

Technology Fundamentals for Business Analytics

Jason Kuruzovich

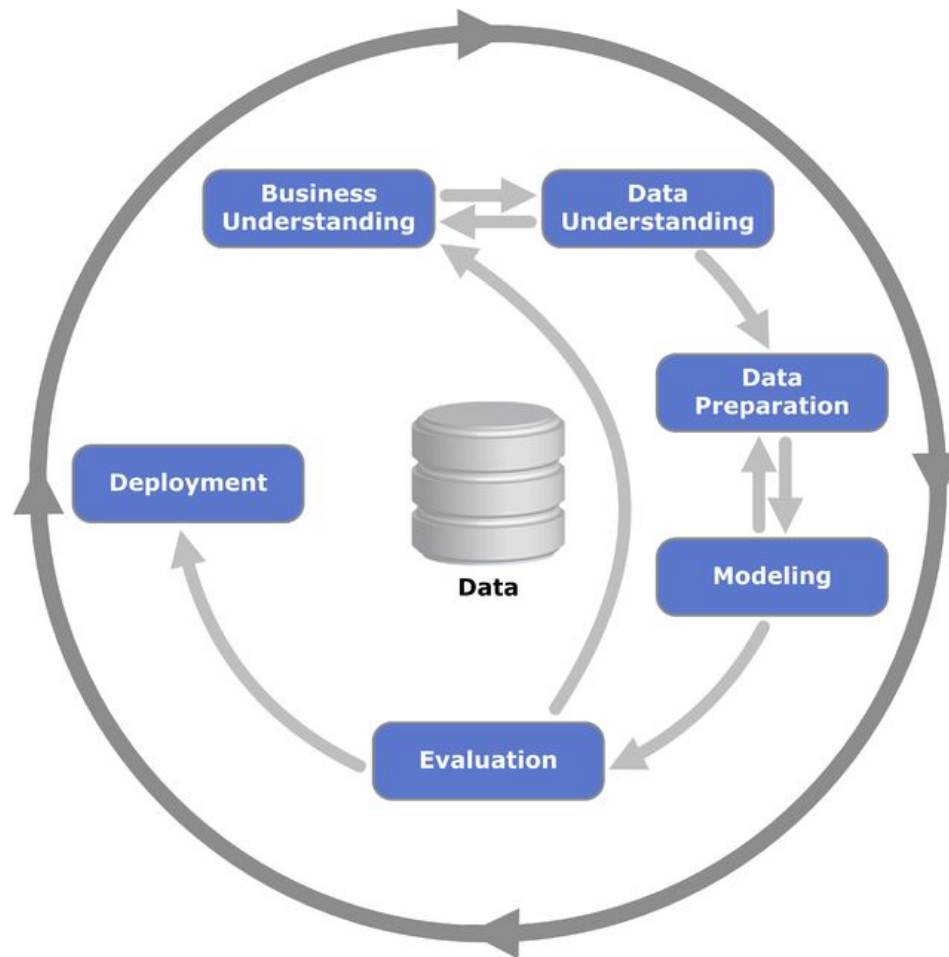
Agenda

- Announcement
- Review the lab
- Understanding data
- IDE/Packages
- Lab 2

Lab 1

- Github Desktop
- Pull Requests
- Emails – Please use Piazza for all communications [including private conversations]
- Piazza – Please start new note when posting an unrelated problem

