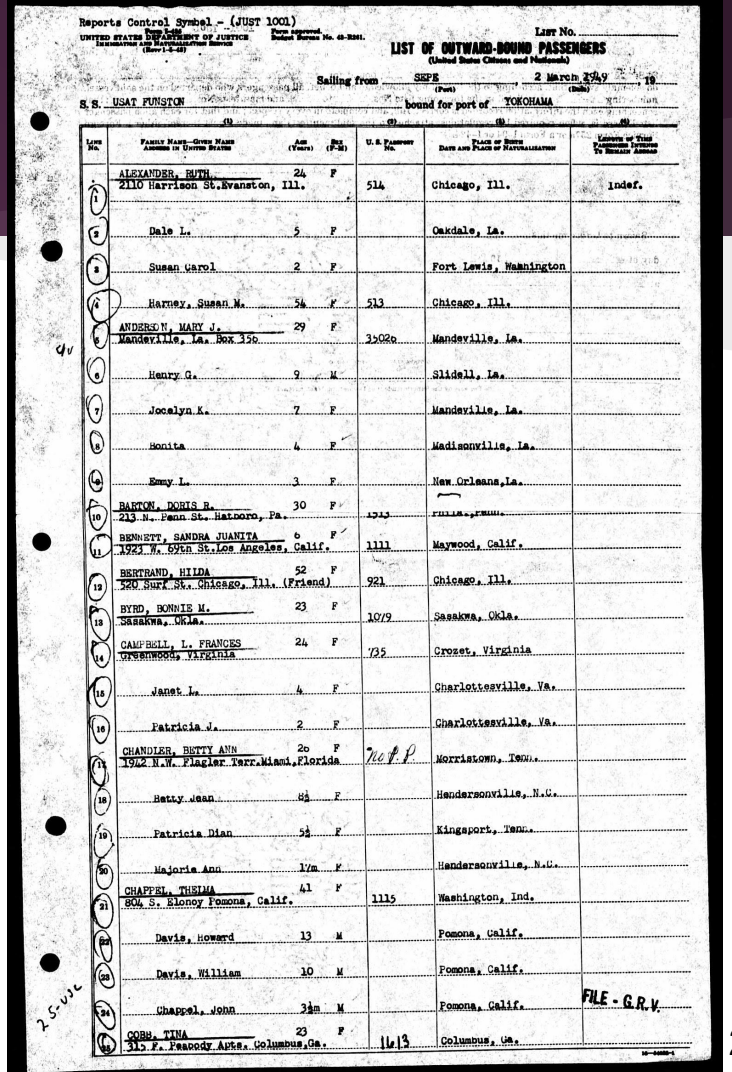
CONFIRM - Clustering Of Noisy Form Images using Robust Metrics

Chris Tensmeyer and Tony Martinez
Machine Learning Laboratory
Brigham Young University

Problem

- Scanned Form Images
- Cluster by Structure
- Noise
- Similar Form Types
- Potentially Large Datasets
  - Not quite solved yet



Reports Control Symbol - (JUST 1001)
Budget Bureau No. 43-R261.
UNITED STATES DEPARTMENT OF JUSTICE
IMMIGRATION AND NATURALIZATION SERVICE
(Rev. 1-2-48)

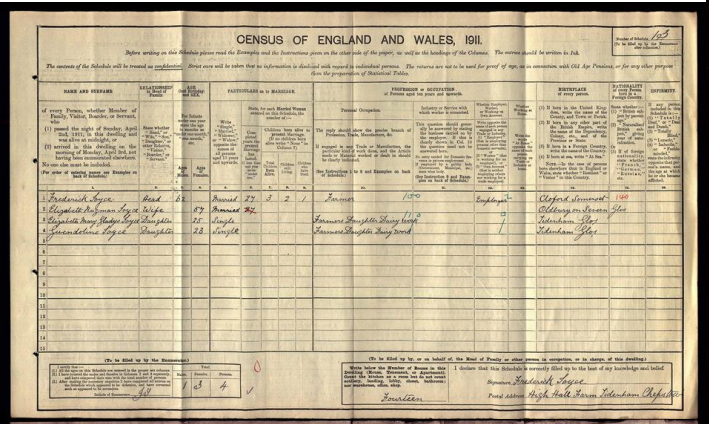
LIST OF OUTWARD-BOUND PASSENGERS
(United States Citizens and Nationals)

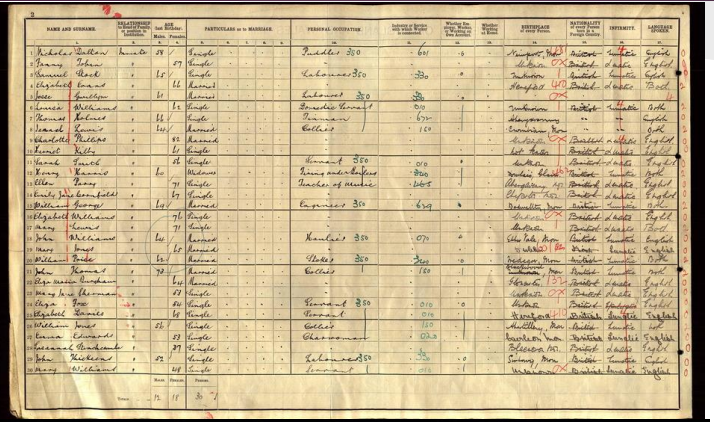
List No. ____

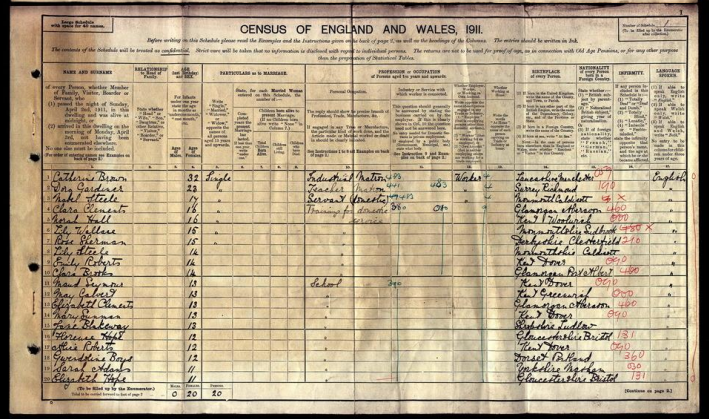
S.S. USAT FUNSTON Sailing from SEPE (Port) 2 March 1949 (Date) 19
bound for port of YOKOHAMA

