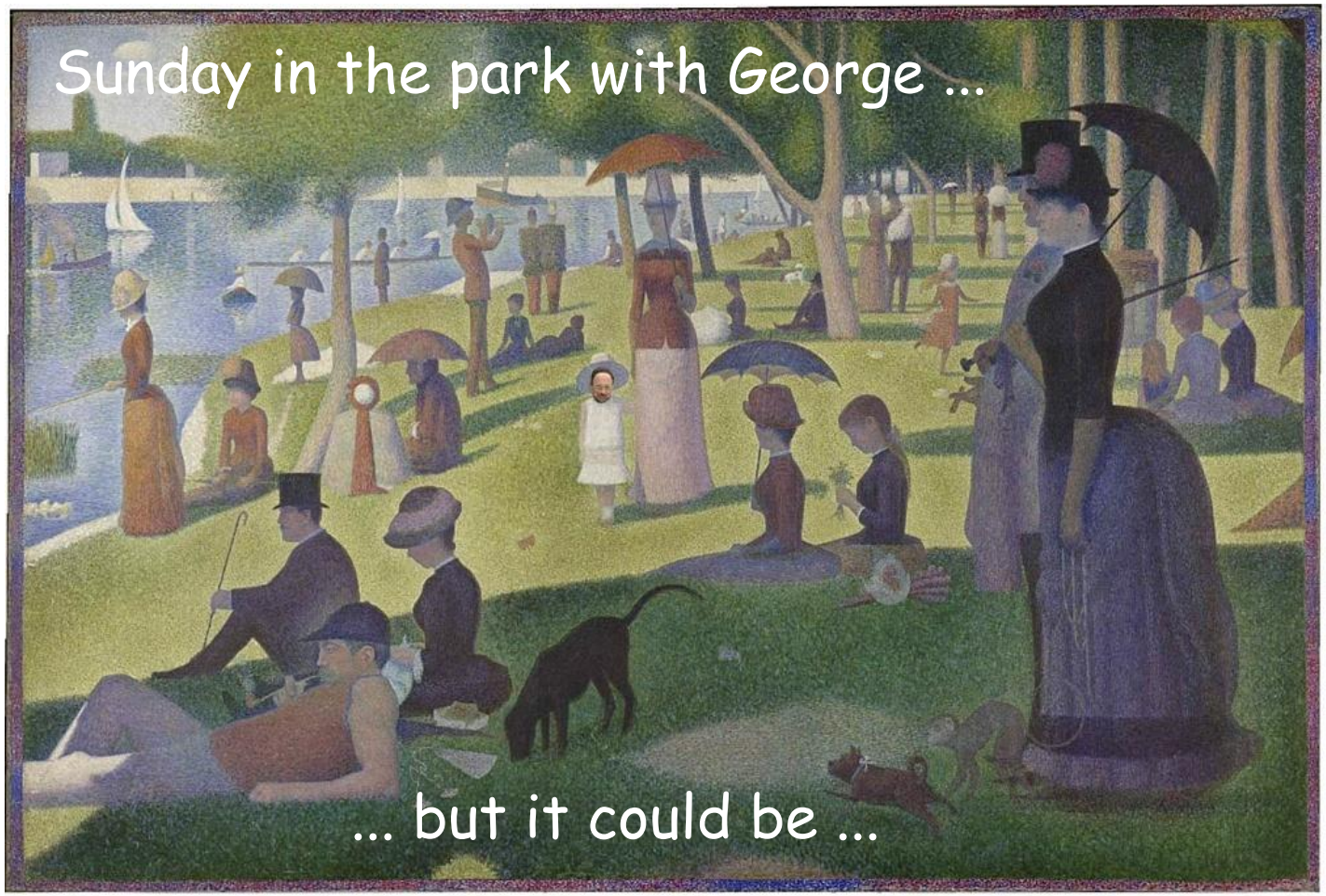
What this talk is not …



Adapted from "A Sunday Afternoon on the Island of La Grande Jatte" by Georges Seurat

What this talk is not …

Curious George

… but it could be …

Adapted from *Curious George* by Margret and H.A. Rey

What we're really talking about

Troy, NY, December 19, 2008

Working with George

Dan Lopresti

Computer Science & Engineering
Lehigh University
Bethlehem, PA, USA

What's the connection?

- Not a former student.
- Never employed by the same institution.

- Just a lucky bystander?

