FIT1043 Introduction to Data Science

Module 4: Data Resources, Processes, Standards and Tools

Lecture 7

Monash University

Assignment 2

- Due 16th September 2018
- Python
- No zip file submission
- Visualization and investigation

Discussion: Data Wrangling Examples

“How we found the worst place to park in New York City” is examples, and a discussion of the complexities of getting data out of New York City:

Danger spots for cycles: *NYPD crash data* obtained by daily download of PDF files followed by (non-trivial) extraction
NB. they now have Excel data to ease the work!

Dirty waterways: *fecal coliform measurements on waterways* from Department of Environmental Protection’s website; extracted from Excel sheets per site; each in a different format

Faulty road markings: parking tickets for fire-hydrants by location from *NYC Open Data portal* need to normalize the addresses supplied

Unit Schedule: Modules

Module	Week	Content
1.	1	Overview and look at projects
	2	(Job) roles, and the impact
2.	3	Data business models / application areas
3.	4	Characterising data and "big" data
	5	Data sources and case studies
4.	6	**Resources and standards**
	7	**Resources case studies**
5.	8	Data analysis theory
	9	Regression and decision trees
	10	Data analysis process
6.	11	Issues in data management
	12	GUEST SPEAKER & EXAM INFO

Standards and Issues
(ePub section 4.5)

- some standards
- open data and open source software
- APIs and SaaS

Some Standards

Semi-Structured Data

Semi-structured data is data that is presented in XML or JSON:

- see some examples *here*
- Note YAML (Yet Another Markup Language), which is just an indentation (easier to read) version of JSON
- standard libraries for reading/writing/manipulating semi-structured data exist in Python, Perl, Java
- don’t need to know all the details of XML (and related Schema languages)

  many good online tutorials, *e.g.* *W3schools.com*

Model Language

PMML ::= Predictive Model Markup Language

PMML provides a standard language for describing a (predictive) model that can be passed between analytic software (e.g. from R to SAS).

- *PMML: An Open Standard for Sharing Models*
- A list of products working with PMML is the *PMML Powered page* on DMG site.

MARS Question

Which of the following statement is FALSE?

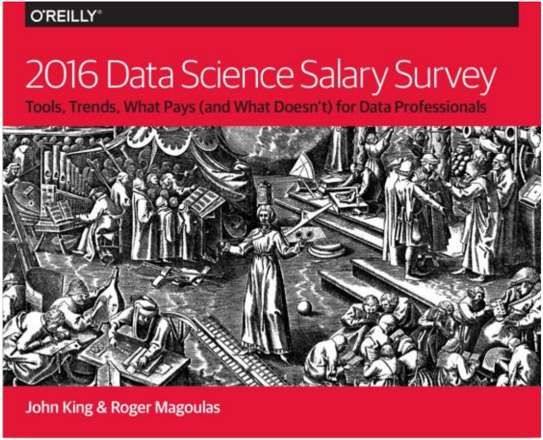
A. PMML is a standard language for describing a predictive model

B. Semi-structured data is data that is presented in XML and JSON

C. JSON is easier to read than YAML

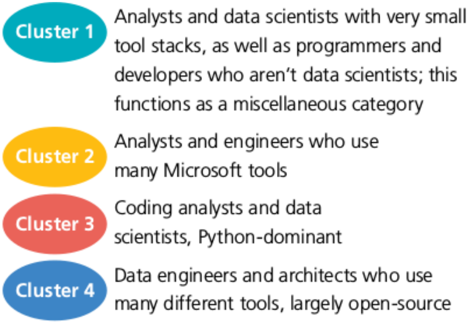
Open data and open source software

Software Usage Survey


O'REILLY
2016 Data Science Salary Survey
Tools, Trends, What Pays (and What Doesn't) for Data Professionals
John King & Roger Magoulas