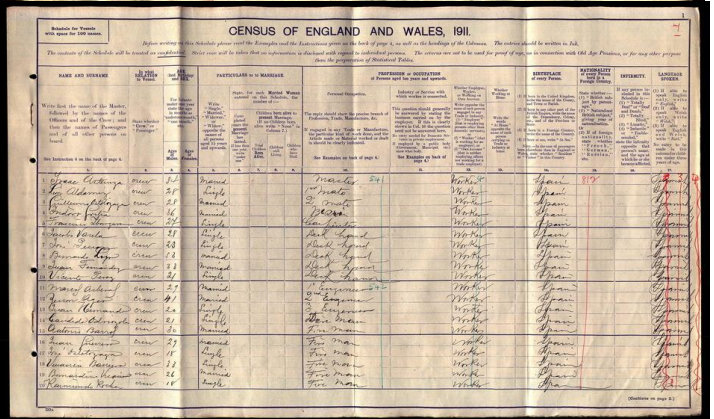
Line No.	Family Name—Given Name Address in United States	Age (Years)	Sex (F-M)	U.S. Passport No.	Place of Birth Date and Place of Naturalization	Length of Time Passengers Intends To Remain Abroad
1	ALEXANDER, RUTH 2110 Harrison St. Evanston, Ill.	24	F	514	Chicago, Ill.	Indef.
2	Dale L.	5	F		Oakdale, La.	
3	Susan Carol	2	F		Fort Lewis, Washington	
4	Harney, Susan M.	34	F	513	Chicago, Ill.	
5	ANDERSON, MARY J. Mandeville, La. Box 356	29	F	3502c	Mandeville, La.	
6	Henry G.	9	M		Slidell, La.	
7	Jocelyn K.	7	F		Mandeville, La.	
8	Monita	4	F		Madisonville, La.	
9	Emy L.	3	F		New Orleans, La.	
10	BARTON, DORIS R. 213 N. Penn St. Hatboro, Pa.	30	F	1213	Phila., Penn.	
11	BENNETT, SANDRA JUANITA 1923 W. 59th St. Los Angeles, Calif.	6	F	1111	Maywood, Calif.	
12	BERTRAND, HILDA 520 Surf St. Chicago, Ill. (Friend)	52	F	923	Chicago, Ill.	
13	BYRD, BONNIE M. Sasakwa, Okla.	23	F	1079	Sasakwa, Okla.	
14	CAMPBELL, L. FRANCES Greenwood, Virginia	24	F	735	Crozet, Virginia	
15	Janet L.	4	F		Charlottesville, Va.	
16	Patricia A.	2	F		Charlottesville, Va.	
17	CHANDLER, BETTY ANN 1922 N.W. Flagler Terr. Miami, Florida	20	F	Not P	Morristown, Tenn.	
18	Betty Jean	8½	F		Hendersonville, N.C.	
19	Patricia Dian	5½	F		Kingsport, Tenn.	
20	Majoria Ann	1½m	F		Hendersonville, N.C.	
21	CHAPPEL, THELMA 604 S. Elonoy Pomona, Calif.	41	F	1115	Washington, Ind.	
22	Davis, Howard	13	M		Pomona, Calif.	
23	Davis, William	10	M		Pomona, Calif.	
24	Chappel, John	3½m	M		Pomona, Calif.	FILE - G.R.V.
25	COBB, TINA 315 E. Peabody Apts. Columbus, Ga.	23	F	1613	Columbus, Ga.	

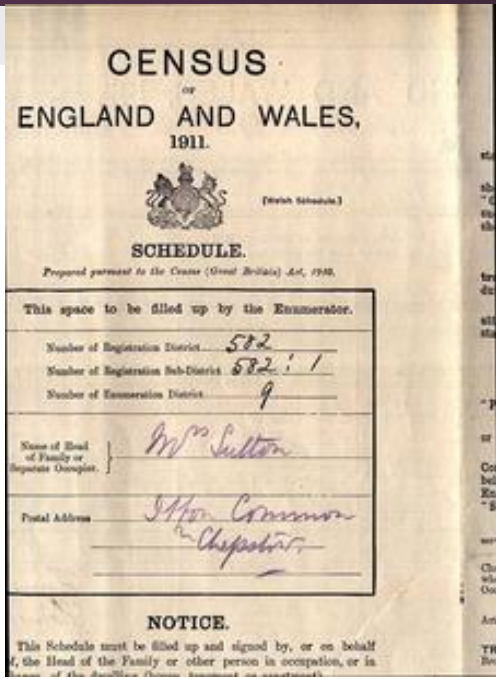
Form Image Clustering


CENSUS OF ENGLAND AND WALES, 1911.




CENSUS OF ENGLAND AND WALES, 1911.


CENSUS OF ENGLAND AND WALES, 1911.


CENSUS OF ENGLAND AND WALES, 1911.

SCHEDULE.

Prepared pursuant to the Census (Great Britain) Act, 1910.

This space to be filled up by the Enumerator.

Number of Registration District 582

Number of Registration Sub-District 582/1

Number of Enumeration District 9

Name of Head of Family or Separate Occupier. Mrs Sutton

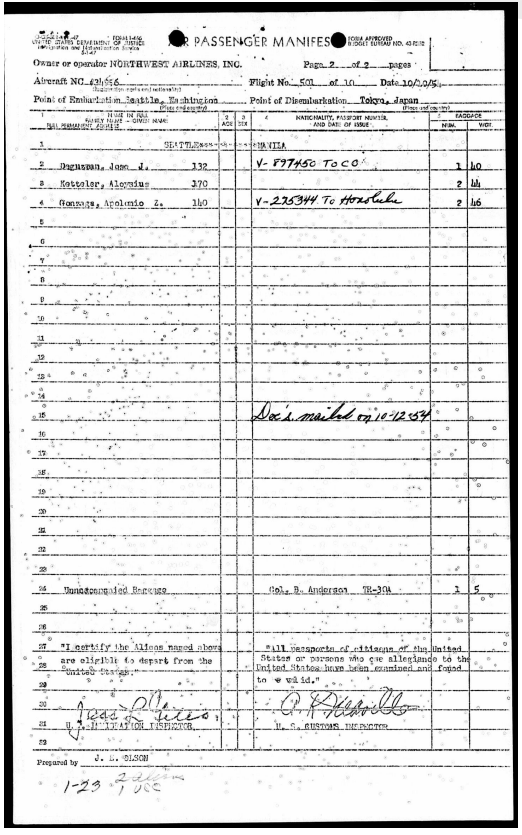
Postal Address The Common Chepstow

NOTICE.

This Schedule must be filled up and signed by, or on behalf of, the Head of the Family or other person in occupation, or in

Form Image Clustering

Washington Passenger Lists Dataset



PASSENGER MANIFEST

Owner or operator NORTHWEST AIRLINES, INC. Page 2 of 2 pages

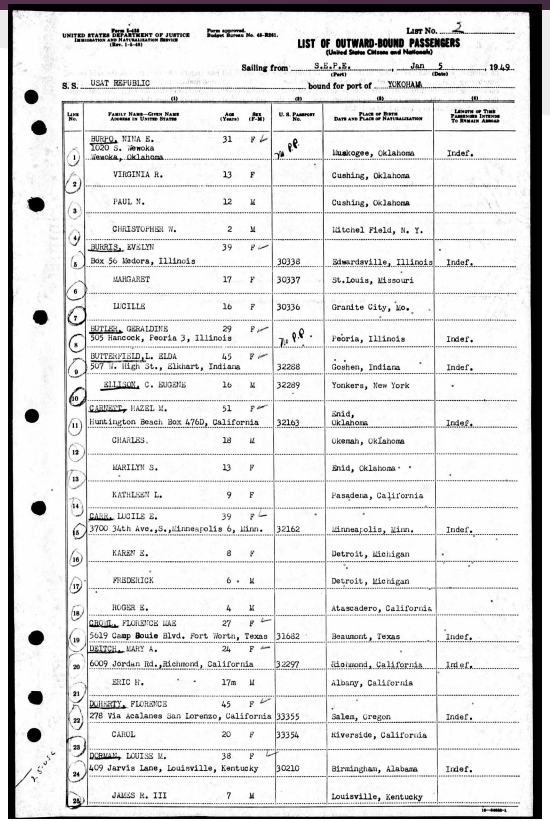
Aircraft NC 42656 Flight No. 501 of 10 Date 10/20/54