Serendipity

serendipity (sɛrən dɪpɪtɪ)
— n
the faculty of making fortunate discoveries by accident

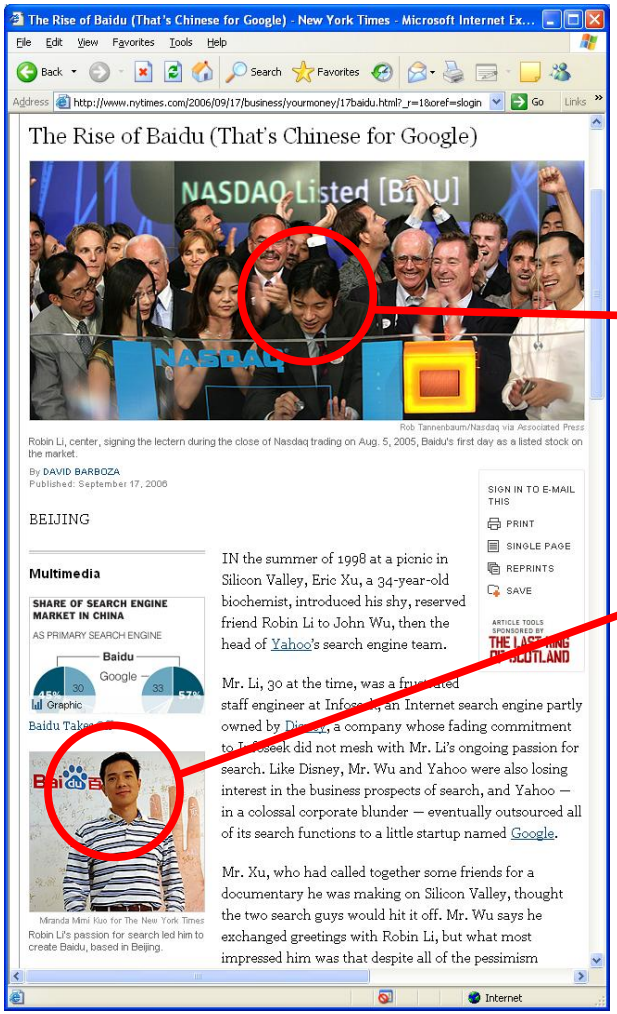
1994 = ancient history!

- "Validation of Simulated OCR Data Sets" G. Nagy, *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*, April 1994, Las Vegas, NV, pp. 127-135.
- "Validation of Document Defect Models for Optical Character Recognition" Y. Li, D. Lopresti, and A. Tomkins, *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*, April 1994, Las Vegas, NV, pp. 137-150.

Papers with George I

- "Spatial Sampling Effects in Optical Character Recognition," D. Lopresti, J. Zhou, G. Nagy, and P. Sarkar, *Proceedings of the Third International Conference on Document Analysis and Recognition*, August 1995, Montréal, Canada, pp. 309-314.
- "Validation of Image Defect Models for Optical Character Recognition," Y. Li, D. Lopresti, G. Nagy, and A. Tomkins, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 2, February 1996, pp. 99-108.
- "Spatial Sampling Effects on Scanned 2-D Patterns," J. Zhou, D. Lopresti, P. Sarkar, and G. Nagy, *Advances in Visual Form Analysis*, C. Arcelli, L. P. Cordella, and G. Sanniti di Baja, eds., Singapore: World Scientific, 1997, pp. 666-675.
- "Spatial Sampling of Printed Patterns," P. Sarkar, G. Nagy, J. Zhou, and D. Lopresti, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, March 1998, pp. 344-351.
- "Automated Table Processing: An (Opinionated) Survey," D. Lopresti and G. Nagy, *Proceedings of the Third IAPR International Workshop on Graphics Recognition*, September 1999, Jaipur, India, pp. 109-134.
- "Issues in Ground-Truthing Graphic Documents," D. Lopresti and G. Nagy, *Proceedings of the Fourth IAPR International Workshop on Graphics Recognition*, September 2001, Kingston, Ontario, Canada, pp. 59-72.

Brush with fame and fortune


The Rise of Baidu (That's Chinese for Google) - New York Times - Microsoft Internet Ex...

The Rise of Baidu (That's Chinese for Google)

Rob Tannenbaum/Nasdaq via Associated Press

Robin Li, center, signing the lectern during the close of Nasdaq trading on Aug. 5, 2005, Baidu's first day as a listed stock on the market.

By DAVID BARBOZA
Published: September 17, 2006

BEIJING

IN the summer of 1998 at a picnic in Silicon Valley, Eric Xu, a 34-year-old biochemist, introduced his shy, reserved friend Robin Li to John Wu, then the head of Yahoo's search engine team.

Mr. Li, 30 at the time, was a frustrated staff engineer at Infoseek, an Internet search engine partly owned by Disney, a company whose fading commitment to Infoseek did not mesh with Mr. Li's ongoing passion for search. Like Disney, Mr. Wu and Yahoo were also losing interest in the business prospects of search, and Yahoo — in a colossal corporate blunder — eventually outsourced all of its search functions to a little startup named Google.

