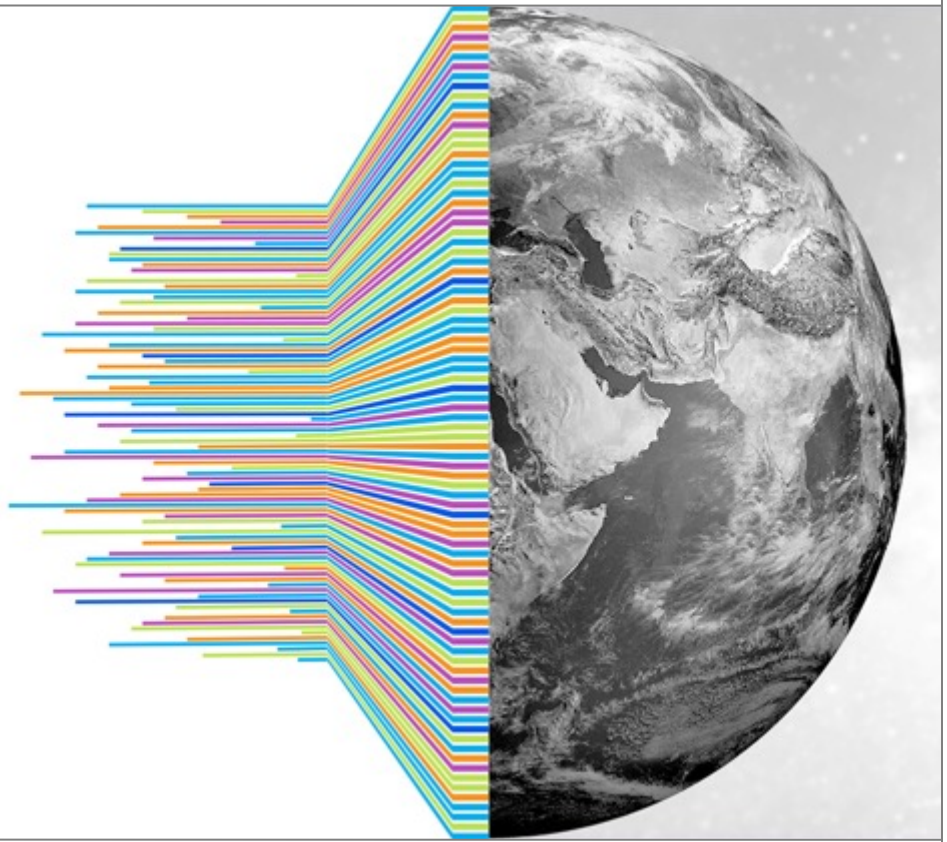
2016 ANALYTICS CHALLENGE
Finalist Presentations - Part 2



Yenny Yang

TERADATA
University Network

ANALYTICS CHALLENGE SESSIONS

- There are a total of 9 Analytics Challenge Finalists
  - 5 finalists presenting in Part ONE
  - 4 finalists presenting in Part TWO
- Each finalist will present for 5-7 minutes
- At conclusion of the session:
  - All teams will remain in the room for additional questions

TERADATA University Network

ANALYTICS CHALLENGE

TUN ANALYTICS CHALLENGE FINALISTS PART TWO

TERADATA PARTNERS CONFERENCE

A06 - University of North Carolina, Charlotte, NC, USA

A07 - University of North Carolina, Charlotte, NC, USA

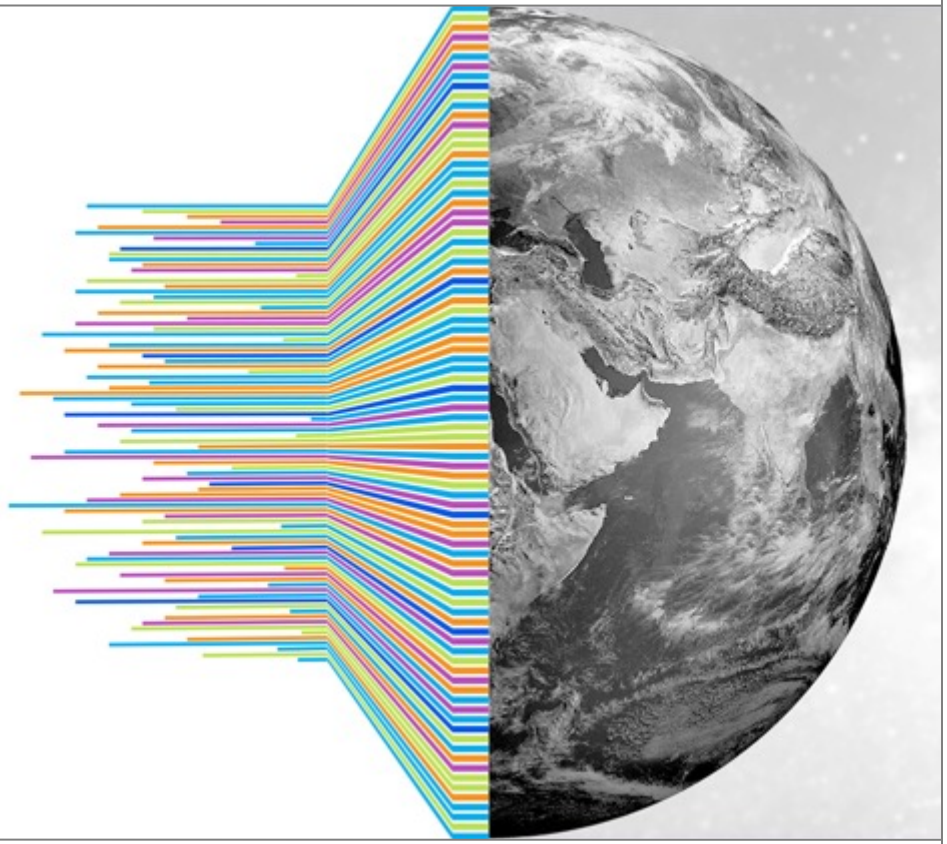
A09 - University of Cincinnati, OH, USA

A10 - University of Oklahoma, Norman, OK, USA

Big Data Analysis

Monisha Niharika Naga
Uddayasri Buddhi

A6 – University of North Carolina at Charlotte



Patents, Revenue & Elections

Objective: The project is intended to find insights from patent filed by the Fortune 500 companies in USPTO & EPO, how it influences the US Presidential Elections in the country.

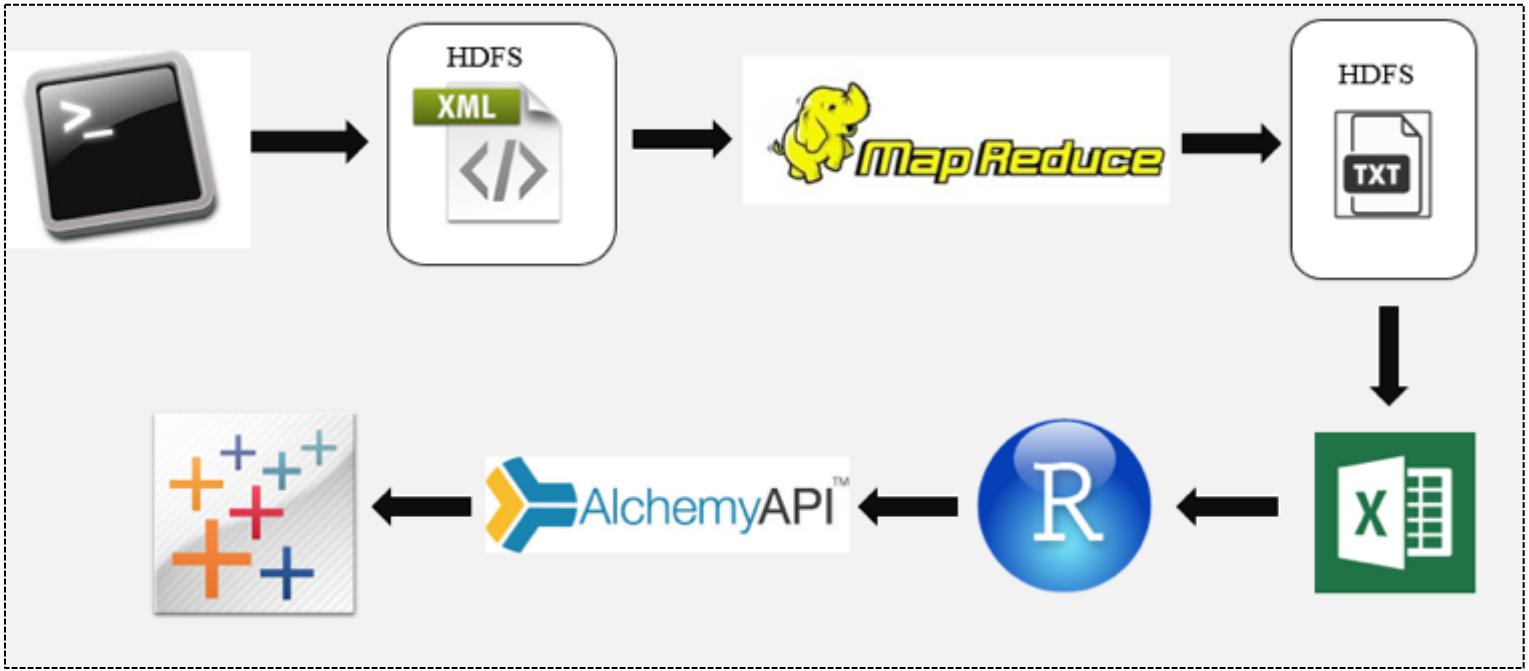
Data Extraction and Wrangling

To achieve our project objective, we require two main datasets.

- Fortune 500 companies' information that includes Patent Published, Revenue, Corporate and Social Responsibility and Headquarter Location from 2005 – 2015.
- Presidential Election information from 1952 – 2012 and Election Sponsorship information.