Point of Embarkation Seattle, Washington Point of Disembarkation Tokyo, Japan

	NAME	AGE	SEX	NATIONALITY, PASSPORT NUMBER, AND DATE OF ISSUE	RACE/COLOR NUM.	RACE/COLOR WGT.
1	SEATTLE			ON VISA		
2	Dagnayan, Jose J.	132		V-597450 TOCO	1	40
3	Kattoler, Alexandre	370			2	44
4	Gonsaga, Apolonio Z.	140		V-275344 To Honolulu	2	46
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15				Doc's mailed on 10-12-54		
16						
17						
18						
19						
20						
21						
22						
23						
24	Unaccompanied Baggage			Col. B. Anderson TR-304	1	5
25						
26						
27	"I certify the aliens named above			"All passports of citizens of the United		
28	are eligible to depart from the United States."			States or persons who owe allegiance to the United States have been examined and found to be valid."		
29						
30						
31	U.S. IMMIGRATION INSPECTOR			U. S. CUSTOMS INSPECTOR		
32						
33						

Prepared by J. H. OLSON

1-23



UNITED STATES DEPARTMENT OF JUSTICE — LIST OF OUTWARD-BOUND PASSENGERS (Aliens and United States Citizens and Nationals) — List No. 2

Sailing from S.F.P.E. (Port), Jan 5 (Date), 1949

S.S. USAT REPUBLIC bound for port of YOKOHAM

Line No.	Family Name—Given Name; Address in United States	Age (Years)	Sex (F or M)	U. S. Passport No.	Place of Birth; Date and Place of Naturalization	Length of Time Previously Visited To Reside Abroad
1	BURPO, NINA E. 1020 S. Muskogee Muskogee, Oklahoma	31	F	74 PP	Muskogee, Oklahoma	Indef.
2	VIRGINIA R.	13	F		Cushing, Oklahoma	
3	PAUL N.	12	M		Cushing, Oklahoma	
4	CHRISTOPHER W.	2	M		Mitchel Field, N. Y.	
5	BURRIS, EVELYN Box 56 Medora, Illinois	39	F	30338	Edwardsville, Illinois	Indef.
6	MARGARET	17	F	30337	St.Louis, Missouri	
7	LUCILLE	16	F	30336	Granite City, Mo.	
8	HUTSELL, GERALDINE 305 Hancock, Peoria 3, Illinois	29	F	74 PP	Peoria, Illinois	Indef.
9	BUTTERFIELD, L. ELDA 307 W. High St., Elkhart, Indiana	45	F	32288	Goshen, Indiana	Indef.
10	KILLION, C. EUGENE	16	M	32289	Yonkers, New York	
11	CARPENT, HAZEL M. Huntington Beach Box 476D, California	51	F	32163	Enid, Oklahoma	Indef.
12	CHARLES	18	M		Okemah, Oklahoma	
13	MARILYN S.	13	F		Enid, Oklahoma	
14	KATHLEEN L.	9	F		Pasadena, California	
15	CARR, LUCILE E. 3700 34th Ave.,S.,Minneapolis 6, Minn.	39	F	32162	Minneapolis, Minn.	Indef.
16	KAREN S.	8	F		Detroit, Michigan	
17	FREDERICK	6	M		Detroit, Michigan	
18	ROGER E.	4	M		Atascadero, California	
19	CRAIG, FLORENCE MAE 5619 Camp Boule Blvd. Fort Worth, Texas	27	F	31682	Beaumont, Texas	Indef.
20	DEITCH, MARY A. 6009 Jordan Rd.,Richmond, California	24	F	32297	Richmond, California	Indef.
21	ERIC H.	17m	M		Albany, California	
22	DOHERTY, FLORENCE 278 Via Acalanes San Lorenzo, California	45	F	33355	Salem, Oregon	Indef.
23	CAROL	20	F	33354	Riverside, California	
24	DORAN, LOUISE M. 409 Jarvis Lane, Louisville, Kentucky	38	F	30210	Birmingham, Alabama	Indef.
25	JAMES H. III	7	M		Louisville, Kentucky	

CENSUS OF ENGLAND AND WALES, 1911.

Number of Schedule 103
(To be filled up by the Enumerator after collection.)

Before writing on this Schedule please read the Examples and the Instructions given on the other side of the paper, as well as the headings of the Columns. The entries should be written in Ink.

The contents of the Schedule will be treated as confidential. Strict care will be taken that no information is disclosed with regard to individual persons. The returns are not to be used for proof of age, as in connection with Old Age Pensions, or for any other purpose than the preparation of Statistical Tables.

NAME AND SURNAME of every Person, whether Member of Family, Visitor, Boarder, or Servant, who (1) passed the night of Sunday, April 2nd, 1911, in this dwelling and was alive at midnight, or (2) arrived in this dwelling on the morning of Monday, April 3rd, not having been enumerated elsewhere. No one else must be included. (For order of entering names see Examples on back of Schedule.)	RELATIONSHIP to Head of Family. State whether "Head," or "Wife," or "Son," "Daughter," or other Relative, "Visitor," "Boarder," or "Servant."	AGE (last Birthday) and SEX. For Infants under one year state the age in months as "under one month," "one month," etc. Ages of Males.	Ages of Females.	PARTICULARS as to MARRIAGE. Write "Single," "Married," "Widower," or "Widow," opposite the names of all persons aged 15 years and upwards.	State, for each Married Woman entered on this Schedule, the number of:— Completed years the present Marriage has lasted. If less than one year write "under one."	Children born alive to present Marriage. (If no children born alive write "None" in Column 7.) Total Children Born Alive.	Children still Living.	Children who have Died.	PROFESSION or OCCUPATION of Persons aged ten years and upwards. Personal Occupation. The reply should show the precise branch of Profession, Trade, Manufacture, &c. If engaged in any Trade or Manufacture, the particular kind of work done, and the Article made or Material worked or dealt in should be clearly indicated. (See Instructions 1 to 8 and Examples on back of Schedule.)	Industry or Service with which worker is connected. This question should generally be answered by stating the business carried on by the employer. If this is clearly shown in Col. 10 the question need not be answered here. No entry needed for Domestic Servants in private employment. If employed by a public body (Government, Municipal, etc.) state what body. (See Instruction 9 and Examples on back of Schedule.)	Whether Employer, Worker, or Working on Own Account. Write opposite the name of each person engaged in any Trade or Industry, (1) "Employer" (that is employing persons other than domestic servants), or (2) "Worker" (that is working for an employer), or (3) "Own Account" (that is neither employing others nor working for a trade employer).	Whether Working at Home. Write the words "At Home" opposite the name of each person carrying on Trade or Industry at home.	BIRTHPLACE of every person. (1) If born in the United Kingdom, write the name of the County, and Town or Parish. (2) If born in any other part of the British Empire, write the name of the Dependency, Colony, etc., and of the Province or State. (3) If born in a Foreign Country, write the name of the Country. (4) If born at sea, write "At Sea." NOTE.—In the case of persons born elsewhere than in England or Wales, state whether "Resident" or "Visitor" in this Country.	NATIONALITY of every Person born in a Foreign Country. State whether:— (1) "British subject by parentage." (2) "Naturalised British subject," giving year of naturalisation. Or (3) If of foreign nationality, state whether "French," "German," "Russian," etc.	INFIRMITY. If any person included in this Schedule is:— (1) "Totally Deaf," or "Deaf and Dumb," (2) "Totally Blind," (3) "Lunatic," (4) "Imbecile," or "Feeble-minded," state the infirmity opposite that person's name, and the age at which he or she became afflicted.
1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.
1 Frederick Joyce.	Head	62		Married	27	3	2	1	Farmer	100	Employer		Cloford Somerset	140	
2 Elizabeth Ruzman Joyce	Wife		57	Married	~~27~~								Oldbury on Severn	Glos	
3 Elizabeth Mary Gladys Joyce	Daughter		25	Single					Farmers Daughter Dairy Work	11 0	a		Tidenham Glos		
4 Gwendoline Joyce	Daughter		23	Single					Farmers Daughter Dairy work		1		Tidenham Glos		
5															
6															
7															
8															
9															
10															
11															
12															
13															
14															
15															

(To be filled up by the Enumerator.)

