



# 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

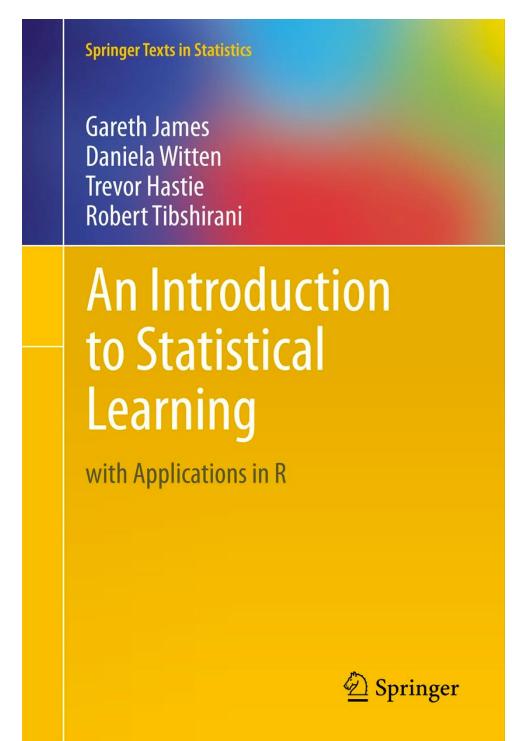
---

*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

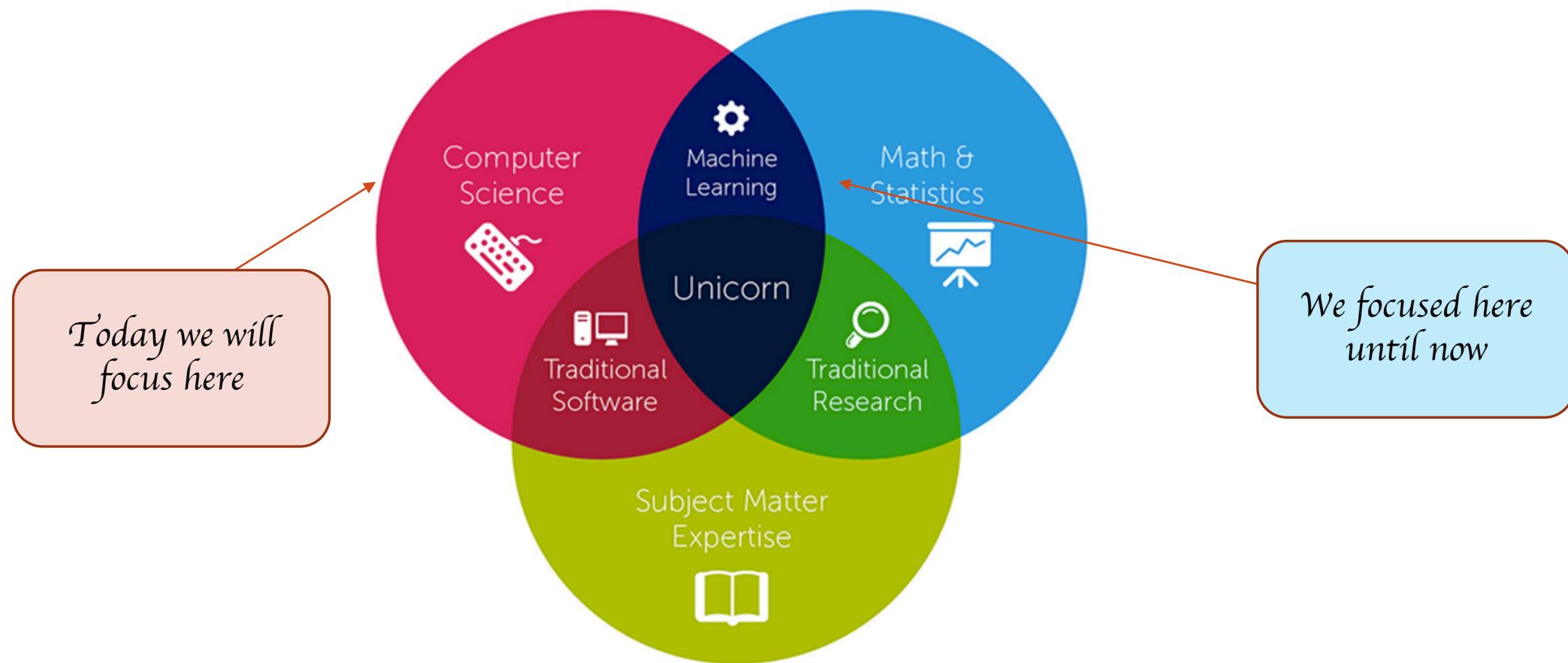
---

*By Saturday Submit*

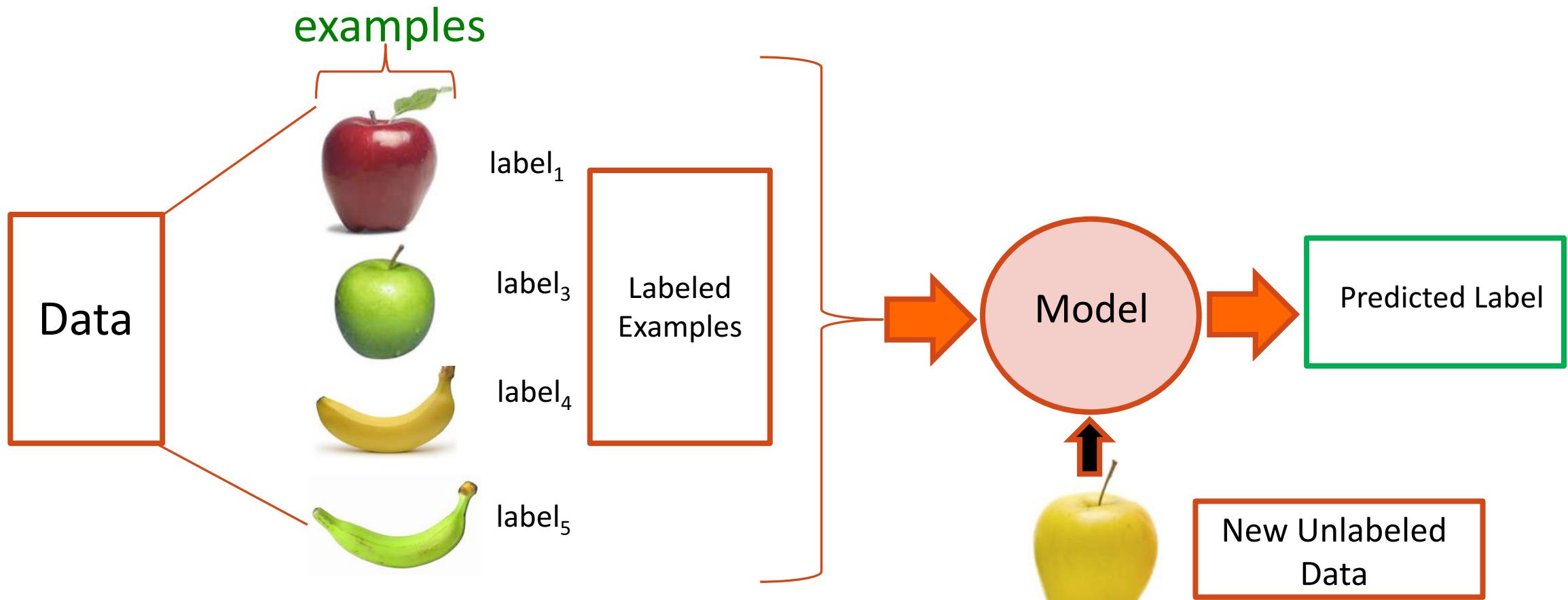
