NUS
National University
of Singapore

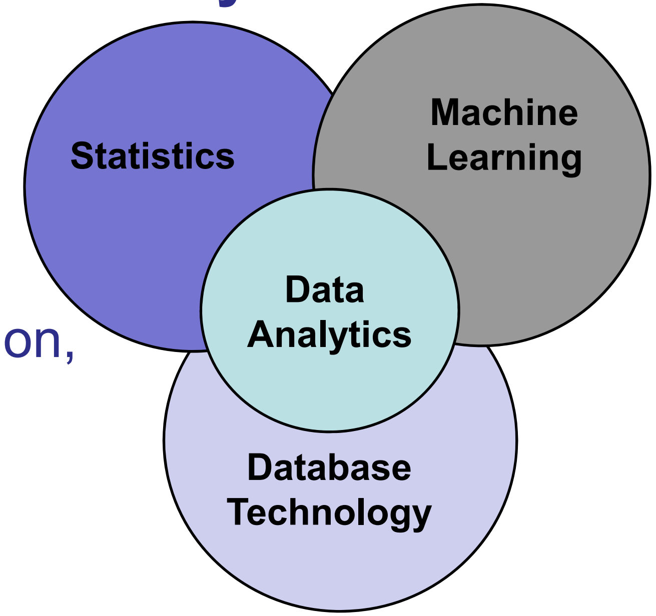
CS5344
Big Data Analytics Technology

Class Information

- **Lecturer: Lee Mong Li**
  - Email: leeml@comp.nus.edu.sg
- **Tutors:**
  - **Gao Qiao** (email: gaoqiao@comp.nus.edu.sg)
  - **Suman Bhoi** (email: e0267909@u.nus.edu)
- **Lectures on Tuesday 1830 – 2030 hrs**
- **Course website @ IVLE**
- **Reference text**
  - Mining of Massive Datasets by J. Leskovec, A. Rajaraman and J.D. Ullman (available online: http:///www.mmds.org)

Course Focus

- **Handle data that cannot fit in main memory**
  - Scalability of algorithms
  - Computing architecture
- **Real world problems**
  - Market basket analysis, Market segmentation, Recommender systems, Spam detection
- **Tools and Techniques**
  - MapReduce/Hadoop, Spark
    - create parallel algorithms to operate on large amount of data
  - Google's PageRank



Assessment

- **100% CA**
- **Lab Assignments (30%)**
- **Written Assessments (30%)**
- **Team-based Project (40%)**

You are reminded ***Plagiarism*** *is a very* ***SERIOUS*** *offence, and disciplinary action (including possibility of expulsion from the university) will be taken against any individual or team found plagiarizing.*

What is Big Data?

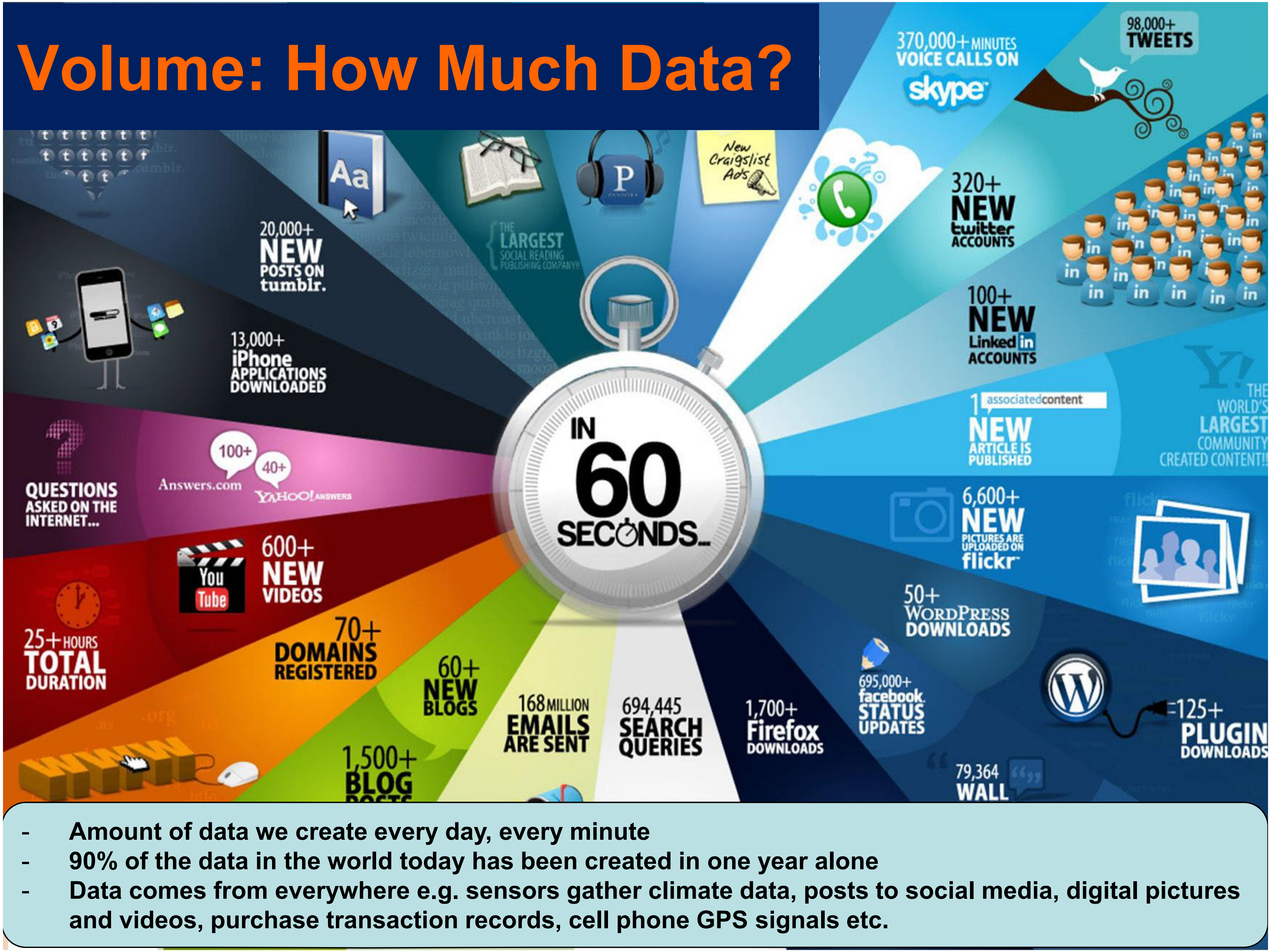
- **Gartner's Definition**

"Big data" is high-volume, -velocity and -variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.