Data Collection

Process Flow:



Data Accuracy & Descriptive Statistics

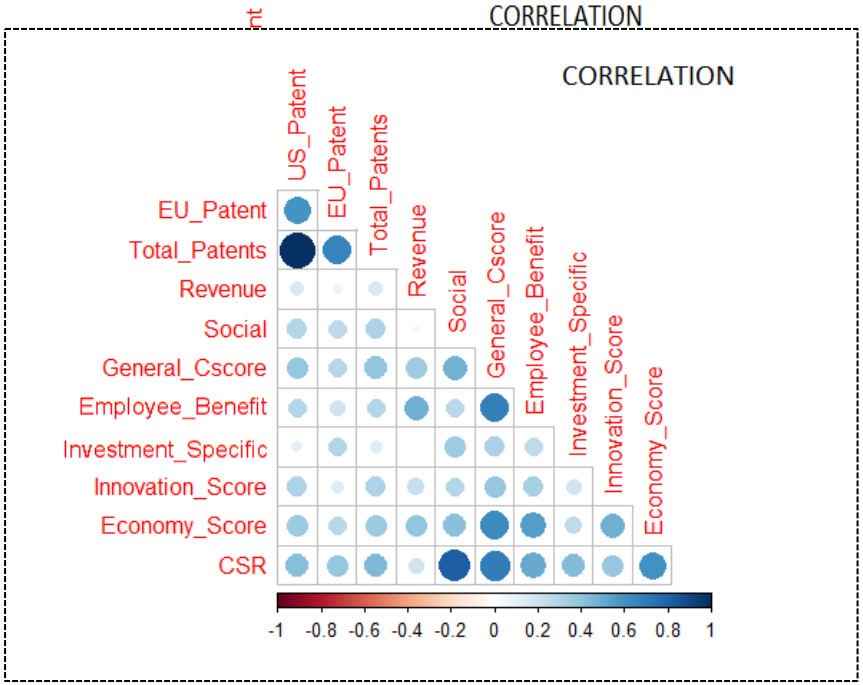
- Accuracy of the data collected using our map-reduce function is **97 – 99%**

CATEGORY	MEAN	STD DEV	MAX
Patent Count	1068	4234	**IBM** - 59720
Revenue	241.28 B	410.11	**Walmart** - 41330.64 B
Social Score	0.35	0.6664	**Xerox Corp** – 4/5
Corporate Score	0.51	0.6382	**General Electric** – 3.4/5
Employee Benefit	0.509	0.6231	**P&G** – 3/5

Relationships

TERADATA PARTNERS CONFERENCE

Correlation:



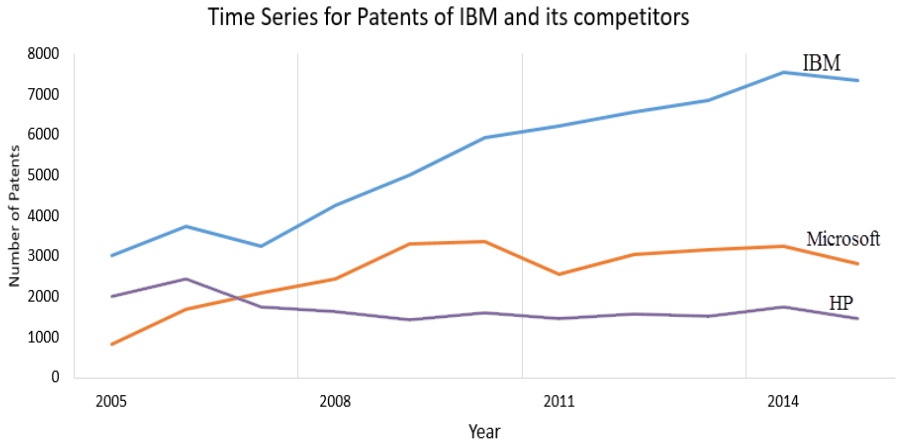
Cluster Analysis:

CLUSTER	IRRESPONSIBLE	SWING	RESPONSIBLE
PROFILE	1	2	3
Apparel and Sales	146.7504917	167.5288	517.9270295
Chemicals and Drug	92.50213215	306.8911	221.2582626
Computing	101.1762586	557.9086	537.9694708
Engineering and Machinery	96.52497925	421.0325	250.4423682
Finance and Banks	140.8683099		1061.119667
Food Products	160.5847444	131.4233	299.5309695
HealthCare	245.1959155	497.347	454.4258627
HouseHold	70.6413415		
Miscellaneous	95.14471968	153.4511	178.931519
Oil, Minerals and Energy	179.8369298	205.4271	630.6325339
Real Estate	85.86601957		
Transportation	110.469451	306.7242	296.1088636

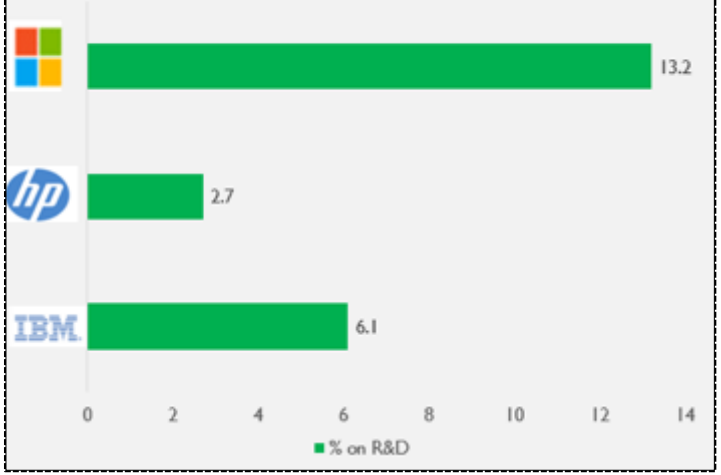
IBM-Highest Patent Count

TERADATA PARTNERS CONFERENCE

Time series for Patents of IBM and its Competitors:



Percentage of Revenue spent on R&D:

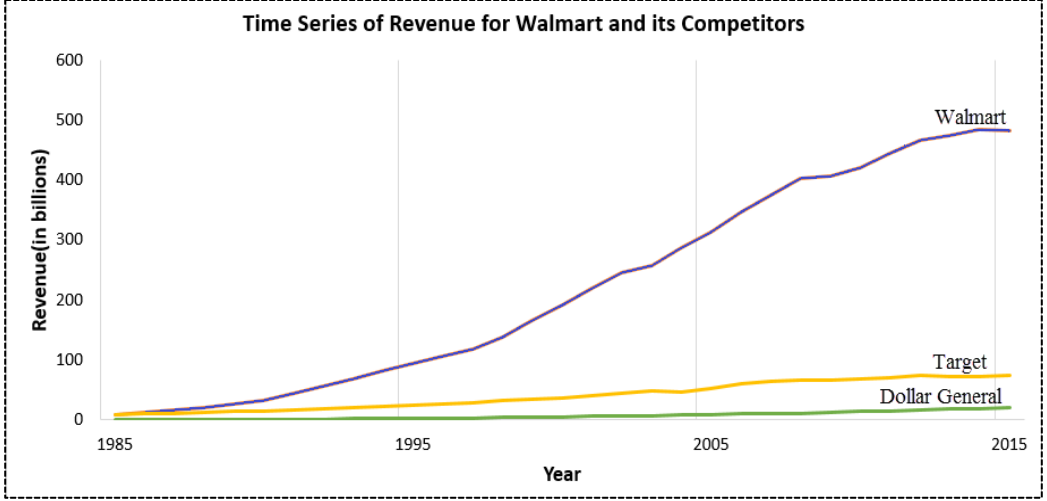


Data. Changes. Everything.

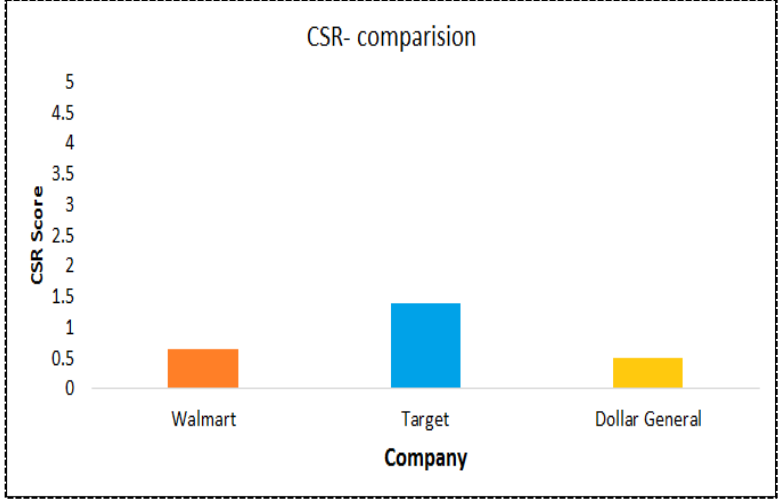
Walmart-Highest Revenue

TERADATA PARTNERS CONFERENCE

Time series of Revenue for Walmart and its Competitors:



Corporate Social Responsibility Comparison:



Data. Changes. Everything.

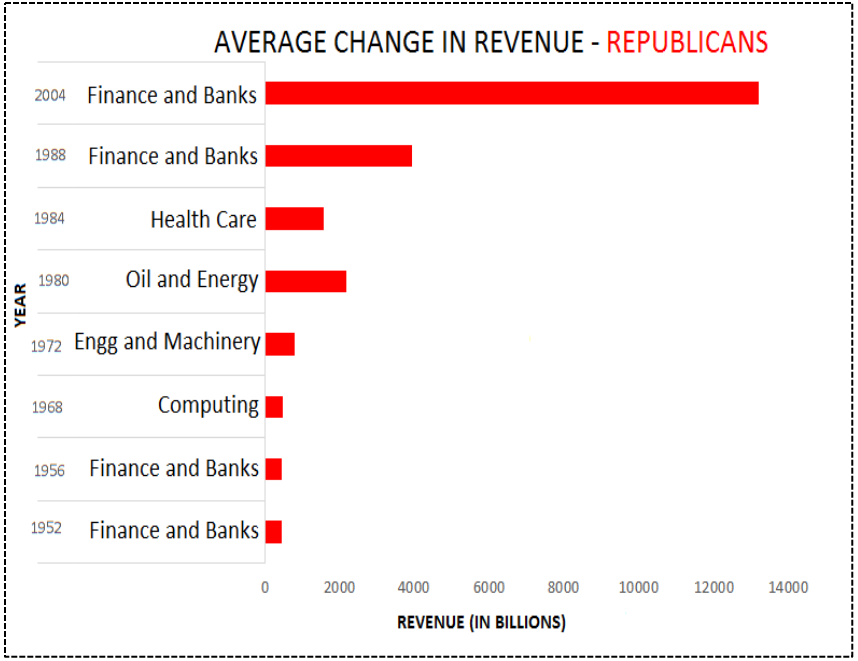
IMPACT OF US PRESIDENTIAL ELECTIONS - REPUBLICANS

TERADATA PARTNERS CONFERENCE

Effect:

- While comparing the impact of election in revenue for the Fortune 500 companies during the republican's period, **Financial and Banking** sector is benefitted.

- There are other sectors like Health Care, Oil and Energy, Computing and Engineering sectors are also benefitted during some terms.



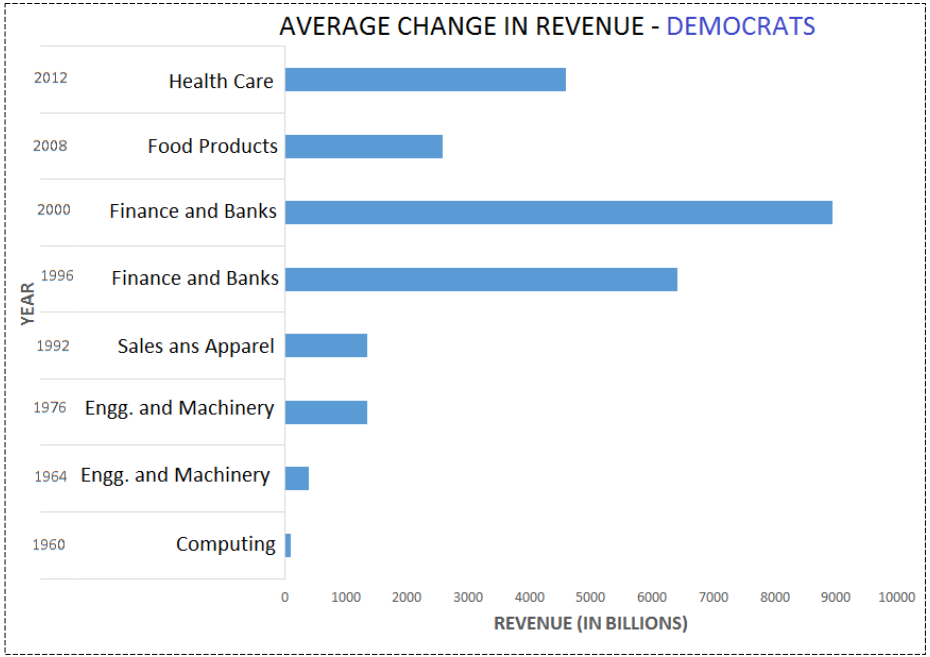
IMPACT OF US PRESIDENTIAL ELECTIONS - DEMOCRATS

TERADATA PARTNERS CONFERENCE

Effect:

- While comparing the impact of election in revenue for the Fortune 500 companies during the democratic period, **Financial and Banking** and **Engineering** sectors are benefitted most of the times.

