A DATA SCIENCE ANABASIS

THIS IS A STORY ABOUT A REAL LIFE DATA SCIENCE PROJECT.

DATA SCIENCE PROJECTS ARE WEIRD.

THEY DON'T FIT WITH HOW COMPANIES WORK.

BUT OCCASIONALLY SOME REAL GEMS FALL OUT UNEXPECTEDLY.

THIS IS A STORY ABOUT ONE OF THOSE GEMS.

AND THE JOURNEY THAT TOOK US THERE.

THE PROBLEM

The Great Eastern Life Assurance Co., Ltd.

(Incorporated in the Straits Settlements.)
HEAD OFFICE: - - SINGAPORE.

OUR TERM INSURANCE POLICIES ARE LAPSING

Policy No.	Assured
70583	Ng Hop Choon
Renewal for 12 months from 20th December 1936	Premium $884.50

Received this 20th day of January 1937.
the above amount being the Premium due as stated, agreeably with the terms of the Policy and those endorsed on the back hereof.

This Policy is not valid nor operative unless this receipt is counter-signed by the Agent of the Company when the said premium is paid.

Horace W. Raffles
Managing Director.

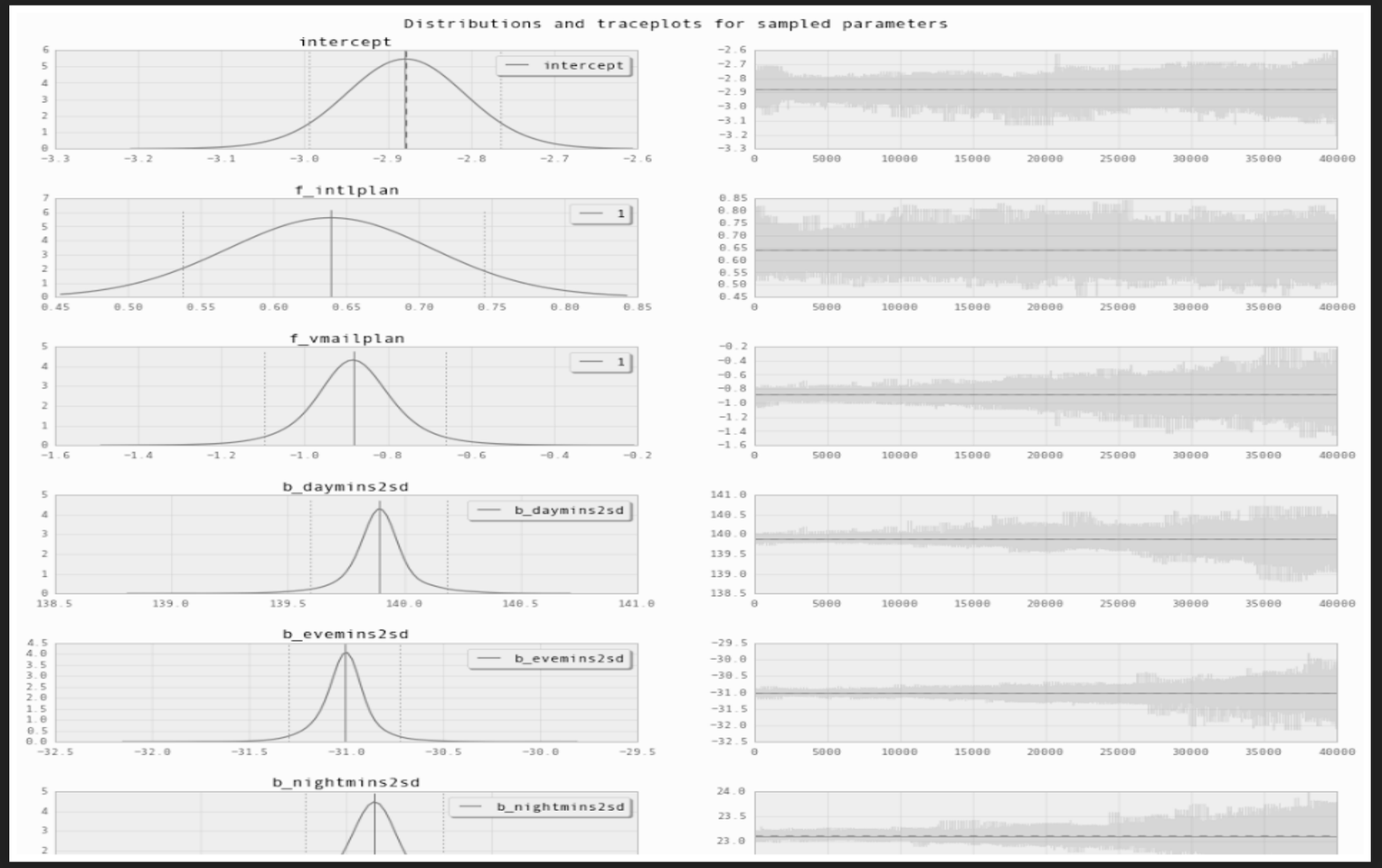
Agent at Head Office

Where the age of the assured has not been proved, it is very desirable that satis-factory evidence of age be produced to the Company, so that the age may be admitted during

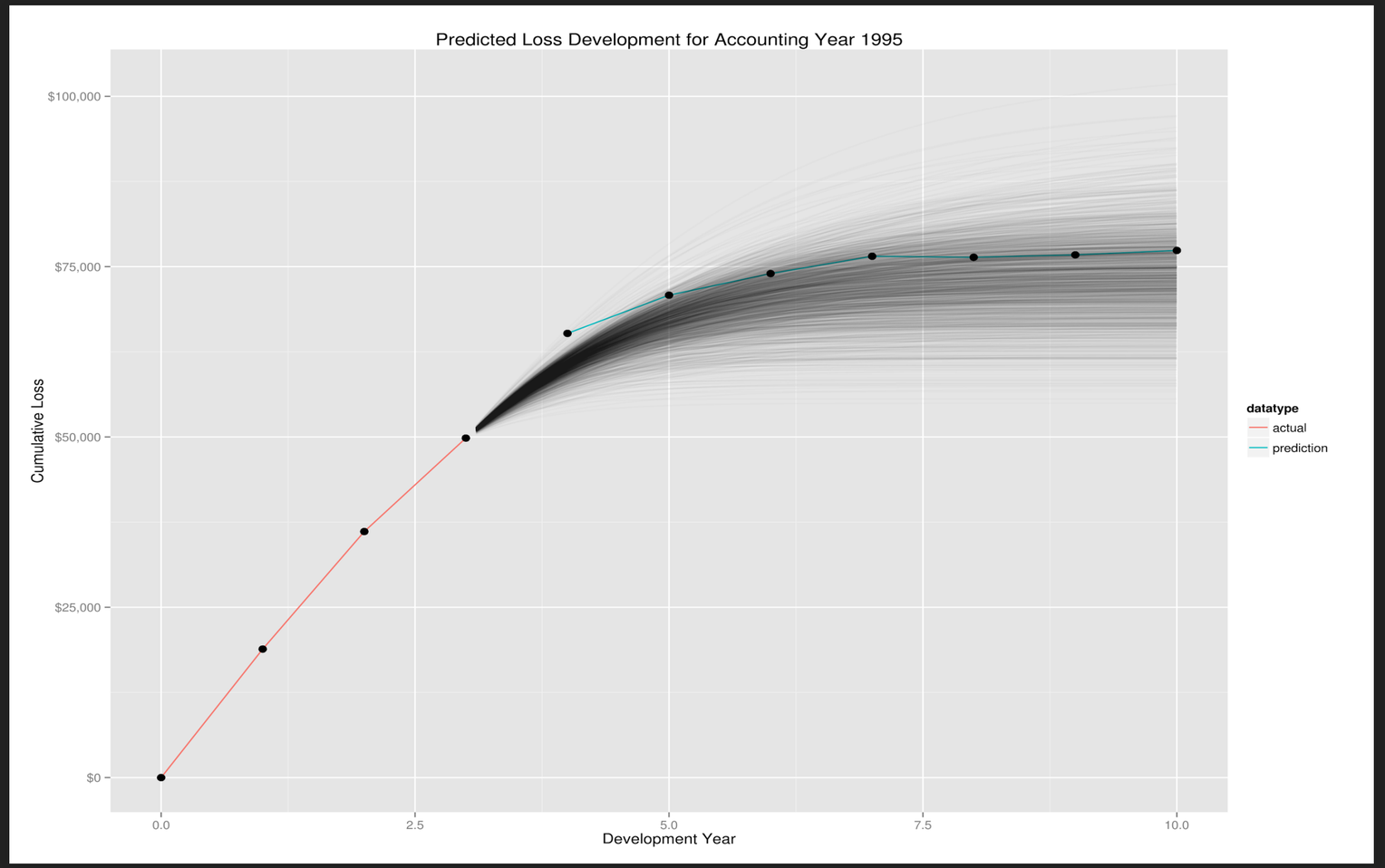
THE BRIEF

- Client: We want you to investigate the causes of lapse for a few hundred thousand policies.
- Us: Great - sounds like a cool project!
- Client: We want to use a Bayesian approach cause we think its cool!..
- ...and we met this cool guy at a conference he invented MCMC...
- ...and he wants a holiday in Ireland and you get to work with him - cool!.
- Us: Cool - what could possibly go wrong!

