Slide 1

How deep is deep & what's next in computational intelligence?

Lambert Schomaker

Artificial Intelligence & Cognitive Engineering Institute
Center for Data Science & Systems Complexity DSSC
University of Groningen

Schomaker (2016) How deep is deep and what is next in computational intelligence? [keynote lecture] International Conference on Frontiers in Handwriting, 23-26 Sept. Shenzhen, China.

 l.r.b.schomaker@rug.nl

Slide 2

Recent advances in ML?

- Neural networks & Deep Learning
- Critical remarks
- Monk: massive shallow but convenient learning

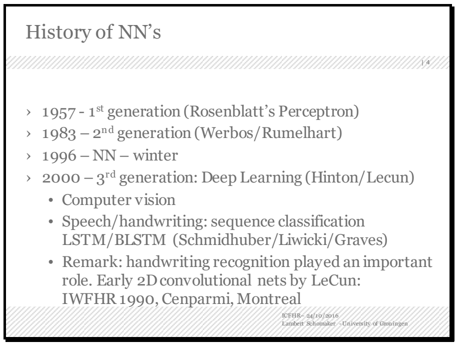
Slide 3

Deep learning / Recent advances in ML

- 'Google self-driving cars'
- Predicting internet user interests ('cookies')
- Twitter-based epidemiology ('flu tweets')
- Create a van Gogh or Munch version of a photograph
- Coloring of B/W movies
- Learning to play Atari Breakout, Pacman etc.
- AlphaGO: computer wins at playing GO
- Improved training (loss function, softmax, ReLU)
- With 1000 hidden layers (Susillo & Abbot, 2015)
- etc.

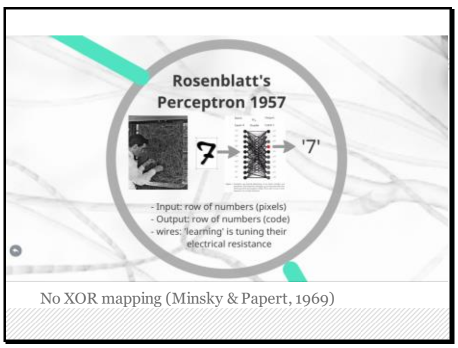
List of current successes in deep learning

Slide 4



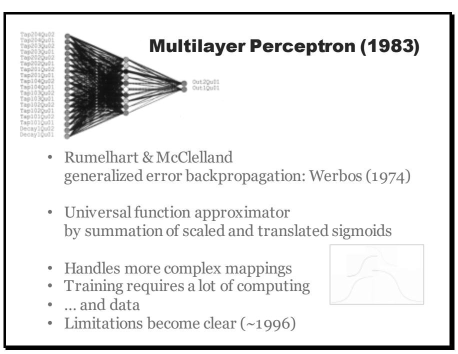
Brief history of NNs

Slide 5



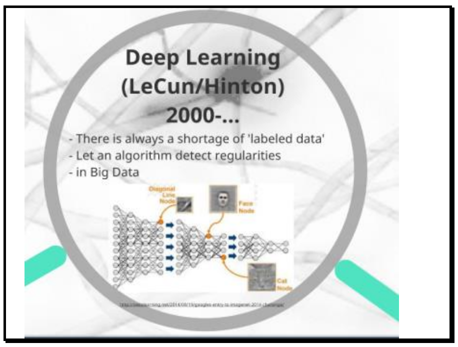
It is remarkable that after Minsky and Papert the rejection of the perceptron was so massive. After all, linear systems with only an input and an output layer still can do a lot and also were in use. Consider for instance Widrow & Hoff, telephone line echo cancellation using a linear system. But indeed, non-linear mappings are impossible and the fact that the output units are thresholded does not introduce a non-linearity in the forward mapping itself.

Slide 6



Finally: non-linear mappings are possible. Rumelhart & McClelland came from psychology. The published two books, one blue, one brown, as a set, with a yellow third book for students. It had a diskette with C code. By 1996 there were many frustrations with MLP. There was not enough labeled data, computers were slow and generalization was not good. The SVM was developed at AT&T by Isabelle Guyon, her husband Bernard Boser and Vapnik, on the basis of the problems in training handwriting recognizers. The bosses at AT&T were not happy with the fact that NNs yielded different solutions from different randomisations.

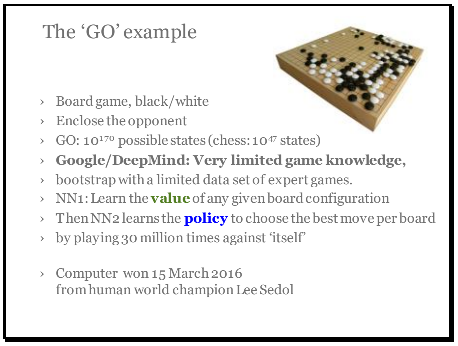
Slide 7



Could have been a better slide, but we all know the drill: CNNs finally have their breakthrough. It must be admitted that for a long time Yan LeCun was the only one with very good results on CNNs. The community was also surprised with the guts of Hinton to publish in Nature about what many were already doing quite extensively: using autoencoders or diabolo MLP for dimensionality reduction because there is no need for labels. But Hinton added some very useful tricks that would ultimately allow for the deep learning revolution. Both researchers were essential, in any case. Hinton, G. E. and Salakhutdinov, R. R **Reducing the dimensionality of data with neural networks.** Science, Vol. 313. no. 5786, pp. 504 - 507, 28 July 2006.

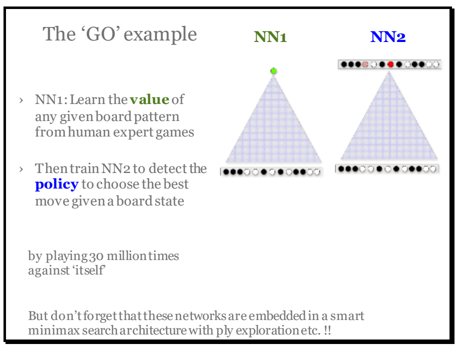
LeCun, Y., Bengio, Y. and Hinton, G. E. **Deep Learning.** Nature, Vol. 521, pp 436-444.

