

# From Paper to Digital

Gaurav Sood

June 11, 2015

When we think about paper . . .

We think about government offices







But paper based storage of information is common

## Libraries and Archives



Health records

Receipts . . .

Small Businesses

And it isn't going away (quickly).

# The Dead Tree Format

## The Dead Tree Format

- Accessible only on location

## The Dead Tree Format

- Accessible only on location
- Typically needs help of another human, who may in turn want money

## The Dead Tree Format

- Accessible only on location
- Typically needs help of another human, who may in turn want **money**
- Hard to copy, distribute



## The Dead Tree Format

- Accessible only on location
- Typically needs help of another human, who may in turn want **money**
- Hard to copy, distribute
- Flammable

## The Dead Tree Format

- Accessible only on location
- Typically needs help of another human, who may in turn want **money**
- Hard to copy, distribute
- Flammable
- Time consuming to find stuff

## The Dead Tree Format

- Accessible only on location
- Typically needs help of another human, who may in turn want **money**
- Hard to copy, distribute
- Flammable
- Time consuming to find stuff  
Google returns average search query in .2 seconds

## The Dead Tree Format

- Accessible only on location
- Typically needs help of another human, who may in turn want **money**
- Hard to copy, distribute
- Flammable
- Time consuming to find stuff  
Google returns average search query in .2 seconds
- Hard to analyze, summarize stored information

## The Dead Tree Format

- Accessible only on location
- Typically needs help of another human, who may in turn want **money**
- Hard to copy, distribute
- Flammable
- Time consuming to find stuff  
Google returns average search query in .2 seconds
- Hard to analyze, summarize stored information
- Hard to track performance, identify anomalous transactions, identify patterns ...

Solved Problem?

## Solved Problem?

- Lots of software:

## Solved Problem?

- Lots of software:
  - Adobe Professional



# Solved Problem?

- Lots of software:
  - Adobe Professional
  - Abbyy FineReader

# Solved Problem?

- Lots of software:
  - Adobe Professional
  - Abbyy FineReader
  - Tesseract

# Solved Problem?

- Lots of software:
  - Adobe Professional
  - Abbyy FineReader
  - Tesseract
- But ...

# Solved Problem?

- Lots of software:
  - Adobe Professional
  - Abbyy FineReader
  - Tesseract
- But ...
  - Still can't handle complex layout, languages other than english etc.

# Solved Problem?

- Lots of software:
  - Adobe Professional
  - Abbyy FineReader
  - Tesseract
- But ...
  - Still can't handle complex layout, languages other than english etc.

*"I found that even native OCR software such as ... the Abbyy Fine Reader **proved utterly incapable of extracting words from scanned images of the texts**, even when those scanned images were of high quality."*

# Solved Problem?

- Lots of software:
  - Adobe Professional
  - Abbyy FineReader
  - Tesseract
- But ...
  - Still can't handle complex layout, languages other than english etc.
  - No information on how well you do (**Quality Metrics**).

# Solved Problem?

- Lots of software:
  - Adobe Professional
  - Abbyy FineReader
  - Tesseract
- But ...
  - Still can't handle complex layout, languages other than english etc.
  - No information on how well you do (**Quality Metrics**).
  - Not scalable

How to Convert Squiggles to Bits?



## How to Convert Squiggles to Bits?

- Take images of paper

## How to Convert Squiggles to Bits?

- Take images of paper
- Within images, find where **relevant** text is located

## How to Convert Squiggles to Bits?

- Take images of paper
- Within images, find where **relevant** text is located
- Find out how the text is laid out

## How to Convert Squiggles to Bits?

- Take images of paper
- Within images, find where **relevant** text is located
- Find out how the text is laid out
- Recognize the characters

Thus Performance Depends on...

- Quality of the scan: spine, contrast etc.

