

FIT1043 Introduction to Data Science

Module 4: Data Resources, Processes, Standards and Tools

Lecture 8

Monash University

Discussion

In the tutorial you used three different tools for data wrangling:

- ▶ SAS
 - ▶ general purpose Data Analytics
 - ▶ strange syntax!
 - ▶ very widely used commercial product
- ▶ DataWrangler
 - ▶ specialised Data Wrangling tool
 - ▶ intuitive Graphical User Interface (GUI)
 - ▶ no coding required!
- ▶ Python
 - ▶ general purpose open-source programming language
 - ▶ contains packages (Pandas) for manipulating data

Note that there are many other tools we could have used

- ▶ R, Matlab, Java, SPSS.

Unit Schedule: This Week

Module	Week	Content
1.	1	overview and look at projects
	2	(job) roles, and the impact
2.	3	data business models
	4	application areas and case studies
3.	5	characterising data and "big" data
	6	data sources and case studies
4.	7	resources and standards
	8	resources case studies
5.	9	data analysis theory
	10	data analysis process
6.	11	issues in data management
	12	data management frameworks

Standards and Issues

(ePub section 4.5)

more on standards and issues

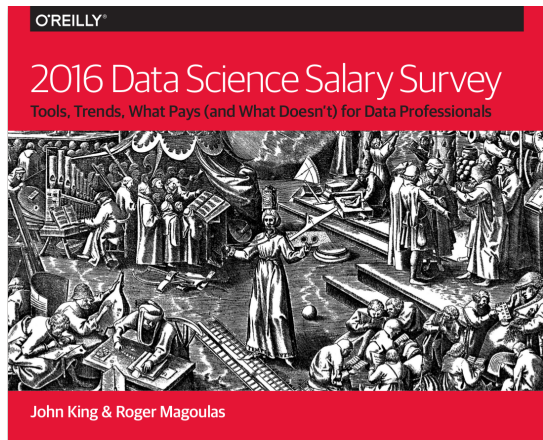
- ▶ some standards
 - ▶ some standards for semi-structured data, data science process and predictive models
- ▶ open data and open source software
 - ▶ critical infrastructure and tools
- ▶ APIs and SaaS
 - ▶ think Web 3.0

Standards and Issues

Open data and open source software

critical infrastructure and tools

Software Usage Survey



[2016 Data Science Salary Survey](#) (DSSS)

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

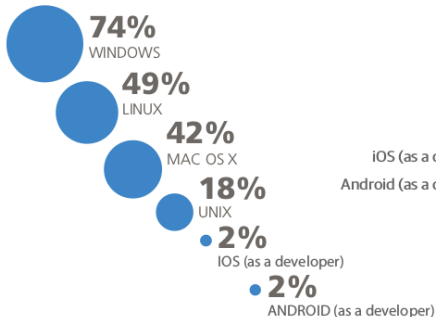
Tools	Cluster			
	1	2	3	4
Windows	86%	92%	48%	55%
SQL	62%	75%	65%	80%
Excel	66%	84%	59%	60%
R	30%	69%	67%	69%
Python	27%	32%	96%	84%
Linux	37%	21%	70%	91%
Mac OS X	26%	23%	70%	67%
MySQL	26%	33%	41%	57%
ggplot	13%	33%	53%	52%
Microsoft SQL Server	32%	51%	17%	27%
Tableau	17%	56%	21%	37%
Scikit-learn	7%	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%

Tools	Cluster			
	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%

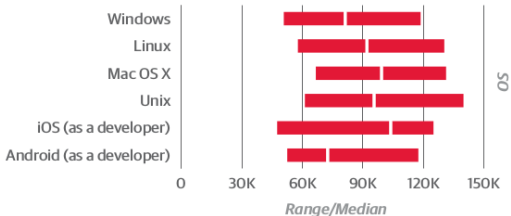
Survey: Operating Systems

OPERATING SYSTEMS (Respondents could choose more than one OS)

SHARE OF RESPONDENTS



SALARY MEDIAN AND IQR (US DOLLARS)



Survey: Number of Tools used

(from 2014 survey)