Mr. Xu, who had called together some friends for a documentary he was making on Silicon Valley, thought the two search guys would hit it off. Mr. Wu says he exchanged greetings with Robin Li, but what most impressed him was that despite all of the pessimism

Multimedia

SHARE OF SEARCH ENGINE MARKET IN CHINA

AS PRIMARY SEARCH ENGINE

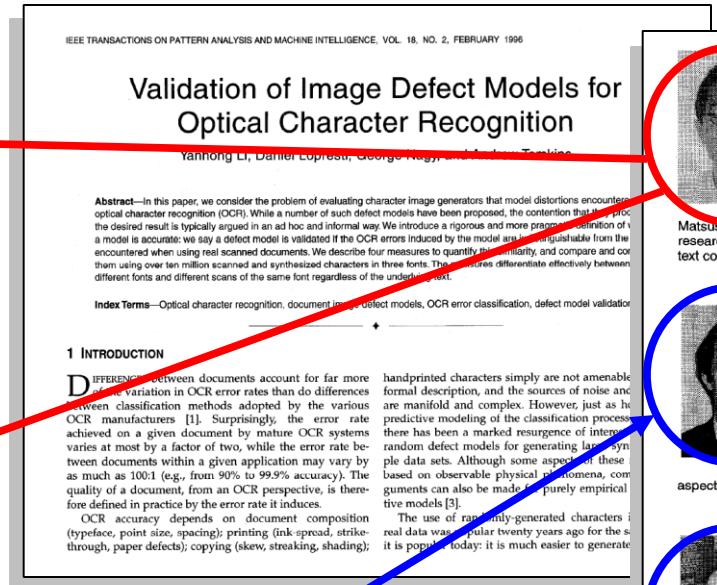
Baidu
Google
Graphic

Miranda Mimi Kuo for The New York Times

Robin Li's passion for search led him to create Baidu, based in Beijing.

From the *New York Times*, Sunday Sept. 17, 2006.

Robin Li, a student who worked with us, is now billionaire founder of Baidu.


IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 18, NO. 2, FEBRUARY 1996

Validation of Image Defect Models for Optical Character Recognition

Yanhong Li, Daniel Lopresti, George Nagy, and Andrew Tomkins

Abstract—In this paper, we consider the problem of evaluating character image generators that model distortions encountered in optical character recognition (OCR). While a number of such defect models have been proposed, the contention that they produce the desired result is typically argued in an ad hoc and informal way. We introduce a rigorous and more pragmatic definition of what it means for a model is accurate: we say a defect model is validated if the OCR errors induced by the model are indistinguishable from those encountered when using real scanned documents. We describe four measures to quantify this similarity, and compare and contrast them using over ten million scanned and synthesized characters in three fonts. These measures differentiate effectively between different fonts and different scans of the same font regardless of the underlying text.

Index Terms—Optical character recognition, document image defect models, OCR error classification, defect model validation.

1 Introduction

Differences between documents account for far more variation in OCR error rates than do differences between classification methods adopted by the various OCR manufacturers [1]. Surprisingly, the error rate achieved on a given document by mature OCR systems varies at most by a factor of two, while the error rate between documents within a given application may vary by as much as 100:1 (e.g., from 90% to 99.9% accuracy). The quality of a document, from an OCR perspective, is therefore defined in practice by the error rate it induces.

OCR accuracy depends on document composition (typeface, point size, spacing); printing (ink spread, strike-through, paper defects); copying (skew, streaking, shading); handprinted characters simply are not amenable to formal description, and the sources of noise and variation are manifold and complex. However, just as human predictive modeling of the classification process, there has been a marked resurgence of interest in random defect models for generating large synthetic data sets. Although some aspects of these models based on observable physical phenomena, comparisons can also be made for purely empirical descriptive models [3].

The use of randomly-generated characters in real data was popular twenty years ago for the same reasons it is popular today: it is much easier to generate

Robin

Me

George

Yanhong Li received the BS degree in information science from Peking University, China, in 1991 and the MS degree in computer science from the State University of New York at Buffalo in 1993. Since 1994, he has been a senior software engineer at GARI Software, IDD Information Services. From 1992 to 1994, he was a research assistant at the Center of Excellence for Document Analysis and Recognition (CEDAR) of SUNY Buffalo. In the summer of 1993, he did an internship at Matsushita Information Technology Laboratory in Princeton, N.J. His research interests include document analysis, information retrieval, text compression, and financial information routing.

Daniel Lopresti received the AB degree in mathematics and engineering from Dartmouth College in 1982 and the PhD degree in computer science from Princeton University in 1987. From 1986 until 1991, he was on the faculty of the Department of Computer Science at Brown University. In 1991 he joined the newly formed Matsushita Information Technology Laboratory as a senior scientist and leader of the Carbon Project. His research interests include pattern matching and recognition, parallel VLSI architectures, and computational aspects of molecular biology.

