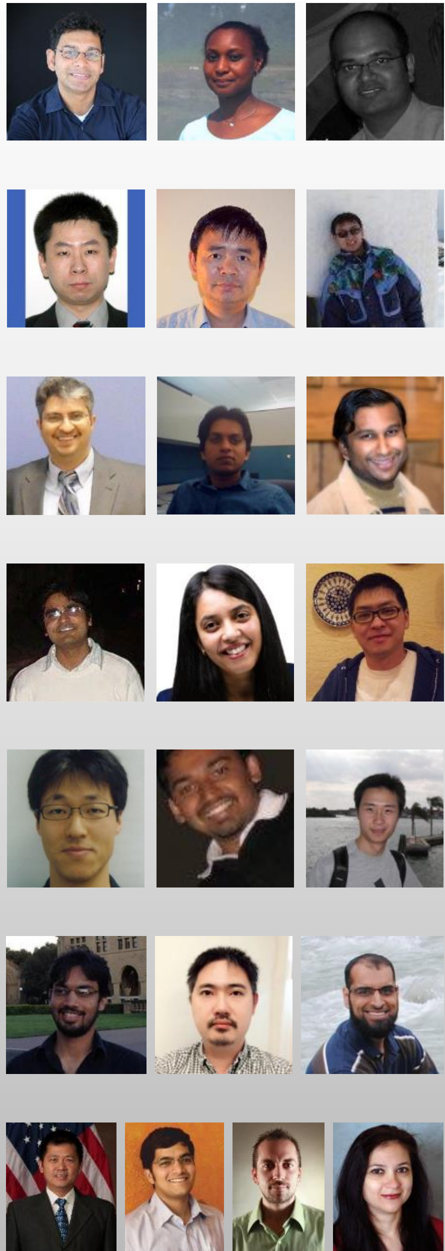
NAVIGATING SCIENTIFIC LITERATURE

A HOLISTIC PERSPECTIVE

Venu Govindaraju



PATTERN RECOGNITION

- Towards a Globally Optimal Approach for Learning Deep Unsupervised Models
- Organizing Multiple Experts for Efficient Pattern Recognition
- Active Pattern Recognition Using Genetic Programming
- A Complexity Framework for Combination of Classifiers in Verification and Identification Systems
- Image Processing using Ontology Concepts for Image Segmentation
- Language Motivated Approaches for Human Action Recognition and Spotting

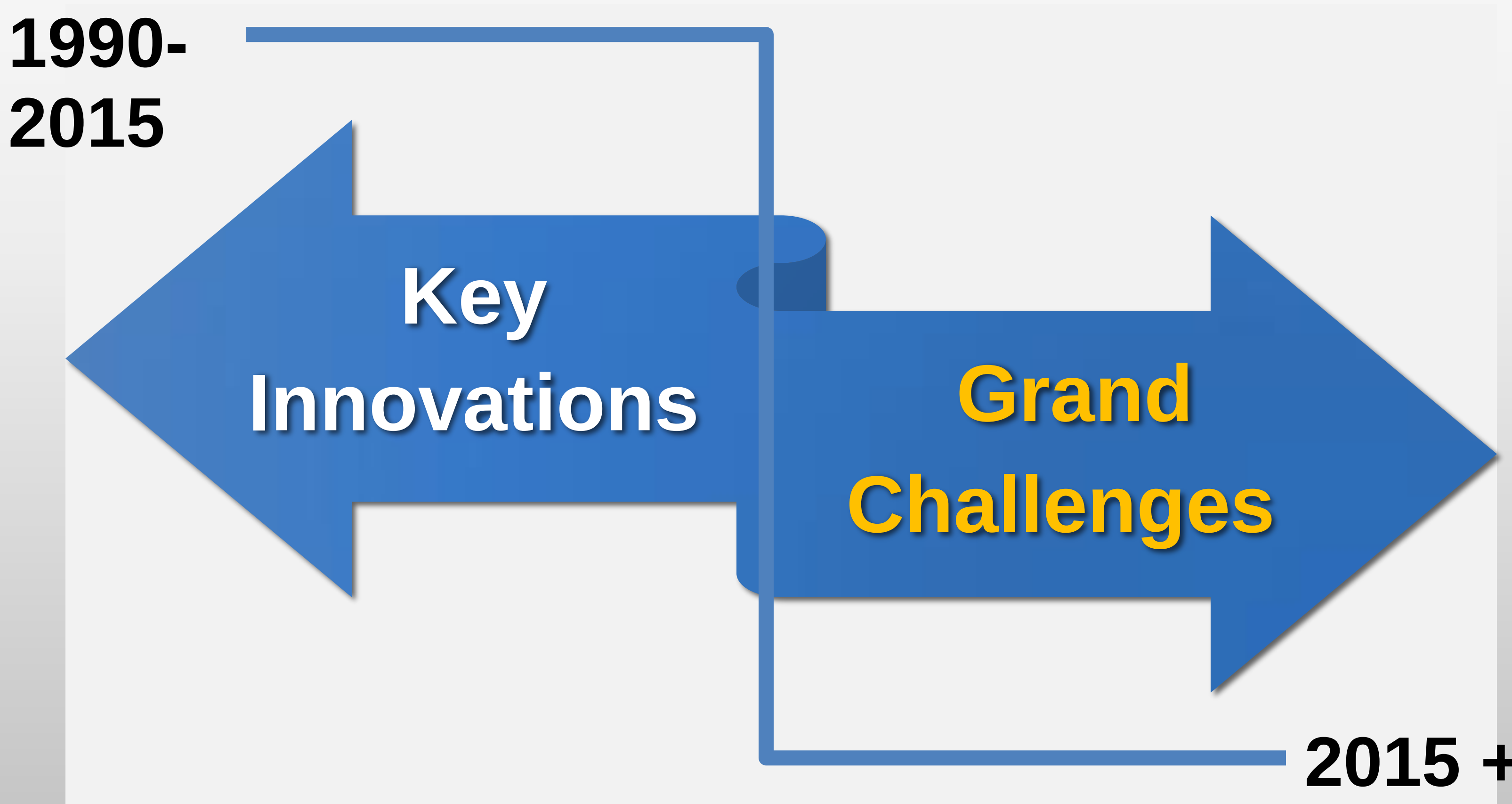
DOCUMENT ANALYSIS

- Intrusion Detection using Spatial Information and Behavioral Biometrics
- Integrating Minutiae Based Fingerprint Matching with Local Correlation Methods
- Integrating Facial Expressions and Skin Texture in Dace Recognition
- Stochastic Modeling of High-level Structures in Handwritten Word Recognition
- Statistical Techniques for Efficient Indexing and Retrieval of Document Images
- Probabilistic Random Field based Text Identification
- Enhancing Cyber Security through the use of Synthetic Handwritten CAPTCHAs
- Language Models and Automatic Topic Categorization for Information Retrieval in Handwritten Documents
- Methods for Biomedical Image Content Extraction Toward Improved Multimodal Retrieval of Biomedical Articles
- A Novel Multi-sample Fusion Methodology for Improving Biometric Recognition
- Enhancement and Retrieval of Low Quality Handwritten Documents
- A Stochastic Framework for Font Independent Devanagari OCR
- A Semi Supervised Framework for Handwritten Document Analysis
- Bayesian Background Models for Retrieval of Handwritten Documents
- Accents in Handwriting: A Hierarchical Bayesian Approach to Handwriting Analysis
- Hierarchical and Dynamic-Relational Models for Handwriting Recognition

BIOMETRICS

- Multilingual Word Spotting in Offline Handwritten Documents
- A Framework for Fingerprint Enhancement and Feature Detection
- Minutia-Based Partial Fingerprint Recognition
- Sequential Pattern Classification without Explicit Feature Extraction
- Automatic Recognition of Handwritten Medical Forms for Search Engines
- Exploiting the Gap between Human and Machine Abilities in Handwriting Recognition for Web Security Applications
- Face Modeling and Biometric Anti-spoofing using Probability Distribution Transfer Learning
- A Framework for Efficient Fingerprint Identification using a Minutiae Tree

