

Beating The Best - The Santander Bank Kaggle

RUser's Group - Calgary

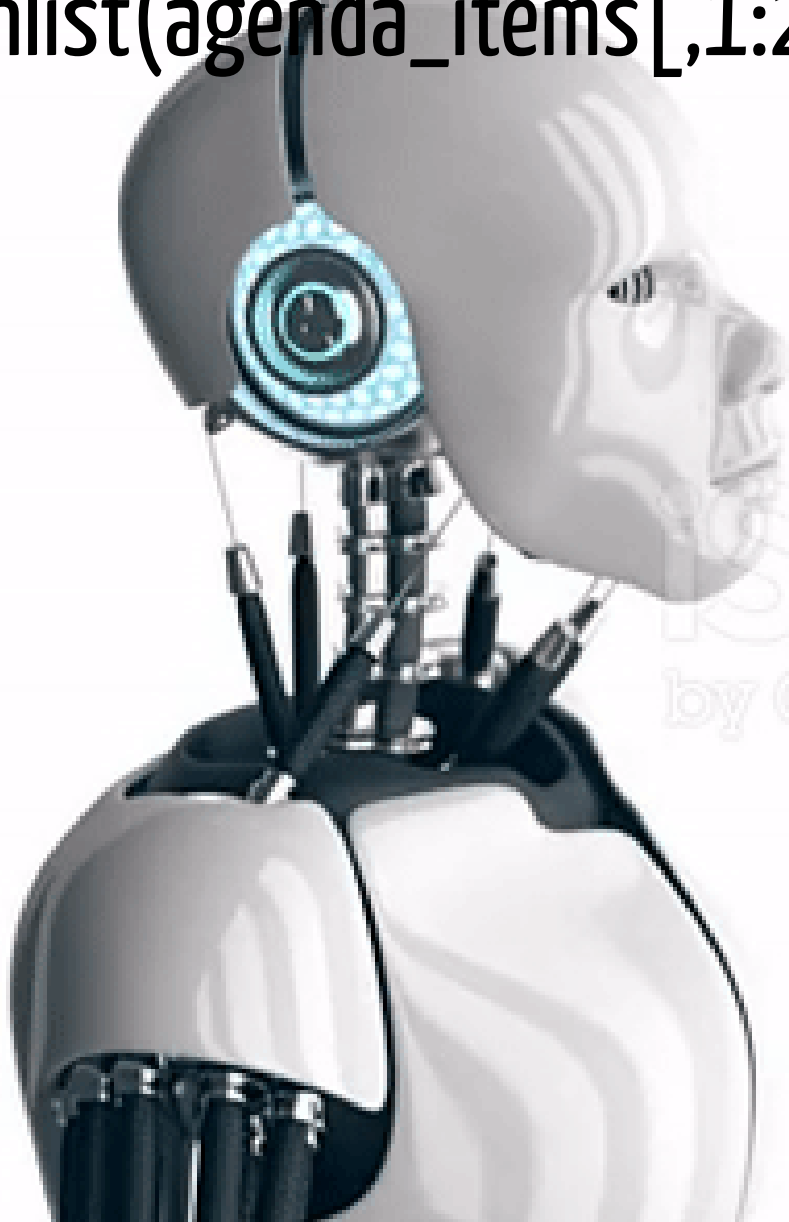
Alastair K. Muir, PhD, MBB

Muir&Associates Consulting

Calgary, Alberta, Canada

March 20, 2019

`unlist(agenda_items[,1:2])`



The Customer Value Project

Kaggle Competition - Pt. 1

- Interesting problem, no explanation
- Leak! Competitors revolt!
- Alastair goes to the cabin with no WiFi

What 4,483 others missed

- The raw data gives vital clues about the process

Kaggle Competition - Pt. 2

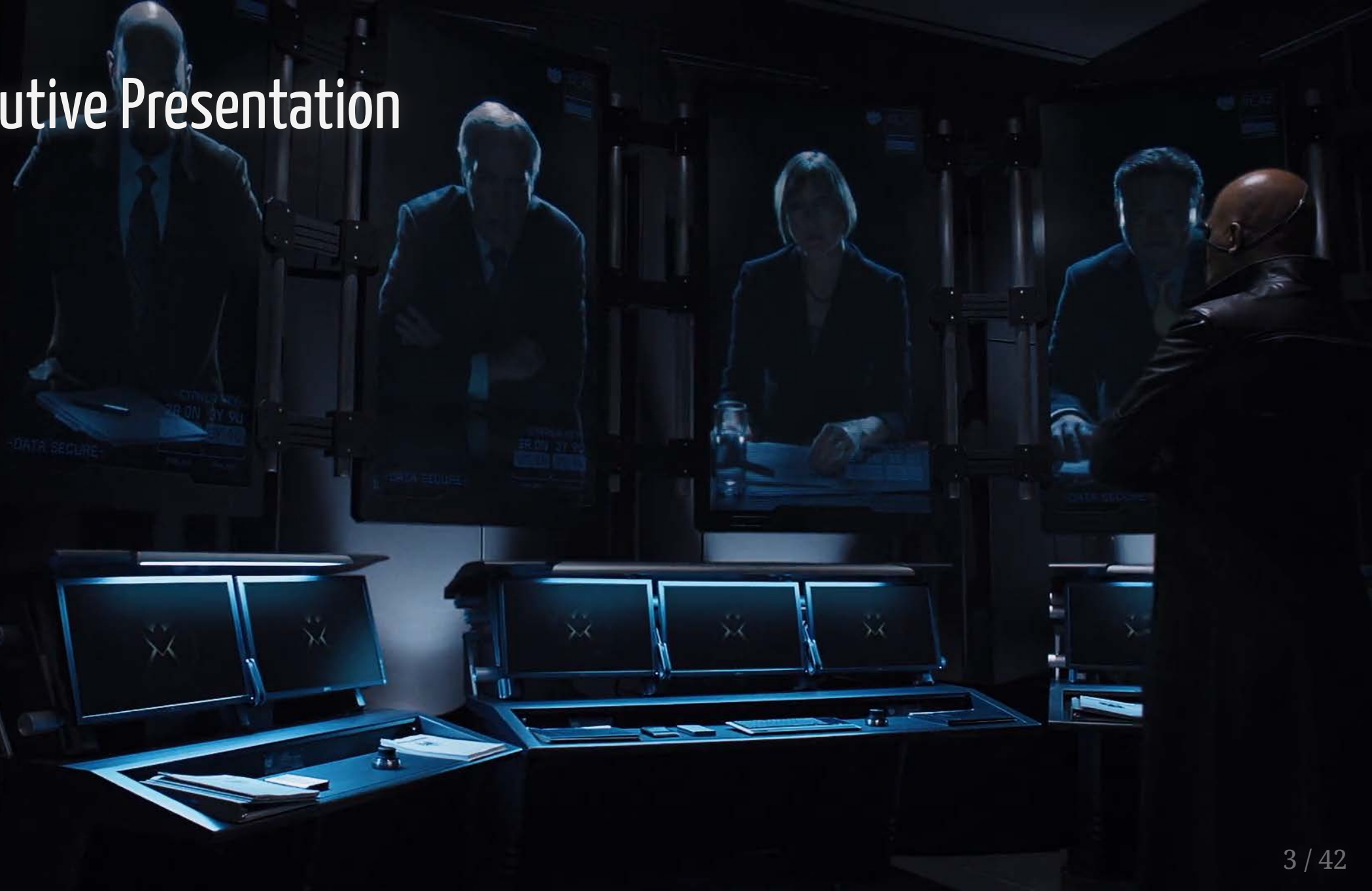
- Winner has code, but no explanation

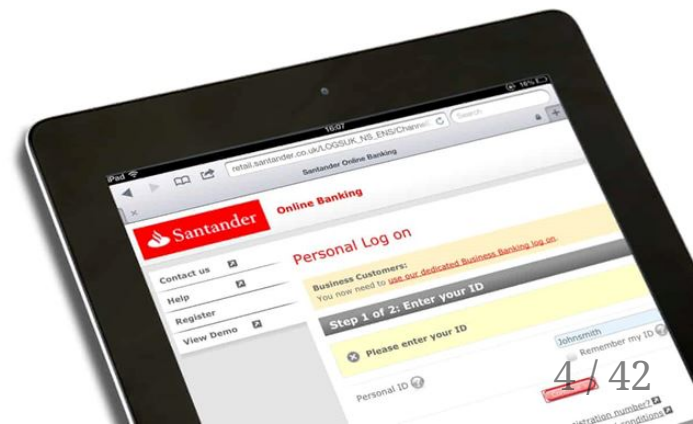
Building a Model That is Useful

Draw for Prizes

- Skill testing question

Executive Presentation





Customer Value Project - Team



Věra Kůrková
VP Transformation



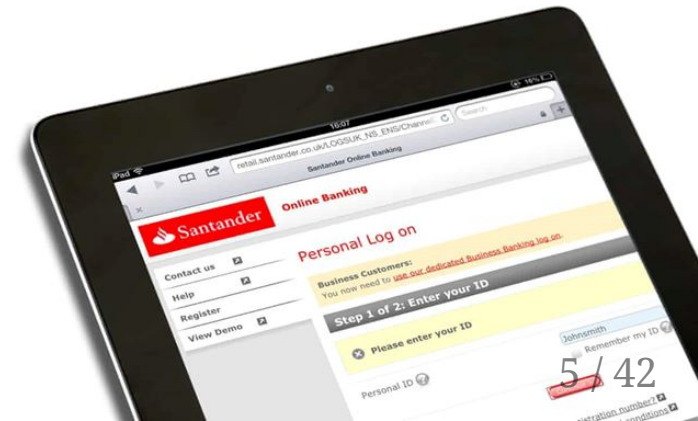
Alastair Muir
Director Data Science







Huang Lu 黄璐
Data Scientist
Mobile and Web
[1] thispersondoesnotexist.com



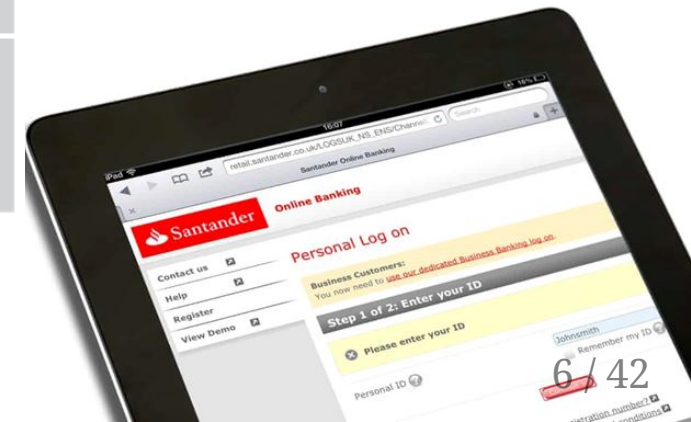
Paul Erdős
Data Scientist
Customer Value