I certify that:—
(1.) All the ages on this Schedule are entered in the proper sex columns.
(2.) I have counted the males and females in Columns 3 and 4 separately, and have compared their sum with the total number of persons.
(3.) After making the necessary enquiries I have completed all entries on the Schedule which appeared to be defective, and have corrected such as appeared to be erroneous.

Initials of Enumerator GH

Total. Males.	Females.	Persons.
1	3	4

0 ✓

(To be filled up by, or on behalf of, the Head of Family or other person in occupation, or in charge, of this dwelling.)

Write below the Number of Rooms in this Dwelling (House, Tenement, or Apartment). Count the kitchen as a room but do not count scullery, landing, lobby, closet, bathroom; nor warehouse, office, shop.

Fourteen

I declare that this Schedule is correctly filled up to the best of my knowledge and belief.

Signature Frederick Joyce

Postal Address High Hall Farm Tidenham Chepstow

CENSUS OF ENGLAND AND WALES, 1911.

Number of Schedule 57
(To be filled up by the Enumerator after collection.)

Before writing on this Schedule please read the Examples and the Instructions given on the other side of the paper, as well as the headings of the Columns. The entries should be written in Ink.

The contents of the Schedule will be treated as confidential. Strict care will be taken that no information is disclosed with regard to individual persons. The returns are not to be used for proof of age, as in connection with Old Age Pensions, or for any other purpose than the preparation of Statistical Tables.

NAME AND SURNAME of every Person, whether Member of Family, Visitor, Boarder or Servant, who (1) passed the night of Sunday, April 2nd, 1911, in this dwelling and was alive at midnight, or (2) arrived in this dwelling on the morning of Monday, April 3rd, not having been enumerated elsewhere. No one else must be included. (For order of entering names see Examples on back of Schedule.)	RELATIONSHIP to Head of Family. State whether "Head" or "Wife," "Son," "Daughter," or other Relative, "Visitor," "Boarder," or "Servant."	AGE (last Birthday) and SEX. For Infants under one year state the age in months as "under one month," "one month," etc. Ages of Males.	Ages of Females.	PARTICULARS as to MARRIAGE. Write "Single," "Married," "Widower," or "Widow," opposite the names of all persons aged 15 years and upwards.	State, for each Married Woman entered on this Schedule, the number of:— Completed years the present Marriage has lasted. If less than one year, write "under one."	Children born alive to present Marriage. (If no Children born alive write "None" in Column 7.) Total Children Born Alive.	Children still Living.	Children who have Died.	PROFESSION or OCCUPATION of Persons aged ten years and upwards. Personal Occupation. The reply should show the precise branch of Profession, Trade, Manufacture, &c. If engaged in any Trade or Manufacture, the particular kind of work done, and the Article made or Material worked or dealt in should be clearly indicated. (See Instructions 1 to 8 and Examples on back of Schedule.)	Industry or Service with which worker is connected. This question should generally be answered by stating the business carried on by the employer. If this is clearly shown in Col. 10 the question need not be answered here. No entry needed for Domestic Servants in private employment. If employed by a public body (Government, Municipal, etc.) state what body. (See Instruction 9 and Examples on back of Schedule.)	Whether Employer, Worker, or Working on Own Account. Write opposite the name of each person engaged in any Trade or Industry, (1) "Employer" (that is employing persons other than domestic servants), or (2) "Worker" (that is working for an employer), or (3) "Own Account" (that is neither employing others nor working for a trade employer).	Whether Working at Home. Write the words "At Home" opposite the name of each person carrying on Trade or Industry at home.	BIRTHPLACE of every Person. (1) If born in the United Kingdom, write the name of the County, and Town, or Parish. (2) If born in any other part of the British Empire, write the name of the Dependency, Colony, etc., and of the Province or State. (3) If born in a Foreign Country, write the name of the Country. (4) If born at sea, write "At Sea." Note.—In the case of persons born elsewhere than in England or Wales, state whether "Resident" or "Visitor" in this Country.	NATIONALITY of every Person born in a Foreign Country. State whether:— (1) "British subject by parentage." (2) "Naturalised British subject," giving year of naturalisation. Or (3) If of foreign nationality, state whether "French," "German," "Russian," etc.	INFIRMITY. If any person included in this Schedule is:— (1) "Totally Deaf" or "Deaf and Dumb," (2) "Totally Blind," (3) "Lunatic," (4) "Imbecile" or "Feeble-minded," state the infirmity opposite that person's name, and the age at which he or she became afflicted	LANGUAGE SPOKEN. (1) If able to speak English only, write "English." (2) If able to speak Welsh only, write "Welsh." (3) If able to speak English and Welsh, write "Both." No entry to be made in this column for children under three years of age.
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	Thomas Williams	Head		71	Married	~~38~~	~~7~~	~~7~~		Plasterer	436		0	At Home	Shirenewton, Mon.			English
2	Rhoda Ann Williams	Wife		68	Married	38	7	7		Midwife	434		1		Newport Mon	481		English
3	Eliza Josephine Williams	Daughter		20	Single									At Home	Shirenewton. Mon			English
4																		
5																		
6																		
7																		
8																		
9																		
10																		
11																		
12																		
13																		
14																		
15																		

9

(To be filled up by the Enumerator.)

I certify that:—
(1.) All the ages on this Schedule are entered in the proper sex columns.
(2.) I have counted the males and females in Columns 3 and 4 separately, and have compared their sum with the total number of persons.
(3.) After making the necessary enquiries I have completed all entries on the Schedule which appeared to be defective, and have corrected such as appeared to be erroneous

Initials of Enumerator. J P D

	Total.		
	Males.	Females.	Persons.
	1	2	3

0

(To be filled up by, or on behalf of, the Head of Family or other person in occupation, or in charge, of this dwelling.)

Write below the Number of Rooms in this Dwelling (House, Tenement or Apartment). Count the kitchen as a room but do not count scullery, landing, lobby, closet, bathroom; nor warehouse, office, shop.

Four 4

I declare that this Schedule is correctly filled up to the best of my knowledge and belief.

Signature Thomas Williams

Postal Address Rose Cottage, Carlwood, Lamwin McAston Mon

4

6

CENSUS OF ENGLAND AND WALES, 1911.

Number of Schedule 123
(To be filled up by the Enumerator after collection.)

Before writing on this Schedule please read the Examples and the Instructions given on the other side of the paper, as well as the headings of the Columns. The entries should be written in Ink.

The contents of the Schedule will be treated as confidential. Strict care will be taken that no information is disclosed with regard to individual persons. The returns are not to be used for proof of age, as in connection with Old Age Pensions, or for any other purpose than the preparation of Statistical Tables.