George Nagy received the BEng. and MEng. degrees from McGill University and the PhD in electrical engineering from Cornell University in 1962 (on neural networks). For the next ten years he conducted research on various aspects of pattern recognition and OCR at the IBM T.J. Watson Research Center in Yorktown Heights. From 1972 to 1985 he was a professor of computer science at the University of Nebraska–Lincoln and worked on remote sensing applications, geographic information systems, computational geometry, and human-computer interfaces. Since 1985 he has been a professor of computer engineering at Rensselaer Polytechnic Institute. He has held visiting appointments at the Stanford Research Institute, Cornell, the University of Montreal, the National Scientific Research Institute of Quebec, the University of Genoa and the Italian National Research Council in Naples and Genoa, AT&T Bell Laboratories, IBM Almaden, McGill University, and the Institute for Information Science Research at

Papers with George II

- "Why Table Ground-Truthing is Hard," J. Hu, R. Kashi, D. Lopresti, G. Nagy, and G. Wilfong, *Proceedings of the Sixth International Conference on Document Analysis and Recognition*, September 2001, Seattle, WA, pp. 129-133.
- "A Nonparametric Classifier for Unsegmented Text," G. Nagy, A. Joshi, M. Krishnamoorthy, Y. Lin, D. Lopresti, S. Mehta, and S. Seth, *Proceedings of Document Recognition and Retrieval XI (IS&T/SPIE International Symposium on Electronic Imaging)*, January 2004, Santa Jose, CA, pp. 102-108.
- "Chipless ID for Paper Documents," D. Lopresti and G. Nagy, *Proceedings of Document Recognition and Retrieval XII (IS&T/SPIE International Symposium on Electronic Imaging)*, January 2005, San Jose, CA, pp. 208-215.
- "Mobile Interactive Support System for Time-Critical Document Exploitation," G. Nagy and D. Lopresti, *Symposium on Document Image Understanding Technology*, November 2005, College Park, MD, pp. 111-119.
- "Match Graph Generation for Symbolic Indirect Correlation," D. Lopresti, G. Nagy, and A. Joshi, *Document Recognition and Retrieval XIII (IS&T/SPIE International Symposium on Electronic Imaging)*, January 2006, San Jose, CA, pages 606706.1-606706.9.

Papers with George III

- "Notes on Contemporary Table Recognition," D. Embley, D. Lopresti, and G. Nagy, *Proceedings of the Seventh IAPR International Workshop on Document Analysis Systems*, H. Bunke and A. L. Spitz, eds., Berlin: Springer-Verlag, 2006, pp. 164-175.
- "Interactive Document Processing and Digital Libraries," G. Nagy and D. Lopresti, *Proceedings of the Second International Conference on Document Image Analysis for Libraries*, April 2006, Lyon, France, pp. 2-11.
- "Table Processing Paradigms: A Research Survey," D. Embley, M. Hurst, D. Lopresti, and G. Nagy, *International Journal on Document Analysis and Recognition*, vol 8, no. 2-3, June 2006, pp. 66-86.
- "A Maximum-Likelihood Approach to Symbolic Indirect Correlation," A. Joshi, G. Nagy, D. Lopresti, and S. Seth, *Proceedings of the Eighteenth International Conference on Pattern Recognition*, August 2006, Hong Kong, pp. 99-103.
- "Multi-Character Field Recognition for Arabic and Chinese Handwriting," D. Lopresti, G. Nagy, S. Seth, and X. Zhang, *Proceedings of the Summit on Arabic and Chinese Handwriting Recognition*, September 2006, College Park, MD, pp. 93-100.

Papers with George IV

- "A Document Analysis System for Supporting Electronic Voting Research," D. Lopresti, G. Nagy, and E. Barney Smith, *Proceedings of the Eighth IAPR International Workshop on Document Analysis Systems*, IEEE Computer Society Press, September 2008, Nara, Japan, pp. 167-174.
- "Ballot Mark Detection," E. Barney Smith, D. Lopresti, and G. Nagy, *Proceedings of the Nineteenth International Conference on Pattern Recognition*, December 2008, Tampa, FL, pages 4 (CD-ROM).
- "Mark Detection from Scanned Ballots," E. Barney Smith, D. Lopresti, and G. Nagy, *Proceedings of Document Recognition and Retrieval XVI (IS&T/SPIE International Symposium on Electronic Imaging)*, January 2009, San Jose, CA, pages 7247-26.01-7247-26.10.
- "Tools for Monitoring, Visualizing, and Refining Collections of Noisy Documents," D. Lopresti and G. Nagy, *Proceedings of the Third Workshop on Analytics for Noisy Unstructured Text Data*, July 2009, Barcelona, Spain, pp. 9-16.
- "Document Photography in Vitro," G. Nagy, B. Clifford, A. Berg, G. Saunders, E. Barney Smith, and D. Lopresti, *Proceedings of the Third International Workshop on Camera-Based Document Analysis and Recognition*, July 2009, Barcelona, Spain, pp. 26-33.