2016 Data Science Salary Survey

Survey: Clusters amongst the Respondents



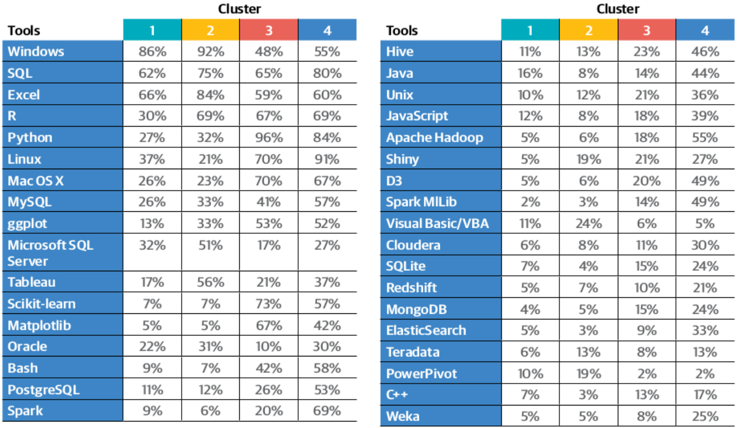
Cluster 1 Analysts and data scientists with very small tool stacks, as well as programmers and developers who aren't data scientists; this functions as a miscellaneous category

Cluster 2 Analysts and engineers who use many Microsoft tools

Cluster 3 Coding analysts and data scientists, Python-dominant

Cluster 4 Data engineers and architects who use many different tools, largely open-source

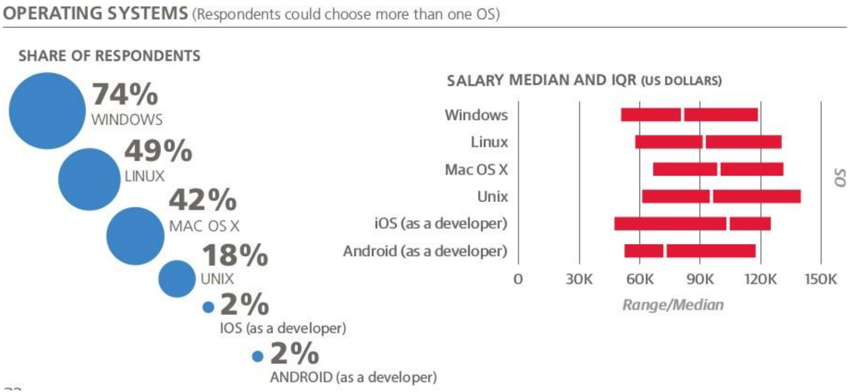
Survey: Commonly Used Software



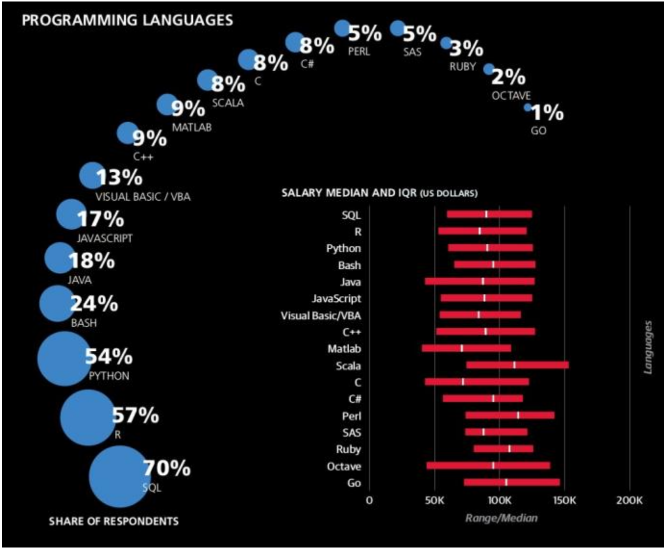
	Cluster			
Tools	1	2	3	4
Windows	86%	92%	48%	55%
SQL	62%	75%	65%	80%
Excel	66%	84%	59%	60%
R	30%	69%	67%	69%
Python	27%	32%	96%	84%
Linux	37%	21%	70%	91%
Mac OS X	26%	23%	70%	67%
MySQL	26%	33%	41%	57%
ggplot	13%	33%	53%	52%
Microsoft SQL Server	32%	51%	17%	27%
Tableau	17%	56%	21%	37%
Scikit-learn	7%	7%	73%	57%
Matplotlib	5%	5%	67%	42%
Oracle	22%	31%	10%	30%
Bash	9%	7%	42%	58%
PostgreSQL	11%	12%	26%	53%
Spark	9%	6%	20%	69%

	Cluster			
Tools	1	2	3	4
Hive	11%	13%	23%	46%
Java	16%	8%	14%	44%
Unix	10%	12%	21%	36%
JavaScript	12%	8%	18%	39%
Apache Hadoop	5%	6%	18%	55%
Shiny	5%	19%	21%	27%
D3	5%	6%	20%	49%
Spark MlLib	2%	3%	14%	49%
Visual Basic/VBA	11%	24%	6%	5%
Cloudera	6%	8%	11%	30%
SQLite	7%	4%	15%	24%
Redshift	5%	7%	10%	21%
MongoDB	4%	5%	15%	24%
ElasticSearch	5%	3%	9%	33%
Teradata	6%	13%	8%	13%
PowerPivot	10%	19%	2%	2%
C++	7%	3%	13%	17%
Weka	5%	5%	8%	25%

