

From Machine Learning to Machine Training

Johann-Mattis List

Forschungsgruppe "Computer-Assisted Language Comparison"
Department of Linguistic and Cultural Evolution
Max Planck Institute for Evolutionary Anthropology
Leipzig, Germany

2022-09-10



MAX-PLANCK-GESELLSCHAFT



Machines on the Rise...



Machines on the Rise

Artificial Intelligence has begun to dominate our lives, proving its growing success by beating humans in games, providing decent translations, solving hard scientific problems, and exercising control in our daily lives.

Games

Go-Game

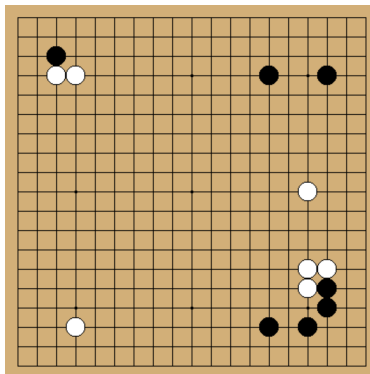


Image from: Wikipedia, "Go (game)"

Games

Go-Game

The screenshot shows the top portion of a Nature journal article page. At the top is the 'nature' logo. Below it are three navigation links: 'Explore content', 'About the journal', and 'Publish with us', each followed by a downward arrow. The breadcrumb trail reads 'nature > articles > article'. The publication date is 'Published: 27 January 2016'. The title is 'Mastering the game of Go with deep neural networks and tree search'. The authors are listed as David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. At the bottom, it says 'Nature 529, 484–489 (2016)' and provides a link to 'Cite this article'. Below that, it shows '421k Accesses', '6364 Citations', '3049 Altmetric', and a link to 'Metrics'.

nature

Explore content ▾ About the journal ▾ Publish with us ▾

[nature](#) > [articles](#) > article

Published: 27 January 2016

Mastering the game of Go with deep neural networks and tree search

[David Silver](#) ✉, [Aja Huang](#), [Chris J. Maddison](#), [Arthur Guez](#), [Laurent Sifre](#), [George van den Driessche](#), [Julian Schrittwieser](#), [Ioannis Antonoglou](#), [Veda Panneershelvam](#), [Marc Lanctot](#), [Sander Dieleman](#), [Dominik Grewe](#), [John Nham](#), [Nal Kalchbrenner](#), [Ilya Sutskever](#), [Timothy Lillicrap](#), [Madeleine Leach](#), [Koray Kavukcuoglu](#), [Thore Graepel](#) & [Demis Hassabis](#) ✉

[Nature](#) 529, 484–489 (2016) | [Cite this article](#)

421k Accesses | 6364 Citations | 3049 Altmetric | [Metrics](#)

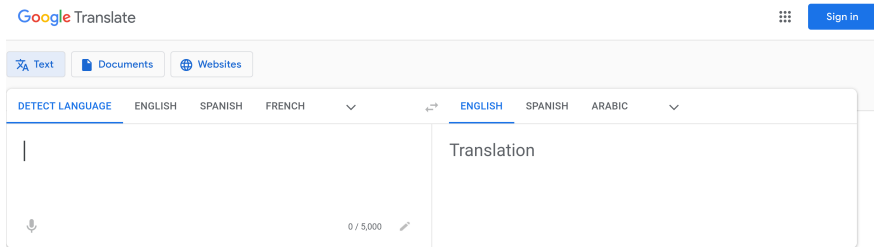
Silver et al. (2021, Nature)

Games

Neural machine learning approaches have advanced so fast of late that computers have started to beat humans in nearly any kind of strategic game by now.

Natural Language

Machine Translation



<https://translate.google.com>

Natural Language

Machine Translation

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova


Google AI Language

{jacobdevlin, mingweichang, kentonl, kristout}@google.com

Devlin et al. (2019)

Natural Language

Language Generation



Contents lists available at [ScienceDirect](#)

Computers in Human Behavior


journal homepage: <http://www.elsevier.com/locate/comphumbeh>

Full length article

Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry

Nils Köbis^{a,b,*}, Luca D. Mossink^a

^a Department of Economics & Center for Experimental Economics and Political Decision Making (CREED), University of Amsterdam, the Netherlands
^b Max Planck Institute for Human Development, Center for Humans & Machines, Germany



Köbis and Mossink (2021)

Natural Language

While it was common to make fun of automated translation services a few years ago, we start to forget that automated translation services are integrated in numerous parts of our lives, and we even start to make active use of them without knowing so, e.g., when booking accommodation or buying products online.

Scientific Problems

Protein Folding Prediction

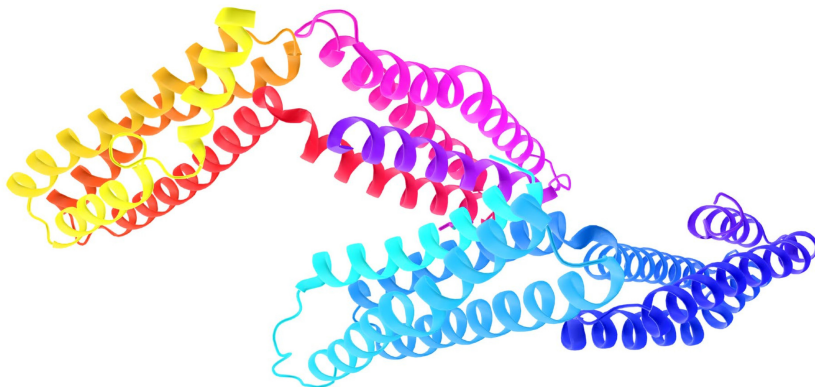


Image from: Callaway (2022, Spektrum — Die Woche)

Scientific Problems

Protein Folding Prediction

Article | [Open Access](#) | [Published: 15 July 2021](#)

Highly accurate protein structure prediction with AlphaFold

[John Jumper](#) , [Richard Evans](#), ... [Demis Hassabis](#)  [+ Show authors](#)

[Nature](#) **596**, 583–589 (2021) | [Cite this article](#)

759k Accesses | **2919** Citations | **3147** Altmetric | [Metrics](#)

Jumper et al. (2021, Nature)

Scientific Problems

Scientific problems that were long thought to be unsolvable neither by humans nor by machines, have recently been tackled and start to change the dynamics of scientific research.