Your Presentations on Twitter Case

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

# Who is a Data Scientist?

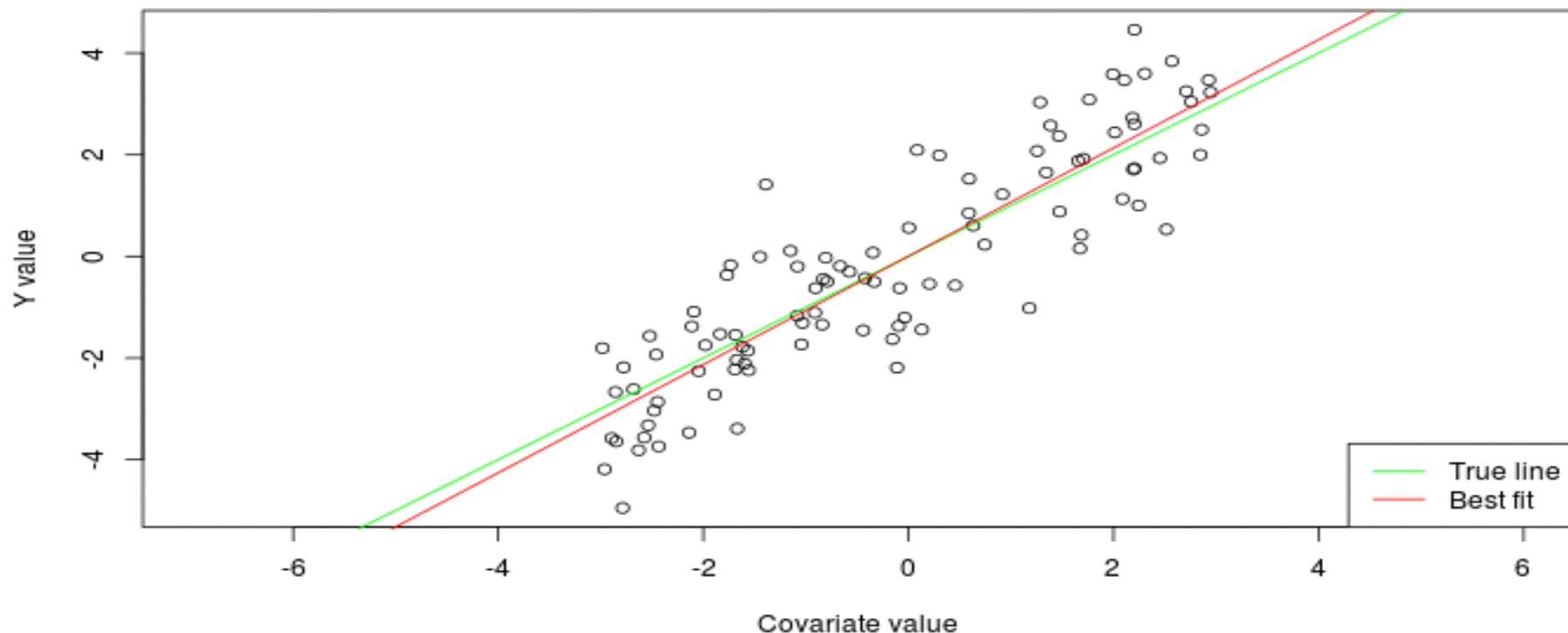


# Classification



# Linear Regression

---



## Select Airports

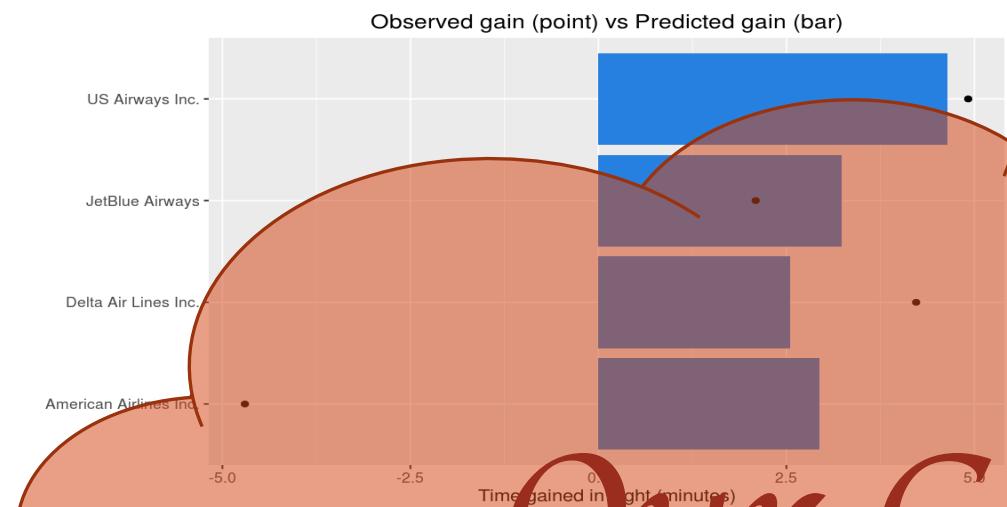
Flight origin