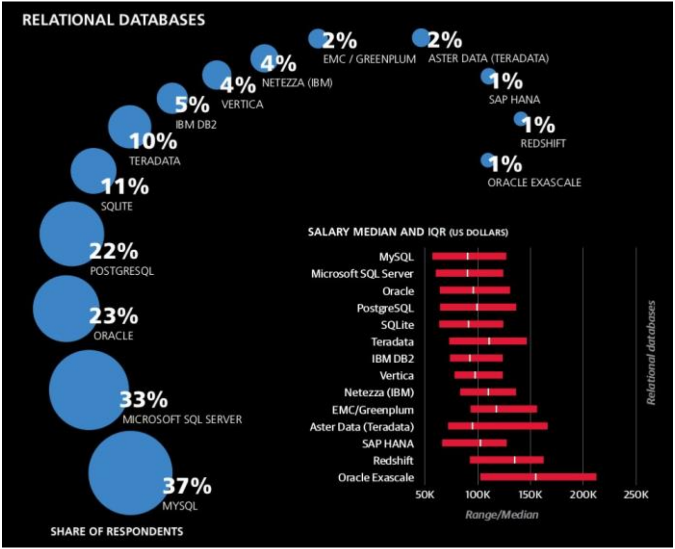
Survey: Operating Systems



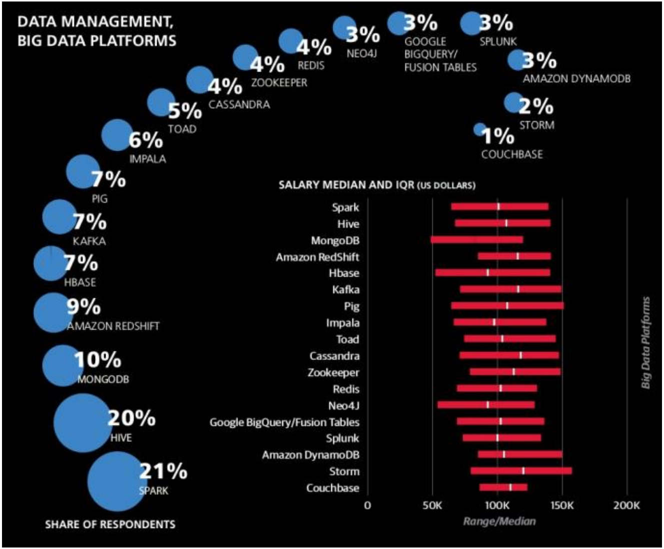
Survey: Programming Languages



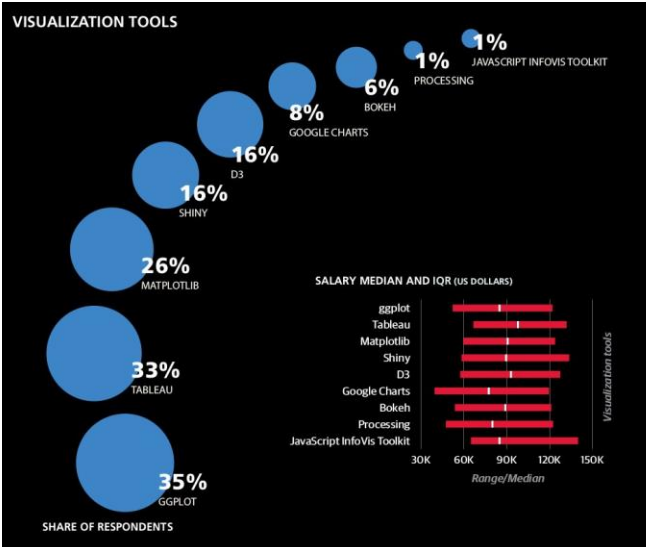
Survey: Relational Databases



Survey: Management and Big Data



Survey: Visualization



Open Source Software Awards

Here's how you learn about which tools are important!

BOSSIE is **B**est **O**pen **S**ource **S**oftware awards, held in September.

