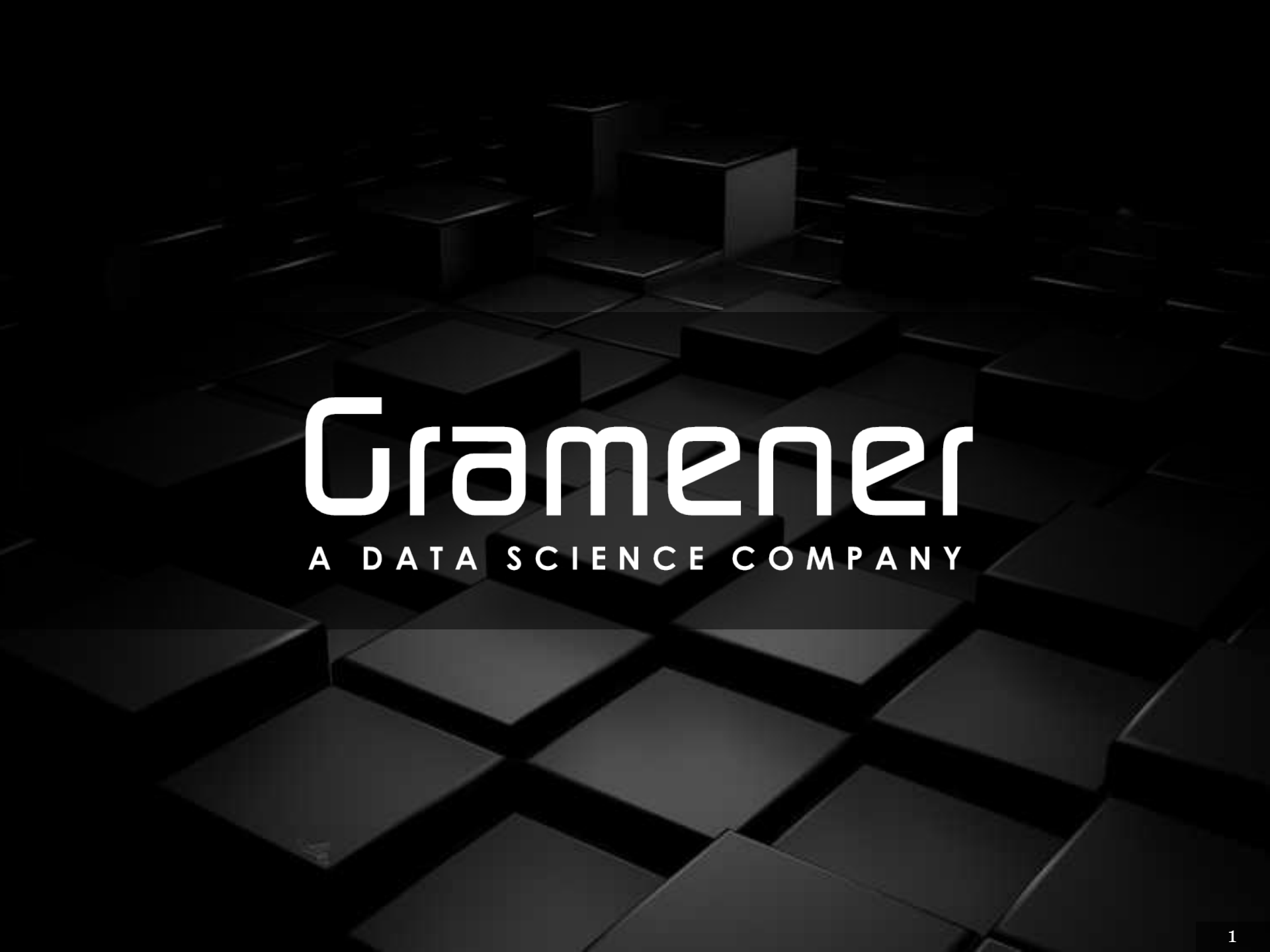

Gramener
A DATA SCIENCE COMPANY

Gramener: Our Offerings

Consulting

Data Consulting

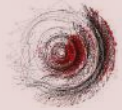
- Data Science Practice Setup
- Strategic Data initiatives roadmap

Design Consulting

- UI-UX Consulting
- Information Design Solution

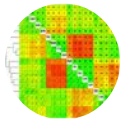
Services

Analytics



Advanced Analytics & Machine Learning

Visualization



Purpose-built Visual Intelligence Apps

Design

Data Infographics & Data Stories

Specialized Trainings: Analytics – Visualization - Design

Products

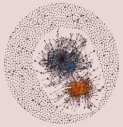
Autolysis

Automated analysis with no bias

Data Explorer

Exploratory Visual drill-down

Link Analyzer



Visual clustering of unstructured networks

Mapping App

Geo insights on custom maps

Platform

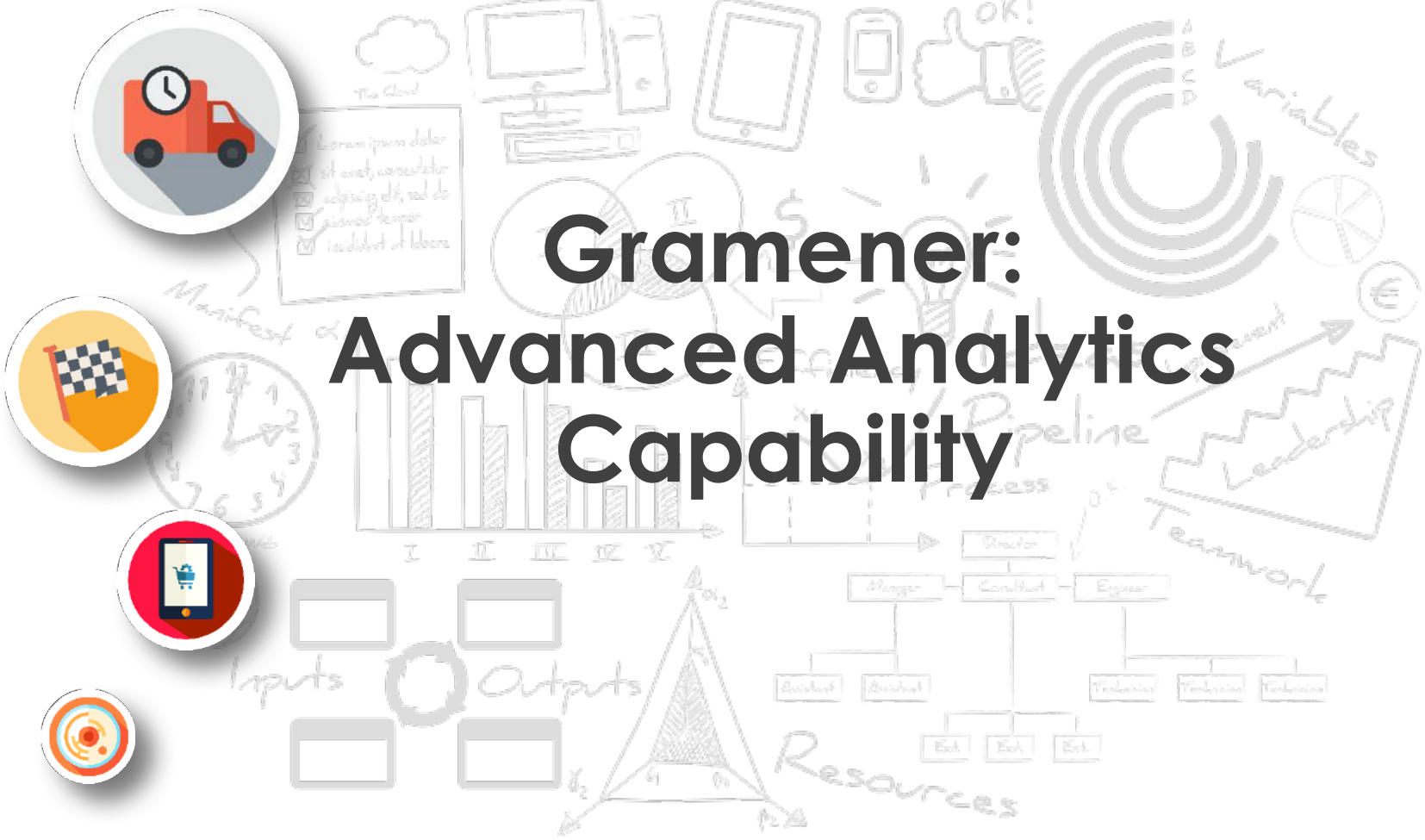
Gramex - A Visual Intelligence Platform

UI Components Library

Visualization Library

Data Handlers Library

Analytics Library



Gramener: Advanced Analytics Capability

1. Advanced Analytics:

- Gramener's Key Differentiators
- Industry-wise Usecases

Gramener's **Key Differentiators** in Analytics

We help define business problems in a **Consultative** mode

We leverage our **Analytics Toolsets** & **Service Frameworks** to provide rapid Insights

We deliver **Recommendations** thru' an interactive **Visual front-end**

Autolysis Tool
Data Exploratory Tool
Network Analytics Tool
Clustering Tool

Analytics is fast-tracked *with our Proprietary Analytical Packages & Toolsets, combined with common tools like R, Python, SAS, Excel*

Our Analytics & Machine learning suite

- Pairwise correlation
- Hierarchical clustering
- Segmented averages
- Comparison of means
- ANOVA
- PCA (Factor analysis)
- Spectral analysis
- Time Series Forecasting
- Multivariate regression
- Discriminant analysis
- Bayesian classification
- Non Linear Programming
- Support vector machines
- Decision trees
- Neural networks
- Markov chains

Industry-wise Use Cases in Analytics

Retail/Consumer Goods

- Merchandising and Market Basket Analysis
- Campaign Management and Loyalty Programs
- Demand Forecasting & Inventory Optimization
- Customer Segmentation

Banking & Financial Services

- Credit Risk Scoring and Security Analytics
- Trade Surveillance
- Fraud/Risk Modeling
- Abnormal claim pattern analysis
- Cross-sell propensity analysis

Media

- Market Mix Modelling
- Ad campaign targeting
- Impressions forecasting
- Show Content analysis
- Inventory forecasting and optimization

Pharma & Life Sciences

- Clinical Trial Analysis
- Disease Pattern Analysis
- Patient Care and Program Quality Analysis
- Supply Chain Planning
- Drug Research and Discovery Analysis

Telecommunications

- Price Optimization
- Customer Churn Prevention
- Network Performance & Optimization
- Market Segmentation
- User Location Analysis

E-commerce & Customer Service

- Clickstream Analytics
- Collaborative Filtering (Recommendation Engine)
- Next best offer/action/product
- Resource Optimization

2. Advanced Analytics Case Studies

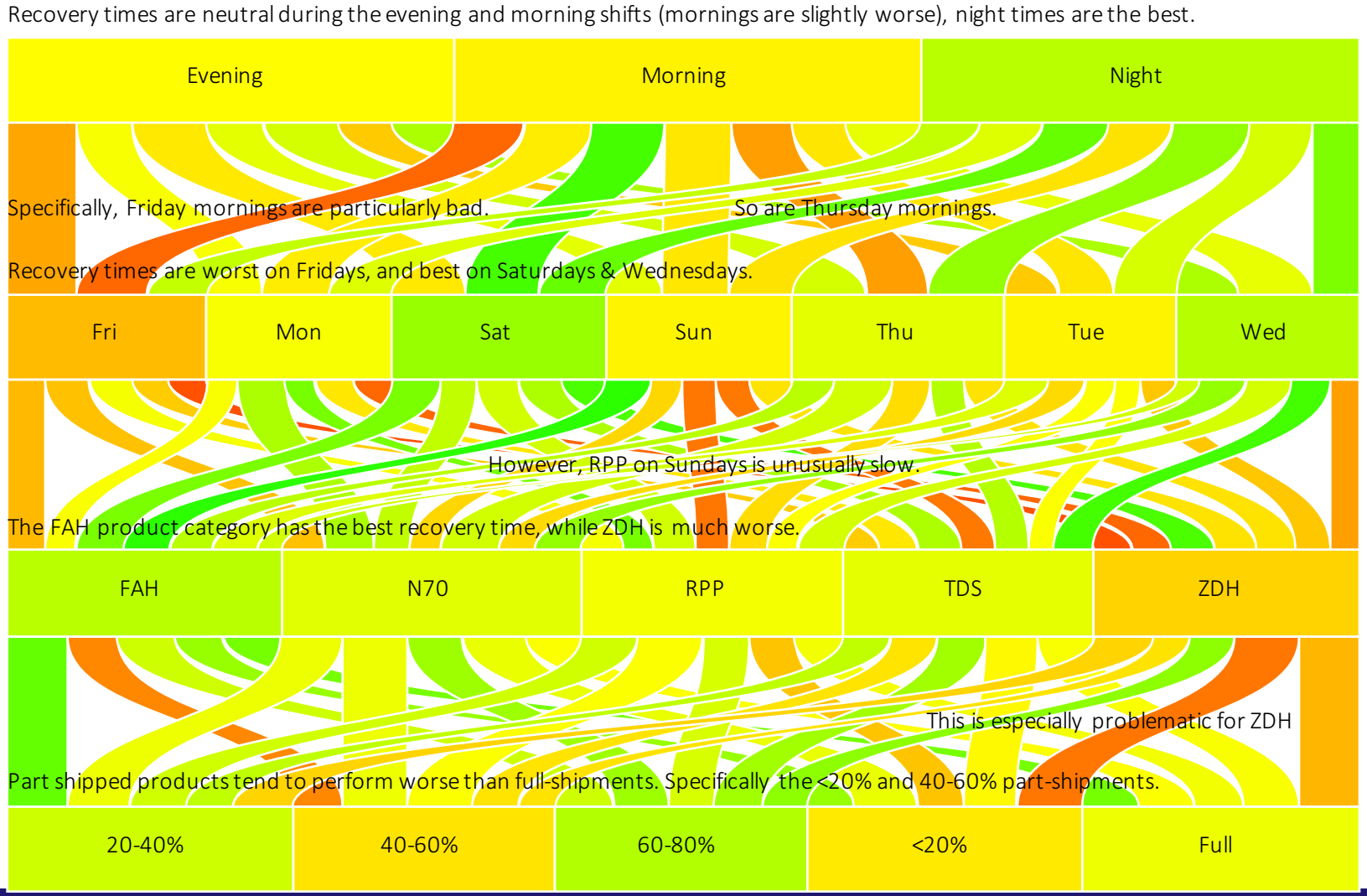
CARGO DELAY

This visualisation measures the recovery time (time from arrival of the flight until delivery), and identifies which factors most influence the recovery time.

This visualisation is part of a suite of analytical techniques we call "grouped means" that allows us to measure the impact of every parameter (shifts, weekdays, etc.) on any measure of interest – recovery time in this case, but this could be extended to revenue, operational efficiency, or ability to cross-sell.

It allows automatically detection of statistically significant flows and highlights only relevant ones to users.

The system therefore analyses all possible patterns, but users only see the insights that matter.



