Looking at the Long Tail

2nd Spinoza Workshop

#SpinozaLongTail

It is time to move
from “Big Data”
to “small data”

Language as a system and language use

image(c)http://www.flickr.com/photos/jvdc/

- The world changes rapidly (Brexit)
- The language system changes slowly (Nexit)
- The relation between the two changes constantly

Small data *example*

Imagine you visit a dear friend for a game of chess. During the game you complain about your white queen being captured too early. While chatting, your friend tells you that his Ana is now already 13 years old, and is beginning with high school in September. After the game, you offer to buy him a beer in O’Neil’s, which is just a 2-3 minutes walk.

Small data *example*

Imagine [2] you visit [11] a dear [8] friend [5] for a game [14] of chess [2]. During the game [14] you complain [2] about your white [25] queen [12] being captured [11] too early. While chatting [4], your friend [5] tells [9] you that his Ana [42] is now already 13 years [4] old [9], and is beginning [11] with [high school [1]] in September. After the game [14], you offer [16] to buy [6] him a beer [1] in O'Neil's [5], which is just a 2-3 minutes [8] walk [17].

2*11*8*5*14*2*14*2*25*12*11*4*5*9*42*4*9*11*1*14*16*6*1*5*8*17

= 2,1185E36 possible interpretations and still missing some

If a machine joined the conversation, what would it understand?

Probably, it would think that you talk about: *capturing* “The White Queen” TV-series, the ANA airways based in Japan *being 13 years old*, and the sport equipment store O’Neil *where you apparently can get a beer* (or the recently retired famous basketball player Shaq O’Neil).

An interpretation that may make sense from a Big Data perspective but that does not make any sense as a combination! Where is the coherence?

Understanding of language is about the Long Tail with many, many small data niches

“Today, a 6 year old has seen less data and read less language than most machines, but still these machines make mistakes that the 6 year old will never make.”

How to create semantic tasks/challenges:

- that force systems to use more ‘intelligence’,
- to understand small data and its details,
- without knowing in advance what these details are.

Looking at the Long Tail

Looking at the Long Tail

Practicalities

Schedule overview

09:30-10:05	Welcome & Introduction
10:05-12:25	Invited talks
12:25-13:25	Lunch
13:25-14:05	Keynote
14:05-17:45	Practical session
17:45-18:00	Wrap-up
18:00-	Drinks

Food & Beverages

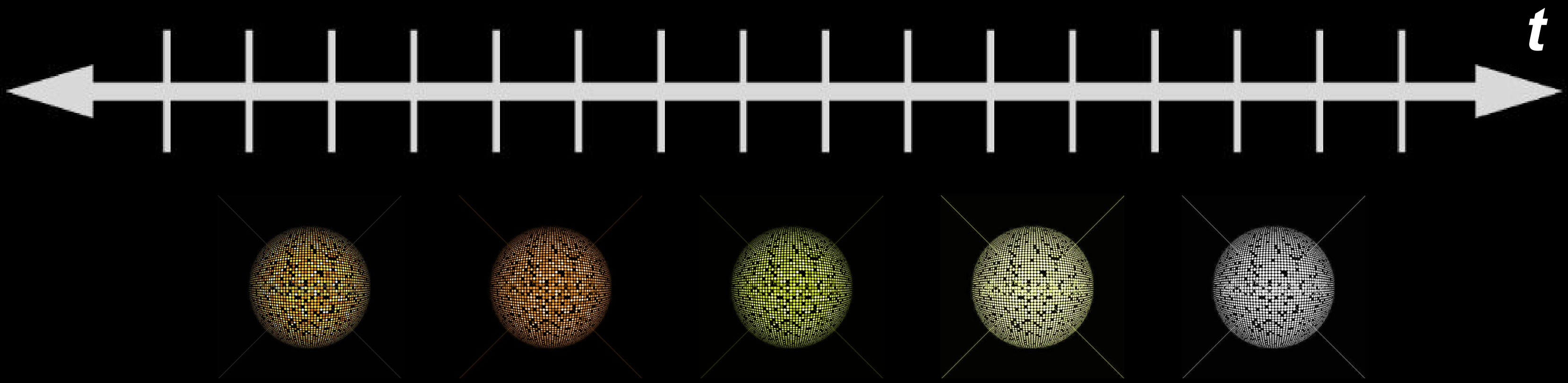
- Coffee
  - Two official coffee breaks (one in the morning + one in the afternoon)
  - Coffee machines available in the hall at any time
- Lunch
  - Will be brought to the forum
- Drinks
  - After the workshop
  - In a nice place nearby the VU (follow us :-))

