Data Science

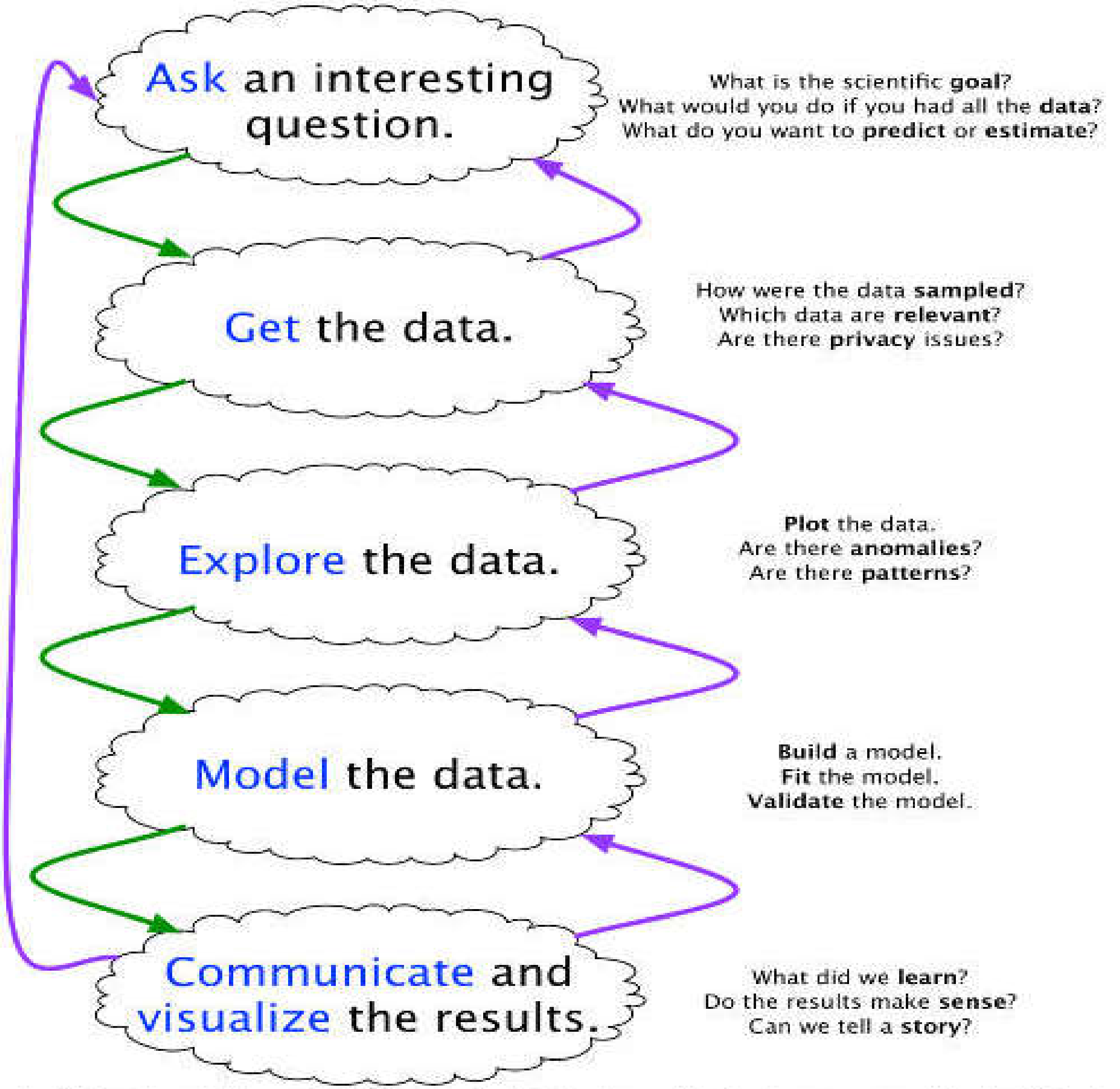
A Discussion on Data

The Data Science Process

- Ask questions
- Data Collection
- Data Exploration
- Data Modeling
- Data Analysis
- Visualization and Presentation of Results

Today we will begin introducing the data collection and data exploration steps.

The Data Science Process



Ask an interesting question.

What is the scientific **goal**?
What would you do if you had all the **data**?
What do you want to **predict** or **estimate**?

Get the data.

How were the data **sampled**?
Which data are **relevant**?
Are there **privacy** issues?

Explore the data.

Plot the data.
Are there **anomalies**?
Are there **patterns**?

Model the data.

Build a model.
Fit the model.
Validate the model.

Communicate and **visualize** the results.

What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?

Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course http://cs109.org/.

Data Collection and Preparation

Data Collection and Preparation

- Before
  - Ask questions <=> Fix sources of data
  - Assess how much data are needed for the analyses

Data Collection and Preparation

- Before
  - Ask questions <=> Fix sources of data
  - Assess how much data are needed for the analyses
- Collect Data
  - By crawling/scraping, automated scripts, crowdsourcing, from 3rd party etc.

