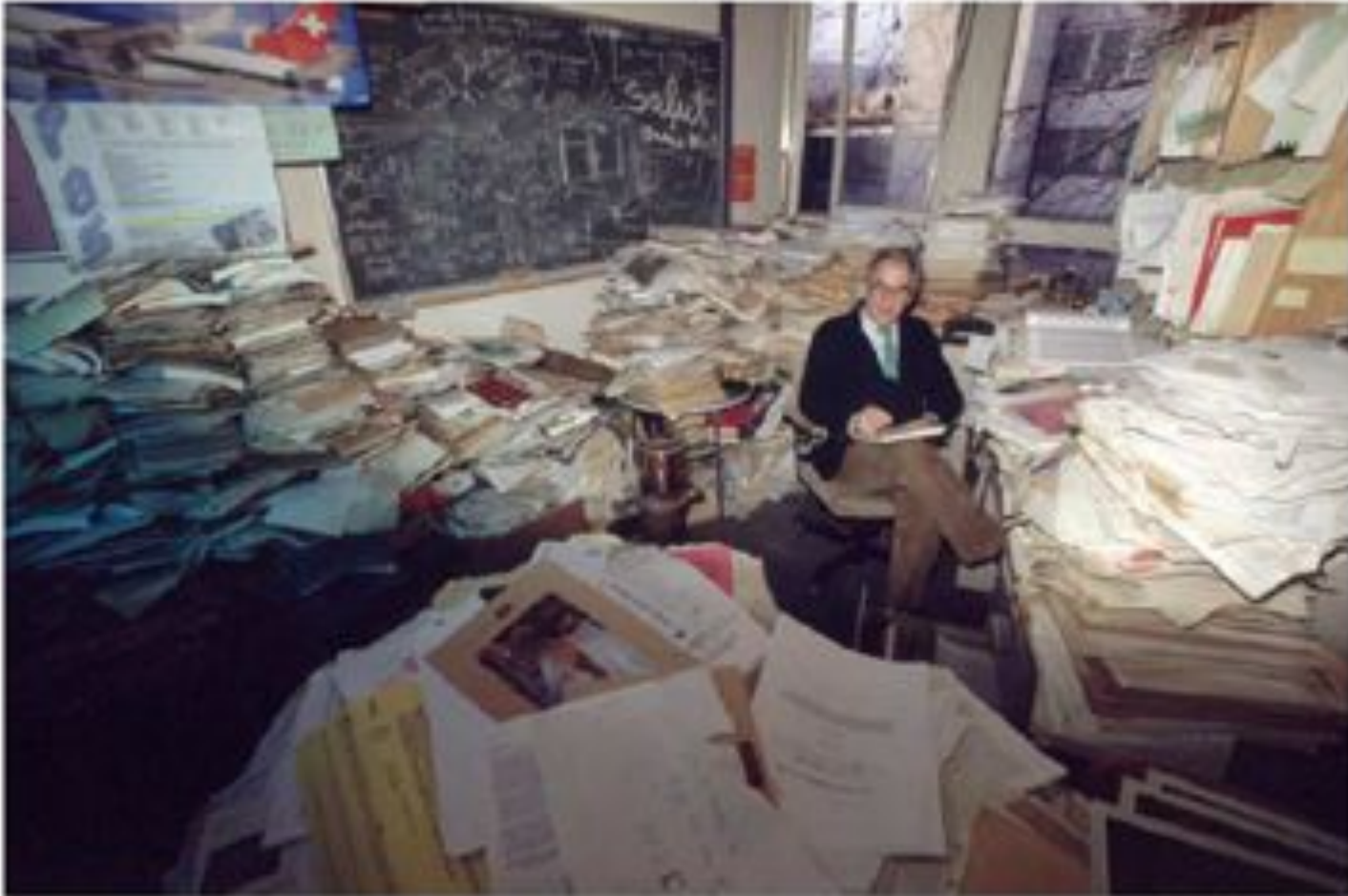


Machine Learning: from flight data to claims management

Xavier Conort
Principal Research Engineer
Institute for Infocomm Research
Agency for Science, Technology
and Research (A*STAR)



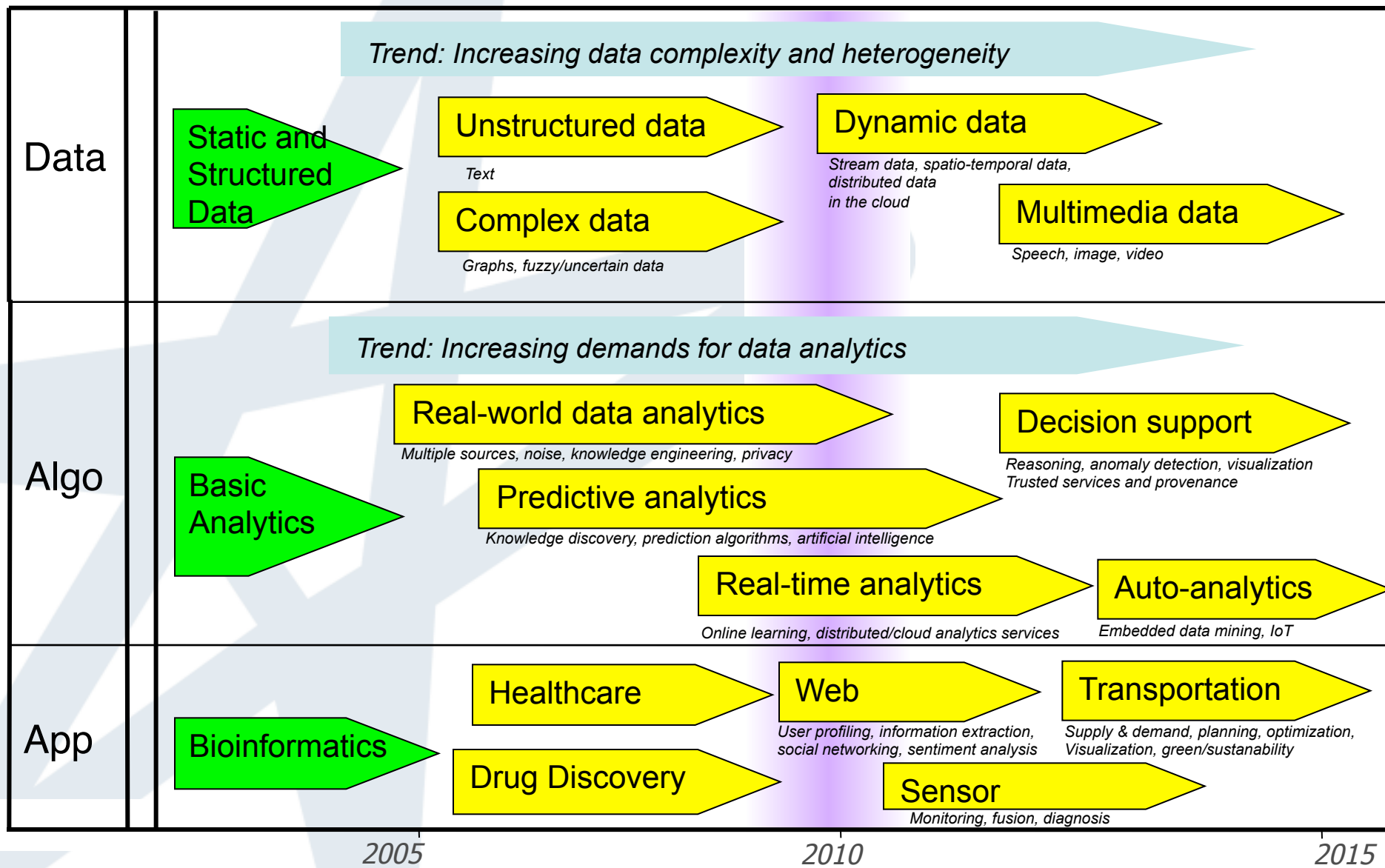
Previously...



