FIT5145 Introduction to Data Science

Module 4

Data Resources, Processes, Standards and Tools

2019 Lecture 8

Monash University

Reminder: NIST Analysis

data sources: where the data comes from

data volume: how much there is

data velocity: how it changes over time

data variety: what different kinds of data there is

data veracity: correctness problems in the data

software: software needed to do the work

analytics: broadly, what sorts of statistical analysis and visualisation needed

processing: broadly, computational requirements

capabilities: broadly, key requirements of the operational system

security/privacy: nature of needs here

lifecycle: ongoing requirements

other: noteable factors

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
	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	GUEST SPEAKER & EXAM INFO

Learning Outcomes (Week 8)

By the end of this week you should be able to:

- Explain about standards we introduce in different aspects of the process of Data Science
- Explain how to access to new data sources through APIs
- Identify how different APIs work
- Describe different software tools and programming languages in data science, and their popularity over time

ASIDE: Mapping Flight Data

24 Hour European Flight Traffic Visualization

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.

FLUX 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

FLUX Question

A vector of ages data was saved to file in the following format:

{"Age":{"0":39,"1":28,"2":44,"3":25,"4":32,"5":33,"6":31,"7":26,"8":22,"9":25,"10":28}}

What format is this?

A. RDF
B. XML
C. JSON
D. CSV

Standards and Issues
Open data and open source software

critical infrastructure and tools

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

machine learning and analytics on top of big data now main stream!

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)

Many state-of-the-art platforms integrated into *Hortonworks*.

Work and Salary Surveys

A number of organisations run salary surveys. These are usually interesting because they also describe what tasks people do and what software they use.

- O'Reilly's Salary Survey: behind login, slides summarised next
  - *2016 Data Science Salary Survey*,
    - really interesting content on software used, ...
  - *2017 European Data Science Salary Survey*,
    - really interesting content on tasks done, coding versus meetings, ..

- Kaggle *state of data science and machine learning*
  - really interesting content on job title, education, methods, barriers, getting started
  - explore this one online!

Tool Number from 2014 Survey

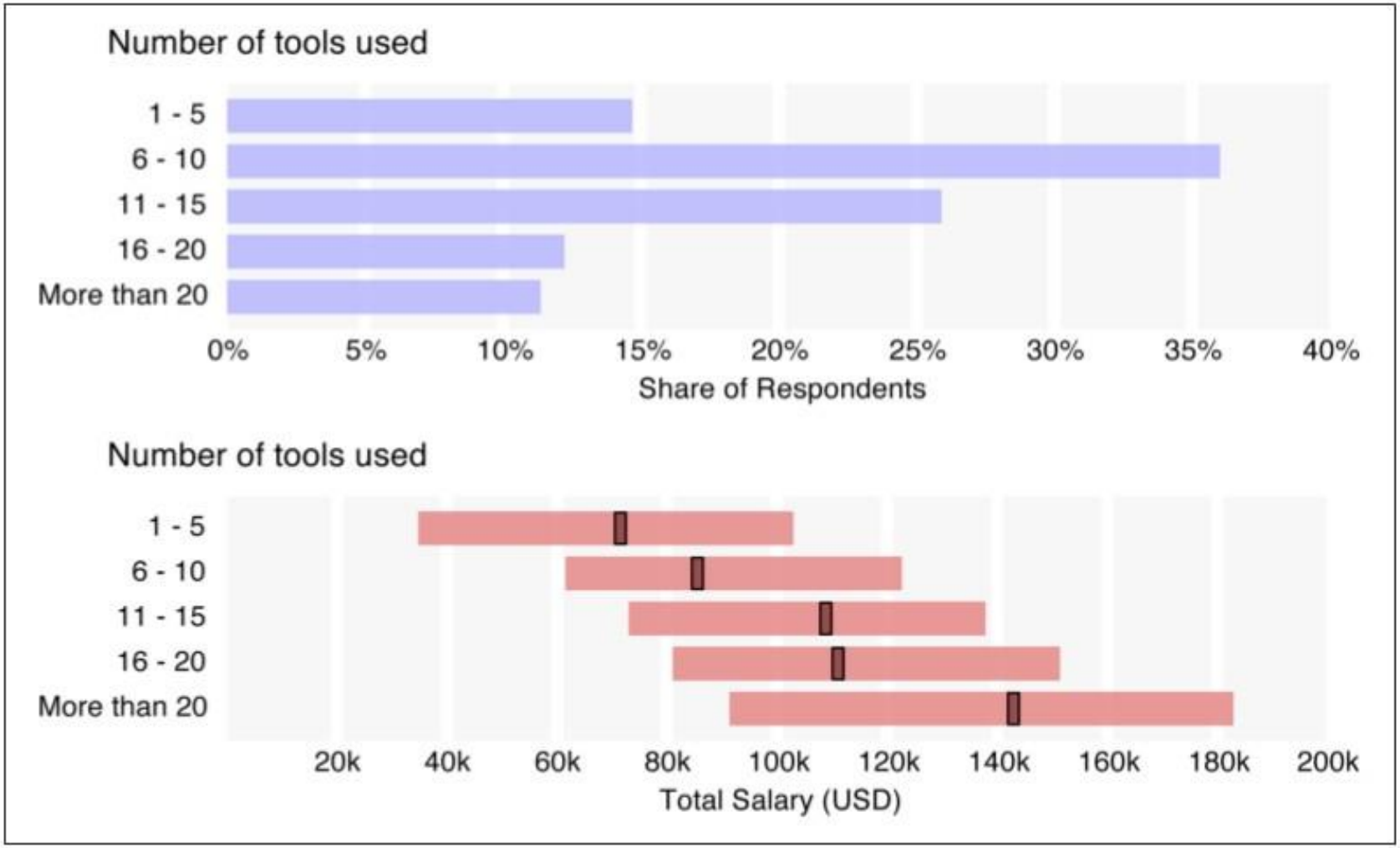
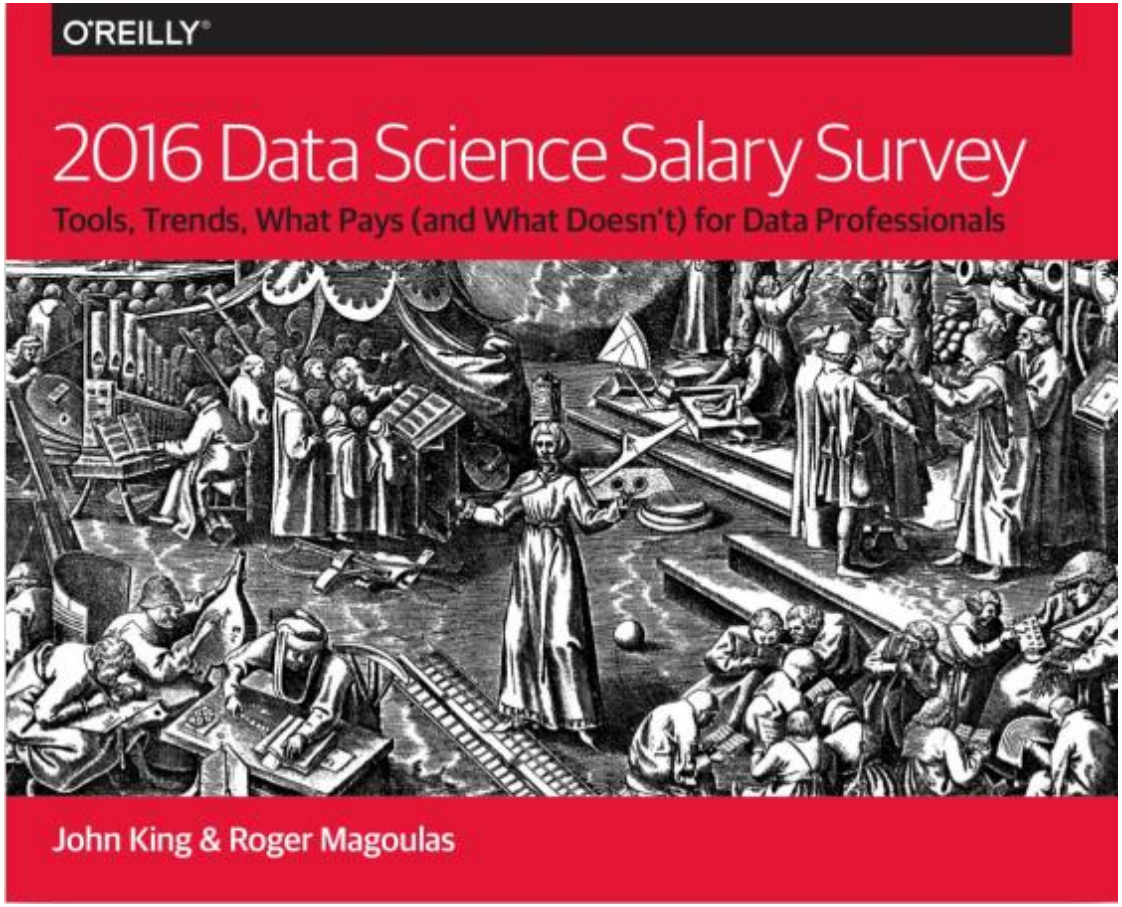


