

Text Mining Tools: Instruments for Scientific Discovery

**Marti Hearst
UC Berkeley SIMS**

**IMA Text Mining Workshop
April 17, 2000**

Outline

- What knowledge can we discover from text?
- How is knowledge discovered from other kinds of data?
- A proposal: let's make a new kind of scientific instrument/tool.

Note: this talk contains some common materials and themes from another one of my talks entitled "Untangling Text Data Mining"

What is Knowledge Discovery from Text?

What is Knowledge Discovery from Text?

- Finding a document?
- Finding a person's name in a document?

This information is already known to the author at least.



~~Needles in Haystacks~~

~~Needlestacks~~

What to Discover from Text?

- What news events happened last year?
- Which researchers most influenced a field?
- Which inventions led to other inventions?



Historical,
Retrospective

What to Discover from Text?

- What are the most common topics discussed in this set of documents?
- How connected is the Web?

Summaries
of the data
itself

- What words best characterize this set of documents' topics?
- Which words are good triggers for a topic classifier/filter?

Features
used in
algorithms

Classifying Application Types

	Patterns	Non-Novel Nuggets	Novel Nuggets
Non-textual data	Standard data mining	Database queries	AI Discovery Systems
Textual data	Computational linguistics	Information retrieval	Real text data mining

The Quandary

- How do we use text to both
 - Find new information not known to the author of the text
 - Find information that is not about the text itself?

Idea: Exploratory Data Analysis

- Use large text collections to gather evidence to support (or refute) hypotheses
 - Not known to author:
Make links across many texts
 - Not self-referential:
Work within the text domain

The Process of Scientific Discovery

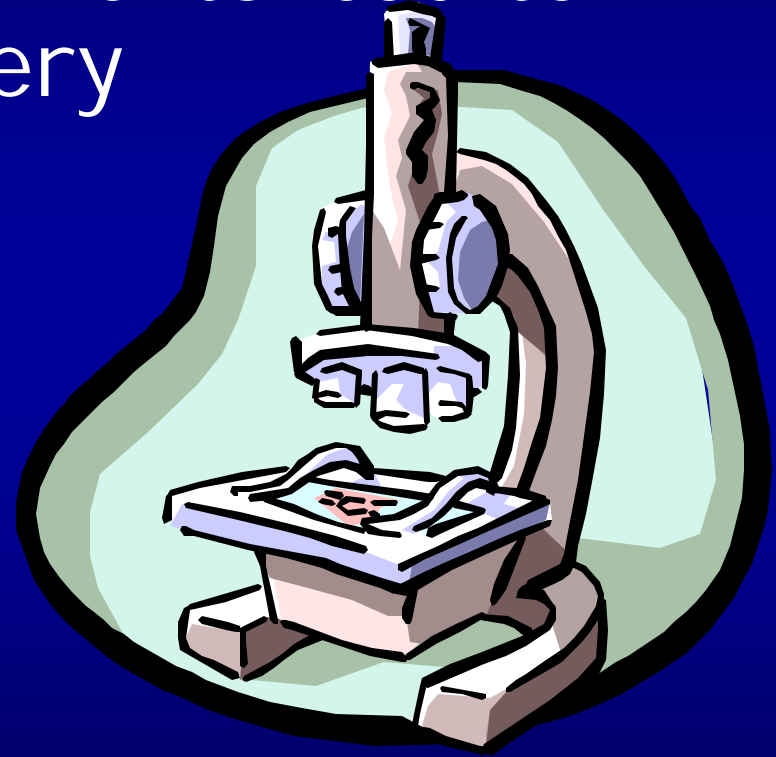
- Four main steps (Langley et al. 87):
 - Gathering data
 - Finding good descriptions of data
 - Formulating explanatory hypotheses
 - Testing the hypotheses

- My Claim:

We can do this with text as the data!

Scientific Breakthroughs

- New scientific instruments lead to revolutions in discovery
 - CAT scans, fMRI
 - Scanning tunneling electron microscope
 - Hubble telescope



- Idea:

Make A New Scientific Instrument!

How Has Knowledge been Discovered in Non-Textual Data?

Discovery from databases involves finding patterns across the data in the records

- Classification

 - » Fraud vs. non-fraud

- Conditional dependencies

 - » People who buy X are likely to also buy Y with probability P

How Has Knowledge been Discovered in Non-Textual Data?

- Old AI work (early 80's):
 - AM/Eurisko (Lenat)
 - BACON, STAHL, etc. (Langley et al.)
 - Expert Systems
- A Commonality:
 - Start with propositions
 - Try to make inferences from these
- Problem:
 - Where do the propositions come from?