Papers with George V

- "Camera-based Ballot Counter," G. Nagy, B. Clifford, A. Berg, G. Saunders, D. Lopresti, and E. Barney Smith, *Proceedings of the Tenth International Conference on Document Analysis and Recognition*, July 2009, Barcelona, Spain, pp. 151-155.
- "Style-Based Ballot Mark Recognition," P. Xiu, D. Lopresti, H. Baird, G. Nagy, and E. Barney Smith, *Proceedings of the Tenth International Conference on Document Analysis and Recognition*, July 2009, Barcelona, Spain, pp. 216-220.
- "Document Analysis Issues in Reading Optical Scan Ballots," D. Lopresti, G. Nagy, and E. Barney Smith, *Proceedings of the Ninth IAPR International Workshop on Document Analysis Systems*, June 2010, Boston, MA, pp. 105-112.
- "Characterizing Challenged 2008 Minnesota Ballots," G. Nagy, D. Lopresti, E. H. Barney Smith, and Z. Wu, *Document Recognition and Retrieval XVIII (IS&T/SPIE International Symposium on Electronic Imaging)*, January 2011, San Francisco, CA.
- "When is a Problem Solved?," D. Lopresti and G. Nagy, to be presented at the *Eleventh International Conference on Document Analysis and Recognition (ICDAR 2011)*, September 2011, Beijing, China.
- "Towards Improved Paper-based Election Technology," E. Barney Smith, D. Lopresti, G. Nagy, and Z. Wu, to be presented at the *Eleventh International Conference on Document Analysis and Recognition (ICDAR 2011)*, September 2011, Beijing, China.

The Debate

"Defect Models are Important to Advance the State-of-the-Art of Optical Character Recognition"

April 1996
Las Vegas, NV

For:

- Henry Baird
- Bob Haralick

Against:

- Dan Lopresti
- George Nagy

Decades of Influence

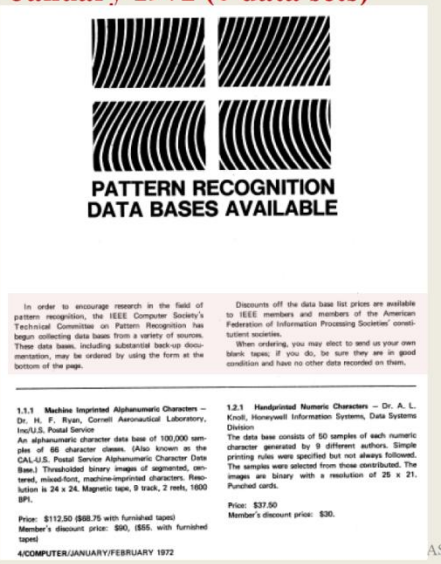
George Nagy graduat
Physics (fencing and
solving Euler's Secon
awarded the PhD at C
Rosenblatt build Tobe
network for speech re
character recognition
claims credit for IBM
reverse sabbatical at t
trains from cats' medi
Department of Compu
where he dabbled in c
1985 he has been Prof
Troy, NY. Nagy's credits in document analysis incl
Casey, "self-corrective" character recognition with
years later with Henry Baird), character recognitior

* Slides available on the DAS 2010 website.

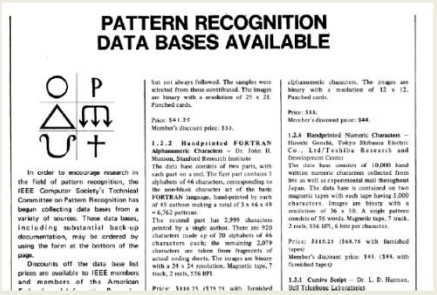
Keynote talk on datasets by George Nagy at DAS 2010 …

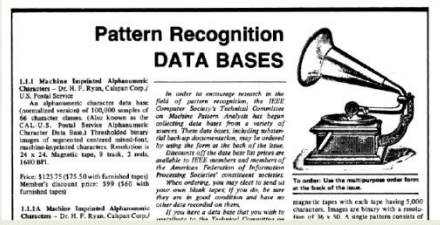
Data Sets advertised in IEEE *Computer*

January 1972 (6 data sets)


PATTERN RECOGNITION
DATA BASES AVAILABLE

January 1973


PATTERN RECOGNITION
DATA BASES AVAILABLE


Pattern Recognition
DATA BASES

AS 2010 (GN)

April 1976 (10 data sets)

A really good bad idea?

