MANCHESTER 1824

The University of Manchester

Role of Text Mining in Research and Innovation

John McNaught
Deputy Director
National Centre for Text Mining
www.nactem.ac.uk
John.McNaught@manchester.ac.uk

NaCTeM
The National Centre for Text Mining

How do we (humans) discover?

- Find, read, learn, analyse a lot
- Ask “What if…?”
- Construct hypotheses, test them
  - Explore many avenues, associations
- Work collaboratively
- Share results and data with others
  - Reproducibility → validation
- Integrate heterogeneous data/information/knowledge
- (vs. Serendipity: by lucky accident)

Barriers to discovery

- Find: *document* oriented, too many hits
- Read: too much to read, even if we find relevant hits
- Learn: too fast growth to keep up, to know most things
- Analyse: duplication of efforts, many new results to *document*
- Construct hypotheses: hard, can't tell which are most promising, or if have missed any
- Share: primary vehicles are *documents* and curated databases (massive curation backlog)
- Integrate: *document* often the key, hard to link in to different worlds of data, information, knowledge

How does TM aid discovery?

- **Find**: more precise, relevant information, *within* and *across* documents
- **Read**: much faster than human
- **Learn**: extracts, packages, links, synthesises, summarises, reduces burden
- **Analyse**: recognises duplication; clusters, classifies, drives semantic author aids
- **Construct hypotheses**: rapidly finds and *ranks* unknown associations for testing
- **Share**: reduces curation effort, complements and validates data bases
- **Integrate**: links documents deeply into worlds of data, information and knowledge

Text Mining in a nutshell

Unstructured Text
(implicit knowledge)

Information Retrieval

Information Extraction

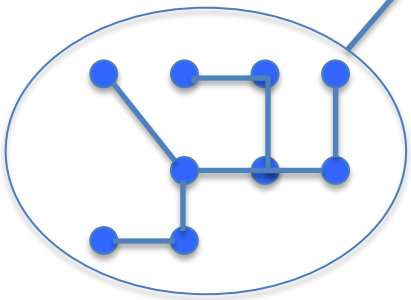
Knowledge Discovery

Semantic Metadata Annotations

Structured Content
(explicit knowledge)

Applications
Semantic search
Data mining

Other data



NaCTeM
The National Centre for Text Mining

What's the difference?

- Human slow, TM system fast
- Human tires, TM system tireless
- Human reads, analyses, …
  - So does the TM system
- There is ***no essential difference*** between human activity and TM system activity
  - But TM activity largely hindered by copyright and licensing

Copyright and licensing

- “The right to read is the right to mine”
- Must take at least one copy to text mine
- Licenses for TM unfit for digital world
  - Pay again for 2nd licence?
  - “Approved project” basis
  - Renew request for every new project (!)
  - Results only for authorised users of institution
    - Prevents national/international TM services & sharing
  - Negotiation with many rights holders: cannot scale

Copyright and licensing

- Myth of TM crashing servers scenario
  - Neylon (PLoS): “If you can’t handle text mining requests, you have no business running a Web server”
- Unrealistic licence constraints, e.g.:
  - Download rate: 1 page every 20 seconds (real offered licence)
  - At this rate, assuming instant download (!), 1m texts each of 10 pages would take over 6 years to download…

The need for full text

- **<8% of scientific claims found in full article appear in its abstract** (Blake 2010, *JBI* 43(2))
- Abstracts deficient on argumentation, discussion, methods, background, …
- Full texts needed to realise full benefits of TM
- Same true for most kinds of documents
  - Especially if wish to link document content deeply to wider knowledge space and other data sources

