

# **INTRO to DATA SCIENCE**

## **LECTURE 6A: DATA MUNGING IN PYTHON**

Rob Hall

DAT9 SF // August 21, 2014

# DATA MUNGING IN PYTHON

- Overview of Data Munging
- Common Scenarios to Expect
- Hands-on Exercise in Python

---

**DATA MUNGING IN PYTHON**

---

# **OVERVIEW OF DATA MUNGING**

# FOR BIG-DATA SCIENTISTS, 'JANITOR WORK' IS KEY HURDLE TO INSIGHTS

*From NYTimes on August 18, 2014:*

“Data wrangling is a huge — and surprisingly so — part of the job,” said Monica Rogati, vice president for data science at Jawbone, whose sensor-filled wristband and software track activity, sleep and food consumption, and suggest dietary and health tips based on the numbers. “It’s something that is not appreciated by data civilians. At times, it feels like everything we do.”



# DATA MUNGING IS OSEMN (AWESOME)

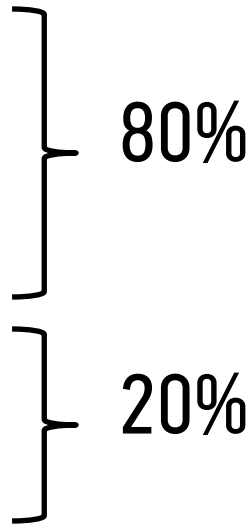
Obtain Data

Scrub Data

Explore

Model Algorithms

interpret Results



Majority of time  
is spent data munging

---

DATA MUNGING IN PYTHON

---

# COMMON SCENARIOS TO EXPECT

# COMMON DATA TYPES

### Structured Data

- CSV
- JSON
- Relational Databases
- XML
- YAML

### Unstructured Data

- Text
- Images
- Audio
- Video

# **COMMON DATA LOADING SCENARIOS**

- Import data from files
- Query data from databases
- Collect data from APIs
- Scrape data from websites



# DATA SCRUBBING SCENARIOS

- Convert data types (e.g., strings to numbers/dates)
- Replace missing data (e.g., impute null values)
- Parse/extract data (e.g., separate first name and last name from single string)
- Clean data
  - Standardize upper and lower cases
  - Correct misspelled words, abbreviations, plurals, conjugations
  - Adjust time zones

---

DATA MUNGING IN PYTHON

---

# HANDS-ON EXERCISE IN PYTHON

# EXERCISE 1: IMPORT JSON FILE

---

11

## KEY OBJECTIVES

---

Extract data from a file and import the data into python environment

## AGENDA

---

*20 mins*

1. Import json package
2. Open file with context manager
3. Implement appropriate json function

## DELIVERABLE

---

List of timestamps

## RESOURCES

---

1. `ga_hw_logins.json` # data file
2. json package documentation
3. stackoverflow example – ‘loading-parsing-json-file- in-python’

## EXERCISE 2: CONVERT STRINGS TO DATETIME

---

12

### KEY OBJECTIVES

---

Convert strings into datetime objects

### AGENDA

---

*10 mins*

1. Import datetime package
2. Build iterator (for loop)
3. Implement appropriate date time function

### DELIVERABLE

---

List of datetime objects

### RESOURCES

---

1. Results from Exercise 1
2. Stackoverflow example - 'converting-string-into-datetime'
3. datetime package documentation

## EXERCISE 3: CREATE SQLITE DATABASE

---

13

### KEY OBJECTIVES

---

Use python to create sqlite3 database and load data into a table

### DELIVERABLE

---

Sqlite3 database with data table

### AGENDA

---

*30 mins*

1. Import sqlite package
2. Create database file
3. Open database connection
4. Create new table
5. Insert values into table

### RESOURCES

---

1. Results from Exercise 2
2. sqlite3 package documentation and examples

---

## EXERCISE 4.: EXECUTE SQLITE QUERY

---

14

### KEY OBJECTIVES

---

Execute query using python to determine date and hour with most timestamps

### AGENDA

---

*10 mins*

1. Write query
2. Use sqlite package to execute query

### DELIVERABLE

---

Peak date and hour

### RESOURCES

---

1. Results from Exercise 3
2. sqlite3 package documentation and examples