Introduction to NLP

Searching for Meaning in Text

Dr. Tony Russell-Rose FBCS CITP CEng

uxlabs

Goldsmiths
UNIVERSITY OF LONDON

2Dsearch

Introductions

Tony Russell-Rose, PhD

Senior Lecturer, Goldsmiths Director, UXLabs

- Led R&D / technology innovation at Microsoft, Canon, Reuters, BT Labs, HP Labs, Oracle
- Visiting Professor of Cognitive Computing & AI, Essex University
- PhD (natural language interfaces), MSc in human-computer interaction, first degree in engineering (human factors)
- 5 patents, 80+ scientific and technical papers on AI/NLP, information retrieval and UX
- Founder, 2Dsearch: next generation advanced search
- Director, UXLabs: UX research & design consultancy
- Honorary Visiting Fellow at City University Centre for Interactive Systems Research
- Founder of Search Solutions conference series. Served as Vice-chair of BCS IRSG, Chair of CIEHF HCI Group
- Blog: http://isquared.wordpress.com

Industry Expertise

- Media & publishing
- Healthcare
- Business intelligence
- eCommerce

Domain Expertise

- UX research & design
- Information architecture
- Search & information retrieval
- AI + machine learning

NLP – a solved problem?

- As humans we do it effortlessly … don’t we?

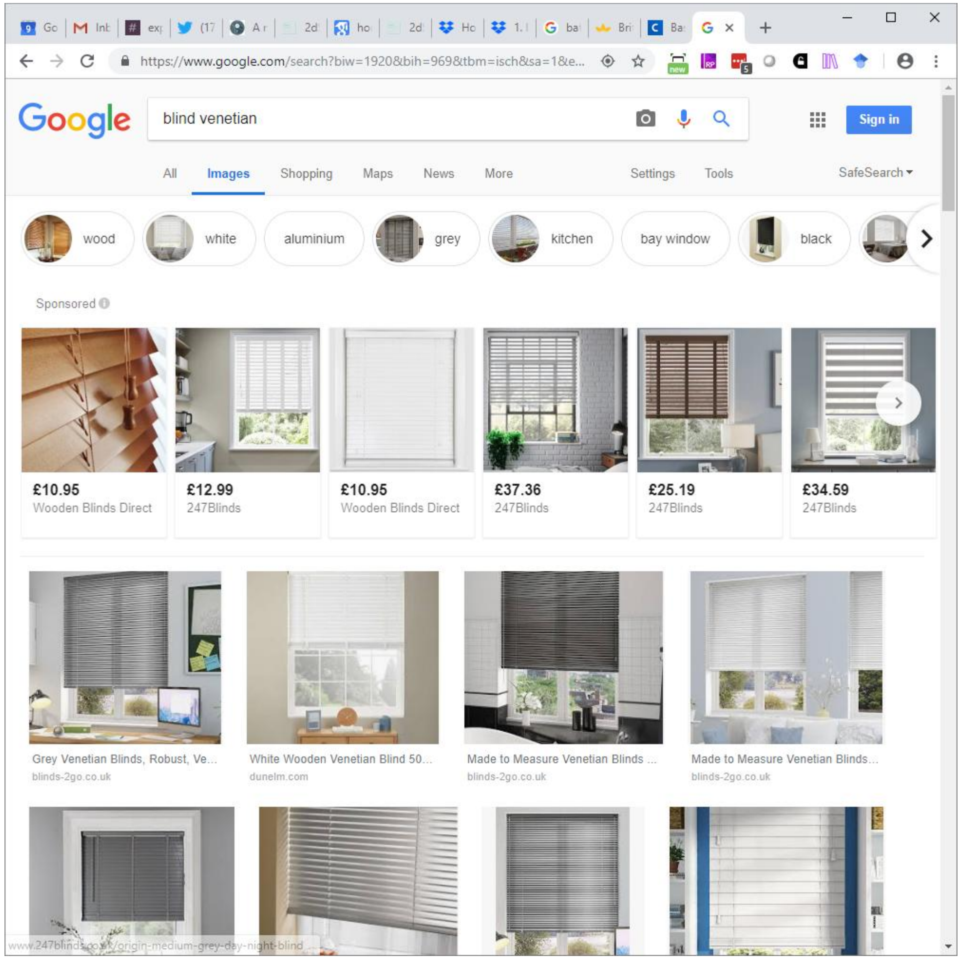
- DRUNK GETS NINE YEARS IN VIOLIN CASE
- PROSTITUTES APPEAL TO POPE
- STOLEN PAINTING FOUND BY TREE
- RED TAPE HOLDS UP NEW BRIDGE
- DEER KILL 300,000
- RESIDENTS CAN DROP OFF TREES
- INCLUDE CHILDREN WHEN BAKING COOKIES
- MINERS REFUSE TO WORK AFTER DEATH

Problems with Text

- Polysemy
  - One word maps to many concepts
  - e.g. *bat*
- Synonymy
  - One concept maps to many words

Problems with Text

- Word order
  - *Venetian blind vs. Blind venetian*



blind venetian

£10.95 Wooden Blinds Direct
£12.99 247Blinds
£10.95 Wooden Blinds Direct
£37.36 247Blinds
£25.19 247Blinds
£34.59 247Blinds

Grey Venetian Blinds, Robust, Ve... blinds-2go.co.uk
White Wooden Venetian Blind 50... dunelm.com
Made to Measure Venetian Blinds ... blinds-2go.co.uk
Made to Measure Venetian Blinds... blinds-2go.co.uk

www.247blinds.co.uk/origin-medium-grey-day-night-blind

Problems with Text

- Language is generative
  - `Starbucks coffee is the best`
  - `The place I like most when I need to feed my caffeine addiction is the company from Seattle with branches everywhere`
- Many different ways to express a given idea
  - Synonymy, paraphrase, metaphor, etc.