Slide 8



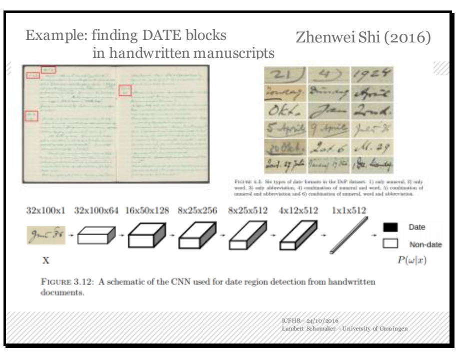
AlphaGO: a real system entails much more than just one single deep net

Slide 9



The point is that Deep Learning in itself is hardly interesting. Only by integrating multiple networks into a functional architecture for the operational stage, they will be useful, as in AlphaGO.
My favorite metaphor for an isolated NN in this respect is the Ferrari engine bolted to a workbench in your shed. Very powerful but utterly useless.

Slide 10



We are currently (evidently) also working on CNN and LSTM. The example here is challenging because of the low prior probability of finding a data in a sea of regions of interest that are not representing a DATE block.

Slide 11

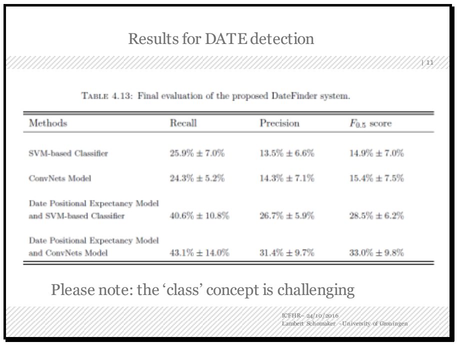
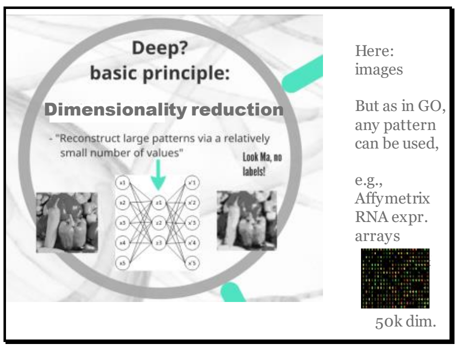


TABLE 4.13: Final evaluation of the proposed DateFinder system.

Methods	Recall	Precision	$F_{0.5}$ score
SVM-based Classifier	25.9% ± 7.0%	13.5% ± 6.6%	14.9% ± 7.0%
ConvNets Model	24.3% ± 5.2%	14.3% ± 7.1%	15.4% ± 7.5%
Date Positional Expectancy Model and SVM-based Classifier	40.6% ± 10.8%	26.7% ± 5.9%	28.5% ± 6.2%
Date Positional Expectancy Model and ConvNets Model	43.1% ± 14.0%	31.4% ± 9.7%	33.0% ± 9.8%

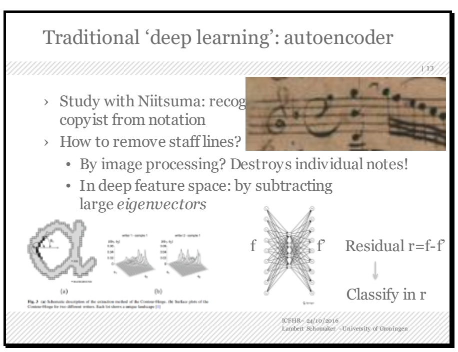
Results on generic DATE block detector. Please note that the prior is very small, 30% is not so bad. Here positional density was used to weigh likelihoods (positional expectancy model)

Slide 12



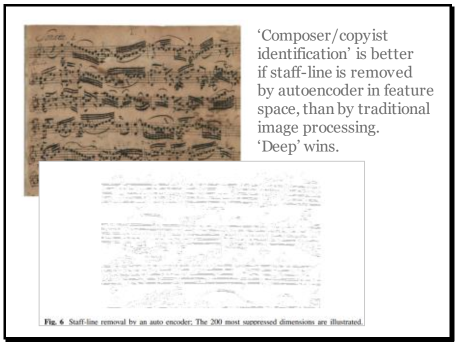
With a dimensionality of 50k, it really becomes useful to do dimensionality reduction. NNs are in many ways convenient, but it is difficult to convince biomedical researchers.

Slide 13



The goal of this study was writer identification on musical script. The staff lines do not represent relevant information. How to get rid of them? In the image or in feature space?

Slide 14

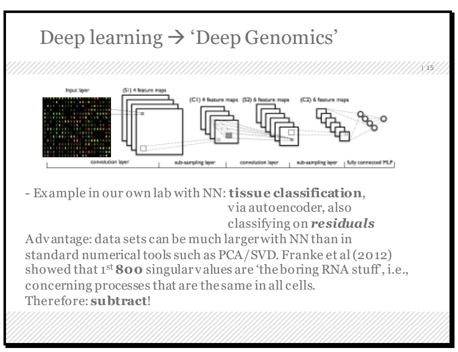


Niitsuma, Masahiro and Schomaker, Lambert, van Oosten, Jean-Paul, Tomita, Yo and Bell, David (2016). Musicologist-driven writer identification in early music manuscripts,Multimedia Tools and Applications, 75(11), pp. 6463—6479

abstract="Recent renew ed interest in computational w riter identification has resulted in an increased number of publications. In relation to historical musicology its application has so far been limited. One of the obstacles seems to be that the clarity of the images from the scans available for computational analysis is often not sufficient. In this paper, the use of the Hinge feature is proposed to avoid segmentation and staff-line removal for effective feature extraction from low quality scans. The use of an auto encoder in Hinge feature space is suggested as an alternative to staff-line removal by image processing, and their performance is compared. The result of the experiment show s an accuracy of 87 {\%} for the dataset containing 84 w riters' samples, and superiority of our segmentation and staff-line removal free approach. Practical analysis on Bach's autograph manuscript of the Well-Tempered Clavier II (Additional MS. 35021 in the British Library, London) is also presented and the extensive applicability of our approach is demonstrated

http://dx.doi.org/10.1007/s11042-015-2583-8

Slide 15



Rudolf Fehrmann, Lude Franke et al. found this. They have several publications, among which one on SSVD (sparse svd) in Nature.

Slide 16

Define ‘deep’!

› Is it the convolutional aspect?

› Is it the number of layers?

› Is it the dimensionality reduction?

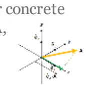
Define ‘deep’

Slide 17

Two notions of ‘deep’

› Deep: is hidden, geometric, averaged over concrete instances, a subspace, etc., like PCA, correlation patterns, non verbal

Yes!



