Text Data to Insights – Creating a storyline with Social Data

Becoming a data ninja

A talk delivered at
Data Visualization - Methods & Tools
IIM Lucknow
28/04/2016

justsmartmandate

Who are we ?

We are a data science company, founded in 2009– with special interest in making the world an intelligent place to live in.

We identify data and bring it to light, making it visible, cohesive, comparable and easy to understand so that it really does support YOU in making the right decisions.

Who am I ?

I am a Practice Lead at JSM for Natural Language Processing & Machine Learning. I have architected multiple solutions in the area of text analytics for multiple industries like finance, healthcare, food & beverages & hospitality.

AREAS WE WORK ON

PHARMA

Sales Pitch Analysis

RETAIL

Predictive + IoT

FINANCE

Competitive Intelligence

F&B

Customer Insights

MR

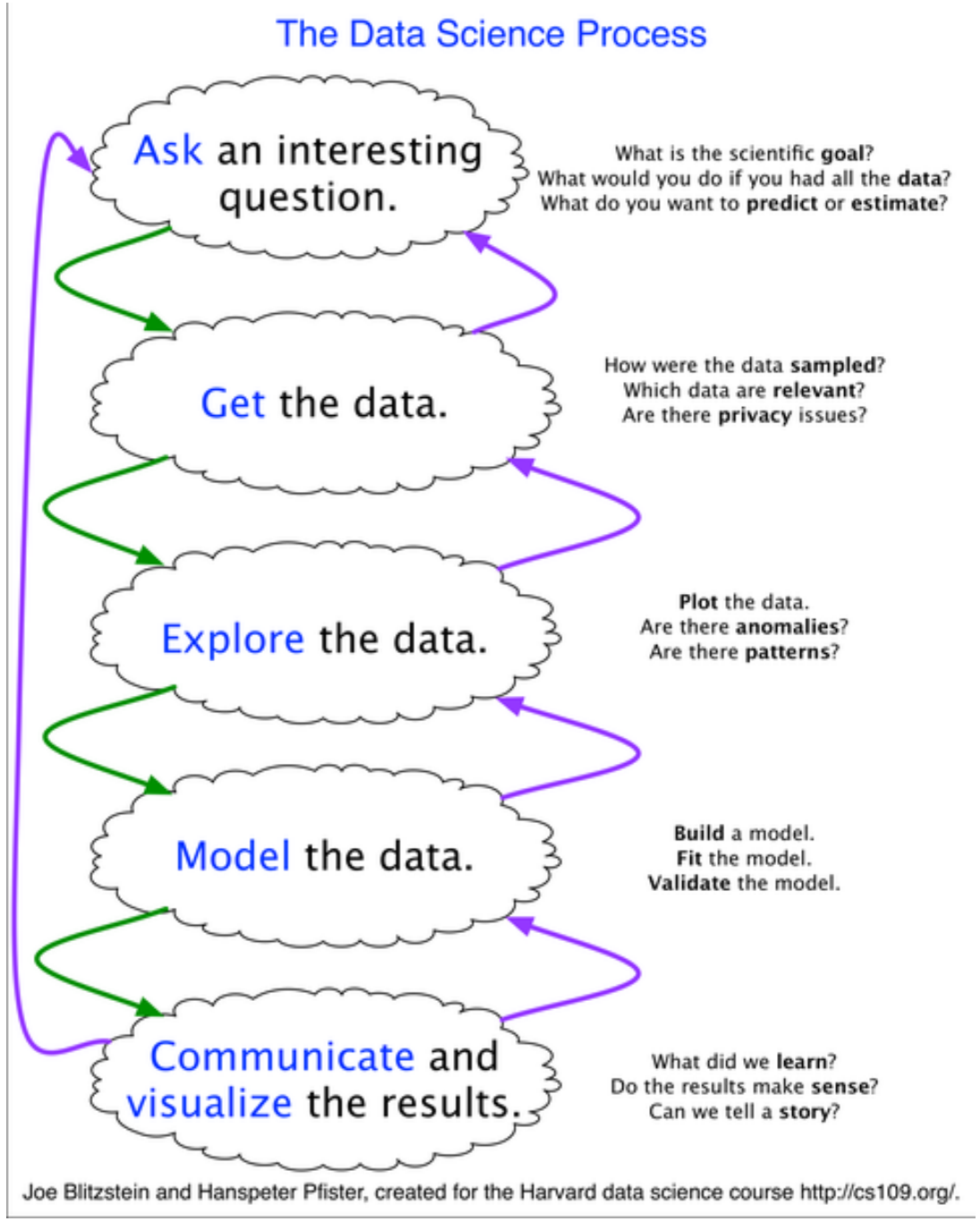
Scoping and Product Evaluation

SaaS

NLP, ML, Text

"Data scientist is
the sexiest job
of the 21st century."

Harvard Business Review


The Data Science Process
Ask an interesting question.
What is the scientific goal?
What would you do if you had all the data?
What do you want to predict or estimate?
Get the data.
How were the data sampled?
Which data are relevant?
Are there privacy issues?
Explore the data.
Plot the data.
Are there anomalies?
Are there patterns?
Model the data.
Build a model.
Fit the model.
Validate the model.
Communicate and visualize the results.
What did we learn?
Do the results make sense?
Can we tell a story?
Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course http://cs109.org/.

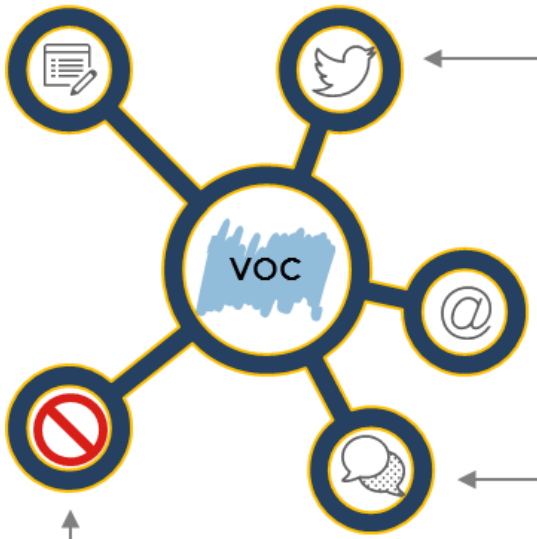
WE WILL TALK OF….

1. ELEMENTS OF STORY TELLING WITH DATA
2. CREATING DATA DRIVEN NARRATIVES
3. DIY – WEB CRAWLING
4. TEXT ANALYTICS– THE DESIGN
5. DIY – TEXT ANALYTICS

BONUS....

HOW TO STOP PAYING THOUSANDS - GET YOUR OWN SOCIAL MEDIA TRACKER AT ZERO COST !

ELEMENTS OF DATA



Internal Reviews & NPS scores, collected via tablet/paper/emails

Social posts, on FB/Twitter, directed towards the brand

Emails, mostly for complaints or seeking information

External reviews – on sites like Zomato, Tripadvisor, Mouthshut

NOISE – Reviews which are either 'fake' or 'planted'

PART 1

DATA VISUALIZATION : HOW TO TELL A STORY

Ideate upon a story

Focus on a story that may interest your audience

Do a fact check !

Double check your data, and that it supports your story.

Limit the number of ideas, not slides

Focus on one or two key statistics from your research

Use visuals and tables (or not)

Call out the data. Highlight. Reduce text.

