

Plan:

1. Discuss the basic types of data used by data scientists
2. Familiarize yourself with common data formats

# Data & Data Types

Shannon E. Ellis, Ph.D  
UC San Diego



Department of Cognitive Science  
[sellis@ucsd.edu](mailto:sellis@ucsd.edu)

**How may data you'll use in  
this course be structured?**

---

# Data Structures Review

## Structured data

- can be stored in database (SQL)
- tables with rows and columns
- requires a relational key
- 5-10% of all data

## Semi-structured data

- doesn't reside in a relational database
- has organizational properties (easier to analyze)
- CSV, XML, JSON

## Unstructured

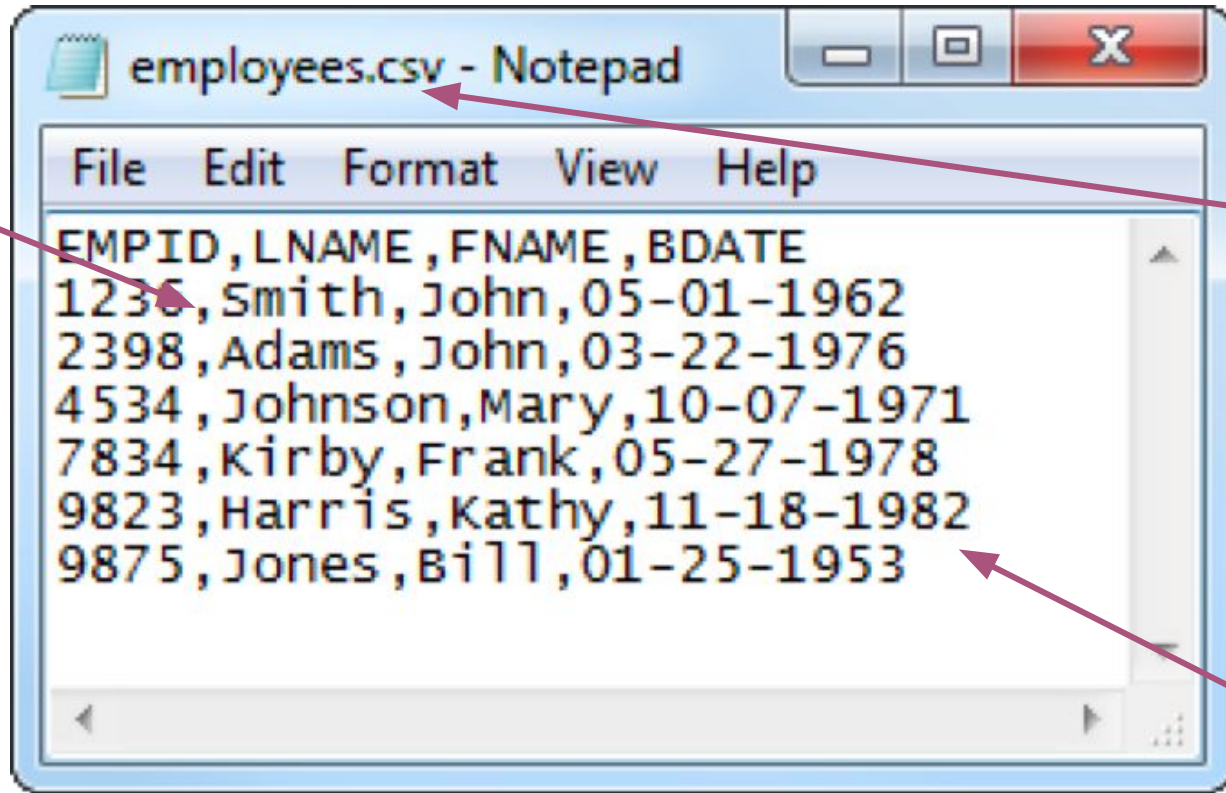
- non-tabular data
- 80% of the world's data
- images, text, audio, videos

# (Semi-)Structured Data

*Data that is stored in such a way that it is easy to search and work with. These data are stored in a particular format that adheres to organization principles imposed by the file format. These are the data structures data scientists work with most often.*

# CSVs

Each column separated by a comma



Has the extension ".csv"

Each row is separated by a new line



# sample\_data



File

Edit

View

Insert

Format

Data

Tools



100%



\$

%

.0



.00



1

*fx*

	A	B	C
1	name	height	blood_type
2	Natasha	5'2"	A-
3	Hassan	6'	B-
4	Chun	5'8"	O

# JSON: key-value pairs

*nested/hierarchical data*

```
{"Name": "Isabela"}
```

key



value

These are all  
nested within  
attributes

These are all  
nested within  
"Good For"

```
"attributes": {  
  "Take-out": true,  
  "Wi-Fi": "free",  
  "Drive-Thru": true,  
  "Good For": {  
    "dessert": false,  
    "latenight": false,  
    "lunch": false,  
    "dinner": false,  
    "breakfast": false,  
    "brunch": false  
  },  
}
```