Problems with Text

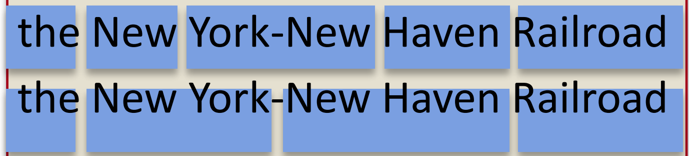
- Language is changing
  - I want to buy a mobile
- Ill-formed input
  - “accomodation office”
- Co-ordination, negation, etc.
  - This is not a talk about neuro-linguistic programming
- Multi-linguality
  - Claudia Schiffer is on the cover of Elle
- Sarcasm, irony, slang, jargon, etc.
  - That was a wicked lecture
  - Yep – the coffee break was the best part

Why else is natural language understanding difficult?

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

segmentation issues



the New York-New Haven Railroad

the New York-New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
Retweet
bromance

world knowledge

Mary and Sue are sisters.

Mary and Sue are mothers.

tricky entity names

Where is *A Bug's Life* playing …

Let It Be was recorded …

… a mutation on the *for* gene …

But that's what makes it fun!

NLP Fundamentals (word level)

- Language is AMBIGUOUS
  - To determine structure, we must resolve ambiguity!

- Lexical analysis (tokenisation)
  - `The cat sat on the mat`
  - `I can't tokenise this sentence`

- Stop word removal
  - No definitive list
  - `The Who, The The, Take That…`
  - `To be or not to be`

NLP Fundamentals (word level)

- Stemming
  - fishing, fished, fish, fisher -> fish
- Lemmatization
  - Linguistically principled analysis
  - Passing -> pass + ING
  - Were -> be + PAST
  - Delegate = de-leg-ate (?)
  - Ratify = rat-ify (?)
- Morphology (prefixes, suffixes, etc.)
  - Gebäudereinigungsfirmenangestellter -> Gebäude + Reinigung + Firma + Angestellter (building + cleaning + company + employee)

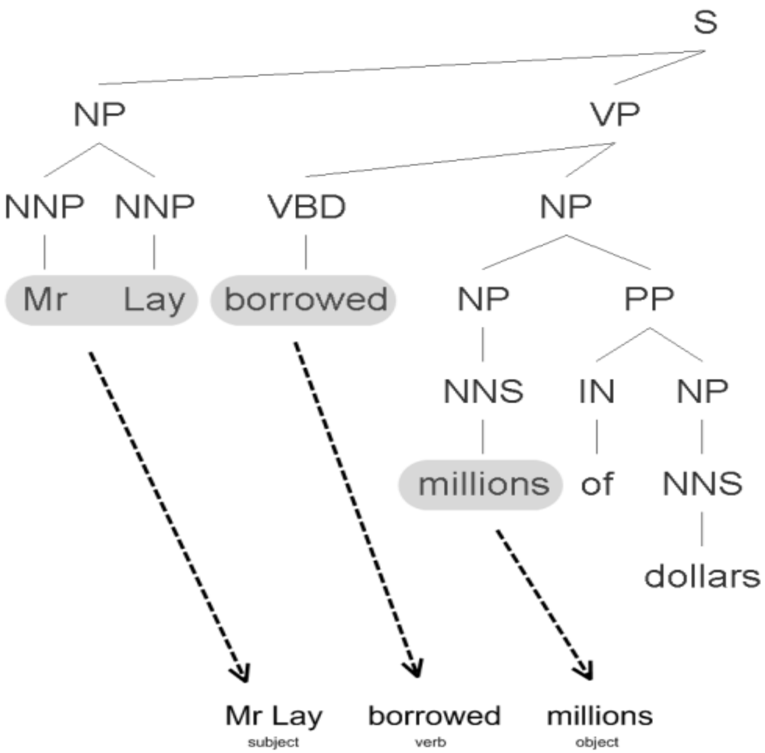
NLP Fundamentals (sentence level)

- Syntax (part of speech tagging)
  - `book -> NOUN, VERB`
  - `that -> DETERMINER`
  - `flight -> NOUN`
  - `Book that flight -> VB DT NN`

- Ambiguity problem
  - `Time flies like an arrow -> ?`
  - `Fruit flies like a banana -> ?`
  - `Eats shoots and leaves -> ?`

NLP Fundamentals (sentence level)

- Parsing (grammar)
  - `I saw a venetian blind`
  - `I saw a blind venetian`
  - `I saw the man on the hill with a telescope`
  - `Rugby is a game played by men with odd-shaped balls`
- Sentence boundary detection
  - `Punctuation denotes the end of a sentence!`
  - `"But not always!", said Fred...`



NLP Fundamentals (paragraph level)

- Anaphora resolution
  - “John dropped a plate. It broke.”
- Anaphora resolution relies on knowledge:
  - *We gave the bananas to the monkeys because* ***they*** *were hungry.*
  - *We gave the bananas to the monkeys because* ***they*** *were ripe.*
  - *We gave the bananas to the monkeys because* ***they*** *were here.*

What do we want?

Anaphora resolution!

When do we want it?

When do we want what?

Information Extraction

Event: Curriculum mtg
Date: Jan-16-2012
Start: 10:00am
End: 11:30am
Where: Gates 159

Subject: **curriculum meeting**
Date: January 15, 2012
To: Dan Jurafsky

Hi Dan, we've now scheduled the curriculum meeting.
It will be in Gates 159 tomorrow from 10:00-11:30.
-Chris

Create new Calendar entry

Named Entity Recognition

- Identification of key concepts,
  - e.g. people, places, organisations, etc.
  - Also postcodes, temporal/numerical expressions, etc.