› Deep: is an exact fact, hidden in a graph, an unexpected explanation, precise, explainable to humans?

Not yet!

As, e.g., in causality inference (‘deep cause’)

Two notions of ‘deep’

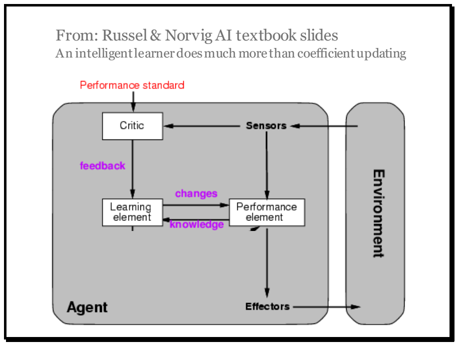
Slide 18

Time to identify what **cannot** be done!

› Deep Learning is no computational intelligence, yet
› Intelligence by proxy: over the supervised labels
› A smart human PhD is always necessary
› No general intelligence: each experiment is a *one-trick pony*
› Extensive, laboratory-based training
› Generalisation to real, new data from new sensors, from new contexts is still difficult: k-fold evaluation is still a scam:
› ‘i.i.d.’ and sampled from one cleaned pool of data yields overly optimistic performance estimates

At 10 minutes

Slide 19

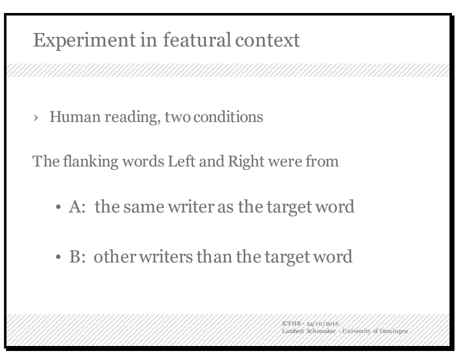


Intelligent learners should be able to know what they don't know and propose experiments (i.e., samplings from the total data) to improve the current learning status.

Slide 20

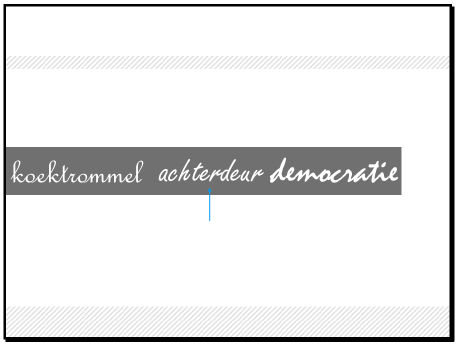
This example can be read quickly by most human subjects. This did not involve a (brain-based HMM) training on these words including the random permutation probabilities. Rather, an opportunistic use is made by the human reader, of the fact that: The first and last letter are ok; and that the other letters only show their presence within the word while their position is irrelevant. All this is done on the fly, in the operational stage, by humans at least.

Slide 21

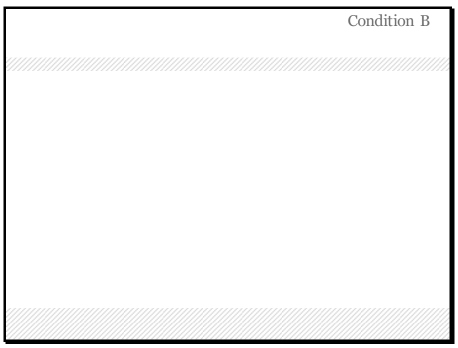


Experiments performed with Hans-Leo Teulings and students in my Nijmegen era (pre 2001).

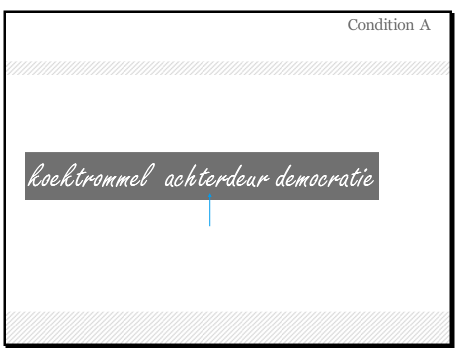
Slide 22



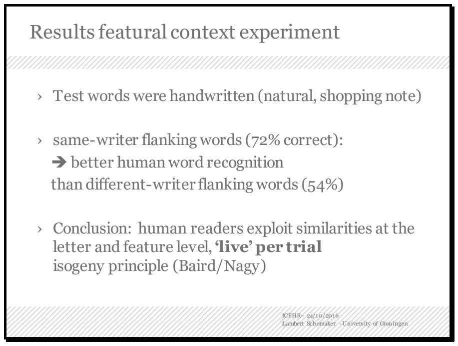
Slide 23



Slide 24



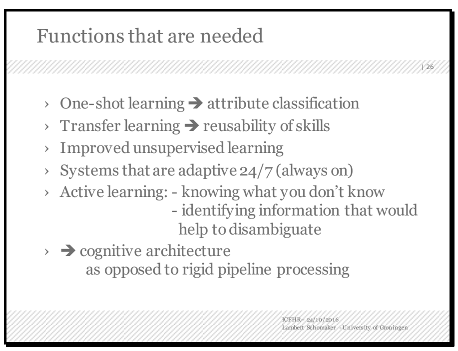
Slide 25



Sriharsha Veeramachaneni and George Nagy (2005). Style context with second-order statistics, IEEE Transactions on Pattern Analysis and Machine Intelligence, 27, 14--22

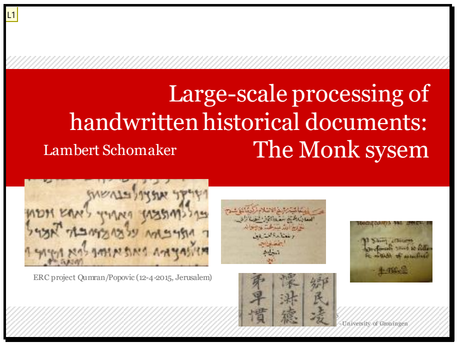
Scaling Up Whole-Book Recognition Pingping Xiu & Henry S. **Baird** Computer Science & Engineering Dept Lehigh University 19 Memorial Drive West, Bethlehem, PA 18017 … Published in:international conference on document analysis and recognition · 2009 Authors:Pingping Xiu · Henry S **Baird**