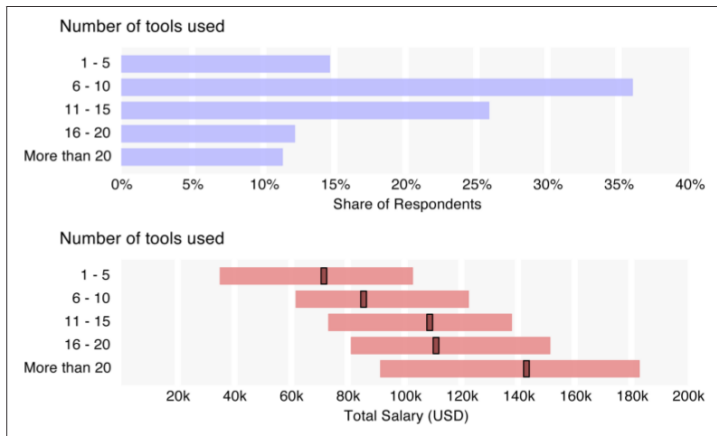
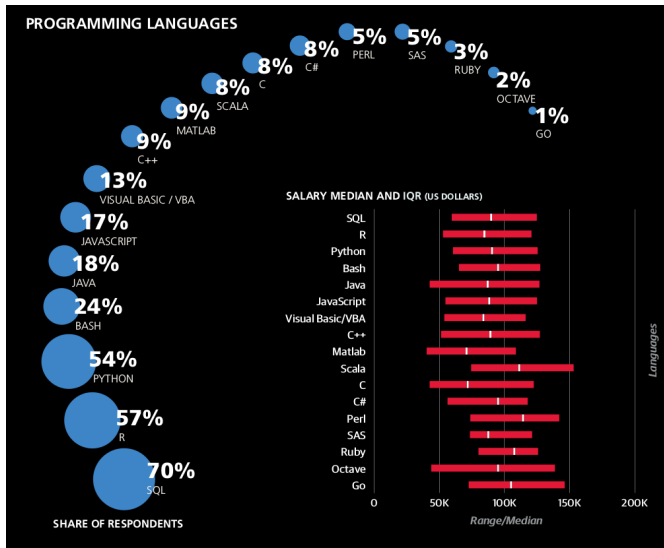
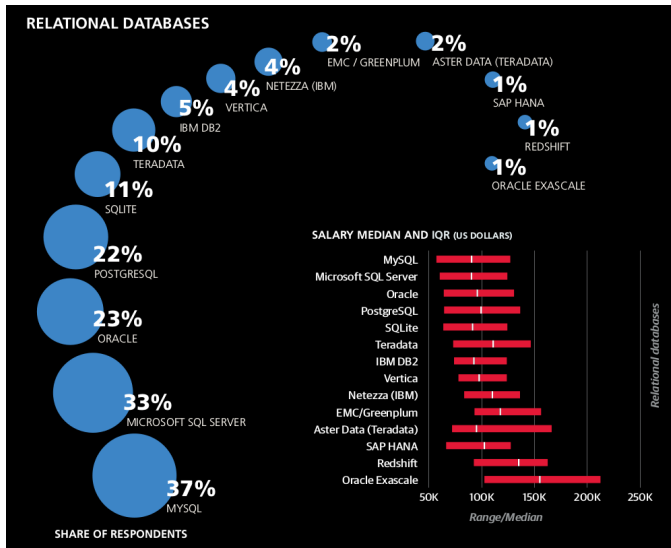


