

# The “Names Game” – Harnessing Inventors’ Patent Data for Economic Research

NSF Workshop  
February 2008



Trajtenberg, Shiff & Melamed  
NBER Working Paper No. 12479

1



- Background
- Stage I – Grouping Similar Names
- Stage II – Computerized Matching Process
- Stage III – Evaluation and Fine-Tuning
- Some Results

2

## Background



- During 1975-1999 the USPTO has granted 2,139,313 patents. These patents (and all the meta-data involved) yield a database consisting of 4,298,457 records
- This database is the product of *Hall, Jaffe & Trajtenberg's* work. It contains data concerning patents, citations, assignees, inventors, patent classes, countries, etc. <http://www.nber.org/patents>
- Previous studies focused mainly on issues such as:
  - Patents rather than inventors
  - Small scale groups of inventors (manual data processing)
  - Large scale groups of inventors (rather “simple” computerized data processing)

3

## Motivation



- Focusing on the inventors, rather than the patents, enables researching issues such as:
  - What is the profile of a successful inventor?
  - Inventors' mobility - which inventors tend to move more often (countries, assignees, technological fields, etc.)
  - How does inventors' mobility affect productivity?
  - Human capital and “Brain Drain”
  - Inventors' mobility and data spillover
  - Working in teams vs. working solo (inventors networks)

4

## The Main Challenge– “Who is who?”



Main questions:

- How can we use this dataset to match inventors?
- How reliable are the names on the patent application form?
- What additional information is available for the matching process and how can we use it?

5

## Under-matching vs. Over-matching



- Avoiding Under-matching (Type I error)
  - Is Gerald Fleischfresser the same person as Gerald Fleischfreser?
  - How do we handle nicknames? (Bob vs. Robert)
  - How do we handle variations in a name's initial? (Tsidon vs. Ziddon)
- Avoiding Over-matching (Type II error)
  - Are Gerald Fleischfresser and Gerald Fleischfreser the same person?
  - Are Robert W. Smith and Robert W. Smith the same person?

6

## Naïve Matching Processes



Matching Process	Inventors
Each record is a different inventor	4,298,457
Records with the exact same first, middle and last name	1,405,318
Records with the exact same first and last name	1,205,403


7

## Steps Towards a Matching Method



1. Extracting all relevant data from the patent application
2. Processing the data into a “usable” form
3. Developing an algorithm for clustering the dataset into inventors
4. Developing a debug mechanism
  - The ability to control each step in the process
  - Forming measurable scales for testing the algorithm’s performance

8



United States Patent  
 Frohman-Bentchkowsky, et. al. 4,203,158  
May 13, 1980

**Electrically programmable and erasable MOS floating gate memory device employing tunneling and method of fabricating same**

**Abstract**  
 An electrically programmable and electrically erasable MOS memory device suitable for high density integrated circuit memories is disclosed. Carriers are tunneled between a floating conductive gate and a doped region in the substrate to program and erase the device. A minimum area of thin oxide (70 Å-200 Å) is used to separate this doped region from the floating gate. In one embodiment, a second layer of polysilicon is used to protect the thin oxide region during certain processing steps.

**Inventors:** Frohman-Bentchkowsky; Dov (Haifa, IL); Mar; Jerry (Sunnyvale, CA); Perlegos; George (Cupertino, CA); Johnson; William S. (Palo Alto, CA).  
**Assignee:** Intel Corporation (Santa Clara, CA).  
 Appl. No. 969,819  
 Filed: Dec. 15, 1978

**Related U.S. Application Data**  
 Continuation-in-part of Ser No. 831,029, Feb. 24, 1978, abandoned.  
**Int. Cl.:** G11C 11/40  
**Current U.S. Cl.:** 365/185.29; 257/321; 326/37; 327/427;  
**Field of Search:** 365/185, 189; 307/238; 357/41, 45, 304


**References Cited | [Referenced By]**

U.S. Patent Documents			
3,500,142	Mar., 1970	Kahng	365/185
4,051,464	Sept., 1977	Huang	365/185

*Primary Examiner:* Fears, Terrell W.