Intensional vs. Extensional

- Database structure:
 - Intensional: The schema
 - Extensional: The records that instantiate the schema
- Current data mining efforts make inferences from the records
- Old AI work made inferences from what would have been the schemata
 - **employees have salaries and addresses**
 - **products have prices and part numbers**

Goal:
Extract Propositions from Text
and Make Inferences

Why Extract Propositions from Text?

- Text is how knowledge at the propositional level is communicated
- Text is continually being created and updated by the outside world
 - So knowledge base won't get stale

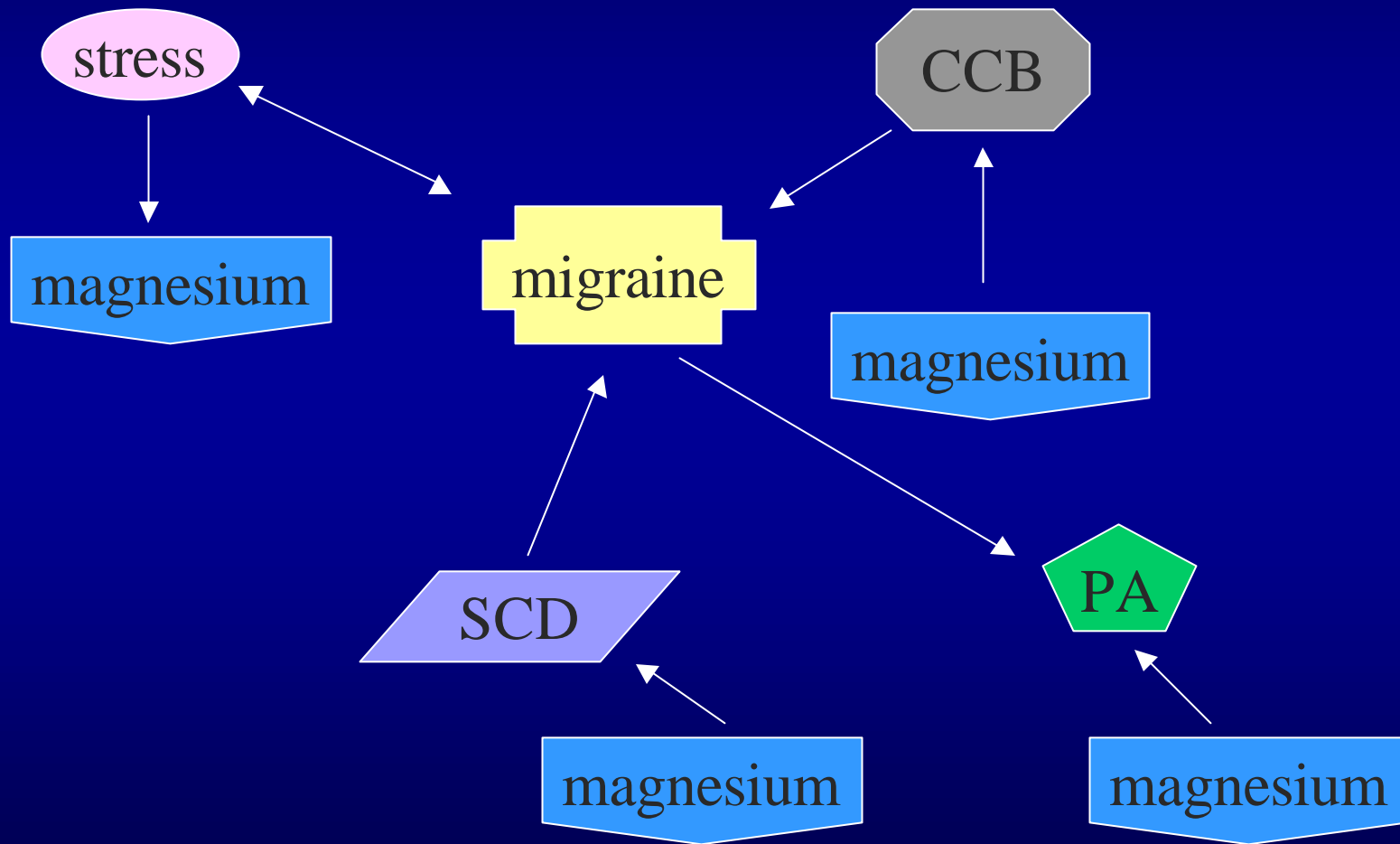
Example: Etiology

- Given
 - medical titles and abstracts
 - a problem (incurable rare disease)
 - some medical expertise
- find causal links among titles
 - symptoms
 - drugs
 - results

