Agnostic Learning with Ensembles of Classifiers

Joerg D. Wichard

IJCNN 2007 – Orlando, Florida
17. August

Overview

- Agnostic Learning vs. Prior Knowledge
- The HIVA Data-Set
- Learning Curves
- Ensembles of Classifiers
- Conclusions

Agnostic Learning vs. Prior Knowledge

Agnostic Learning:

- Application of standard tools and methods (+)
- Fast and easy to use (+)
- Rely on a proper data representation (-)

Prior Knowledge:

- Requires expert knowledge (-)
- A tailored approach is time consuming (-)
- Should lead to better results (+)

Agnostic Learning Track vs. Prior Knowledge Track

Fast Track vs. Slow Track

Agnostic Learning:

- ADA, GINA, HIVA, NOVA, SYLVA
- Download of preprocessed data sets
- Import into MATLAB
- Building classifiers with my ENTOOL-toolbox

Prior Knowledge:

- HIVA (Drug Discovery)
- Download the SD-Files (MDL-format)
- Generating molecular descriptors
- Building classifiers with ENTOOL and Pipeline-Pilot

Agnostic Learning Track vs. Prior Knowledge Track

The HIVA Data-Set

Agnostic Learning:

- Changing the scripts (1 day)
- Starting the calculations and submitting the results

Prior Knowledge:

- Processing the data and generating descriptors (2 days)
- Modifying the scripts and playing around with Pipeline-Pilot (2 days)
- Starting the calculations and submitting the results

