HOW IS DATA SCIENCE GOING TO IMPROVE INSURANCE?

MICHAEL CRAWFORD &
JONATHAN SEDAR

APPLIED AI LTD

@APPLIED_AI

APPLIED.AI


HOW ARE WE *ALREADY* USING
DATA SCIENCE TO IMPROVE
INSURANCE?

THREE STORIES FROM OUR EXPERIENCES OVER THE PAST TWO YEARS.

-

1. CURATING EXTERNAL DATA TO BETTER UNDERSTAND CUSTOMERS
2. BRINGING STATISTICAL MODELING INTO THE BOARDROOM
3. APPLYING MORE EXOTIC MACHINE LEARNING TO THE BUSINESS

APPLIED AI IS A DATA SCIENCE CONSULTANCY

WE HELP LEADING INSURERS CREATE A COMPETITIVE ADVANTAGE THROUGH **APPLIED ARTIFICIAL INTELLIGENCE**

WHO ARE WE?

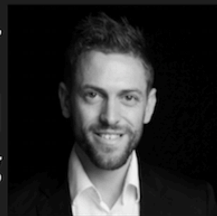
A SMALL BUT EXPERIENCED TEAM WITH EXPERTISE IN STATISTICS, MACHINE LEARNING, ACTUARIAL SCIENCE, SOFTWARE DEVELOPMENT, FINANCIAL SYSTEMS AND CONSULTANCY.

Michael Crawford

Actuarial Science | Financial Systems | Software Dev

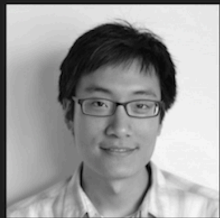
Jonathan Sedar

Machine Learning | Physical Sciences | Consulting



A Team of Expert Practitioners

Quantitative Finance | Statistics | Software Development | Insurance




FOUNDING MYTH
WE MET IN THE PUB FOLLOWING AN R MEETUP
... NOT A MILLION MILES AWAY FROM THIS
WE BONDED OVER BEER & GENERAL GEEKINESS
AND THOUGHT THAT MACHINE LEARNING & INSURANCE WERE A NATURAL FIT
LETS ASK INSURANCE COMPANIES


BINGO!

SO WHAT HAPPENED NEXT?

PART ONE

CURATING EXTERNAL DATA TO BETTER UNDERSTAND CUSTOMERS

-

OR WHAT HAPPENED WHEN WE CUT EXPERIAN OUT OF THE LOOP

-

CLUSTERING

INTROSPECTION

VISUALISATION

BACKGROUND

WE WORK MAINLY WITH INSURANCE COMPANIES

THEY DON'T HAVE A REPUTATION FOR BEING EXCITING

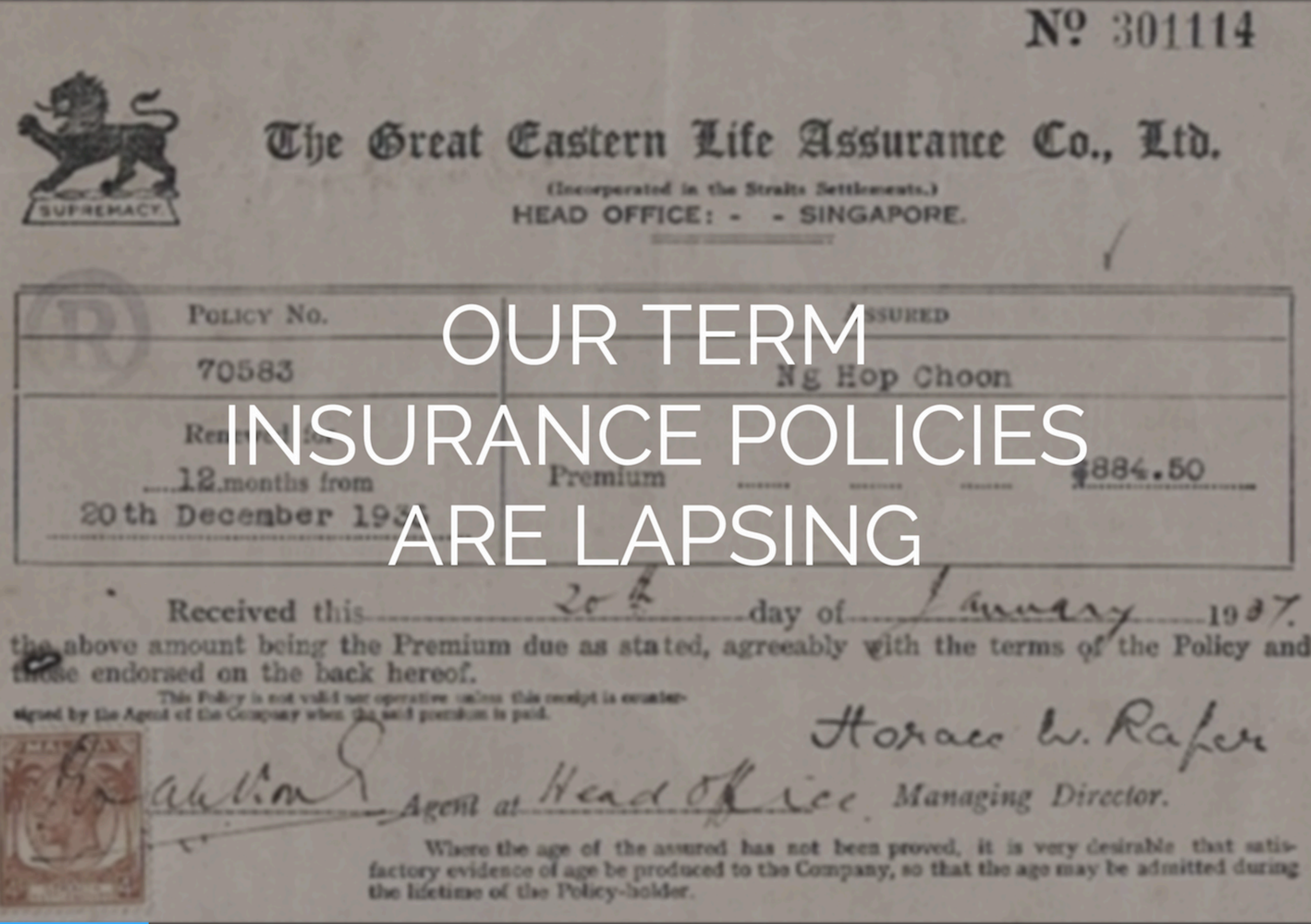
BUT FROM A DATA SCIENCE POINT OF VIEW...


IT'S QUITE
INTERESTING!
INDIANA JONES
and the
TEMPLE OF DOOM


The
mplete
RIPPING YARNS
2 DISC SET
By Michael Palin
and Terry Jones
HERE'S A STORY

THE PROBLEM

OUR TERM INSURANCE POLICIES ARE LAPSING



Nº 301114

SUPREMACY

The Great Eastern Life Assurance Co., Ltd.

(Incorporated in the Straits Settlements.)

HEAD OFFICE: - - SINGAPORE.

POLICY No.	ASSURED
70583	Ng Hop Choon
Renewed for 12 months from 20th December 193	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 countersigned by the Agent of the Company when the said premium is paid.

Horace W. Raper

Agent at Head Office Managing Director.

Where the age of the assured has not been proved, it is very desirable that satisfactory evidence of age be produced to the Company, so that the age may be admitted during the lifetime of the Policy-holder.

TERM INSURANCE

Extra Interesting!

WHEN AN INSURANCE COMPANY SETS A POLICY UP:

- IT PAYS A COMMISSION TO THE BROKER WHO SOLD THE POLICY
- IT MAY HAVE TO SEND YOU FOR A MEDICAL - AND PAY
- IT INCURS ADMINISTRATION & REGULATORY EXPENSES

IT'S 2 YEARS BEFORE THE INSURER IS IN THE BLACK

SO THEY REALLY WANT YOU TO STICK AROUND

THE THING IS

Still Really Interesting!