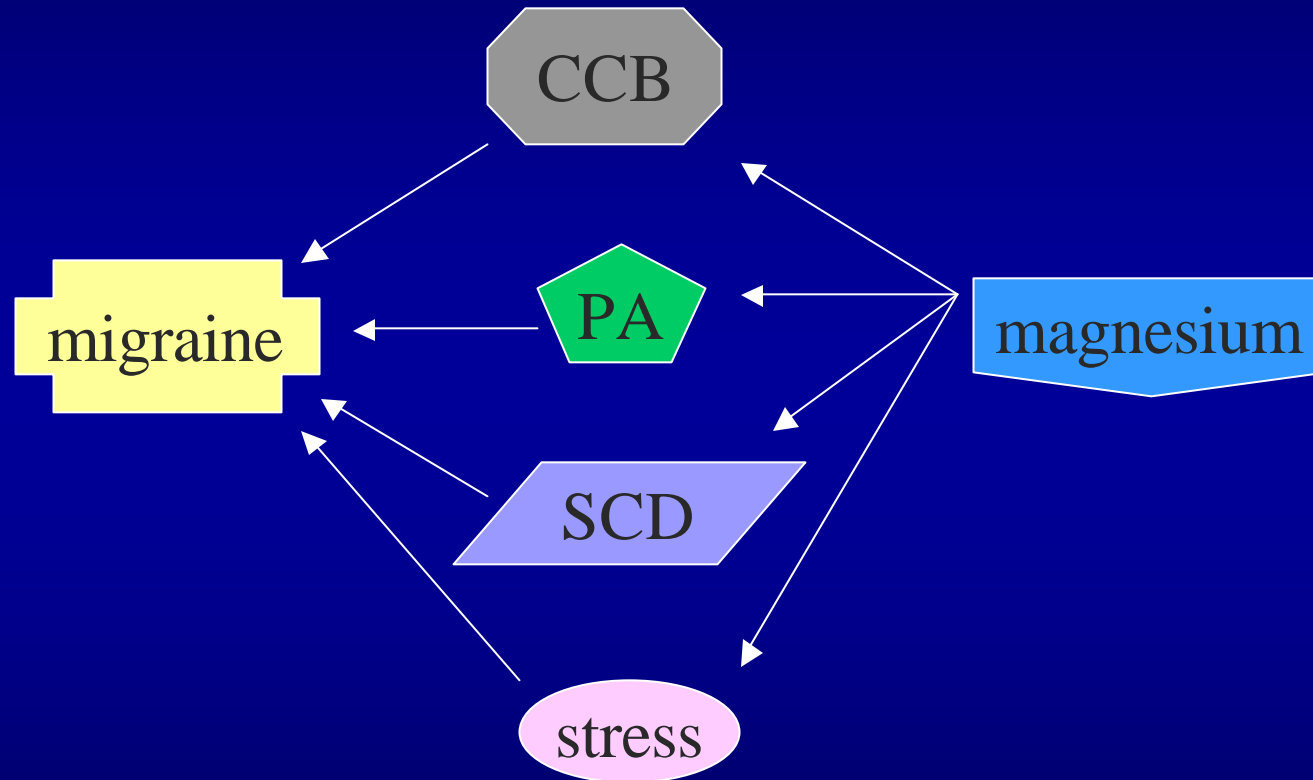
Swanson Example (1991)

- Problem: Migraine headaches (M)
 - stress associated with M
 - stress leads to loss of magnesium
 - calcium channel blockers prevent some M
 - magnesium is a natural calcium channel blocker
 - spreading cortical depression (SCD) implicated in M
 - high levels of magnesium inhibit SCD
 - M patients have high platelet aggregability
 - magnesium can suppress platelet aggregability
- All extracted from medical journal titles

Gathering Evidence



Gathering Evidence



Swanson's TDM

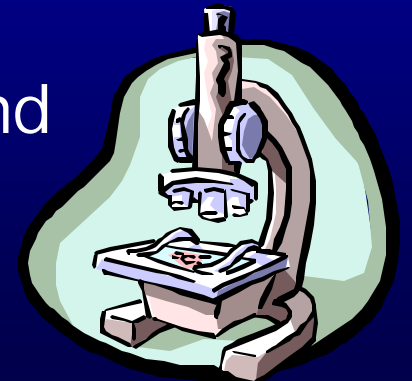
- Two of his hypotheses have received some experimental verification.
- His technique
 - Only partially automated
 - Required medical expertise
- Few people are working on this.

