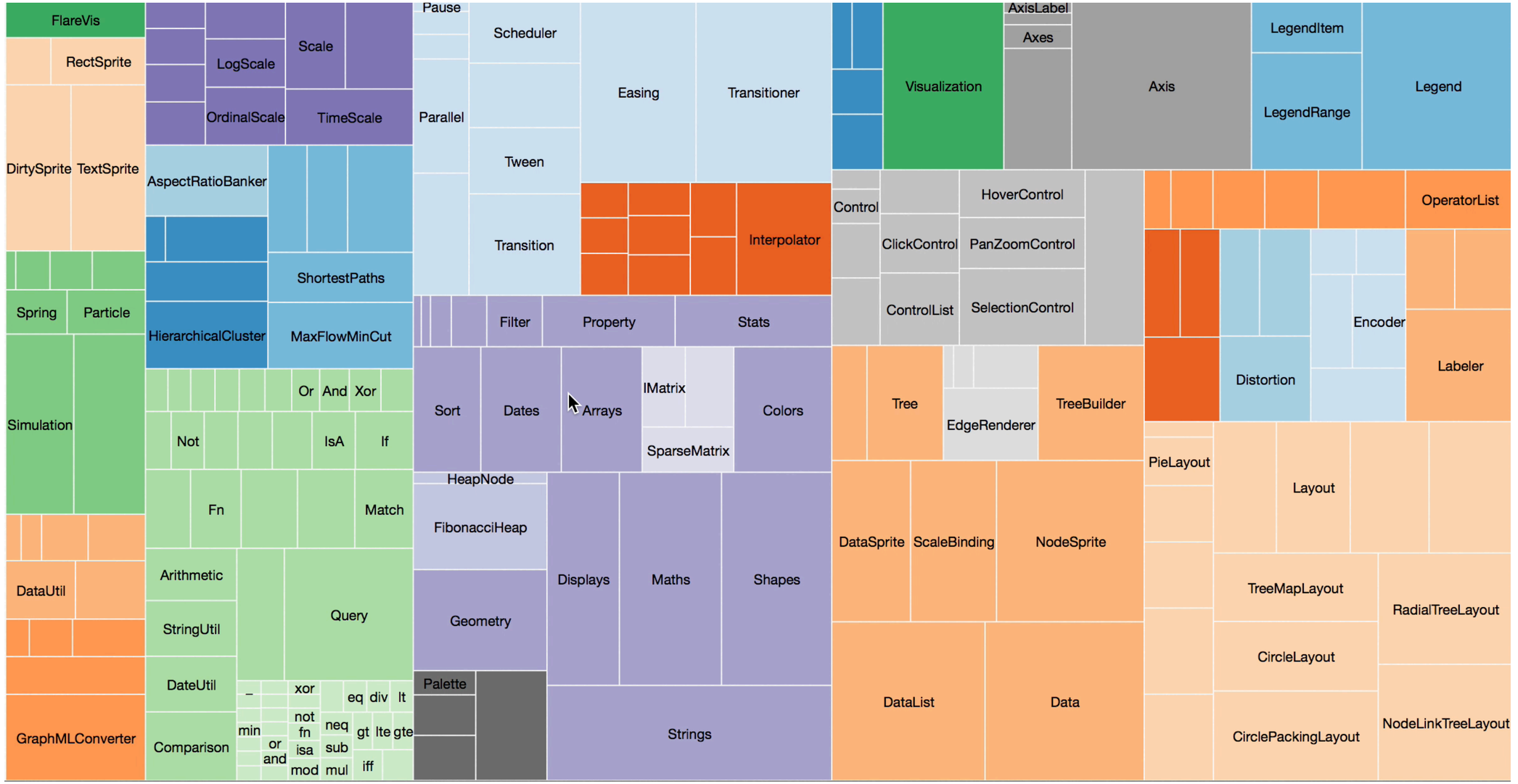
First, we wanted to make something like this.



Then, we wanted to make something like this.



Then, we wanted to make something like this.



① CAN WE ACCESS THE MORE INDIVIDUALIZED PATIENT DATA?