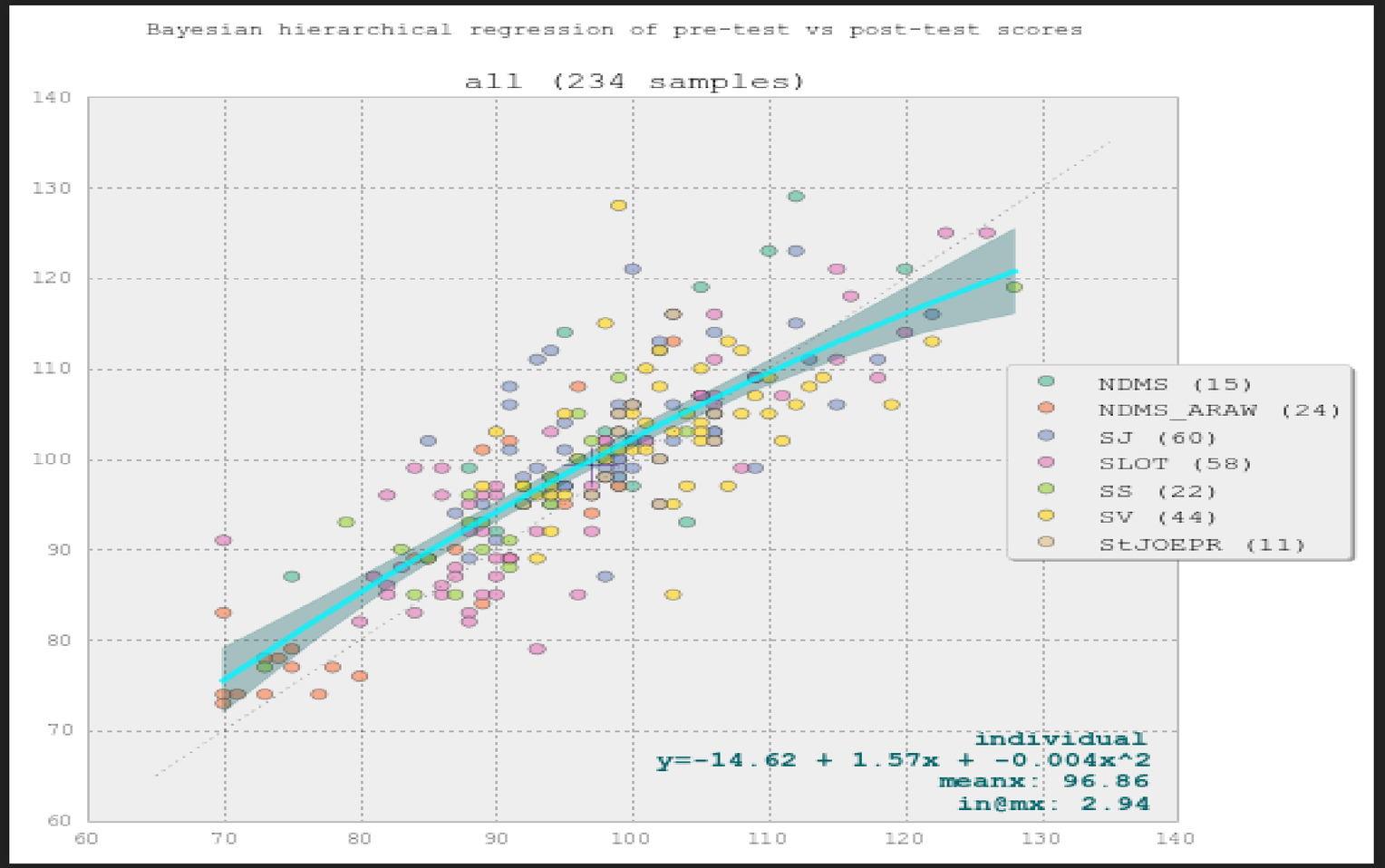
LET THE MODELLING COMMENCE



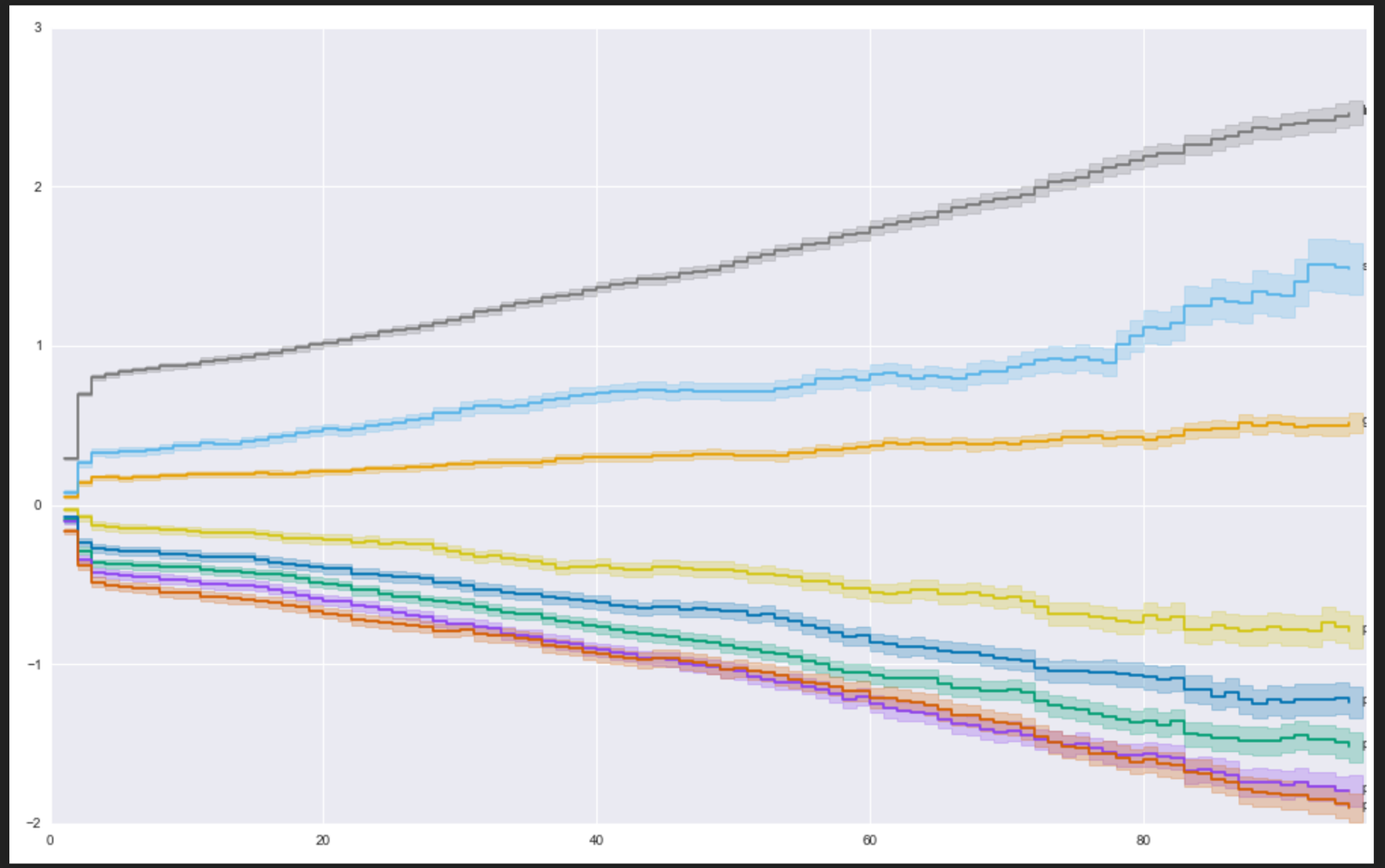
LET THE MODELLING COMMENCE



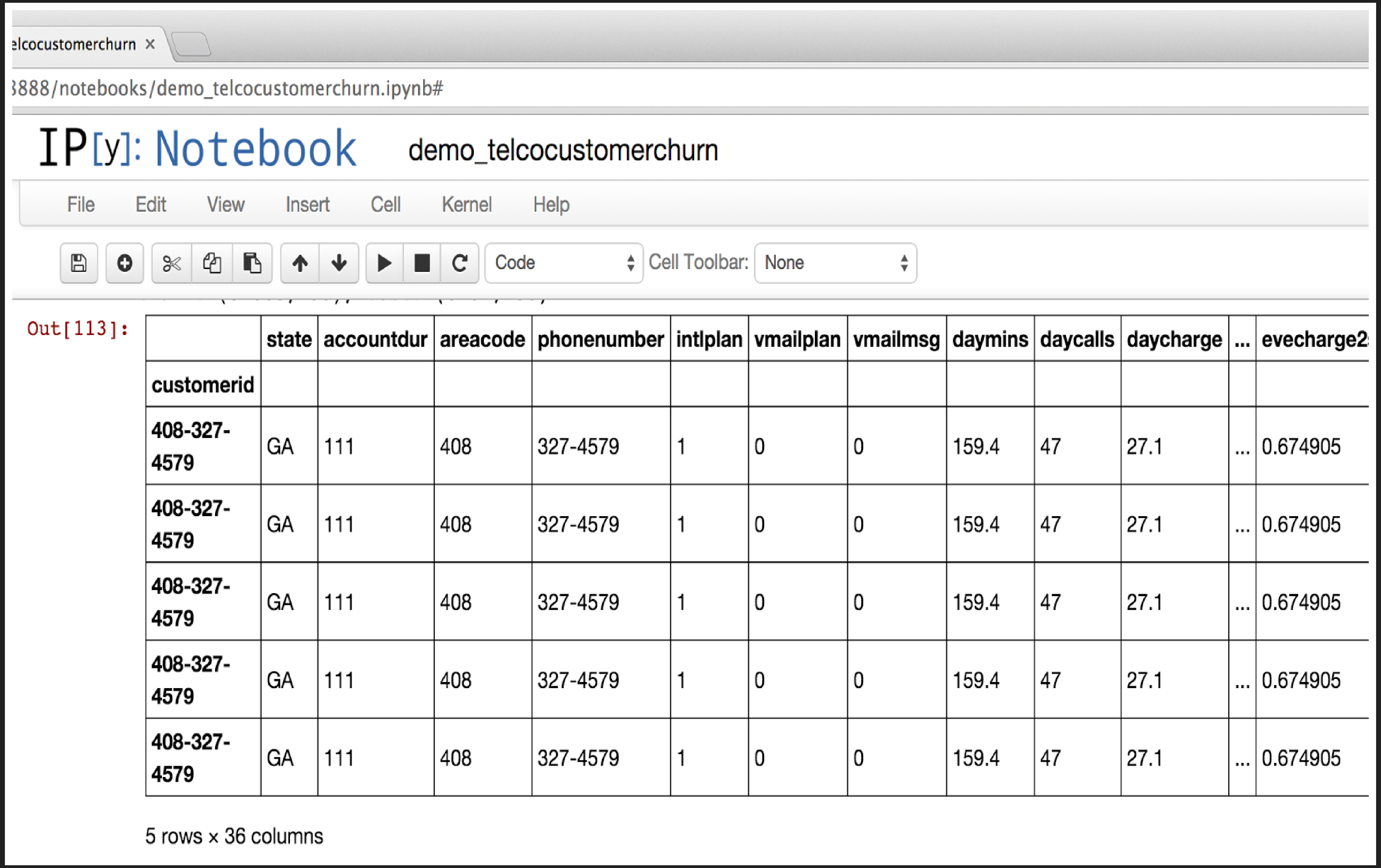
LET THE MODELLING COMMENCE



LET THE MODELLING COMMENCE



LET THE MODELLING COMMENCE



elcocustomerchurn ×

3888/notebooks/demo_telcocustomerchurn.ipynb#

IP[y]: Notebook demo_telcocustomerchurn

File Edit View Insert Cell Kernel Help

Code Cell Toolbar: None

Out[113]:

	state	accountdur	areacode	phonenumber	intlplan	vmailplan	vmailmsg	daymins	daycalls	daycharge	...	evecharge2
customerid												
408-327-4579	GA	111	408	327-4579	1	0	0	159.4	47	27.1	...	0.674905
408-327-4579	GA	111	408	327-4579	1	0	0	159.4	47	27.1	...	0.674905
408-327-4579	GA	111	408	327-4579	1	0	0	159.4	47	27.1	...	0.674905
408-327-4579	GA	111	408	327-4579	1	0	0	159.4	47	27.1	...	0.674905
408-327-4579	GA	111	408	327-4579	1	0	0	159.4	47	27.1	...	0.674905

5 rows × 36 columns

A COUPLE OF MONTHS LATER


AWKWARD QUESTIONS

WHAT HAPPENED?

- The Bayesian approach just didn't work.
- We were overawed by academic fire power...
- ...and spent too long flogging a dead horse.

Lesson 1: When you're in a hole stop digging...

...trust your judgement and go back to the...


N individuals
Select
individuals
Poisson
Binomial
Bernoulli
Normal
Gaussian
Exponential
Uniform
Gamma
Beta
Chi-squared
Dirichlet
Multinomial
Multivariate Normal
Student-t
Cauchy
Laplace
Pareto
Weibull
Log-normal
Inverse Gamma
Geometric
Negative Binomial
Hypergeometric
Discrete Uniform
Categorical
Wishart
Inverse Wishart
Logistic
Gumbel
Rayleigh
Triangular
Von Mises
Wrapped Cauchy
Wrapped Normal
Skew Normal
Skew-t
Generalized Extreme Value
Generalized Pareto
Inverse Chi-squared
Inverse Gaussian
Kumaraswamy
Noncentral Chi-squared
Noncentral F
Noncentral t
Power Function
Rice
Truncated Normal
Truncated Skew Normal
Truncated Student-t
Truncated Cauchy
Truncated Laplace
Truncated Pareto
Truncated Weibull
Truncated Log-normal
Truncated Inverse Gamma
Truncated Geometric
Truncated Negative Binomial
Truncated Hypergeometric
Truncated Discrete Uniform
Truncated Categorical
Truncated Wishart
Truncated Inverse Wishart
Truncated Logistic
Truncated Gumbel
Truncated Rayleigh
Truncated Triangular
Truncated Von Mises
Truncated Wrapped Cauchy
Truncated Wrapped Normal
Truncated Skew Normal
Truncated Skew-t
Truncated Generalized Extreme Value
Truncated Generalized Pareto
Truncated Inverse Chi-squared
Truncated Inverse Gaussian
Truncated Kumaraswamy
Truncated Noncentral Chi-squared
Truncated Noncentral F
Truncated Noncentral t
Truncated Power Function
Truncated Rice

ON A POSITIVE NOTE

- We learned a lot about...
  - The problem
  - The data
  - Bayesian statistics - always useful
  - The inherent computational complexities
  - Ninja level Python