Training Humans to Read Like Machines
(or, The Lazy Researcher's Approach to Perfect OCR)

Dan Lopresti and George Nagy (if he agrees)

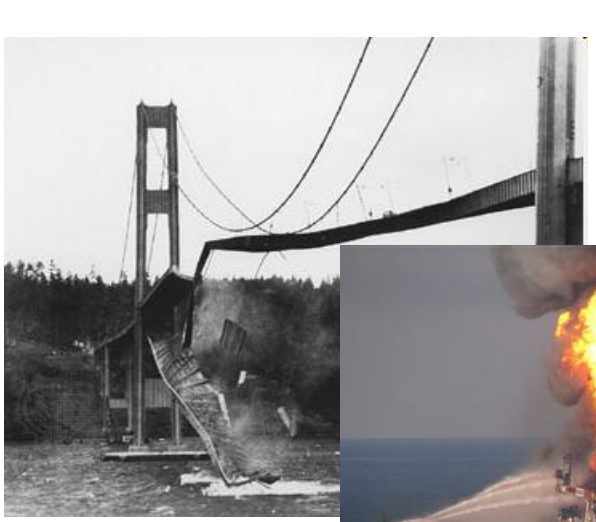
- For decades, document analysis researchers have labored with tremendous effort and unbridled enthusiasm in desperate attempts to raise accuracy rates for optical character recognition to 100%.
- Success has proved to be elusive for all but the cleanest of documents typeset using standard fonts, i.e., boring cases that present absolutely no challenge and that even a moderately-talented trained monkey could handle with one paw tied behind its back.

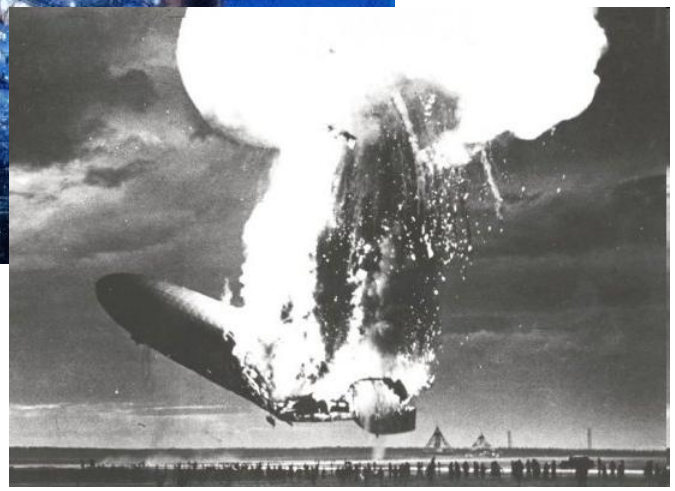
A really bad good idea?

- Tired of seeing the field perpetuate this exercise in futility, in this work we propose a novel, radical, earth-shaking, ground-breaking, revolutionary, radical idea.
- We posit that if it is too hard to solve the problem, it is always possible to change the problem and thereby make it easier to solve.
- Our thesis is that if we train humans to read like machines – to make all of the same mistakes that our current computer algorithms make when processing a typical page image – we will instantly achieve 100% OCR accuracy with no additional research effort required.

Support to back up our claims

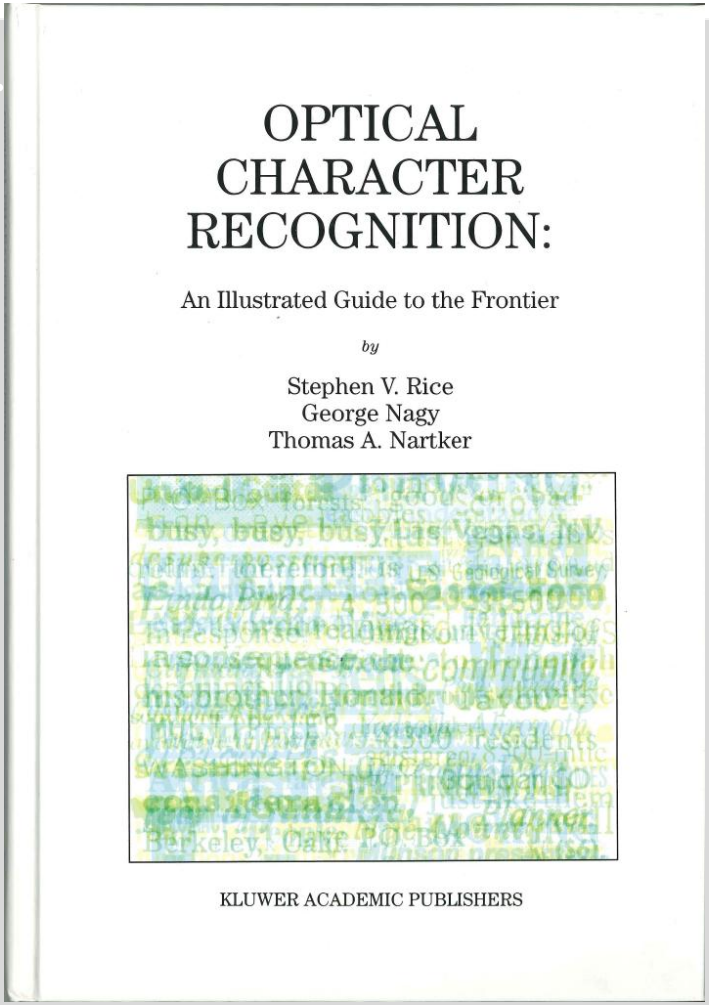
- We are quite certain this is feasible because humans are typically very smart and infinitely adaptable – making mistakes comes naturally to our species.



