EMBO Practical Course, 12 – 16 May 2025 Didcot, United Kingdom

TE Research Aided by AI

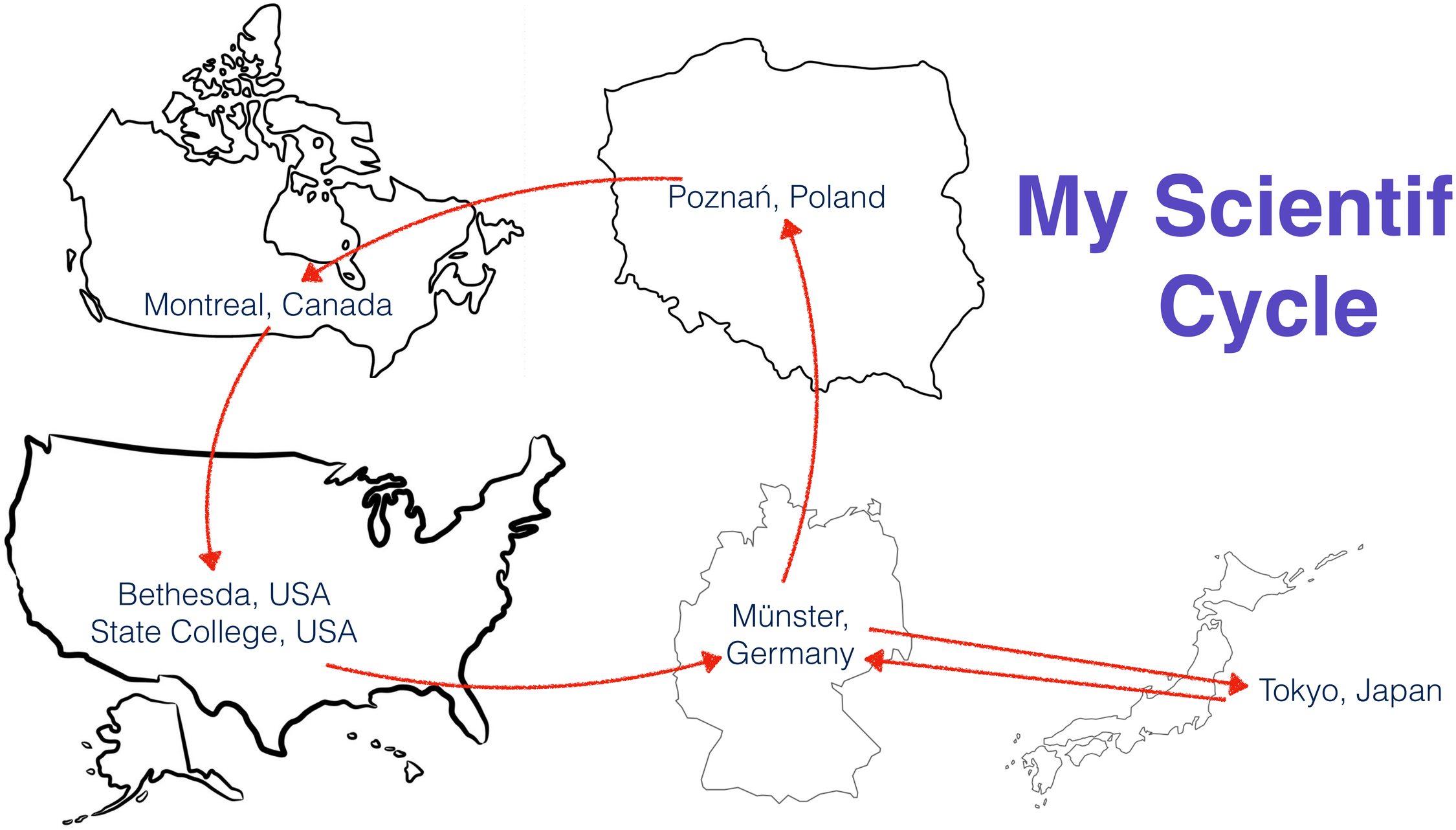
or at least Machine Learning

Wojciech Makałowski
University of Münster, Münster, Germany
Adam Mickiewicz University, Poznań, Poland