Now



Evolution of Data and Analytics



Why GLMs are getting less popular in a Big Data world

GOOD	BAD	UGLY
Recognized as a standard in the banking and insurance industry	Need to pre-process data (missing values, outliers, dimension reduction)	GLMs is prone to overfitting while used with large amount of features or features with a large number of categories
Accommodate responses with skewed distributions	GLMs do not automatically capture complexity in the data. It can take weeks or months to go through the GLM iterative modelling process	
Simple mathematical formula easy to implement and easy to interpret		

GLMs are still the technique of choice in insurance as most companies exploit a limited number of risk factors and actuaries know well their data.

However, in a Big Data environment where the competitive battleground consists of analyzing massive amounts of both structured and unstructured to gain real-time insights, **Machine Learning based techniques became the techniques of choice of many industries.**

Who is using Machine Learning (ML)?



Andrew Ng
Stanford University

- Co-founder of Coursera
- Among the 100 most influential people in the world according to Time

“In the past decade, machine learning has given us self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome. Machine learning is so pervasive today that you probably use it dozens of times a day without knowing it. Many researchers also think it is the best way to make progress towards human-level AI.”

Why do I or data scientists like to use ML?

- Help to capture **complexity in the data**. According to Andrew Ng, “Machine learning is the science of getting computers to act without being explicitly programmed”.
- Allow to work with **high dimensionality** thanks to various techniques to combat over-fitting such as
 - Penalties in the loss function
 - Random examples sampling
 - Random features sampling
 - Wisdom of crowds (ensemble of independent weak learners)

ML Techniques toolkit includes

- Regularized Generalized Linear Model
- Random Forest (RF)
- Gradient Boosting Machine (GBM)
- Support Vector Machine (SVM)
- Neural networks
- Collaborative filtering techniques for recommendation engine
-



Data Analytics in I²R



Institute for
Infocomm Research



Agency for
Science, Technology
and Research



Jurong Town
Corp



Singapore
Tourism Board



Sentosa
Development
Corp



Competition
Commission of
Singapore



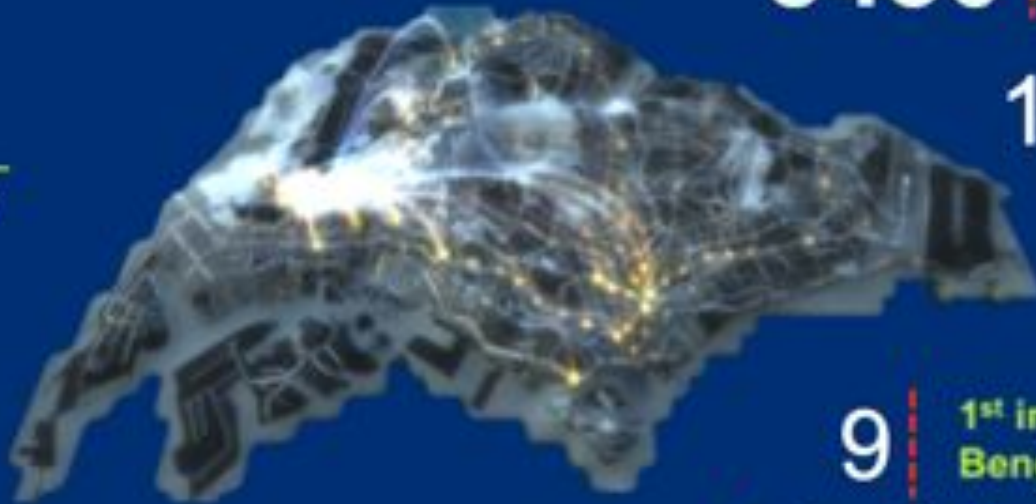
Department
of Statistics



Energy Market
Authority

I²R

Singapore's
Largest ICT
Research
Institute



600

Staff today

5430

Research
Publications

1220

Technology
Disclosures

216

Patents

9

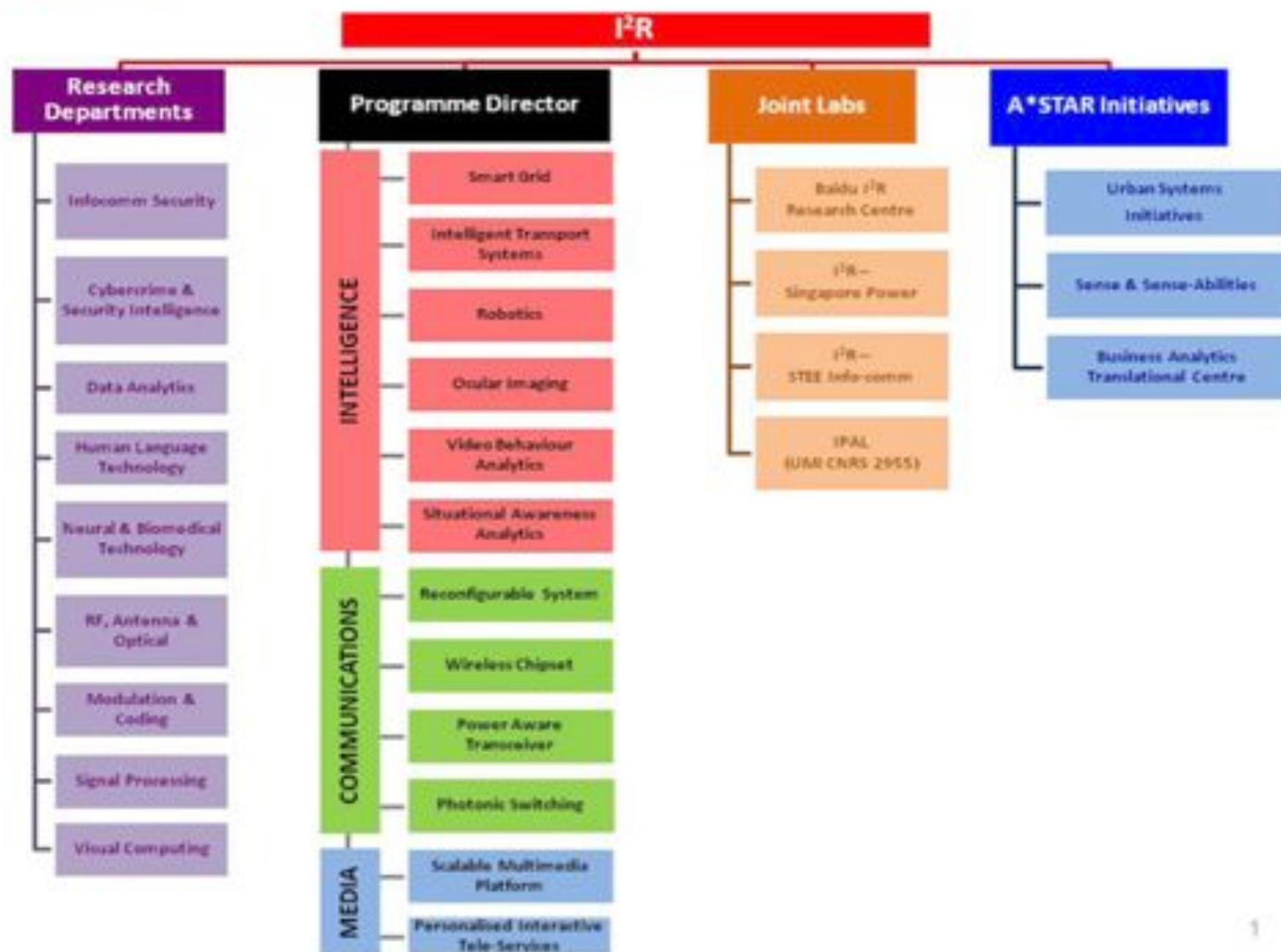
1st in International
Benchmarking Competitions

