



MONASH University

Information Technology

FIT5190 Introduction to IT Research Methods

Lecture 3

Literature Review – Setting the Context

Slides prepared by

David Green, Frada Burstein, Jacques Steyn, Geoff Webb, Chung-Hsing Yeh

Learning objectives

- Understand
 - the need to know the context of a research study
 - the need for a literature review
 - the main elements of a good literature review
- Be able to
 - define the scope of a literature review
- Know
 - how to search for information
 - how to cite and reference correctly

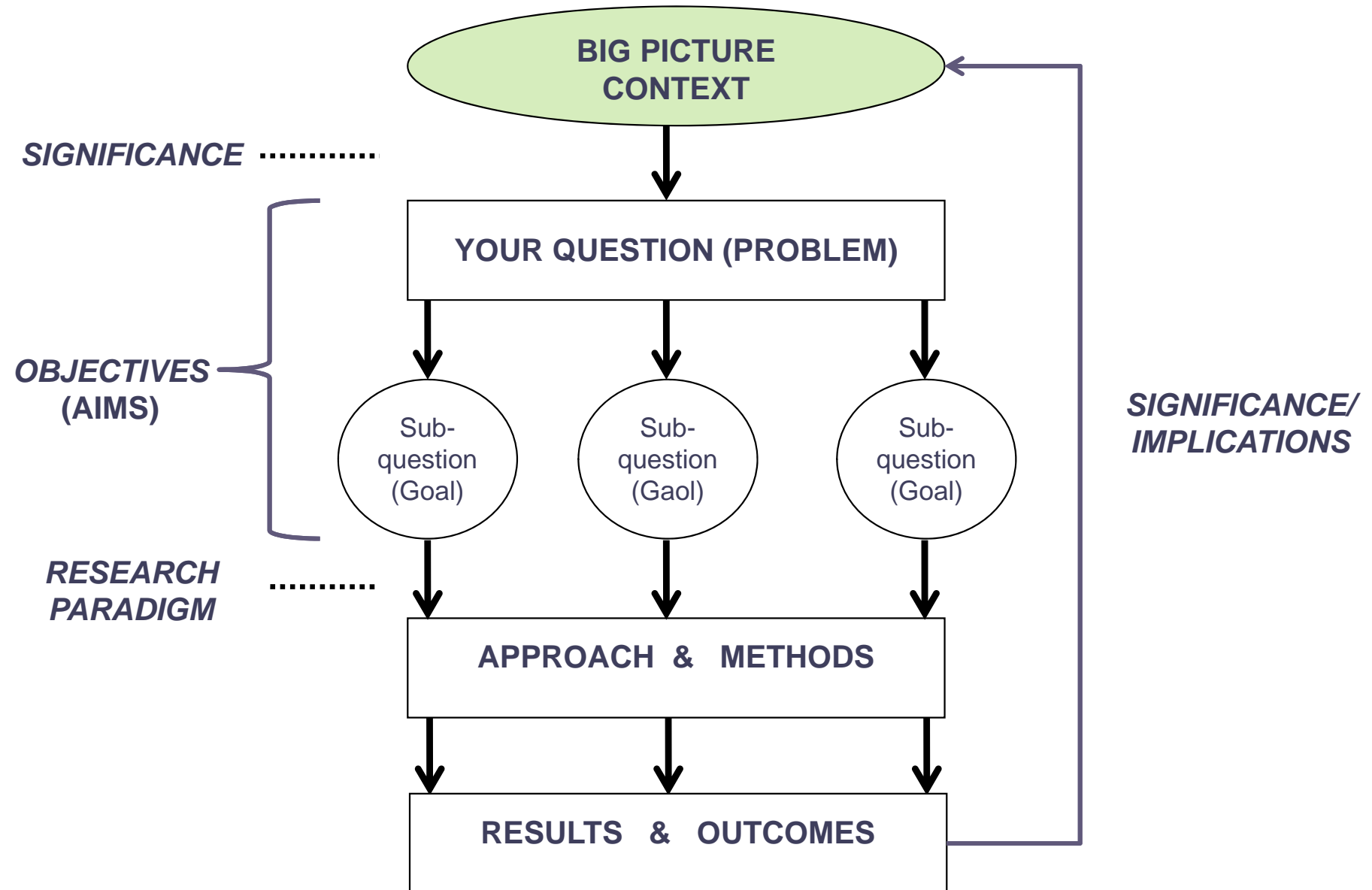
Overview

- In this lecture we address a vital early stage in any research project - understanding the academic literature that provides the background and context for a study.
- The lecture will cover both thematic and bibliometric analysis.
- In addition, the nature and quality of published evidence will be covered.

Overview (continued)

- The tutorial introduces a powerful literature reference and citation management software - EndNote.
- The tutorial also involves practical exercises in identifying major journals for a particular research area, searching for publications on given topics, analysis of journal papers, and identifying of the impact of these papers.

Research centres around questions



What is the context?

- Research questions sit within a context
 - Social, technical or other needs that inspired the question
 - Previous research
 - Relevant paradigms and methodologies
 - Potential implications
- These contexts provide the big picture
 - Implications for what you do and how you do it.
 - Implications for how you explain
 - Why it is important (why you are doing it)
 - What you have done

Example - context

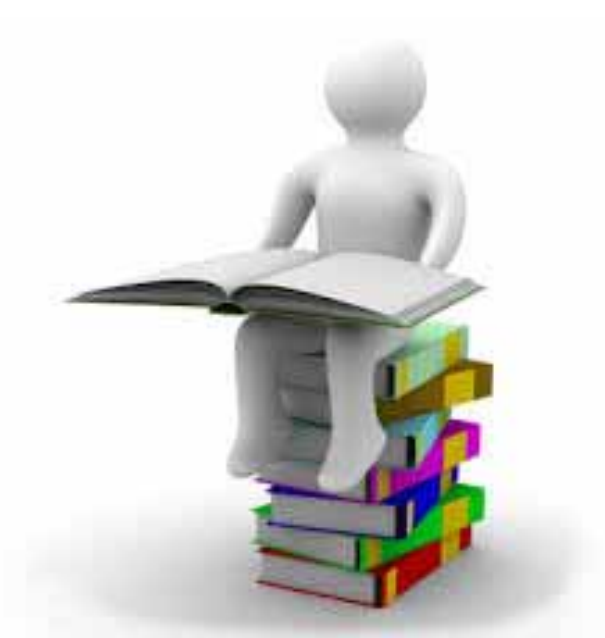
- Question
 - How to make a better wheel?
- Bigger Picture/Context
 - Economic: Need to reduce transport costs
 - Social: Easier to maintain
 - Technical: Efficiency, durability, materials, construction
 - Environment: Reduce wear and tear
 - Previous research: Existing wheel technologies

Another example - context

- Question
 - How to make a database that answers queries faster?
- Bigger Picture/Context
 - Economic: Business efficiency and profitability
 - Technical: Flaws or gaps in existing methods
 - Human: Useability, Interface

Literature review

- Purpose
 - Set the context of a project
- A literature review ...
 - Explains motivation
 - Provides background
 - Lays out the theoretical basis
 - Explains the project's significance
- Two main kinds of review
 - Context for a particular study (e.g. thesis)
 - Survey of research within a field



Literature review for a thesis

- Give the background needed to understand the study
 - Terminology
 - Techniques
- Explain the rationale for the study
 - Questions and issues that underlie the project
 - Outlines important issues, needs and gaps
 - Sets out the theoretical and practical framework
- Related research
 - Synthesis of relevant published work
 - Summarize recent advances and their implications
 - Critically evaluate contribution and shortcomings of previous research
- A review is NOT a textbook or primer on a subject
 - Critical discussion and independent thought
 - Sets out the author's vision of the field

What makes a good review?

