Introduction to Data

DSLA COURSE

ROHIT PADEBETTU

Course Assignments

Programming Assignments

Reading Assignments

Presentation Assignments

Technical Skills Assignments

Writing Assignments

Twitter Case

Twitter User Analysis

Build a set of tools to later (next week) use to build a Shiny App

Tools:

1. Build an LDA model using tweets downloaded from a few distinct topics
2. Given a twitter user (fairly famous users)
    1. Based on his tweets, guess the topics he is interested in and compare it to his description
    2. Do a sentiment analysis of the timeline of his/her tweets and present it
    3. Display a word cloud or bag of words of what the twitter user uses
    4. Find and compare his top 5 best followers
3. Given a tweet, estimate the top 3 topics it might be classified as and show it graphically

Technical Assignment

Teach yourself Shiny

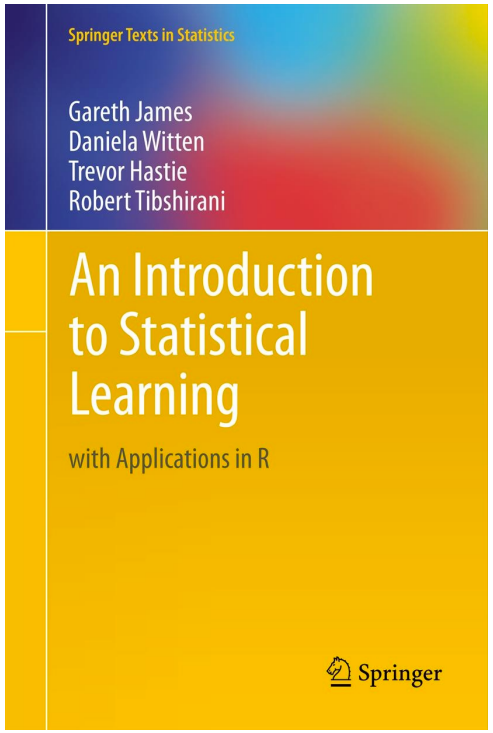
https://shiny.rstudio.com/tutorial/

Programming Assignment

Code Submission on Github for Twitter Case

Reading Assignment

Read Chapter 6: Linear Model Selection & Regularization



Writing Assignment

Submit by Saturday
Written Report (not to exceed 15 pages) on Twitter Case

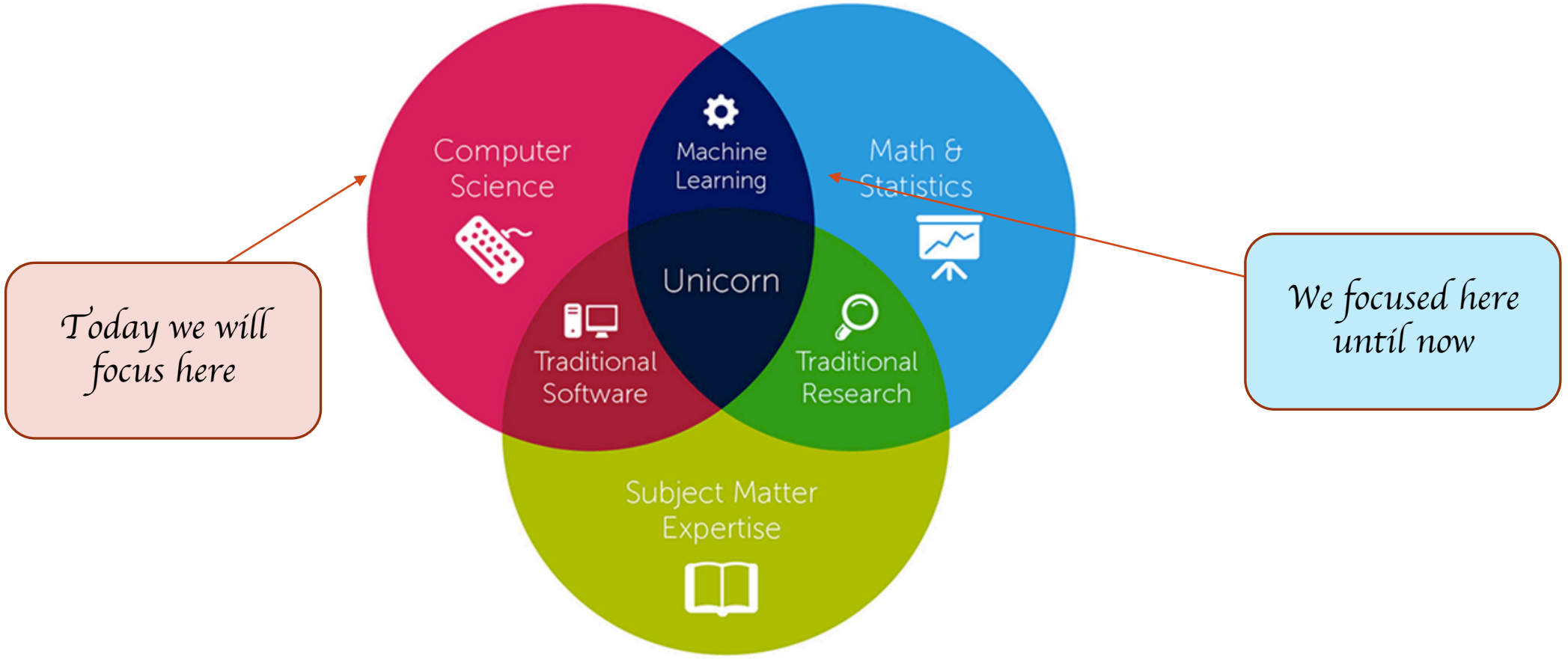
Presentation Assignment

By Saturday Submit

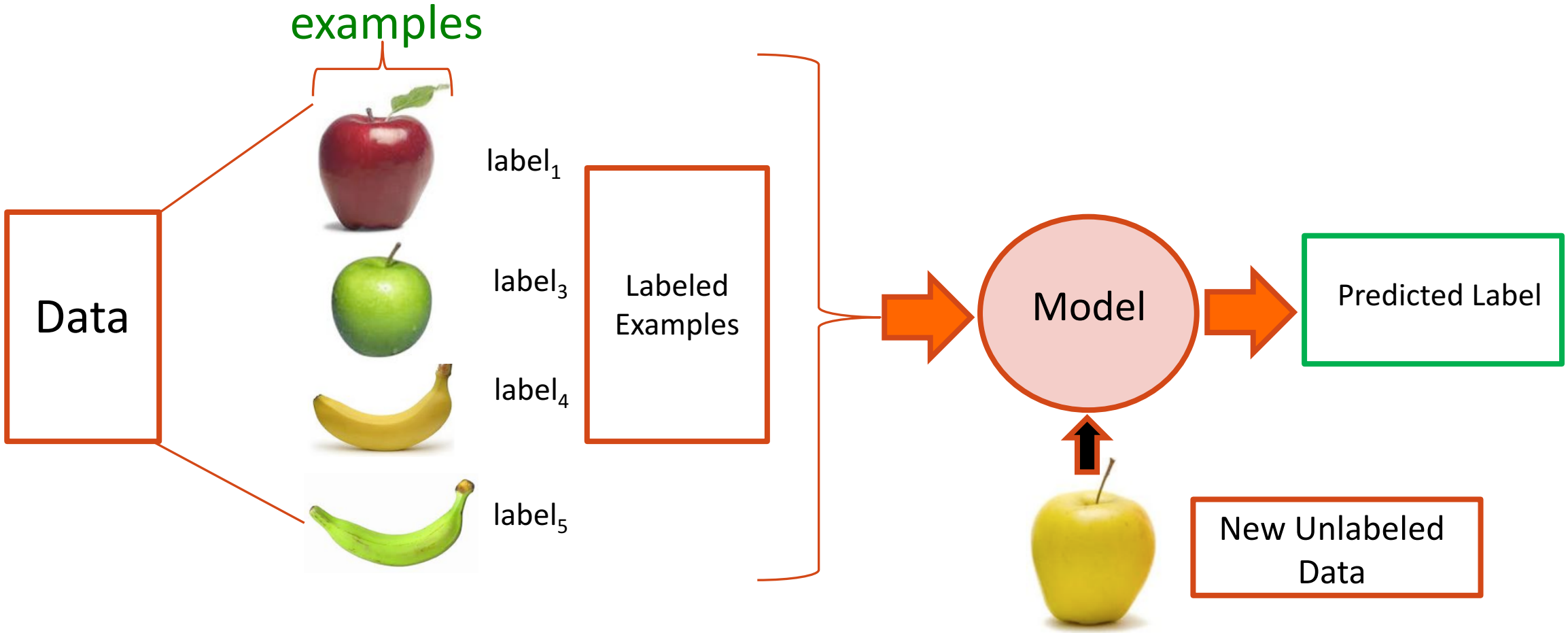
Your Presentations on Twitter Case

1. Technical Presentation
2. Business Presentation (Not to exceed 5 slides)

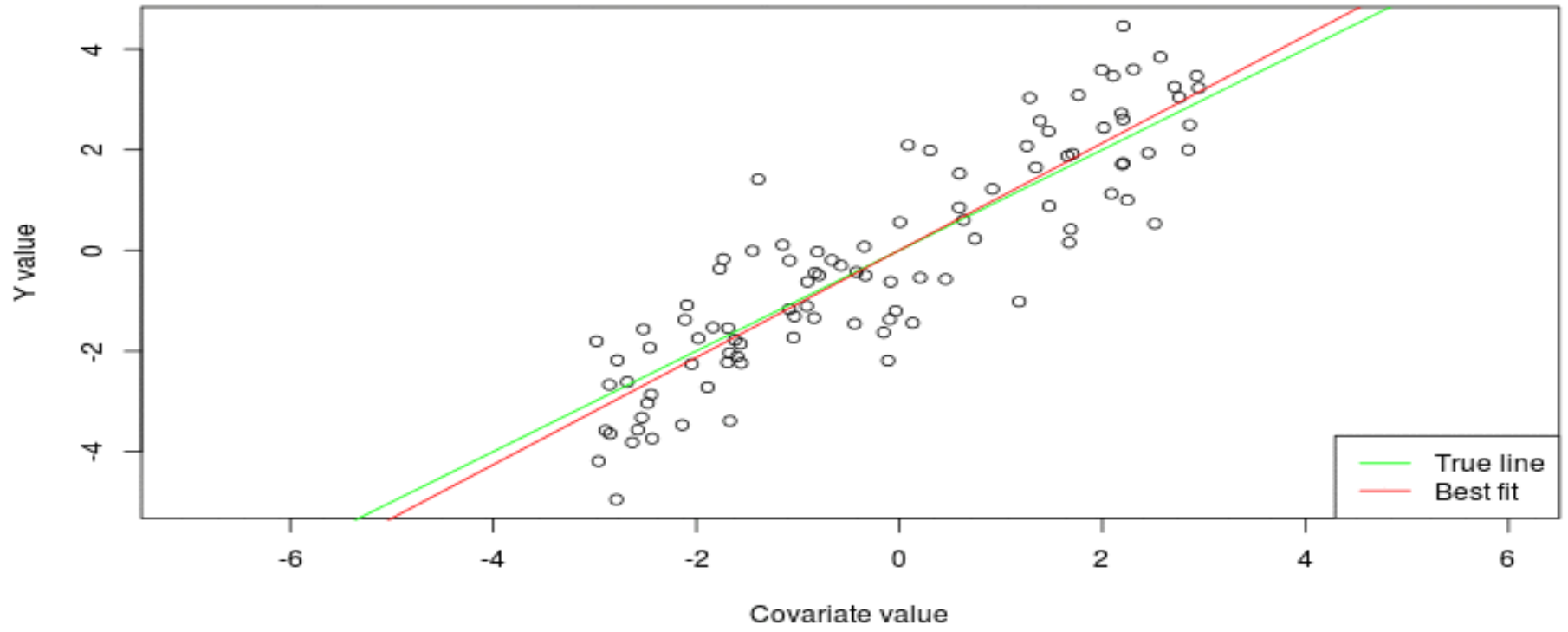
Who is a Data Scientist?