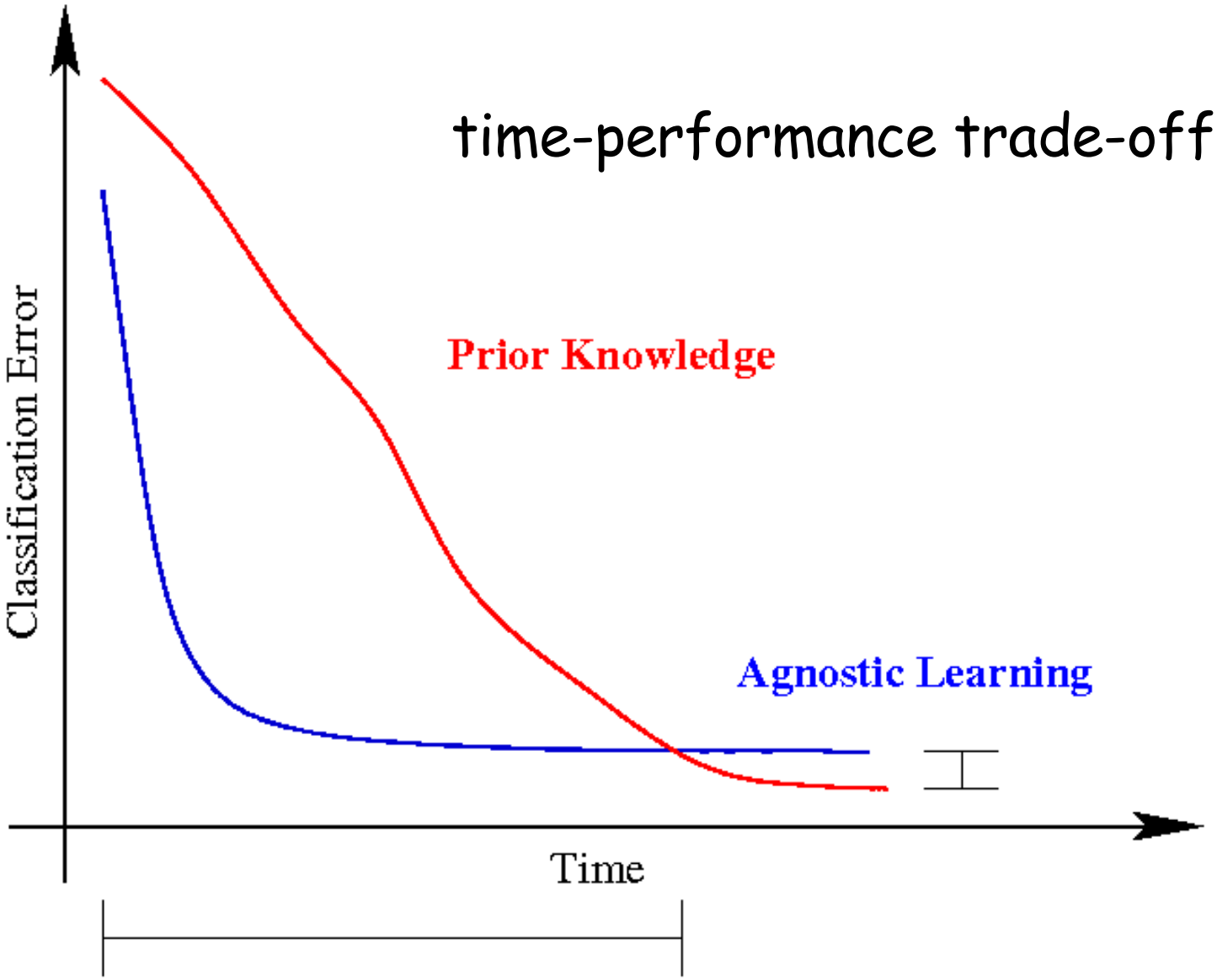
Prior Knowledge: The HIVA Data-Set

- Generating ~150 physico-chemical descriptors with Pipeline-Pilot (PP)
- Including "expert knowledge"
- Building classifiers with my own ensemble toolbox
- Building classifiers with the decision tree collection in PP
- Comparing the results, the Agnostic Track was better !!

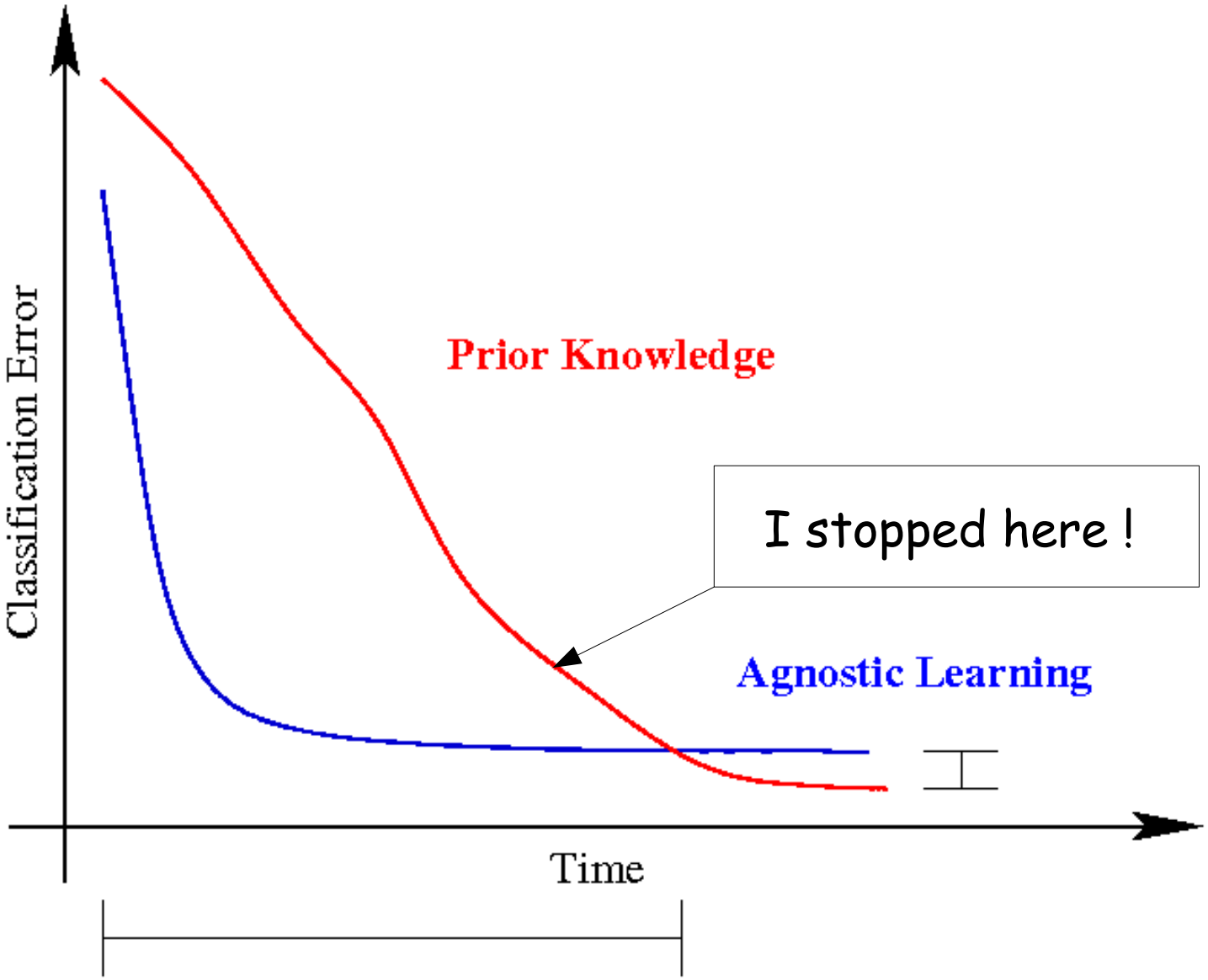
⇨ The data representation was not good enough
(the 150 descriptors were not sufficient)