Telecommunication

“ Churn of customers is a particularly severe problem in the telecom industry.

The challenge is to identify the propensity of churn upto a month in advance, even before a customer moves out, so that proactive interventions can begin”

Churn Prediction for a Telecom Operator

Background & Objective

Customer churn is a well noted problem in telecom industry today. One of the leading telecom operator in the country wanted to predict the churn of customers a full month before the customer actually moved out. Gramener executed this engagement as a Predictive modeling engagement to build models that identify customers with reasonable accuracy, so that proactive retention interventions could be taken up.

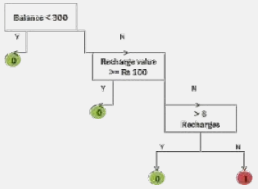
Gramener Approach

Exploratory Analysis & influencers

Exploratory business analysis performed to identify influencers & create additional derived metrics & derived dimensions

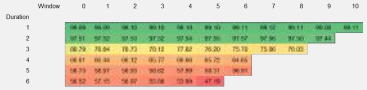
Linear Discriminant Parameters

Using selective metrics, models were built on Linear Classification like Decision trees, Linear Discriminant Parameters



Non – Linear Prediction Models

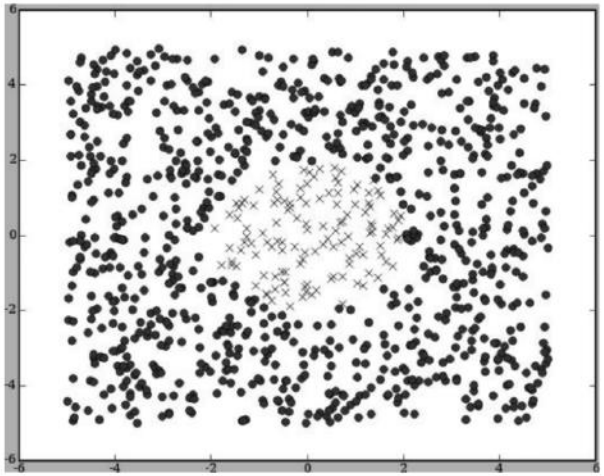
With refined set of metrics non-linear families of models were built: Neural Networks, Random Forests & Support Vector Machines

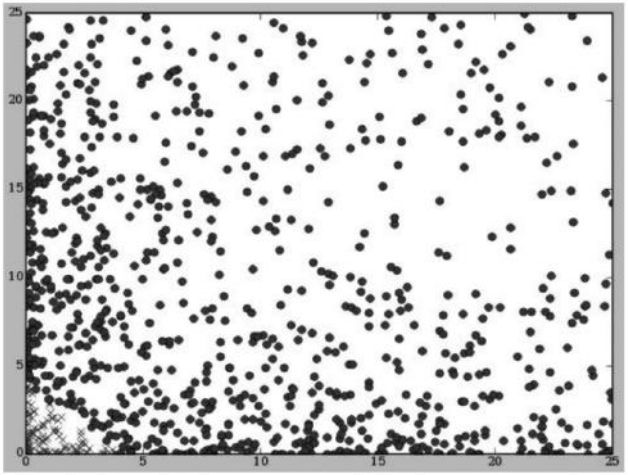


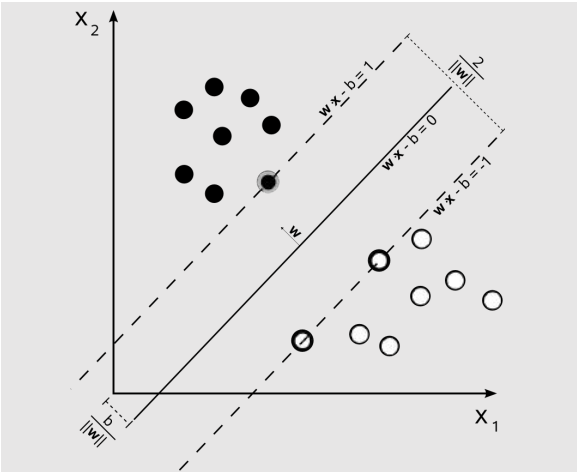
Predictive Intervention to arrest Churn

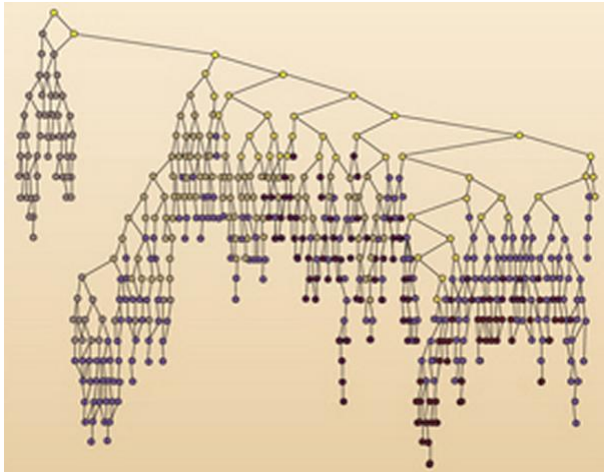
The best model was implemented & compared with a control set. Targeted promotions for predicted set yielded ~60% drop in churn

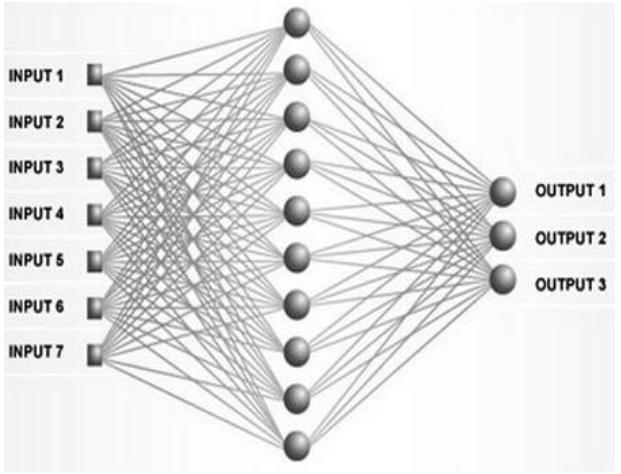
Actual	Prediction: No churn	Prediction: Churn
No churn	OK	WASTED Marketing cost Rs 40
Churn	MISSED Acquisition cost Rs 80	OK











MODELS	COST PER CUST.	IMPROVEMENT	MISSED	WASTED
Random Forest/ SVM/ Neural Nets	~2.0-3.0	~40-50%	~1-2%	~2-3%

Corporate Decisions

“Understanding stakeholder behavior is a major challenge for companies.

Intelligence in this area is essential for companies, in order to understand and possibly predict the patterns of voting by their shareholders”

Predictive modelling of stakeholder behavior

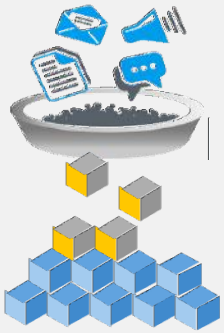
Background & Objective

A US based business conglomerate welcomes their shareholders to participate in corporate decision making. This is achieved by shareholders voting to let know of their opinion on various issues. This business conglomerate would like to use past data and predict the voting behavior of these shareholders.

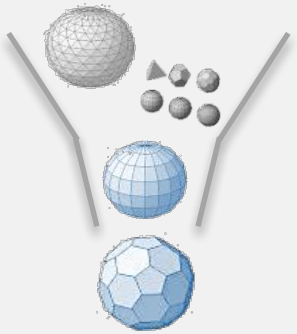
Gramener Approach

Consolidating business recognized variables & data in order to establish significance

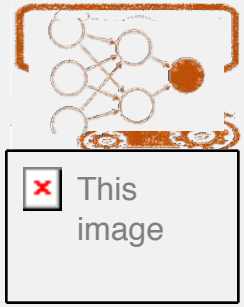
Supplementing extrinsic variables as additional influencers helps enrich the potential predictors



Dimensionality reduction through statsitical techniques by prioritizing impactful variables



Data Modelling & Choice of Algorithms to predict voting behavior & enable proactive intervention



Marketing & Media Planning

“What do I do to increase my audience viewership time”

“ What price point should I offer to my channel advertisers”

“ What should be the ideal media spend on my brand”

Impressions forecasting for a TV audience measurement agency

Client Background & Challenge

- The client was interested to project TV Audience views of advertisements in the Hair Care category which could be used to plan optimized slotting of ads by Advertisers across channels of different genres and across various time bands of the day
- Client had provided last 4 years of data at an ad spot level for all ads in the Hair care category (eg. Oils, shampoo). Gramener was asked to build models to accurately forecast ad impressions for selected channels

Gramener Approach

Understand distribution and consistency of data

- *Data provided was at an individual ad spot level, containing impressions gathered, ad duration, description, etc.*
- *The initial steps involved identifying channels and timebands with high consistency in terms of presence of ads at a weekly level, in order to facilitate accurate forecasting.*

Identify potential derived metrics and engineered features

- *Several external datasets were mined to include additional information related to event presence, holidays, program performance, season etc.*
- *Thru Text mining a list of key celebrities appearing in the ad was identified whose impact on impressions was measured.*

Test causal and time series based forecasting methods

- *A number of different models, both causal and time series based, were evaluated for forecasting*
- *Causal models include linear regression and regressional ARIMA.*
- *Time series models include moving averages, exponential smoothing, neural networks, and ensemble model collections.*

Identified best performing models

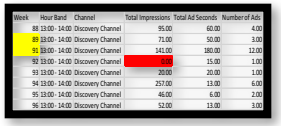
- *Best fit models for each channel - timeband were chosen based on performance of accuracy metrics (MAPE, adjusted MAPE, Rsquare) against training and test data.*
- *Forecast projects for using best-fit model were compiled, and presented in both aggregated and decomposed forms to the client*

Outcomes

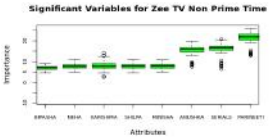
- Forecasting of impressions was performed for 8 channels across 8 different genres for 2 time-bands, one from Prime Time(20:00 – 21:00) , one from Non Prime Time(13:00 – 14:00)
- Channels chosen: B4U Music, Discovery Channel, India TV, Movies OK, MTV, Pogo TV, TLC, Zee TV.
- Neural networks, exponential smoothing over state space models were found to have best forecast performance
- The top 6 channels by model accuracy were then selected for the final aggregation and aggregated and decomposed forecasts were shared
- Overall accuracy values: 84.20% for training sample, 67.48% for testing sample.

Missing Value Treatment, Variable Selection and Statistical Modeling Output

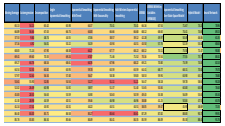
Missing Value & Outlier Analysis



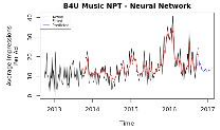
Significant Variable Selection



Causal & Time Series Model Selection



Forecasts & Plots



Additional Insights

Top Advertisers

The top 4 Hair Care advertisers - HUL , Marico, L'Oréal, and P&G account for 56% of all ads

Top Genres

94% of all Hair Care ads are placed in the GEC, Movies, News, or Music Channel Genres.

Top Brands

The top 6 brands by average impressions per ad on TV are all for hair oil products.

Advertiser Level Clustering based on buying behavior for a leading media company

Client Background & Challenge

- A leading Indian Media Conglomerate comprising 58 channels wanted to increase revenue through sales of Ad slots by analyzing advertiser patterns across the country by offering personalized deals
- Objective was to accurately classify advertisers as per their buying behavior across all channels irrespective of size, category, region or time of the year

Gramener Approach

Understand Data granularity and buying behavior of advertisers

- *The Viewership metrics, buying metrics and revenue metrics per advertiser for FY16 was used for analysis*
- *Performed Descriptive analysis for advertisers across different indictors to understand buying behavior across different channels as well as industry sectors*

Scoping metrics and advertisers

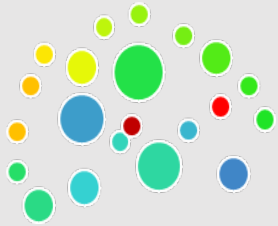
- *Missing Value, Bivariate & Correlation analysis was performed on the data to understand anomalies and relationship of variables*
- *Business inputs and data consistency checks were used to scope variables and advertisers for clustering*

Customer Segmentation Analysis

- *Unsupervised machine learning algorithms such as K-means and hierarchical clustering were implemented to classify advertisers*
- *Models with best classification metrics and accuracies was shortlisted to generate optimal clusters*

Visualizing the clusters

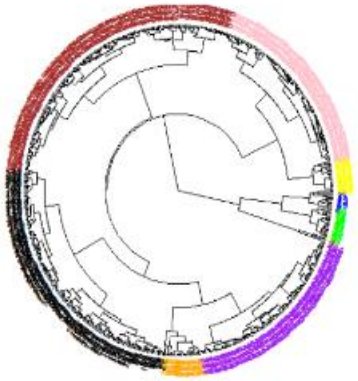
- *The clusters were visualized and profiled based on their attributes so as to group an optimal mix of advertisers*



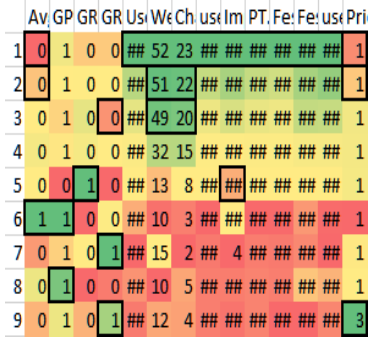
Outcomes

- Advertisers who typically get high regional GRPs tend to buy less number of channels – lower overall presence
- Slots during Impact shows and Prime Time shows are extensively bought during the festive season which contribute to a strong % of revenue (~85% +)
- To deliver targeted offers to media company's clients, advertisers were optimally divided into 9 clusters
- Specifically, three metrics were used to define the characteristics of each cluster : **Buying metrics**(Weeks bought, Channels bought, total seconds bought) to categorise them into large, medium, small & tiny clients, **Viewership metrics**(Avg. TVR, total GRPS,GRPs per genre) & **Revenue metrics**(price points, total revenue, festive revenue etc.)

Clustering & Profiling



Cluster customers to minimize overlap



Profile clusters to interpret advertiser characteristics

Geographical Clustering of India Districts for a media conglomerate

Client Background & Challenge

- One of the leading media company wanted to create segments of all districts in India based on their underlying similarities so that they can effectively build & broadcast TV shows accordingly.
- Final objective was to create an interactive web interface which will allow user to select the relevant attributes of interest and cluster parameters to obtain segmentations of the regions.

Gramener Approach

Collect, understand and clean the data

- *Collected district level data from different censuses like Household Census, Socio-Economic Caste Census, Religious Census and from District of India website.*
- *Cleaned and pre-processed the data to create a usable set of columns for the analysis*

Creation of a Platform for clustering

- *Created a customized tool to apply different types of clustering algorithms (like K-means, HDBSCAN etc). and visualize the district clusters in India Map.*
- *Link to the tool* https://gramener.com/cluster/

Try different segments and algorithm

- *Business relevant segments like 'Urban Districts', 'Rural Districts', 'All districts' were iterated and used to get relevant clusters for client*
- *Introduced development chain in the platform: Segregating districts based on Education, Skills and Wealth using decision tree type clustering.*

Compile visualized key insights.