Humanize

Focus on things people care about

Make it insightful and helpful

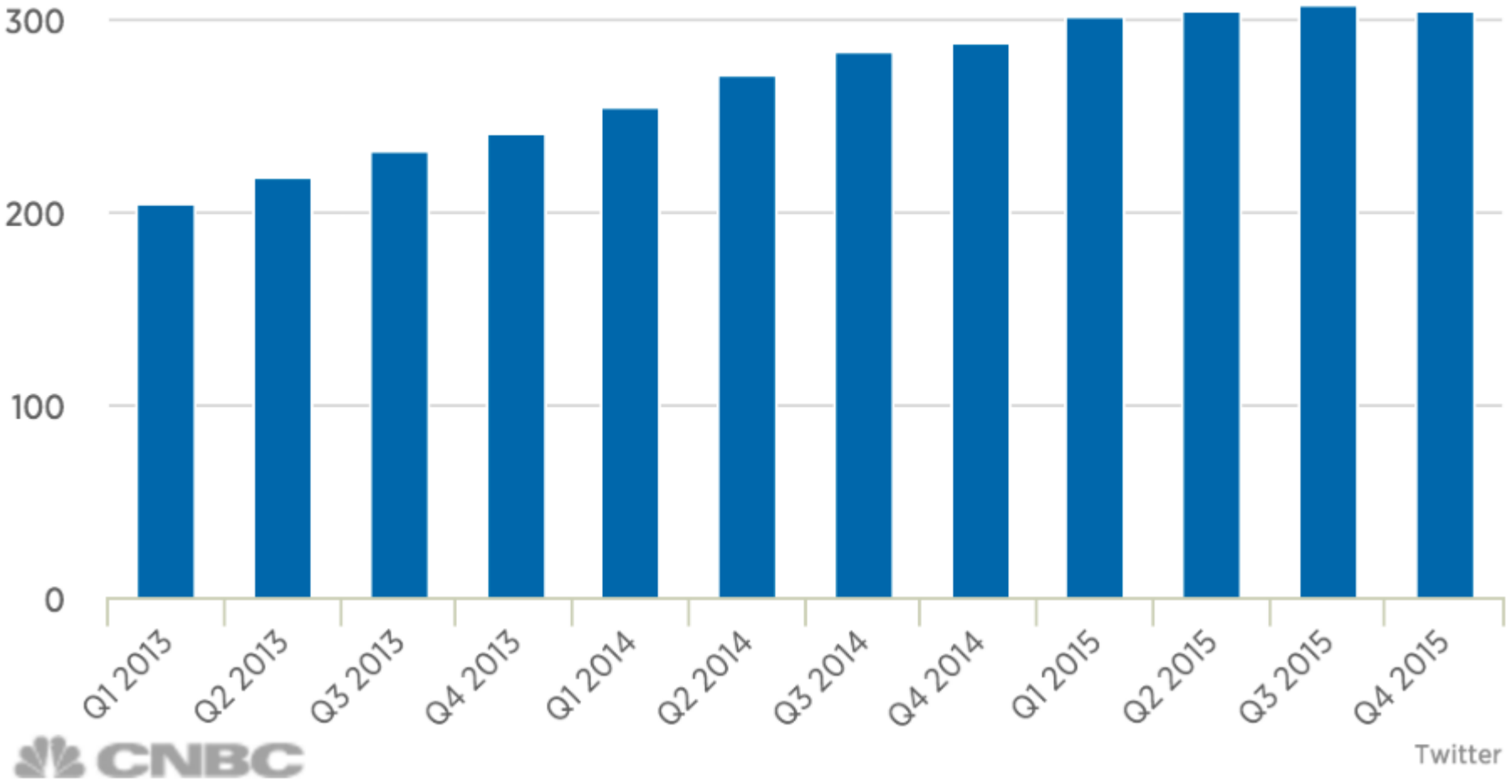
Focus on things people care about

PART 2

CREATING A DATA DRIVEN NARRATIVE

TRENDS

Typically these stories focus on how something is rising or falling over time. However, even a flattening trend can be a major story.



The obvious next question is “Why?”

COMPARISONS

Compare it with a peer


2013
2014
2015
Twitter*
Facebook
1,110
1,155
1,189
1,228
1,276
1,317
1,350
1,393
1,441
1,490
204
218
232
241
255
271
284
288
301
304
Q1'13
Q2'13
Q3'13
Q4'13
Q1'14
Q2'14
Q3'14
Q4'14
Q1'15
Q2'15

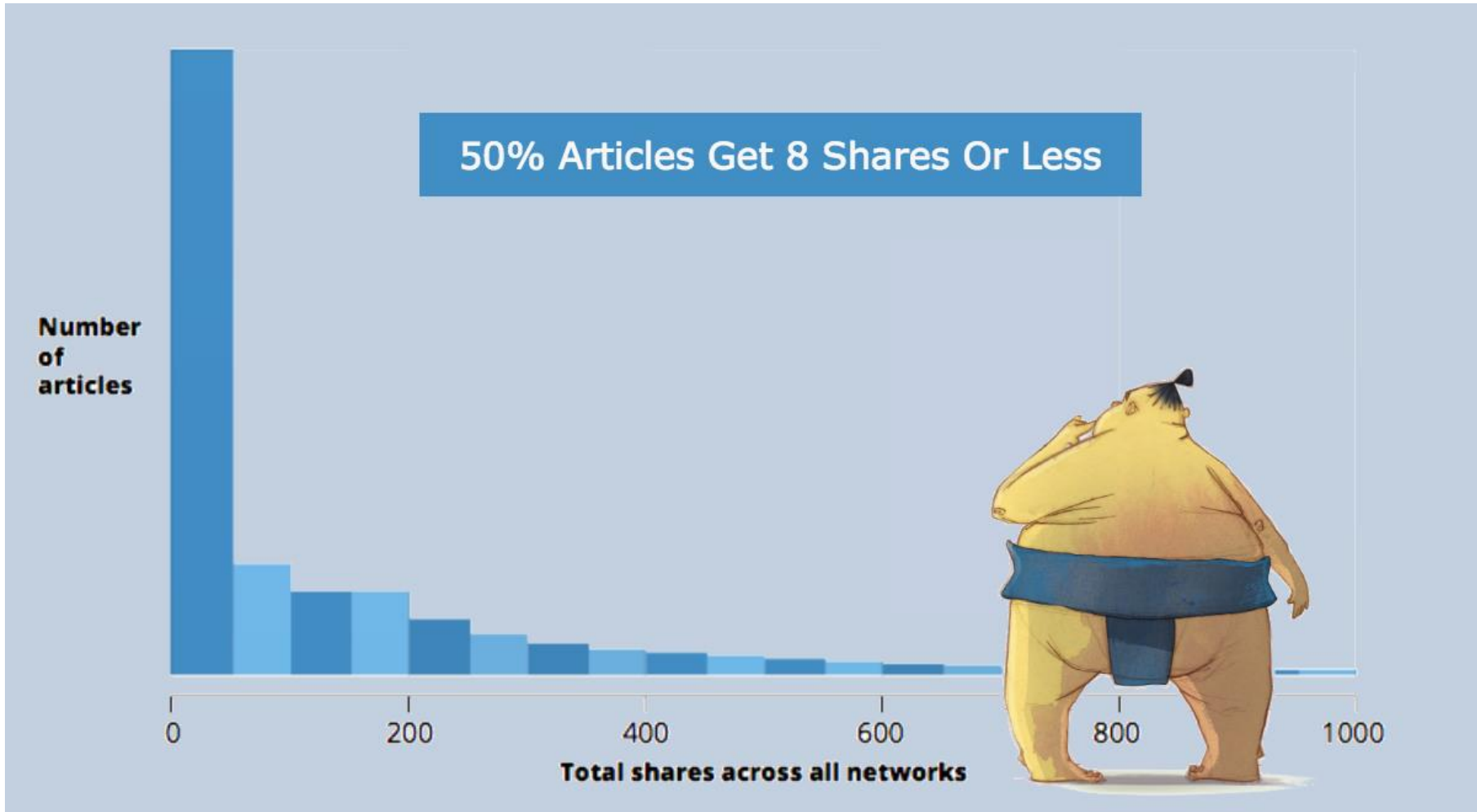
TABLES :/

Necessary Evil

Author	Number Articles Published	Average shares per content marketing article
Brian Sutter	8	6,304
Rand Fishkin	13	3,852
Kevan Lee	19	3,262
Lindsay Kolowich	57	2,782
Kelsey Libert	20	2,239
Carly Stec	20	2,083