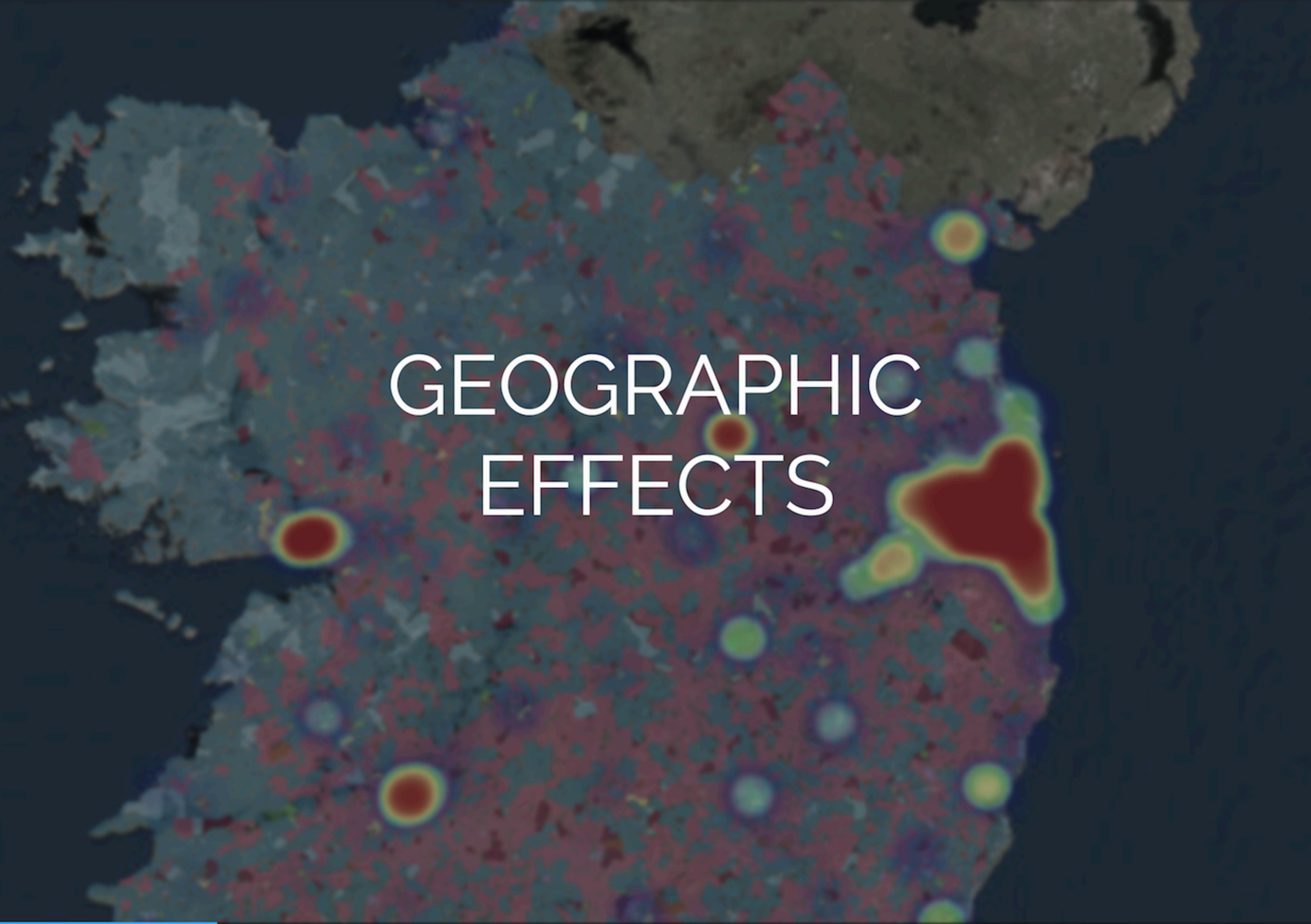
PEOPLE DON'T WANT TO STICK AROUND!

- IN THE RECESSION THEY WERE DROPPING LIKE FLIES
- WE WERE ASKED IF WE COULD FIGURE OUT WHY
- AND TRY FIND WAYS TO REDUCE IT
  - We used survival analysis (of which more later)
  - After a few weeks we had a good model

ALONG THE WAY WE NOTICED SOMETHING


THE ATTRITION RATE
WAS SKY-HIGH IN
NEW ESTATES

GEOGRAPHIC EFFECTS




AND
SOCIOECONOMIC
EFFECTS

HOW CAN WE USE THESE EFFECTS?

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

WE CALLED EXPERIAN...

IT WAS A SHORT CONVERSATION

I'D LIKE SOCIOECONOMIC INFORMATION FOR 250K ADDRESSES

...

HOW MUCH!?


WE CAN DO IT
CHEAPER & BETTER
OURSELVES

GEOCODING

FIRST LET'S GEOCODE 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

IN IRELAND A CENSUS IS DONE EVERY FIVE YEARS

- THE AMOUNT OF INFORMATION IN IT IS ASTOUNDING
- APPROXIMATELY 700 FEATURES
- COVERING 15 THEMES

LET'S LOOK AT THE DATA 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
8	Principal Status		


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 THE POWER WE WERE LOOKING FOR!

WE REALISED THAT WE COULD DO WHAT EXPERIAN DOES

BUT ALSO:

- 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 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

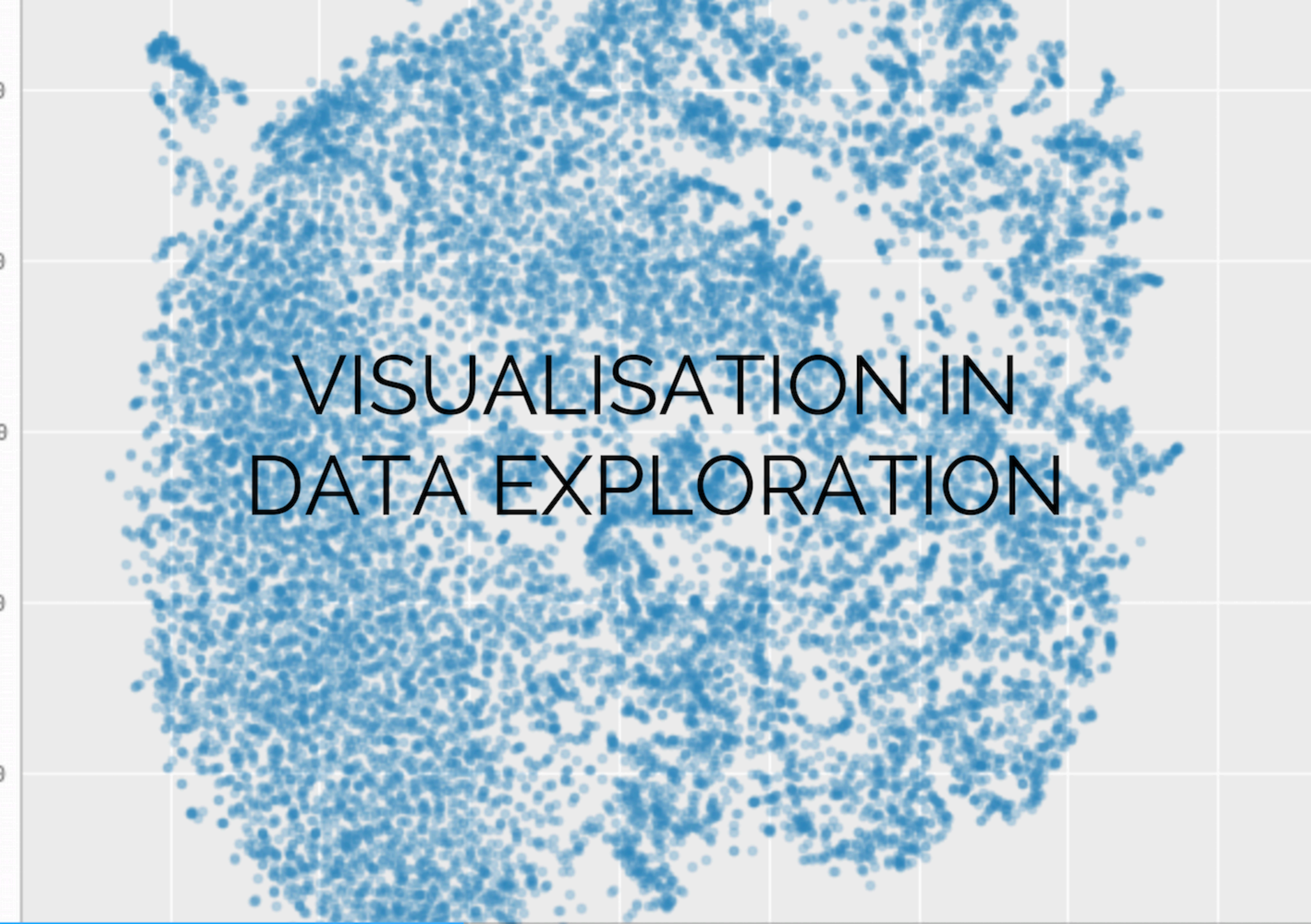
REDUCES THE SIZE OF THE PROBLEM BY DESCRIBING THE DATA IN A NEW SET OF AXES

- 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)

$$M = \begin{bmatrix} M_{1,1} & M_{1,2} \\ M & M \end{bmatrix}$$

SVD LETS US SHRINK DOWN TO 100 COLUMNS

- AND RETAIN 80% OF THE FULL INFORMATION
- THIS IS STILL PRETTY HARD TO VISUALISE

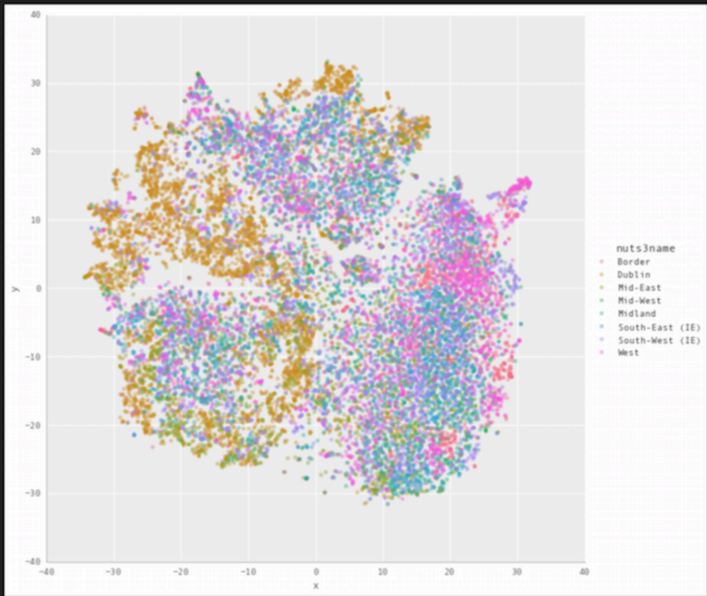

VISUALISATION IN
DATA EXPLORATION

HUMAN INTERPRETATION IS OFTEN VITAL

WE WANT TO SEE STRUCTURE IN THE DATA, BUT VIEWING 100 DIMS IS *STILL* TRICKY

- USE UNSUPERVISED LEARNING
- T-DISTRIBUTED STOCHASTIC NEIGHBOR EMBEDDING (T-SNE)
- CREATE A 2D REPRESENTATION OF ND SPACE