- On Average though there is no sectors that benefitted always from the election in the previous 50 years, there are other sectors like Sales and Apparel, Computing sectors which are also benefitted during some terms.



Impact of US Presidential Elections

TERADATA PARTNERS CONFERENCE

- The data collected had less relationships between them.
- Some of the analytics techniques like factoring, associations did not much useful information .
- Election has some influence on the revenue of certain type of firms, however it depends on other factors also.

Thank You

TERADATA PARTNERS CONFERENCE

Questions/Comments

Email: mnaga@uncc.edu

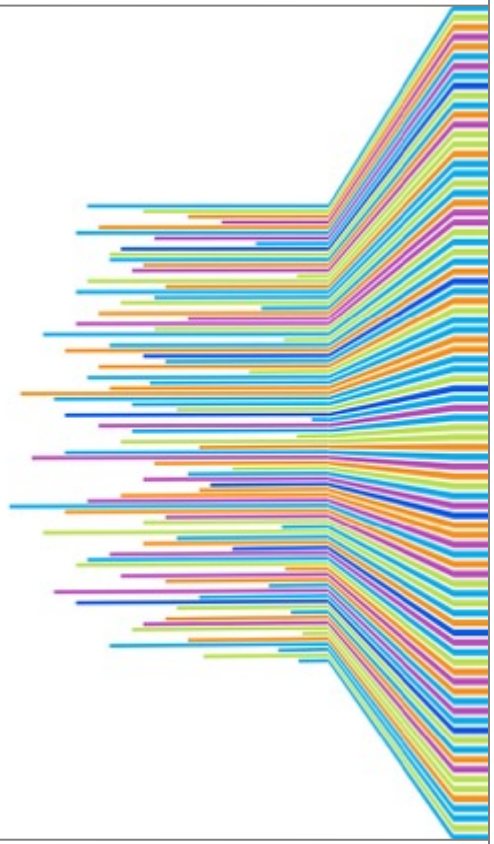
Follow Me

Twitter @

Rate This Session

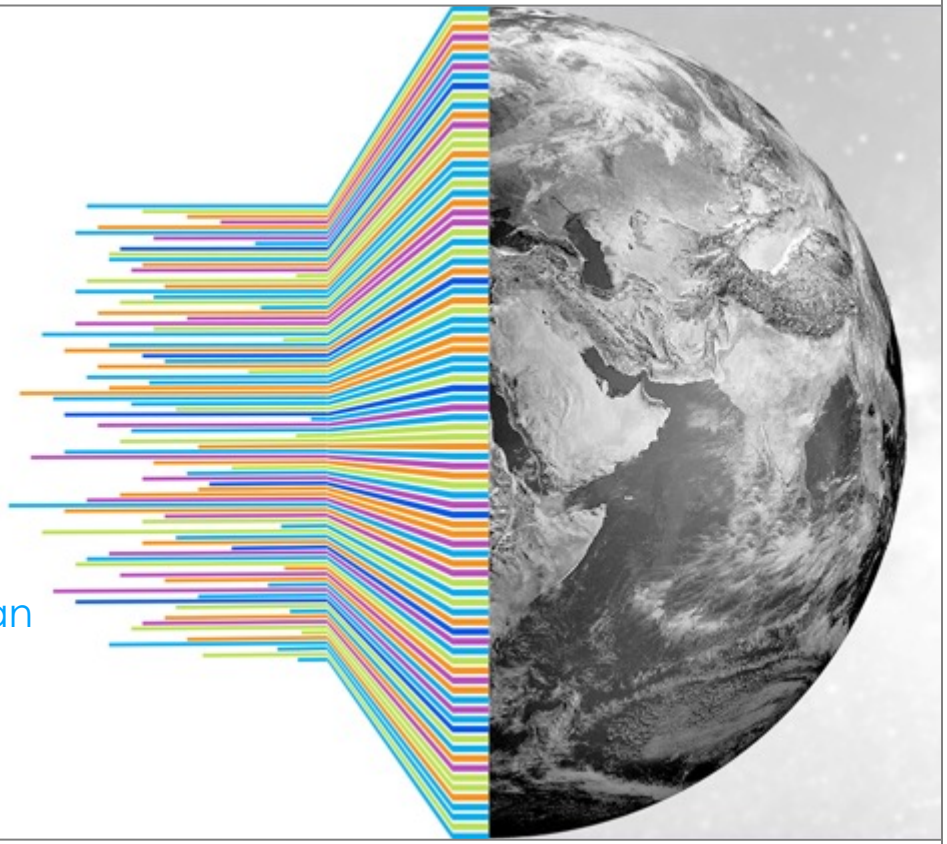
with the PARTNERS Mobile App

Remember To Share Your Virtual Passes



UNC CHARLOTTE

Smartphone User Trace Analytics



Anand Niranjan
Swaminathan Tirunelveli Vijayaraghavan

A7 – University of North Carolina at Charlotte

Problem Statement

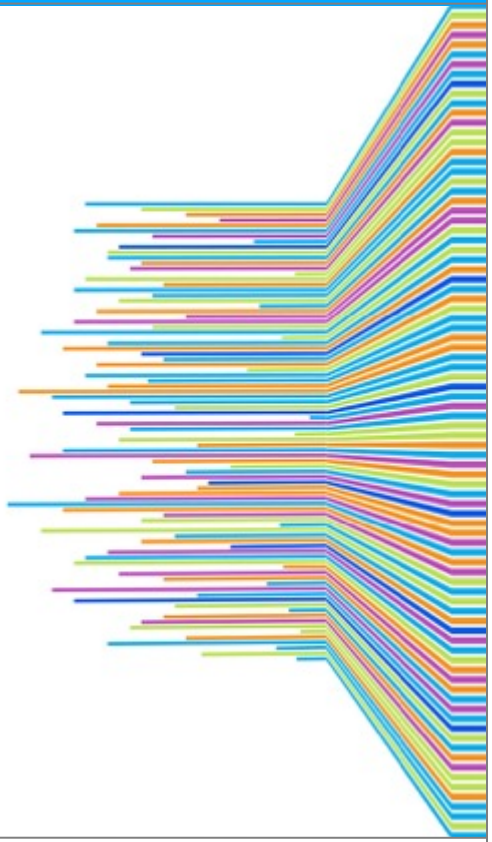
TERADATA PARTNERS CONFERENCE

Objective

- To achieve security in smartphones beyond single entry authentication

Challenging Tasks

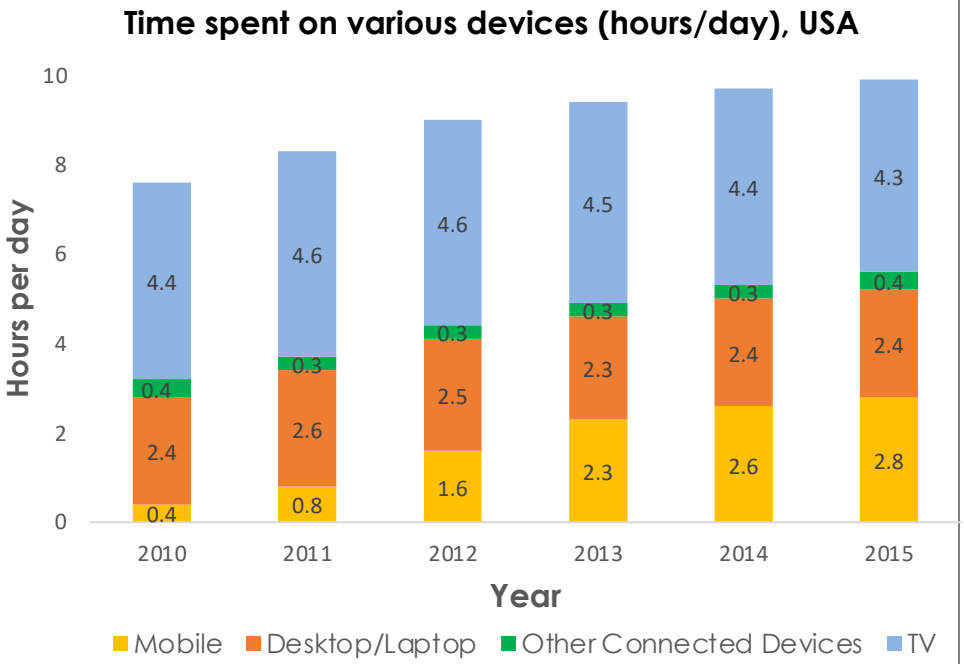
- Segregate actual user of smartphone from unauthorized users
- Analyze the patterns of user behavior on different smartphones
- Predict the accuracy of touch traces for validation
- Ensure minimum investment to develop this model



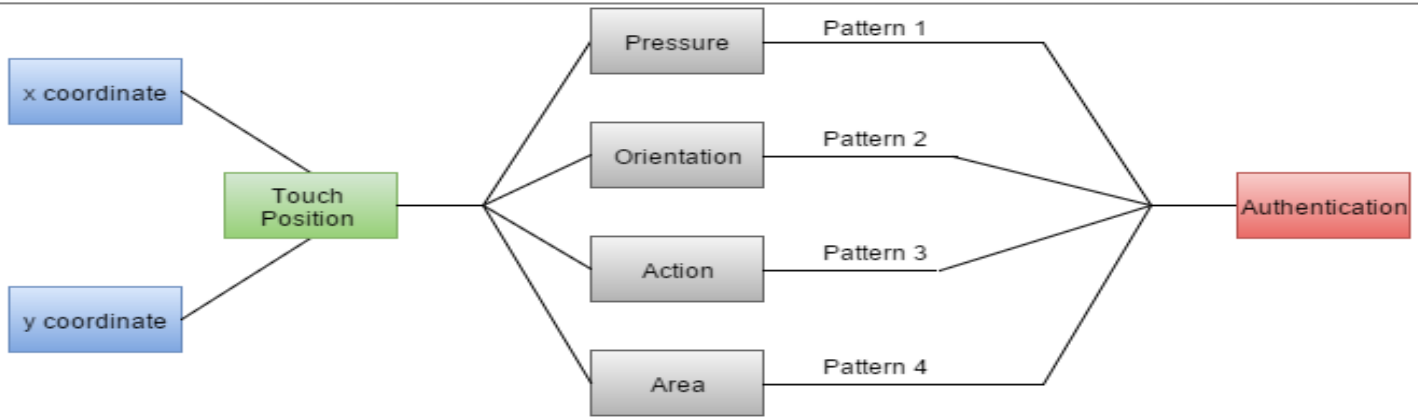
Smartphone & Security

Nowadays, security of one's smartphone in turn determines the security of his bank accounts, photos, documents, emails and so on.

- Stats shows mobile usage increases every year when compared to other devices.
- As storage of sensitive data increases in phones; breach & threat also increases
  - 1.02 billion records breached in 2014
  - 5.2 million smartphones were lost or stolen in the U.S. in 2014.