18

Contributions to
Industry Standards

681

Industry Collaborations

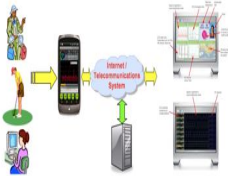




*From Data to Knowledge
to \$ - Extracting
Hidden
Insight*

**Machine
Learning**

*Large-Scale,
Distributed
Analytics for
BIG Data,
Mobile Data
Mining*

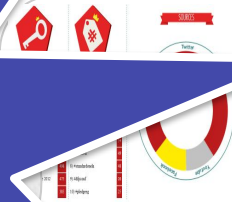


*High-Speed
Data Streams
on the Cloud
Privacy
Preserving
Analytics*



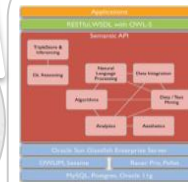
**I²R DA dpt =
30+ Data
Scientists and
10+ R&D
Engineers**

*Information
Extraction
from Text,
Natural
Language
Processing*



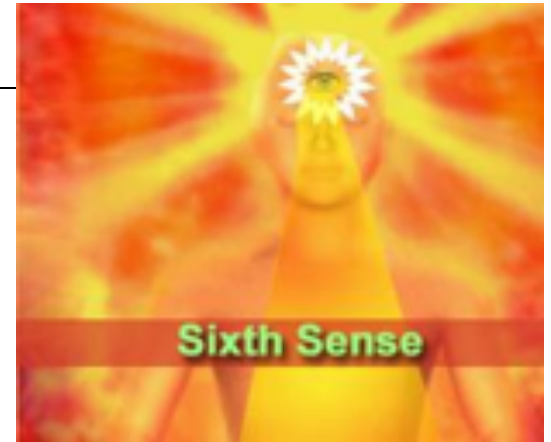
**Semantic
Computing**

*Semantic
Services
and
Reasoning*



Our Mission

- **DATA MINING**: exploration & analysis
 - of *large quantities of data*
 - by *automatic means*
 - to discover *actionable* patterns & rules



Attracting MNC Investments into Singapore

I²R - VISA Collaboration

Innovative system to enhance e-commerce payment security

- Detects activities of fraudulent merchants
- Collects evidence of criminal activities



Attracting MNC Investments into Singapore

I²R - Boeing Collaboration



Source: Randy's Journal by Randy Tinsell

Joint Development of Integrated Sensory Platform

- Part of Intelligent Factory Alert System
- Improve the production and assembly processes
- Successfully introduced in production line of Boeing 777



Attracting MNC Investments into Singapore

Baidu – I²R Research Centre (BIRC)



Baidu and A*Star open joint lab here

Both giant will use I2R's language technology to target S.E. Asian users

The joint lab will focus on research in natural language processing, machine learning, and semantic search. The lab will be located at the I2R Research Centre, which is a collaboration between Baidu, I2R, and A*STAR.

Baidu's First Overseas Joint Lab with I²R 26th Jul 2012.

Roadmap to research South-East Asian languages, and other language processing.

Baidu would use I²R's technologies for its semantic search engine called "box computing".



Research Impact & Recognition Awards & Benchmarking (2011-2013)

- Over 50+ Top-Tier Publications Between 2010-2012
- **First Place** in GE's FlightQuest - 2013
- **First Place** in PAKDD Churn Prediction - 2012
- **First Place** in Fraud Detection in Mobile Advertising - 2012
- **First Place** in Don't Get Kicked Challenge – 2012
- **First Place** in Mobile Activity Recognition Challenge – 2011
- **Second Place** in Merck Molecular Challenge - 2012
- **Second Place** in Online Product Sales Competition- 2012
- **Second Place** in NIST Entity Linking Competition - 2011
- **Second Place** in Give Me Some Credit Competition - 2011
- **Third Place** in Time-Series Forecasting Competition - 2012
- **Third Place** in IEEE Services Cup – 2012



GE Flight Quest

36

f Share

110

t Tweet

38

in Share

0

Submit

1

+1

5

reddit

**Tomio Geron**, Forbes Staff

Covering start-ups, social and venture capital.

+ Follow (517)

f Follow 18k

TECH | 11/29/2012 @ 2:12PM | 1,880 views

GE Uses Crowdsourcing To Solve Air Travel Delays And Healthcare



3 comments, 1 called-out

+ Comment Now

+ Follow Comments

GE is a large company that creates and builds a wide range of products, but the industrial giant is now looking for a little outside help. The company is looking for some Silicon Valley-style innovation, launching two developer crowdsourced "Quests" that challenge scientists to create new ways to solve longstanding air travel and healthcare problems.



Announced at today's GE's "Minds and Machines" event, the two Quests are organized through the startup Kaggle and for the Hospital Quest will be

Kaggle: “The social data geeks fight club”

In 2010, Anthony Goldbloom took the SIGKDD and Netflix’s model and made it viable for all type of companies



+ **Investors** including PayPal co-founder Max Levchin, Google Chief Economist Hal Varian and Yuri Milner, put \$11 million into Kaggle last November.

Bloomberg Businessweek

Kaggle's Contests: Crunching Numbers for Fame and Glory

The startup's competitions lure PhDs and whiz kids to solve companies' data problems



TECH | 9/13/2012 @ 8:41AM | 1,420 views

World's Top Data Scientists Open Doors To Big-Cash Contests

+ [Comment now](#)

Xavier Conort is a French-born actuary who runs a consultancy in Singapore. When he's not assessing



The I²R team for GE Flight Quest



Xavier is a French actuary with more than 15 years of working experience in France, Brazil, China and Singapore. He gained experience in Machine Learning by running Gear Analytics, a Singapore-based consultancy and by competing on Kaggle and is currently the #1 ranked data scientist on Kaggle overall. Xavier earned a MSc in Actuarial Science and Statistics from Paris Tech and Paris VII University and is a Chartered Enterprise Risk Actuary. He recently joined the Institute for Infocomm Research, I2R, in Singapore. [Website »](#)



Hong Cao (Singapore) is currently a data analytics scientist in Institute for Infocomm Research (I2R) of Singapore's Agency of Science, Technology and Research (A*STAR), since 2011. He received the first-class honors B.Eng. and Ph.D degrees from Nanyang Technological University, Singapore, in 2001 and 2011, respectively. His current research interests include mobile data analytics, time series data mining, machine learning and multimedia forensics. His previous work in image forensics received the best paper award in IWDW 2010 and the honorary mention in ISCAS 2010. Recently, he also led teams to win international and local benchmarking challenges such as Opportunity Activity Recognition Challenge 2011 and Up-Singapore Hackathon 2012. He currently serves as secretary for IEEE signal processing society, Singapore section. [Website »](#)