Slide 26



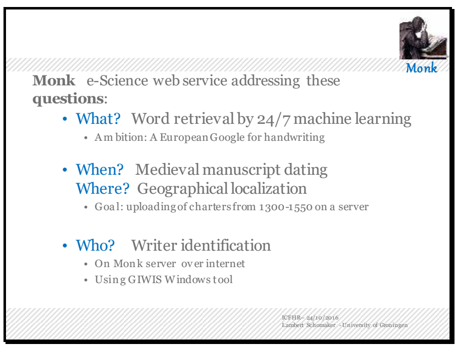
Timing: You really should be halfway now

Slide 27



At 20 minutes

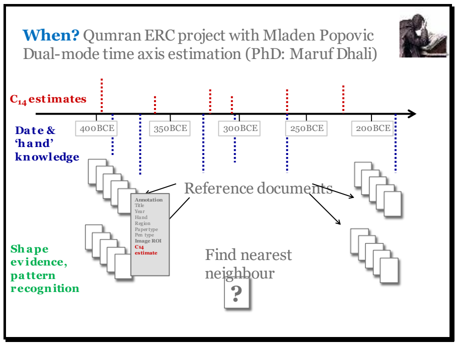
Slide 28



Slide 29

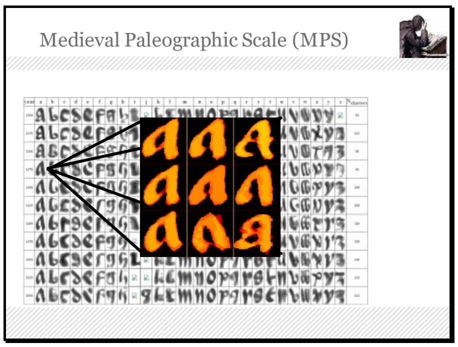
What? (is written): recognition and retrieval of text

Slide 30



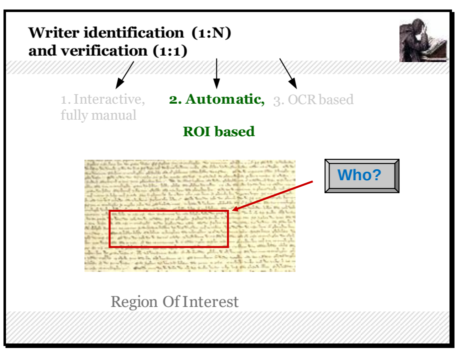
When? (has it been written)

Slide 31



In the MPS project (PhD student: Sheng He), we developed textural methods for dating of acts. However, also individual segmented characters show paleographical developments in a traditional manner.

Slide 32



Who? (wrote it)

It is much less work to just define a region of interest for writer identification (or dating)

Slide 33

Researchers & Monk

- Marius Bulacu, Axel Brink, Katrin Franke
- Ralph Niels, Louis Vuurpijl, Jean-Paul van Oosten, Sheng He, Jan Burgers, Petros Samara, Olarik Surinta
- "The Nijmegen Handwriting Group 1984-1993"
- Netherlands Forensic Institute: Ton Broeders, Wil Fagel, Elisa van den Heuvel
- *Isabelle Guyon, Rejean Plamondon*
- 2016: Maruf Dhali, Mahya Ameryan, Sukalpa Chanda
- **Users (humanities researchers)**:
  - Jinna Smit, Mark Aussems, Masahiro Niitsuma, Mladen Popovic, Daniel Stoekl, Jetze Touber, Grace Fong, Elaine Treharne, Dominique Stutzmann, Andreas Weber, Maxim Romanov, many others

People who has worked on Monk or had an influence on its development.

Also humanities scholars: They are high

Slide 34

Case: Optical Character Recognition

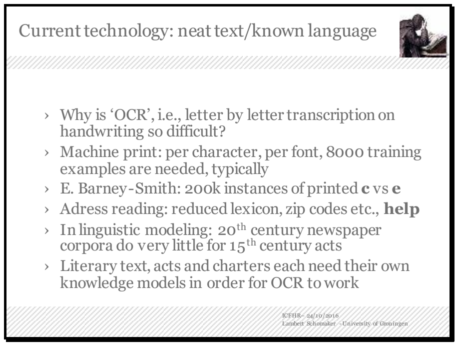
- "The process of segmenting a text image into individual c h a r a c t e r images and classifying each as being a letter in an alphabet"
- Impressive results on contemporary printed text in machine fonts: with some linguistic postprocessing results are close to 100%
- Solved?

Slide 35

OCR ?

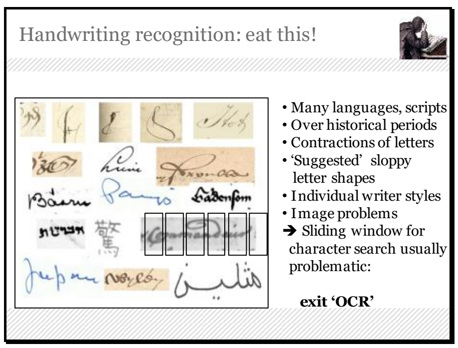
- 'OCR' on historic documents does not work well
- On *handwritten* manuscripts it doesn't work, at all!
- Problems:
  - image quality
  - unknown character shapes
  - unknown statistical language models
- However: pattern recognition and machine learning make enormous progress these days!
- Which methods? How to apply them?

Slide 36



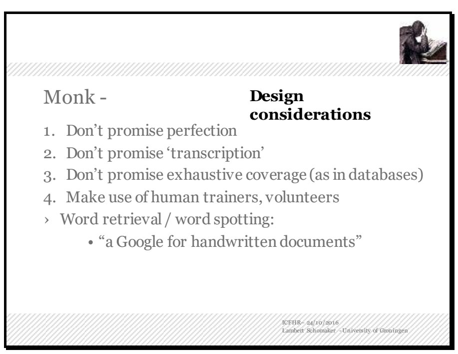
For example: What would a TREC corpus do for medieval administrative text?
In any case, shape recognition needs to be strong if such additional sources cannot be used.

Slide 37



Exit OCR as: Exit the methods that assume identifiable individual characters in the input stream for all letters of the intended word.

