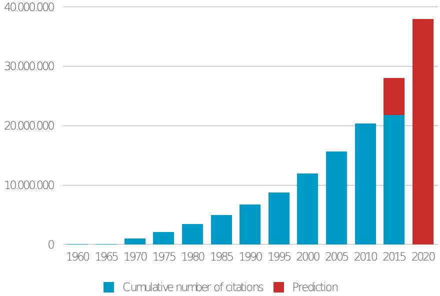
Algorithms for resource-constrained domain-specific knowledge management

Bachelor's thesis

Uli Köhler

September 24, 2015

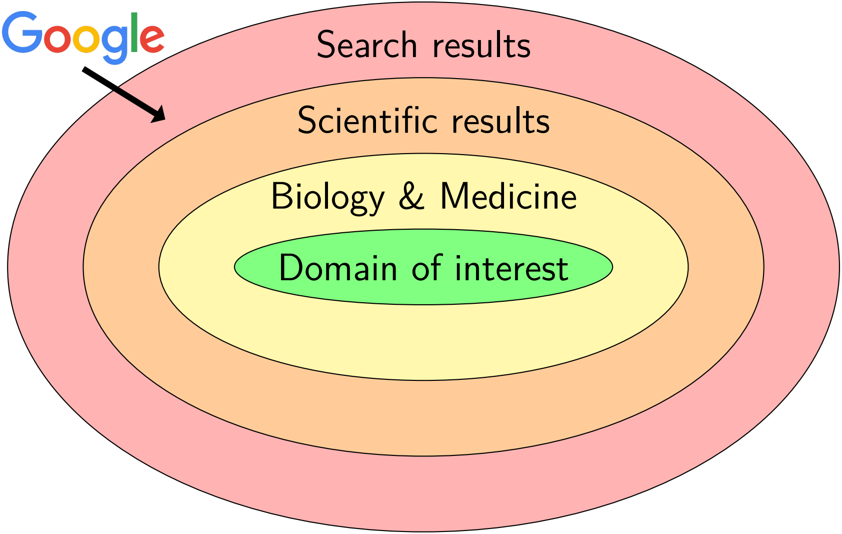
Text mining



- Growing number of publications
- Quick access to information required

→ Text mining

Search engines



Google

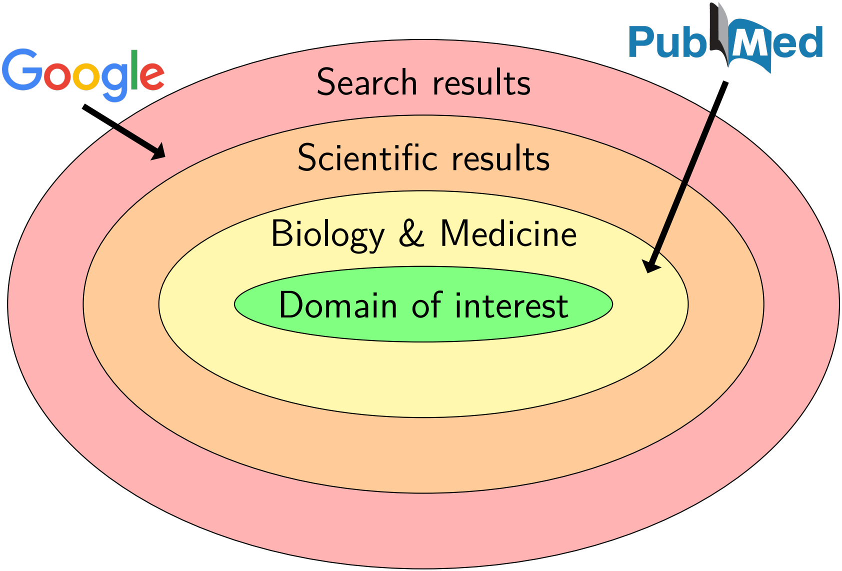
Search results

Scientific results

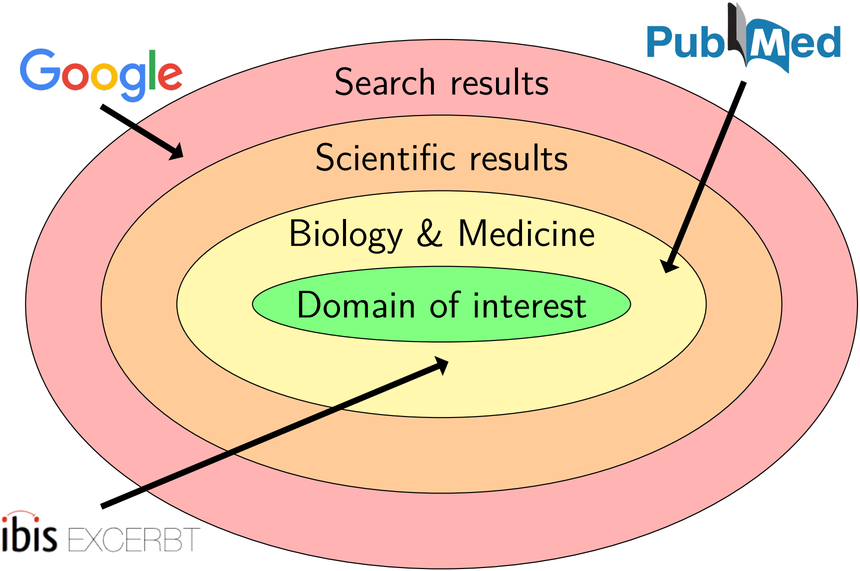
Biology & Medicine

Domain of interest

Search engines



Search engines



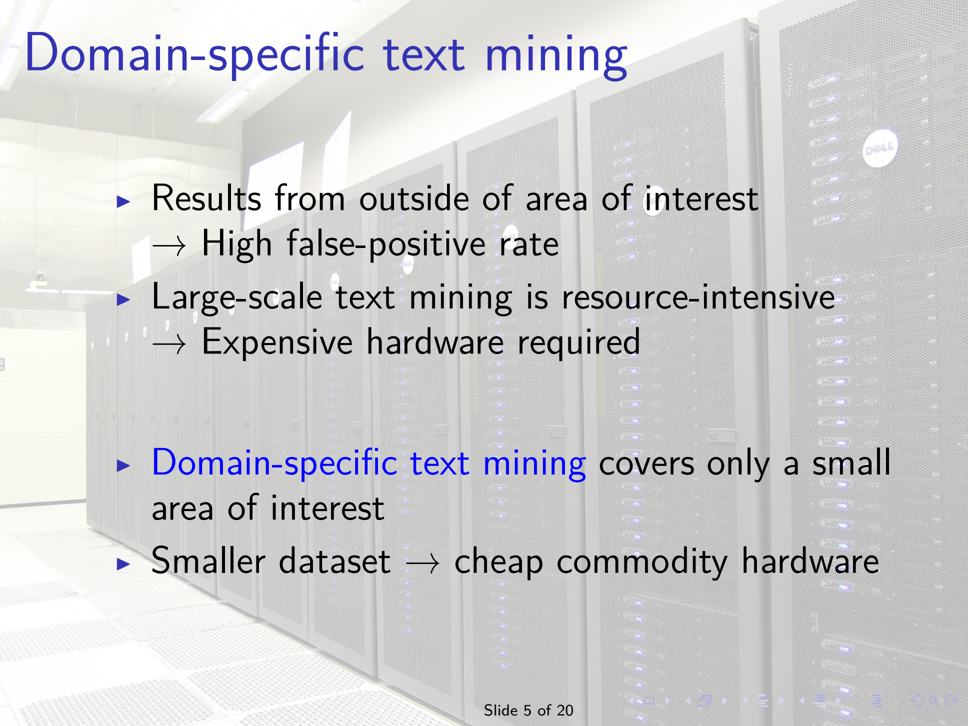
Search results

Scientific results

Biology & Medicine

Domain of interest

Domain-specific text mining



- Results from outside of area of interest
  → High false-positive rate
- Large-scale text mining is resource-intensive
  → Expensive hardware required

- Domain-specific text mining covers only a small area of interest
- Smaller dataset → cheap commodity hardware

TRANSLATRON

- **Transla**tional Bioinformatics **T**ool with **r**elatime **on**tology
- A simple tool for domain-specific text mining
- Easy to use — low hardware requirements
- Web-based user interface
- Real-time search in corpus and ontology
- Named entity recognition (**NER**)

