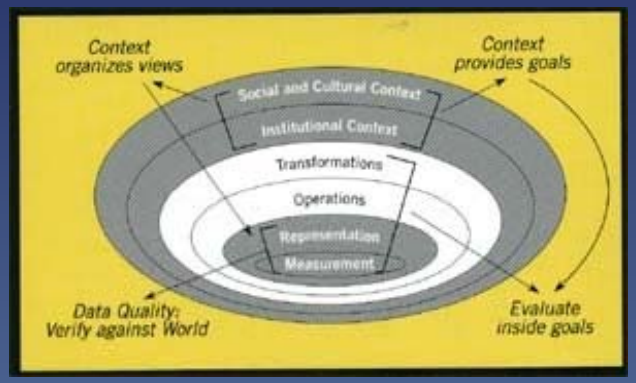
A difference that makes a difference

Reflections on 30 + years

in the field of Spatial Data Quality

Nicholas Chrisman
Sciences géomatiques, Université Laval
Québec, Québec, CANADA
Scientific Director, GEOIDE Network

Why the title?



A story of forgetting a citation…

A difference that makes a difference…

Catchy?

Challenge for ISSDQ 2009

Panel on Monday afternoon

N. Chrisman (moderator)

Questions from conference organizers:

- **Top five achievements in SDQ research**
- **Top five failures in SDQ research**
- **Top five research opportunities**

My opportunity to set the stage…

2 achievements, 3 failures, unlimited opportunities

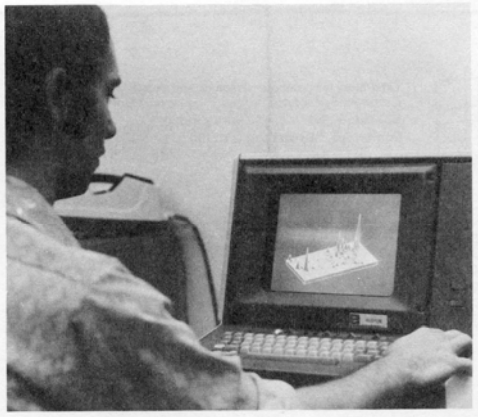
(plus or minus a few here or there…)

Outline of presentation

- Back to 30 years ago (but not for long!)
  - A few of the achievements and failures
- Revisiting the I word: Information
  - Connections to SDQ
  - Directions for future research