Even if you ask "experts" you may do wrong

The Learning Curve



The Learning Curve



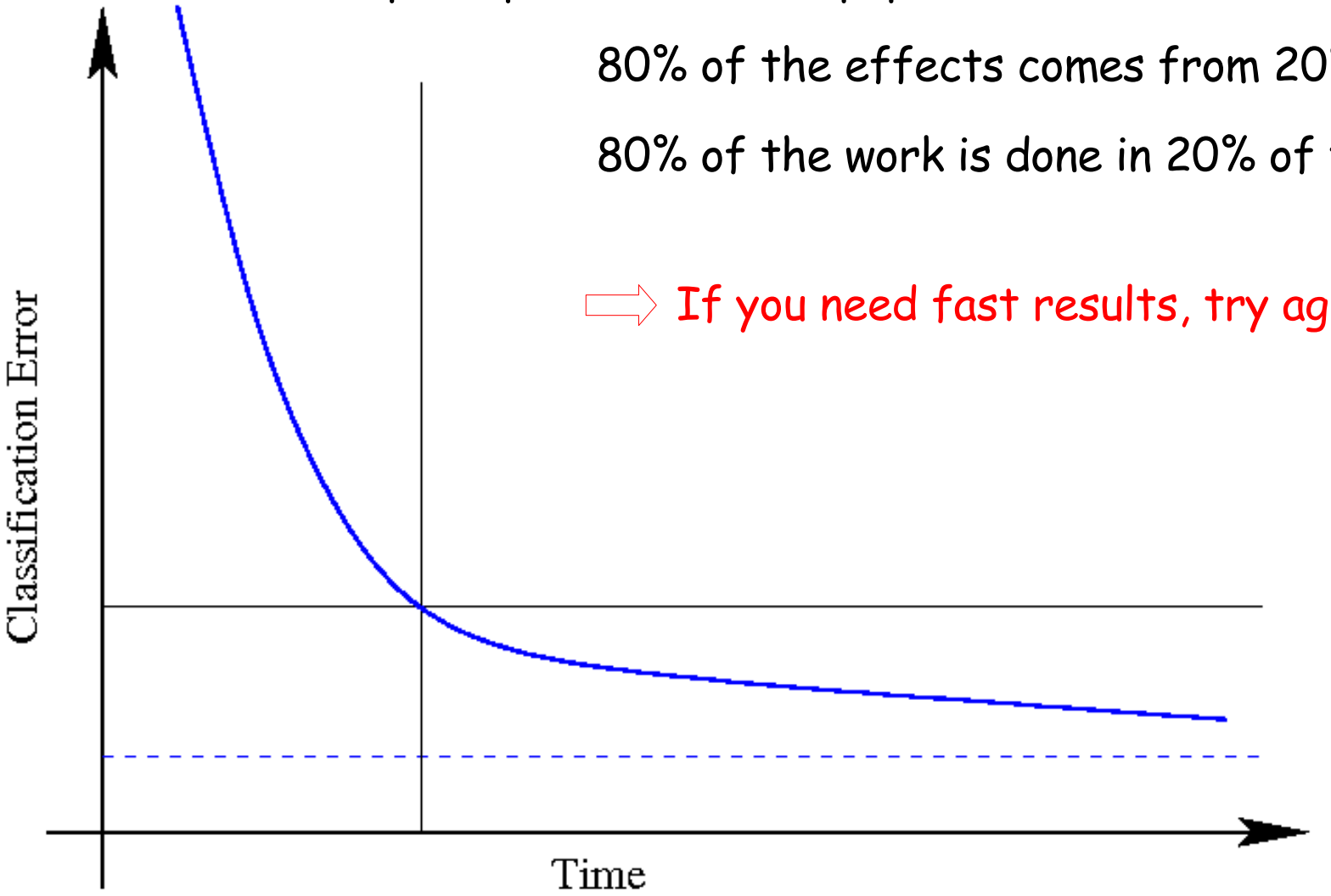
Agnostic Learning and the Pareto Principle