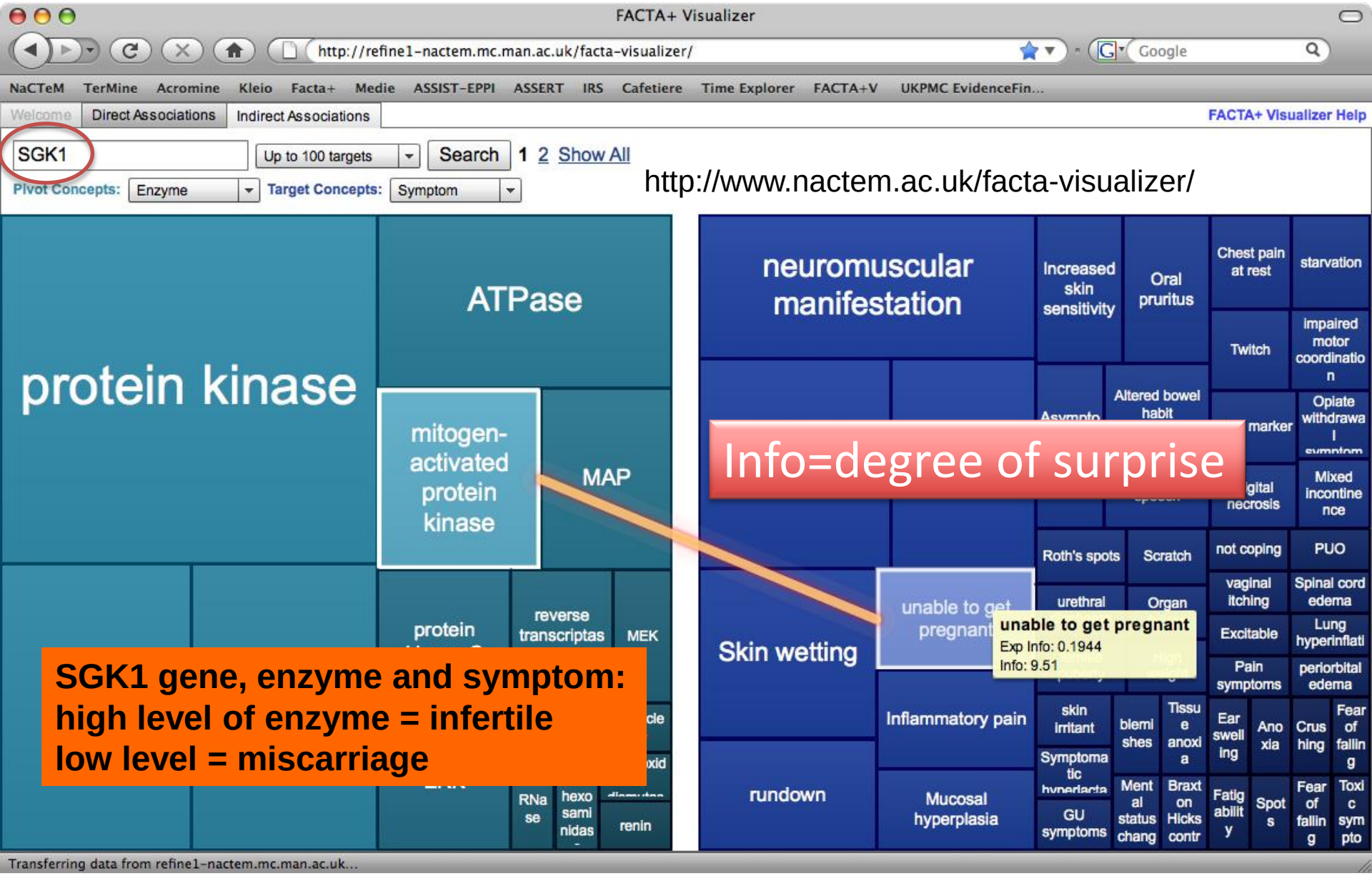
Nuances of language

- Most messages are nuanced
- Argumentation, rhetorical intent
- Speculation/hedging
  - Probable, possible,…
  - Suggest, indicate, …
  - May, might, would, …
- Manner: slightly, rapidly, greatly, …
- Polarity (negative, positive): no, never, …
- Such knowledge required for: discourse analysis, opinion mining, …
- If not taken into account, then results can be invalid and misleading

