Lecture1: Introduction to Knowledge Technology

COMP90049
Knowledge Technologies

Sarah Erfani and Karin Verspoor

July, 2017

THE UNIVERSITY OF MELBOURNE

Who and where

Lecturers:
Dr Sarah Erfani, DMD Level 07, Room 7.14,
sarah.erfani@unimelb.edu.au
Mr Jeremy Nicholson, DMD Level 10, Room 10.03,
nj@unimelb.edu.au

Tutors:
Jeremy Nicholson (Head Tutor), DMD, nj@unimelb.edu.au
Andrey Scherbakov, andreas@softwareengineer.pro
Oscar Fabian Correa Guerrero, oscarcg@student.unimelb.edu.au
Qingyu Chen, qingyuc1@student.unimelb.edu.au
Daniel Ma, xingjunm@student.unimelb.edu.au

Lectures:
Tuesdays, 4:15pm - 5:15pm, Public Lecture Theatre in Old Arts
Fridays, 12:00pm - 1:00pm, Charles Pearson Theatre in the Eastern Resource Centre (ERC)

Who and where

Tutorials: (per your registration) Start in Week 2

Lecture Materials: Lecture slides available on LMS, lectures recorded on Lecture Capture

Feedback

- During/after lecture
- Tutorials
- Discussion board
- Consultation sessions
- Assignment feedback
- Sarah/Jeremy office (by announcement or by appointment)

Prerequisites

Subjects:

- COMP20003 (433-253/298) Algorithms and Data Structures
- COMP90038 (433-521) Algorithms and Complexity

Skills:

- Data structures & algorithms coding in C, C++, Python, Java or similar.
- Assignments to be completed in any programming language. (Elementary C and scripts to be used in lectures.)
- Familiarity with formal mathematical notation.
- Basic understanding of statistics and information theory helpful but not essential.

This subject does not include programming language tuition.

Assessment

- Assessment: 50% final exam, 10% mid-semester test, and 40% project
- Requirements: 30/60 exam hurdle (final and mid-semester test), 20/40 project hurdle, and 50/100 overall
- Projects:
  Project 1 will be released in week 3 and due in week 7.
  Project 2 will be released in week 8 and due in week 11.
  (Dates to be confirmed in project specification on subject LMS site)
  You are expected to complete these *individually*.
  We will discuss the project in more detail over the coming weeks.

(Note that the non-teaching week is between weeks 9 and 10.)

COMP90049 Subject overview from handbook

"Much of the world's knowledge is stored in the form of unstructured data (e.g., text) or implicitly in structured data (e.g., databases).

"In this subject students will learn algorithms and data structures for extracting, retrieving and storing explicit knowledge from various data sources, with a focus on the web.

"Topics include: data encoding and markup, web crawling, clustering, regular expressions, pattern mining, Bayesian learning, instance-based learning, document indexing, database storage and indexing, and text retrieval."

Learning objectives

On successful completion of the subject, students should be able:

- "To apply knowledge and skills in many fields that need extensive data analysis."
- "To describe and apply the fundamentals of knowledge systems, including data acquisition and aggregation knowledge extraction, text retrieval, machine learning and data mining"

What the subject covers

- Exposure to a range of computing technologies for:
  - Making use of uncertain, irregular, or complex data.
  - Accomplishing tasks that may not be well-specified or well-understood.
  - Supporting humans who are engaged in discovery or decision-making.
- A broader understanding of the kinds of things that can – and can't – be accomplished computationally.
- Insight into some research activities in computing, why they are undertaken, and how.

Content

Week 1-5:

- Basic text processing
- Pattern and string matching, spelling correction
- Web and text search

Week 6-12:

- Machine learning
- Clustering, classification
- Data mining

Along the way:

- Measurement of effectiveness (Evaluation)
- Some interesting algorithms, a little theory
- Bayesian reasoning
- Insights into current research

Beyond the scope of this subject

Far more knowledge technology topics are out than are in!

- Computational modelling: traffic, medical, climate, . . .
- General approximation and reasoning techniques for computing solutions in the presence of formal intractability. (But we do look at a couple of specific examples.)
- Natural language processing, machine translation.
- Image analysis, image matching.