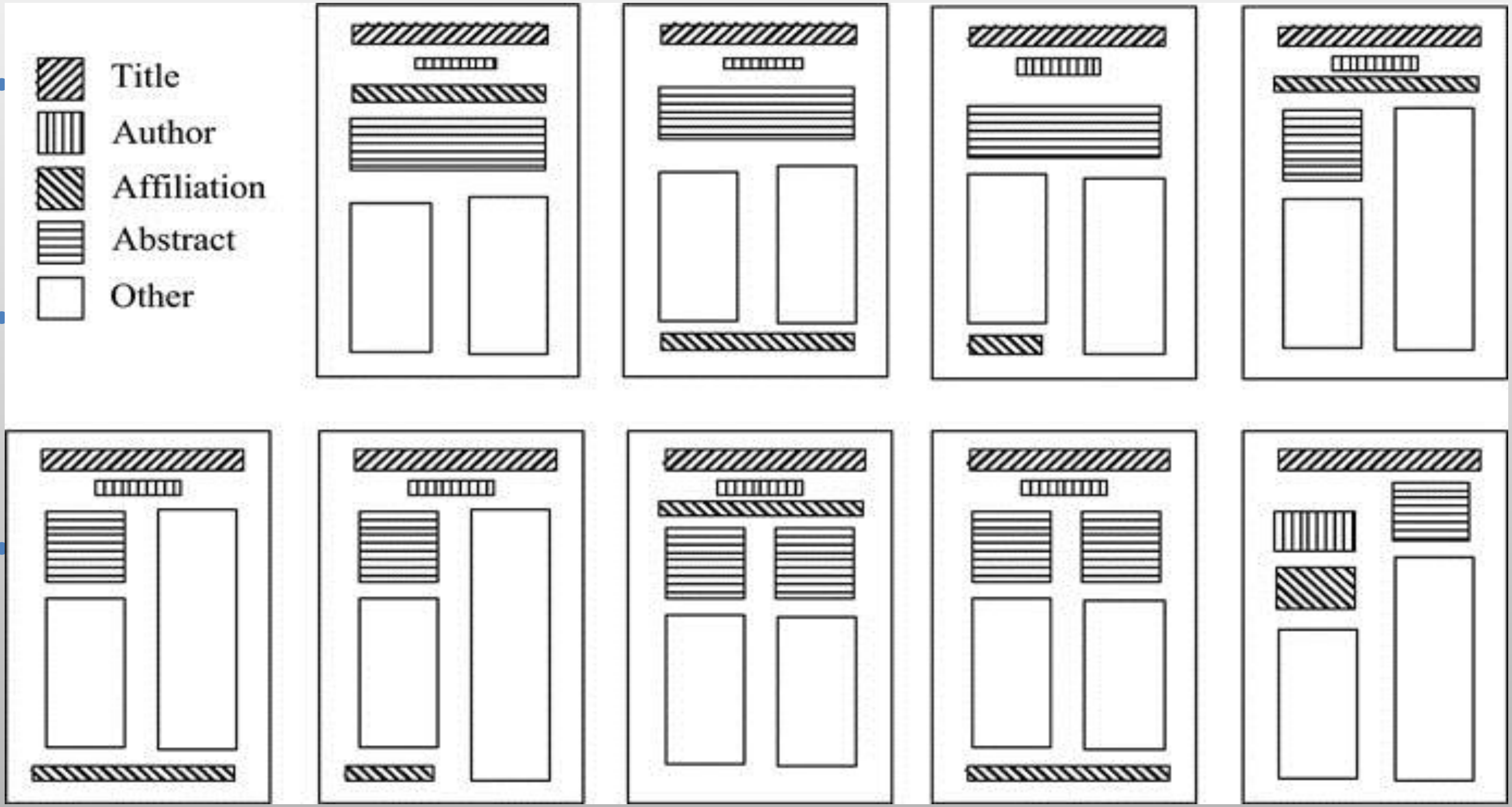

handwriting
recognition
framework
retrieval
identification
fingerprint
documents
models
using
image
pattern
efficient
feature
information
hierarchical
detection
biometric
stochastic
analysis
minutiae
language
based
approach
modeling
biomedical
extraction
security
automatic
handwritten
methods
bayesian
integrating
document
enhancement
word
human
spotting
learning


1990-
2015
Key
Innovations
Grand
Challenges
2015 +

Old Order - DIA

UW English Document Image Database

(Phillips, Technical report, 1996, citations: 29)



Scientific Process

nanos gigantum humeris insidentes

1676 letter of Isaac Newton: “ If I have seen further it is by standing on the shoulders of giants.”

Past/Current → **Future**

LEARN **DISCOVER**

STATE OF THE ART

DATA

Data
- Numbers
- Symbols

Information
- Processed data
- Who/What/When

Knowledge
- Insights
- Theory
- Experiment
- Framework

Understanding
- Analysis
- Meaning

Wisdom
- Judgment

KNOWLEDGE

When knowledge becomes data

Source (4Vs of Big Data) : IBM

Scientific Literature

- 2009 estimate: 50 million articles; 28 thousand journals
- 1.8M articles added every year.

PubMed www.pubmed.gov A service of the U.S. National Library of Medicine and the National Institutes of Health	23 million articles (Just biomedical literature)	**Volume**
Scopus The largest abstract and citation database of peer-reviewed literature.	45 million articles	**Variety**
THOMSON REUTERS WEB OF KNOWLEDGE℠	40 million articles	
Google scholar	Unknown (peer reviewed only)	**Veracity**

BIG DATA!

Roughly, papers double every 10-15 years !

Velocity

[Meadows, 1998, p.16]

Big Data Side Effects

Challenges for ICDAR Community

4Vs

4Vs		4Rs
Volume	→	References
Velocity	→	Reinvention
Variety		Replicability
Veracity	→	Reputation

4Rs

References
Volume Challenge

The Royal Society (1662)
First journal : *Journal of Philosophical Transactions (1665)*
Le Journal des Sçavans (1665)

Scientists believe they are only reading 40% of the relevant literature.
Faraday reported the same problem already in 1826 !!

[Meadows, 1998], page 211, and Faraday is quoted on page 19

"50% of papers are never read by anyone other than their authors, referees and journal editors."….

"90% of papers are never cited …"

[Smithsonian.com, 2007 study]

Reinvention

Velocity Challenge

``in some disciplines it is occasionally easier to repeat an experiment than it is to determine that the experiment has already been done." [Garvey, 1979, p.8].

Replicability

Veracity Challenge

- ***Nullius in verba***

  “On the word of no one" or "Take nobody's word for it"

SCIENCE is in crisis, just when we need it most. Two years ago, C. Glenn Begley and Lee M. Ellis reported in Nature that they were able to replicate only six out of 53 “landmark” cancer studies. Scientists now worry that many published scientific results are simply not true.