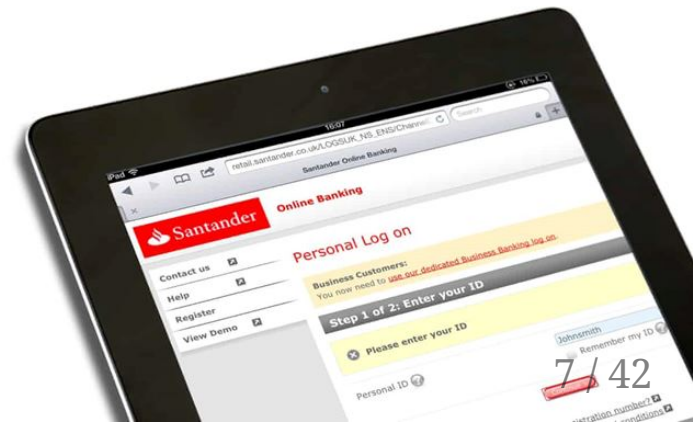
Our Bank's Strategic Position in the Value Chain Spectrum

	Manufacturing focus	Hybrid	Distribution focus
 Ambition	Best-in-class production and processing of banking products	Growth in select core markets via distinct products, customer segments, geographies	Best-in-class client insight, and management of channels and relationships
 Where to play	Best-in-class solutions for specific customer segments, including other banks	Distinct choices of products, according to customer segments and geography	Full product suite, bundled and tailored to the sector and size of the customer with white-label solutions
 How to win	Economies of scale; high fixed costs require large volumes to hold down unit costs	Both scale and scope; manufacturing in core local markets and distribution in select overseas markets	Economies of scope; high cost of acquiring clients makes a large share of wallet essential
 Examples	Black Rock, State Street, Goldman Sachs, parts of JPMorgan Chase	Many large banks in their home markets	Community banks, smaller overseas branches

[1] BIAN & Company (Banking Industry Architecture Network)



Customer Value Project - Predict Customer Value

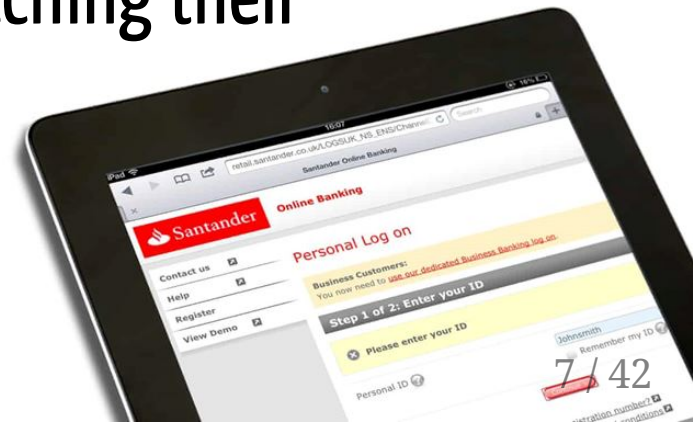


Customer Value Project - Predict Customer Value

Managing customer value and relationships means we must anticipate customer needs in a concrete, simple and personal way. With so many choices for financial services, this need is greater now than ever before.

We can't read their minds, but we must understand how individual customers make decisions.

We can determine what influences their decisions by watching their behaviour when they interact with us.



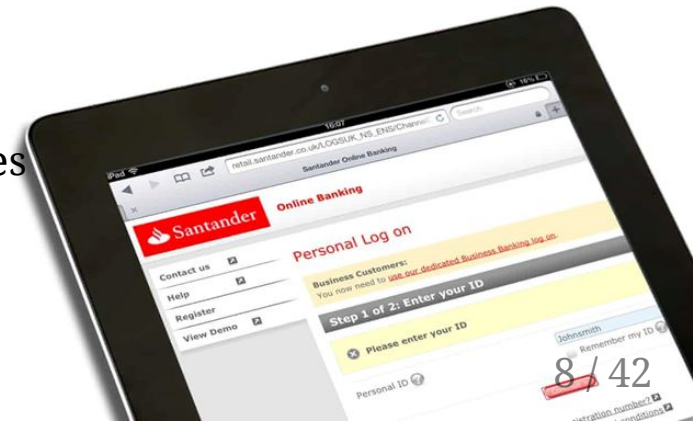
Customer Value Project - Our Solution:

Accurately predicts timing of a customer's next interaction with the bank

- Tracking a customer's behaviour in real time allows us to anticipate their future choices.
- This also gives us insight into the decision making process for each individual customer.
- We sponsored an international Kaggle data science competition with \$60,000 in prizes.
- Our team's solution is more accurate than all 4,484 international teams including the number 1 rated Kaggle in the world.

Scalable to all users and services

- Defines new **Customer Value Metrics** to use for marketing campaigns, service bundling, and new services introduction
- Applicable to mobile, web, call centre, and teller channels
- Coordinates with campaign and service offerings rollouts
- Follows customer information privacy and vendor information sharing policies
- Integrates with internal data sources and existing bank core services



Customer Value Project - Rollout

Project Sponsor - Věra Kůrková, VP Transformation

Consulting and Communication

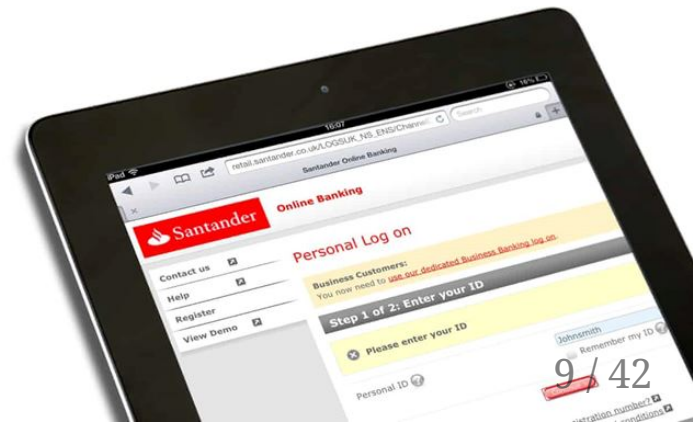
- Transformation and Strategy Planning - project rollout integration
- Operations and Marketing - new and ongoing campaign assessment; web, mobile, and teller channels
- Risk - assessment of cloud computing solution
- HR - program training phasing
- Customer Services - Customer privacy and vendor information guidelines
- IT - data feeds and cloud implementation

Reporting - Monthly report to executive, semiweekly to directors and project sponsor

Team - PM + three resources from Data Science group for four months

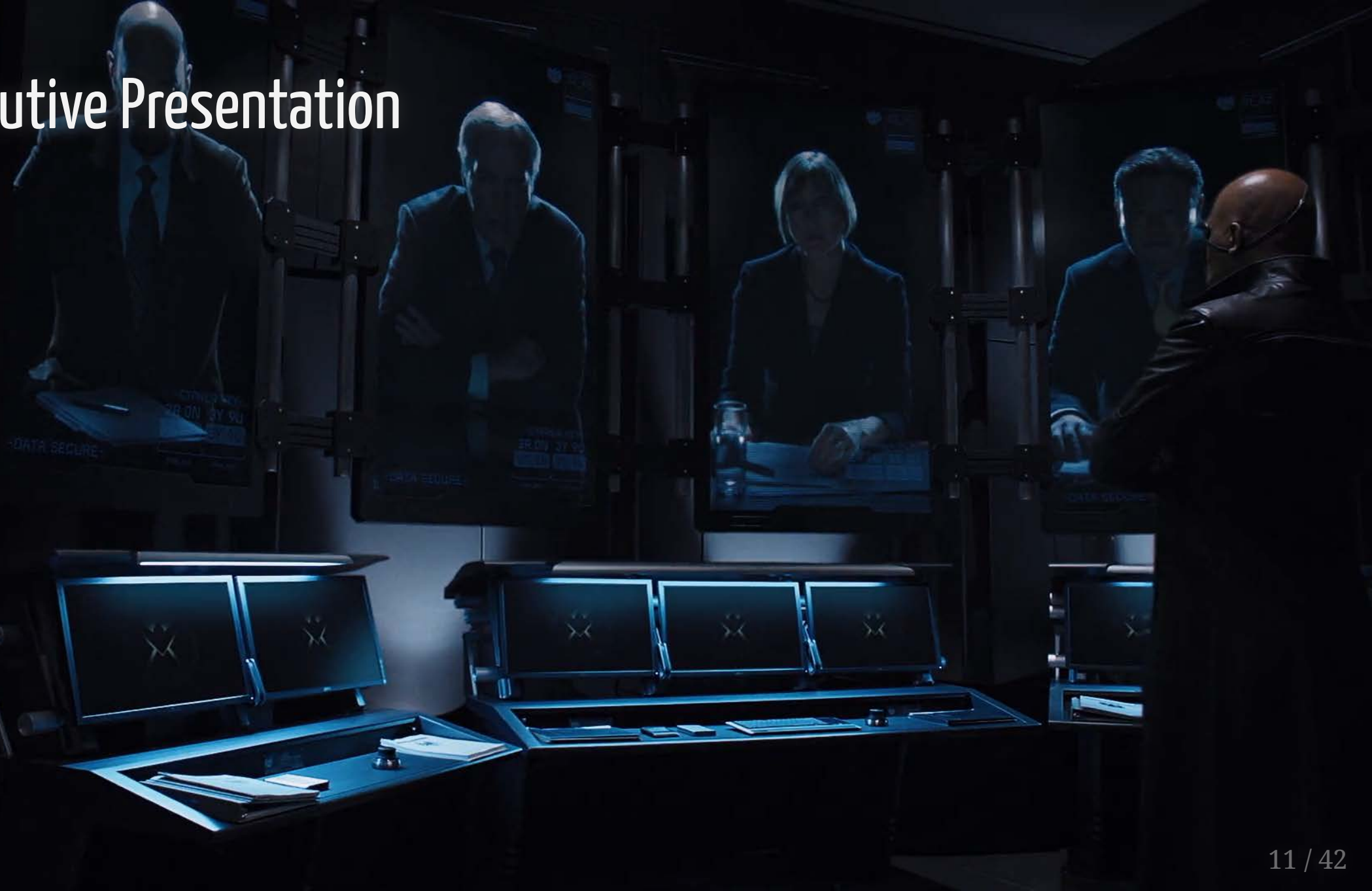
Other Transformation Projects that may be impacted or augmented

- Customer transaction prediction (Who will make a transaction?)
- Product recommendation (Can you pair products with customers?)
- Customer satisfaction (Which customers are happy customers?)