#SpinozaLongTail

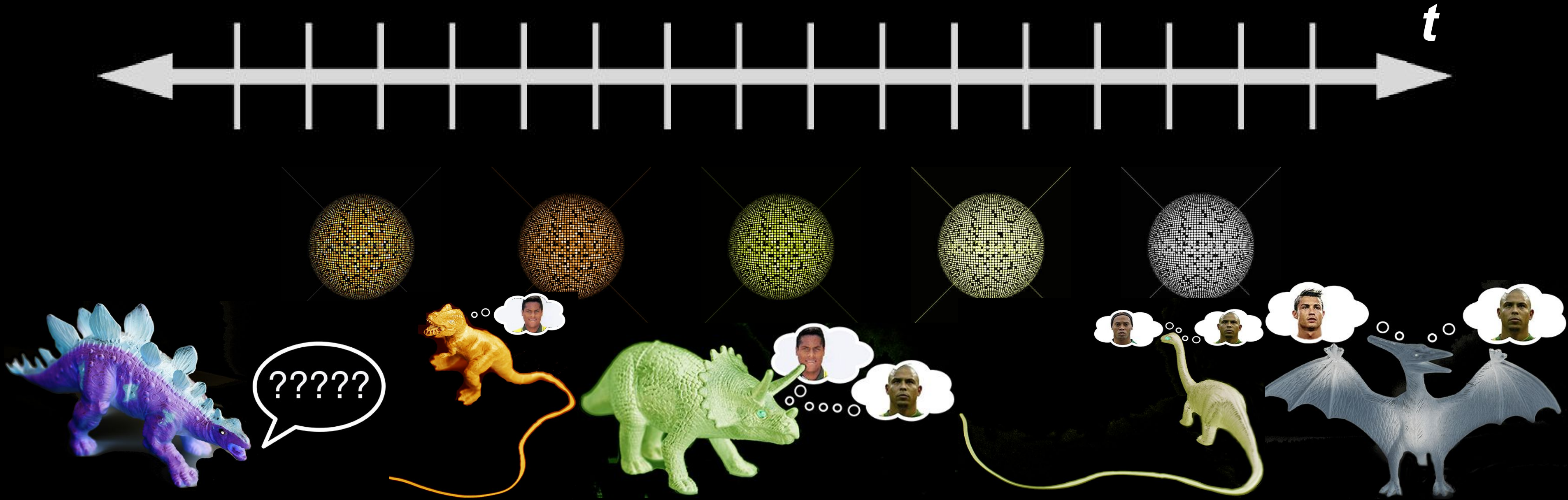
the Long Tail

World(t)