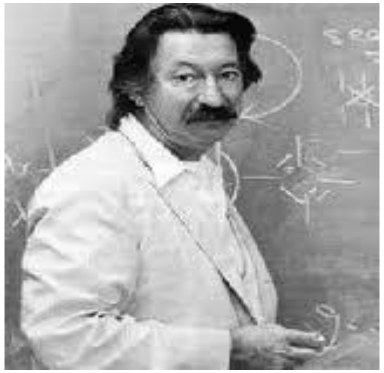
- *"Mexico has been trying to stage a recovery since the beginning of this year and it's always been getting ahead of itself in terms of fundamentals," said Matthew Hickman of Lehman Brothers in New York."*

Persons	Organisations	Cities	Countries
Matthew Hickman	Lehman Brothers	New York	Mexico

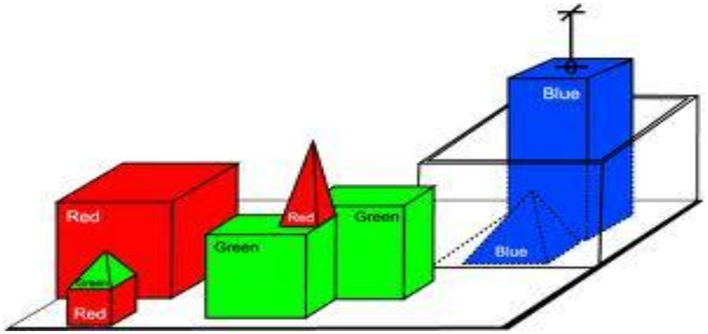
NLP Applications

- Human-computer interfaces (chatbots etc.)
- Text classification
- Text summarisation
- Machine translation
- Speech recognition & synthesis
- Natural language generation
- Text mining
- Question answering
- Sentiment Analysis

Early NLP Systems



- ELIZA
  - Wiezenbaum 1966
  - Simple pattern matching
- SHRDLU
  - Winograd 1970
  - Natural language understanding
  - Comprehensive grammar of English

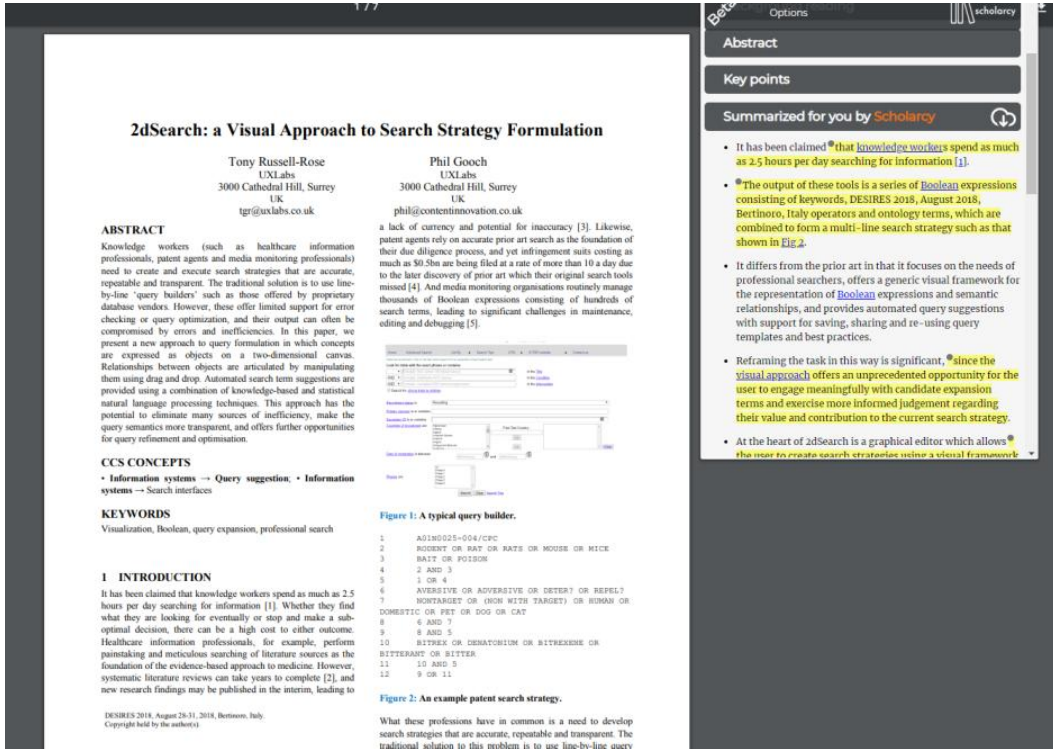


Text Categorization

- Media, publishing, libraries...
  - Classify news stories, papers, books...
- Spam detection
- Search engines:
  - classify query intent, e.g. search Google for 'LOG313'

Summarisation

- Summarization
  - single-document vs. multi-document
- Search result snippets
- Word processing tools
- Research / analysis tools



2dSearch: a Visual Approach to Search Strategy Formulation

Tony Russell-Rose
UXLabs
3000 Cathedral Hill, Surrey
UK
tgr@uxlabs.co.uk

Phil Gooch
UXLabs
3000 Cathedral Hill, Surrey
UK
phil@contentinnovation.co.uk

ABSTRACT

Knowledge workers (such as healthcare information professionals, patent agents and media monitoring professionals) need to create and execute search strategies that are accurate, repeatable and transparent. The traditional solution is to use line-by-line 'query builders' such as those offered by proprietary database vendors. However, these offer limited support for error checking or query optimization, and their output can often be compromised by errors and inefficiencies. In this paper, we present a new approach to query formulation in which concepts are expressed as objects on a two-dimensional canvas. Relationships between objects are articulated by manipulating them using drag and drop. Automated search term suggestions are provided using a combination of knowledge-based and statistical natural language processing techniques. This approach has the potential to eliminate many sources of inefficiency, make the query semantics more transparent, and offers further opportunities for query refinement and optimisation.

CCS CONCEPTS

• **Information systems → Query suggestion**; • **Information systems →** Search interfaces

KEYWORDS

Visualization, Boolean, query expansion, professional search

1 INTRODUCTION

It has been claimed that knowledge workers spend as much as 2.5 hours per day searching for information [1]. Whether they find what they are looking for eventually or stop and make a sub-optimal decision, there can be a high cost to either outcome. Healthcare information professionals, for example, perform painstaking and meticulous searching of literature sources as the foundation of the evidence-based approach to medicine. However, systematic literature reviews can take years to complete [2], and new research findings may be published in the interim, leading to a lack of currency and potential for inaccuracy [3]. Likewise, patent agents rely on accurate prior art search as the foundation of their due diligence process, and yet infringement suits costing as much as $0.5bn are being filed at a rate of more than 10 a day due to the later discovery of prior art which their original search tools missed [4]. And media monitoring organisations routinely manage thousands of Boolean expressions consisting of hundreds of search terms, leading to significant challenges in maintenance, editing and debugging [5].