- More than a list of papers describing their content
- Provides frameworks for understanding
 - the key themes
 - the main approaches
 - the important controversies and
 - the evolution of knowledgewithin a field of research
- Critically evaluates work
 - Fits it into contexts and frameworks
 - Identifies issues, controversies, gaps, flaws and directions for future research

What makes a good review?

A bad review

- Presents old ideas
- Just lists people's results
- Repeats ideas of others
- Reads like a textbook
- Presents a narrow view

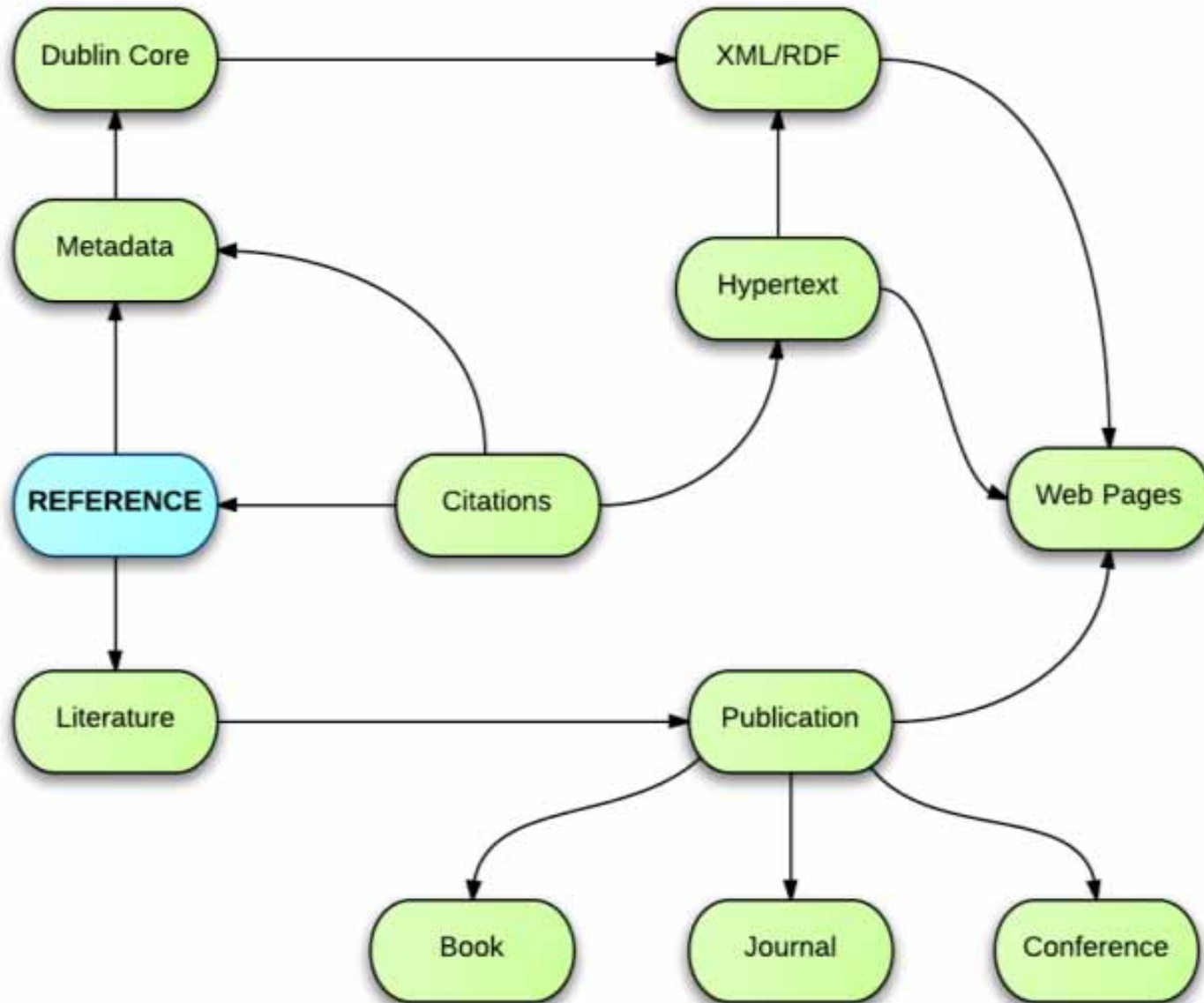
A good review

- Presents a fresh outlook
- Identifies gaps and flaws
- Presents independent ideas
- Critically evaluates work
- Gives wider perspective

Concept networks

- Show relationships between ideas
- In a literature review
 - Useful tools to map out relevant ideas and issues
- Concept networks include:
 - Semantic networks
 - Mind maps
 - Entity relationships diagrams
 - Object models

Example: mind map for “reference”



Finding sources

Starting points

- Published articles
 - Reference list
 - Key words
- Related topics
 - Online searches
 - **Monash Library – Online resources - Databases**
 - Google, Google Scholar, Web of Science, Scopus, ProQuest (ABI/Inform), IEEE Xplore, CiteSeer.



Example

A source article

The emergence of social consensus in simulation studies with Boolean networks

David G. Green, Tania G. Leishman¹ and Suzanne Sadedin

Clayton School of Information Technology,
Monash University, Clayton 3800, Australia
david.green@infotech.monash.edu.au

Abstract

Interactions between individuals play a crucial part in the maintenance of social order and group identity. We examine three Boolean network simulations that explore the roles of network structure and peer influence in social cohesion and consensus. These studies reveal several unifying characteristics of social opinion formation, including phase transitions, positive feedback and cluster formation. Results suggest that there is an upper limit to the size of social networks within which peer interactions can produce universal consensus via the formation of self-maintaining local clusters. In large societies, both enforcement and peer pressure are likely to be needed to achieve full social cohesion. In the absence of enforcement, peer pressure can either reinforce cooperative behaviour, or flip the entire society into rebellion. In poorly connected social networks disobedience is consistently high. However, provided a social network is well-connected, even a small incidence of punishment ensures conformity. Similarly, social connectivity slows the spread of conflicting mass media influences until media opinions reach a threshold of prevalence, but above this threshold, social connectivity facilitates the spread of media opinions. Overall, the results show that the frequency and distribution of interactions among individuals are powerful drivers of the dynamics of social consensus and dissent.

1. Introduction

One of civilization's great achievements has been to create large and relatively stable communities that may consist of many millions of individuals. To achieve order in large, permanent social groups of this kind requires a high degree of cooperation by individuals. The importance of social order is illustrated by situations in which it breaks down. Riots, anarchy and other outbreaks of civil unrest highlight the fragile nature of social cooperation and cohesion. Social consensus can reduce these conflicts. At the same time, maintaining diversity of opinions is important for creativity, innovation and cultural evolution.

