

Text Data to Insights – Creating a storyline with Social Data

Becoming a data ninja

A talk delivered at
Data Visualization - Methods & Tools
IIM Lucknow
25/04/2016

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

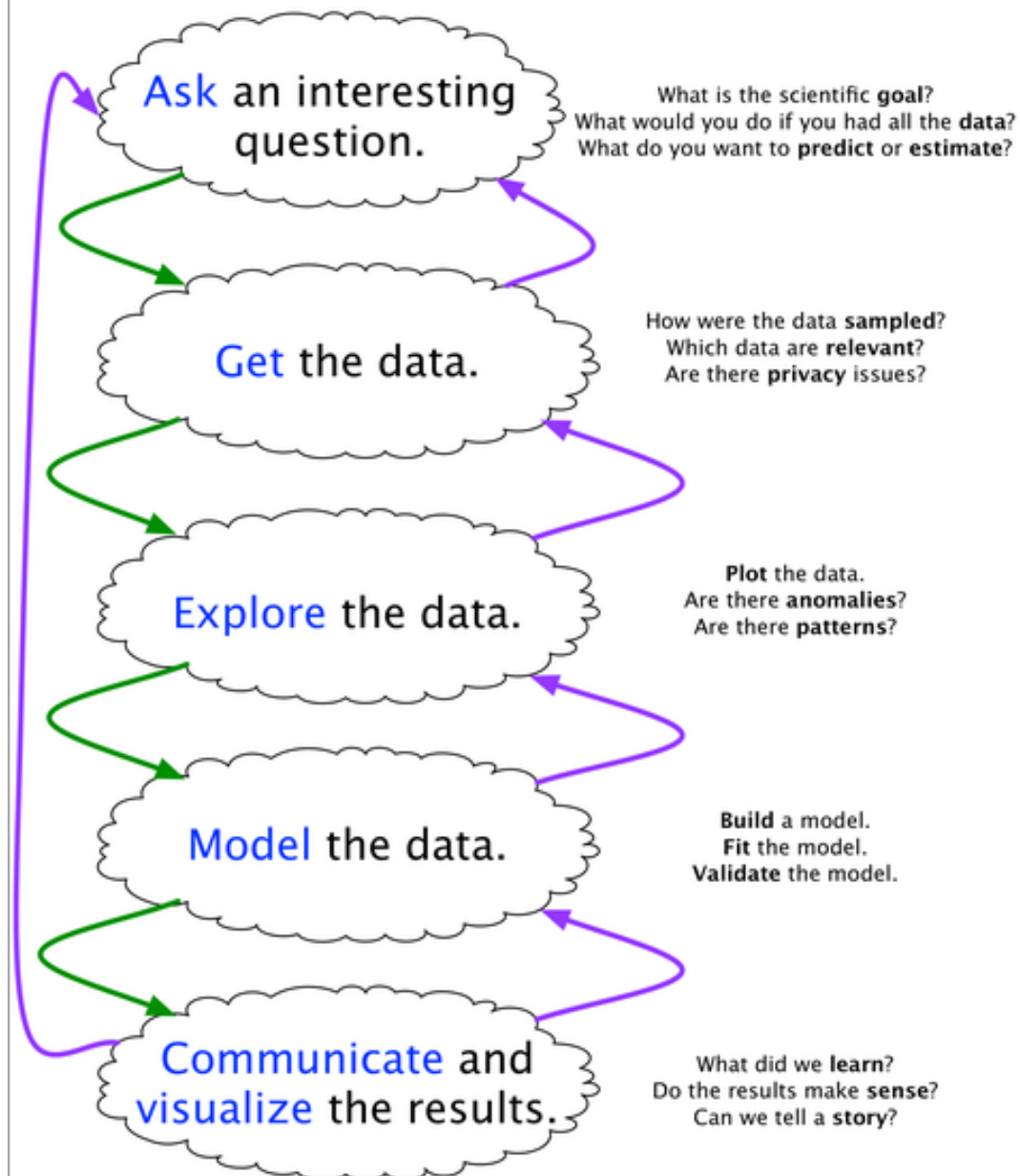
NLP, ML, Text

“Data scientist is
the sexiest job
of the 21st century.”

Harvard Business Review



The Data Science Process



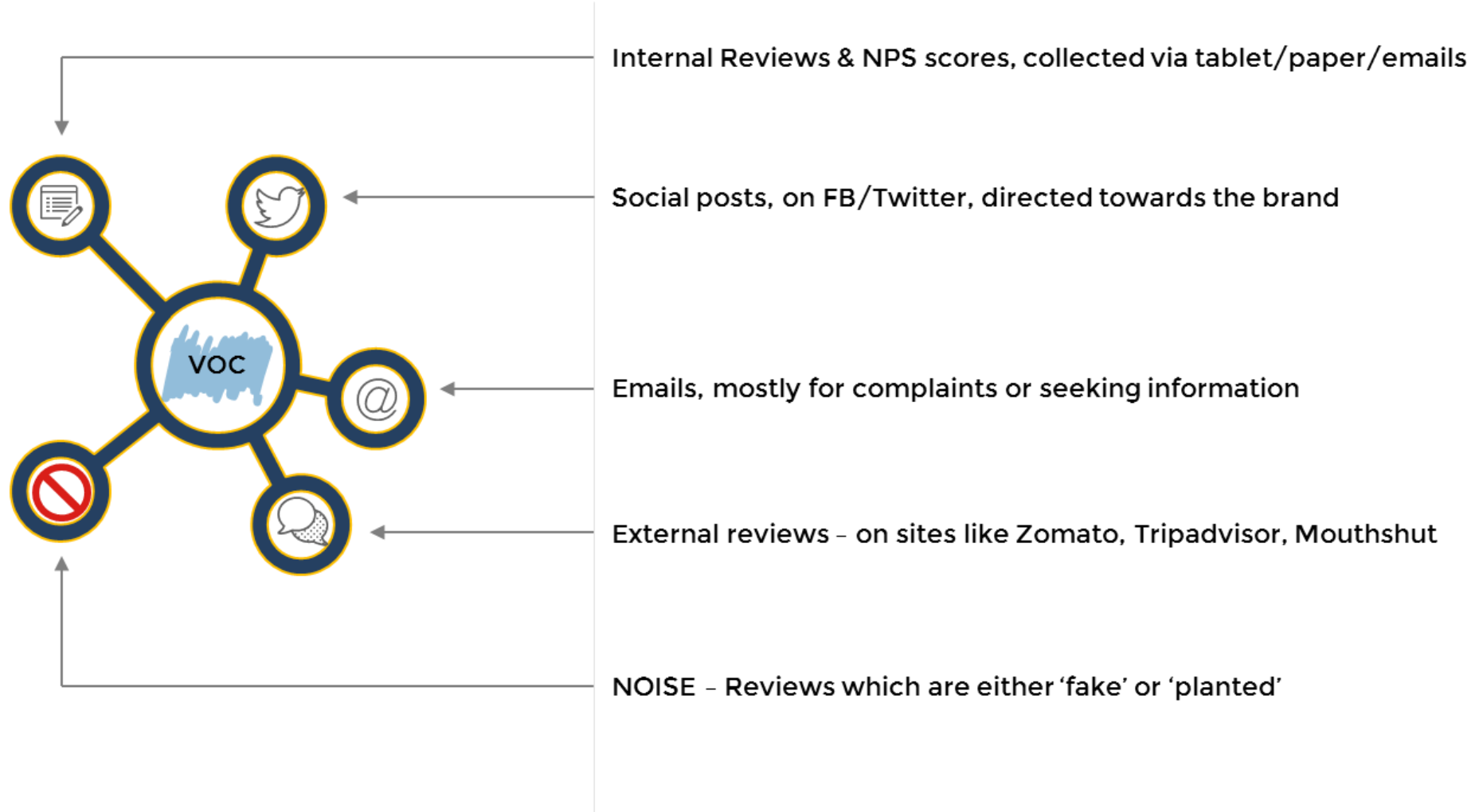
WE WILL TALK OF...