Finding unknown associations

- Need massive amounts of full text to find unknown associations, generate hypotheses
- Must go across collections: silos irrelevant to researcher
- Must go across disciplines: cognate and distant – all can shed light
- Discoveries made in lab: information often available in literature many years before, but unsuspected as not explicitly written down

Reproducing a finding - reported (11/2011) in *Nature Medicine* - with FACTA+, using MEDLINE ***prior to date***



Finding evidence from full text

- Example: system for Europe PubMedCentral
- Beyond full text search and pattern matching
- Deeply analyse documents off-line
- Index relationships
- Key off search term, to dynamically generate from indexed relationships *questions that have known answers*
  - Not auto-completion
- BUT: Restricted to small Open Access subset…

Europe PubMed Central labs

EvidenceFinder: a new way to discover

81,792,696 Sentences about genes, proteins, diseases & metabolites

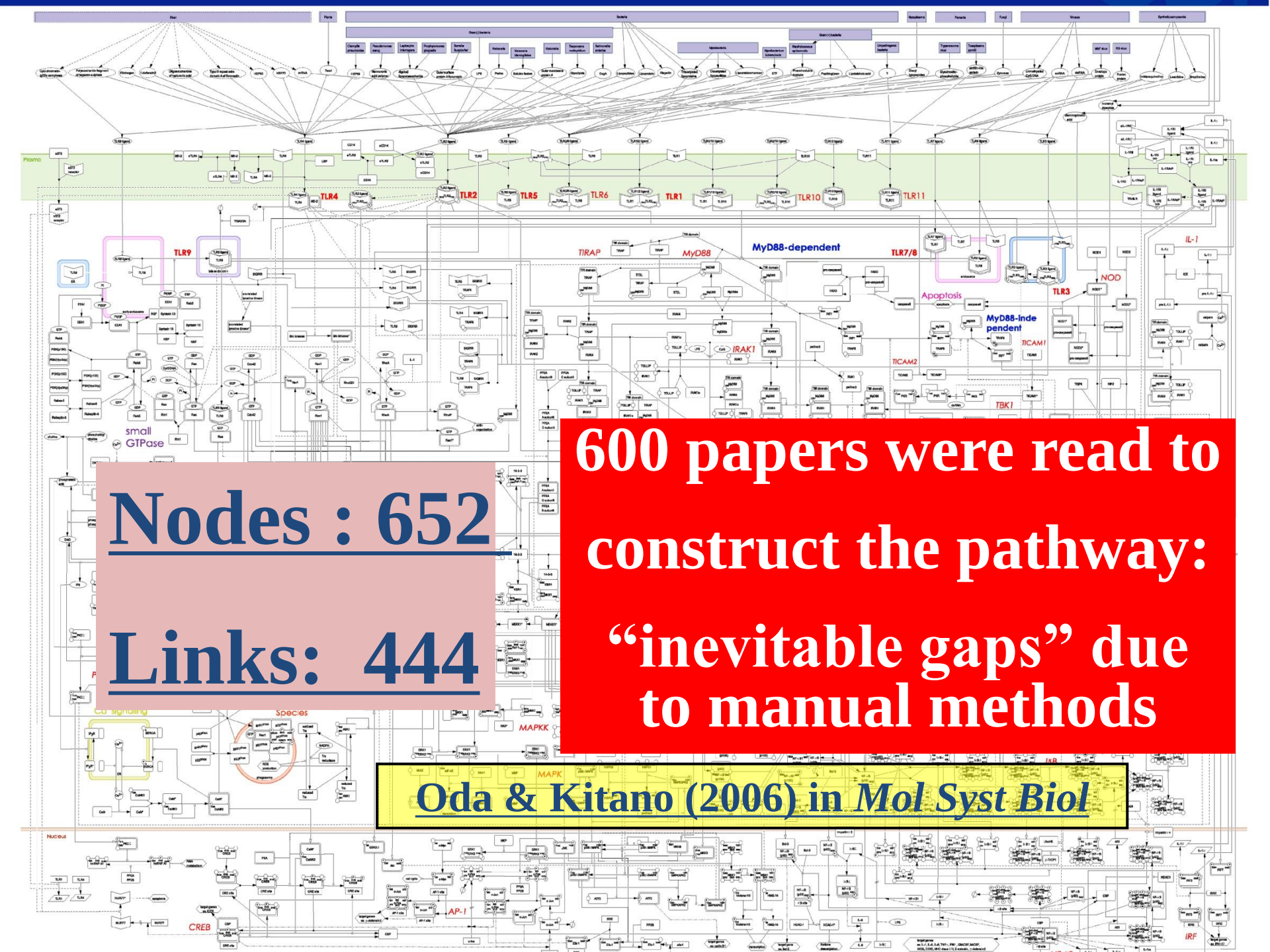
How can you tell if an article is relevant to you in your listed search results? Are t