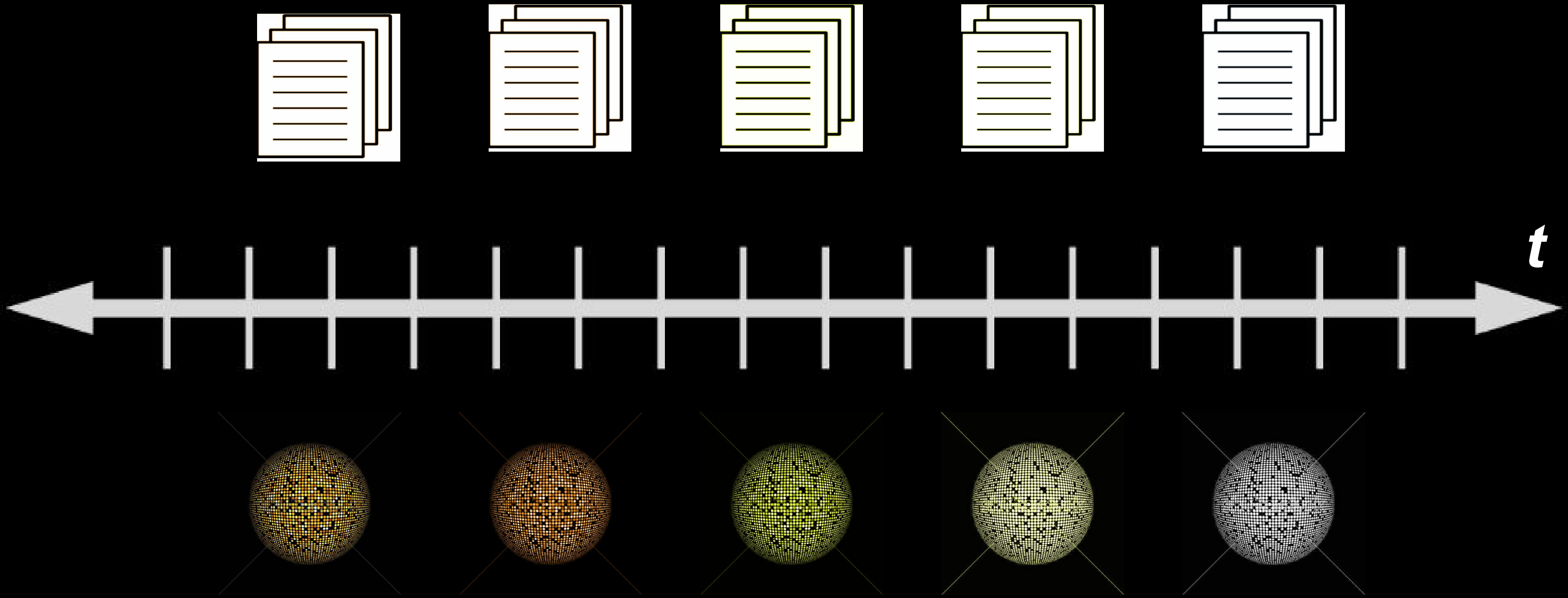
Dinos born in different generations live in different worlds, i.e. what they know about the world depends on the time they live in.



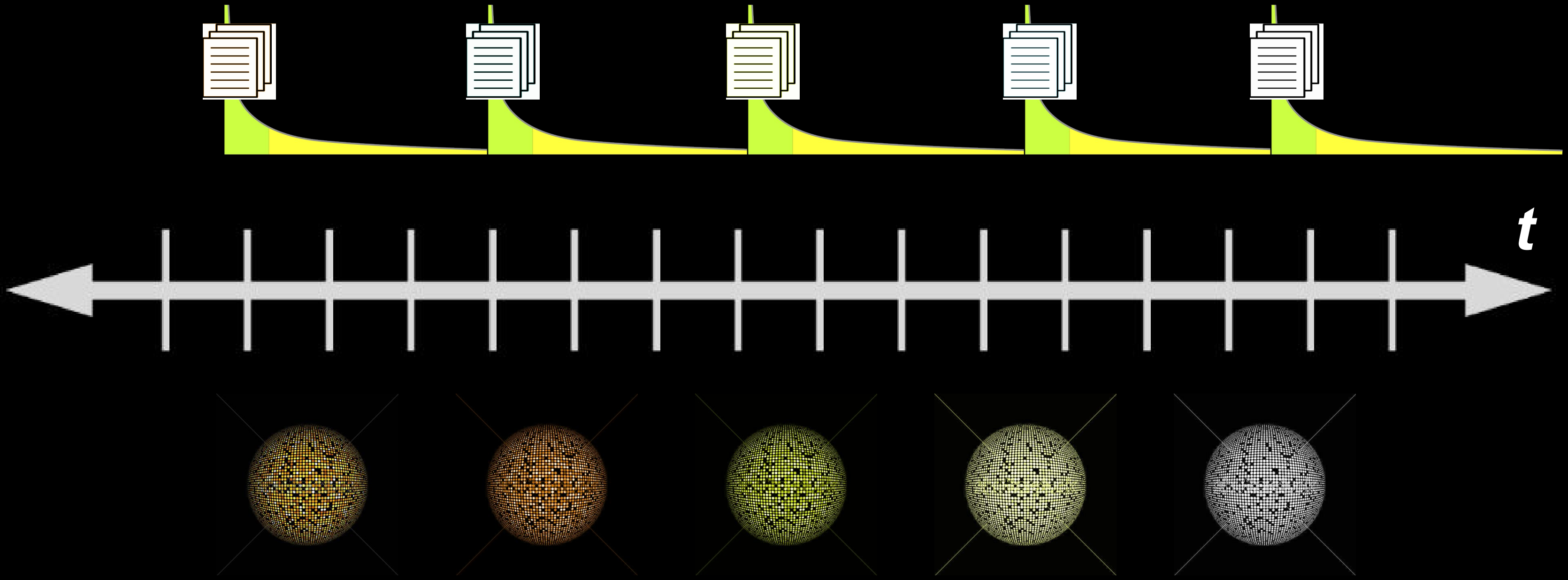
Every generation of dinos has its own Ronaldo



