National Centre for Text Mining
NaCTeM

e-science and data mining workshop

John Keane
Co-Director, NaCTeM

john.keane@manchester.ac.uk
School of Informatics, University of Manchester

What is text mining?

Text mining attempts to discover new, previously unknown information by applying techniques from data mining, information retrieval, and natural language processing:

- To identify and gather relevant textual sources
- To analyse these to extract facts involving key entities and their properties
- To combine the extracted facts to form new facts or to gain valuable insights.

Why text mining?

- Users overwhelmed by amount of text; 80% of information in textual form
  - Results of scientific experiments also reproduced in textual form
  - Critical information missed
    - Only 12% of TOXLINE users find what they want
  - Lost in irrelevant documents

Why text mining? (cont.)

- Biological information
  - Disseminated over huge amount of documents
  - Dynamic (e.g. discovery of new genes)
  - Databases and controlled vocabularies encode only fraction of information and interactions
    - Significant error rate in manually curated DBs
  - No agreement on naming (inconsistent terminology)
  - $800M for 12 years for discovery of new drug
    - Reduced to 10 years if all relevant information known

Problem and Purpose

Main problems?

- Growing number and size of documents
- Language variability and ambiguity
- Difficult to assimilate/locate new knowledge without automated help

Main purposes of TM?

- Make information buried in text accessible
- Integrate information across articles
- Help interpret experimental data
- Update databases (semi)automatically