OVERLAY KNOWN CLASSES

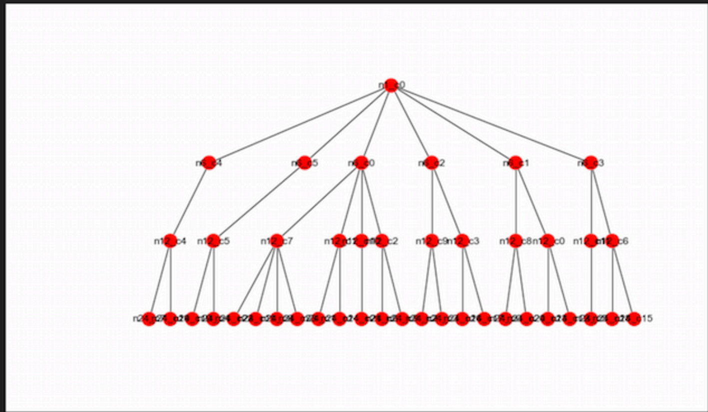


- Instantly see grouping in the regions
- The tSNE was fitted using full census data, but *not* the region id
- Yet we see the regions have similarity

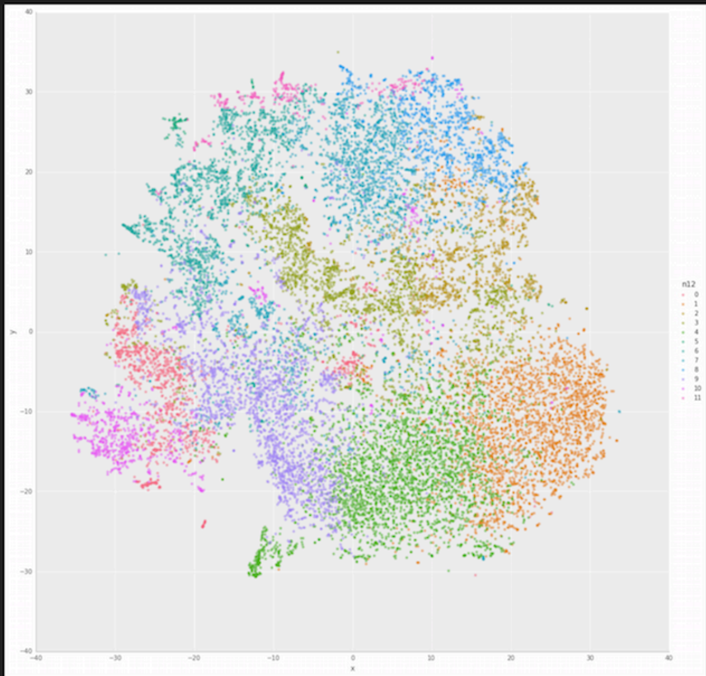
A GOOD HINT THERE'S REAL STRUCTURE IN THE DATA

AGGLOMERATIVE HIERARCHICAL CLUSTERING

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



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

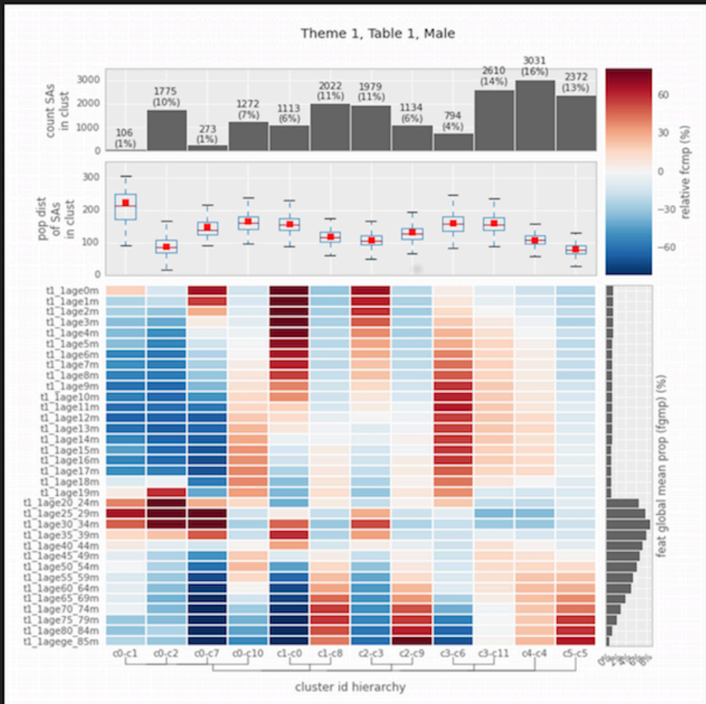

DUBLIN


WOW


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


SECOND TRICK:
PONDERING

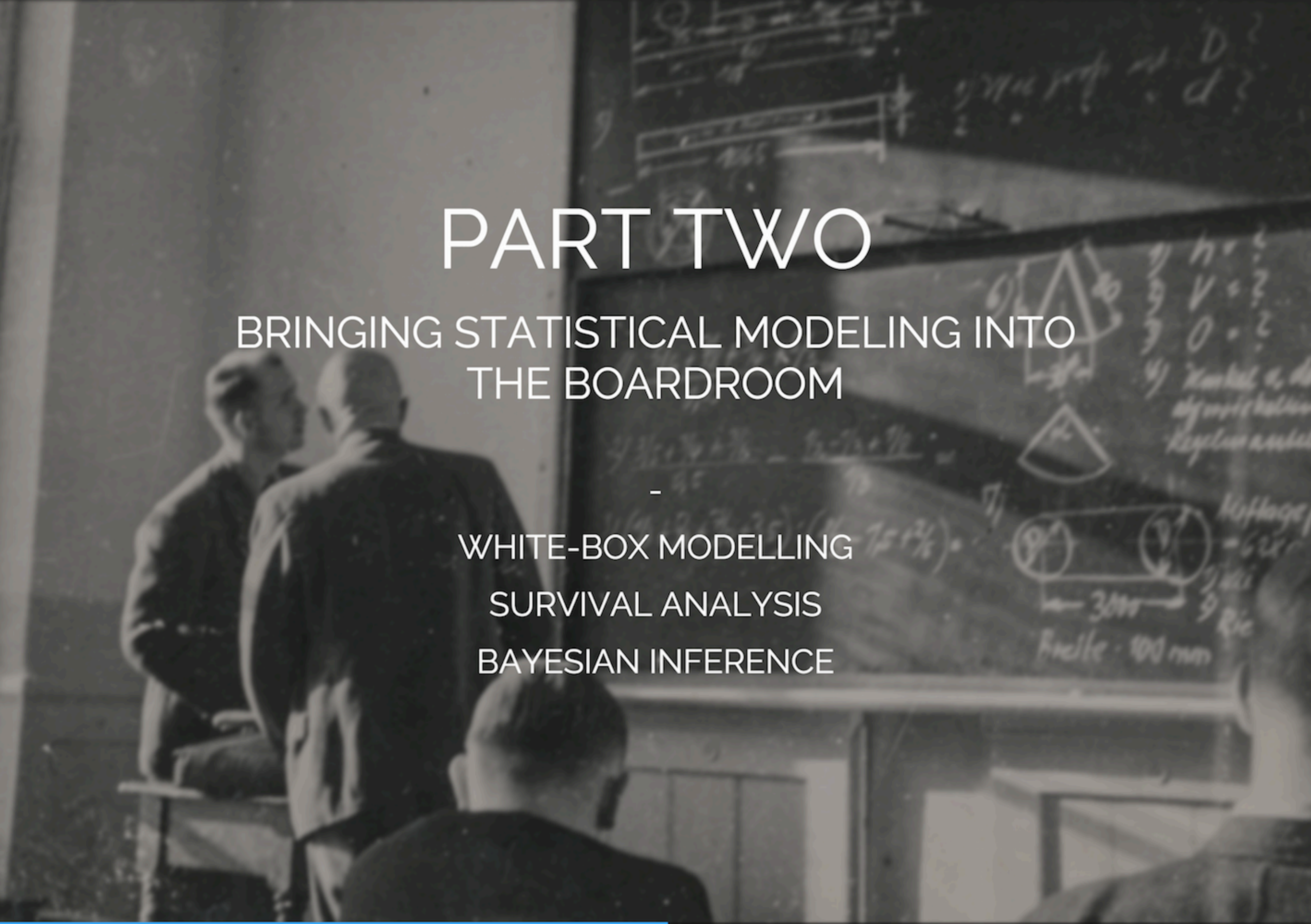
INTERPRETING THE CLUSTERS



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

RESULT

VALUABLE GEO-SOCIOECONOMIC INDICATORS FOR USE IN OTHER PROJECTS



PART TWO

BRINGING STATISTICAL MODELING INTO THE BOARDROOM

-

WHITE-BOX MODELLING

SURVIVAL ANALYSIS

BAYESIAN INFERENCE

WHITE-BOX MODELLING

THE ABILITY TO INTROSPECT THE MODEL FUNCTION

"WHAT'S OUR CHURN RATE?"