# Thus Performance Depends on...

## – Quality of the scan: spine, contrast etc.

### Alabama—Cable Systems

#### Pay Service 2

Pay Units: 71 (11/30/88).  
Programming (via satellite): Disney Channel.  
Fee: \$7.00 monthly.

#### Pay Service 3

Pay Units: 338 (11/30/88).  
Programming (via satellite): HBO.  
Fee: \$7.00 monthly.

#### Pay Service 4

Pay Units: 173 (11/30/88).  
Programming (via satellite): Showtime.  
Fee: \$7.00 monthly.

Local advertising: Yes. Rates: \$50.00/Minute; \$35.00/30 Seconds.  
Equipment: Scientific-Atlanta headend; C-COR amplifiers; Times Fiber cable; Scientific-Atlanta satellite antenna; Scientific-Atlanta satellite receivers.  
Miles of plant: 80.0 (coaxial). Homes passed: 3,000.  
Manager: Roth Hook. Chief technician: Larry Junkin.  
Ownership: Northland Communications Corp. (MSO).

**ALLGOOD**—Brookridge Cable Special Purpose Partnership, Suite 404, 7901 Stoneridge Dr., Pleasanton, CA 94588-3600. Phone: 510-463-1919. County: Blount. ICA: ALD165.

TV Market Ranking: 40. Franchise award date: N.A. Franchise expiration date: N.A. Began: February 1, 1989.  
Channel capacity: N.A. Channels available but not in use: N.A.

#### Basic Service

Subscribers: 107 (12/01/90).  
Programming (received off-air): WD88 (F) Beesmer; WBQ (P), WBRC-TV (F), WTO (I), WVTM-TV (N) Birmingham; WHNT-TV (C) Huntsville-Decatur.  
Programming (via satellite): WTBS (I) Atlanta; WGN-TV (W) Chicago; Turner Classic Movies.  
Fee: N.A.

Programming (via satellite): WGN-TV (W) Chicago; Disney Channel; Family Channel.  
Fee: \$4.65 monthly.

#### Pay Service 1

Pay Units: 98 (06/01/96).  
Programming (via satellite): HBO.  
Fee: \$11.95 monthly.

#### Pay Service 2

Pay Units: 62 (06/01/96).  
Programming (via satellite): The Movie Channel.  
Fee: \$10.95 monthly.

Miles of plant: 33.0 (coaxial).  
Manager: Freddy A. Arenobio. Chief technician: Daryl Bunn.  
Franchise fee: 3% of gross.  
Ownership: Falcon Cable TV (MSO).

**ANDALUSIA**—TV Cable Co. of Andalusia Inc., Box 34, 213 Dunson St., Andalusia, AL 36420-3705. Phone: 334-222-6464. Fax: 334-222-7226. County: Covington. ICA: ALD043.

TV Market Ranking: Outside TV Markets. Franchise award date: January 1, 1963. Franchise expiration date: September 1, 2017. Began: March 1, 1965.

Channel capacity: 40 (not 2-way capable). Channels available but not in use: 4.

#### Basic Service

Subscribers: 4,150 (06/21/92).  
Programming (received off-air): W0HN (A), WTVY (C) Ootham; WQIQ (F) Dozier; WEAR-TV (A) Mobile-Pensacola; WAKA (C), WSPA (N) Montgomery-Selma; 1 FM. Programming (via satellite): WTBS (I) Atlanta; WGN-TV (W) Chicago; A & E; BET; C-SPAN; CNBC; CNN; Country Music TV; Discovery Channel; ESPN; Family Channel; Lifetime; MTV; Nashville Network; Nickelodeon; Odyssey; The Weather Channel; Turner Network TV; USA Network.  
Current originations: Time-weather; public service announcements.  
Fee: \$20.00/installation; \$14.00 monthly.

TV Market Ranking: Below 100. Franchise award date: N.A. Franchise expiration date: N.A. Began: May 1, 1961.

Channel capacity: 42 (not 2-way capable). Channels available but not in use: None.