Control in Daily Life

Facial Recognition



Golembiewski (2022, New York Times)

Control in Daily Life

Facial Recognition

Cell Reports

Open access

REPORT | [VOLUME 40, ISSUE 8, 111257, AUGUST 23, 2022](#)

Look-alike humans identified by facial recognition algorithms show genetic similarities

[Ricky S. Joshi](#) ¹⁰ • [Maria Rigau](#) ¹⁰ • [Carlos A. García-Prieto](#) ¹⁰ • ... [Xavier Binefa](#) • [Alfonso Valencia](#) • [Manel Esteller](#)  ¹¹  • [Show all authors](#) • [Show footnotes](#)

Open Access • DOI: <https://doi.org/10.1016/j.celrep.2022.111257> •



Check for updates

Joshi et al. (2022, Cell Reports)

Control in Daily Life

Facial recognition, voice recognition, and many more tasks which were originally entrusted only to humans are now being routinely done by machines.

Mind the Machines!



The Quantitative Turn



The Quantitative Turn

letters to nature

**Language-tree divergence times
support the Anatolian theory
of Indo-European origin**

Russell D. Gray & Quentin D. Atkinson

*Department of Psychology, University of Auckland, Private Bag 92019,
Auckland 1020, New Zealand*

2002

2004

2006

2008

2010

2012

2014

The Quantitative Turn

letters to nature

Language-tree divergence times support the Anatolian theory of Indo-European origin

Russell D. Gray & Quentin D. Atkinson

Department of Psychology, University of Auckland, Private Bag 92019, Auckland 1020, New Zealand



2002

2004

2006

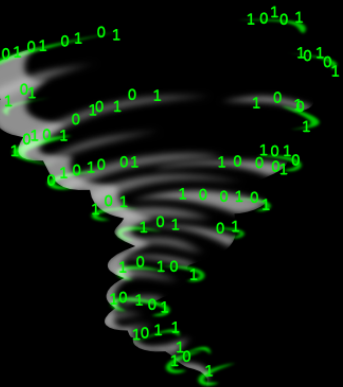
2008

2010

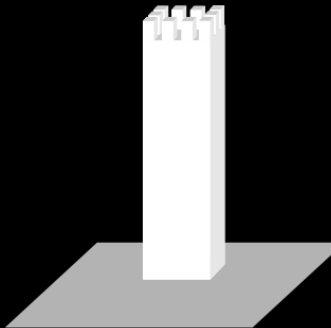
2012

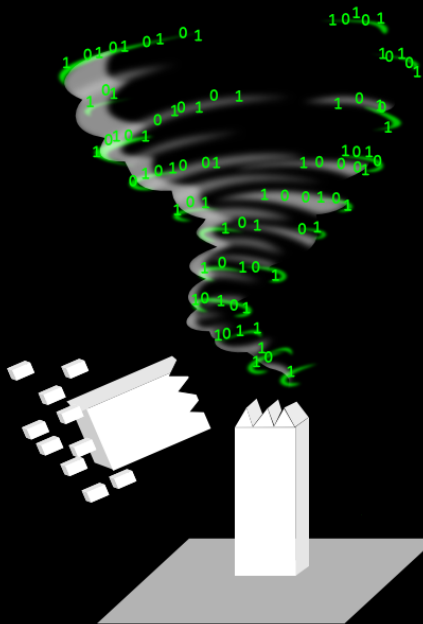
2014

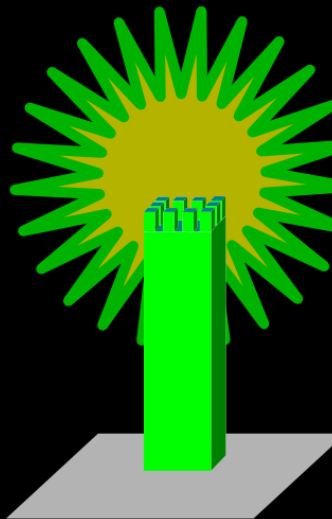
9 / 34



? ?
! ?







The Quantitative Turn

Even if not yet acknowledged by all linguists who actively practice historical language comparison, the quantitative turn has greatly changed the field over the past years. Nowadays, most calls for positions in the field of historical linguistics and linguistic typology require at least some quantitative skills, and it is getting more and more common to defend claims on the phylogeny of languages with the help of quantitative analyses rather than with qualitative arguments based on shared innovations.

The Stochastic Turn

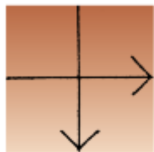
[...] it was at the 1985 workshop [...] that Fred Jelinek uttered the now immortal phrase *“Every time we fire a phonetician/linguist, the performance of our system goes up”*. (Moore 2005: 1)



The Stochastic Turn

Altaic as Fact or Fantasy?

DIACHRONICA
VOLUME 36 NUMBER 1 2022



JOHN BENJAMINS PUBLISHING COMPANY

Significance testing of the Altaic family

Author(s): Andrea Ceolin¹

⊕ View Affiliations

Source: [Diachronica](#), Volume 36, Issue 3, Sept 2019, p. 299 - 336

DOI: <https://doi.org/10.1075/dia.17007.ceo>

Version of Record published : 17 Sept 2019

[« Previous Article](#) | [Table of Contents](#) | [Next Article »](#)

Ceolin (2019, Diachronica)

The Stochastic Turn

Altaic as Fact or Fantasy?

Evolutionary Human Sciences (2021), 3, e32, page 1 of 10
doi:10.1017/ehs.2021.28



RESEARCH ARTICLE

Permutation test applied to lexical reconstructions partially supports the Altaic linguistic macrofamily

Alexei S. Kassian^{1*} , George Starostin^{2,3*} , Ilya M. Egorov¹ , Ekaterina S. Logunova⁴
and Anna V. Dybo⁵ 

¹School of Advanced Studies in the Humanities, The Russian Presidential Academy of National Economy and Public Administration, Moscow, Russia, ²Institute for Oriental and Classical Studies, National Research University Higher School of Economics, Moscow, Russia, ³Santa Fe Institute, New Mexico, USA, ⁴Institute of Linguistics, Russian State University for the Humanities, Moscow, Russia and ⁵Institute of Linguistics, Russian Academy of Sciences, Moscow, Russia

*Corresponding authors. E-mails: a.kassian@gmail.com, gstarst1@gmail.com.

Kassian et al. (2021, *Evolutionary Human Sciences*)

The Stochastic Turn

Altaic as Fact or Fantasy?