- 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



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

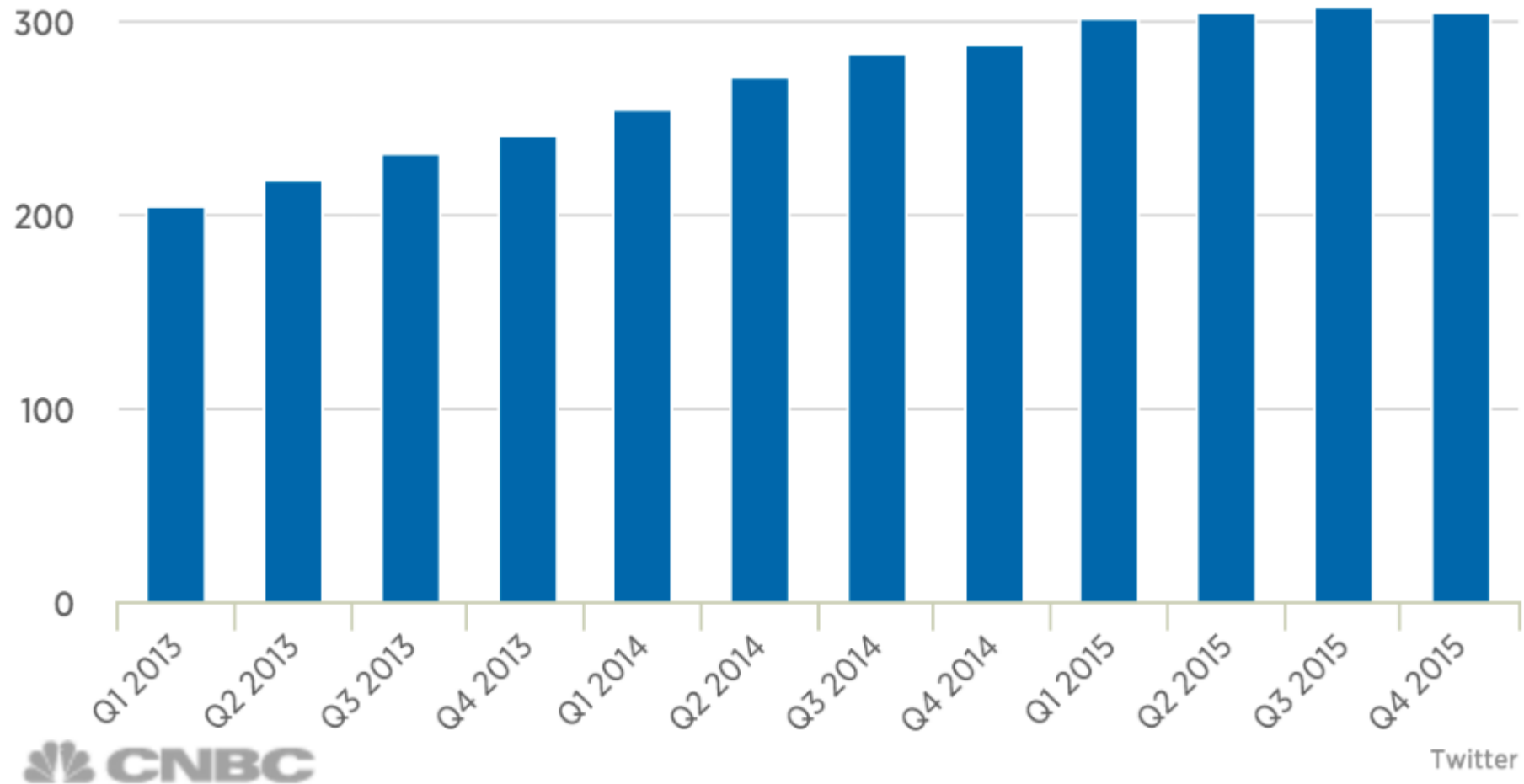
Focus on things people care about

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



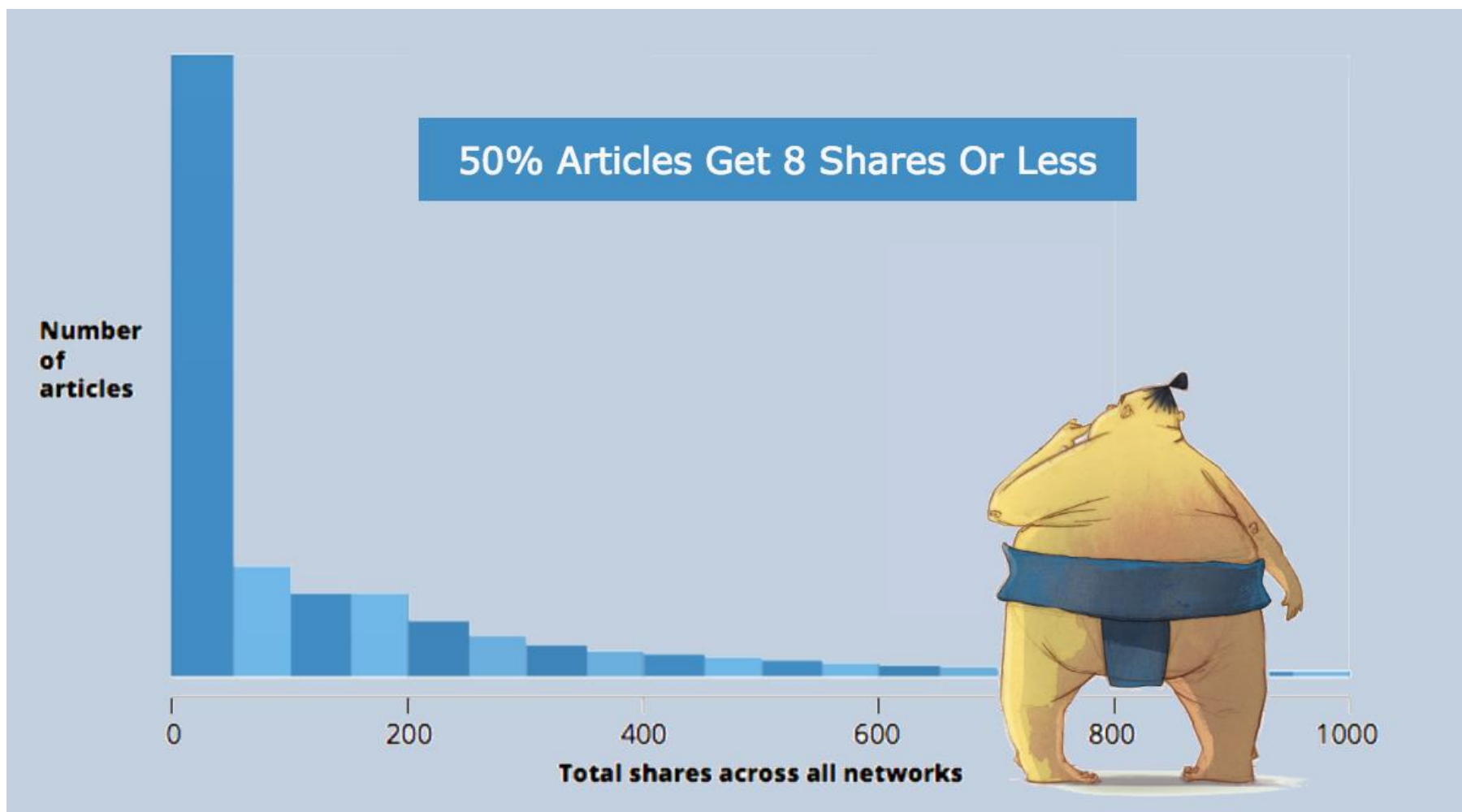
TABLES :/

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





GENDER: { female }

Ms. Prachi Tamanna

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