THE CRISP-DM PROCESS MODEL



Cross Industry Standard Process for Data Mining

https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

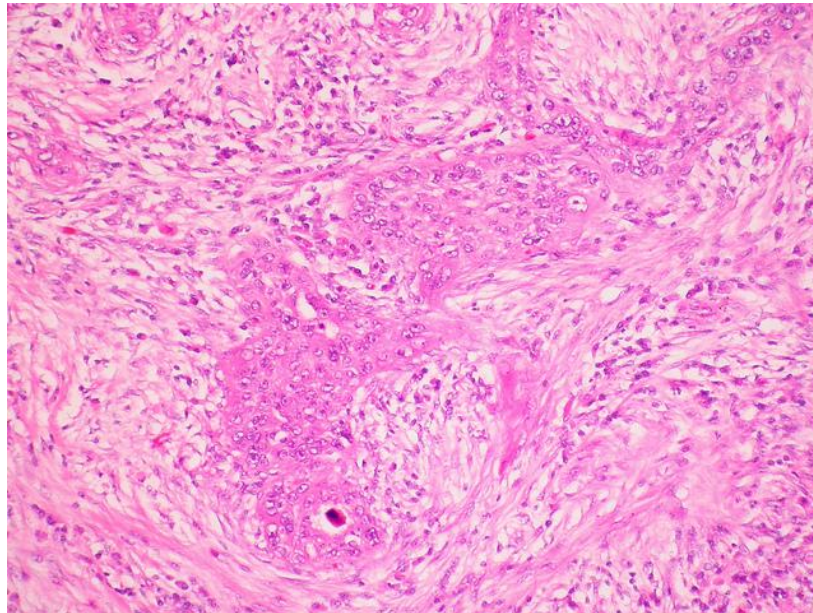
Let's Talk About Data

Consider a Single Variable

Cancer = TRUE

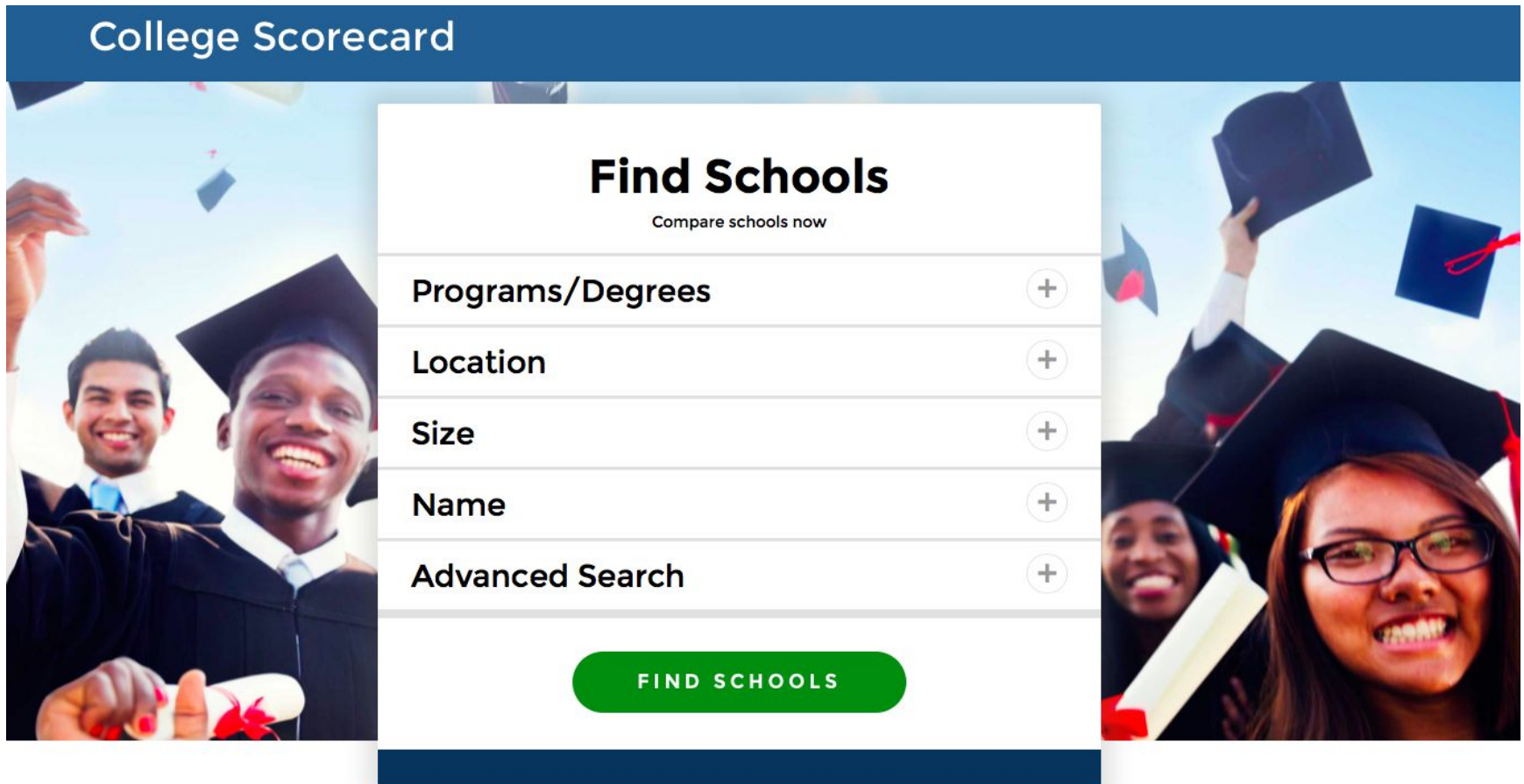
Cancer = Mucoepidermoid carcinoma

Cancer =



Data In the News...

Open Data For Students



Information for individuals searching for education data.

<https://youtu.be/hgqG6NuQRIU>

Why would the government care
about citizen's access to data?

DATA.GOV



[DATA](#) [TOPICS ▾](#) [IMPACT](#) [APPLICATIONS](#) [DEVELOPERS](#) [CONTACT](#)

The home of the U.S. Government's open data

Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more.

GET STARTED

SEARCH OVER [164,199 DATASETS](#)



<http://www.socrata.com>

Open Data Platforms

- Open data platforms like <http://www.socrata.com> provide ease of interacting with data via API or download

College Scorecard

- With college scorecard, Government has made the shift from providing raw data to providing unbiased decision making tools (Information Apps)
- The licenses is also public domain, meaning anyone can fork and use the code!!!!
- <http://jkuruzovich.github.io/newedinc/>

Pair Exercise

Find an available dataset

- <http://www.data.gov> (free and open)
- <https://www.quandl.com> (free and paid)

On Piazza (Class 2 data)

1. In what format is the data provided?
2. What information is there that would lead to data understanding?
3. Link to data

Data Types and Structures

Files/HDFS

Key-value Database

Relational Database

Document Database

Triplestore

Graph Database

Application (API)