DESIRES 2018, August 28-31, 2018, Bertinoro, Italy.
Copyright held by the author(s).

Figure 1: A typical query builder.

```
1       A01M0025-004/CPC
2       RODENT OR RAT OR RATS OR MOUSE OR MICE
3       BAIT OR POISON
4       2 AND 3
5       1 OR 4
6       AVERSIVE OR ADVERSIVE OR DETER? OR REPEL?
7       NONTARGET OR (NON WITH TARGET) OR HUMAN OR
DOMESTIC OR PET OR DOG OR CAT
8       6 AND 7
9       8 AND 5
10      BITREX OR DENATONIUM OR BITREXENE OR
BITTERANT OR BITTER
11      10 AND 5
12      9 OR 11
```

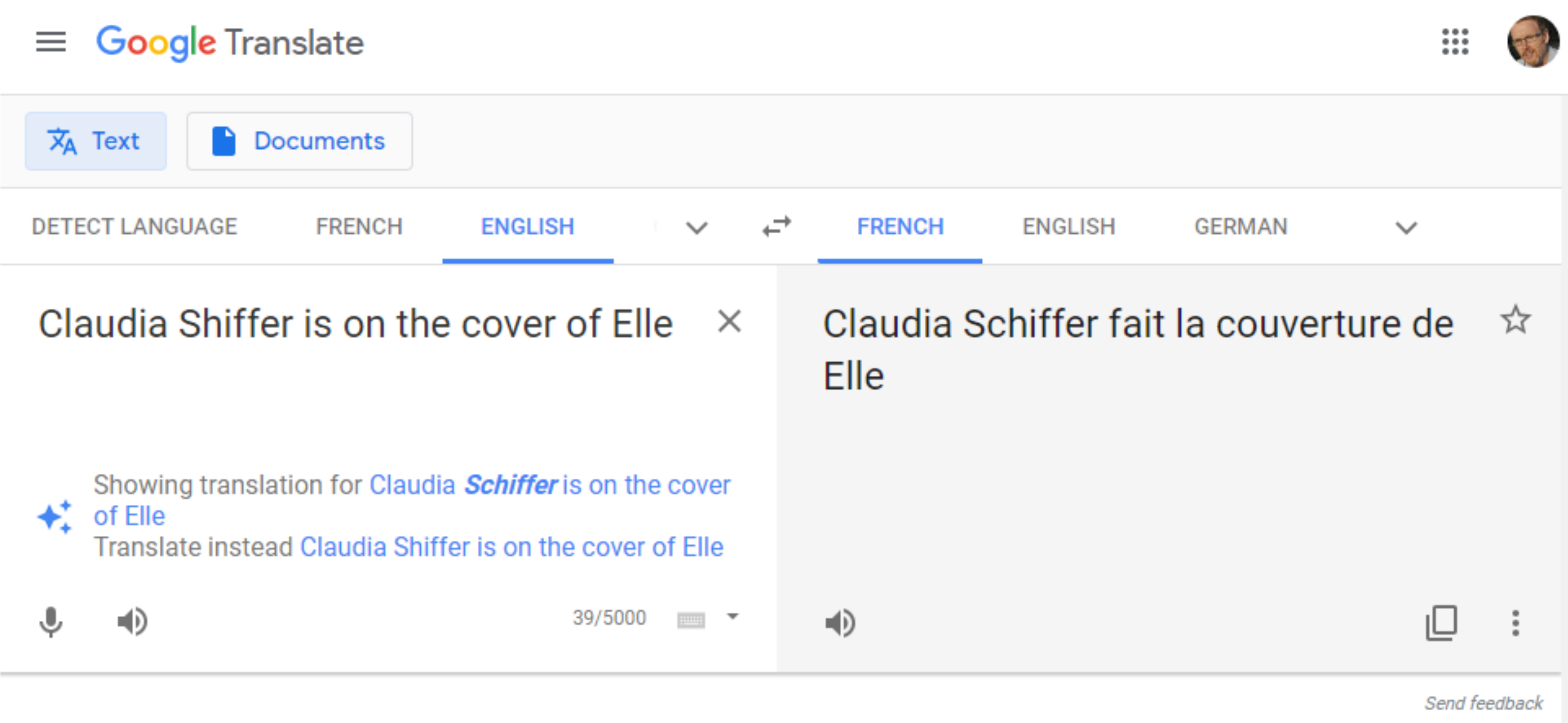
Figure 2: An example patent search strategy.

What these professions have in common is a need to develop search strategies that are accurate, repeatable and transparent. The traditional solution to this problem is to use line-by-line query

Summarized for you by Scholarcy

- It has been claimed that knowledge workers spend as much as 2.5 hours per day searching for information [1].
- The output of these tools is a series of Boolean expressions consisting of keywords, DESIRES 2018, August 2018, Bertinoro, Italy operators and ontology terms, which are combined to form a multi-line search strategy such as that shown in Fig 2.
- It differs from the prior art in that it focuses on the needs of professional searchers, offers a generic visual framework for the representation of Boolean expressions and semantic relationships, and provides automated query suggestions with support for saving, sharing and re-using query templates and best practices.
- Reframing the task in this way is significant, since the visual approach offers an unprecedented opportunity for the user to engage meaningfully with candidate expansion terms and exercise more informed judgement regarding their value and contribution to the current search strategy.
- At the heart of 2dSearch is a graphical editor which allows the user to create search strategies using a visual framework

Machine translation



Claudia Shiffer is on the cover of Elle

Claudia Schiffer fait la couverture de Elle

Showing translation for Claudia ***Schiffer*** is on the cover of Elle
Translate instead Claudia Shiffer is on the cover of Elle

Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation, 2016.