Data Collection and Preparation

- Before
  - Ask questions <=> Fix sources of data
  - Assess how much data are needed for the analyses
- Collect Data
  - By crawling/scraping, automated scripts, crowdsourcing, from 3rd party etc.
- Preprocess
  - Cleaning Data
    - Formatting
    - Handling missing values
    - Removing repetitions (optional) etc.
  - Data Encoding

Data Collection and Preparation

- Before
  - Ask questions <=> Fix sources of data
  - Assess how much data are needed for the analyses
- Collect Data
  - By crawling/scraping, automated scripts, crowdsourcing, from 3rd party etc.
- Preprocess
  - Cleaning Data
    - Formatting
    - Handling missing values
    - Removing repetitions (optional) etc.
  - Data Encoding
- Annotate (if required)
  - Validate annotations

Data Collection and Preparation

- Before
  - Ask questions <=> Fix sources of data
  - Assess how much data are needed for the analyses
- Collect Data
  - By crawling/scraping, automated scripts, crowdsourcing, from 3rd party etc.
- Preprocess
  - Cleaning Data
    - Formatting
    - Handling missing values
    - Removing repetitions (optional) etc.
  - Data Encoding
- Annotate (if required)
  - Validate annotations
- Exploratory Data Analysis
  - Examining data to observe patterns, find issues using Statistics and Visualization

Let’s start from the basic!

Let’s start from the basic!

What are Data?

What are Data?

- “A datum is a single measurement of something on a scale that is understandable to both the recorder and the reader. Data are multiple such measurements .”

What are Data?

- “A datum is a single measurement of something on a scale that is understandable to both the recorder and the reader. Data are multiple such measurements .”
- Everything is data!

Where do data come from?

Where do data come from?

- Internal Sources
  - Already collected by or is part of the overall data collection of an organization
  - Business-centric data that is available in the organization database to record day to day operations; data obtained from scientific experiments etc.

Where do data come from?

- Internal Sources
  - Already collected by or is part of the overall data collection of an organization
  - Business-centric data that is available in the organization database to record day to day operations; data obtained from scientific experiments etc.
- Existing External Sources
  - Available and ready to read format from an outside source for free or for a fee
  - Public govt. databases, stock market data

Where do data come from?

- Internal Sources
  - Already collected by or is part of the overall data collection of an organization
  - Business-centric data that is available in the organization database to record day to day operations; data obtained from scientific experiments etc.
- Existing External Sources
  - Available and ready to read format from an outside source for free or for a fee
  - Public govt. databases, stock market data
- External Sources Requiring Collection Efforts
  - Available from external sources but acquisition requires special processing
  - Printed data; website data

Where do data come from?

- API(Application Program Interface)
  - Using a prebuilt set of functions developed by a company to access their services. Often paid.
  - Google Map API, Facebook API, Twitter API

Where do data come from?

- API(Application Program Interface)
  - Using a prebuilt set of functions developed by a company to access their services. Often paid.
  - Google Map API, Facebook API, Twitter API
- RSS(Rich Site Summary)
  - Summarizes frequently updated online content in standard format. Free to read if the site has one.
  - News related sites, blogs

Where do data come from?

- API(Application Program Interface)
  - Using a prebuilt set of functions developed by a company to access their services. Often paid.
  - Google Map API, Facebook API, Twitter API
- RSS(Rich Site Summary)
  - Summarizes frequently updated online content in standard format. Free to read if the site has one.
  - News related sites, blogs
- Web Scraping
  - Using software, scripts or by hand extracting data from what is displayed on a page or what is contained in the HTML file.

Web Scraping

- Older govt. or smaller news sites might not have APIs for accessing data, or publish RSS feeds or have databases for download.
- One doesn’t want to pay for API or database access!
- Should we do it?
    - For exploration or publishing the analysis or product
    - See terms of services
    - Check privacy concerns for websites and their clients

Types of Data

Types of Data

- What kind of values are in the data - Simple or atomic?

Types of Data

- What kind of values are in the data - Simple or atomic?
- Numeric
  - Integers, floats

Types of Data

- What kind of values are in the data - Simple or atomic?
- Numeric
  - Integers, floats
- Boolean
  - Binary or true-false values

Types of Data

- What kind of values are in the data - Simple or atomic?
- Numeric
  - Integers, floats
- Boolean
  - Binary or true-false values
- Strings
  - Sequences or symbols

Types of Data

- It is important to distinguish between classes of variables or attributes based on the type of values they can take on.

Types of Data

- It is important to distinguish between classes of variables or attributes based on the type of values they can take on.
- Quantitative variable
    - Discrete – A finite number of values are possible in any bounded interval
        - Number of siblings
    - Continuous – An infinite number of values are possible in any bounded interval
        - Temperature, salary

Types of Data

- It is important to distinguish between classes of variables or attributes based on the type of values they can take on.
- Quantitative variable
  - Discrete – A finite number of values are possible in any bounded interval
    - Number of siblings
  - Continuous – An infinite number of values are possible in any bounded interval
    - Temperature, salary

Types of Data

- It is important to distinguish between classes of variables or attributes based on the type of values they can take on.
- Quantitative variable
  - Discrete – A finite number of values are possible in any bounded interval
    - Number of siblings
  - Continuous – An infinite number of values are possible in any bounded interval
    - Temperature, salary

Types of Data

- It is important to distinguish between classes of variables or attributes based on the type of values they can take on.
- Quantitative variable
  - Discrete – A finite number of values are possible in any bounded interval
    - Number of siblings
  - Continuous – An infinite number of values are possible in any bounded interval
    - Temperature, salary
- Categorical variable
  - Nominal – No inherent order among the values
    - What kind of pet do you have?
  - Ordinal – Categories represent some kind of order
    - User Rating {1,2, 3, 4, 5}, {Agree, Slightly Agree, Neither, Slightly Disagree, Disagee}

Types of Data

- It is important to distinguish between classes of variables or attributes based on the type of values they can take on.
- Quantitative variable
  - Discrete – A finite number of values are possible in any bounded interval
    - Number of siblings
  - Continuous – An infinite number of values are possible in any bounded interval
    - Temperature, salary
- Categorical variable
  - Nominal – No inherent order among the values
    - What kind of pet do you have?
  - Ordinal – Categories represent some kind of order
    - User Rating {1,2, 3, 4, 5}, {Agree, Slightly Agree, Neither, Slightly Disagree, Disagee}