RELATIONSHIPS

Machine Learning, Statistics


50% Articles Get 8 Shares Or Less
Number of articles
0
200
400
600
800
1000
Total shares across all networks

GENDER: { female }

Ms. Prachi Tamanna

The service was good. The staff was courteous and polite. 4/5

TOPIC: { service }

SENTIMENT: { positive }

KEYWORDS: { good service, courteous staff, polite staff }

PART 3

Some Examples

Lunchbox

www.lunchbox.subcortext.com

Lunchbox

Jane Doe

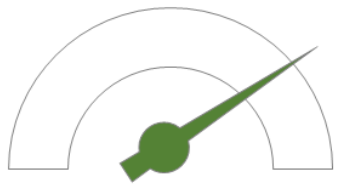
Select City | Select Store | Select Source | Select Gender | 01/01/2016 to 30/11/2016

12,000	6500	4000	1500	Food
Mentions	Internal Reviews	Via Zomato	Via Tripadvisor	Most talked about

FEW WEEKS THAT WERE

You received **12000** mentions, out of which **6500** were from your internal repository. **Ambience** was the most talked about topic. Overall, guests were **positive** about your brand. Bravo !

SHARE NOT NOW



MOOD METER

wide_menu **starters**

favorite_menu bright **pricey**

chicken_tikka rajma_masala

good_service perfect_menu

brilliant_ambience **expensive**

good_decor tasty_food fast_service daal_tadka

dim_lighting **best_chicken**

REVIEW SNIPPETS

Chicken Tikka	★★★★★
Mutton Seekh	★★★★★
Non-veg starters	★★★★☆
Rajma Masala	★★★★☆
Daal Tadka	★★★☆☆

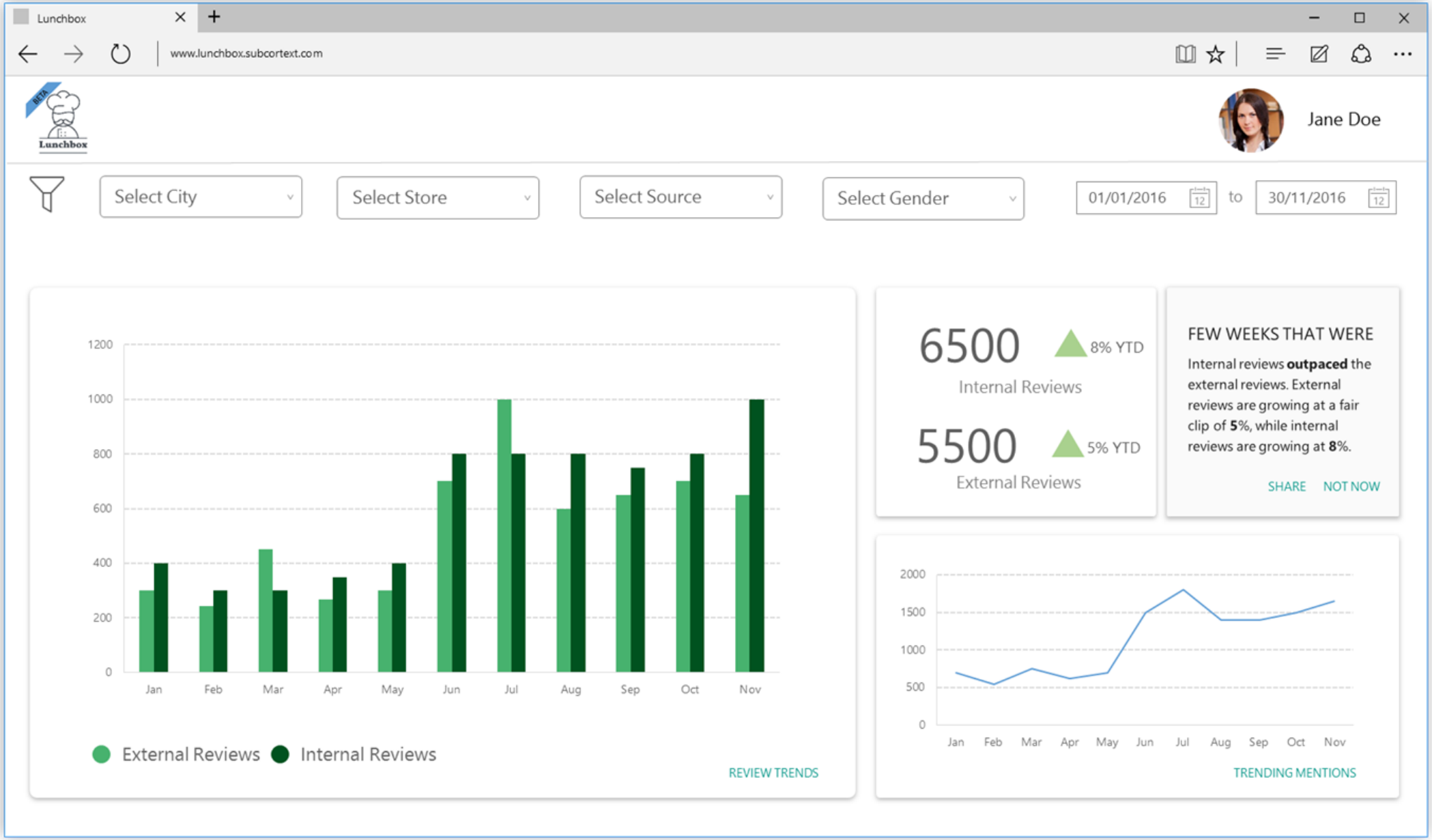
MOST TALKED ABOUT DISHES

92% ▲6.2% YTD

Customer Satisfaction

Ambience Pricing

Good vs Bad


Lunchbox
www.lunchbox.subcortext.com
BETA
Lunchbox
Jane Doe
Select City
Select Store
Select Source
Select Gender
01/01/2016
to
30/11/2016
1200
1000
800
600
400
200
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
External Reviews
Internal Reviews
REVIEW TRENDS
6500
8% YTD
Internal Reviews
5500
5% YTD
External Reviews
FEW WEEKS THAT WERE
Internal reviews outpaced the external reviews. External reviews are growing at a fair clip of 5%, while internal reviews are growing at 8%.
SHARE
NOT NOW
2000
1500
1000
500
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
TRENDING MENTIONS

Lunchbox

www.lunchbox.subcortext.com

Lunchbox

Jane Doe

Select City | Select Store | Select Source | Select Gender | 01/01/2016 to 30/11/2016

Good | Okay | Needs Improvement

Show 5

However, the RICE AUR RAJMA MASALA were amazing, I loved them.

The non veg ones were significantly better.

The fish was out of this world. It was soft and just perfect.

The roti and the naan were lovely!

Be it chicken, sea food or veggies you just can't go wrong with any starter.