Executive Presentation



Executive Presentation

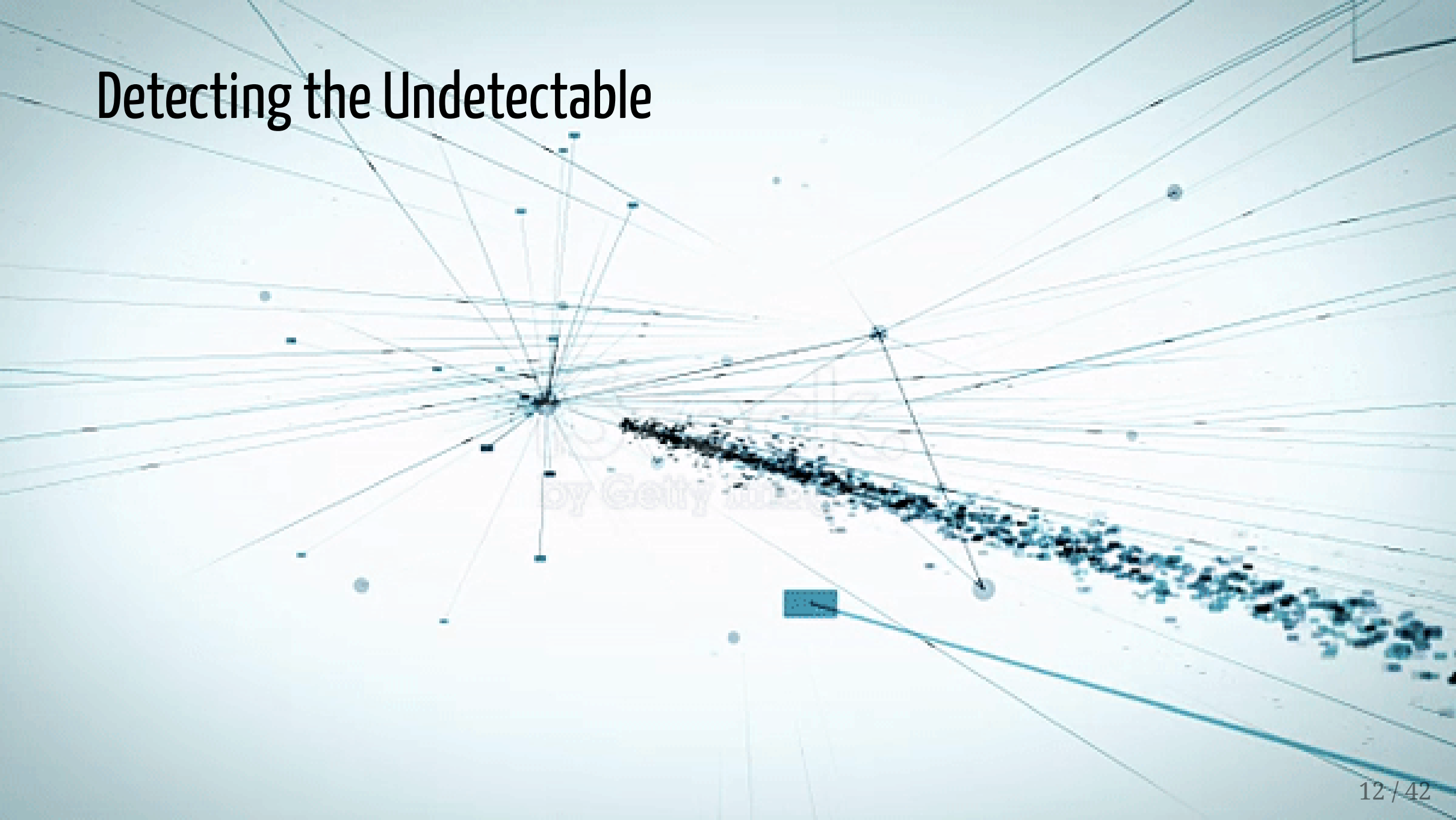
Include

- Introduce the team, give credit
 - What is the problem?
 - Why does the problem exist?
 - Why is this important now?
 - What could we do about it?
 - What should we do about it?
 - Is this solution the best choice now?
 - Do we have your endorsement?
- Rehearse, try to anticipate questions (only when asked, though)
- Identify the barriers to implementation
- Rough Return on Investment timeline, if applicable
- Have a communication plan involving the stakeholders

Avoid

- "model", "theoretical"; "technical", "new", "game changing", "cutting edge"
- Your PhD/Masters degrees
- Your experience at Facebook, Cambridge Analytica, Landsbanki...
- Details of your new favourite algorithm
- Describing the logical, sequential process of getting to the solution. They aren't listening to understand the methodology, that's your job(s)
- Acronyms
- A progress report that could have been handled by an email

Detecting the Undetectable



The Journey from Data to Insight

Data

Data, in and by themselves, do not directly create knowledge. In the age of **Big Data**, the availability of vast amounts of information can coexist with the **absence of knowledge**.

Analysis, Interpretation and Knowledge

It is in when we interpret data that knowledge is created. My focus is on **bridging the gap between data and knowledge creation**. That gap is filled by **statistical analysis** and **evidence based decision making**.

The Journey

An **algorithm** is process or set of rules to be followed in calculations or other problem solving operation. An algorithm can be a series of functions operating in sequence.

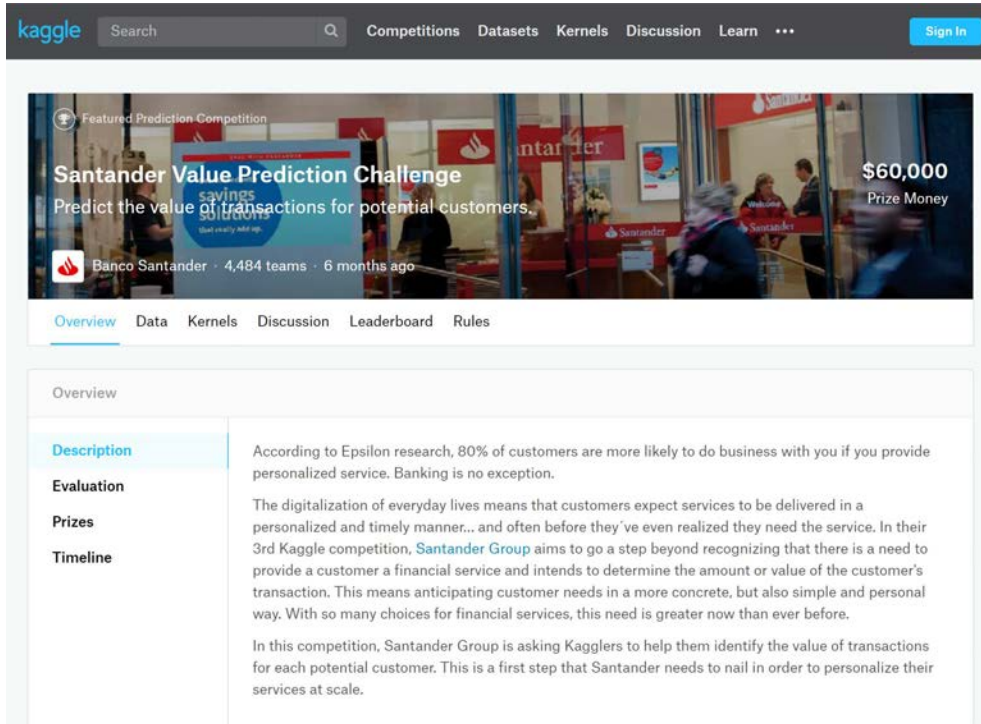
The Journey Begins

According to Epsilon research, **80% of customers are more likely to do business with you if you provide personalized service.** Banking is no exception.

The digitalization of everyday lives means that **customers expect services to be delivered in a personalized and timely manner... and often before they've even realized they need the service.** In their 3rd Kaggle competition, Santander Group aims to go a step beyond recognizing that there is a need to provide a customer a financial service and intends to determine the amount or value of the customer's transaction. **This means anticipating customer needs in a more concrete, but also simple and personal way. With so many choices for financial services, this need is greater now than ever before.**

In this competition, Santander Group is asking Kagglers to help them identify the value of transactions for each potential customer. **This is a first step that Santander needs to nail in order to personalize their services at scale.**

The Santander Value Prediction Challenge



The screenshot shows the Kaggle interface for the Santander Value Prediction Challenge. At the top, the Kaggle logo and navigation links (Search, Competitions, Datasets, Kernels, Discussion, Learn, Sign In) are visible. The main banner features the competition title "Santander Value Prediction Challenge" with a subtitle "Predict the value of transactions for potential customers." and a prize of "\$60,000 Prize Money". Below the banner, the competition is organized by "Banco Santander" with "4,484 teams" and "6 months ago" remaining. The "Overview" tab is selected, showing a table with sections: Description, Evaluation, Prizes, and Timeline. The Description section contains text about personalized banking services and the competition's goal.

Section	Content
Description	According to Epsilon research, 80% of customers are more likely to do business with you if you provide personalized service. Banking is no exception.
Evaluation	The digitalization of everyday lives means that customers expect services to be delivered in a personalized and timely manner... and often before they've even realized they need the service. In their 3rd Kaggle competition, Santander Group aims to go a step beyond recognizing that there is a need to provide a customer a financial service and intends to determine the amount or value of the customer's transaction. This means anticipating customer needs in a more concrete, but also simple and personal way. With so many choices for financial services, this need is greater now than ever before.
Prizes	In this competition, Santander Group is asking Kagglers to help them identify the value of transactions for each potential customer. This is a first step that Santander needs to nail in order to personalize their services at scale.
Timeline	

train.csv 62.1 MB; 4,459 observations(96.5% are zero); 4,993 variables

test.csv 967.2 MB; 49,342 observations(96.5% are zero); 4,992 variables

One participant achieved an RMSE($\ln(\text{target})$) on test data of 0.70 while hundreds of others were only just getting to 1.6-1.7.

LEAK, LEAK, LEAK

This competition will test your ability to perform without relying on any domain knowledge. While we cannot disclose the nature of these methods or the extent of possible leakage, the host has made the decision to continue to run the competition as is.

The Santander Value Prediction is a Murdoch Mystery



There is a mysterious situation...

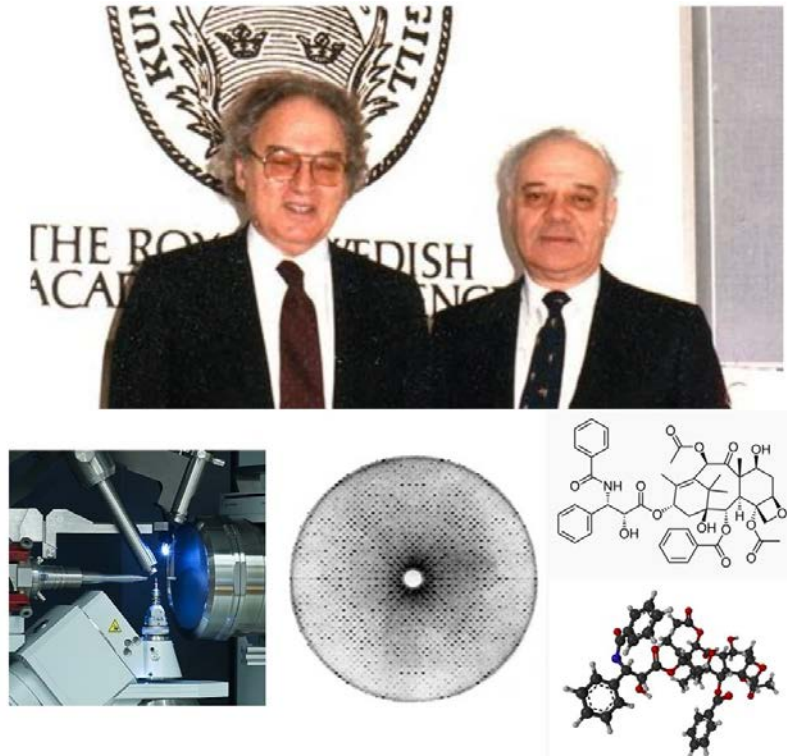
- It requires a systematic, objective investigation
- We have to search for clues and leads
- There are many false leads and dead ends
- It involves research into the latest technology and applying it in new ways
- The project features a handsome and intelligent lead

The search for clues begins...