SEVERAL DIFFERENT WAYS TO MEASURE CHURN

PARTICULARLY IMPORTANT IN LIFE INSURANCE

"WHAT'S OUR PROJECTED CREDIT RISK?"

A FAR MORE JUICY QUESTION

... AND IT NEEDS A *DEFENSIBLE* ANSWER

EUREKA LIFE INSURANCE COMPANY of Baltimore, Md.

INFANTILE BONUS ADDITIONAL POLICY

Table A

mount payable if death occur after policy has been in force the following periods. Weekly Premium Ten Cent

Under six Months	Six Months	One Year	Two Years	Three Years	Four Years	Five Years	Six Years	BONUS ADDITIONS		
								Paid up at end of Years	Bonus Additions	Maximum Amount
$16	$ 30	$ 50	$ 70	$ 90	$110	$160	[illegible]	38	$28	$268
30	50	70	90	110	160	240		40	30	270
40	70	90	110	160	240			42	32	272
50	90	110	160	240				[illegible]	34	274
60	110	160	240					46	36	276
70	130	240						48	38	278
80	240							50	40	280
90	240							52	42	282
100	240							54	44	284

After policy has been in force ten years, an amount equal to ten times the weekly premium will be added to unt payable at death, and on each anniversary of policy thereafter an equal amount will be added until the maxim unt has been reached; after which, the policy becomes fully paid up and no further premiums will be required.

CONVENTIONAL METHODS

MS EXCEL

RUNNING AVERAGES & POINT-ESTIMATES

VERY LIMITED INTROSPECTION, EXPLANATION AND FORECASTING

TIME-TO-EVENT MODELING

"HOW MUCH TIME ELAPSED BETWEEN EVENTS?"

- HALF-LIFE OF A RADIOACTIVE COMPOUND
- DURATION OF A PITSTOP IN F1
- DURATION BETWEEN A CUSTOMER JOINING AND QUITTING A CONTRACT

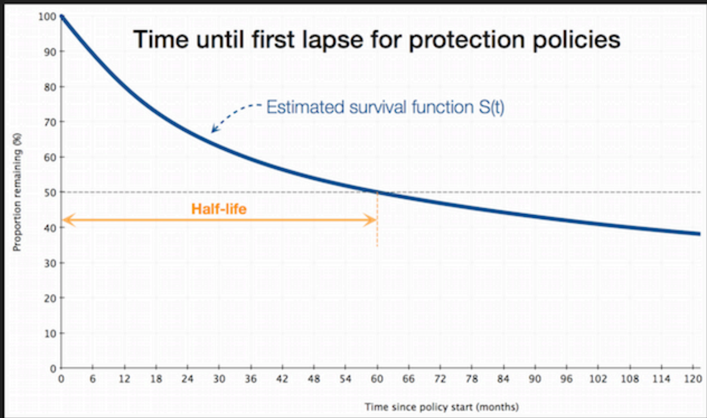
SURVIVAL ANALYSIS

A WELL-STUDIED PART OF MEDICAL STATISTICS

BUT NOT USUALLY DISCUSSED IN MACHINE LEARNING

SEEMS TO HAVE BEEN OVERLOOKED

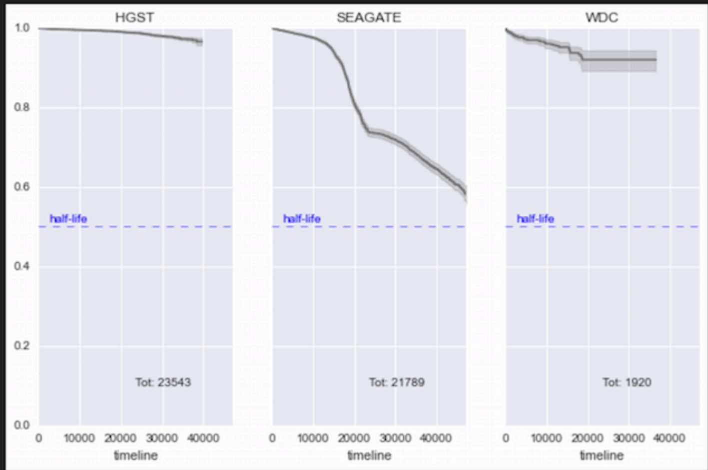
THE BASIC IDEA:



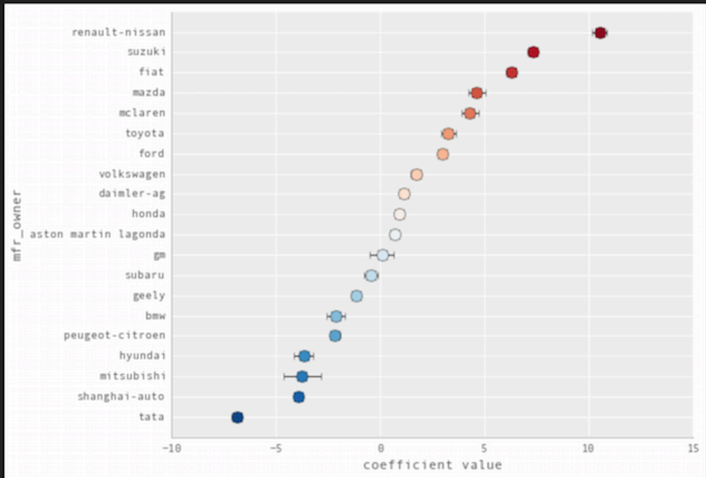
- Take a cohort at some start date
- Count the prop. remaining at subsequent dates
- Simply draw or fit the line

SIMPLEST VERSION: KAPLAN-MEIER

- Simple count-based description of events
- Draw the line per group
- Prediction not possible



A MODEL-BASED VERSION: COX PROPORTIONAL HAZARDS



- Fit semi-parametric linear model
- Learn effect of feature values
- Make predictions on new data

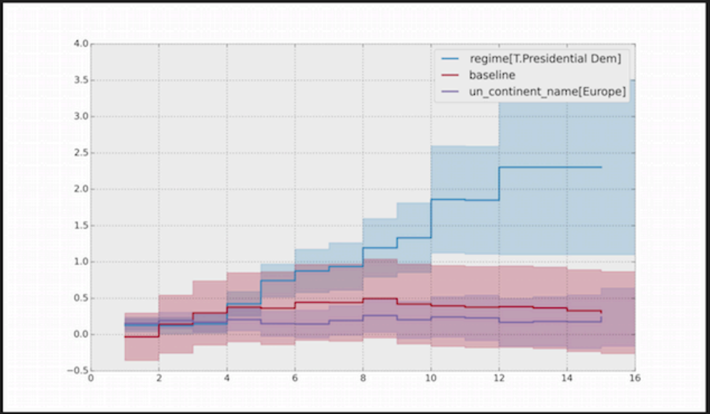
THIS IS ALREADY SUFFICIENT TO ANSWER

WHAT IS THE EXPECTED RANGE OF SURVIVAL OVER TIME?

CAN I PREDICT THE SURVIVAL RATE?

WHAT ARE THE DRIVERS OF THE SURVIVAL RATE?

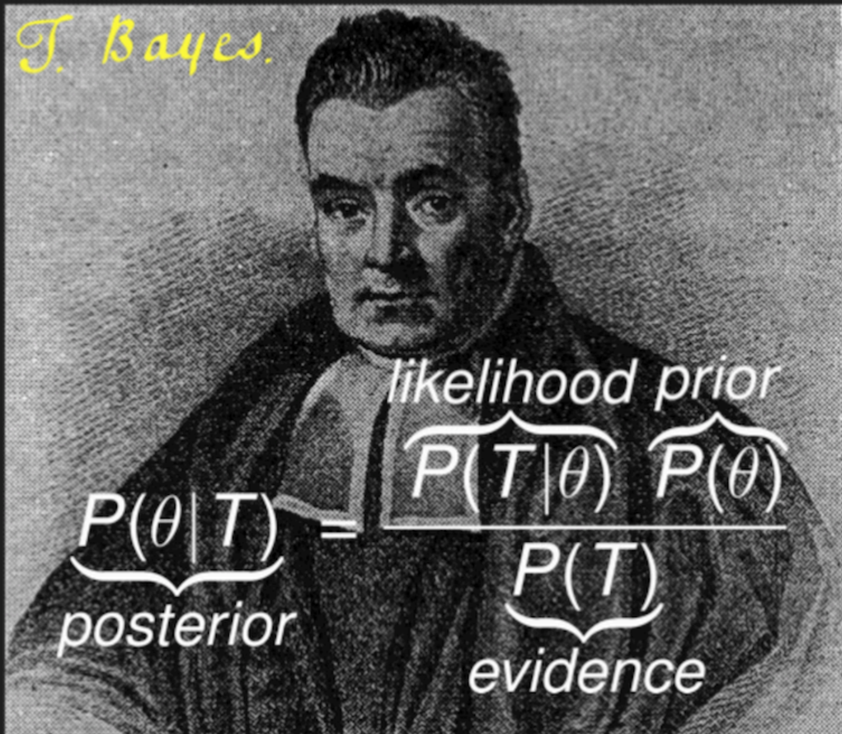
INCREASE THE COMPLEXITY TO BETTER FIT THE REAL-WORLD