# JSON



# Extensible Markup Language (XML): nodes, tags, and elements

*nested/hierarchical data*

A **node**

`$node`

An **opening tag**

`<tag>`

An **element**

`<tag2> more content </tag2>`

`<tag3> more content </tag3>`

`</tag>`

A **closing tag**

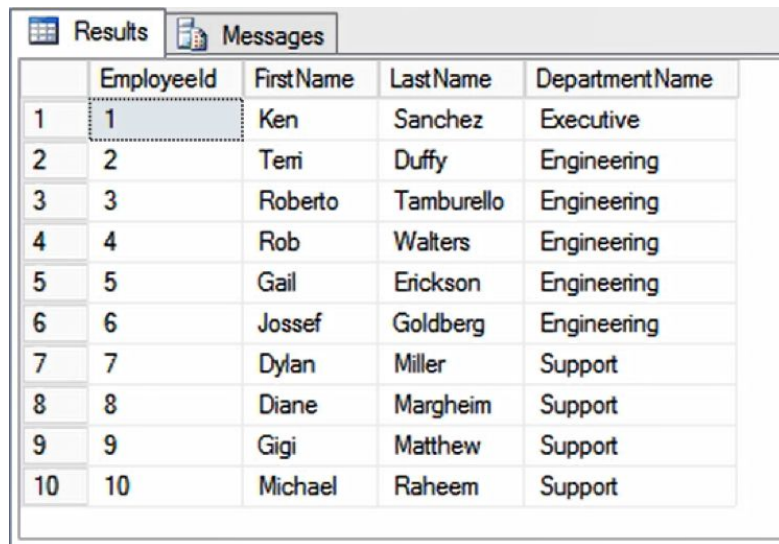
XML

```
<?xml version="1.0" encoding="UTF-8"?>
<customers>
  <customer>
    <customer_id>1</customer_id>
    <first_name>John</first_name>
    <last_name>Doe</last_name>
    <email>john.doe@example.com</email>
  </customer>
  <customer>
    <customer_id>2</customer_id>
    <first_name>Sam</first_name>
    <last_name>Smith</last_name>
    <email>sam.smith@example.com</email>
  </customer>
  <customer>
    <customer_id>3</customer_id>
    <first_name>Jane</first_name>
    <last_name>Doe</last_name>
    <email>jane.doe@example.com</email>
  </customer>
</customers>
```