Example: *The* $\boxed{\textit{PrP}^{\textit{Sc}}}$ $\boxed{\textit{prion}}$ *causes ovine* $\boxed{\textit{prion diseases}}$

Algorithms for domain-specific text mining

- Conventional algorithms built for large-scale datasets:
    - Hundreds of gigabytes of RAM available
    - Hundreds of terabytes of disk space available
    - Clustered architecture

Algorithms for domain-specific text mining

- Conventional algorithms built for large-scale datasets:
  - Hundreds of gigabytes of RAM available
  - Hundreds of terabytes of disk space available
  - Clustered architecture
- Novel algorithms required for domain-specific approaches

Algorithms in *TRANSLATRON*

- *YakDB* High-performance database
- *PRIMORDIAL* text indexing
- *PRAISER* distributed indexing
- *PERSIST* single-token indexing
- *PRESIDE* real-time prefix search
- *PRO-PANE* priority-based result ordering
- ***FiT-NESS* named entity recognition**
- *WESTSIDE* client interface

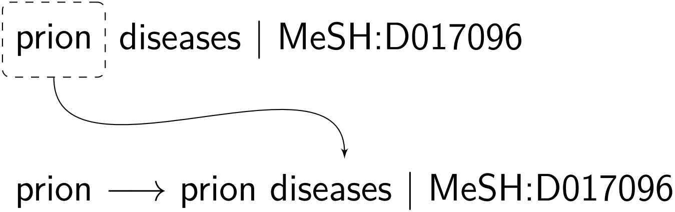
FiT-NESS

- **Fi**rst-**T**oken-based **N**amed **E**ntity **S**election **S**cheme
- Trivial: *Single-token entities* like *BRCA1*
- Hard: *Multi-token entities* like *prion diseases*

FiT-NESS

- **F**i**rst-T**oken-based **N**amed **E**ntity **S**election **S**cheme
- Trivial: *Single-token entities* like *BRCA1*
- Hard: *Multi-token entities* like *prion diseases*
- *FiT-NESS aproach*:
  - Ignore everything but the first token
  - When we find a hit, check if subsequent tokens match the entity