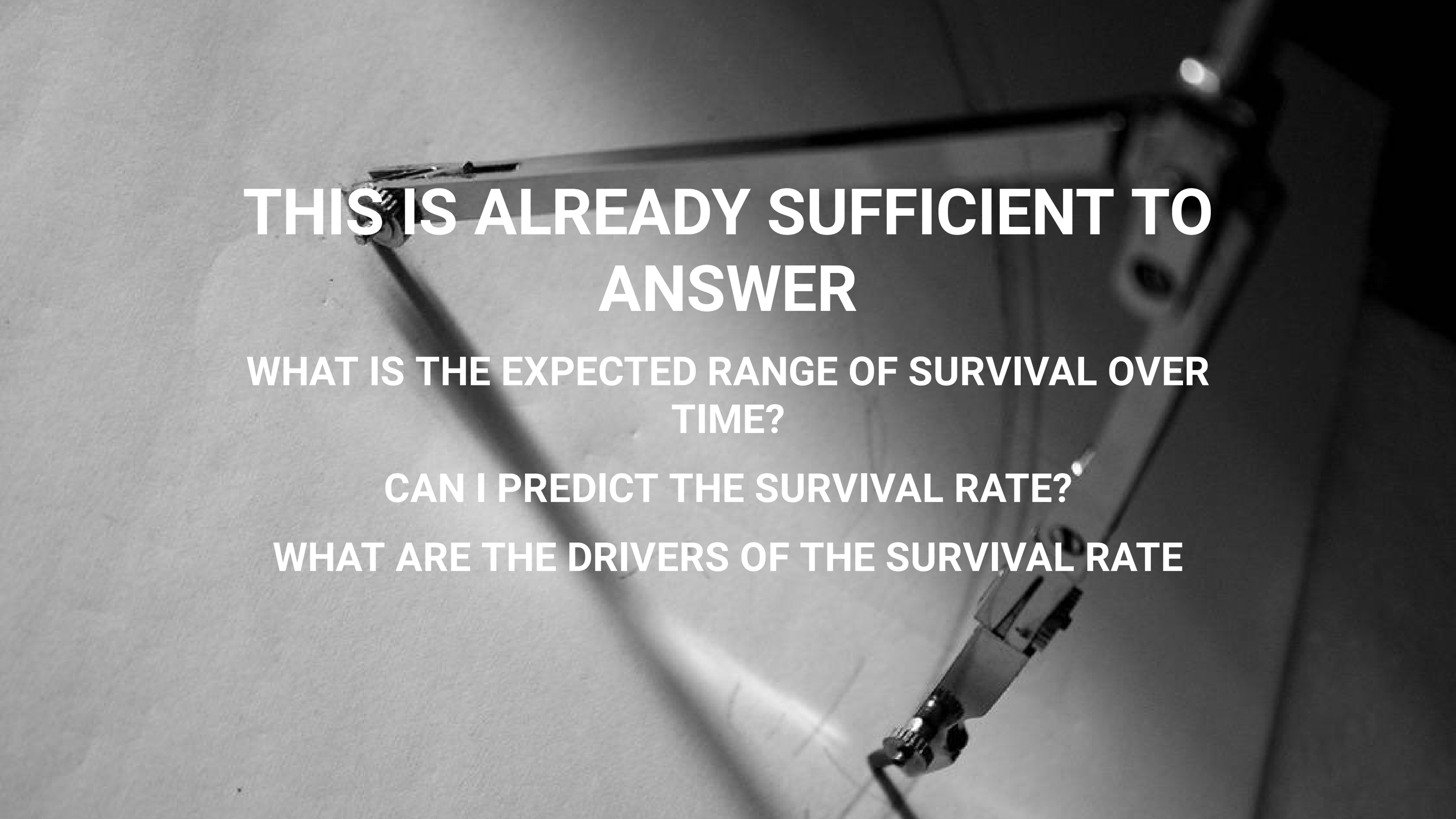
Lesson 2: Try and avoid doing computer science in the office

WHAT WE DID NEXT

WHILE I WAS THERE

- I looked up a really smart lady I'd met at R in Insurance...
- ...who runs the Risk Center at University of Barcelona.
- She told me that our approach would not work and...
- ...gave me a paper on Time Varying Survival Analysis
- I thought what's the worst that can happen?


OP SUCCESS



THIS IS ALREADY SUFFICIENT TO ANSWER

WHAT IS THE EXPECTED RANGE OF SURVIVAL OVER TIME?

CAN I PREDICT THE SURVIVAL RATE?

WHAT ARE THE DRIVERS OF THE SURVIVAL RATE

RESULTS OF SURVIVAL ANALYSIS

- We now had a good grasp on what makes policies lapse
- But the company had limited information on their clients
- Things like:
  - Age, Sex, Smoker
  - Term & Premium
  - Policy Type & Broker
  - Premium payment history

THE COMPANY HAD A CREDIT RISK PROBLEM

WHAT DID WE LEARN?

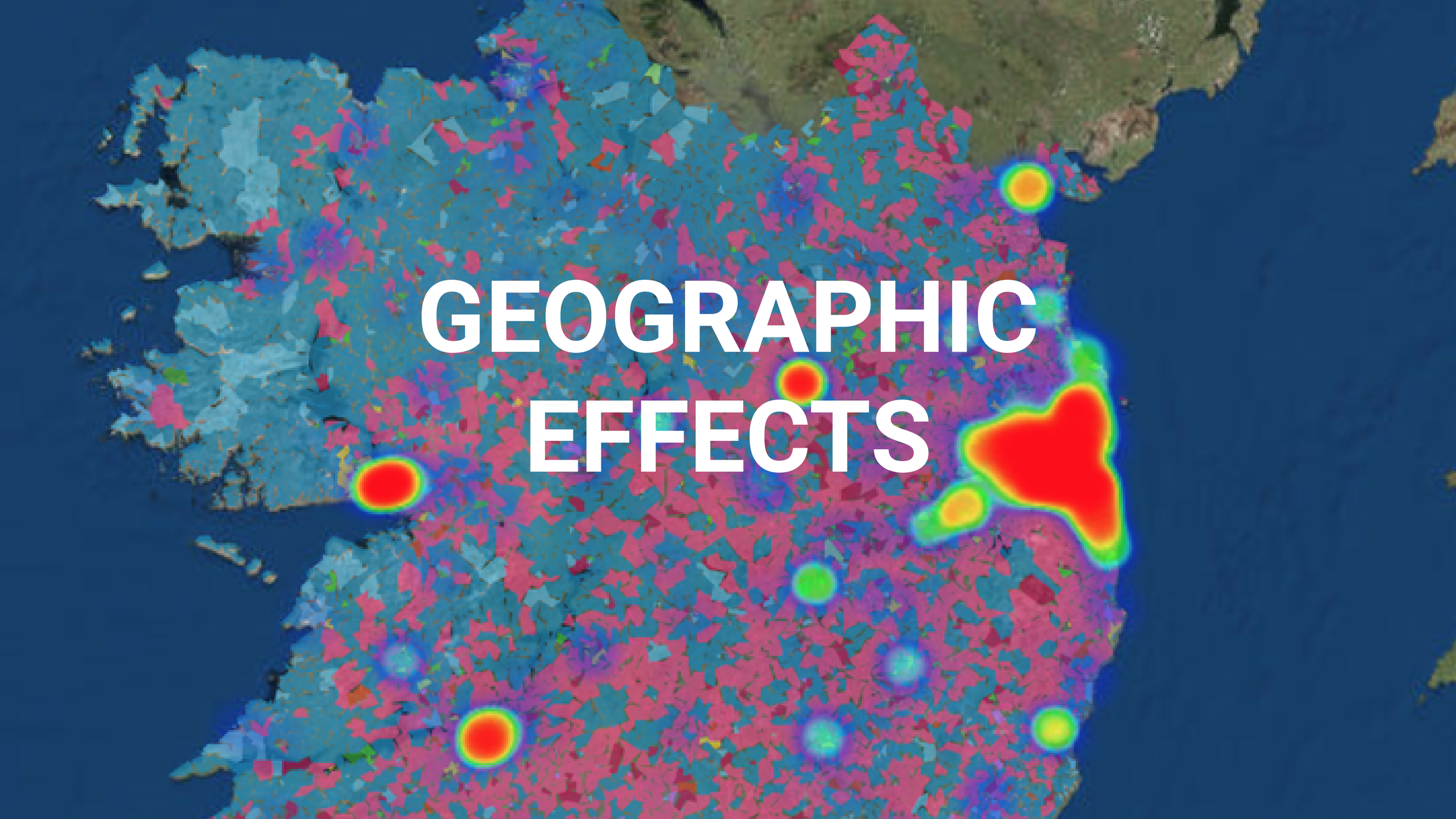
- Talk to everybody and filter later
- I can't *emphasise* this next point enough...
- ...go to...
  - Meetups
  - Technical User Groups
  - Conferences
- You'll learn a lot from other people...
- ...and every now and then a random conversation will save your life

ARE WE DONE YET?

ALONG THE WAY WE HAD NOTICED A FEW THINGS


THE ATTRITION RATE
WAS SKY-HIGH IN
NEW ESTATES


IT SEEMS THERE ARE


GEOGRAPHIC
EFFECTS

AND SOCIOECONOMIC EFFECTS

HOW CAN WE USE THEM?

- TO ENCOURAGE CUSTOMERS TO STAY
- TO HELP PRICE RISK
- IDENTIFY NEW MARKETS

SO WE CALLED EXPERIAN

IT WAS A SHORT CONVERSATION

HI, I'D LIKE SOCIOECONOMIC INFORMATION FOR 250K ADDRESSES

HOW MUCH!

OKAY SO THAT'S A NON-RUNNER

GEOCODING

- **FIRST LET'S GEO-CODE OUR ADDRESSES**
- **WE HAD TWO CHOICES:**
  - Use Google - which you pay for
  - Use Nominatim - FOSS / roll your own

WE TRIED BOTH

FOR IRELAND, GOOGLE IS BETTER

MAINLY BECAUSE ...

IRISH ADDRESSES ARE PATHOLOGICAL!

BUT WE NOW HAVE A LAT / LONG FOR EACH CLIENT

SHOPPING FOR DATA

THERE'S A LOT OF IT AROUND

WE LOOKED AT THE CENSUS DATA

IN IRELAND A CENSUS IS DONE EVERY FIVE YEARS

THE AMOUNT OF INFORMATION IN IT IS ASTOUNDING

APPROXIMATELY 700 FEATURES

COVERING 15 THEMES

WHAT DATA IS AVAILABLE

CENSUS DATA THEMES

Theme	Subject	Theme	Subject
1	Sex, Age & Migration	9	Social class
2	Ethnicity & Religion	10	Education
3	Irish Language	11	Commuting
4	Families	12	Health
5	Private Households	13	Occupation
6	Housing	14	Industries
7	Hospitals & Prisons	15	PC & Internet

