FIT5145 Introduction to Data Science

Module 6

Data Curation and Management

2019 Lecture 11

Monash University

Reminders

- SETU time, see Moodle and go to SETU – Unit Evaluation link
- Week 11 you submit your Assignment 4
  - reports are due Friday
- Week 12 you present your Assignment 5 slides in tutorial
  - slides are due Monday morning (7:55am) so that tutors can assemble
  - talks will be given in alphabet order ... be ready or loose 5%
  - if seeking special consideration, you must post a video to Youtube

Discussion: Regression in iPython

In the last tutorial we investigated important ideas from Machine Learning theory:

- linear and polynomial regression
- model classes
  - e.g. polynomial regression versus Legendre polynomials
- model complexity
  - controlled by order of polynomial (# parameters)
- overfitting and underfitting
- ensembling

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	**GUEST SPEAKER & EXAM INFO**

Issues in Data Curation and Management (ePub section 6.1)

overview of issues

- confidentiality and ethics
- compliance and governance
- data management

Issues in Data Curation and Management

Confidentiality and Ethics

issues around confidentiality

Terminology

For our purposes, we define:

- Privacy as having control over how one shares oneself.
  - e.g. closing the blinds in your living room
- Confidentiality as information privacy, how information about an individual is treated and shared.
  - e.g. excluding others from viewing your search terms or browsing history
- Security as the protection of data, preventing it from being improperly used
  - e.g. preventing hackers from stealing credit card data
- Ethics as the moral handling of data.
  - e.g. not selling on other's private data to scammers
- Implicit data that is not explicitly stored but inferred with reasonable precision from available data
  - see *"Private traits and attributes are predictable..."*

Confidentiality

See: *"The curly fry conundrum: Why social media 'likes' say more than you might think"* by Jennifer Golbeck (TED) – see 1:00 to 3:40

e.g. Target®predicting which women are pregnant based on their purchases

- Many things can be predicted from Facebook "likes"
- Homophily (tendency to associate with similar individuals) is important for enabling prediction
- We often don't own or manage corporate/internet/app data about ourselves
- The source data critical for advertisers so we cannot expect companies to be banned/excluded from using it
- So how can we manage confidentiality?

Confidentiality, cont.

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

- for many apps/websites, you must accept their privacy data sharing policies to use their services fully;
- the interface for selecting privacy preferences should move away from individual Internet platforms and be put into the hands of individual consumers;
- user could have an open source agent that broker their confidentiality preferences
- but would that be feasible and would businesses ever agree?

Politics of Confidentiality

See: *"Four political camps in the big data world"* by Cathy O'Neil (blog)

1. Corporations: want to use data for business advantage;
   - opposing consumers
2. Security conscience: concerned with freedom, liberty, mass surveillence;
   - opposing intelligence orgs like NSA
3. Open data: want open accessability, support FOI requests
   - opposing security experts concerned with leaks
4. Big data and civil rights: concerned about big data and citizens;
   - opposing data brokers selling consumer data

Facebook and Personal Data

See: *Facebook Doesn't Tell Users Everything* and *Facebook Privacy: Social Network Buys Data*

Facebook buys 3rd party data (from brokers) to glean a user's activity, income, etc.

- keeps upwards of 52,000 features about users, many provided to advertisers
- bought data used as a complementOracle's Datalogix,
- it is public, offline data, e.g., from Oracle's *Datalogix*,
- but is not revealed to users

Facebook and Voting

See: *"Can Facebook influence an election result?"* by Michael Brand (ex-Monash, opinion on ABC news via *The Conversation*)
and also *How Facebook could swing the election* by Caitlin Dewey (article, Washington Post)

- *implicit data:* Facebook can predict who you will vote for
- their "I voted" button encourages people to vote (as they see which of their friends have)
- studies show it significantly increased voting in 2010 US election
- they can therefore subtly affect your voting
- could Facebook deploy "I voted" button selectively to favour certain parties in certain areas?

Population-level Prediction

See *"Machine logic: our lives are ruled* by big tech's 'decisions by data' ", and see *"If prejudice lurks among us,* can our analytics do any better?"

Predictive models built on large populations are used to filter/make key life decisions like release from jail, treatment in hospital, getting a loan, news/videos you see (e.g., Facebook) ...