Intermediate
Structured
Data

Delimited File

JSON File

XML File

R

{Vector, MATRIX, DATA
FRAME}

Python

{JSON, List, NumPy {Array}

Pandas {Data Frame}}

SAS

{Vector, Matrix Data
Frame}

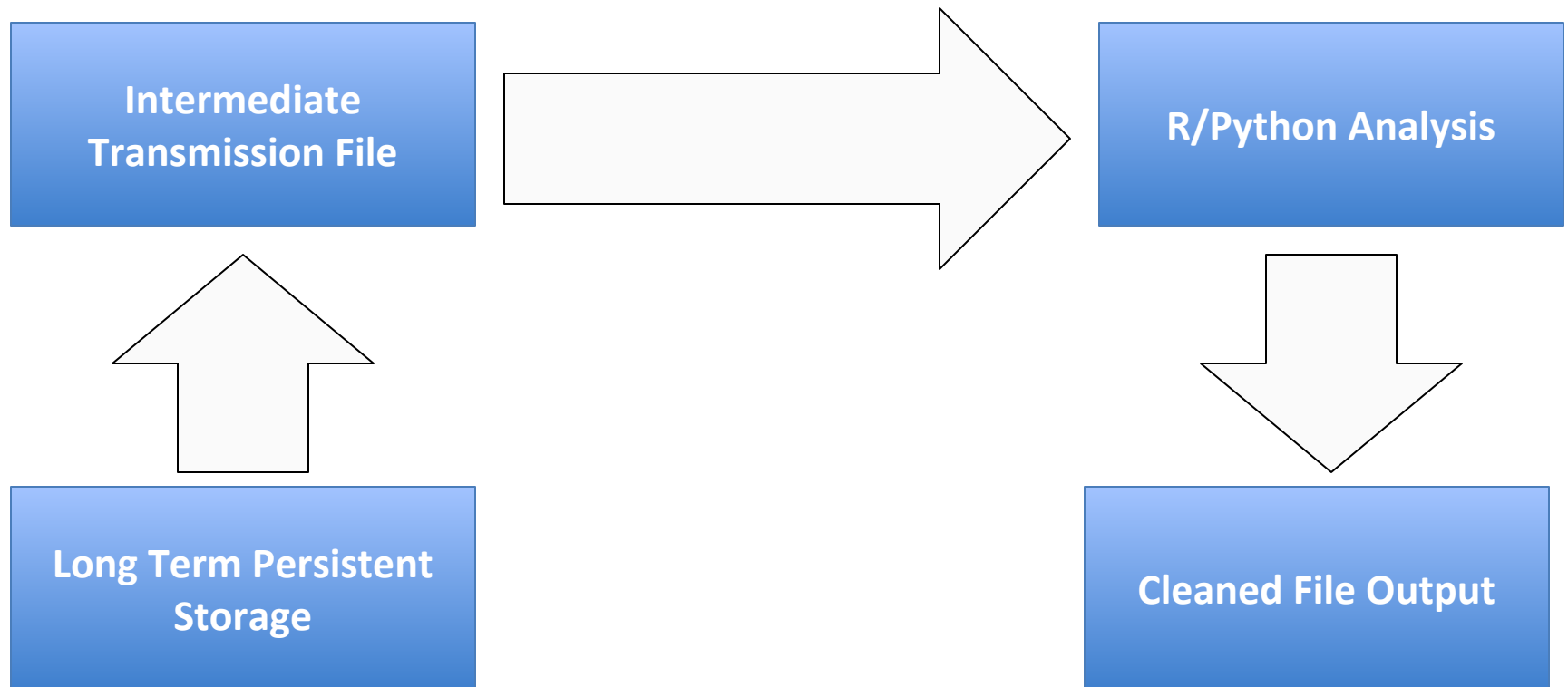
R/PYTHON/SAS PACKAGES FACILITATE CONNECTIONS BETWEEN STAGES

**Long Term
Persistent Storage**

**Intermediate
Transmission
File**

**In Memory
Storage for
Analysis**

Typical Analysis

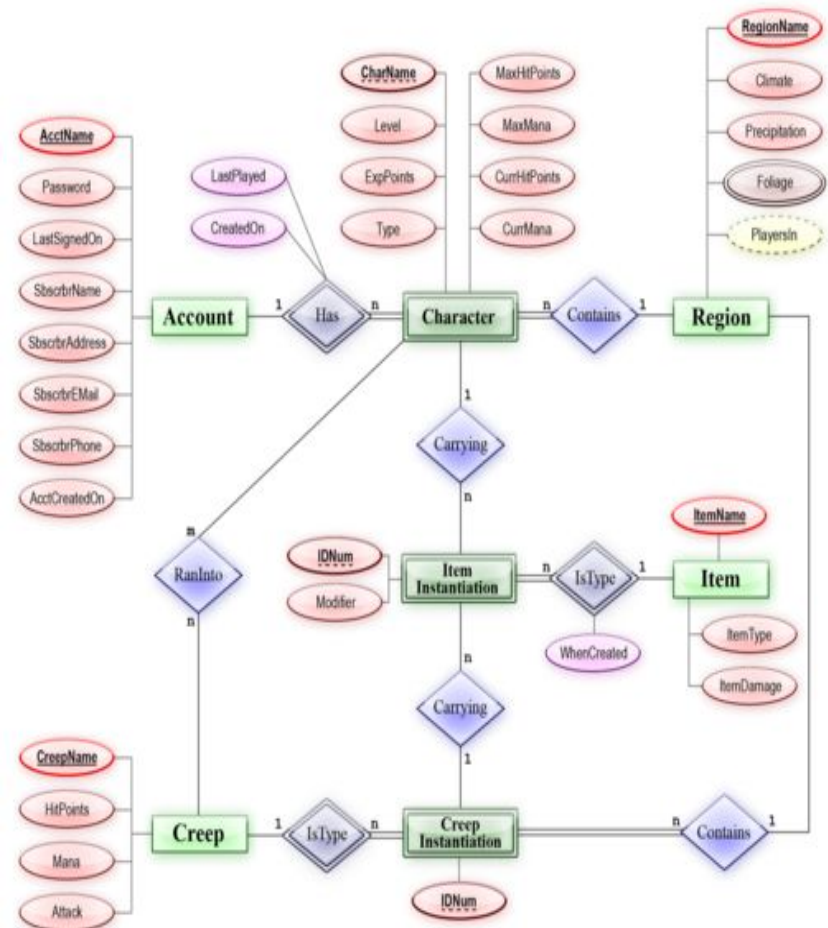


Long Term Persistent Storage

- Designed to facilitate persistent storage and retrieval of data
- Employ functionality to select only desired data and even do some processing of data