REVIEW SAMPLES - FOOD

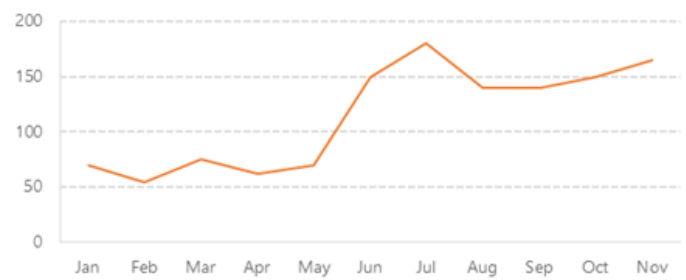
1256 3% YTD

Internal Reviews

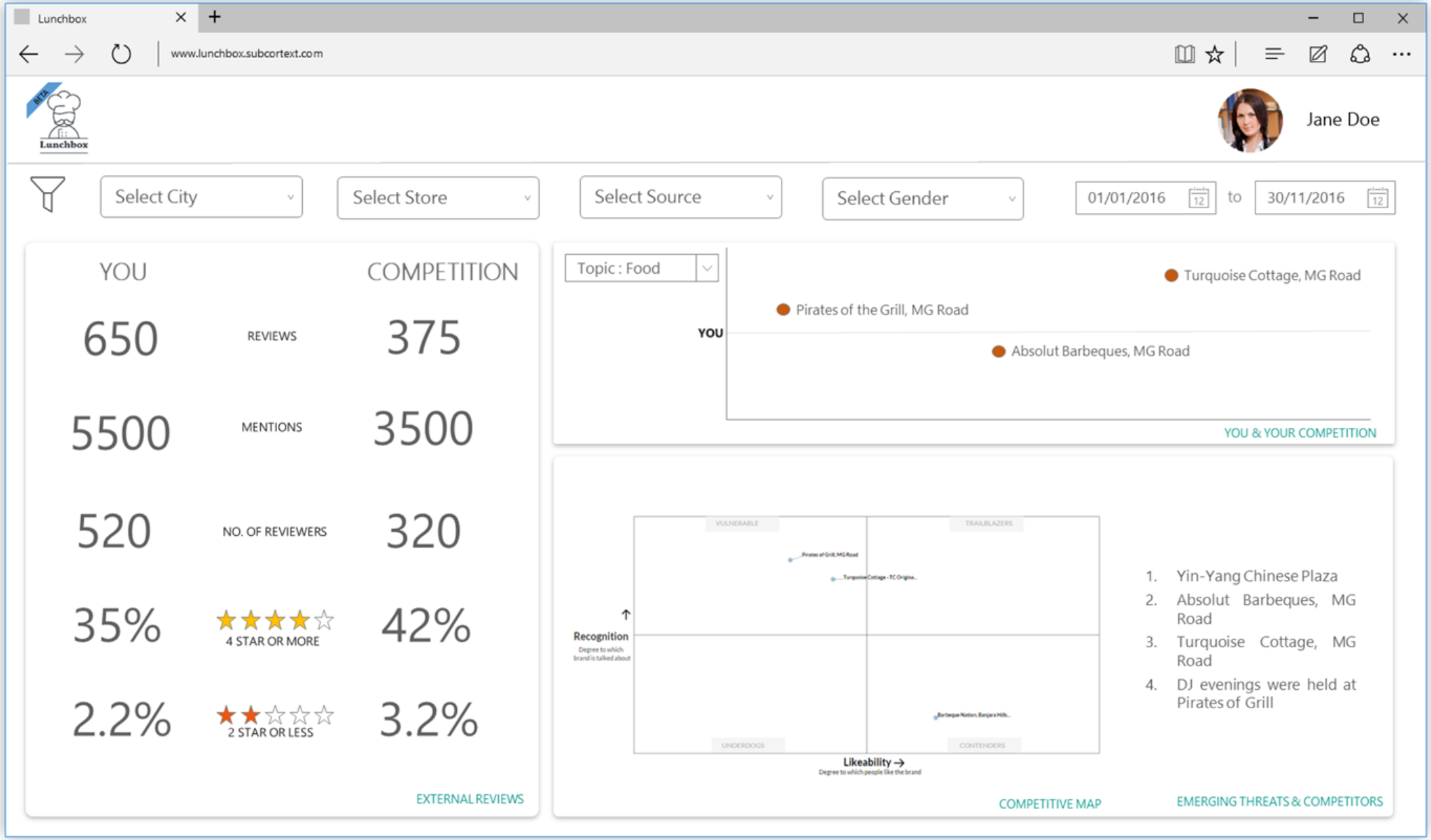
1045 1% YTD

External Reviews

MOOD METER



TRENDING MENTIONS


Lunchbox
www.lunchbox.subcortext.com
Lunchbox
Jane Doe
Select City
Select Store
Select Source
Select Gender
01/01/2016
to
30/11/2016
YOU
COMPETITION
650
REVIEWS
375
5500
MENTIONS
3500
520
NO. OF REVIEWERS
320
35%
4 STAR OR MORE
42%
2.2%
2 STAR OR LESS
3.2%
EXTERNAL REVIEWS
Topic : Food
YOU
Turquoise Cottage, MG Road
Pirates of the Grill, MG Road
Absolut Barbeques, MG Road
YOU & YOUR COMPETITION
Recognition
Likeability
COMPETITIVE MAP
1. Yin-Yang Chinese Plaza
2. Absolut Barbeques, MG Road
3. Turquoise Cottage, MG Road
4. DJ evenings were held at Pirates of Grill
EMERGING THREATS & COMPETITORS

PART 4

Scraping data from the web

CLIENT BRIEF

CLIENT BUDGET

WHAT DO CLIENTS WANT

TOOLS WE PLAY WITH

OPEN SOURCE Inexpensive	**DATABASES** Fast & Scalable	**INSIGHTS** Python, R
TECHNIQUES Latest yet tested	**VISUALIZATIONS** D3, GCharts, Tableau	**MANAGEMENT** Basecamp