Datasets as Time-bound World Proxies



Big Data can be a bad Proxy



The datasets fail to represent everything in the world,

and over-represent some things.

The Long Tail Phenomenon of Disambiguation (1)

In theory, any lexical expression can refer to any meaning of the world at that time (and vice versa).

People may be able to use the full range of expressions and interpretation of a language. But, they are not aware of it and only use a specific set in relation to a real life situation.

This balances the trade-off between:

- using many different expressions, and
- resolving extreme ambiguity of a small set of expressions
- contextual competition (are you stating the obvious)

The Long Tail Phenomenon of Disambiguation (2)

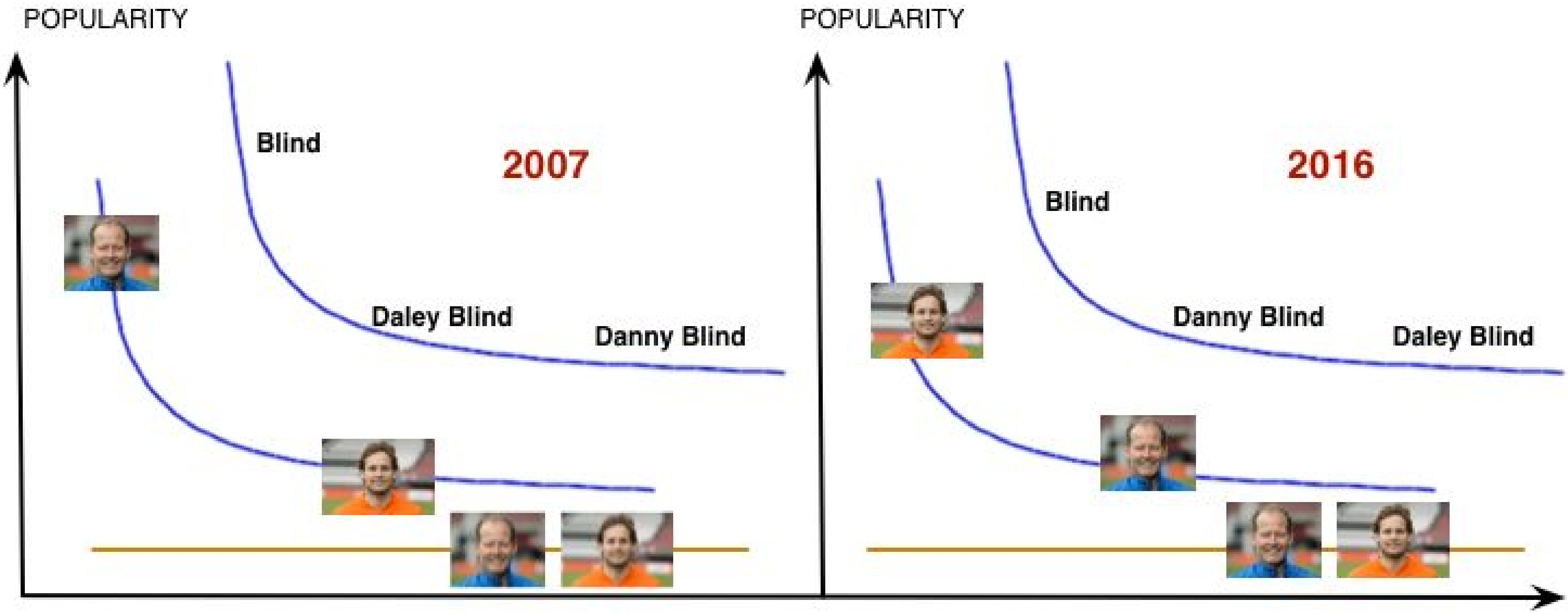
The distribution of the lexical expressions and the denoted meanings in evaluation datasets both follow Zipf’s Law.