Classification



Linear Regression



Select Airports

Flight origin

JFK

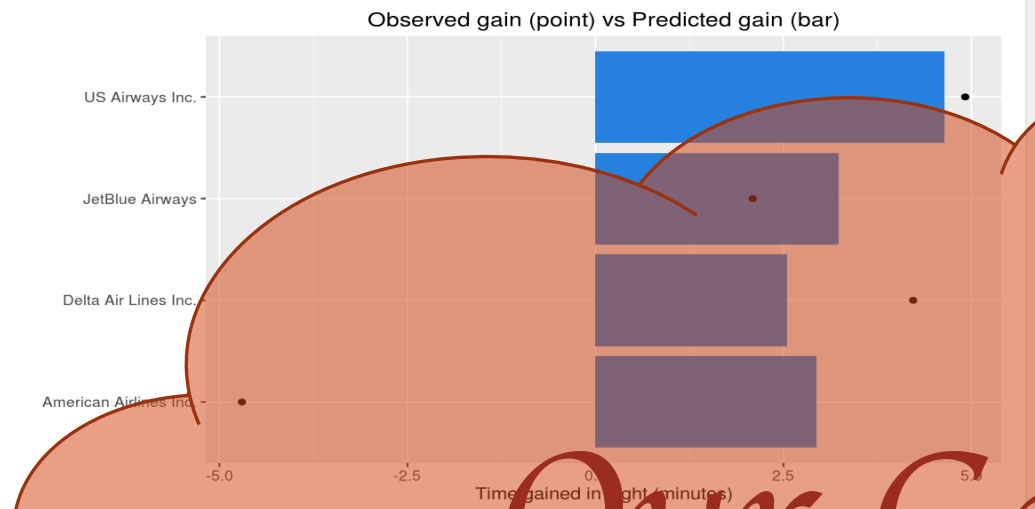
Flight destination

LAS

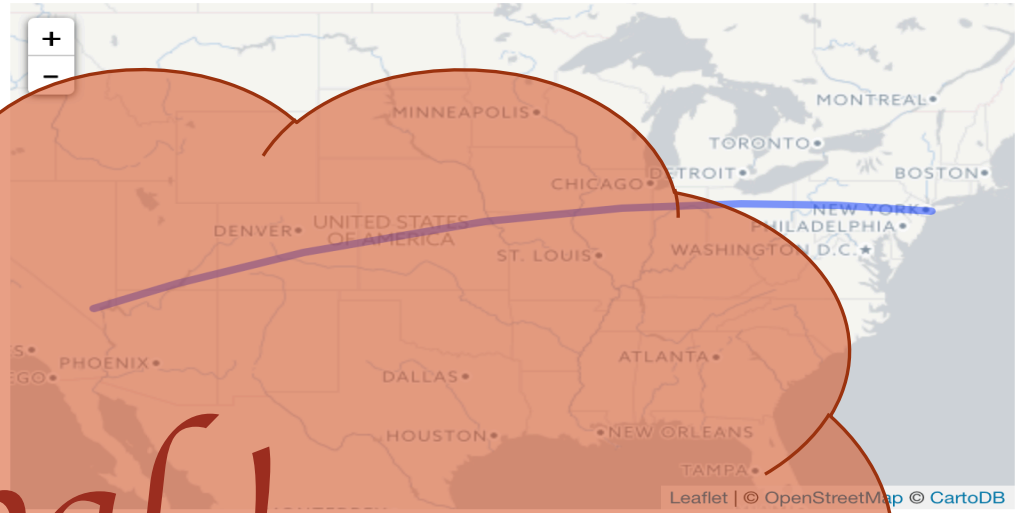
Background

Given that your flight was delayed by 15 minutes or more, what is the likelihood your airline carrier will make up time in route? Some of the most signficant factors for making up time are flight distance and airline carrier. The data model behind this dashboard is based on flights from NYC airports in 2013.

Observed versus predicted time gain



Route



Data details

Show 10 entries

Search:

	airline	flights	distance	avg_dep_delay	avg_arr_delay	avg_gain	pred_gain
1	JetBlue Airways	286	2248	61.7	59.6	2.1	3.2
2	American Airlines Inc.	97	2248	57.2	61.9	-4.7	2.9
3	Delta Air Lines Inc.	227	2248	58.7	54.4	4.2	2.5
4	US Airways Inc.	192	2248	57.4	52.5	4.9	4.6

Showing 1 to 4 of 4 entries

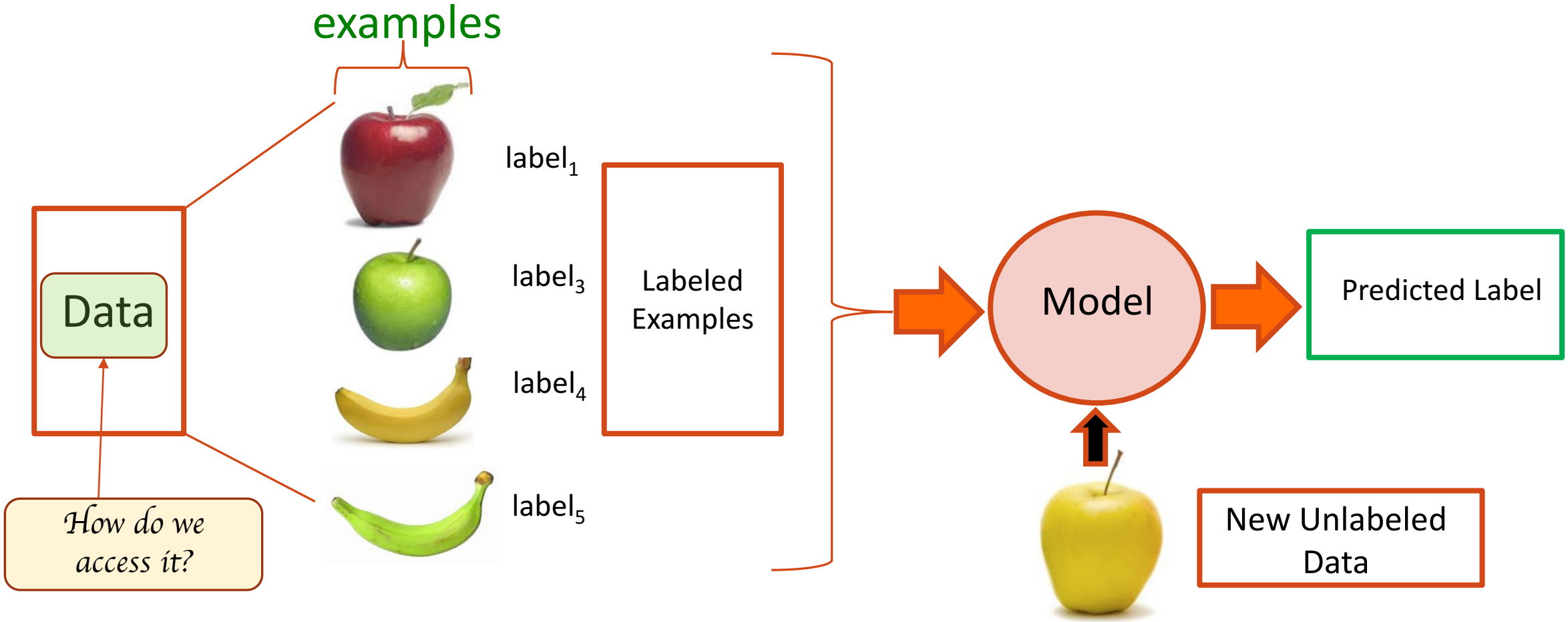
Previous 1 Next

Our Goal !