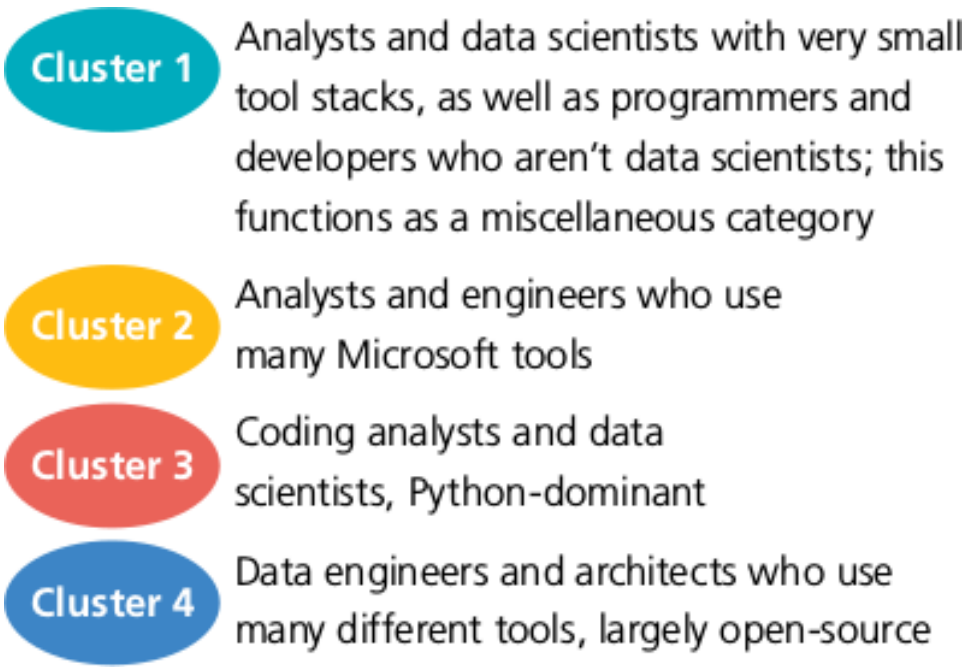
Figure 1-13. Number of tools used

Software Usage Survey



2016 Data Science Salary Survey

Survey: Clusters amongst the Respondents

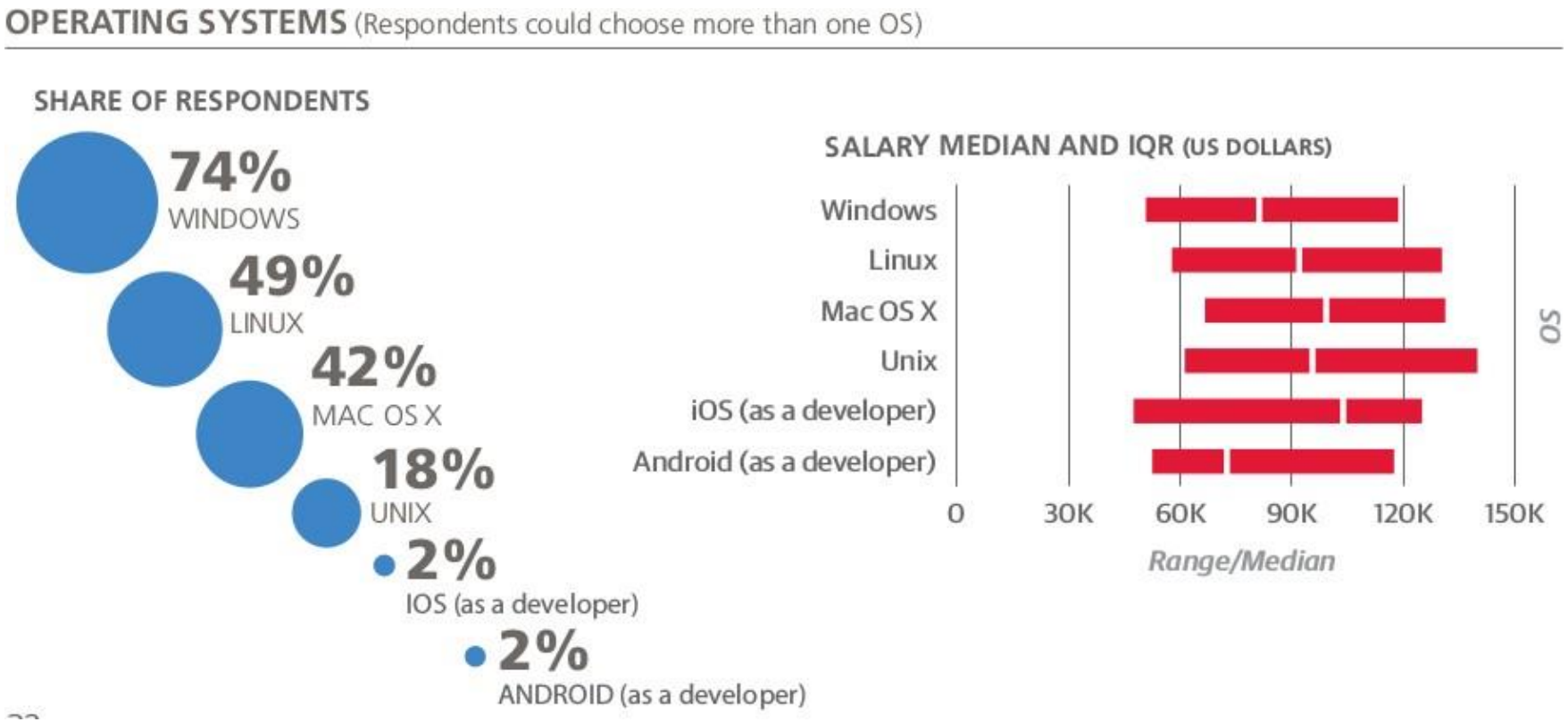


Survey: Commonly Used Software

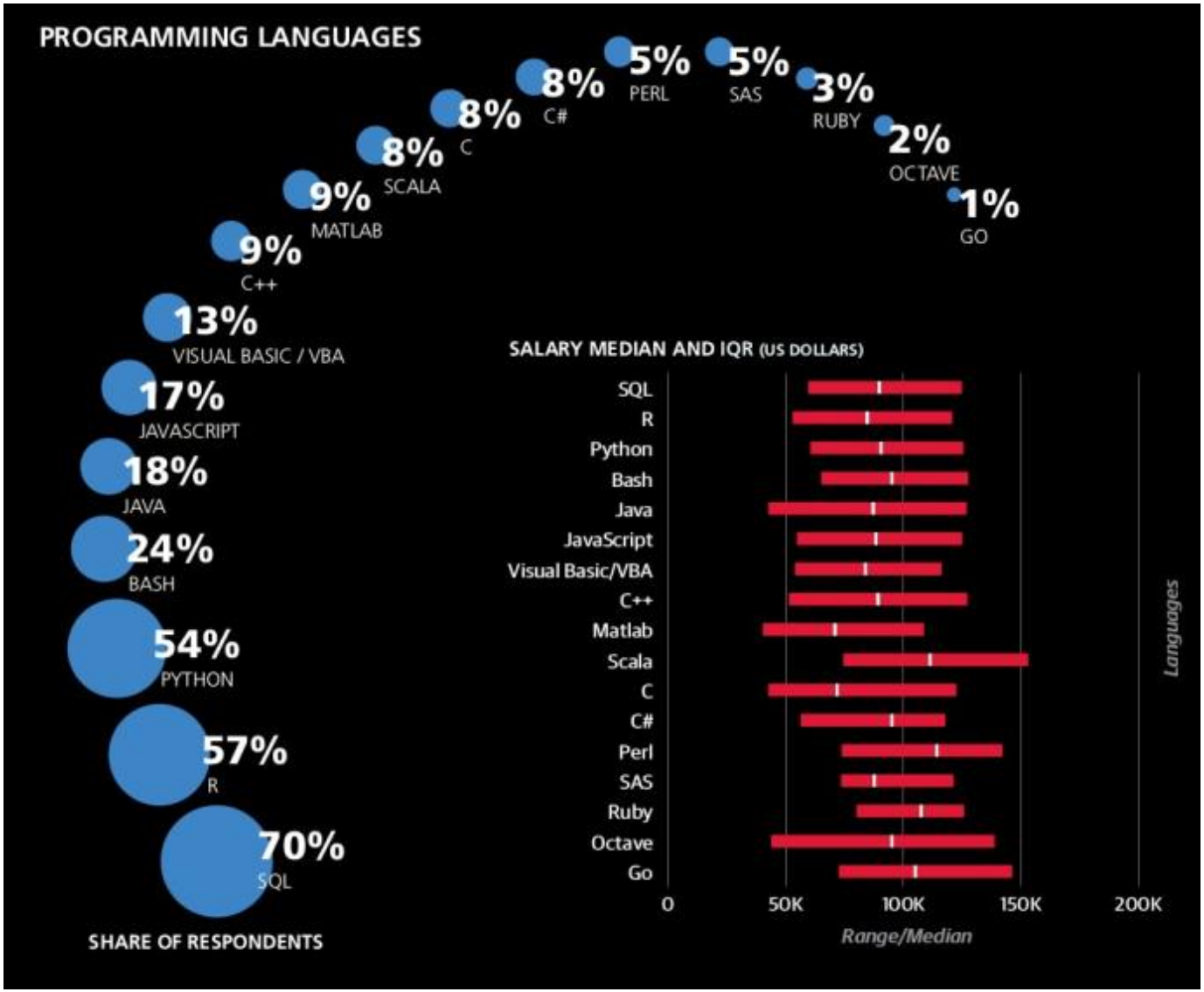
Tools	Cluster 1	2	3	4
Windows	86%	92%	48%	55%
SQL	62%	75%	65%	80%
Excel	66%	84%	59%	60%
R	30%	69%	67%	69%
Python	27%	32%	96%	84%
Linux	37%	21%	70%	91%
Mac OS X	26%	23%	70%	67%
MySQL	26%	33%	41%	57%
ggplot	13%	33%	53%	52%
Microsoft SQL Server	32%	51%	17%	27%
Tableau	17%	56%	21%	37%
Scikit-learn	7%	7%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