Low Cost Data Collection

+

Comprehensive Analytics

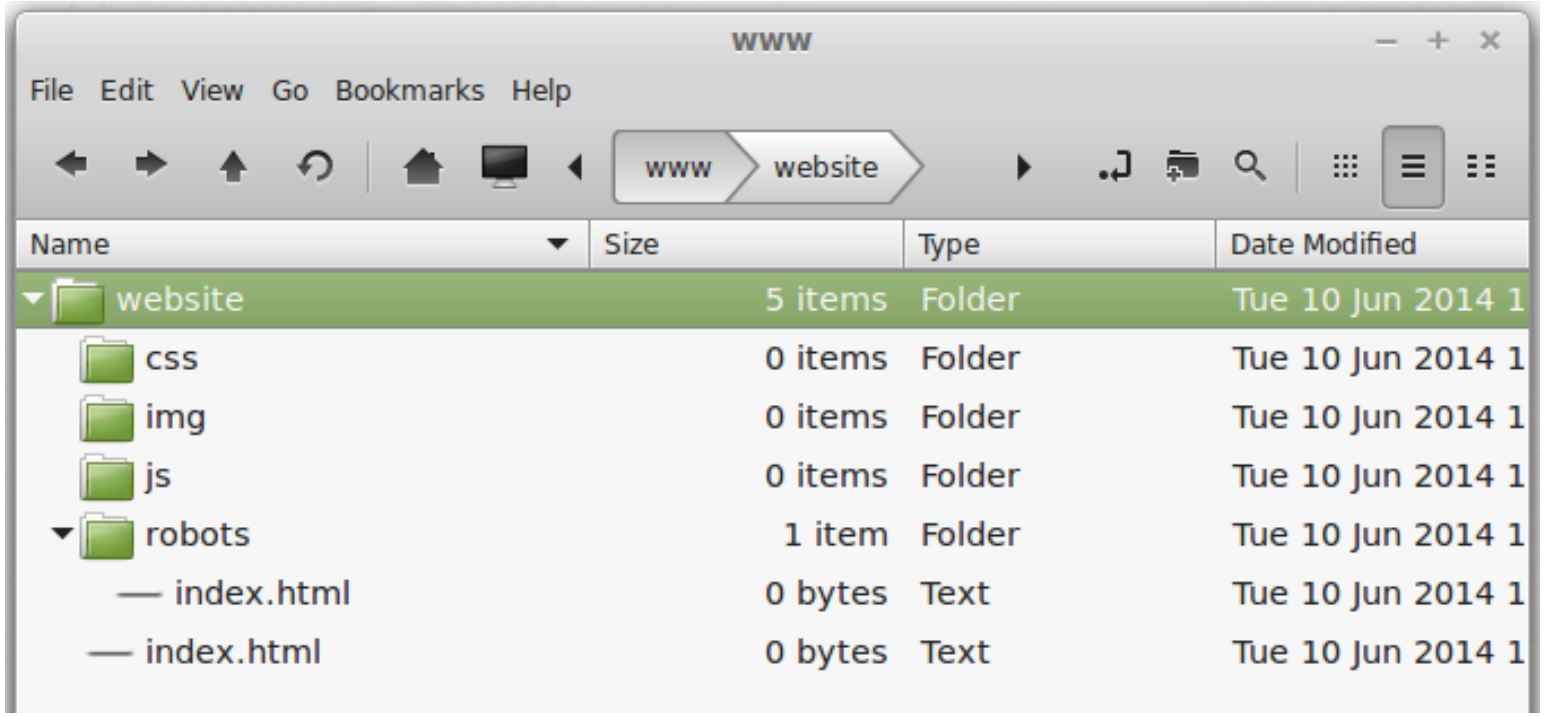
Basic Structure of a HTML document

```
<html>
   <title>Page title</title>
   <head></head>
    <body>
         ****Content comes here****
     </body>
</html>
```

HTML PAGES

• HTML pages are like textbooks – content, titles, subtitles, paragraphs and so on

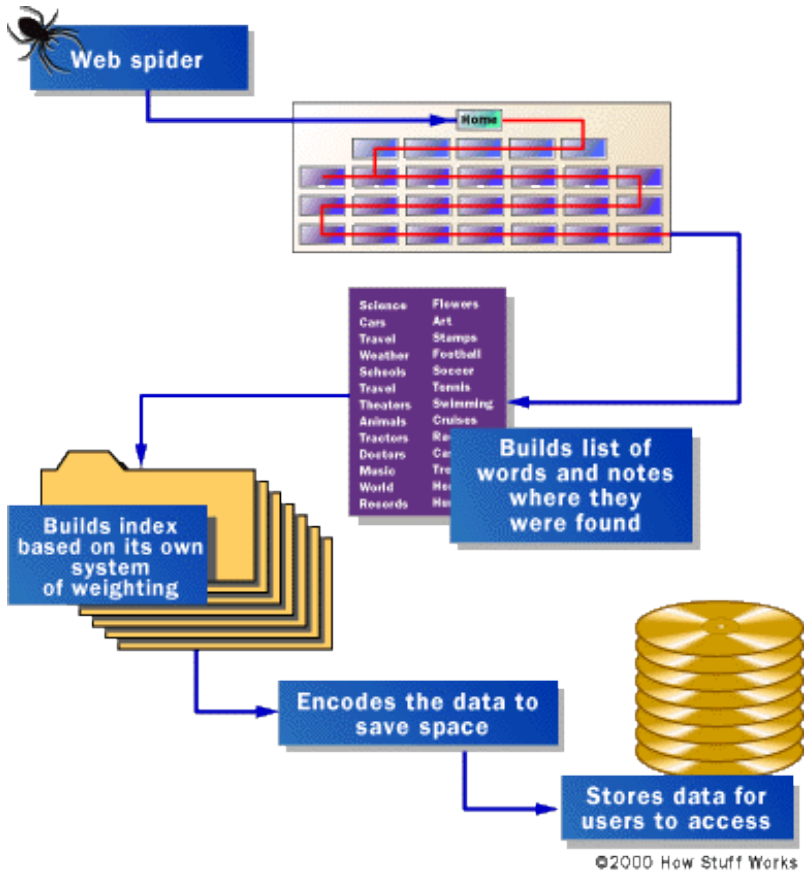
•Javascript adds interactivity to the HTML pages



OVERVIEW– WEB CRAWLING

A **crawler** is a program that visits **Web** sites and reads their pages and other information in order to create entries for a search engine index.

The major search engines on the **Web** all have such a program, which is also known as a "**spider**" or a "**bot.**"



BUT, YOU ARE MANAGERS.

DO YOU NEED THIS ?

YES !

You don’t need to write a code, promise.*

* T&Cs apply

Tool 1

Import.io

INTRODUCTION

- BEST of the lot
- Gives great flexibility – just click and extract
- Most of the sites are compatible
- Easy CSV/Google Docs Export
- Provides APIs for regular data updates
- Low training time

USE CASES

You want to monitor feedback

http://www.consumercomplaints.in/snapdeal-com-b100038

USE CASES

You want to get all the brands for your qnr.

http://www.amazon.in/Smartphones/b/ref=nav_shopall_sa_menu_mobile_smartphone?ie=UTF8&node=1805560031

USE CASES

You want to stay ahead of competition

http://cashkaro.com/

USE CASES

You want to create pricing strategy

http://www.shopclues.com/mobiles/unboxed-mobiles.html

Tool 2

webscraper.io

INTRODUCTION

- Works where Import.io fails
- Bit buggy, but does a god job of providing flexible choices of data extraction
- Most of the sites are compatible
- Easy CSV Export
- NO APIs for regular data updates
- Moderate learning curve

LET'S GET OUR HANDS DIRTY

PRECAUTIONS

- Don’t scrape too fast or you will get banned
- Respect **robots.txt**
- Extract only what you need
- Don’t overload their servers
- Don’t take data what’s not yours – only the data in public domain