- *BOSSIE awards 2015 for Big Data* and *BOSSIE awards 2016 for Big Data*
- BOSSIE awards 2017 for *machine learning and deep learning tools* and for *databases and analytics tools*

Open Source Software Awards, cont.

2015: big data tools, Spark and “elastic” processing, scalable ML and databases, stream/real-time processing (ML, search, analysis, storage, time-series), security

2016: big data tools, pipelines, TensorFlow, distributed IR (Solr), NoSQL analytics, stream analytics, graph database

2017: big data and analytics tools, GPU acceleration, real-time SQL, more Spark, Solr, R, graph databases

2017: ML tools, deep learning, scalable prediction, Python, gradient boosting, TensorFlow

Popular Open Source Projects

Let's have a look at what all these Open Source Projects doing

1. *Apache Hadoop Distributed File System (HDFS)*
2. *Apache Hadoop YARN*
3. *Apache Spark*
4. *Apache Cassandra* (distributed NoSQL, wide-column store)
5. *Apache HBase* (distributed NoSQL, wide-column store)
6. *Apache Hive* (distributed SQL)
7. *Apache Mahout* (distributed linear algebra with GPU)
8. *Apache Pig* (data flow and data analysis on top of Hadoop)
9. *Apache Storm* (distributed real-time computation)
10. *Apache Tez* (dataflow for Hive and Pig)

APIs and SaaS

REST API Terminology

API: **A**pplication **P**rogrammer **I**nterface

Routines providing programatic access to an application.

REST: **RE**presentational **S**tate **T**ransfer

a stateless API usually running over HTTP

Watch a simple introduction to REST-based APIs in this video: *REST API concepts and examples* by WebConcepts

SaaS: **S**oftware **a**s **a** **S**ervice

The provisioning of software in a Web browser and/or via an API over the Web as a subscription service.

MARS Question

Name a popular data/information API.

Example APIs

Many companies are exposing their data **and their website functionality** as APIs for others to make use of:

- *Facebook API*
- *Twitter API*
- *LinkedIn API*
- *Google Maps API*
- *Youtube API*
- *Amazon Advertising API*
- *TripAdvisor API*

SaaS Examples

- Email systems (Google, Microsoft Office365),
- File sharing systems(Dropbox, Box, Microsoft One drive, Google drive ..)
- Business systems (Salesforce, Servicenow, ..)

Why SaaS

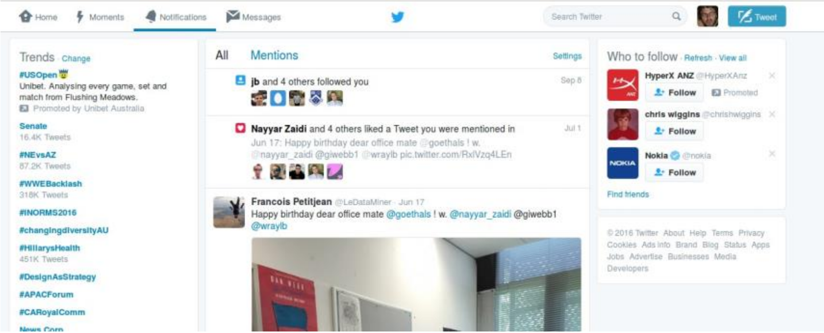
- Pay as you go
- Scale up/down
- Low maintenance
- Performance, better infrastructure

Disadvantage: data privacy

Case Studies of Data

(ePub section 4.8)

Twitter



Trends · Change

#USOpen
Unibet. Analysing every game, set and match from Flushing Meadows.
Promoted by Unibet Australia

Senate
16.4K Tweets

#NEvsAZ
87.2K Tweets

#WWEBacklash
318K Tweets

#INORMS2016

#changingdiversityAU

#HillarysHealth
451K Tweets

#DesignAsStrategy

#APACForum

#CARoyalComm

News Corp

jb and 4 others followed you — Sep 8

Nayyar Zaidi and 4 others liked a Tweet you were mentioned in — Jul 1
Jun 17: Happy birthday dear office mate @goethals ! w. @nayyar_zaidi @giwebb1 @wraylb pic.twitter.com/RxlVzq4LEn

Francois Petitjean @LeDataMiner · Jun 17
Happy birthday dear office mate @goethals ! w. @nayyar_zaidi @giwebb1 @wraylb