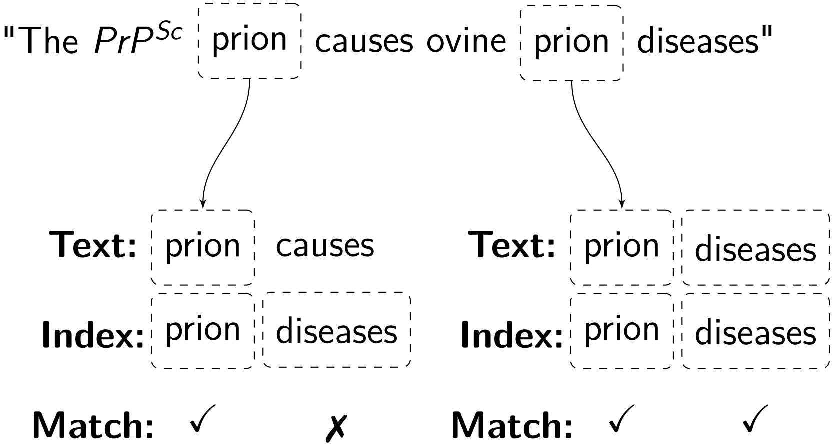
FiT-NESS II



prion diseases | MeSH:D017096

prion ⟶ prion diseases | MeSH:D017096

FiT-NESS III

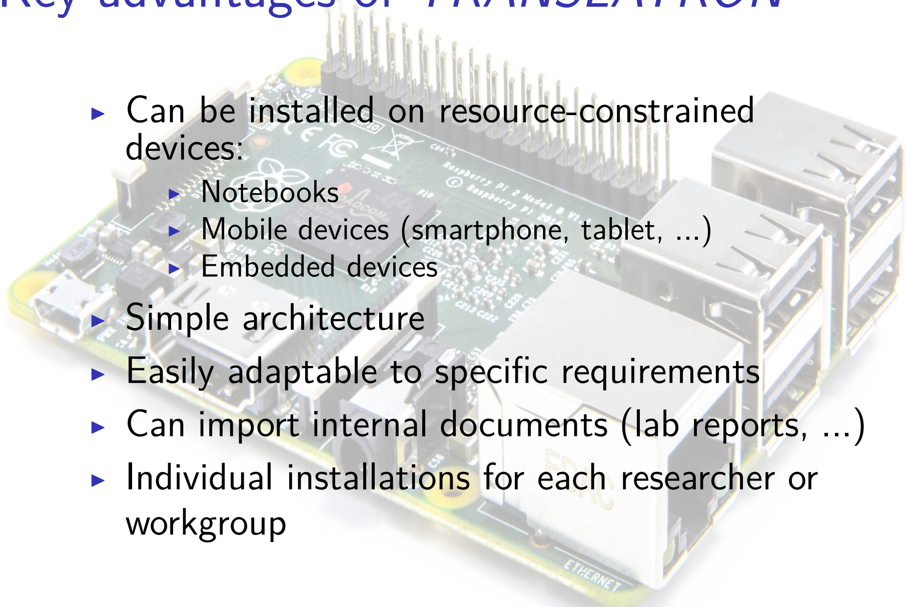


"The *PrP*^{Sc} prion causes ovine prion diseases"

Text:	prion	causes
Index:	prion	diseases
Match:	✓	✗

Text:	prion	diseases
Index:	prion	diseases
Match:	✓	✓

Key advantages of *TRANSLATRON*

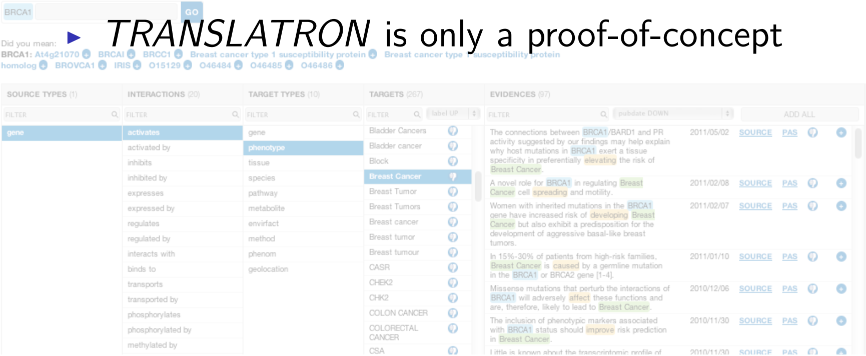


- Can be installed on resource-constrained devices:
  - Notebooks
  - Mobile devices (smartphone, tablet, ...)
  - Embedded devices
- Simple architecture
- Easily adaptable to specific requirements
- Can import internal documents (lab reports, ...)
- Individual installations for each researcher or workgroup

Live demonstration

Outlook & conclusion

- *TRANSLATRON* is only a proof-of-concept



BRCA1 GO

Did you mean:

BRCA1: At4g21070 BRCAI BRCC1 Breast cancer type 1 susceptibility protein Breast cancer type 1 susceptibility protein homolog BROVCA1 IRIS O15129 O46484 O46485 O46486

SOURCE TYPES (1)

FILTER

gene

INTERACTIONS (20)

FILTER

activates
activated by
inhibits
inhibited by
expresses
expressed by
regulates
regulated by
interacts with
binds to
transports
transported by
phosphorylates
phosphorylated by
methylated by

TARGET TYPES (10)

FILTER

gene
phenotype
tissue
species
pathway
metabolite
envirfact
method
phenom
geolocation

TARGETS (267)

FILTER label UP

Bladder Cancers
Bladder cancer
Block
Breast Cancer
Breast Tumor
Breast Tumors
Breast cancer
Breast tumor
Breast tumour
CASR
CHEK2
CHK2
COLON CANCER
COLORECTAL CANCER
CSA

EVIDENCES (97)

FILTER pubdate DOWN ADD ALL

The connections between BRCA1/BARD1 and PR activity suggested by our findings may help explain why host mutations in BRCA1 exert a tissue specificity in preferentially elevating the risk of Breast Cancer. 2011/05/02 SOURCE PAS

A novel role for BRCA1 in regulating Breast Cancer cell spreading and motility. 2011/02/08 SOURCE PAS

Women with inherited mutations in the BRCA1 gene have increased risk of developing Breast Cancer but also exhibit a predisposition for the development of aggressive basal-like breast tumors. 2011/02/07 SOURCE PAS