Demo – to Recap

NYC 2013 Flights Case

Classification

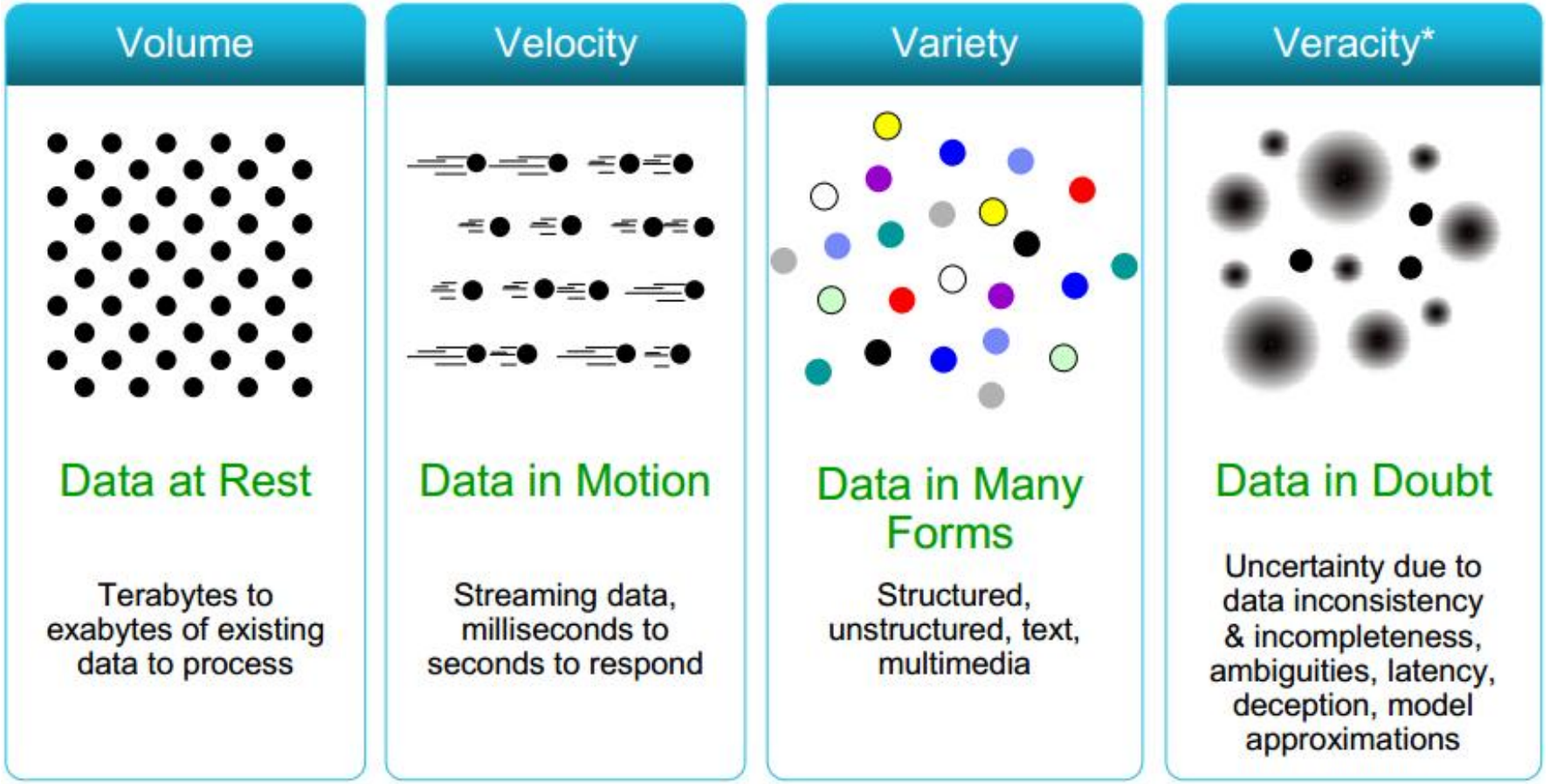


Data in all Shapes & Sizes

Data in all Shapes & Sizes

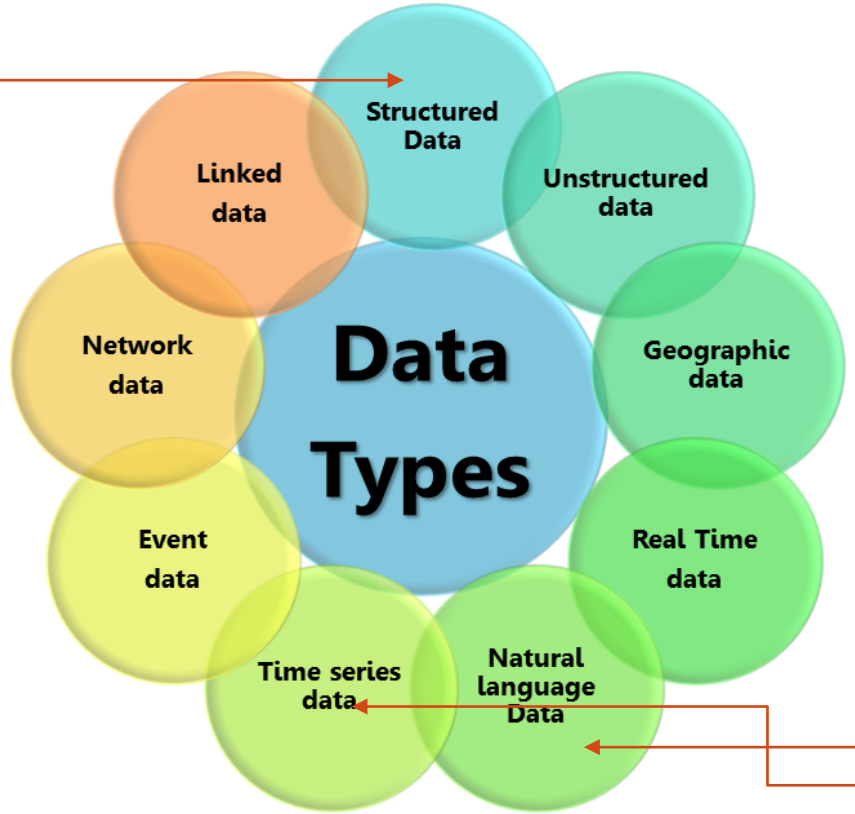
Leads to

Characteristics of Data



Types of Data

Let us begin by focusing here



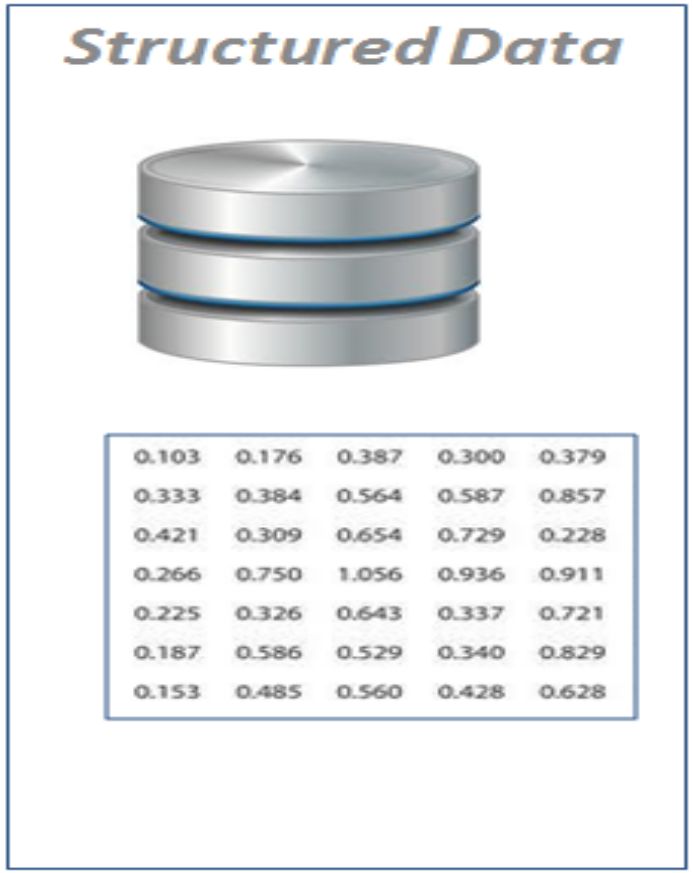
Did we see these two?