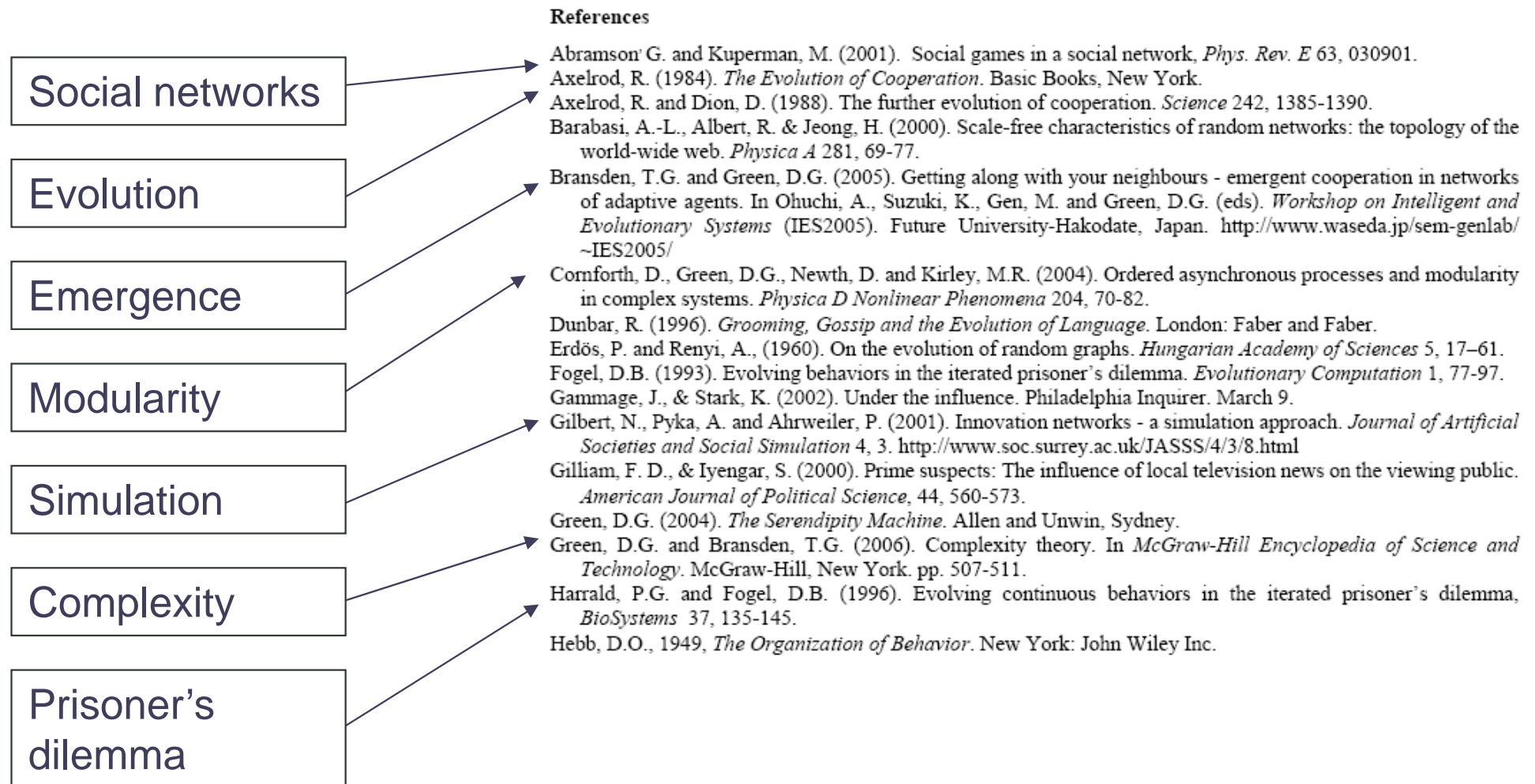
Many models of the emergence of social order approach the issue from the perspective of game theory. Game

Reference:

Green, D.G., Leishman, T.G. and Sadedin, S. (2006). The emergence of social consensus in simulation studies with Boolean networks. In Takahashi, S., Sallach, D., Rouchier, J. (eds). *Proceedings of the First World Congress on Social Simulation*, Kyoto. Vol. 2, pp. 1-8

References


Relevant terms?



Reference:

Green, D.G., Leishman, T.G. and Sadedin, S. (2006). The emergence of social consensus in simulation studies with Boolean networks. In Takahashi, S., Sallach, D., Rouchier, J. (eds). *Proceedings of the First World Congress on Social Simulation*, Kyoto. Vol. 2, pp. 1-8

A Google Scholar search

 **Advanced Scholar Search** [Advanced Search Tips](#) | [About Google Scholar](#)

Find articles
with **all** of the words
with the **exact phrase**
with **at least one** of the words
without the words
where my words occur

Innovation networks

in the title of the article ▾

100 results ▾

Search Scholar

Author
Return articles written by

"N Gilbert"

e.g., "PJ Hayes" or McCarthy

Publication
Return articles published in

e.g., J Biol Chem or Nature

Date
Return articles published between

—

e.g., 1996

Subject Areas

☒ Return articles in all subject areas.
☐ Return only articles in the following subject areas:

- ☐ Biology, Life Sciences, and Environmental Science
- ☐ Business, Administration, Finance, and Economics
- ☐ Chemistry and Materials Science
- ☐ Engineering, Computer Science, and Mathematics
- ☐ Medicine, Pharmacology, and Veterinary Science
- ☐ Physics, Astronomy, and Planetary Science
- ☐ Social Sciences, Arts, and Humanities

A Google Scholar search (continued)



[Web](#) [Images](#) [Video](#) [News](#) [Maps](#) [more »](#)

allintitle: "Innovation networks" author:"N Gilb

Search

[Advanced Scholar Search](#)

[Scholar Preferences](#)

[Scholar Help](#)

Scholar All articles - [Recent articles](#)

Results 1 - 11 of 11 for allintitle: "Innovation networks" author:"N Gilbert". (0.09 seconds)

[All Results](#)

[N Gilbert](#)

[P Ahrweiler](#)

[A Pyka](#)

[J Vaux](#)

[Innovation Networks-A Simulation Approach](#) - [Check for full text](#) - [all 7 versions »](#)

N Gilbert, A Pyka, P Ahrweiler - Journal of Artificial Societies and Social Simulation, 2001 - [jasss.soc.surrey.ac.uk](#)

© Copyright JASSS JASSS logo. Nigel Gilbert, Andreas Pyka and Petra Ahrweiler (2001).

Innovation Networks - A Simulation Approach. ... **Innovation networks**. ...

[Cited by 31](#) - [Related Articles](#) - [Cached](#) - [Web Search](#)

[CITATION] **Innovation Networks** by Design: The Case of Mobile VCE

J Vaux, **N Gilbert** - **Innovation Networks**: Theory and Practice, Cheltenham, Edward ... , 2002

[Cited by 7](#) - [Related Articles](#) - [Web Search](#)

[CITATION] Simulating Knowledge Dynamics in **Innovation Networks**

P Ahrweiler, A Pyka, **N Gilbert** - Industry and Labor Dynamics: The Agent-based Computational ... , 2004

[Cited by 4](#) - [Related Articles](#) - [Web Search](#)

[PDF] [Innovation Networks: A Policy Model](#) - [all 3 versions »](#)

N Gilbert, P Ahrweiler, A Pyka, REI Glen - SEIN Project paper, 1999 - [rand.org](#)

Page 1. **Innovation networks**: a policy model **Innovation networks**: a policy model Nigel Gilbert Nigel Gilbert University of Surrey ...