NAME AND SURNAME	RELATIONSHIP to Head of Family.	AGE (last Birthday) and SEX.		PARTICULARS as to MARRIAGE.					PROFESSION or OCCUPATION of Persons aged ten years and upwards.				BIRTHPLACE of every person.	NATIONALITY of every Person born in a Foreign Country.	INFIRMITY.
		For Infants under one year state the age in months as "under one month," "one month," etc.		Write "Single," "Married," "Widower," or "Widow," opposite the names of all persons aged 15 years and upwards.	State, for each Married Woman entered on this Schedule, the number of:—				Personal Occupation.	Industry or Service with which worker is connected.	Whether Employer, Worker, or Working on Own Account.	Whether Working at Home.			
of every Person, whether Member of Family, Visitor, Boarder, or Servant, who (1) passed the night of Sunday, April 2nd, 1911, in this dwelling and was alive at midnight, or (2) arrived in this dwelling on the morning of Monday, April 3rd, not having been enumerated elsewhere. No one else must be included. (For order of entering names see Examples on back of Schedule.)	State whether "Head," or "Wife," "Son," "Daughter," or other Relative, "Visitor," "Boarder," or "Servant."	Ages of Males.	Ages of Females.		Completed years the present Marriage has lasted. If less than one year write "under one."	Children born alive to present Marriage. (If no children born alive write "None" in Column 7.) Total Children Born Alive.	Children still Living.	Children who have Died.	The reply should show the precise branch of Profession, Trade, Manufacture, &c. If engaged in any Trade or Manufacture, the particular kind of work done, and the Article made or Material worked or dealt in should be clearly indicated. (See Instructions 1 to 8 and Examples on back of Schedule.)	This question should generally be answered by stating the business carried on by the employer. If this is clearly shown in Col. 10 the question need not be answered here. No entry needed for Domestic Servants in private employment. If employed by a public body (Government, Municipal, etc.) state what body. (See Instruction 9 and Examples on back of Schedule.)	Write opposite the name of each person engaged in any Trade or Industry, (1) "Employer" (that is employing persons other than domestic servants), or (2) "Worker" (that is working for an employer), or (3) "Own Account" (that is neither employing others nor working for a trade employer).	Write the words "At Home" opposite the name of each person carrying on Trade or Industry at home.	(1) If born in the United Kingdom, write the name of the County, and Town or Parish. (2) If born in any other part of the British Empire, write the name of the Dependency, Colony, etc., and of the Province or State. (3) If born in a Foreign Country, write the name of the Country. (4) If born at sea, write "At Sea." NOTE.—In the case of persons born elsewhere than in England or Wales, state whether "Resident" or "Visitor" in this Country.	State whether:— (1) "British subject by parentage," (2) "Naturalised British subject," giving year of naturalisation. Or (3) If of foreign nationality, state whether "French," "German," "Russian," etc.	If any person included in this Schedule is:— (1) "Totally Deaf," or "Deaf and Dumb," (2) "Totally Blind," (3) "Lunatic," (4) "Imbecile," or "Feeble-minded," state the infirmity opposite that person's name, and the age at which he or she became afflicted.
1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.

CENSUS OF ENGLAND AND WALES, 1911.

Number of Schedule 57
(To be filled up by the Enumerator after collection.)

Before writing on this Schedule please read the Examples and the Instructions given on the other side of the paper, as well as the headings of the Columns. The entries should be written in Ink.

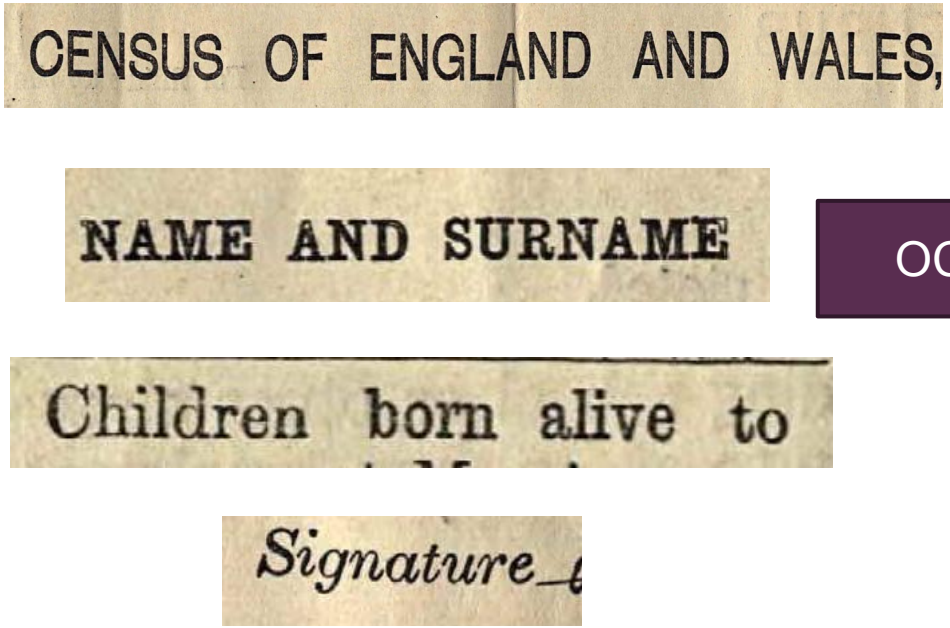
The contents of the Schedule will be treated as confidential. Strict care will be taken that no information is disclosed with regard to individual persons. The returns are not to be used for proof of age, as in connection with Old Age Pensions, or for any other purpose than the preparation of Statistical Tables.

NAME AND SURNAME	RELATIONSHIP to Head of Family.	AGE (last Birthday) and SEX.		PARTICULARS as to MARRIAGE.					PROFESSION or OCCUPATION of Persons aged ten years and upwards.				BIRTHPLACE of every Person.	NATIONALITY of every Person born in a Foreign Country.	INFIRMITY.	LANGUAGE SPOKEN.
		For Infants under one year state the age in months as "under one month," "one month," etc.		Write "Single," "Married," "Widower," or "Widow," opposite the names of all persons aged 15 years and upwards.	State, for each Married Woman entered on this Schedule, the number of:—				Personal Occupation.	Industry or Service with which worker is connected.	Whether Employer, Worker, or Working on Own Account.	Whether Working at Home.				
of every Person, whether Member of Family, Visitor, Boarder or Servant, who (1) passed the night of Sunday, April 2nd, 1911, in this dwelling and was alive at midnight, or (2) arrived in this dwelling on the morning of Monday, April 3rd, not having been enumerated elsewhere. No one else must be included. (For order of entering names see Examples on back of Schedule.)	State whether "Head" or "Wife," "Son," "Daughter," or other Relative, "Visitor," "Boarder," or "Servant."	Ages of Males.	Ages of Females.		Completed years the present Marriage has lasted. If less than one year, write "under one."	Children born alive to present Marriage. (If no Children born alive write "None" in Column 7.) Total Children Born Alive.	Children still Living.	Children who have Died.	The reply should show the precise branch of Profession, Trade, Manufacture, &c. If engaged in any Trade or Manufacture, the particular kind of work done, and the Article made or Material worked or dealt in should be clearly indicated. (See Instructions 1 to 8 and Examples on back of Schedule.)	This question should generally be answered by stating the business carried on by the employer. If this is clearly shown in Col. 10 the question need not be answered here. No entry needed for Domestic Servants in private employment. If employed by a public body (Government, Municipal, etc.) state what body. (See Instruction 9 and Examples on back of Schedule.)	Write opposite the name of each person engaged in any Trade or Industry, (1) "Employer" (that is employing persons other than domestic servants), or (2) "Worker" (that is working for an employer), or (3) "Own Account" (that is neither employing others nor working for a trade employer).	Write the words "At Home" opposite the name of each person carrying on Trade or Industry at home.	(1) If born in the United Kingdom, write the name of the County, and Town, or Parish. (2) If born in any other part of the British Empire, write the name of the Dependency, Colony, etc., and of the Province or State. (3) If born in a Foreign Country, write the name of the Country. (4) If born at sea, write "At Sea." NOTE.—In the case of persons born elsewhere than in England or Wales, state whether "Resident" or "Visitor" in this Country.	State whether:— (1) "British subject by parentage," (2) "Naturalised British subject," giving year of naturalisation. Or (3) If of foreign nationality, state whether "French," "German," "Russian," etc.	If any person included in this Schedule is:— (1) "Totally Deaf" or "Deaf and Dumb," (2) "Totally Blind," (3) "Lunatic," (4) "Imbecile or "Feeble minded," state the infirmity opposite that person's name, and the age at which he or she became afflicted.	(1) If able to speak English only, write "*English*." (2) If able to speak Welsh only, write "*Welsh*." (3) If able to speak English and Welsh, write "*Both*." No entry to be made in this column for children under three years of age.
1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.