Structured vs Unstructured Data

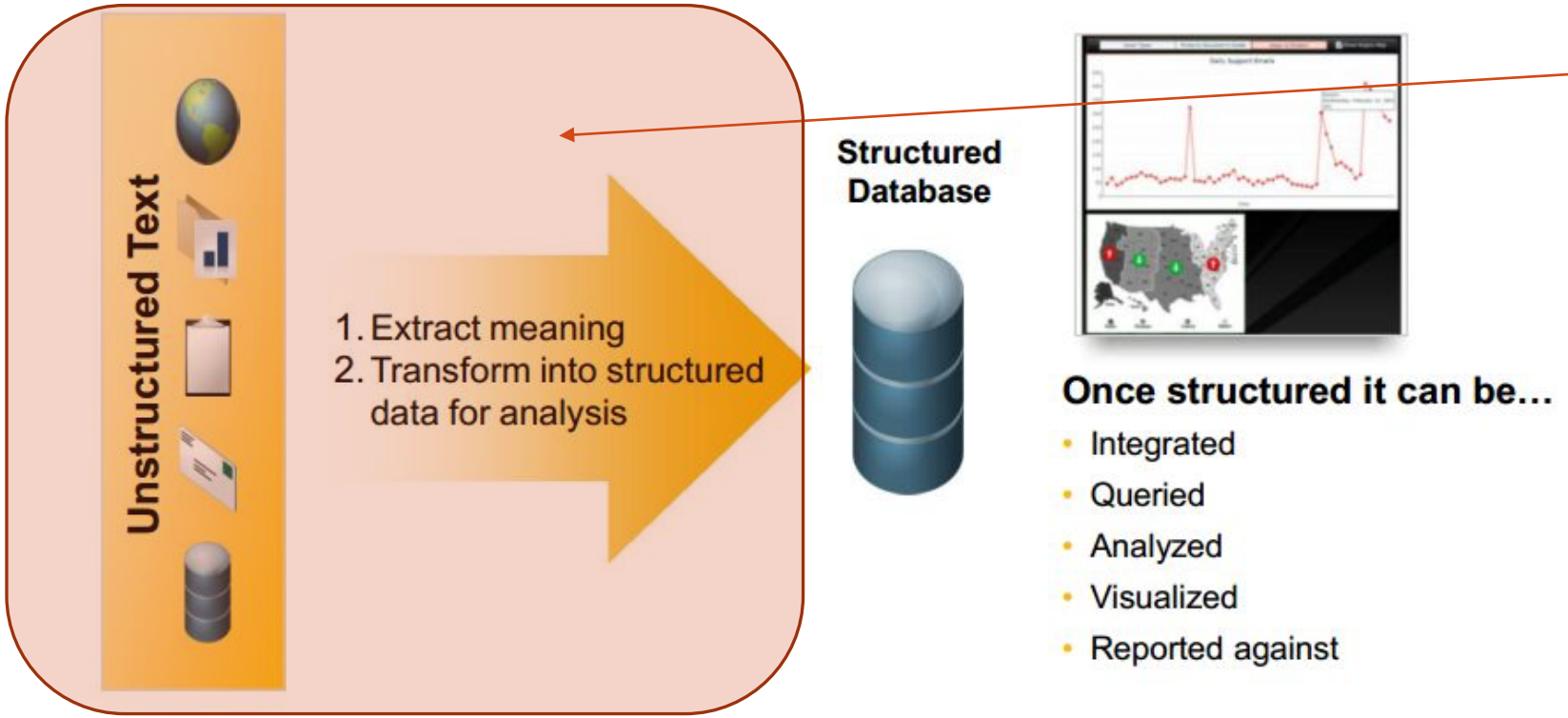
For the most part, structured data refers to information with a high degree of organization

Unstructured data refers to data which for the most part does not have a pre-defined data model

Structured vs Unstructured Data



Unstructured -> Structured Data



Once structured it can be…

- Integrated
- Queried
- Analyzed
- Visualized
- Reported against

You already did this for Twitter case, right?

Structured Data

Structured Data at a Glance

Characteristics of Structured Data

- High organized
- Clearly defined
- Easy to access
- Easy to analyze

Examples of Structured Data

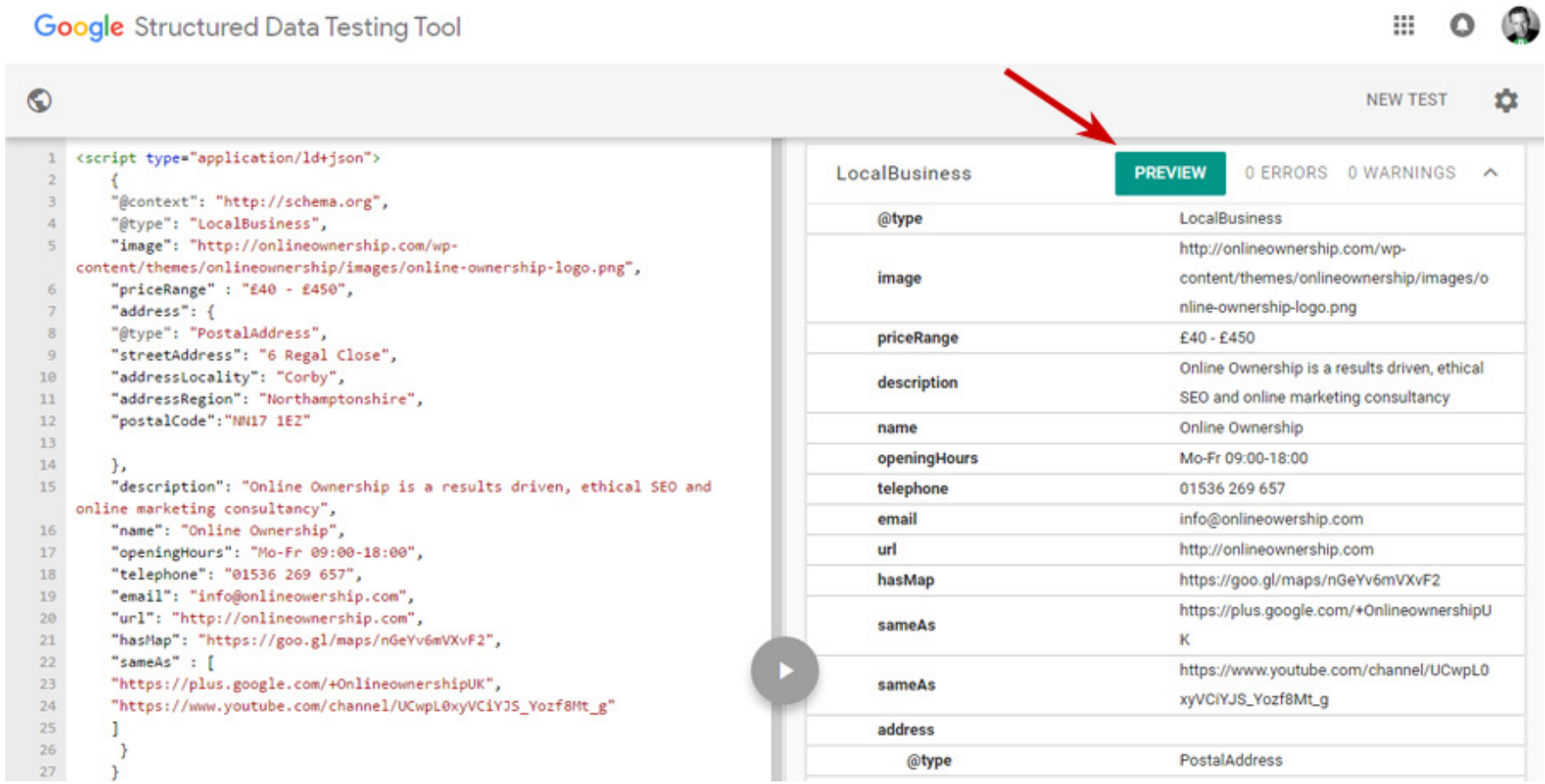
- Name
- Age
- Gender
- Address
- Phone number
- Currency
- Dates
- Billing info