In 15%-30% of patients from high-risk families, Breast Cancer is caused by a germline mutation in the BRCA1 or BRCA2 gene [1-4]. 2011/01/10 SOURCE PAS

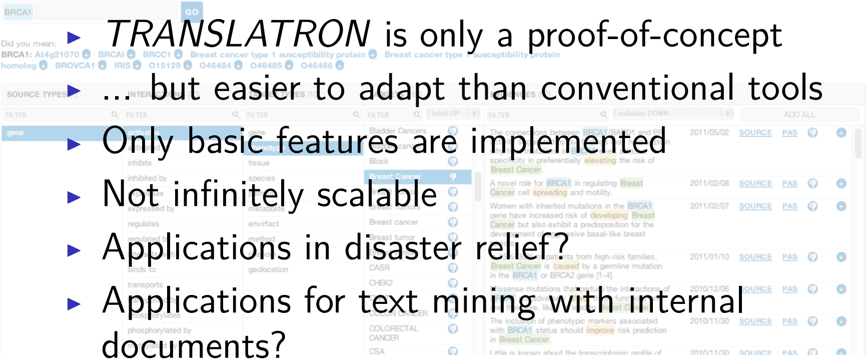
Missense mutations that perturb the interactions of BRCA1 will adversely affect these functions and are, therefore, likely to lead to Breast Cancer. 2010/12/06 SOURCE PAS

The inclusion of phenotypic markers associated with BRCA1 status should improve risk prediction in Breast Cancer. 2010/11/30 SOURCE PAS

Little is known about the transcriptomic profile of 2010/11/30 SOURCE PAS

Outlook & conclusion

- *TRANSLATRON* is only a proof-of-concept
- … but easier to adapt than conventional tools
- Only basic features are implemented
- Not infinitely scalable
- Applications in disaster relief?
- Applications for text mining with internal documents?



The connections between BRCA1/BARD1 and PR … specificity in preferentially elevating the risk of Breast Cancer.	2011/05/02	SOURCE	PAS
A novel role for BRCA1 in regulating Breast Cancer cell spreading and motility.	2011/02/08	SOURCE	PAS
Women with inherited mutations in the BRCA1 gene have increased risk of developing Breast Cancer but also exhibit a predisposition for the development of aggressive basal-like breast tumors.	2011/02/07	SOURCE	PAS
In 15%-30% of patients from high-risk families, Breast Cancer is caused by a germline mutation in the BRCA1 or BRCA2 gene [1-4].	2011/01/10	SOURCE	PAS
Missense mutations that perturb the interactions of BRCA1 … adversely affect … functions … are therefore likely to lead to Breast Cancer.	2010/12/06	SOURCE	PAS
The inclusion of phenotypic markers associated with BRCA1 status should improve risk prediction in Breast Cancer.	2010/11/30	SOURCE	PAS
Little is known about the transcriptomic profile of	2010/11/30	SOURCE	PAS

Acknowledgements

Mathias C. Walter

Prof. Dr. Hans-Werner Mewes

... and many others ...

Thank you for your attention!

References and sources available at

`https://github.com/ulikoehler/Bachelor`
`https://github.com/ulikoehler/Translatron`
`https://github.com/ulikoehler/YakDB`

Thesis & talk available at `http://techoverflow.net`
Contact: ukoehler@techoverflow.net

Questions?

Image sources

http://www.case.edu/med/nutrition/images/pubmed-logo.jpg

http://mips.helmholtz-muenchen.de/excerbt

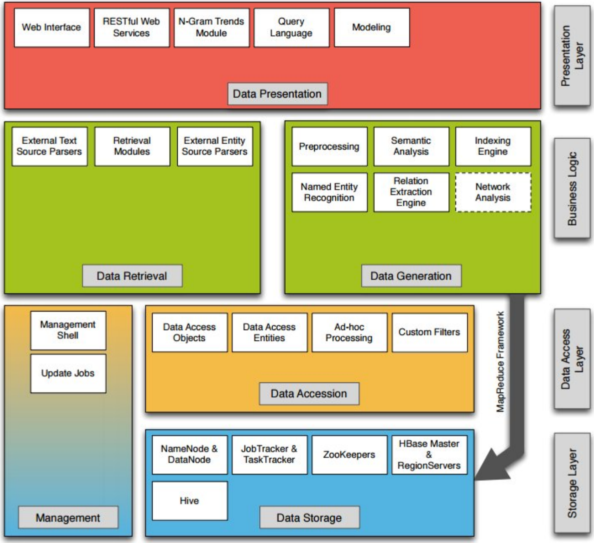
http://www.raspberrypi.org/blog/raspberry-pi-2-on-sale/