CONFIRM

- Form Specific Features
  - OCR Text
  - Rule Lines
- Matching algorithms
- Pick T random forms as templates
- Encode forms by matching with templates
- Learn form-form similarity metric

Extract Text Features



OCR

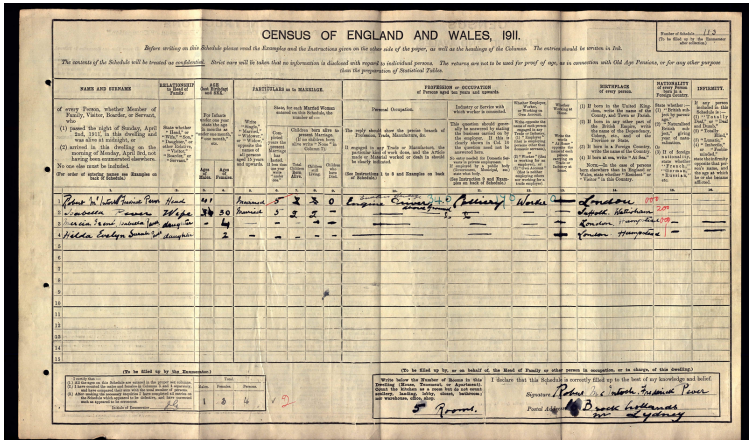
Noisy
Transcription of Text

“CENSUS OF ENGLAND AND WALES”
“NAME AND SURNAME”
“Children bom alive to”
“Signature”

Text Matching

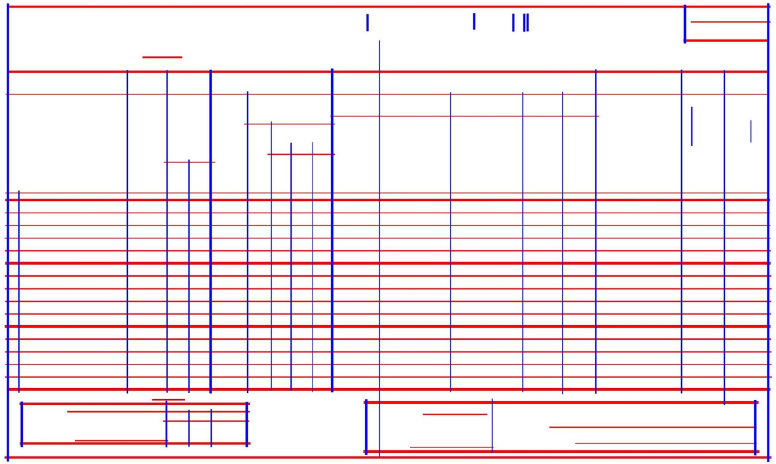
- Correspondence problem
- Do greedy matching
  - Similar location and content
  - Prefix/Suffix matching
- Edit Distance
  - Number of character edits to transform one string into another.

Extract Rule Lines



CENSUS OF ENGLAND AND WALES, 1911.

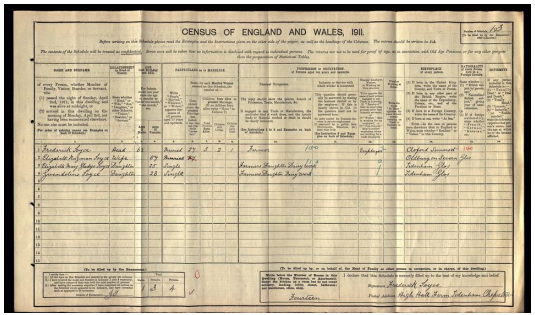
Line Extraction



Rule Line Matching

- Horz/Vert as sequences of parallel lines
- Edit Distance between sequences
  - Also shows how to merge two sequences
- Operations include
  - Match Deletion
  - Contains Overlap
  - Transpose Connect

Matching Vector



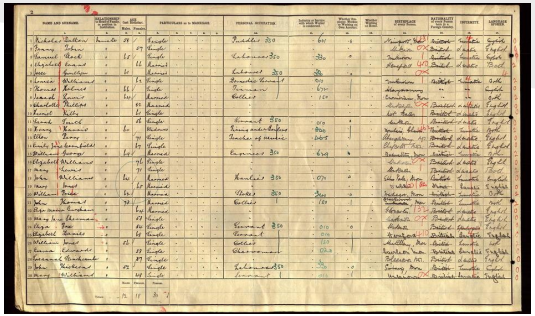
Query Form

Match Vectors

(0, 1, 1, …)

(1, 1, 0.5, 0.75, …)

(0, 0, 0.25, 0, …)



Template Form

(0, 1, 1, …, 1, 1, 0.5, 0.75, …, 0, 0, 0.25, 0, …)

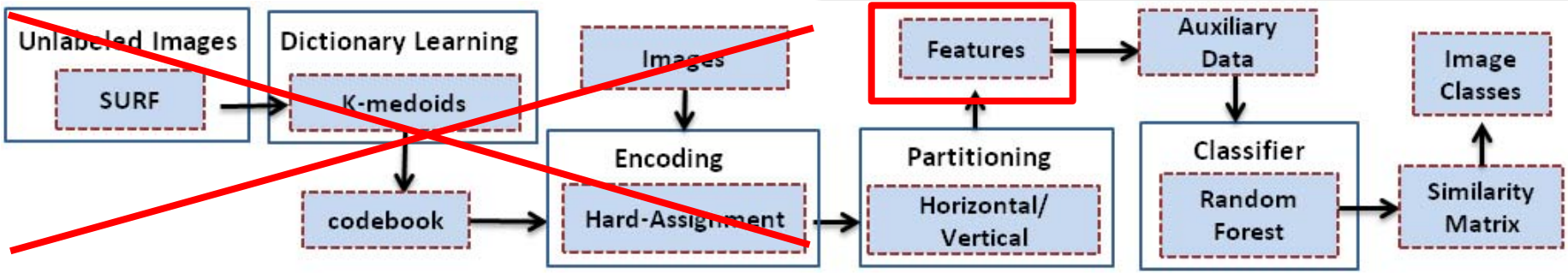
New Feature Vector

Text

Horizontal Lines

Vertical Lines

Kumar and Doermann



- Create codebook of SURFs
- Encode images as histogram of codewords
- Learn similarity matrix using Random Forest