Contradictory Results

 [Follow this preprint](#)

Triangulation fails when neither linguistic, genetic, nor archaeological data support the Transeurasian narrative

 Zheng Tian,  Yuxin Tao,  Kongyang Zhu,  Guillaume Jacques,  Robin J. Ryder,  José Andrés Alonso de la Fuente,  Anton Antonov,  Ziyang Xia,  Yuxuan Zhang,  Xiaoyan Ji,  Xiaoying Ren,  Guanglin He,  Jianxin Guo,  Rui Wang,  Xiaomin Yang,  Jing Zhao,  Dan Xu,  Russell D. Gray,  Menghan Zhang,  Shaoqing Wen,  Chuan-Chao Wang,  Thomas Pellard

doi: <https://doi.org/10.1101/2022.06.09.495471>

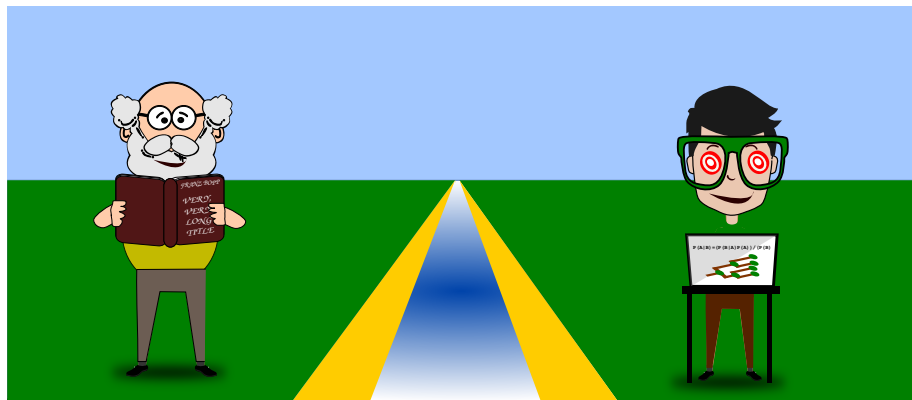
This article is a preprint and has not been certified by peer review [what does this mean?].

Tian et al. (2022, bioRxiv preprint)

The Stochastic Turn

Although results are often still contradictory, and scholars tend to be far away from reaching a common agreement, stochastic arguments play an increasingly important role in studies on historical language comparison. This role enjoys a doubtful authority that feeds upon the idea that numbers cannot lie, which we find in other parts of science.

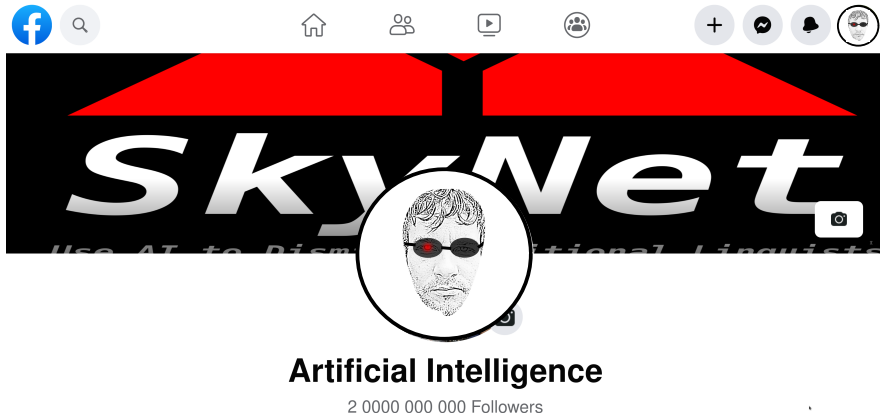
Classical Linguistics Under Threat?



Linguistics Under Threat?

There are a lot people who fear for the future of classical historical linguistics in the light of the new methods of computational language comparison. They argue that some tasks could never be adequately carried out by machines, they also support a romantic view on science, in which scholars are lonely persons spending most of their time in libraries, and they find their views supported in the obvious shortcomings of the new methods which are often even accompanied by a rather arrogant attitude from the practitioners of computational historical linguistics.

Meet the Machines



Background

Failures of AI

DOI: 10.1002/widm.1278

WILEY



WIREs

DATA MINING AND KNOWLEDGE DISCOVERY

FOCUS ARTICLE

Facial feature discovery for ethnicity recognition

Cunrui Wang^{1,2} | Qingling Zhang² | Wanquan Liu³ | Yu Liu¹ | Lixin Miao¹

Wang et al. (2018, Data Mining and Knowledge Discovery)

Background

Failures of AI

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 36, NO. 9, SEPTEMBER 2014

The Hidden Sides of Names—Face Modeling with First Name Attributes

Huizhong Chen, *Student Member, IEEE*, Andrew C. Gallagher, *Senior Member, IEEE*, and
Bernd Girod, *Fellow, IEEE*

Chen et al. (2014, IEEE)

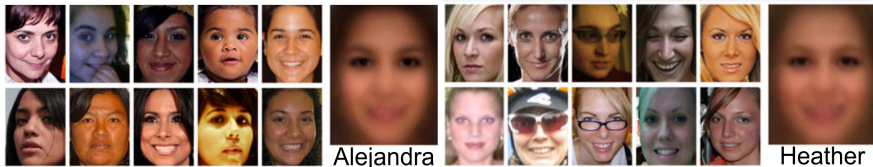
Background

Failures of AI

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 36, NO. 9, SEPTEMBER 2014

The Hidden Sides of Names—Face Modeling with First Name Attributes


Huizhong Chen, *Student Member, IEEE*, Andrew C. Gallagher, *Senior Member, IEEE*, and Bernd Girod, *Fellow, IEEE*








Chen et al. (2014, IEEE)




Background

Failures of AI

RESEARCH-ARTICLE • 

Are we really making much progress? A worrying analysis of recent neural recommendation approaches

Authors:  Maurizio Ferrari Dacrema,  Paolo Cremonesi,  Dietmar Jannach [Authors](#)

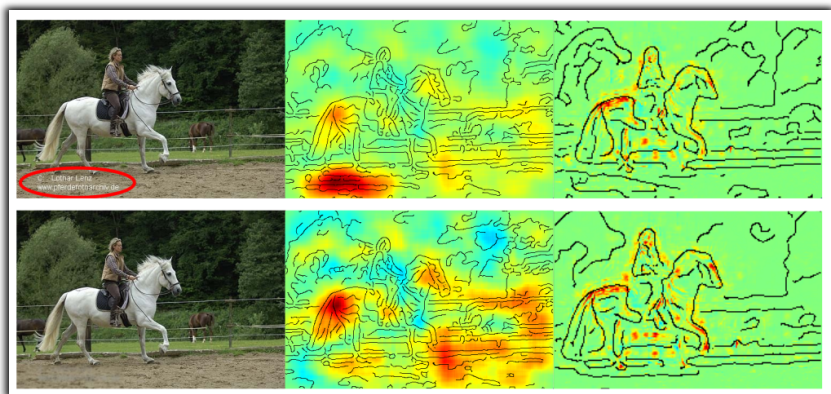
[Info & Claims](#)