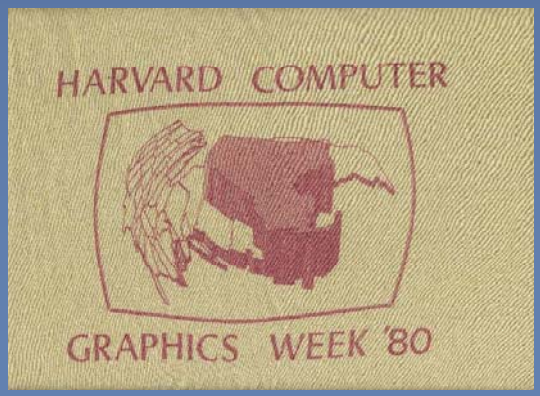
About 30 years ago

Chrisman fiddles with high technology (Carousel slide projector), *Manilla 1976*



JIM LITTLE AT THE CONTROLS

Chrisman and Morehouse ca 1979



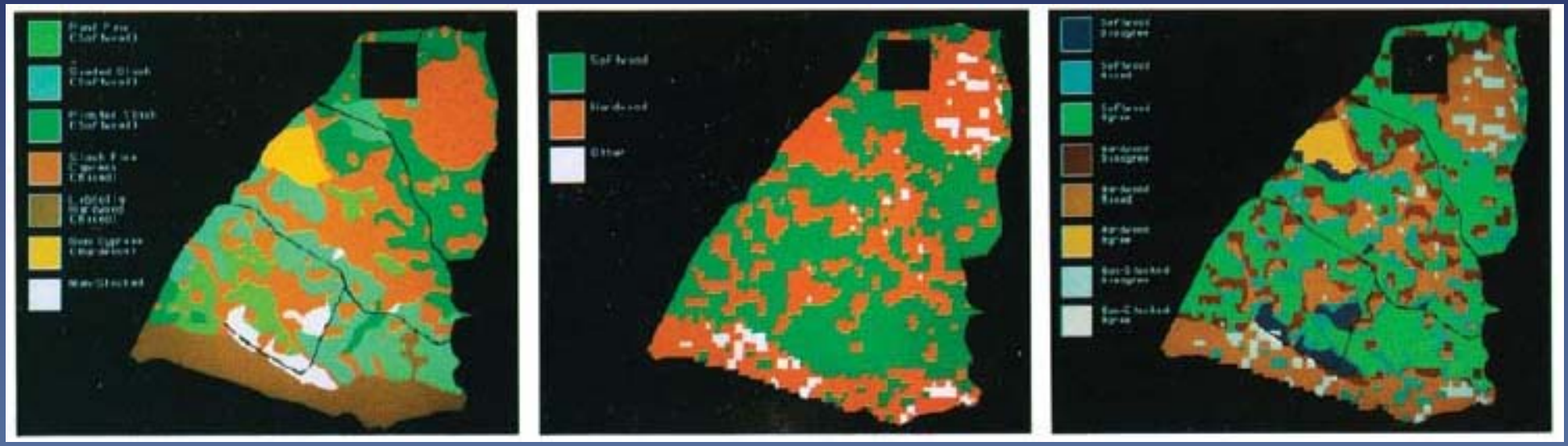
Where were we in 1979?

Able to make maps,
using expensive equipment,
that looked like what could
be made by hand…

AND

Rudimentary GIS analysis -
polygon overlay

The challenge:
following an overlay…



1. Slivers (recognized as huge plague in CGIS)
2. How to distinguish between results and inevitable errors?

– ***Plus or minus what?***

– Here, Map 1 (forest inventory) overlaid on Map 2 (Landsat classification) => Map 3 (dark areas disagreement / change)

– *(1978: St Regis project, Harvard LCGSA)*

Slivers:

- What is a sliver?
  - o Small, inconsequential *(maybe)*
  - o Thin *(actually not)*
  - o Caused by two independent versions of same line…

The solution (quick fix?) we implemented: filter the overlay- ***fuzzy tolerance***

- ➢Covers up the variability;
- ➢no error estimate

All wrapped together

- An achievement
  - Software reduced slivers from eating up the whole database
- A flaw
  - loss of data (potential information)
- An opportunity
  - where to mine for forgotten treasure

A side issue: Experimental approach in SDQ

- A significant proportion of papers show empirical results, but how many are experimental?
  - (with deliberately introduced error)
- for example: Lester, M.K. and Chrisman, N.R., 1991: *Not all slivers are skinny: A comparison of two methods for detecting positional errors in categorical maps,* Proceedings GIS/LIS 91, 2:648-658.

An opportunity

- Why work to reduce error?
  - Go make some well-controlled errors.
  - See if they match the empirical case.

Chrisman PhD 1982

- Grappled with errors in overlaid maps
  - Epsilon model (buffer around lines)
    - *error bounds on overlay results*
  - Multivariate categorical data analysis using the estimated error matrix as link equation
    - *totally impractical…*

One major failure in SDQ

Comparison of Old Growth mapped by USFS (rows) versus Wilderness Society (columns)
Mt. Baker-Snoqualmie National Forest (from Norheim,1996)

	Low elevation old growth	Other ancient forest	High elevation old growth	Other than ancient forest	Total
Low elevation old growth	209,935	69,168	7,714	96,834	383,650
% of array	12%	4%	0%	6%	22%
% of column	71%	50%	5%	8%	
% of row	55%	18%	2%	25%	
High elevation old growth	6,689	4,371	108,664	118,964	238,689
% of array	0%	0%	6%	7%	14%
% of column	2%	3%	75%	10%	
% of row	3%	2%	46%	50%	
Other than old growth	77,428	64,574	29,182	932,192	1,103,376
% of array	4%	4%	2%	54%	64%
% of column	26%	47%	20%	81%	
% of row	7%	6%	3%	84%	
Total	294,052	138,113	145,559	1,147,990	1,725,715
% of array	17%	8%	8%	67%	

- Results of map overlay are still reported without error estimates.