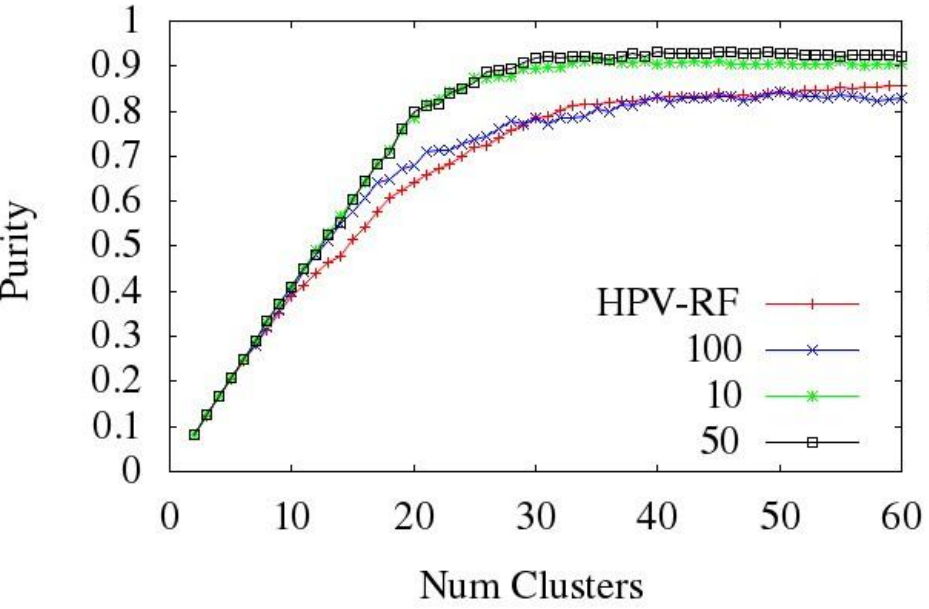
http://docs.opencv.org/trunk/doc/py_tutorials/py_feature2d/py_surf_intro/py_surf_intro.html

Experiment

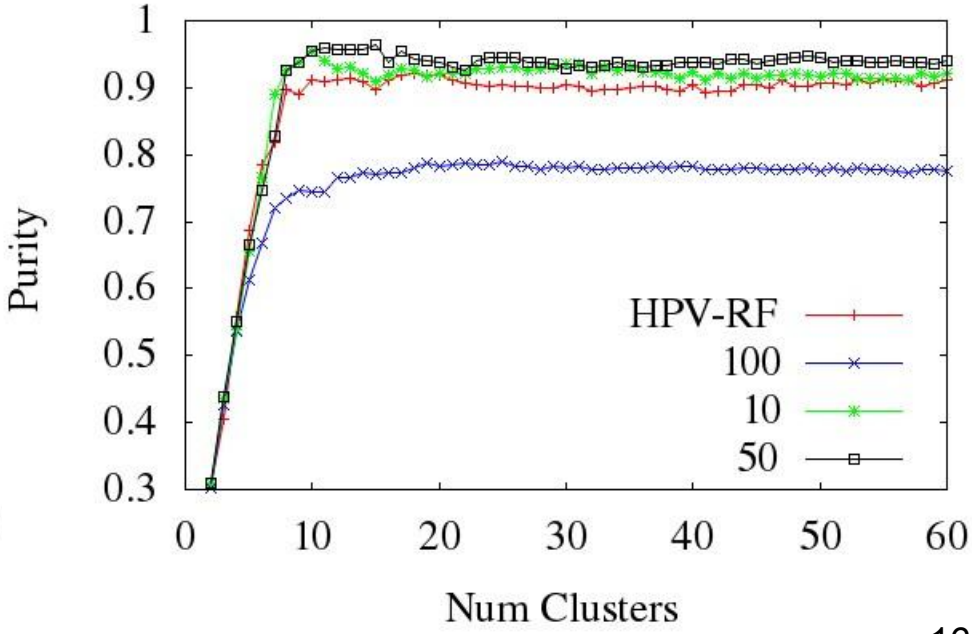
- CONFIRM vs Kumar & Doermann
  - Vary number of random templates
  - Vary types among random templates
- Historical Datasets
  - Census
  - Death Certificates
  - Passenger Lists
- Metric: Purity