② Generate our own statistics → differentiate between good + bad case scenarios (makes more sense)

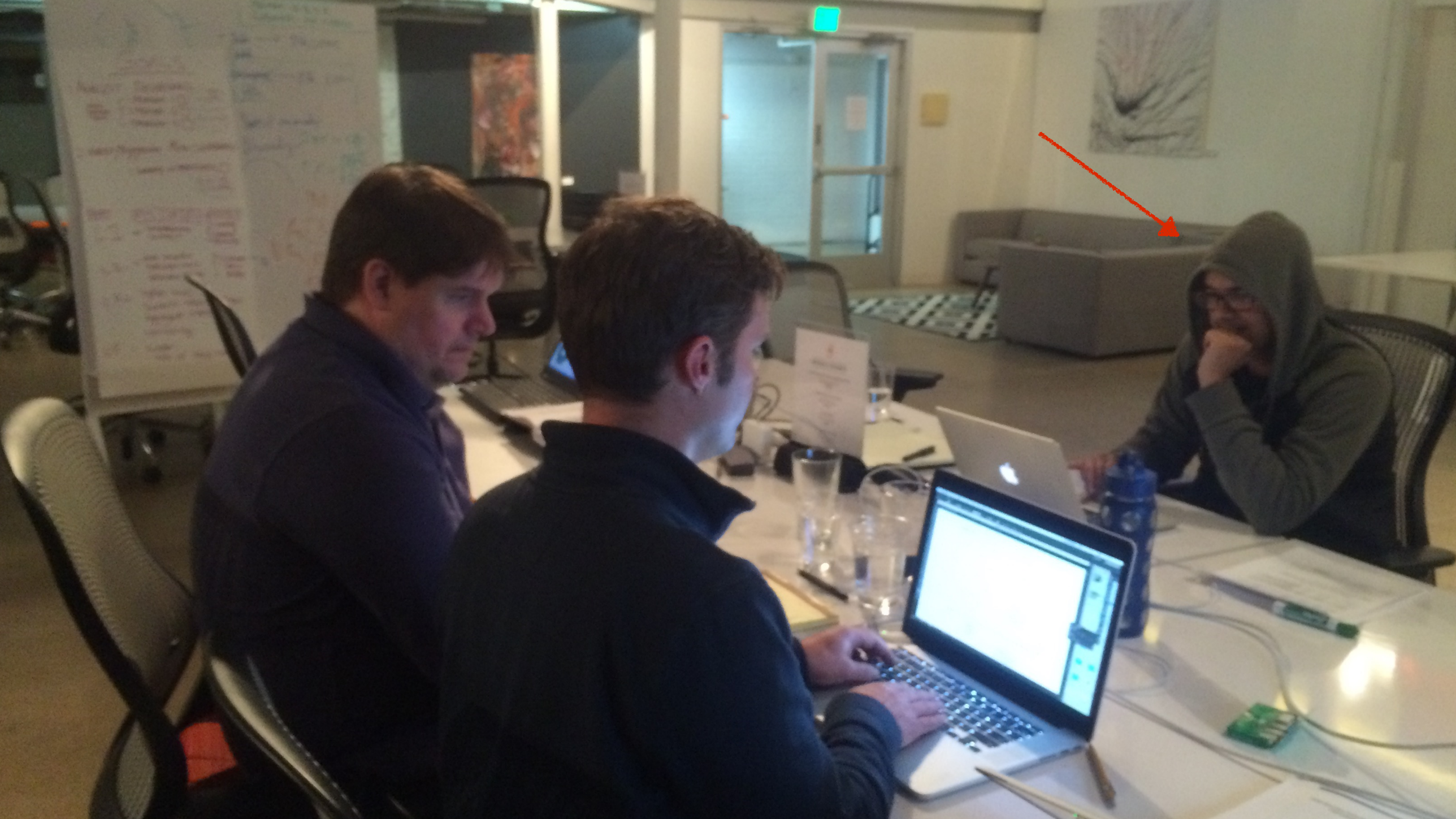
SANITY CHECKS!!! (examples)