NY Times 2014

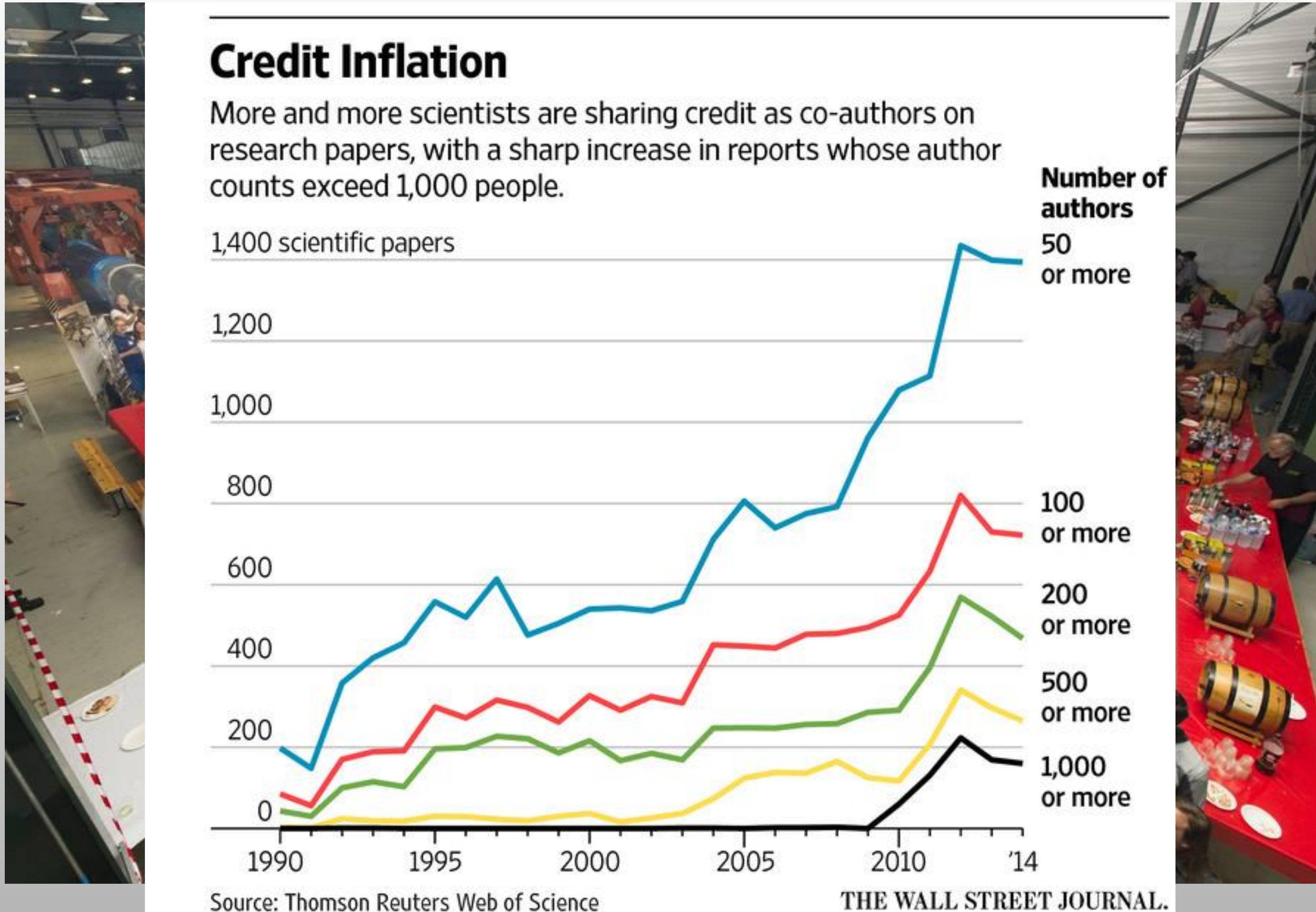
Reputation
Veracity Challenge

- How Many Scientists Does It Take to Write a Paper?

Scientific journals see a spike in number of contributors; 24 pages of alphabetized co-authors.

The Wall Street Journal, August 10, 2015

Challenges


Credit Inflation
More and more scientists are sharing credit as co-authors on research papers, with a sharp increase in reports whose author counts exceed 1,000 people.
Number of authors
1,400 scientific papers
1,200
1,000
800
600
400
200
0
50 or more
100 or more
200 or more
500 or more
1,000 or more
1990
1995
2000
2005
2010
'14
Source: Thomson Reuters Web of Science
THE WALL STREET JOURNAL.

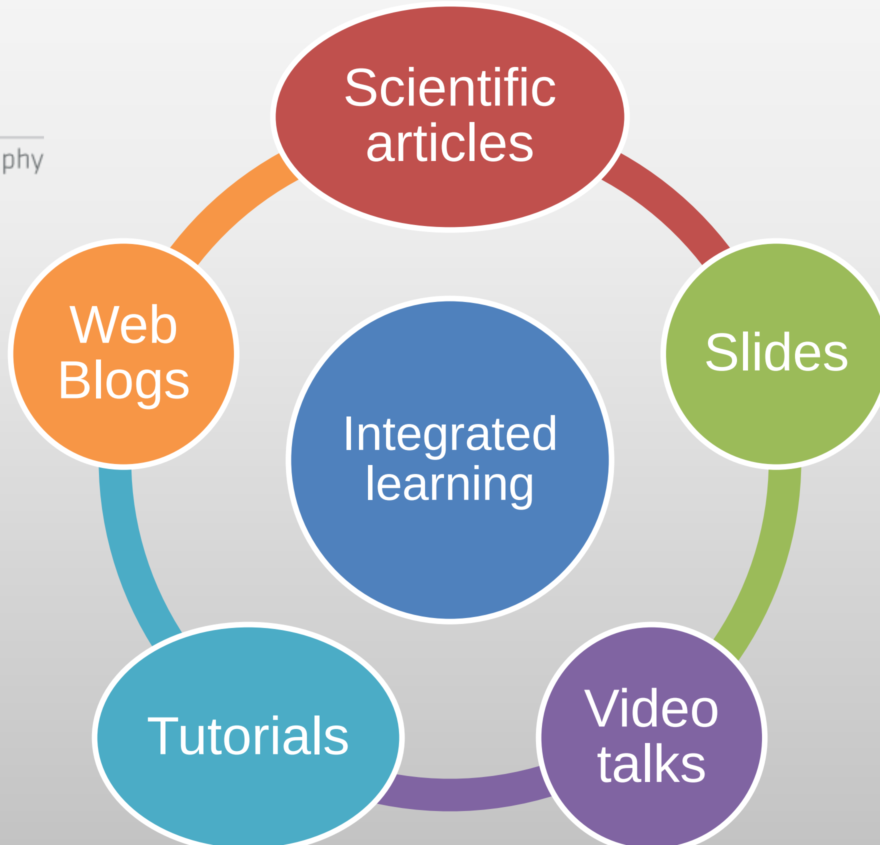
GRAND CHALLENGE

BIG DATA

VOLUME	VELOCITY	VARIETY	VERACITY
			
References	Reinvention		Replicability
Cognitive burden	Reinventing wheel		Authenticity

Addressing the Cognitive Burden

Volume



Addressing Reinventing the wheel ?

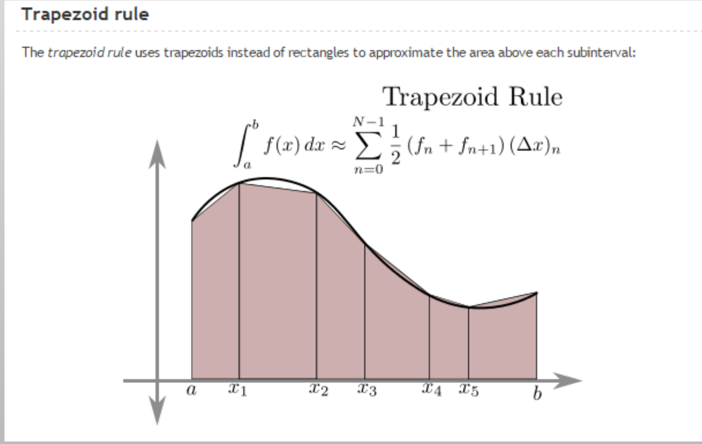
Velocity

- **Least square with linear constraints:** one type of quadratic program in mathematics

$$\begin{array}{ll} \text{minimize} & \|Ax-b\|_2^2 \\ \text{subject to} & l_i \le x_i \le u_i, i = 1, \dots, n \end{array}$$

- **Isotonic regression:** in statistics

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n} w_i(x_i - a_i)^2 \\ \text{subject to} & x_i \ge x_j for(i,j) \in E \end{array}$$