#### Basic Service

Subscribers: 35,217 (01/04/96).  
Programming (received off-air): WJSU-TV (C) Aniston; WGNX (C), WSB-TV (A) Atlanta; WBRC-TV (F), WVTM-TV (N) Birmingham; WRAL-TV (C) Gadsden; WCIO (F) Mount Cheaha State Park.  
Programming (via satellite): WTBS (I) Atlanta.

Current originations: Channel guide; classified ads.  
Fee: \$63.89 installation; \$5.19 monthly.

#### Expanded Basic Service

Subscribers: 33,719 (06/27/94).  
Programming (via satellite): A & E; BET; C-SPAN; CNBC; CNN; Country Music TV; Discovery Channel; E! Entertainment TV; ESPN; Family Channel; Headline News; Learning Channel; Lifetime; MTV; Nashville Network; Nickelodeon; Odyssey; QVC; The Weather Channel; Trinity Scisig Network; Turner Classic Movies; Turner Network TV; USA Network; VH1.  
Fee: \$13.39 monthly.

#### Expanded Basic Service 2

Subscribers: N.A.  
Programming (via satellite): American Movie Classics; Fox Sports South.  
Fee: \$1.50 monthly.

#### Pay Service 1

Pay Units: 3,398 (06/27/94).  
Programming (via satellite): Cinemax.  
Fee: \$17.36 installation; \$10.00 monthly.

#### Pay Service 2

Pay Units: 2,353 (06/27/94).  
Programming (via satellite): Disney Channel.  
Fee: \$17.36 installation; \$10.00 monthly.

#### Pay Service 3

Pay Units: 11,431 (06/27/94).  
Programming (via satellite): HBO.

#### Basic Service

Subscribers: 342 (05/01/96).  
Programming (received off-air): WQIQ (F) Dozier; WALA-TV (F), WEAR-TV (A), WJTC (I), WKRG-TV (C), WPMI (N) Mobile-Pensacola.  
Programming (via satellite): WTBS (I) Atlanta; WGN-TV (W) Chicago; Turner Classic Movies.  
Fee: N.A.  
Ownership: Torrence Cable Inc. (MSO).

**ARAB**—Charter Communications Inc., 904 Ross Rd., Albertville, AL 35950. Phone: 205-878-3802. 800-239-5111. Fax: 205-878-8287. County: Marshall. Also serves Marshall County (portions), Union Grove. ICA: ALD030.

TV Market Ranking: 96. Franchise award date: N.A. Franchise expiration date: N.A. Began: December 12, 1968.

Channel capacity: 38. Channels available but not in use: None.

#### Basic Service

Subscribers: 4,019 (02/16/96).  
Programming (received off-air): WBRC-TV (F), WVTM-TV (N) Birmingham; WTSP (C) Gadsden; WAAY-TV (A), WAFF (N), WHQ (P), WHNT-TV (C), WZLX (F) Huntsville-Decatur.

Programming (via satellite): WTBS (I) Atlanta; CNN; Comedy Central; Country Music TV; Discovery Channel; ESPN; ESPN2; Family Channel; Fox Sports South; Headline News; Home Shopping Network; Learning Channel; Lifetime; MTV; Nashville Network; Nickelodeon; Prevue Channel; The Weather Channel; Turner Network TV; USA Network; VH1.  
Current originations: Public access.  
Fee: \$20.50 installation; \$10.95 monthly.

#### Expanded Basic Service

Subscribers: 3,907 (02/16/96).  
Programming: N.A.  
Fee: \$15.90 monthly.

Channel capacity: 45 (2-way capable; 2-way). Channels available but not in use: 3.

#### Basic Service

Subscribers: 7,342 (07/14/95).  
Programming (received off-air): WAFF (N), WHQ (P), WHNT-TV (C) Huntsville-Decatur.