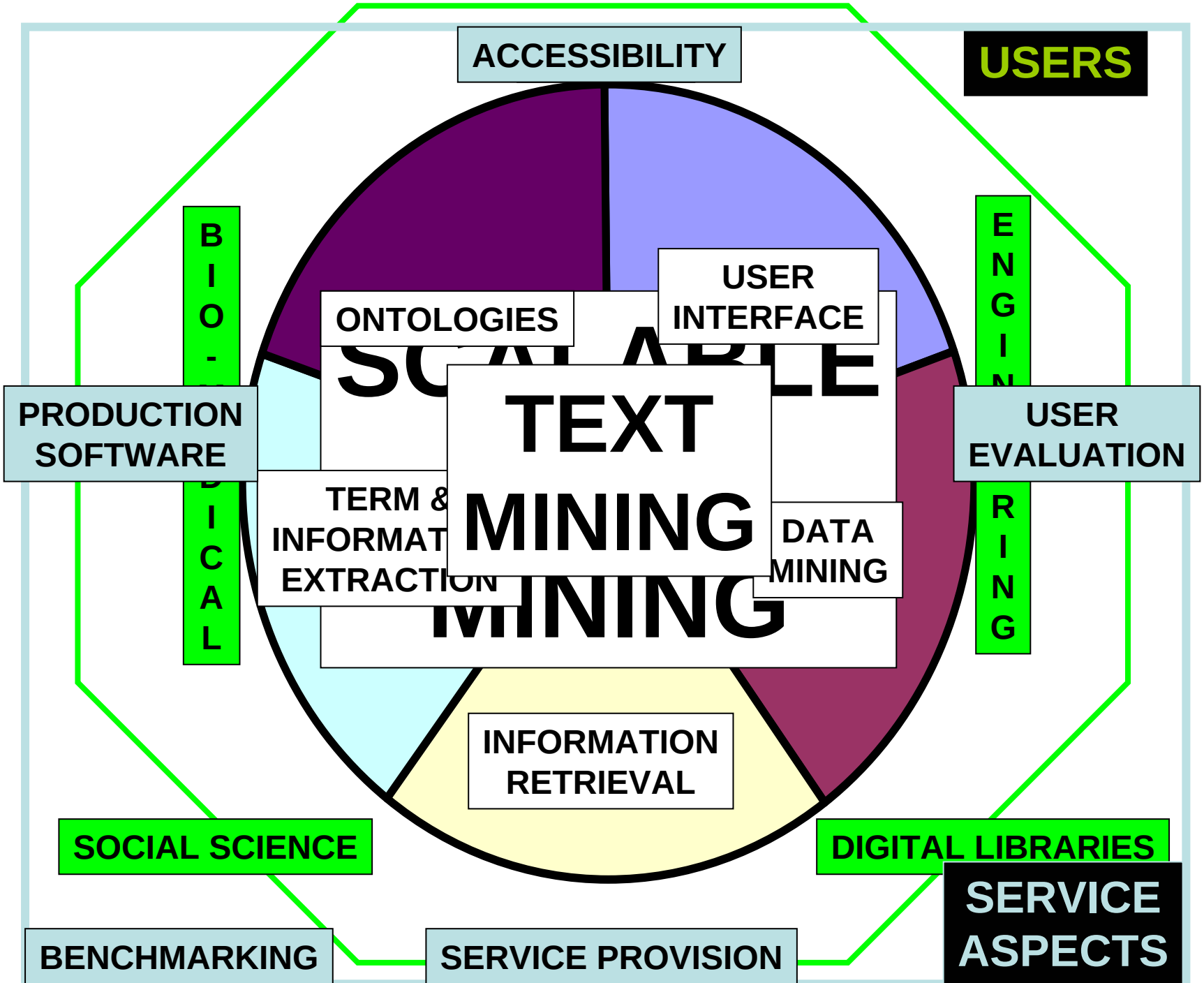
A simple motivating example

- Blagosklonny & Pardee argued information necessary to understand some processes related to p53 gene implicitly available for **10 years** in the literature before finally clarified

Blagosklonny, M., Pardee, A., 2001: *Conceptual Biology: Unearthing the Gems*, Nature 416: 373.

A vision for the future

- DBs with accurate, valid, exhaustive, rapidly updated data
- Drug discovery costs slashed; animal experimentation reduced through early identification of unpromising paths
- New insights gained through integration and exploitation of experimental results, DBs, and scientific knowledge
- Product development archives and patents yield new directions for R&D
- Searching yields *facts* rather than documents to read


ACCESSIBILITY
USERS
ONTOLOGIES
USER INTERFACE
TEXT MINING
TERM & INFORMATION EXTRACTION
DATA MINING
INFORMATION RETRIEVAL
PRODUCTION SOFTWARE
USER EVALUATION
SOCIAL SCIENCE
DIGITAL LIBRARIES
SERVICE ASPECTS
BENCHMARKING
SERVICE PROVISION

Text mining tasks and resources

- Information retrieval
  - Gather, select, filter, documents that may prove useful
  - Find what is known
- Information extraction
  - Partial, shallow language analysis
  - Find relevant entities, facts about entities
  - Find only what looking for
- Mining
  - Combine, link facts
  - Discover new knowledge, find new facts
- Resources:
  - ontologies, lexicons, terminologies, grammars, annotated corpora (machine learning, evaluation)

National Centre for Text Mining

- First such centre in the world
- Funding: JISC, BBSRC, EPSRC
  - £1M over 3 years
- Consortium investment
  - >800K including new chair in TM
- Location:
  - Manchester Interdisciplinary Biocentre (MIB)
- Focus: Bio, then extend to other domains
- To move towards self-sustainability
  - Extend services to industry

Consortium

- Manchester, Liverpool, Salford
- Service activity run by MIMAS (National Centre for Dataset Services), within MC (Manchester Computing)
- Self-funded partners
  - San Diego Supercomputing Center
  - UCalifornia, Berkeley
  - UGeneva
  - UTokyo
- Strong industrial & academic support
  - Astrazeneca, Xerox, EBI, Wellcome Trust, Sanger Institute, IBM, Unilever, ELRA, NowGEN, bionow

Competence within the consortium

- Manchester: TM (IE); standards; bioinformatics; parallel & distributed DM, HPC; HCI; ontologies & standards for semantic web; e-Science and GRID; curation of biodatabases
- Salford: Bio-TM; terminology; visualisation
- Liverpool: digital archives & IR; bioinformatics
- Manchester Computing: service provision; national data services (MIMAS); Supercomputing (HPC CSAR)

