Lars Kai Hansen
lkai@dtu.dk

Machine Learning as a Service – what can be automated?

Learning from past workflows
Mission of MLaaS
The no free lunch theorem
Decision making
Representation & deep learning

Towards general workflows for key big data tasks

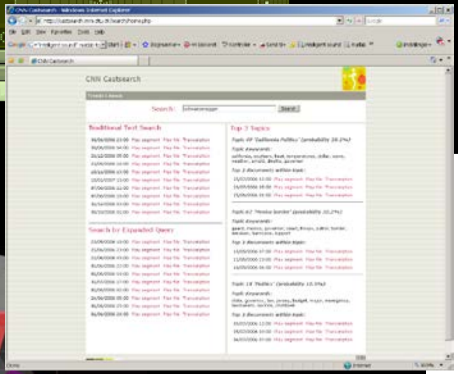
Data driven services require integrated workflows!

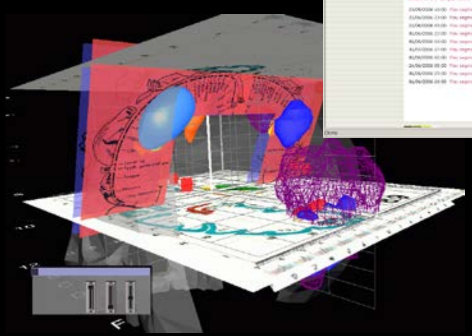
DTU ML audio demos

MIRocket (2006)
CastSearch (2007)
Muzeeker(2009)
CoSound (2015)

Medical search engines

Brede Search neuroinformatics (2005)
FindZebra, diagnostic queries (2013)





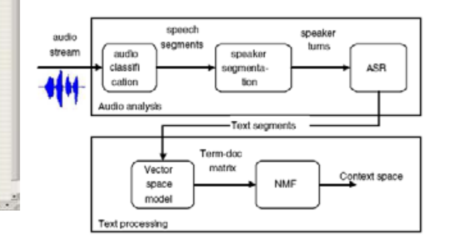
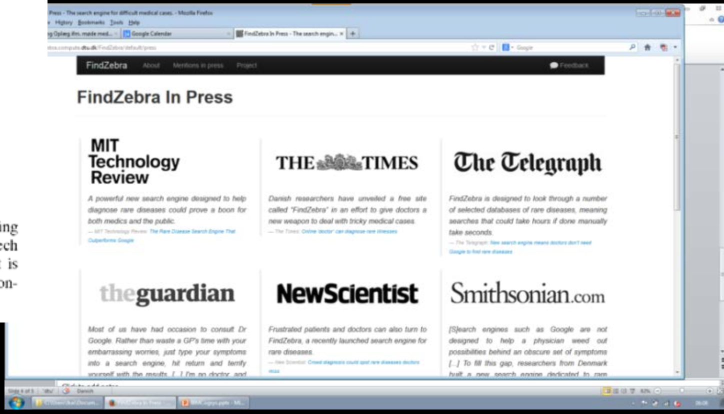


Fig. 1. The system setup. The audio stream is first processed using audio segmentation. Segments are then using an automatic speech recognition (ASR) system to produce text segments. The text is then processed using a vector representation of text and apply non-negative matrix factorization (NMF) to find a topic space.



Machine Learning as a Service

DataRobot

BUTLER ANALYTICS

10+ Machine Learning as a Service Platforms - Mar 21, 2016

ABM

Google Developers

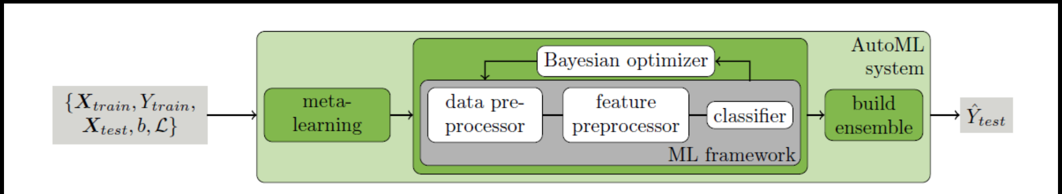
bigml

Algorithms.io a LUMENDATA company

Data Analysis and Model Building Fees $0.42 per Hour

amazon web services

Commercial services, often strong focus on machine learning - limited focus on workflows - Increasing interest in data science: AutoML Competitions…



Feurer M, Klein A, Eggensperger K, Springenberg J, Blum M, Hutter F. Efficient and robust automated machine learning. In Advances in Neural Information Processing Systems 2015 pp. 2962-2970.)