- **Information assets characterized by 3Vs**
  - **High-volume (Terabytes → Zettabytes)**
  - **High-velocity (Batch → Streaming data)**
  - **High-variety (Structured → Semistructured & unstructured)**

Data becomes BIG when the volume, velocity or variety EXCEEDS the abilities of our IT systems to ingest, store, analyze and process it to derive actionable intelligence in a TIMELY manner.

Volume: How Much Data?

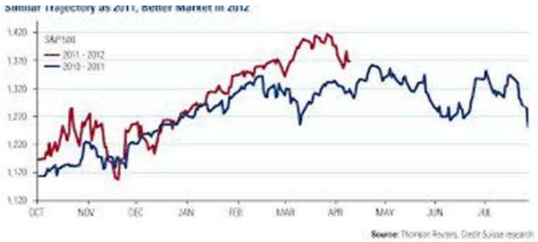


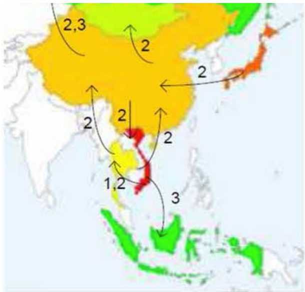
IN 60 SECONDS...
370,000+ MINUTES VOICE CALLS ON skype
98,000+ TWEETS
320+ NEW twitter ACCOUNTS
100+ NEW LinkedIn ACCOUNTS
1 associatedcontent NEW ARTICLE IS PUBLISHED
THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!
6,600+ NEW PICTURES ARE UPLOADED ON flickr
50+ WORDPRESS DOWNLOADS
125+ PLUGIN DOWNLOADS
695,000+ facebook STATUS UPDATES
79,364 WALL
1,700+ Firefox DOWNLOADS
694,445 SEARCH QUERIES
168 MILLION EMAILS ARE SENT
1,500+ BLOG POSTS
60+ NEW BLOGS
70+ DOMAINS REGISTERED
25+ HOURS TOTAL DURATION
600+ NEW VIDEOS
You Tube
QUESTIONS ASKED ON THE INTERNET...
100+ Answers.com
40+ YAHOO! ANSWERS
13,000+ iPhone APPLICATIONS DOWNLOADED
20,000+ NEW POSTS ON tumblr.
THE LARGEST SOCIAL READING PUBLISHING COMPANY!!
New Craigslist Ads

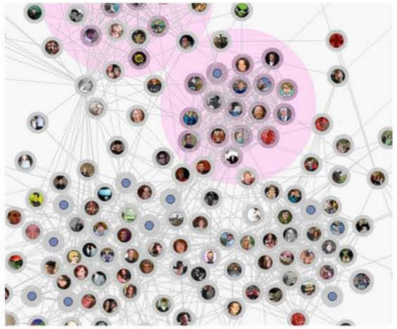
- Amount of data we create every day, every minute
- 90% of the data in the world today has been created in one year alone
- Data comes from everywhere e.g. sensors gather climate data, posts to social media, digital pictures and videos, purchase transaction records, cell phone GPS signals etc.

Variety: What Kind of Data?

- Relational databases
- Transactional databases
- XML databases
- Spatial databases
- Temporal databases
- Text databases and multimedia databases
- Graph databases







Relationships between people

YouTube

twitter

YAHOO! ANSWERS

Do not fit into a data warehouse, into neat tables of columns and rows.
Better place in Hadoop Distributed File System (HDFS) or in non-relational NoSQL databases.

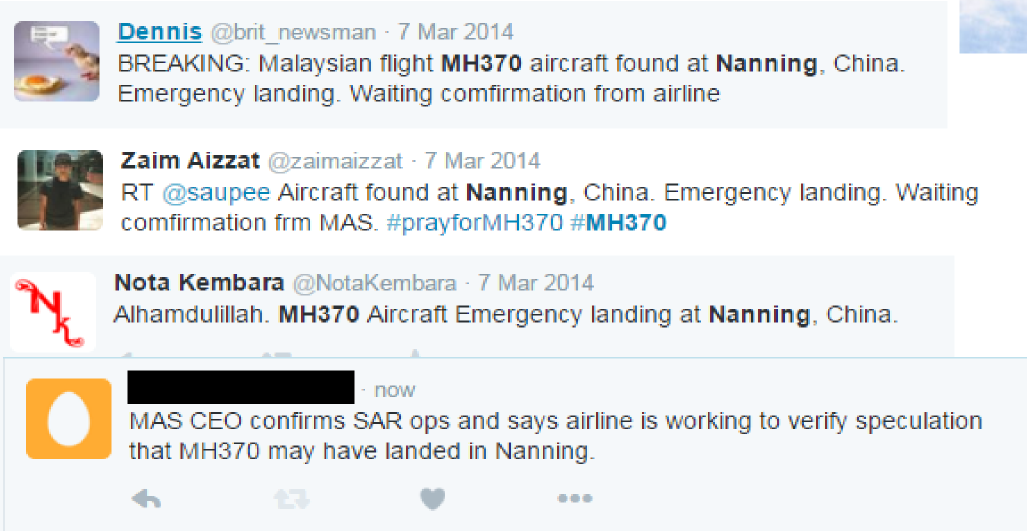
Velocity: At What Speed?

- Pace at which data flows in from sources
- Bursts of activities
- Real-time analysis
  - Late decisions → Missed opportunities

Fourth V - Veracity

NUS National University of Singapore

- How accurate or trustworthy is the data?
- Bias, inconsistencies
- Reliability of data source



Dennis @brit_newsman · 7 Mar 2014
BREAKING: Malaysian flight **MH370** aircraft found at **Nanning**, China. Emergency landing. Waiting comfirmation from airline

Zaim Aizzat @zaimaizzat · 7 Mar 2014
RT @saupee Aircraft found at **Nanning**, China. Emergency landing. Waiting comfirmation frm MAS. #prayforMH370 #**MH370**

Nota Kembara @NotaKembara · 7 Mar 2014
Alhamdulillah. **MH370** Aircraft Emergency landing at **Nanning**, China.

[redacted] · now
MAS CEO confirms SAR ops and says airline is working to verify speculation that MH370 may have landed in Nanning.

Why Big Data?

- Can collect cheaply, due to automation

$5 million vs $500

Price of fastest supercomputer in 1975 and iPhone with comparable performance

- Can store cheaply, due to falling media prices

$600 to buy a disk drive that can store all of the world's music

- Can create **Value**
  - Turn 12 terabytes of tweets created each day into improved product sentiment analysis
  - Convert 350 billion meter readings to better predict power consumption
  - Find communication patterns of successful projects in emails
  - Analyze elevator logs to predict vacated real estate
  - Scrutinize 5 million trade events created each day to identify potential fraud (time-sensitive, sometimes 2 minutes is too late)
  - Monitor 100's of live video feeds from surveillance cameras to target points of interest (new insights when you link and analyse different data types together)

Why Big Data?