One Approach: The LINDI Project

Linking Information for New Discoveries

Three main components:

- Search UI for building and reusing hypothesis seeking strategies.
- Statistical language analysis techniques for extracting propositions from text.
- Probabilistic ontological representation and reasoning techniques



LINDI

- First use category labels to retrieve candidate documents,
- Then use language analysis to detect causal relationships between concepts,
- Represent relationships probabilistically, within a known ontology,
- The (expert) user
 - Builds up representations
 - Formulates hypotheses
 - Tests hypotheses **outside** of the text system.

Objections

- Objection:
 - This is GOF NLP, which doesn't work
- Response:
 - GOF NLP required hand-entering of knowledge
 - Now we have statistical techniques and very large corpora

Objections

- Objection:
 - Reasoning with propositions is brittle
- Response:
 - Yes, but now we have mature probabilistic reasoning tools, which support
 - » Representation of uncertainty and degrees of belief
 - » Simultaneously conflicting information
 - » Different levels of granularity of information

Objections

- Objection:
 - Automated reasoning doesn't work
- Response
 - We are not trying to automate all reasoning, rather we are building new powerful **tools** for
 - » Gathering data
 - » Formulating hypotheses

Objections

- Objection:
 - Isn't this just information extraction?
- Response:
 - IE is a useful tool that can be used in this endeavor, however
 - » It is currently used to instantiate pre-specified templates
 - » I am advocating coming up with entirely new, unforeseen "templates"

Traditional Semantic Grammars

- Reshape syntactic grammars to serve the needs of semantic processing.
- Example (Burton & Brown 79)
 - Interpreting "What is the current thru the CC when the VC is 1.0?"

```
<request> := <simple/request> when <setting/change>  
<simple/request> := what is <measurement>  
<measurement> := <meas/quant> <prep> <part>  
<setting/change> := <control> is <control/value>  
<control> := VC
```

- Resulting semantic form is:

```
(RESETCONTROL (STQ VC 1.0) (MEASURE CURRENT CC))
```

Statistical Semantic Grammars

- Empirical NLP has made great strides
 - But mainly applied to syntactic structure
- Semantic grammars are powerful, but
 - Brittle
 - Time-consuming to construct
- Idea:
 - Use what we now know about statistical NLP to build up a probabilistic grammar

Example: Statistical Semantic Grammar

- To detect causal relationships between medical concepts
 - Title:

Magnesium deficiency implicated in increased stress levels.
 - Interpretation:

<nutrient><reduction> related-to
<increase><symptom>
 - Inference:

» Increase(stress, decrease(mg))

Example: Using Semantics + Ontologies

- acute migraine treatment
- intra-nasal migraine treatment

Example: Using Semantics + Ontologies

- [acute migraine] treatment
- intra-nasal [migraine treatment]

We also want to know the *meaning* of the attachments,
not just which way the attachments go.

Example: Using Semantics + Ontologies

- acute migraine treatment
- <severity> <disease> <treatment>
- intra-nasal migraine treatment
- <Drug Admin Routes> <disease> <treatment>

Example:

Using Semantics + Ontologies

- acute migraine treatment
- <severity> <disease> <treatment>
- <severity> <Cerebrovascular Disorders> <treatment>
- intra-nasal migraine treatment
- <Drug Admin Routes> <disease> <treatment>
- <Administration, Intranasal> <disease> <treatment>

Problem: which level(s) of the ontology should be used?
We are taking an information-theoretic approach.

The User Interface

- A general search interface should support
 - History
 - Context
 - Comparison
 - Operator Reuse
 - Intersection, Union, Slicing
 - Visualization (where appropriate)
- We are developing such an interface as part of a general search UI project.

Summary

- Let's get serious about discovering new knowledge from text
- This will build on existing technologies
- This also requires new technologies

Summary

- Let's get serious about discovering new knowledge from text
 - We can build a **new kind of scientific instrument** to facilitate a whole new set of scientific discoveries
 - Technique: linking propositions across texts (Jensen, Harabagiu)

Summary

- This will build on existing technologies
 - Information extraction (Riloff et al., Hobbs et al.)
 - Bootstrapping training examples (Riloff et al.)
 - Probabilistic reasoning

Summary

- This also requires new technologies
 - Statistical semantic grammars
 - Dynamic ontology adjustment
 - Flexible search UIs