Slide 38



Slide 39

Monk - Design considerations

- Word based:
  - "a Google for handwritten documents"
- The word is a reliable chunk of information with many shape features: redundancy

- Big Data: With sufficient data, there is always a reasonable response on a query

Slide 40

Monk's world model:

- Institutes
  - Collections
    - Books (i.e., documents)
      - Pages
        - Paragraphs
          - Lines
            - Word zones and characters
              - Pixels

Slide 41

Coarse overview of **current** contents Monk, 400+ documents, 75k page scans, 1.8M lines, 777k labels

- KdK Dutch administrative 1893-1906 20 books
- Dutch Admirality 1760-1823 5 books
- Printed, Elzevirium, 1616
- Qumran scrolls (2463 scans)
- Middelduits (example document)
- Accounts, 1425 Gelria
- Schepenbank Louvain, 1421-1559, 3 books
- Colonial diary (1932)
- Municipal year report 1855
- 20k illuminated initials
- Russian handwritten newspaper 1672
- Scholarly correspondence 1674-1682
- Chronicon Boemorum 1201
- Homiliarum Opatovicense 1150
- Resoluties Staten Generaal 1627
- Medieval charters 1300-1550 in 25yr periods
- Witch trial 1605
- Chancellery Philippe le Bel 13xx
- Hieratic (Egyptian)
- Hieroglyphs
- 'Beowulf' and related, 4 books
- Ming Qing Chinese poetry , 300+ documents
- Arabic document, 291 scans
- Charlotte Perkins-Stetson diaries 1883- (682p.)
- Wittenbergsches (fraktur machine print)
- Etc. ...

Cooperations with:

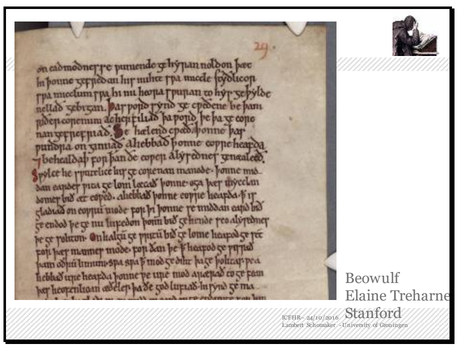
- *Huygens institute, Sorbonne,*
- *Harvard,*
- *Stanford,*
- *Czech National Library,*
- *Dutch National Library, McGill Univ.,*
- *Utrecht Univ.,*
- *Univ. Uppsala*
- *City archive Louvain,*
- *Leiden Univ.*
- *CNRS*

Note that almost all documents are handwritten, but some machine printed text is also in Monk. This entails difficult material such as German fraktur, Arabic and printed hieroglyphs.

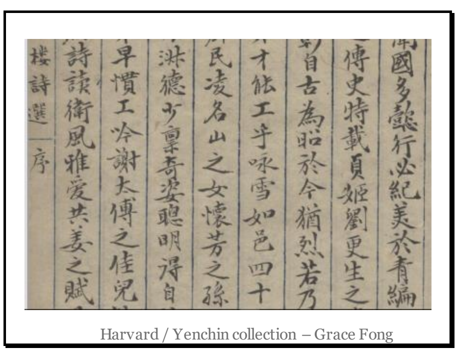
Slide 42

While binarisation is usually too destructive, Otsu-based contrast enhancement works well, especially if local Otsu (or other method) is used.

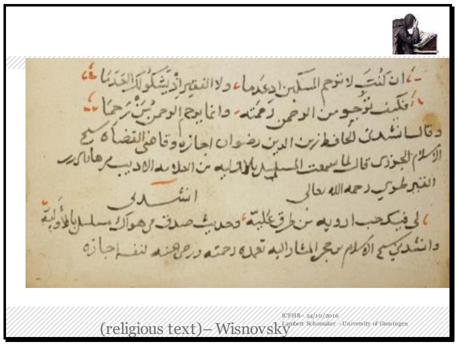
Slide 43



Slide 44



Slide 45



Slide 46

[صاحب إفريقية ينسلخ من الإمارة ويتصوّف]

وفيها قَوِيَت أمور أبي عبدالله الشّيعيّ بالمغرب، فصنع صاحب إفريقية صُنْع محمد بن يَعْفُر ملك اليمن، فانسلخ من الإمارة، وأظهر توبةً، ولبِس الصُّوف، وردَّ المَظَالم، وخرج إلى الرّوم غازياً. فقام بعده ابنه أبو العبّاس(١).

وكان خروج إبراهيم بن أحمد صاحب إفريقية(٢) منها وركوبه البحر سنة تسعٍ وثمانين، فوصل إلى صِقِلّية، ومنها إلى طَبَرْمِين، فافتتحها، ثمّ حاصر كنيسة، فمرض بإسهال، ومات في ذي القعدة. وكانت ولايته ثمانية وعشرين عاماً ونصف(٤)، ودفِن بصِقِلّية(٥).

(١) أنظر هذا الخبر في:
تاريخ الطبري ٩٦/١٠، والعيون والحدائق ج ٤ ق ١٨٢/١، وتجارب الأمم ٣٢/٥، والكامل ٥٢٢/٧، والبداية والنهاية ٩٥/١١.

(٢) هو الأمير إبراهيم بن محمد: كما في: العيون والحدائق ج ٤ ق ١٦٥/١، وهو: عبدالله بن إبراهيم بن أحمد أبو العباس، كما في: الحلّة السّيراء لابن الآبار ١٧٤/١ الذي وصفه بأنه «كان شجاعاً بطلاً ذا بصر بالحروب والتدبير، عاقلاً أديباً عالماً، له نظر في الجدل وعناية باللغة والآداب..». وذكره ابن عذاري بكنيته ولم يُسمّه فقال: أبو العباس بن إبراهيم بن أحمد: «وأظهر التقشّف، والجلوس على الأرض، وإنصاف المظلوم، وجالس أهل العلم، وشاورهم، وكان لا