User Based Security in Smartphones



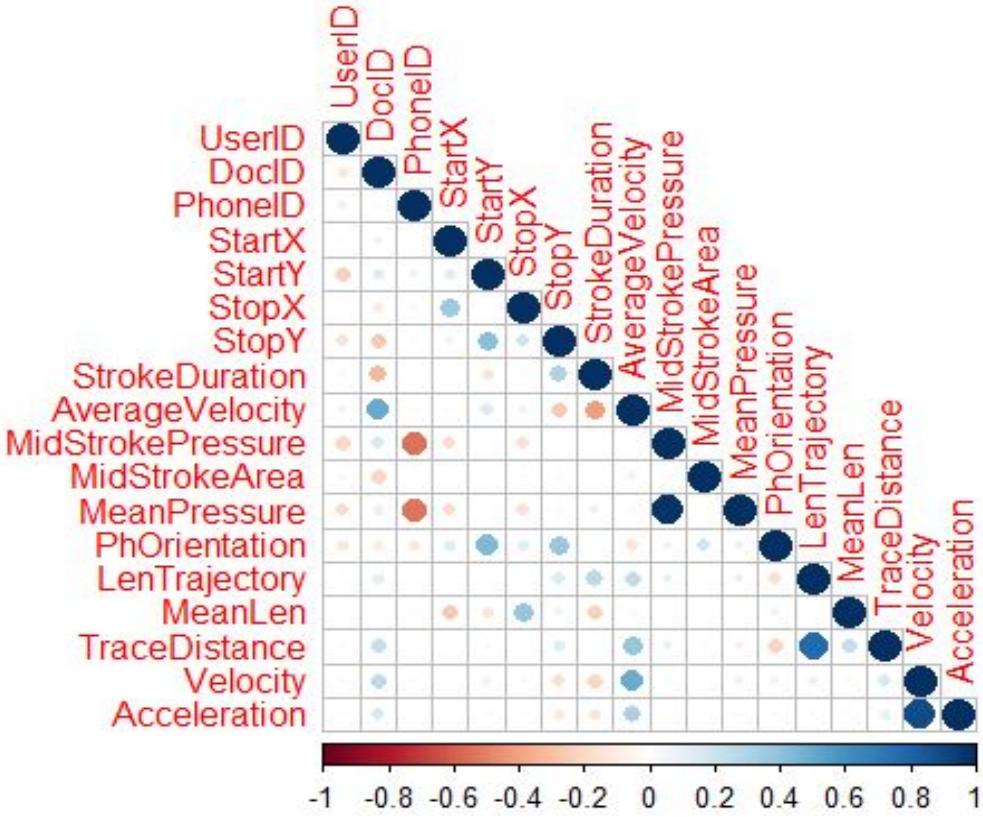
- Touch patterns of a particular user varies significantly that is sufficient to define his own security policy.
- Though this method is user specific, it varies from device to device for the same user.
- Pressure applied, phone orientation, area coverage by fingers differs for every user.
- The policies are formulated based on the usage of apps like Wikipedia article, image comparison game and answer questionnaire.

Feature Correlation & Inverse Relation

TERADATA PARTNERS CONFERENCE

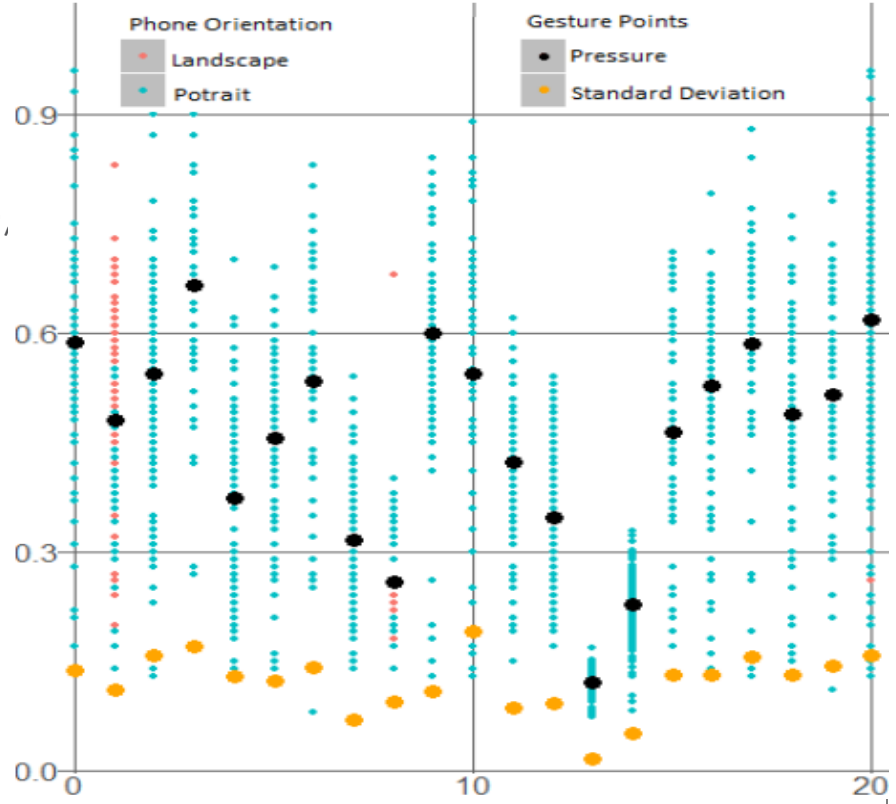
Features that can differentiate actual user of smartphone from unauthorized users:

- Position – x, y coordinates
- Pressure applied
- Area covered
- Length of trajectory
- Duration of stroke
- Velocity
- Acceleration
- Phone orientation

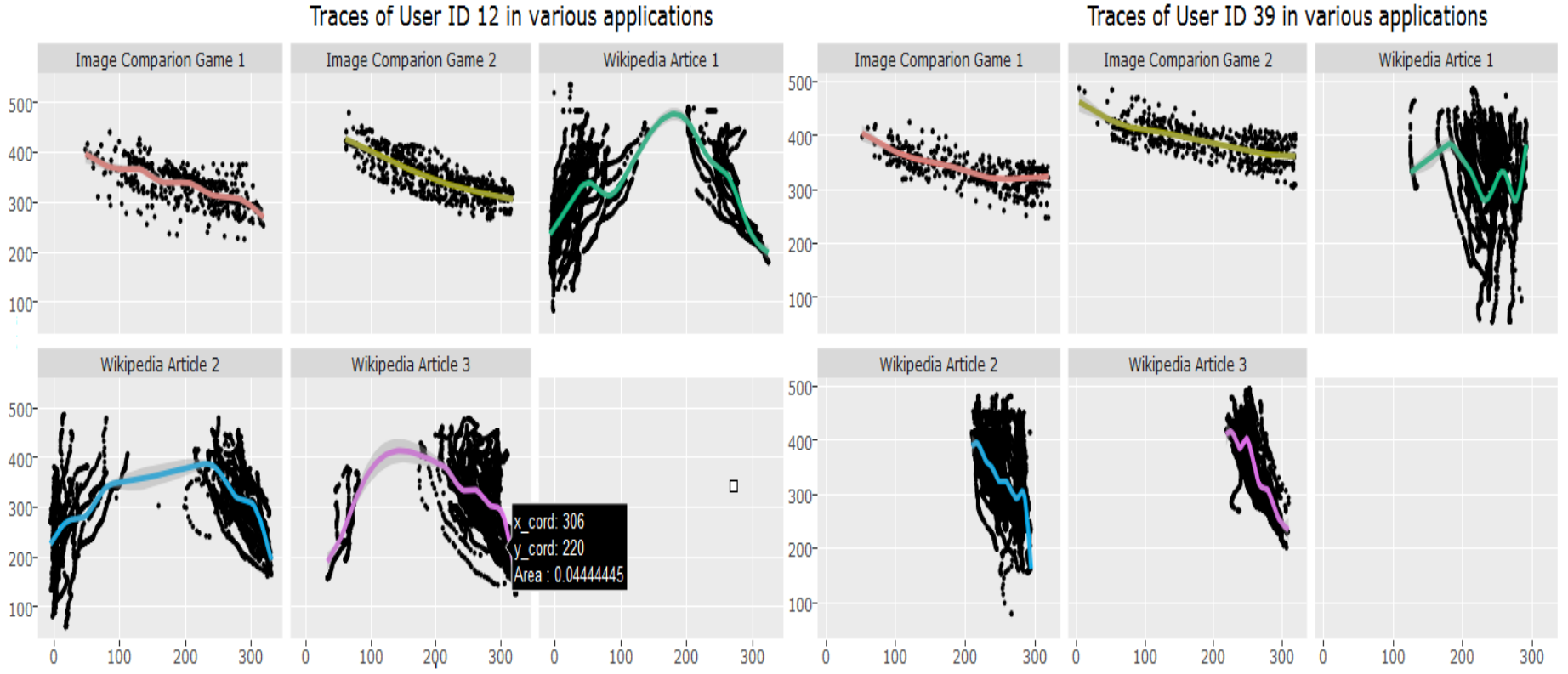


Touch ID and Touch Patterns

- "Average iPhone user unlocks device 80 times per day", says Apple.
- 89% of total unlocks comes from Touch ID, this shows the reliability of the security system.
- Apple has spent 32% more in R&D in 2013 (iPhone 5s released) compared to its previous year.
- Though touch sensors are the most efficient security mechanism, it is not suitable for all range of smartphones.