Clifton Phua (Singapore) is currently a Senior Consultant in SAS Institute Pte Ltd, since Feb 2013. He was formerly a data analytics scientist in Infocomm Research (I2R) of Singapore's Agency of Science, Technology and Research (A*STAR) for more than 5 years. Clifton has published more than 30 technical papers on various industry applications of data analytics, such as fraud detection and activity recognition. Recently, his teams have won challenges such as Fraud Detection in Mobile Advertising Competition 2012 and Up-Singapore Hackathon 2012. Clifton's PhD and bachelor's (first class honors) degrees are from Clayton School of Information Technology, Monash University, Australia. [Website »](#)



Ghim-Eng Yap (Singapore) is a Principal Investigator with the Institute for Infocomm Research (I2R) in the Agency for Science, Technology and Research (A*STAR), Singapore. He received his Bachelor Degree (First Class Honors) from Nanyang Technological University (NTU) of Singapore, and was awarded the A*STAR Scholarship to complete his PhD degree from NTU in 2009. During the course of his work, Ghim-Eng has led multiple I2R R&D projects in the areas of recommender systems and privacy preserving audience measurement, and he is currently the I2R Principal Investigator for Data Privacy. Recently, his team won the First Prize in the Fraud Detection in Mobile Advertising Competition 2012. As a Scientist and Inventor, Ghim-Eng has published widely in various areas of data analytics including recommender systems and Bayesian statistics, and he holds a A*STAR patent on a method for privacy-preserving data aggregation.



Hon Nian "Kenny" Chua (Singapore) is currently a data analytics scientist in Institute for Infocomm Research (I2R) of Singapore's Agency of Science, Technology and Research (A*STAR) since March 2012. Prior to that, he was a postdoctoral fellow at Roth Lab in the University of Toronto and the Harvard Medical School, where he worked on applications of machine learning in biology and biotechnology. During his PhD he worked on the prediction of protein functions, complexes, and interaction networks, and led a team that won a protein subnetwork prediction challenge at the DREAM2 conference in 2007. He was awarded the A*STAR graduate scholarship in 2003 and the A*STAR postdoctoral fellowship in 2008. He received his bachelor degree and his PhD from the National University of Singapore.

What convinced us to enter the competition



Novelty of the data

flight data
usually not
available to
the public!

- **Complex:** many influencing factors
- **Huge \$:** 22 billion per year
- **Potential to benefit** to multiple parties:
 - **Airline:** reduced buffer time, better fuel planning
 - **Airport:** reduced gate congestion and idle time in crew management
 - **Passengers and logistic companies:** saving of travelling time



The Flight Data

- **Big** with over **300 GB** uncompressed;
- **87 days** of concurrent flight data for the **entire set of US domestic flights**;
- Over **20,000 flights per day**;
- Each flight associated with large amount of data in different types:
 - Flight history: scheduled/actual flights stats and events;
 - ASDI: flight route plan, actual trajectories, etc.
 - ATSCC: airport delays, deicing, ground delays, etc.
 - METAR: Weather reports by the weather stations
 - Other Weather: weather phenomena, wind, turbulence, etc.
 - Info. on weather stations

GE Flight Quest - Phase 1 (179 teams, 242 participants, 3073 entries)

I²R's Winning Formula

Competition Data 26,000 US domestic flights and weather data x 87 days

Feature Engineering



258 Features Extracted

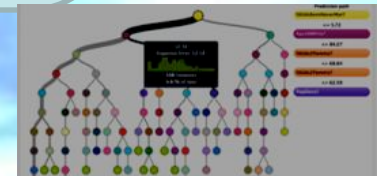
Feature Selection



84 Features Selected Only 58 used for predicting Runway Arrival Time

Machine Learning

Mixture of Prediction Models GBM and RF models



Prediction of Runway and Gate Arrival Time on Test Data (unseen)



Winning Results



Average 40%-45% less errors for gate and runway arrival time, respectively, compared to the standard industry benchmark



Statistics on very recent flights

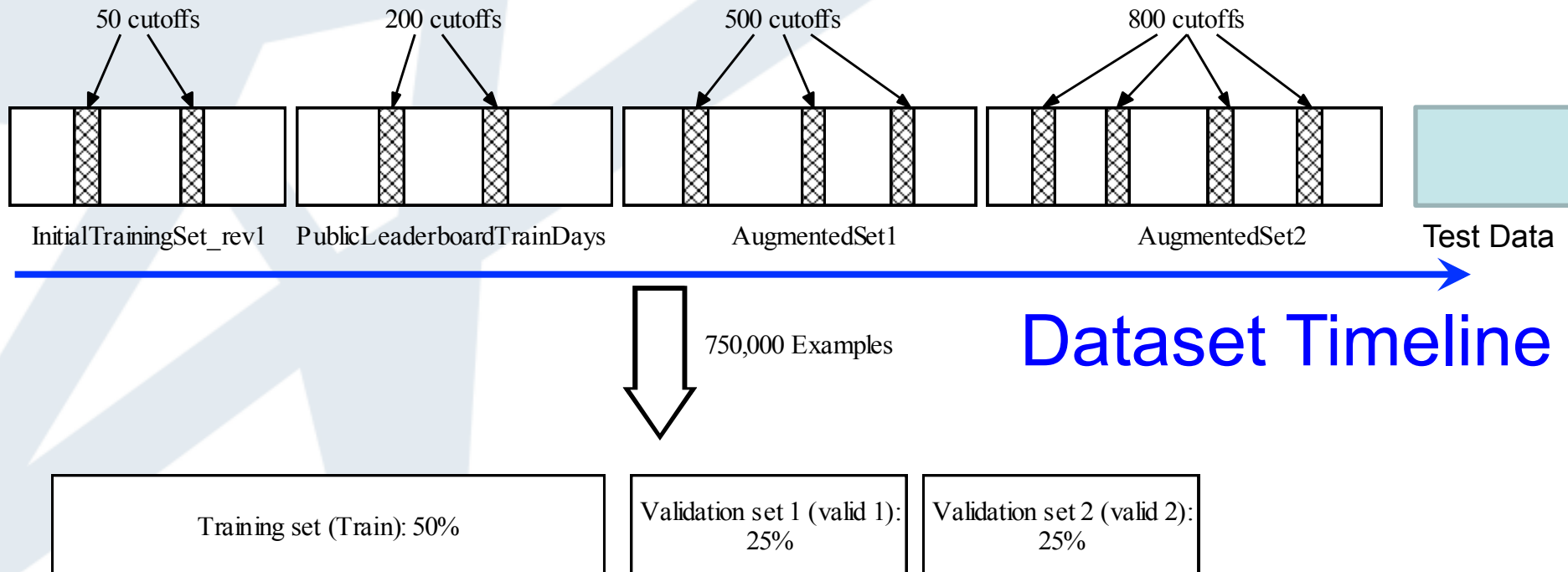


ined



Setup for Training and Validation Data

- 4 Flight datasets for our modelling and validation
- 1550 cut-offs with 750,000 flight examples
- We engineer and compute 258 features for each example
- Random 50% for training, 25% each for validation 1 and 2