Long Term Persistent Storage

- Relational Databases normalize data, storing in separate tables related information accessed via keys

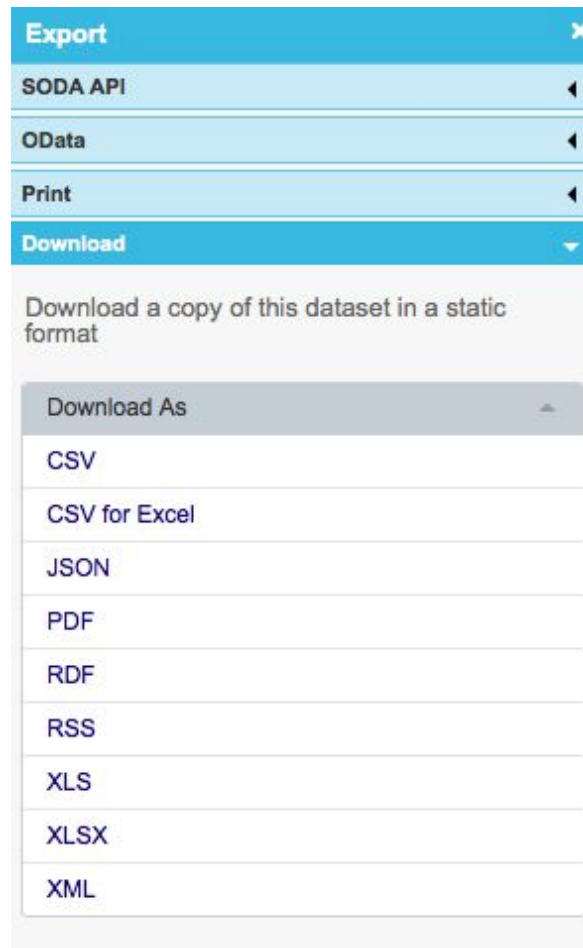


Document Oriented Databases

- MongoDB is one of several document oriented databases that provides an object based way of storing data

```
{
  "firstName": "Jane",
  "lastName": "Doe",
  "address": {
    "streetAddress": "1 North Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021"
  },
  "phoneNumbers": [
    {
      "type": "home",
      "number": "555-111-2222"
    },
    {
      "type": "mobile",
      "number": "555-121-1212"
    }
  ],
}
```

Intermediate Storage



<https://data.cms.gov/Medicare/Inpatient-Prospective-Payment-System-IPPS-Provider/97k6-zzx3>

Intermediate Storage: Sample CSV

From the Titanic Kaggle Assignment

<https://www.kaggle.com/c/titanic>

```
PassengerId,Survived,Pclass,Name,Sex,Age,SibSp,Parch,Ticket,Fare,Cabin,Embarked
1,0,3,"Braund, Mr. Owen Harris",male,22,1,0,A/5 21171,7.25,,S
2,1,1,"Cumings, Mrs. John Bradley (Florence Briggs Thayer)",female,38,1,0,PC 17599,71.2833,C85,C
3,1,3,"Heikkinen, Miss. Laina",female,26,0,0,STON/O2. 3101282,7.925,,S
4,1,1,"Futrelle, Mrs. Jacques Heath (Lily May Peel)",female,35,1,0,113803,53.1,C123,S
5,0,3,"Allen, Mr. William Henry",male,35,0,0,373450,8.05,,S
6,0,3,"Moran, Mr. James",male,,0,0,330877,8.4583,,Q
7,0,1,"McCarthy, Mr. Timothy J",male,54,0,0,17463,51.8625,E46,S
8,0,3,"Palsson, Master. Gosta Leonard",male,2,3,1,349909,21.075,,S
9,1,3,"Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)",female,27,0,2,347742,11.1333,,S
10,1,2,"Nasser, Mrs. Nicholas (Adele Achem)",female,14,1,0,237736,30.0708,,C
```