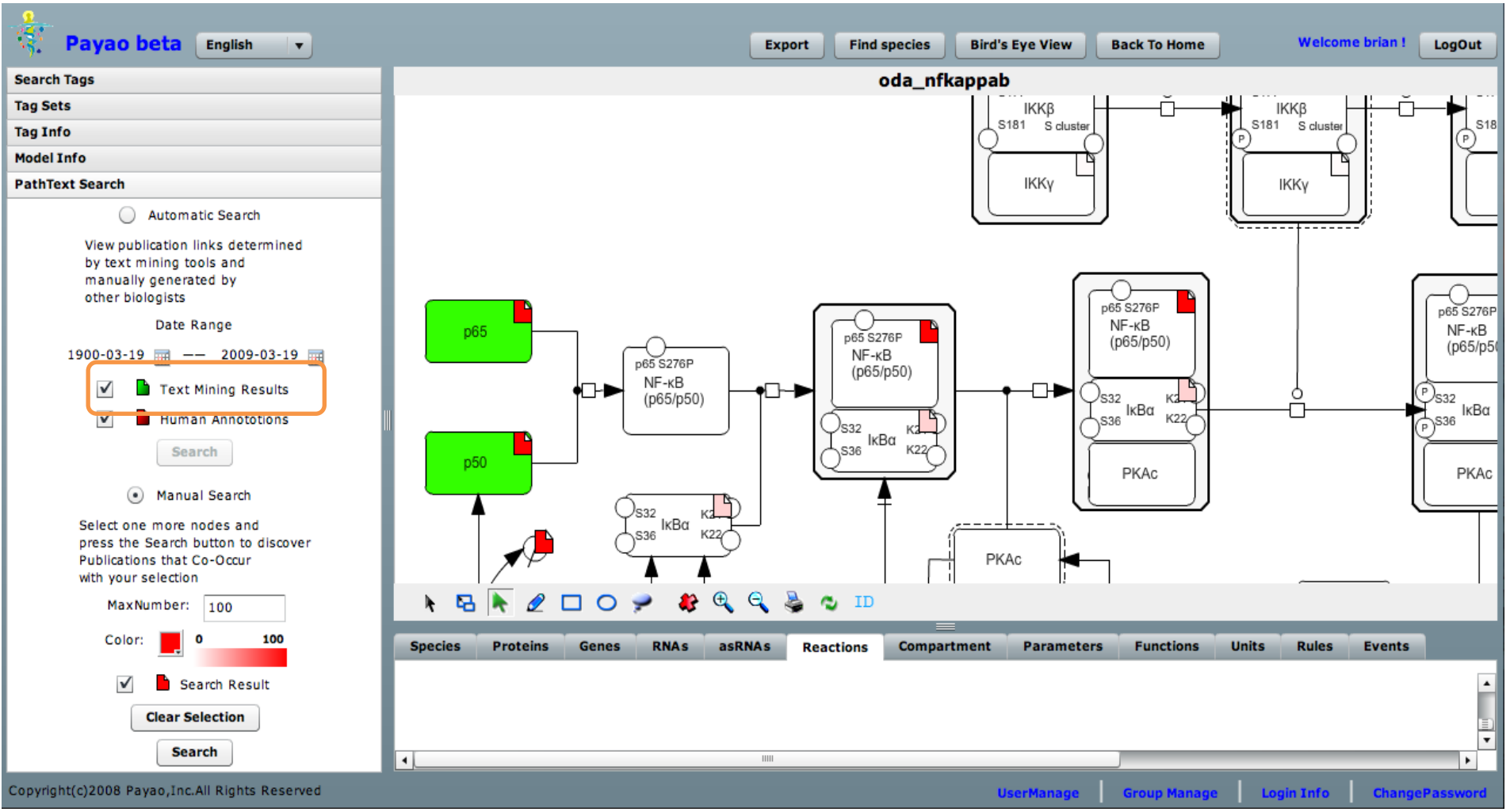
Europe PMC's EvidenceFinder enriches your literature exploration by suggesting questions alongside your search results, providing a way to find information buried in full text articles that is directly relevant to you.
This helps you identify articles and research that you might have overlooked through direct key word searching.