Types of Data

- It is important to distinguish between classes of variables or attributes based on the type of values they can take on.
- Quantitative variable
  - Discrete – A finite number of values are possible in any bounded interval
    - Number of siblings
  - Continuous – An infinite number of values are possible in any bounded interval
    - Temperature, salary
- Categorical variable
  - Nominal – No inherent order among the values
    - What kind of pet do you have?
  - Ordinal – Categories represent some kind of order
    - User Rating {1,2, 3, 4, 5}, {Agree, Slightly Agree, Neither, Slightly Disagree, Disagree}

Types of Data

- What kind of values are in the data – Compound, composed of a bunch of atomic types?
- Date and time
    - Compound value with a specific structure
- Lists
    - A sequence of values
- Dictionaries
    - A collection of key-value pair

How data are represented and stored?

How data are represented and stored?

- Unstructured Data
  - Text, Images, Video, Audio

How data are represented and stored?

- Unstructured Data
  - Text, Images, Video, Audio
- Tabular Data
  - A dataset is a two-dimensional table, where each row typically represents a single data record and column represents one type of measurement
  - Csv, tsp, xlsx

How data are represented and stored?

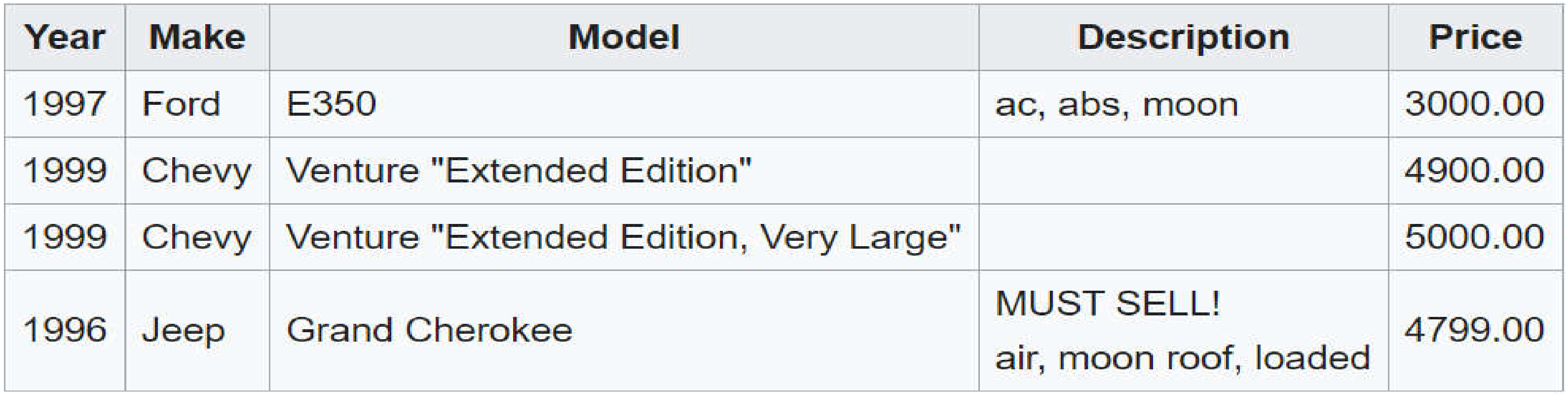
- Unstructured Data
  - Text, Images, Video, Audio
- Tabular Data
  - A dataset is a two-dimensional table, where each row typically represents a single data record and column represents one type of measurement
  - Csv, tsp, xlsx
- Structured Data
  - Each data record is presented in a form of a complex, multi-tiered, dictionary
  - Json, xml etc.

How data are represented and stored?

- Unstructured Data
  - Text, Images, Video, Audio
- Tabular Data
  - A dataset is a two-dimensional table, where each row typically represents a single data record and column represents one type of measurement
  - Csv, tsp, xlsx
- Structured Data
  - Each data record is presented in a form of a complex, multi-tiered, dictionary
  - Json, xml etc.
- Semi-structured data
  - Not all records are represented by the same set of keys or some data records are not represented using the key-value pair structure.