RecSys '19: Proceedings of the 13th ACM Conference on Recommender Systems • September 2019 • Pages 101–109 • <https://doi.org/10.1145/3298689.3347058>

Dacrema et al. (2019, ACM, RecSys)

Background

Failures of AI



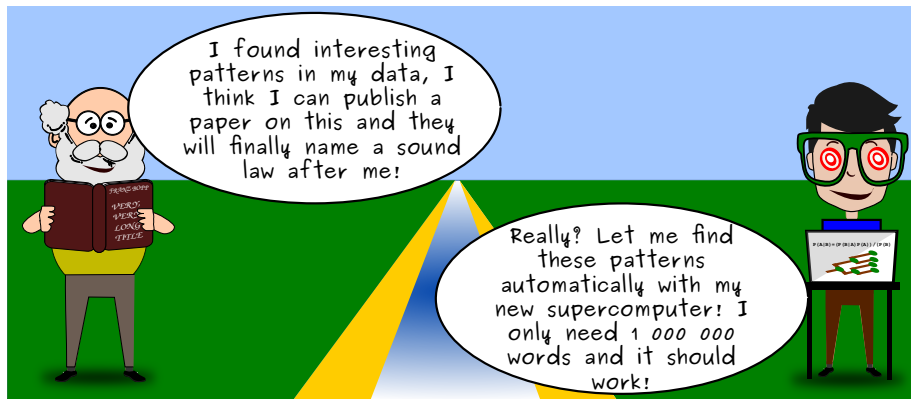
Lapuschkin et al. (2016)

Background

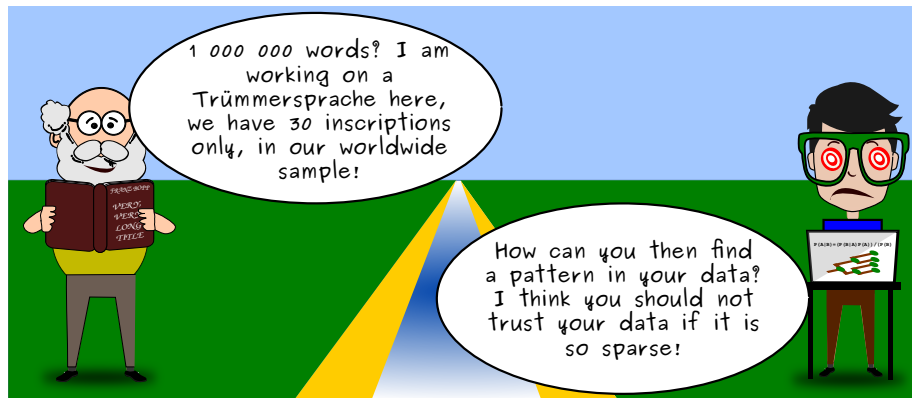
In defense of AI skeptics, we find numerous cases where machine learning methods were (1) applied in an extremely naive or even harmful way, (2) not as efficient as their creators wanted their audience to believe, or (3) configured in a wrong way due to the use of problematic training and development data.

→ Blind trust in AI has never been a good idea!

Data



Data

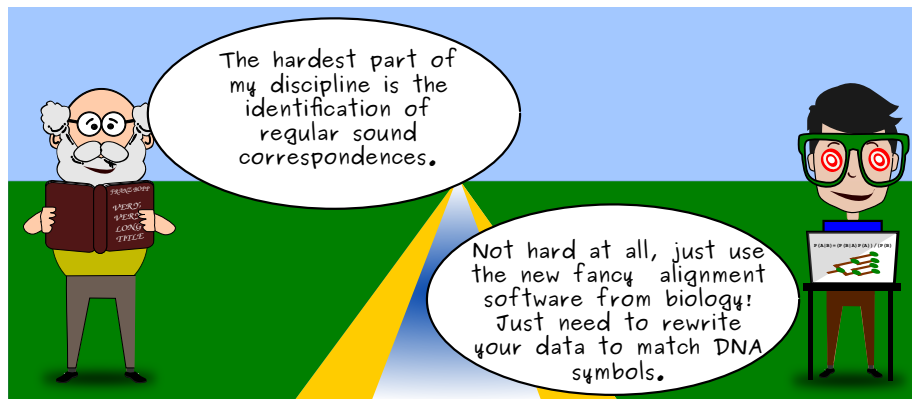


Data

For many people practicing machine learning, the only aspect of data that counts is their size. That there are techniques that work well with sparse data and that data quality may often be more important than data quantity is still mostly ignored when it comes to applying machine learning methods to problems in comparative linguistics.

→ **We need a qualitative turn with respect to our data!**

Modeling

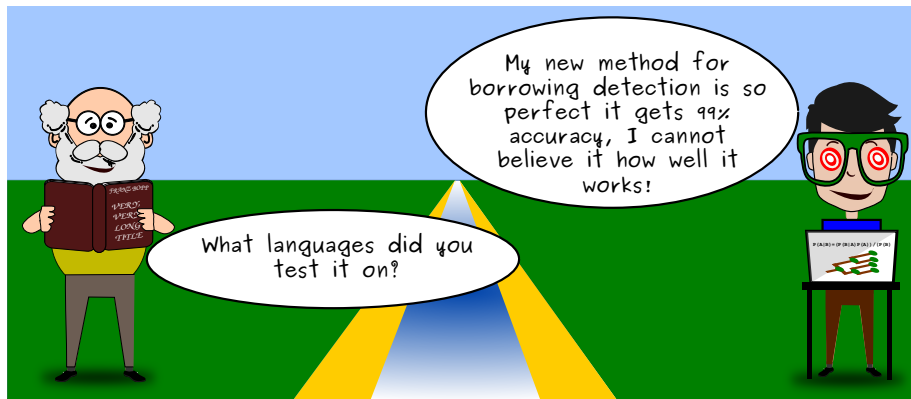


Modeling

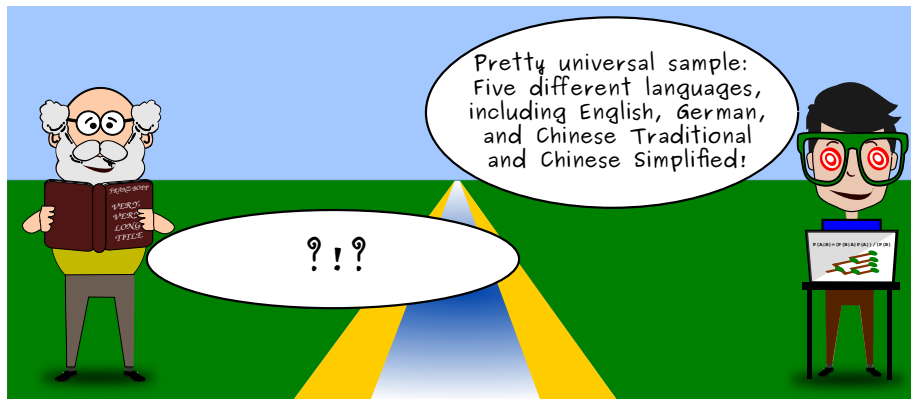
We often find the attitude among machine learning practitioners that the way in which data are represented or the way in which data are modeled is much less important, given the incredible power of machine learning approaches. That machine learning methods might profit from a careful discussion about data representation and modeling is often ignored.