Features Selection

We select **58 features** for predicting **runway arrival time** and **84** for predicting **gate arrival time** including

- **Statistics and attributes of the flight** (type of aircraft, airline, scheduled times, delay at departure, route distance, route changes, groundspeed, altitude, gate assigned...)
- **Statistics of the very recent flights**, that arrived at/ left the arrival airport one hour before cutoff or scheduled at similar period as the current example
- **Weather conditions**
 - weather conditions during the flight
 - weather at arrival (current and forecasts)

Our modelling strategy

Estimation of runway arrival time

$$E2RA = \max(\text{cutoff}, ERA - \text{ERA_error})$$

Estimation of gate arrival time

$$EGA = \max(\text{cutoff}, E2RA + \text{pred_taxi_time})$$

We modelled and predicted these 2 terms:

- **ERA_error** = ERA (the latest Estimated Runway Arrival information in flight history events before the time of *cutoff*) - actual runway arrival time
- **taxi_time** = actual gate arrival time – runway arrival time

Machine Learning Algorithms

We used tree-based ensemble approaches:

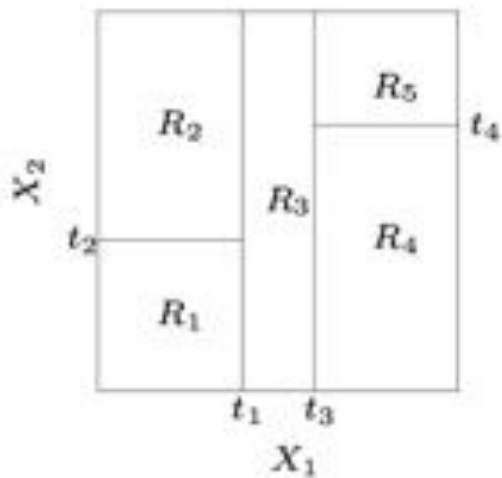
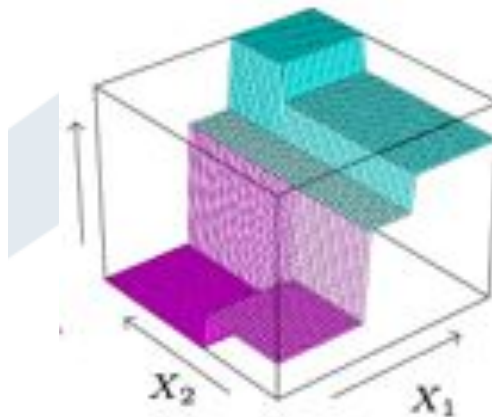
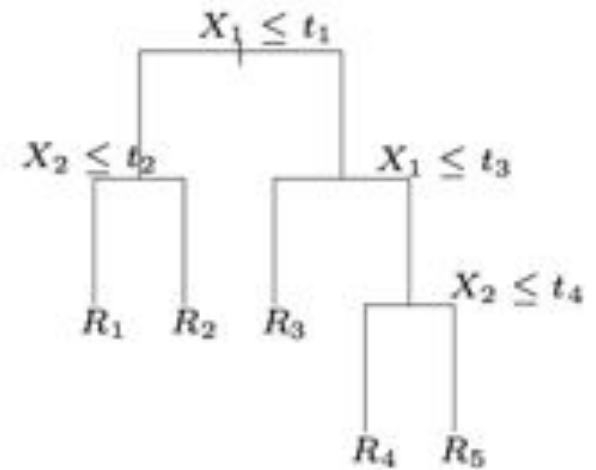
- ***Random Forest*** – Bagging of Decision / Regression Trees. Pioneered by Leo Breiman in 2001
- ***Gradient Boosting Machines*** – Boosting with Regression Trees. Pioneered by Jerome Friedman in 1999

How did we choose Machine Learning Algorithms?

- Our choice was motivated by the fact that both Gradient Boosting Machine and Random Forest can work with both categorical and numerical features, can capture automatically complexity in the data that we were not aware of, are not sensitive to outliers or monotone transformation and require little effort to fine tune the hyper parameters
- The additional strength of GBM over Random Forest in R is that it can work with missing data and categorical features with high number of categories.
- The key weakness of GBM is it is slow and difficult to parallelize unlike Random Forest.

Regression trees are good to detect interactions

Regression trees which partition the feature space into a set of rectangles and then produce a multitude of local interactions.

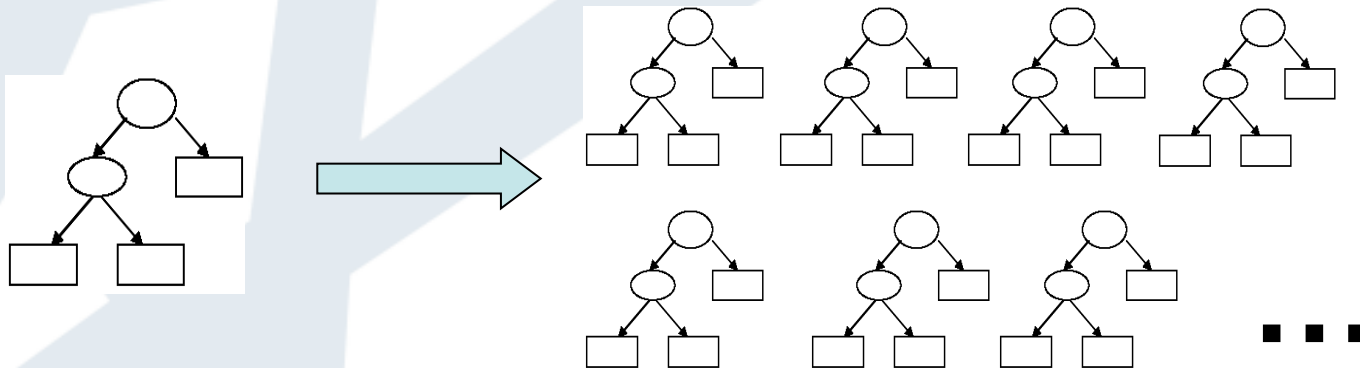


How does Gradient Boosting Machine work?

- Gradient Boosting Machines use boosting and decision trees techniques:
 - The boosting algorithm learns step by step slowly and gradually increases emphasis on poorly modelled observations. It minimizes a loss function (the deviance, as in GLMs) by adding, at each step, a new simple tree whose focus is only on the residuals
 - The contributions of each tree are shrunk by setting a learning rate very small (and < 1) to give more stable fitted values for the final model
 - To further improve predictive performance, the process uses random subsets of data to fit each new tree (bagging).

How does Random Forest work?