Trapezoid rule: calculus 17th century

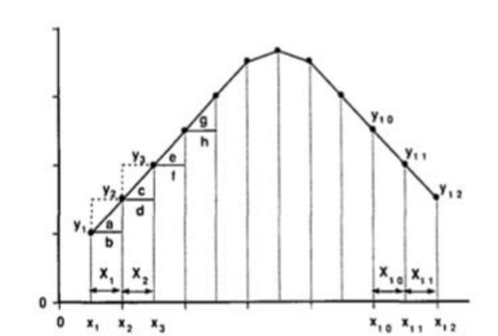
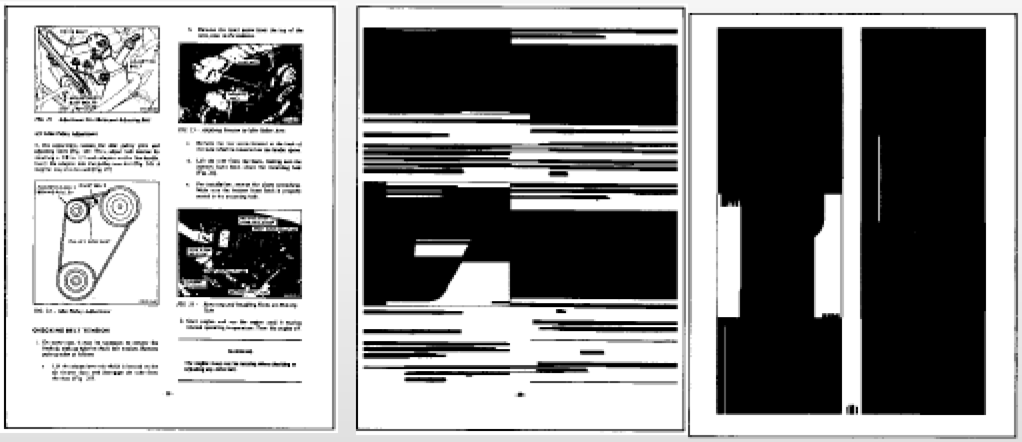


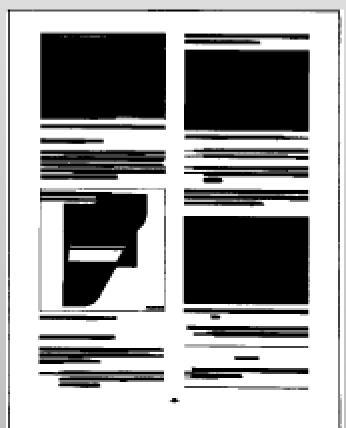
Figure 1—*Total area under the curve is the sum of individual areas of triangles a, c, e, and g and rectangles b, d, f, and h.*

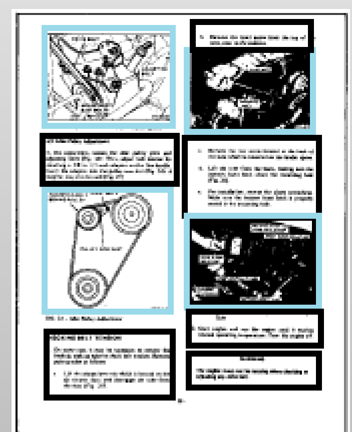
Tai's Model, 1994, 254 citations

Addressing Replicability

Velocity







IBM Journal 1982

- **Dataset** – UNLV/ISRI
  - 64 pages, 6796 blocks
- **Heuristics parameters**
  - Vertical: 15 pixels
  - Horizontal: 50/30 pixels
- **Classes:**
  - Text, Table, Caption, Figs
- **Classifier:**
  - Support Vector Machines
- **Accuracy:**
  - 91.73 % at block level

Addressing Authenticity

Veracity

Datasets

- Public
- Benchmark
- Published

Reputation

- Authors
- Lab
- Journal

Experiments

- Comparative results
- CODE available !!

Citations

- Only Counts ?

Veracity

All citations are not equal

- **Which citation is more trustworthy?**

object classification [3]. However, the same level of success has not been obtained for generative tasks, despite numerous efforts [13, 24, 28].

Table 2. Results on MNIST dataset.

Method	Paper	Error rate[%]
CNN	[32]	0.40
CNN	[26]	0.39
MLP	[5]	0.35
CNN committee	[6]	0.27
MCDNN	this	**0.23**

Sentiment analysis:
Targeted NLP

We are unable to replicate the results from paper [14]

area of speech recognition, with breakthrough results (Dahl *et al.*, 2010; Deng *et al.*, 2010; Seide *et al.*, 2011a; Mohamed *et al.*, 2012; Dahl *et al.*, 2012; Hinton *et al.*, 2012) obtained by several academics as well as researchers at industrial labs

Veracity

Dataset linkages

MNIST: 60k training, 10k testing images
"Gradient-based learning applied to document recognition", Lecun et al 1998 (Citations: 3547)

Ciresan et al. 2012

descent with an annealed learning rate. During training, images are continually translated, scaled and rotated (even elastically distorted in case of characters), whereas only the original images are used for validation. Training ends once the validation error is zero or when the learning rate reaches its predetermined minimum. Initial weights are drawn from a uniform random distribution in the range $[-0.05, 0.05]$.

Table 2. Results on MNIST dataset.

Method	Paper	Error rate[%]
CNN	[32]	0.40
CNN	[26]	0.39
MLP	[5]	0.35
CNN committee	[6]	0.27
MCDNN	this	**0.23**

Training on automatically augmented dataset:
"During training the digits are randomly distorted …
The MCDNN has a very low 0.23% error rate"

such as unsupervised pre-training [29, 24, 2, 10] or carefully prewired synapses [27, 31].

(3) The DNN of this paper (Fig. 1a) have 2-dimensional layers of winner-take-all neurons with overlapping receptive fields whose weights are shared [19, 1, 32, 7]. Given some input pattern, a simple max pooling technique [27] determines winning neurons by partitioning layers into quadratic regions of local inhibition, selecting the most active neuron of each region. The winners of some layer represent a smaller, down-sampled layer with lower resolution, feeding the next layer in the hierarchy. The approach is inspired by Hubel and Wiesel's seminal work on the cat's primary visual cortex [37], which identified orientation-selective *simple cells* with overlapping local receptive fields and *complex cells* performing down-sampling-like operations [15].

(4) Note that at some point down-sampling automatically leads to the first 1-dimensional layer. From then on, only trivial 1-dimensional winner-take-all regions are possible, that is, the top part of the hierarchy becomes a standard multi-layer perceptron (MLP) [36, 18, 28]. Receptive fields and winner-take-all regions of our DNN often are (near-)minimal, e.g., only 2x2 or 3x3 neurons. This results in (near-)maximal depth of layers with non-trivial (2-dimensional) winner-take-all regions. In fact, insisting on minimal 2x2 fields automatically defines the entire deep architecture, apart from the number of different convolutional kernels per layer [19, 1, 32, 7] and the depth of the plain MLP on top.

(5) Only winner neurons are trained, that is, other neurons cannot forget what they learnt so far, although they may be affected by weight changes in more peripheral layers. The resulting decrease of synaptic changes per time interval corresponds to biologically plausible reduction of energy consumption. Our training algorithm is fully online, i.e. weight updates occur after each gradient computation step.

(6) Inspired by microcolumns of neurons in the cerebral cortex, we combine several DNN columns to form a Multi-column DNN (MCDNN). Given some input pattern, the predictions of all columns are averaged:

$$y^i_{MCDNN} = \frac{1}{N} \sum_j^{\#columns} y^i_{DNN_j} \quad (1)$$

where i corresponds to the i^{th} class and j runs over all DNN. Before training, the weights (synapses) of all columns are randomly initialized. Various columns can be trained on the same inputs, or on inputs preprocessed in different ways. The latter helps to reduce both error rate and number of columns required to reach a given accuracy. The MCDNN architecture and its training and testing pro-

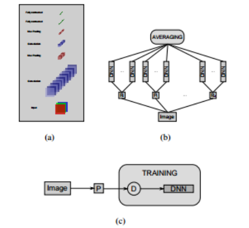


Figure 1. (a) DNN architecture. (b) MCDNN architecture. The input image can be preprocessed by $P_0 - P_{n-1}$ blocks. An arbitrary number of columns can be trained on inputs preprocessed in different ways. The final predictions are obtained by averaging individual predictions of each DNN. (c) Training a DNN. The dataset is preprocessed before training, then, at the beginning of every epoch, the images are distorted (D block). See text for more explanations.