But physically there is no Zipfian distribution between the meanings.

Any of these worlds contains a set of unique concepts and instances, each existing physically exactly once.

On instance level, any entity, concept, or event is as prominent as any other.

Tendencies between the long tails of expressions and meanings



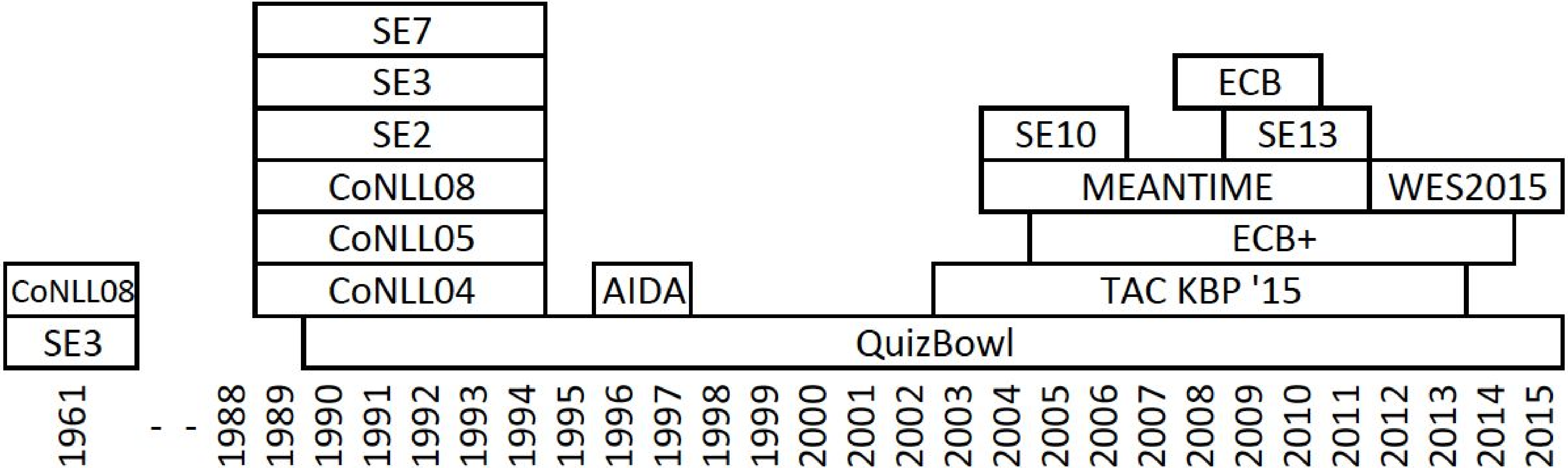
Dataset Analysis

Our analysis shows that the existing disambiguation datasets exhibit:

- Low ambiguity (~1.0)
- Low variance (~1.0)
- High dominance
- Temporal bias towards data from 1961 or 1990s