common discrepancies

① TOOL TO FIND ANOMALIES

② WAYS TO GENERATE THESE anomalies

③ HOW TO DISPLAY THAT



868 observations of 3 variables

210	phis	1555	Px codes are invalid for >=.1% of cases.
211	phis	2264	Attending physician entered <=98% cases.
212	phis	1336	Financial data missing rate. (CTC Flag=No and >=1%)
213	phis	1867	Payer-Tertiary >= 1% unknown.
214	phis	1612	PDx code is invalid for >=.5% of cases.
215	phis	1997	Country code invalid value count
216	phis	2009	Reported to Chg based ECMO count difference >=3x median
217	phis	1222	Billing Number duplicate count
218	phis	2660	Discharge Hour = Medical Discharge Hour for 100% of cases
219	phis	1696	APGAR 10-minute score is missing for >=98% of neonates
220	phis	1324	Billed charges unmapped rate >$50K or >.2% of billed chgs
221	phis	1275	CTC unspecified code charges-Supply
222	phis	1223	Encounter duplicate count (same BN & demographics)
223	phis	2155	PDx code = symptom for >2x PHIS median rate.
224	minisentinel	1	all_l2_n_records_dupes.sas7bdat
225	minisentinel	2	cod_l1_cod.sas7bdat
226	minisentinel	3	cod_l2_patidmatch.sas7bdat
227	minisentinel	4	cod_l3_causet.sas7bdat
228	minisentinel	5	cod_l3_cod.sas7bdat
229	minisentinel	6	cod_l3_codet.sas7bdat
230	minisentinel	7	cod_l3_confidence.sas7bdat
231	minisentinel	8	cod_l3_n_patid.sas7bdat
232	minisentinel	9	cod_l3_source.sas7bdat
233	minisentinel	10	dem_l2_patidmatch.sas7bdat
234	minisentinel	11	dem_l3_ageyrsdist1.sas7bdat
235	minisentinel	12	dem_l3_ageyrsdist2.sas7bdat

219,884 observations of 5 variables

	measure_id	set_id	measure_name	result_name	value
1	1	1	all_l2_n_records_dupes.sas7bdat	name=COUNT	1.487070e+05
2	1	1	all_l2_n_records_dupes.sas7bdat	name=COUNT DISTINCT	1.487070e+05
3	1	1	all_l2_n_records_dupes.sas7bdat	name=PROBABILITY OF DUPLICATES	0.000000e+00
4	1	2	all_l2_n_records_dupes.sas7bdat	name=COUNT	4.188355e+06
5	1	2	all_l2_n_records_dupes.sas7bdat	name=COUNT DISTINCT	4.188355e+06
6	1	2	all_l2_n_records_dupes.sas7bdat	name=PROBABILITY OF DUPLICATES	0.000000e+00
7	1	3	all_l2_n_records_dupes.sas7bdat	name=COUNT	3.736901e+06
8	1	3	all_l2_n_records_dupes.sas7bdat	name=COUNT DISTINCT	3.736901e+06
9	1	3	all_l2_n_records_dupes.sas7bdat	name=PROBABILITY OF DUPLICATES	0.000000e+00
10	1	4	all_l2_n_records_dupes.sas7bdat	name=COUNT	4.762000e+03
11	1	4	all_l2_n_records_dupes.sas7bdat	name=COUNT DISTINCT	4.762000e+03
12	1	4	all_l2_n_records_dupes.sas7bdat	name=PROBABILITY OF DUPLICATES	0.000000e+00
13	1	5	all_l2_n_records_dupes.sas7bdat	name=COUNT	5.925453e+06
14	1	5	all_l2_n_records_dupes.sas7bdat	name=COUNT DISTINCT	5.925453e+06
15	1	5	all_l2_n_records_dupes.sas7bdat	name=PROBABILITY OF DUPLICATES	0.000000e+00
16	1	6	all_l2_n_records_dupes.sas7bdat	name=COUNT	1.000000e+05
17	1	6	all_l2_n_records_dupes.sas7bdat	name=COUNT DISTINCT	1.000000e+05
18	1	6	all_l2_n_records_dupes.sas7bdat	name=PROBABILITY OF DUPLICATES	0.000000e+00
19	1	7	all_l2_n_records_dupes.sas7bdat	name=COUNT	4.762000e+03
20	1	7	all_l2_n_records_dupes.sas7bdat	name=COUNT DISTINCT	4.762000e+03
21	1	7	all_l2_n_records_dupes.sas7bdat	name=PROBABILITY OF DUPLICATES	0.000000e+00
22	1	8	all_l2_n_records_dupes.sas7bdat	name=COUNT	1.888325e+06
23	1	8	all_l2_n_records_dupes.sas7bdat	name=COUNT DISTINCT	1.888011e+06
24	1	8	all_l2_n_records_dupes.sas7bdat	name=PROBABILITY OF DUPLICATES	1.662849e-04

Displayed 1000 rows of 219,884 (218,884 omitted)

- can the measures be categorized?
- how to categorize the magnitude of the anomalies

HOW TO STRATIFY THE DATA?

Size vs count
↓ ↓
relativity children

WHAT IS THE SIZE OF THE SCALE REPRESENTING?