→ **We need to discuss how to model our data!**

Testing



Testing



Testing

Many methods which have been popularly debated and propagated have never been rigorously tested on a diverse sample of datasets. Even if tests are shown along with a method, scholars often think it is enough to evaluate a method on a few common Indo-European languages.

→ **We need to work on realistic test scenarios!**

Train the Machines



Computer-Assisted Language Comparison



consistency

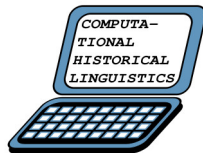


efficiency

accuracy



flexibility



Computer-Assisted Language Comparison



consistency



efficiency



accuracy



flexibility



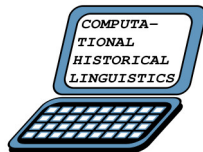
Computer-Assisted Language Comparison



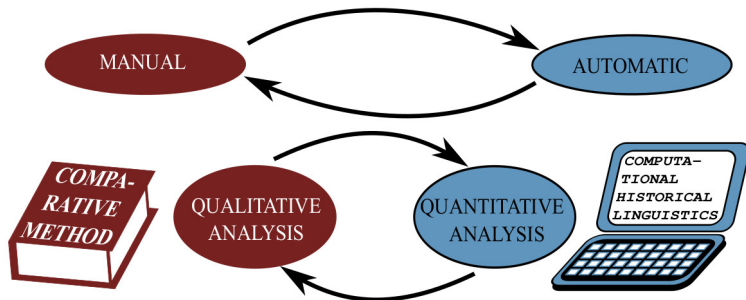
accuracy
flexibility



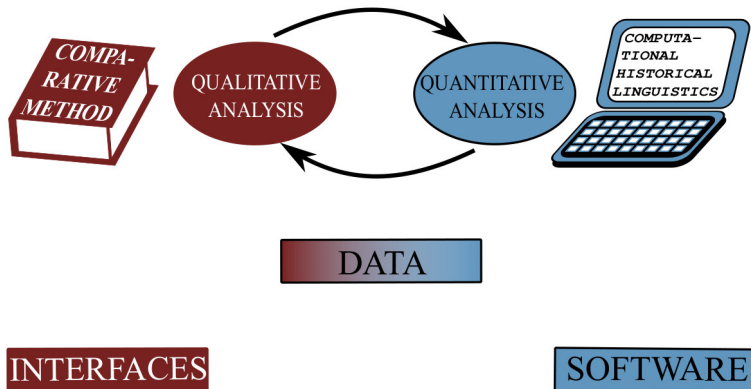
consistency
efficiency



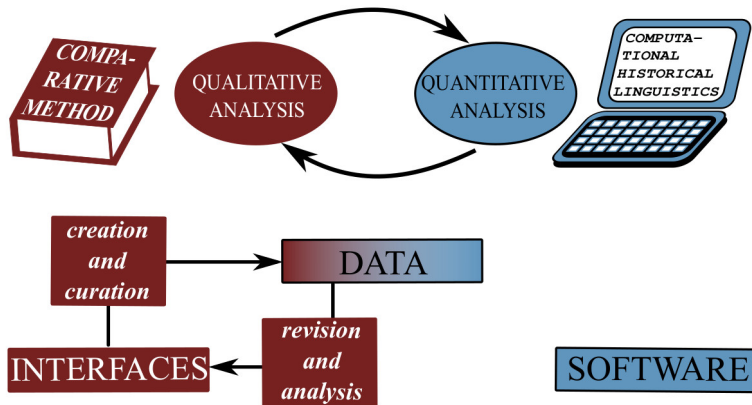
Computer-Assisted Language Comparison



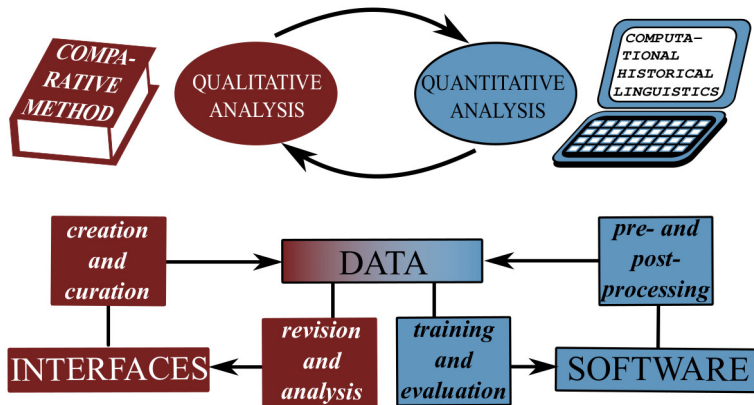
Computer-Assisted Language Comparison



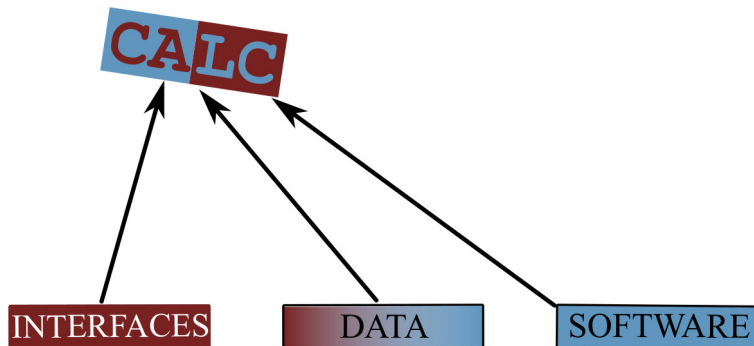
Computer-Assisted Language Comparison



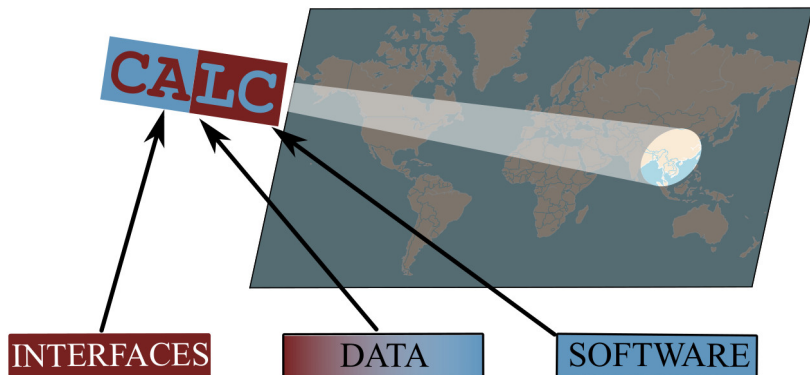
Computer-Assisted Language Comparison



Computer-Assisted Language Comparison

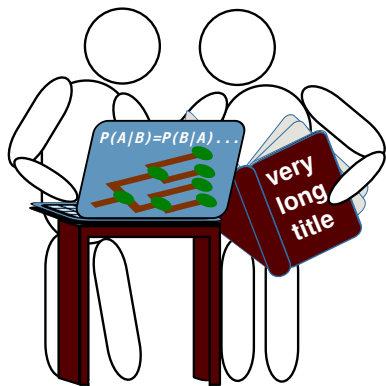


Computer-Assisted Language Comparison



Computer-Assisted Language Comparison

- Funding: ERC Starting Grant (2017-2022)
- Host Institution: MPI-EVA (Leipzig)
- Team: 2 Post-Docs, 4 Docs (2 financed by project, 2 financed externally), PI
- Goal: establish a framework for CALC and show how to apply it to the Sino-Tibetan language family.



<https://digling.org/calc/>

Computer-Assisted Language Comparison

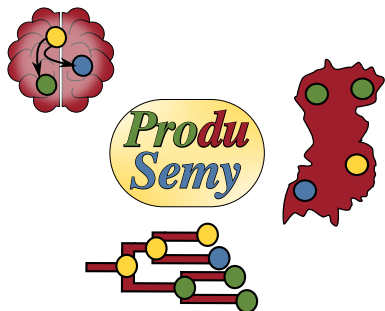
- Funding: MPG Extension Grant (2022-2024)
- Host Institution: MPI-EVA (Leipzig)
- Team: 1 Post-Docs, 1 Doc, PI
- Goal: extend the CALC framework to include typology and cognition.



<https://digling.org/calc/>

Computer-Assisted Language Comparison

- Funding: ERC Consolidator Grant (2023-2027)
- Host Institution: MPI-EVA (Leipzig)
- Team: 3 Post-Docs, 3 Docs, PI
- Goal: Use the CALC framework to investigate lexical compositionality from different perspectives.



<https://digling.org/calc/>

Problem Solving

- A identify the **core class** of your problem (modeling, inference, analysis)
- B look at **existing qualitative solutions**
- C formalize the problem in a way that allows you to **test it**
- D qualitative solutions are often holistic, do not hesitate to **specify sub-problems**