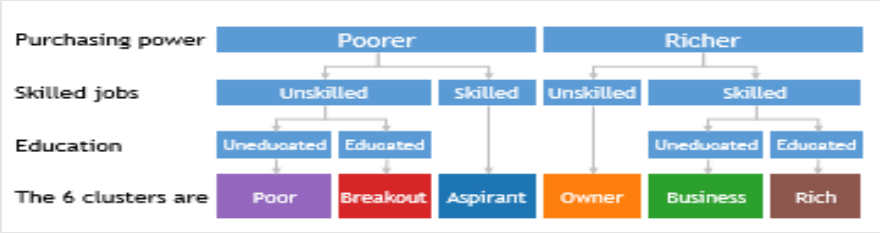
- *Valuable insights from the clustering exercise thru profiling of different segments was presented through a detailed presentation*
- *The Clustering platform developed for the engagement was also showcased with interactivity and visualizations*

Outcomes

- Client treated the Hindi-speaking markets as a homogenous entity from a channel content perspective
- To deliver targeted content, segmentation of the geographical layout of India was done into 6 clusters based on their demographic behaviour.
- Specifically, three composite indices were created based on the economic development lifecycle:

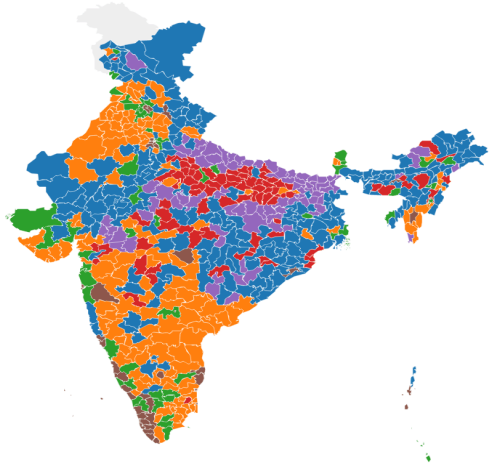
Education (literacy, higher education) that leads to...**Skilled jobs** (in mfg or services) that leads to...**Purchasing power** (higher income, asset ownership)

- Districts were divided (at the average cut-off) by:



- Offering targeted content to these clusters will reach a more homogenous demographic population leading to higher penetration and impact of the channels of the media company

Geographical Clustering - Visuals



Cluster	Aspirants	Landlords	Businessmen	Breaking out	Poor	Rich
Size	212	153	60	81	84	50
Total population	1,539,434	2,030,596	2,440,379	1,767,330	1,952,765	2,404,104
People per household	3.58	3.56	3.35	4.28	4.24	3.16
Rural %	81.5%	68.9%	47.2%	86.2%	88.3%	28.4%
Female %	48.6%	48.7%	48.0%	48.4%	48.3%	49.0%
Literacy %	60.0%	66.2%	71.2%	60.3%	47.7%	79.5%
SC+ST %	37.5%	29.9%	26.3%	34.2%	36.9%	17.6%
Workers %	42.9%	42.9%	39.9%	38.3%	39.6%	37.8%

Marketing Mix Modeling for a multinational CPG client

Client Background & Challenge

- The project was commissioned by the media planning team of the client.
- The intent was to understand key indicators driving sales and spends of the different marketing mediums to plan optimal spends for a food product in the Breakfast category
- Gramener was partnered to study the dynamic changing markets and external factors influencing Sales, using advanced modeling techniques to measure return of investments (ROI)

Gramener Approach

Business understanding & Data Collection

- *High level business metrics and financials were tracked using annual reports. Sales volatility based on seasons, Competitor performance were all assessed.*
- *Monthly Sales, Spends through each of the mediums, Survey Scores and GRPs were provided.*
- *Monthly Average Temperature, Consumer Price Index, Google Trends data was obtained as a part of data collection process*

AdStock optimization & Interaction Effects

- *Ad-stock was applied to all the marketing channels.*
- *Linear programming (LP) optimization algorithm was used to obtain weights maximizing the correlation of marketing spends with revenue*
- *Interaction Variables i.e combined effect variables of TV + Print , TV + Digital etc. were also formulated in addition to the main effects*

Statistical Modeling

- *Several Linear regression models were constructed with all combinations of variables.*
- *Models with lower MAPE, higher Adjusted – R Square and with coverage of all spend mediums were selected*
- *Models were refined further based on the business inputs and 2-3 scenarios were modeled for evaluation*

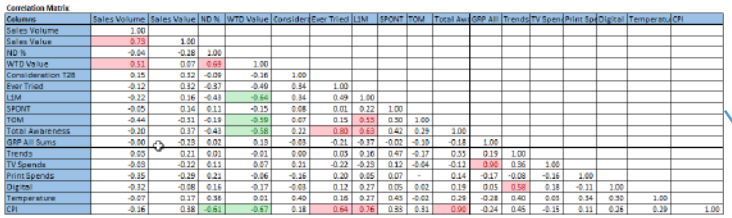
ROI quantification

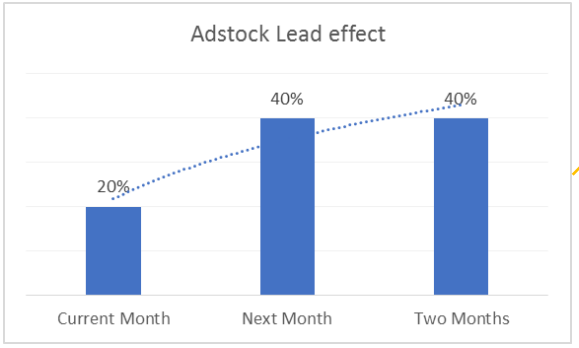
- *Revenue contribution from each of the medium(TV, Radio, Digital) were attributed based on the Model coefficient percentages*
- *ROI across mediums was quantified based on model outputs, Spend, Sales and profit %*

Outcomes

- Models with high accuracy levels (97%) were obtained for the Brand in the breakfast category.
- Modeling Scenarios included Spends, Media Interaction effects Seasonality factors, Competitor indicators and Socio Economic variables as significant contributing factors.
- ROIs were quantified for the mediums TV, Digital and Print.
- Total Sales of the Brand were attributed to Base, Incremental and External factors.
- Basis this engagement, client has better visibility in the returns being generated by their marketing vehicles and other drivers impacting the sales. Leveraging this analysis, optimal spend and effort allocation can be planned.

Correlations, AdStock, and ROIs





Scenario Planning

Scenario -1

Medium	ROI
TV	0.07
Digital	0.32
Print	0.21

Scenario -2

Medium	ROI
TV	0.04
Digital	0.17
Print	0.12

Capacity forecasting

> *Forecasting using advanced but uncomplicated algorithms improved resource consumption outlook by up to 90%*

Leveraging the improved visibility for planning and incident reduction will offer cascaded benefits

Demand & Capacity Management of Servers for a leading pharma Company

Client Background & Challenge

- The project was commissioned by Global IT Infrastructure & Services team of the client
- Future utilization of servers was the primary concern to plan capacities in advance
- Client was using a primitive method to forecast servers utilization with not so accurate results, led to less visibility into resource consumption and lower quality in planning and procurement of servers resources

Gramener Approach

Understand Data granularity and utilization of Servers

- *Utilizations of each server is available at a day level across 3 parameters: CPU,RAM and DISK*
- *Performed Descriptive analysis of utilizations across different category of servers to understand utilizations across all 3 parameters*

Forecast future utilizations

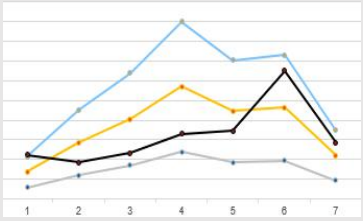
- *Treated utilizations in time series(TS) format and different TS forecast models right form simple moving averages to ARIMA and Holt-Winters were applied, accuracies were examined between each model*
- *Model with best accuracy was implemented to forecast utilizations*

Classify servers using Smart Assistance tags

- *Tags are created from the amalgamation of forecasted utilizations, existed trend component and derived slope values of each server*
- *These Tags help to identify those servers likely to breach threshold utilizations in future and help the user to act cognitively*

Bring the forecast to life on a Dashboard

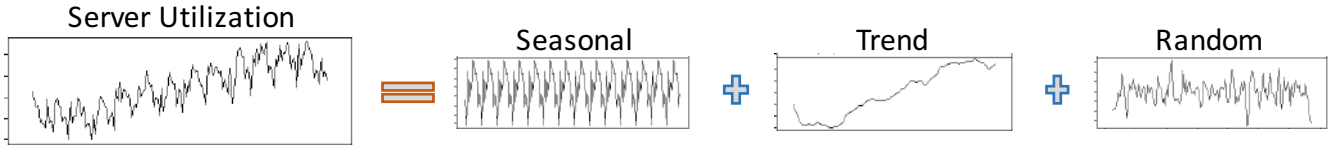
- *On a Dashboard Visualize the utilizations forecast and classification of high to low attention servers based on smart assistance tags*



Outcomes

- Both, week and month level utilizations were forecasted for 9000 servers
- All the servers utilization by CPU, RAM and DISK were forecasted for a week at an accuracy of 80%
- Gramener forecast was 90% more accurate on average than method that was used by client
- Servers classification by smart assistance tags had an accuracy of 83% and the monitoring team was able to quickly action on high, medium and low risk servers, with very high predictability

Server Utilization Decomposed as Time-series



Smart Assistance Tags

Tag: Critical

Servers Expected to have >80% utilization in coming 1 week

Tag: At-Risk

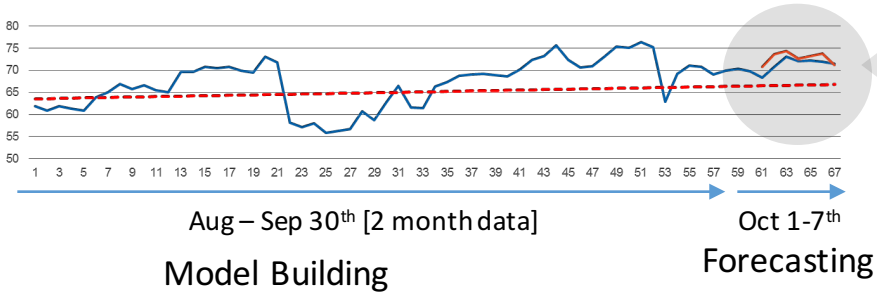
Servers close to threshold (60-80% Utilizations) & steady growth

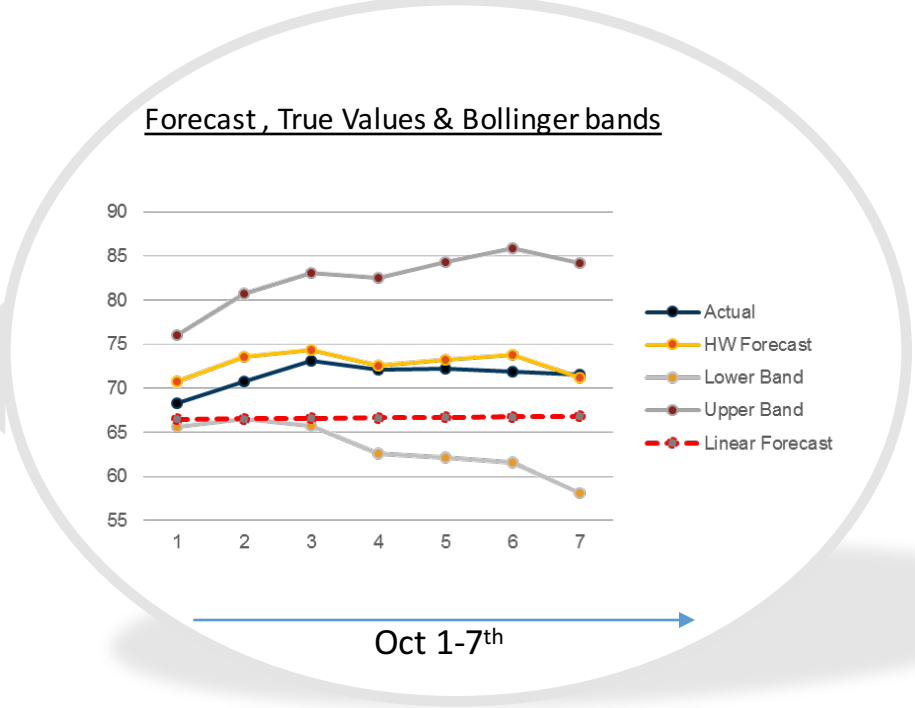
Tag: Green

Servers close to threshold & on downward trend

Takeaway: Client teams had better visibility in resource consumption and could plan for timely procurement to avoid resource shortfalls