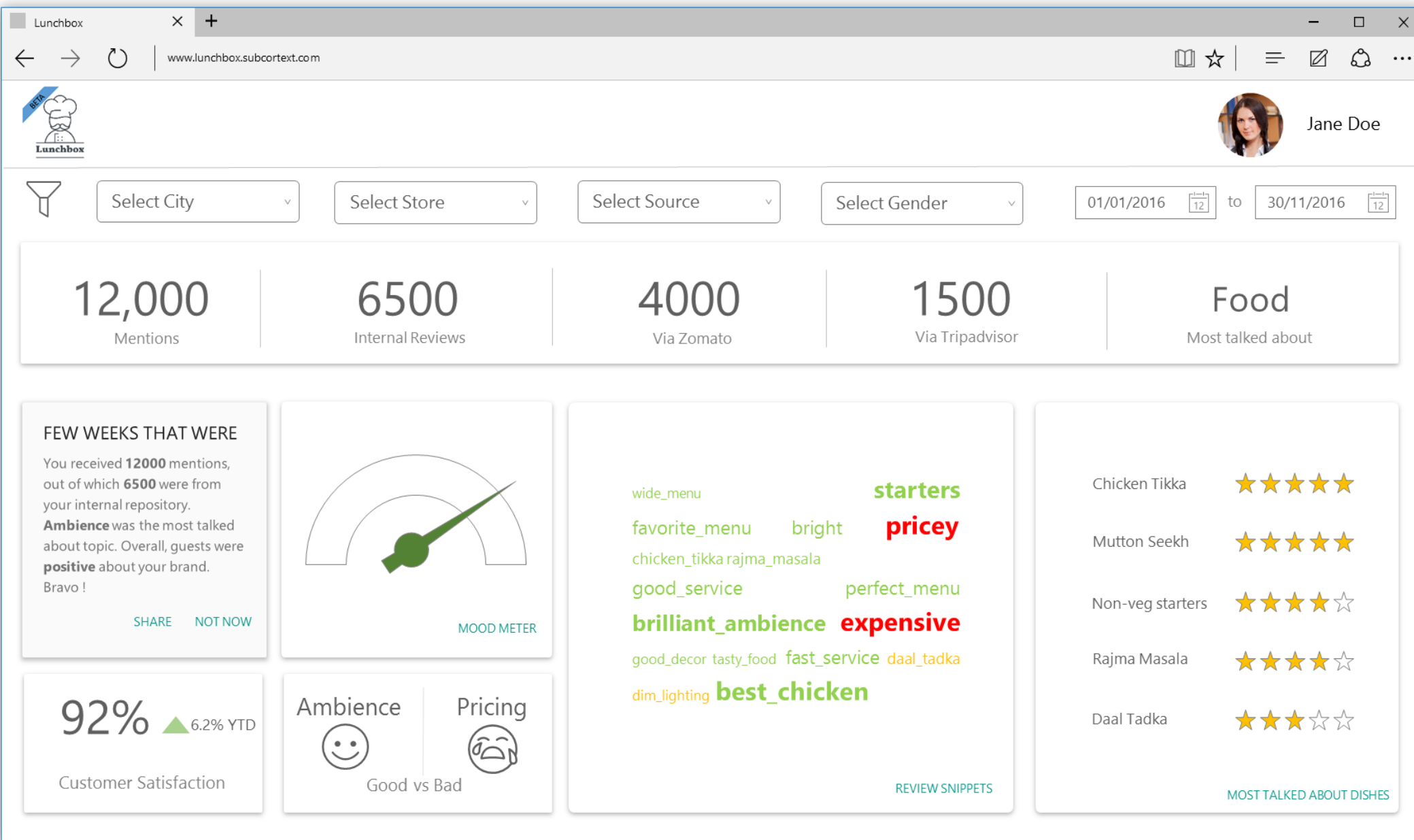
TOPIC: { service }

SENTIMENT: { positive }

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

PART 3

Some Examples



12,000

Mentions

6500

Internal Reviews

4000

Via Zomato

1500

Via Tripadvisor

Food

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



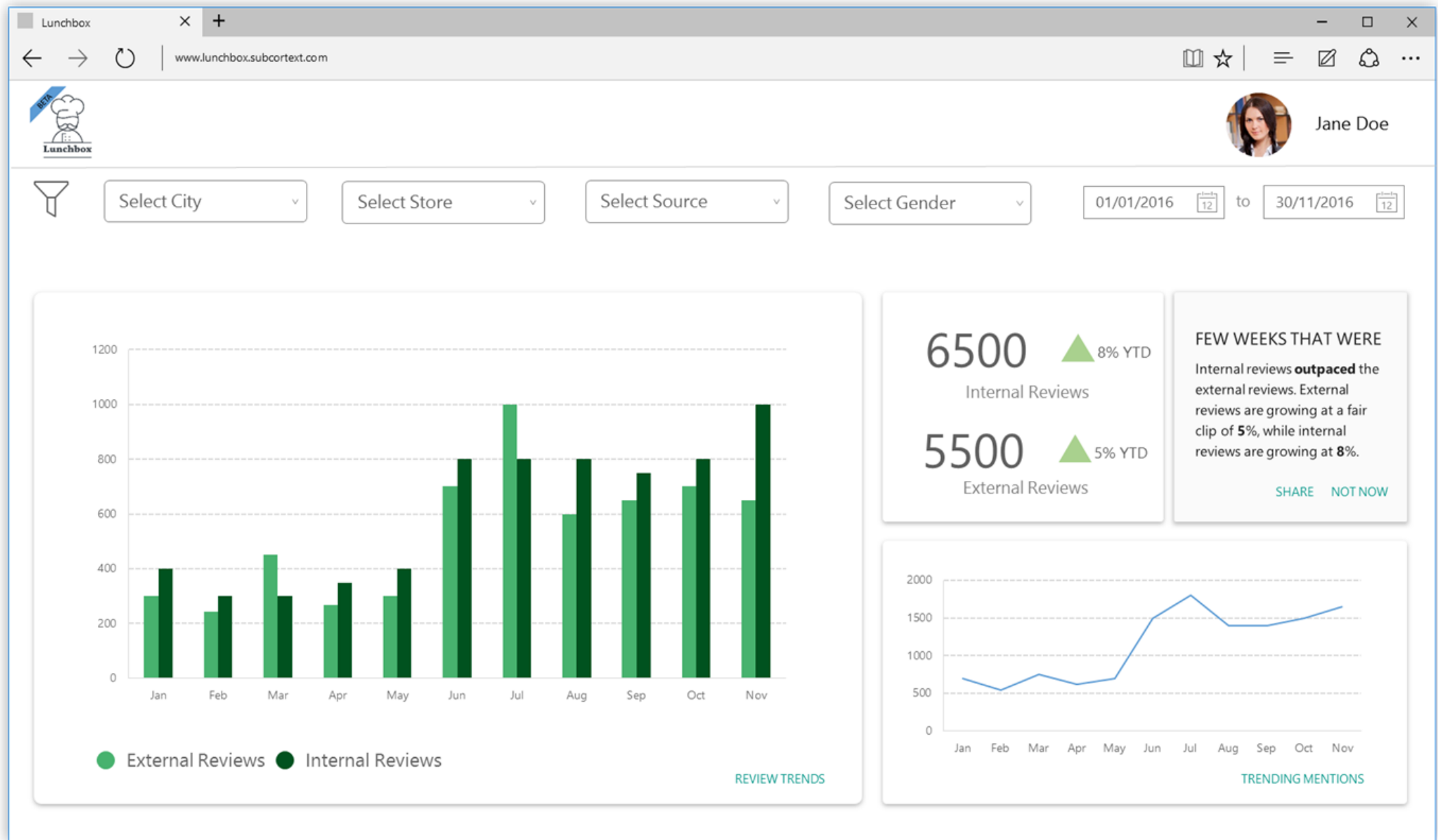
Good

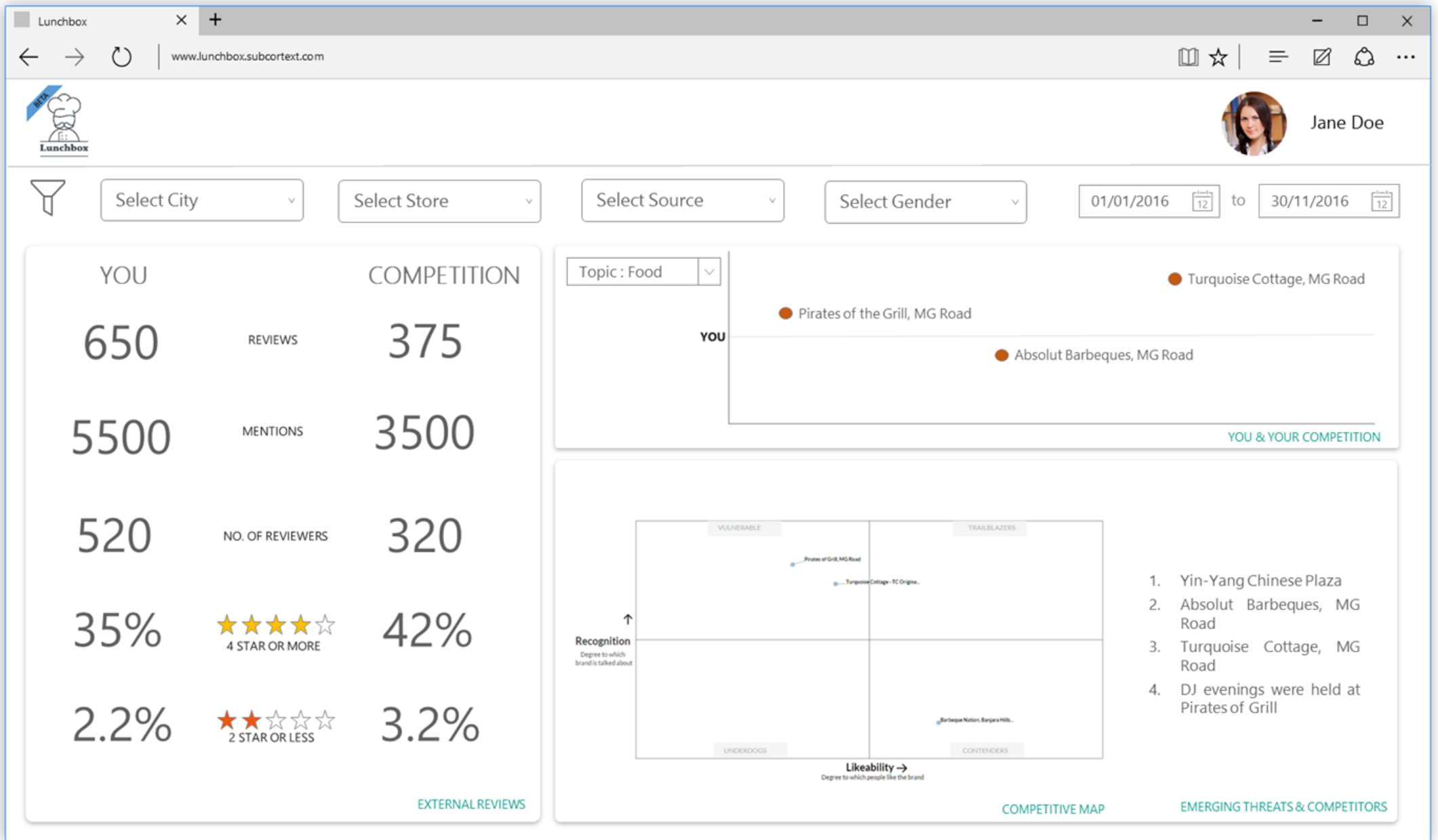
Pricing



Bad

Good vs Bad