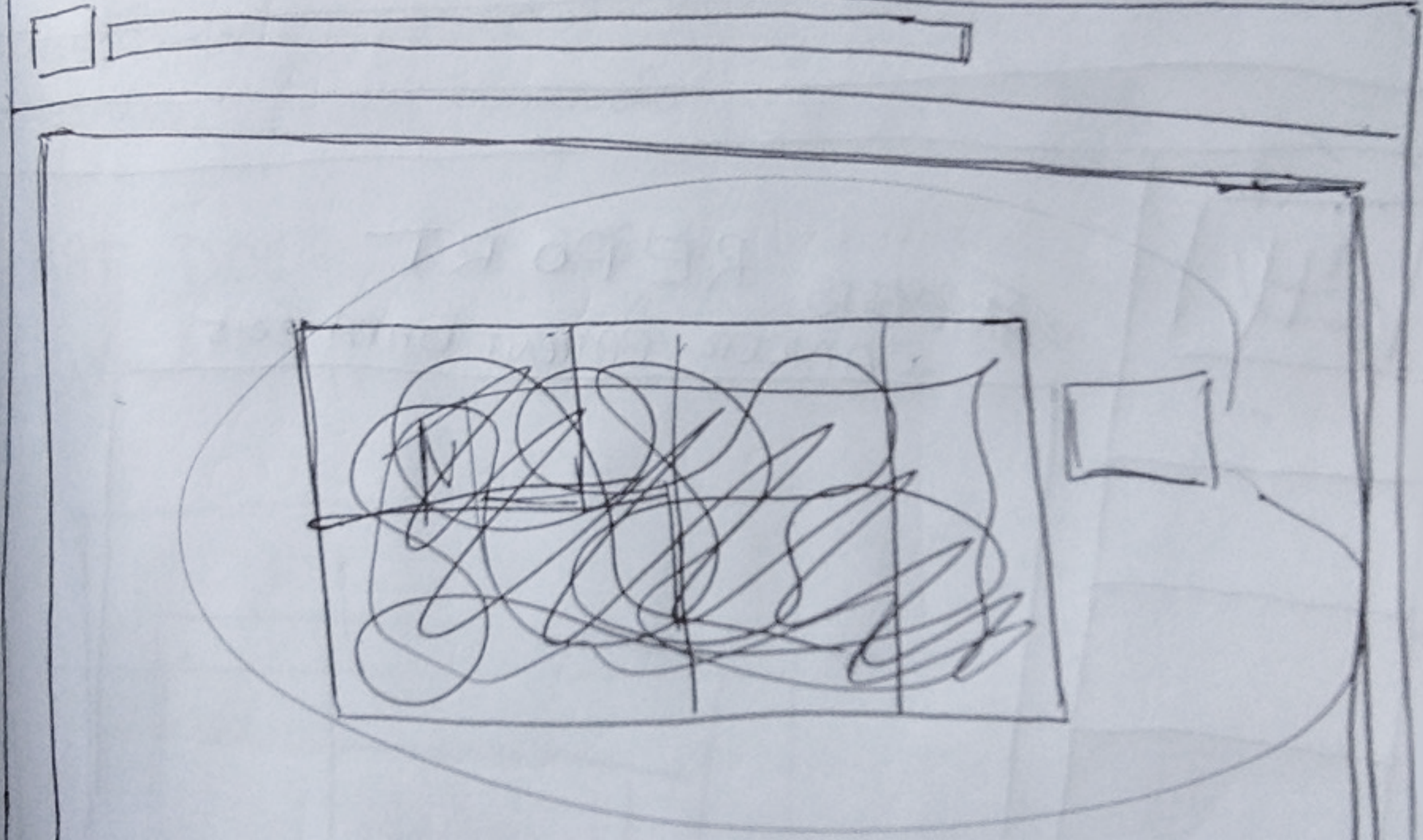
~~Categorical~~

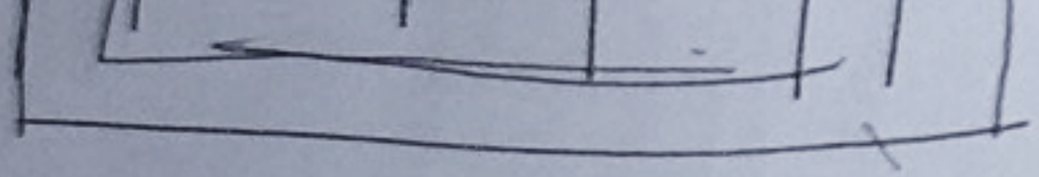
(1) WHAT QUADRANT?