. . . and many others.

Texts and references

There is no prescribed text. You may find these useful.

- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze (2008), *Information Retrieval*, Cambridge University Press. Freely available at `informationretrieval.org`

- Pang-Ning Tan, Michael Steinbach and Vipin Kumar (2005) *Introduction to Data Mining*, Addison-Wesley.

- Cathy O'Neil, Rachel Schutt. *Doing Data Science: Straight Talk from the Frontline.* (available as eBook: `http://shop.oreilly.com/product/0636920028529.do`)

- Ian Witten, Eibe Frank, Mark Hall *Data Mining: Practical Machine Learning Tools and Techniques* `http://www.cs.waikato.ac.nz/ml/weka/book.html`

- Anand Rajaraman and Jeff Ullman *Mining of Massive Datasets* `http://infolab.stanford.edu/~ullman/mmds.html`

Lecture 1: Introduction to Knowledge Technology

What's in a database?

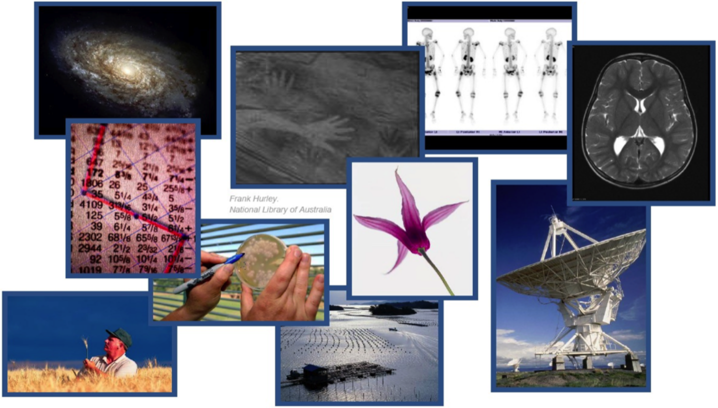
What’s in a database?

Okay, **data**, obviously.

What's in a database?

- transactional data (e.g., consumer purchases)
- sensor data (e.g., weather data)
- measurements (e.g., laboratory data)
- accounting data
- . . .

Data is everywhere



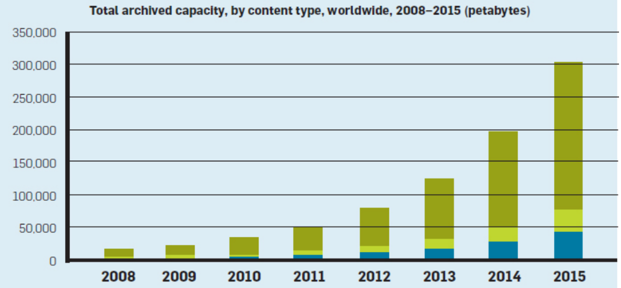
Frank Hurley.
National Library of Australia

What’s in a database?

- transactional data
- sensor data
- measurements
- images
- free text: short (e.g. notes, emails, tweets) or long (e.g. papers, blog entries)
- videos
- maps
- recipes
- . . .

Maybe the better question is, What’s **not** in a database?

Some Data on Data



	2008	2009	2010	2011	2012	2013	2014	2015
Unstructured	11,430	16,737	25,127	39,237	59,600	92,536	147,885	226,716
Database	1,952	2,782	4,065	6,179	9,140	13,824	21,532	32,188
Email	1,652	2,552	4,025	6,575	10,411	16,796	27,817	44,091

Source: Vasant Dhar "Data Science and Prediction" (2013) Communications of the ACM, Vol. 56 No. 12, Pages 64-73 doi:10.1145/2500499

What to do with all that data?

Most data generated by humans and computers today is for consumption by computers

Facilitated by:

- database schemas
- mark-up languages
- programmatic APIs

Database querying, and basic computational processing of data, asks:

What data satisfies a given pattern?

- retrieval of records
- linking data across multiple data sources
- descriptive statistics
- report generation, summaries
- visualisations

Uses of computation

What is computation for?