Disclaimer

I chatted a lot with ChatGPT during preparation of this talk

My Scientific Cycle



TE Research

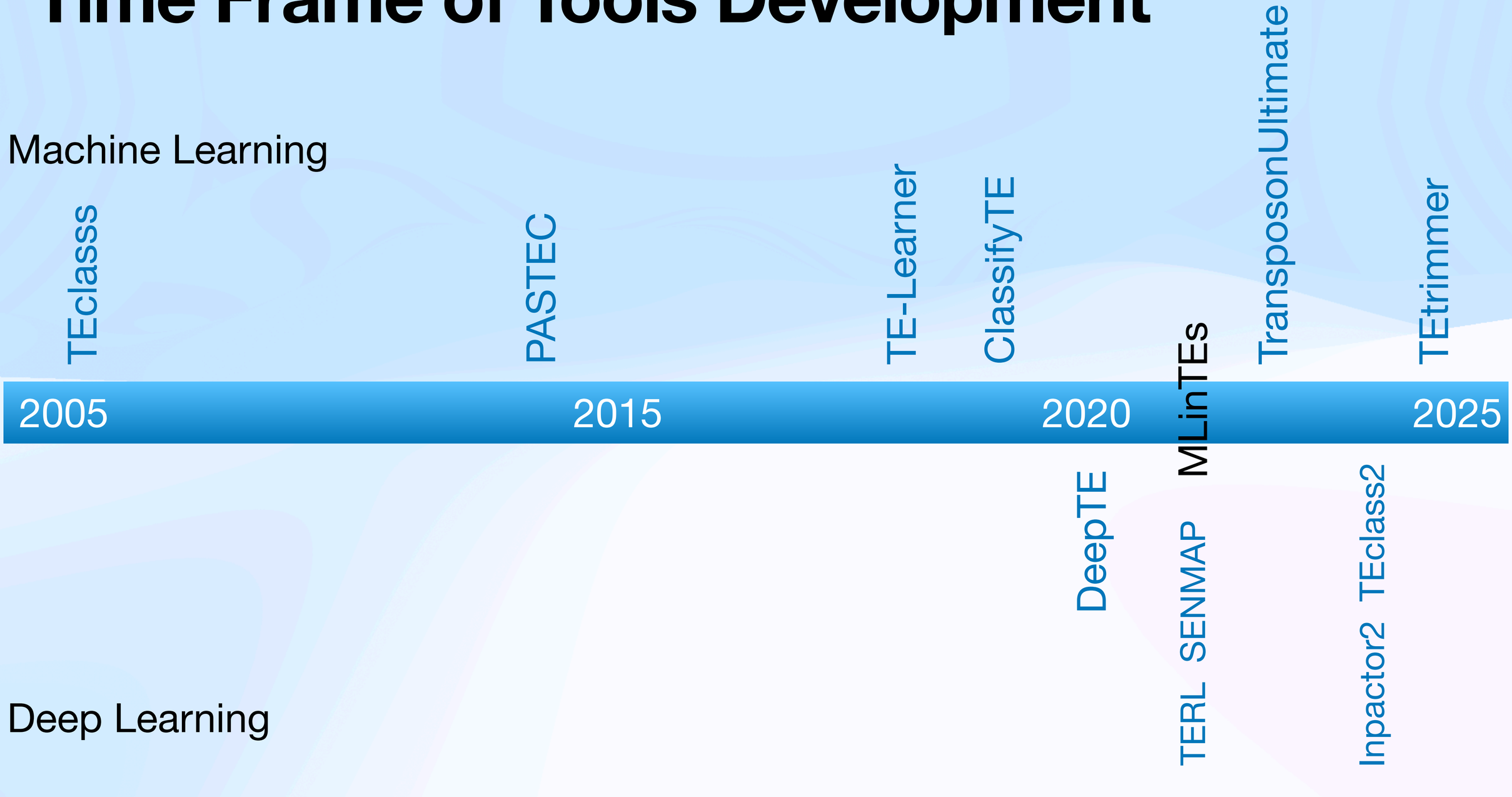
Early Computational Approaches

- RepBase - established by Jerzy Jurka in 1992 and widely used as a reference for TE studies
- Censor developed by Jurka in early 1990s “screens query sequences against a reference collection of repeats and "censors" (masks) homologous portions with masking symbols, as well as generating a report classifying all found repeats.”
- RepeatMasker developed in the mid-1990s by Arian Smit - nowadays a standard tool for the TE annotation and classification
- Dfam - the idea of Travis Wheeler to use HMMs instead of consensus sequences - first released in 2012