3. Experiments

We evaluate our architecture on various commonly used object recognition benchmarks and improve the state-of-the-art on all of them. The description of the DNN architecture used for the various experiments is given in the following way: 2x48x48-100C5-MP2-100C5-MP2-100C4-MP2-300N-100N-6N represents a net with 2 input images of size 48x48, a convolutional layer with 100 maps and 5x5 filters, a max-pooling layer over non overlapping regions of size 2x2, a convolutional layer with 100 maps and 4x4 filters, a max-pooling layer over non overlapping regions of size 2x2, a fully connected layer with 300 hidden units, a fully connected layer with 100 hidden units and a fully connected output layer with 6 neurons (one per class). We use a scaled hyperbolic tangent activation function for convolutional and fully connected layers, a linear activation function for max-pooling layers and a softmax activation function for the output layer. All DNN are trained using on-line gradient descent with an annealed learning rate. During training, images are continually translated, scaled and rotated (even elastically distorted in case of characters), whereas only the original images are used for validation. Training ends once the validation error is zero or when the learning rate reaches

3.1. MNIST

The original MNIST digits [19] are normalized such that the width or height of the bounding box equals 20 pixels. Aspect ratios for various digits vary strongly and we therefore create six additional datasets by normalizing digit width to 10, 12, 14, 16, 18, 20 pixels. This is like seeing the data from different angles. We train five DNN columns per normalization, resulting in a total of 35 columns for the entire MCDNN. All 1x29x29-20C4-MP2-40C5-MP3-150N-10N DNN are trained for around 800 epochs with an annealed learning rate (i.e. initialized with 0.001 multiplied by a factor of 0.993/epoch until it reaches 0.00003). Training a DNN takes almost 14 hours and after 500 training epochs little additional improvement is observed. During training the digits are randomly distorted before each epoch (see Fig. 2a for representative characters and their distorted versions [7]). The internal state of a single DNN is depicted in Figure 2b, where a particular digit is forward propagated through a trained network and all activations together with the network weights are plotted.

Results of all individual nets and various MCDNN are summarized in Table 1. MCDNN of 5 nets trained with the same preprocessor achieve better results than their constituent DNNs, except for original images (Tab. 1). The MCDNN has a very low 0.23% error rate, improving state of the art by at least 34% [5, 7, 25] (Tab. 2). This is the first time an artificial method comes close to the ≈0.2% error rate of humans on this task [21]. Many of the wrongly classified digits either contain broken or strange strokes, or have wrong labels. The 23 errors (Fig. 2c) are associated with 20 correct second guesses.

We also trained a single DNN on all 7 datasets simultaneously which yielded worse result (0.52%) than both MCDNN and their individual DNN. This shows that the improvements come from the MCDNN and not from using more preprocessed data.

Table 2. Results on MNIST dataset.

Method	Paper	Error rate[%]
CNN	[32]	0.40
CNN	[26]	0.39
MLP	[5]	0.35
CNN committee	[6]	0.27
MCDNN	this	**0.23**

How are the MCDNN errors affected by the number of preprocessors? We train 5 DNNs on all 7 datasets. A MCDNN 'y out-of-7' (y from 1 to 7) averages $5y$ nets trained on y datasets. Table 3 shows that more preprocessing results in lower MCDNN error.

We also train 5 DNN for each odd normalization, i.e. W11, W13, W15, W17 and W19. The 60-net MCDNN performs (0.24%) similarly to the 35-net MCDNN, indicating that additional preprocessing does not further improve recognition.

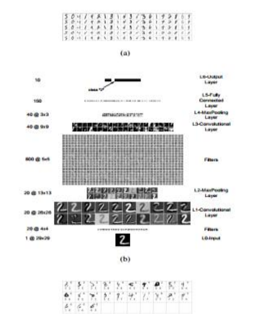


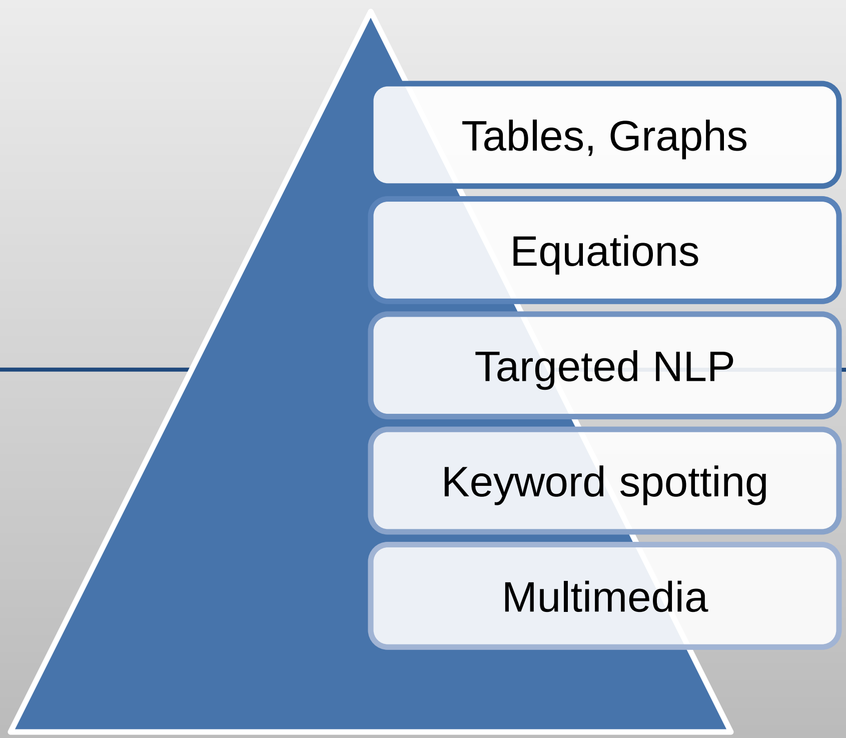
Figure 2. (a) Handwritten digits from the training set (top row) and their distorted versions after each epoch (second to fifth row). (b) DNN architecture for MNIST. Output layer not drawn to scale; weights of fully connected layers not displayed. (c) The 23 errors of the MCDNN, with correct label (up right) and first and second best predictions (down left and right).

Table 3. Average test error rate [%] of MCDNN trained on y preprocessed datasets.

y	# MCDNN	Average Error[%]
1	7	0.33±0.07
2	21	0.27±0.02
3	35	0.27±0.02
4	35	0.26±0.02
5	21	0.25±0.01
6	7	0.24±0.01
7	1	0.23

We conclude that MCDNN outperform DNN trained on the same data, and that different preprocessors further decrease the error rate.