Programming (via satellite): VHS; WGN-TV (W) Chicago; A & E; BET; C-SPAN; Country Music TV; Discovery Channel; ESPN; Family Channel; South; Headline News; Learning Channel; MTV; Nashville Network; Nickelodeon; Nostalgia Channel; QVC; The Weather Channel; Network; Turner Network TV; VH1.

Current originations: Classified ads.  
Fee: \$21.50 installation; \$16.00 converter; \$32.00 activation.

#### Pay Service 1

Pay Units: 1,111 (07/14/95).  
Programming (via satellite): Channel.  
Fee: \$9.95 monthly.

#### Pay Service 2

Pay Units: 1,159 (07/14/95).  
Programming (via satellite): Channel.  
Fee: \$9.95 monthly.

#### Pay Service 3

Pay Units: 2,344 (07/14/95).  
Programming (via satellite): Channel.  
Fee: \$9.95 monthly.

#### Pay Service 4

Pay Units: 2,267 (07/14/95).  
Programming (via satellite): Channel.  
Fee: \$9.95 monthly.

#### Pay Service 5

Pay Units: 2,218 (07/14/95).  
Programming (via satellite): Channel.  
Fee: \$9.95 monthly.

#### Pay Service 6

Pay Units: 2,218 (07/14/95).  
Programming (via satellite): Channel.  
Fee: \$9.95 monthly.

#### Pay Service 7

Pay Units: 2,218 (07/14/95).  
Programming (via satellite): Channel.  
Fee: \$9.95 monthly.

Thus Performance Depends on...

- Quality of the scan: spine, contrast etc.
- Complexity of the layout





# Thus Performance Depends on...

- Quality of the scan: spine, contrast etc.
- Complexity of the layout

WBQ (P)  
TV (P)  
PM (P)  
BS (P)  
Al  
ter  
Ch

CO.

no. 904  
Phone  
loc. 205-  
1  
series  
ord. fax  
tve. fax  
L. Began

table but

ABRQ-TV  
WLP (P)  
4, WHQ  
antville

IS (P)  
nity Mo-  
ESPN2  
h; Head-  
network;  
r; Nash-  
Char-  
Network

Channel capacity: 45 (2-way capable; operating 2-way). Channels available but not in use: 5.

**Basic Service**  
Subscribers: 7,342 (07/14/95).  
Programming (received off-air): WAAY-TV (A), WAFF (N), WHQ (P), WHNT-TV (C), WDOX (F) Huntsville-Decatur; WSMV (N) Nashville.  
Programming (via satellite): WTBS (I) Atlanta; WGN-TV (W) Chicago; A & E; American Movie Classics; BET; C-SPAN; CNBC; CNN; Country Music TV; Discovery Channel; ESPN; Family Channel; Fox Sports South; Headline News; Learning Channel; Lifetime; MTV; Nashville Network; Nick at Nite; Nickelodeon; Nostalgia Good TV; Odyssey; QVC; The Weather Channel; Trinity Bostg Network; Turner Network TV; USA Network; VH1.  
Current origination: Classified ads; local live.  
Fee: \$21.50 installation; \$16.99 monthly; \$1.00 converter; \$32.00 additional installation.

**Pay Service 1**  
Pay Units: 1,111 (07/14/95).  
Programming (via satellite): The Movie Channel.  
Fee: \$5.95 monthly.

**Pay Service 2**  
Pay Units: 1,159 (07/14/95).  
Programming (via satellite): Cinemax.  
Fee: \$9.95 monthly.

**Pay Service 3**  
Pay Units: 2,344 (07/14/95).  
Programming (via satellite): Disney Channel.  
Fee: \$9.95 monthly.

**Pay Service 4**

**Basic Service**  
Subscribers: 241 (09/01/96).  
Programming (received off-air): WDBB (F) Bessemer; WBQ (P), WSMG (C), WBRC-TV (F), WTTD (I), WVTM-TV (N) Birmingham.  
Programming (via satellite): WTBS (I) Atlanta; WGN-TV (W) Chicago; CNN; Discovery Channel; ESPN; Family Channel; Headline News; Nashville Network; Turner Network TV; USA Network.  
Fee: \$24.12 monthly.