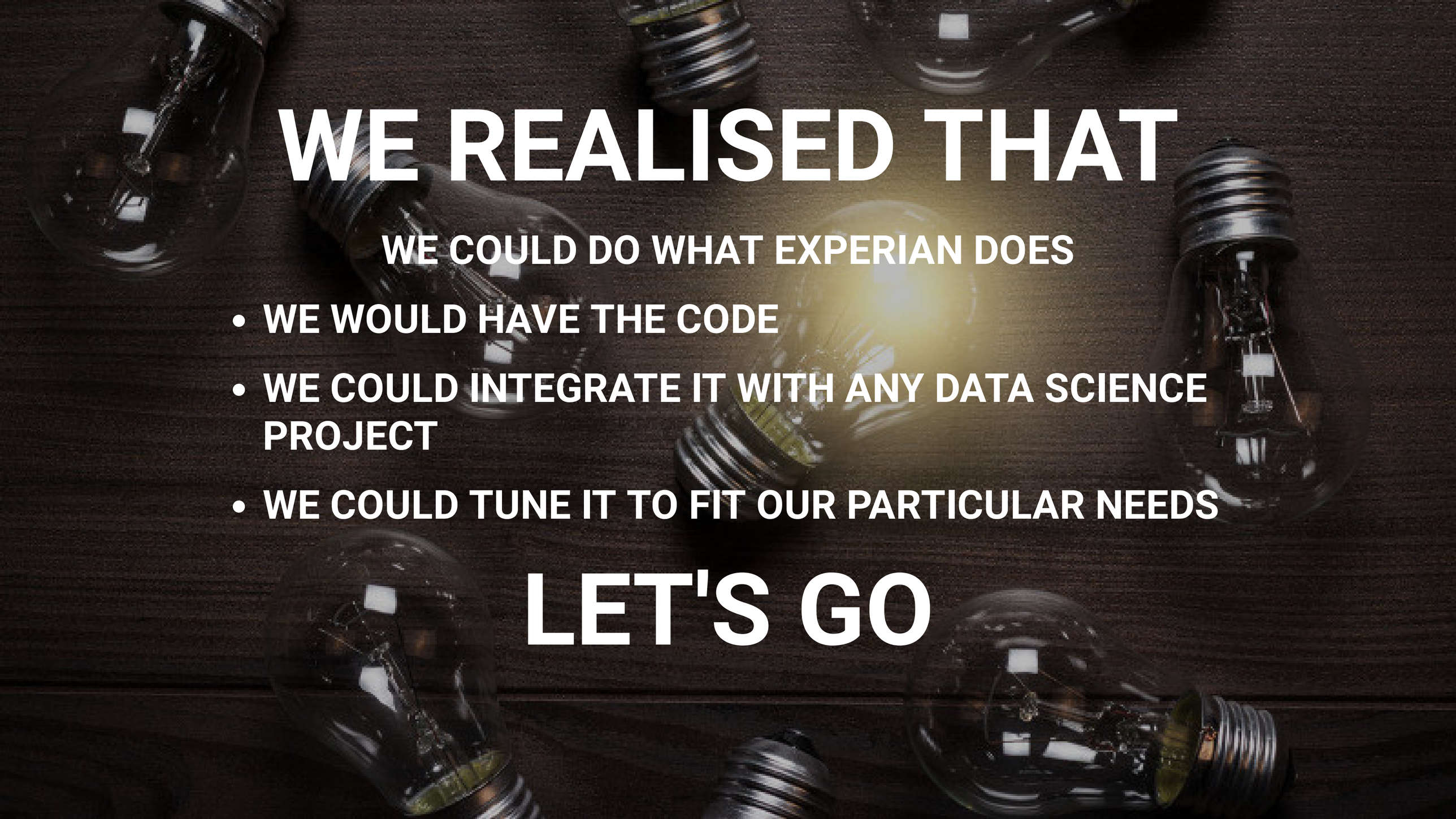

IT GETS BETTER

SMALL AREA MAPS

THE SMALLEST OUTPUT AREA FOR CENSUS DATA

- ~20,000 SMALL AREAS COVERING IRELAND
- EACH COVERS APPROX. 200 PEOPLE
- EACH CENSUS FEATURE AVAILABLE AT THIS LEVEL

THIS IS TERRIFYING



WE REALISED THAT

WE COULD DO WHAT EXPERIAN DOES

- WE WOULD HAVE THE CODE
- WE COULD INTEGRATE IT WITH ANY DATA SCIENCE PROJECT
- WE COULD TUNE IT TO FIT OUR PARTICULAR NEEDS

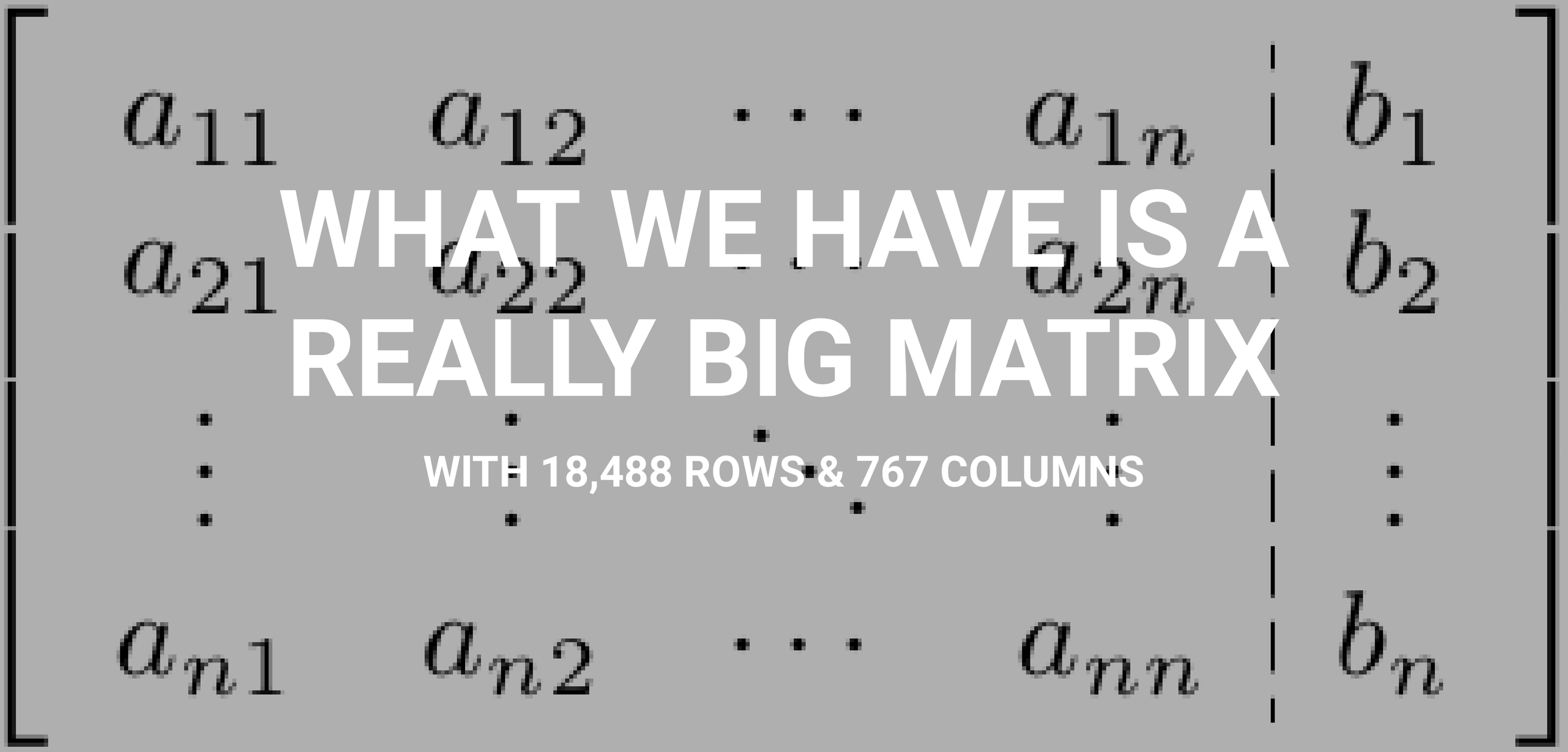
LET'S GO


LET'S LOOK AT THE DATA

NOT A TRIVIAL TASK

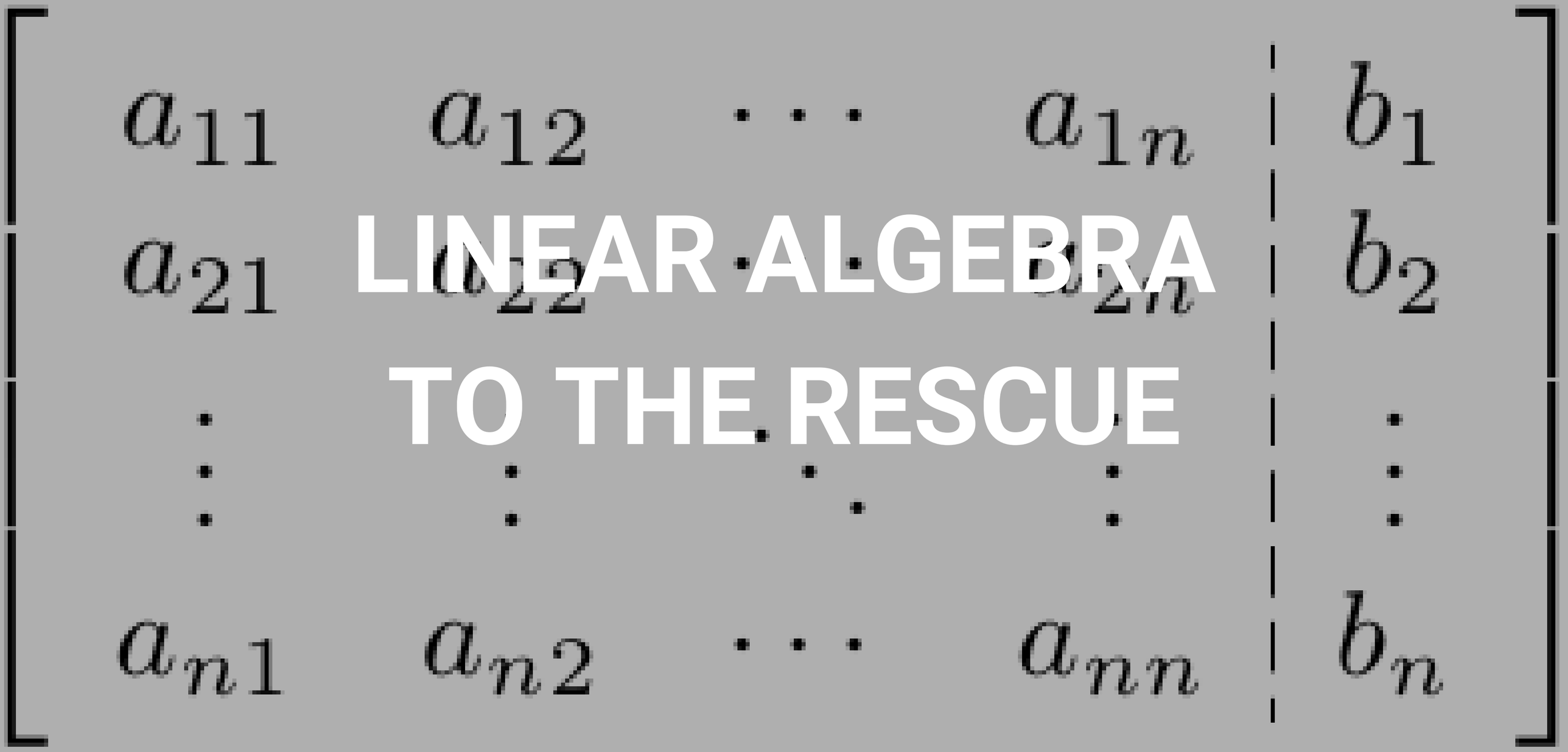
IT'S HARD TO MAKE SENSE OF THIS MUCH DATA:

- THERE ARE 18,488 SMALL AREA MAPS
- EACH SMALL AREA MAP IS REPRESENTED BY A ROW
- EACH ROW HAS 767 ENTRIES ONE FOR EACH FEATURE