Sources of Structured Data

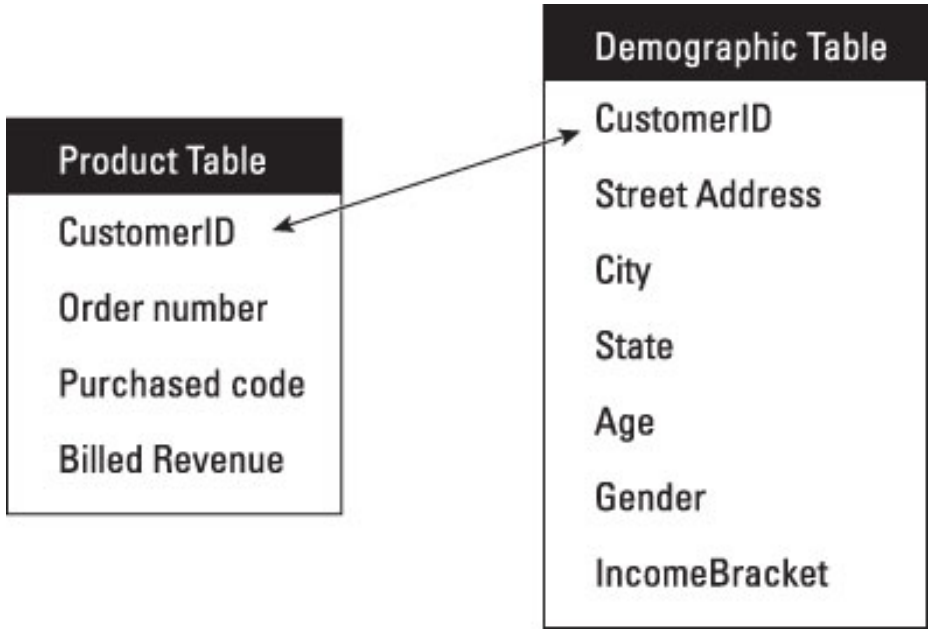
- SQL databases
- Spreadsheets
- Sensors
- Medical Devices
- Online Forms
- Point of Sales Systems
- Web and Server Logs

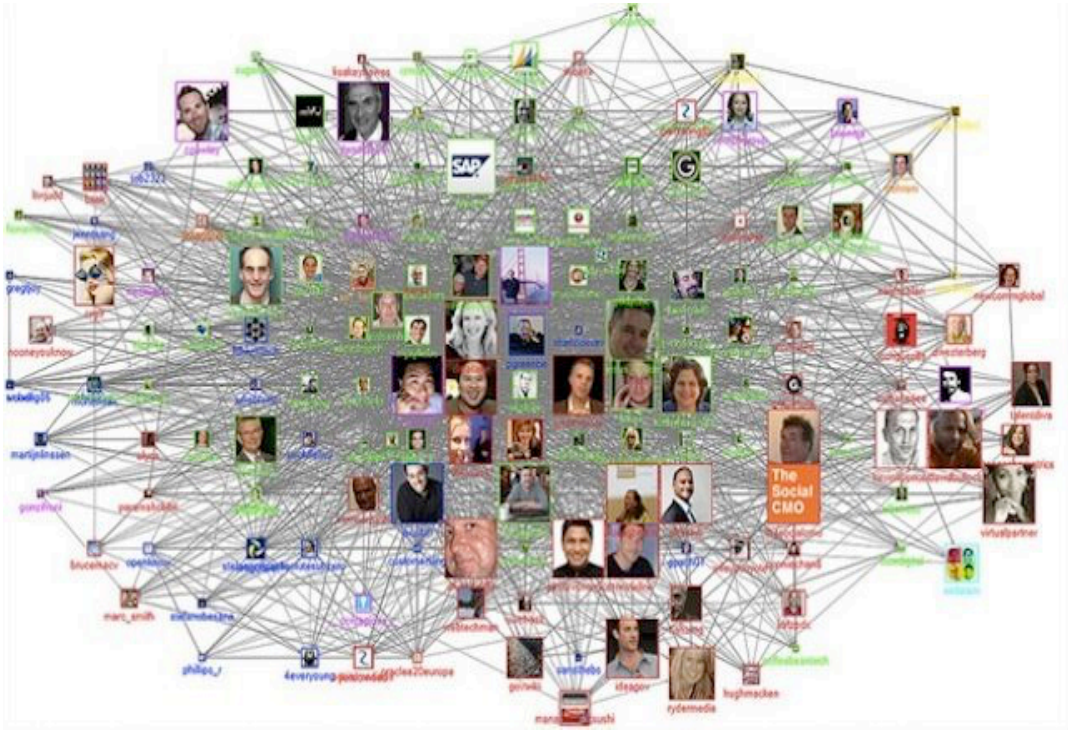
First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotitaw	28

