FIT1043 Introduction to Data Science

Module 6

Data Management

Lecture 11

Slides by Mahsa Salehi
Acknowledgement: Wray Buntine

Discussion: Regression in iPython

- In week 10 tutorial we investigated important ideas from Machine Learning theory:
- Polynomial models
  - e.g. polynomial regression versus Legendre polynomials
- model complexity
  - controlled by order of polynomial (# parameters)
- overfitting and underfitting
- ensembling

Discussion: Prediction with BigML

- In week 11 tutorial you made use of a commercial product BigML for building simple predictive models using Decision Regression trees
- BigML:
  - Example of a modern Machine Learning Tool provided as an online service
  - Emphasis onuser-interface and making model building simple from a graphical interface perspective
  - Combines Decision/Regression Tree Ensembles, Clustering, Frequent Itemset Mining, and Outlier Detection models
  - Provides fewer classification/regression models in comparison to Weka, R, Python (Scikit learn)

Unit Schedule: This Week

Module	Week	Content
1.	1 2	Overview and look at projects (Job) roles, and the impact
2.	3	Data business models / application areas
3.	4 5	Characterising data and “big” data Data sources and case studies
4.	6 7	Resources and standards Resources case studies
5.	8 9 10	Data analysis theory Regression and decision trees Data analysis process
6.	11 12	**Issues in data management + Unit overview and exam info for Clayton** **Guest lecture (Clayton)** **Unit Overview and exam info (Malaysia)**

Why manage data?

- The data is very valuable, data collection is usually time consuming and hard
- Large amount of data and documents are being generated with high growth rate
- Multiple sources of data (general business documents, ERP systems etc)

What is Data management?

Data management is the development, execution and supervision of plans, policies, programs and practices that *control, protect, deliver and enhance the value of data and information assets*.

Data management is important

See “How to avoid a data management nightmare”, a video created by NYU Health Sciences Library (Youtube)

Data management in research

- Information privacy where human are involved is important

- See "Managing Research Data" from Digital Curation Centre in the UK

Issues in Data Management (ePub section 6.1)

Data management plan in an organisation

Deals with issues:

- integration and data warehousing [See Data Warehousing- an Overview (Youtube, 2:44-5:33)]
- replication and persistence
- standardising the vocabulary used across the organisation, e.g., job titles
- security

Privacy versus Confidentiality

- Privacy is (for our purposes) having control over how one shares oneself with others.
  - e.g. closing the blinds in your living room
- Confidentiality is information privacy, how information about an individual is treated and shared.
  - e.g. excluding others from viewing your search terms or browse history
- Security as the protection of data, preventing it from being improperly used
  - e.g. preventing hackers from stealing credit card data

Social media and the loss of confidentiality

- See: "The curly fry conundrum: Why social media 'likes' say more than you might think" by Jennifer Golbeck (TED)
- Target is predicting which women is pregnant from their purchases

- many things can be predicted from Facebook "likes"

- Implicit data ::= data not explicitly stored but inferred with reasonable precision from available data

Confidentiality, cont.

- See: "*Empower consumers to control their privacy* in the Internet of Everything" by Carla Rudder (blog)

- For many apps or services, you must either accept their data sharing policies or you can't use their services fully