Much of computer science concerns our attempts to coerce computers into accomplishing *tasks*, loosely definable as:

- An identified source of data.
- An identified context or situation.
- A desired outcome.

The data may be created for the task, or might be derived from the physical world – transformed, by a device, into bits from entities or events in our universe.

A context might be a specific piece of hardware or operating system, or might be assumptions such as “the numbers represent prices” or “the text is in ASCII”.

An outcome might be a number, an action, a list of results, . . .

Uses of computation

Computers and algorithms were originally developed to solve what might be called *concrete* tasks. For example (tiny selection):

- Compute a missile trajectory.
- Crack a code (decryption).
- Do accountancy over financial data.
- Operate a camera (focus, exposure), store the image.
- Guide a cutting tool, operate an assembly line.
- Map mouse movements to cursor movements.

In common: the task is well-defined, we can assess whether the solution is correct.

In these tasks, the data is transformed in a mechanical way or leads to a mechanical action, but only in a very limited way do they enhance our (that is, human) knowledge.

Hence – not “knowledge technologies”.

From Data to Information

Not

What data satisfies a given pattern?

but

What patterns satisfy this data?

(Actually, we want to find *interesting* and *robust* patterns that satisfy the data.)

Finding patterns

Patient	Age	#Medications	Complication
1	52	7	Yes
2	57	9	Yes
3	43	6	Yes
4	33	6	No
5	35	8	No
6	49	8	Yes
7	58	4	No
8	62	3	No
9	48	0	No
10	37	6	Yes

Source: Vasant Dhar "Data Science and Prediction" (2013) Communications of the ACM, Vol. 56 No. 12, Pages 64-73

doi:10.1145/2500499

Finding patterns

Patient	Age	#Medications	Complication
1	52	7	Yes
2	57	9	Yes
3	43	6	Yes
4	33	6	No
5	35	8	No
6	49	8	Yes
7	58	4	No
8	62	3	No
9	48	0	No
10	37	6	Yes

Age >= 37
AND
#Medications >= 6
→
Complication = Yes (100% confidence)

Source: Vasant Dhar "Data Science and Prediction" (2013) Communications of the ACM, Vol. 56 No. 12, Pages 64-73

doi:10.1145/2500499

Knowledge tasks

Consider tasks where the data is irregular or unreliable, or the outcome is not well-defined:

- Compression of an image.
- List of answers to a typical web query.
- Translation between languages.
- Identification of what a health condition might be caused by; identification of a treatment.
- Finding an "optimal" route between two locations. (Optimal? Distance, time, stress, fuel?)
- Deciding what movie to watch.

"What movie to watch?" (Or music to buy, or place to visit, or . . .)

This is not a computational task – but we do use computers to *mediate* between us and data, in helping to reach a decision.

Context is critical: the origin of the data, the consumer of the output.

These use, produce, or enhance human knowledge.

Knowledge task example: Translation

Why is translation between languages not well-defined?

Because translation leads to loss of meaning, and loss of nuance and "feel".

Test: ask a human to translate from English to some other language; ask another human to translate back.

I tried via Chinese:

In: *Consider tasks where the outcome is not well-defined*

Out: *Consider work where the result is unclear*

Why would we expect a machine to do better?

What *is* a correct translation?

Paraphrasing Julian Barnes (London Review of Books, 18 Nov 2010), imagine reading a famous 1850s French novel in English.

What do you want? Probably, that it provoke the same reactions in you as in a French reader.

But what about the topical references that only a French speaker would know?

Or the glaring errors concerning English culture that a French speaker wouldn't notice?

And what English? An attempt at 1850s English, with disused forms of expression, or modernised?