Data contains Value and Knowledge

POPULAR SCIENCE
THE FUTURE NOW
THE CONTROL CENTERS
Using Data to Feed the World, Solve Cold Cases, Battle Malware, Predict Our Fate
OFFICER ALGORITHM
Can a Crime Be Prevented Before It Begins?
NEW WAYS OF SEEING
A Gallery of Extraordinary Infographics
PLUS
Juan Enriquez Reprograms Life
James Gleick Unsplits the Bit
AND
Lawrence Weschler Questions the Cloud
SPECIAL ISSUE
DATA IS POWER
HOW INFORMATION IS DRIVING THE FUTURE

Big Data Analytics

- From raw data to actionable information
- Data needs to be
  - Stored
  - Managed
  - and **ANALYZED**

Discover - Do we really know what we have?

Explore - How do different data relate to each other?

Iterative - What are the actual relationships?

Data Mining ≈ Big Data ≈
Predictive Analytics ≈ Data Science

Data Analytics/ Data Mining

- **Discover patterns and models that are**
  - **Valid:** hold on new data with some certainty
  - **Useful:** should be possible to act on the item
  - **Unexpected:** non-obvious to the system
  - **Understandable:** humans should be able to interpret the pattern

Data Mining Tasks

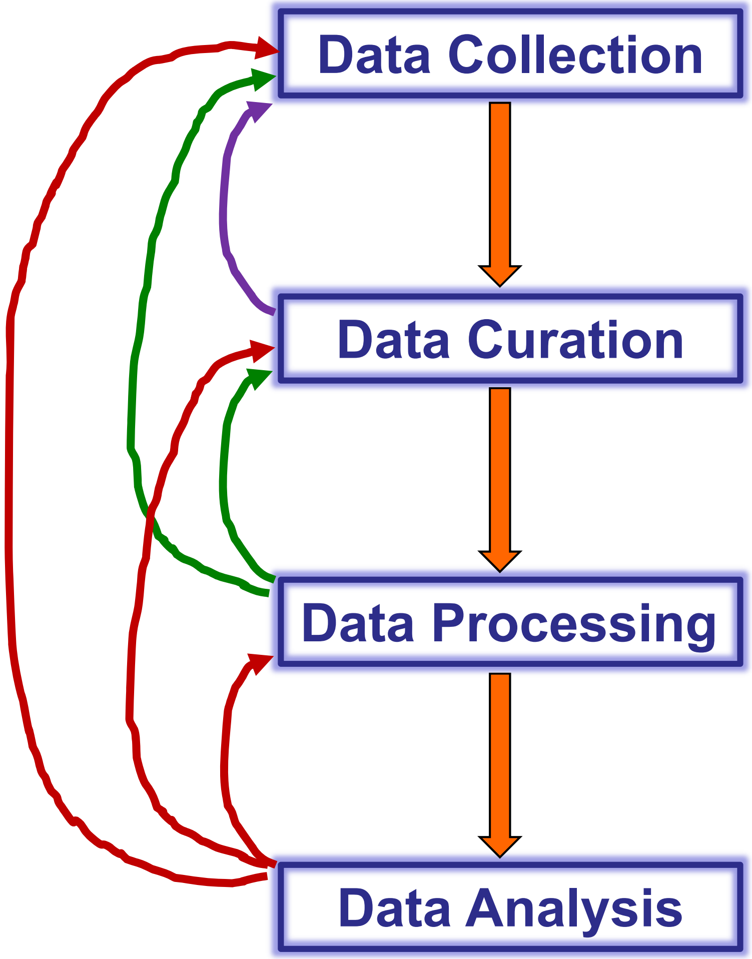
- **Descriptive methods**
  - Find human-interpretable patterns that describe the data
  - Example: **Clustering**

- **Predictive methods**
  - Use some variables to predict unknown or future values of other variables
  - Example: **Recommender systems**

movielens

helping you find the *right* movies

Big Data Analytics Pipeline



Acquire data from different sources

Clean, format, integrate with other datasets, store in database

Run queries (aggregate), plot graphs

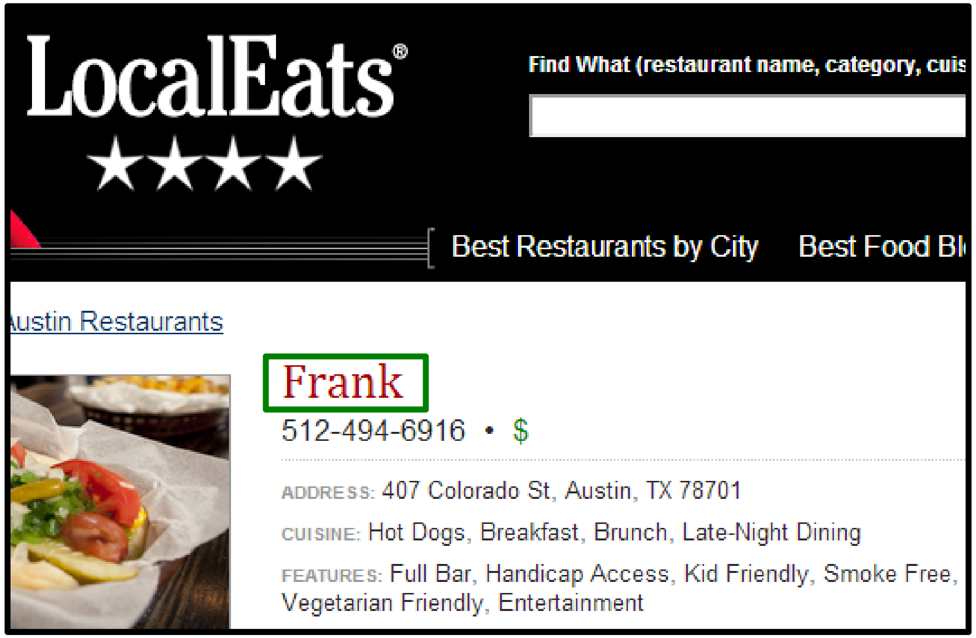
Examine trends and anomalies, understand results

Data Integration and Cleaning

Garbage in ➡ Garbage out

- Quality of results relates directly to quality of data
- 50% to 70% of analytics process effort is spent on data integration and cleaning
- Problems: duplicate records, entity resolution, conflict resolution, missing values, outliers, etc