Figure 1-13. Number of tools used

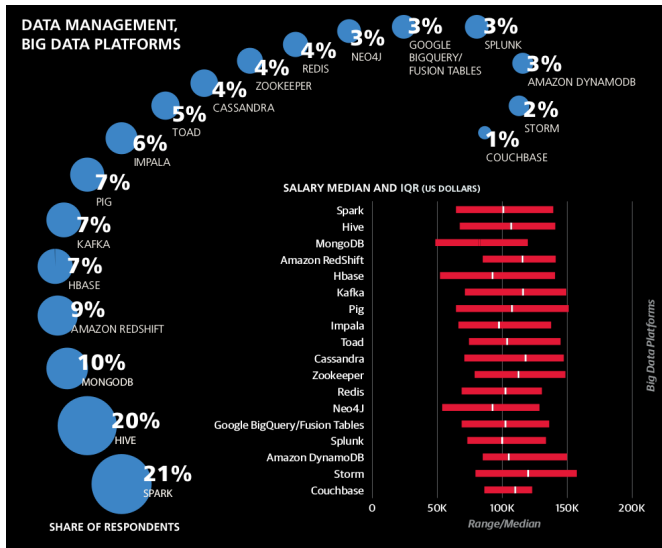
Survey: Programming Languages



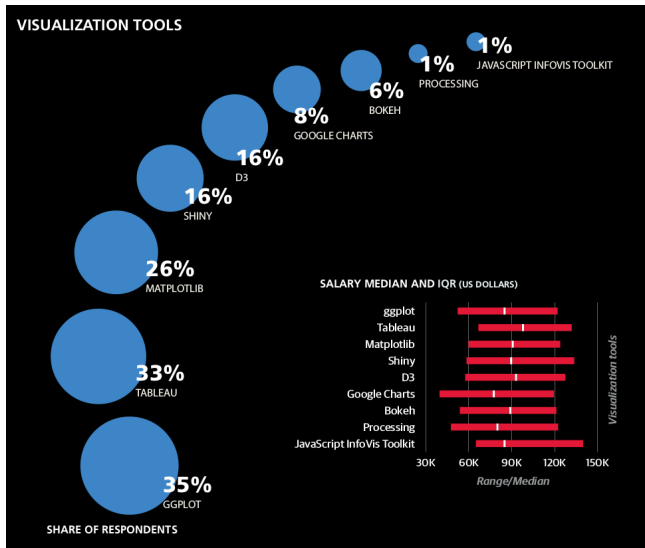
Survey: Relational Databases



Survey: Management and Big Data



Survey: Visualization



Open Source Software: Examples

Prize winning Open Source platforms for managing big data:

- ▶ [*BOSSIE Awards for Big Data 2015*](#)
- ▶ Similar awards also for applications:
 - ▶ [*BOSSIE Awards for Applications 2015*](#)

Many of the state-of-the-art platforms are integrated in:

- ▶ [*Hortonworks Data Platform*](#)

Popular Open Source Projects

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

1. [*Hadoop Distributed File System \(HDFS\)*](#)
2. [*Apache Hadoop YARN*](#)
3. [*Apache Cassandra*](#)
4. [*Apache HBase*](#)
5. [*Apache Hive*](#)
6. [*Apache Mahout*](#)
7. [*Apache Pig*](#)
8. [*Apache Spark*](#)
9. [*Apache Storm*](#)
10. [*Apache Tez*](#)

Standards and Issues

APIs and SaaS

think Web 3.0

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.

The API Economy

Companies provide functionality via APIs so that others can make use of their data and services:

- ▶ *The Application Economy: A New Model for IT* (CISCO)
- ▶ *ProgrammableWeb API Category: Data*
- ▶ *Top 30 Predictive Analytics 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](#)

Case Studies of Data and Standards

(ePub section 4.8)

look at some examples of standardised data collections

Twitter

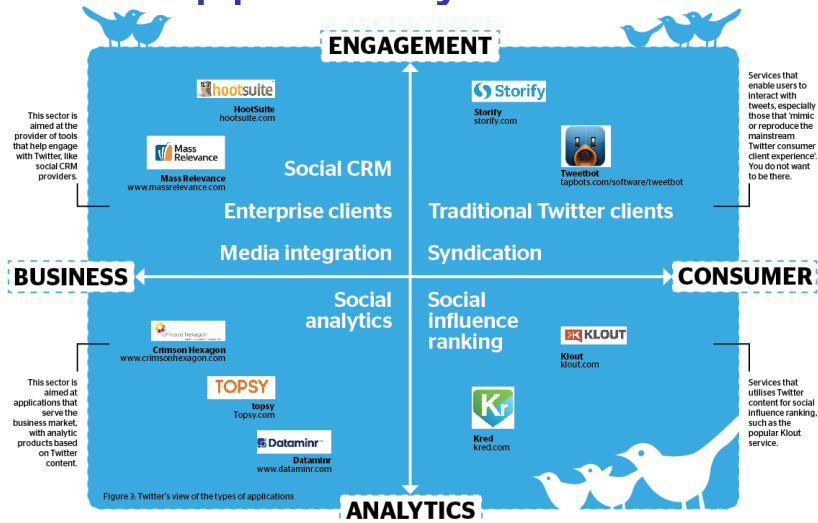


- ▶ microblogging with attached media
- ▶ big corporate use
- ▶ also has information about users, their follower network, locations, hashtags, emojis+emoticons, ...

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<br>
  <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 App Ecosystem



from Gadgetdaily.xyz

Twitter Developer API

See [Twitter's developer platform](#)

- ▶ library interfaces for Java, C++, Javascript, Python, Perl, PHP, Ruby, ...
- ▶ allows other applications to manage Twitter data for users
- ▶ extensive developer policy
- ▶ lots of [example case studies](#)

Freebase and DBPedia

Freebase:

- ▶ an example of a graph database we looked at earlier
- ▶ graph can be represented in RDF which is triples of URIs
- ▶ now owned by Google, currently read-only (may be decommissioned soon)
- ▶ used by others as a knowledge-base in many text processing pipelines:
 - ▶ e.g., using [TextRazor](#) to extract meaning from text

DBpedia:

- ▶ aim to extract all structured content from information in Wikipedia
- ▶ open source project

Medical Data Dictionaries

A service of the U.S. National Library of Medicine | National Institutes of Health