Flight destination

## 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 significant 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



Data details

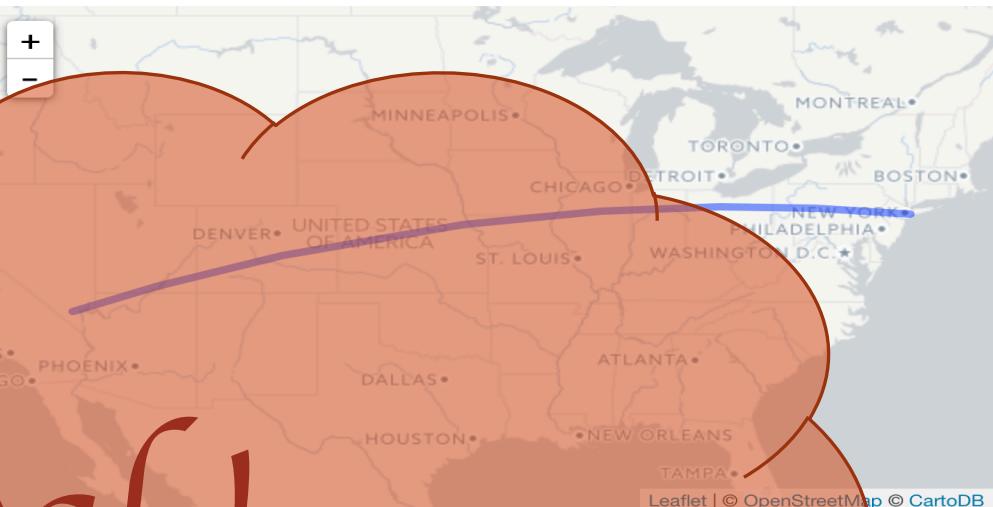
Show 10 entries

	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

Route



Search:

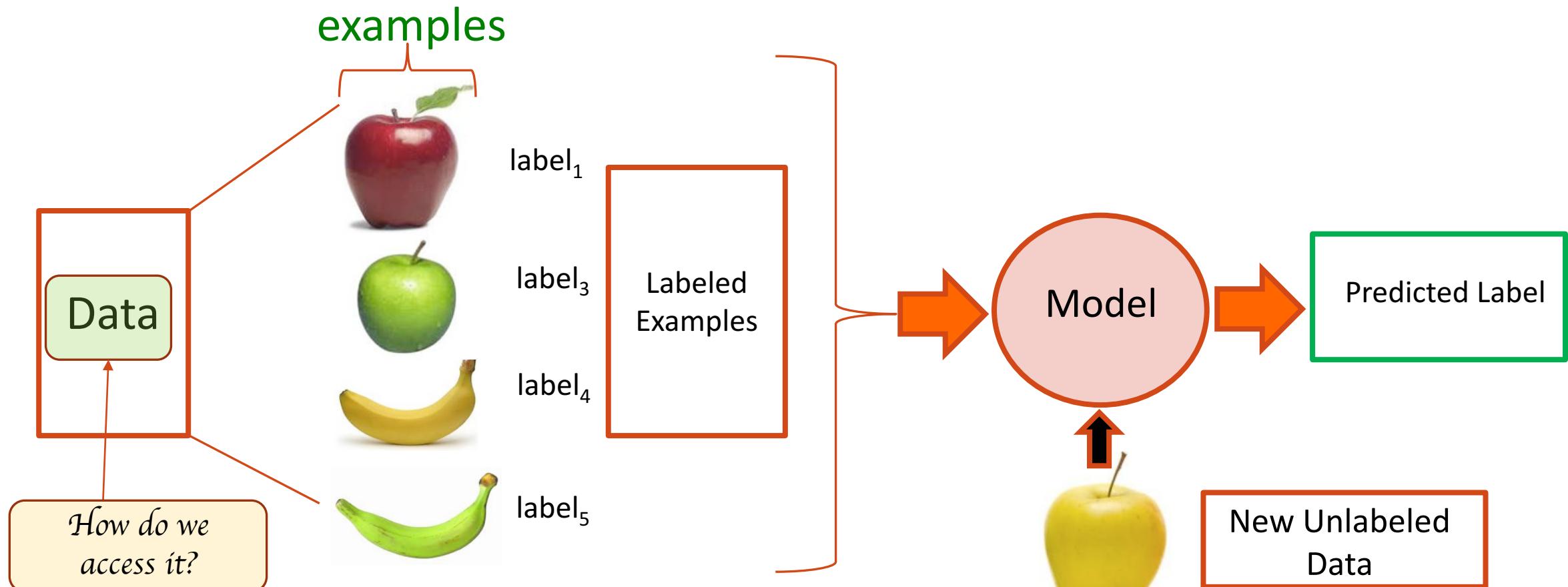
avg\_gain pred\_gain

# Demo – to Recap

---

*NYC 2013 Flights Case*

# Classification



# Data in all Shapes & Sizes

Beauty comes in all shapes & sizes. Small, large, circle, square, thin crust, thick crust, stuffed crust, extra toppings.



som~~e~~ecards  
user card

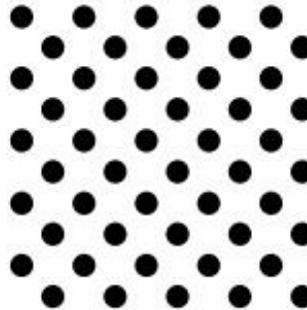
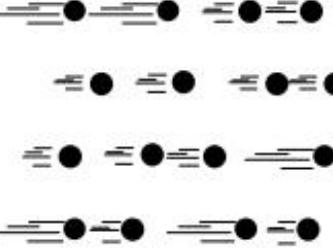
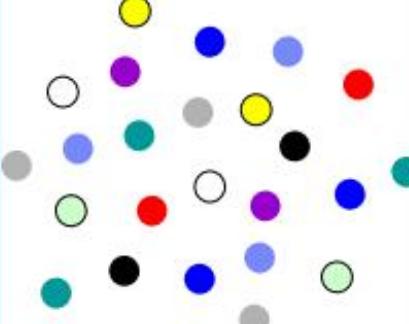
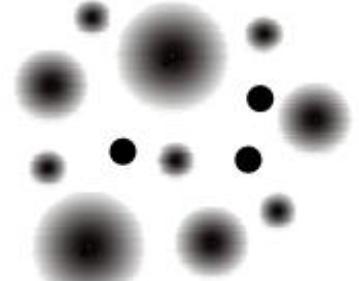
# Data in all Shapes & Sizes