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

+

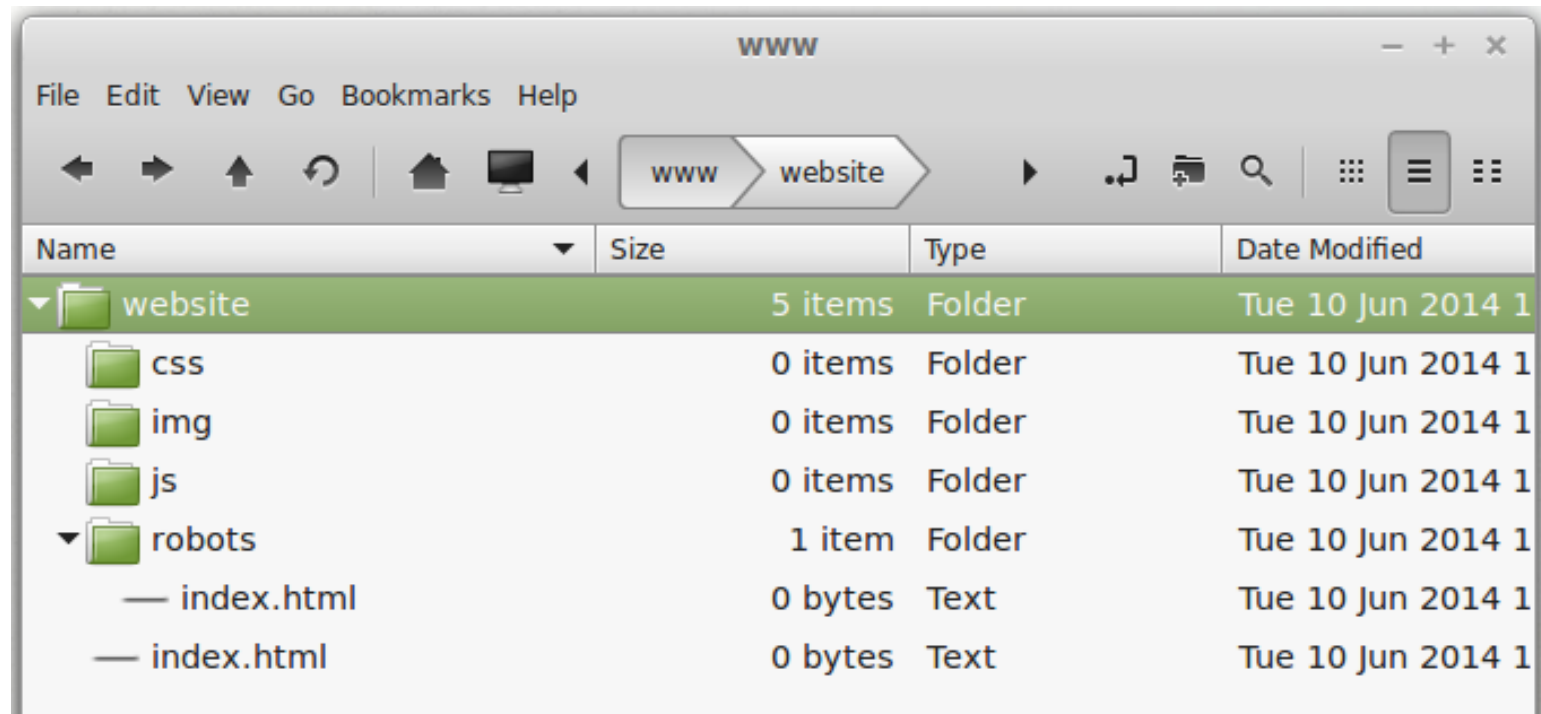
Comprehensive Analytics

Basic Structure of a HTML document

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

HTML PAGES

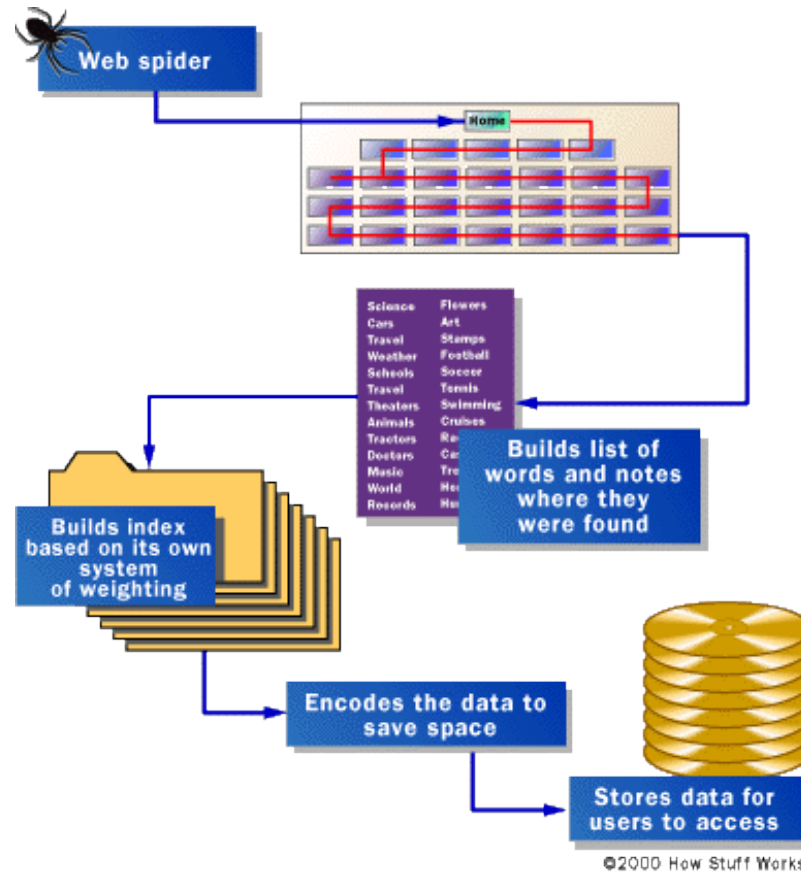
- HTML pages are like textbooks – content, titles, subtitles, paragraphs and so on
- Javascript adds interactivity to the HTML pages



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.consumercomplaints.in/snapdeal-com-b100038>

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?

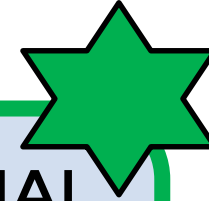
WHAT WOULD YOU WANT TO TARGET?

TRANSACTIONAL CONVERSATIONS

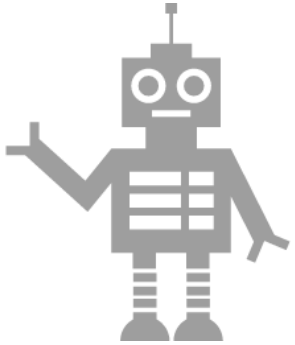
Users talk about
current, events,
share cat videos
and engage in
trivial gossip

INFORMATIONAL CONVERSTIONS

Users engage with
the brand to air
appreciation or
complaints.