Dhahabi.TarikhIslam.Tadmuri/tiz21 – Maxim Romanov

Slide 47

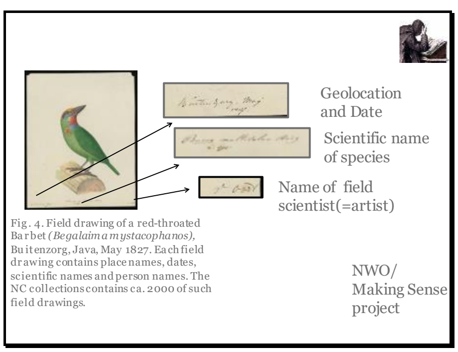
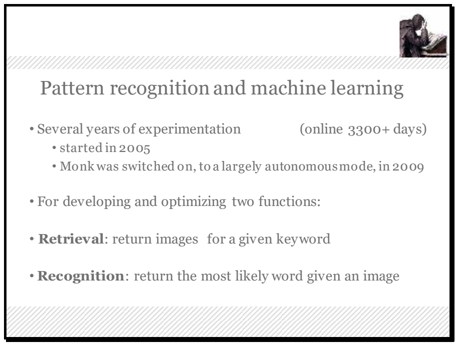
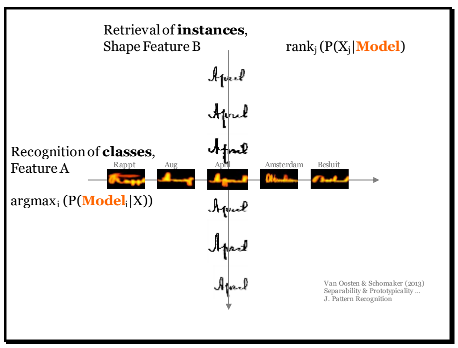


Fig. 4. Field drawing of a red-throated Barbet *(Begalaima mystacophanos)*, Buitenzorg, Java, May 1827. Each field drawing contains place names, dates, scientific names and person names. The NC collections contains ca. 2000 of such field drawings.

Slide 48

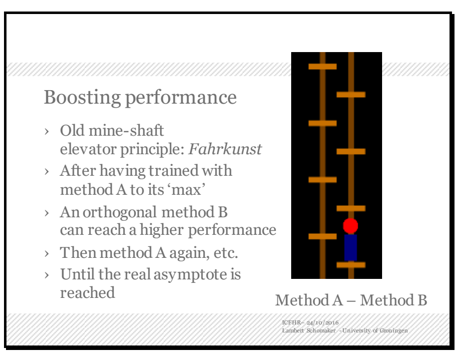


Slide 49



Contrary to expectation, a good classifier for recognition (in terms of recall and precision, e.g.) is not guaranteed to provide an a posteriori likelihood that is useful for intuitive ranking. For retrieval, other methods may be more applicable. The distance to a centroid is more informative in this respect than the distance to a separating boundary.

Slide 50

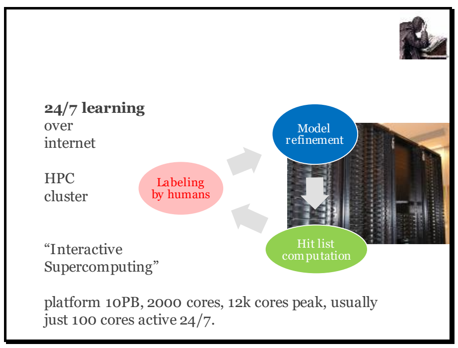


http://w w w.sciencedirect.com/science/article/pii/S0031320313003786

Jean-Paul van Oosten and Lambert Schomaker(2014).
Separability versus prototypicality in handw ritten w ord-image retrieval ,
Pattern Recognition, 47(3),
pp. 1031 - 1038,
issn = "0031-3203",
abstract = "Hit lists are at the core of retrieval systems. The top ranks are important, especially if user feedback is used to train the system. Analysis of hit lists revealed counter-intuitive instances in the top ranks for good classifiers. In this study, w e propose that tw o functions need to be optimised: (a) in order to reduce a massive set of instances to a likely subset among ten thousand or more classes, separability is required. How ever, the results need to be intuitive after ranking, reflecting (b) the prototypicality of instances. By optimising these requirements sequentially, the number of

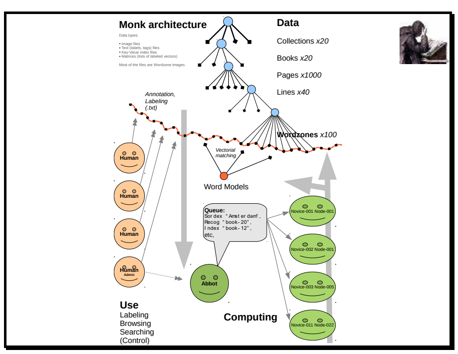
distracting images is strongly reduced, followed by nearest-centroid based instance ranking that retains an intuitive (low-edit distance) ranking. We show that in handwritten word-image retrieval, precision improvements of up to 35 percentage points can be achieved, yielding up to 100% top hit precision and 99% top-7 precision in data sets with 84 000 instances, while maintaining high recall performances. The method is conveniently implemented in a massive scale, continuously trainable retrieval engine, Monk. "

Slide 51



Users look at lines of text or at hit lists, a word model is computed, a new ranking is computed, presented to the user(s), and so on. In the early stage of the Monk development, Blue Gene (12k cores) was used for doing a grid search on optimal MLP configurations. However, centroid search proved to be much more convenient such that the current computational requirements are less demanding.

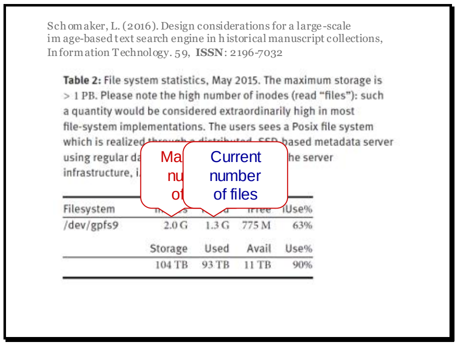
Slide 52



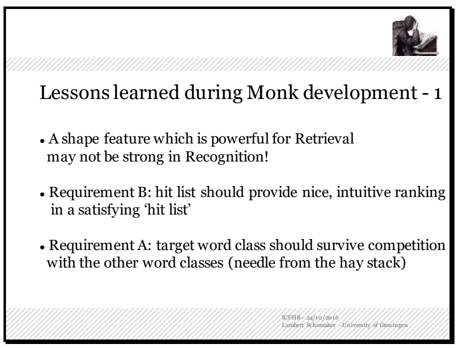
Labeling induces changes in the knowledge state, and leads to retraining, using a queing of jobs in HPC.

As another metaphor, the Monk system is like a blossoming tree presenting flowers (word zone candidates) to bees (human labelers). The large surface that is needed was provided by the IBM gpfs file system.