PART 5

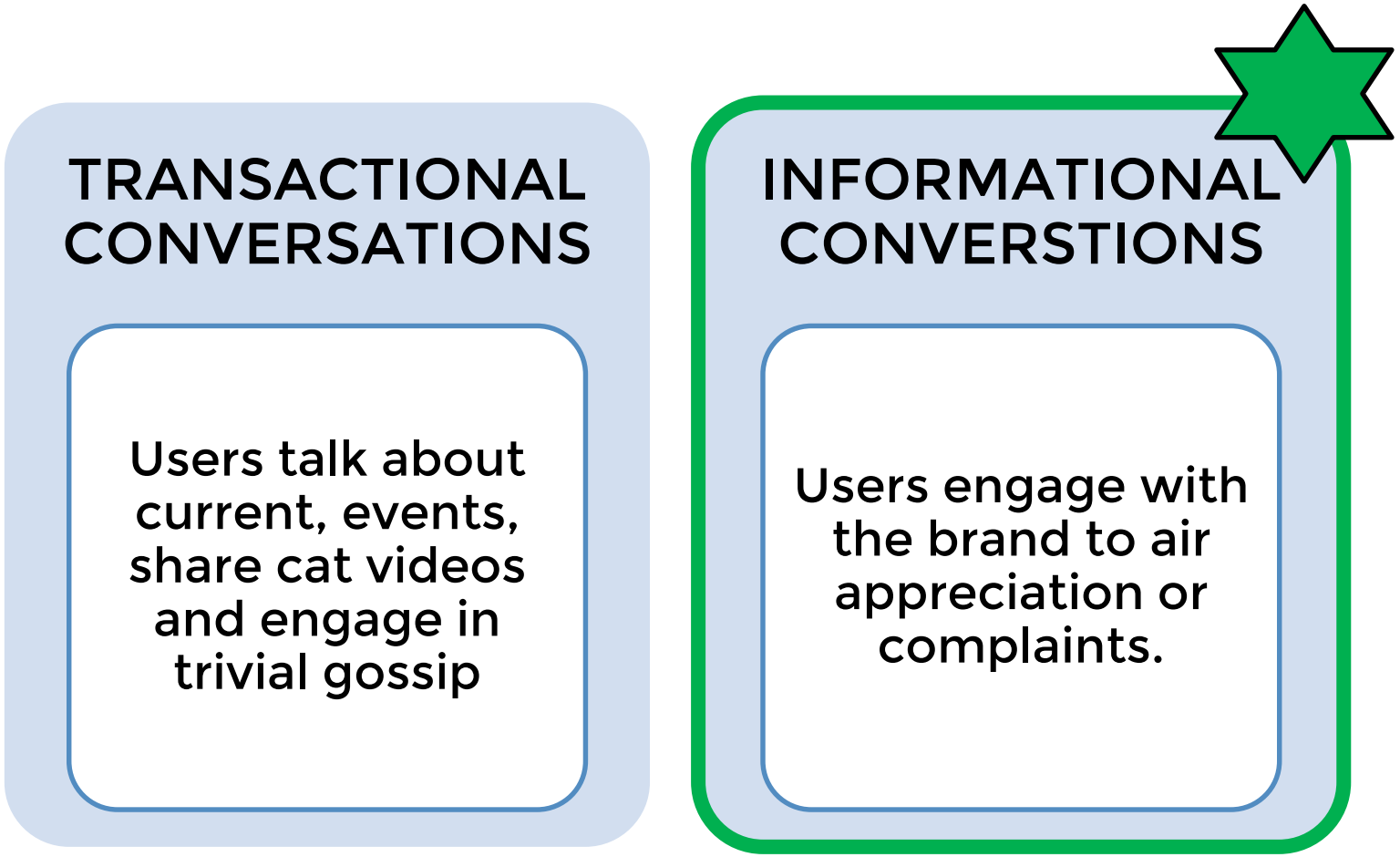
Starting off with Text Analytics

WHAT WOULD YOU WANT?

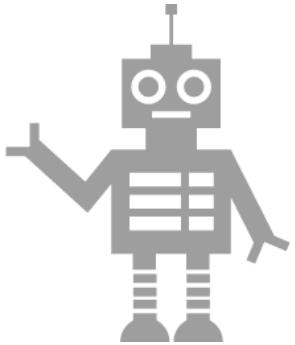
As a brand owner with significant investments in social media, the usual questions you might have in mind…

- Is the brand exuding same attributes I intended it to be?
- Is my internet presence helping me ?
- Can I measure my ROI for the money I spent?
- What are the measurable metrics for effective social media management?
- When can I exploit emerging trends for my brand?
- How can I understand my customers better?

WHAT WOULD YOU WANT TO TARGET?



BRIEF TERMINOLOGY



You build an algorithm, machine learns patterns, machine predicts, rinse & repeat.

MACHINE LEARNING

Analyzing unstructured text, assign structure, load into a BI/program to visualize

TEXT ANALYTICS

FOCUS AREAS

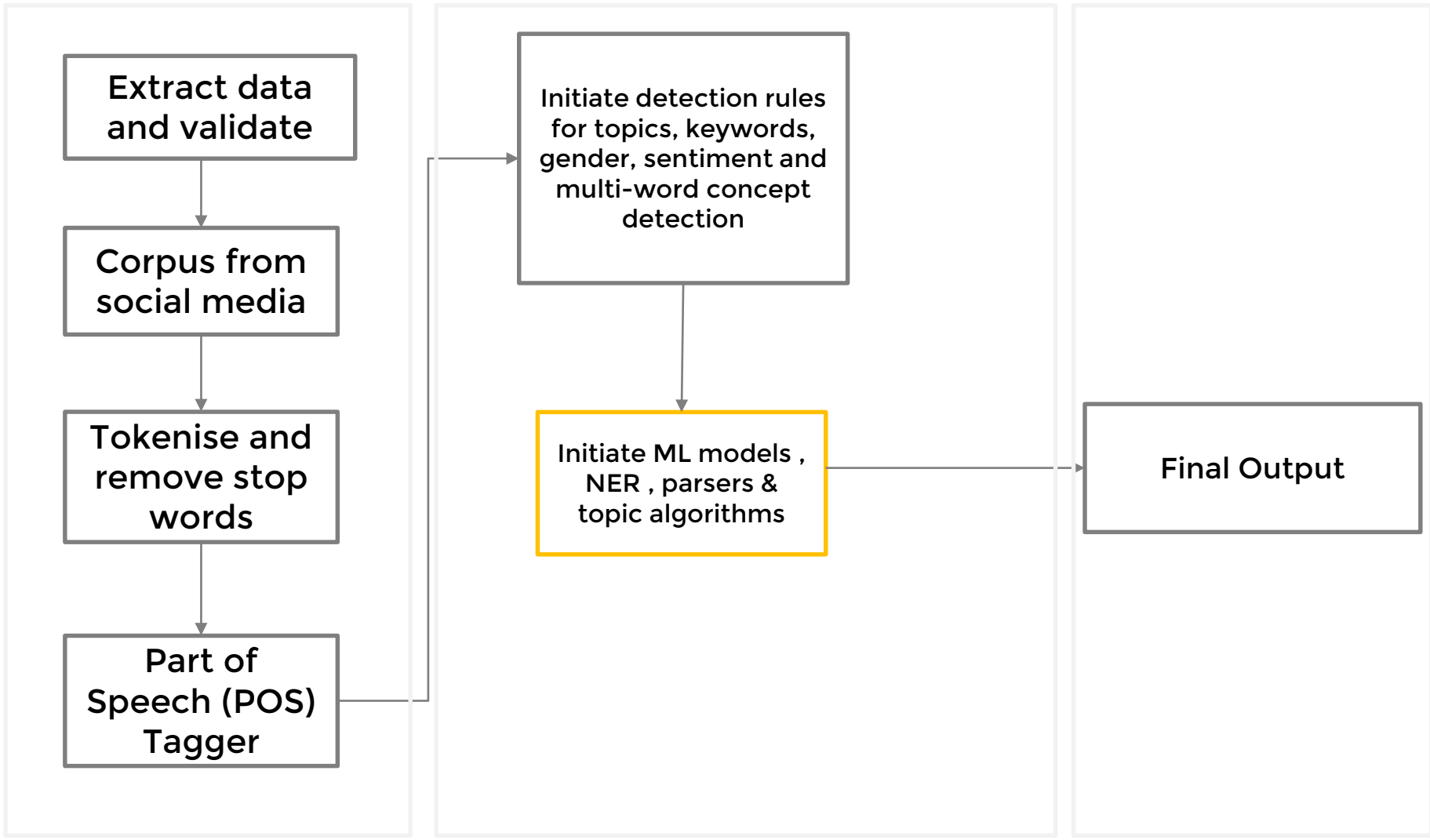
TOPICS

KEYWORDS

SENTIMENT

POINT OF SALES

APPROACH


Extract data and validate
Corpus from social media
Tokenise and remove stop words
Part of Speech (POS) Tagger
Initiate detection rules for topics, keywords, gender, sentiment and multi-word concept detection
Initiate ML models , NER , parsers & topic algorithms
Final Output
PRE - PROCESSING
PARSING & ANALYSIS
OUTPUT

The Samsung Galaxy S6 offers a killer mobile payments service, leaves Apple for dead

DUNCAN RILEY | MARCH 2ND

READ MORE

Tweet 18 | +1 2 | Like 2 | Share 9

Samsung Co. Ltd's launch of the Galaxy S6 Sunday shipped with one feature that is a killer in the mobile payments spaces and leaves Apple Inc. for dead, and that's a service Samsung has simply called Pay.

