Text Mining: Where do we come from? Where do we go?

Stan Matwin, CRC

Institute for Big Data Analytics

DALHOUSIE UNIVERSITY
Inspiring Minds
Faculty of Computer Science

AI'2014, Montréal

Plan

- A tiny bit of history, and…
- Where we are today
- A **burning** question….
- …and discussion around it
- Some current work related to the question…
- A peek into the future

Text mining (~ text analytics)

- Classification, clustering, <u>information extraction</u>, topic identification, sentiment analysis,…
- ie: unstructured text → database record (frame); e.g.

  business news → corporate_takeover

  `company_A, company_B, share_price, CEO, …`

Very brief history

- Early NLP (60s-80s) : syntax-driven, semantic approaches,
  - frame semantics based on case grammars [Fillmore 68]
- Statistical approaches (90s-today)
  - info extraction e.g. Conditional Random Fields [Lafferty, McCallum, Pereira 01]
  - topic identification, e.g. Latent Dirichlet Allocation [Blei, Ng Jordan 03]
- Deep learning models [Mnih,Hinton 09], [Morin, Bengio 05],[Bordes et al. 12], [Mikolov 13]

The **burning question**: relationship between Knowledge and Data

- Qualitative
- Quantitative

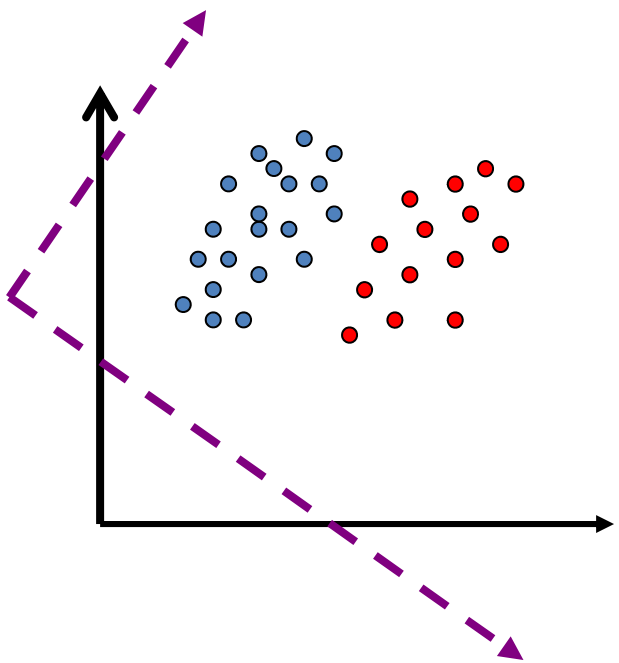
Knowledge replaces data? – yes (1993)

- [Clark, Matwin 93]
- Learning from data and from the model of the world
- The use of the model in learning is equivalemt to having additional examples

	model-example equivalence
ore-grinder model	+50%
waterflow model	+20%

- Other advantages
  - Comprehensibility
  - Noise resistance

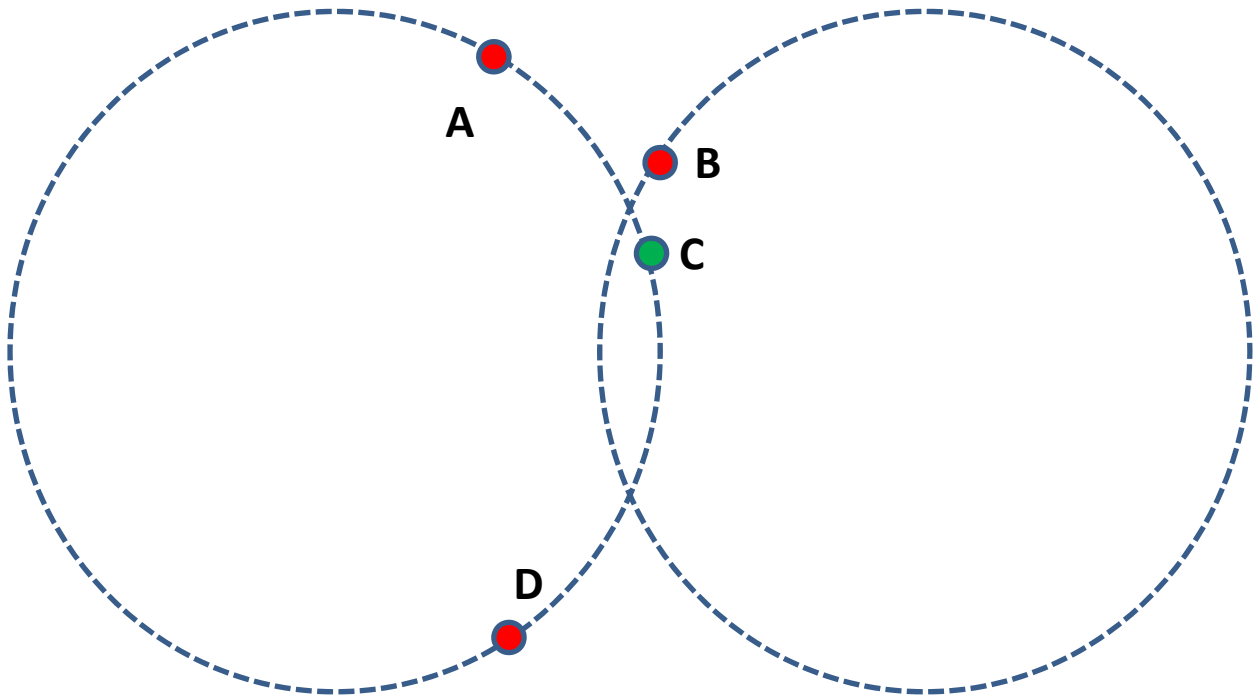
- “You can’t learn something unless you already almost know it”
- But, PCA