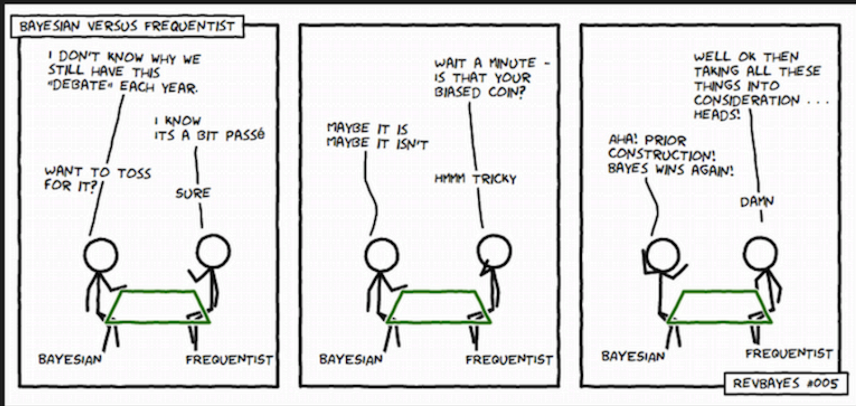
- Survival Regression
- Allow hazards to vary with time: Aalen Additive
- Bayesian approach: Gaussian Processes


THE BEST MODELS
PRESERVE
UNCERTAINTY

BAYESIAN INFERENCE

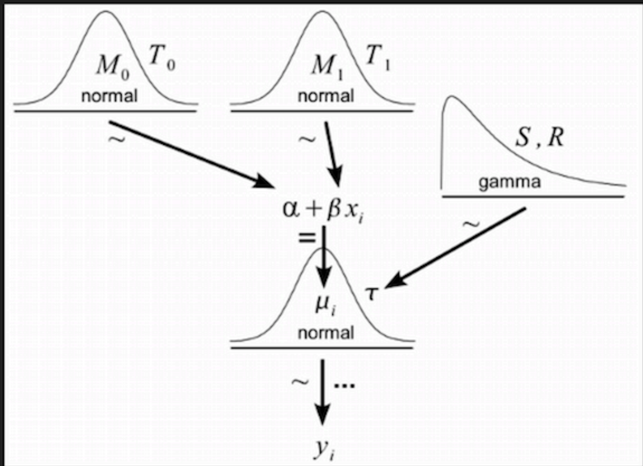

T. Bayes.
$$\underbrace{P(\theta|T)}_{\text{posterior}} = \frac{\overbrace{P(T|\theta)}^{\text{likelihood}}\ \overbrace{P(\theta)}^{\text{prior}}}{\underbrace{P(T)}_{\text{evidence}}}$$

A NATURAL ALTERNATIVE TO TRADITIONAL (FREQUENTIST) STATISTICS



MADE POSSIBLE THROUGH COMPUTATIONAL POWER

FAR TOO MUCH DETAIL FOR HERE



FAR TOO MUCH DETAIL FOR HERE

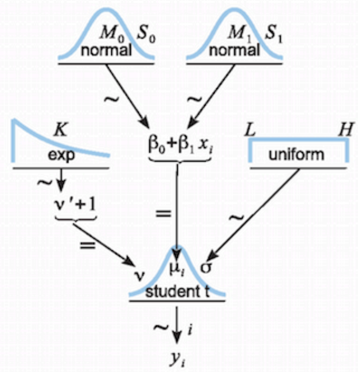
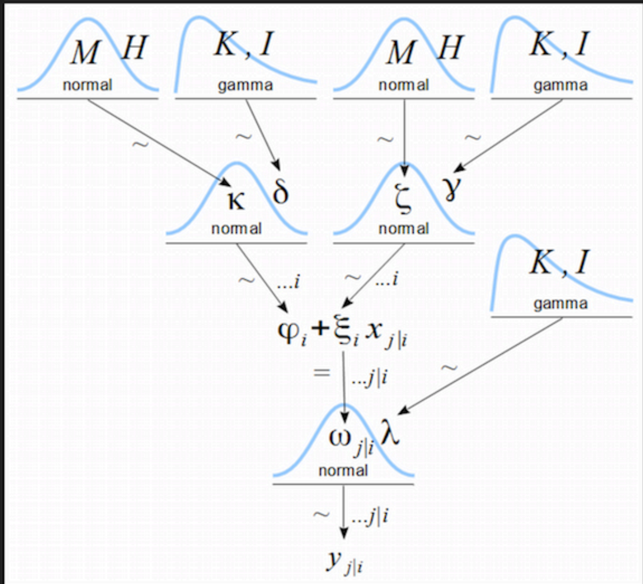
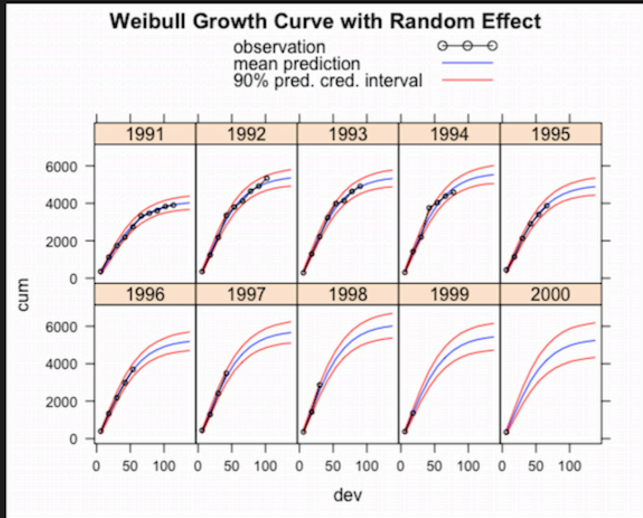


Figure 17.2: A model of dependencies for robust linear regression. The datum, y_i at the bottom of the diagram, is distributed around the central tendency μ_i, which is a linear function of x_i. Compare with Figure 16.11 on p. 437. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

FAR TOO MUCH DETAIL FOR HERE

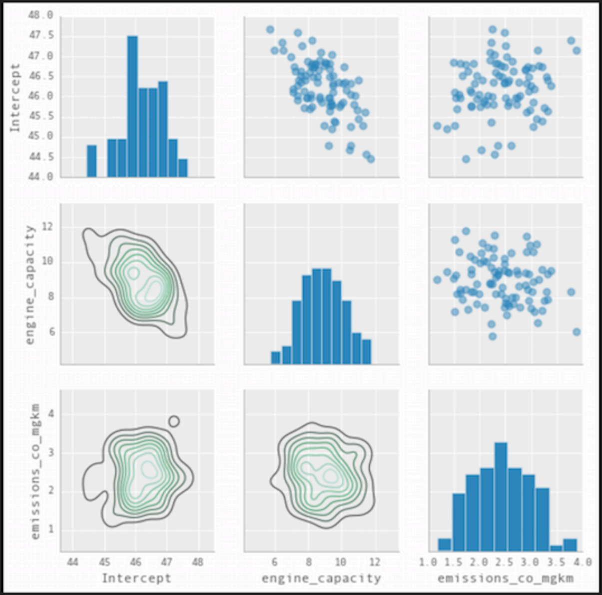


FAR TOO MUCH DETAIL FOR HERE

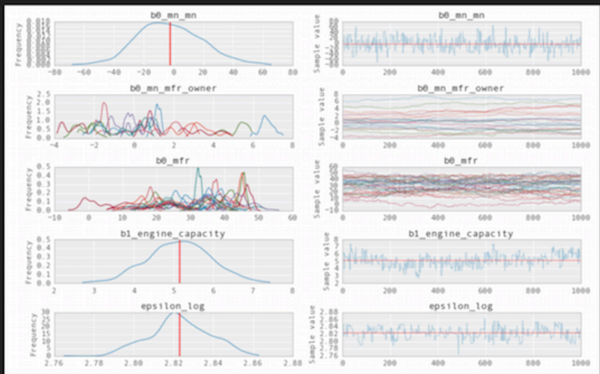


BIGGEST WIN: MODEL INTROSPECTION

- Design structure of model to reflect real processes & relationships
- Inform the model using prior knowledge
- Use qualified results and predictions
- Maintain uncertainty



STATE OF THE ART ACADEMIC TOOLS & TECHNIQUES OPEN TO ALL



- Basic non-sampling frameworks: scikit-learn, arm, nlme etc
- Advanced model frameworks and samplers: STAN, PyMC3, emcee
- Probabilistic graphical models: gRain, pgmpy


THIS TRULY DOES OFFER
EXTRAORDINARY
POWER

ALLOWING US TO CREATE MODELS THAT:

ACCURATELY REFLECT REAL-WORLD PROCESSES

PRESERVE UNCERTAINTY IN PREDICTION

WORK WITH SMALL AND OBSCURE DATASETS

AUTOMATICALLY IMPROVE WITH MORE DATA

ALLOWING US TO ANSWER QUESTIONS LIKE:

KNOWING THE YEARLY CUMULATIVE POLICY CLAIMS IN MY INSURANCE BOOK OVER THE PAST 10 YEARS, HOW MUCH SHOULD I SET ASIDE FOR CLAIMS RESERVES FOR THIS YEAR'S NEW COHORT, AND TO WHAT RANGE EITHER SIDE?

ALLOWING US TO ANSWER QUESTIONS LIKE:

BASED ON PRIOR APPLICATIONS FOR LIFE INSURANCE POLICIES, CAN I DETERMINE WHICH PROSPECTIVE CUSTOMERS TO SEND FOR MEDICAL EXAMINATIONS (EXPENSIVE, NOT OFTEN NECESSARY), DEFINE BETTER QUESTIONS TO ASK IN THE PRE-SCREENING, AND CHANGE MY BELIEFS BASED ON NEW EVIDENCE?