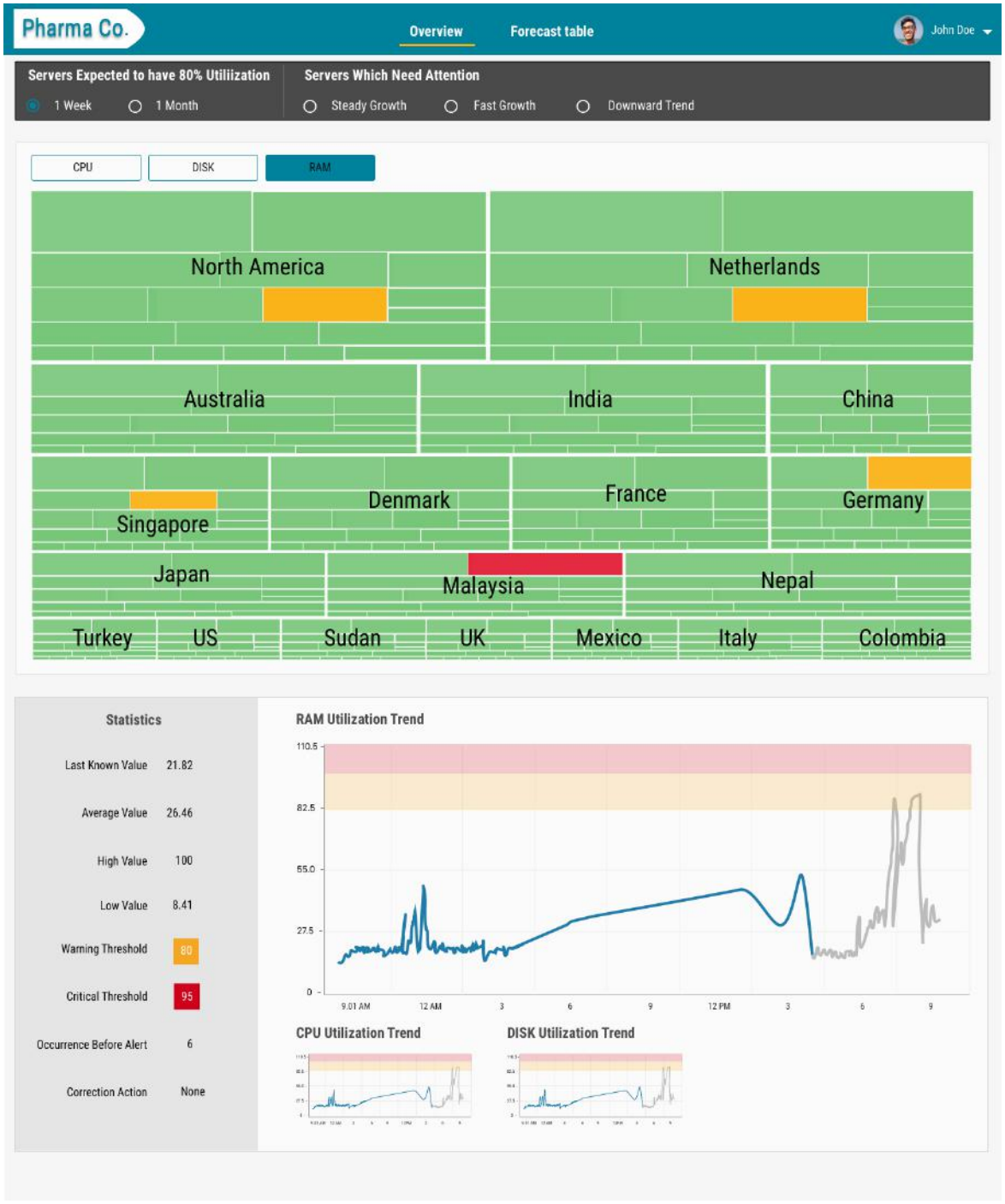
A Sample Forecasting result



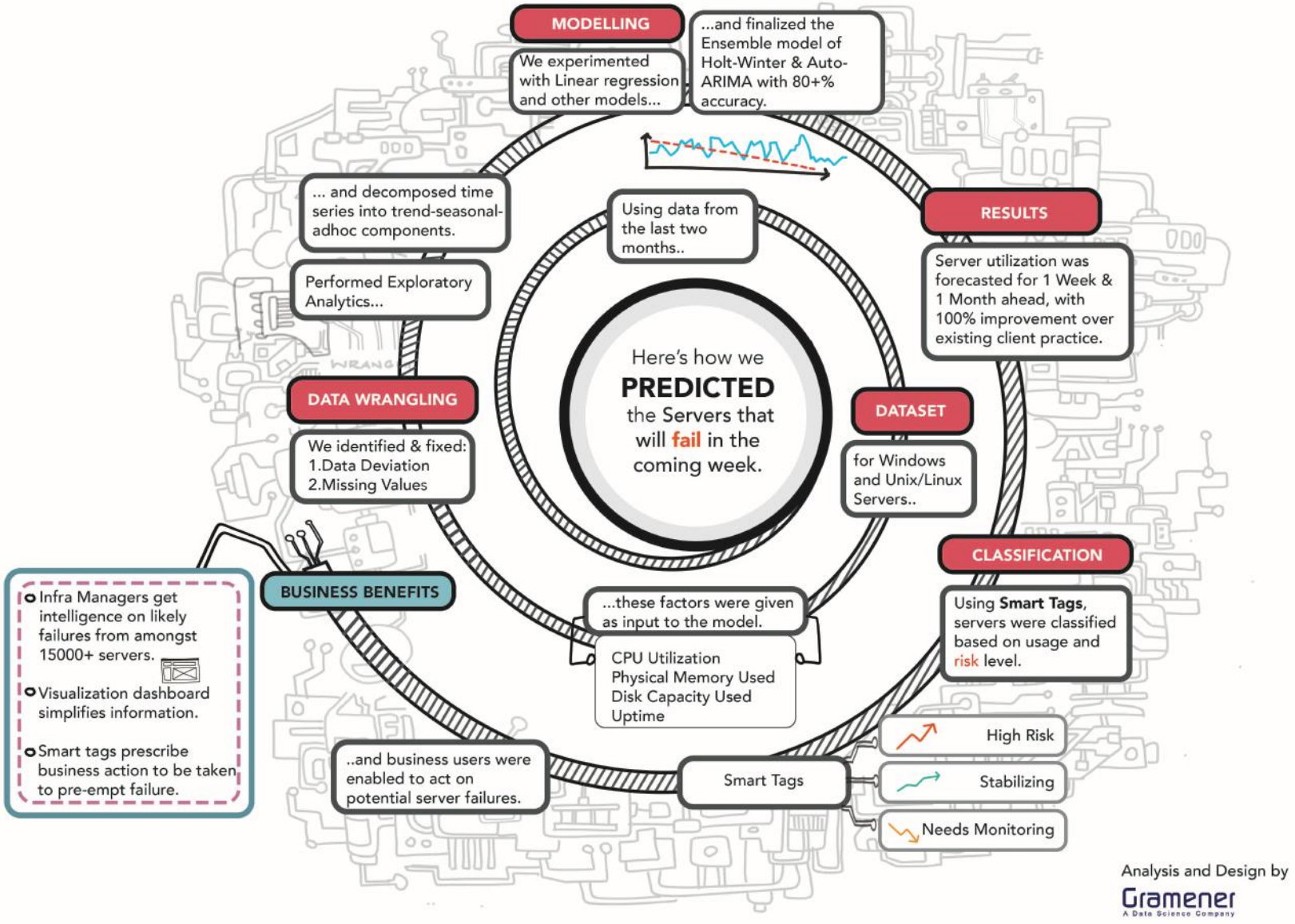


Smart Assistance Tags for actionability

Plan	Servers which are expected to have >80% utilization in coming 1 week	Servers which are expected to have >80% utilization in coming 1 month
Monitor	Servers which need attention – close to thresholds & steady growth	
Understand	Servers with fastest growth rate	Servers close to threshold but downward trend


Pharma Co.
Overview
Forecast table
John Doe
Servers Expected to have 80% Utiliization
1 Week
1 Month
Servers Which Need Attention
Steady Growth
Fast Growth
Downward Trend
CPU
DISK
RAM
North America
Netherlands
Australia
India
China
Singapore
Denmark
France
Germany
Japan
Malaysia
Nepal
Turkey
US
Sudan
UK
Mexico
Italy
Colombia
Statistics
Last Known Value 21.82
Average Value 26.46
High Value 100
Low Value 8.41
Warning Threshold 80
Critical Threshold 95
Occurrence Before Alert 6
Correction Action None
RAM Utilization Trend
110.5
82.5
55.0
27.5
0
9.01 AM
12 AM
3
6
9
12 PM
3
6
9
CPU Utilization Trend
DISK Utilization Trend

Server Failure Forecasting – Data Infographic



EMERGENCY

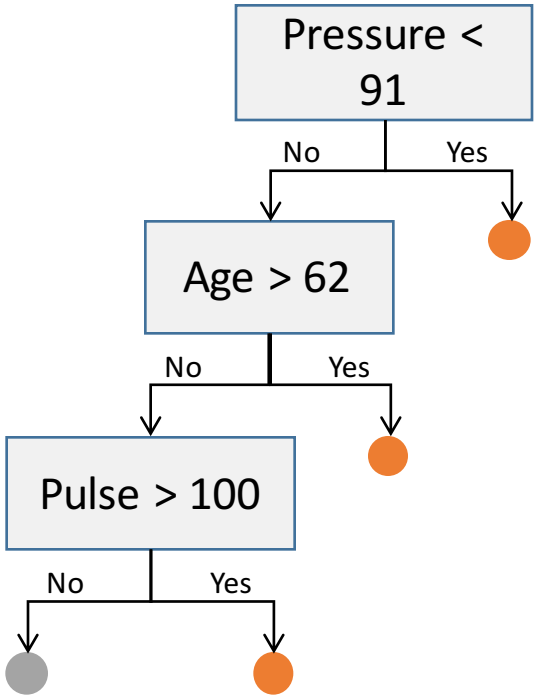
A man is rushed to a hospital in the throes of a heart attack.

The nurse needs to decide whether the victim should be admitted into emergency care.

Although this decision can save or cost a life, the nurse must decide using only the available cues, and within a few seconds – preferably using some fancy statistical software package.

A simple set of easily measurable parameters, with high accuracy were initially identified through the **Predictive modelling process**.

This led to speeding up of the decision on whether to admit a patient or not to critical care, for emergency treatment. This helps the nurse take quick decisions statistically rather than taking a decision by wit.



The entire solution was delivered as **Prescriptive analytics** – "advising best course of action, given a situation by analyzing, predicting and recommending on-the-fly, a specific business decision".

Thus, a step by step pre-defined process identifying the causes and the remedies will help in saving lives.

RECRUITING BASED ON PERFORMANCE DRIVERS

A large Government organization conducts an assessment used for recruitment purposes. Their question was: what demographic and behavioural drivers affect performance?

Our assessment solution uses machine learning to determine which factors have the strongest influence on an outcome.

For example, the table alongside shows that the college and previous salary are the best predictors of overall performance.

While experience is not a strong factor, it is a good predictor of communication ability, and moderately of closure.

Number of job hops is a strong indicator of initiative.

This allowed our client to:

- Increase conversion ratio 2.4 times
- Fill targets 40% faster
- Hire candidates whose post-employment performance was better

Factor	Overall ▼	Communication	Initiative	Creativity	Closure
College	11.0%	6.6%	18.8%	9.9%	7.9%
Previous salary	10.6%	8.7%	17.4%	10.3%	8.5%
Family background	10.5%	4.3%	18.4%	10.3%	7.7%
Level of education	9.1%	5.7%	14.9%	7.3%	7.9%
Extra-curriculars	5.2%	4.6%	5.7%	5.0%	6.5%
Reading habits	4.7%	2.5%	8.2%	4.1%	3.8%
Age	4.1%	3.1%	8.0%	3.3%	2.9%
Recommendation	3.7%	2.0%	7.1%	3.5%	2.5%
Psychometric profile	3.4%	1.6%	7.2%	2.3%	2.4%
Experience	3.4%	6.5%	5.9%	3.2%	4.1%
Background relevance	3.3%	1.3%	6.4%	3.1%	2.2%
Grooming	3.3%	3.0%	4.7%	4.0%	2.9%
Family income	3.3%	1.8%	5.7%	2.6%	3.0%
Location preference	3.1%	0.8%	6.4%	2.5%	2.6%
Management experience	2.8%	2.9%	5.6%	2.7%	3.3%
Idle time	2.7%	0.9%	5.0%	2.7%	2.1%
# job hops	2.4%	1.5%	8.3%	2.2%	0.6%
Background check	2.1%	1.5%	5.3%	2.0%	1.1%
Culture fit	1.3%	1.3%	2.1%	1.2%	0.3%
Time management	0.8%	2.2%	1.2%	0.9%	0.9%
Gender	0.7%	0.2%	1.9%	0.1%	0.5%

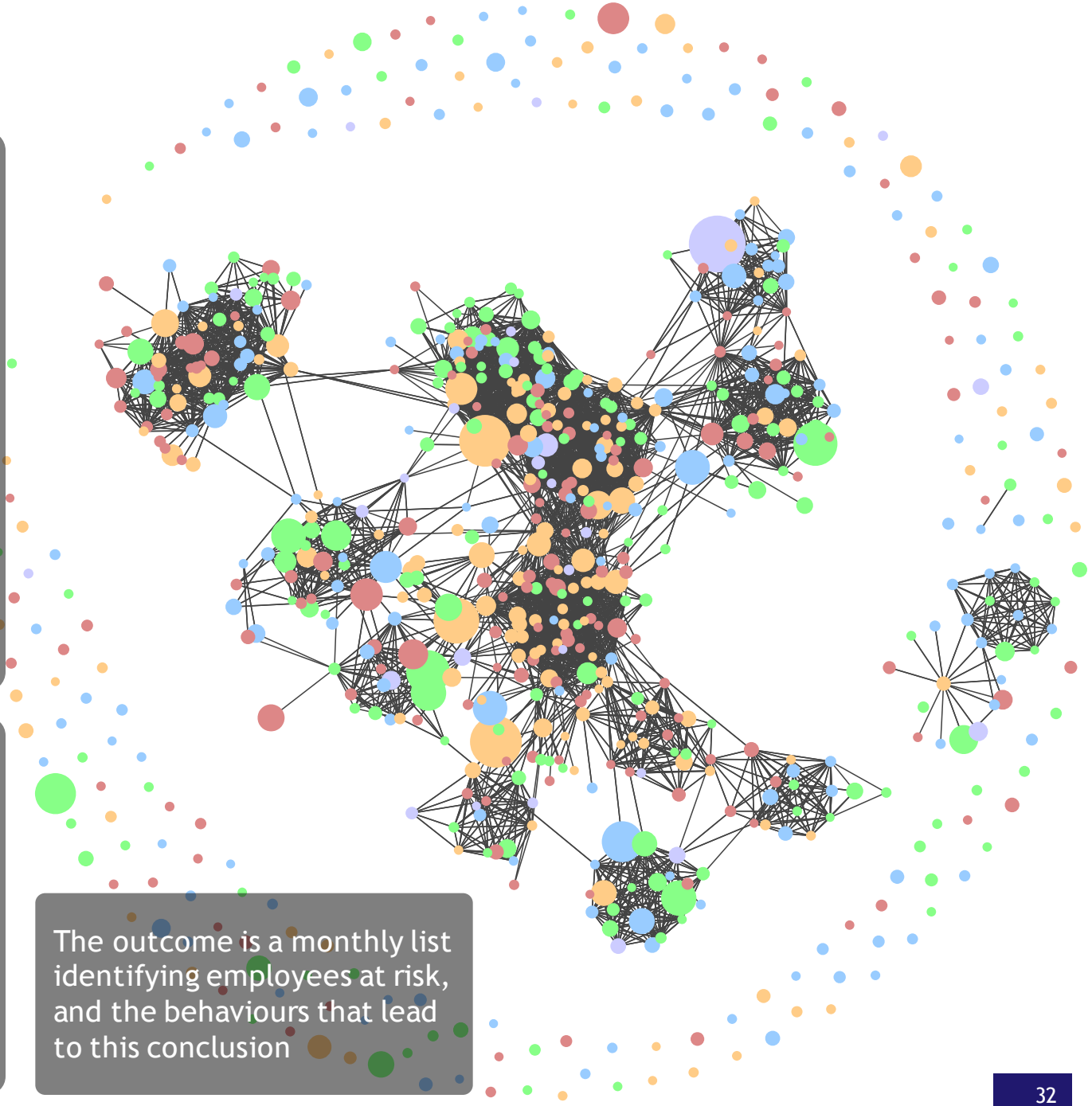
Anonymised summary of factors driving performance in assessment

Attrition Defence

A manufacturing firm asked the question: "How can we predict which employees will leave me next?"

One part of the answer is to take the network of email traffic among employees. The ones in close contact, exchanging emails with an alumnus are likely candidates for attrition.

The firm was able to put in place a retention and defence mechanism for these employees.