- Where to get knowledge?
- In what form?
  - Symbolic?
  - Probabilistic?
  - How to embed it in domain-specific kernels?

- So: thou should use knowledge in AI systems…
- Knowledge will bring structure…and understanding

Who's closer to C, A or B?



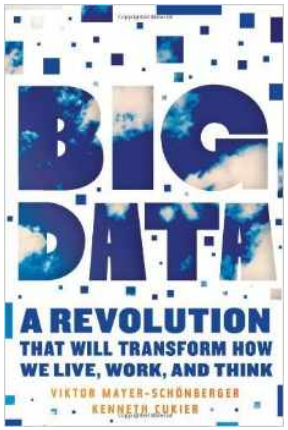
Statistician: B

Knowledge-based AI-er: C

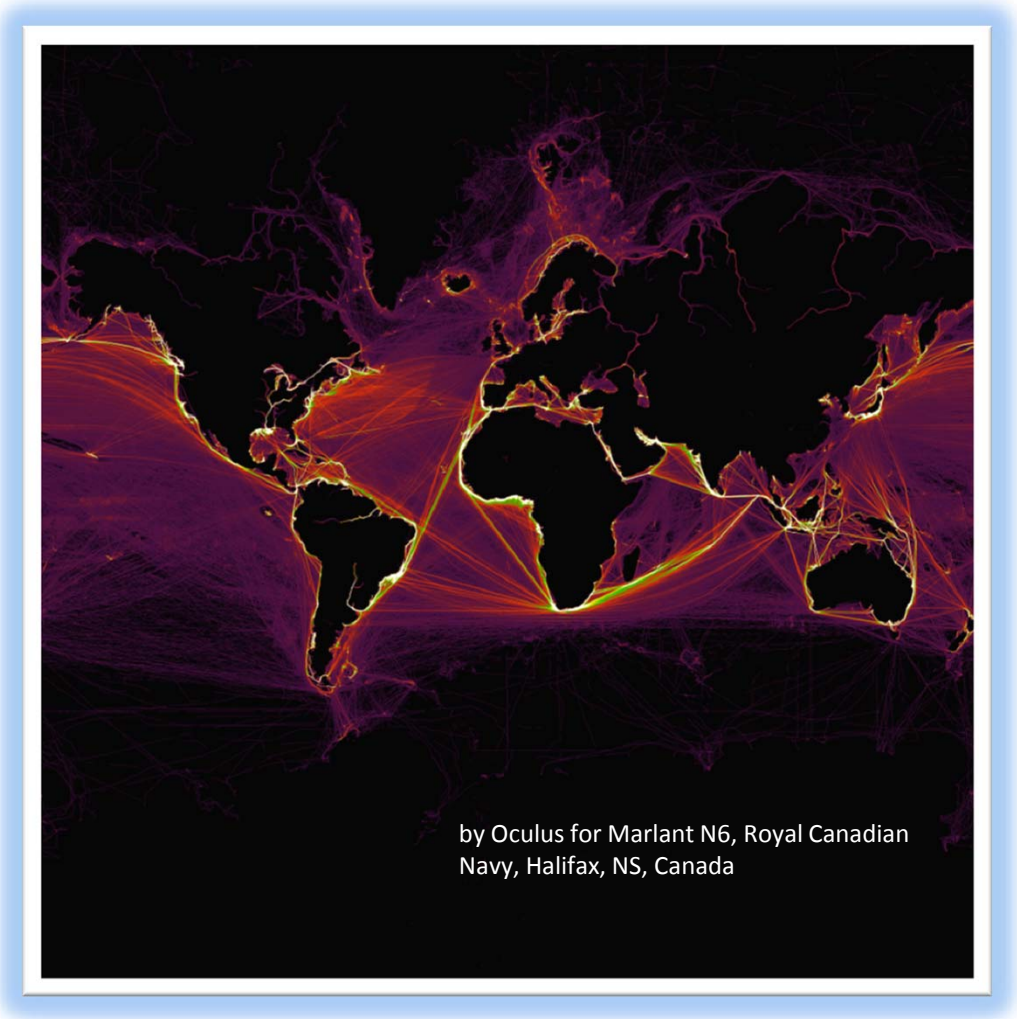
Data replaces knowledge (Big Data - 2011)