**16 Claims, 14 Drawing Figures**

9



- Background
- Stage I – Grouping Similar Names
- Stage II – Computerized Matching Process
- Stage III – Evaluation and Fine-Tuning
- Some Results

10

## Names Handling



- “Cleaning up” names
  - Removing all non-letters characters (including spaces)
    - O’Brien → OBrien, Jean-Jaques → JeanJacques
    - Same for prefixes (De, Van der, Von, etc.)
- Removing surnames and middle names
- Rewriting in capital letters
- Soundex

11

## Names Handling - Summary



12

## Names Coding Using Soundex



- Invented in 1850 for the US Census
- The algorithm:
  - Each letter (except for the initial) is coded into a digit according to a given table
  - The letters A,E,H,I,O,U,W,Y are ignored
  - If the current letter is coded the same as the letter before, it is also ignored
  - The name is coded until it is four character long (an initial plus three digits)
  - Prefixes may be ignored (depends on the user)

13

## Name Coding Using Soundex



Score	Letter(s)
1	B F P V
2	C G J K Q S X Z
3	D T
4	L
5	M N
6	R
None	A E H I O U W Y

14

## Name Coding Using Soundex



- Changing the Soundex code to 7 characters
  - Original Soundex coding  
Winterheimer Sylvester → W536 S412
  - Our Soundex coding  
Winterheimer Sylvester → W536560 S412360
  - Affects ~700,000 records
  - Example

Grosmann Klaus	G625500 K420000
Grosman Klaus	
Grossman Klaus	

15

## Name Coding Using Soundex



- Further examples

Garcia David	G620000 D130000
Greig David	
Gross David	

Brook William	B620000 W450000
Bryg William	
Byers William	

16

## Soundex - Summary



- Pros
  - A tested method for names standardization and coding
  - Solves most spelling problems (except for initials)
  - A decent solution for western names
- Cons
  - Found to be problematic with East-Asian names (~25% of the record)
  - Sensitive for mistakes in the initial

17

- Background
- Stage I – Grouping Similar Names
- Stage II – Computerized Matching Process
- Stage III – Evaluation and Fine-Tuning
- Some Results



18

## The Matching Process - Outline



- After grouping the names in to p-sets (“potential sets for match) using the Soundex algorithm:
- Each record is compared to all other records within the p-set, and only to records within the p-set
- Each pair of records is compared and assigned with a score depending on the matching criteria
- For each pair there is a relevant threshold. If the score exceeds this threshold the records are assigned with the same ID.
- The process is transitive
- The inherent problem – assigning a cardinal measure to an ordinal relationship

19

## Name Frequencies



Differentiating between “frequent” and “rare” names

Main issues:

- Is the name sample in our database a random sample?
- Dealing with the multi-nationality of the inventors (hence, different name frequencies for each country)

Our solution – Setting a cutoff value. Names that appear in our dataset more than this value are considered “frequent”, otherwise they are considered “rare”

20

## Other Frequencies



Differentiating between “frequent” and “rare” cities,  
companies & patent classes:

Hitachi vs. a small startup, Tokyo vs. a small town

Our solution – Setting cutoff values for  
assignee, city and patent class

21

## Determining the Cutoff Values



- Example for determining cutoff value for cities:  
x – number of patents originating in a city (city’s size)  
f(x) – number of cities of the same size  
 $y=x*f(x)$  – total amount of patents from cities of size x
- The threshold is the median of y’s distribution

Category	Cutoff	Median
Soundex Code	16	23
City	1,382	1,382
Assignee	500	1,540
Patent Class	18,861	18,861

22

## The Matching Process



- Selecting appropriate criteria for comparing and matching within each p-set:
  - Exact same address (street + city (or ZIP) + country)
  - Self citation
  - Partners
  - Middle name, Surname, etc.
  - Assignee
  - City
  - Patent Class
- Setting score for each criteria and comparing the accumulated score to a threshold

23

## The Matching Process - Thresholds



- Scores are meaningless without the thresholds
- Setting thresholds according to the name similarity:

Names' Similarity	Threshold
Exact same first and last name (6 character Soundex code is equal to an exact name)	100
Exact same last name or the same Soundex code (more than 2 digits, less than 5)	120
Other	180

24

## The Matching Process – Scoring Scheme



- Not all criteria are as informative

Our solution – Categorizing criteria by their “strength” and scoring accordingly

Group	Criterion	Score
1	full exact name, self citation, partners	120
2	middle name, initial of middle name (rare name), small assignee (rare name), small city (rare name)	100
3	Small class (rare name), large assignee (frequent name), large city (rare name)	80
4	large class (rare name), surname, initial of middle name (frequent name)	50

25

- Background
- Stage I – Grouping Similar Names
- Stage II – Computerized Matching Process
- Stage III – Evaluation and Fine-Tuning
- Some Results



26

## Benchmark Israeli Inventors Set (BIIS)



- The Israeli Inventors: Inventors with at least one patent applied with an address in Israel
- Starting point:
  - Over 6,000 potential inventors
  - Over 13,000 records
  - Gathering all other records yielded over 18,000 records
- After a through data analysis and manual matching of the set:
  - 6,023 inventors
  - 15,316 records