Based on technology from LoopPay, a startup Samsung acquired February 18th, Pay has two killer features that Apple doesn't provide: it offers both near field communications (NFC) and support for magnetic stripe cards, meaning it will work with legacy point of sales (POS)

Author	Value	Type	Sentiment
Duncan Riley	Samsung Galaxy S6	Entity	Positive
Duncan Riley	Apple	Entity	Negative
Duncan Riley	LoopPay	Entity	Neutral
Duncan Riley	mobile payments	Keyword	Positive
Duncan Riley	point of sales	Keyword	Positive

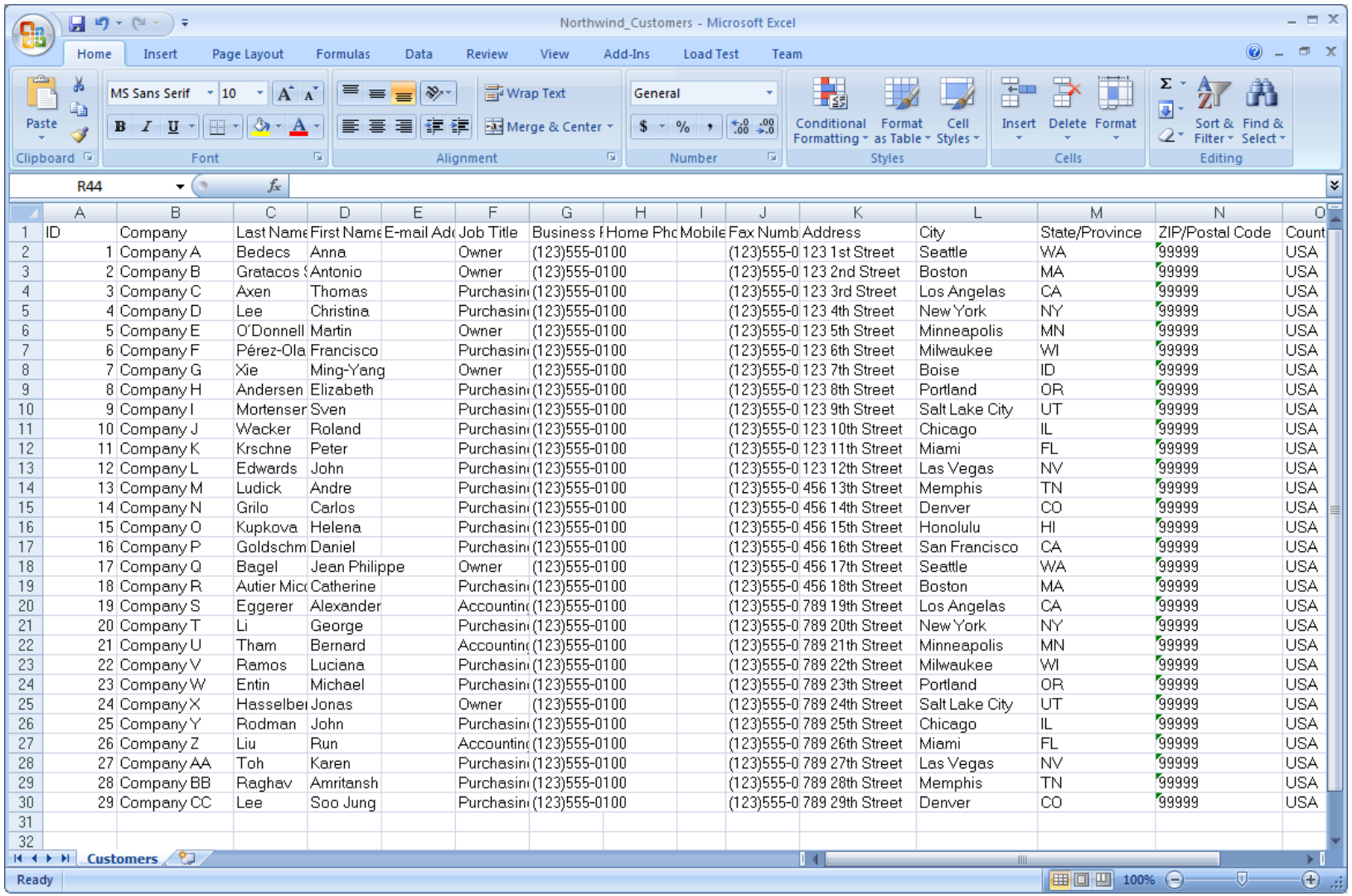
Actual blog post parsed through our SmartText Engine

Structuring data from free flowing text is easy to use by existing reporting and business intelligence software.
Insights from the final reports can now be used for decision-making by the PR firm and their client

BUT HEY ! THIS NEEDS ME TO WRITE A CODE.

YOU LIAR !

Excel



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	ID	Company	Last Name	First Name	E-mail Ad	Job Title	Business F	Home Pho	Mobile	Fax Numb	Address	City	State/Province	ZIP/Postal Code	Count
2	1	Company A	Bedecs	Anna		Owner	(123)555-0100			(123)555-0	123 1st Street	Seattle	WA	99999	USA
3	2	Company B	Gratacos S	Antonio		Owner	(123)555-0100			(123)555-0	123 2nd Street	Boston	MA	99999	USA
4	3	Company C	Axen	Thomas		Purchasin	(123)555-0100			(123)555-0	123 3rd Street	Los Angelas	CA	99999	USA
5	4	Company D	Lee	Christina		Purchasin	(123)555-0100			(123)555-0	123 4th Street	New York	NY	99999	USA
6	5	Company E	O'Donnell	Martin		Owner	(123)555-0100			(123)555-0	123 5th Street	Minneapolis	MN	99999	USA
7	6	Company F	Pérez-Ola	Francisco		Purchasin	(123)555-0100			(123)555-0	123 6th Street	Milwaukee	WI	99999	USA
8	7	Company G	Xie	Ming-Yang		Owner	(123)555-0100			(123)555-0	123 7th Street	Boise	ID	99999	USA
9	8	Company H	Andersen	Elizabeth		Purchasin	(123)555-0100			(123)555-0	123 8th Street	Portland	OR	99999	USA
10	9	Company I	Mortenser	Sven		Purchasin	(123)555-0100			(123)555-0	123 9th Street	Salt Lake City	UT	99999	USA
11	10	Company J	Wacker	Roland		Purchasin	(123)555-0100			(123)555-0	123 10th Street	Chicago	IL	99999	USA
12	11	Company K	Krschne	Peter		Purchasin	(123)555-0100			(123)555-0	123 11th Street	Miami	FL	99999	USA
13	12	Company L	Edwards	John		Purchasin	(123)555-0100			(123)555-0	123 12th Street	Las Vegas	NV	99999	USA
14	13	Company M	Ludick	Andre		Purchasin	(123)555-0100			(123)555-0	456 13th Street	Memphis	TN	99999	USA
15	14	Company N	Grilo	Carlos		Purchasin	(123)555-0100			(123)555-0	456 14th Street	Denver	CO	99999	USA
16	15	Company O	Kupkova	Helena		Purchasin	(123)555-0100			(123)555-0	456 15th Street	Honolulu	HI	99999	USA
17	16	Company P	Goldschm	Daniel		Purchasin	(123)555-0100			(123)555-0	456 16th Street	San Francisco	CA	99999	USA
18	17	Company Q	Bagel	Jean Philippe		Owner	(123)555-0100			(123)555-0	456 17th Street	Seattle	WA	99999	USA
19	18	Company R	Autier Mic	Catherine		Purchasin	(123)555-0100			(123)555-0	456 18th Street	Boston	MA	99999	USA
20	19	Company S	Eggerer	Alexander		Accountin	(123)555-0100			(123)555-0	789 19th Street	Los Angelas	CA	99999	USA
21	20	Company T	Li	George		Purchasin	(123)555-0100			(123)555-0	789 20th Street	New York	NY	99999	USA
22	21	Company U	Tham	Bernard		Accountin	(123)555-0100			(123)555-0	789 21th Street	Minneapolis	MN	99999	USA
23	22	Company V	Ramos	Luciana		Purchasin	(123)555-0100			(123)555-0	789 22th Street	Milwaukee	WI	99999	USA
24	23	Company W	Entin	Michael		Purchasin	(123)555-0100			(123)555-0	789 23th Street	Portland	OR	99999	USA
25	24	Company X	Hasselber	Jonas		Owner	(123)555-0100			(123)555-0	789 24th Street	Salt Lake City	UT	99999	USA
26	25	Company Y	Rodman	John		Purchasin	(123)555-0100			(123)555-0	789 25th Street	Chicago	IL	99999	USA
27	26	Company Z	Liu	Run		Accountin	(123)555-0100			(123)555-0	789 26th Street	Miami	FL	99999	USA
28	27	Company AA	Toh	Karen		Purchasin	(123)555-0100			(123)555-0	789 27th Street	Las Vegas	NV	99999	USA
29	28	Company BB	Raghav	Amritansh		Purchasin	(123)555-0100			(123)555-0	789 28th Street	Memphis	TN	99999	USA
30	29	Company CC	Lee	Soo Jung		Purchasin	(123)555-0100			(123)555-0	789 29th Street	Denver	CO	99999	USA
31															
32															

```
= ISNUMBER(SEARCH($N$1,H2))
```