WHAT WE HAVE IS A REALLY BIG MATRIX

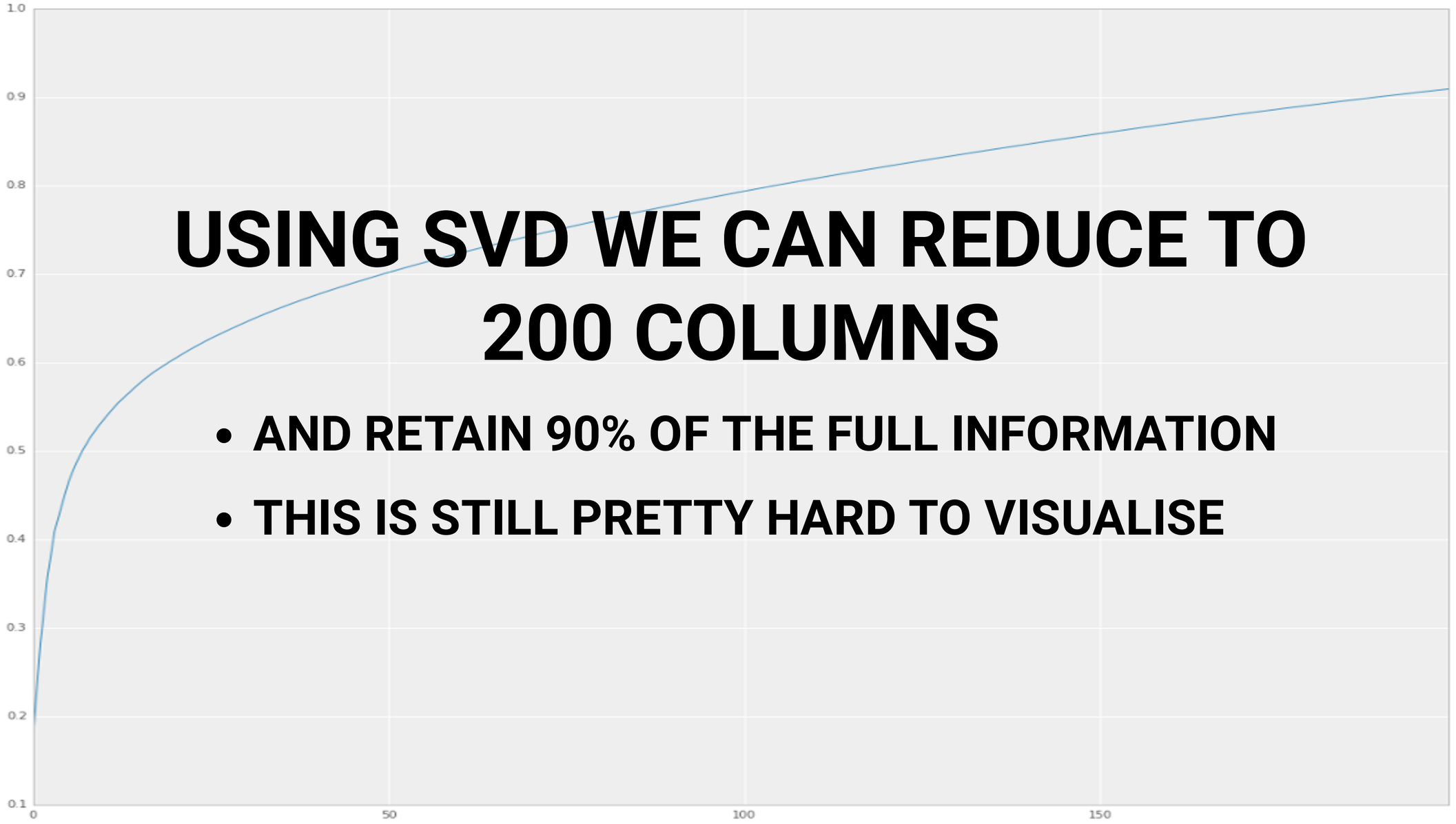
WITH 18,488 ROWS & 767 COLUMNS


LINEAR ALGEBRA
TO THE RESCUE

SINGULAR VALUE DECOMPOSITION

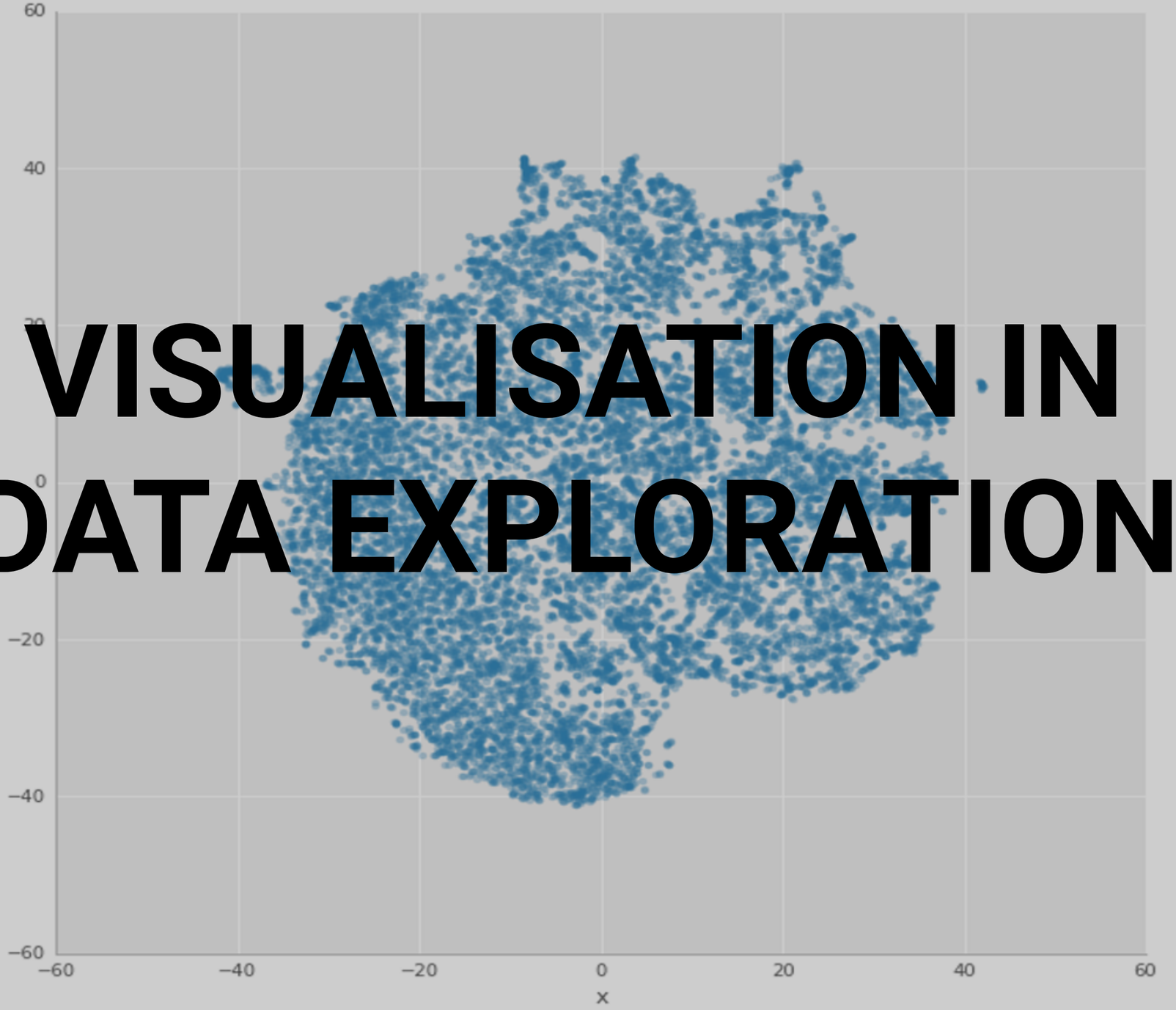
REDUCES THE SIZE OF THE PROBLEM BY DESCRIBING THE DATA IN A NEW SET OF AXES

- FEATURES OFTEN PARTIALLY CORRELATE
- IF YOU KNOW ONE COLUMN, YOU PARTIALLY KNOW THE OTHER
- SO WE DESCRIBE BOTH USING A SINGLE COLUMN (WITH SOME MINOR LOSS OF INFO)