Analyzing Touch Traces of Users

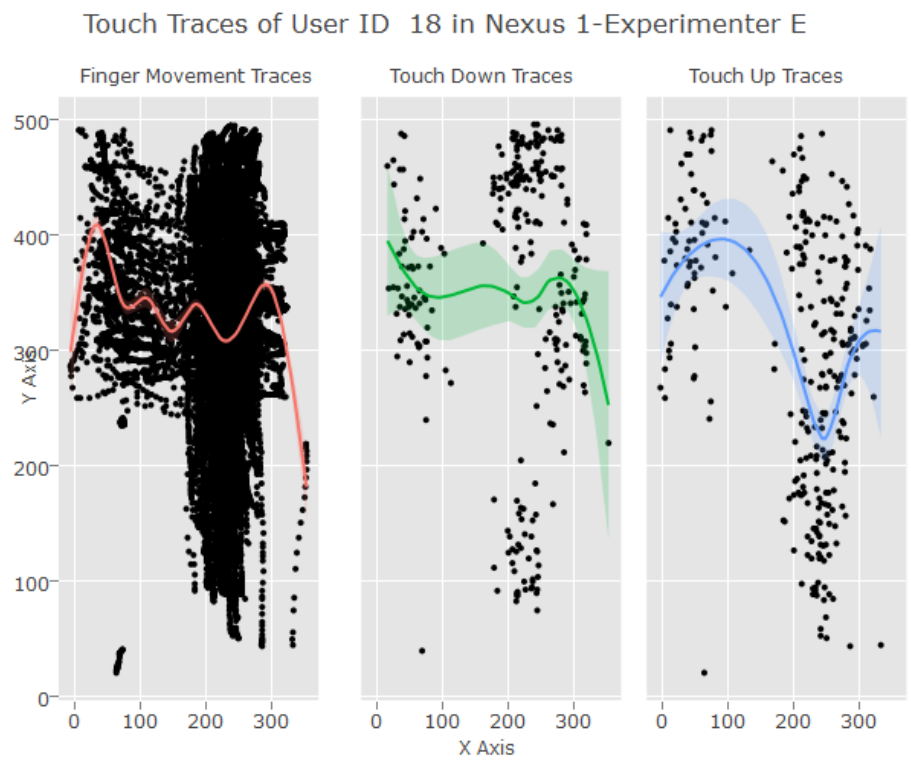


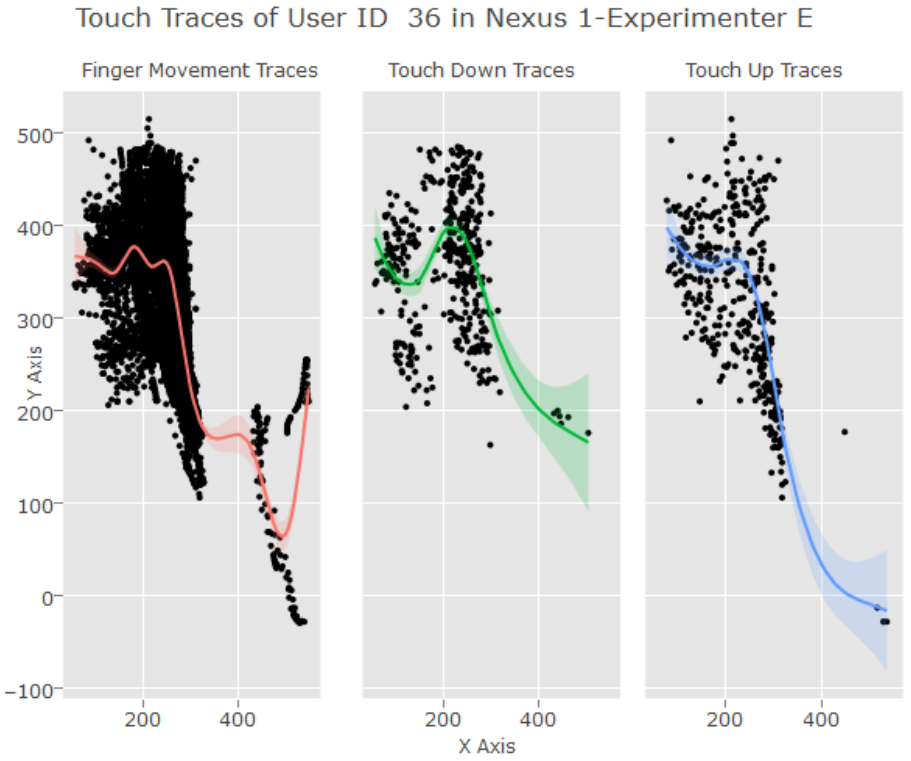
One for all – Security Policy

- The security policy defined for one user using his trace patterns, can be used for other applications.
- In recent times, many applications like drop box, google drive, bank of America and other sensitive apps use touch ID for authentication.
- But this is specific to devices, that has fingerprint sensors enabled. With our security method, these applications can use the touch pattern security policy.

Touch Traces Analysis in Smartphone

TERADATA PARTNERS CONFERENCE





Data. Changes. Everything.

Secure Your Data & Apps

The proposal has some limitations as in time to learn user behavior, additional storage & cost. We recommend the following to overcome:

- Store sensitive documents in a security app and secure by pin.
- Keep personal apps like e-mail, social media apps inside the security app and access from that app to keep data encrypted.
- Other functionality like remotely lock the device, remote factory reset and phone location tracking if lost or stolen

Conclusion

Validation

- Smartphone security can be achieved by continuously validating the touch traces of user's touch traces.

Feature Extraction

- Extracting more features is essential to improve the accuracy of authenticity so that the access of smartphone by unauthorized uses can be forbidden.

References

- http://www.mariofrank.net/paper/touchalytics.pdf
- Dataset - http://www.mariofrank.net/touchalytics/

Thank You

TERADATA PARTNERS CONFERENCE

Questions/Comments

Email: aniranja@uncc.edu

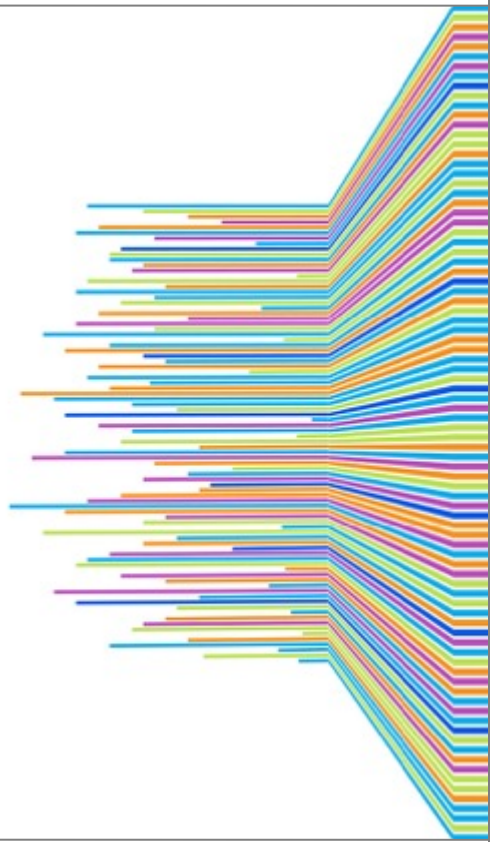
Follow Me

Twitter @ niranjan9

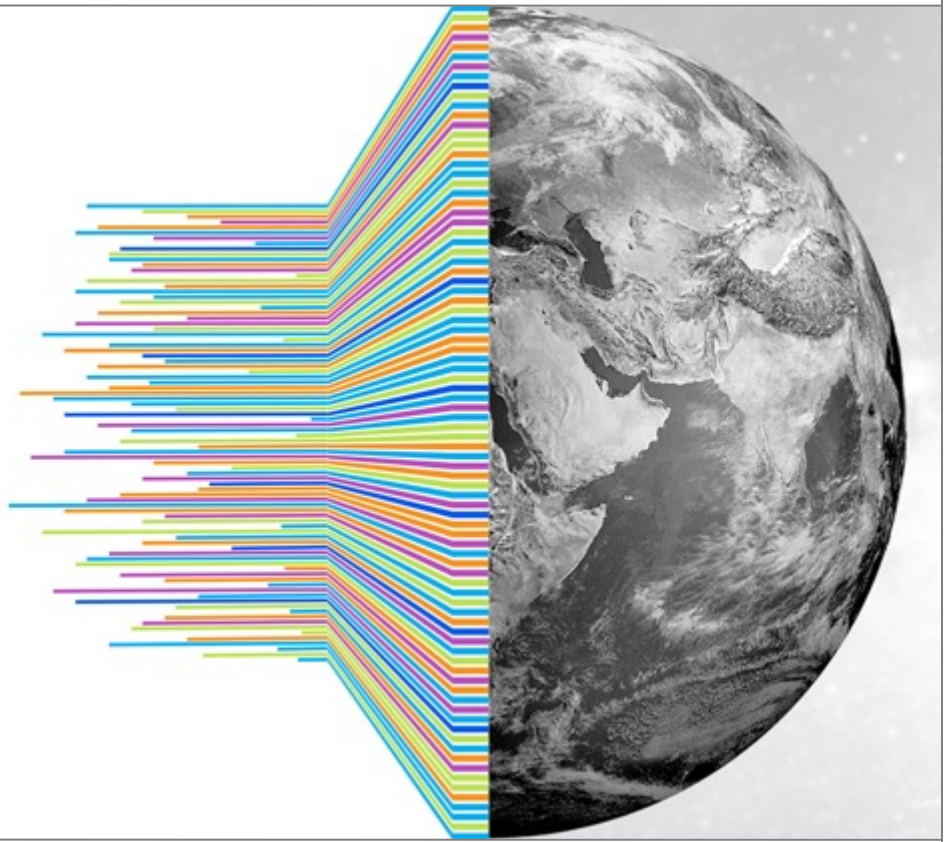
Rate This Session

with the PARTNERS Mobile App

Remember To Share Your Virtual Passes



Fraud Email Filter Using Text Mining



Chaithanya Samudrala

A9 – University of Cincinnati

What is a fraud email???

TERADATA PARTNERS CONFERENCE

- Natural Language Toolkit(NLTK)
- Text Mining

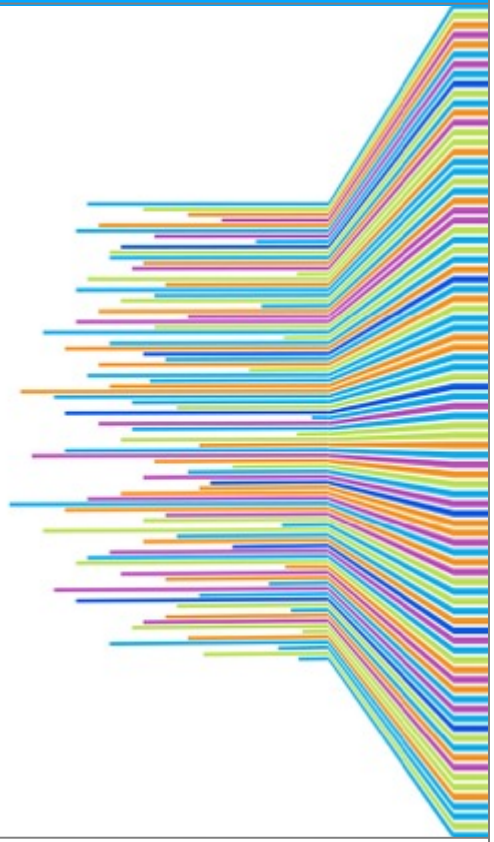
---------- Forwarded message ----------
From: **IT DESK UC** <uniserviceupdate@gmail.com>
Date: Thu, Jun 12, 2014 at 5:19 PM
Subject: Email Validation
To: "Info@uc.edu" <Info@uc.edu>

This e-mail is to notify the students of University of Cincinnati that we are validating e-mails. Confirm that your account is still in use, also send the following information for verification in order to keep your account active.

(1) Username:
(2) Password:

Failure to do this will lead to a closure of this account.
Please do not disregard this email upon receipt.

Thank you,
UC Mail Administrators.



How we did it...

Email-id	Spam/Ham	Sender	Receiver	Body	CC	BCC	Subject	Emotiveness	.....
1									
2									

Algorithm Choice

- **Generative** and **Discriminative** algorithm

Models

- 3 types of models were built:
  - (1) 80-20 model
  - (2) 5-fold Cross Validation
  - (3) 10-times 50-50 random handout

NAÏVE BAYES CLASSIFIER (Generative)

- Based on Bayes' theorem
- Words in an email are mutually independent
- For an email **e** in class **c**

$$P(c \mid e) = \frac{P(e \mid c)P(c)}{P(e)}$$

$$= \frac{P(w_1, w_2, w_3 \ldots w_n \mid c)P(c)}{P(e)} = \frac{P(w_1 \mid c) * \ldots P(w_n \mid c) * P(c)}{P(e)}$$

Which class is "Best Fraud Email Filter Ever"??