[1] Murdoch Mysteries, CBC

[2] Shouldn't we all have a theme song?

Data Distributions Give Vital Information

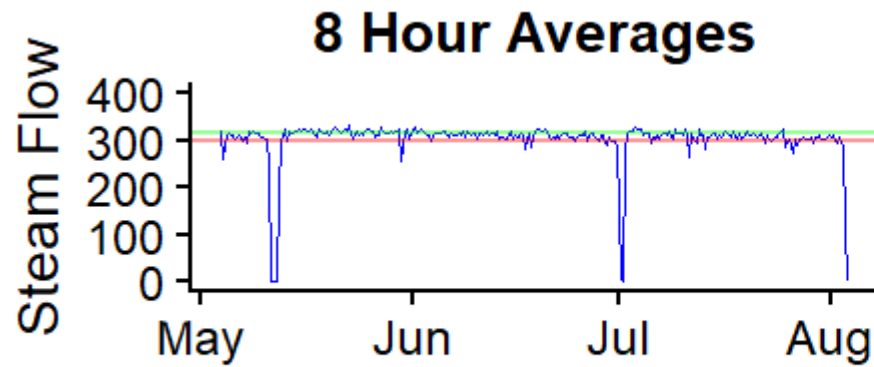


- Herb Hauptman and Jerome Karle awarded the 1985 Nobel Prize in Chemistry for techniques in determining the 3D structure of molecules
- The intensities of diffraction patterns are proportional to squares of the Fourier transform of the electron densities
 - Electron density gives you a diffraction pattern
 - The diffraction pattern does **not** give you electron density directly

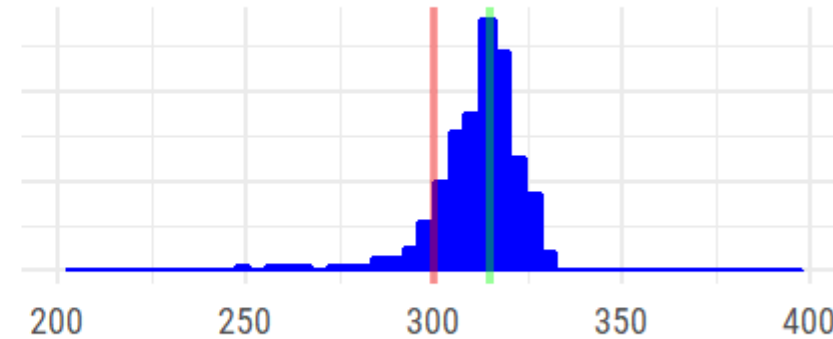
$$F_{hkl} = \sum_j f_j \cos[2\pi(hx_j + ky_j + lz_j)] \\ + i \sum_j f_j \sin[2\pi(hx_j + ky_j + lz_j)]$$

- Statistical properties of intensity distributions allow the determination of phases

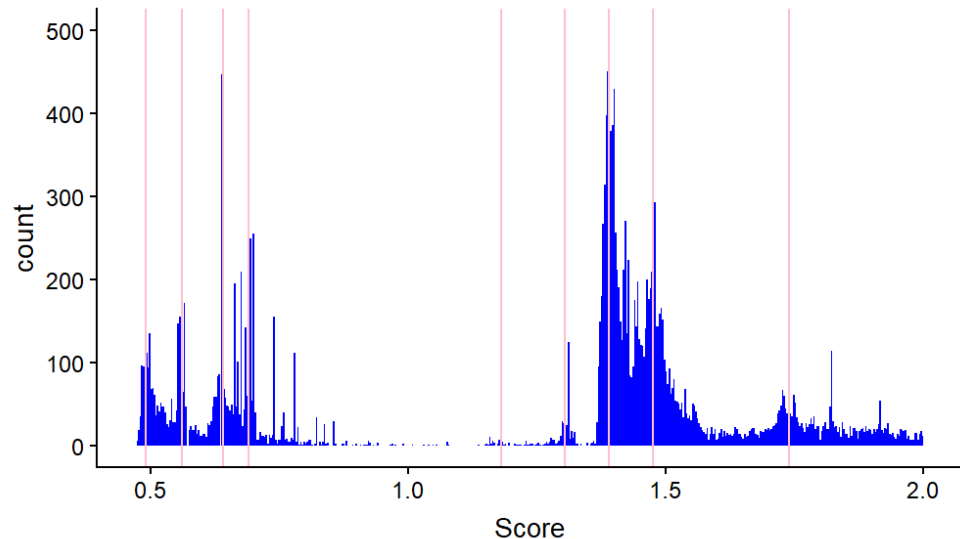
Averaging Data Hides Vital Clues: Use All the Data



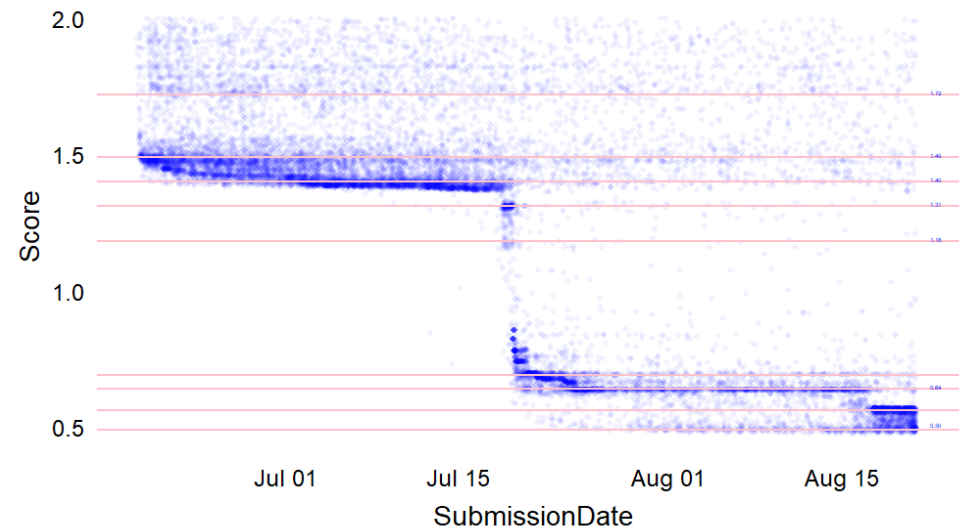
Target - 315, Lower Spec 300



The Kaggle Leaderboard - Dr. Ogden Conducts a Post-mortem



Different subgroups are present in this competition. Width of subgroups corresponds to hyperparameter tuning and random starting positions.

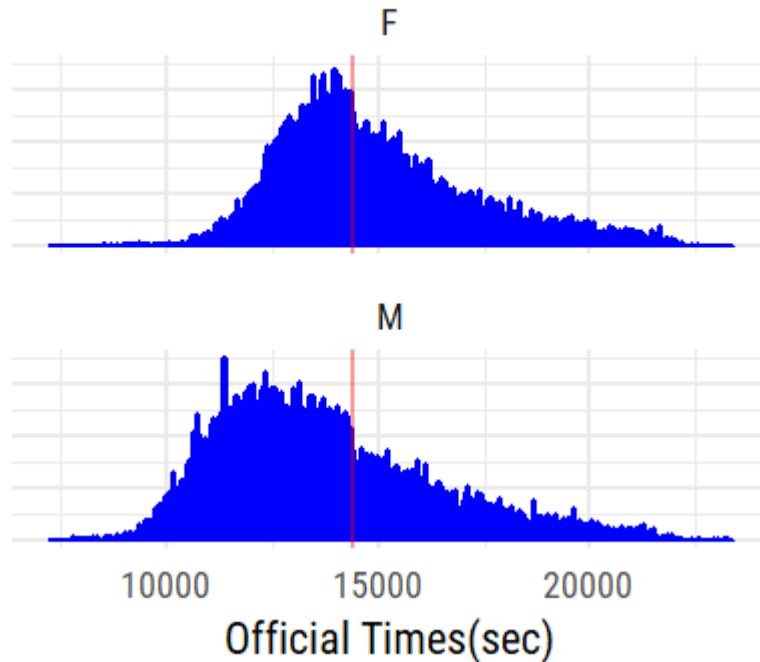


Different subgroups evolve over time as kernels are published and adopted by other Kagglers.

Clearly, throwing a very sparse matrix of 4,459 observations of 4,992 variables at a stacked **LightGBM, XGBoost, CatBoost** model is not the best approach (RMSE = 1.53).

Examine Data Distributions for Clues

2017 Boston Marathon



Why is there a kink at 14,400 seconds?

Where else does this apply?

- Accounts payable behaviour - over \$1B/year
- Raw diamond size (fraud detection)
- Medical malpractice suit - estimate of damages
- Seismic risk with hydraulic fracturing in Montney formation
- UO_2 production - 20% increase in capacity
- \$430M yearly maintenance budget forecast - board of directors
- Power plant steam generation - operations and maintenance (5% increase)
- Oil sands processing plant feed rate increase (\$95M/year)
- Financial transactions (money laundering detection)
- Capital expense risk forecast (Wall Street credit rating, over \$1B/year)
- Crisis Centre call volume (operations)
- Supply chain - order to remittance