- Hegelian principle: transition
  `quantity → quality`

- Why would that be?
  - [Mayer-Schonberger, Cukier 2013]: “end of theory”, “correlation instead of causality”, “demise of the expert”



Large enough data will bring structure…



AIS signal data only, no knowledge of geography etc.

AI example – Machine Translation

- Rule-based MT: Systran (Toma, 1988)/Babel Fish 2001
- Google Translate (“GT”, Och 2006)

Google Translate – Statistical MT [Och 2006]

- Corpus-based: exploits corpora of existing **aligned** translations
- SMT = Language Model + Statistical Decision Theory: $\hat{e}(f) = argmin_e \sum_{e'} L(\mathrm{e}, e', f) * Pr(e|f)$
- *Pr(e|f)* is modelled using *p(e|f)* **acquired from data** (corpora)
- Data is **BIG:**
  - 200M in [Och 2006]
  - 5M-4,800M words

Tools

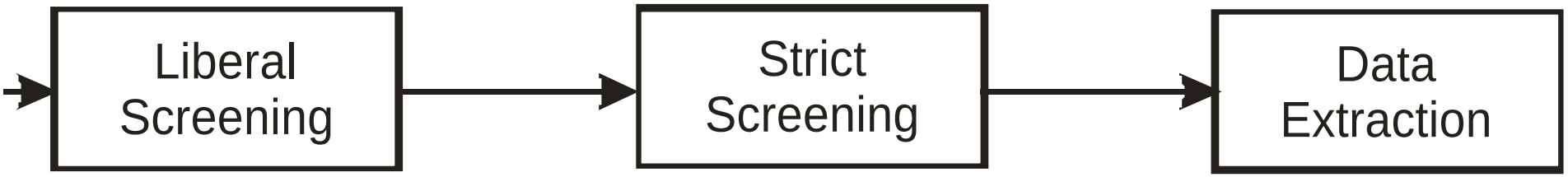
- EM
- Advanced heuristic search (beam)
- Bayesian modeling
- Markovian modeling
- optimization

Recent use of deep learning techniques for text mining and SMT

- Learning meaning representations *relation(subject, object)* from WordNet, Wikipedia, etc. [Bordes et al . 11]
- [Mikolov 13]
  - Learning a large bilingual dictionary from
    - A small dictionary
    - A learned model for the tow languages
    - A liner mapping between mdels

Is Knowledge Really Helpful for a Given Text Mining Task [Matwin et al. JAMIA 2010]?

Systematic Reviews:



~10^4	~10^3	~10^1
abstracts:	full papers:	fragments:

- classification
- Eventually, information extraction

Representation

- Bag of words: obvious shortcomings
- Co-occurrence [Pedersen, Razavi]
- Knowledge-based distance
- t-SNE approach to visualization [van der Maaten, Hinton 2008]

Co-occurrence

- Capturing non-local semantic relationships between words in a corpus
- Used on its own as a text representation [Liu et al. 12], [Matwin, De Koninck, Razavi 10].

Other knowledge-based distances

- MeSH distance (ontology-based, hence semantic) - MMTX tool extracts MeSH terms from text
- OMIOTIS distance – combines geometric distance (tf-idf median) with sematic distance (WordNet based)