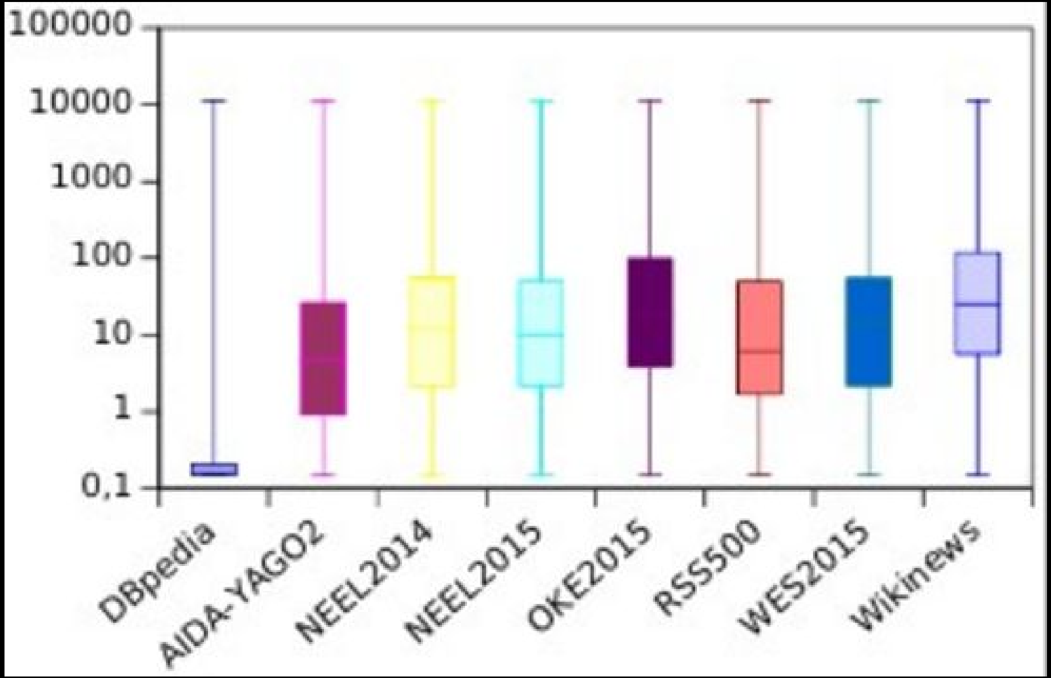
Semantic overfitting: what `world' do we consider when evaluating disambiguation of text? (Under review)

Datasets: How *old* is the data?



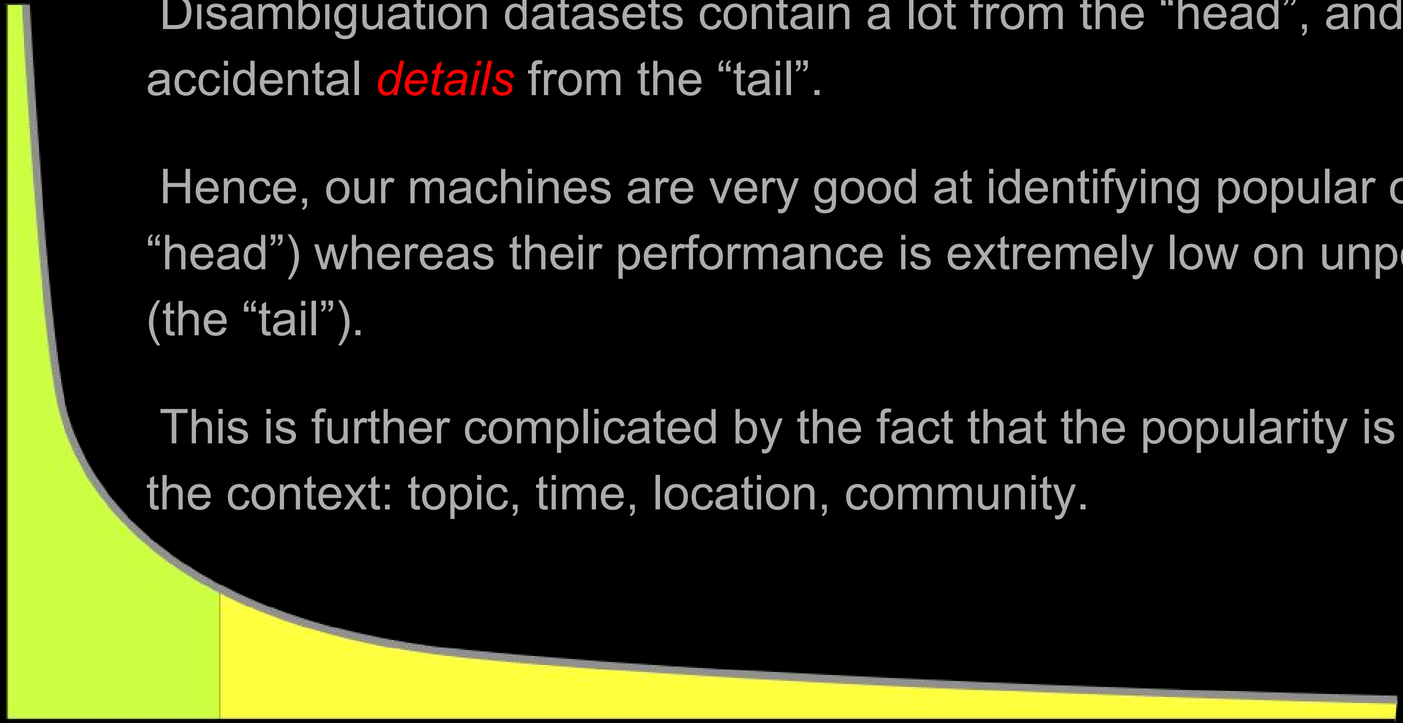
Semantic overfitting: what `world' do we consider when evaluating disambiguation of text? (Under review)