BUT REMEMBER:
IT'S ONLY A MODEL

PART THREE

APPLYING MORE EXOTIC MACHINE LEARNING ACROSS THE BUSINESS

-

NATURAL LANGUAGE PROCESSING

UNSUPERVISED FEATURE REDUCTION

PATTERN MATCHING AND PREDICTION

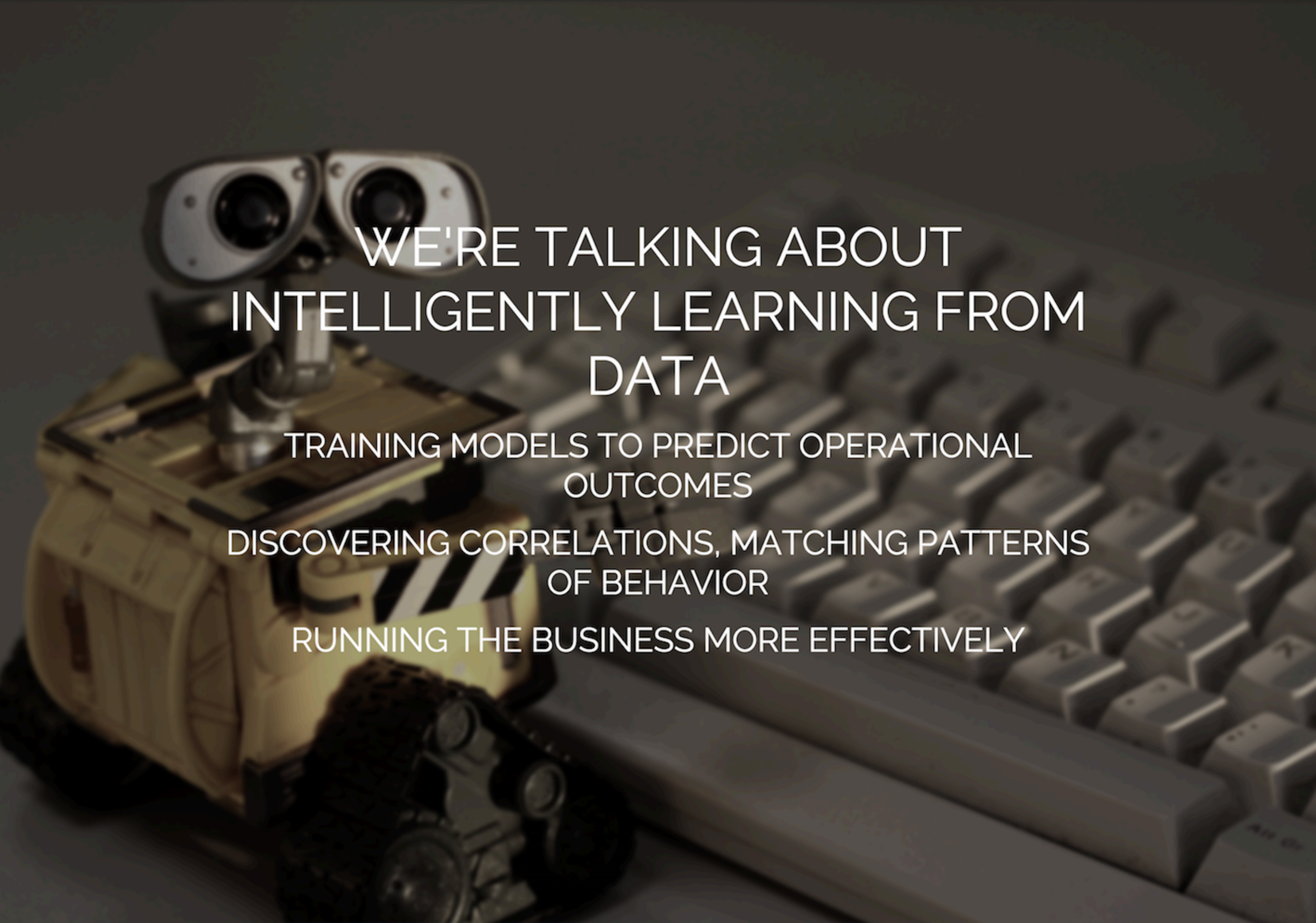

"MACHINE LEARNING" SOUNDS
OMINOUS
VAGUE
MONEYPIT

"OUR COMPANY HAS BEEN USING DATA FOR YEARS"