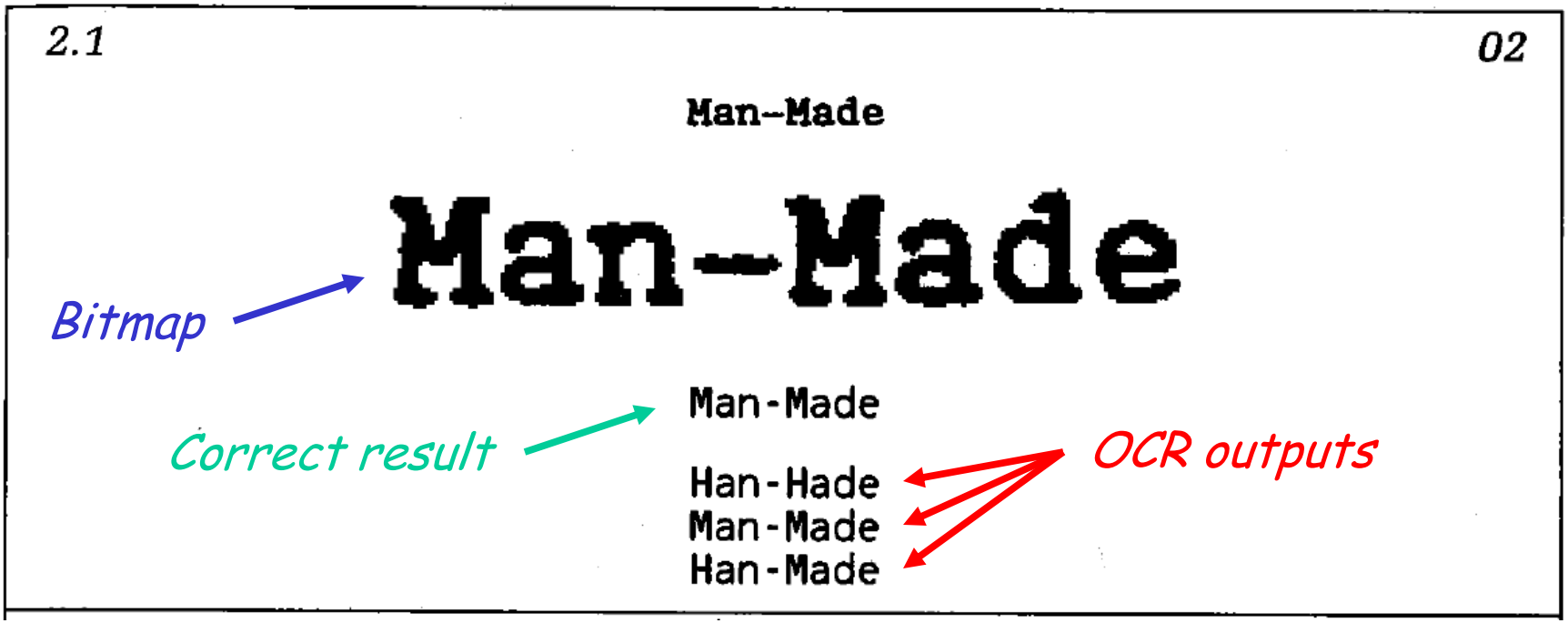


Proof? We don't need no proof.