Who to follow · Refresh · View all

HyperX ANZ @HyperXAnz — Promoted

chris wiggins @chrishwiggins

Nokia @nokia

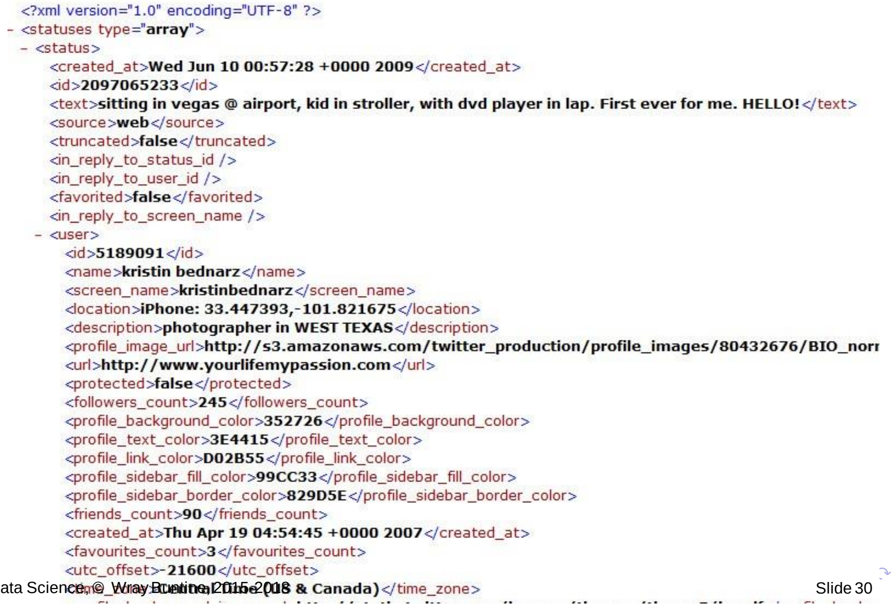
Find friends

© 2016 Twitter About Help Terms Privacy Cookies Ads info Brand Blog Status Apps Jobs Advertise Businesses Media Developers

Twitter is the most famous microblogging platform

- with big corporate use
- contains lots of metadata: information about users, their follower network, locations, hashtags, emojis+emoticons, ...

Sample Twitter XML Data



```
   <?xml version="1.0" encoding="UTF-8" ?>
 - <statuses type="array">
  - <status>
      <created_at>Wed Jun 10 00:57:28 +0000 2009</created_at>
      <id>2097065233</id>
      <text>sitting in vegas @ airport, kid in stroller, with dvd player in lap. First ever for me. HELLO!</text>
      <source>web</source>
      <truncated>false</truncated>
      <in_reply_to_status_id />
      <in_reply_to_user_id />
      <favorited>false</favorited>
      <in_reply_to_screen_name />
    - <user>
        <id>5189091</id>
        <name>kristin bednarz</name>
        <screen_name>kristinbednarz</screen_name>
        <location>iPhone: 33.447393,-101.821675</location>
        <description>photographer in WEST TEXAS</description>
        <profile_image_url>http://s3.amazonaws.com/twitter_production/profile_images/80432676/BIO_norr
        <url>http://www.yourlifemypassion.com</url>
        <protected>false</protected>
        <followers_count>245</followers_count>
        <profile_background_color>352726</profile_background_color>
        <profile_text_color>3E4415</profile_text_color>
        <profile_link_color>D02B55</profile_link_color>
        <profile_sidebar_fill_color>99CC33</profile_sidebar_fill_color>
        <profile_sidebar_border_color>829D5E</profile_sidebar_border_color>
        <friends_count>90</friends_count>
        <created_at>Thu Apr 19 04:54:45 +0000 2007</created_at>
        <favourites_count>3</favourites_count>
        <utc_offset>-21600</utc_offset>
        <time_zone>Central Time (US & Canada)</time_zone>
```

Twitter Developer API

See *Twitter's developer platform*

- library interfaces for Java, C++, Javascript, Python, Perl, PHP, ...
- allows other applications to manage Twitter data for users
- extensive developer policy

Unit Schedule: Next Week

Module	**Week**	**Content**
1.	1	Overview and look at projects
	2	(Job) roles, and the impact
2.	3	Data business models / application areas
3.	4	Characterising data and "big" data
	5	Data sources and case studies
4.	6	Resources and standards
	7	Resources case studies
5.	8	**Data analysis theory**
	9	**Regression and decision trees**
	10	**Data analysis process**
6.	11	Issues in data management
	12	GUEST SPEAKER & EXAM INFO