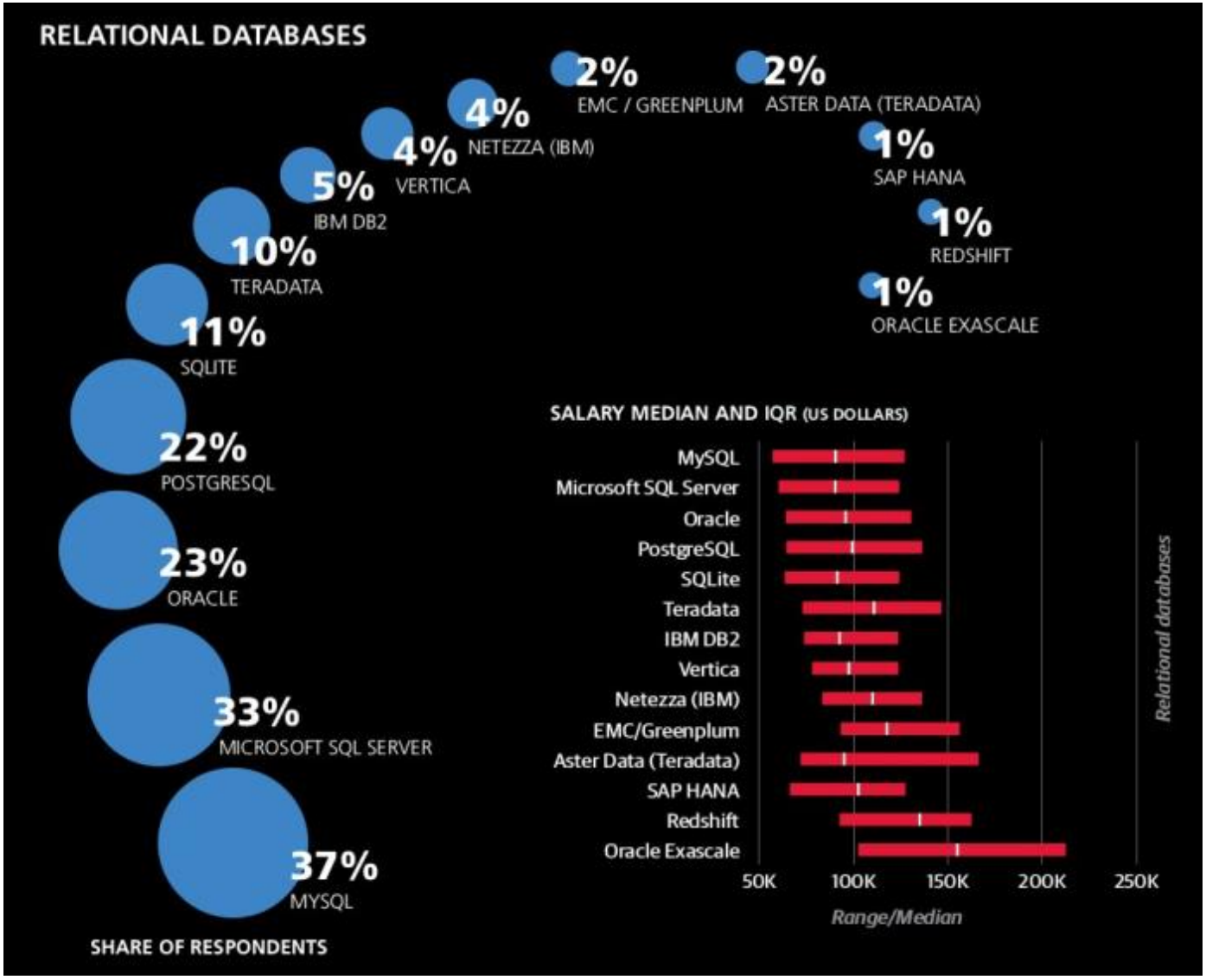
Survey: Operating Systems



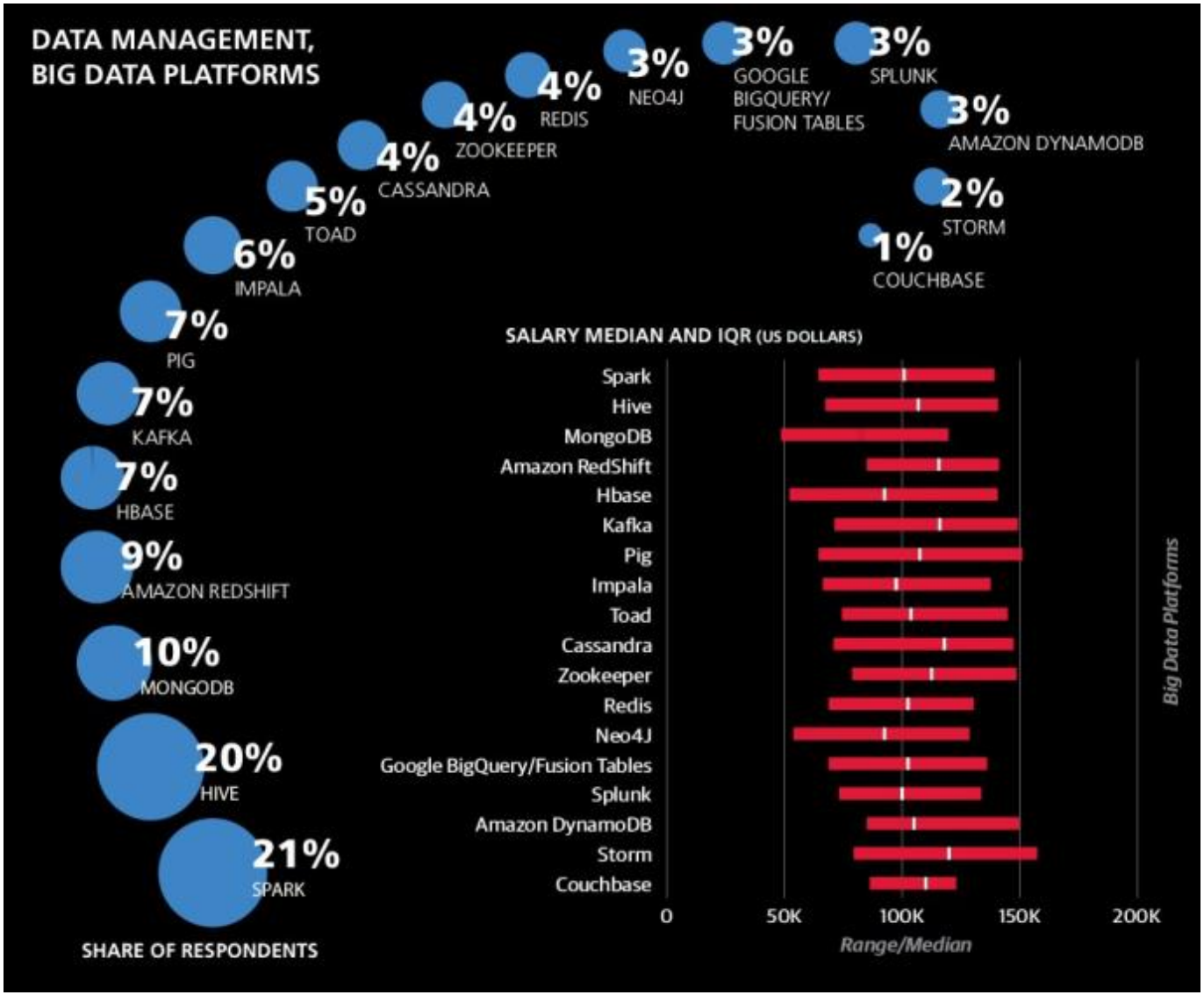
Survey: Programming Languages



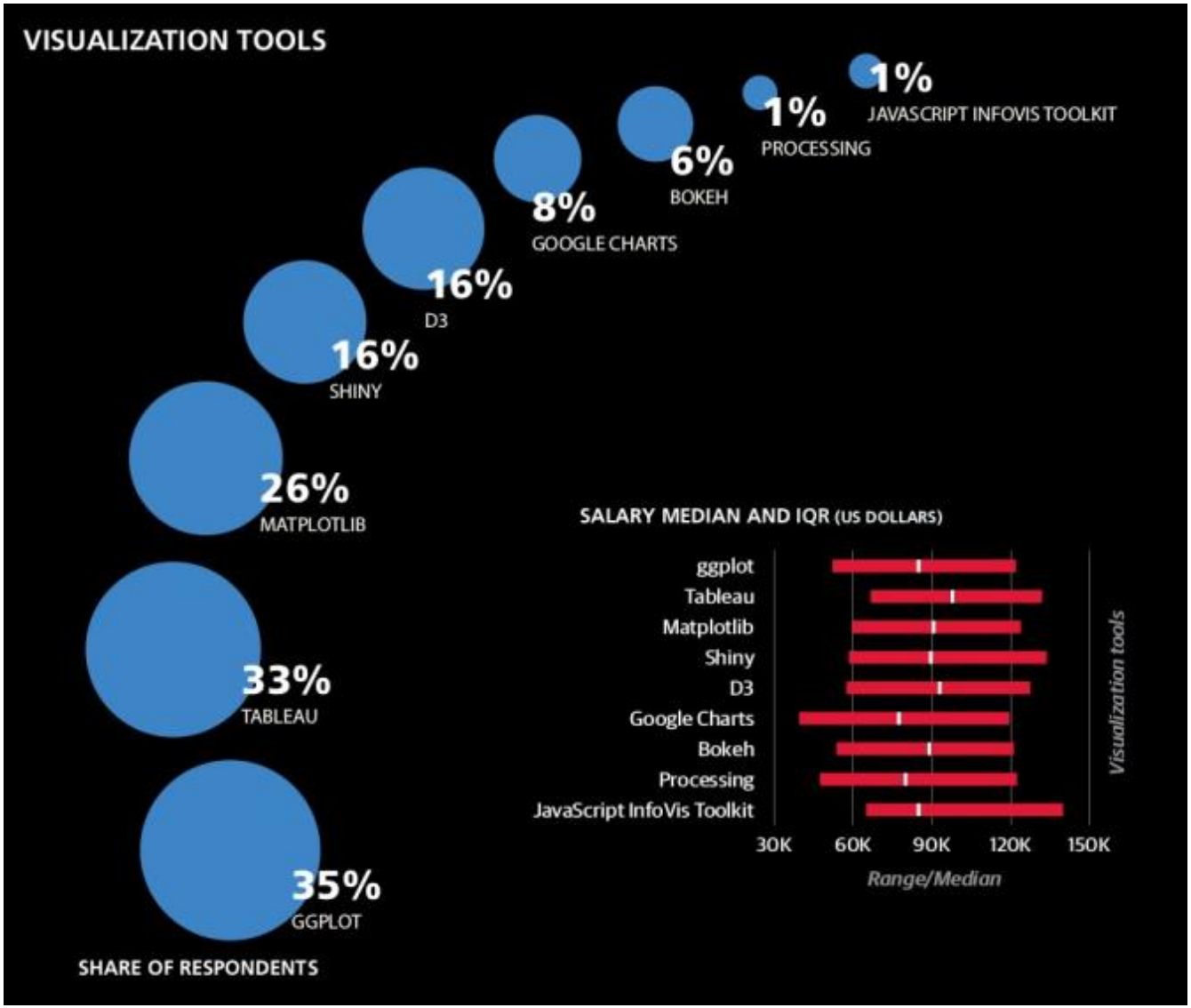
Survey: Relational Databases



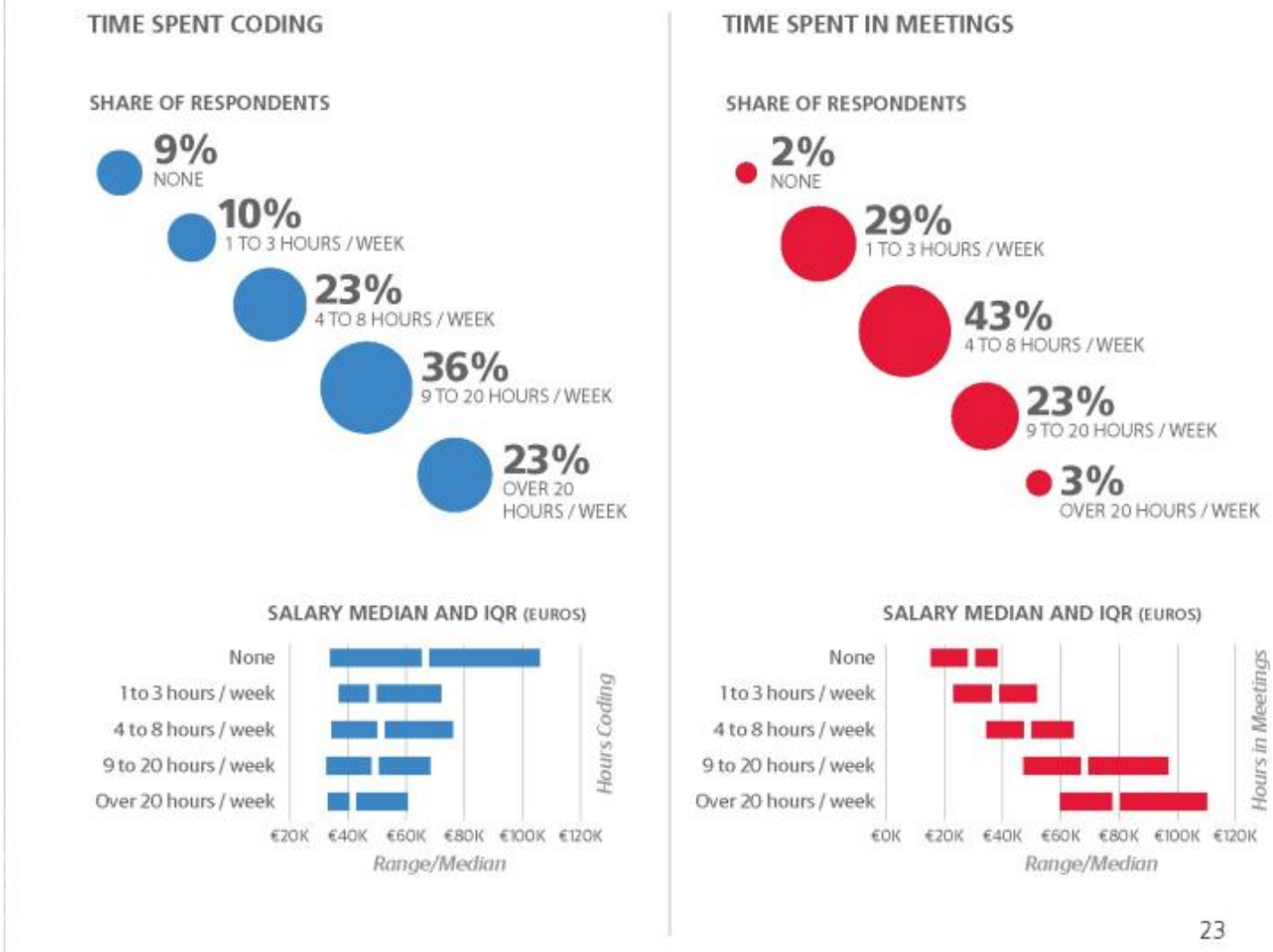
Survey: Management and Big Data



Survey: Visualization

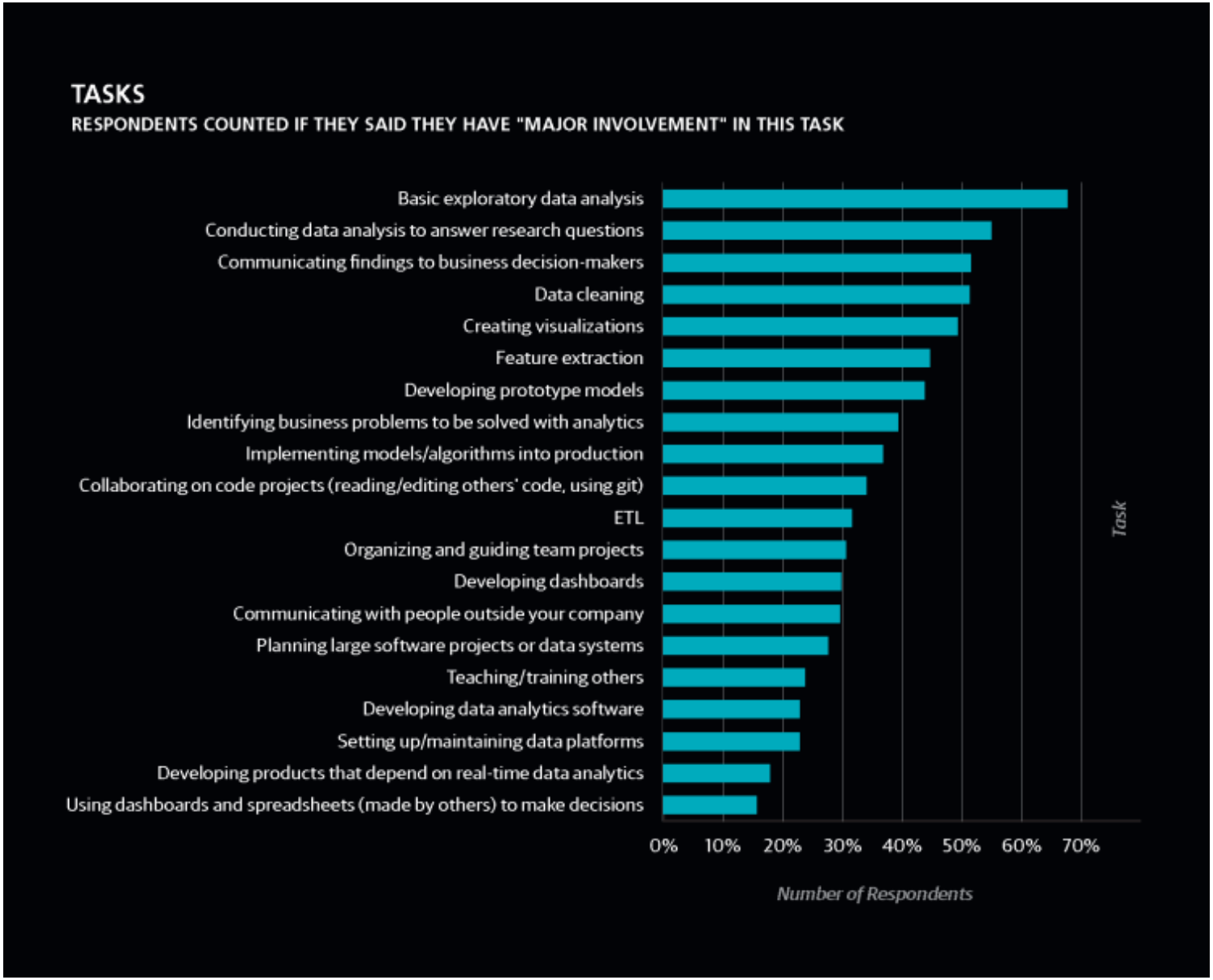


Coding versus Meetings

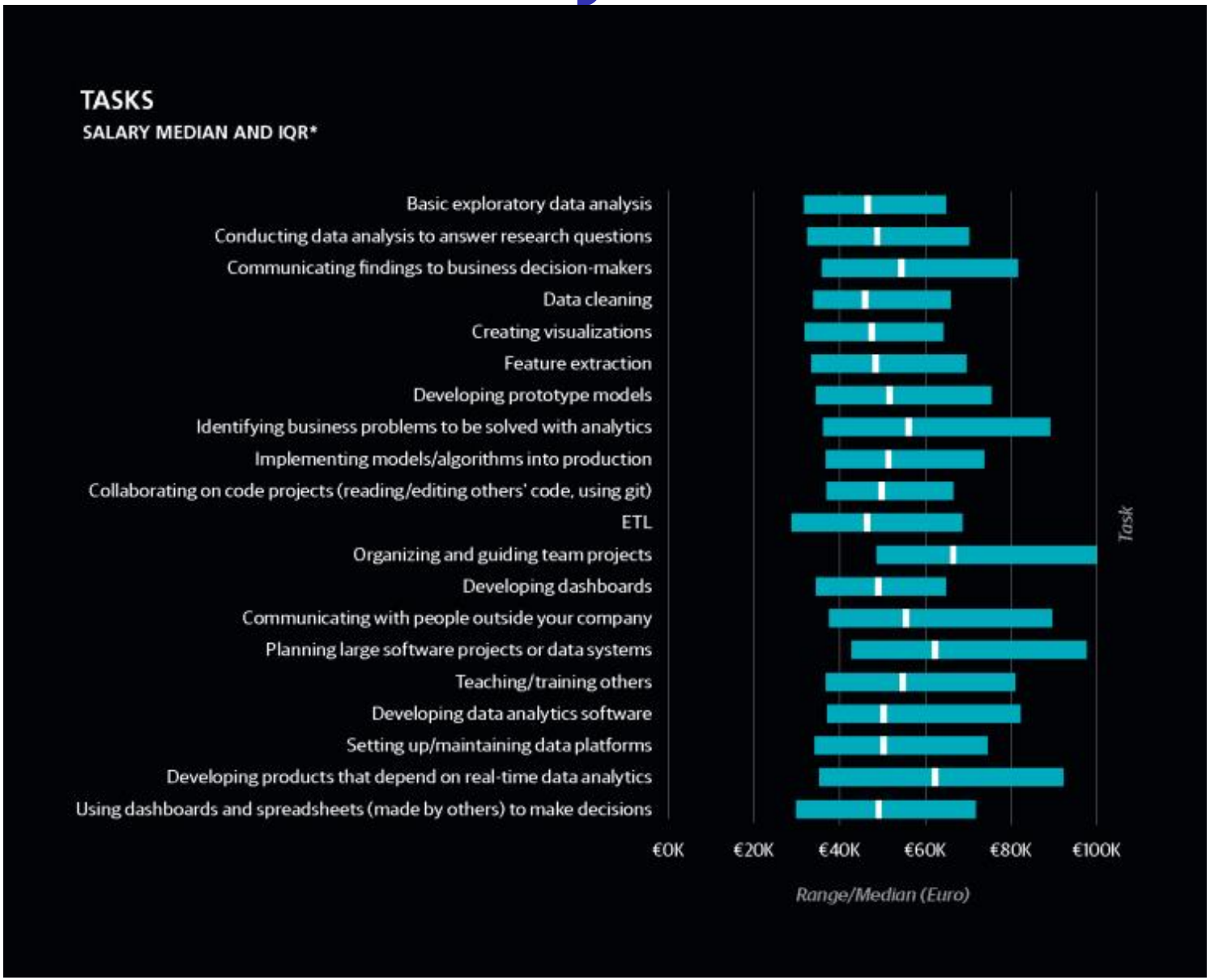


Career Choices

Tasks – Time


TASKS
RESPONDENTS COUNTED IF THEY SAID THEY HAVE "MAJOR INVOLVEMENT" IN THIS TASK
Basic exploratory data analysis
Conducting data analysis to answer research questions
Communicating findings to business decision-makers
Data cleaning
Creating visualizations