We use MMTX and Omiotis in t-SNE

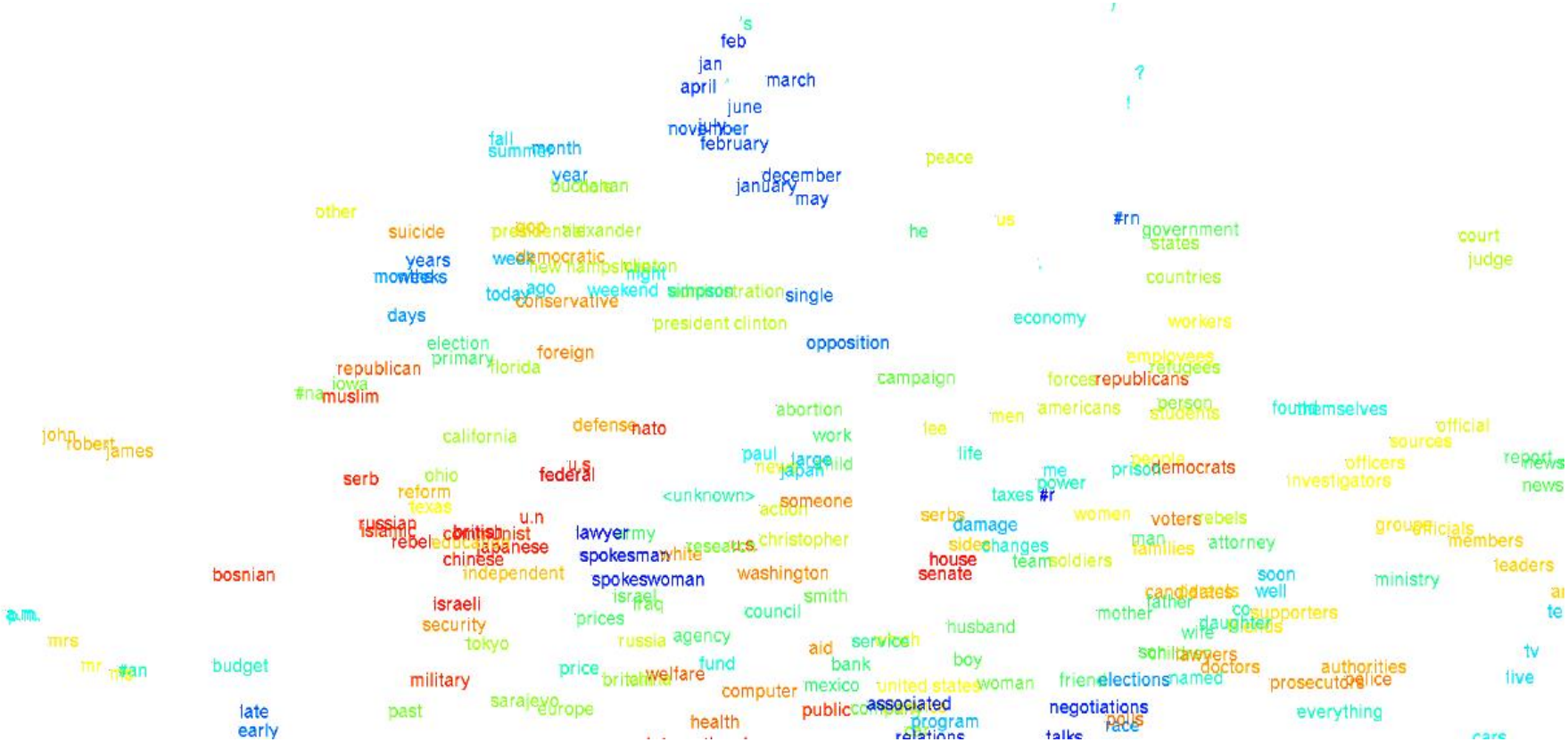
- Hypothesis: semantic (knowledge-based) approaches should give better (clearer) visualizations

Visualization of text data (documents)

with Dr. Vera Sazonova

- Multi-dimensional scaling MDS
- Stochastic Neighbor Embedding t-SNE "t-SNE is capable of capturing much of the local structure of the high-dimensional data very well, while also revealing global structure such as the presence of clusters at several scales."

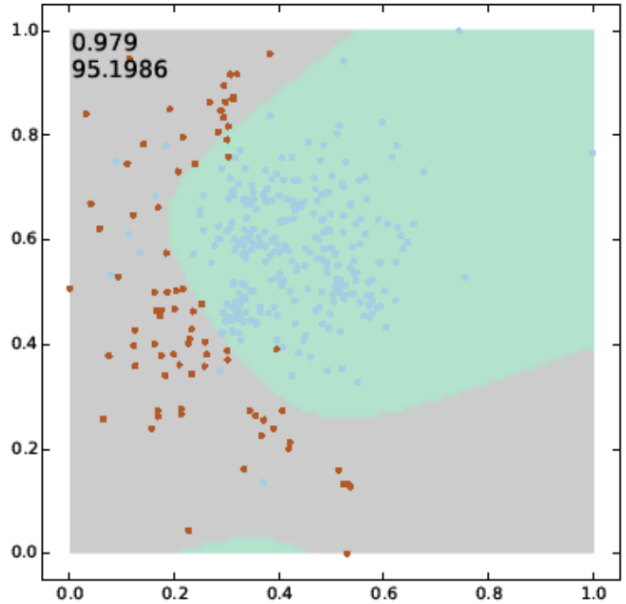
1000 most common words from news reports, 100 features (from [van der Maaten, Hinton 2008])