Datasets: How *dominant* is the data?



Marieke van Erp, Pablo Mendes, Heiko Paulheim, Filip Ilievski, Julien Plu, Giuseppe Rizzo and Joerg Waitelonis (2016). Evaluating Entity Linking: An Analysis of Current Benchmark Datasets and a Roadmap for Doing a Better Job. In Proceedings of LREC 2016.

Problem Statement: The Long Tail *DeTail*



Disambiguation datasets contain a lot from the “head”, and only accidental *details* from the “tail”.

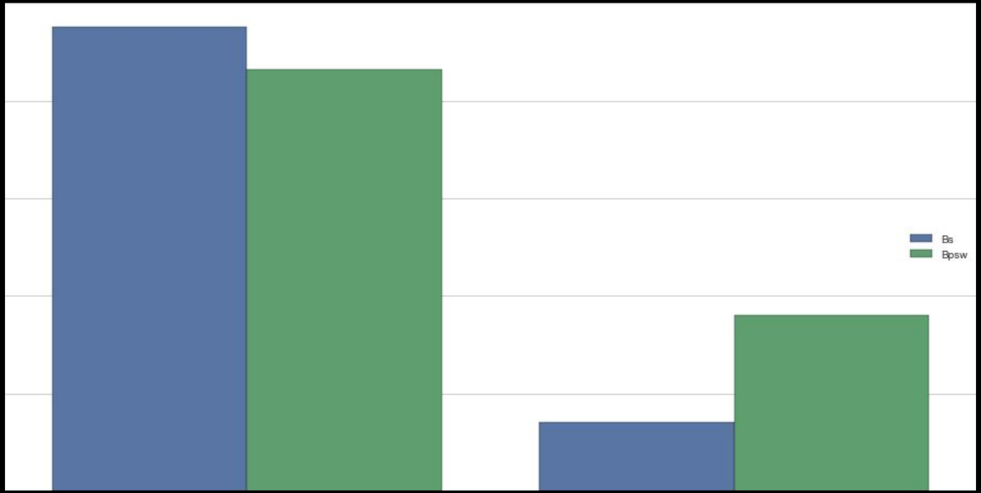
Hence, our machines are very good at identifying popular objects (the “head”) whereas their performance is extremely low on unpopular objects (the “tail”).

This is further complicated by the fact that the popularity is determined by the context: topic, time, location, community.

Systems: Our analysis

For WSD, we tried to improve on the disambiguation of the less frequent senses

Outcome: better performance on the tail causes worse performance on the head!



Marten Postma, Ruben Izquierdo, Eneko Agirre, German Rigau, Piek Vossen (2016). Addressing the MFS Bias in WSD systems. In Proceedings of LREC 2016.

Goal(s) of the workshop

We aim to propose this task as a “Long Tail Shared Disambiguation Task” to the next call for SemEval-2018 tasks, which is expected late 2016/early 2017.

In addition, we plan to propose a workshop for ACL 2017.