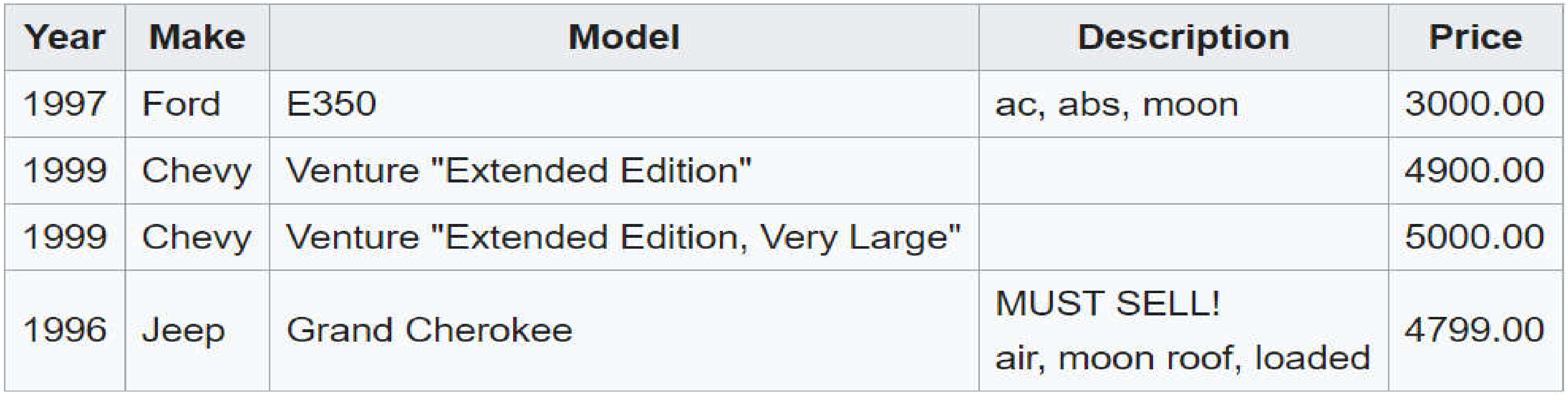
CSV, Comma Separated Values

CSV, Comma Separated Values

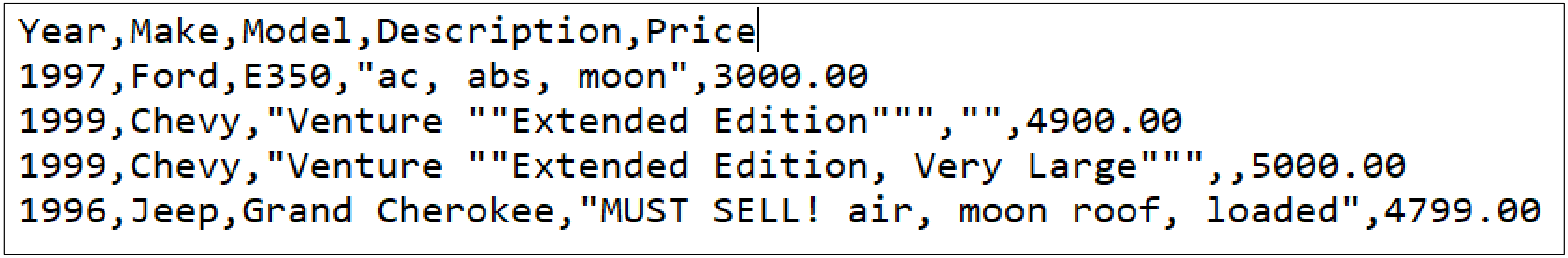


Year	Make	Model	Description	Price
1997	Ford	E350	ac, abs, moon	3000.00
1999	Chevy	Venture "Extended Edition"		4900.00
1999	Chevy	Venture "Extended Edition, Very Large"		5000.00
1996	Jeep	Grand Cherokee	MUST SELL! air, moon roof, loaded	4799.00

CSV, Comma Separated Values



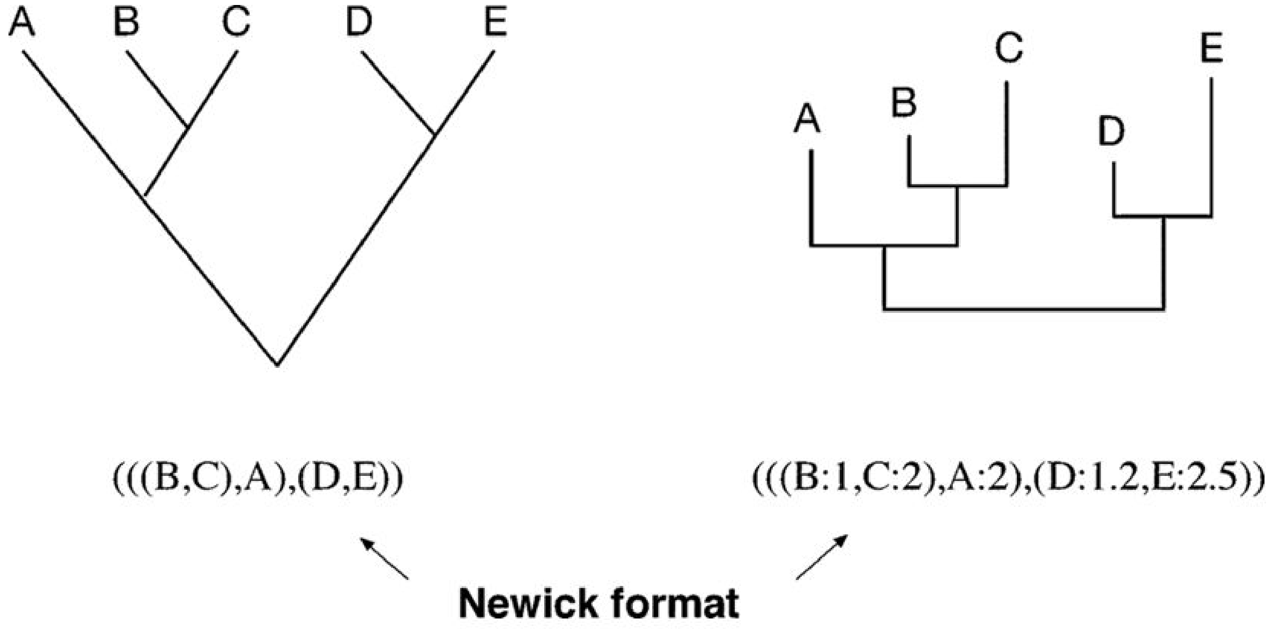
Year	Make	Model	Description	Price
1997	Ford	E350	ac, abs, moon	3000.00
1999	Chevy	Venture "Extended Edition"		4900.00
1999	Chevy	Venture "Extended Edition, Very Large"		5000.00
1996	Jeep	Grand Cherokee	MUST SELL! air, moon roof, loaded	4799.00



```
Year,Make,Model,Description,Price
1997,Ford,E350,"ac, abs, moon",3000.00
1999,Chevy,"Venture ""Extended Edition""","",4900.00
1999,Chevy,"Venture ""Extended Edition, Very Large""",,5000.00
1996,Jeep,Grand Cherokee,"MUST SELL! air, moon roof, loaded",4799.00
```