Feature extraction
Developing prototype models
Identifying business problems to be solved with analytics
Implementing models/algorithms into production
Collaborating on code projects (reading/editing others' code, using git)
ETL
Organizing and guiding team projects
Developing dashboards
Communicating with people outside your company
Planning large software projects or data systems
Teaching/training others
Developing data analytics software
Setting up/maintaining data platforms
Developing products that depend on real-time data analytics
Using dashboards and spreadsheets (made by others) to make decisions
Task
0%
10%
20%
30%
40%
50%
60%
70%
Number of Respondents

Tasks – Salary



Standards and Issues

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.

FLUX Question

Name a popular data/information API.

Example Data/Information APIs

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

- *Facebook API*
- *Twitter API*
  - e.g. *search tweets*
- *LinkedIn API*
- *Google Maps API*
- *Youtube API*
  - e.g. *documentation*
- *Amazon Advertising API*
- *TripAdvisor API*
- *New York Times API*

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* (see #4)
- *20+ Machine Learning as a Service Platforms*

And for something completely different:

- *The Sharing Economy | Bullish* (on TechCrunch)
  - these companies are huge users of data science!

Example Processing APIs or Web Services

Some companies are exposing their **tools/services** as APIs or browser based tools for others to make use of:

- *Azure Machine Learning Studio*
- *Figure-Eight Human in the Loop ML* with croudsourcing support
- *Watson REST API* for semantic web, metadata, entity analysis in text
- *Google Cloud Prediction API*
  - is closing down in April 2018, and they will focus on cloud solutions

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 and Standards
(ePub section 4.8)

look at some examples of standardised data collections

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, and decommissioned
- 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
- effectively replaced Freebase