Data from Different Sources



LocalEats®

Find What (restaurant name, category, cuis

Best Restaurants by City Best Food Bl

ustin Restaurants

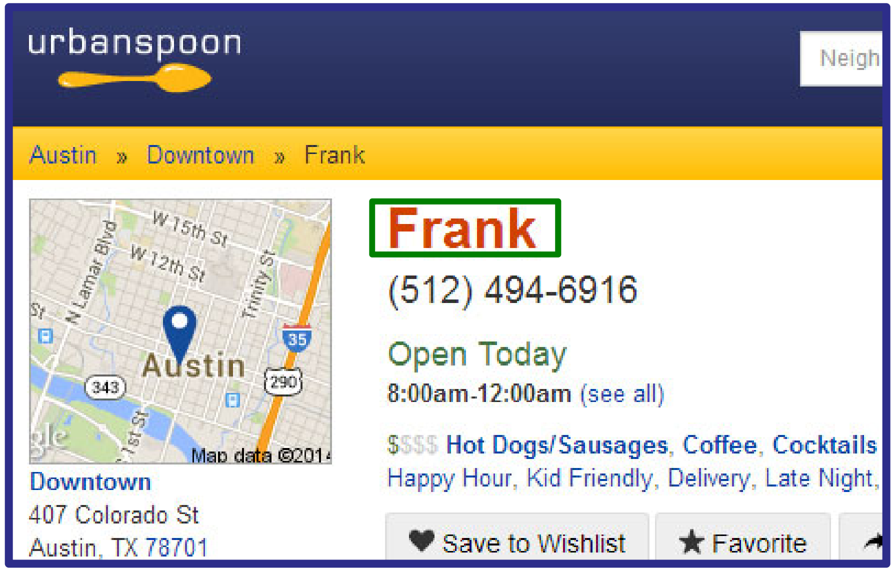
Frank

512-494-6916 • $

ADDRESS: 407 Colorado St, Austin, TX 78701

CUISINE: Hot Dogs, Breakfast, Brunch, Late-Night Dining

FEATURES: Full Bar, Handicap Access, Kid Friendly, Smoke Free, Vegetarian Friendly, Entertainment



urbanspoon

Neigh

Austin » Downtown » Frank

Frank

(512) 494-6916

Open Today

8:00am-12:00am (see all)

$$$$ Hot Dogs/Sausages, Coffee, Cocktails

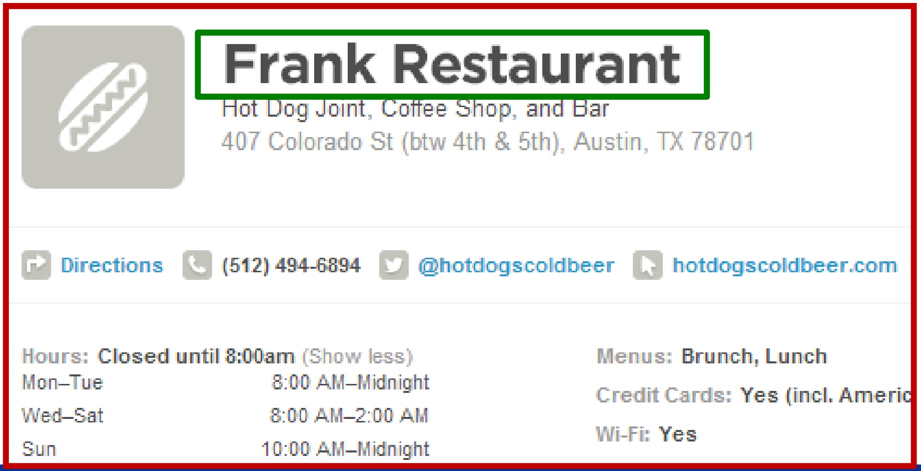
Happy Hour, Kid Friendly, Delivery, Late Night,

Downtown

407 Colorado St

Austin, TX 78701

♥ Save to Wishlist ★ Favorite



Frank Restaurant

Hot Dog Joint, Coffee Shop, and Bar

407 Colorado St (btw 4th & 5th), Austin, TX 78701

Directions (512) 494-6894 @hotdogscoldbeer hotdogscoldbeer.com

Hours: Closed until 8:00am (Show less)