**Pay Service 1**  
Pay Units: N.A.  
Programming (via satellite): Disney Channel; HBO.  
Fee: \$7.95 monthly (Disney). \$9.95 monthly (HBO).  
Miles of plant: 34.0 (coaxial). Homes passed: 664.  
State manager: Bill Flowers. Technical manager: Ken Bryant.  
Ownership: Galaxy Cablevision (MSO).

**ASHFORD**—Galaxy Cablevision, Box 8, Highland, AL 36345-0008. Phone: 334-693-2610. 800-365-6988. Fax: 334-693-2291.  
County: Houston. Also serves Avon, Cowarts, Houston County (portions), Webb. ICA: AL0167.  
TV Market Ranking: Below 100. Franchise award date: N.A. Franchise expiration date: N.A. Began: N.A.  
Channel capacity: 37. Channels available but not in use: None.

**Basic Service**  
Subscribers: 235 (07/01/96).  
Programming (received off-air): WLTZ (N), WRSL (C), WYVM (A) Columbus; WDOH (A), WTVY (C) Dothan; WBQ (P) Dozier;

**ASHVILLE**—St. Clair Cablevision, Box 932, Fayette, AL 35555. Phone: 205-932-7264. County: St. Clair. Also serves Springville, St. Clair County (portions), Steele. ICA: AL0166.  
TV Market Ranking: 40 (Asheville, Springville, portions of St. Clair County). Below 100 (portions of St. Clair County, Steele). Franchise award date: N.A. Franchise expiration date: N.A. Began: January 1, 1988.  
Channel capacity: 42. Channels available but not in use: 4.

**Alabama—Cable Systems**

**NEXT LEVEL™** Broadband solutions at every level!  
Broadband Networks Group

Programming (via satellite): WTBS (I) Atlanta; WGN-TV (W) Chicago; ESPN; Family Channel.  
Fee: \$20.00 installation; \$11.00 monthly; \$12.50 additional installation.

**Pay Service 1**  
Pay Units: 180 (01/01/88).  
Programming (via satellite): HBO.  
Fee: \$8.00 installation; \$10.00 monthly.  
Local advertising: No.  
Equipment: Tocom headend; Magnevox amplifiers; Times Fiber cable; Eagle traps; Antenna Technology satellite receivers.  
Miles of plant: 27.0 (coaxial). Homes passed: 1,300.  
Manager: Cary Manning.  
Ownership: James Cable Partners (MSO).  
Note: Current information not available.

TV Market Ranking: 96. Franchise award date: N.A. Franchise expiration date: N.A. Began: October 24, 1986.  
Channel capacity: 40. Channels available but not in use: None.

**Basic Service**  
Subscribers: 7,862 (06/01/96).  
Programming (received off-air): WAAY-TV (A), WAFF (N), WHQ (P), WHNT-TV (C), WDOX (F) Huntsville-Decatur; alband FM.  
Programming (via satellite): A & E; American Movie Classics; BET; C-SPAN; CNBC; Comedy Central; Country Music TV; ESPN; Family Channel; Fox Sports South; Home Shopping Network; Knowledge TV; Learning Channel; Lifetime; MTV; Nickelodeon; QVC; The Weather Channel; Travel Channel; Trinity Bostg Network; Turner Network TV; USA Network; VH1.  
Current origination: Time-weather; news-ticker; stock ticker; bulletin board; message wheel.  
Fee: \$19.99 monthly.

**Expanded Basic Service**  
Subscribers: 7,707 (06/01/96).  
Programming (via satellite): CNN; Discovery Channel; Headline News.  
Fee: \$1.65 monthly.