https://www.raspberrypi.org/blog/raspberry-pi-2-on-sale/

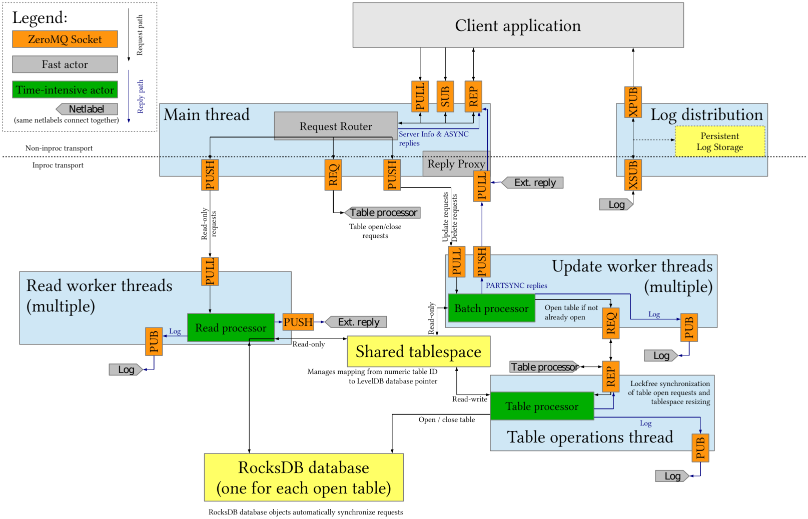
http://www.depts.ttu.edu/hpcc/

Wachinger: Next Generation Knowledge Extraction from Biomedical Literature with Semantic Big Data Approaches

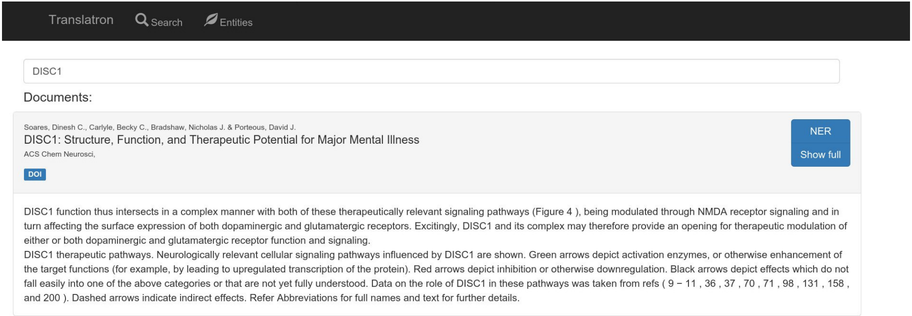
Excerbt architecture



YakDB architecture



Translatron demo



DISC1

Documents:

Soares, Dinesh C., Carlyle, Becky C., Bradshaw, Nicholas J. & Porteous, David J.

DISC1: Structure, Function, and Therapeutic Potential for Major Mental Illness

ACS Chem Neurosci,

DISC1 function thus intersects in a complex manner with both of these therapeutically relevant signaling pathways (Figure 4), being modulated through NMDA receptor signaling and in turn affecting the surface expression of both dopaminergic and glutamatergic receptors. Excitingly, DISC1 and its complex may therefore provide an opening for therapeutic modulation of either or both dopaminergic and glutamatergic receptor function and signaling.

DISC1 therapeutic pathways. Neurologically relevant cellular signaling pathways influenced by DISC1 are shown. Green arrows depict activation enzymes, or otherwise enhancement of the target functions (for example, by leading to upregulated transcription of the protein). Red arrows depict inhibition or otherwise downregulation. Black arrows depict effects which do not fall easily into one of the above categories or that are not yet fully understood. Data on the role of DISC1 in these pathways was taken from refs (9 − 11 , 36 , 37 , 70 , 71 , 98 , 131 , 158 , and 200). Dashed arrows indicate indirect effects. Refer Abbreviations for full names and text for further details.