Fitting a Distribution to Data - Parameters

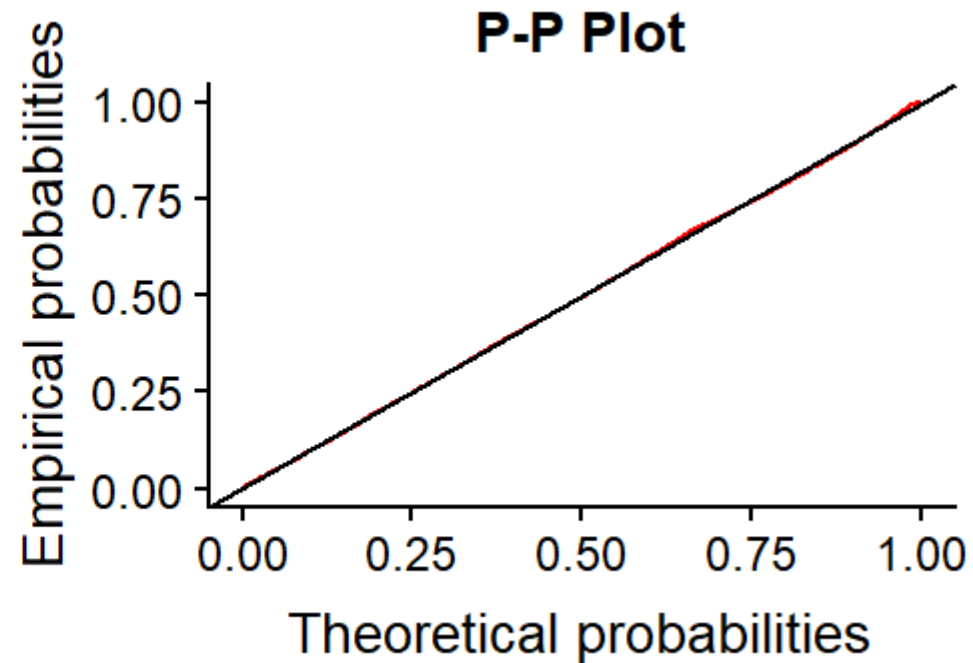
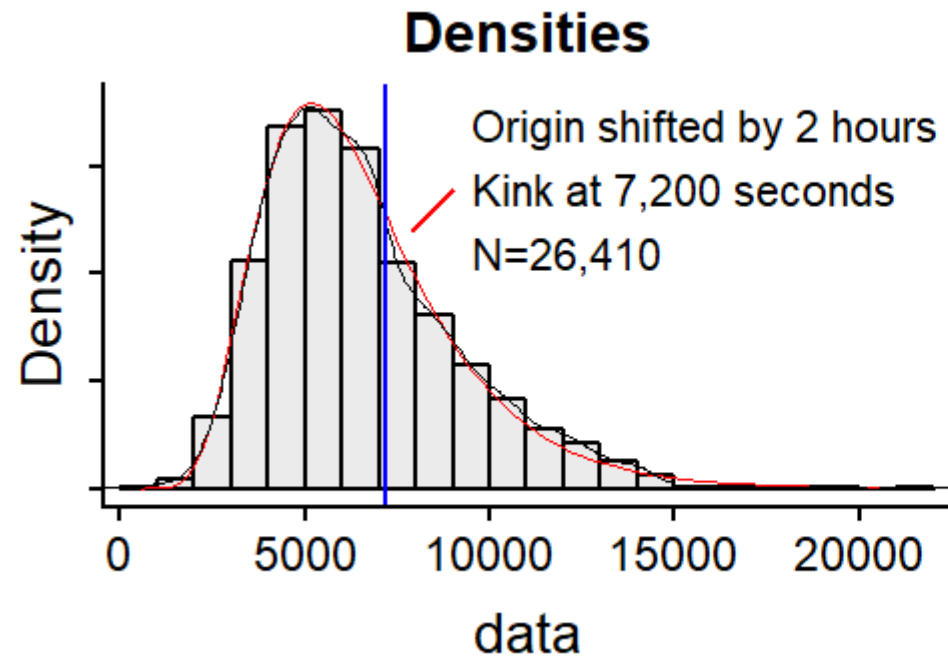
```
library(data.table)
library(hms)
library(fitdistrplus)
library(ggplot2)
marathon_results_2017 <- data.table::fread(file = "figures/boston_marathon/marathon_results_2017")
marathon_times <- dplyr::select(marathon_results_2017, c("Official Time", "M/F"))
marathon_times$`Official Time` <- hms::parse_hms(marathon_times$`Official Time`)
marathon_times_male <- marathon_times %>%
  dplyr::filter(`M/F` == "M") %>%
  dplyr::select("Official Time")

male_times <- as.numeric(unlist(marathon_times_male),
                           units="days")
params <- fitdist(male_times-7200, distr = "lnorm", method = "mle")
params
```

```
## Fitting of the distribution ' lnorm ' by maximum likelihood
## Parameters:
##           estimate Std. Error
## meanlog 8.7083360 0.003305627
## sdlog   0.3971982 0.002337364
```

Fitting a Distribution to Data - Checking the Fit

- unimodal, bimodal, multimodal?
- log-normal, normal, Weibull, exponential?



A good fit to the best fit distribution is shown by a straight line in the P-P plots.

The diagram illustrates the relationships between various probability distributions, categorized into discrete and continuous types. The nodes are arranged in a hierarchical structure, with arrows indicating the direction of the relationships.

Discrete Distributions (Top):

- Geometric** (Self-loop, connected to Negative Binomial)
- Negative Binomial** (connected to Geometric, Poisson, Binomial)
- Poisson** (Self-loop, connected to Negative Binomial, Binomial)
- Binomial** (connected to Poisson, Bernoulli, Beta-Binomial, Hypergeometric)
- Beta-Binomial** (connected to Discrete Uniform, Binomial)
- Hypergeometric** (connected to Binomial)
- Bernoulli** (connected to Binomial)
- Discrete Uniform** (connected to Beta-Binomial)

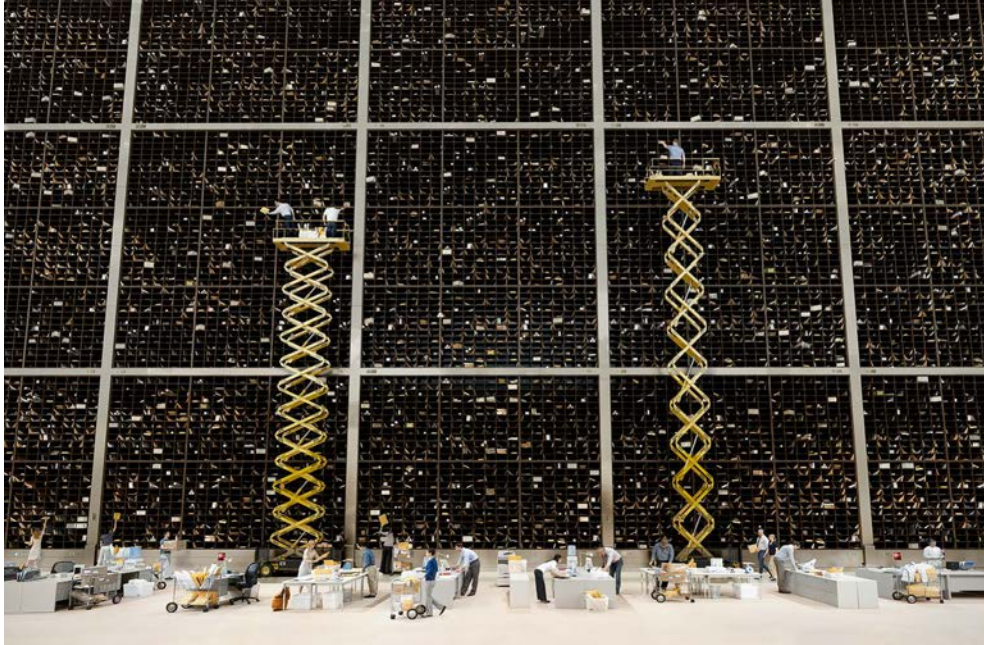
Continuous Distributions (Bottom):

- Lognormal** (Self-loop, connected to Normal)
- Normal** (Self-loop, connected to Lognormal, Beta, Gamma, Chi-squared, Standard Normal)
- Beta** (connected to Normal, Gamma, Uniform)
- Gamma** (connected to Normal, Beta, Chi-squared, Exponential)
- Chi-squared** (Self-loop, connected to Normal, Gamma, Standard Normal, Snedecor F, Exponential)
- Standard Normal** (connected to Normal, Chi-squared, Cauchy, Student t)
- Cauchy** (Self-loop, connected to Standard Normal, Student t)
- Student t** (connected to Standard Normal, Cauchy, Snedecor F)
- Snedecor F** (Self-loop, connected to Chi-squared, Student t, Exponential)
- Uniform** (connected to Beta, Gamma, Exponential)
- Exponential** (Self-loop, connected to Gamma, Chi-squared, Snedecor F, Weibull, Double Exponential)
- Weibull** (connected to Exponential)
- Double Exponential** (connected to Exponential)

The diagram shows a complex network of relationships, with many distributions having multiple parents or children. For example, the Normal distribution is a central node, connected to several other distributions. The Exponential distribution is also a central node, connected to many other distributions.

- Financial transactions - log-normal
 - Benford analysis(fraud detection)
- Boston marathon times - overlapping mixture of shifted log-normal
- Nuclear waste decay - overlapping mixture of exponentials
- Time gap between arrival times - exponential
- Tonnage of oil-sands loads to the crusher - overlapping normal(Gaussian)
- Customer service requests per day - Poisson
- Time to failure - Weibull
 - infant mortality
 - wearing out
- **Time for a customer to respond** - Weibull

S12:E13 Muirdoch and the Undetectable

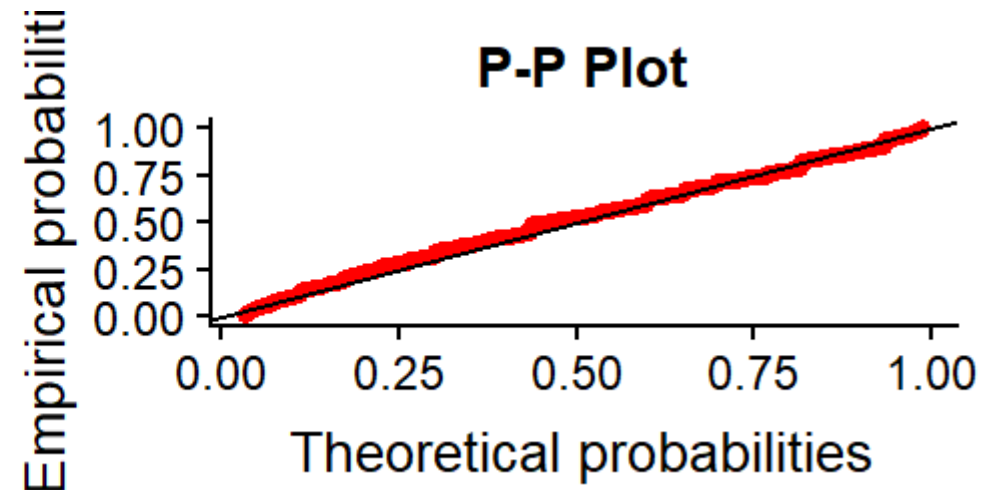
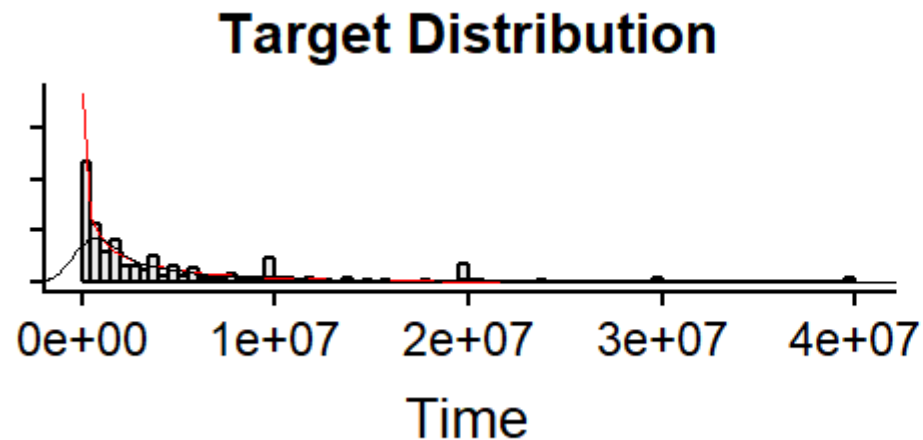


- **train.csv**
 - 62.1 MB
 - 4,459 observations
 - 4,993 variables
- **test.csv**
 - 967.2 MB
 - 49,342 observations
 - 4,992 variables
- 96.5% values are zero
- all column names are hashed
- all row names are hashed
- column order is randomized
- row order is randomized
- variable assignment within subgroups is scrambled