**Expanded Basic Service 2**  
Subscribers: 7,543 (06/01/96).

## Thus Performance Depends on...

- Quality of the scan: spine, contrast etc.
- Complexity of the layout
- Font

## Thus Performance Depends on...

- Quality of the scan: spine, contrast etc.
- Complexity of the layout
- Font
- Language

## Thus Performance Depends on...

- Quality of the scan: spine, contrast etc.
- Complexity of the layout
- Font
- Language
- Hardware and Software (duh!)

# OCR

- Make images

# OCR

- Make images
- Detect Text

# OCR

- Make images
- Detect Text

**Alabama—Cable Systems**

# OCR

- Make images
- Detect Text

**Alabama—Cable Systems**

- Segment “Characters”



# OCR

- Make images

- Detect Text

**Alabama—Cable Systems**

- Segment “Characters”

**Alabama—Cable Systems**

# OCR

- Make images

- Detect Text

**Alabama—Cable Systems**

- Segment “Characters”

**Alabama—Cable Systems**

- Classify “Characters”

# OCR

- Make images

- Detect Text

**Alabama—Cable Systems**

- Segment “Characters”

**Alabama—Cable Systems**

- Classify “Characters”

**A** ⇔ A   **a** ⇔ a

# Mechanics

- Detect Text

# Mechanics

- Detect Text
  - Supervised Learning

# Mechanics

- Detect Text
  - Supervised Learning
  - Blobs with text, Blobs without

# Mechanics

- Detect Text
  - Supervised Learning
  - Blobs with text, Blobs without
  - But size of a blob is an issue

# Mechanics

- Detect Text
  - Supervised Learning
  - Blobs with text, Blobs without
  - But size of a blob is an issue
- Character Segmentation



# Mechanics

- Detect Text
  - Supervised Learning
  - Blobs with text, Blobs without
  - But size of a blob is an issue
- Character Segmentation
  - Supervised Learning

# Mechanics

- Detect Text
  - Supervised Learning
  - Blobs with text, Blobs without
  - But size of a blob is an issue
- Character Segmentation
  - Supervised Learning
  - Letters (and **Ligatures**) versus Splits

# Mechanics

- Detect Text
  - Supervised Learning
  - Blobs with text, Blobs without
  - But size of a blob is an issue
- Character Segmentation
  - Supervised Learning
  - Letters (and **Ligatures**) versus Splits
- Classify Characters (and Ligatures)
  - Supervised Learning
  - A versus B versus C...

## Supervised Learning

- Classified (training) data

# Supervised Learning

- Classified (training) data
- Estimate a model

# Supervised Learning

- Classified (training) data

- Estimate a model

$\text{logit}[p(\text{spam})] = \alpha + f'\beta$  where  $f$  is frequencies.

# Supervised Learning

- Classified (training) data
- Estimate a model
  - Predict class (e.g. Blobs with or without text) using features (pixel by pixel rgb)
  - Use cross-validation to tune the parameters

## Supervised Learning

- Classified (training) data
- Estimate a model
- Predict classes of unseen data (groups of pixels)



## Paper to Digital Pipeline

- Take images of paper
- Within images, find where **relevant** text is located
- Find out how the text is laid out
- Recognize the characters

## Paper to Digital Pipeline

- Take images of paper
- Within images, find where **relevant** text is located
- Find out how the text is laid out
- Recognize the characters
- **Every step is error prone**

Optimize all steps w.r.t final error rate.

Optimize all steps w.r.t final error rate. How to deal with errors that remain

## How to Fix Errors

- How confident are you that...

## How to Fix Errors

- How confident are you that...
  - An area has **relevant** text

# How to Fix Errors

- How confident are you that...
  - An area has **relevant** text
  - Split is correct

# How to Fix Errors

- How confident are you that...
  - An area has **relevant** text
  - Split is correct
  - Right character (or ligature) is recognized



# How to Fix Errors

- How confident are you that...
  - An area has **relevant** text
  - Split is correct
  - Right character (or ligature) is recognized
- Flag low confidence areas, splits, characters...