[Cited by 3](#) - [Related Articles](#) - [View as HTML](#) - [Web Search](#)

[CITATION] Simulating **Innovation Networks**

A Pyka, NG Gilbert, P Ahrweiler - **Innovation Networks**: Theory and Practice, Cheltenham, Edward ... , 2002

[Cited by 3](#) - [Related Articles](#) - [Web Search](#)

[SIMULATING KNOWLEDGE DYNAMICS IN INNOVATION NETWORKS \(SKIN\)](#) - [all 10 versions »](#)

P AHRWEILER, A PYKA, **N GILBERT** - Industry And Labor Dynamics: The Agent-based Computational ... , 2004 - [books.google.com](#)

SIMULATING KNOWLEDGE DYNAMICS IN **INNOVATION NETWORKS** 1 (SKIN) PETRA AHRWEILER Research

Centre Media and Politics, Institute for Political Science, University of ...

[Cited by 1](#) - [Related Articles](#) - [Web Search](#)

OPEN

A Google Scholar search (continued)

Google Scholar BETA

[Web](#) [Images](#) [Video](#) [News](#) [Maps](#) [more »](#)

[Advanced Scholar Search](#)
[Scholar Preferences](#)
[Scholar Help](#)

Scholar Results 1 - 31 of 31 citing [Gilbert: Innovation Networks-A Simulation Approach](#). (0.16 seconds)

All Results

[H DAWID](#)
[K Wersching](#)
[R Garcia](#)
[N Gilbert](#)
[???](#)

[AGENT-BASED MODELS OF INNOVATION AND TECHNOLOGICAL CHANGE](#) - [all 4 versions »](#)
H DAWID - Handbook of Computational Economics, 1996 - books.google.com
Chapter 25 AGENT-BASED MODELS OF INNOVATION AND TECHNOLOGICAL CHANGE HERBERT
DAWID * Department of Business Administration and Economics and Institute
of Mathematical Economics, Bielefeld University, Bielefeld, Germany e-mail: ...
[Cited by 22](#) - [Related Articles](#) - [Web Search](#)

[Innovation and knowledge spillover with geographical and technological distance in an agentbased ...](#) - [all 3 versions »](#)
K Wersching - 2005 - savannah-simulations.com
The paper introduces an agent-based simulation model to study the ... Keywords:
Innovation, Learning, Knowledge Spillover, Agent-based ... * BIELEFELD
UNIVERSITY, Department of Business Administration and Economics, PO Box 10 ...
[Cited by 6](#) - [Related Articles](#) - [View as HTML](#) - [Web Search](#)

[Uses of Agent-Based Modeling in Innovation/New Product Development Research](#) - [Check for full text](#) - [all 3 versions »](#)
R Garcia - Journal of Product Innovation Management, 2005 - Blackwell Synergy
Little has been written in the new product development literature about the
simulation technique agent-based modeling, which is a by-product of recent
explorations into complex adaptive systems in other disciplines. ...
[Cited by 7](#) - [Related Articles](#) - [Web Search](#)

[Agent-based social simulation: dealing with complexity](#) - [all 5 versions »](#)
N Gilbert - Retrieved from <http://www.complexityscience.org/NoE/ABSS-...> - ccsr.ac.uk
While the idea of computer simulation has had enormous influence on most areas
of science, and even on the public imagination through its use in computer games
such as SimCity, it took until the 1990s for it to have a significant ...
[Cited by 4](#) - [Related Articles](#) - [View as HTML](#) - [Web Search](#)

[????????????????](#)
???, ??? - ?????, 2003 - ???
21?5?2003?10?????Studies in Science of Science
Vo1.21 No.5 Oct.2003 ????:1003—2053(2003105—0546—06
????????????????,??? ...
[Cited by 4](#) - [Related Articles](#) - [Web Search](#)

OPEN

A Google Scholar search (continued)

Agent-based social simulation: dealing with complexity

Nigel Gilbert

Centre for Research on Social Simulation
University of Surrey
Guildford
UK

n.gilbert@soc.surrey.ac.uk

18 December 2004

While the idea of computer simulation has had enormous influence on most areas of science, and even on the public imagination through its use in computer games such as SimCity, it took until the 1990s for it to have a significant impact in the social sciences. The breakthrough came when it was realised that computer programs offer the possibility of creating 'artificial' societies in which individuals and collective actors such as organisations could be directly represented and the effect of their interactions observed. This provided for the first time the possibility of using experimental methods with social phenomena, or at least with their computer representations; of directly studying the emergence of social institutions from individual interaction; and of using computer code as a way of formalising dynamic social theories. In this chapter, these advances in the application of computer simulation to the social sciences will be illustrated with a number of examples of recent work, showing how this new methodology is appropriate for analysing social phenomena that are inherently complex, and how it encourages experimentation and the study of emergence.

Start of the downloaded article



Start of its reference list



References

- Anderson, J. R. and Lebiere, C. (1998) *The Atomic Components of Thought*. Erlbaum, Mahwah, NJ.
- Arthur, W.B. (1989) 'Competing Technologies, Increasing Returns, and Lock-In by Historical Events', *Economics J.* vol.116, p.99
- Axelrod, R. (1995) 'A model of the emergence of new political actors'. In N. Gilbert & R. Conte (Eds.), *Artificial Societies*. London: UCL Press.
- Beck, K. (1999) *Extreme Programming Explained*. Addison-Wesley, Boston, MA.
- Booch, G., Rumbaugh, J. and Jacobson, I. (2000) *The Unified Modeling Language User Guide*. 6th print edn. Addison-Wesley, Reading, MA.
- Breiger, R., Kathleen Carley and Philippa Pattison (eds.) (2003) *Dynamic Social Network Modelling and Analysis: Workshop Summary and Papers* Washington: The National Academies Press.
- Bruch, Elizabeth E. and Robert D. Mare (2003) 'Neighborhood Choice and Neighborhood Change'. Presented at the Annual Meeting of the Population Association of America, Minneapolis, GA, May 2003.

Primary literature

- Original sources of data/ideas
 - Periodicals (journals)
 - Conference proceedings
 - Thesis/dissertations
 - Reports
 - Government publications
 - Trade/commercial literature
 - Patents
 - Standards

Secondary literature

- Material distilled from primary sources
 - Books
 - Review articles
 - Reference material (e.g. encyclopaedias)

Guides to the literature

- Resources to assist in retrieval of information from primary and secondary sources:
 - Abstracting and indexing journals
 - Guides to abstracting and indexing journals
 - Bibliographies
 - Bibliographies of bibliographies
 - Unit guides
 - General guides

Some starting resources

- Libraries
 - Subject keyword indexes
 - <https://my.monash.edu.au/library>
- Google Advanced Search
 - http://www.google.com.au/advanced_search
 - Useful for finding links to general sources
 - Beware of credibility and bias!!
 - Be aware of the “filter bubble”
 - Google lists first what it “thinks” you want to see
- Wikipedia
 - <http://en.wikipedia.org>
 - Useful as first step and overview articles
 - NOT definitive, comprehensive or reliable!
 - Should NOT be quoted as an authoritative source

Citation searches

- Google Scholar

<http://scholar.google.com>

- Web of Science (ISI)

<https://login.ezproxy.lib.monash.edu.au/login?qurl=http%3a%2f%2fisiknowledge.com>

- Scopus

<https://login.ezproxy.lib.monash.edu.au/login?qurl=http%3a%2f%2fwww.scopus.com>

- CiteSeer

<http://citeseer.ist.psu.edu>

- DBLP

<http://www.informatik.uni-trier.de/~ley/db/>

– DBLP is a useful source of bibliographic information

Journals and conferences

- Most IT primary literature in journals and conferences
- Journal
 - Periodical publication
 - Peer reviewed
 - Papers subject to revision and refinement
 - Papers submitted when research is ready for publication
- Conference
 - Meeting where researchers present their research to one another
 - Papers are published in proceedings
 - May be peer reviewed
 - Papers submitted to deadline
 - Limited opportunity for revision