Hierarchies, e.g. Newick Tree

Hierarchies, e.g. Newick Tree



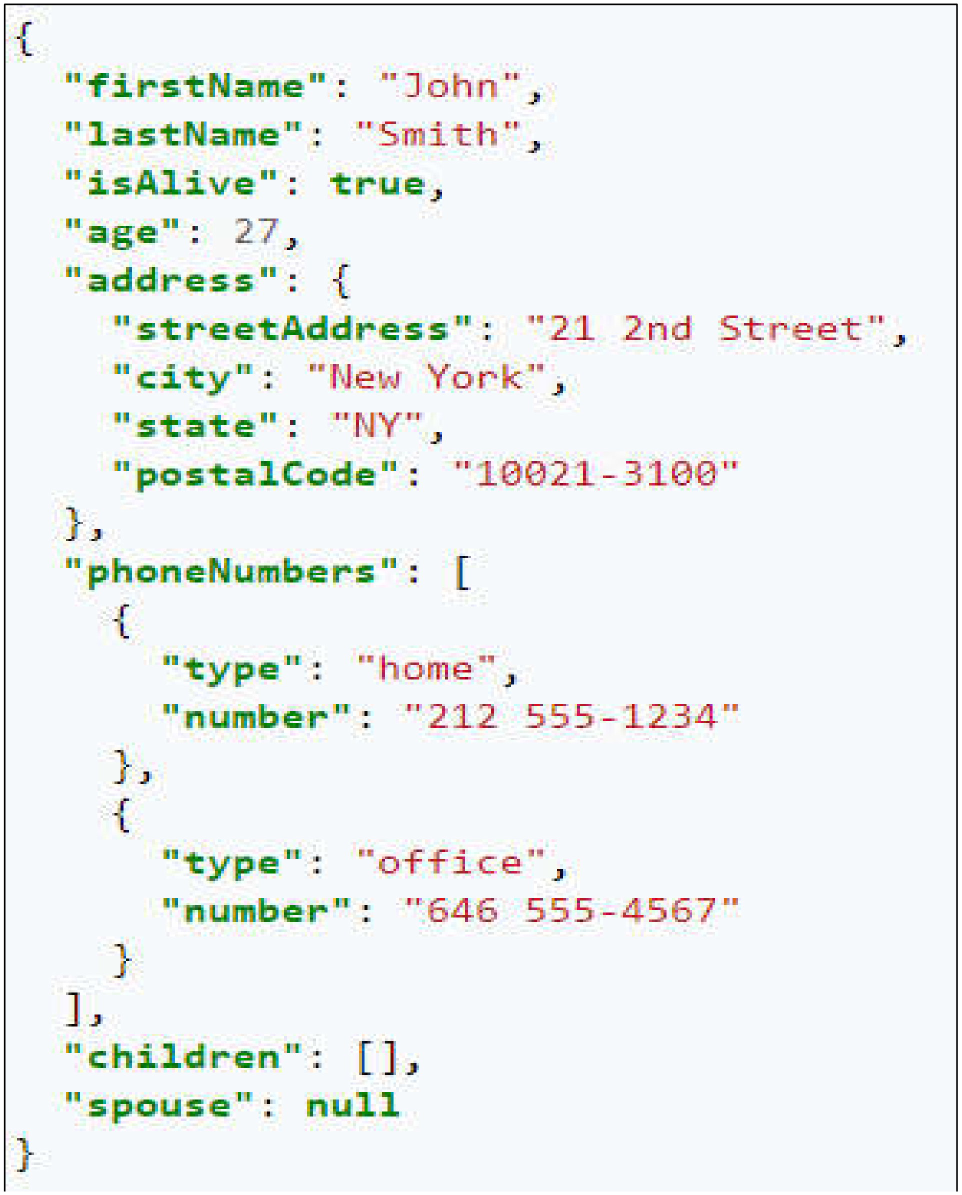
(((B,C),A),(D,E))

(((B:1,C:2),A:2),(D:1.2,E:2.5))