Mon–Tue	8:00 AM–Midnight
Wed–Sat	8:00 AM–2:00 AM
Sun	10:00 AM–Midnight

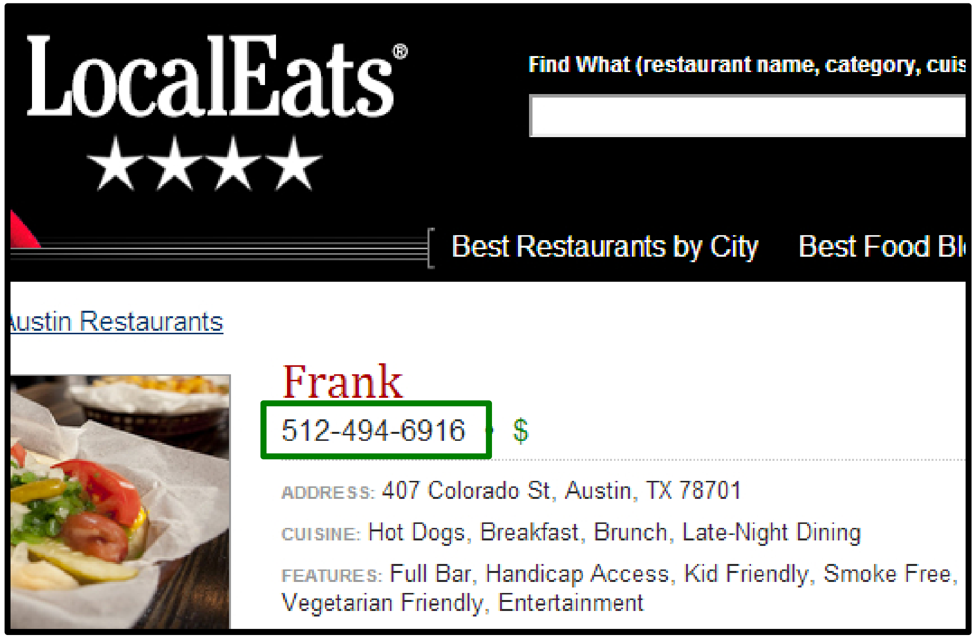
Menus: Brunch, Lunch

Credit Cards: Yes (incl. Americ

Wi-Fi: Yes

- **Different name representations**

Data from Different Sources



LocalEats®

★★★★

Find What (restaurant name, category, cuis

Best Restaurants by City   Best Food Bl

ustin Restaurants

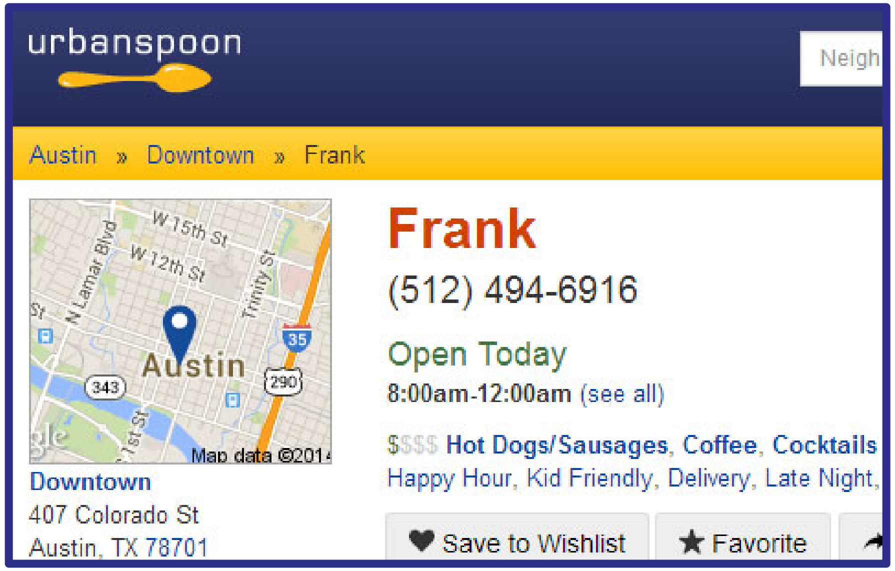
Frank

512-494-6916 · $

ADDRESS: 407 Colorado St, Austin, TX 78701

CUISINE: Hot Dogs, Breakfast, Brunch, Late-Night Dining

FEATURES: Full Bar, Handicap Access, Kid Friendly, Smoke Free, Vegetarian Friendly, Entertainment



urbanspoon

Neigh

Austin » Downtown » Frank

Frank

(512) 494-6916

Open Today

8:00am-12:00am (see all)

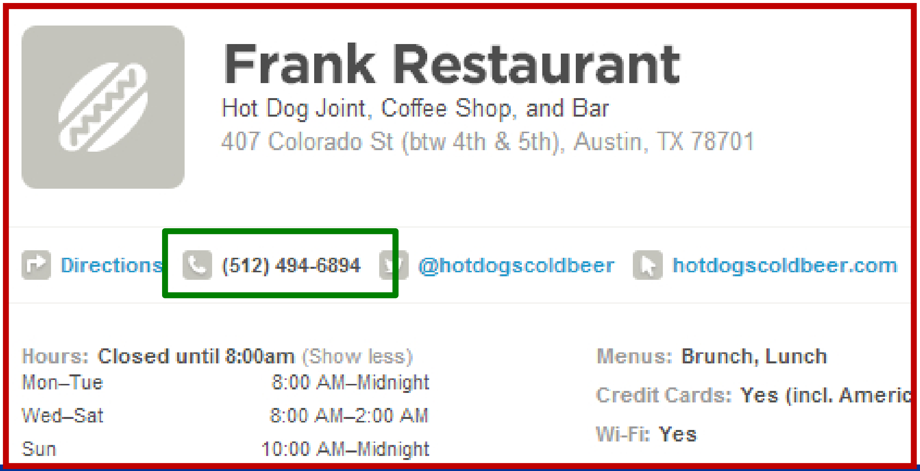
$$$$ Hot Dogs/Sausages, Coffee, Cocktails

Happy Hour, Kid Friendly, Delivery, Late Night,

Downtown

407 Colorado St

Austin, TX 78701



Frank Restaurant

Hot Dog Joint, Coffee Shop, and Bar

407 Colorado St (btw 4th & 5th), Austin, TX 78701

Directions   (512) 494-6894   @hotdogscoldbeer   hotdogscoldbeer.com

Hours: Closed until 8:00am (Show less)