Publication quality

- Both journals and conferences vary in quality
 - Conference participants pay fees
 - Some journals charge publication fees
 - Acceptance rates vary greatly
 - Some conferences and journals accept almost any submission
- ISI Journal Citation Reports – impact factor
 - Science citation index (SCI)
 - Social science citation index (SSCI)



Publication quality (continued)

- Journal and conference rankings
- Note that ranks are subjective and may vary with discipline
 - <http://lamp.infosys.deakin.edu.au/era/>
 - <http://www.journal-ranking.com/>
 - <http://webdocs.cs.ualberta.ca/~zaiane/htmldocs/ConfRanking.html>
 - <http://core.edu.au/index.php/categories/conference%20rankings>
 - <http://core.edu.au/index.php/categories/journals>
 - <http://www.acphs.org.au/index.php/is-journal-ranking>

Publication quality (continued)

- Motivations for publishing vary greatly
 - Academics and many students under great pressure to publish
 - Fraud sometimes occurs
 - Same work published many times
 - Other people's work copied
 - Results fabricated
- Researchers' training and knowledge of the field vary
 - Work may be methodologically flawed
 - Conclusions may be unjustified
 - Work may ignore the state-of-the-art

Indicators of publication quality

- Rank of publication venue
 - But still some rubbish published in good venues and some good papers in poor venues
- Citation counts
 - But these can be inflated
 - Self citation
 - Communities tend to cite one another
 - Citations depend on the work coming to the community's attention

Format of a paper

First page

- Essential metadata

Bibliographic details

Mach Learn (2012) 86:233–272
DOI 10.1007/s10994-011-5263-6

Learning by extrapolation from marginal to full-multivariate probability distributions: decreasingly naive Bayesian classification

Title

Authors

Geoffrey I. Webb · Janice R. Boughton · Fei Zheng · Kai Ming Ting · Houssam Salem

Received: 8 December 2009 / Accepted: 15 September 2011 / Published online: 13 October 2011
© The Author(s) 2011

Abstract

Abstract Averaged n -Dependence Estimators ($AnDE$) is an approach to probabilistic classification learning that learns by extrapolation from marginal to full-multivariate probability distributions. It utilizes a single parameter that transforms the approach between a low-variance high-bias learner (Naive Bayes) and a high-variance low-bias learner with Bayes optimal asymptotic error. It extends the underlying strategy of Averaged One-Dependence Estimators (AODE), which relaxes the Naive Bayes independence assumption while retaining many of Naive Bayes' desirable computational and theoretical properties. $AnDE$ further relaxes the independence assumption by generalizing AODE to higher-levels of dependence. Extensive experimental evaluation shows that the bias-variance trade-off for Averaged 2-Dependence Estimators results in strong predictive accuracy over a wide range of data sets. It has training time linear with respect to the number of examples, learns in a single pass through the training data, supports incremental learning, handles directly missing values, and is robust in the face of noise. Beyond the practical utility of its lower-dimensional variants, $AnDE$ is of interest in that it demonstrates that it is possible to create low-bias high-variance generative learners and suggests strategies for developing even more powerful classifiers.

Keywords

Keywords Bayesian learning · Classification learning · Probabilistic learning · Averaged one-dependence estimators · Naive Bayes · Semi-naive Bayesian learning · Learning without model selection · Ensemble learning · Feating

Contacts

Editor: Peter Flach.
G.I. Webb (✉) · J.R. Boughton · F. Zheng · K.M. Ting · H. Salem
Faculty of Information Technology, Monash University, Clayton, VIC, 3809, Australia
e-mail: Geoff.Webb@monash.edu
J.R. Boughton
e-mail: Janice.Boughton@monash.edu
K.M. Ting
e-mail: Kaiming.Ting@monash.edu

Format of a paper

Body

Introduction

1 Introduction

This paper presents a family of learning algorithms that utilize a predefined function to extrapolate from observed marginal distributions to the full multivariate distribution of interest. This stands in contrast to the majority of learning algorithms that instead seek to fit a model directly to the observed multivariate probability distribution. Whereas learning is sometimes cast as a problem of searching through a space of hypotheses to find one that best fits the training data (Mitchell 1982), this new approach does not employ search and does not perform model selection.

The members of this new family of algorithms have a unique combination of features that is well suited to many applications. We discuss these features in more detail below. Notable amongst them are training complexity linear with respect to the number of training examples; single pass learning; direct capacity for incremental learning; and accuracy that is competitive with the state-of-the-art. They are of further theoretical interest because they demonstrate that it is possible to create low bias generative learners.

The family contains algorithms that range from low variance coupled with high bias through to high variance coupled with low bias. Successive members of the family will be best suited to differing quantities of data, starting with low variance for small data, with successively lower bias but higher variance suiting ever increasing data quantities (Brain and Webb 2002). The asymptotic error of the lowest bias variant is Bayes optimal.

One member of this family of algorithms, naive Bayes (NB), is already well known. A second member, Averaged One-Dependence Estimators (AODE) (Webb et al. 2005), has enjoyed considerable popularity since its introduction in 2005 (Nikora 2005; Camporelli 2006; Flikka et al. 2006; Orhan and Altan 2006; Lasko et al. 2006; Hunt 2006; Ferrari and Aitken 2006; Birzele and Kramer 2006; Kunchevaa et al. 2007; Lau et al. 2007; Masegosa et al. 2007; Wang et al. 2007; Garcia et al. 2008; Tian et al. 2008; Kurz et al. 2009; Leon et al. 2009; Shahri and Jamil 2009; Simpson et al. 2009; Affendey et al. 2010; García-Jiménez et al. 2010; Hopfgartner et al. 2010; Liew et al. 2010). The work presented in this paper arises from the realization that NB and AODE are but two instances of a family of algorithms, which we call *AnDE*.

In Sect. 2 we explain the underlying learning strategy, and define the *AnDE* family of algorithms. The *AnDE* family of algorithms build upon the method pioneered by AODE (Webb et al. 2005). In Sect. 3 we discuss how the *AnDE* algorithms relate to Feating (Ting et al. 2011), a generic approach to ensembling that also builds upon techniques pioneered by AODE. In Sect. 4 we present an extensive evaluation of the *AnDE* family of algorithms, comparing their performance to relevant Bayesian techniques, to Feating and to the state-of-the-art Random Forests classifier. Section 5 presents conclusions and directions for future research.

2 The *AnDE* family of algorithms

We wish to estimate from a training sample T of t classified objects the probability $P(y | \mathbf{x})$ that an example $\mathbf{x} = \langle x_1, \dots, x_s \rangle$ belongs to class y , where x_i is the value of the i th attribute and $y \in \{c_1, \dots, c_k\}$. We use \bar{v} to denote the average number of values per attribute. These and other elements of notation are listed in Table 1.