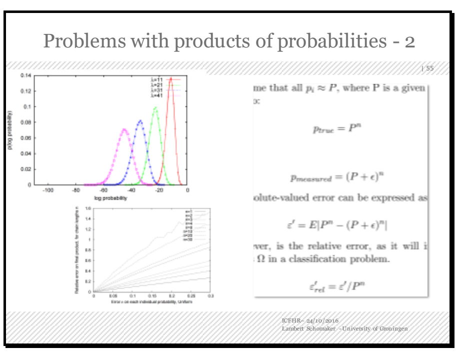
Slide 53



Slide 54

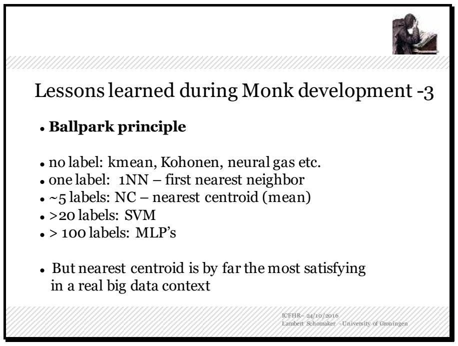


Slide 55

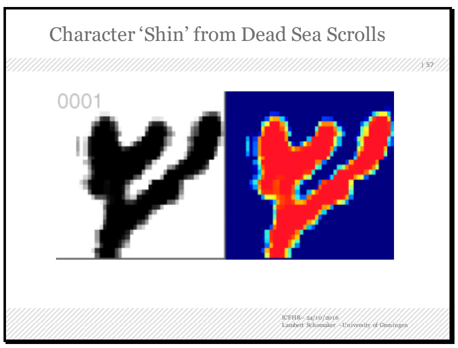


Draft can be found on arXiv, I am still working on this.

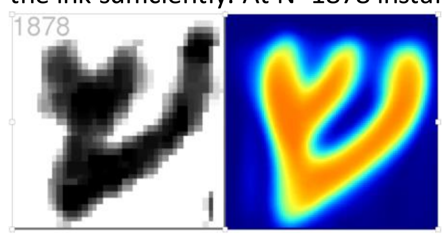
Slide 56



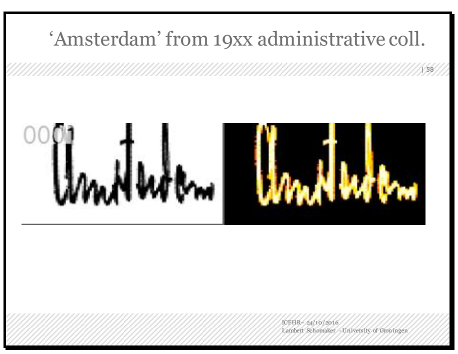
Slide 57



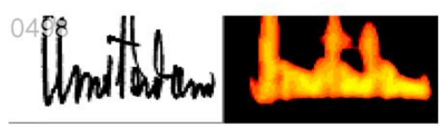
Movie of Shin converging to a stable probability landscape for ink. Normalisation on the basis of center of gravity and standard deviation of the radius, times a factor such as 2.5 to cover the ink sufficiently. At N=1878 instances:



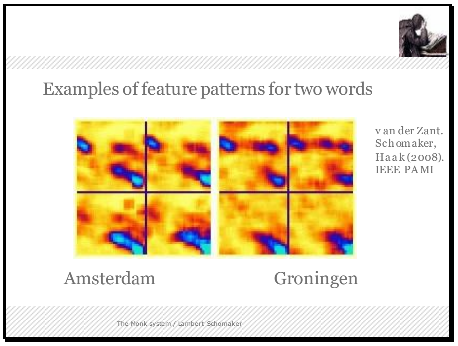
Slide 58



Movie for the word Amsterdam, much less samples but still converging to a stable mean. This principle yields an attractor, also in other feature spaces, of which the inertia prevents drift in an on-line (24/7) learning setup. At N=498:

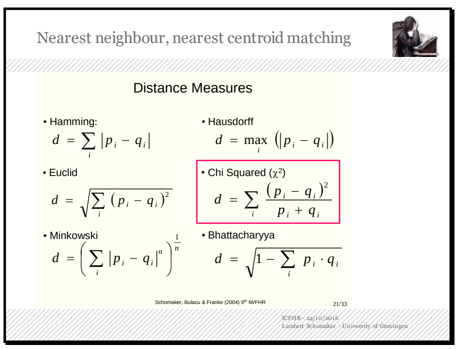


Slide 59



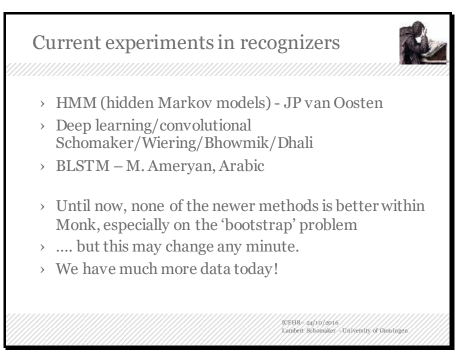
The paper in PAMI 2008 gave us the confidence that whole-word feature approaches can be very powerful. The Serre/Poggio neural network was a bit complicated and Matlab based, so I developed a more technical feature method for words in this period, that was generic enough to handle a wide range of handwritten scripts. Not only in averaged images but also in averaged feature spaces, big data or more than a hundred examples per class yield very stable models.

Slide 60

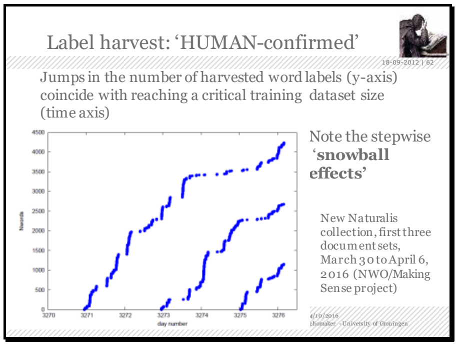


It is advisable to used Bhattacharrya or Chisquare on feature vectors that represent probability. I am sorry if that reduces the shine of your more complicated method. Up to 10% improvements in performance with respect to Euclidean or Manhattan can be observed in a wide range of pattern recognition problems.