# XML

# Relational Databases: A set of interdependent tables

1. Efficient Data Storage
2. Avoid Ambiguity
3. Increase Data Privacy

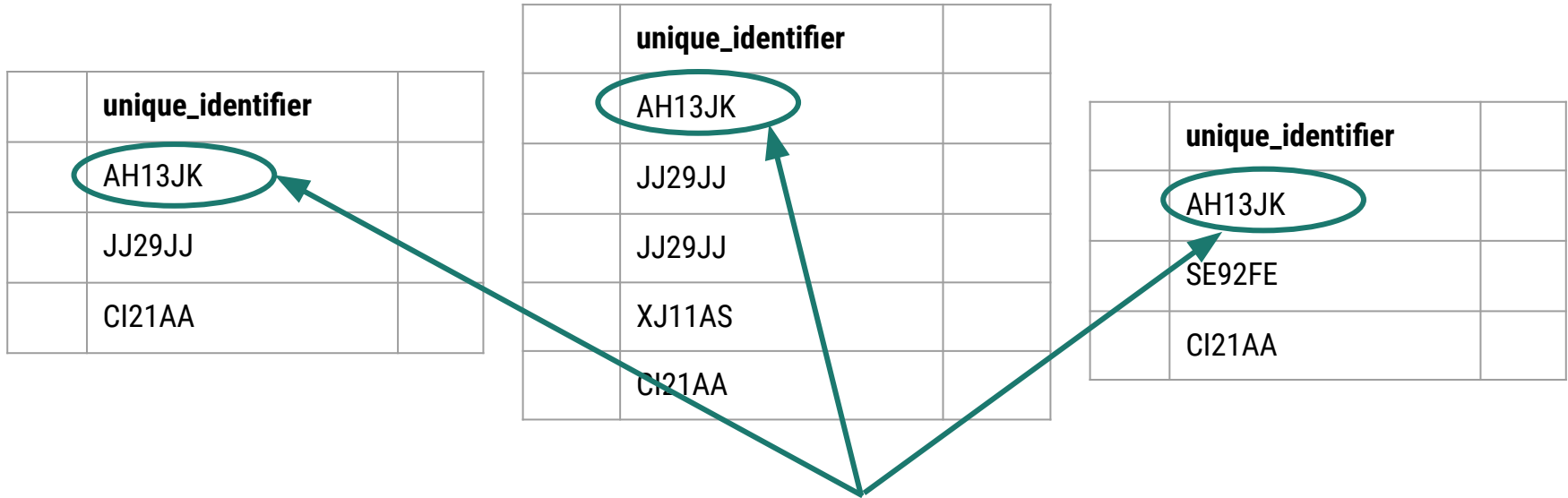


A screenshot of a database application window. At the top, there are two tabs: 'Results' (active) and 'Messages'. Below the tabs is a table with five columns: 'EmployeeId', 'FirstName', 'LastName', and 'DepartmentName'. The table contains 10 rows of data. The first row is highlighted with a blue background. The data is as follows:

	EmployeeId	FirstName	LastName	DepartmentName
1	1	Ken	Sanchez	Executive
2	2	Teri	Duffy	Engineering
3	3	Roberto	Tamburello	Engineering
4	4	Rob	Walters	Engineering
5	5	Gail	Erickson	Engineering
6	6	Jossef	Goldberg	Engineering
7	7	Dylan	Miller	Support
8	8	Diane	Margheim	Support
9	9	Gigi	Matthew	Support
10	10	Michael	Raheem	Support

relational database

# Information is stored across tables



entries are *related* to one another by their unique identifier

relational database

## restaurant

name	id	address	type
Taco Stand	AH13JK	1 Main St.	Mexican
Pho Place	<b>JJ29JJ</b>	192 Street Rd.	Vietnamese
Taco Stand	XJ11AS	18 W. East St.	Fusion
Pizza Heaven	CI21AA	711 K Ave.	Italian

## health inspections

id	inspection_date	inspector	score
AH13JK	2018-08-21	Sheila	97
<b>JJ29JJ</b>	2018-03-12	D'eonte	98
<b>JJ29JJ</b>	2018-01-02	Monica	66
XJ11AS	2018-12-16	Mark	43
CI21AA	2018-08-21	Anh	99

## rating

id	stars
AH13JK	4.9
<b>JJ29JJ</b>	4.8
XJ11AS	4.2
CI21AA	4.7

relational database

## restaurant

name	id	address	type
Taco Stand	AH13JK	1 Main St.	Mexican
Pho Place	JJ29JJ	192 Street Rd.	Vietnamese
Taco Stand	XJ11AS	18 W. East St.	Fusion
Pizza Heaven	CI21AA	711 K Ave.	Italian

## health inspections

id	inspection_date	inspector	score
AH13JK	2018-08-21	Sheila	97
JJ29JJ	2018-03-12	D'eonte	98
JJ29JJ	2018-01-02	Monica	66
XJ11AS	2018-12-16	Mark	43
CI21AA	2018-08-21	Anh	99

## rating

id	stars
AH13JK	4.9
JJ29JJ	4.8
XJ11AS	4.2
CI21AA	4.7

Two different restaurants with the same name will have different unique identifiers

# relational database

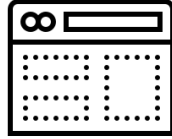
# Unstructured Data

*Some datasets record information about the state of the world, but in a more heterogeneous way. Perhaps it is a large text corpus with images and links like Wikipedia, or the complicated mix of notes and test results appearing in personal medical records.*

# Unstructured Data Types



Text files and  
documents



Websites and  
applications



Sensor  
data



Image  
files



Audio  
files



Video  
files



Email  
data



Social media  
data





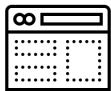
Positive:  
70%

Negative:  
20%

Neutral:  
10%



Text:  
Sentiment Analysis



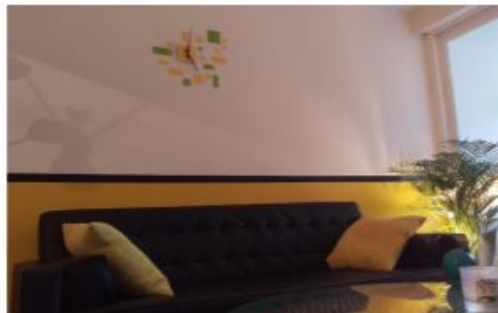
# PYTHON

---

# BEAUTIFULSOUP WEB SCRAPING



## Bedroom Or Not?



“The left two photos were correctly predicted as bedrooms; The right two photos were correctly predicted NOT as bedrooms.”

# Data Structures Review

## Structured data

- can be stored in database (SQL)
- tables with rows and columns
- requires a relational key
- 5-10% of all data

## Semi-structured data

- doesn't reside in a relational database
- has organizational properties (easier to analyze)
- CSV, XML, JSON

## Unstructured

- non-tabular data
- 80% of the world's data
- images, text, audio, videos