Dialog systems

- Chatbots
- Smart speakers
- Smartphone assistants
- Call handling systems
  - Travel
  - Hospitality
  - Banking

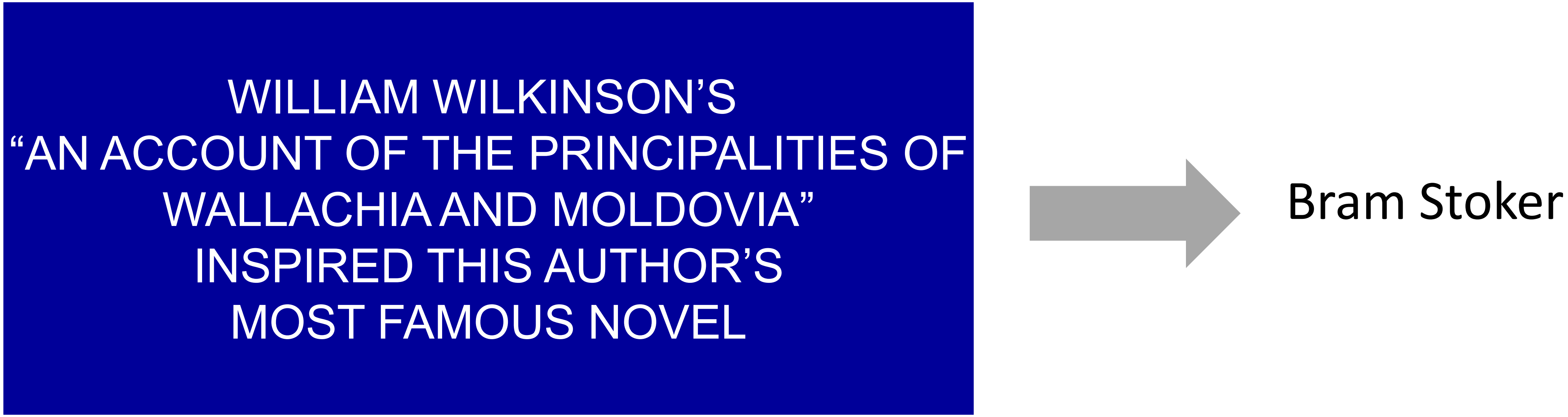
What can I help you with?

Text Mining

- Analogy with *Data Mining*
  - Discover or infer new knowledge from unstructured text resources
- A <-> B and B <-> C
  - Infer A <-> C?
  - e.g. link between migraine headaches and magnesium deficiency
- Applications in life sciences, media/publishing, counter terrorism, competitive intelligence

Question Answering: IBM's Watson

- Won Jeopardy on February 16, 2011!



WILLIAM WILKINSON'S
"AN ACCOUNT OF THE PRINCIPALITIES OF
WALLACHIA AND MOLDOVIA"
INSPIRED THIS AUTHOR'S
MOST FAMOUS NOVEL

Bram Stoker

Sentiment Analysis

- Identify and extract *subjective* information
  - Predict stock market movements
- Sub-tasks:
  - Identify *polarity*, e.g. of movie reviews
    - e.g. positive, negative, or neutral
  - Identify emotional states
    - e.g. angry, sad, happy, etc.
  - Subjectivity/objectivity identification
    - E.g. "fact" from opinion
  - Feature/aspect-based
    - Differentiate between specific features or aspects of entities

Language Technology

mostly solved

Spam detection

Let's go to Agra! ✓

Buy V1AGRA … ✗

Part-of-speech (POS) tagging

ADJ	ADJ	NOUN	VERB	ADV
Colorless	green	ideas	sleep	furiously.

Named entity recognition (NER)

PERSON		ORG		LOC
Einstein	met with	UN	officials in	Princeton

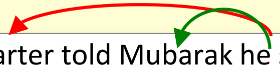
making good progress

Sentiment analysis

Best roast chicken in San Francisco! 👍

The waiter ignored us for 20 minutes. 👎

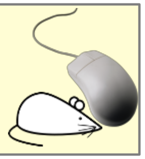
Coreference resolution



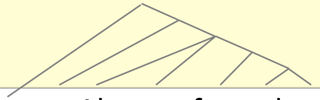
Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my ***mouse***.



Parsing



I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕… ⇨

The 13th Shanghai International Film Festival…

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up
The S&P500 jumped
Housing prices rose

⇨ Economy is good

Dialog

Where is Citizen Kane playing in SF?

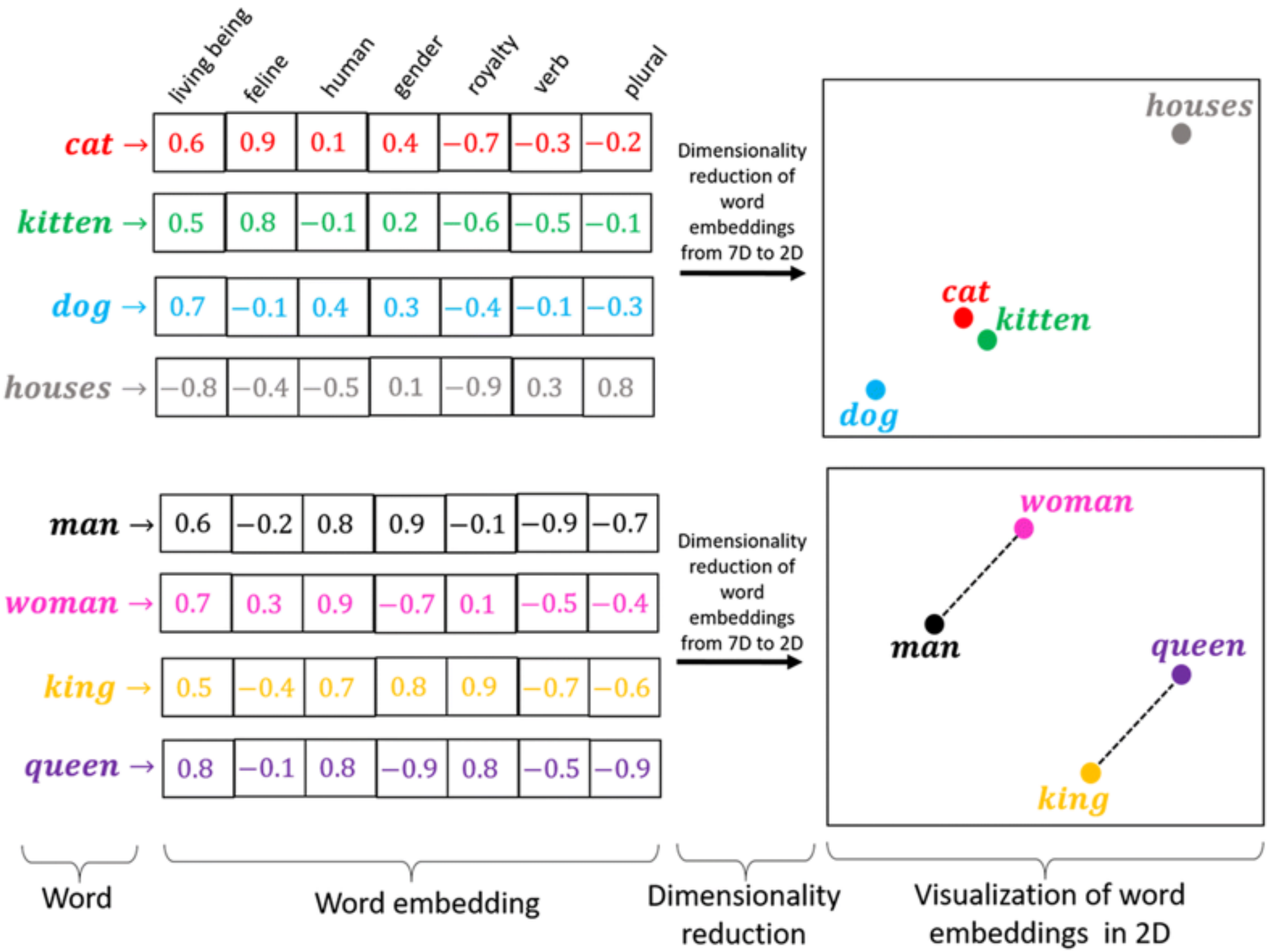
Castro Theatre at 7:30. Do you want a ticket?