This is augmented with additional signals:

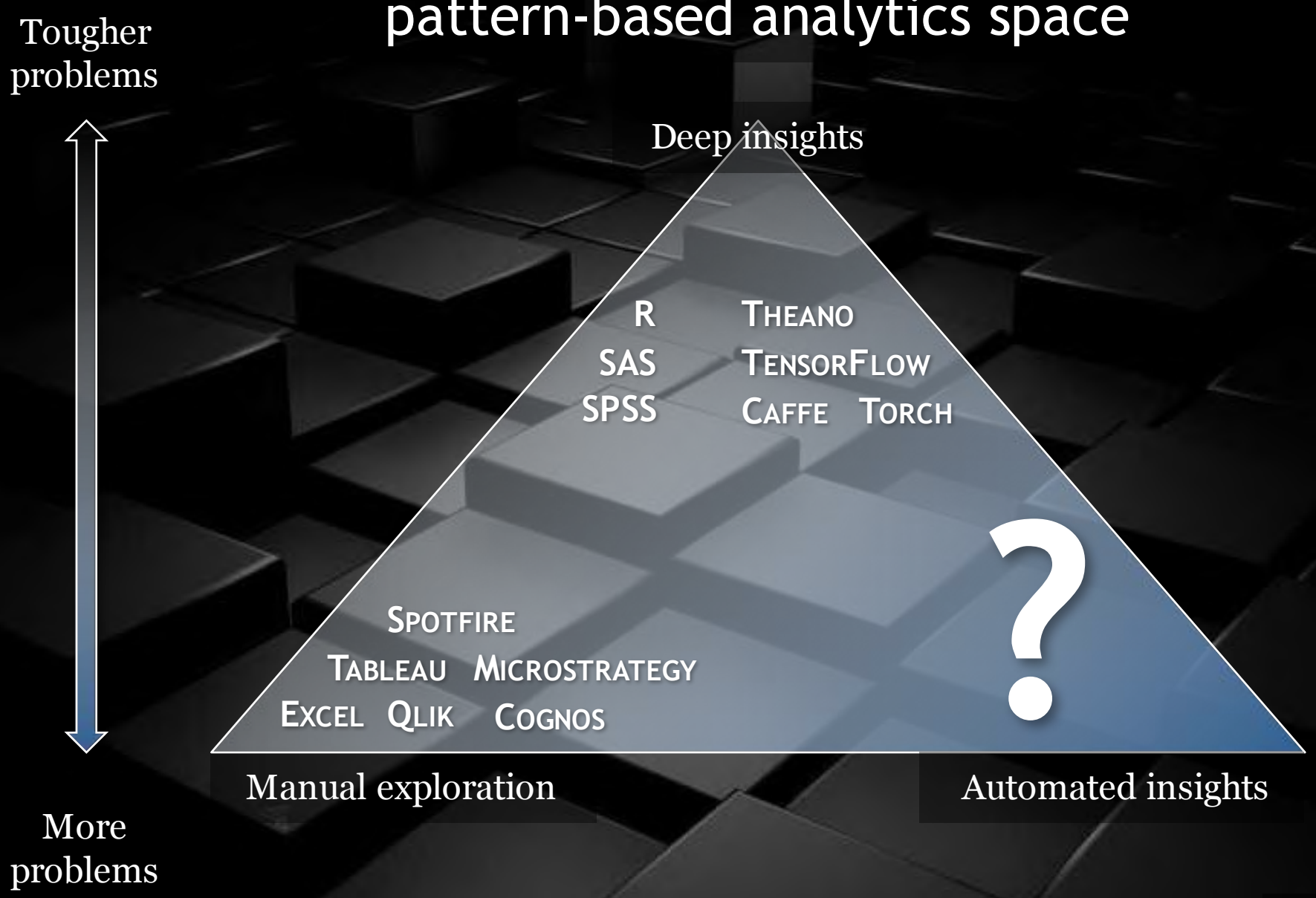
- Disengaged employees
- Active on LinkedIn
- Dip in performance
- Atypical browsing
- Collateral downloads
- Peer feedback
- Reduced working hours
- Increased sick leave

The outcome is a monthly list identifying employees at risk, and the behaviours that lead to this conclusion

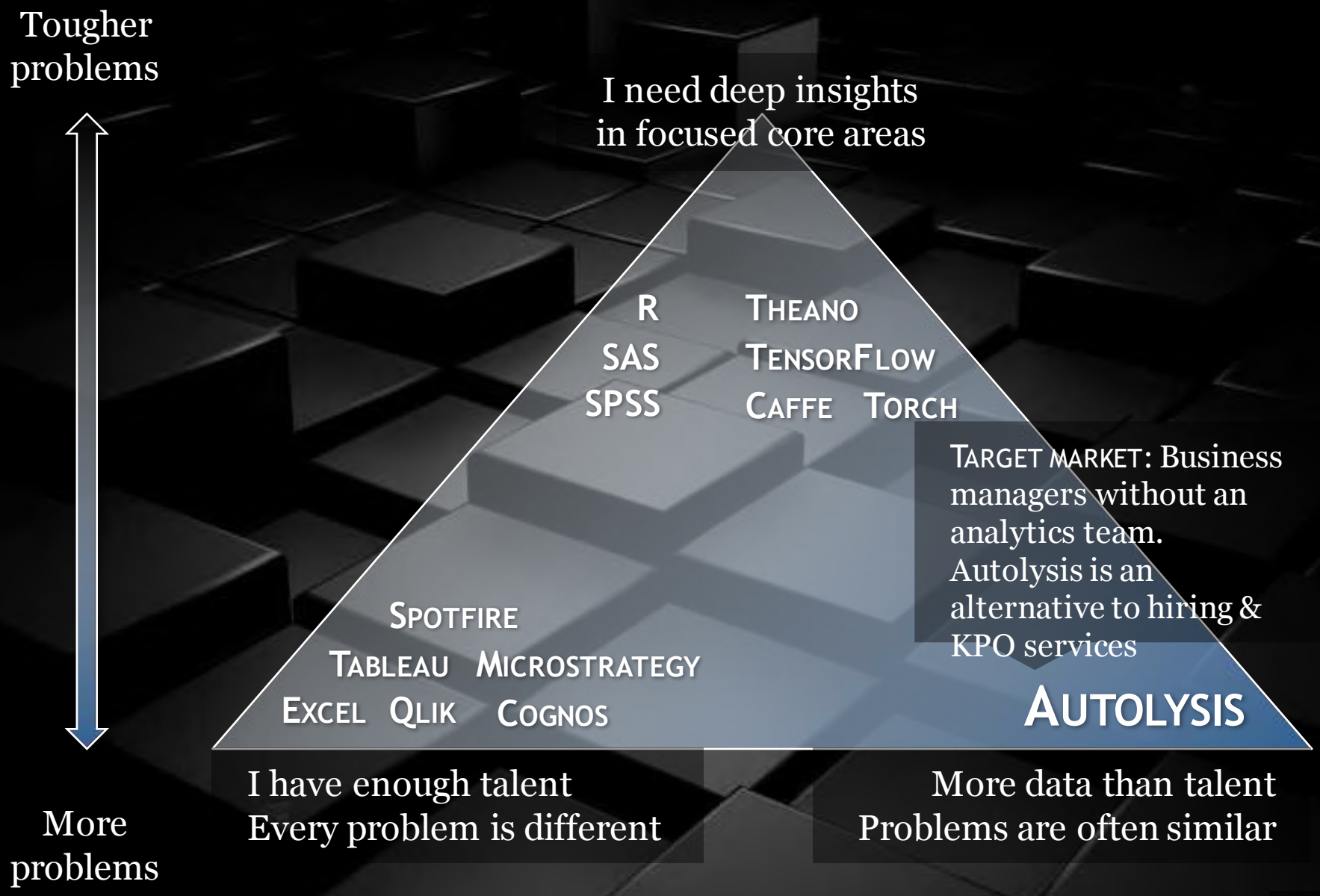
3. Gramener's Analytics Accelerators:

- Autolysis Tool,
- Clustering Tool,
- Data Exploratory Tool,
- Network Analytics Tool

There is a tool gap in automated pattern-based analytics space



This gap leaves a critical space open



Autolysis is still in development, but already has clients and successful engagements across verticals

Global Financial Co

Asset Management

There are millions of system processes that run globally, and a proportion of these fail.

Can we **predict which one will fail**? Why they fail? And how we correct it?

Top India Broadcaster

Media

Every serial's TVR is carefully monitored along with the brands, cast, moods, scenes, location, etc.

Can we **predict what drives TV ratings**? Who should be cast together? What moods?

Top India Poultry Co

Agriculture

The mortality of poultry is a significant driver of profitability. This is typically caused by the weather.

Can we identify **what actionable factors reduce mortality** in poultry?

Global Pharma Co

Pharma

The operations are managed through tickets that follow a carefully managed closure cycle.

Can we identify **what drives delays in ticket closure**?

Global Fin Services Co

Financial Services

Shareholders receive thousands of messages a month from companies, and proxy votes get lost in the mix.

Can we see **what improves proxy voting**? Which medium, segments and messages?

Indian Govt Edu Board

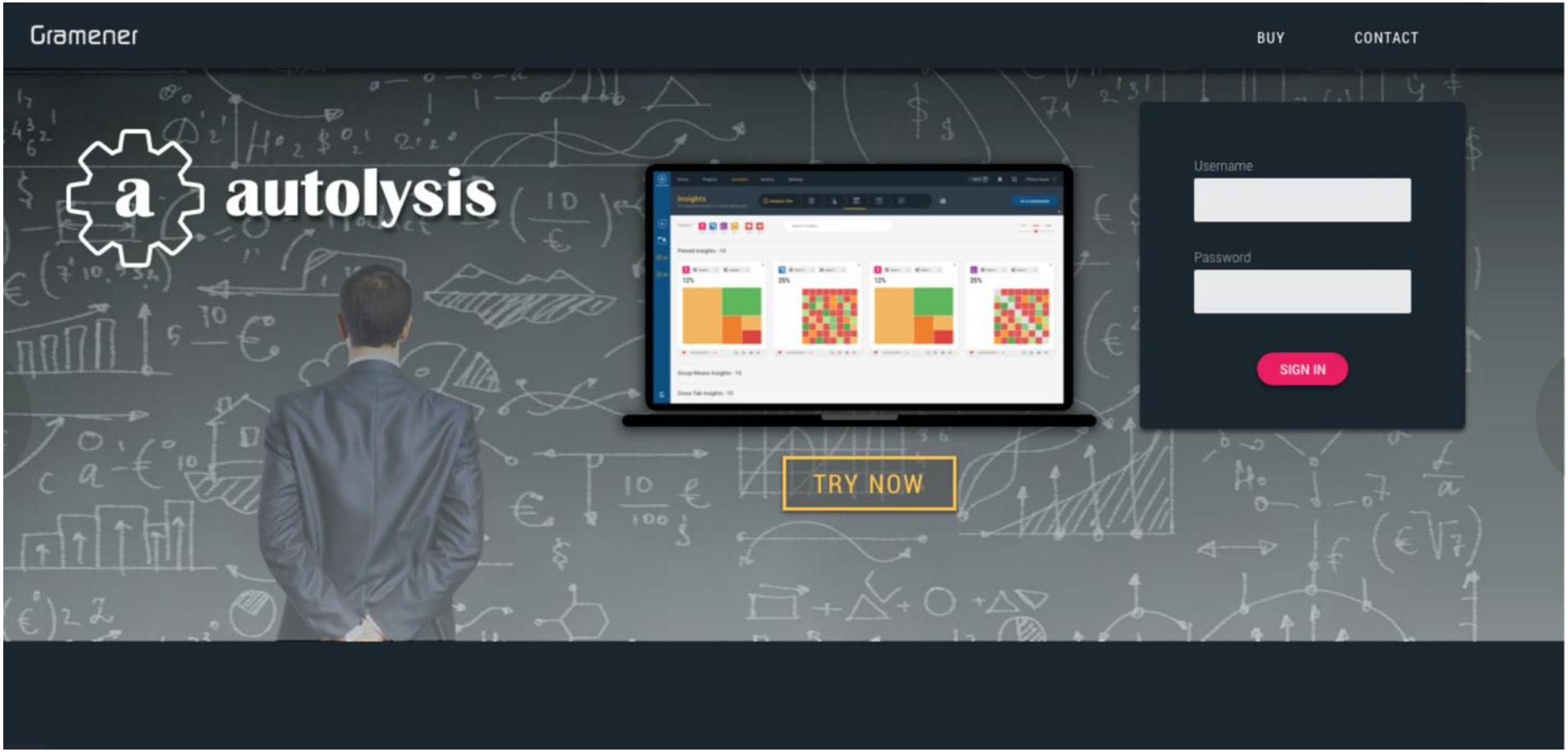
Government

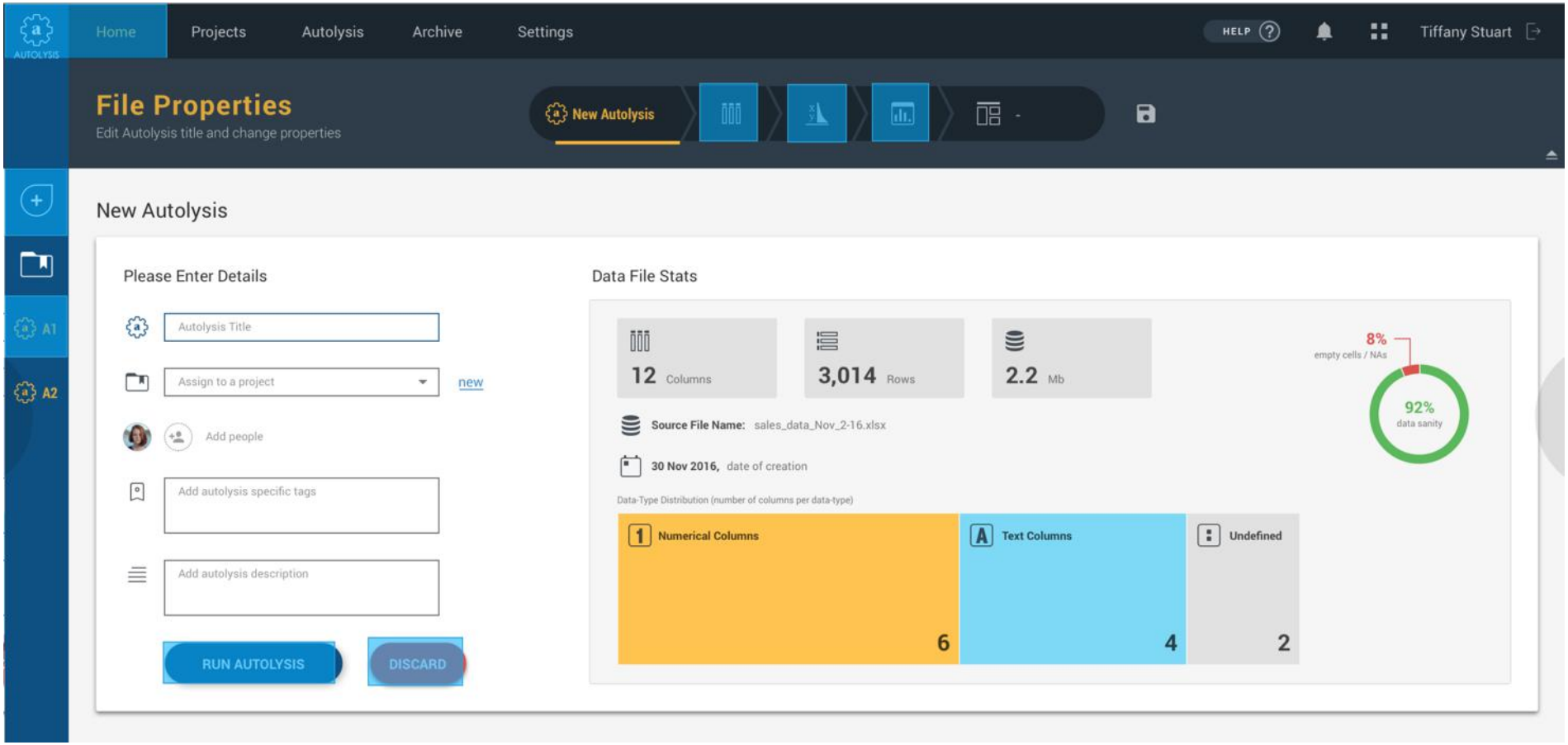
Students performance is a factor of education, demographics and behavior, and monitored for a large sample cross-country.