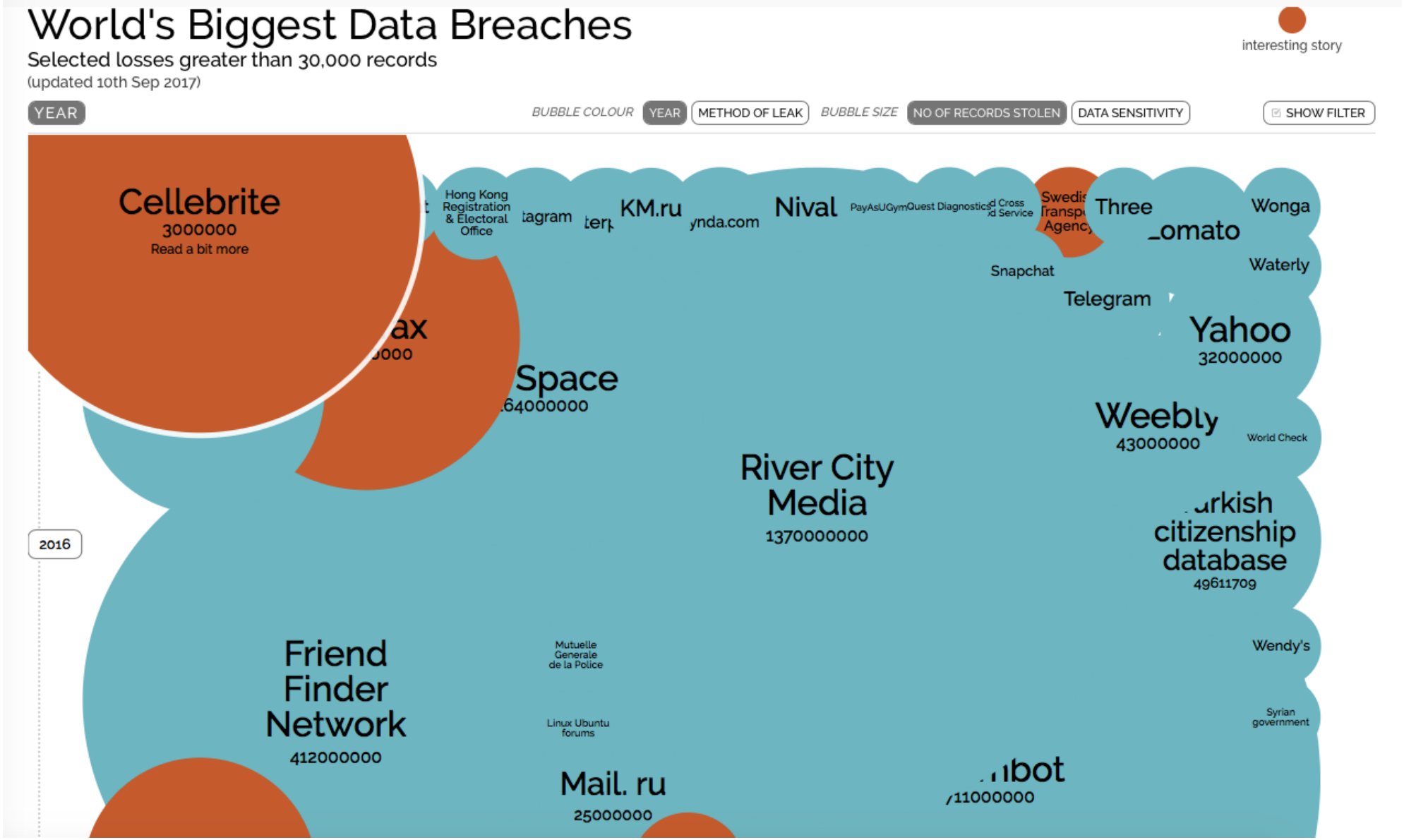
- There could be an agent to interact in a narrative form with individual consumers,
  - For instance the app might ask: 'Are you willing to share your health data with company X?'

Compliances and Regulations

- Ethics: the moral handling of data
- There should be regulations in place to ensure that confidentiality is protected
- The process of ensuring you meet regulations is called compliance

Compliances and Regulations Example

- PCI (Payment Card Industry) standard
  - Aims to reduce credit card fraud
    - By placing specific regulations, e.g., credit card information should be stored only in encrypted format.
  - Companies who handle credit card have to comply with PCI standards
  - Audit (validation of compliance) is done annually



Here is the [link](#).

Data Management and Data Science

- medical informatics: for predicting fungal infections from nursing notes, the team needs to abide by confidentiality and security

- internet advertising: what implicit and explicit data is stored about a user

Unit Schedule: Next Week

- Clayton
  - Guest lecture
  - Prashant Bhatnagar from Microsoft Australia (Principal Delivery Pursuit Lead- Data and Artificial Intelligence)

- Malaysia:
  - Quick overview of what we learnt in this unit
  - Exam format and discussion