Pass all (3) of them

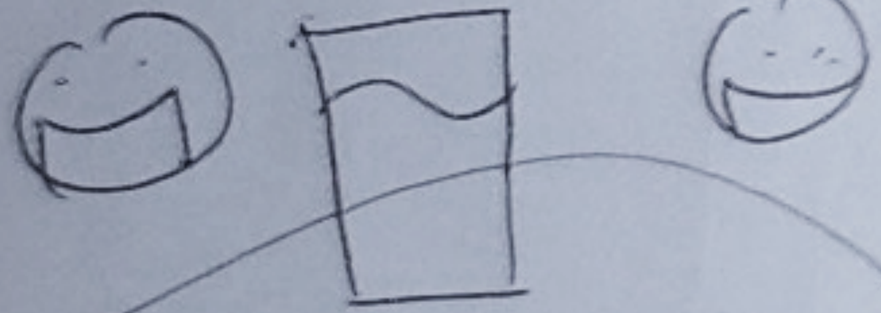
(2)

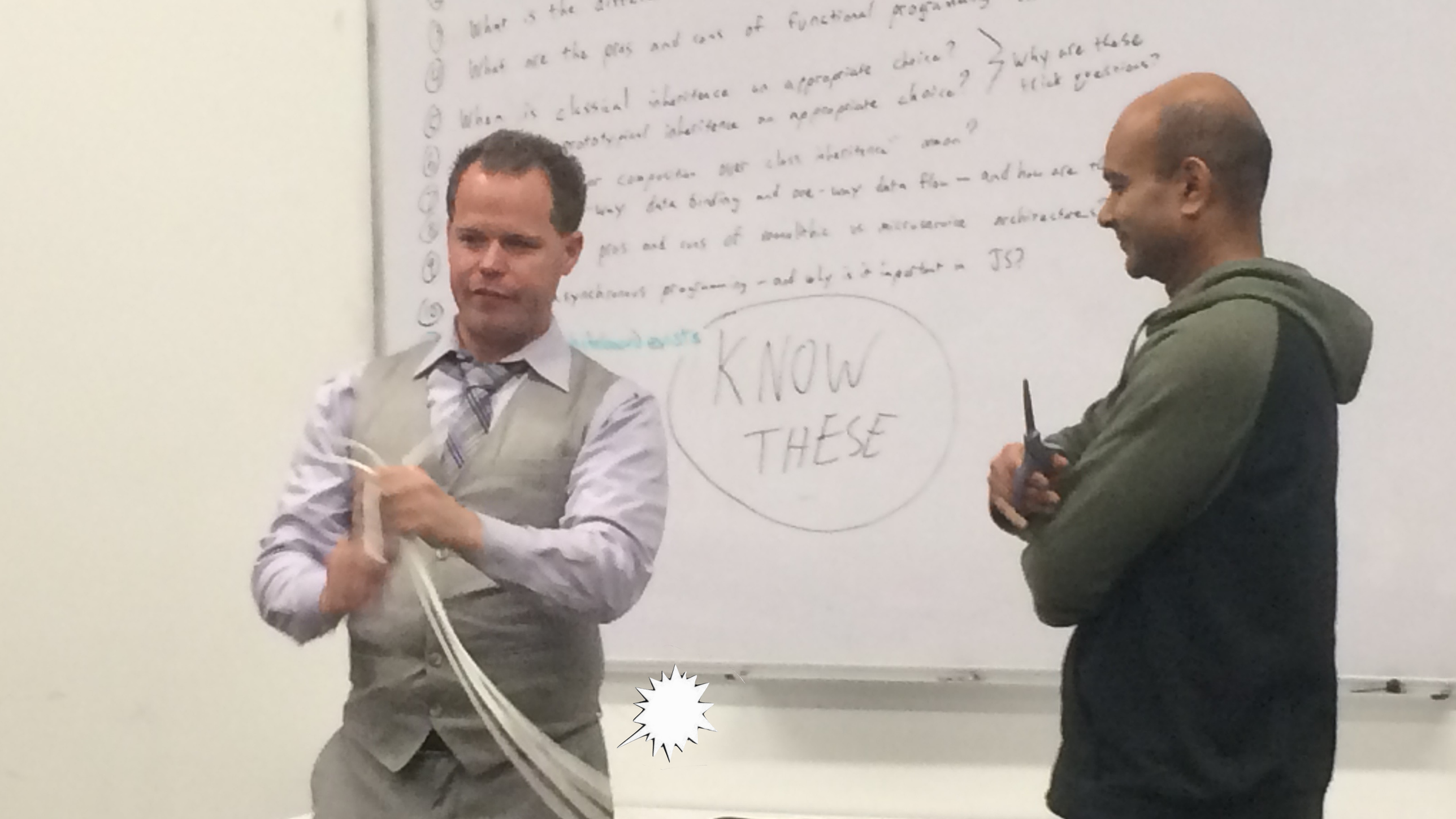
TOOL - ACCESSED THROUGH SOME SORT OF AC



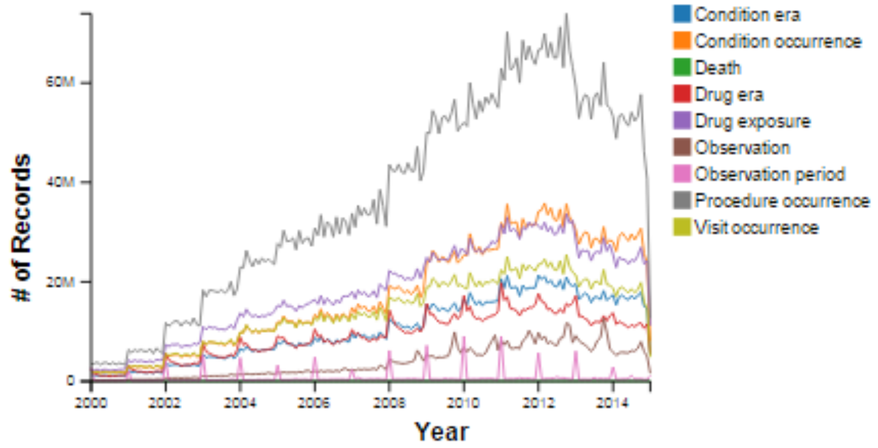


Level → Demographic → Anomaly