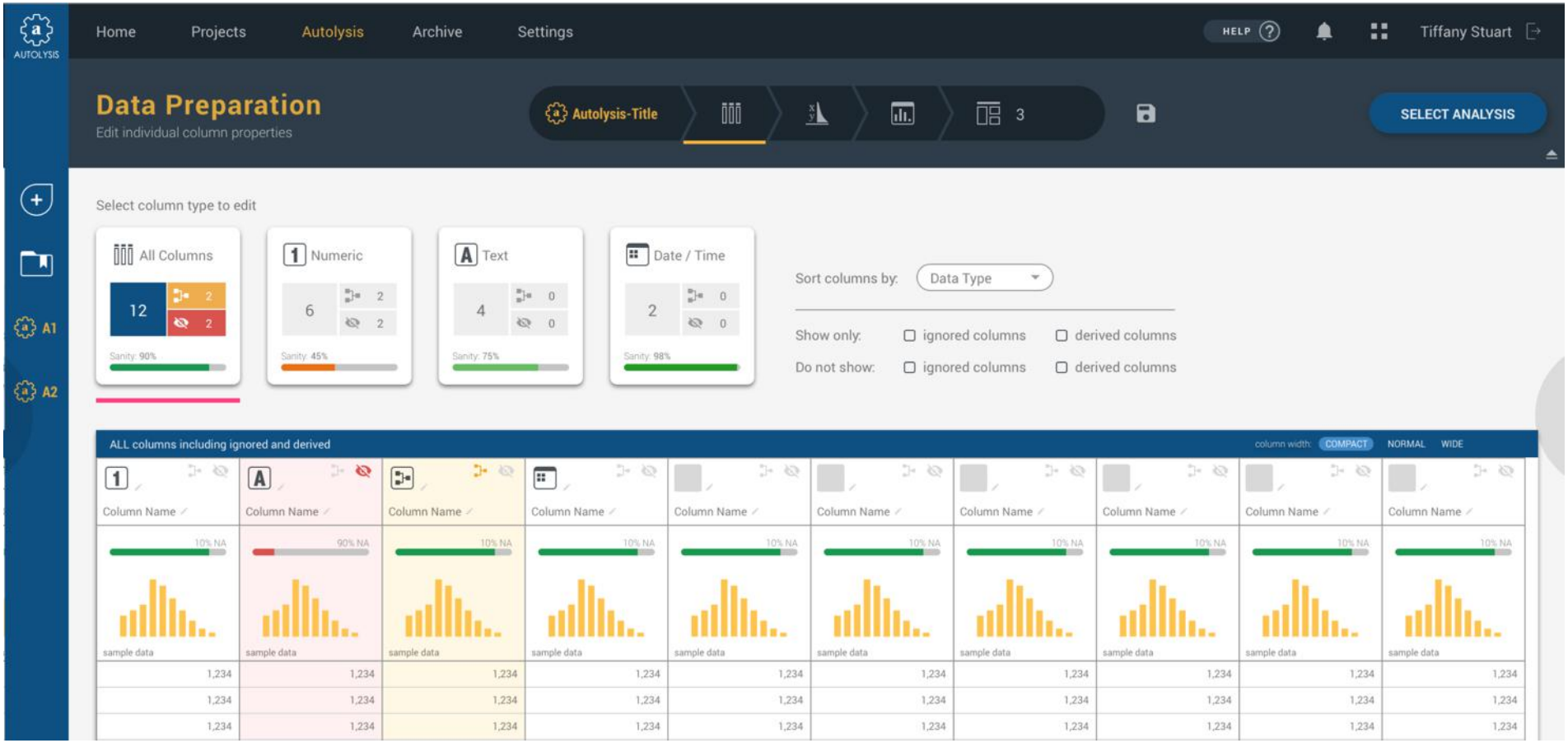
Can we see **what improves proxy voting**? Which medium, segments and messages?

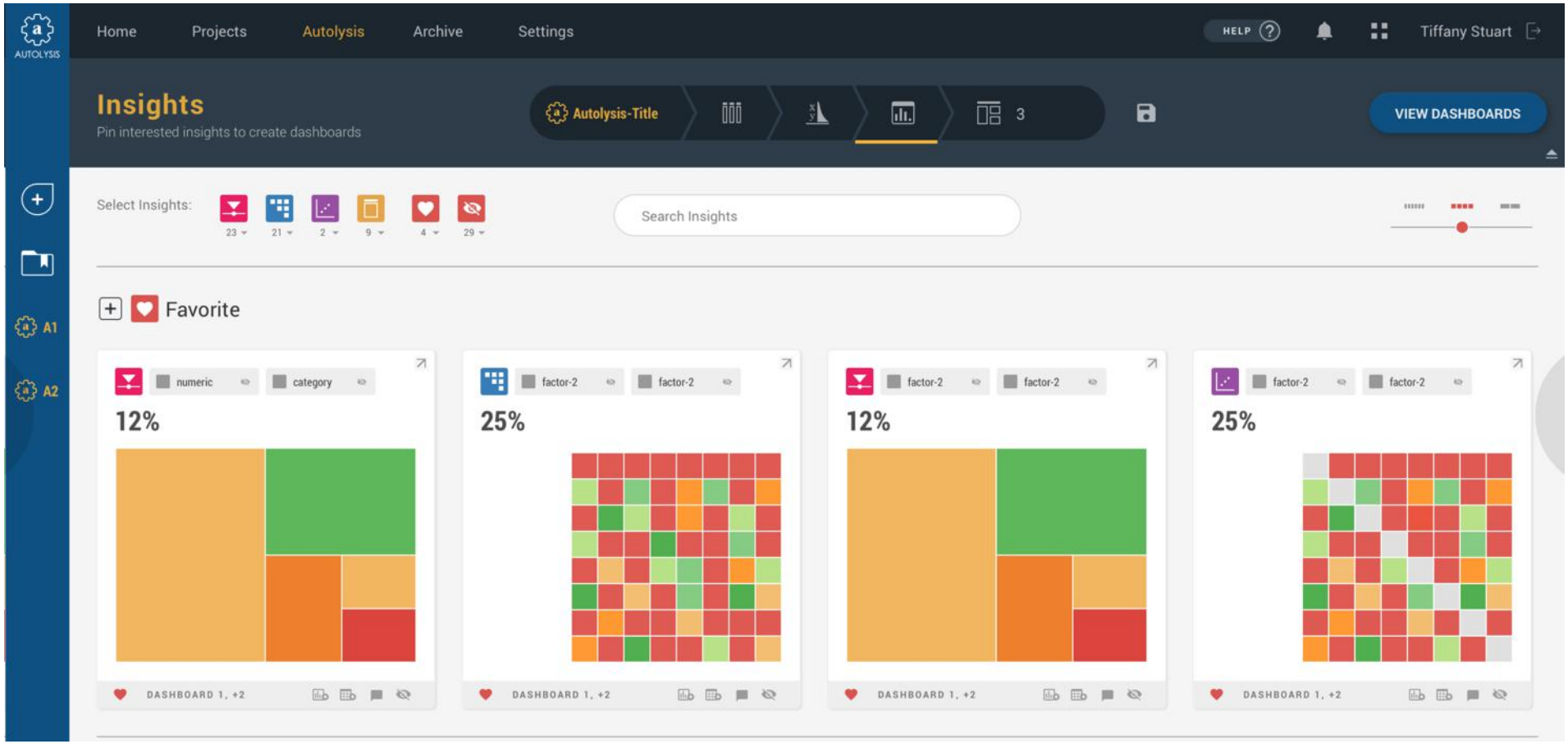
This is a subset of clients who used Autolysis

GRAMENER AUTOLYSIS – AUTOMATED ANALYSIS TOOL

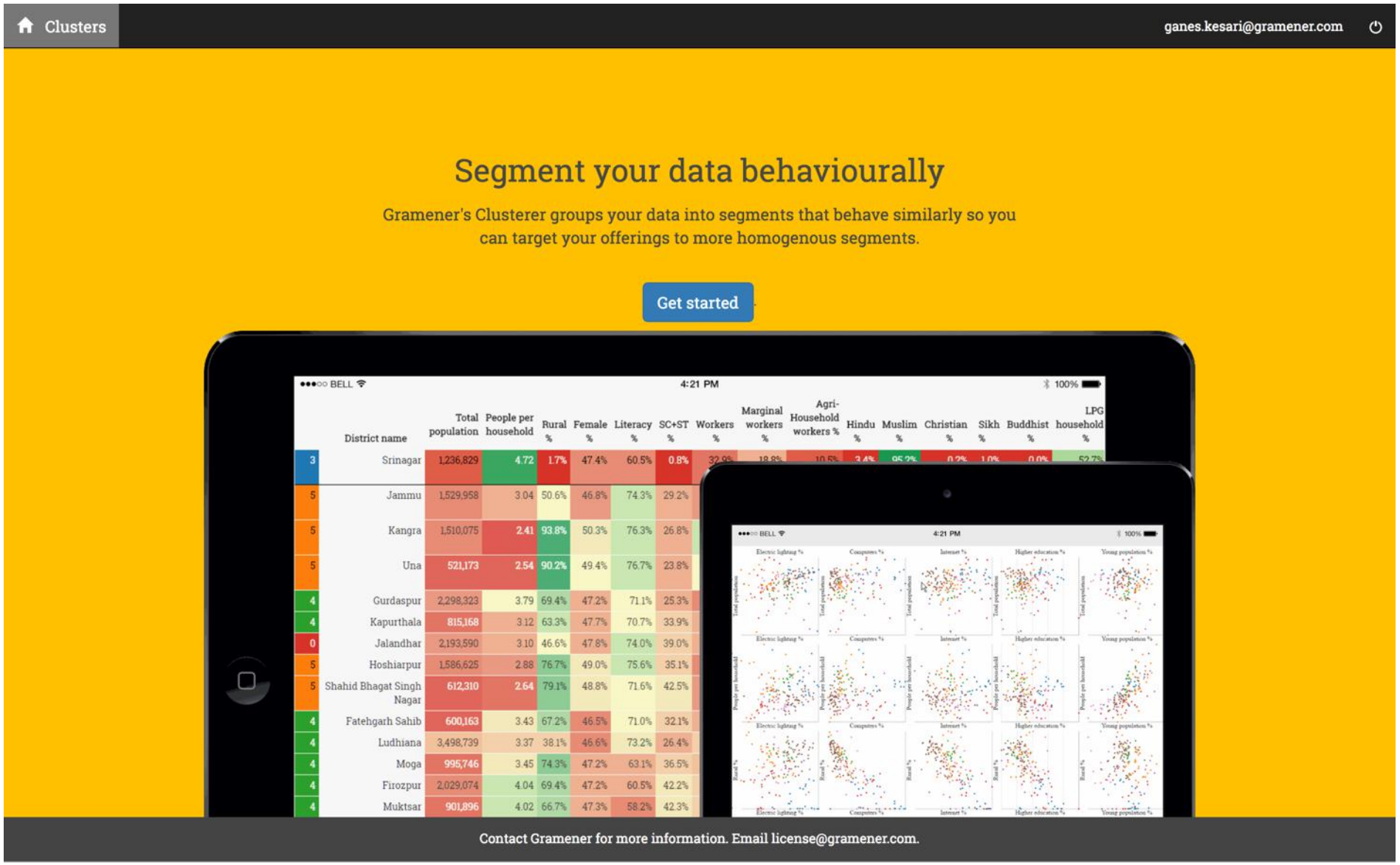



Home
Projects
Autolysis
Archive
Settings
HELP
Tiffany Stuart
File Properties
Edit Autolysis title and change properties
New Autolysis
New Autolysis
Please Enter Details
Autolysis Title
Assign to a project
new
Add people
Add autolysis specific tags
Add autolysis description
RUN AUTOLYSIS
DISCARD
Data File Stats
12 Columns
3,014 Rows
2.2 Mb
8%
empty cells / NAs
92%
data sanity
Source File Name: sales_data_Nov_2-16.xlsx
30 Nov 2016, date of creation
Data-Type Distribution (number of columns per data-type)
Numerical Columns
6
Text Columns
4
Undefined
2


AUTOLYSIS
Home
Projects
Autolysis
Archive
Settings
HELP
Tiffany Stuart
Data Preparation
Edit individual column properties
Autolysis-Title
3
SELECT ANALYSIS
A1
A2
Select column type to edit
All Columns
12
Sanity: 90%
Numeric
6
Sanity: 45%
Text
4
Sanity: 75%
Date / Time
2
Sanity: 98%
Sort columns by:
Data Type
Show only:
ignored columns
derived columns
Do not show:
ignored columns
derived columns
ALL columns including ignored and derived
column width:
COMPACT
NORMAL
WIDE
Column Name
10% NA
90% NA
sample data
1,234


AUTOLYSIS
Home
Projects
Autolysis
Archive
Settings
HELP
Tiffany Stuart
Insights
Pin interested insights to create dashboards
Autolysis-Title
3
VIEW DASHBOARDS
Select Insights:
23
21
2
9
4
29
Search Insights
A1
A2
Favorite
numeric
category
12%
DASHBOARD 1, +2
factor-2
factor-2
25%
DASHBOARD 1, +2
factor-2
factor-2
12%
DASHBOARD 1, +2
factor-2
factor-2
25%
DASHBOARD 1, +2

GRAMENER CLUSTERING TOOL



We distilled all demographic data into 32 relevant features

1. Source the data

We identified data from many demographic and behavioural datasets and mapped them to districts defined in Census 2011

- **Primary Census Abstract** — Population data, working population, gender, etc
- **Household Census** — Asset ownership, condition of households, etc
- **Socio-Economic Caste Census** — Caste information, economic and occupation data
- **Religious Census** — Religion beliefs followed by individuals
- **Districts of India** — Age group data, purchasing power parity information

2. Select the columns

From this universe of columns, we removed columns that were duplicated or irrelevant, and cleaned the remaining data

~400 columns → Select columns that are relevant to TV viewership → 115 columns

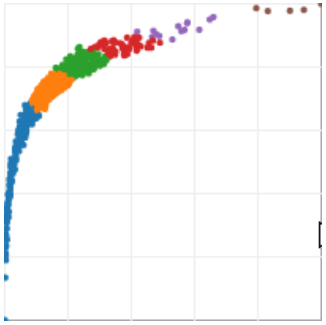
Ignore repeated permutations
Ignore irrelevant attributes
Cleanse data to fill missing values

3. Generate features

Features are columns or derived columns that are not auto-correlated, and hence better suited for clustering

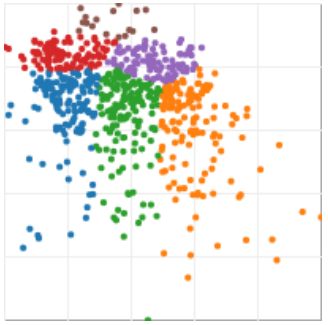
For example, this shows the distribution of workers to total population. The higher the population, the higher the number of workers, though there is a small spread across the districts.