- E search for inspiration in **neighboring disciplines** by looking for similar processes
- F accept a **qualitative or semi-automatic solution** for inference, but make sure the results are also machine-readable
- G insist on **transparent output** to allow experts to review the results

Problem Solving

Our general strategy for problem solving is pragmatic. We aim to get the best out of machine learning techniques, but we do not trust the methods blindly, so we know we need to start from the problems we want to solve, and pay attention to the representation of our data and to the desired results we want to achieve.

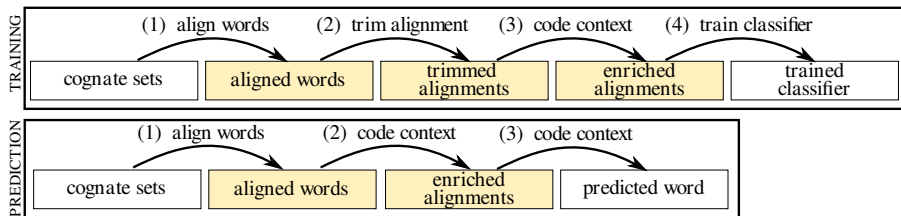
Problem Solving

Where possible, we aim for explicit algorithmic solutions, since these have the advantage of being faster in their application and also closer to the traditional scholarship in the field. Machine learning is welcome, but only where we are confident we have the capacities to *train* the methods properly with good test data and clear data models and data representations.

Examples: Supervised Phonological Reconstruction

Task	Based on existing reconstructions for cognate sets in a given language family, train a model that can create reconstructions from cognate sets that have not been encountered before by the model.
Solution	Our method for correspondence pattern inference (List 2019) can be extended in a dedicated workflow by which we use machine learning approaches to “predict” proto-forms for aligned cognate sets.

Examples: Supervised Phonological Reconstruction



List et al. (2022, ACL Workshop)

Examples: Supervised Phonological Reconstruction

Trimming

	1	2	3	4	5	6	7
Latin	k	-	e:	n	a:	r	ε
	↑	↑	↑	↑	↑	↑	↑
Romanian	tʃ	-	i	n	a	-	-
Spanish	θ	-	e	n	a	r	-
Portuguese	s	j	-	-	a	ɹ	-

List et al. (2022, ACL Workshop)

Examples: Supervised Phonological Reconstruction

Trimming

	1	2	3	4	5	6
Latin	k	-	e:	n	a:	r.ɛ
	↑	↑	↑	↑	↑	↑
Romanian	tʃ	-	i	n	a	-
Spanish	θ	-	e	n	a	r
Portuguese	s	j	-	-	a	ɾ

List et al. (2022, ACL Workshop)

Examples: Supervised Phonological Reconstruction

Prediction

	<i>Ro</i>	<i>Sp</i>	<i>Pt</i>	<i>P</i>	<i>S</i>	<i>Ini</i>		<i>Lt</i>
1	tʃ	θ	s	1	C	^	→	k
2	-	-	j	2	C	-	→	-
3	i	e	-	3	v	-	→	e:
4	n	n	-	4	C	-	→	n
5	a	a	a	5	v	-	→	a:
6	-	r	ɹ	6	C	\$	→	r.ɛ

List et al. (2022, ACL Workshop)

Examples: Supervised Phonological Reconstruction

A New Framework for Fast Automated Phonological Reconstruction Using Trimmed Alignments and Sound Correspondence Patterns