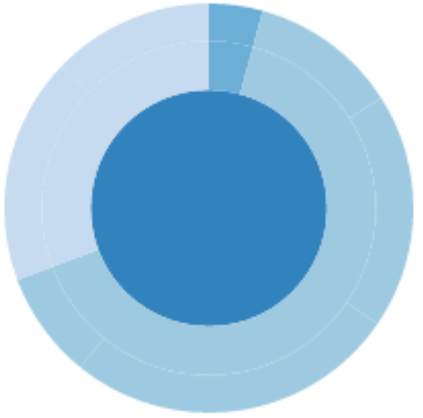



KNOW
THESE

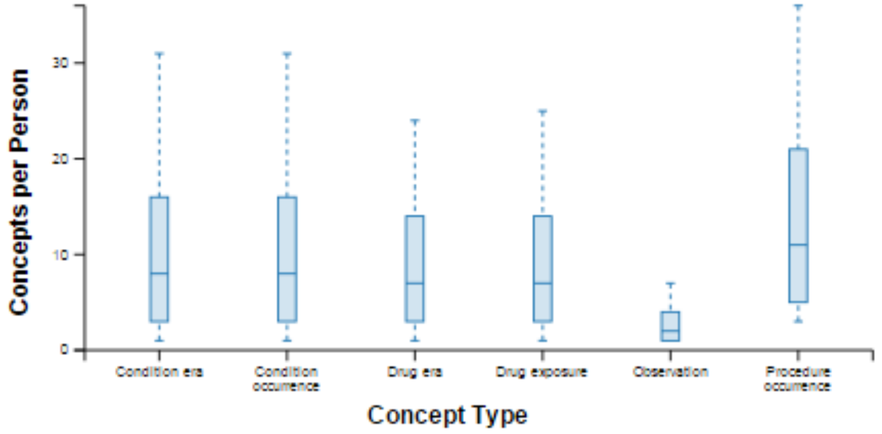
Multi-Series Line (Date Axis)



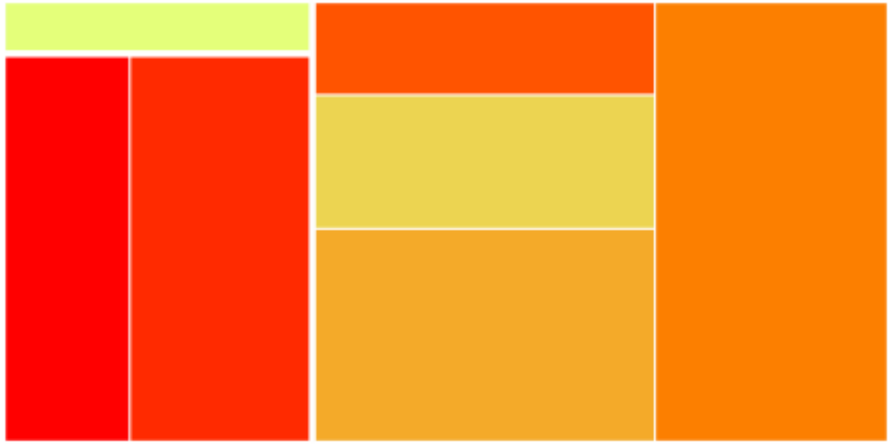
Sunburst



Boxplot



Treemap



And then, we made a prototype.