Slide 61



Slide 62

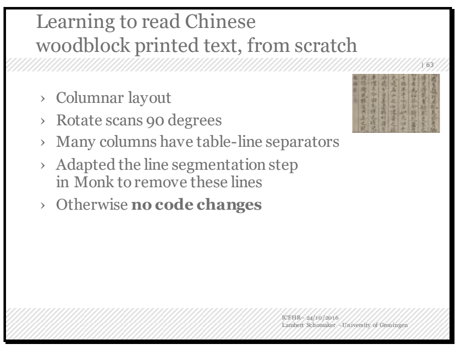


Each trace represents the harvest curve for a book

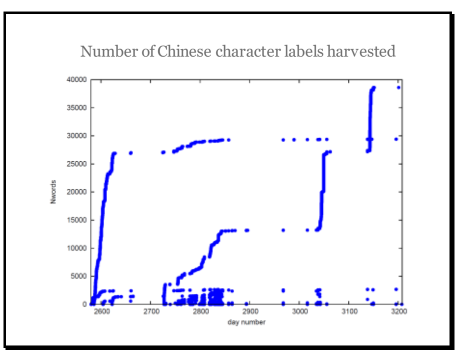
Please note phase transitions: at some point during label harvesting, the visual word models become successful in attracting new unseen instances and provide a clean hit list that is easy to confirm by the human users:

Phase transitions in the training!

Slide 63

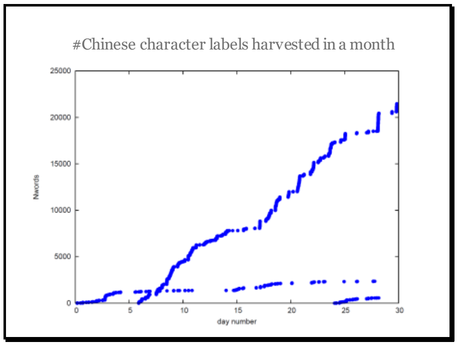


Slide 64

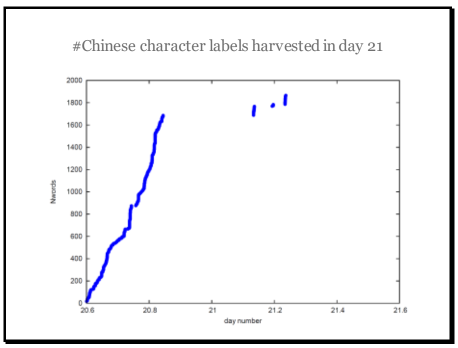


Woodblock-printed documents (higher curves) elicit more labels than handwriting, as can be expected. Still, this category is considered 'difficult' by many researchers in Chinese script recognition and by the users in the humanities.

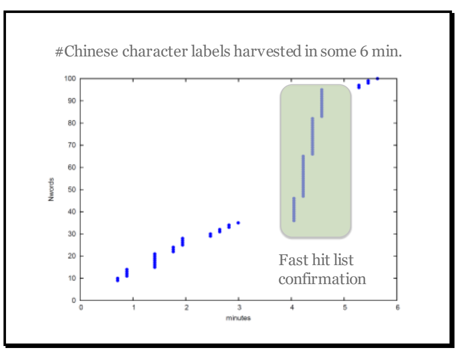
Slide 65



Slide 66

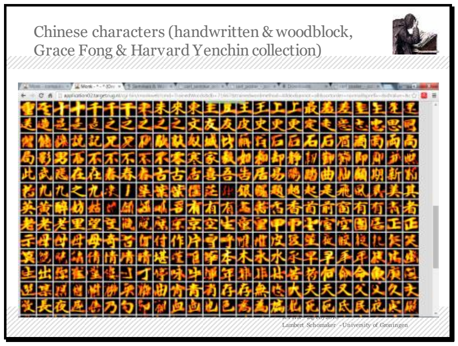


Slide 67



This slides shows a **phase transition** due to easy labeling of a correct hit list with about 60 correctly recognized and top-ranked instances.

Slide 68



Catalogue of mixed printed and written Chinese characters.

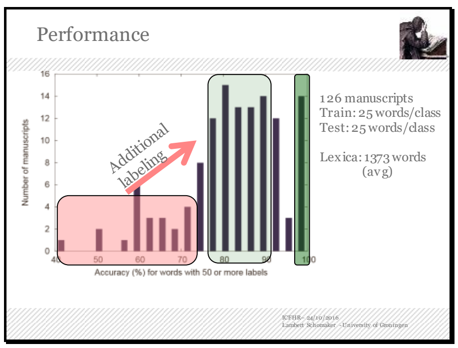
Slide 69

Baron is the target word. Words can be individually confirmed or per visible hit list as a whole.

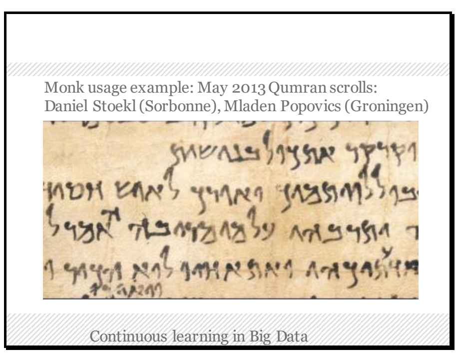
Slide 70

Slide 71

Slide 72



Slide 73



Slide 74

Qumran scrolls: 2400 photographs

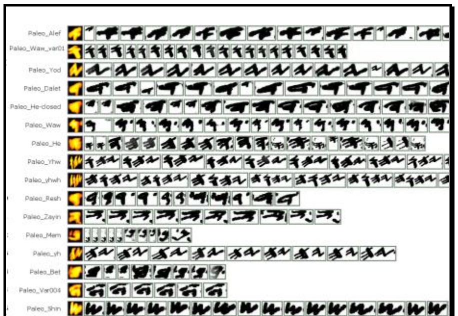
- Using **Monk** for character labeling
- With Daniel Stoekl and Mladen Popovics
- Using its 24/7 machine-learning cycle:
  - Label ➔ Train
    ➔ Label some More ➔ Train
    ➔ Easily label Many
- **Thousands** of characters 'mined' out of the Qumran collection of photographs in just **two weeks**, with very **little effort** in human labeling

Slide 75

Slide 76

Continuous learning in Big Data

Slide 77



Slide 78

Allograph harvest in Qumran collection

Slide 79

Conclusions

- Deep learning is a powerful concept
- But it is not enough, for building autonomous and intelligent agents
- The challenge is to design systems that handle unseen problems
- Part of deep-learning success may be just caused by the amount of data: better comparative evaluation is needed
- Iterative recognition and ranking works great!

- Engineered neural-network architectures vs
-

Engineered features: we're still not there, either way!

Slide 80

We consider hieratic script on papyrus the most difficult material in Monk at the moment. Hieroglyphs are easier to recognize, but for interpretation a stochastic grammar needs to be trained (2D).