OUR NEXT FRONTIER



Tables Analysis

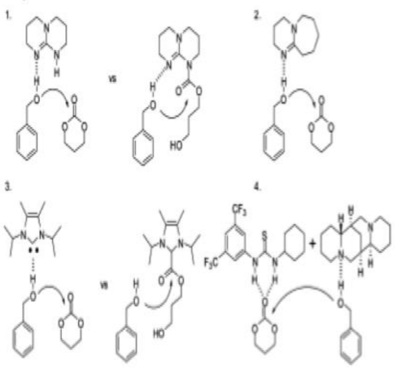
4Vs

Table 2. Characterization of Macroinitiators and PTMC Block Copolymers

macroinitiator	[TMC]/[MI][a]	conversion (%)	DP_{PTMC}[b]	M_n[c]	PDI	T_g[d]
PEO_{110}-OH[e]	50	75	40	9500	1.03	−34 °C (PTMC) 52 °C (T_m PEO)
PS_{80}-OH[f]	50	>99	52	21900	1.08	−32 °C (PTMC) 91 °C (PS)
$PDMA_{70}$-OH[g]	50	>99	45	14800	1.06	−12 °C
$PMMA_{140}$-OH[h]	100	>99	102	30800	1.11	−36 °C (PTMC) 107 °C (PMMA)
$P2VP_{90}$-OH[i]	50	>99	50	20200	1.09	−36 °C (PTMC) 77 °C (P2VP)

[a] Targeted degree of polymerization. [b] Experimentally determined degree of polymerization by ^{1}H NMR. [c] Obtained by GPC in THF. [d] Scan rate of 10 °C/min, second heating run. [e] Poly(ethylene oxide) (DP = 125; M_n = 5 kg mol^{-1}; and PDI = 1.03). [f] Polystyrene (DP = 80; M_n = 8.3 kg mol^{-1}; and PDI = 1.07). [g] Poly(*N*,*N*-dimethylacrylamide) (DP = 70; M_n = 7.1 kg mol^{-1}; and PDI = 1.08). [h] Poly(methyl methacrylate) (DP = 140; M_n = 14.5 kg mol^{-1}; and PDI = 1.12). [i] Poly(2-vinylpyridine) (DP = 90; M_n = 9.2 kg mol^{-1}; and PDI = 1.06).

Scheme 4. Various Catalysts and Their Mechanism in the Polymerization of TMC[a]



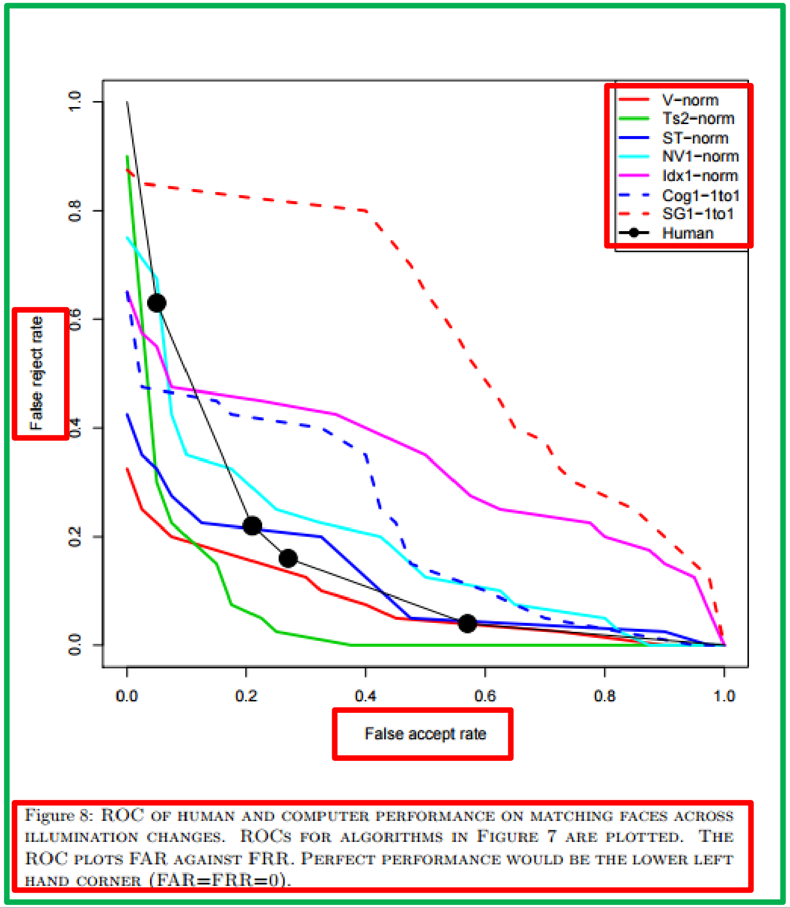
[a] (1) 1, (2) **2** or **3**, (3) **4** or **5**, and (4) **6** or **7**.

such. Technical quality 4,4′-bisazo(4-cyanopentan-1-ol) (containing ~30% water by weight) was bought from Langfang Hawk Ltd. (China) and dissolved in methylene chloride, and the organic layer was separated and dried over $MgSO_4$, filtered, and evaporated in vacuo. The resulting solid was recrystallized twice from methylene chloride/hexanes yielding off-white crystals. The hydroxy-functionalized alkoxyamine, 2,2,5-trimethyl-3-(4′-*p*-hydroxymethylphenylethoxy)-4-phenyl-3-azahexane,[35] for NMP and the hydroxy-functionalized RAFT-agent, 4-cyano-4-((thiobenzoyl)sulfanyl)-pentan-1-ol,[36] were prepared according to literature procedures. Hydroxyfunctional PS and PDMA were prepared by NMP, whereas hydroxyfunctional PMMA[37] and P2VP[38] were prepared by RAFT polymerization according to literature procedures. Macroinitiators from NMP and RAFT polymerizations as well as commercially avaible poly(ethylene oxide) (Fluka) were dried in a vacuum oven and further dried by coevaporation of dry distilled toluene 3 times before transferring to a glovebox for assembly of the ROP reaction. ^{1}H− NMR spectra were obtained on a Bruker Avance 400

- Extract data
- Compare data
- Headers
- Merged columns and rows
- Color
- Caption?
- Reference in text?
- Such tables in other articles?

Graphs Analysis

4Vs



- Type of graph? x-y plot, bar graph etc.
- Labels on the axes?
- Number of curves?
- Color?
- Curves intersect?
- Actual data points?
- Legend?
- Figure number?
- Caption?
- Reference in text?
- Such graphs in other articles?

Equations Analysis

Velocity

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

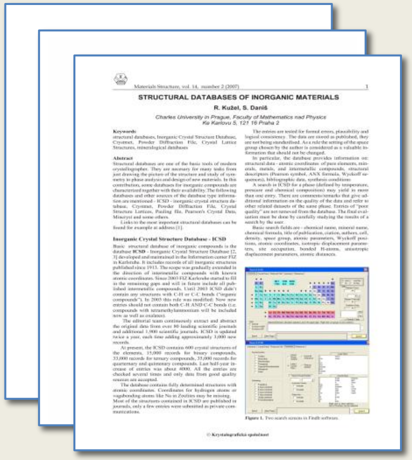
Domain awareness
Matrix representation
Operators
Symbols

Document awareness
Dependent variables independent variables regression coefficients, error

$$h_i = \beta_1 t_i + \beta_2 t_i^2 + \varepsilon_i$$

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$



Query Interface

Face Recognition FAR (0.0-0.2) vs FRR

☑ CVPR ☐ Science ☐ Nature

☑ Advanced Search Options

☑ X-Y Plots ☐ Tables ☑ Figures

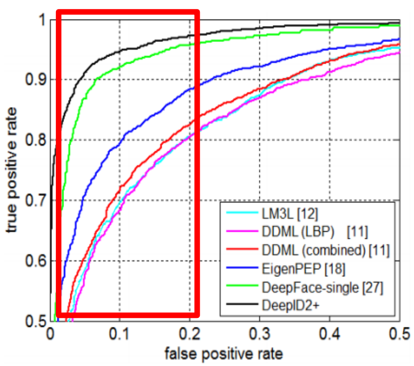


Figure 5: ROC of face verification on YouTube Faces. Best viewed in color.

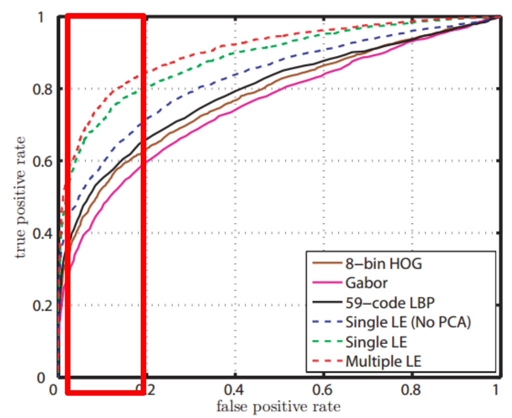


Figure 7. ROC curve comparison between our LE descriptors and existing descriptors.