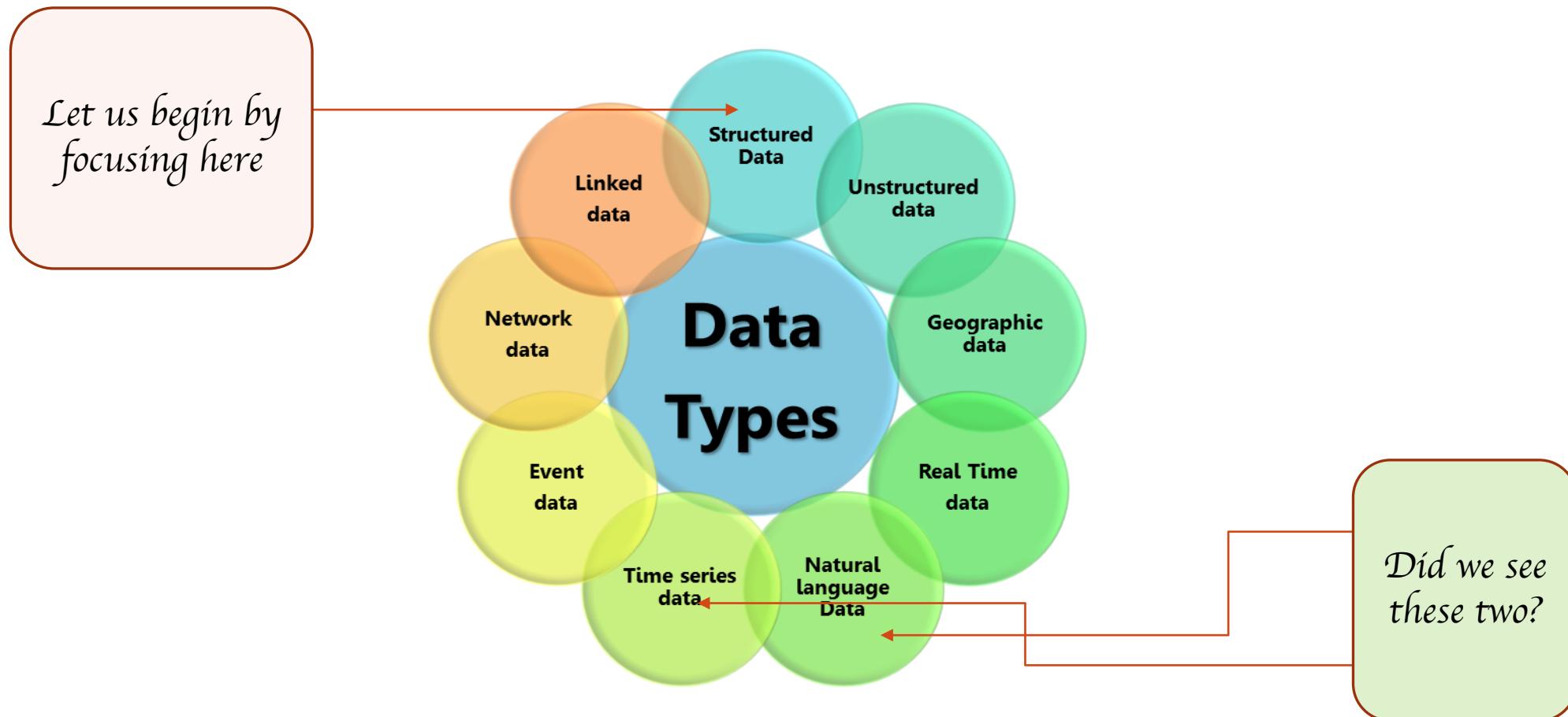
*Leads to*



# Characteristics of Data

Volume	Velocity	Variety	Veracity*
 <p><b>Data at Rest</b></p> <p>Terabytes to exabytes of existing data to process</p>	 <p><b>Data in Motion</b></p> <p>Streaming data, milliseconds to seconds to respond</p>	 <p><b>Data in Many Forms</b></p> <p>Structured, unstructured, text, multimedia</p>	 <p><b>Data in Doubt</b></p> <p>Uncertainty due to data inconsistency &amp; incompleteness, ambiguities, latency, deception, model approximations</p>

# Types of Data



# 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

## *Structured Data*

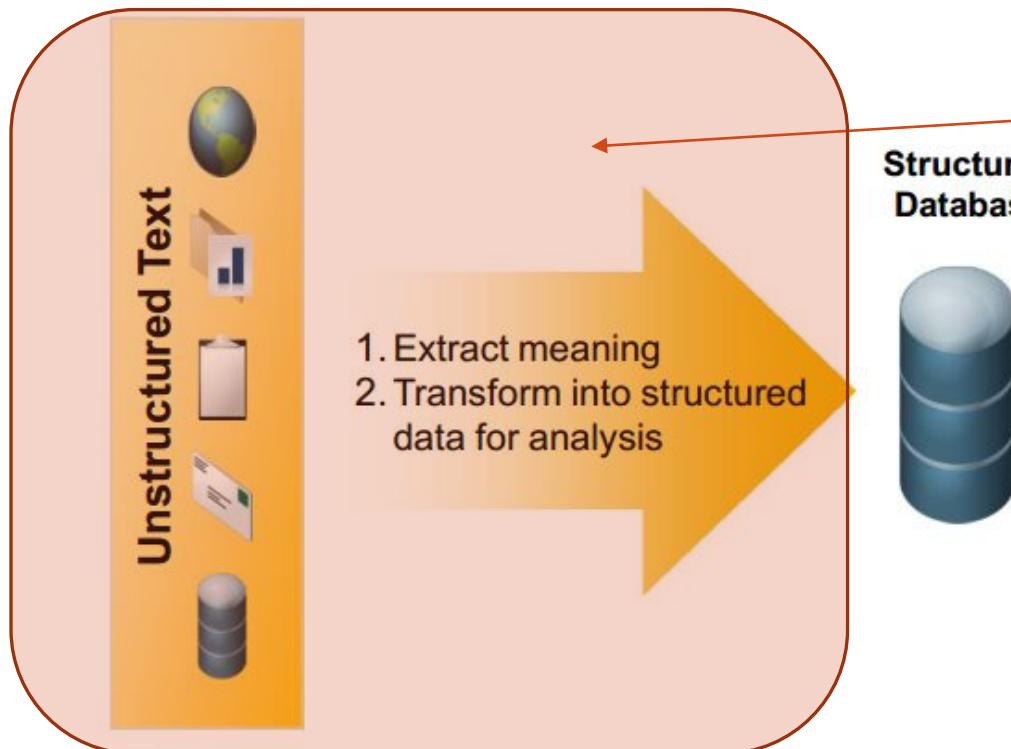


0.103	0.176	0.387	0.300	0.379
0.333	0.384	0.564	0.587	0.857
0.421	0.309	0.654	0.729	0.228
0.266	0.750	1.056	0.936	0.911
0.225	0.326	0.643	0.337	0.721
0.187	0.586	0.529	0.340	0.829
0.153	0.485	0.560	0.428	0.628