Johann-Mattis List

DLCE
MPI-EVA
Leipzig

mattis_list@eva.mpg.de

Robert Forkel

DLCE
MPI-EVA
Leipzig

robert_forkel@eva.mpg.de

Nathan W. Hill

Trinity Centre for Asian Studies
University of Dublin
Dublin

nathan.hill@tcd.ie

List et al. (2022, ACL Workshop)

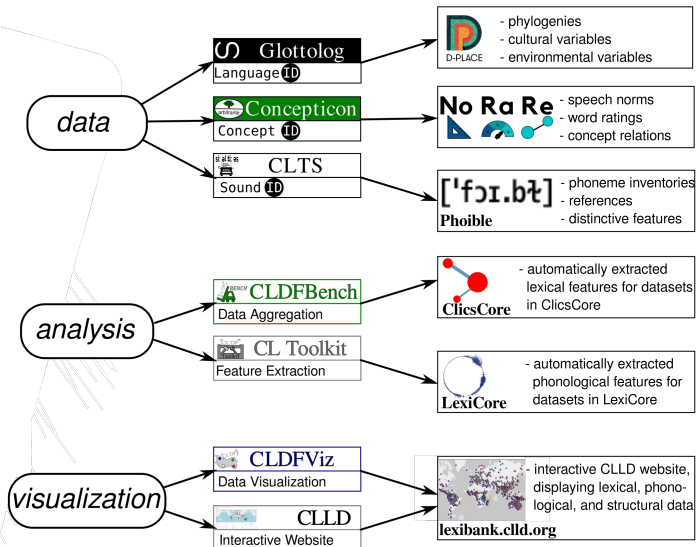
Examples: Standardizing Lexical Data

Scholars have produced large amounts of cross-linguistic datasets in the form of dictionaries and wordlists. These data have so far not been sufficiently standardized. If standardized, we could use the data for numerous tasks in qualitative and quantitative comparative linguistics.

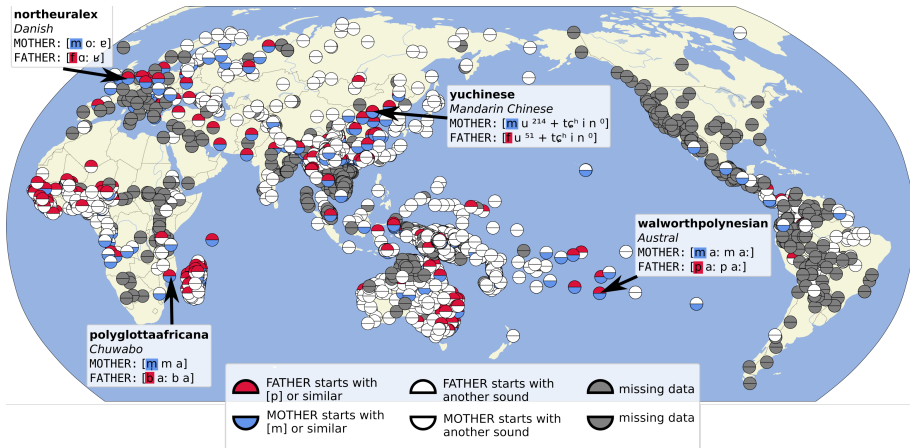
Examples: Standardizing Lexical Data

With the Lexibank repository (List et al. 2022), we have created a first version of a large repository that provides standardized versions of various lexical datasets published over the last two centuries. We standardize the data by converting individual datasets to Cross-Linguistic Data Formats (CLDF, Forkel et al. 2018). We link languages to Glottolog (Hammarström et al. 2022), concepts to Concepticon (List et al. 2021), and speech sounds to the Cross-Linguistic Transcription Systems (CLTS) reference catalogue (Anderson et al. 2018). In this way, we can *aggregate* lexical data in standardized transcriptions for almost 2000 language varieties and *analyze* the data automatically and manually.

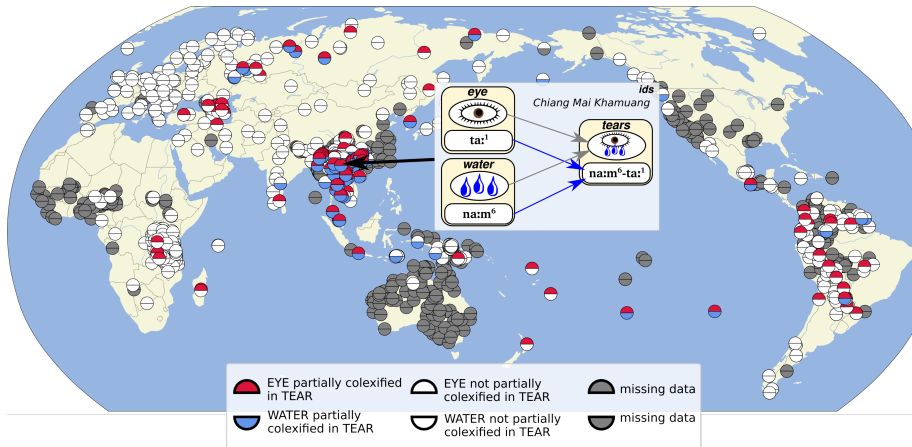
Examples: Standardizing Lexical Data



Examples: Standardizing Lexical Data



Examples: Standardizing Lexical Data



Examples: Standardizing Lexical Data

Lexibank datasets have already been successfully used by colleagues (Winter et al. 2022) and are currently being investigated by several teams (independently from the team of creators). More and more datasets are regularly added to the large Lexibank repository. On the long run, we expect the data to revolutionize our knowledge about lexical typology, language contact, and cognitive aspects of lexical coding.

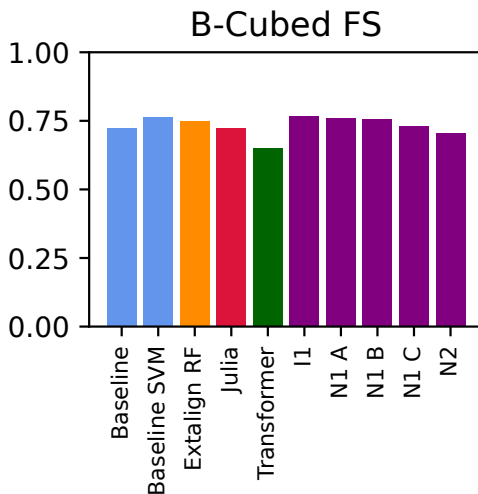
Examples: Shared Task on Word Prediction

Similar to the “prediction” of a proto-language based on cognates from descendant languages, we can predict how a word in a given language should sound from cognates in neighboring languages (Bodt and List 2022). We extended this idea by organizing a *shared task* in which we invited specialists in computational linguistics to provide new solutions to a new test set based on a large number of language families taken from the Lexibank collection (List et al. 2022).