In Memory Storage

- Python, R, and SAS each have different method of operating and storing data for analysis

IDE/Packages and R/Python/SAS

Working with RSTUDIO

The image shows the RStudio desktop application. The main editor window contains an R script with SPARQL queries and R code to execute them. The right-hand pane is divided into 'Workspace' and 'History' tabs, showing loaded data objects. The bottom-left pane is the 'Console', displaying startup messages and help text. The bottom-right pane is the 'Browser/Plot/Help/Packages' pane, which is currently empty.

Scripts

```
1 # Define the data.gov endpoint
2 endpoint <- "http://services.data.gov/sparql"
3
4 # create query statement
5 query <-
6   "PREFIX dgp1187: <http://data-gov.tw.rpi.edu/vocab/p/1187/>
7   SELECT ?ye ?fi ?ac
8   WHERE {
9     ?s dgp1187:year ?ye .
10    ?s dgp1187:fires ?fi .
11    ?s dgp1187:acres ?ac .
12   }"
13
14 # Step 2 - Use SPARQL package to submit query and save results to a data frame
15 qd <- SPARQL(endpoint,query)
16 df <- qd$results
17
18 # Step 3 - Prep for graphing
```

R Objects

Data	
goog	26 obs. of 8 variables
goog_sub	25 obs. of 8 variables
Values	
m_full	lm[13]
m_step	lm[13]

Console

Natural language support but running in an English locale

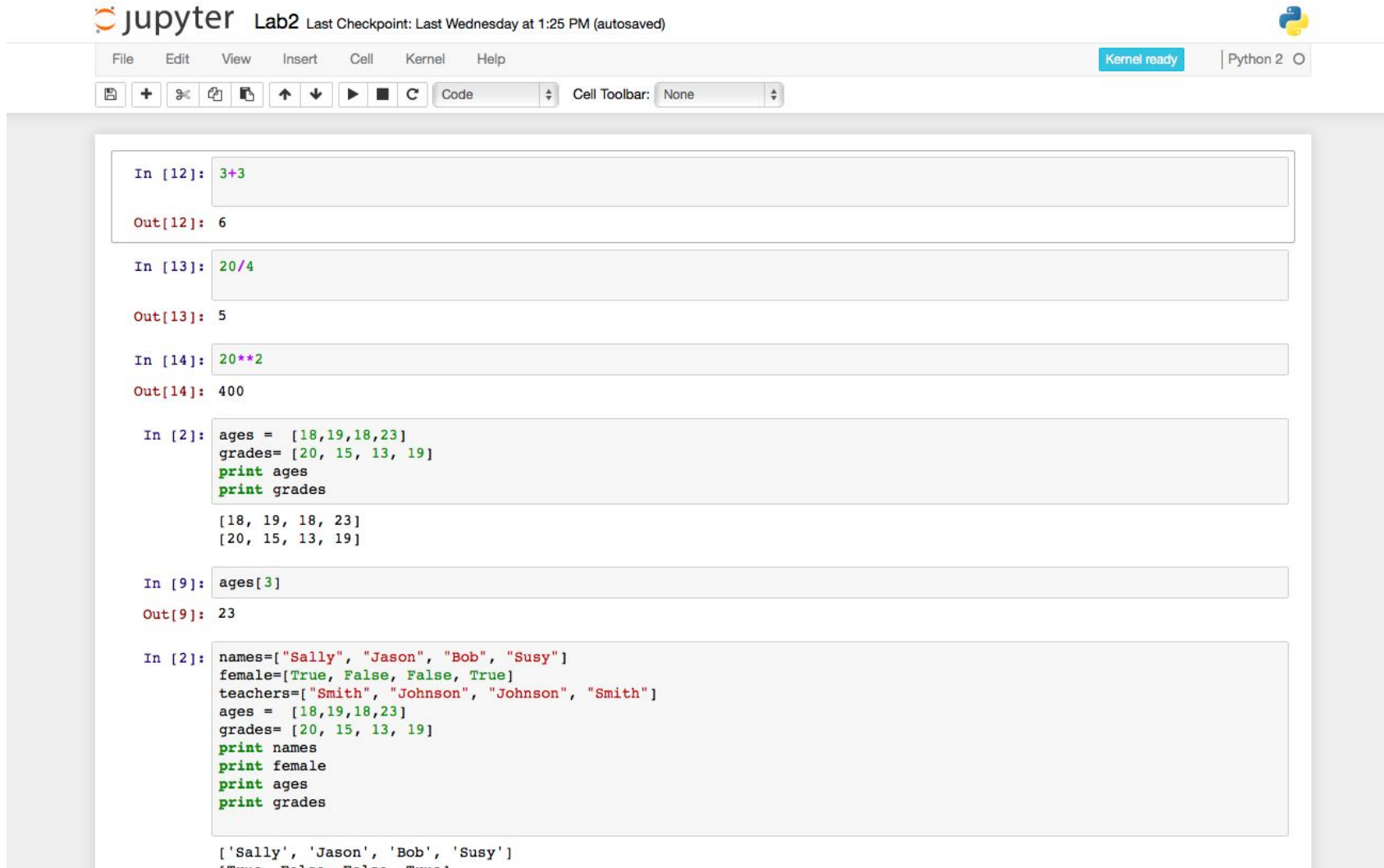
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