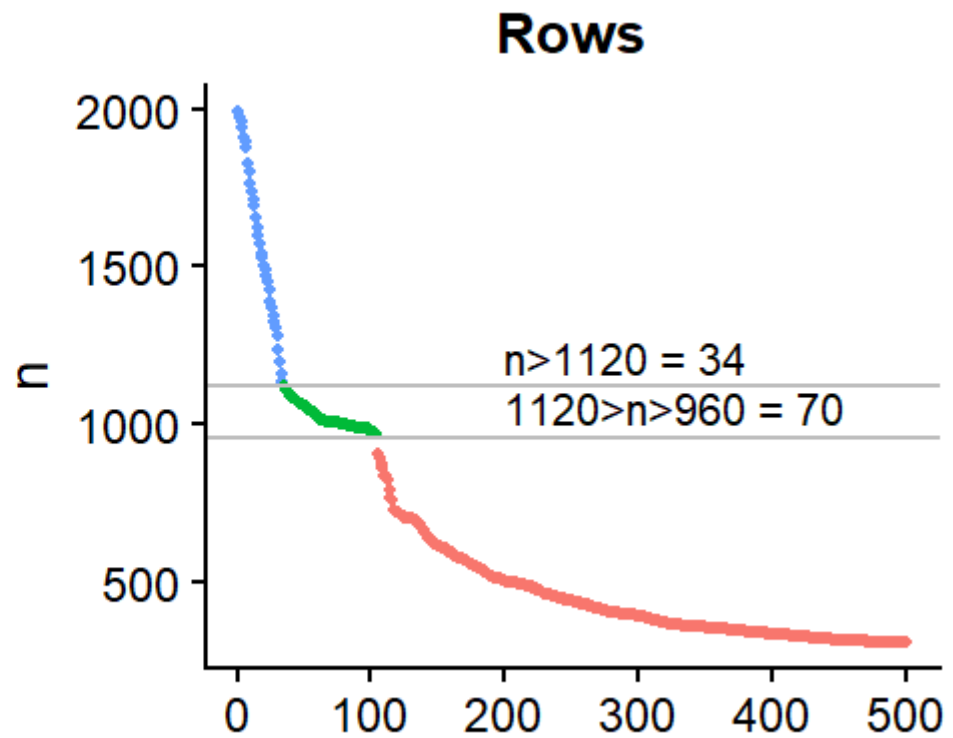
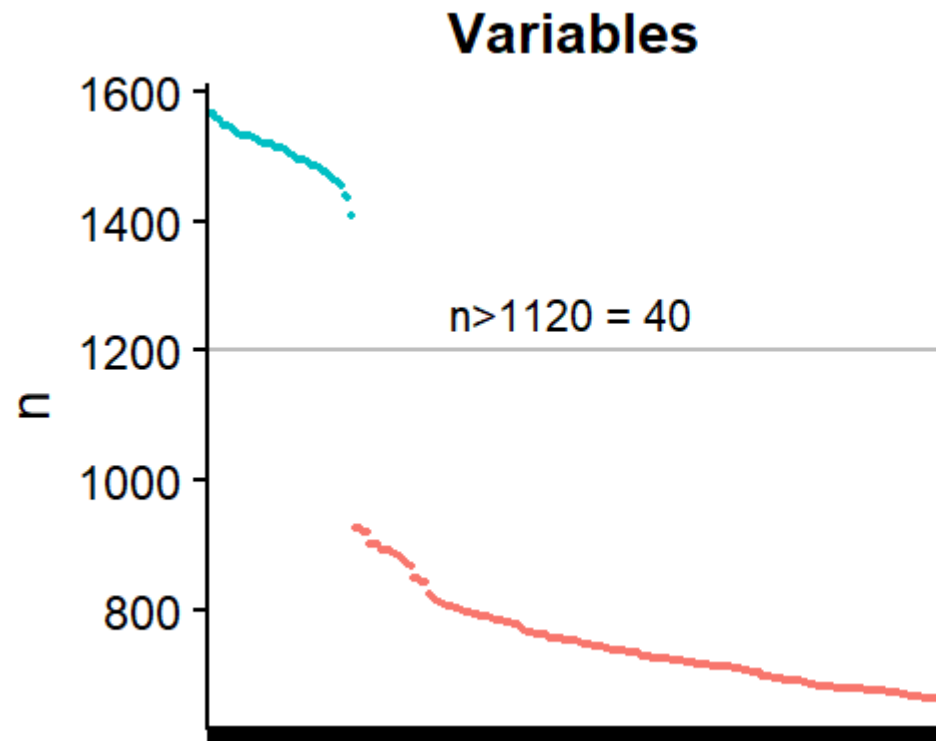
Distribution of Target - Weibull

- No distribution fits well until we realize NAs are encoded as zero.
- The *shape* parameter of less than unity for the target means the probability that the customer will respond **increases** with time - **they are motivated**.
- All variables show a similar distribution.

```
## Fitting of the distribution ' weibull ' by maximum likelihood
## Parameters:
##           estimate Std. Error
## shape 6.775485e-01          NA
## scale 4.570632e+06          NA
```



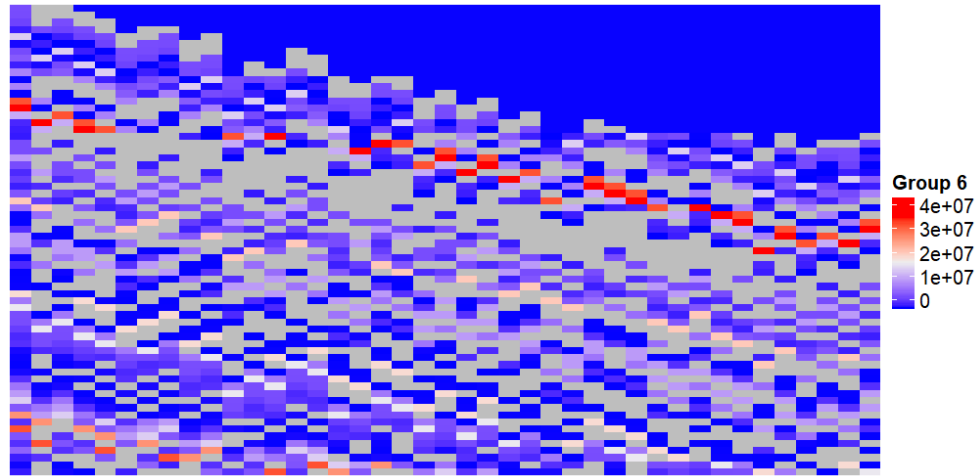
Sort Variables and Observations by n(non-NA)



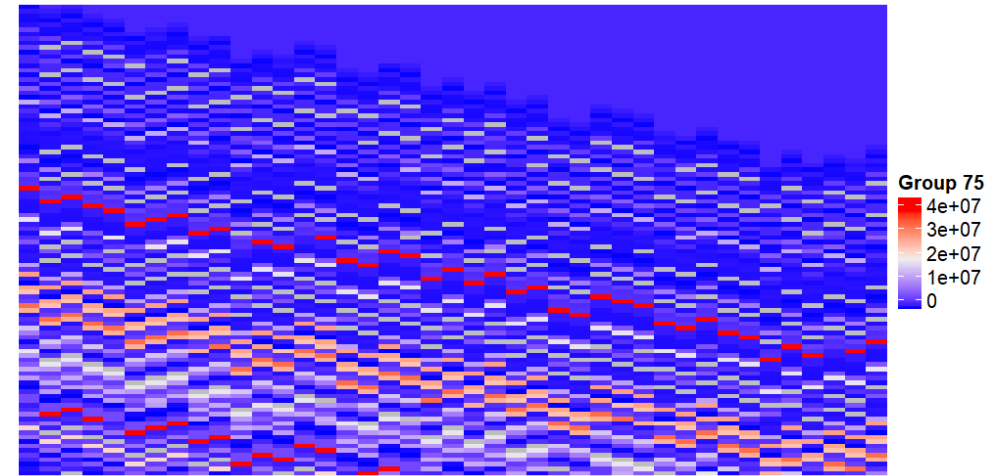
There are subgroups of data with high counts of non-NAs in both observations and variables. These form **Data Bricks**.

Data Bricks Have Internal Structure

User 6 Campaign 0 Brick with 66 Observations



User 75 campaign 0 Brick with 104 Observations

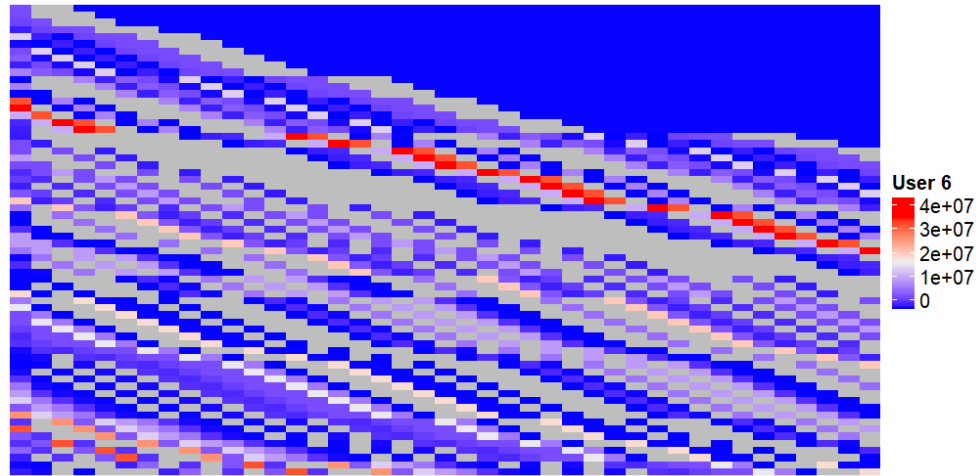


Interesting patterns are emerging

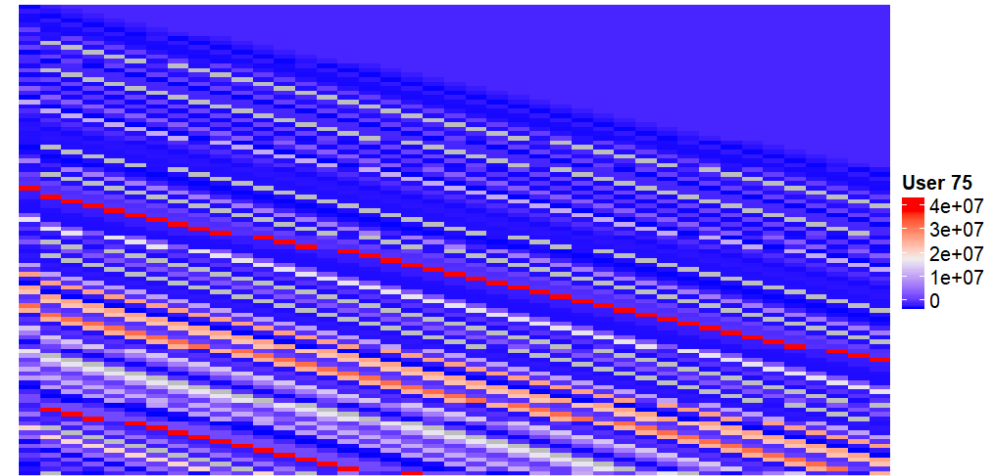
- Adjacent rows values are **exactly** the same, but shifted sideways
- Additional columns are found by brute force searching with a range of lags
- Additional rows are also found by brute force in a similar manner
- Short list candidates are drawn from columns and rows with a large number of order independent matches