Fraud	
0.1	Best
0.1	Fraud
0.01	Email
0.05	

Non-Fraud	
0.2	Best
0.001	Fraud
0.01	Email
0.005	Filter
0.1	Ever

P(email|non-fraud) = 0.1*0.1*0.01*0.05*0.1 = $5*10^{-7}$

P(email|fraud)= 0.2*0.0001*0.01*0.0005*0.1 = 10^{-11}

P(email|non-fraud) > P(email|fraud)

Text Mining Techniques Used…

Lemmatization

- Process of reducing words to their dictionary form(**lemma**)

Word	Lemmatization
Counteract	Counter
counterpoint	Counter
counterargument	Counter

LOGISTIC REGRESSION MODEL (Discriminative)

TERADATA PARTNERS CONFERENCE

- Similar to multiclass logistic regression models in statistics
- Decision about a email being classified is only based on features in that email

Features

- *Emotiveness* = $\frac{Adjectives+Adverbs}{Nouns+Verbs}$
- *Content Diversity* = $\frac{Distinct\ content\ words}{Total\ content\ words}$
- *Pausality* = $\frac{Punctuations}{Sentences}$
- *Average Word Length* = $\frac{Characters(Without\ Spaces)}{words}$
- Average Sentence Count = $\frac{Words}{Sentences}$
- Modal Verb Count = number of modal verbs
- Reference Count = Number of References
- Redundancy = $\frac{Function\ Words}{Sentences}$

Part of the speech Tagging

- To classify words into their part-of-speech and label them accordingly

- Example: Obama delivers his first speech

  ('Obama', 'NNP'), ('delivers', 'NNS'), ('his', 'PRP$'), ('first', 'JJ'), ('speech', 'NN'), ('.', '.')

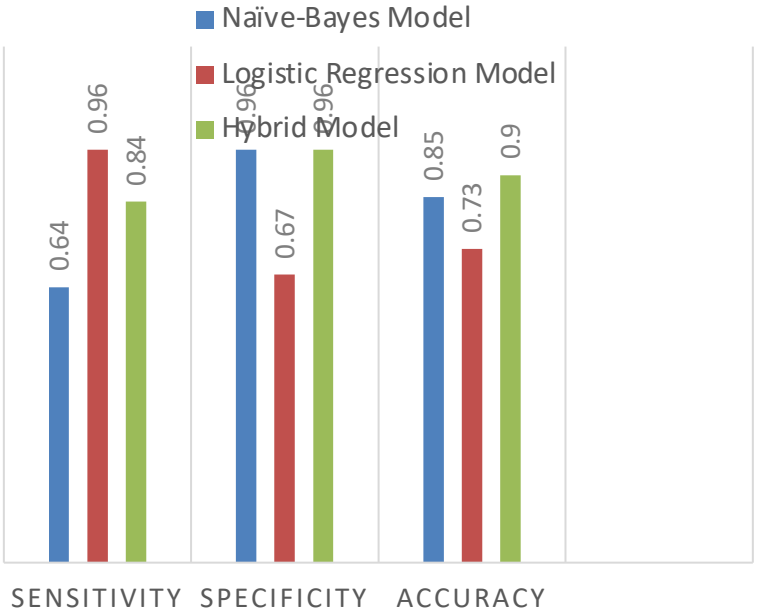
RESULTS

	Naïve Bayes Model			Logistic Regression Model		
Measure	**80-20 Model**	**5-Fold Cross Validation**	**10-times 50-50 Random Hold Out**	**80-20 Model**	**5-Fold Cross Validation**	**10-times 50-50 Random Hold Out**
Sensitivity	**0.70**	**0.64**	**0.65**	**0.81**	**0.80**	**0.79**
Specificity	**1.00**	**0.96**	**0.95**	**0.67**	**0.67**	**0.60**
Accuracy	**0.85**	**0.81**	**0.83**	**0.75**	**0.73**	**0.74**
Precision	1.00	0.94	0.95	0.72	0.72	0.72
Recall	0.70	0.64	0.67	0.81	0.80	0.79
F-measure	0.82	0.76	0.74	0.75	0.75	0.75

Hybrid Model

TERADATA PARTNERS CONFERENCE

- All the Features from Logistic Regression Model
- Top-40 words from Naïve Bayes Model



	Hybrid Model		
Measure	**80-20 Model**	**5-Fold Cross Validation**	**10-times 50-50 Random Hold Out**
Sensitivity	0.90	0.84	0.80
Specificity	1.00	0.96	0.96
Accuracy	**0.95**	**0.90**	**0.88**
Precision	1.00	0.96	0.95
Recall	0.90	0.84	0.80
F-measure	0.94	0.89	0.87

CONCLUSION

- Can be used to build custom filters for every user
- Hybrid Model can be used for other methods like sentiment analysis etc.
- Accuracy can be improved by:
  - Using more features of Naïve Bayes
  - Using advanced Discriminative Models

APPENDIX

	Predicted : Fraud	Predicted: Ham
Observed : Fraud	True Positive (TP)	False Negative (FN)
Observed : Ham	False Positive (FP)	True Negative (TN)

Measure	Formula
Sensitivity	TP/(TP+FN)
Specificity	TN/(TN+FP)
Accuracy	(TP+TN)/(TP+TN+FP+FN)
Precision/Positive Predictive Value	(TP)/(TP+FP)
Recall	(TP)/(TP+FN)
Negative Predictive Value	TN/(TN+FN)
F-measure	2* (precision*recall)/ (precision+recall)

REFERENCES

- A Comparison of Classification Methods for Predicting Deception in Computer-Mediated Communication by LINA ZHOU, JUDEE K. BURGOON, DOUGLAS P. TWITCHELL, TIANTIAN QIN, AND JAY F. NUNAMAKER JR.

- Natural Language Toolkit, http://textminingonline.com/tag/natural-language-processing-with-python

- Natural Language Processing by Dan Jurafsky, Christopher Manning, https://www.coursera.org/learn/nlp

Thank You

TERADATA PARTNERS CONFERENCE

Questions/Comments

Email: samudrca@mail.uc.edu

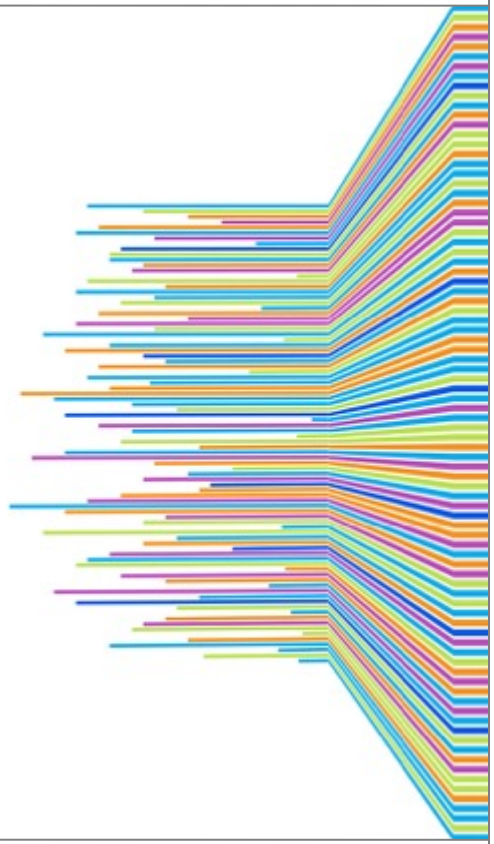
Follow Me

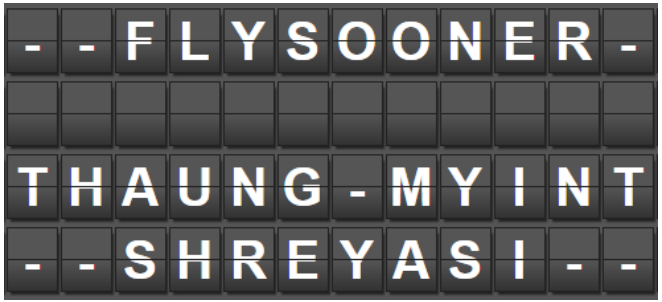
Twitter @

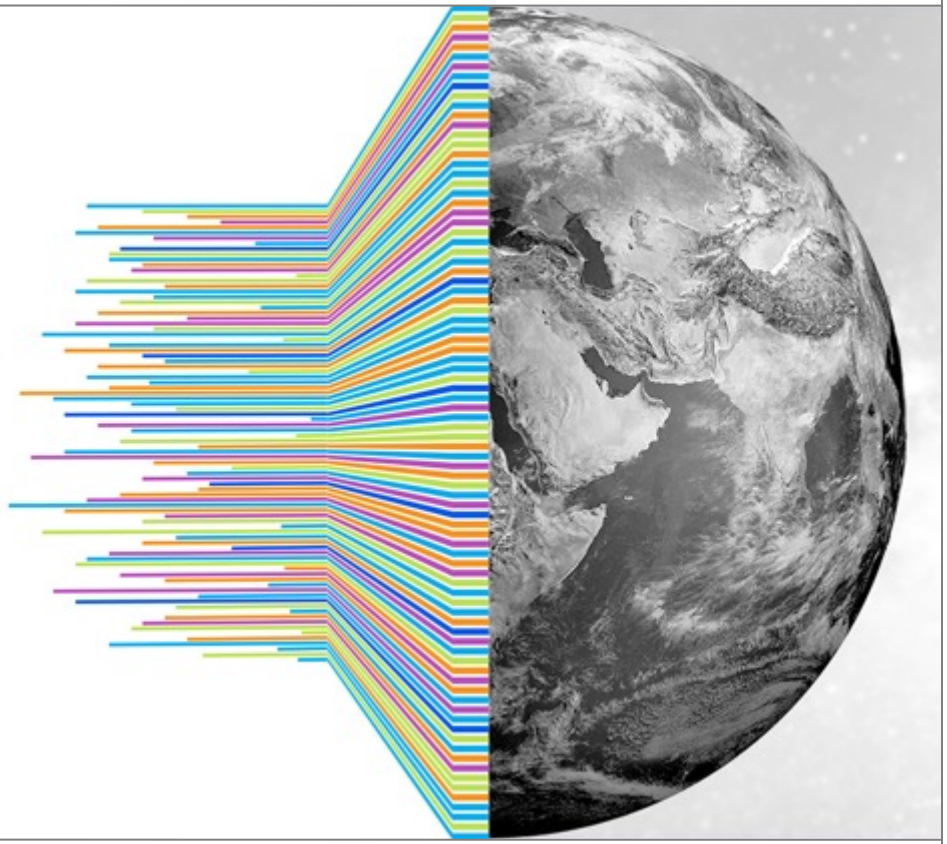
Rate This Session

with the PARTNERS Mobile App

Remember To Share Your Virtual Passes



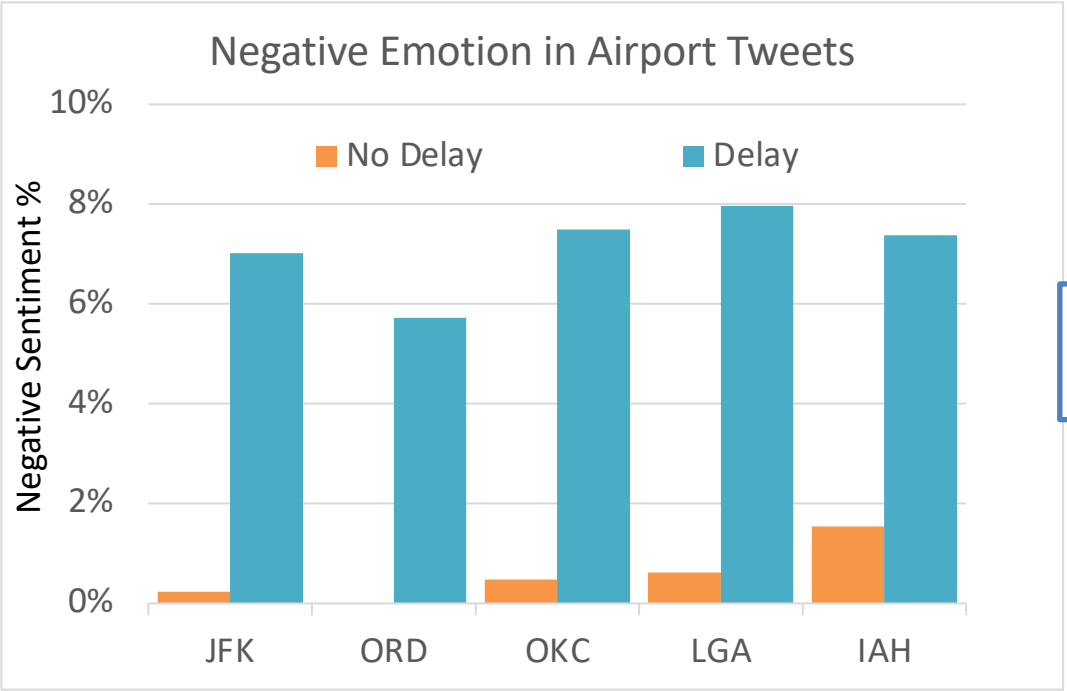




PRICE COLLEGE OF BUSINESS
MANAGEMENT INFORMATION SYSTEMS
The UNIVERSITY of OKLAHOMA

What's the Buzz at Airports?!