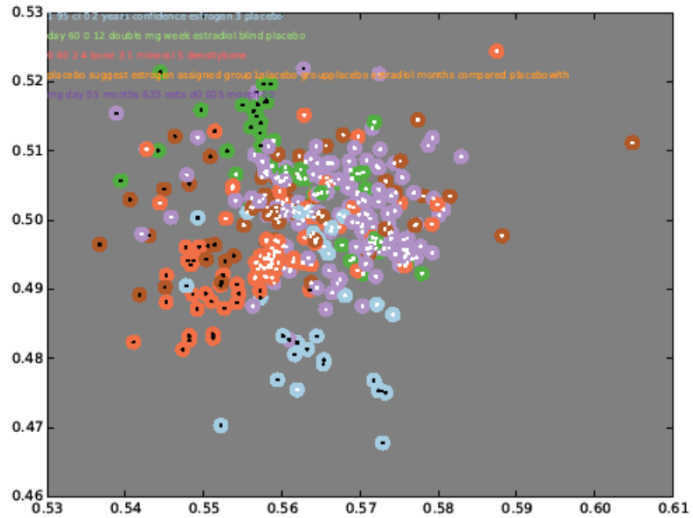
classes

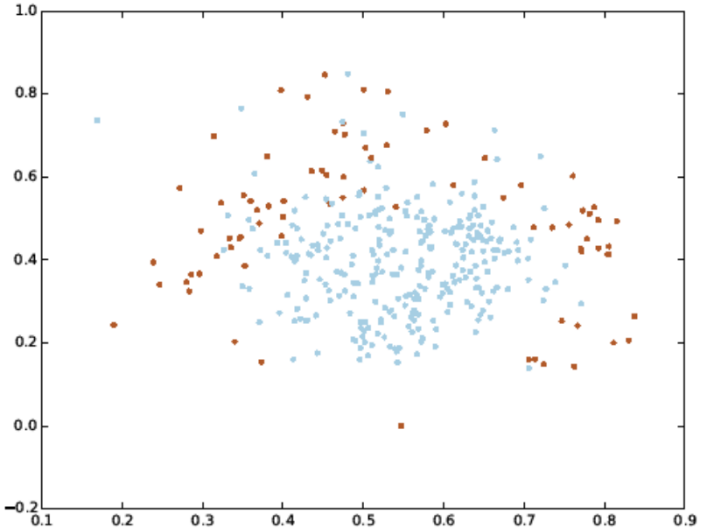
Cosine distance

LDA topics

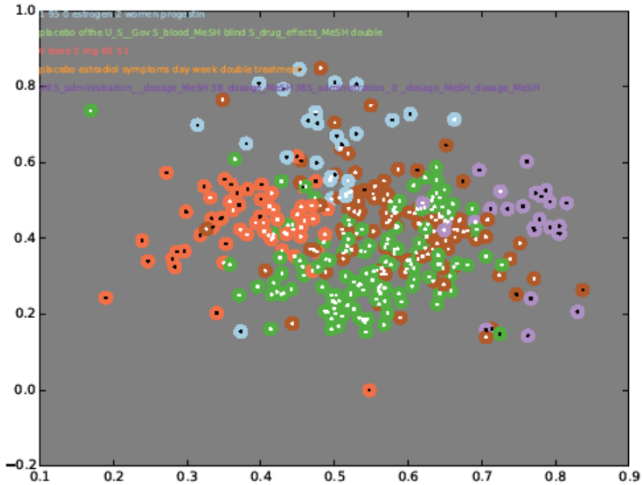


Text and abstract only



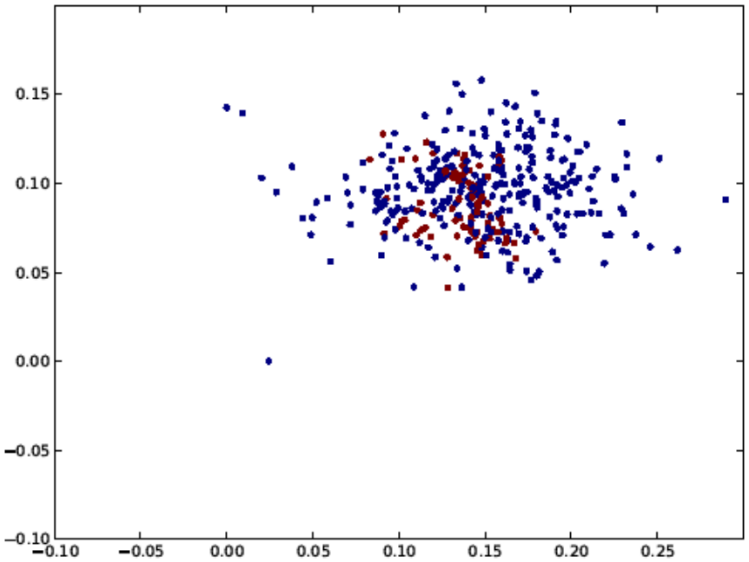


Text, abstract and MeSH

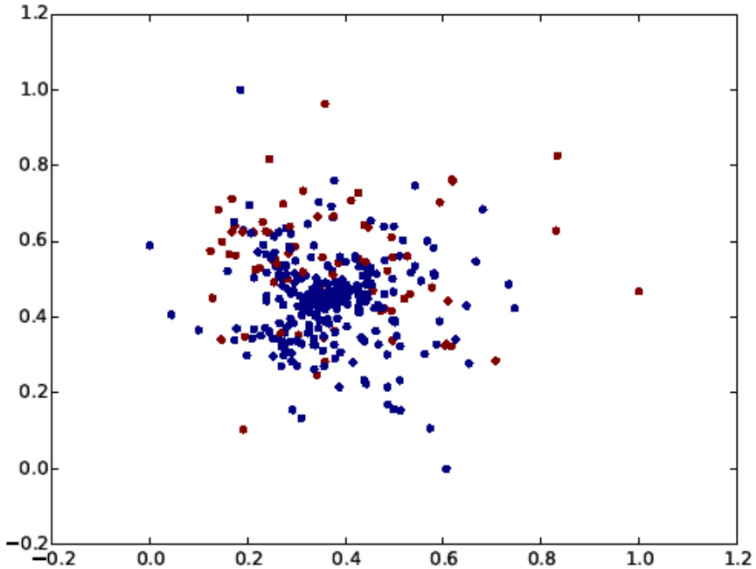


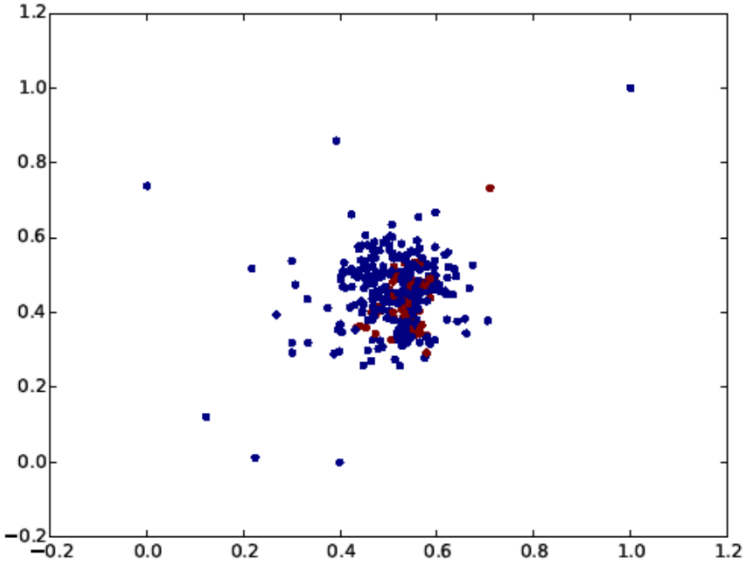
Cosine distance