BONUS !

PART 6

Create your own free social media listening dashboard !

Caravel | Security | Sources | Slices | Dashboards | admin user

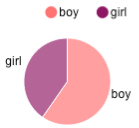
Births

Participants

1.11M

-22.5% over 5Y

Genders

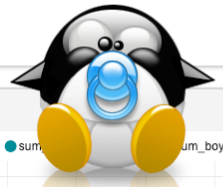


Name Cloud

Title

Birth Names Dashboard

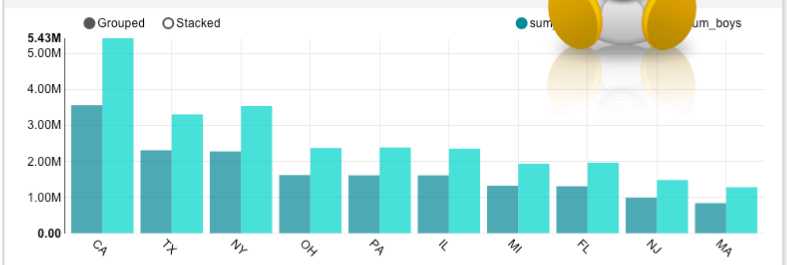
The source dataset came from [here]



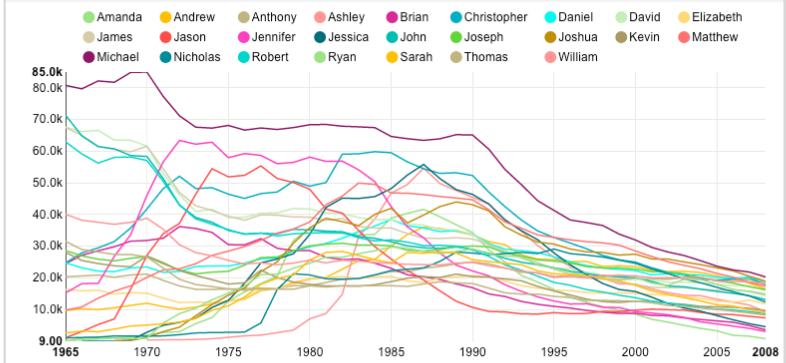
Pivot Table

	sum__num			
state	CA	FL	IL	MA
name				
All	8998550.0	3280653.0	3971838.0	21
Michael	259809.0	102070.0	134250.0	80
Christopher	185674.0	79368.0	72462.0	49
David	203216.0	62192.0	77826.0	44
James	122151.0	64316.0	72096.0	39
John	127360.0	56425.0	71680.0	47
Matthew	141032.0	51145.0	69172.0	46
Jennifer	159368.0	50954.0	70922.0	37

Genders by State



Trends



Girls

name	sum__num
Jennifer	1.34M
Jessica	997k
Ashley	789k
Sarah	745k
Amanda	720k
Elizabeth	713k
Melissa	665k
Michelle	659k
Kimberly	648k
Stephanie	628k

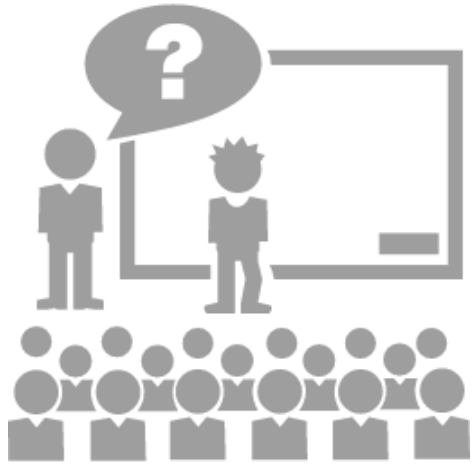
Boys

name	sum__num
Michael	2.47M
Christopher	1.73M
David	1.57M
James	1.51M
John	1.43M
Matthew	1.36M
Robert	1.31M
Daniel	1.16M
Joseph	1.11M
William	1.11M

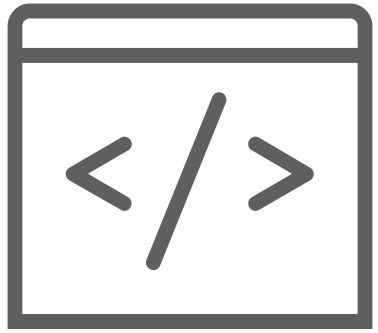
Caravel's main goal is to make it easy to slice, dice and visualize data. It empowers users to **perform analytics at the speed of thought.**

- A quick way to **intuitively visualize datasets** by allowing users to create and share interactive dashboards
- **A rich set of visualizations to analyze your data**, as well as a flexible way to extend the capabilities
- An extensible, **high granularity security model** allowing intricate rules on who can access which features, and integration with major authentication providers (database, OpenID, LDAP, OAuth & REMOTE_USER through Flask AppBuiler)
- A simple semantic layer, allowing to **control how data sources are displayed in the UI**, by defining which fields should show up in which dropdown and which aggregation and function (metrics) are made available to the user
- Deep integration with Druid allows for Caravel to stay **blazing fast** while slicing and dicing large, realtime datasets

LET’S GET OUR HANDS DIRTY, AGAIN !



QUESTIONS ?



https://goo.gl/VRKvFc

OR

https://github.com/manasRK/IIM_Lucknow_MRSI_April

#connect

MANAS RANJAN KAR

 manas@jsm.email

 +91-9971 420 188

 www.unlocktext.com