Platforms & Toolkits

- Many commercial vendors
- Freely available:
  - GATE (Sheffield University)
  - Stanford CoreNLP
  - Apache OpenNLP
  - TextBlob
  - Textacy
  - Spacy
  - Nltk
- Also: RapidMiner, R, Weka

Distributional approaches

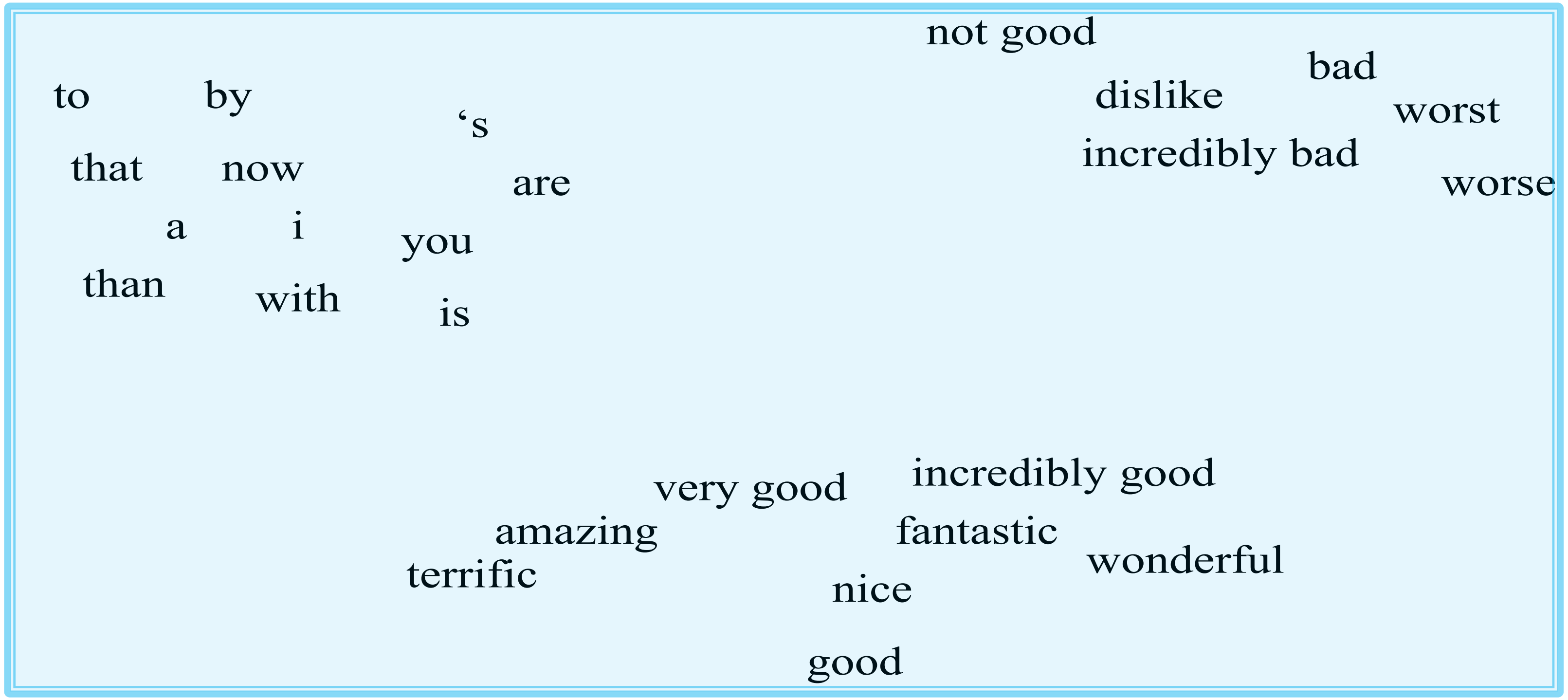
- Word meaning is defined by context
  - “You **shall know a word** by the **company it keeps”**
- Context-free word embeddings
  - Word2vec, GloVe, FastText
- Bidirectional approaches
  - ELMo, BERT, etc.