This does not lead to clear clustering behaviour.



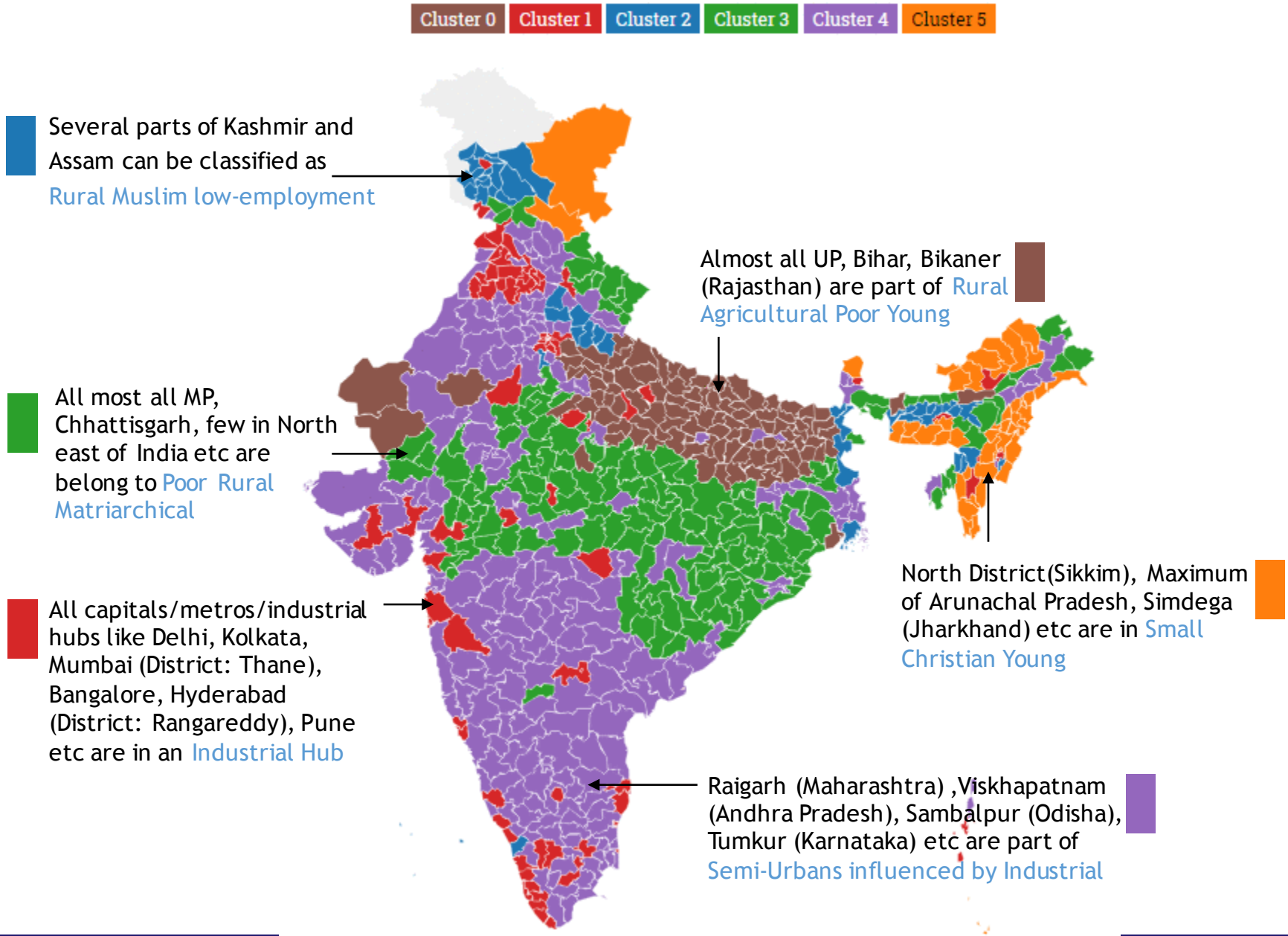
Replacing number of workers with the **percentage** of workers shows a much more distributed spread, and leads to clearer clusters.

By taking relevant ratios and combinations, we generated 32 features of relevance to geo-clustering.



32 features

This is a colour-coded map of the 6 All-India clusters



THESE FEATURES WERE USED TO CLUSTER URBAN DISTRICTS AS WELL

Extract all districts

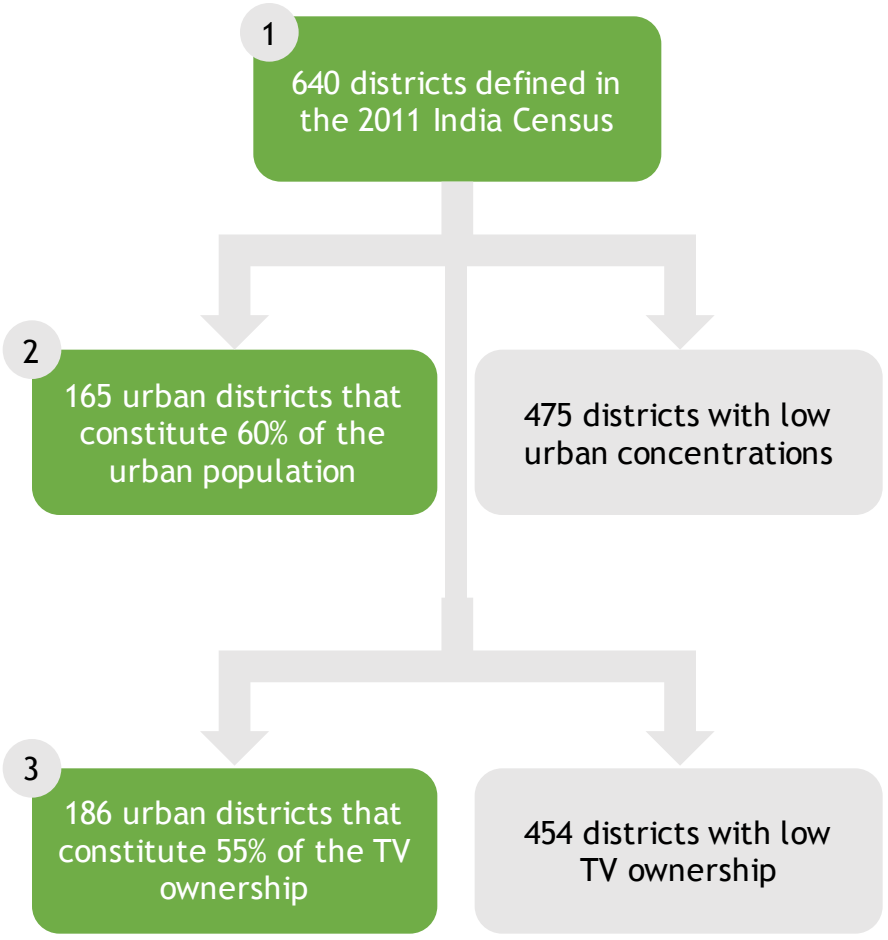
The original 2011 India Census has a total of 640 districts that acts as the first dataset for clustering.

Extract urban districts

Since urban districts are an area of focus, we also cluster the districts that have a high urban population

Extract TV districts

Districts with high TV ownership are also an area of focus.



33.3% of households are classified as living in urban regions. We took the districts with highest urban density (e.g. Delhi, Hyderabad, Chennai have 100%, Diu and Jammu are at ~50%, etc.)

The top 165 districts have a population that matches India's urban population, and also covers over 60% of the urban population.

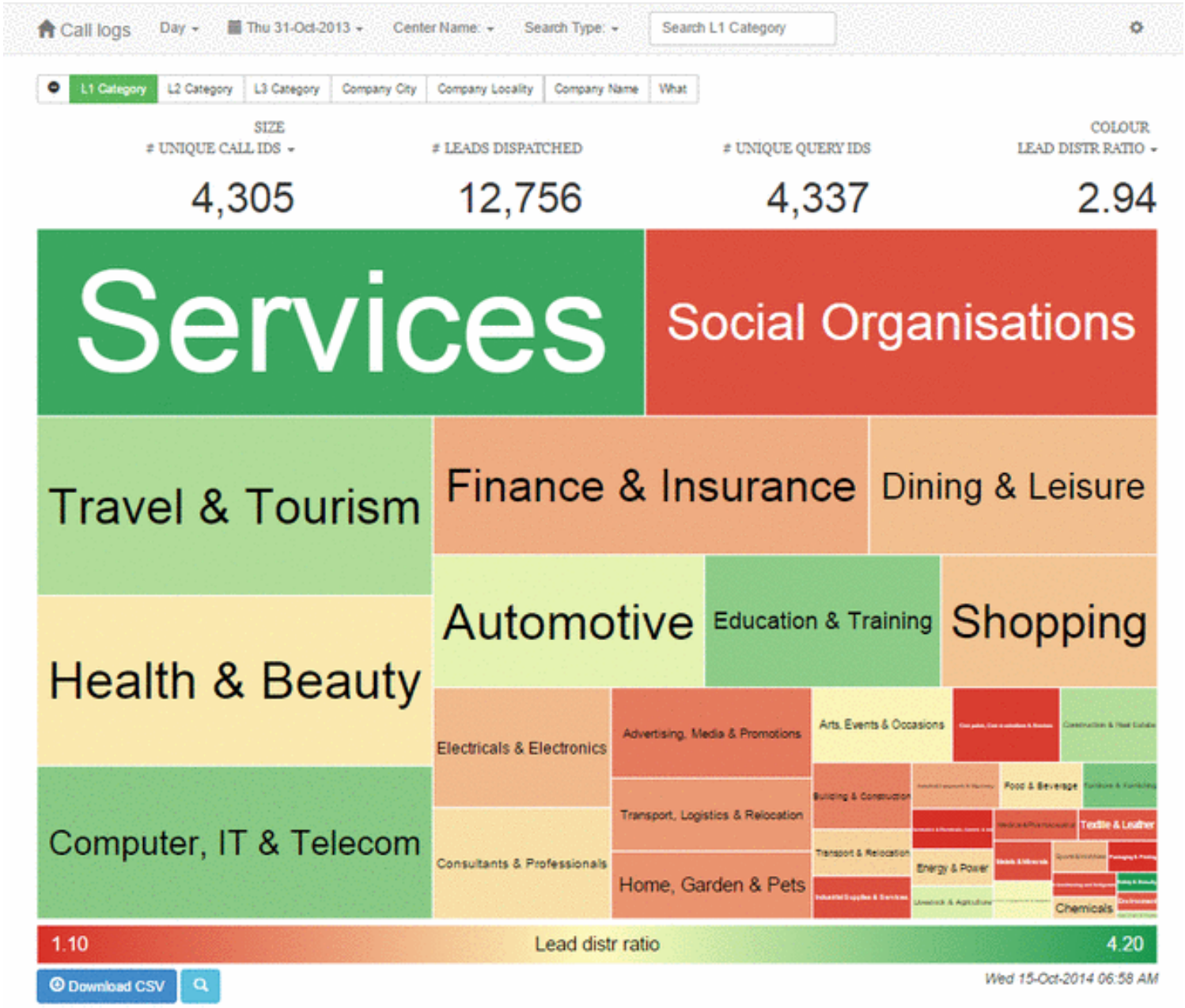
This occurs at a urban % cutoff of 34.2%, i.e. only districts with at least 34.2% urban households are considered.

35.2% of households own a TV. We took the districts with TV density (e.g. Chennai has 80%, Delhi has 50%, etc.)

The top 186 districts have a population that matches India's TV owning population, and also covers over 55% of the TV households.

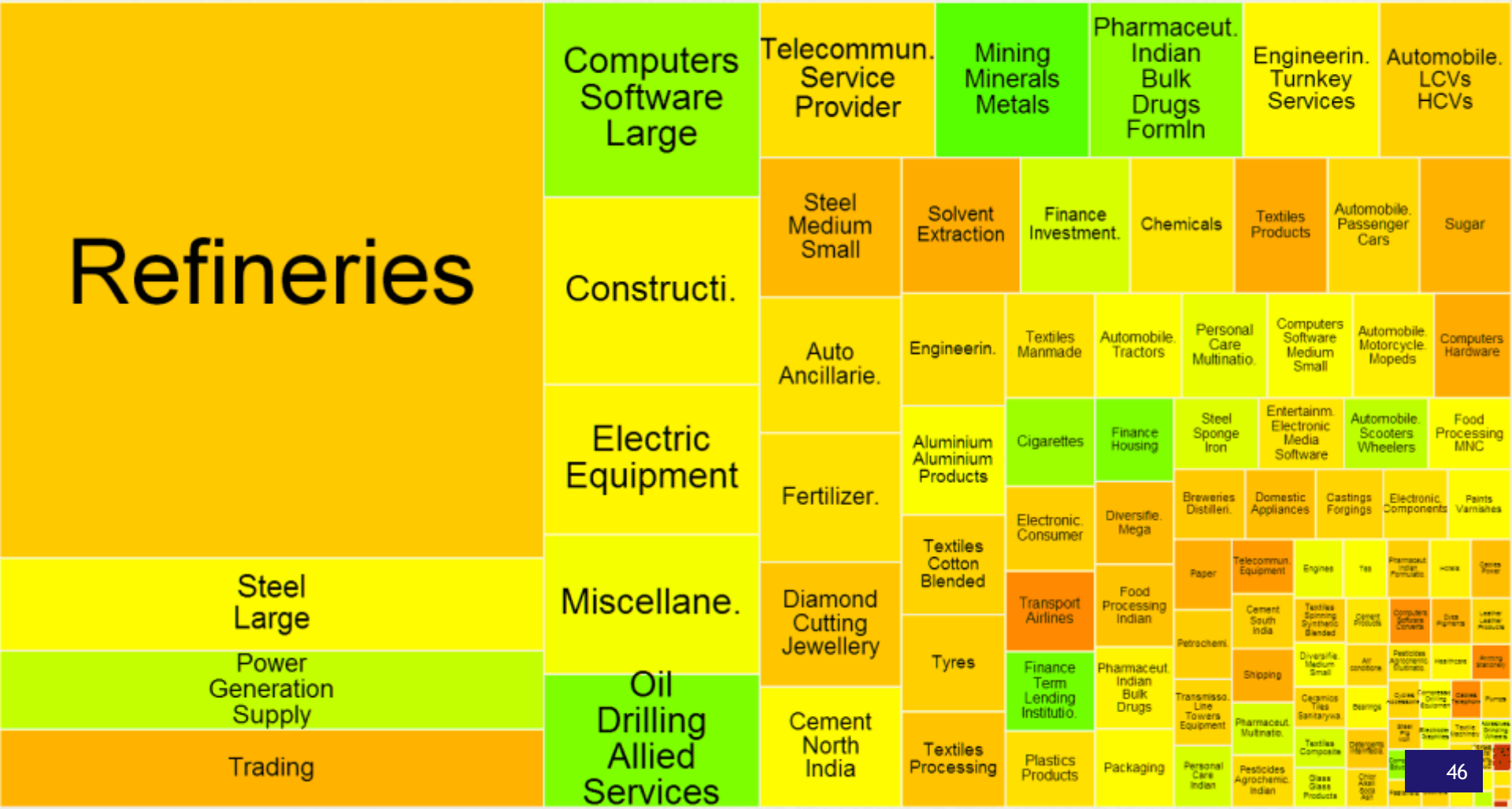
This occurs at a TV % cutoff of 40.3%, i.e. only districts with at least 40.3% TV ownership are considered.