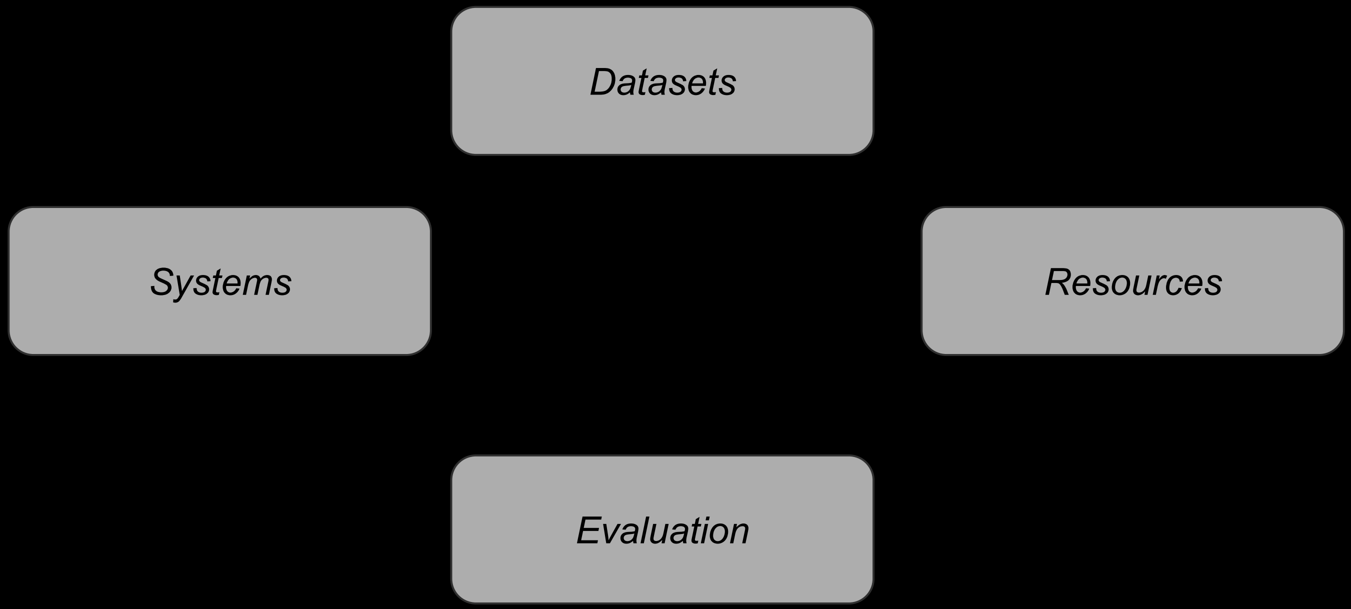
Goal(s) of the workshop

We want systems to resolve extreme ambiguity within specific context:

- Discriminate one context from another
- Use more semantics: coherence, logic, inferencing, comprehensive
- Combine semantic layers and subtasks
- Use the complete document and more (external knowledge)
- Answer more complex questions, e.g. quantification and identity
- Be smarter than a 6 year old
- Can explain why something is an answer

We do not want: ***yet another domain task***

Looking at the Long Tail: What can be done?



Datasets

Systems

Resources

Evaluation

#1 Datasets

What kind of datasets are needed for the long tail disambiguation task?

- Properties
- Multi-task
- Are current ones sufficient?
- Optimal acquisition methods

#1 Datasets: *Property-driven data*

MEANTIME -> Multi-task corpus

Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begoña Altuna, Marieke van Erp, Anneleen Schoen and Chantal van Son (2016). MEANTIME, the NewsReader Multilingual Event and Time Corpus. In Proceedings of LREC 2016.

Dutch SemCor -> Balanced WSD corpus

Piek Vossen, Rubén Izquierdo, and Attila Görög (2013). DutchSemCor: in quest of the ideal sense-tagged corpus. RANLP.

ECB+ -> Increased ambiguity for event coreference

Agata Cybulska and Piek Vossen (2014). Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. In LREC 2014.

#2 Resources

What kind of knowledge is needed for the long tail disambiguation task?

- Long tail knowledge bases
- Contextual knowledge bases
- Locating appropriate knowledge

#3 Evaluation

How should we evaluate the long tail disambiguation task(s)?

- Optimal evaluation metric(s)
- Generalizability over disambiguation tasks
- Incentivizing context- and long tail-aware systems

#4 Systems

What are the requirements for a system to perform well on the long tail disambiguation task(s)?

- Existing systems
- Multi-task approach
- Long tail performance
- Sustainable systems

Thanks!