Sort the Columns Within Data Bricks

User 6 Campaign 0 Brick with 66 Observations



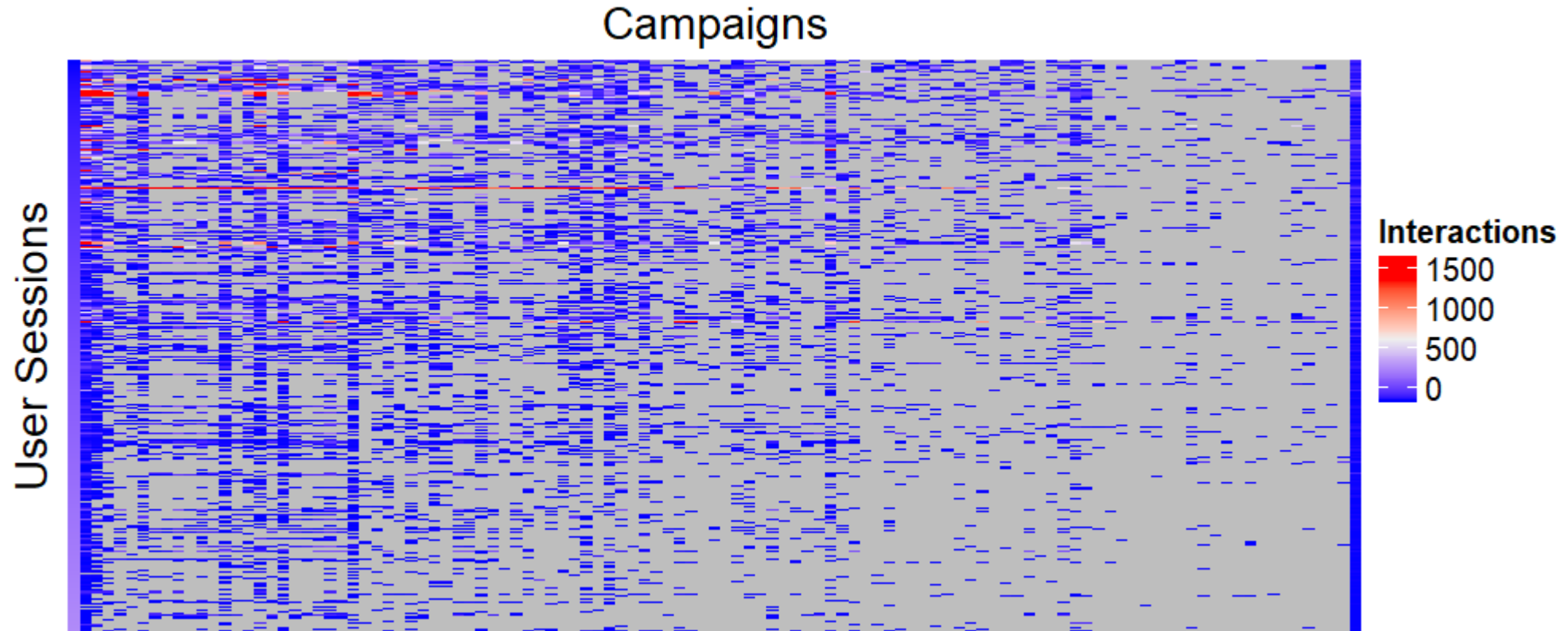
User 75 campaign 0 Brick with 104 Observations



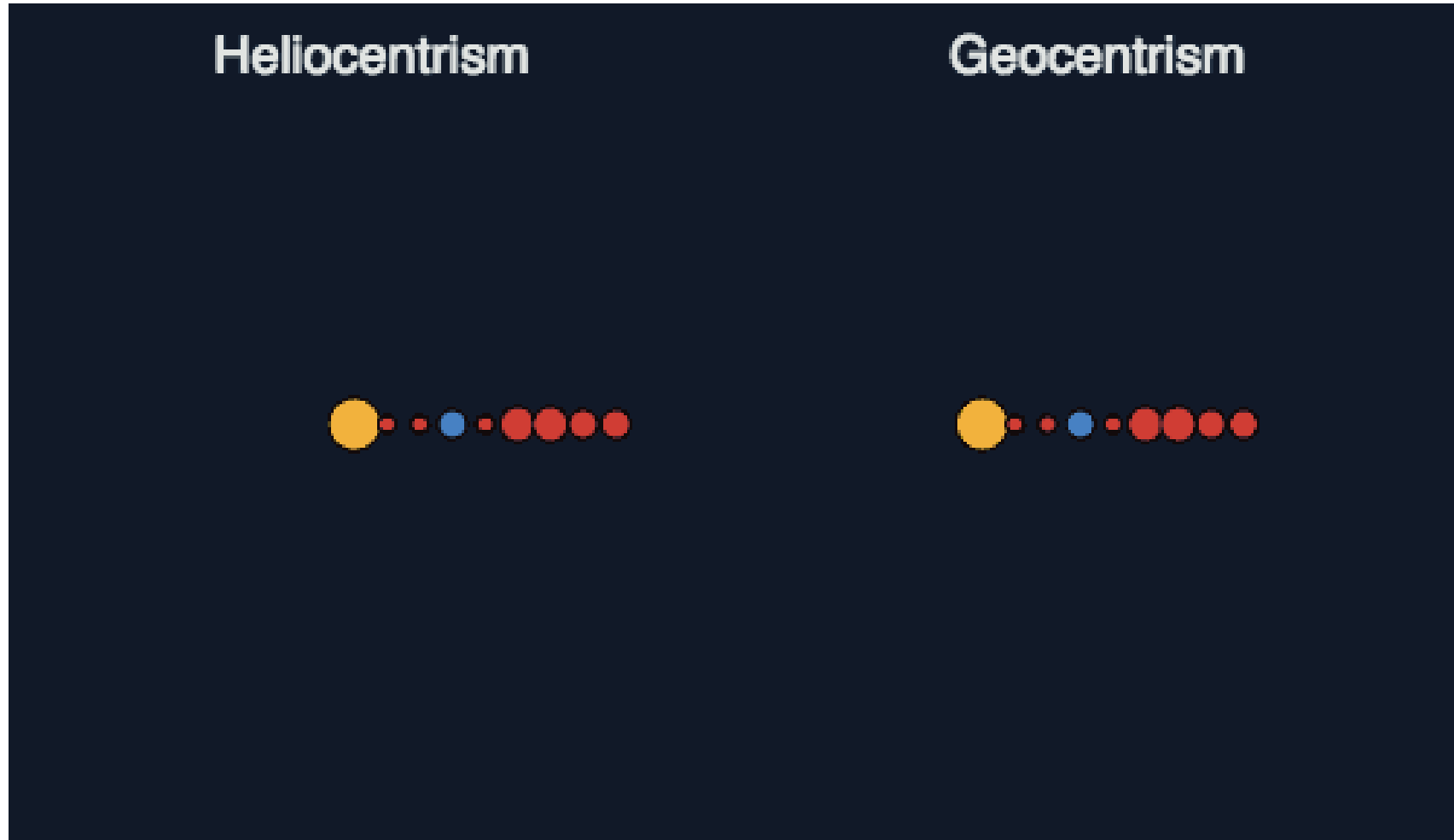
Row to row matches with single and double lags are emerging

- Data bricks are 40 variables wide and come in a range of heights
- Groups of 40 variables are **campaigns**
- Groups of observations are customer **user sessions**

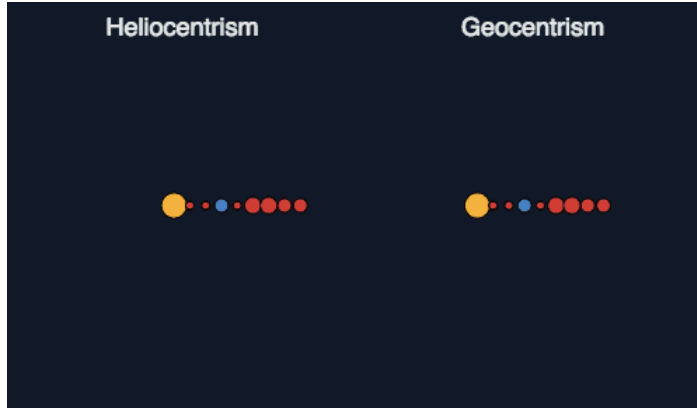
Data Bricks Form a Wall of Data



Which Problem Would You Prefer to Solve?



Which Problem is Better to Solve?

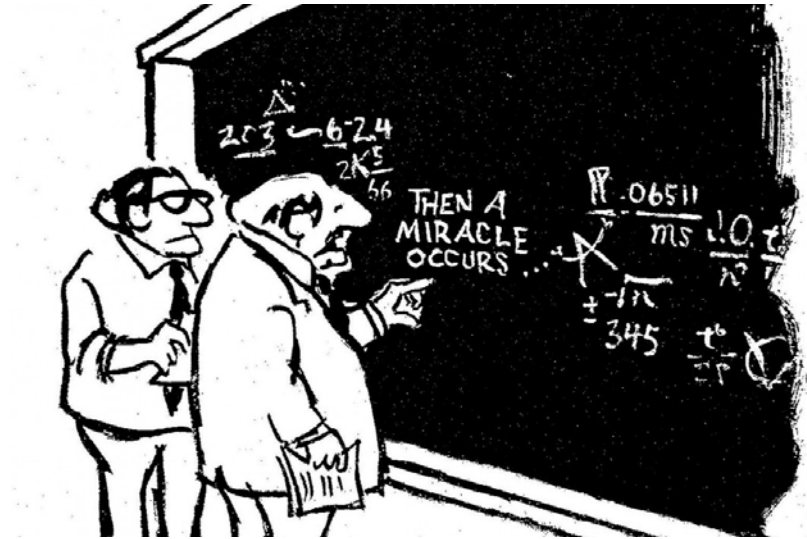


Now doubt that with enough data, both these solutions would have arithmetically acceptable solutions, that is, low RMSE.

... but, even in 1543, Nicolaus Copernicus' knew how to make a model easier to parameterise and interpret. The heliocentric model:

- has far fewer parameters than the geocentric model
- requires less data to define and converges faster during refinement (training)
- is **more interpretable**
- shows enough clarity that others can make profound insight into the underlying physical process
 - Galileo in 1632
 - Sir Isaac Newton 1687
- could get your book banned by The Church and have you placed under house arrest

Define the Features of the Problem with Domain Experts



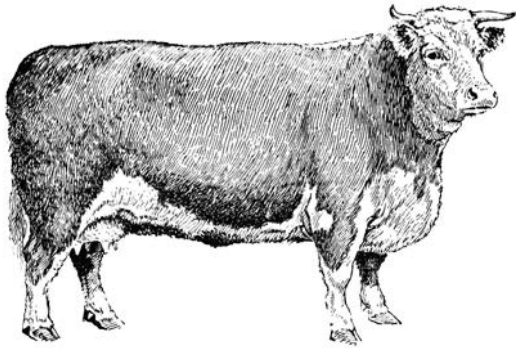
The overall solution algorithm is a multistep process and includes presenting the final optimization function with suitable data.