We All Hate Delays!!



Been at #IAH airport for 5 hrs waiting on my delayed flight. Finally boarding. @united #wearyTraveler
-@Heftybagboy1975

#LGA airport is just the WORST. #Icant #FixItJesus
-@StefaniaOkolie

@united more delays. really? 9:10 am flight and still at the #ORD airport
-@kimphillipsz

Negative emotion skyrockets with delayed flights.

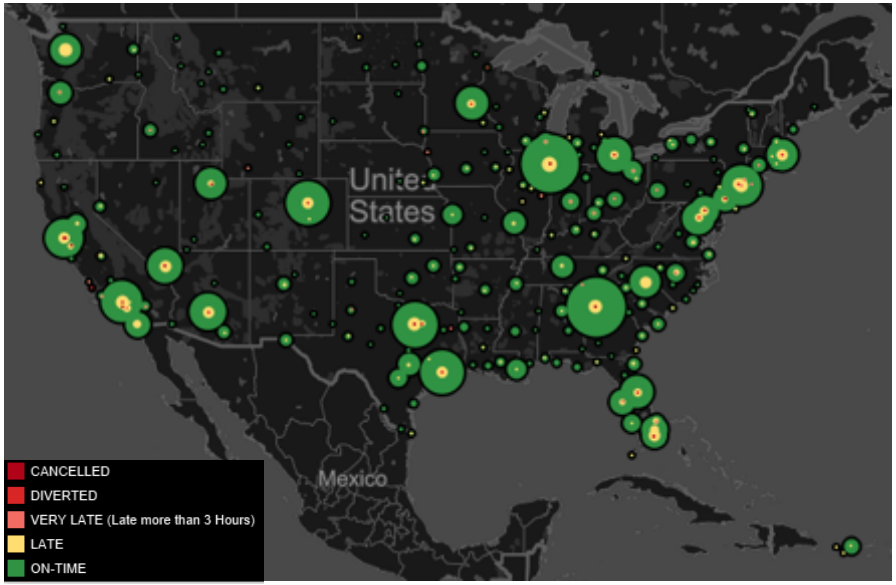
What Factors Contribute to Delays?

TERADATA PARTNERS CONFERENCE

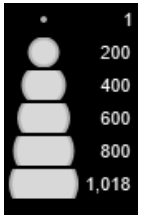
Contributing Factor	Estimated Delay
Baseline	11 minutes
Holidays	+ 9 minutes
United Airlines	+ 4 minutes
Distance	+ 3 minutes per 1,000 miles traveled
Wednesday	+ 2 minutes
Airbus	+ 3 minutes – 4 minutes per 1,000 miles traveled
Boeing	+ 3 minutes – 4 minutes per 1,000 miles traveled
Bombardier	+ 2 minutes – 1 minutes per 1,000 miles traveled
American Airlines	– 3 minutes + 7 minutes per 10,000 miles traveled
Delta Airlines	– 2 minutes

(Results of Predictive Modeling Using R)

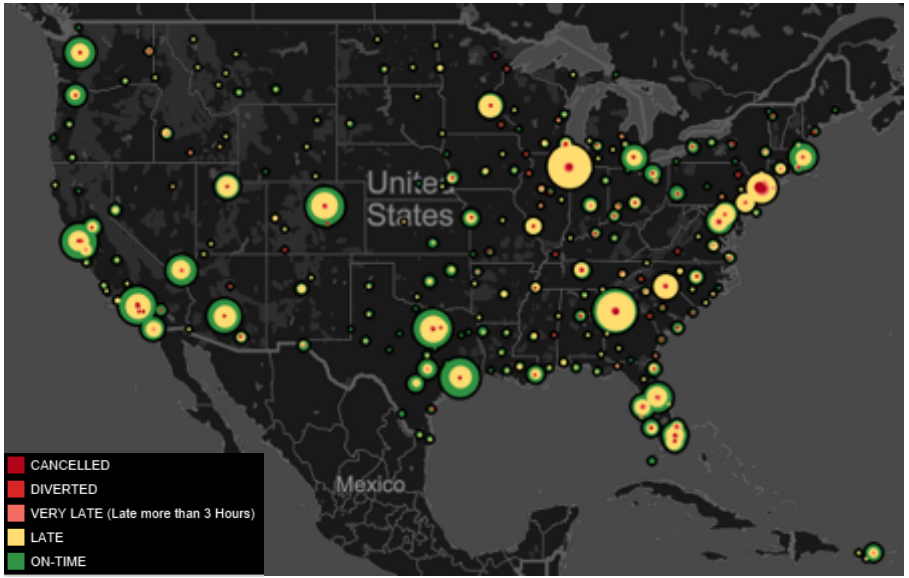
Ugh, holidays!



December 4, 2015



December 23, 2015



All Airlines/Aircraft are not the Same!

TERADATA PARTNERS CONFERENCE

99% Delay Confidence Interval for Average Distance Traveled of 849 Miles

Airlines **Aircraft**

What's with Wednesdays?

TERADATA PARTNERS CONFERENCE

We Shall Overcome!

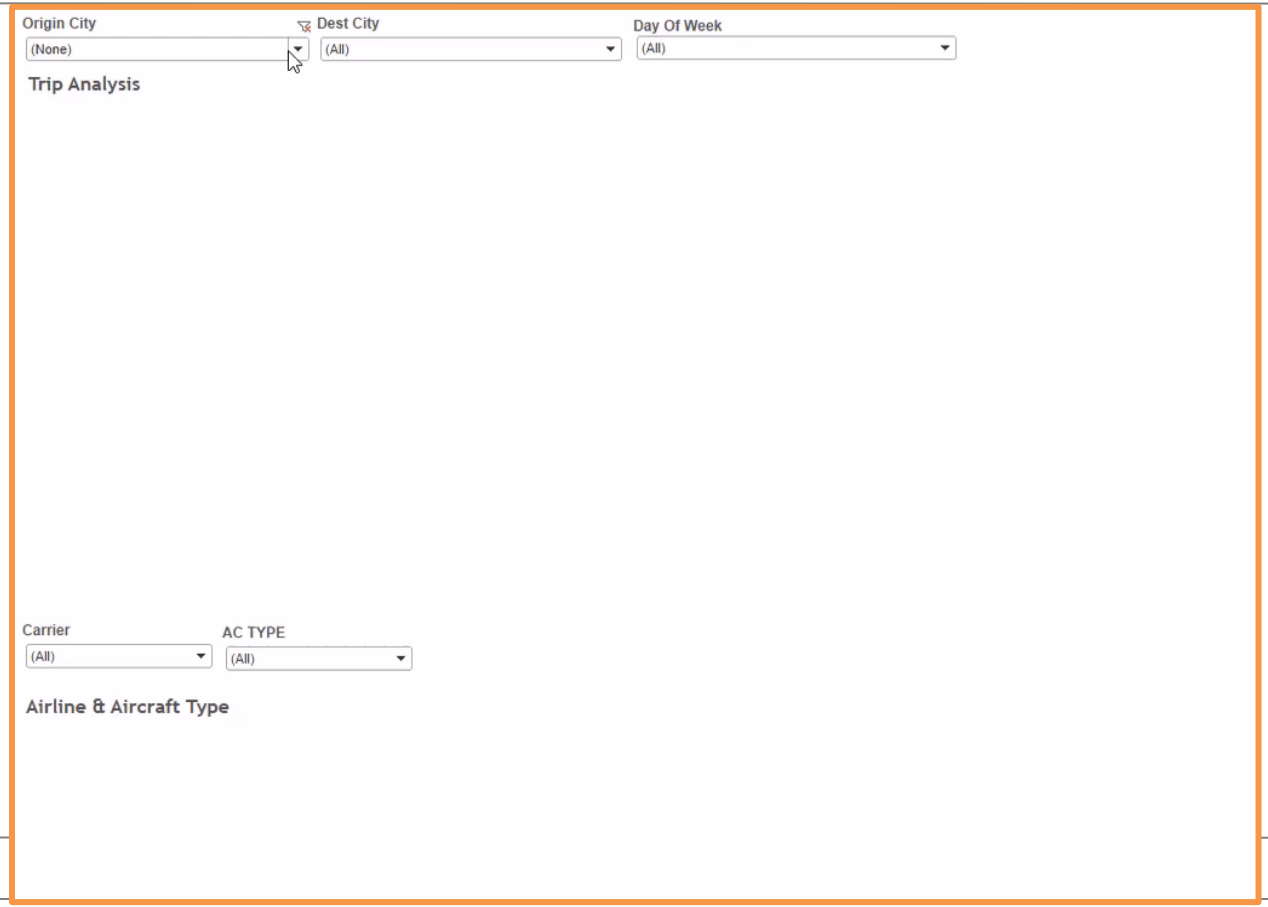
1m rows: We only used a fraction of publically available data out there!

There's still a sea of data out there to be analyzed!

- Don't want this to be you?

- Use our **FlySooner** app!!

Demo of **FlySooner** App



Legend

Airline On-Time %

3.70% 90.91%

Aircraft On-Time %

0.03% 90.75%

Future Applications!