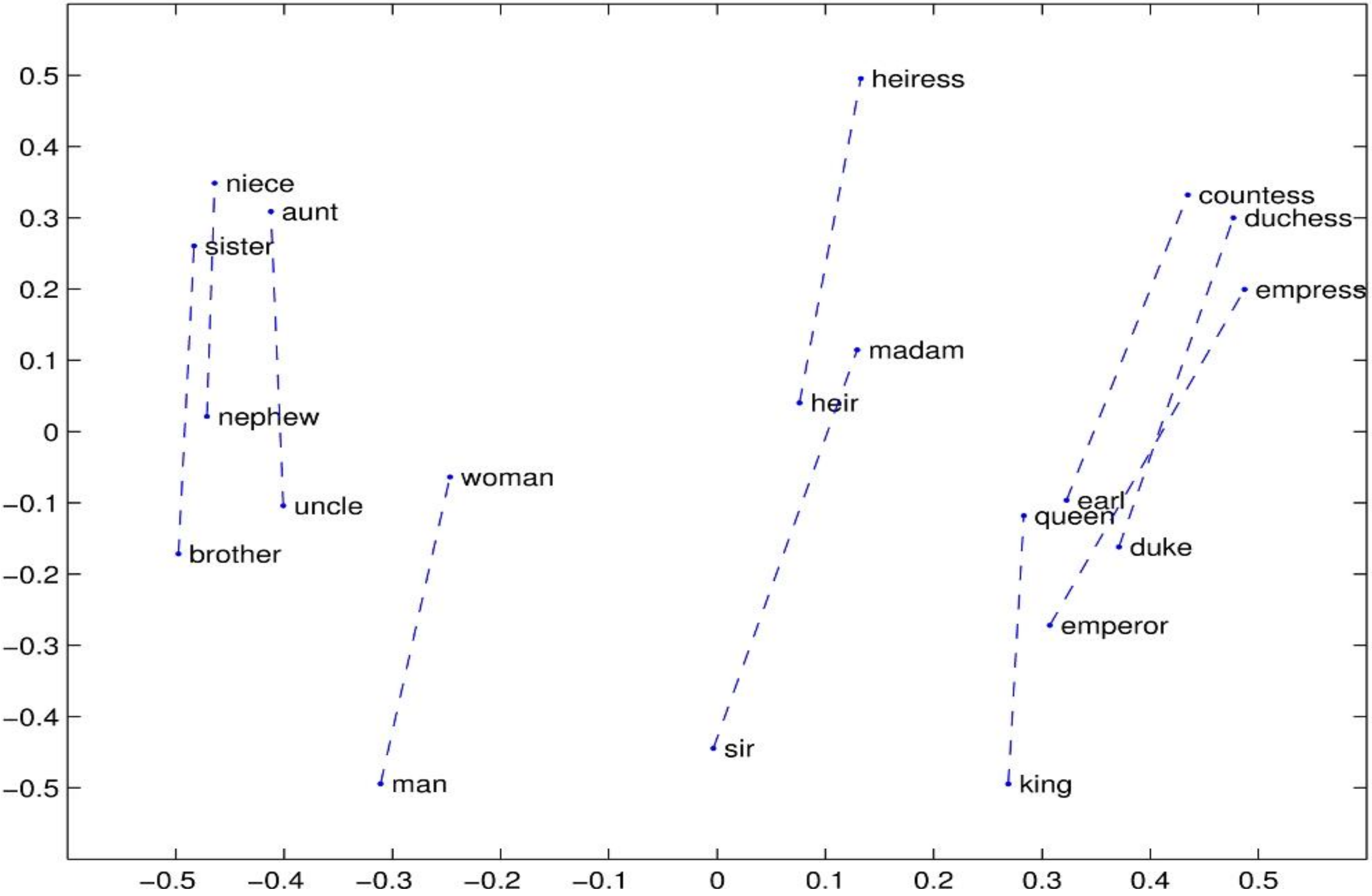


	living being	feline	human	gender	royalty	verb	plural
cat →	0.6	0.9	0.1	0.4	−0.7	−0.3	−0.2
kitten →	0.5	0.8	−0.1	0.2	−0.6	−0.5	−0.1
dog →	0.7	−0.1	0.4	0.3	−0.4	−0.1	−0.3
houses →	−0.8	−0.4	−0.5	0.1	−0.9	0.3	0.8
man →	0.6	−0.2	0.8	0.9	−0.1	−0.9	−0.7
woman →	0.7	0.3	0.9	−0.7	0.1	−0.5	−0.4
king →	0.5	−0.4	0.7	0.8	0.9	−0.7	−0.6
queen →	0.8	−0.1	0.8	−0.9	0.8	−0.5	−0.9