Twitter

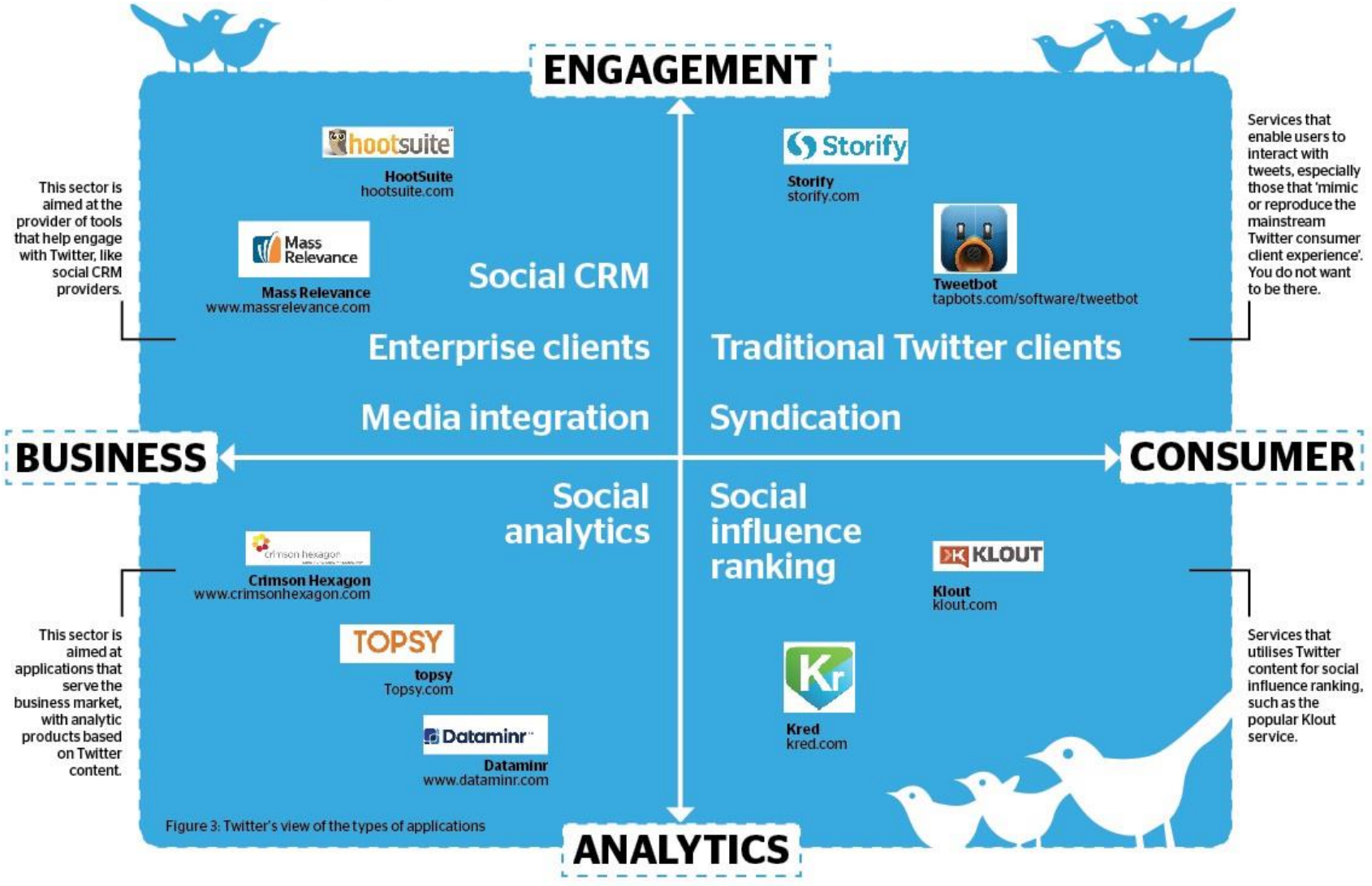
Twitter is the most famous microblogging platform

- 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 App Ecosystem



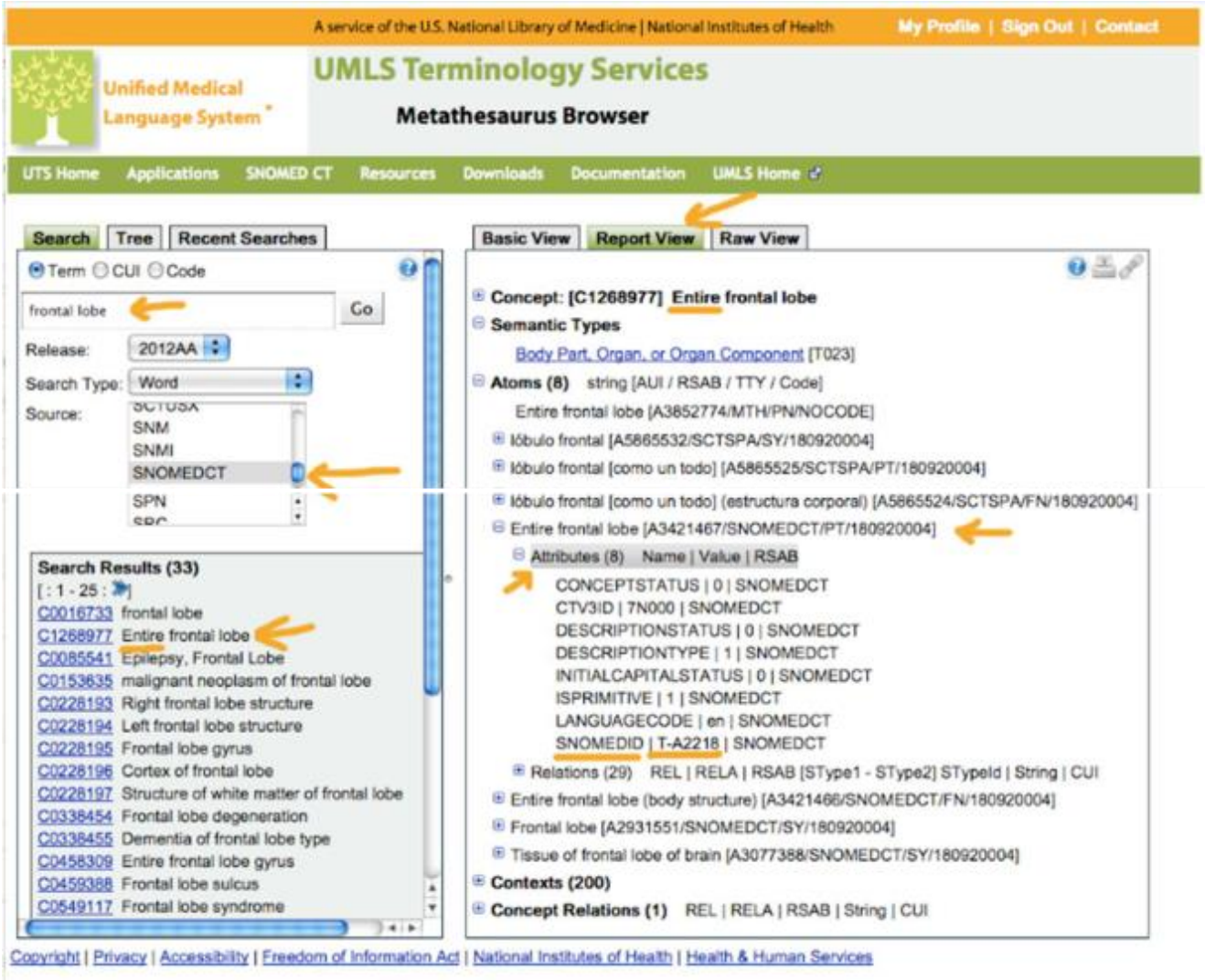
from Gadgetdaily.xyz

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
- see *search API doc*
- lots of *example case studies*

Medical Data Dictionaries



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*
- *Global Patent Index* provided by the EPO
- *Semantic Scholar* for research article search

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)*

  e.g. *SD Private Health Insurance*
- *BioGrid Australia*
  - curated for research use and usually require getting approval to use

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
	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	GUEST SPEAKER & EXAM INFO