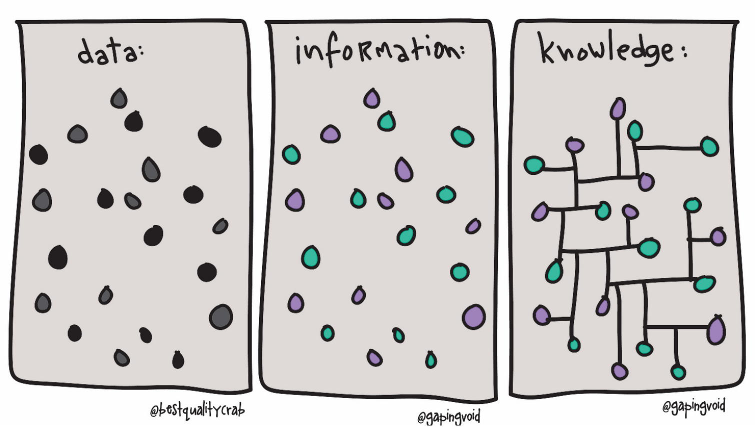
And what judgements about class and education? (The two societies were not equivalent.)

Are trousers held up by braces, or pants held up by suspenders?

From Data to Wisdom

Data serves as the raw material for creation of new knowledge.

From Data to Wisdom



Characteristics of knowledge technologies

Knowledge technologies tend to be either fairly general (e.g., machine learning algorithms) or fairly specific (e.g., machine translation).

General: the data must be transformed to suit the axioms or assumptions of the method, in a rigorous way.

Specific: detailed understanding of the task is used to drive development of the method, perhaps by drawing on a toolkit of components and of solutions to similar problems.

- A specialized problem: parse a particular language.
- An approximate problem: assign a document to a category.
- A general problem: find features of the data items that discriminate between categories.

A broader perspective

Knowledge technologies are occasionally transformational.

Innovations of this kind are often designed to assist or augment an existing activity, but the consequence is that they displace it entirely.

Think of:

- The impact of social networking on email. (NB: slack is explicitly designed to reduce email)
- The impact of search on libraries and encyclopaedias.
- The impact of blogging (and tweeting) on reportage and newspapers.

Computational thinking

Finding solutions to tasks requires application of computational thinking:

- How should the data be represented?
- How should it be manipulated?
- What heuristics or simplifications can be safely applied?
- Can the problem be transformed or rearranged in a way that usefully changes the costs?
- Does it have properties that let it be addressed by sorting?
- Does it have properties that let it be addressed by searching?
- It is possible to eliminate the need to consider global properties, allowing a focus on local properties? That is, does all of the data have to be considered holistically, or can it be divided in some way?
- How will a solution behave as the data approaches boundary conditions? (Increase or decrease in number of errors; data items unique or frequently repeated; as item size or item number grows, . . .)

Thinking tools for knowledge technologies

Consider *effectiveness* rather than *correctness*.
(Can a document ranking possibly be “correct”?)

Identify features and characteristics that can be quantified.

Identify approximations to the task.

Consider whether the outcome is likely to seem plausible or appealing.

Whether it makes sense to consider training data from which tailored solutions can be automatically learnt. (Which may make a solution easy, but may make it difficult to gain insight into the problem.)

Ask: What does signal look like? What does noise look like?
What would a human do, given sufficient stamina and memory?
What output would a human produce?

Is a human part of the loop in some way? How is the output to be consumed?

Example: All of these questions apply to aspects of web search.

Methods for data analysis

Supervised learning

- Classification
  predicting a discrete class
- Regression
  predicting a numeric quantity

Unsupervised learning

- Association
  detecting associations between features
- Information organisation; Clustering
  grouping similar instances into clusters
- Reinforcement learning
- Recommender systems
- Anomaly/outlier detection

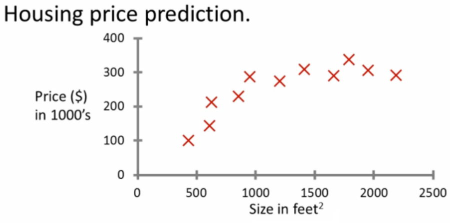
Example: Supervised Learning (Classification)

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
⋮	⋮	⋮	⋮	⋮

Given information about current weather conditions and the forecast, can we determine whether we will go out to play?

Example: Supervised Learning (Regression)

Can we predict housing prices?



A friend has a house which is 750 square feet – how much can he expect to get?

(draw a straight line vs. fit a curve)

Summary

- What is a database?
- What is knowledge technology?
- What are some of its challenges?
- How to get from data to wisdom?

Next: Document representation and string processing