## *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	
<b>Characteristics of Structured Data</b>	
• High organized	
• Clearly defined	
• Easy to access	
• Easy to analyze	
<b>Examples of Structured Data</b>	
• Name	
• Age	
• Gender	
• Address	
• Phone number	
• Currency	
• Dates	
• Billing info	
<b>Sources of Structured Data</b>	
• 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	50
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itottawa	28

▼ Person (1) Alles in orde ✓

Person	
name:	Nick van de Veerdonk
jobTitle:	WordPress developer & Designer
address [PostalAddress]:	
streetAddress:	Plesmanlaan 36
postalCode:	1421XR
addressLocality:	Uithoorn
telephone:	+31614805219
addressCountry [Country]:	
name:	The Netherlands

# Structured Data

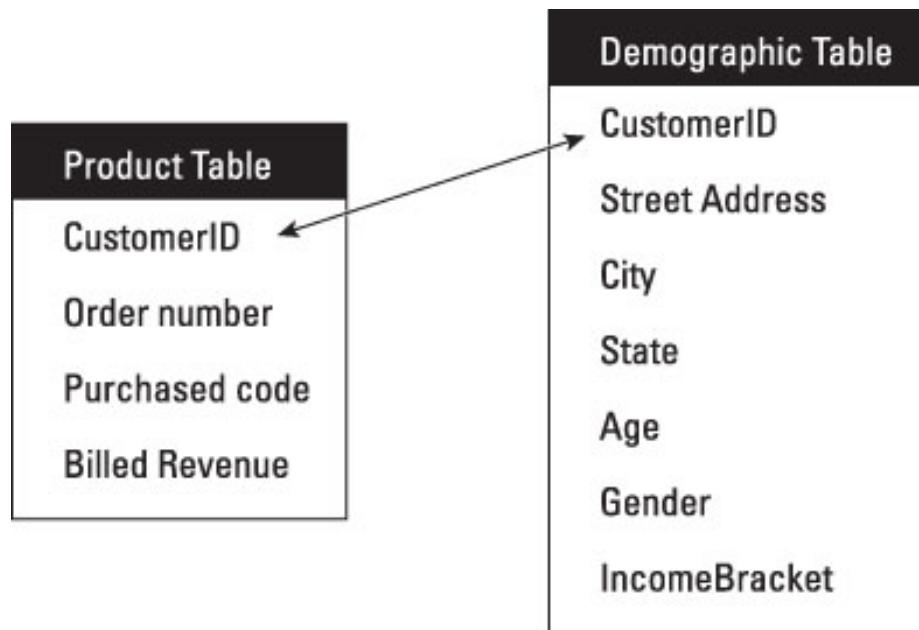
Google Structured Data Testing Tool

```

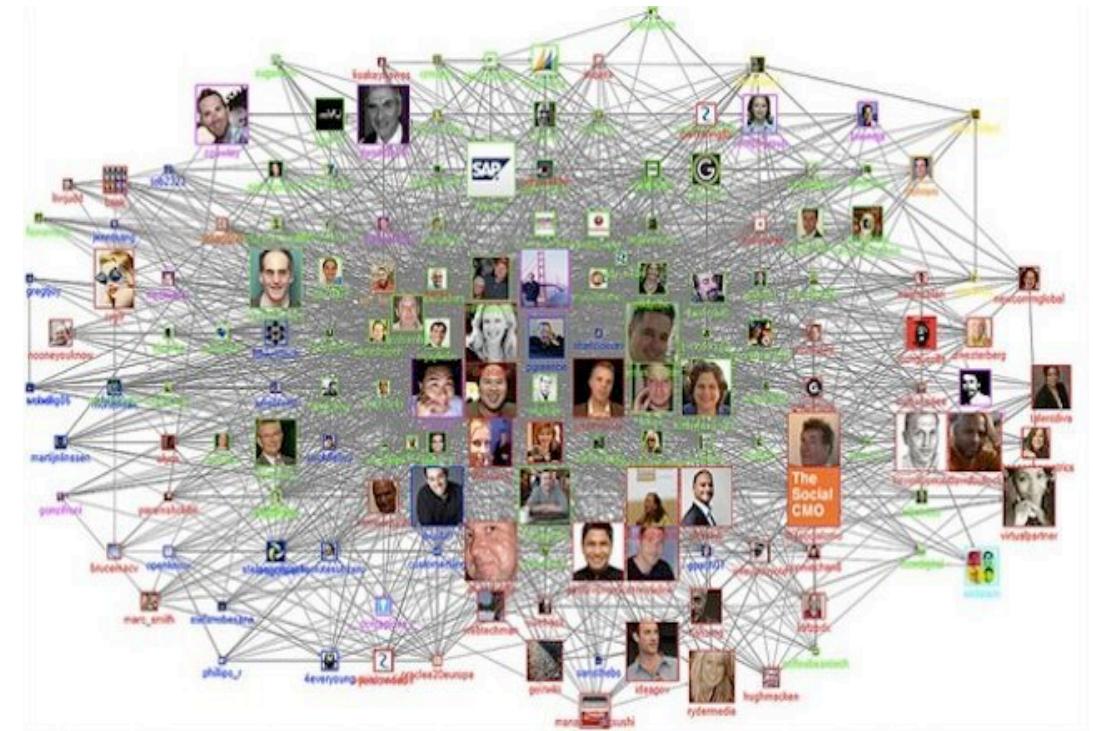
1 <script type="application/ld+json">
2 {
3   "@context": "http://schema.org",
4   "@type": "LocalBusiness",
5   "image": "http://onlineownership.com/wp-
content/themes/onlineownership/images/online-ownership-logo.png",
6   "priceRange" : "£40 - £450",
7   "address": {
8     "@type": "PostalAddress",
9     "streetAddress": "6 Regal Close",
10    "addressLocality": "Corby",
11    "addressRegion": "Northamptonshire",
12    "postalCode": "NN17 1EZ"
13  },
14  "description": "Online Ownership is a results driven, ethical SEO and
online marketing consultancy",
15  "name": "Online Ownership",
16  "openingHours": "Mo-Fr 09:00-18:00",
17  "telephone": "01536 269 657",
18  "email": "info@onlineownership.com",
19  "url": "http://onlineownership.com",
20  "hasMap": "https://goo.gl/maps/nGeYv6mVxvF2",
21  "sameAs" : [
22    "https://plus.google.com/+OnlineownershipUK",
23    "https://www.youtube.com/channel/UCwpL0xyVCiYJS_Yozf8Mt_g"
24  ]
25 }
26 }
27 
```