Newick format

JSON, JavaScript Object Notation

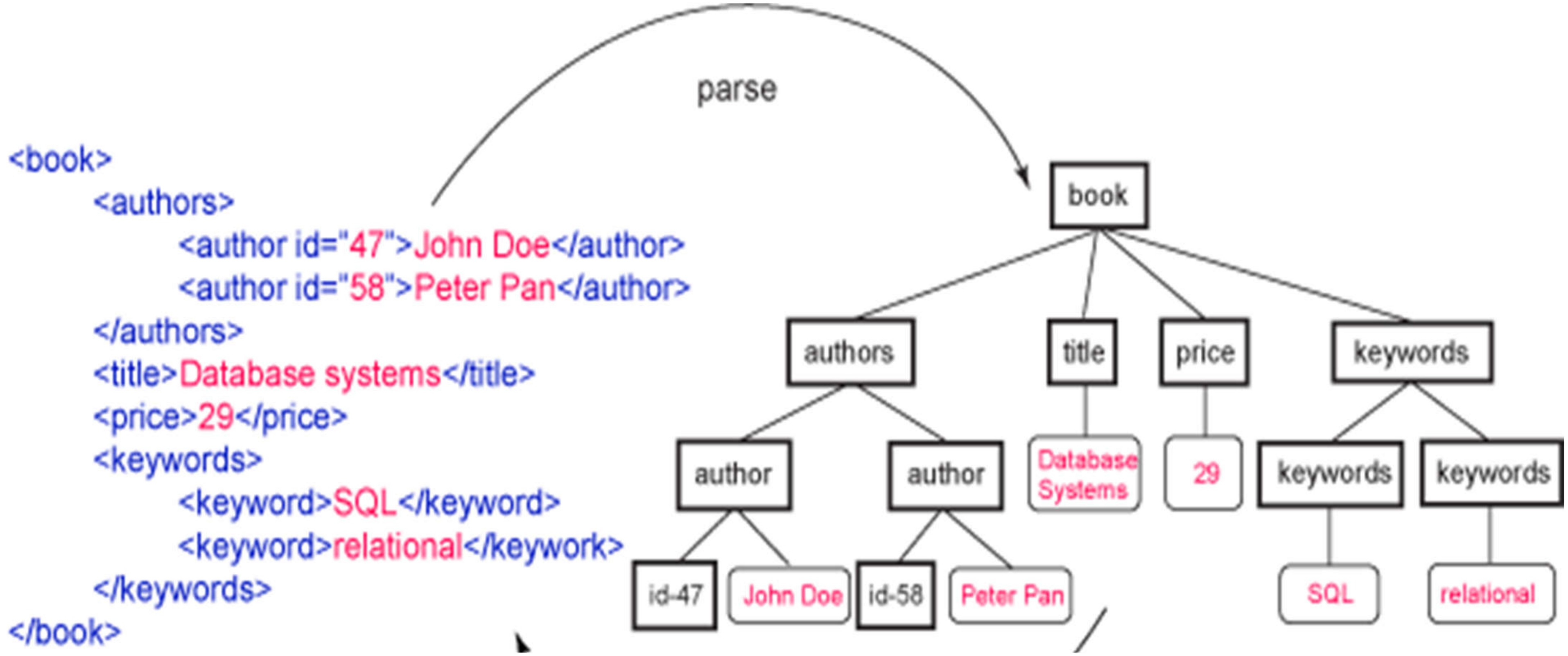
JSON, JavaScript Object Notation



```
{
  "firstName": "John",
  "lastName": "Smith",
  "isAlive": true,
  "age": 27,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021-3100"
  },
  "phoneNumbers": [
    {
      "type": "home",
      "number": "212 555-1234"
    },
    {
      "type": "office",
      "number": "646 555-4567"
    }
  ],
  "children": [],
  "spouse": null
}
```

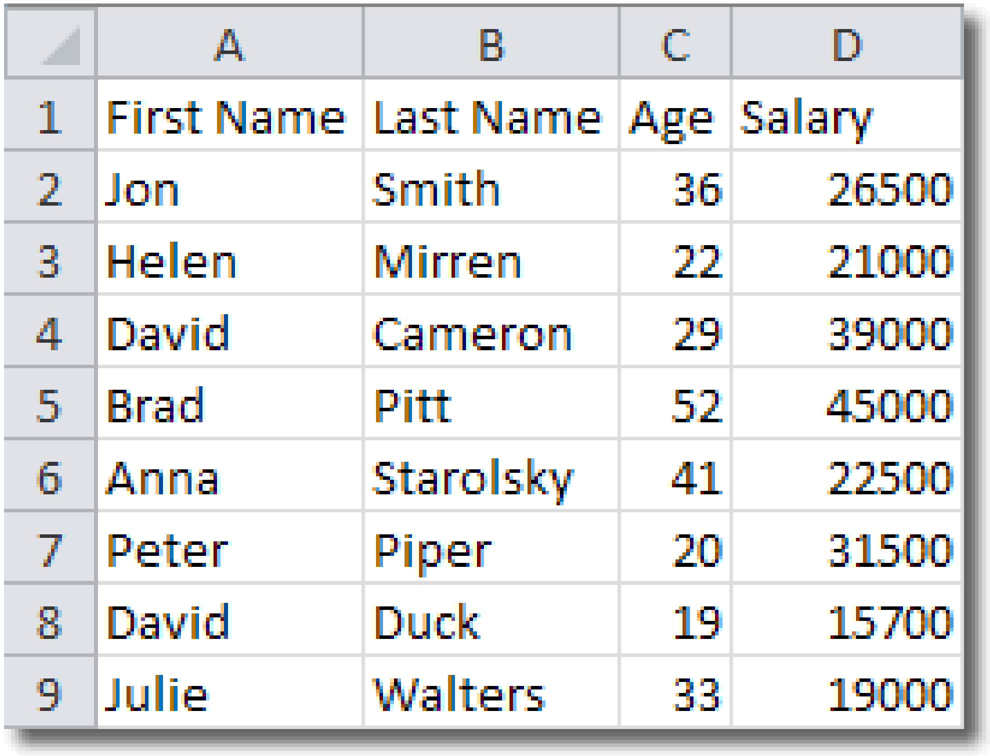
XML,
Extensible Markup Language

XML,
Extensible Markup Language



```
<book>
        <authors>
                <author id="47">John Doe</author>
                <author id="58">Peter Pan</author>
        </authors>
        <title>Database systems</title>
        <price>29</price>
        <keywords>
                <keyword>SQL</keyword>
                <keyword>relational</keywork>
        </keywords>
</book>
```

Tabular Data



	A	B	C	D
1	First Name	Last Name	Age	Salary
2	Jon	Smith	36	26500
3	Helen	Mirren	22	21000
4	David	Cameron	29	39000
5	Brad	Pitt	52	45000
6	Anna	Starolsky	41	22500
7	Peter	Piper	20	31500
8	David	Duck	19	15700
9	Julie	Walters	33	19000

Tabular Data

- We expect each record or observation to represent a set of measurements of a single object or event.