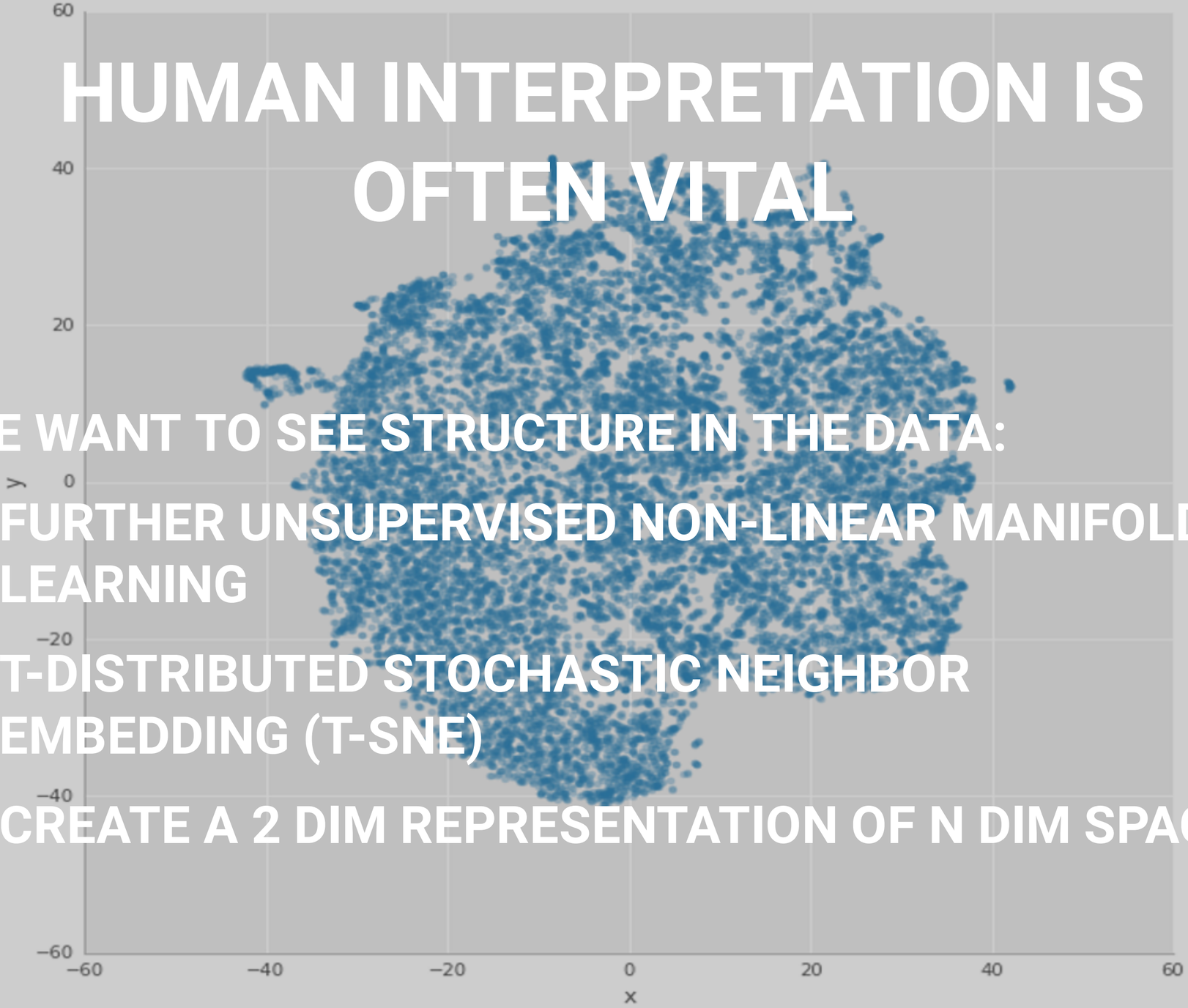
USING SVD WE CAN REDUCE TO 200 COLUMNS

- AND RETAIN 90% OF THE FULL INFORMATION
- THIS IS STILL PRETTY HARD TO VISUALISE



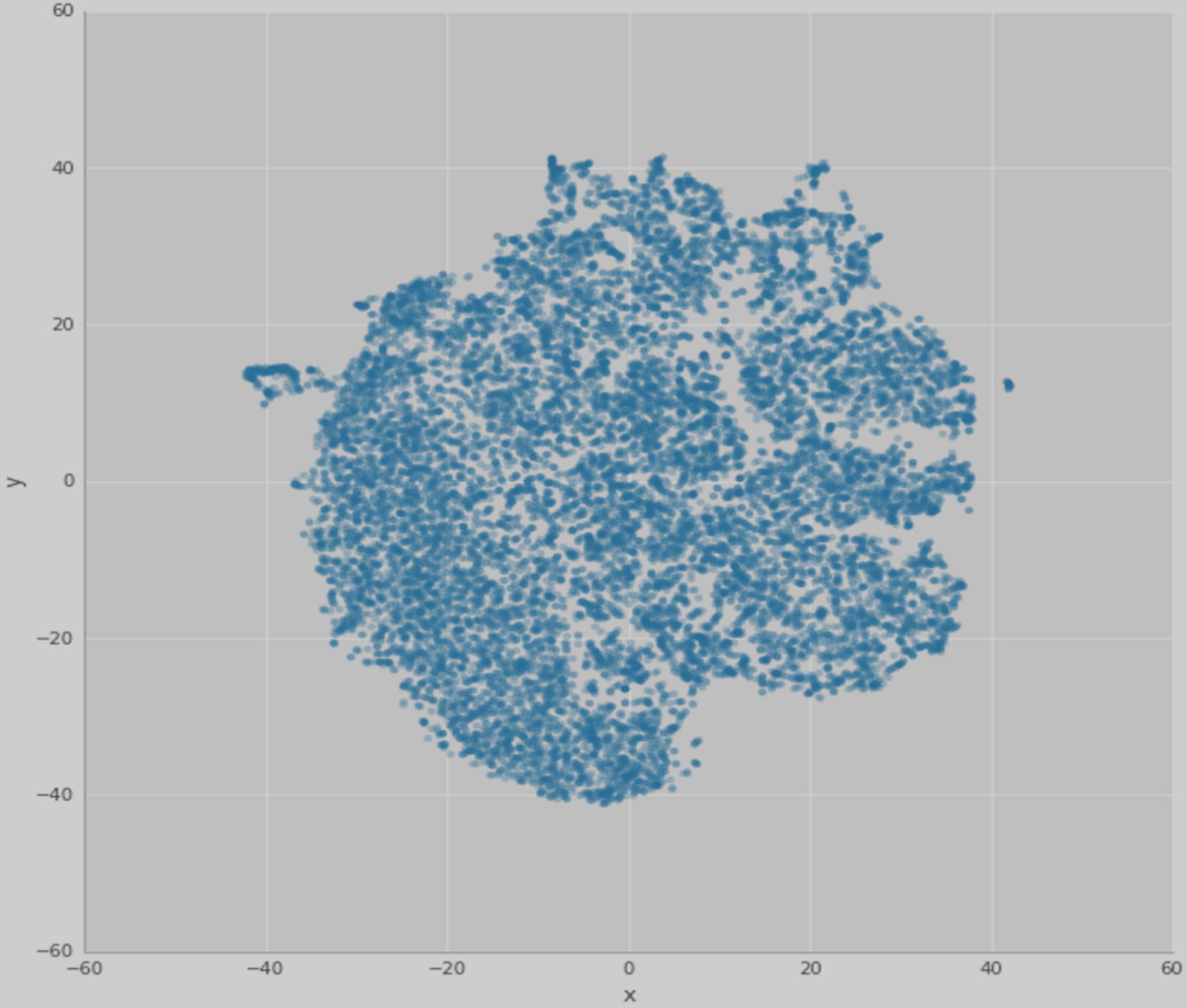
VISUALISATION IN DATA EXPLORATION

HUMAN INTERPRETATION IS OFTEN VITAL



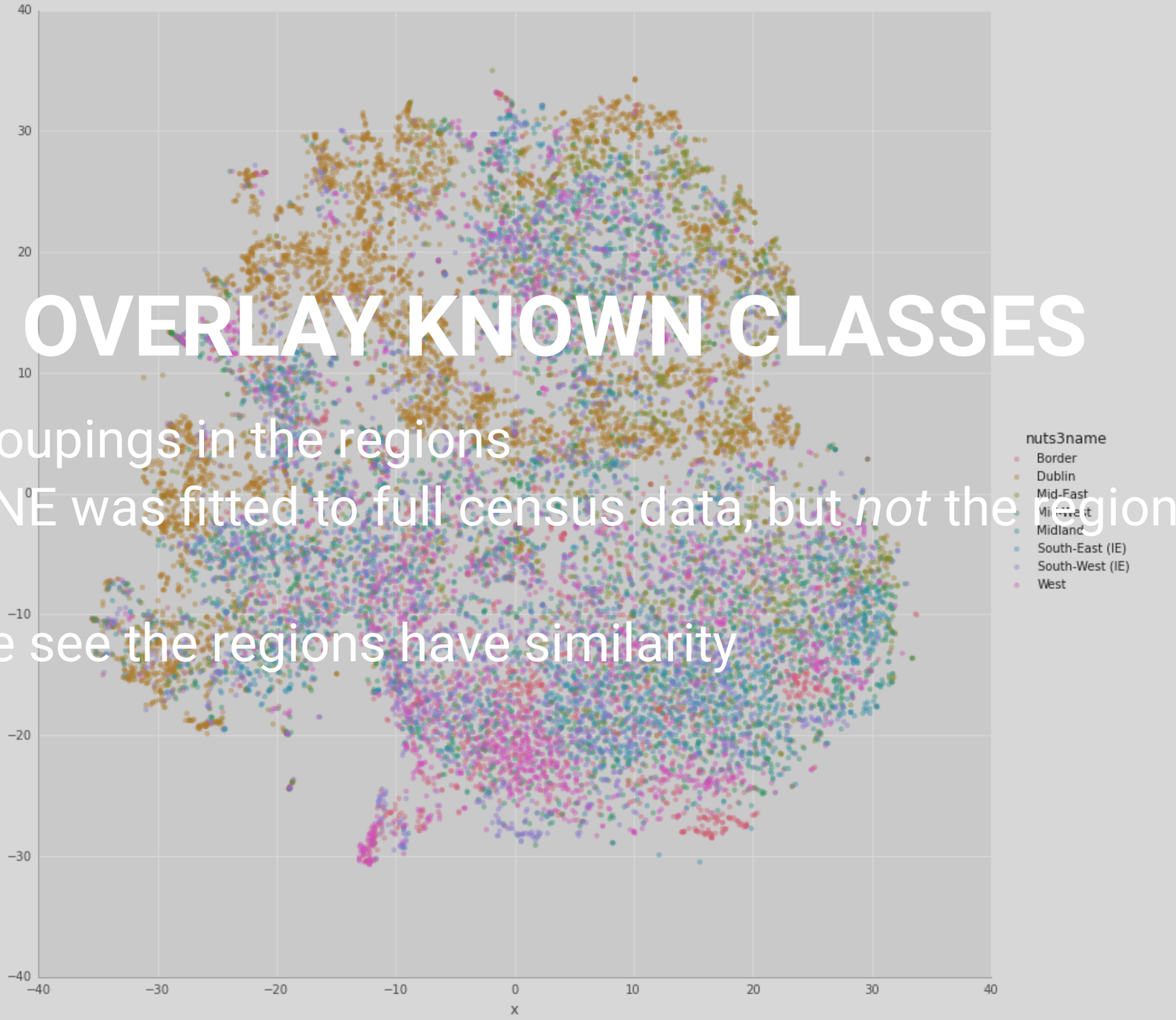
WE WANT TO SEE STRUCTURE IN THE DATA:

- FURTHER UNSUPERVISED NON-LINEAR MANIFOLD LEARNING
- T-DISTRIBUTED STOCHASTIC NEIGHBOR EMBEDDING (T-SNE)
- CREATE A 2 DIM REPRESENTATION OF N DIM SPACE


60
40
20
0
−20
−40
−60
y
−60
−40
−20
0
20
40
60
x

OVERLAY KNOWN CLASSES

- Groupings in the regions
- tSNE was fitted to full census data, but *not* the region id
- We see the regions have similarity