~~SQL & SPREADSHEETS~~

~~BI DASHBOARDS~~

~~EXPERT / RULE-BASED SYSTEMS~~

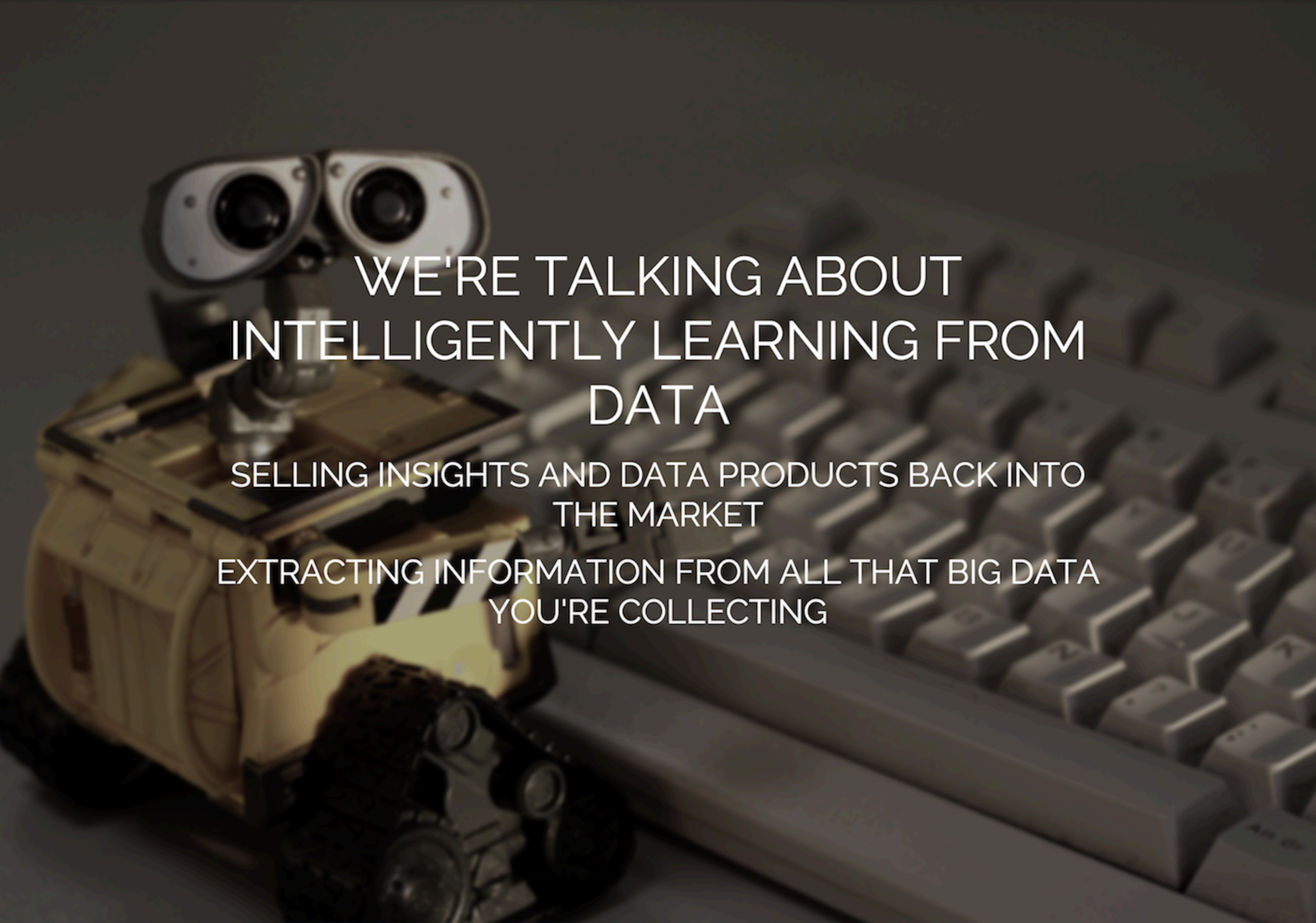


WE'RE TALKING ABOUT INTELLIGENTLY LEARNING FROM DATA

TRAINING MODELS TO PREDICT OPERATIONAL OUTCOMES

DISCOVERING CORRELATIONS, MATCHING PATTERNS OF BEHAVIOR

RUNNING THE BUSINESS MORE EFFECTIVELY



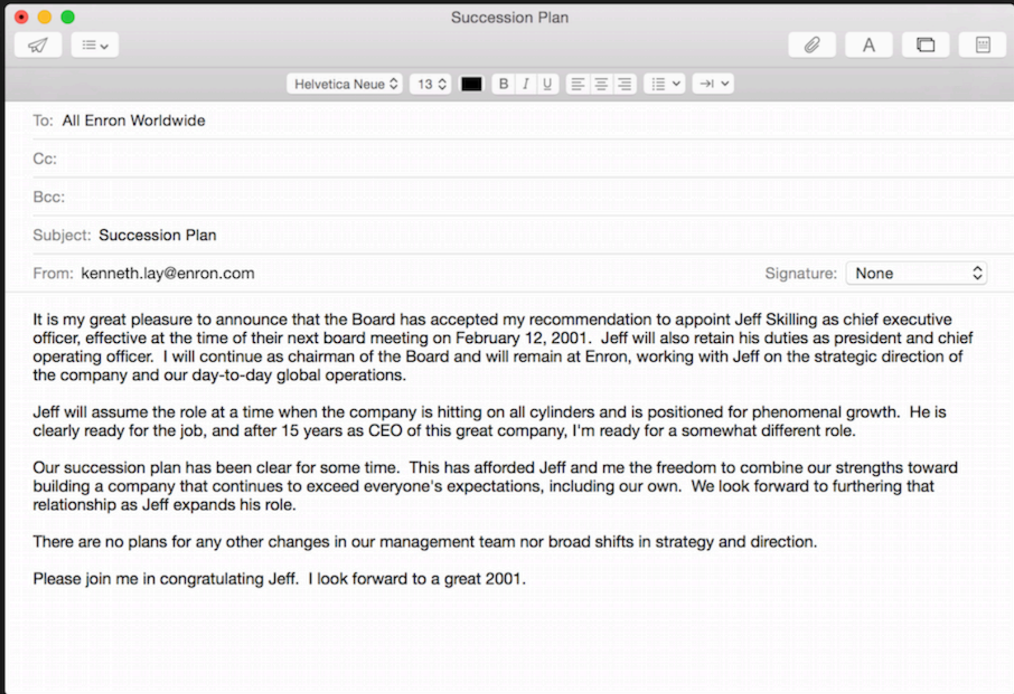
WE'RE TALKING ABOUT INTELLIGENTLY LEARNING FROM DATA

SELLING INSIGHTS AND DATA PRODUCTS BACK INTO THE MARKET

EXTRACTING INFORMATION FROM ALL THAT BIG DATA YOU'RE COLLECTING


THERE'S A HUGE AMOUNT OF DATA ALREADY INSIDE THE BUSINESS
IT JUST TAKES A LITTLE PUTTING TOGETHER

NLP UPON UNCONVENTIONAL DATA SOURCES



Succession Plan

To: All Enron Worldwide

Cc:

Bcc:

Subject: Succession Plan

From: kenneth.lay@enron.com

Signature: None

It is my great pleasure to announce that the Board has accepted my recommendation to appoint Jeff Skilling as chief executive officer, effective at the time of their next board meeting on February 12, 2001. Jeff will also retain his duties as president and chief operating officer. I will continue as chairman of the Board and will remain at Enron, working with Jeff on the strategic direction of the company and our day-to-day global operations.

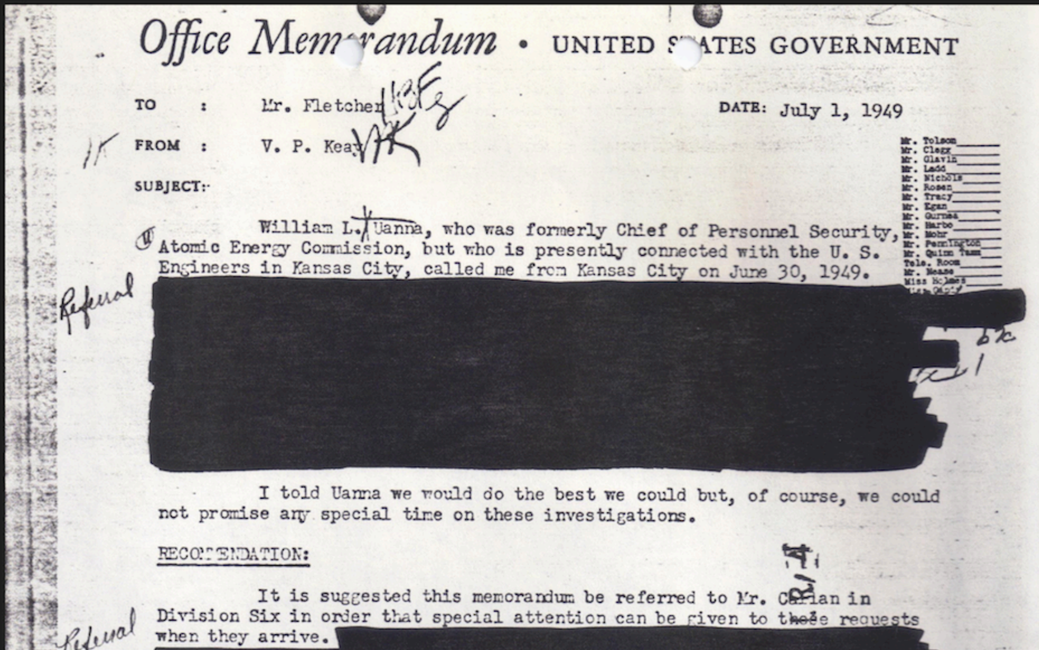
Jeff will assume the role at a time when the company is hitting on all cylinders and is positioned for phenomenal growth. He is clearly ready for the job, and after 15 years as CEO of this great company, I'm ready for a somewhat different role.

Our succession plan has been clear for some time. This has afforded Jeff and me the freedom to combine our strengths toward building a company that continues to exceed everyone's expectations, including our own. We look forward to furthering that relationship as Jeff expands his role.

There are no plans for any other changes in our management team nor broad shifts in strategy and direction.

Please join me in congratulating Jeff. I look forward to a great 2001.

AFTER CAREFUL CLEANING AND PREPROCESSING



Office Memorandum • UNITED STATES GOVERNMENT

TO : Mr. Fletcher

DATE: July 1, 1949

FROM : V. P. Keay

SUBJECT:

Mr. Tolson
Mr. Clegg
Mr. Glavin
Mr. Ladd
Mr. Nichols
Mr. Rosen
Mr. Tracy
Mr. Egan
Mr. Gurnea
Mr. Harbo
Mr. Mohr
Mr. Pennington
Mr. Quinn Tamm
Tele. Room
Mr. Nease
Miss Holmes

Referral

William L. Uanna, who was formerly Chief of Personnel Security, Atomic Energy Commission, but who is presently connected with the U. S. Engineers in Kansas City, called me from Kansas City on June 30, 1949. [redacted]

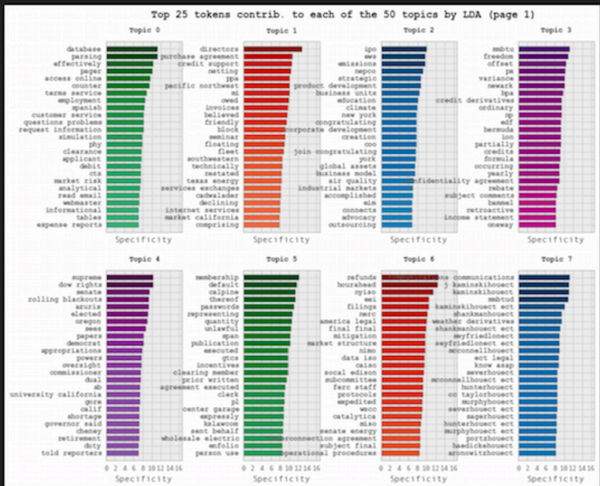
I told Uanna we would do the best we could but, of course, we could not promise any special time on these investigations.

RECOMMENDATION:

It is suggested this memorandum be referred to Mr. Carlan in Division Six in order that special attention can be given to these requests when they arrive. [redacted]

Referral

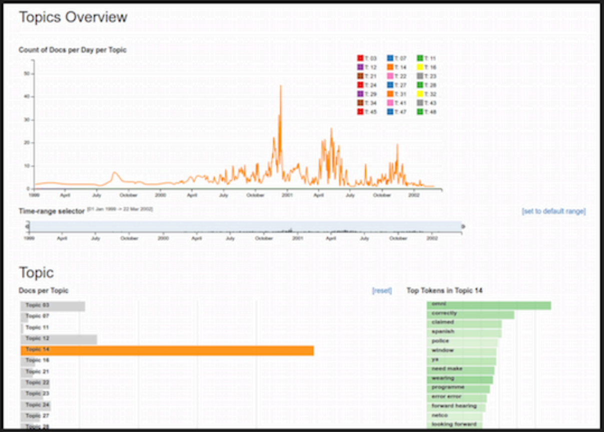
MODEL THE TOPICS OF CONVERSATION WITHIN A CORPUS



- Topic modelling (Latent Dirichlet Allocation)
- Cluster words into topics by their co-occurence
- Gain a new way to define the contents of documents

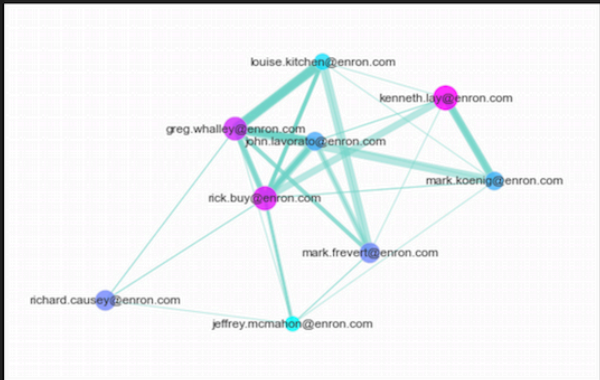
USE TOPICS AS A METHOD TO INTROSPECT THE HEARTBEAT OF COMMUNICATIONS

- View trends of topics by time
- Correlate topics to internal / external events
- Use topics in other models e.g. predict customer satisfaction & longevity according to topics




CAN WE USE SOCIAL
& ORGANISATIONAL
STRUCTURE TOO?