The process for solving these types of problems often requires a series of function after function. As we define each function, we gain more insight.

I think we need a bit more detail in this step

Can We Use Distribution Characteristics as Features?

Andrew Ng Says So



Applied Machine
Learning

It's Feature Engineering

O RLY?

Alastair Muir

[1] *Deep Learning* - Stanford [🔗](#)



Coming up with features is difficult, time-consuming, and requires expert knowledge. **Applied machine learning** is basically feature engineering

Andrew Ng

Why Features Should Make Sense

- Less "black box model" resistance from management
- Easier to assess model risk and bias
- You can build in observational behaviours that are logical and testable
- Fewer variables produces a model that is more general and robust
- Fewer variables produces a model that converges faster



42? Is that all you have after four and a half million years?

I think the problem is that you never actually know what the question is. You have to know what the question actually is, in order to know what the answer means.

Well, can you please tell us the question?

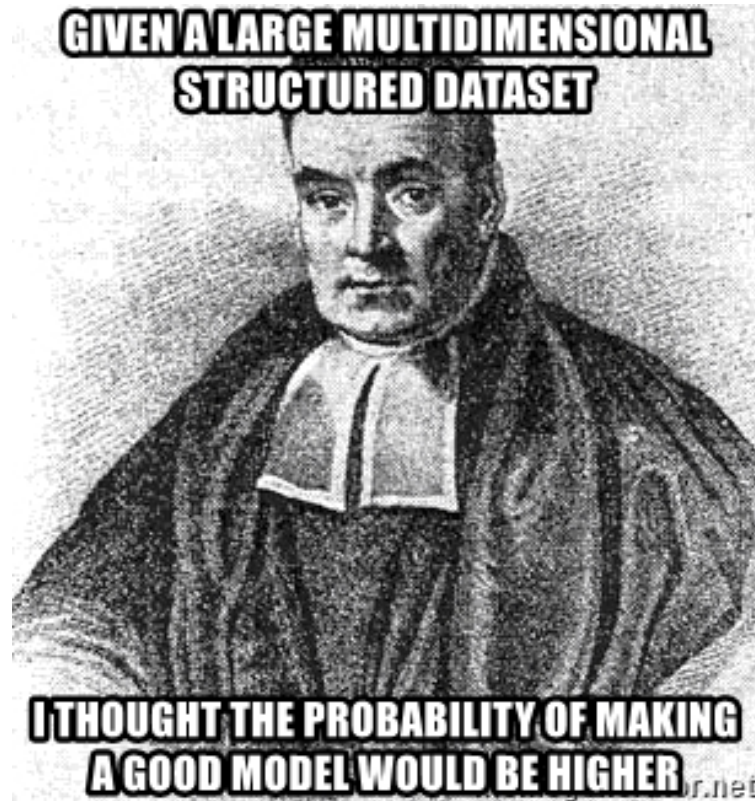
... tricky

Now We Know Where We Are Going



- The target value is the $n+2$ event of the user history in the most recent campaign
- The row and column data are scrambled in time order and can be sorted
- This data is a collection of histories of customer interactions within campaigns
- The data values are coded times until next interaction
- Data are reported to a maximum of 40 recent customer interactions per campaign
- Zeros were identified as NAs within campaign histories
- Extrapolating past history to the present explains the "leaks"
- The "leaks" can be used to augment the training set
- ~5,000 variables are a collection of about 125 40-value maximum data collection events

Finally! Build, Train, and Test Some Models



The target value is the time taken on the final interaction with a customer for the most recent campaign

- We have the times for the last 40 interactions with the customers
- This dataset is a collection of interaction histories for the same customers on more than 100 campaigns
- We have engineered a number of features to describe the probability distributions to measure the nature of the customer interactions in the most recent campaign
- Features include; number of interactions, minimum, maximum, skewness, kurtosis, interaction count for each campaign, the most recent campaign, and the user's "average" behaviour.
- We have constructed a customer profile for the "usual" behaviour for these metrics for all past campaigns.

The Final(ish) Model(s)

XGBoost

- 10 fold cross validation
- bootstrap aggregating twice
- 206 features
 - ntrees=100, num_threads=24, booster="gbtree", eta=0.01, min_child_weight=4, depth=5, subsample=0.50, colsample_bytree=0.50, colsample_bylevel=1.0
- $4,476 + 7,837 = 12,313$ observations in the training set
- 837 seconds, dual quad processor, 3.6GHz
- **RMSE = 1.306 -> 0.51789(with leak)**

h2o.io

- Stacked model
 - GBM(Gradient Boosting Machine)
 - DeepLearning grid
 - XRT(Extremely Randomized Trees)
 - DRF(Distributed Random Forest)
 - GLM(Generalized Linear Model) grid
- 419 features
- 12,313 training set in the training set
- 3,600 seconds, dual quad processor, 3.6GHz
- RMSE = 1.32 -> 0.531(est)

TensorFlow and Tensorboard

- 2 dense layers with dropout
- 2 LSTM layers with dropout
- Similar results to above - not pursued

Leaderboard

Public Leaderboard

Private Leaderboard

The private leaderboard is calculated with approximately 51% of the test data.
This competition has completed. This leaderboard reflects the final standings.

In the money

Gold

Silver

Bronze

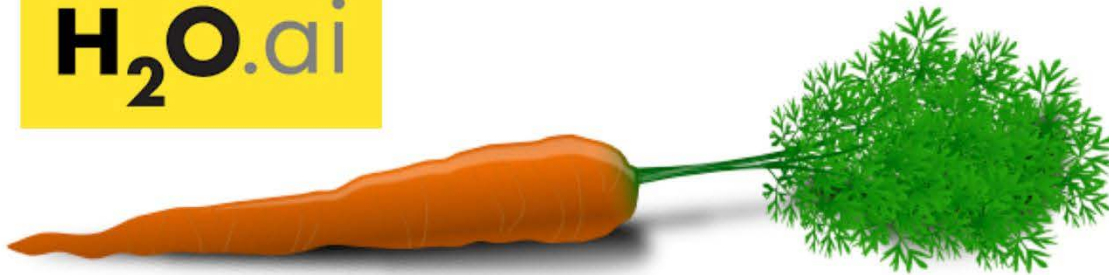
#	Δ pub	Team Name	Kernel	Team Members	Score
1	\uparrow 8	ML-eak			0.51980
2	\uparrow 50	adilism			0.52383
3	\uparrow 36	anatoly			0.52430
4	\uparrow 17	chenhan zhang			0.52496
5	\uparrow 98	Vladimir Larin [ods.ai]			0.52567

21 submissions for [Alastair](#) Sort by

All Successful Selected

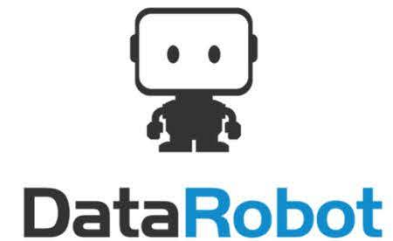
Submission and Description	Private Score	Public Score
xgb1_plus_LEAK.csv 6 months ago by Scrooge 10 fold XGBoost boost with leak	0.51789	0.49013
my_xgb1_plus_leak_20180312a.csv 3 days ago by Scrooge 10 fold XGBoost twice bagged with 4 more leaks	0.51980	0.49651
my_xgb1_plus_leak_20180312.csv 3 days ago by Scrooge 10 fold XGboost bagged twice with four more leak values	0.51985	0.49602
my_xgb_plus_leak_20190308.csv 7 days ago by Scrooge XGBoost 10 fold twice bagged 7837 leaks	1.20009	1.23404
submission20180717_2.csv 8 months ago by Scrooge Trimmed variables and observations	1.36150	1.42075

Tools of the Trade



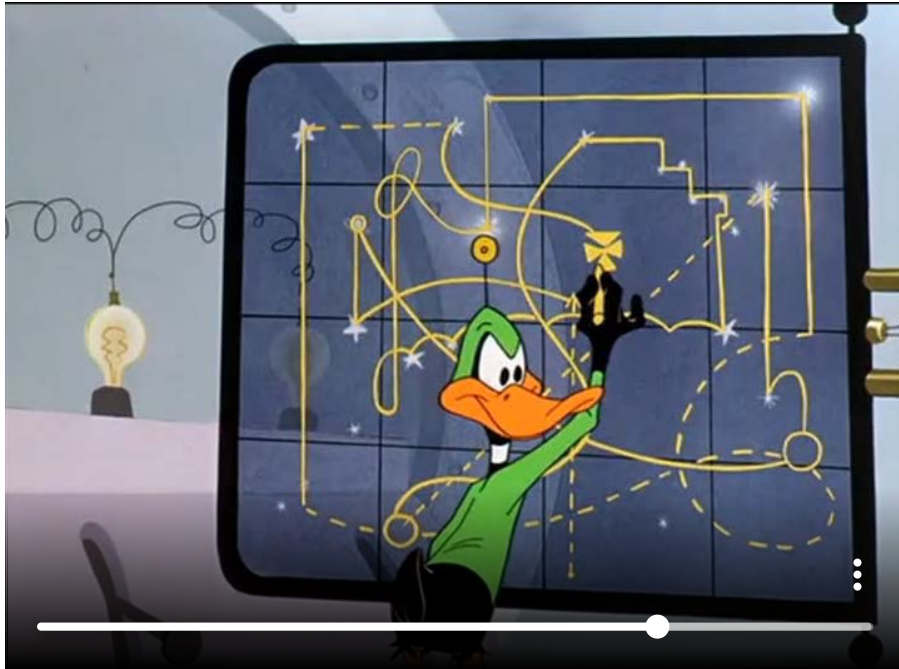
Caret package in R

data.table
fitdistrplus
ComplexHeatmap
Xaringan
xgboost
mclust



tl;dr

How the project went



[1] Too long; didn't read

How projects should go

- Talk to the stakeholders about the data and the process that generated them. The goal is to **understand the problem**.
- Get really good at `data.table`, `cdata`, `tidyverse`, and `ggplot2`.
- Check every assumption (zeros, missing data scenarios).
- Construct features that mean something, not just ones that work.
- Models with smaller number of variables are more stable, generalizable, and **explainable**.
- Make a simple baseline.
- Construct a full pipeline for trying out ideas (eg, autoML from **h2o.io**, keras with **TensorFlow**).
- Graph everything - look for patterns.
- No free lunch theorem, don't just use XGBoost or stacked models.
- **Call me**

Artificial Intelligence-Machine Learning

Presentation to Management and Investors

Skill testing question - Name the movie



Thank you

Alastair Kerr Muir

AlastairKerrMuir@gmail.com 

www.linkedin.com/in/alastairkerrmuir 

*This analysis, presentation and graphs
were produced in using **R**, a programming language
and software environment for statistical computing,
and the **RMarkdown** and the **Xaringan** packages.*