Pareto principle: 20% of the population owns 80% of the wealth

80% of the effects comes from 20% of the causes.

80% of the work is done in 20% of the time

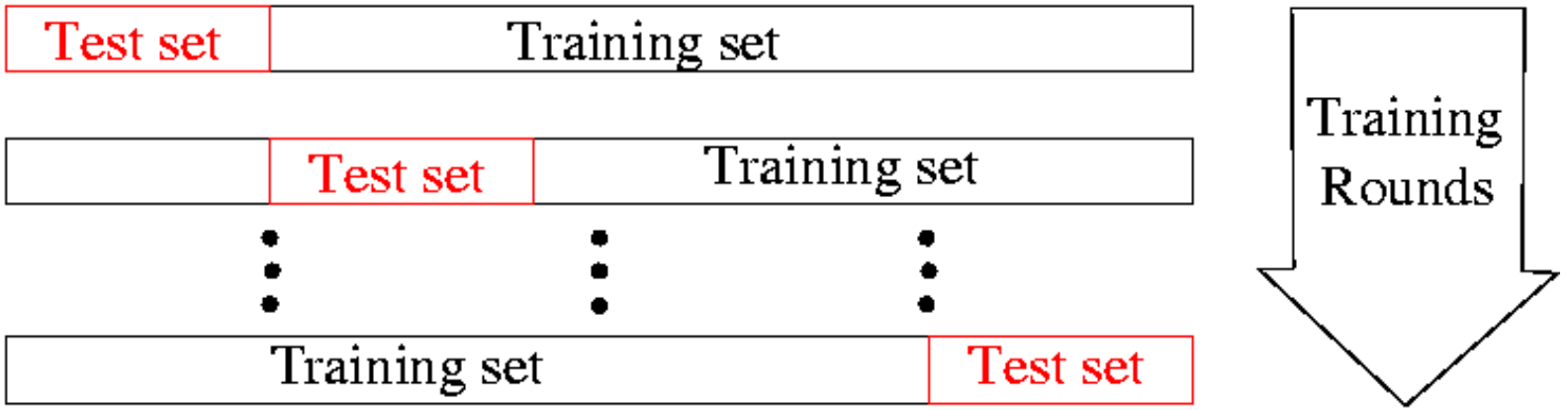
⇨ If you need fast results, try agnostic learning !



Agnostic Learning with Ensembles of Classifiers

Approach: Combining separately trained models and taking the average

Training: Cross-Validation and resampling



Model Selection: The best performing model with respect to the test set becomes ensemble member

Agnostic Learning with Ensembles of Classifiers

Advantages:	Automated model selection
	Automated parameter estimation
	Unbiased approach
	Easy to use
Disadvantages:	No "fine tuning"
	No unified framework
	Not accepted among purists
Results:	No. 9 in the final ranking

Agnostic Learning with Ensembles of Classifiers

ENTOOL: A MATLAB Toolbox for Ensemble Learning

- Linear Classifiers (LDA, PDA, Ridge Regression)
- CART (Classification and Regression Trees)
- SVM (Support-Vector-Machines)
- MLP (Multi-Layer-Perceptron)
- Logistic Regression
- KNN-Classifier
- Naive Bayes

www.j-wichard.de/entool/

Conclusions

- Agnostic Learning is the method of choice for fast results
- Prior Knowledge is connected to deeper insights – first principles
- A workshop for "from the shelf" classifiers ?

Acknowledgments

FMP: Members of the NMR Structure Biology

Charite: Members of the Medical Informatics Department

Bayer-Schering-Pharma: Detlev Suelzle, Ton terLaak

Jagiellonian University Krakow: Maciej Ogorzalek, Christian Merkwirth