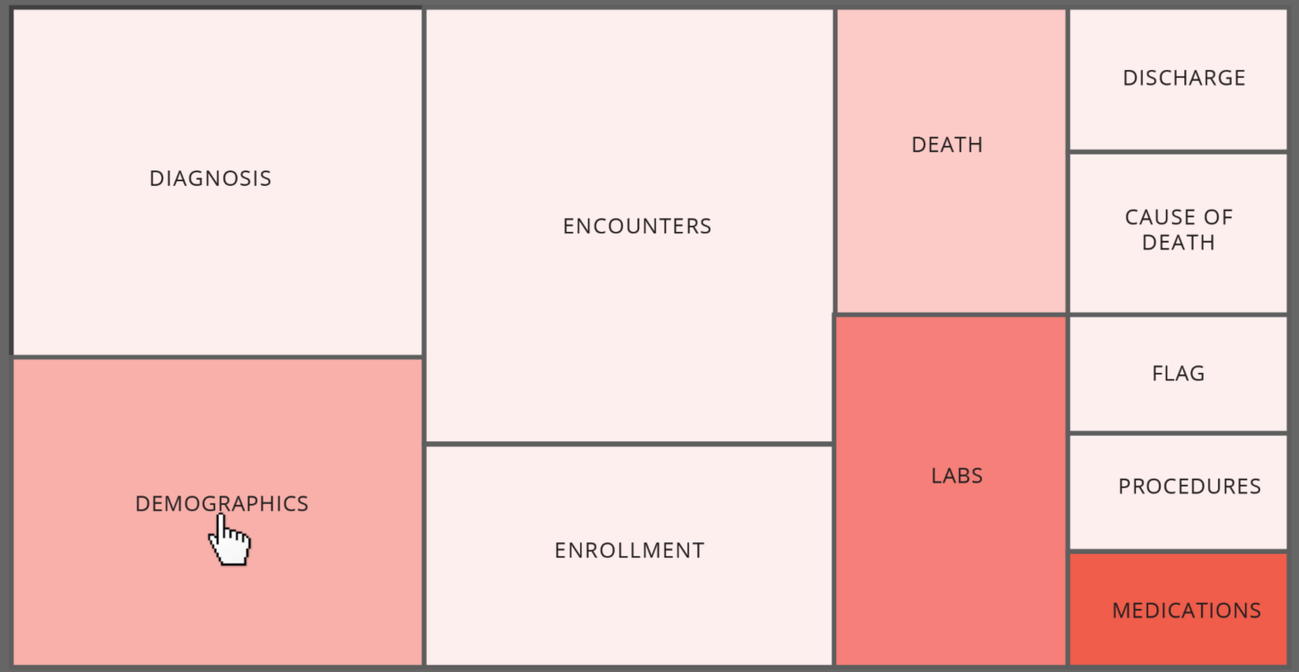
HTTPS://INVIS.IO/GK4ULWS2E

DATA QUALITY MEASURE

- SUMMARY
- COMPLETENESS
- FIDELITY
- PLAUSABILITY

SUMMARY ANALYSIS

DATA QUALITY OVERVIEW

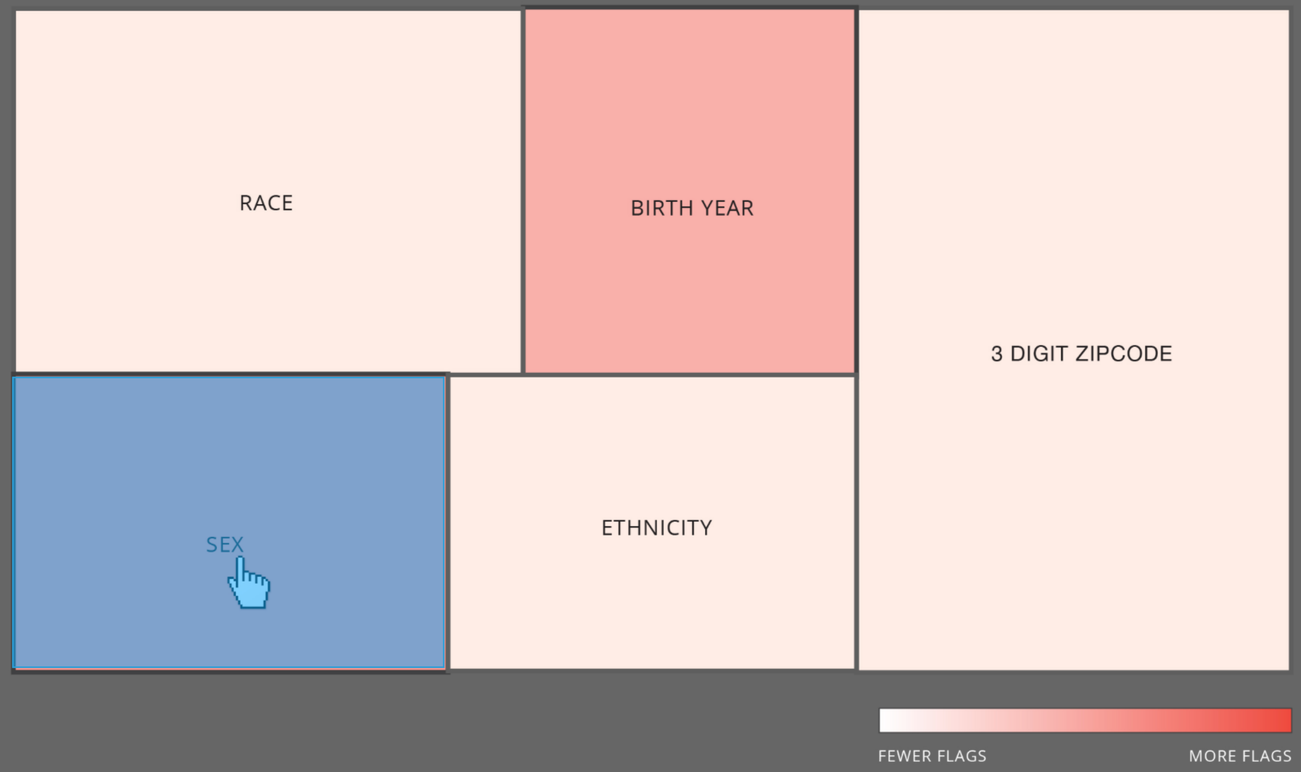


DATA QUALITY MEASURE

SUMMARY ANALYSIS

- SUMMARY
- COMPLETENESS
- FIDELITY
- PLAUSABILITY

DEMOGRAPHIC DATA OVERVIEW