An RF is a collection of regression trees such that each tree has been trained on a bootstrapped dataset with a random selection of the input variables

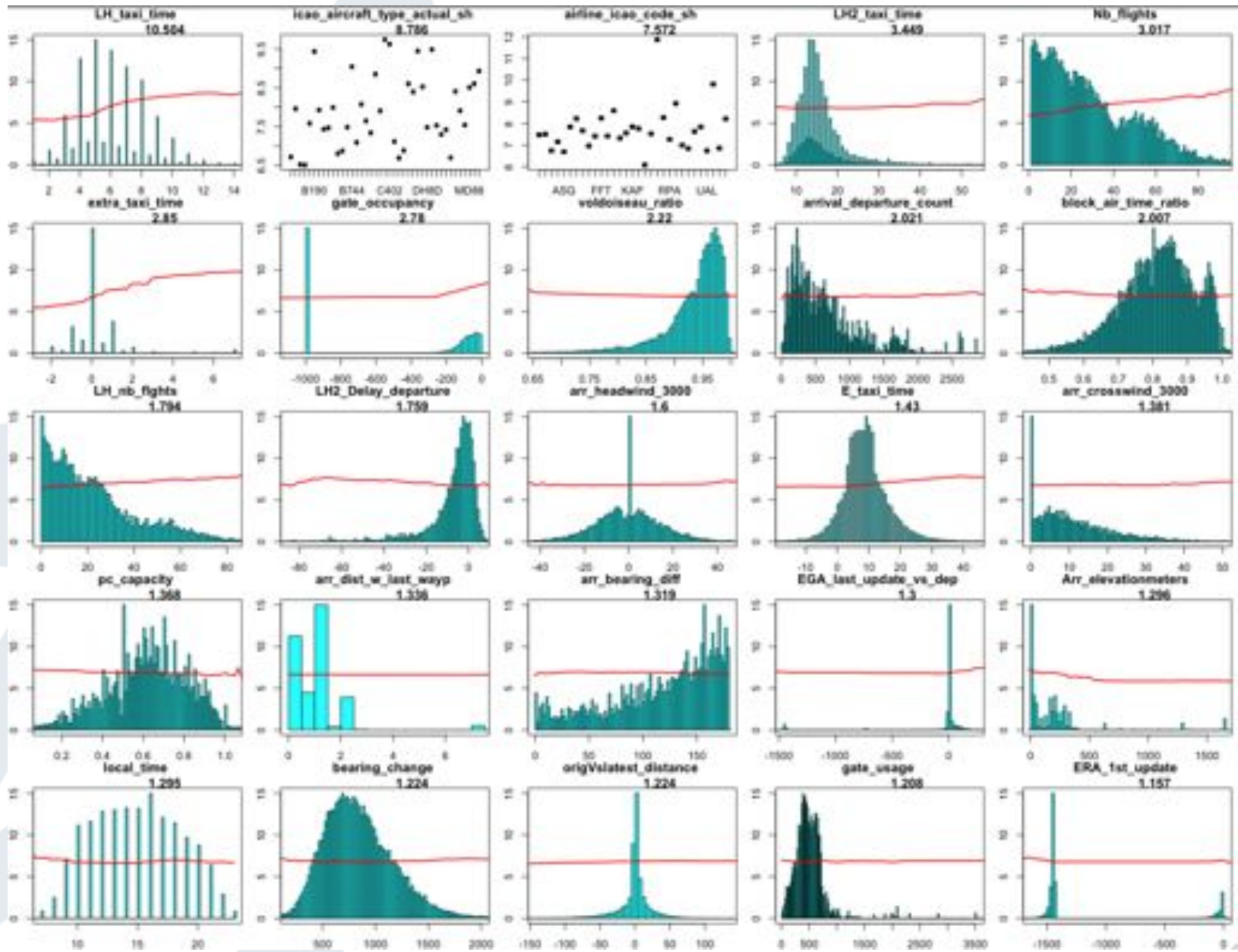


Top Features for Predicting Gate Arrival Timing (84 used our model)

$EGA = \max(\text{cutoff}, E2RA + \text{pred_taxi_time})$ → We model and predict this term

Feature Name	Importance (RF)	Description
LH_taxi_time	480,893	Median taxi time per airport one hour before cutoff (after having discarded redirected/diverted flights)
Nb_flights	223,006	Number of flights scheduled at the same airport and at the same hour as the example
icao_aircraft_type_actual_sh2	222,902	icao_aircraft_type_actual where categories with less than 900 examples are replaced by blank
airline_icao_code_sh	218,225	Airline_icao_code where categories with less than 100 examples are replaced by blank
extra_taxi_time	173,711	Median deviation of taxi_time at the gate level from the median at the airport level

GBM's Partial Dependence plots for taxi time





Predictive Modelling in the insurance industry

Current state

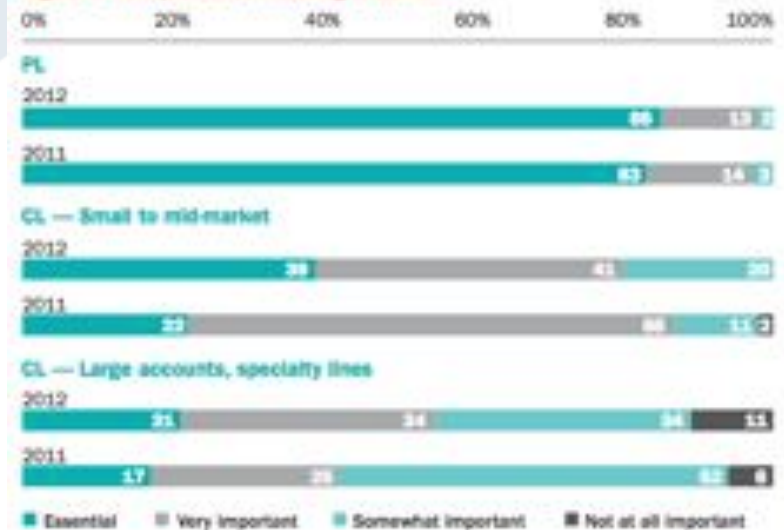
Insurers are very good at risk-centric analytics.

According to a Towers Watson survey in 2012:

- Almost all U.S. personal lines respondents indicate they view sophisticated underwriting and risk selection as essential or very important.
- the % of small to mid-market commercial participants that described it as essential grew by 17 points in 2012
- Smaller carriers face loss of market share and adverse selection as companies that have implemented predictive models target better risks and price more accurately.

- **The vast majority of carriers continue to rely on generalized linear models (GLMs). A majority of carriers indicate that they have adopted or will soon adopt principal components analyses and decision-tree analyses.**

Figure 1. How predictive modeling is viewed



Source: 2012 Predictive Modeling Benchmarking Survey: Advances in Implementation published by Towers Watson

Increased attention for Big Data in the insurance industry



- Big Data analytics are used in many industries to:
- crunch rapidly a wide variety of data (structured and unstructured data)
 - detect automatically complex, non-linear signals

Insurers increasingly recognize benefits of Big Data solutions over traditional approaches. It enables them to:

- increase predictive accuracy
- shorten the modeling process cycle
- exploit data they discarded before
- exploit new data and opportunities faster and keep ahead from competitors

One well known Big Data problem in insurance



Telematic applications (known as Usage Based insurance or Pay As You Go) are becoming increasingly important for auto insurers in US and UK although in the early stages of telematics, insurers leveraged only a tiny portion of the massive amounts of data available.