- ML algorithms do the filtering
- ML algorithms can also produce prejudice (i.e., are biased)
- decisions made on mass, not personalised
- decisions are centralised (who writes the algorithms?)
- perhaps this is OK

Issues in Data Curation and Management

Compliance and Governance

how and why an organisation deals with data

Regulations and Compliance

- Regulations devised by various government bodies: taxation, medical care, securities and investments, work health and safety, employment, corporate law.
- they need to check companies for their compliance
- Auditing

  systematic and independent examination of books, accounts, documents and vouchers of an organization to ascertain how far they present a true and fair view

- Regulatory compliance:

  that organisations ensure that they are aware of and take steps to comply with relevant laws and regulations.

- auditing data and records are a good source for Data Science

What is Data Governance?

See *"What is Data Governance?"* by Rand Secure Data (Youtube)

See *"What is Data Governance?"* by Intricity (Youtube)

Data Governance

Supporting and handling:

- ethics, confidentiality
- security
- consolidation and quality-assurance (e.g. link all customer related information together)
- persistence (backups and recoverability)
- regulatory compliance
- organisation policy compliance
- organisation business outcomes

which may include handling the steps in the data science and/or big data value chain

Security Example

See
"Target CEO ousted as boards focus on cyber risk mitigation"

- Target (retail chain) had credit card data stolen by hackers
- the CEO was subsequently ousted!
- data security now taken seriously at the board level
- data security important to customers
  - e.g. Google treats search terms as extremely confidential

AI/ML for Deep Fakes

Increasingly realistic synthetic faces generated by variations on Generative Adversarial Networks (GANs) from 2014-2017 (taken from report on next slide).

Machine learning image processing software lets one create deep fakes (which are supposedly high quality fakes):

- splice a celebrity's face onto a porn video
- create a video with President Trump declaring war on France

Malicious Use of AI/ML

see *The Malicious Use of Artificial Intelligence:* Forecasting, Prevention, and Mitigation, e.g.,

- faking digital media
- faking interactions (phone calls, teleconferences)
- cyber-attacks taking over autonomous vehicles
- spoofing autonomous weapons systems

Mitigations:

- formal verification, exploring vulnerabilities, etc.
- effective licensing of technologies
- education, norms and standards in data science and AI
- policies

FLUX Question

Data governance does NOT dealing with:

A. archiving
B. anthropomorphic
C. legal compliance
D. privacy issues

Issues in Data Curation and Management

Data Management

managing to achieve governance, etc.

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, cont.

See *"Top 10 Mistakes in Data Management"* a tutorial from Intricity (a data management company) (Youtube)

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

Data Management and Data Science

Examples of data management issues arising in data science projects:

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

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

retailing: conduct market intelligence on new products; put together data from different divisions (brands) within the company.

predictive medical system: implementation may need changing standard operating procedure for staff

FLUX Question

Data management for a medical application includes:

A. developing security tools
B. developing custom streaming database solutions for medical data
C. developing a policy for user privacy
D. analysing the big data

Frameworks for Data Management (ePub section 6.2)

management of data

- Digital Curation Centre
- Australian Public Service
- science/research lifecycle models
- NIST reference architecture

Contexts for Data Management

Science: reproducibility and credibility of scientific work, producing artifacts of knowledge, creating scientific data

Business: governance, compliance, information privacy, *etc.*

Curation: *e.g.* museums and libraries, preservation, maintenance, *etc.*

Government: a unique legislative environment that regulates them (*e.g.*, "transparency"), archiving, FOIs, support data infrastructure, *etc.*

Medicine: significant privacy issues, conflicting corporate financial constraints, government regulations and furthering of medical science

Digital Curation Centre

About:

> *The Digital Curation Centre (DCC) is a world-leading centre of expertise in digital information curation with a focus on building capacity, capability and skills for research data management across the UK's higher education research community.*

See *"The DCC Curation Lifecycle Model"* by DCC (PDF)

Australian Public Service

Background:

the creation, collection, management, use and disposal of agency data is governed by a number of legislative and regulatory requirements, government policies and plans

- data needs to be authentic, accurate and reliable
- strong governance framework
- sensible risk management and a focus on information security, privacy management
- clear and transparent privacy policies and provide ethical leadership