Structured Data



Relational Data





Stored in Traditional Relational Databases

Stored in Newer Graph Databases

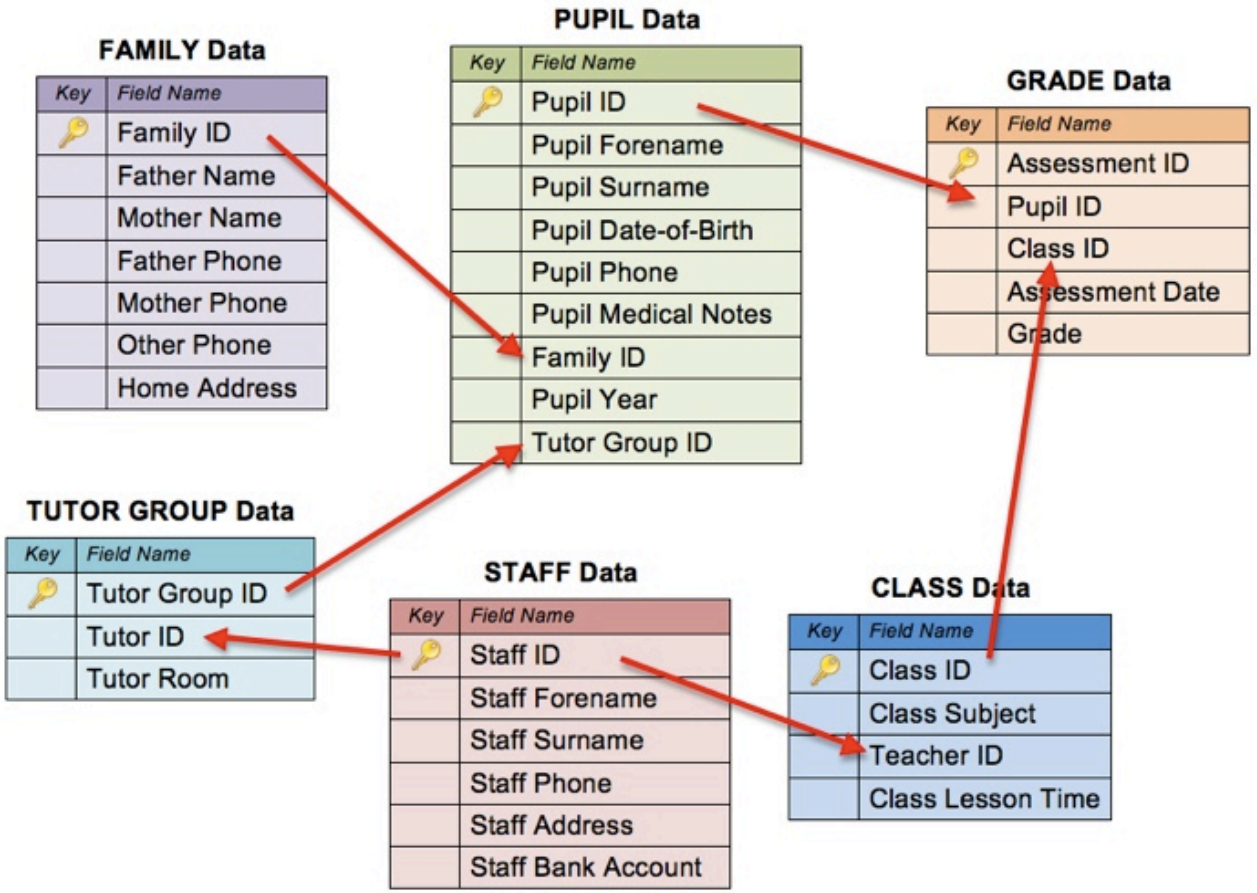
Relational Database (RDBMS)

Vendors

Before

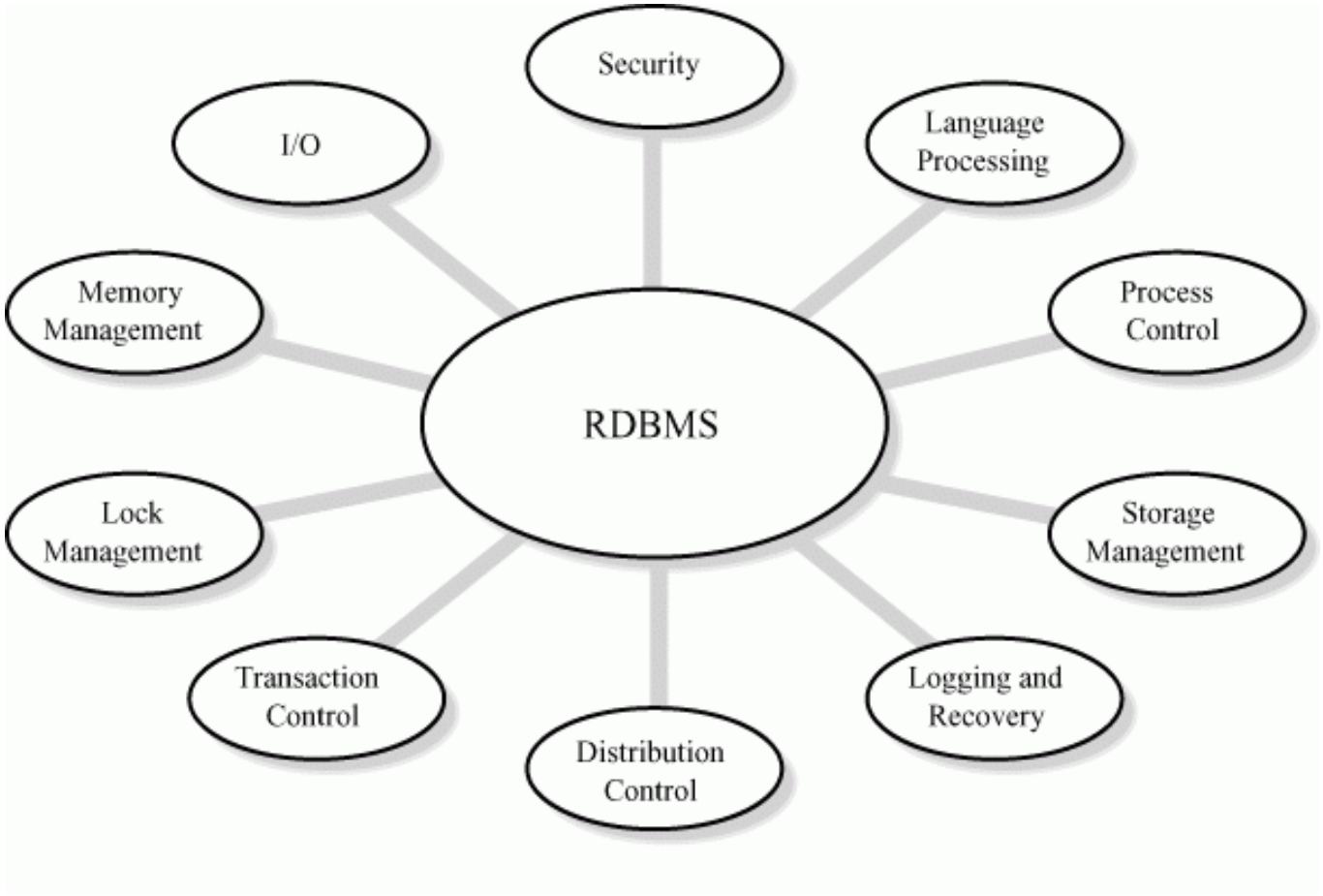
After

Relational Database (RDBMS)



Data Model

Relational Database (RDBMS)



A - Atomicity

All or Nothing Transactions

C - Consistency

Guarantees Committed Transaction State

I - Isolation

Transactions are Independent

D – Durability

Committed Data is Never Lost

(c) http://blog.sqlauthority.com

Relational Databases in Cloud

Amazon Relational Database Service

*RDS is a **managed Relational database service** that is simple to deploy, easy to scale, reliable and cost-effective*

- Choice of Database Engines
- Managed Service
- Easy to Scale and Operate
- High Performance
- High Availability

Data access – How ?