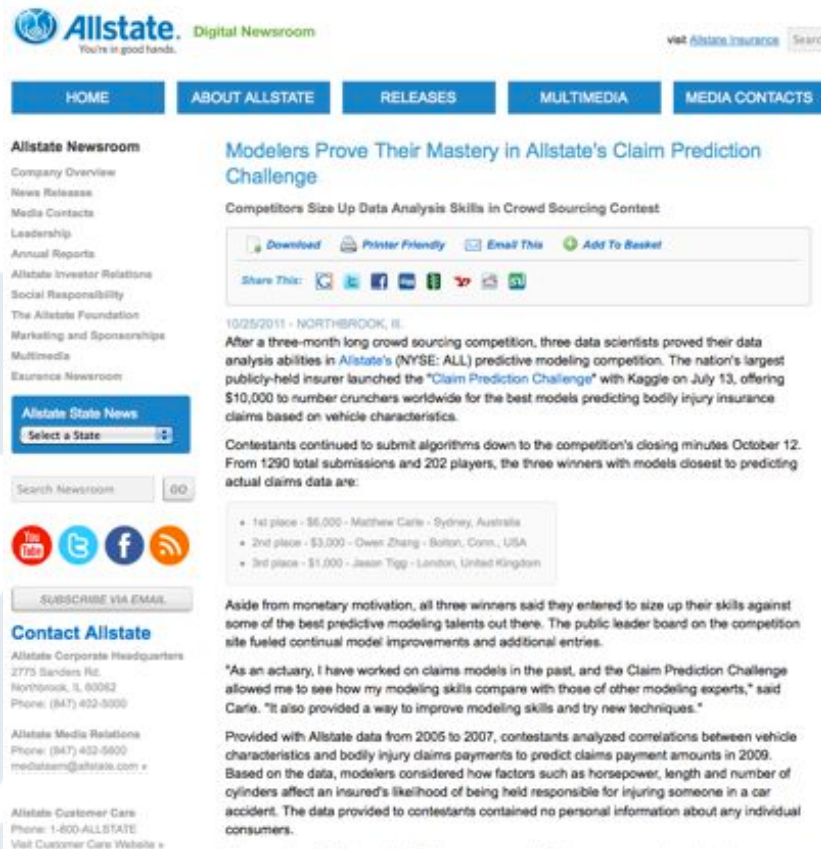
It is very likely that we will see in the future

- More Pay As You Go products thanks to cheaper implementation cost (proliferation in the use of sensor networks and low-cost communications) and increase awareness of customers
- More Advanced Big Data solutions in insurers to fully exploit information collected such as driver behavior, vehicle performance, and location factors.

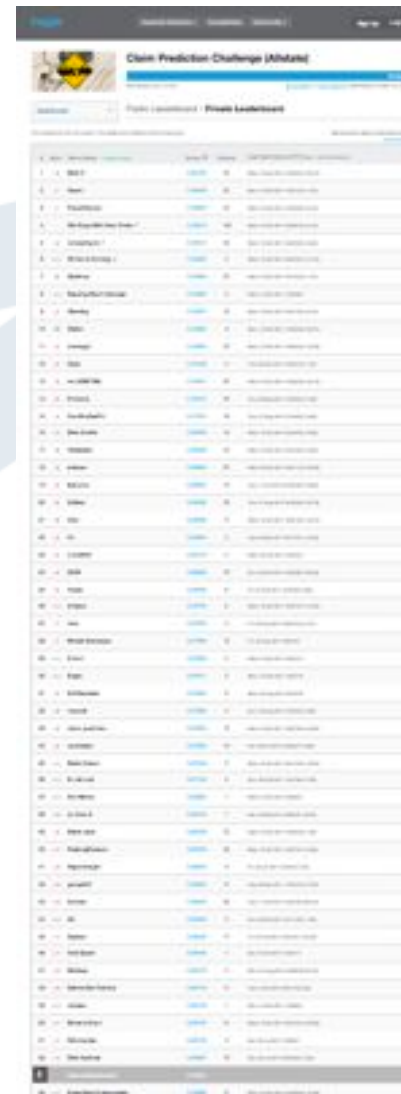


Data Analytics for pricing and underwriting

Machine Learning vs Allstate's benchmark in predicting bodily injury claims



The screenshot shows the Allstate Digital Newsroom website. The header includes the Allstate logo and navigation links: HOME, ABOUT ALLSTATE, RELEASES, MULTIMEDIA, and MEDIA CONTACTS. The main content area features an article titled "Modelers Prove Their Mastery in Allstate's Claim Prediction Challenge" with a sub-headline "Competitors Size Up Data Analysis Skills in Crowd Sourcing Contest". The article text describes a three-month crowd sourcing competition where data scientists predicted bodily injury insurance claims based on vehicle characteristics. It lists the winners: 1st place - \$6,000 - Matthew Carle - Sydney, Australia; 2nd place - \$3,000 - Owen Zhang - Bolton, Conn., USA; 3rd place - \$1,000 - Jason Tigg - London, United Kingdom. A quote from Matthew Carle is included: "As an actuary, I have worked on claims models in the past, and the Claim Prediction Challenge allowed me to see how my modeling skills compare with those of other modeling experts." The article also mentions that contestants analyzed correlations between vehicle characteristics and bodily injury claims payments to predict claims payment amounts in 2009. The footer contains contact information for Allstate Corporate Headquarters, Allstate Media Relations, and Allstate Customer Care.



The screenshot shows the Kaggle Claim Prediction Challenge (Allstate) leaderboard. The table lists the top 50 participants, their scores, and their names. The top three participants are:

Rank	Score	Participant
1	0.20156	Matthew Carle
2	0.19875	Owen Zhang
3	0.19875	Jason Tigg

Machine Learning (Gradient Boosting Machine, ...) Best score: 0.20156

GLM: Allstate's benchmark score achieved with same data 0.05933

Use of non traditional data in UW



Insurance data

Very personal finance

Marketing information offers insurers another way to analyse risk

Jun 2nd 2012 | From the print edition

A SWISS life-settlements firm called Rigi Capital Partners (RCP) recently considered buying the life-insurance policy of an elderly woman apparently suffering from dementia. RCP would take over payment of the policy premiums and receive the full death benefit when she passed away. Her medical records revealed that she had forgotten even her son's birthday. But Robin Willi, RCP's owner, searched Facebook to find out more about her. Her profile suggested she had a vibrant social life, not dementia. Reckoning that she was much healthier than she wanted to appear, Mr Willi did not offer to buy her policy.

Medical records can be deceptive, incomplete or expensive to analyse. Mr Willi thinks that publicly available data will be increasingly useful in helping insurers distinguish the aerobics enthusiasts from the couch potatoes. His firm is small enough to be able to search for data manually. For bigger insurers, software is now being developed by technology firms such as Allfinanz and TCP LifeSystems to sift through all the marketing data that might help them identify tomorrow's cancer patients or accident victims.



Make sure you pay in cash.

Underwriting (UW) in Health Insurance can be costly and time consuming

According to Tim Hill of Milliman, “at least two big American life insurers already waive medical exams for some prospective customers partly because marketing data suggest that they have healthy lifestyles”

According to the Economist, those companies exploit data aggregated by marketing firms about individuals from records of things like prescription-drug and other retail sales, product warranties, consumer surveys, magazine subscriptions and, in some cases, credit-card spending !