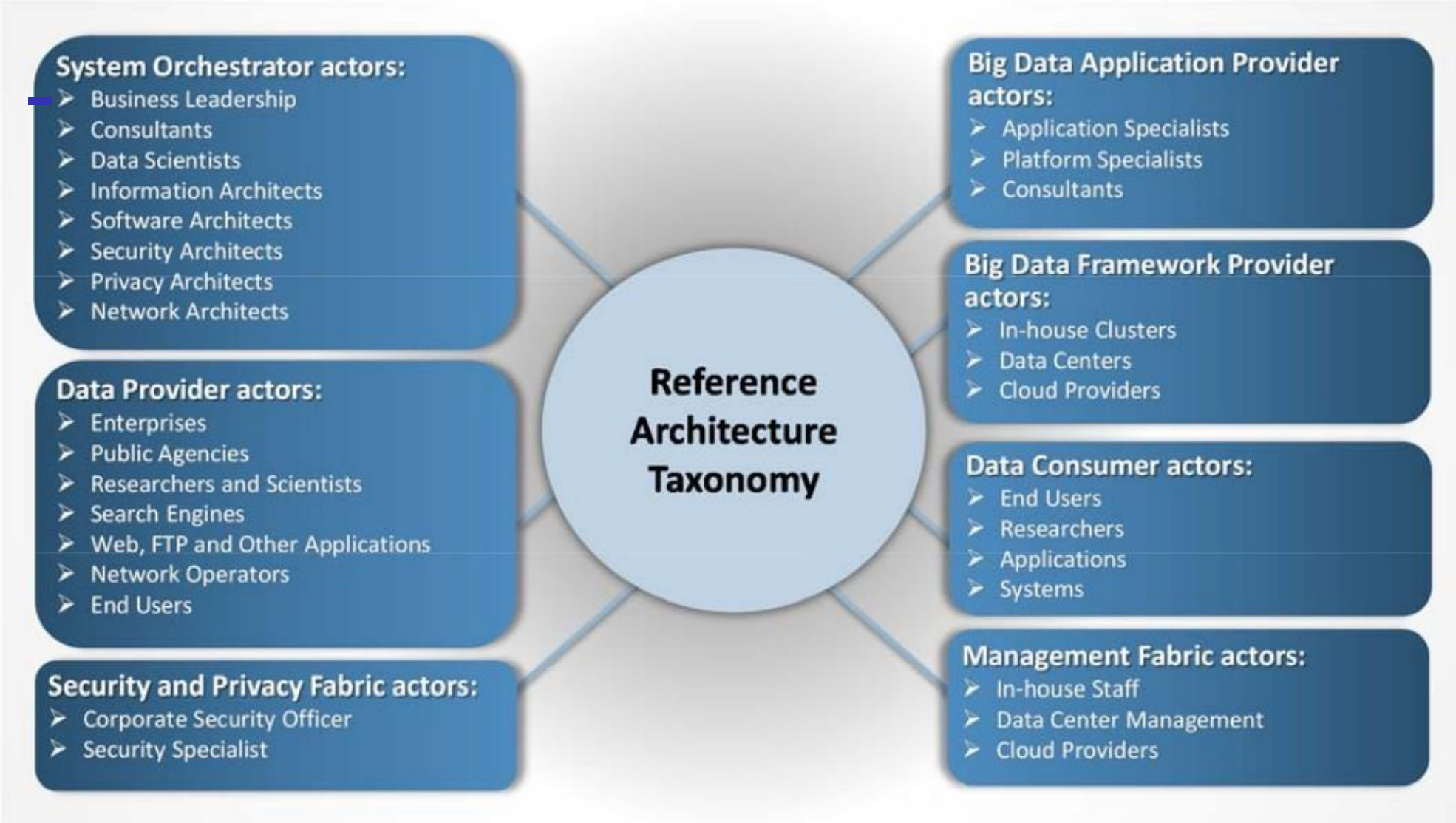
Data Observation Network for Earth

About:

is a US funded research organisation with an interest in data collection for environmental sciences

- an example of science/research lifecycle models
- sponsors public data collection
- guidelines related to standard data curation

See *"Example Data Management Plan: Rio Grande Basin Hydrologic Geodatabase Compendium"* by DataONE (PDF)



NIST Reference Architecture showing actors and roles in data management

Tutorial this Week: Prediction with BigML

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 on user-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: 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**