Tabular Data

- We expect each record or observation to represent a set of measurements of a single object or event.
- Each type of measurement is called a variable or an attribute (feature/Predictor) of the data.

Tabular Data

- We expect each record or observation to represent a set of measurements of a single object or event.
- Each type of measurement is called a variable or an attribute (feature/Predictor) of the data.
- The number of attributes is called the dimension.

Tabular Data

- We expect each record or observation to represent a set of measurements of a single object or event.
- Each type of measurement is called a variable or an attribute (feature/Predictor) of the data.
- The number of attributes is called the dimension.
- We expect each table to contain a set of records or observations of the same kind of object or event.

Common issues with the data

- Missing values
  - How to fill in?

Common issues with the data

- Missing values
  - How to fill in?
- Wrong values
  - How to detect and correct?

Common issues with the data

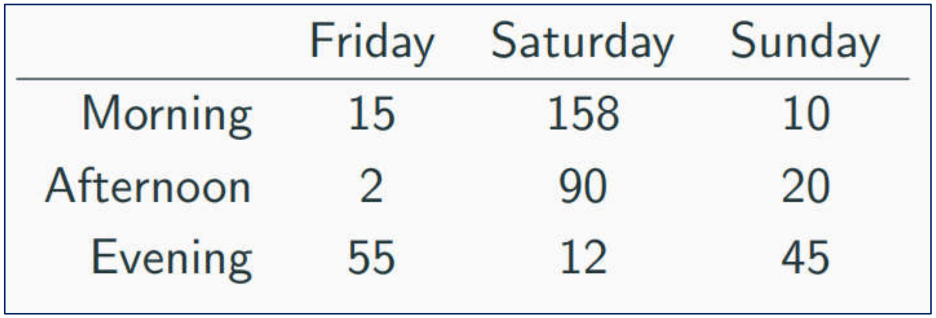
- Missing values
  - How to fill in?
- Wrong values
  - How to detect and correct?
- Messy format
  - Convert

Common issues with the data

- Missing values
    - How to fill in?
- Wrong values
    - How to detect and correct?
- Messy format
    - Convert
- Not usable
    - The data cannot answer the question asked

Handling Messy Data

- The following is a table accounting for the number of product deliveries over a weekend

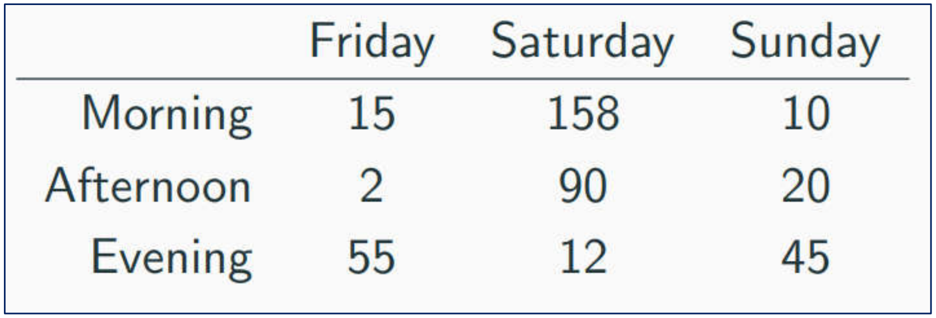


	Friday	Saturday	Sunday
Morning	15	158	10
Afternoon	2	90	20
Evening	55	12	45

- What are the variables?
- What object or event are we measuring?
- What's the issue? How to fix this?

Handling Messy Data

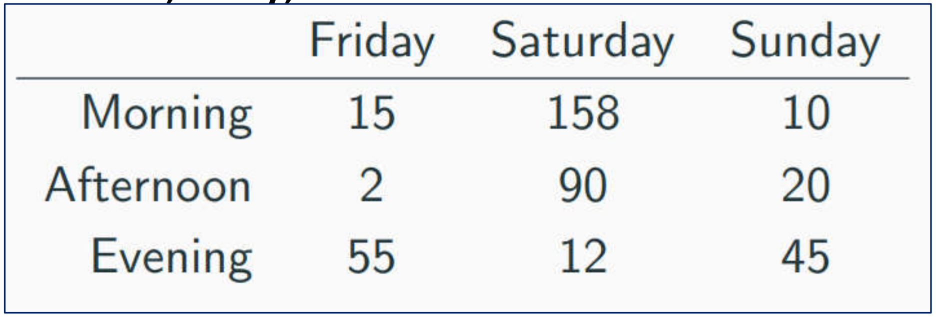
- Measuring individual deliveries
- The variables are Time, Day, Number of Products



	Friday	Saturday	Sunday
Morning	15	158	10
Afternoon	2	90	20
Evening	55	12	45

Handling Messy Data

- Measuring individual deliveries
- The variables are Time, Day, Number of Products