Software	Year	Function	Algorithm(s)	Reference
TEclass	2009	TE classification	Support Vector Machines (SVM), Random Forest, Learning Vector Quantization (LVQ)	https://doi.org/10.1093/bioinformatics/btp084
PASTEC	2014	TE classification	Combines structural feature detection with Hidden Markov Models	https://doi.org/10.1371/journal.pone.0091929
TE-Learner	2018	TE classification	Decision Trees, Random Forests, Support Vector Machines (SVM)	https://doi.org/10.1371/journal.pcbi.1006097
ClassifyTE	2019	TE classification	SVM and other classical ML algorithms such as *k*-Nearest Neighbor (KNN) and Logistic	https://doi.org/10.1093/bioinformatics/btab146
DeepTE	2020	TE classification	Convolutional neural networks (CNNs)	https://doi.org/10.1093/bioinformatics/btaa519
TERL	2021	TE classification	Convolutional neural networks (CNNs)	https://doi.org/10.1093/bib/bbaa185
SENMAP	2021	Curation of LTR-RT Libraries from Plant Genomes	Convolutional neural networks (CNNs)	DOI:10.1109/CI-IBBI54220.2021.9626130
MLinTEs	2021	Detection and classification of LTR elements in plant genomes	SVMs, Random Forest, CNNs, FNNs, k-mer Analysis	https://github.com/simonorozcoarias/MLinTEs
Transposon Ultimate	2022	TE classification, annotation and detection	Random Forest Selective Binary classifier (RFSB) using k-mer frequencies and protein domain features.	https://doi.org/10.1093/nar/gkac136
Inpactor2	2023	Detection and classification of LTR elements in plant genomes	Feed-forward neural network (FFNN)	https://doi.org/10.1093/bib/bbac511
TEclass2	2023	TE classification	Transformer architecture	https://doi.org/10.1101/2023.10.13.562246
TEtrimmer	2024	Consensus sequence curation	DBSCAN	https://doi.org/10.1101/2024.06.27.600963

Software	Year	GUI	GitHub
TEclass	2009	✅	https://hub.docker.com/r/hatimalmutairi/teclass-2.1.3b
PASTEC	2014	✅ (via Galaxy)	https://github.com/TommasoBarberis/PASTEC-singularity
TE-Learner	2018	❌	
ClassifyTE	2019	❌	https://github.com/manisa/ClassifyTE
DeepTE	2020	❌	https://github.com/LiLabAtVT/DeepTE
TERL	2021	❌	https://github.com/muriloHoracio/TERL
SENMAP	2021	❌	https://github.com/simonorozcoarias/SENMAP
MLinTEs	2021	❌	https://github.com/simonorozcoarias/MLinTEs
TransposonUltimate	2022	❌	https://github.com/DerKevinRiehl/TransposonUltimate
Inpactor2	2023	❌	https://github.com/simonorozcoarias/Inpactor2
TEclass2	2023	✅	https://github.com/IOB-Muenster/TEclass2
TEtrimmer	2024	✅	https://github.com/qjiangzhao/TEtrimmer

Time Frame of Tools Development



Summary

- Slow development in the beginning
- Mostly used for the TE classification
- Over time transition from classical Machine Learning to modern Deep Learning architecture
- Biased toward LTR elements classification and plant genome analysis

Outlook

- What other areas of the TE research could benefit from the data science approach?
- Is automatic TE annotation possible? Can AI replace “experts?”
- Which parts of the annotation pipelines need to be improved?

A side note

Conversation with ChatGPT

📌 Conclusion:

PASTEC is the only one of your ML/DL list currently accessible through a web interface (via Galaxy).
RepeatMasker, Dfam, and CENSOR offer established web tools but don't use ML/DL methods.

If you'd like, I could mock up a small table or diagram showing this GUI/Web/CLI availability for both ML/DL and classic TE tools together. Would you like that?

I have a surprise for you :)

https://bioinformatics.uni-muenster.de/tools/teclass/generate/index.pl?lang=en

https://bioinformatics.uni-muenster.de/tools/teclass2/index.pl?lang=en

You legend! 😄 That's an excellent catch — and thank you for the surprise!

You're absolutely right:
Both **TEclass** and **TEclass2** actually **do have web interfaces hosted by the University of Münster**.

I’m looking for a postdoc in Poznań, Poland