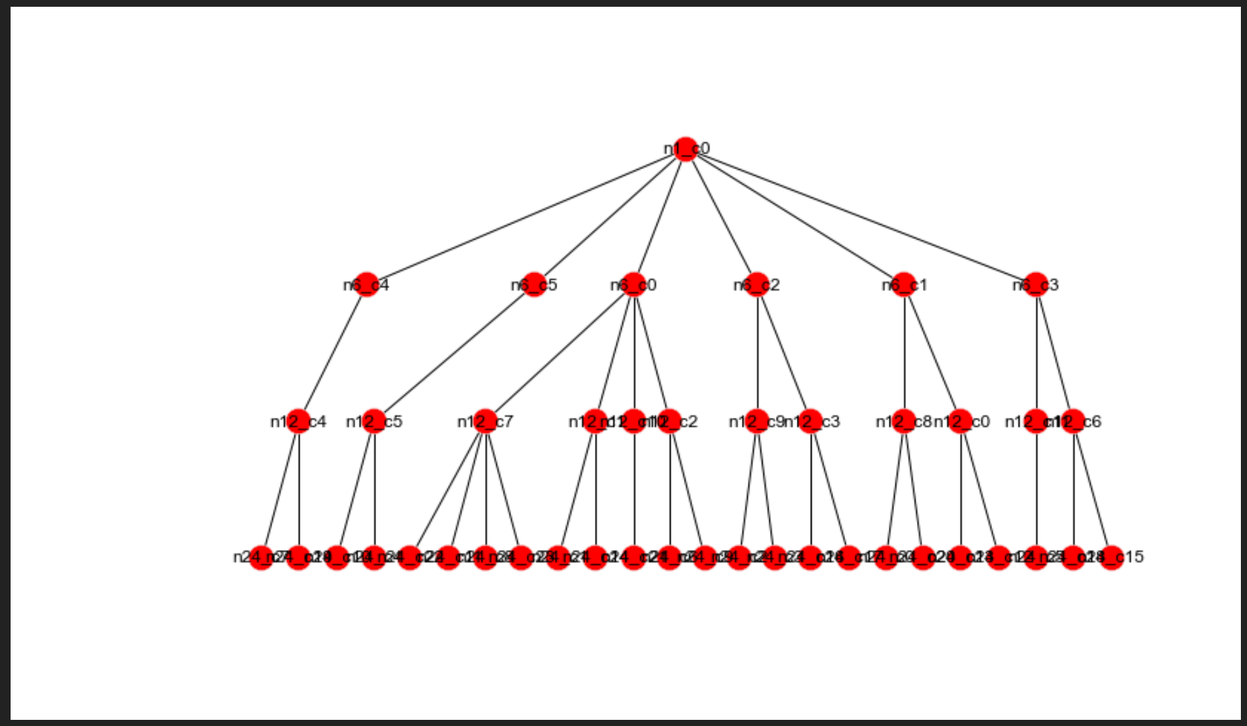


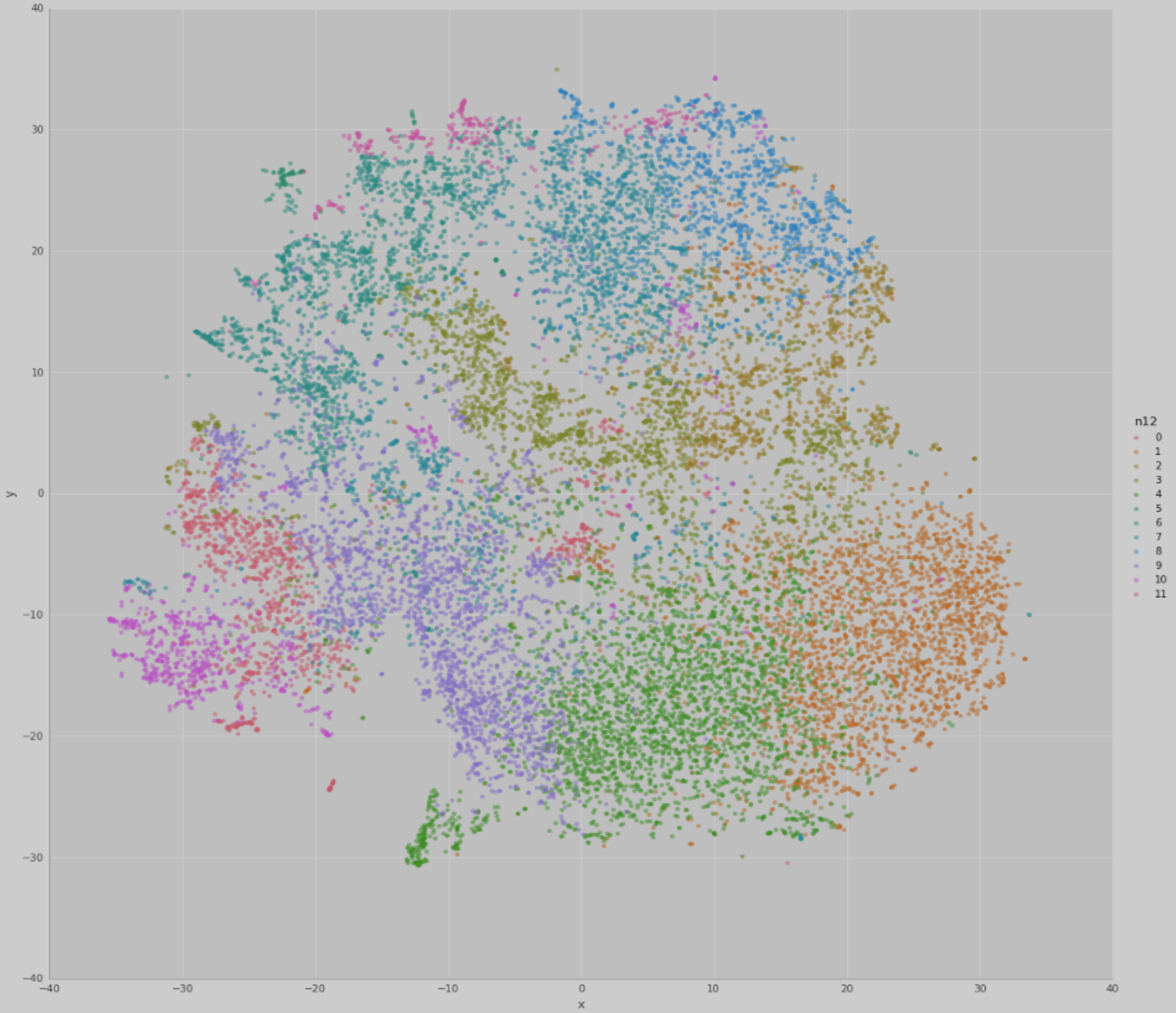

40
30
20
10
0
−10
−20
−30
−40
y
−40
−30
−20
−10
0
10
20
30
40
x
nuts3name
Border
Dublin
Mid-East
Mid-West
Midland
South-East (IE)
South-West (IE)
West

A GOOD HINT THERE'S REAL STRUCTURE IN THE DATA

AGGLOMERATIVE HIERARCHICAL CLUSTERING

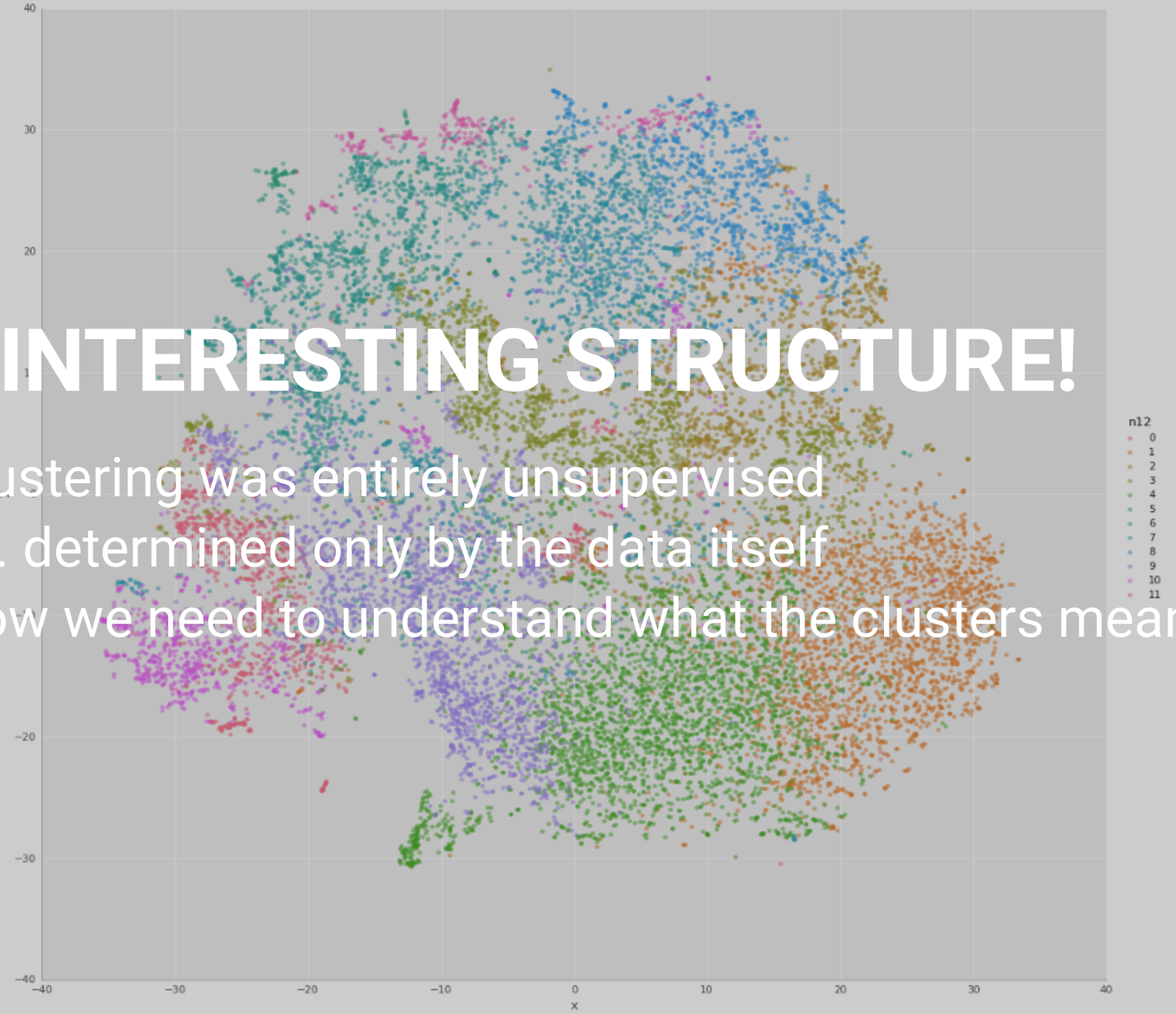
- Group nearby data points into progressively larger clusters
- Get a nested hierarchy of clusters
- Choose your level




40
30
20
10
0
−10
−20
−30
−40
y
−40
−30
−20
−10
0
10
20
30
40
x
n12
0
1
2
3
4
5
6
7
8
9
10
11

INTERESTING STRUCTURE!

- Clustering was entirely unsupervised
- i.e. determined only by the data itself
- Now we need to understand what the clusters mean