## How to Fix Errors

- How confident are you that...
  - An area has **relevant** text
  - Split is correct
  - Right character (or ligature) is recognized
- Flag low confidence areas, splits, characters...
- Get humans to identify the correct classes

# How to Fix Errors

- How confident are you that...
  - An area has **relevant** text
  - Split is correct
  - Right character (or ligature) is recognized
- Flag low confidence areas, splits, characters...
- Get humans to identify the correct classes
- Use that knowledge to fix other errors

## Fixing Character Recognition Errors

- Search and Replace

## Fixing Character Recognition Errors

- Search and Replace
- OCR makes certain kinds of errors (| is mistaken for an l)

## Fixing Character Recognition Errors

- Search and Replace
- OCR makes certain kinds of errors (| is mistaken for an l)
- Compare against a corpora (dictionary) and replace

## Fixing Character Recognition Errors

- Search and Replace
- OCR makes certain kinds of errors (I is mistaken for an l)
- Compare against a corpora (dictionary) and replace
- But replace with what?

## Fixing Character Recognition Errors

- Search and Replace
- OCR makes certain kinds of errors (l is mistaken for an I)
- Compare against a corpora (dictionary) and replace
- But replace with what?
- standd -> strand, stand, stood, or sand?



## Edit Distance

- How similar are two strings?

## Edit Distance

- How similar are two strings?
- Typically refers to minimum edit distance

## Edit Distance

- How similar are two strings?
- Typically refers to minimum edit distance
- Minimum number of editing operations (Insertion, Deletion, Substitution) to convert one string to another.

## Edit Distance

- How similar are two strings?
- Typically refers to minimum edit distance
- Minimum number of editing operations (Insertion, Deletion, Substitution) to convert one string to another.
- Levenshtein Distance, substitution cost = 2

## Edit Distance

- How similar are two strings?
- Typically refers to minimum edit distance
- Minimum number of editing operations (Insertion, Deletion, Substitution) to convert one string to another.
- Levenshtein Distance, substitution cost = 2
- You can implement this at word level so Microsoft Corp. is 1 away from Microsoft.

## Supervised Learning

- But edit distance isn't context aware. Use surrounding words.

## Supervised Learning

- But edit distance isn't context aware. Use surrounding words.
- How likely is a certain word within a phrase?

## Supervised Learning

- But edit distance isn't context aware. Use surrounding words.
- How likely is a certain word within a phrase?
- ~ Contemporary spelling correction algorithms



## Supervised Learning

- But edit distance isn't context aware. Use surrounding words.
- How likely is a certain word within a phrase?
- $\sim$  Contemporary spelling correction algorithms
- A bigram model of language: given previous word, probability of next word

## Supervised Learning

- But edit distance isn't context aware. Use surrounding words.
- How likely is a certain word within a phrase?
- $\sim$  Contemporary spelling correction algorithms
- A bigram model of language: given previous word, probability of next word
- But good training data is paramount.

## Supervised Learning

- Training data is 'similar data' (topic model) and data from human computation

## Supervised Learning

- Training data is 'similar data' (topic model) and data from human computation
- Estimate a model based on similar data

## Supervised Learning

- Training data is 'similar data' (topic model) and data from human computation
- Estimate a model based on similar data
- Use stochastic gradient descent to continue to tweak parameters based on human computation

## Supervised Learning

- Training data is 'similar data' (topic model) and data from human computation
- Estimate a model based on similar data
- Use stochastic gradient descent to continue to tweak parameters based on human computation
- Human computation parallelized, data for costlier (most duplicated low confidence strings, errors in recognition correlated) errors prioritized

## Supervised Learning

- Training data is ‘similar data’ (topic model) and data from human computation
- Estimate a model based on similar data
- Use stochastic gradient descent to continue to tweak parameters based on human computation
- Human computation parallelized, data for costlier (most duplicated low confidence strings, errors in recognition correlated) errors prioritized
- Calculate error rate against trained random