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

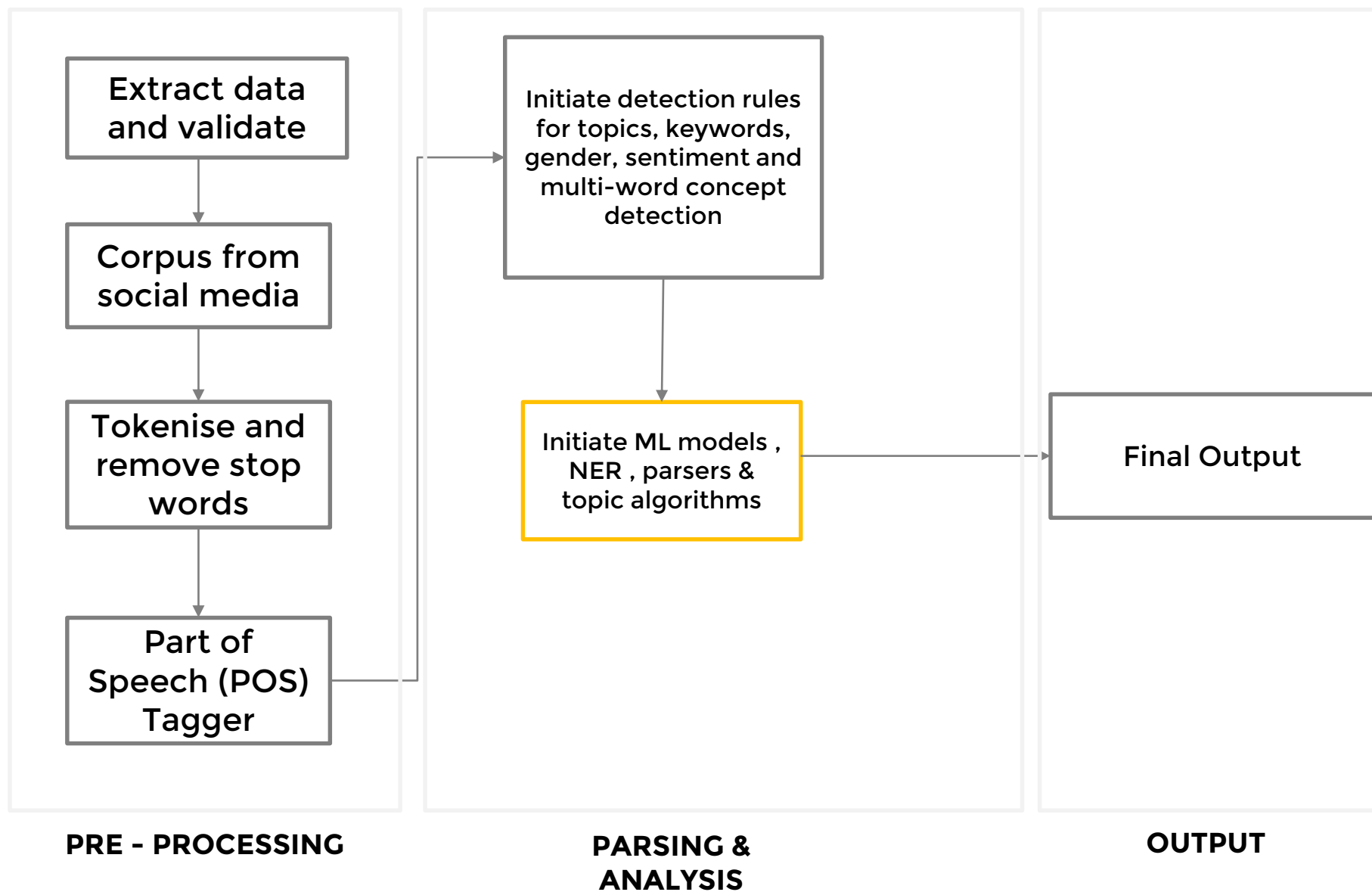


SENTIMENT



POINT OF SALES

APPROACH



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 in 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

Northwind_Customers - Microsoft Excel

HomeInsertPage LayoutFormulasDataReviewViewAdd-InsLoad TestTeam

Clipboard

Paste

MS Sans Serif10

Font

Alignment

General

\$%>

Number

Conditional Formatting

Format as Table

Cell Styles

Styles

Insert

Delete

Format

Cells

Editing

R44

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
	ID	Company	Last Name	First Name	E-mail Address	Job Title	Business Phone	Home Phone	Mobile	Fax Number	Address	City	State/Province	ZIP/Postal Code	Country
1	1	Company A	Bedecs	Anna		Owner	(123)555-0100			(123)555-0123	1st Street	Seattle	WA	99999	USA
2	2	Company B	Gratacos	Antonio		Owner	(123)555-0100			(123)555-0123	2nd Street	Boston	MA	99999	USA
3	3	Company C	Axen	Thomas		Purchasing	(123)555-0100			(123)555-0123	3rd Street	Los Angeles	CA	99999	USA
4	4	Company D	Lee	Christina		Purchasing	(123)555-0100			(123)555-0123	4th Street	New York	NY	99999	USA
5	5	Company E	O'Donnell	Martin		Owner	(123)555-0100			(123)555-0123	5th Street	Minneapolis	MN	99999	USA
6	6	Company F	Pérez-Ola	Francisco		Purchasing	(123)555-0100			(123)555-0123	6th Street	Milwaukee	WI	99999	USA
7	7	Company G	Xie	Ming-Yang		Owner	(123)555-0100			(123)555-0123	7th Street	Boise	ID	99999	USA
8	8	Company H	Andersen	Elizabeth		Purchasing	(123)555-0100			(123)555-0123	8th Street	Portland	OR	99999	USA
9	9	Company I	Mortensen	Sven		Purchasing	(123)555-0100			(123)555-0123	9th Street	Salt Lake City	UT	99999	USA
10	10	Company J	Wacker	Roland		Purchasing	(123)555-0100			(123)555-0123	10th Street	Chicago	IL	99999	USA
11	11	Company K	Krschne	Peter		Purchasing	(123)555-0100			(123)555-0123	11th Street	Miami	FL	99999	USA