From the definition of conditional probability we have

$$P(y | \mathbf{x}) = P(y, \mathbf{x}) / P(\mathbf{x}) \quad (1)$$

Format of a paper

Conclusion

Table 10 Win/draw/loss: $AnDE$, $n = 0, 1$ and 2 , vs RF10 and RF100 on all 62 data sets

		$AnDE$ vs RF10		$AnDE$ vs RF100	
		W/D/L	p	W/D/L	p
A2DE	Bias	18/1/43	0.001	22/2/38	0.026
	Variance	57/0/5	<0.001	45/1/16	<0.001
	Zero-one loss	42/0/20	0.004	36/3/23	0.059
	RMSE	40/0/22	0.015	35/0/27	0.187
AODE	Bias	16/0/46	<0.001	20/0/42	0.004
	Variance	57/1/4	<0.001	47/0/15	<0.001
	Zero-one loss	41/0/21	0.008	33/1/28	0.304
	RMSE	39/0/23	0.028	34/0/28	0.263
NB	Bias	14/1/47	<0.001	16/1/45	<0.001
	Variance	56/0/6	<0.001	51/0/11	<0.001
	Zero-one loss	33/0/29	0.352	30/1/31	0.500
	RMSE	30/0/32	0.450	28/0/34	0.263

Table 11 Win/draw/loss: $AnDE$, $n = 0, 1$ and 2 , vs RF10 and RF100 on the ten largest data sets

		$AnDE$ vs RF10		$AnDE$ vs RF100	
		W/D/L	p	W/D/L	p
A2DE	Zero-one loss	2/0/8	0.055	1/1/8	0.020
	RMSE	3/0/7	0.172	2/0/8	0.055
AODE	Zero-one loss	0/0/10	0.001	0/0/10	0.001
	RMSE	1/0/9	0.011	0/0/10	0.001
NB	Zero-one loss	0/0/10	0.001	0/0/10	0.001
	RMSE	0/0/10	0.001	0/0/10	0.001

for very large training data, in the absence of any prior knowledge of the nature of the multi-variate probability distribution that the data embodies, Random Forests are likely to achieve lower error than an $AnDE$ classifier, although the data quantity at which this is achieved will be ever greater as the dimensionality of $AnDE$ is increased.

However, Random Forests' error advantage for large data comes at a cost in training time. Figure 6 shows the training and classification times for AODE, A2DE, RF10 and RF100. It is apparent that, overall, RF100 has very high training times. While A2DE's training time does approach RF100's for high dimensional data, for small data and low dimensional data its training times are competitive with RF10. On the other hand, A2DE requires substantially more classification time on average than Random Forests. This requirement grows greatly with high-dimensional data. A2DE will not be feasible for classification of large numbers of high-dimensional objects. In contrast, its classification time is very competitive on low-dimensional data.

5 Conclusions and directions for future research

$AnDE$ provides an attractive framework for developing machine learning techniques. A single parameter n controls a bias-variance trade-off such that $n = a$ provides a classifier whose

Format of a paper

References

References

- Affendey, L., Paris, I., Mustapha, N., Sulaiman, M., & Muda, Z. (2010). Ranking of influencing factors in predicting students' academic performance. *Information Technology Journal*, 9(4), 832–837.
- Birzele, F., & Kramer, S. (2006). A new representation for protein secondary structure prediction based on frequent patterns. *Bioinformatics*, 22(21), 2628–2634.
- Brain, D., & Webb, G. I. (2002). The need for low bias algorithms in classification learning from large data sets. In *Proceedings of the sixth European conference on principles of data mining and knowledge discovery (PKDD)* (pp. 62–73). Berlin: Springer.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.
- Camporelli, M. (2006). *Using a Bayesian classifier for probability estimation: analysis of the AMIS score for risk stratification in myocardial infarction*. Diploma thesis, Department of Informatics, University of Zurich.
- Cerquides, J., & Mántaras, R. L. D. (2005). Robust Bayesian linear classifier ensembles. In *Proceedings of the sixteenth European conference on machine learning* (pp. 70–81).
- Cestnik, B. (1990). Estimating probabilities: a crucial task in machine learning. In *Proceedings of the ninth European conference on artificial intelligence* (pp. 147–149). London: Pitman.
- Domingos, P., & Pazzani, M. J. (1996). Beyond independence: conditions for the optimality of the simple Bayesian classifier. In *Proceedings of the thirteenth international conference on machine learning* (pp. 105–112). San Mateo: Morgan Kaufmann.
- Fayyad, U. M., & Irani, K. B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the thirteenth international joint conference on artificial intelligence* (pp. 1022–1029). San Mateo: Morgan Kaufmann.
- Ferrari, L. D., & Aitken, S. (2006). Mining housekeeping genes with a naive Bayes classifier. *BMC Genomics*, 7(1), 277.
- Flikka, K., Martens, L., Vandekerckhove, J., Gevaert, K., & Eidhammer, I. (2006). Improving the reliability and throughput of mass spectrometry-based proteomics by spectrum quality filtering. *Proteomics*, 6(7), 2086–2094.
- Flores, M., Gámez, J., Martínez, A., & Puerta, J. (2009). GAODE and HAODE: two proposals based on AODE to deal with continuous variables. In *Proceedings of the 26th annual international conference on machine learning* (pp. 313–320). New York: ACM.
- Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, 29(2), 131–163.
- Garcia, B., Aler, R., Ledezma, A., & Sanchis, A. (2008). Protein-protein functional association prediction using genetic programming. In *Proceedings of the tenth annual conference on genetic and evolutionary computation* (pp. 347–348). New York: ACM.
- García-Jiménez, B., Juan, D., Ezkurdia, I., Andrés-León, E., & Valencia, A. (2010). Inference of functional relations in predicted protein networks with a machine learning approach. *PLoS ONE*, 4, e9969.
- Hopfgartner, F., Urruty, T., Lopez, P., Villa, R., & Jose, J. (2010). Simulated evaluation of faceted browsing based on feature selection. *Multimedia Tools and Applications*, 47(3), 631–662.
- Hunt, K. (2006). *Evaluation of novel algorithms to optimize risk stratification scores in myocardial infarction*. PhD thesis, Department of Informatics, University of Zurich.
- Jiang, L., & Zhang, H. (2006). Weightily averaged one-dependence estimators. In *PRICAI 2006: trends in artificial intelligence* (pp. 970–974).
- Kuncheva, L. I., Vilas, V. J. D. R., & Rodríguez, J. J. (2007). Diagnosing scrapie in sheep: a classification experiment. *Computers in Biology and Medicine*, 37(8), 1194–1202.
- Kurz, D., Bernstein, A., Hunt, K., Radovanovic, D., Erne, P., Siudak, Z., & Bertel, O. (2009). Simple point-of-care risk stratification in acute coronary syndromes: the AMIS model. *British Medical Journal*, 35(8), 662.
- Langley, P., & Sage, S. (1994). Induction of selective Bayesian classifiers. In *Proceedings of the tenth conference on uncertainty in artificial intelligence* (pp. 399–406). San Mateo: Morgan Kaufmann.
- Lasko, T. A., Atlas, S. J., Barry, M. J., & Chueh, K. H. C. (2006). Automated identification of a physician's primary patients. *Journal of the American Medical Informatics Association*, 13(1), 74–79.
- Lau, Q. P., Hsu, W., Lee, M. L., Mao, Y., & Chen, L. (2007). Prediction of cerebral aneurysm rupture. In *Proceedings of the nineteenth IEEE international conference on tools with artificial intelligence* (pp. 350–357). Washington: IEEE Computer Society.
- Leon, A., et al. (2009). EcID: A database for the inference of functional interactions in *E. coli*. *Nucleic Acids Research*, 37, D629 (Database issue).
- Liew, C., Ma, X., & Yap, C. (2010). Consensus model for identification of novel PI3K inhibitors in large chemical library. *Journal of Computer-Aided Molecular Design*, 24(2), 131–141.