One positive outcome

- Widespread recognition of SDQ
- The standards process *(1982- 2082?)* included Data Quality as a part of Metadata
  - (of course the metadata standard is pretty toothless…)

A key to Data Quality
The I word

- The middle word in GIS
- Not the one argued over...
- Yet crucial.

Extract from a paper:

Annals AAG, 2006

Order from Noise:

Towards a social theory of information

Barbara Poore, US Geological Survey

Nick Chrisman, GEOIDE

Two conflicting models of information

- Both arise from Cybernetics movement in late 1940s; Macy Conferences
- Shannon [Bell Labs] (publicized by Weaver)
  - Transmission model: encoding as bits
- Norbert Wiener [MIT, Lincoln Labs]
  - Control of aircraft gunnery, cybernetics
  - Feedback as main issue

Shannon's boxes and arrows

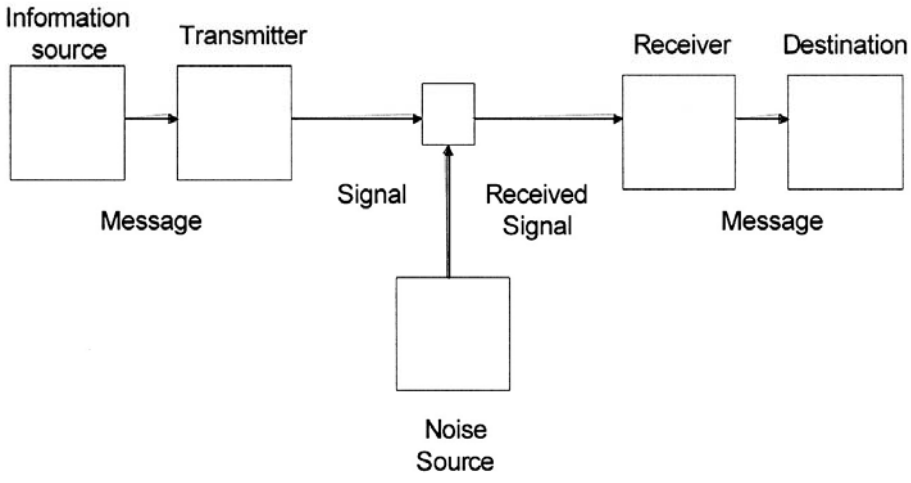
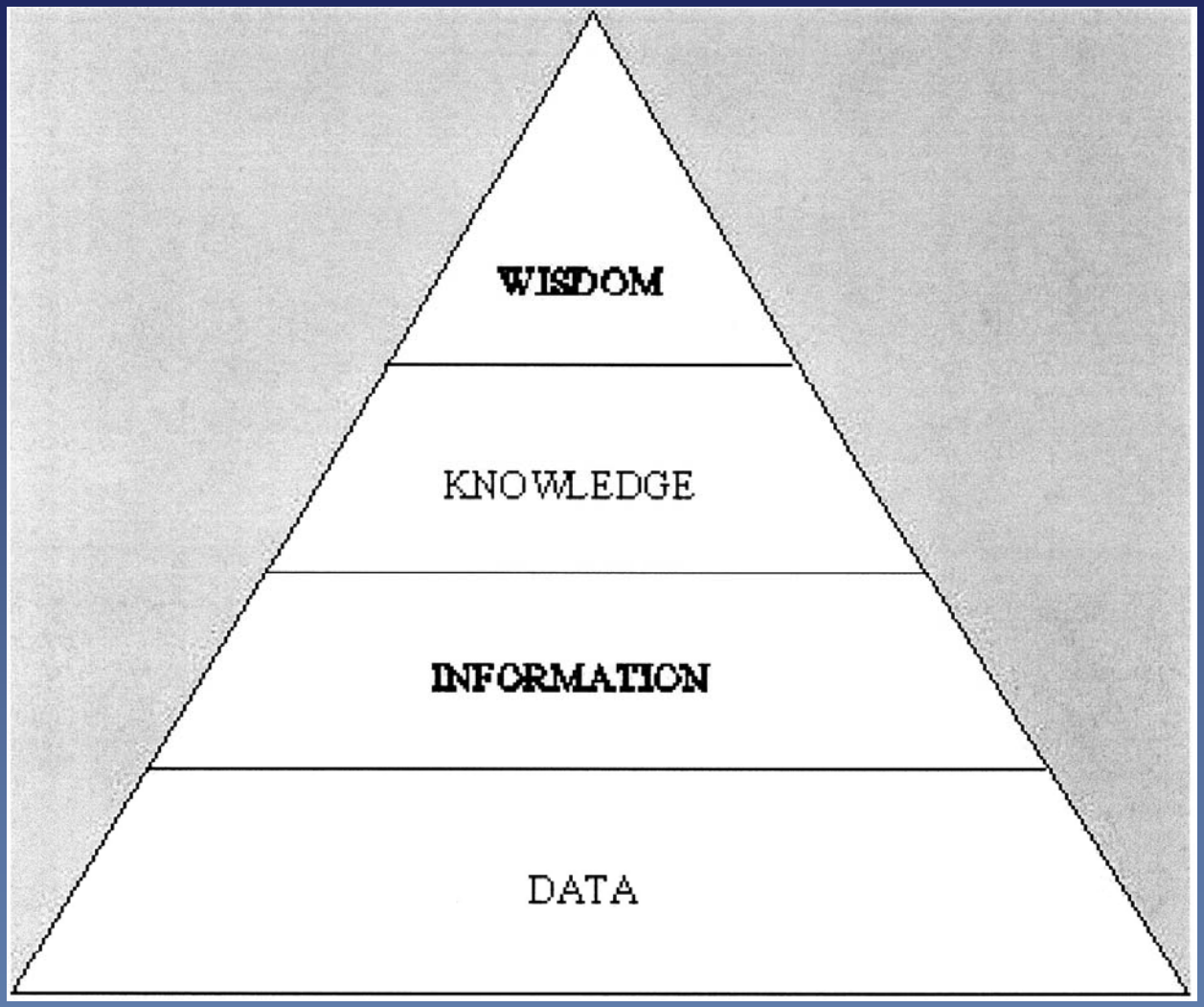