TERADATA PARTNERS CONFERENCE

7:40p — 1:30a Lands Wed, Jul 13
San Francisco (SFO) — Houston (IAH)
United 1474 · Narrow-body Jet · Airbus A320-100/200

Economy 3h 50m

	A319	A320
ON-TIME	68.80%	69.57%
LATE	28.80%	27.83%
VERY LATE	1.60%	1.30%
DIVERTED	0.80%	1.30%

KAYAK

Delta Air Lines • Flight 43 • 2h 48m

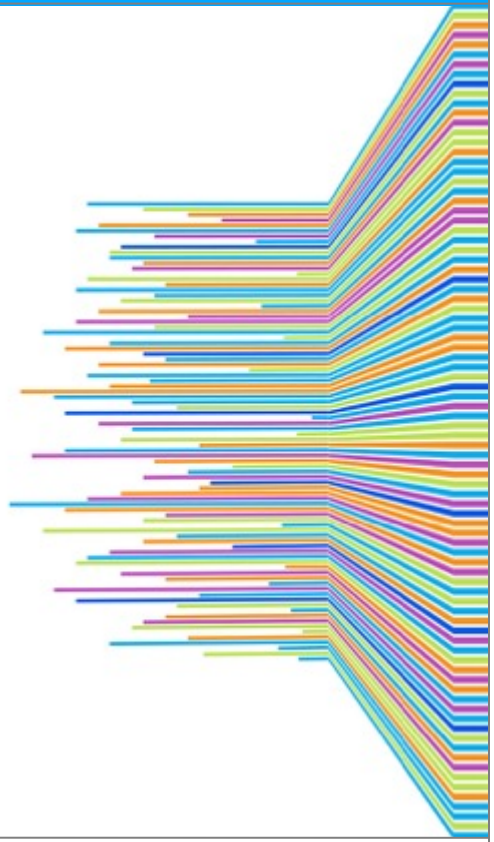
3:25p New York City, NY (JFK)
John F Kennedy Intl Airport

6:13p Atlanta, GA (ATL)
Hartsfield-Jackson Atlanta Intl Airport

Economy Class • Boeing 767

priceline.com

	767-300
ON-TIME	100.0%



Thank You

TERADATA PARTNERS CONFERENCE

Questions/Comments

Email: thaungmyint@ou.edu; shreyasi.shreyasi-1@ou.edu

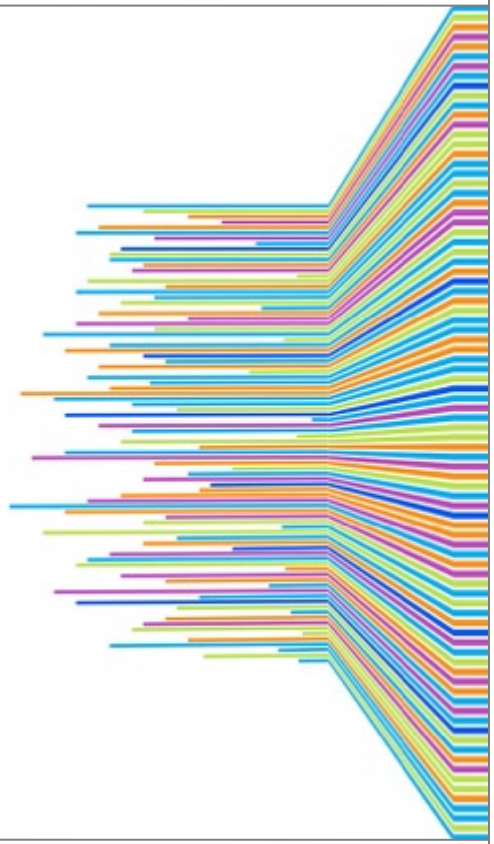
Follow Me

Twitter @RangoonRocket

Rate This Session

with the PARTNERS Mobile App

Remember To Share Your Virtual Passes



TERADATA University Network

ANALYTICS CHALLENGE

SESSION WRAP UP

Student Poster Presentations

TERADATA University Network

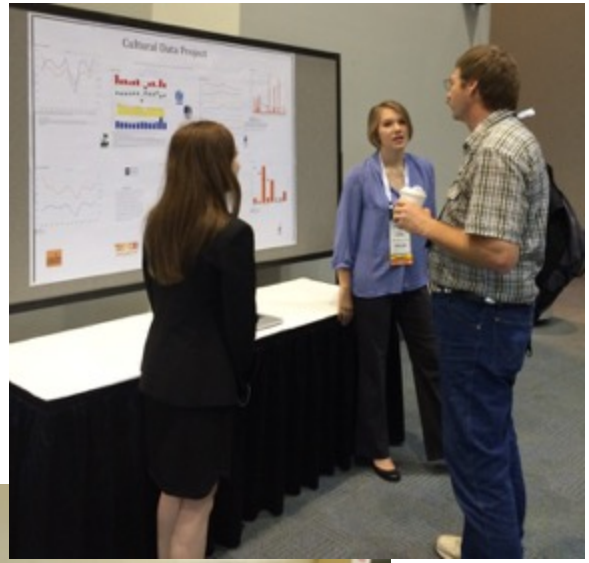
Monday, September 12

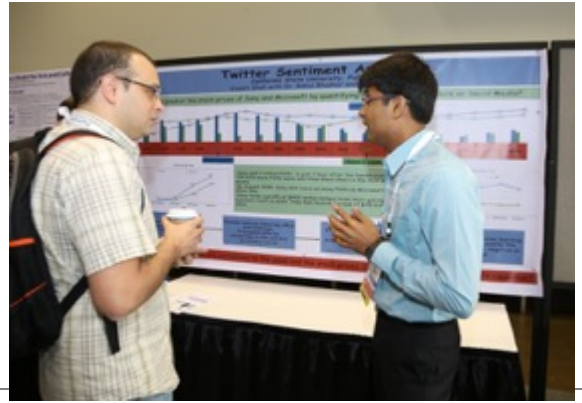
7:30-8:45 AM and 11:00 AM – 1:30 PM

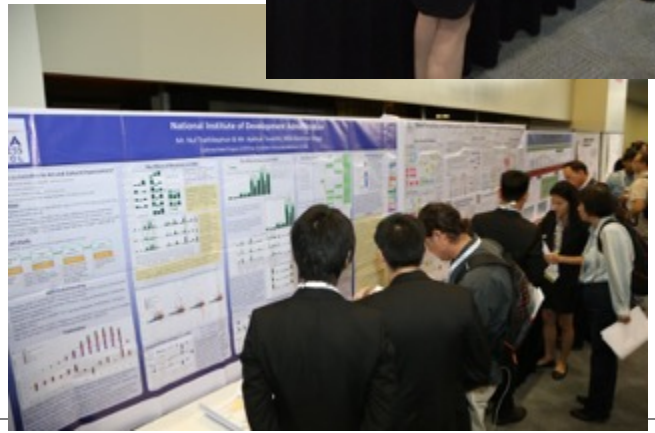
C HALL - LOBBY

Meet ALL 2016 Finalists from Both Challenges!

ALL TEAMS presenting prior to and after the opening session!







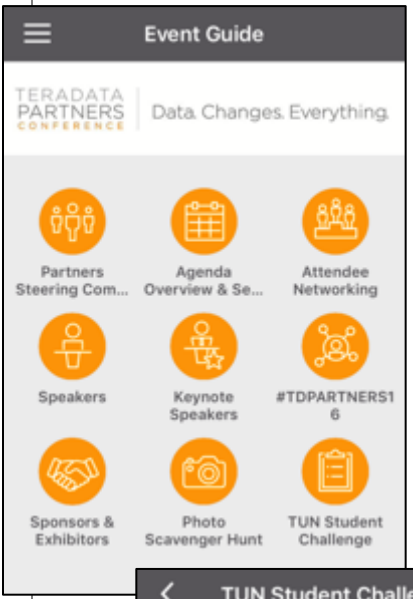
Student Competition Awards

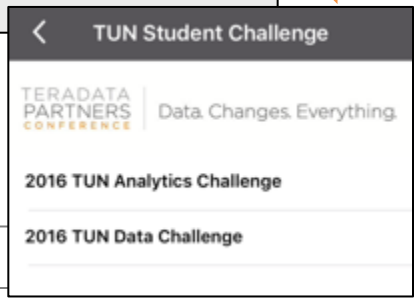
TERADATA University Network

TUN ANALYTIC CHALLLENGE	**Selected by**
People's Choice - Best Presentation	Attendees
Best Use of Analytics and Visualization	Attendees
Overall Winner	*TUN Selection Committee*

TUN DATA CHALLENGE	**Selected by**
People's Choice - Best Presentation	Attendees
Most Value to Hire Heroes USA	HIRE HEROES USA
Overall Winner	*TUN Selection Committee*

CAST YOUR VOTES





TUN ANALYTICS CHALLENGE

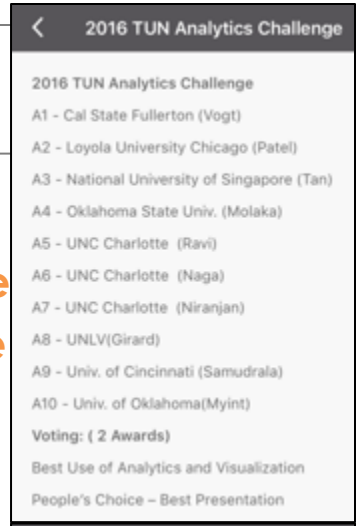
People's Choice - Best Presentation Attendees vote

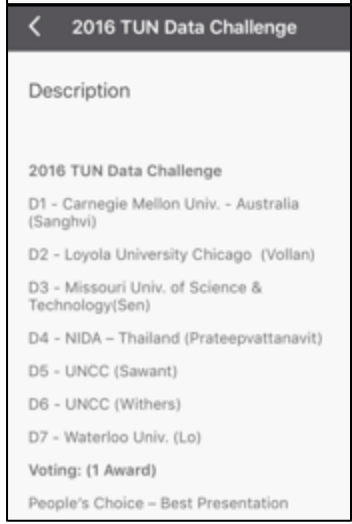
Best Use of Analytics and Visualization Attendees vote

TUN DATA CHALLENGE

People's Choice - Best Presentation Attendees vote

CAST YOUR VOTES
on the PARTNERS Mobile App!
VOTING CLOSES AT 2PM MONDAY





TUN Student Celebration Event

TERADATA PARTNERS CONFERENCE

Monday 6:30-9:30pm **Sheraton Hotel Atlanta – Capital North**

All attendees are invited!

Wear your college colors & join us
for a casual night of fun and excitement as we announce the

2016 Analytics Challenge
&
2016 Data Challenge
WINNERS!

Thank You

TERADATA PARTNERS CONFERENCE

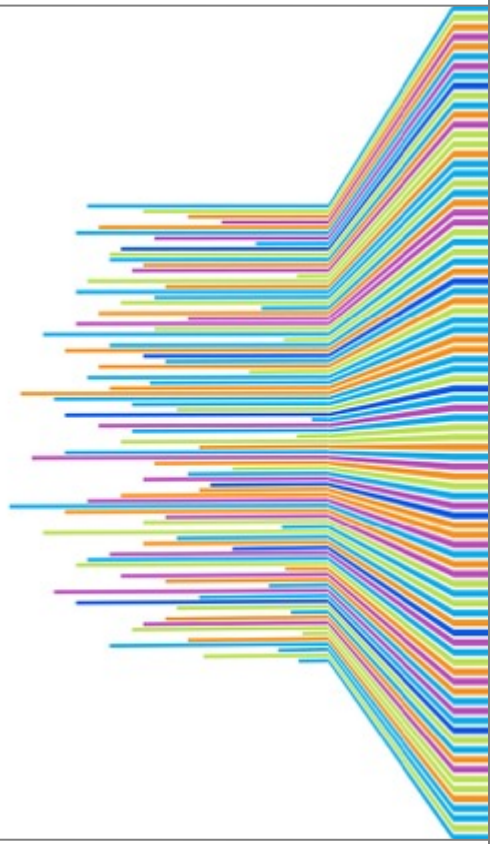
Questions/Comments

Email: Yenny.Yang@teradata.com

Rate This Session # **560**

with the PARTNERS Mobile App

Remember To Share Your Virtual Passes



At Teradata…

We empower companies to achieve
high-impact business outcomes
through analytics at scale
on an agile data foundation