GRAMENER DATA EXPLORATORY TOOL



GRAMENER DATA EXPLORATORY TOOL

Here are all public Indian companies, grouped by Industry. The size of the box indicates revenue (2012) and the colour indicates net profit (red is low, green is high). Click on the group to see companies below.



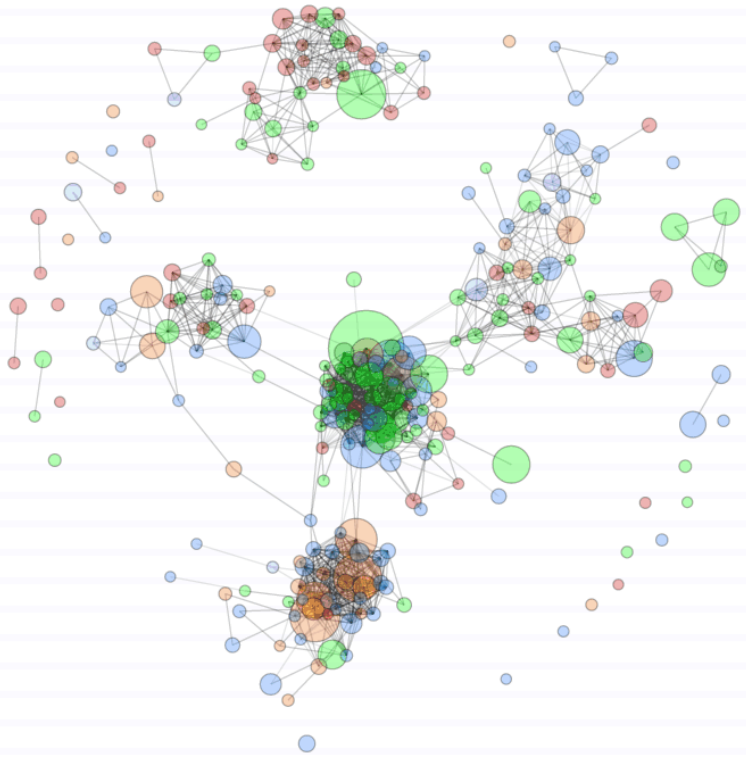
GRAMENER NETWORK ANALYTICS TOOL

Software ecosystem

Explore alternatives to various software used on a day-to-day basis. This is a visual interface to alternativeto.net. Search for a software, filter by any tag, or adjust the size of the network to explore interesting structures. Here are some interesting searches:

Google | Microsoft | Music | Productivity | Open source | iPad | Android | Windows games | Dropbox | Photoshop

Search | model | platform | tag



VLC media player (3328) Open Source Mac iPhone iPad Windows Android + Tablet Linux BSD Haiku VLC media player is an lightweight, open-source multimedia player that supports nearly all digital audio and video formats (e.g. H.264 ,MKV, TS (but not from Technisat DigiCorder), MPEG-2, mp3, AVI, MPEG-4...

Last.fm (1124) Freemium Mac iPhone Windows Windows Phone Android Linux Online Last.fm is a music community website that offers personalized internet radio, using a recommendation system called "Audioscrobbler" to build a detailed profile of users based on their music tastes...

Audacity (1075) Open Source Mac Windows Linux Audacity is free, open source software for recording and editing sounds. You can use Audacity to record live audio, convert tapes and records, edit sound files, change the speed or pitch of a recording and...

Grooveshark (1042) Freemium Mac iPhone Windows Android Online Symbian S60 HP webOS Blackberry Grooveshark is a free, internationally-available music streaming service and song recommendation engine. It allows users to search for, stream, and upload music, free of charge.

foobar2000 (895) Free Windows foobar2000 is an advanced freeware audio player for the Windows platform. Some of the basic features include full unicode support, ReplayGain support and native support for several popular audio formats.

The Social Tale of Two Cities: Bangalore & Singapore

Gramener

Recruiting top quality developers is always a problem. We decided to use an algorithmic approach and pulled out the social network of developers on **Github** (a social network for open source code).

In this visualisation, each circle is a person. The size of the circle represents the number of followers. **Larger circles have more followers** (but not in proportion – it's a log scale.)

The circle's colour represents the city the programmer's live in. This visual is a slice showing the tale of two cities: Bangalore and Singapore

Two people are connected if one follows the other. This leads to a clustering of people in the form of a network.

Here, you can see that **Bangalore** and **Singapore** are reasonably well connected cities. Bangalore has more developers, but Singapore has more popular ones (larger circles).

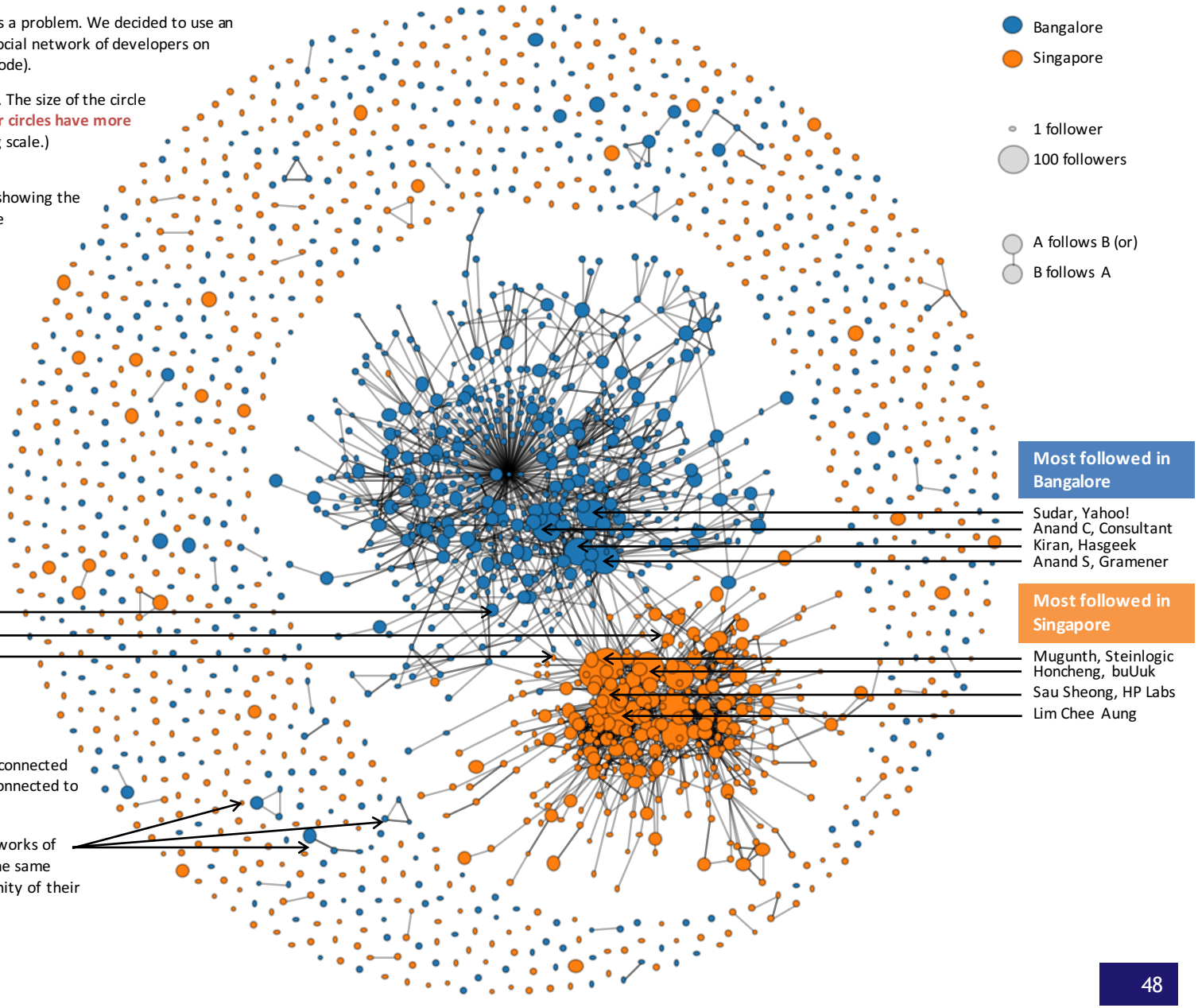
However, the interaction between Bangalore and Singapore are few and far between. But for a few people across both cities, like:

Ciju Cherian
Lin Junjie
Amudhi Sebastian

... etc.

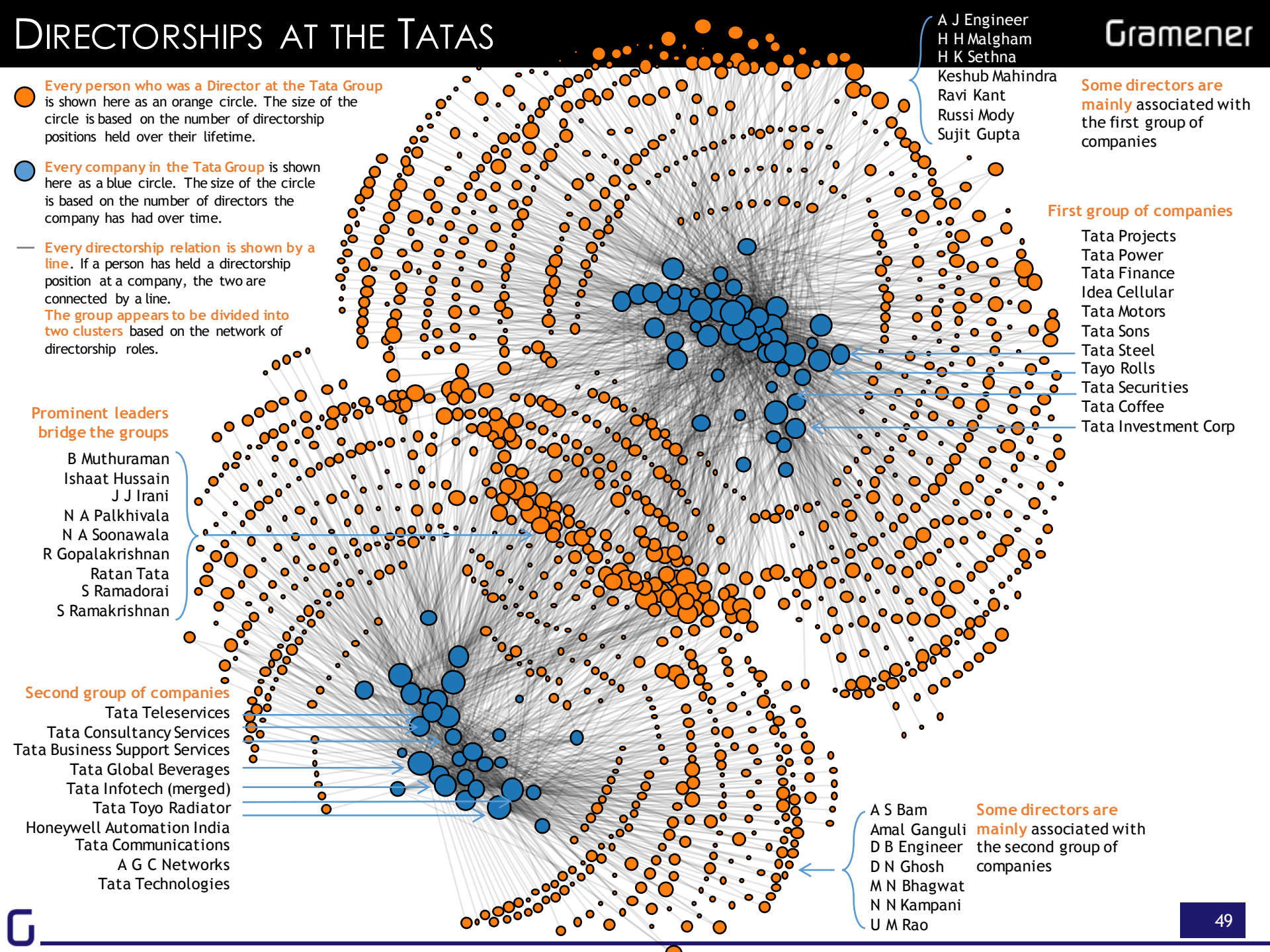
There are, of course, a number of smaller independent circles – people who are not connected to others in the same city. (They may be connected to people in other cities.)

Apart from this, there are a few small networks of connected people – often people within the same company or start-up – who form a community of their own.



Directorships at the Tatas

Gramener



Contact Gramener

5000 Birch St, Newport Beach,
California 92660, USA.

+1 949 878 0703

contact@gramener.com

"Vishala", 51 Kalingarayan Street,
Ramnagar, Coimbatore 641 009, INDIA

+91 422 223 1010

contact@gramener.com

9/2, 2nd Floor, Survey 64,
HUDA Techno Enclave, Phase 2
Madhapur, Hyderabad 500081
Telangana, INDIA

040-67642100

contact@gramener.com

320-A, I Floor, C.M.H. Road,
Above The Wearhouse, Indiranagar,
Bangalore 560 038, INDIA.

080 41225398

contact@gramener.com