Figure 1. After Shannon and Weaver. 1949. The Mathematical Theor y of Communication. Chicago , University of Illinois Press p. 34

Invariance through transmission;

Meaning not a part of model

Model of refinement



Wiener promotes the mythology

DIKW persists to this day

Gillette's World Corporation (1910)

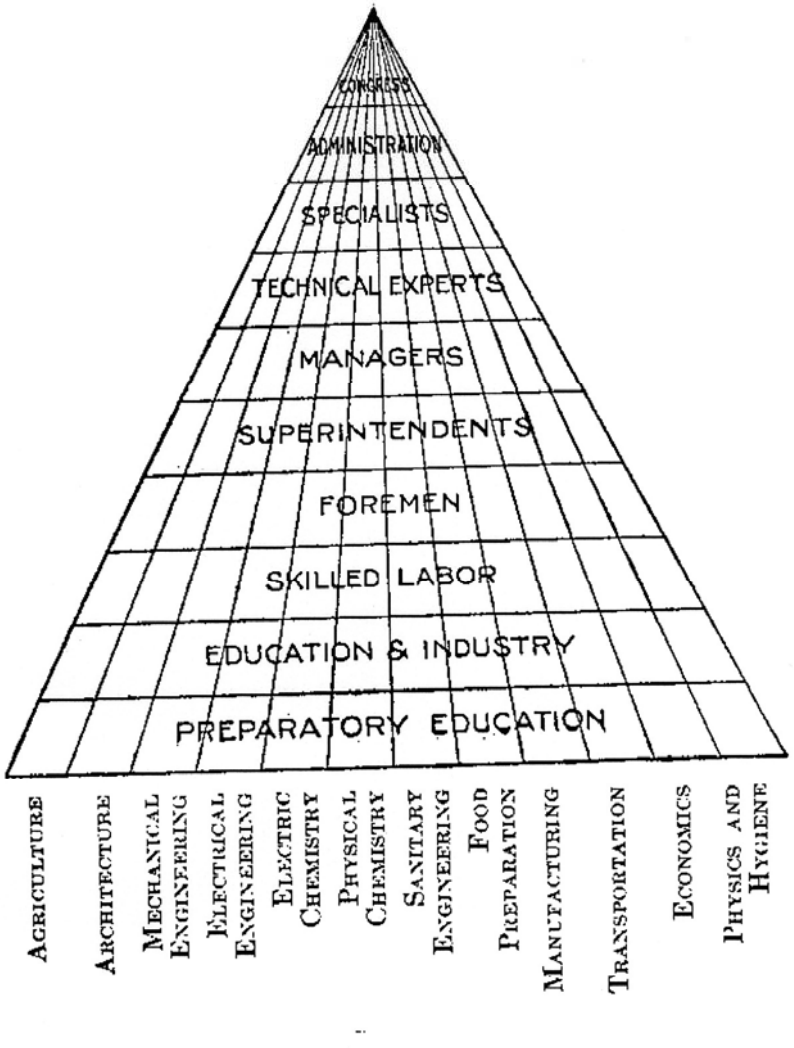


Figure 3. World Corporation from Gillette 1910 p. 85

Goodchild's Tower of Babel

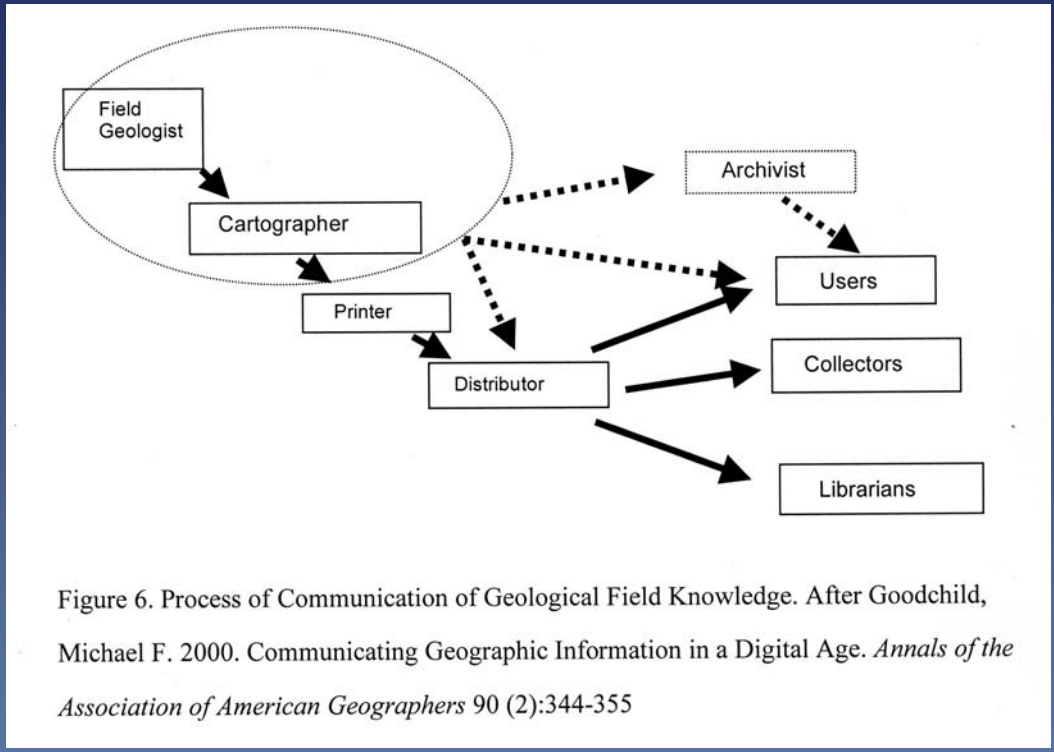


Figure 6. Process of Communication of Geological Field Knowledge. After Goodchild, Michael F. 2000. Communicating Geographic Information in a Digital Age. *Annals of the Association of American Geographers* 90 (2):344-355