Referencing

- Referencing normally consists of 2 parts:
 - **Citations:** Links to information within text

and webb 2002). The asymptotic error of the lowest bias variant is Bayes optimal.

One member of this family of algorithms, naive Bayes (NB), is already well known. A second member, Averaged One-Dependence Estimators (AODE) (Webb et al. 2005), has enjoyed considerable popularity since its introduction in 2005 (Nikora 2005; Camporelli 2006; Flikka et al. 2006; Orhan and Altan 2006; Lasko et al. 2006; Hunt 2006; Ferrari and Aitken 2006; Birzele and Kramer 2006; Kunchevaa et al. 2007; Lau et al. 2007; Masegosa et al. 2007; Wang et al. 2007; Garcia et al. 2008; Tian et al. 2008; Kurz et al. 2009; Leon et al. 2009; Shahri and Jamil 2009; Simpson et al. 2009; Affendey et al. 2010; García-Jiménez et al. 2010; Hopfgartner et al. 2010; Liew et al. 2010). The work presented in this paper arises from the realization that NB and AODE are but two instances of a family of algorithms.

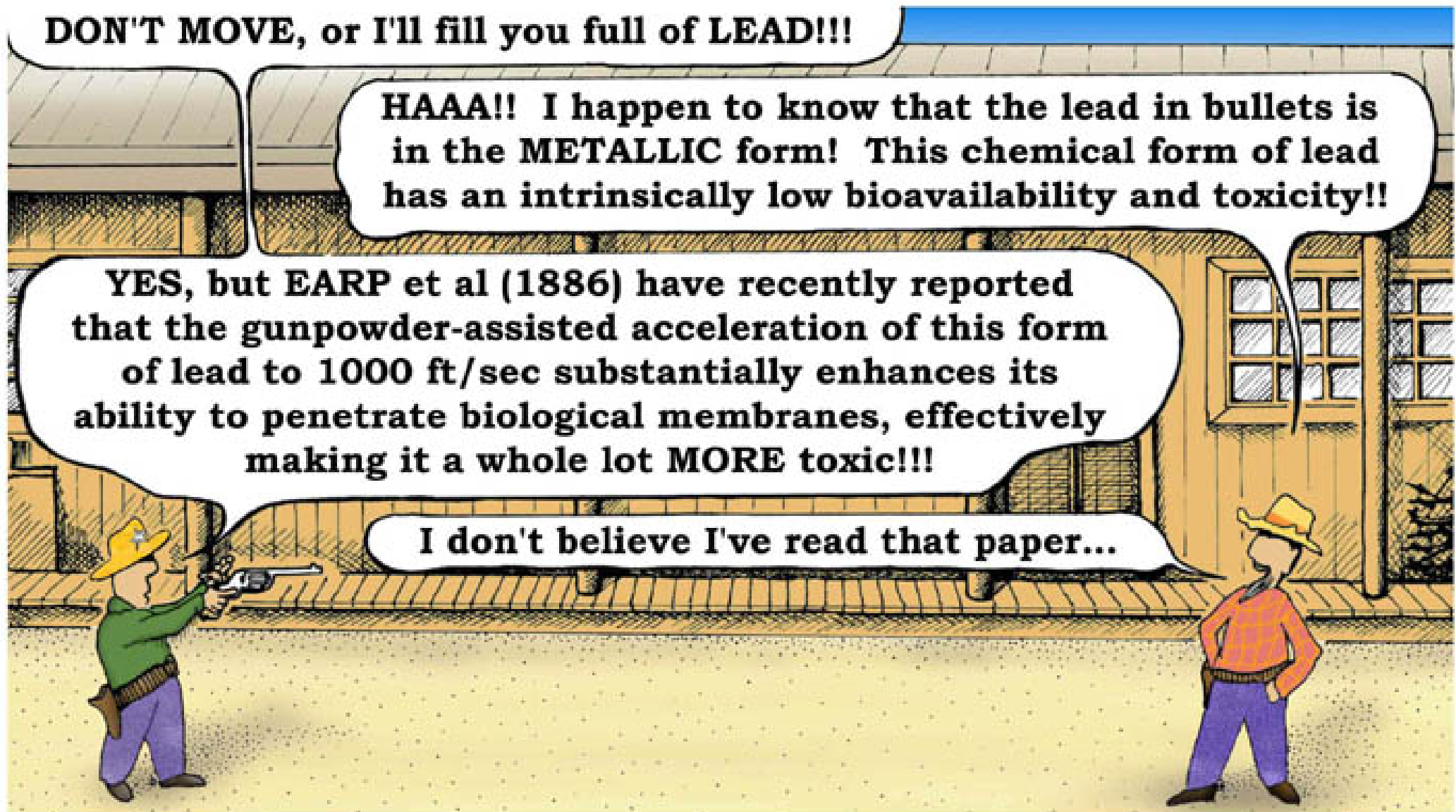
- **References:** Details of the sources

References

Affendey, L., Paris, I., Mustapha, N., Sulaiman, M., & Muda, Z. (2010). Ranking of influencing factors in predicting students' academic performance. *Information Technology Journal*, 9(4), 832–837.

Birzele, F., & Kramer, S. (2006). A new representation for protein secondary structure prediction based on frequent patterns. *Bioinformatics*, 22(21), 2628–2634.

Brain, D., & Webb, G. I. (2002). The need for low bias algorithms in classification learning from large data sets. In *Proceedings of the sixth European conference on principles of data mining and knowledge*



ENVIRONMENTAL SCIENTISTS IN THE WILD WEST

Source:

Nearing Zero (Science cartoons)
<http://www.lab-initio.com/>

References and bibliography

- Follow the journal style exactly.
- References are published works that are readily accessible.
- References are those specifically referenced (cited) in the text of the article.
- Do not reference anything that is common knowledge.
- A bibliography lists works that are found to be relevant to the research topic. Bibliographies are seldom necessary or even allowed in scholarly articles.

Purposes of references

References are used primarily to

- Acknowledge significant sources;
- Articulate where the field is positioned currently;
- Place your research work in context;
- Demonstrate your understanding of the existing state of knowledge;
- **Support your research work.**

Use of references

- **Use references to**

- Justify and support your arguments;
- Allow you to make comparisons with other research;
- Express matters better than you could have done so;
- Demonstrate your familiarity with your field of research.

- **Do not use references to**

- Impress your readers with the scope of your reading;
- Replace the need for you to explore your own thoughts;
- Litter your writing with names and quotations.

Referencing styles

- **Two main styles are widely used:**
 - **Author-Date** (Harvard and APA)
 - e.g. (Brown, 2008)
 - **Numbered** (Vancouver [..], IEEE [..])
 - e.g. [1]
 - See the unit Moodle site for
 - Example papers with difference referencing styles.
 - The differences between the Harvard and APA (American Psychological Association) systems.