Structured Query Language

SQL

"ess-que-ell" *OR* *"sequel"*

Start Page | Schema Editor | SQL introduction.sql

Execute | Cancel | Functions... | Client/NoLock

Schema | Queries | Snippets

spam

- Notification channels (0)
- Schemas (12)
  - information_schema
  - pg_catalog
  - pg_temp_1
  - pg_temp_2
  - pg_temp_3
  - pg_temp_4
  - pg_toast
  - pg_toast_temp_1
  - pg_toast_temp_2
  - pg_toast_temp_3
  - pg_toast_temp_4
  - public
    - Domains (0)
    - Functions (0)
    - Links (0)
    - Sequences (0)
    - Tables (15)
      - airlines
      - airports
      - bigdata_need
      - bigdata_need_query2
      - cartable
      - data
      - flights_2008
      - flights_2014
      - flights_jan_2015
      - idahohousing
      - iris
      - nyc_flights_2013
      - order_details
      - stock_history

```sql
SELECT table_name
FROM information_schema.tables
WHERE table_schema = 'public'

SELECT *
FROM information_schema.columns
WHERE table_name = 'flights_2014'

SELECT count(*) FROM flights_2014;

SELECT *
FROM flights_2014
WHERE day_of_week = 3
ORDER BY distance DESC
LIMIT 10;


SELECT day_of_week,avg(dep_delay) AS avg_dep_delay,
                   max(arr_delay) AS max_arr_delay,
                   avg(distance) AS avg_distance,
                   avg(air_time) AS avg_air_time
FROM flights_2014
WHERE day_of_week BETWEEN 1 AND 7
GROUP BY day_of_week
ORDER BY day_of_week
LIMIT 10;
```

	day_of_week	avg_dep_delay	max_arr_delay	avg_distance	avg_air_time
1	1	11.3637014673212215	1793	793.2987378055318869	110.9867766031015242
2	2	9.9987494257567250	1621	786.8784815217905645	110.3293813892337676
3	3	10.5848063873202676	1940	790.0396894563054883	110.5705763207250778
4	4	12.2345236185144908	2444	794.3418916225561540	110.9513488883614021
5	5	11.8200642302935955	2017	795.1303428331188768	110.7921680001990524
6	6	8.5440782957724453	1697	831.0373802081544327	114.5324871821306341
7	7	9.4967667809410245	1707	806.5960979903183218	112.0352708701659696

Number of records: 7 Number of fields: 5 Query time: 2.762 second(s)

select day_of_week,avg(dep_delay) as avg_dep_delay, max(arr_delay) as max_arr_delay,...here day_of_week between 1 and 7 group by day_of_week order by day_of_week limit 10;

All (12) | Logs (14) | Tunes (13) | Warnings

```
Rohits-MacPro:Downloads rohitpittu$ ssh rohit@rohit-lubuntu
rohit@rohit-lubuntu's password:
Welcome to Ubuntu 16.04.2 LTS (GNU/Linux 4.4.0-87-generic i686)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

31 packages can be updated.
4 updates are security updates.

Last login: Fri Jul 28 10:26:39 2017 from 192.168.1.5
rohit@rohit-lubuntu:~$ psql spam
psql (9.5.7)
Type "help" for help.

spam=# select day_of_week,avg(dep_delay) as avg_dep_delay,
spam-#                    max(arr_delay) as max_arr_delay,
spam-#                    avg(distance) as avg_distance,
spam-#                    avg(air_time) as avg_air_time
spam-# from flights_2014
spam-# where day_of_week between 1 and 7
spam-# group by day_of_week
spam-# order by day_of_week
spam-# limit 10;
 day_of_week |    avg_dep_delay    | max_arr_delay |     avg_distance     |     avg_air_time
-------------+---------------------+---------------+----------------------+----------------------
           1 | 11.3637014673212215 |          1793 | 793.2987378055318869 | 110.9867766031015242
           2 |  9.9987494257567250 |          1621 | 786.8784815217905645 | 110.3293813892337676
           3 | 10.5848063873202676 |          1940 | 790.0396894563054883 | 110.5705763207250778
           4 | 12.2345236185144908 |          2444 | 794.3418916225561540 | 110.9513488883614021
           5 | 11.8200642302935955 |          2017 | 795.1303428331188768 | 110.7921680001990524
           6 |  8.5440782957724453 |          1697 | 831.0373802081544327 | 114.5324871821306341
           7 |  9.4967667809410245 |          1707 | 806.5960979903183218 | 112.0352708701659696
(7 rows)

spam=#
```

```
Rohits-MacPro:Downloads rohitpittu$ ssh rohit@rohit-lubuntu
rohit@rohit-lubuntu's password:
Welcome to Ubuntu 16.04.2 LTS (GNU/Linux 4.4.0-87-generic i686)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

31 packages can be updated.
4 updates are security updates.

Last login: Fri Jul 28 10:26:39 2017 from 192.168.1.5
rohit@rohit-lubuntu:~$ psql spam
psql (9.5.7)
Type "help" for help.

spam=# select day_of_week,avg(dep_delay) as avg_dep_delay,
spam-#                    max(arr_delay) as max_arr_delay,
spam-#                    avg(distance) as avg_distance,
spam-#                    avg(air_time) as avg_air_time
spam-# from flights_2014
spam-# where day_of_week between 1 and 7
spam-# group by day_of_week
spam-# order by day_of_week
spam-# limit 10;
 day_of_week |    avg_dep_delay    | max_arr_delay |     avg_distance      |     avg_air_time
-------------+---------------------+---------------+-----------------------+----------------------
           1 | 11.3637014673212215 |          1793 | 793.2987378055318869 | 110.9867766031015242
           2 |  9.9987494257567250 |          1621 | 786.8784815217905645 | 110.3293813892337676
           3 | 10.5848063873202676 |          1940 | 790.0396894563054883 | 110.5705763207250778
           4 | 12.2345236185144908 |          2444 | 794.3418916225561540 | 110.9513488883614021
           5 | 11.8200642302935955 |          2017 | 795.1303428331188768 | 110.7921680001990524
           6 |  8.5440782957724453 |          1697 | 831.0373802081544327 | 114.5324871821306341
           7 |  9.4967667809410245 |          1707 | 806.5960979903183218 | 112.0352708701659696
(7 rows)

spam=#
```

Demo

Flights using Relational Databases

Data

DSLA
DATA SCIENCE LEARNING ACADEMY

Break Time