Browser/Plot/Help/Packages

Working with IPython Notebooks



The screenshot displays the JupyterLab interface. At the top, the header shows the Jupyter logo, the text "Lab2", and the status "Last Checkpoint: Last Wednesday at 1:25 PM (autosaved)". On the right, there is a Python logo and a "Kernel ready" status indicator. Below the header is a menu bar with options: File, Edit, View, Insert, Cell, Kernel, and Help. A toolbar follows, containing icons for saving, adding cells, undo, redo, and running code, along with a "Code" dropdown and a "Cell Toolbar" set to "None".

The main workspace contains several code cells:

- Cell 1:** Input: `In [12]: 3+3`; Output: `Out[12]: 6`
- Cell 2:** Input: `In [13]: 20/4`; Output: `Out[13]: 5`
- Cell 3:** Input: `In [14]: 20**2`; Output: `Out[14]: 400`
- Cell 4:** Input:

```
In [2]: ages = [18,19,18,23]
grades= [20, 15, 13, 19]
print ages
print grades
```

; Output:

```
[18, 19, 18, 23]
[20, 15, 13, 19]
```
- Cell 5:** Input: `In [9]: ages[3]`; Output: `Out[9]: 23`
- Cell 6:** Input:

```
In [2]: names=["Sally", "Jason", "Bob", "Susy"]
female=[True, False, False, True]
teachers=["Smith", "Johnson", "Johnson", "Smith"]
ages = [18,19,18,23]
grades= [20, 15, 13, 19]
print names
print female
print ages
print grades
```