Results - Census

Balanced

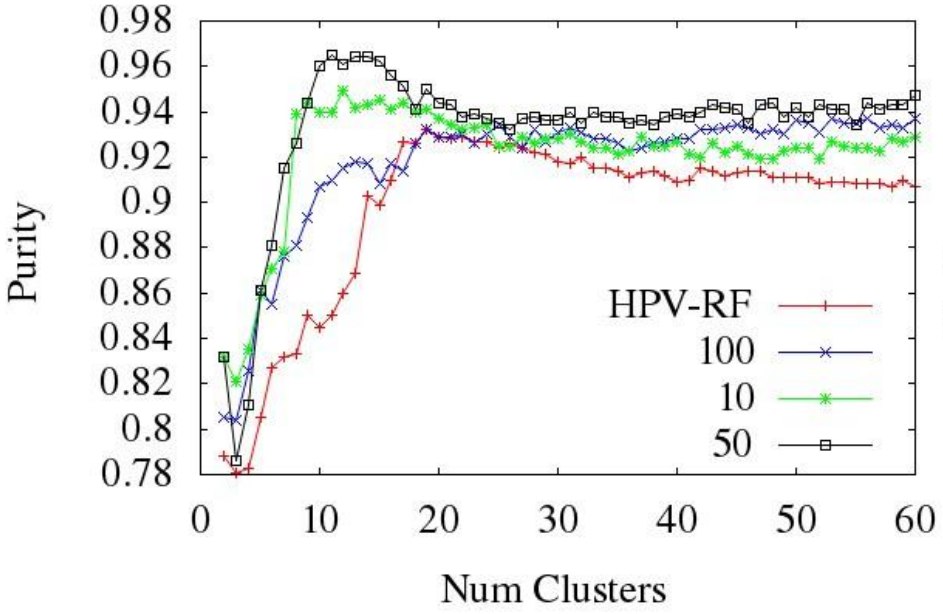


Skewed

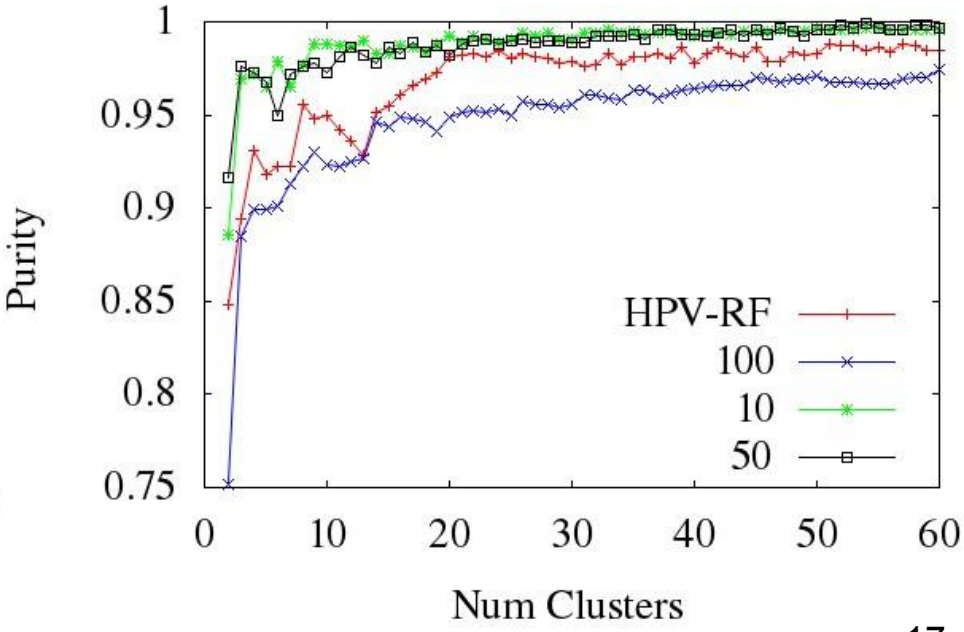


Results - Cont

Death Certificates

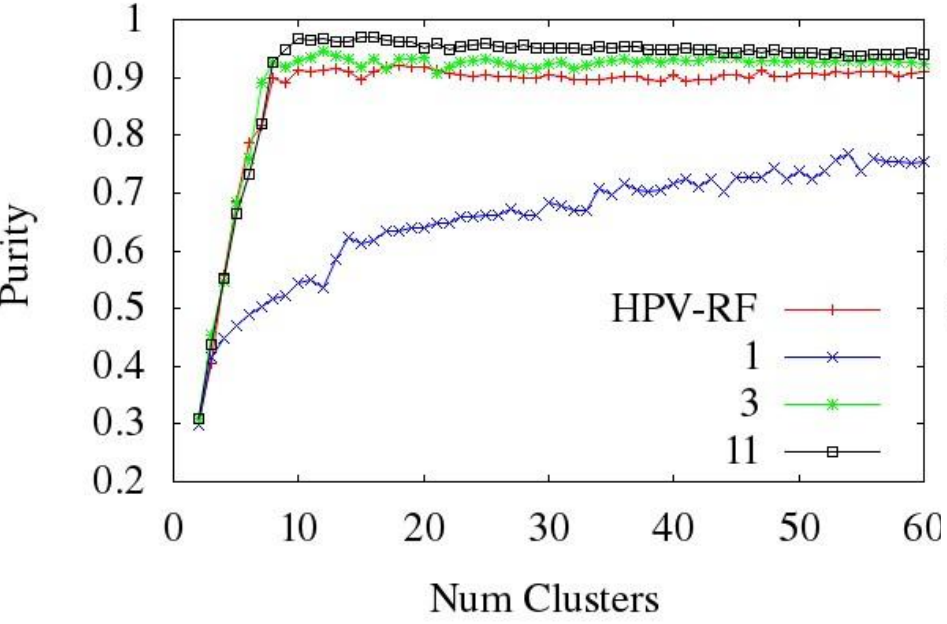


Passenger Lists

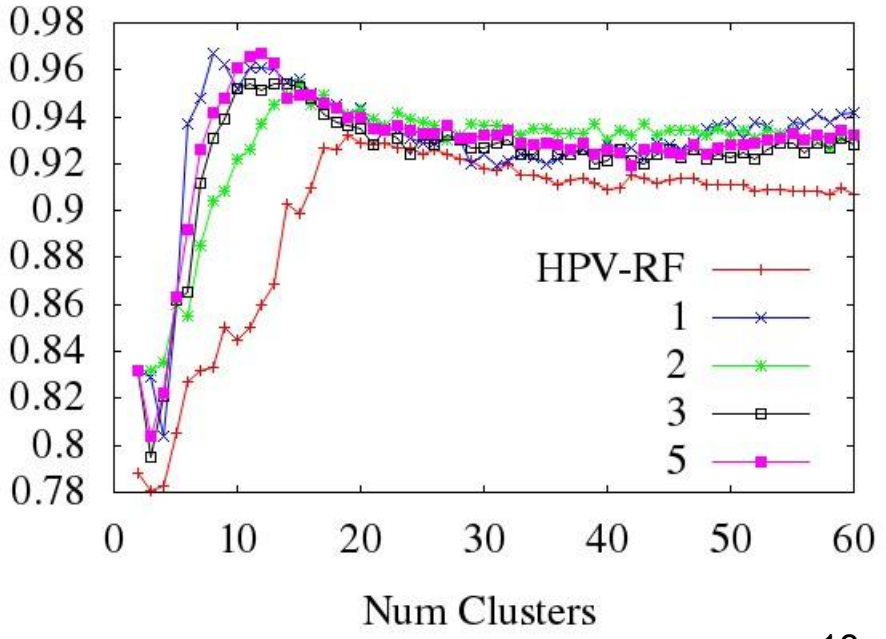


Results - Num Types

Census - 11 total types



Death Certs - 5 total types

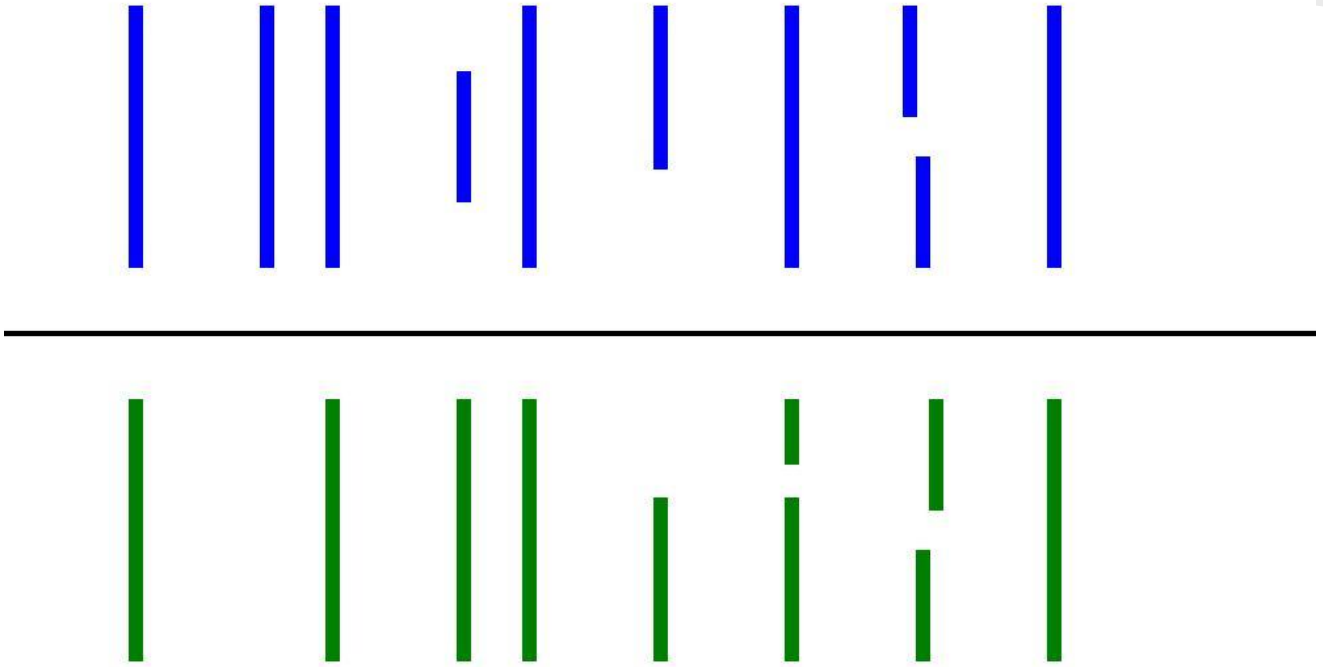


Future Work

- Detect and split impure clusters
- Bootstrap to classification
- Scale to millions of images
- Other features

Questions

Line Edit Distance - Aligned



Line Edit Distance - Offset