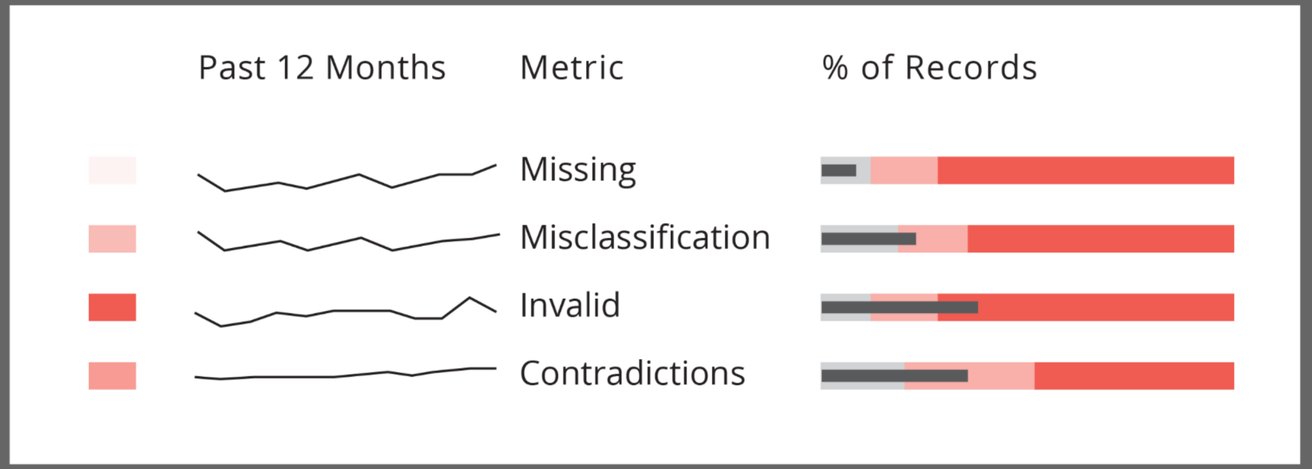


DATA QUALITY MEASURE

- SUMMARY
- COMPLETENESS
- FIDELITY
- PLAUSABILITY

SUMMARY ANALYSIS

SEX ANALYSIS



GOALS

1. ANALYST DASHBOARD

3 DIFF MEASURE TYPES

1. MEASURE 1
2. MEASURE 2
3. MEASURE 3

TARGETS

PRIORITIES

2. ANALYST/RESEARCHER PI META DASHBOARD

1. SUMMARY OF MEASURES

META DQ SUMMARY

→ Fitness of use ←