Omiotis distance

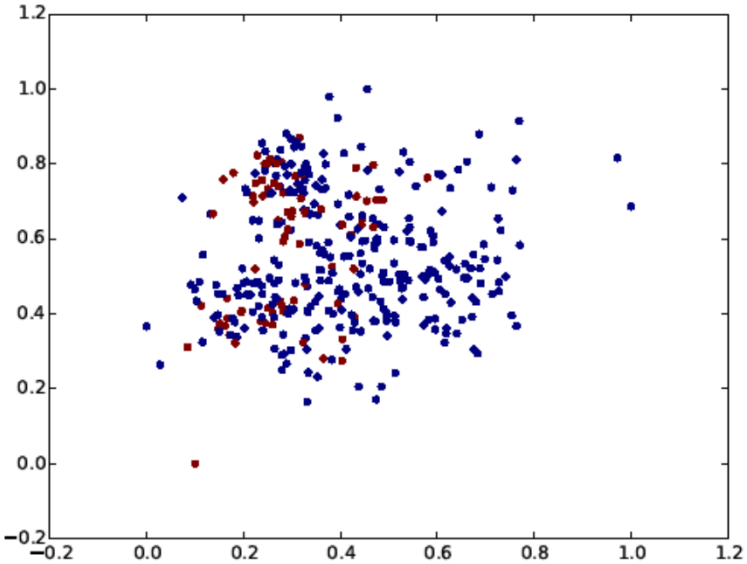


Wiki, Text, abstr., MeSH



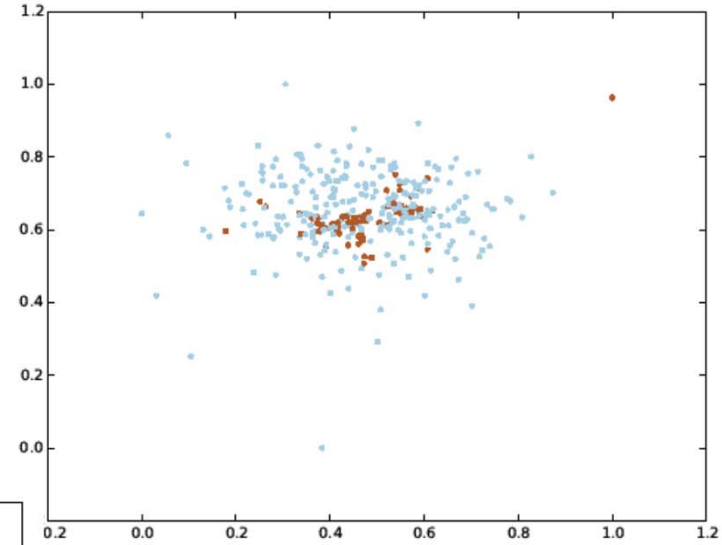


MeSH, MMTX, path Dist.

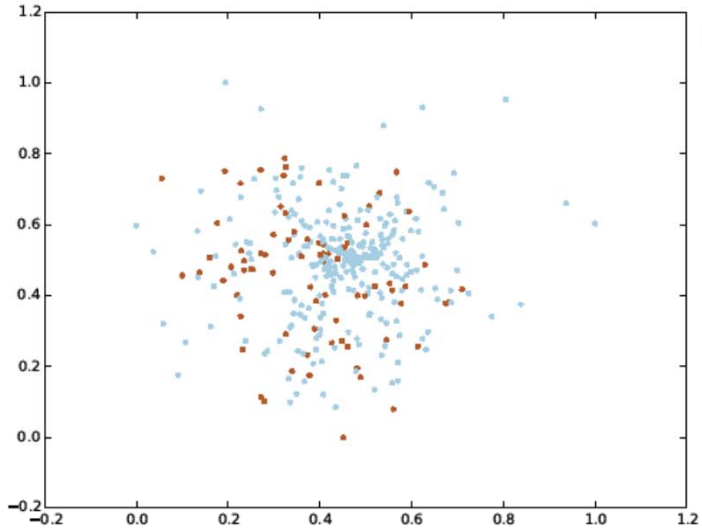


Word2Vector results

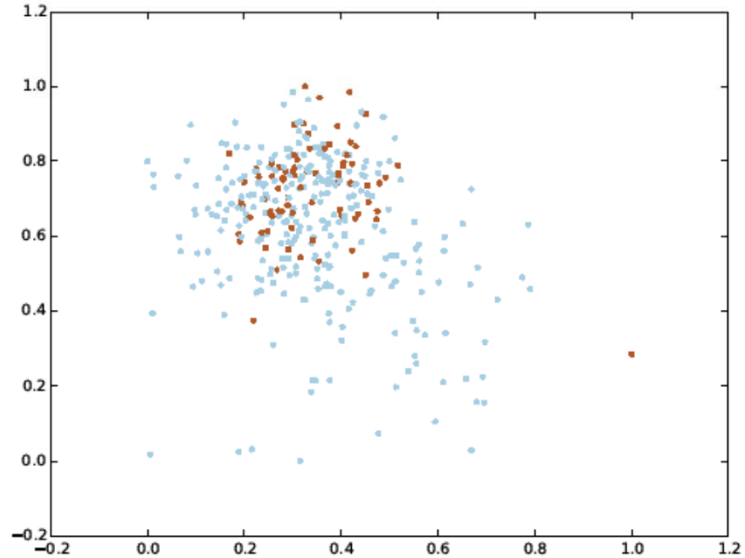
- Baseline



- Omiotis



- Vectors added



- Despite many trials, no evidence that knowledge-based distances result in better visualization

So what's the future for TM? Some guesses…

- Non-local representations
  - Co-occurrence → LDA
  - Semi-relational – eg frequency of co-occurrence of neighbouring n-grams [Giannakopoulos 12]
  - Neural networks-"smoothened" bag of words [Mikolov 13]
- Good document visualization techniques [Stasko-JIGSAW]