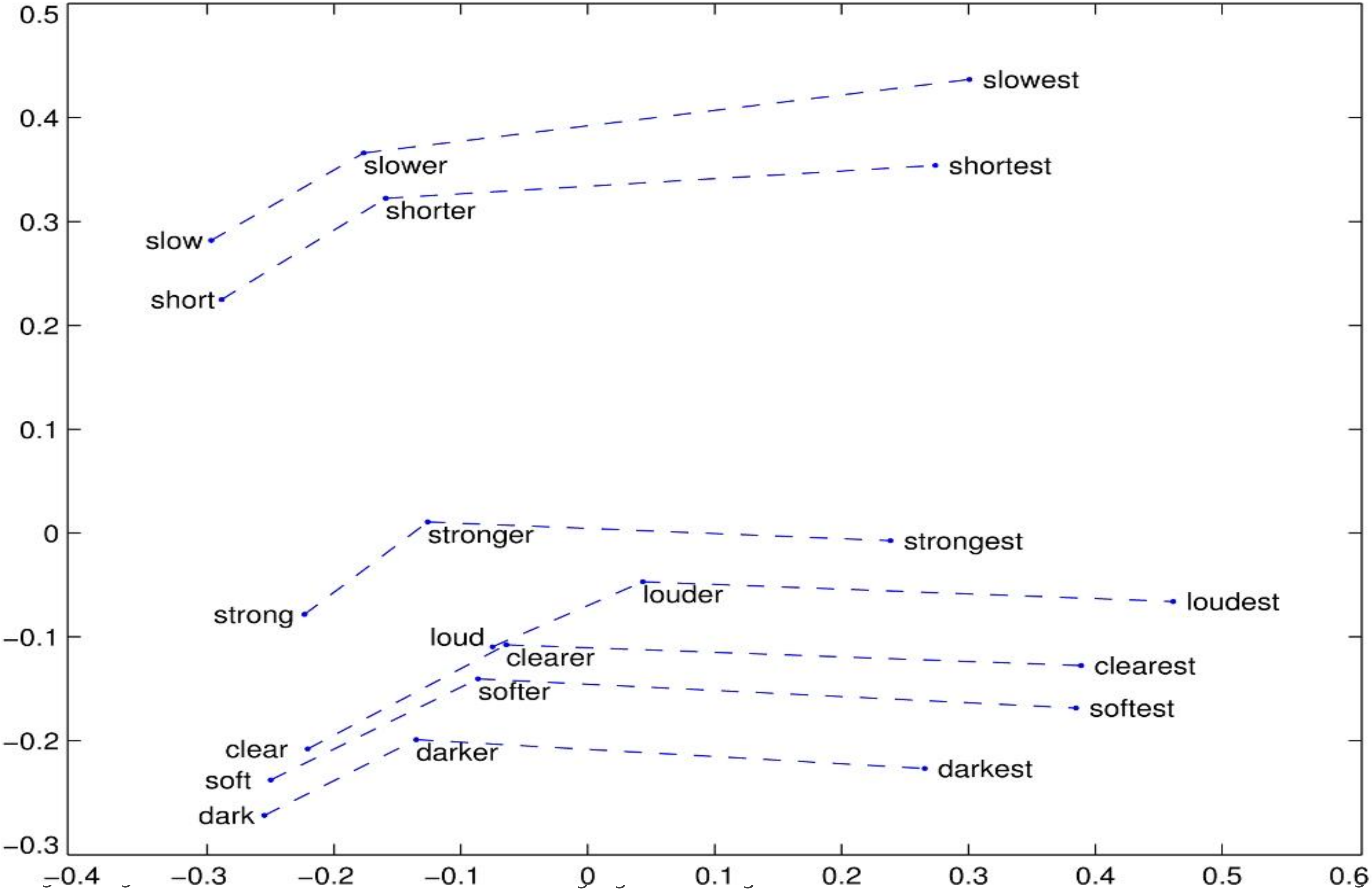
Dimensionality reduction of word embeddings from 7D to 2D
houses
cat
kitten
dog
Dimensionality reduction of word embeddings from 7D to 2D
woman
man
queen
king
Word
Word embedding
Dimensionality reduction
Visualization of word embeddings in 2D

Similarity can be projected in 2D



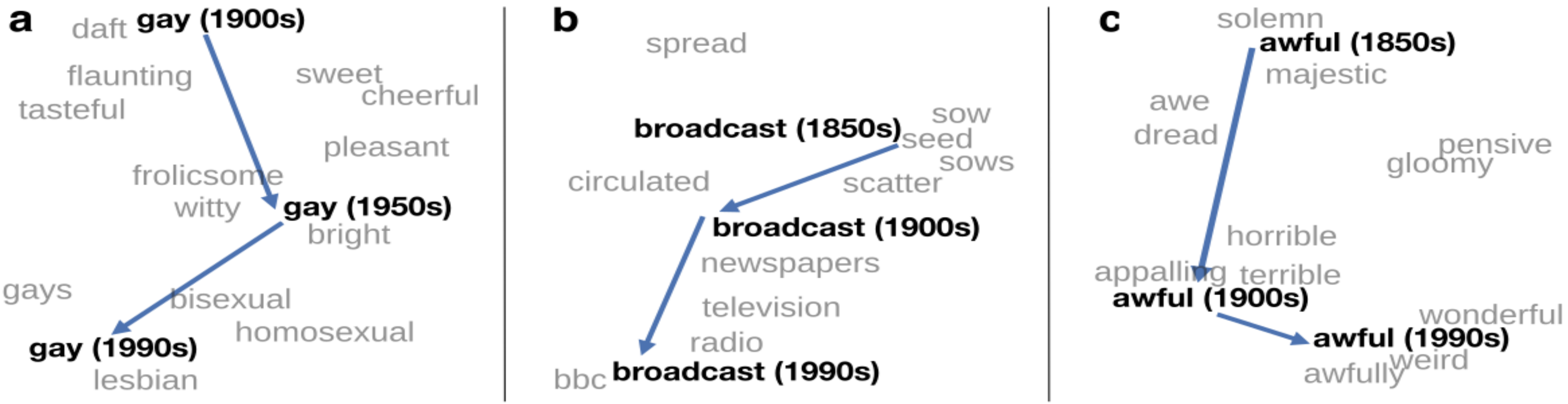


heiress
niece
aunt
sister
countess
duchess
empress
madam
heir
nephew
woman
uncle
earl
queen
duke
brother
emperor
sir
man
king
0.5
0.4
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.5
−0.4
−0.3
−0.2
−0.1
0
0.1
0.2
0.3
0.4
0.5


slowest
slower
shortest
shorter
slow
short
stronger
strongest
strong
louder
loudest
loud
clearer
clearest
softer
softest
clear
darker
darkest
soft
dark
0.5
0.4
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
−0.4
−0.3
−0.2
−0.1
0
0.1
0.2
0.3
0.4
0.5
0.6

Diachronic changes

~30 million books, 1850-1990, Google Books data



Embeddings reflect cultural biases

- Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." In *Advances in Neural Information Processing Systems*, pp. 4349-4357. 2016.
- Paris : France :: Tokyo : x?
  - x = Japan
- father : doctor :: mother : x?
  - x = nurse
- man : computer programmer :: woman : x?
  - x = homemaker

Conclusions

- Language is ambiguous!
  - Ambiguity permeates all levels
- NLP has a (relatively) long history
  - Recent years have seen remarkable progress
- NLP applications are all around us
- The tooling, resources and datasets are now much more freely available
  - Universities & textbooks -> Blogs, YouTube, Medium, Coursera, Meetups, Github, etc…

Thank you

Tony Russell-Rose, PhD FBCS CITP CEng
Director, UXLabs | Founder, 2Dsearch
Senior Lecturer, Goldsmiths

- Web: uxlabs.co.uk, 2dsearch.com
- Email: tgr@uxlabs.co.uk
- Blog: http://isquared.wordpress.com
- LinkedIn: http://uk.linkedin.com/in/tonyrussellrose
- Twitter: @tonygrr