12	12	Company L	Edwards	John		Purchasing	(123)555-0100			(123)555-0123	12th Street	Las Vegas	NV	99999	USA
13	13	Company M	Ludick	Andre		Purchasing	(123)555-0100			(123)555-0456	13th Street	Memphis	TN	99999	USA
14	14	Company N	Grilo	Carlos		Purchasing	(123)555-0100			(123)555-0456	14th Street	Denver	CO	99999	USA
15	15	Company O	Kupkova	Helena		Purchasing	(123)555-0100			(123)555-0456	15th Street	Honolulu	HI	99999	USA
16	16	Company P	Goldschm	Daniel		Purchasing	(123)555-0100			(123)555-0456	16th Street	San Francisco	CA	99999	USA
17	17	Company Q	Bagel	Jean Philippe		Owner	(123)555-0100			(123)555-0456	17th Street	Seattle	WA	99999	USA
18	18	Company R	Autier	Michelle		Purchasing	(123)555-0100			(123)555-0456	18th Street	Boston	MA	99999	USA
19	19	Company S	Eggerer	Alexander		Accounting	(123)555-0100			(123)555-0789	19th Street	Los Angeles	CA	99999	USA
20	20	Company T	Li	George		Purchasing	(123)555-0100			(123)555-0789	20th Street	New York	NY	99999	USA
21	21	Company U	Tham	Bernard		Accounting	(123)555-0100			(123)555-0789	21th Street	Minneapolis	MN	99999	USA
22	22	Company V	Ramos	Luciana		Purchasing	(123)555-0100			(123)555-0789	22th Street	Milwaukee	WI	99999	USA
23	23	Company W	Entin	Michael		Purchasing	(123)555-0100			(123)555-0789	23th Street	Portland	OR	99999	USA
24	24	Company X	Hasselber	Jonas		Owner	(123)555-0100			(123)555-0789	24th Street	Salt Lake City	UT	99999	USA
25	25	Company Y	Rodman	John		Purchasing	(123)555-0100			(123)555-0789	25th Street	Chicago	IL	99999	USA
26	26	Company Z	Liu	Run		Accounting	(123)555-0100			(123)555-0789	26th Street	Miami	FL	99999	USA
27	27	Company AA	Toh	Karen		Purchasing	(123)555-0100			(123)555-0789	27th Street	Las Vegas	NV	99999	USA
28	28	Company BB	Raghav	Amritansh		Purchasing	(123)555-0100			(123)555-0789	28th Street	Memphis	TN	99999	USA
29	29	Company CC	Lee	Soo Jung		Purchasing	(123)555-0100			(123)555-0789	29th Street	Denver	CO	99999	USA
30															
31															
32															