FIRST TRICK: MAP THEM

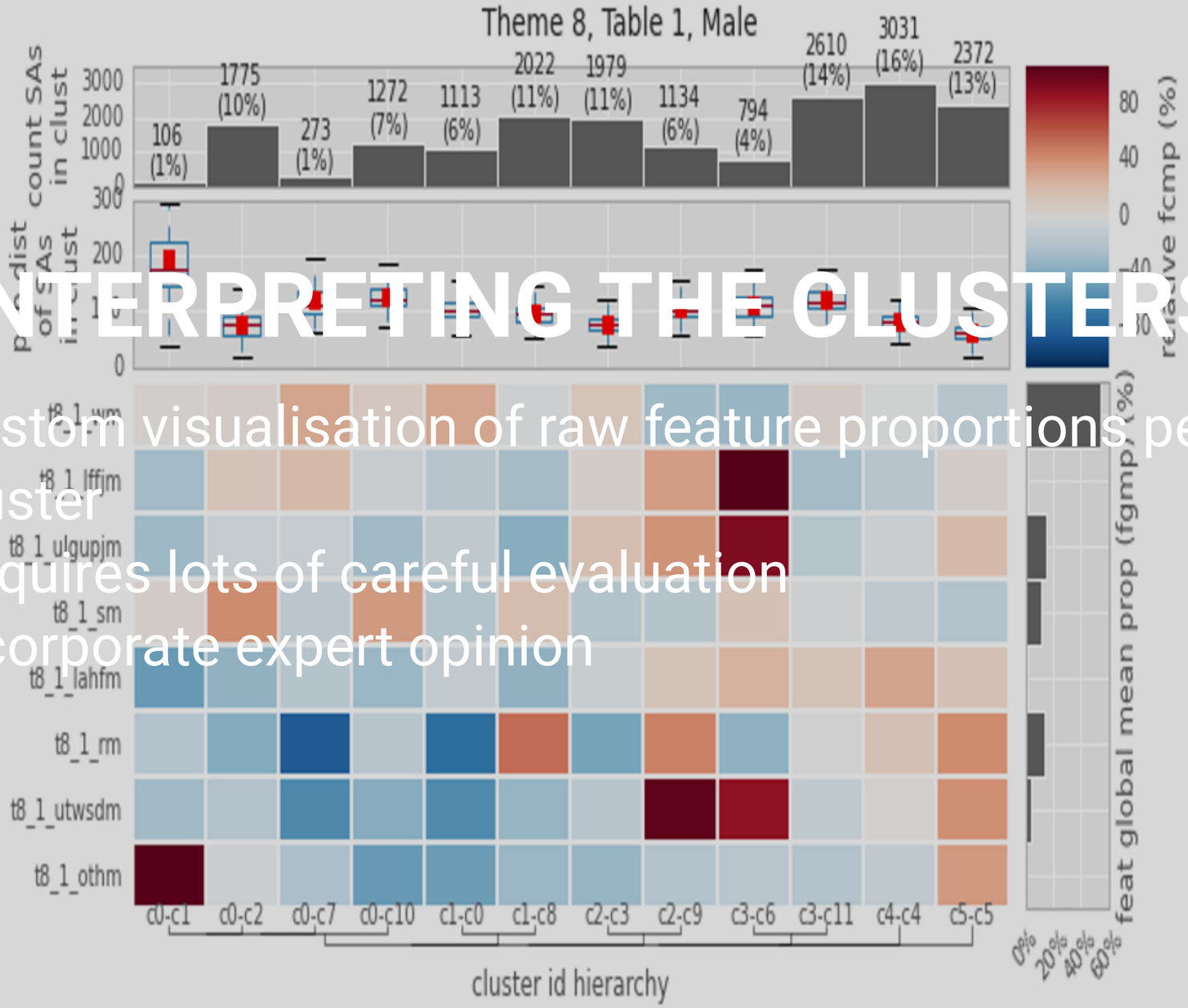
COLOR THE SMALL AREAS BY CLUSTER


IRELAND

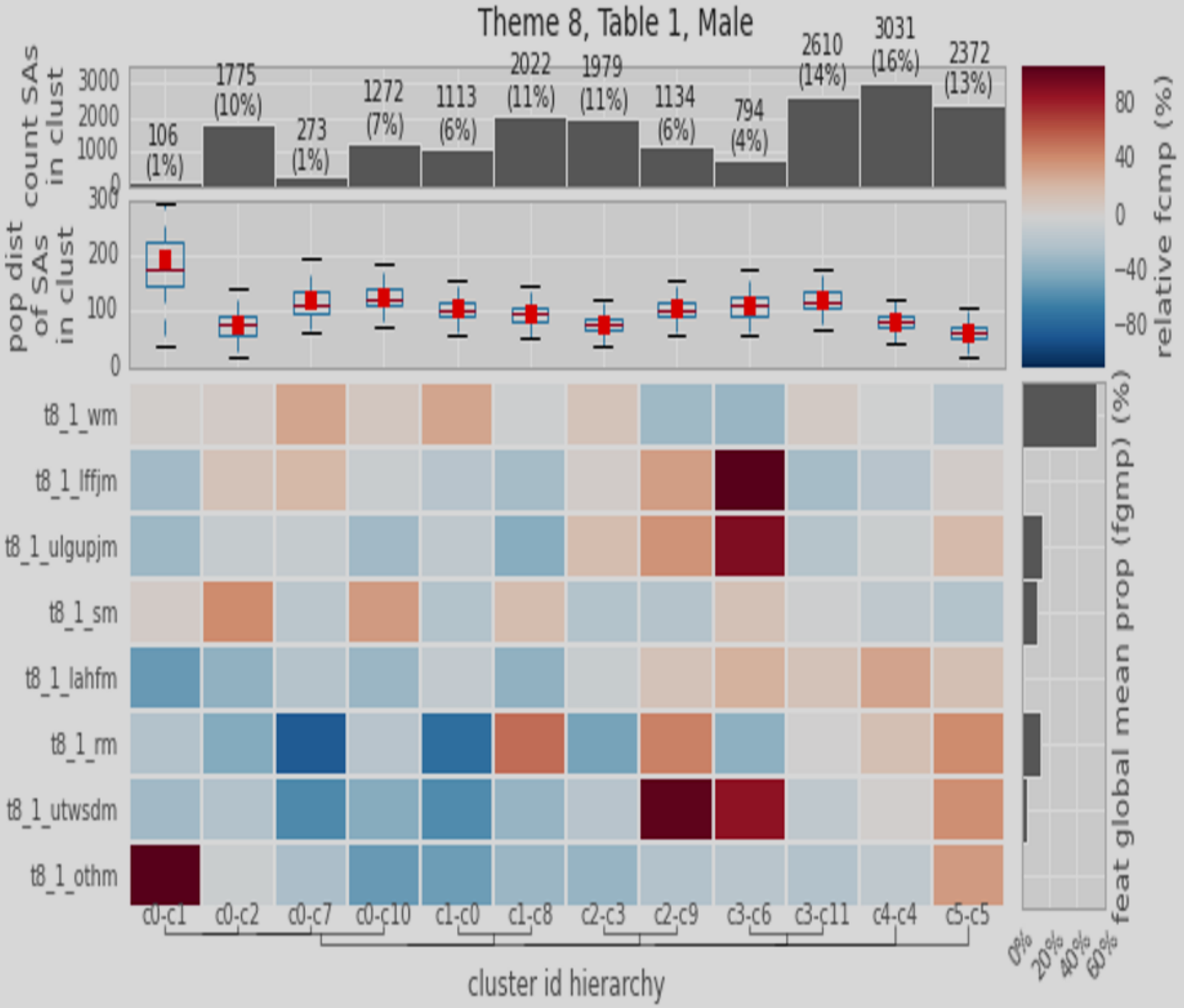

DUBLIN

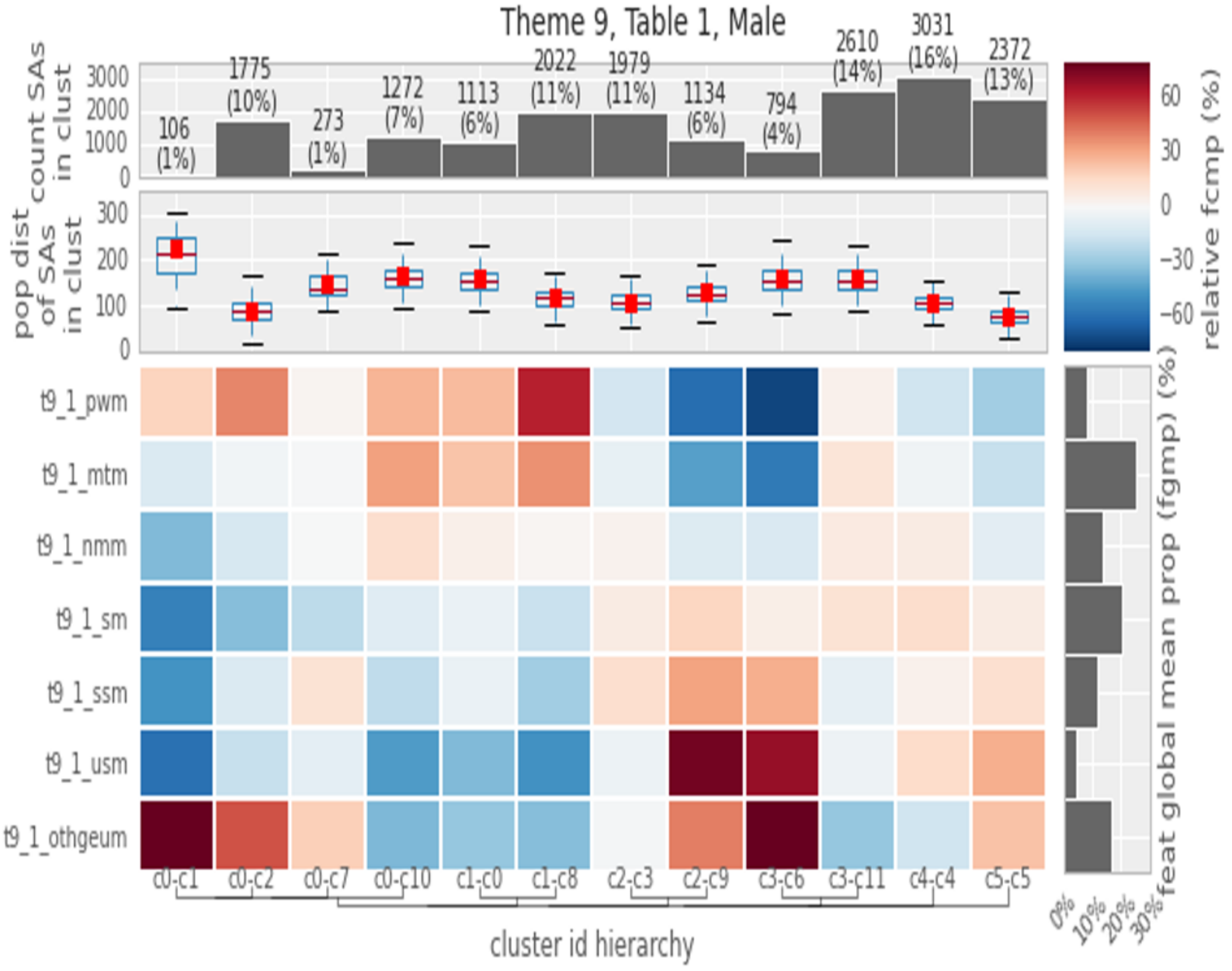

BUT WHAT DO THE CLUSTERS MEAN?
IF YOU LIVE IN IRELAND YOU CAN MAKE A GOOD GUESS...

INTERPRETING THE CLUSTERS



- Custom visualisation of raw feature proportions per cluster
- Requires lots of careful evaluation
- Incorporate expert opinion


Theme 8, Table 1, Male
count SAs in clust
3000
2000
1000
0
106 (1%)
1775 (10%)
273 (1%)
1272 (7%)
1113 (6%)
2022 (11%)
1979 (11%)
1134 (6%)
794 (4%)
2610 (14%)
3031 (16%)
2372 (13%)
pop dist of SAs in clust
300
200
100
0
relative fcmp (%)
80
40
0
-40
-80
t8_1_wm
t8_1_lffjm
t8_1_ulgupjm
t8_1_sm
t8_1_lahfm
t8_1_rm
t8_1_utwsdm
t8_1_othm
c0-c1
c0-c2
c0-c7
c0-c10
c1-c0
c1-c8
c2-c3
c2-c9
c3-c6
c3-c11
c4-c4
c5-c5
cluster id hierarchy
feat global mean prop (fgmp) (%)
0%
20%
40%
60%

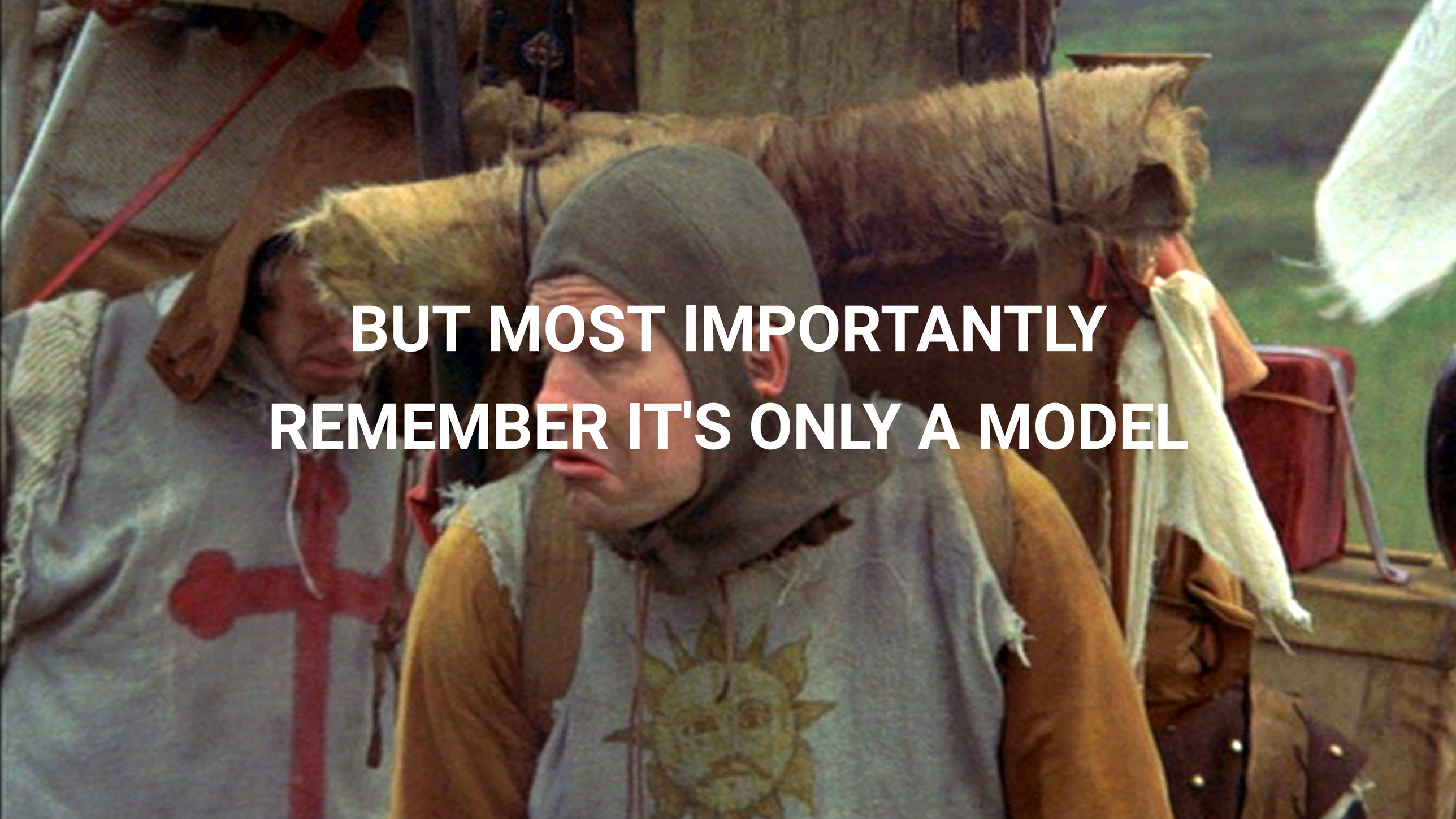

Theme 9, Table 1, Male
count SAs in clust
pop dist of SAs in clust
106 (1%)
1775 (10%)
273 (1%)
1272 (7%)
1113 (6%)
2022 (11%)
1979 (11%)
1134 (6%)
794 (4%)
2610 (14%)
3031 (16%)
2372 (13%)
relative fcmp (%)
t9_1_pwm
t9_1_mtm
t9_1_nmm
t9_1_sm
t9_1_ssm
t9_1_usm
t9_1_othgeum
feat global mean prop (fgmp) (%)
c0-c1
c0-c2
c0-c7
c0-c10
c1-c0
c1-c8
c2-c3
c2-c9
c3-c6
c3-c11
c4-c4
c5-c5
cluster id hierarchy

RESULT

VALUABLE GEO-SOCIOECONOMIC INDICATORS FOR USE IN OTHER PROJECTS

IN SUMMARY

DATA SCIENCE IS APPLICABLE THROUGHOUT THE INSURANCE BUSINESS


BUT MOST IMPORTANTLY
REMEMBER IT'S ONLY A MODEL

QUESTIONS?

Nah, we're running outta time. Seriously.

THANKS FOR LISTENING