My Profile | Sign Out | Contact

Unified Medical Language System™

UMLS Terminology Services

Metathesaurus Browser

UMLS Home | Applications | SNOMED CT | Resources | Downloads | Documentation | UMLS Home

Search | Tree | Recent Searches

Term ☐ CUI ☐ Code

frontal lobe

Release: 2012AA

Search Type: Word

Source: OCTUBA
SNM
SNMI
SNOMEDCT
SPN
SPC

Search Results (33)
[: 1 - 25 :]

- C0016733 frontal lobe
- C1268977 Entire frontal lobe
- C0085541 Epilepsy, Frontal Lobe
- C0153635 malignant neoplasm of frontal lobe
- C0226193 Right frontal lobe structure
- C0226194 Left frontal lobe structure
- C0226195 Frontal lobe gyrus
- C0226196 Cortex of frontal lobe
- C0226197 Structure of white matter of frontal lobe
- C0338454 Frontal lobe degeneration
- C0338455 Dementia of frontal lobe type
- C0458309 Entire frontal lobe gyrus
- C0459388 Frontal lobe sulcus
- C0549117 Frontal lobe syndrome

Basic View | **Report View** | **Raw View**

Concept: [C1268977] Entire frontal lobe

Semantic Types
Body Part, Organ, or Organ Component [T023]

Atoms (8) string [AUI / RSAB / TTY / Code]

- Entire frontal lobe [A3852774/MT/HP/NNOCODE]
- lóbulo frontal [A5865532/SC/SPA/SY/180920004]
- lóbulo frontal [como un todo] [A5865525/SC/SPA/PT/180920004]
- lóbulo frontal [como un todo] (estructura corporal) [A5865524/SC/SPA/FN/180920004]
- Entire frontal lobe [A3421467/SNOMEDCT/PT/180920004]

Attributes (8) Name | Value | RSAB

- CONCEPTSTATUS | 0 | SNOMEDCT
- CTV3ID | 7N000 | SNOMEDCT
- DESCRIPTIONSTATUS | 0 | SNOMEDCT
- DESCRIPTIONTYPE | 1 | SNOMEDCT
- INITIALCAPITALSTATUS | 0 | SNOMEDCT
- ISPRIMITIVE | 1 | SNOMEDCT
- LANGUAGECODE | en | SNOMEDCT
- SNOMEDID | T-A2218 | SNOMEDCT

Relations (29) REL | RELA | RSAB [SType1 - SType2] STypeId | String | CUI

- Entire frontal lobe (body structure) [A3421466/SNOMEDCT/FN/180920004]
- Frontal lobe [A2931551/SNOMEDCT/SY/180920004]
- Tissue of frontal lobe of brain [A3077388/SNOMEDCT/SY/180920004]

Contexts (200)

Concept Relations (1) REL | RELA | RSAB | String | CUI

Copyright | Privacy | Accessibility | Freedom of Information Act | National Institutes of Health | Health & Human Services

The Unified Medical Language System (UMLS)

Medical Data Dictionaries, cont.

ICD: the **I**nternational **C**lassification of **D**iseases

- ▶ used to classify diseases and other health problems
- ▶ based on health and vital records
- ▶ for example:
 - ▶ *Pneumonia due to Streptococcus pneumoniae*

Medical Data Dictionaries, cont.

Other Medical Dictionaries:

- ▶ [SNOMED CT](#)
 - ▶ Systematized Nomenclature of Medicine Clinical Terms
- ▶ [Gene Ontology](#)
 - ▶ concepts for describing gene function

Usage of Medical Dictionaries:

- ▶ controlled vocabularies
- ▶ semantic data exploration
- ▶ clinical surveillance
- ▶ decision support

Publishing Repositories

- ▶ PUBMED, we have seen before
- ▶ [ACM Digital Library](#)
- ▶ Patent databases (for WIPO, USPTO, EPO, *etc.*), *e.g.*,
[Global Patent Search Network](#)

News and Event Registry

Event Registry

- ▶ collect news article globally, process and organise as events
- ▶ perform concept and event identification
- ▶ create a document database for inspection
- ▶ sometimes news stored as NewsML

Government Data

- ▶ US Government's [Data.GOV](#)
- ▶ [NYC Open Data](#)
- ▶ [Australia's Urban Intelligence Network \(AURIN\)](#)
- ▶ [BioGrid Australia](#)

Unit Schedule: Next Week

Module	Week	Content
1.	1	overview and look at projects (job) roles, and the impact
	2	
2.	3	data business models application areas and case studies
	4	
3.	5	characterising data and "big" data data sources and case studies
	6	
4.	7	resources and standards resources case studies
	8	
5.	9	data analysis theory data analysis process
	10	
6.	11	issues in data management data management frameworks
	12	