Mon–Tue	8:00 AM–Midnight
Wed–Sat	8:00 AM–2:00 AM
Sun	10:00 AM–Midnight

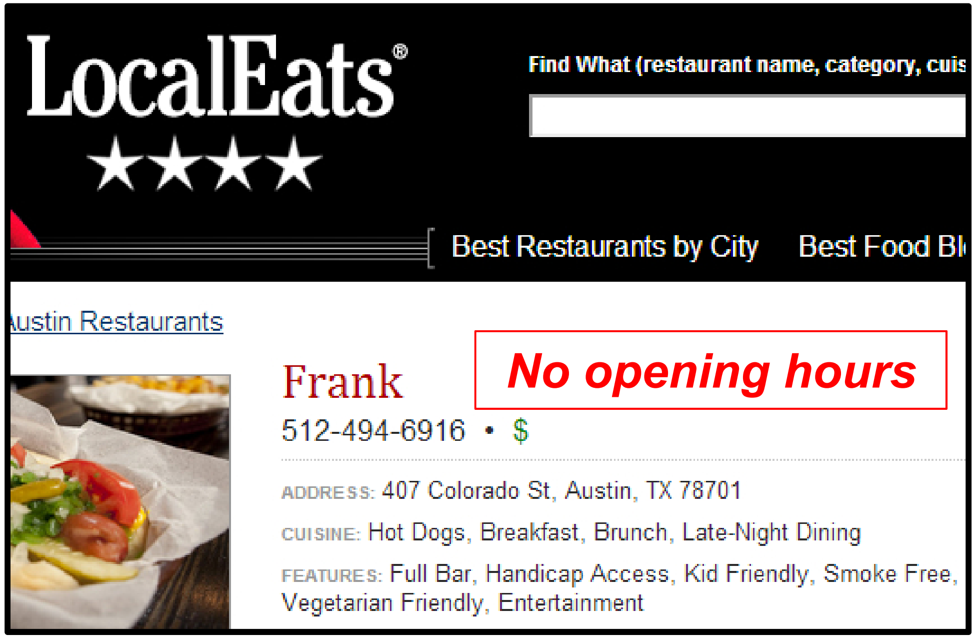
Menus: Brunch, Lunch

Credit Cards: Yes (incl. Americ

Wi-Fi: Yes

- **Erroneous attribute values**

Data from Different Sources



LocalEats®
★★★★

Find What (restaurant name, category, cuis

Best Restaurants by City   Best Food Bl

ustin Restaurants

Frank

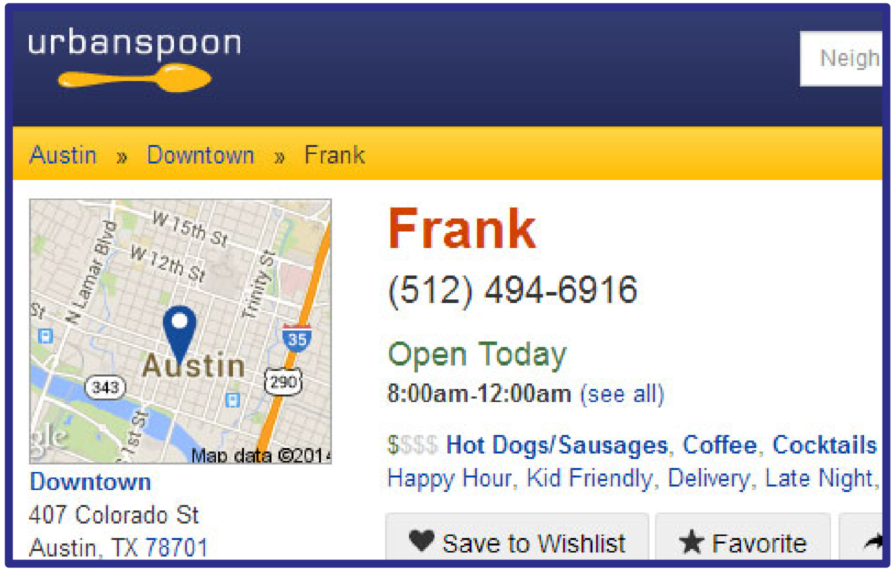
512-494-6916 • $

ADDRESS: 407 Colorado St, Austin, TX 78701

CUISINE: Hot Dogs, Breakfast, Brunch, Late-Night Dining

FEATURES: Full Bar, Handicap Access, Kid Friendly, Smoke Free, Vegetarian Friendly, Entertainment

No opening hours



urbanspoon

Neigh

Austin » Downtown » Frank

Frank

(512) 494-6916

Open Today

8:00am-12:00am (see all)

$$$$ Hot Dogs/Sausages, Coffee, Cocktails

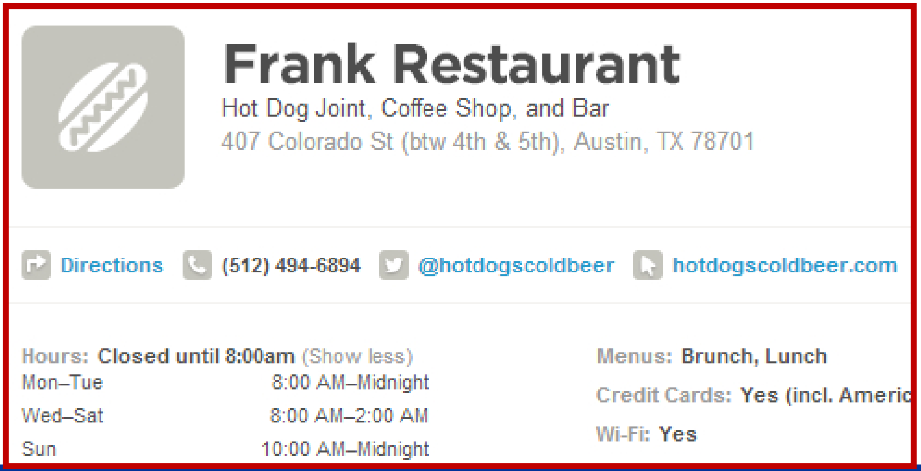
Happy Hour, Kid Friendly, Delivery, Late Night,

Downtown

407 Colorado St

Austin, TX 78701

♥ Save to Wishlist   ★ Favorite



Frank Restaurant

Hot Dog Joint, Coffee Shop, and Bar

407 Colorado St (btw 4th & 5th), Austin, TX 78701

Directions   ☎ (512) 494-6894   @hotdogscoldbeer   hotdogscoldbeer.com

Hours: Closed until 8:00am (Show less)