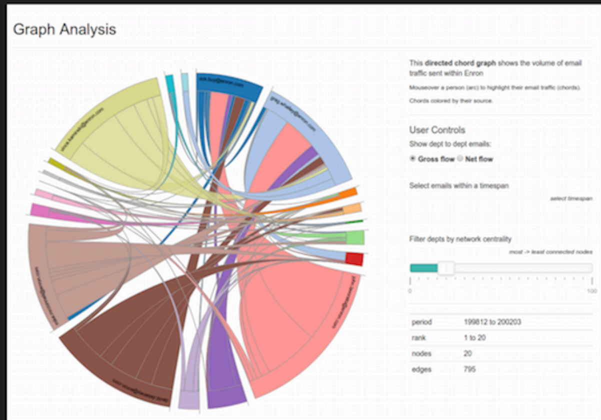
USE THE FROM: AND TO: TO CREATE A NETWORK / GRAPH



- Nodes are email accounts / people
- Edges are the communications between them
- Overall counts / frequency / time of day / season

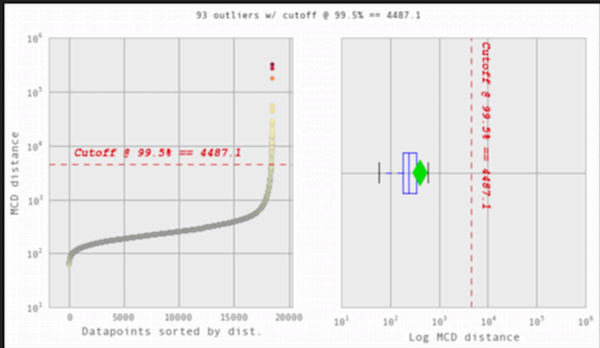
ATTACH THOSE TOPICS FROM BEFORE TO GAIN A RICH PICTURE OF COMMUNICATIONS

- Non-invasive introspection of the state of the organisation
- Relevant to networks of brokers, employees, customers, reinsurers etc.




A FEW MORE THINGS
NOT YET MENTIONED
THING
THING

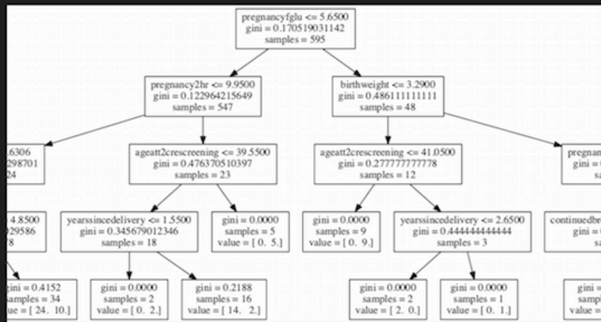
OUTLIER / FRAUD DETECTION



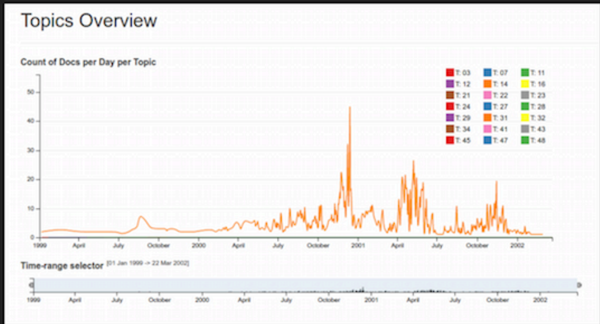
- Describing data in a vector space and assigning outlier flags to unusual points
- (or use graphs / trees)
- Triage these first in fraud or operations investigations

SUPERVISED CLASSIFICATION

- Good old-fashioned prediction
- Product-market fit
- Operational improvements
- Meeting compliance etc.



INTERACTIVE DASHBOARDING



- Surface descriptive & predictive insights to the business to enable better decision making
- Lightweight Javascript frameworks are powerful and easy to use


WE'VE ONLY
SCRATCHED THE
SURFACE TODAY
BUT MANY COMPANIES ARE ALREADY DOING THIS

IN SUMMARY

DATA SCIENCE IS APPLICABLE THROUGHOUT THE INSURANCE BUSINESS

-

WE ARE PROVING THIS EVERY DAY THROUGH OUR PROJECTS

1. CURATING EXTERNAL DATA TO BETTER UNDERSTAND CUSTOMERS
2. BRINGING STATISTICAL MODELING INTO THE BOARDROOM
3. APPLYING MORE EXOTIC MACHINE LEARNING ACROSS THE BUSINESS


THE TOOLS ARE POWERFUL AND LARGELY FREE


THERE'S A WEALTH OF
INFORMATION ALREADY INSIDE
THE BUSINESS


AND WITH INTELLIGENT CURATION, HIGH QUALITY EXTERNAL DATA IS READILY AVAILABLE

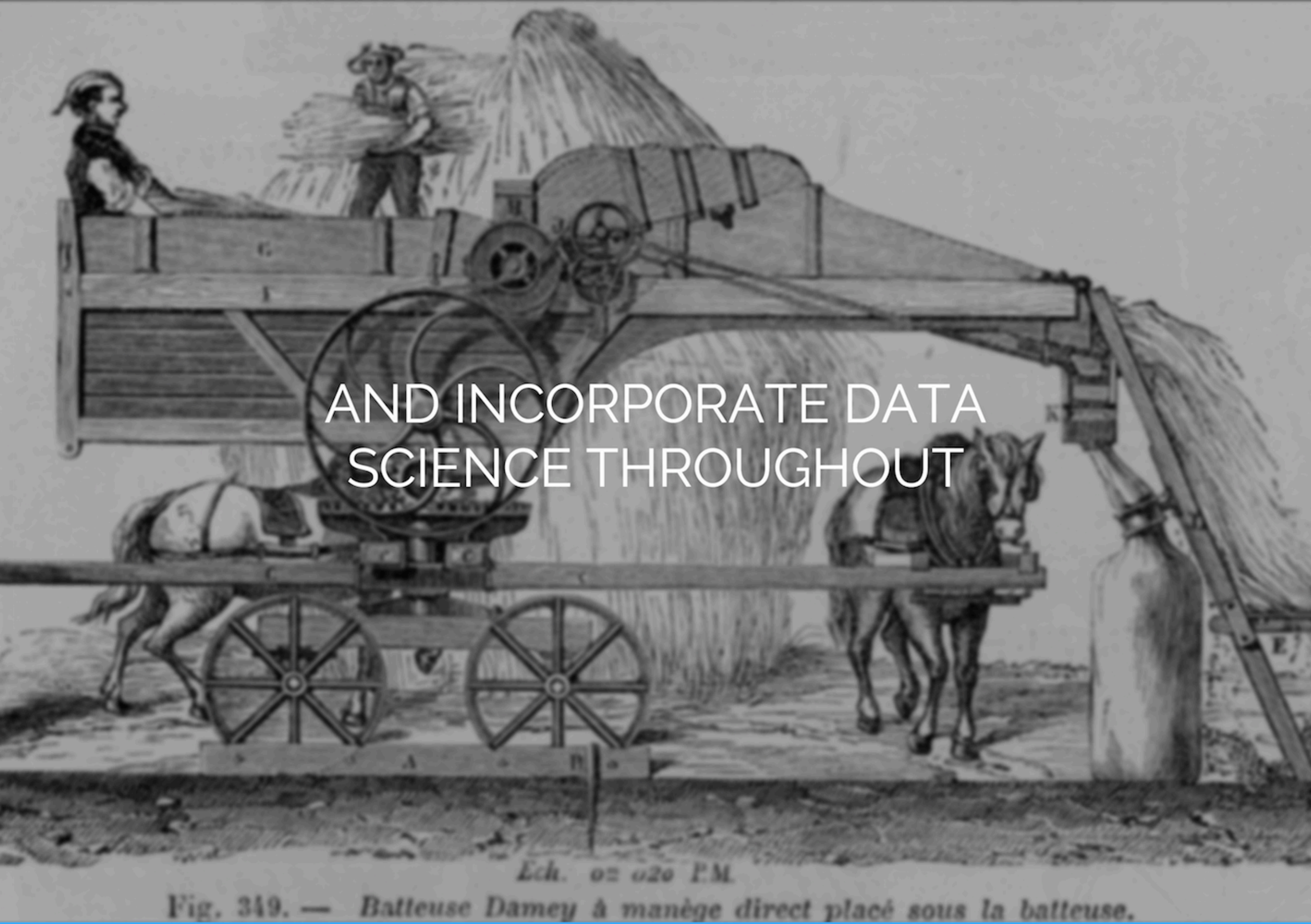

SPEND MONEY ON THE TEAM


ENABLE THEM TO HONE & SHARE
SKILLS & EXPERTISE

PLOUGH INSIGHTS BACK INTO THE BUSINESS

Ech. 0= 020 P.M.

Fig. 349. — Batteuse Damey à manège direct placé sous la batteuse.


AND INCORPORATE DATA
SCIENCE THROUGHOUT
Fig. 349. — Batteuse Damey à manège direct placé sous la batteuse.


APPLIED
AI