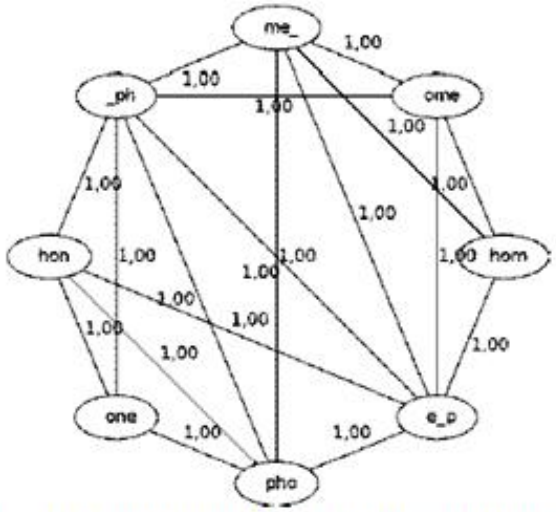
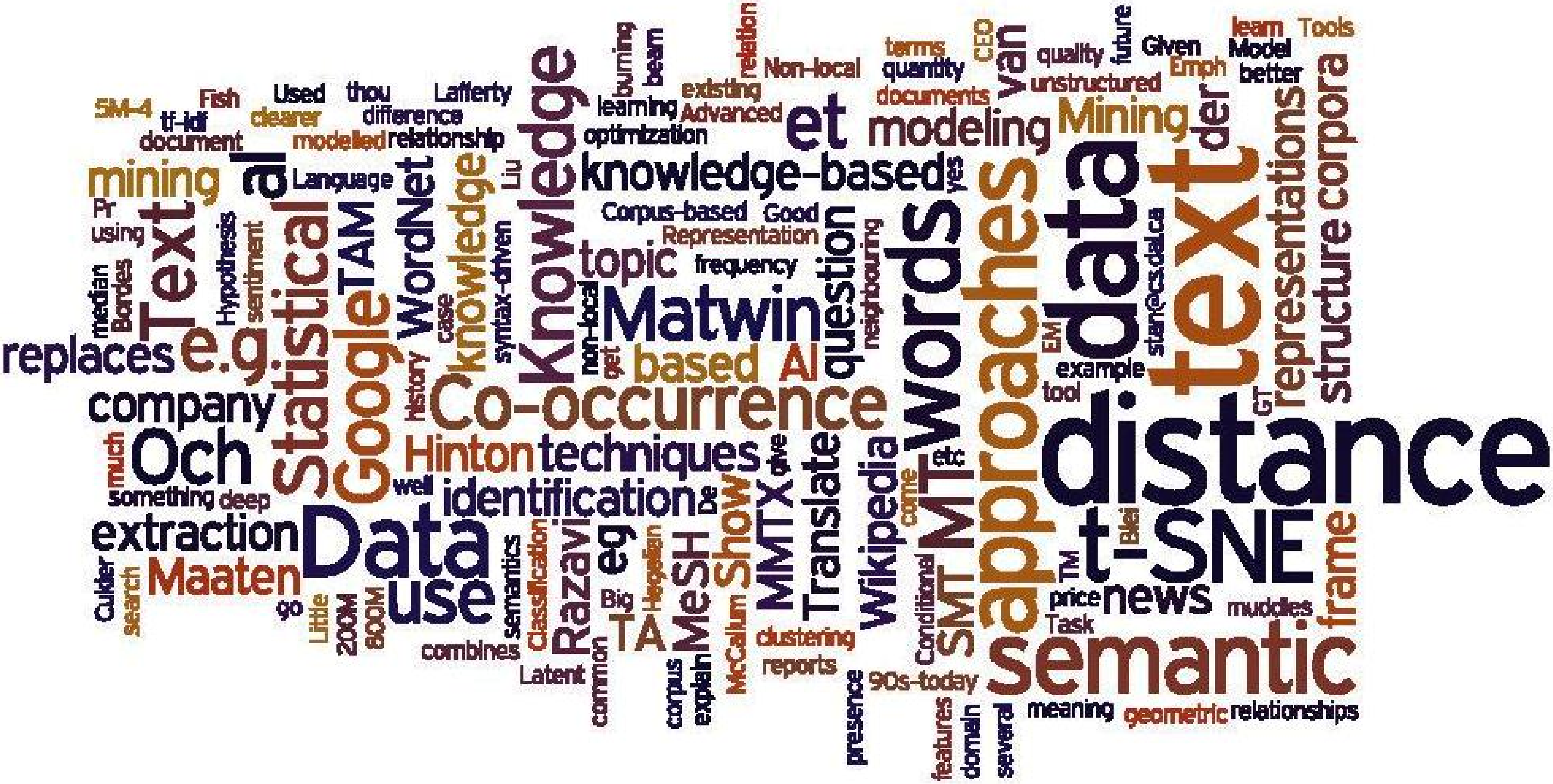


Figure 1: Tri-gram graph of "home_phone" string.

- Massive corpora/big data approaches (eg to question answering)
  - Google N-grams
- Use of Wikipedia for ML/NLP
- From MT to interpretation



stan@cs.dal.ca