- Instead, we shall conduct a brief but revealing demonstration using this classic tome:



The task: read like a machine



The task: read like a machine

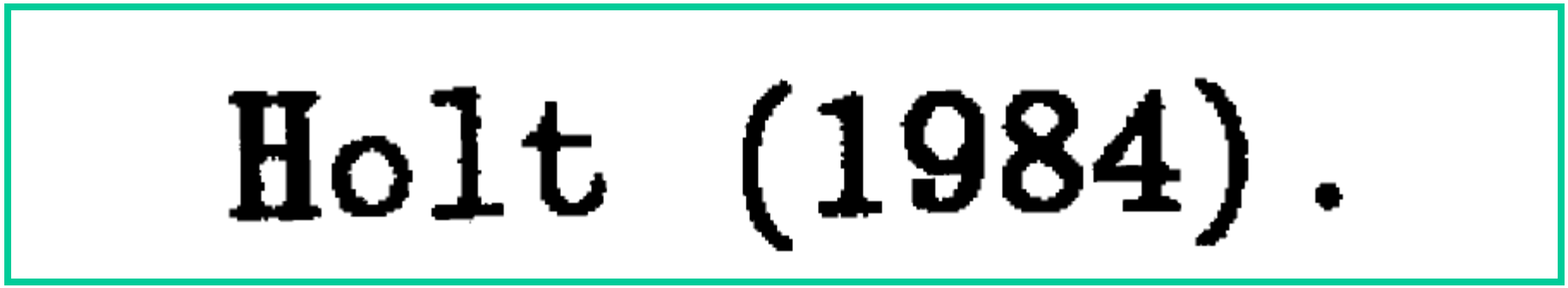
What would a machine output for this bitmap?

(A) Unnamed Mineral

(B) I think that I shall never see …

(C) Unnamed Nineral ← *Correct answer*

The task: read like a machine



What would a machine output for this bitmap?

(A) Bolt (1984). ← *Correct answer*

(B) Holt (1984).

(C) … a poem as lovely as a tree,

The task: read like a machine

What would a machine output for this bitmap?

(A) A penny saved is a penny earned.

(B) McGovern

(C) McGovem ← *Correct answer*

The task: read like a machine

What would a machine output for this bitmap?

(A) Imaging Experts

(B) A stitch in time saves nine.

(C) Iiiij~in~ L\1)(4~ ← *Correct answer*

The task: read like a machine

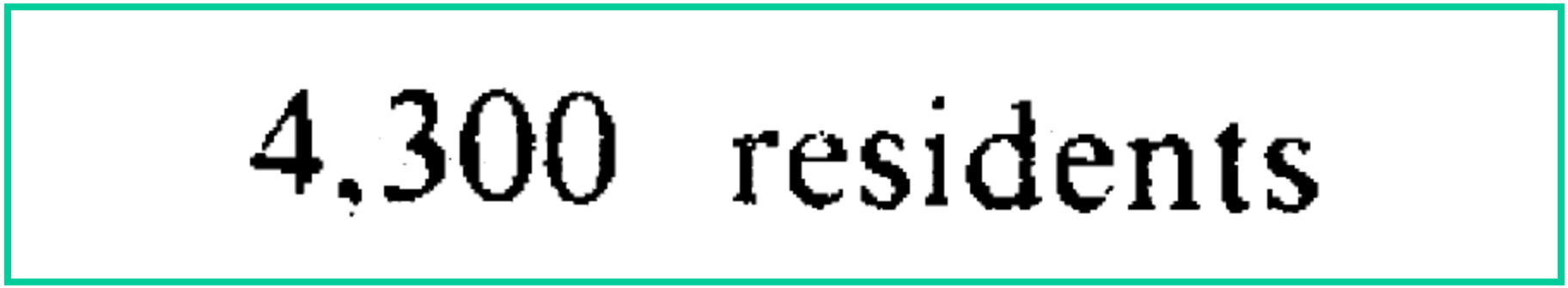
What would a machine output for this bitmap?

(A) May you live in interesting times.

(B) Grear Lakes ← *Correct answer*

(C) Great Lakes

The task: read like a machine



What would a machine output for this bitmap?

(A) 4.300 residents ← *Correct answer*

(B) 4,300 residents

(C) A miss is as good as a mile.

The task: read like a machine

What would a machine output for this bitmap?

(A) l have learned ← *Correct answer*

(B) I have learned

(C) What, me worry?

The task: read like a machine

What would a machine output for this bitmap?

(A) just like them

(B) justlike them ← *Correct answer*

(C) A rolling stone gathers no moss.

The task: read like a machine

Do you think you could learn to make the same mistakes a machine would make?

(A) Yes, I already make those same mistakes.

(B) Yes, I'm smarter than a dumb machine.

(C) Yes, anything you say, just stop talking!

Voilà!

Perfect OCR!!!

(or "Perfect OCRlll")

A final Haiku

Farewell RPI
George Nagy is retired
He is all ours now

Congratulations and best wishes, George and Jill, for a long, healthy, enjoyable, fulfilling retirement!