http://europepmc.org/

TM helps map biological systems

- Re-construction of signalling pathways and networks
- Important to understand how biological systems function
- Laborious manual construction
  - Hard to manually find indirect associations to help construct networks
- E.g. Manual map of how immune system recognises pathogens


Nodes : 652
Links: 444
600 papers were read to construct the pathway:
“inevitable gaps” due to manual methods
Oda & Kitano (2006) in Mol Syst Biol
TLR4
TLR2
TLR5
TLR6
TLR1
TLR10
TLR11
TLR9
TLR7/8
TLR3
MyD88-dependent
MyD88-independent
TIRAP
MyD88
IRAK1
TICAM1
TICAM2
TBK1
Apoptosis
NOD
IL-1
small GTPase
MAPKK
MAPK
IκB
AP-1
CREB
IRF
Plasma
Nucleus



Now: TM tools help construction by discovering **associations** (Kemper et al. (2010), *Bioinformatics* 26(12))

NaCTeM
The National Centre for Text Mining

Systematic reviewing

- For policy development, guidelines, best practice, evidence based public health,…
- SR can take 12-18 months
- *Basic* TM methods proven to reduce time spent in searching/screening by 75%
  - More certain not missing key material

Take home messages

- Text mining saves lives
  - Metaphorically, when you're drowning in data
  - Literally, by radically transforming knowledge discovery to enable unsuspected associations and evidence to be discovered that then drive research and innovation
- Text mining improves lives
  - More productive researchers
  - Faster, more accurate evidence-based policy development for societal benefit

NaCTeM
The National Centre for Text Mining

Why have we waited so long for a copyright exception for TM?

"Dr Bush urges that men of science should turn to the massive task of making more accessible our bewildering store of knowledge. Now, says Dr Bush, instruments are at hand which, if properly developed, will give man access to and command over the inherited knowledge of the ages." (**1945**)

http://www.theatlantic.com/doc/194507/bush

NaCTeM
The National Centre for Text Mining

Useful resources

- Recent JISC report on the value and benefits of text mining
  - http://www.jisc.ac.uk/publications/reports/2012/value-and-benefits-of-text-mining.aspx
- Reichman & Okediji (2012) When Copyright Law and Science Collide: Empowering Digitally Integrated Research Methods on a Global Scale. *Minnesota Law Review*.

http://www.minnesotalawreview.org/wp-content/uploads/2012/08/ReichmanOkediji_MLR1362.pdf