27

## Using the BIIS



- Comparing the computerized matching process (CMP) to the manual one (BIIS) using indexes built for testing the “goodness of fit” (GOFI)
- $GOFI_1$  - how close is the match in the different method?

$$GOFI_1 \equiv \text{Mean} \left[ \frac{|B_{ij} \cap C_{ij}|}{|B_{ij} \cup C_{ij}|} \right], \quad i = 1, \dots, N_{IL}$$

- $GOFI_2$  - in which method is the under-matching more dominant?

$$GOFI_2 \equiv \text{Mean} \left[ \frac{|B_{ij} \cap C_{ij}|}{|B_{ij}|} \right], \quad GOFI_2 \equiv \text{Mean} \left[ \frac{|B_{ij} \cap C_{ij}|}{|C_{ij}|} \right]$$

28

## Using the BIIS



- $GOFI_3$

$$GOFI_3(i, j) = \begin{cases} 1 & [|C_{ij}| - |B_{ij} \cap C_{ij}|] \neq 0 \text{ or } [|B_{ij}| - |B_{ij} \cap C_{ij}|] \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$GOFI_3 \equiv \sum_{i,j} GOFI_3(i, j)$$

29

## Comparing CMP to BIIS



	CMP	BIIS
Patents	9,155	
Records	15,306	
Original Names	6,314	
P-sets	5,858	
Unique Inventors	6,900	6,023
Average patents per inventor	2.21	2.54
$GOFI_1$	0.88	
$GOFI_2$	0.99	0.89

30

## Comparing CMP to BIIS



- From GOFI<sub>3</sub> we found that 5,081 records were matched differently - there's a difference of about 15% in the number of inventors
- The process performs well in avoiding over-matching:
  - 73 inventors over-matched (parallel to 196 inventor in the BIIS)
  - Some errors found in the BIIS
- High incidence for under-matching
  - 780 inventors were split into 1,871 inventors
  - The reasons – too little in common (~50%), spelling mistakes in names (~33.3%), other spelling mistakes (~16.7%)

31

## Comparing CMP to BIIS



- Conclusions (cont.)
  - Lower bound for Type I errors – 7-8% (records having too little in common)

Still, we must recall that this is not necessarily a random sample. Applying the process on a different sample might have resulted differently

32

## Average Mean Score (AMS)



- An additional index for matching quality – the average mean score

$$AMS_i \equiv \frac{\sum_{j=1}^{m_i} \text{pairwise score}_j}{m_i}, \quad m_i = \frac{N_i(N_i - 1)}{2}$$

- Calculated for each inventor after the matching is completed
  - Example 1: score (A,B) = 310, score (A,C)=150, score (B,C) = 80, → **AMS=180**
  - Example 2: score (A,B) = 120, score (A,C)=120, score (B,C) = 0, → **AMS=80**

33

## Average Mean Score (AMS)



- Calculating the AMS for the entire process

$$AMS = \frac{\sum_i N_i AMS_i}{N}, \quad N = \sum_i N_i$$

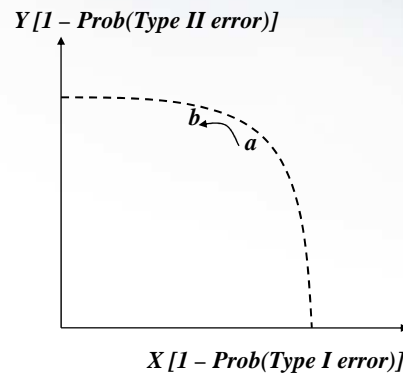
- Any “partial derivative” change in the CMP will change this score, and give an indication whether it is an improvement or not

34

## Average Mean Score (AMS)



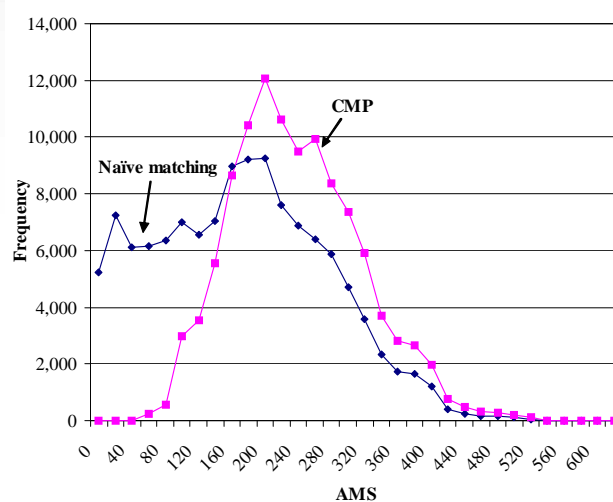
- The Substitution curve of the matching process:



- An inherent trade-off between under-matching and over-matching, moving along the curve

35

## Average Mean Score (AMS)



Means – 235 (CMP) vs. 171 (Naïve)

36

## East-Asian Inventors



- Soundex was invented for use of American (and western) names
- 1,102,459 (25.6%) records for East-Asian inventors (China, Hong Kong, Japan, Korea, Singapore & Taiwan)
- Many inventors, who originate in East-Asian countries apply from non East-Asian countries
- The way Soundex ignores vowels (plus h, w and y) generates a major problem when coding East-Asian names:
- East-Asian names tend to be coded into shorter Soundex codes, therefore bigger p-sets and higher probability for over-matching

37

## East Asian Inventors



- Problems assessing the similarity of East-Asian names and nicknames
  - How close are “Takeshi” and “Takashi”? “Kazo” and “Kazuo”?
- Problems assessing East-Asian names’ frequencies
  - What is the Japanese “Robert Smith”?

38

## East Asian Inventors



- Examples:
  - All the following last names: Cho, Choe, Choi, Cha, Choy, Chioo, Chiou, Chiu, Chae are Soundex coded into **C000000**
  - The biggest p-set in the data set is **S300000 K200000** consists of 1753 records (137 “Sato Koichi”s, 130 “Sato Kozo”s, 123 “Sato Kauzo”s and many more...)

39

## East Asian Inventors



- One possible solution may be not coding East-Asian using Soundex but using an alternative method
- The solution chosen was handling East-Asian names in a slightly different manner:
  - Flagging East-Asian names
  - Forcing stricter conditions for a match (e.g. Soundex codes must be longer than 2, both first and last must be exact, etc.)

40



- Background
- Stage I – Grouping Similar Names
- Stage II – Computerized Matching Process
- Stage III – Evaluation and Fine-Tuning
- Some Results

41

## Some Results



- Most frequent inventors' names:
  - Robert Smith – 271 inventors (749 records)
  - David Smith – 227 inventors (643 records)
  - Robert Miller – 176 inventors (588 records)
- Top patent holders:
  - George Spector (since 1976) – 715 patents
  - Shunpei Yamazaki (since 1979) – 605 patents
  - Donald Weder (since 1978) – 466 patents

42

## Some Results – Inventors Clustering



Matching Method	Inventors
Each record a different inventor	4,298,457
<b>CMP</b>	<b>1,632,532</b>
Same first, middle & last name	1,405,318
Same first & last name	1,205,403
<b>Same Soundex code</b>	<b>630,887</b>

43

## Some Results – Number of Patents



Patents	Inventors	%
1	983,859	60.27
2-5	497,480	30.49
6-9	80,835	4.95
10-50	67,565	4.14
50+	2,402	0.15
<b>Total</b>	<b>1,632,441</b>	<b>100</b>

44

## Some Results – Frequent Cities



City	Patents
Tokyo	135,910
Yokohama	74,577
Kanagawa	47,695
Kawasaki	40,615
Osaka	33,360
Houston (TX)	26,241
San Jose (CA)	22,573
Rochester (NY)	18,452
Austin (TX)	17,910
Saitama	16,768

45

## Some Results – Frequent Assignees



Assignee	Patents
HITACHI, LTD	70,921
IBM, LTD	63,311
CANON KABUSHIKI KAISHA	52,994
GENERAL ELECTRIC COMPANY	38,297
BAYER AKTIENGESELLSCHAFT	37,200
TOSHIBA CORPORATION	36,290
MATSUSHITA ELECTRIC INDUSTRIAL CO., LTD.	32,316
mitsubishi denki kabushiki kaisha	30,604
BASF AKTIENGESELLSCHAFT	27,806
EASTMAN KODAK COMPANY	27,720

46

## Some Results – Frequent Classes



Assignee	Patents
Drug, bio-affecting and body treating compositions	163,051
Stock material or miscellaneous articles	90,736
Chemistry: molecular biology and microbiology	76,919
Radiation imagery chemistry: process, composition, or product thereof	68,628
Drug, bio-affecting and body treating compositions	61,405
Measuring and testing	54,696
Internal-combustion engines	49,513
Active solid-state devices (e.g., transistors, solid-state diodes)	46,379
Semiconductor device manufacturing: process	45,832
Radiant energy	44,521

47

**Thank You!**



48