Data Analytics for customer analysis

Opportunities to gain critical insights on customers

Insurers have access to a large amount of data :

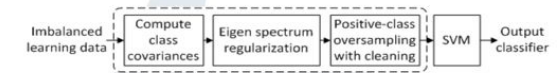
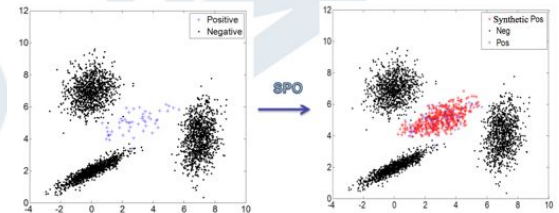
- Internal information scattered across organizational silos: CRM, policy, billing, claim systems, complaints, contact center logs, customer survey...
- external data: bancassurance partners data, social media sites, customer comments, web navigation, demographics, data aggregated by marketing firms

This information could help them to develop a much deeper understanding of:

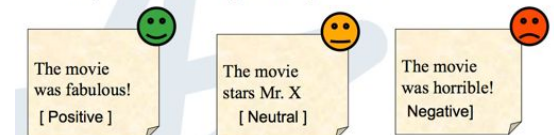
- what the customer wants, needs, thinks and
- how he will behave in the future

Several I²R techniques could help in this matter

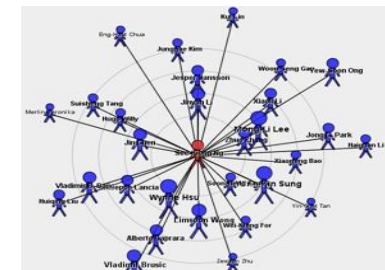
- Classification techniques:
 - to model customers retention
 - and identify customers most likely to respond to specific campaigns
- Text mining: to do sentiment and opinion mining
- Collaborative filtering: to develop recommendation engine for cross selling activities
- Sequential pattern mining: to capture value from transactional data
- Social Network Analysis: to identify those most influential people



- Identify the orientation of opinion in a piece of text



- Extract features/aspects of opinion targets.





Data Analytics for P&C Back Office

Invaluable information buried in Back Office systems



FROM THE MARCH 2012 ISSUE OF CLAIMS MAGAZINE • [SUBSCRIBE!](#)

Text Mining In the P&C Back Office

Manage Claims and Customers with Critical Insights

BY KARTHIK BALAKRISHNAN, JANINE JOHNSON
February 29, 2012 • Reprints



In today's challenging global economy, insurers face the difficult task of controlling costs and allocating resources and capital effectively, all in the face of intense, rising competition. Generating additional business lift more efficiently is a high priority, and insurers are increasingly relying on new data sources and advanced analytics to do so.

Text mining is one such capability that property and casualty (P&C) insurers are increasingly using to uncover critical insights from unstructured data sources. Using text mining, carriers are able to identify high-impact opportunities selectively, control losses and costs, allocate resources, and optimize financial outcomes. In the back office, text mining allows them to gain a better



Complaints, location data, adjuster notes, first notice of loss, emails, call center logs, medical records, police reports and accident descriptions can provide invaluable information to improve customer satisfaction and bottom line.

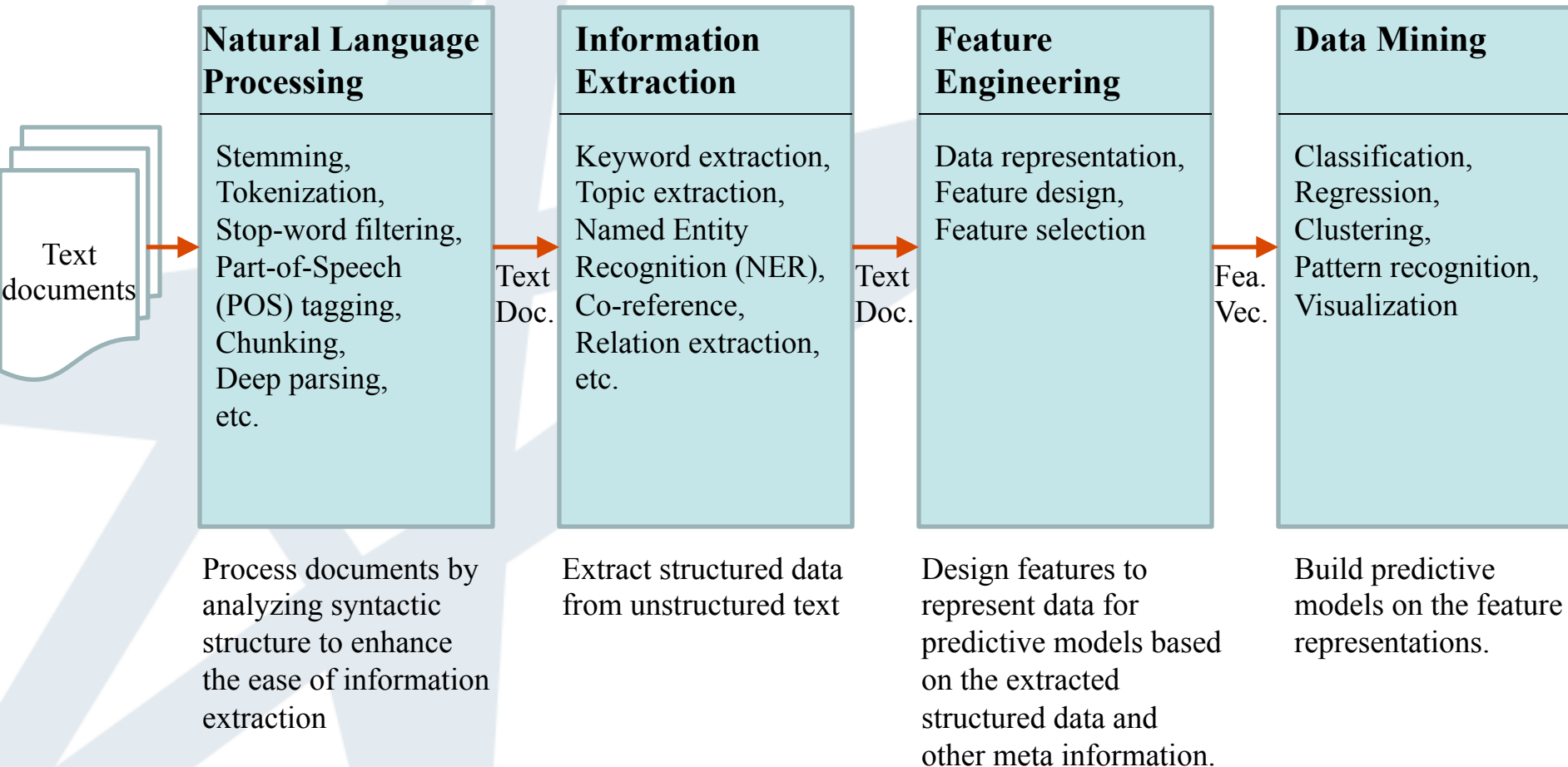
Text mining and predictive analytics can exploit this information for :

- resolving customers' issues quickly
- combating fraud
- assigning the right claim adjusters to the right case
- minimizing missed recovery opportunities
- accelerating settlements
- reducing reserves uncertainty
- managing litigation wisely
- identifying dissatisfied claimants

Why text analytics is likely to bring value in claims handling

Claim adjuster	Text analytics and scoring engine	Benefit
Needs time to read claim files	Reads, sorts and scores claims file in a fraction of second	Accelerates claims settlement, assigns the right claim adjusters to the right case faster and increases opportunity to minimize claims and litigation cost
Sees only a fraction of the company's claims and might have difficulty in accessing or evaluating information not in the claims file	Learns from all claim files, connects the dots and can systematically collect relevant information outside claim files (such as UW info, service agreement with providers...)	Suspicious claims have better chance to be flagged, case reserves will be better set and risk of claims overpaid will be reduced
Has limited memory capacity, might be overloaded by other cases and be too unexperienced to conduct cost/benefit analysis	Remembers full claim history and treats all claims in a consistent manner	Improves the claim workflow and minimizes missed recovery opportunities

Components of Text Mining





Thank you!