Communication models

Attempts to combine transmission and refinement; a missing element

A social model of information

Goguen: *an interpretation of a configuration of signs for which some social group is accountable.*

Bateson: *“a difference that makes a difference”*

So, here is the citation:

Bateson, Gregory (1972). *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology*. University Of Chicago Press. ISBN 0-226-03905-6

And some context:

Gregory Bateson, a psychologist, married Margaret Mead. Participated with Shannon, Weaver, von Neumann, Weiner, and other in the Macy Conferences (1940s-50s) – which defined « cybernetics » and modern information theory. Count Korzybski (« The map is not the territory ») was not far away…

Bateson's experiment

no consent form needed!

Hold your hand perfectly still, palm upwards and resting comfortably on a table. With your other hand, drop a small coin into the palm.

You will feel the impact, and if the coin is cold, you will feel the coldness of the metal. Soon however, you will feel nothing. The nerve cells don't bother repeating themselves. They will only report to the brain when something changes.

Information is difference.

But a difference that makes a difference

But there are differences and differences…

As you dropped the coin into your palm,
your eyes told you automatically,
without your brain even asking,
what the value of the coin was;
but you were probably not aware what date it was minted.

This is because *(unless you are a numismatist)* the **value of the coin makes a difference to you** whereas its **date doesn't**.

A matter of perspective

Bateson continued with a story of lizard detecting an incoming insect on a leaf.

What is information for the lizard is not information for you, and what is information for you is not information for the lizard.

Perspective defines **what counts** as information at all, perspective defines **to whom** the information makes a difference.

And hence to Goguen's social interpretation of signs- and the importance of responsibility…

Back to Spatial Data Quality

- How the information model matters:
  - Treated as a “thing”, information is capital, a precious resource
  - BUT, information is the result of work, interactions of sensors and humans
  - Data Quality is NOT inherent, but the result of perspective (people, purpose, place…)

Limitations of transformational viewpoint

- Tobler's view on transformations:
  - focus on invariance; what is preserved
- BUT, it is exactly the opposite:
  - what is deliberately lost;
  - what does not matter and what does;
  - the difference that makes a difference

And a consequence (or two)

- van Foerster: « *you can turn a library upside down, but not a drop of information will flow out* »
- All writing requires an audience; meaning is interpreted by the reader
- SDQ: Metadata is not an end in itself
  - requires perspective of use

Crucial issue: responsibility

- Data Quality depends on ‘producers’ taking on the work of documentation - metadata
- and Users taking on the duty to read it and act accordingly
  - *one without the other is fruitless*

Understanding perspective

- The error matrix Norheim constructed:
  - Forest Service: Old Growth
  - Wilderness Society: Ancient Forest

- *will not be resolved automatically by some semantic ontology tool…*

The social moves to the center

- Recent ontology efforts mostly limited to ‘consenting parties’ – the easy stuff
- Data Quality research must seek the conflict, understand how perspective matters…
- *Build new tools to explore differences.*

Role of Networks

- Developments depend on mixing up perspectives.
- Interaction between disciplines is necessary.
- Network research groups have a greater chance to learn.

Conclusion

- Achievements and failures often come together; and point toward future challenges.
- Knowledge networks permit the interaction.

- ***Best wishes to ISSDQ for a fruitful interchange of perspectives***

Disclaimer

- The preceeding slide show was neither accurate, nor complete, and its quality is entirely in the hands of the audience.
- Any resemblence of the fictional characters referenced to well-known academics is not coincidental.
- No animals were hurt in the making of this entertainment.