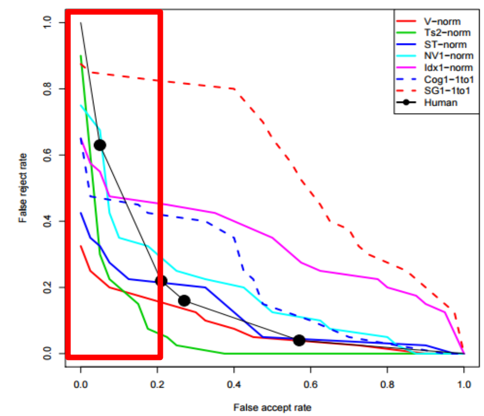


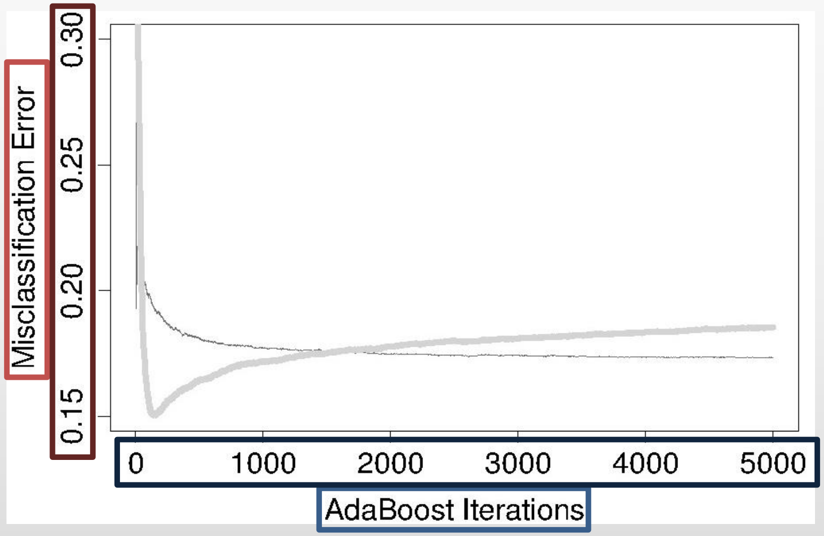
Figure 8: ROC of human and computer performance on matching faces across illumination changes. ROCs for algorithms in Figure 7 are plotted. The ROC plots FAR against FRR. Perfect performance would be the lower left hand corner (FAR=FRR=0).

Original Paper

Original Paper

Original Paper

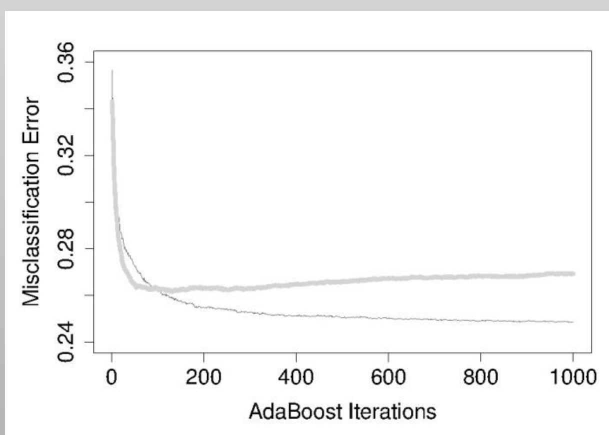
Query:

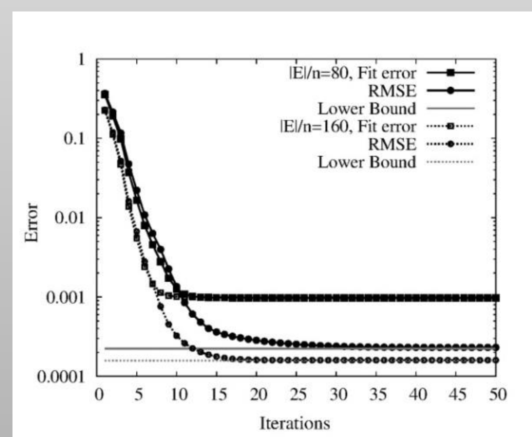


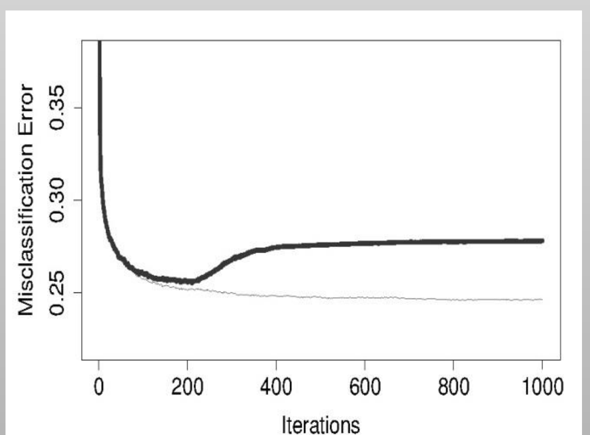
Line graph

- X Axis
  - Label: AdaBoost iterations
  - Range: 0-5000
  - Unit: -
- Y Axis
  - Label: Misclassification Error
  - Range: 0.15-0.30
  - Unit: -
- Legend: -
- Number of lines: 2