The mission of MLaaS

Basic machine learning functions are well defined, have well-understood representations and learning curves. Large number of open source tools are available to the data scientist

<u>Prediction</u>, "supervised learning" p(action|measurement)
replacing a human operator, diagnostics,quality control, monitoring etc etc

<u>Structure</u>, "unsupervised learning" p(measurement|structure)
exploratory analysis and simulation, novelty detection, cleaning

Tuning complexity to given application still a challenge…
<u>MLaaS: Help data scientist to tune, validate, understand ML</u>

DTU Courses: <u>02450 Introduction to ML</u>,
02457 Non-linear signal processing,
02460 Advanced ML, 02456 Deep learning
02807 Computational Tools for Big Data
02901 Topics in advanced ML (PhD)

MOOCS galore…

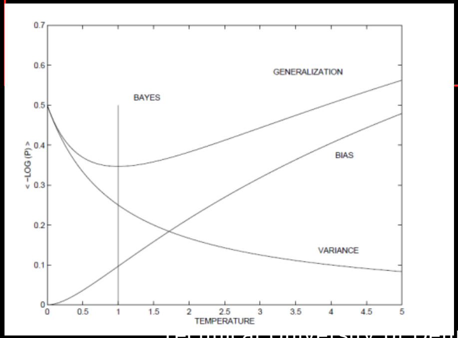
No free lunch in data science

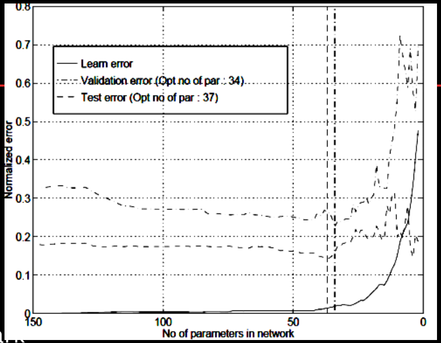
Bad news: **Without modeling "a priori" you can learn nothing!**
Without priors, you are left with "uniform inference" (Wolpert, D.H., 1996. The lack of a priori distinctions between learning algorithms. *Neural computation*, *8*(7), pp.1341-1390.)

Good News: **Well-understood means for optimizing models for relevant & "physically" relevant priors/representations – typical tuning complexity**

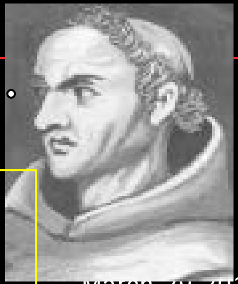
Bayes vs Cross-validation - **Bayes is optimal in the rare case when you got the prior right (no complexity tuning….**

Cross-validation optimizes when prior is uncertain….
Often used forms of prior information used:
Smoothness in feature space, sparsity / dimensional, transfer learning…





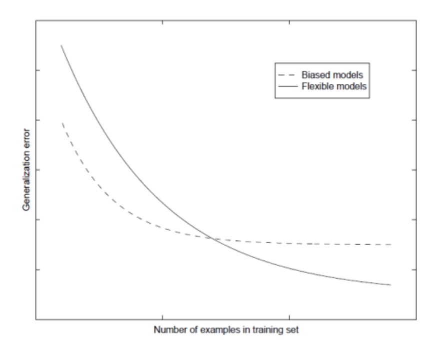
Hansen, L.K. Bayesian Averaging is Well Tempered. In *NIPS* 1999:265-71.

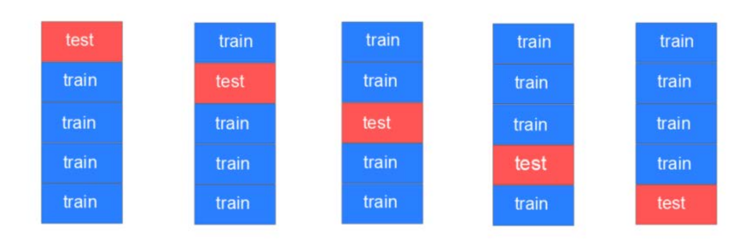


Decision support p(decision | measurement)

Classification, signal detection, discrimination

- Generalization: Predcition errors on test
- Training, validation, test simulated by resampling





Tuning: Complexity / smoothness of classifier, calibration of probabilities for risk assesment

Potential for automatic tuning: High, good examples provided by Hutter group:

”In this work we introduce a robust new AutoML system based on scikit-learn (using 15 classifiers, 14 feature preprocessing methods, and 4 data preprocessing methods, giving rise to a structured hypothesis space with 110 hyperparameters). This system, which we dub auto-sklearn, improves on existing AutoML methods by automatically taking into account past performance on similar datasets, and by constructing ensembles from the models evaluated during the optimization”
Feurer M, Klein A, Eggensperger K, Springenberg J, Blum M, Hutter F. Efficient and robust automated machine learning. In Advances in Neural Information Processing Systems 2015 pp. 2962-2970.)