LocalBusiness		PREVIEW	0 ERRORS	0 WARNINGS
@type	LocalBusiness			
image	http://onlineownership.com/wp- content/themes/onlineownership/images/o nline-ownership-logo.png			
priceRange	£40 - £450			
description	Online Ownership is a results driven, ethical SEO and online marketing consultancy			
name	Online Ownership			
openingHours	Mo-Fr 09:00-18:00			
telephone	01536 269 657			
email	info@onlineownership.com			
url	http://onlineownership.com			
hasMap	https://goo.gl/maps/nGeYv6mVxvF2			
sameAs	https://plus.google.com/+OnlineownershipU K			
sameAs	https://www.youtube.com/channel/UCwpL0 xyVCiYJS_Yozf8Mt_g			
address				
@type	PostalAddress			

# Relational Data



*Stored in Traditional Relational Databases*



*Stored in Newer Graph Databases*

# Relational Database (RDBMS)

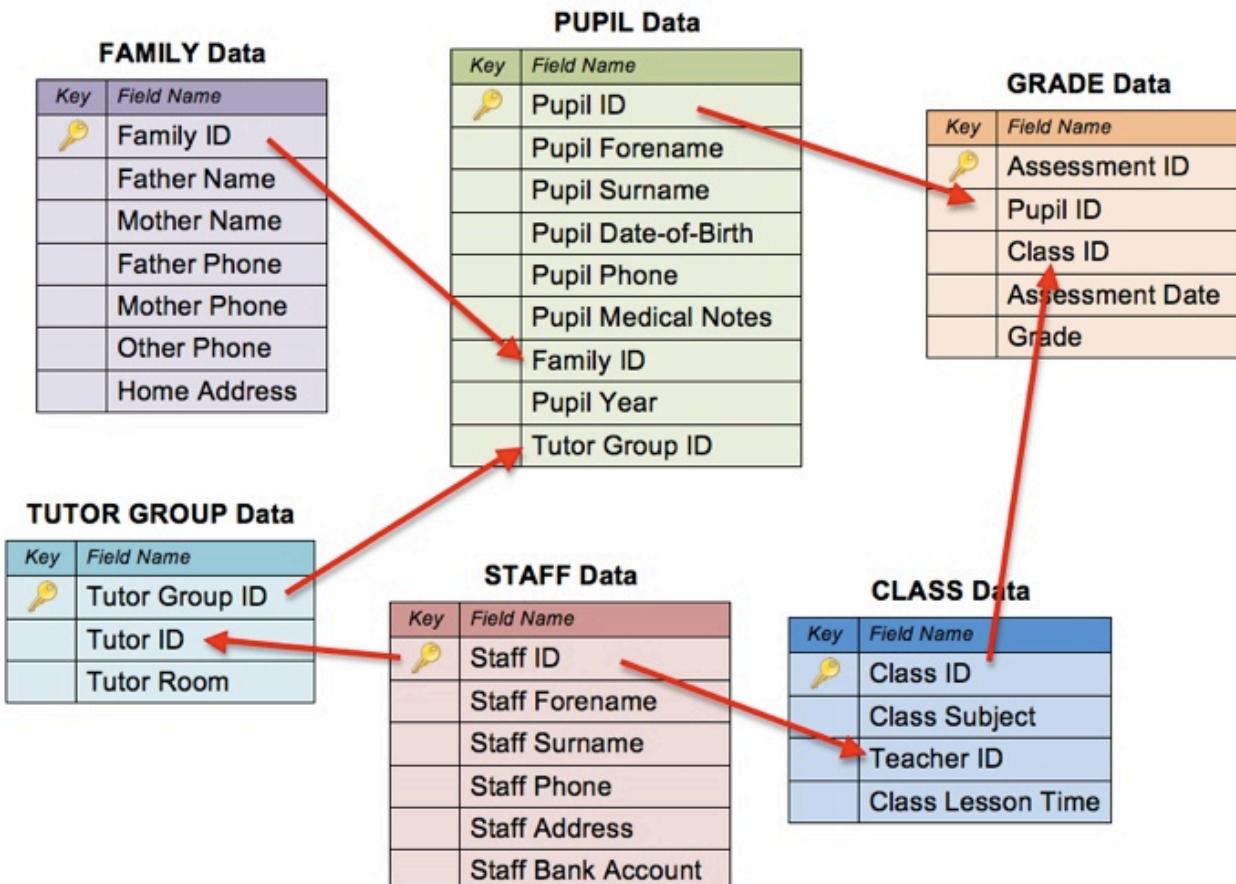


*Before*



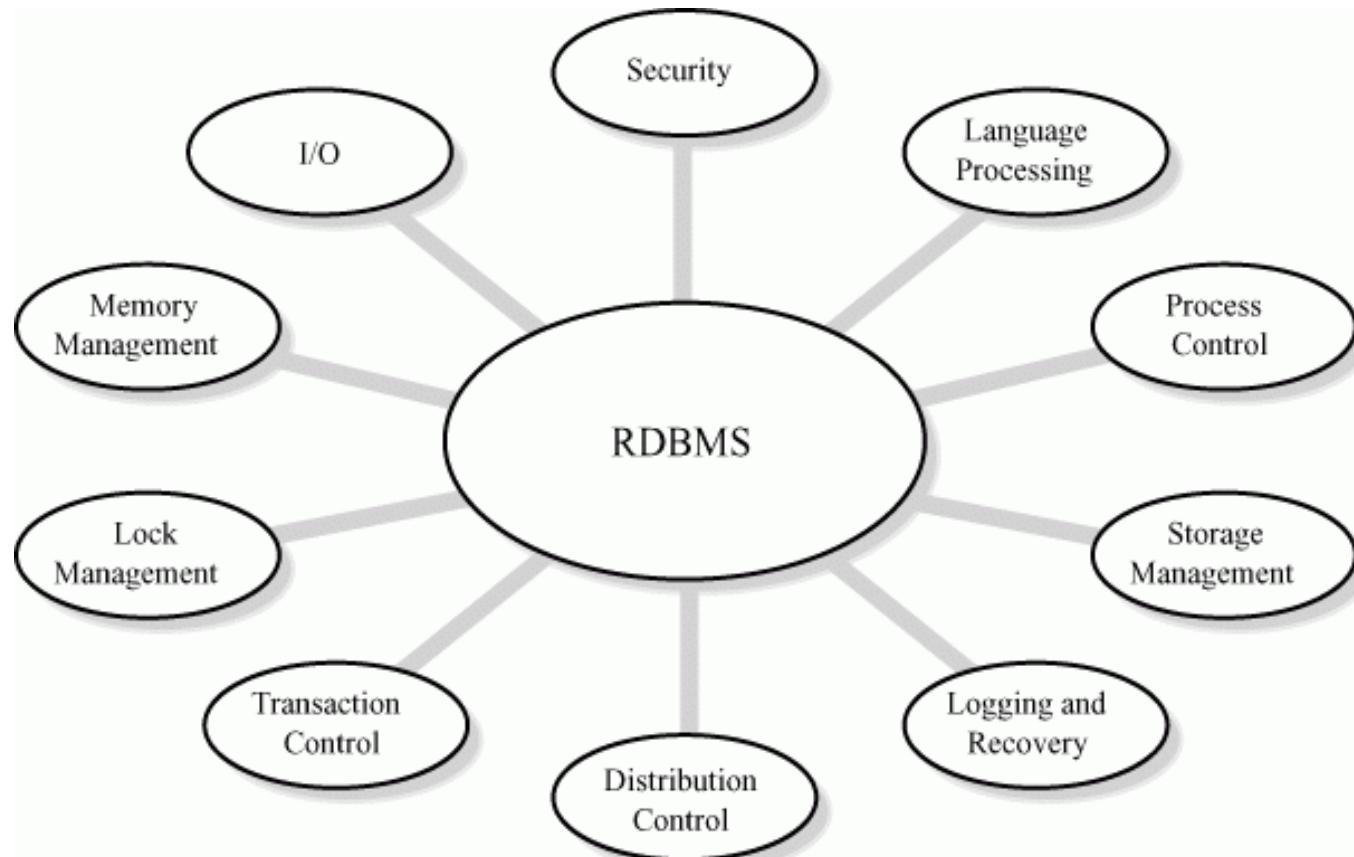
*After*

# 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*



Amazon Relational Database Service (RDS)

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 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
      - stocks

```

1 SELECT table_name
2 FROM information_schema.tables
3 WHERE table_schema = 'public'
4
5 SELECT *
6 FROM information_schema.columns
7 WHERE table_name = 'flights_2014'
8
9 SELECT count(*) FROM flights_2014;
10
11 SELECT *
12 FROM flights_2014
13 WHERE day_of_week = 3
14 ORDER BY distance DESC
15 LIMIT 10;
16
17
18 SELECT day_of_week,avg(dep_delay) AS avg_dep_delay,
19           max(arr_delay) AS max_arr_delay,
20           avg(distance) AS avg_distance,
21           avg(air_time) AS avg_air_time
22 FROM flights_2014
23 WHERE day_of_week BETWEEN 1 AND 7
24 GROUP BY day_of_week
25 ORDER BY day_of_week
26 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=#
```

```
~/Downloads — rohit@rohit-lubuntu: ~ — ssh rohit@rohit-lubuntu — ttys000
```

```
~/R Projects/DSLA Course/Week 5 — rohit@rohit-lubuntu: ~ — -bash — ttys001
```

```
[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

---



**Break Time**