Customers

Ready

100%

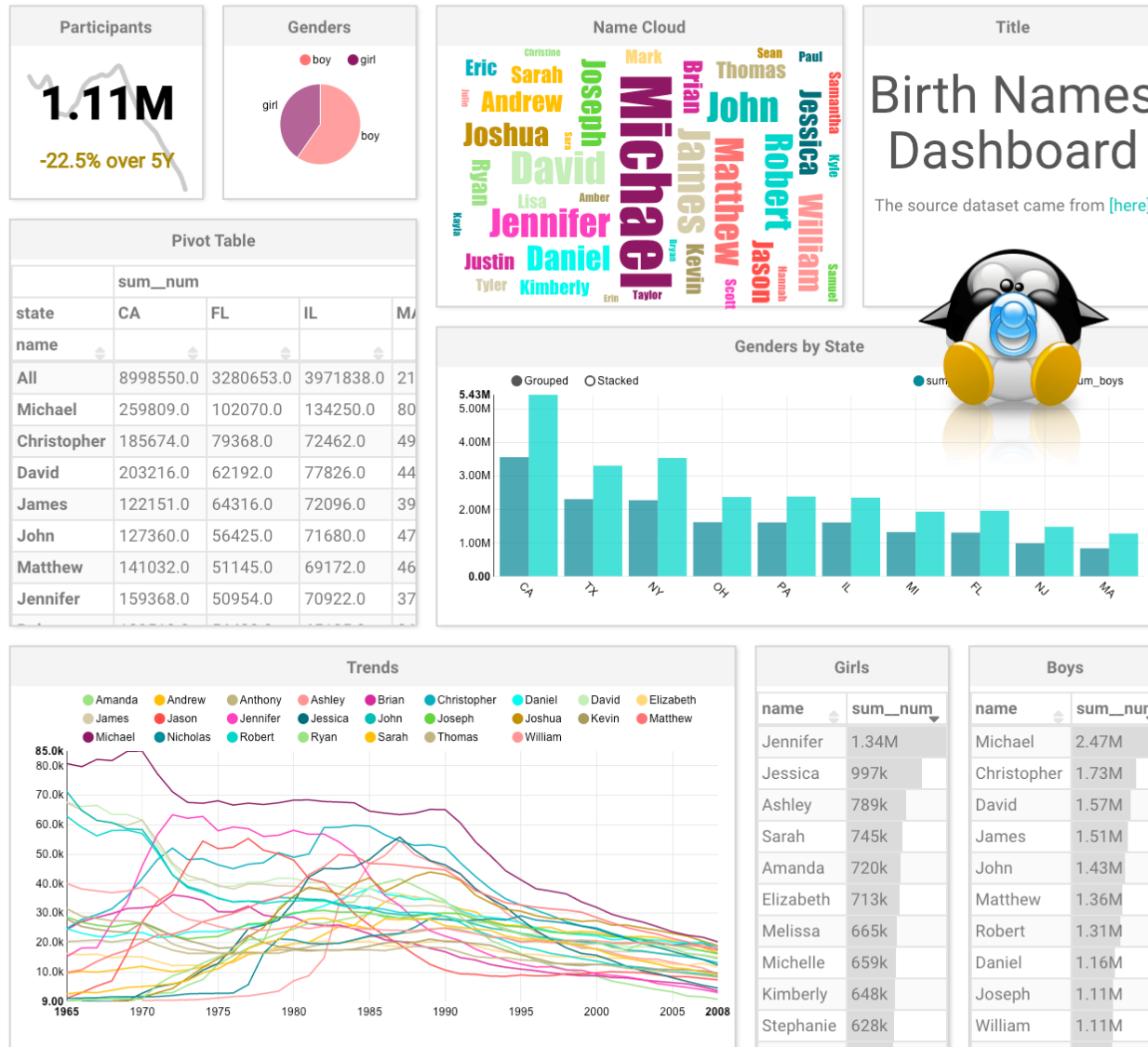
= ISNUMBER(SEARCH(\$N\$1,H2))

BONUS !

PART 6

Create your own free social media listening dashboard !

☆ Births

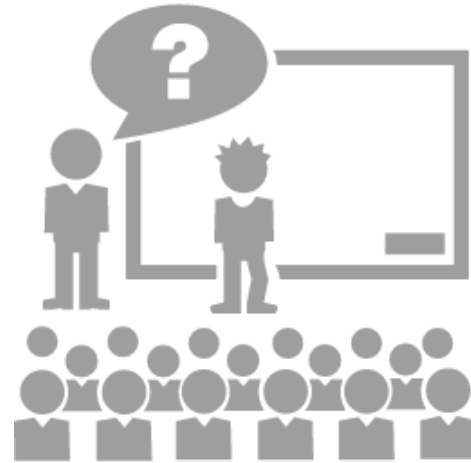


Caravel's main goal is to make it easy to slice, dice and visualize data. It empowers users to **perform analytics at the speed of thought.**

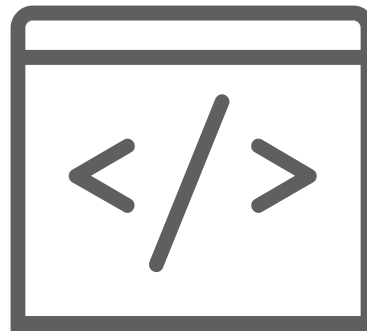
- A quick way to **intuitively visualize datasets** by allowing users to create and share interactive dashboards
- **A rich set of visualizations to analyze your data**, as well as a flexible way to extend the capabilities
- An extensible, **high granularity security model** allowing intricate rules on who can access which features, and integration with major authentication providers (database, OpenID, LDAP, OAuth & REMOTE_USER through Flask AppBuilder)
- 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