Mon–Tue	8:00 AM–Midnight
Wed–Sat	8:00 AM–2:00 AM
Sun	10:00 AM–Midnight

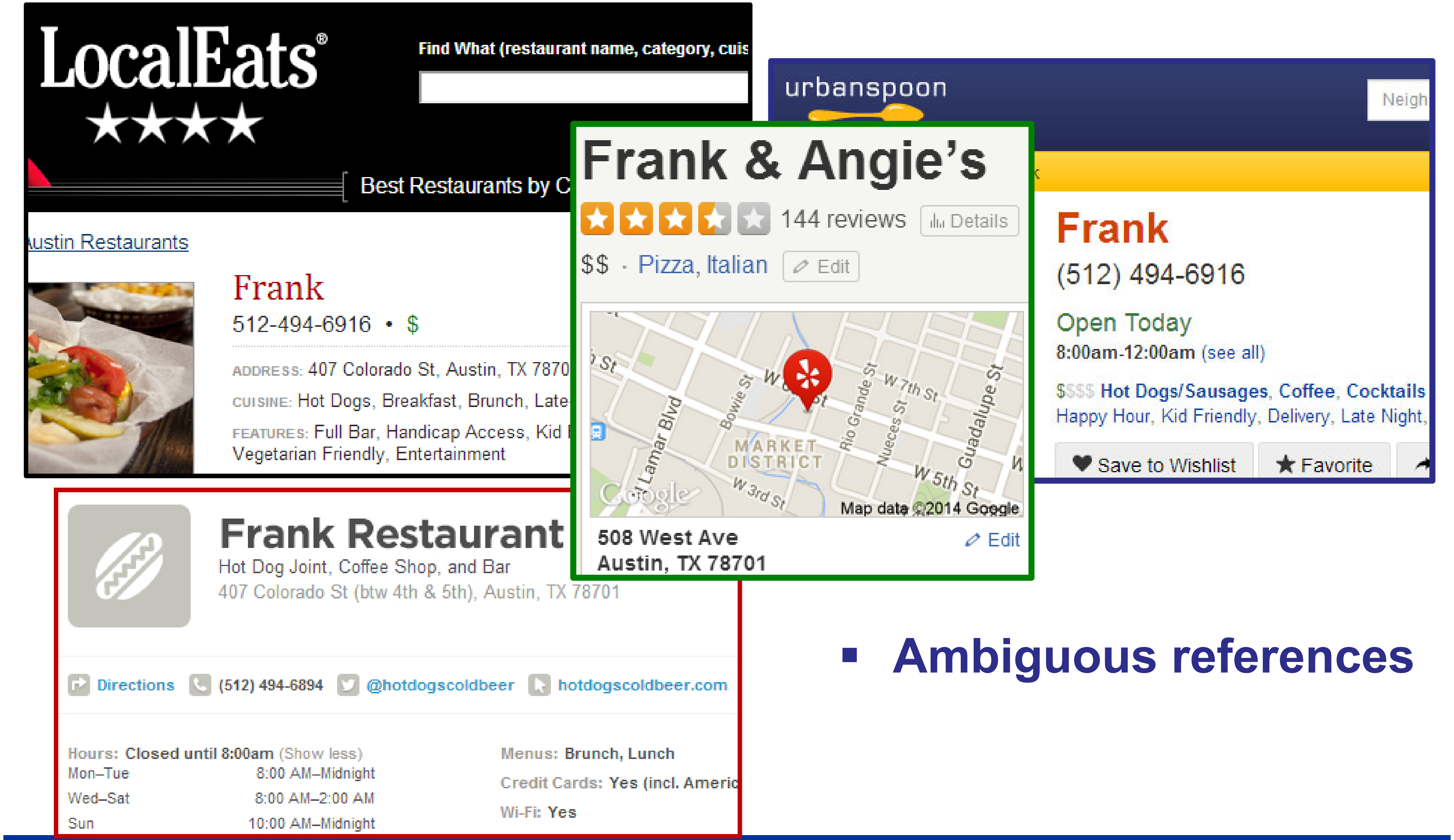
Menus: Brunch, Lunch

Credit Cards: Yes (incl. Americ

Wi-Fi: Yes

- **Incomplete information**

Data from Different Sources



LocalEats®

★★★★

Find What (restaurant name, category, cuis

Best Restaurants by C

ustin Restaurants

Frank

512-494-6916 • $

ADDRESS: 407 Colorado St, Austin, TX 7870

CUISINE: Hot Dogs, Breakfast, Brunch, Late

FEATURES: Full Bar, Handicap Access, Kid F
Vegetarian Friendly, Entertainment

urbanspoon

Neigh

Frank

(512) 494-6916

Open Today

8:00am-12:00am (see all)

$$$$ Hot Dogs/Sausages, Coffee, Cocktails
Happy Hour, Kid Friendly, Delivery, Late Night,

♥ Save to Wishlist ★ Favorite

Frank & Angie's

★★★★☆ 144 reviews Details

$$ · Pizza, Italian Edit

508 West Ave
Austin, TX 78701

Edit

Frank Restaurant

Hot Dog Joint, Coffee Shop, and Bar

407 Colorado St (btw 4th & 5th), Austin, TX 78701

Directions ☎ (512) 494-6894 @hotdogscoldbeer hotdogscoldbeer.com

Hours: Closed until 8:00am (Show less)