Retrieve similar graphs





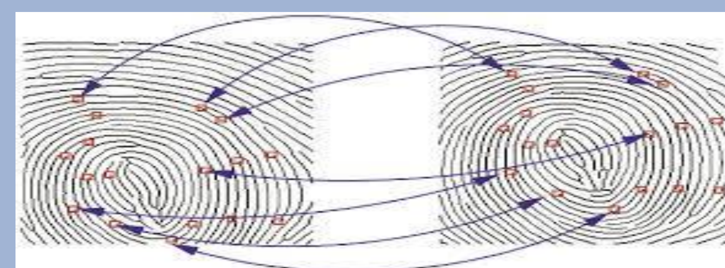


Holistic View

Linkages

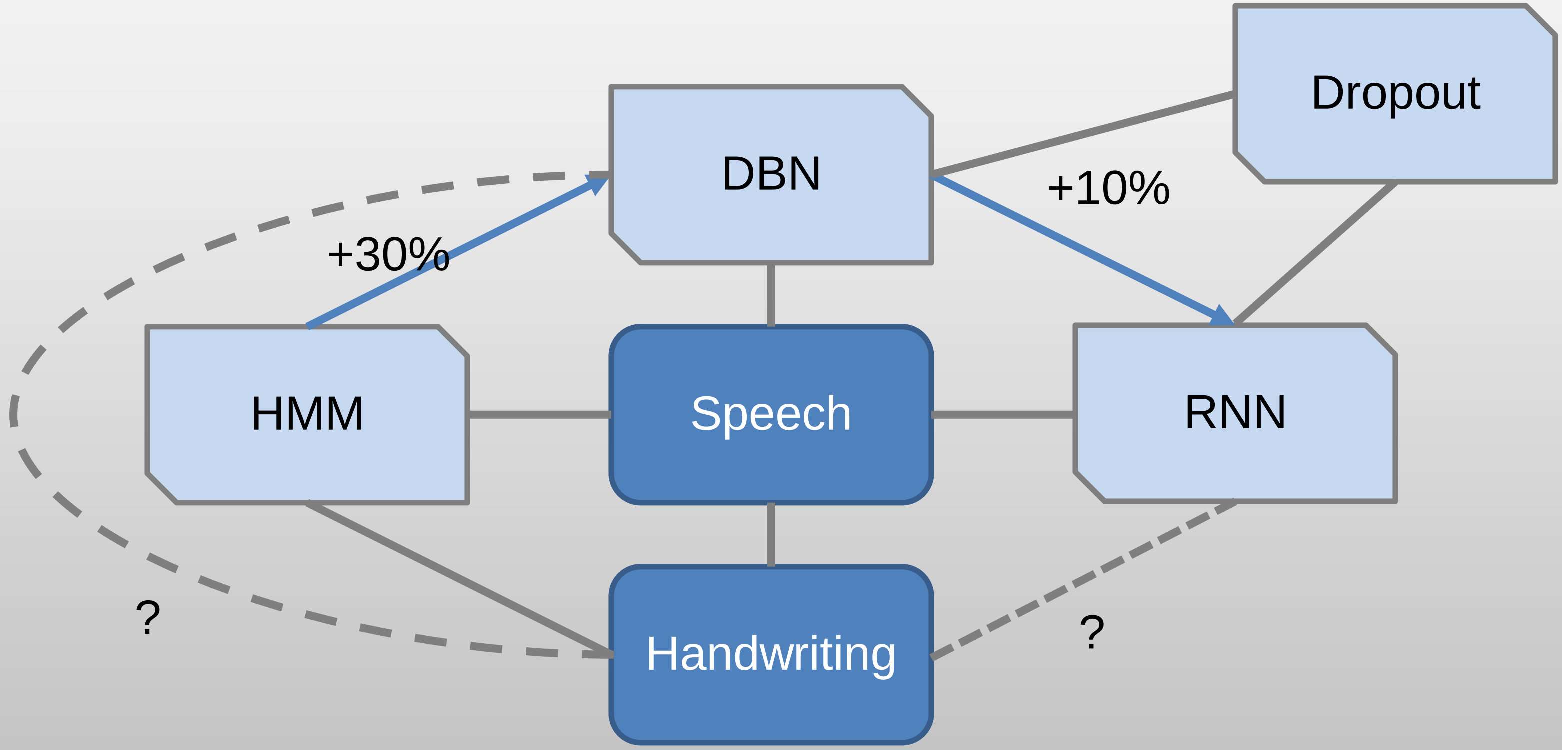
SUPERVISED

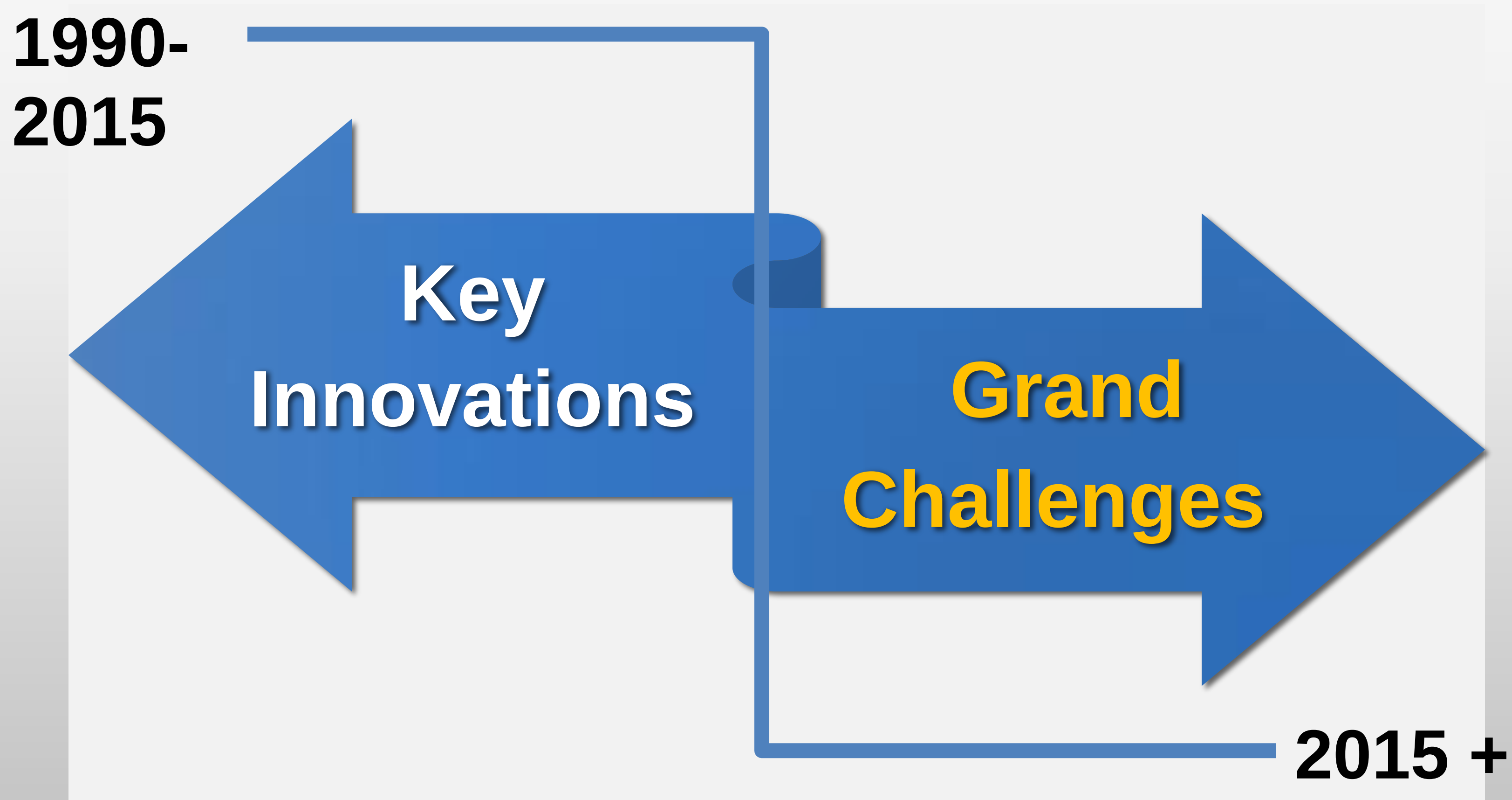
SEMI-SUPERVISED



- Online cursive handwriting recognition using speech recognition methods; , John Makhoul, Richard Schwartz, and George Chou ICASSP 1994

Accelerated Discovery

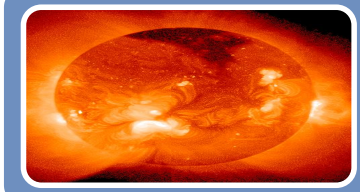



1990-
2015
Key
Innovations
Grand
Challenges
2015 +

Handwriting Recognition

Key Innovations

Lexicons



Fusion

Retrieval

Security

Summary

Grand Challenges

- 4Vs of Scientific Big Data
- 4 Rs: References, Reinvention, Replicability, Reputation

Grand Opportunities

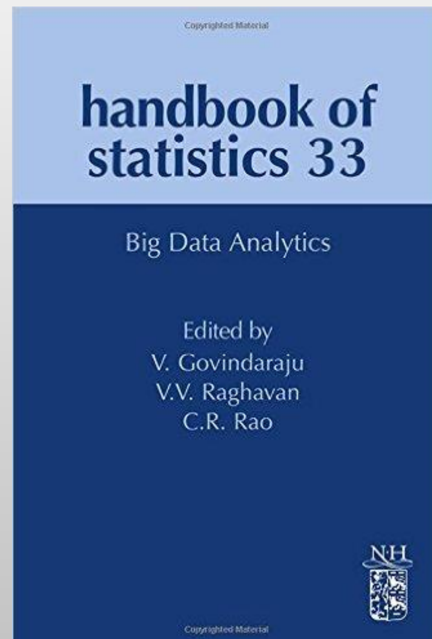
- Accelerated Discovery : Supervised linkages, heuristics;
- Integrate learning channels

Key Innovations

- Handwriting Recognition: Lexicons; Fusion; Retrieval; Security

Special Thanks to
All my students and colleagues

especially to colleagues
Srirangaraj Setlur and Ifeoma Nwogu



Venu Govinaraju, Ifeoma Nwogu, and Srirangaraj Setlur, "Document Informatics for Scientific Learning and Accelerated Discovery", *Handbook of Statistics (33): Big Data Analytics*, pp. 4-28, Elsevier, 2016.

Thank You

venu@cubs.buffalo.edu