Examples: Shared Task on Word Prediction

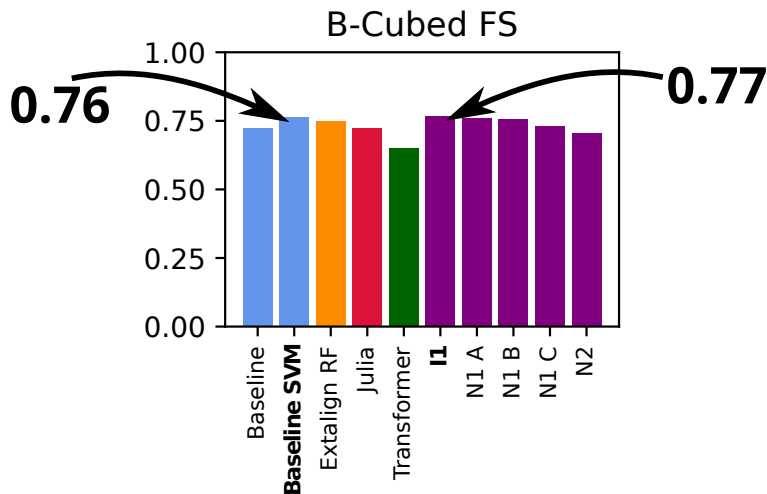
Teams	Four teams, proposed 7 systems, two teams worked with neural network approaches, two teams provided alternative workflows based on phylogenetic methods or alignments.
Baseline	We used our method for automated phonological reconstruction using trimmed and enriched alignments as a baseline.
Training Data	The data shared for the development and training of the methods consisted of 10 datasets covering 7 language families, all taken from Lexibank.
Surprise Data	The surprise data shared only a short time before submission of results was due consisted of 10 datasets covering 6 language families.

Examples: Shared Task on Word Prediction



List et al. (2022, SIGTYP Workshop)

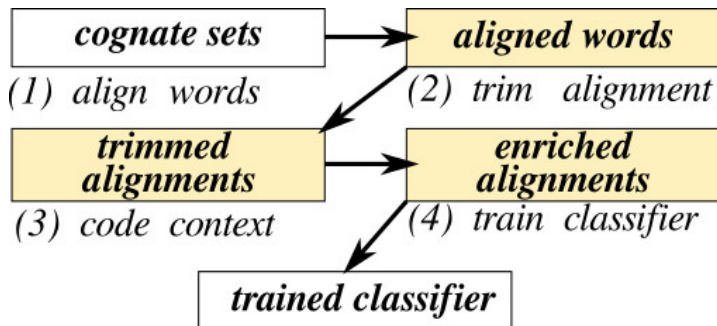
Examples: Shared Task on Word Prediction



List et al. (2022, SIGTYP Workshop)

Examples: Shared Task on Word Prediction

Baseline SVM (List et al. 2022)



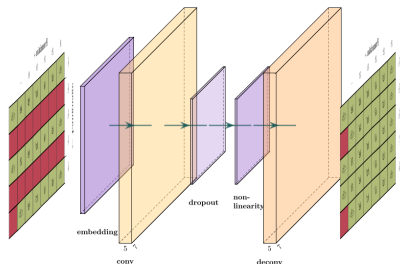
List et al. (2022, SIGTYP Workshop)

Examples: Shared Task on Word Prediction

I2 (Kirov et al. 2022)

		Pronunciation →					
		p_1	p_2	p_3	p_4	p_5	...
↓ Language	l_1	<S>	/s/	/i/	/n/	/d/	/e/
	l_2						
	l_3	<S>	/s/	/i/	/n/	/e/	/r/
	l_t						
	l_k	<S>	/s/	/i/	/r/	/e/	</S>

		Pronunciation →					
		p_1	p_2	p_3	p_4	p_5	...
↓ Language	l_1	<S>	/s/	/i/	/n/	/d/	/e/
	l_2	<S>	/s/	/i/	/n/	/a/	</S>
	l_3	<S>	/s/	/i/	/n/	/e/	/r/
	l_t	<S>	/s/	/i/	/n/	/e/	/r/
	l_k	<S>	/s/	/i/	/r/	/e/	</S>



List et al. (2022, SIGTYP Workshop)

Examples: Shared Task on Word Prediction

We find that the neural methods by Kirov et al. (2022) outperform all other methods. Their methods were carefully selected and adjusted from existing methods that solve tasks similar to the cognate reflex prediction task. Interestingly, our baseline, not supposed to play a major role in the shared task, performed not only surprisingly well, but needs only a very small amount of the computation time needed by all other methods proposed in the shared task. Both results emphasize the importance of *carefully designed, linguistically informed* workflows.

Examples: Shared Task on Word Prediction

The SIGTYP 2022 Shared Task on the Prediction of Cognate Reflexes

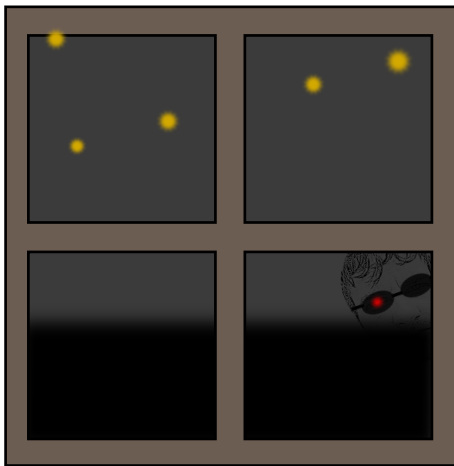
Johann-Mattis List^{III} Ekaterina Vylomova^Θ Robert Forkel^{III}

Nathan W. Hill^Λ Ryan D. Cotterell^Θ

^{III}MPI-EVA Leipzig ^ΘUniversity of Melbourne ^ΛUniversity of Dublin ^ΘETH Zürich
mattis_list@eva.mpg.de

List et al. (2022, SIGTYP Workshop)

Outlook



Outlook

- We need to unify qualitative and quantitative approaches in comparative linguistics.
- Classical comparative linguists should obtain some basic training in formal and computational approaches.
- Computational comparative linguists should obtain increased training in classical comparative linguistics.
- Computational methods based on machine learning should always be *carefully designed* and *linguistically informed*.
- Big Data won't solve all problems in comparative linguistics.
- Representatives of machine learning approaches in linguistics should take classical linguistic approaches seriously.



Thanks for your attention!