Mon–Tue	8:00 AM–Midnight
Wed–Sat	8:00 AM–2:00 AM
Sun	10:00 AM–Midnight

Menus: Brunch, Lunch

Credit Cards: Yes (incl. Americ

Wi-Fi: Yes

- **Ambiguous references**

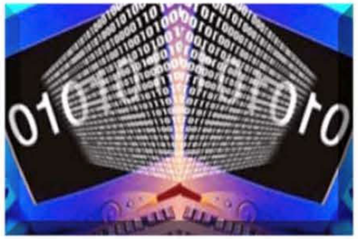
Application of Big Data Analytics

Smarter Healthcare

Multi-channel

Finance

Log Analysis



Homeland Security

Traffic Control

Telecom

Search Quality

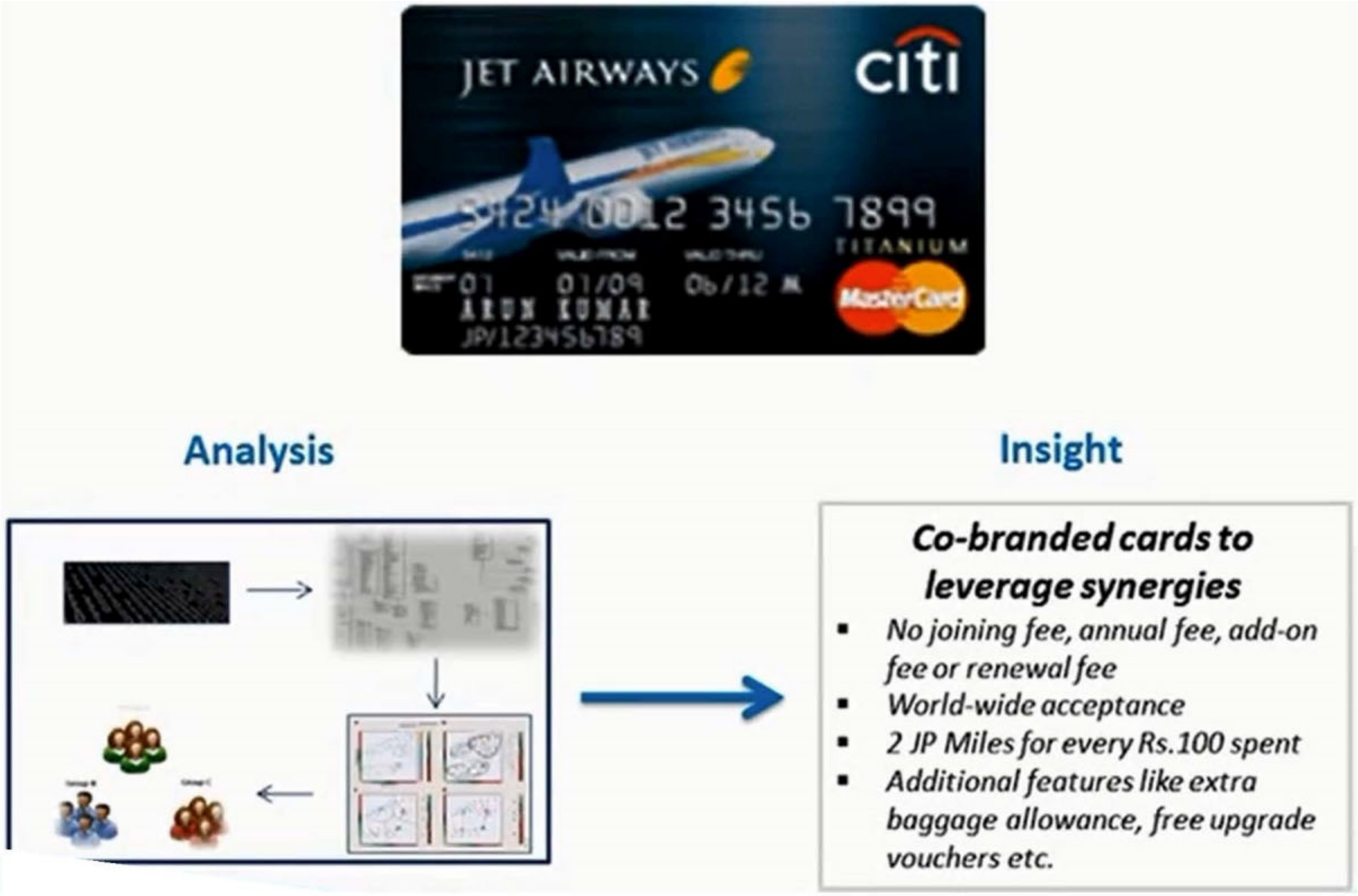
Manufacturing

Trading Analytics

Fraud and Risk

Retail: Churn, NBO

Acquiring Better Customers



Analysis

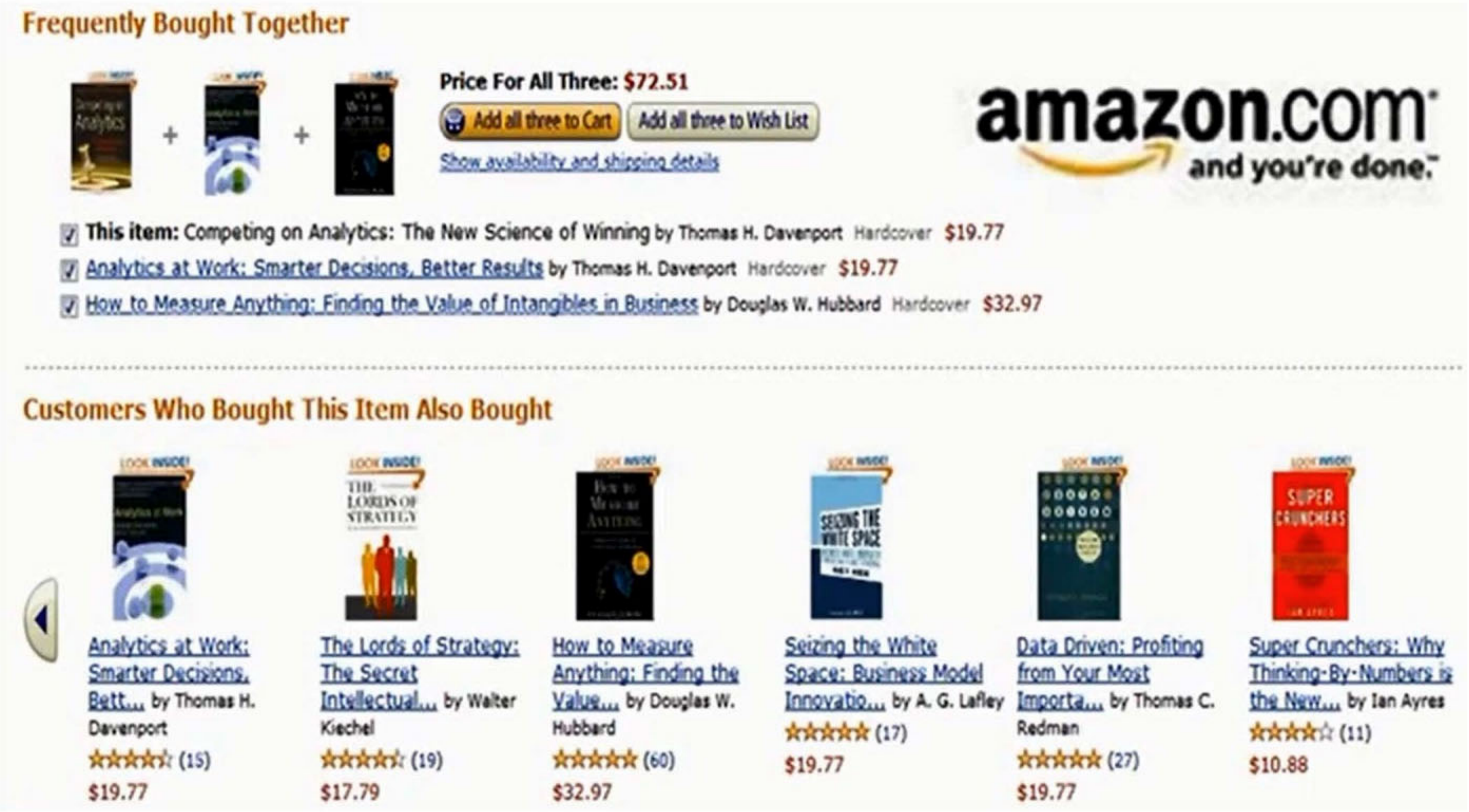
Insight

Co-branded cards to leverage synergies

- No joining fee, annual fee, add-on fee or renewal fee
- World-wide acceptance
- 2 JP Miles for every Rs.100 spent
- Additional features like extra baggage allowance, free upgrade vouchers etc.

Source: https://www.youtube.com/watch?v=BfoJgoItd4M

Improving Customer Experience



Frequently Bought Together

Price For All Three: $72.51

Show availability and shipping details

☑ **This item:** Competing on Analytics: The New Science of Winning by Thomas H. Davenport Hardcover $19.77

☑ Analytics at Work: Smarter Decisions, Better Results by Thomas H. Davenport Hardcover $19.77

☑ How to Measure Anything: Finding the Value of Intangibles in Business by Douglas W. Hubbard Hardcover $32.97

Customers Who Bought This Item Also Bought

Analytics at Work: Smarter Decisions, Bett... by Thomas H. Davenport ★★★★½ (15) $19.77

The Lords of Strategy: The Secret Intellectual... by Walter Kiechel ★★★★½ (19) $17.79

How to Measure Anything: Finding the Value... by Douglas W. Hubbard ★★★★★ (60) $32.97

Seizing the White Space: Business Model Innovatio... by A. G. Lafley ★★★★★ (17) $19.77

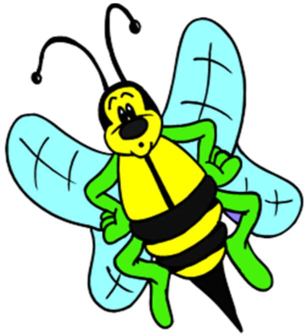
Data Driven: Profiting from Your Most Importa... by Thomas C. Redman ★★★★★ (27) $19.77

Super Crunchers: Why Thinking-By-Numbers is the New... by Ian Ayres ★★★★☆ (11) $10.88

Source: https://www.youtube.com/watch?v=BfoJgoItd4M

Summary

- Lots of Buzz



- With good reason
  - Great potential
  - Many challenges

Lab 1 (5%)

- **Get started with Spark**
- **Compile and execute a simple Spark program**
  - WordCount
- **Write your own Spark program**
  - Count the number of words that begin with each letter
- **Due: Tuesday, 29 January**
- **Submit to IVLE Lab 1 Folder by 6 PM**

Apache Spark

- **Big Data is diverse and messy.**
- **Typical pipeline**
  - MapReduce for data loading and batch processing
  - Exploratory SQL-like queries
  - Iterative machine learning
- **Specialized engines create complexity and inefficiency**
  - Users must stitch together disparate systems
- **Spark is a unified engine for distributed data processing**
  - Programming model similar to MapReduce
  - Data-sharing abstraction called **Resilient Distributed Databases (RDDs)** to capture range of processing workloads that previously need separate engines (SQL, machine learning, graph processing)