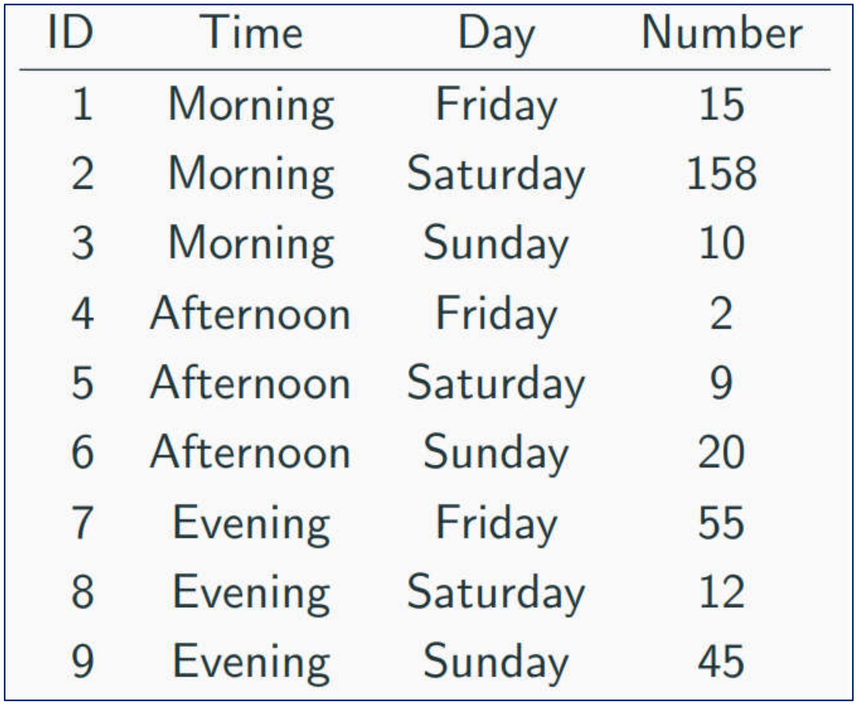


	Friday	Saturday	Sunday
Morning	15	158	10
Afternoon	2	90	20
Evening	55	12	45

- Problem
  - Each column header represents a single value rather than a variable
  - Row headers are hiding the day variable
  - The values of the variable “Number of products” is not recorded in a single column.

Handling Messy Data

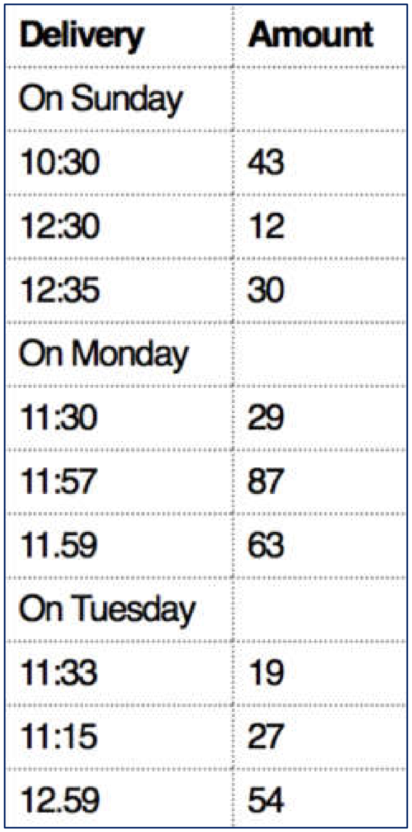
- We need to reorganize the information to make explicit the event were observing and the variables associated to the event.



ID	Time	Day	Number
1	Morning	Friday	15
2	Morning	Saturday	158
3	Morning	Sunday	10
4	Afternoon	Friday	2
5	Afternoon	Saturday	9
6	Afternoon	Sunday	20
7	Evening	Friday	55
8	Evening	Saturday	12
9	Evening	Sunday	45

More Messiness

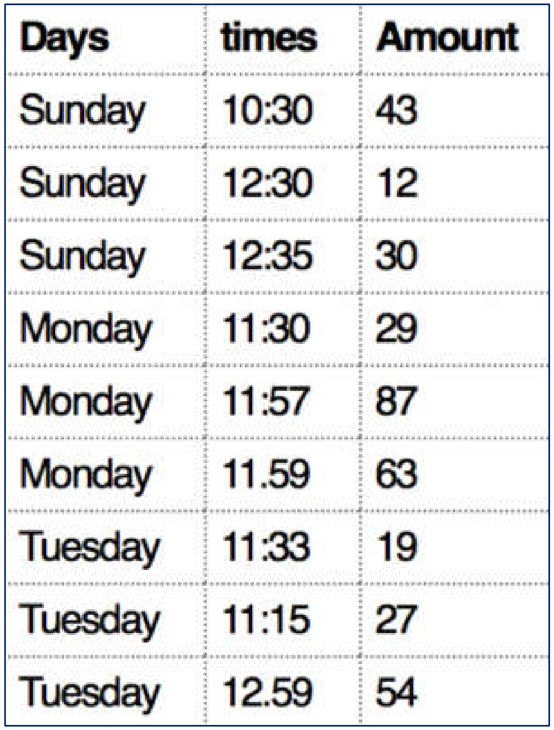
- What object or event we are measuring?
- What are the variables in this dataset?
- How do we fix it?



Delivery	Amount
On Sunday	
10:30	43
12:30	12
12:35	30
On Monday	
11:30	29
11:57	87
11.59	63
On Tuesday	
11:33	19
11:15	27
12.59	54

More Messiness

- Were measuring Individual Deliveries
- The variables are
  - Time, Day, Number of Products



Days	times	Amount
Sunday	10:30	43
Sunday	12:30	12
Sunday	12:35	30
Monday	11:30	29
Monday	11:57	87
Monday	11.59	63
Tuesday	11:33	19
Tuesday	11:15	27
Tuesday	12.59	54

Common Causes of Messiness

- Column headers are values, not variable names
- Variables are stored in both rows and columns
- Multiple variables are stored in one column
- Multiple types of experimental units are stored in same table

Tabularized Data

- In general, we want each file to correspond to a dataset
- We want to tabularize the data
- Each column to represent a single variable
- Each row to represent a single observation

Example Datasets

- UCI Machine Learning Data Repository
  - https://archive.ics.uci.edu/ml/datasets

data collection for dissertation?
COMPLETE.
quickmeme.com