Some useful references

- **Monash Library tutorials**
 - <http://lib.monash.edu/tutorials/citing/>
 - <http://www.lib.monash.edu.au/tutorials/citing/infotech.html>
 - <http://www.lib.monash.edu.au/tutorials/citing/compsci.html>
- **Vancouver System**
 - Commonly used in medical and scientific journals
 - <http://lib.monash.edu/tutorials/citing/vancouver.html>
- **American Psychological Association (APA)**
 - Common in life sciences
 - <http://lib.monash.edu/tutorials/citing/apa.html>
- **Chicago Manual of Style**
 - <http://lib.monash.edu/tutorials/citing/chicago.html>

Chicago Manual of Style

- **Documentary style (Footnote style)**

- <http://library.williams.edu/citing/styles/chicago1.php>
- Commonly used in humanities, arts, history
- Citations use **numbers**, e.g. [5]
 - “The connectivity avalanche [5] is the source ...”
- References in order cited, e.g.
[5] Nairn, T., Faces of Nationalism: Janis Revisited, Verso, London and New York, 1997.

- **Author-date style**

- <http://library.williams.edu/citing/styles/chicago2.php>
- Commonly used in social sciences and sciences
- Citations use **author and date**, e.g. Nairn (1997).
- References in alphabetical order. e.g.

Nairn, T. (1997). Faces of Nationalism: Janis Revisited, Verso, London and New York.

Problems with citing

- Time consuming to track down and copy details of each reference, e.g. exact journal name, volume, issue and page numbers, publication year, authors' initials, etc.
 - helps to keep a database of all papers that you read
 - get in the habit of collecting bibliographic information at the time you access a paper
 - can often download bibliographic information from publisher's site
 - can download bibliographic information from Google Scholar, but may not be correct

Problems with citing

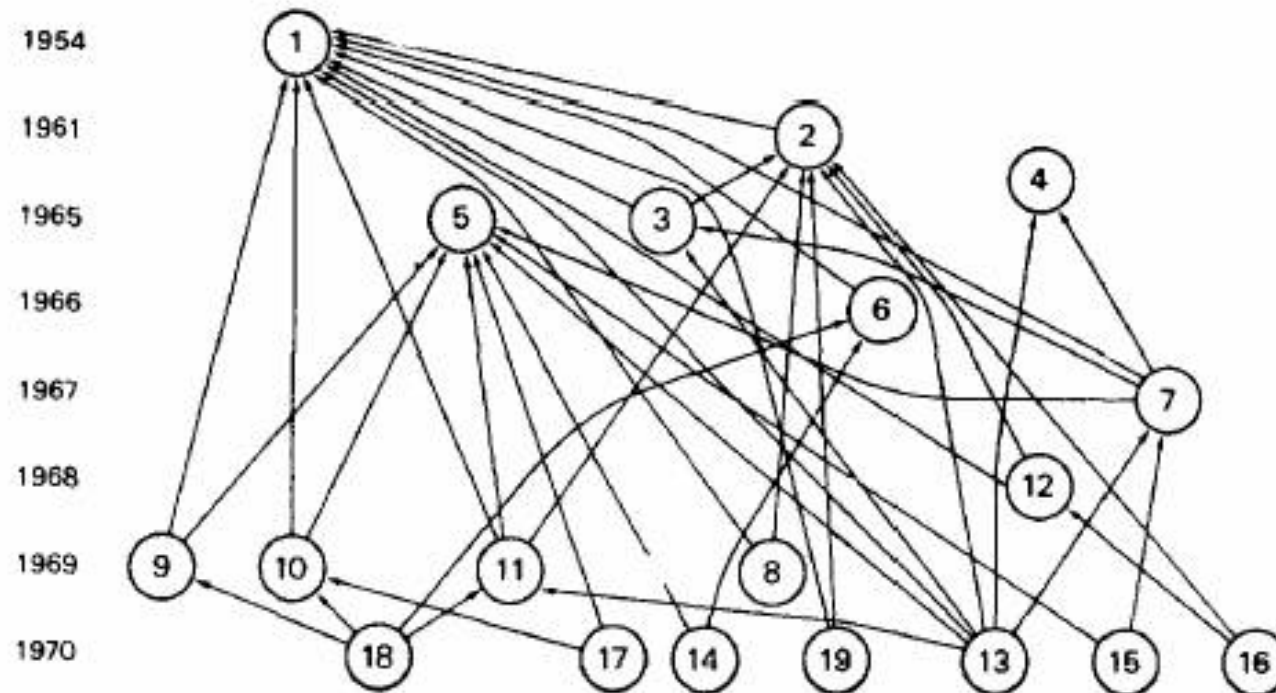
Almost every journal uses a slightly different variation on the basic referencing styles

- Tedious to learn and apply details of each different style
- Referencing in citation order is difficult
- Time-consuming to reformat
- Much easier to use automated tools
 - EndNote (A reference and citation management software)
 - BibTeX



Citation networks

- Ideas form concept networks, citations can help to reveal them
- Citations in articles form networks ...Cawkell (1971) p. 814
...



Cawkell, A. E. (1971). Science Citation Index: Effectiveness in locating articles in the anaesthetics field: 'Perturbation of ion transport'. *British Journal of Anaesthesia*, 43: 814.

Impact of research



- The influence of a study on a field of research or broader community
- Often measured by the number of citations
 - Note that the older a paper is, the more citations it should get
- Data derived from Google Scholar, ISI, Citeseer
- **Impact factor** (ISI Journal Citation Reports)
 - Indicator of a journal's impact
 - Average number of citations in the reference year per article published in the two previous years

Reading a book

- Look for an **introduction**, concluding chapter, **abstract or executive summary** and read it briefly.
- Read the **table of contents** and identify any chapters relevant to your research, again starting from the introduction and/or conclusion.
 - You can find your way through a chapter or section by using the subheadings.
- Look for **an index** for specific points you are interested in (terms, people, events, etc.) and locate them in the text from the index.
- In the text itself, key points are often highlighted, or placed in the first or last paragraphs.
 - The **first and last sentences of paragraphs** are often used to indicate and summarise their contents.

Reading a paper

- An **abstract** should tell you what a paper is about
- The **introduction** and **conclusion** should summarise all the major points
- You may not need to read the body to determine whether a paper is relevant to your project

Critical reading



- Go beyond mere description by **offering opinions**, and making a response, to what has been written.
- Relate different writing to each other, indicating their differences and contradictions, and highlighting what they are lacking.
- Do not take what is written at face value.
- Strive to be explicit about the values and theories which inform and colour reading and writing.
- Take up **alternative views** and positions in research writing.

– Blaxter et al. (2010), p. 118.

Summary

- Read widely
 - Build your research upon the state-of-the-art
- Read critically
 - Research papers are written by people, often students like you, and are ultimately expressions of their opinions, not statements of fact...
- Keep a database of references
 - You will need the bibliographic information when you write your thesis
 - EndNote
 - BibTeX

Readings

- **Essential**

- Monash Library tutorials on referencing styles

- <http://lib.monash.edu/tutorials/citing/>

- **Recommended**

- Tutorials on Literature reviews (see also the unit Moodle site)

- <http://www.monash.edu.au/lis/lonline/writing/general/lit-reviews/index.xml>

- <http://www.writing.utoronto.ca/advice/specific-types-of-writing/literature-review>

- <http://library.ucsc.edu/ref/howto/literaturereview.html>

- Blaxter, L., Hughes, C., Tight, M. (2010). How to Research, 4th Edition. Open University Press, Maidenhead, Chapter 4.

- Creswell (2009). Chapter 2.

- Neuman (2000). Chapter 16 (or equivalent chapter in a later edition).

- Smith, A.J. (1990). The task of the referee. IEEE Computer, 23 (4), 65-73.

- Webster, J., Watson, R.T. (2002). Analyzing the past to prepare for the future: Writing a literature review. MIS Quarterly, 26 (2), xiii-xxiii.