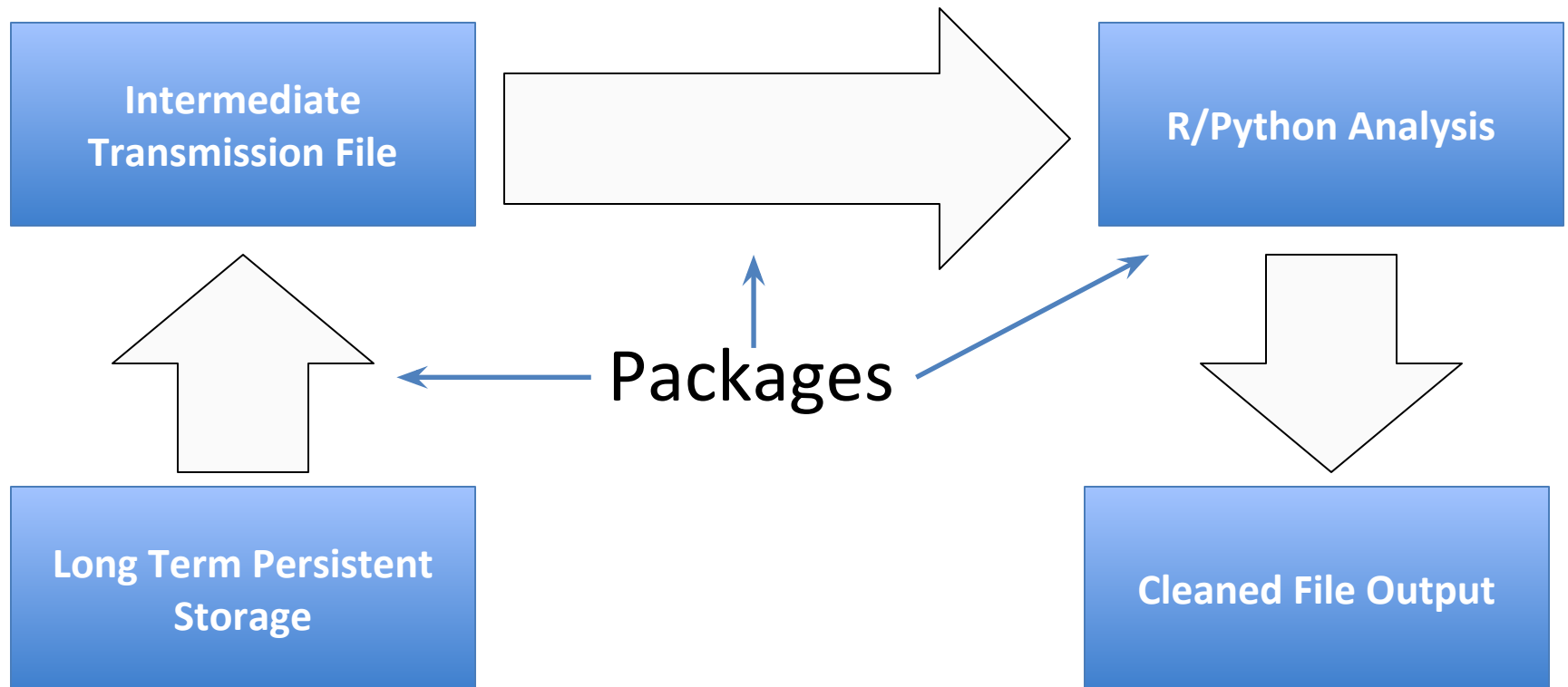
; Output:

```
['Sally', 'Jason', 'Bob', 'Susy']
[True, False, False, True]
```


Packages

- Packages allow access to additional functionality not contained in the original language

Packages



R and Packages

- Packages are collections of **R** functions, data, and compiled code from <https://cran.r-project.org>
- External packages only have to be installed once, but they have to be loaded each time they are used
- You can create your own packages and contribute them back to the ecosystem
- Top packages for R <http://www.r-statistics.com/2013/06/top-100-r-packages-for-2013-jan-may/>

R and Packages

#This installs the MySQL driver it only needs to be run once.

#Install latest binary for the operating system

```
install.packages('RMySQL')
```

#Install the package from source code

```
install.packages('RMySQL', type="source")
```

#This loads the library for use by subsequent lines of the script.

```
library('RMySQL')
```

Python Packages

- Python package manager provides access to many python packages
- <https://pypi.python.org/pypi/pip>
- Installed via command line tool `pip`
- Top packages for data science
 - NumPy <http://www.numpy.org>
 - Pandas <http://pandas.pydata.org>
 - Matplotlib <http://matplotlib.org>

Python Packages

- Python package manager provides access to many python packages
- <https://pypi.python.org/pypi/pip>
- Installed via command line tool `pip`
- Top packages for data science
 - NumPy <http://www.numpy.org>
 - Pandas <http://pandas.pydata.org>
 - Matplotlib <http://matplotlib.org>

Installing packages

This is a Bash magic command.

```
In: []      %%bash  
          pip install twitter pandas nltk
```

```
In: []      import io  
          import json  
          import twitter
```

Python PIP

- Alternate method is *vagrant ssh* into the machine and then enter *pip package name*

Lab 2