Self-funded Partners

- San Diego Supercomputing Centre: SKIDL toolkit for DM of high dimensional datasets; distributed and parallel computing
- SIMS, UCalifornia, Berkeley: probabilistic semantic grammar TM
- UTokyo: Bio TM; IE, GENIA, ontologies
- ISSCO, UGeneva: Standards-based evaluation methodologies for TM tools

Services

- Establish a high quality service provision for the UK academic community
  - Identify the 'best of breed' TM tools;
- Types of services
  - Facilitate access to TM tools, resources & support
  - Offer on-line use of resources and tools (also to guide and instruct)
  - Offer one-stop shop for complete, end-to-end processing

Critical aspects of service provision

- Scalability: most critical aspect of TM tool usage for a national service

  Consortium strong in distributed, high performance, GRID activity
- Focus on user-need related development whilst using experts to feed-in research results
  - need to separate research from TM service provision

Core development

- Creation of an infrastructure for TM
  - Standards based infrastructure for components and datasets to allow efficient distributed computing
  - Portal access
- Support for information retrieval
  - Use of a digital library system (Cheshire) to harvest and index data with improved index term weighting
  - Combine SKIDL (data mining toolkit, SDSC) with latent semantic analysis and probabilistic retrieval to extract text fragments for IE

Core development

- Support for inter-component communication, term management and IE
  - Produce a common annotation scheme to support UK TM, further developed from EU project http://www.crim.co.umist.ac.uk/parmenides/
  - Collaboration ISO TC37/SC4
- Term management
  - ATRACT workbench (Lion BioScience / EBI)
- Information Extraction
  - Manchester CAFETIERE

Core development

- Support for data grid technologies
  - Provide a means to connect to heterogeneous resources
- User interfaces, scientific data integration and mediation
  - Support user to set up data mining scenarios (via wizards)
- Support for advanced visualisation

Exemplary Outcomes

- Specific collaboration with EBI/SIB Swiss-Prot
  - Semi-auto produce highly precise terminology
  - Demonstrate precision of facts and associations mined from text that can be linked to terminology.
- Curation facilitation via text mining
  - Enhances access to, and retrieval of, literature
  - Keep DB links current/consistent with literature
  - Extracts pertinent facts from literature

    EBI tests suggest ~10% improvement over manual
  - Evident link with work of Digital Curation Centre
- Improve/construct/validate ontologies
- Evaluation for users/performance evaluation
- Component standardisation/ API/infrastructure/Common Annotation Scheme

Conclusion

- Creation of NaCTeM responds to a widely-felt need for text mining to support research
- Critical to engage with
  - Users in academia and (eventually) industry
  - Providers of bioinformatics infrastructure and resources such as biodatabases
- TM can be used also in teaching & learning
  - Problem-based learning
  - CAL support (extraction of facts and examples)
  - Exploration of link between student experimental results and results reported in literature

Background reading

Introductory

http://www.sims.berkeley.edu/~hearst/text-mining.html

BioLink SIG (workshop at ISMB 04 in Glasgow)

http://www.pdg.cnb.uam.es/BioLink/

BIONLP.org

http://www.ccs.neu.edu/home/futrelle/bionlp/

Tutorials

http://www.ccs.neu.edu/home/futrelle/bionlp/psb2001tutorials.html

http://www.3rdmill.com/pdf/MassBioTech%20BrfgPDF.pdf

Appendix: Types of services

- Access to TM tools developed from leading research
- Access to a selection of commercial text mining tools
- Access to ontology libraries
- Access to large and varied data sources
- Access to a library of data filtering tools
- On-line tutorials, briefings and white papers
- On-line advice to match specific requirements to TM solutions
- On-line TM & packaging of results involving GRID-based flexible composition of tools, resources and data by users
- TM tool trials and evaluations
- Collaborative development / enhancement of TM tools, annotated corpora and ontologies