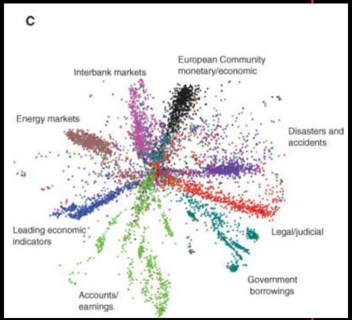
Structure & representation learning p(measurement | structure)

Typical use cases:

Outlier detection, cleaning ...flag data with low relative probability density

Feature engineering, representations in terms of sparse independent components

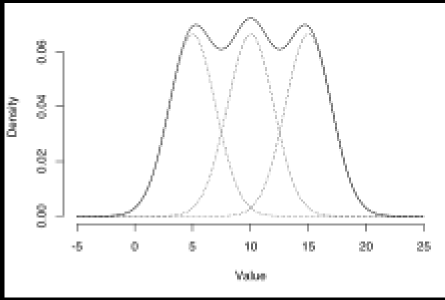
Pretraining for neural networks: Hinton's revelation
(Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. "Reducing the dimensionality of data with neural networks." *Science* 313.5786 (2006): 504-507.)



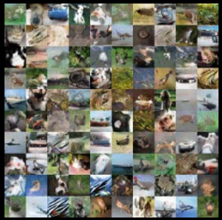
Tuning: Complexity of structure
(e.g. #clusters, latent variable dimensionality,…)

Potential for automatic tuning: High, work is ongoing (F. Zdyb)

Gaussian (and other) mixtures



Generative adverserial networks



Non-parametrics / kernel methods

Workflows for prototypical applications

DABAI case studies: Data cleaning & outlier detection

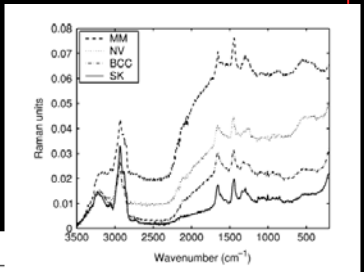
Flattening of data structures to form basic ”feature space” (Domain expert)

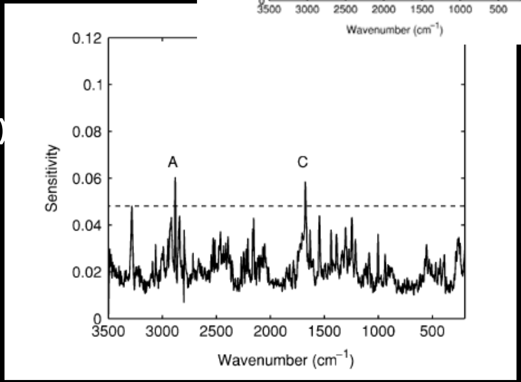
Structure learning in feature space to find local simple distributions, e.g. clustering. (AutoML)

Interpret local distributions / clusters using prototypes (Domain expert)

Find relative to local outliers (AutoML)

Evaluate prototypes and outliers (Domain expert)





DABAI case studies: Classification

Flattening of data structures to form basic ”feature space” (Domain expert)

Train and cross-validate classifier. (AutoML.)

Visualize feature relevance (AutoML)

Evaluate feature relevance (Domain expert)

DABAI case studies: Missing data handling

Flattening of data structures to form basic ”feature space”. (Domain expert)

Train and cross-validate generative models for which marginal distributions can be obtained (AutoML).

Infer latent variables for missing records (AutoML)

Sigurdsson, S., Philipsen, P.A., Hansen, L.K., Larsen, J., Gniadecka, M. and Wulf, H.C., 2004. Detection of skin cancer by classification of Raman spectra. *IEEE transactions on biomedical engineering*, *51*(10):1784-1793.

Data science
with machine learning as a service

<u>Data driven services require integrated workflows!</u>

Multiple steps … some interactive, some automated

Interactive: Data preparation / exploration

Automate: Machine learning representations

Automate: Machine learning decision support systems

Interactive: Model interpretation / decision making