Automatic Methods for Coding Historical Occupation Descriptions to Standard Classifications

Graham Kirby, Jamie Carson, Fraser Dunlop, Chris Dibben, Alan Dearle, Lee Williamson, Eilidh Garrett, Alice Reid

digitisingscotland@lscs.ac.uk

digitisingscotland.cs.st-andrews.ac.uk

University of St Andrews

THE UNIVERSITY *of* EDINBURGH

UNIVERSITY OF CAMBRIDGE

digitising SCOTLAND understanding Scotland's people

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

wellcome trust

Motivation

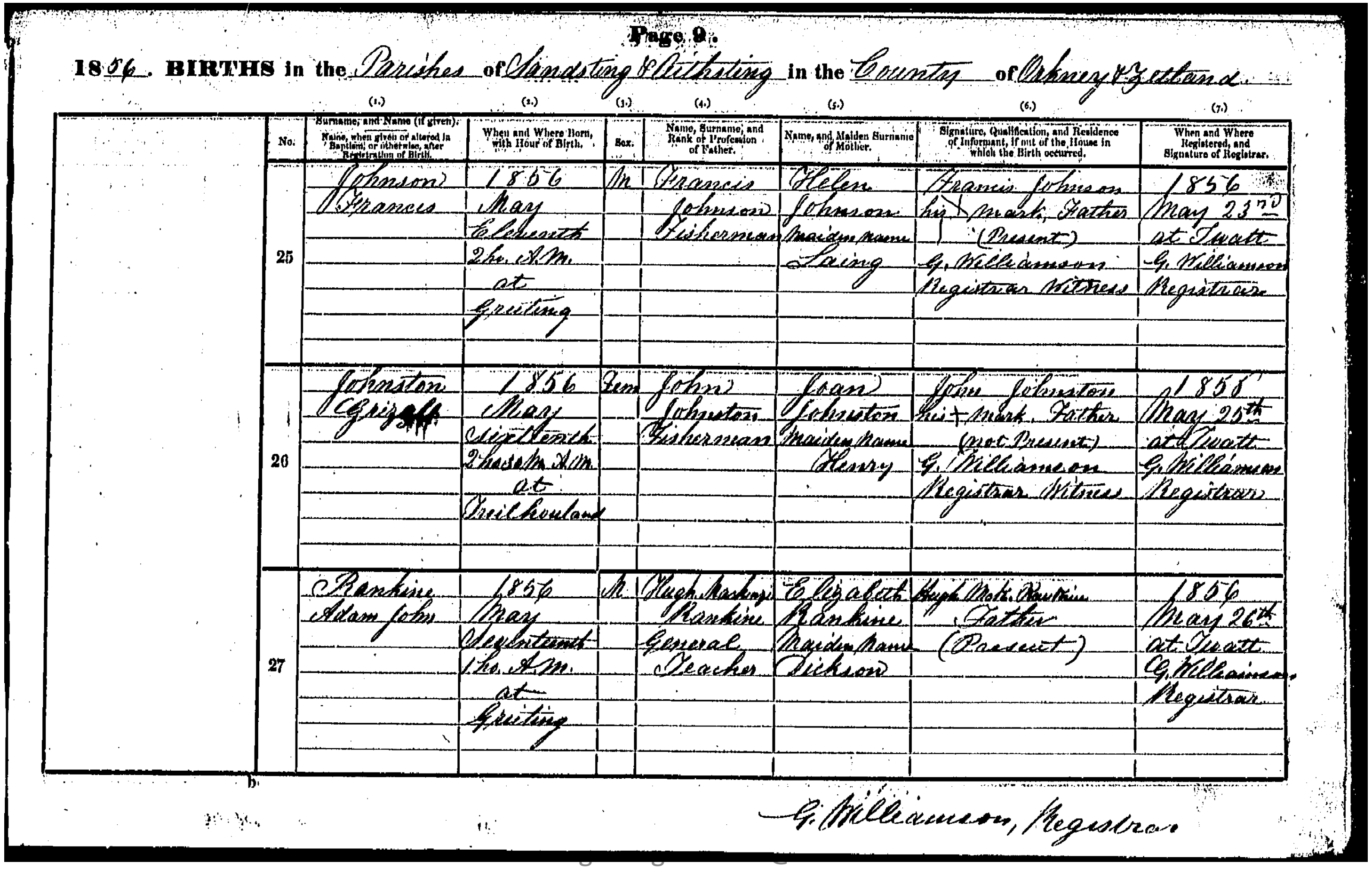
- Increasing number of digitised registration records for the 19th and 20th centuries.
- Varying forms of data
- Scale of data prevents manual analysis

Challenges

- Significant methodological issues:
  – How can we consistently code occupational data so that researchers can explore changing patterns and trends?
  – How can we automate this process so that the majority of records do not need to be manually coded?

Digitising Scotland

- Records of births, marriages and deaths recorded in Scotland from 1855 to present day.



Page 9.

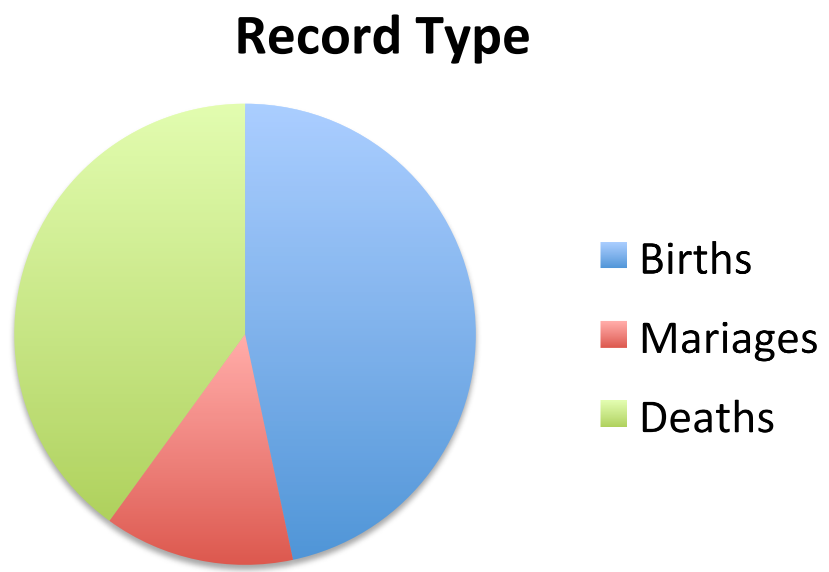
1856. BIRTHS in the Parishes of Sandsting & Aithsting in the County of Orkney & Zetland.

No.	(1.) Surname, and Name (if given). Name, when given or altered in Baptism, or otherwise, after Registration of Birth.	(2.) When and Where Born, with Hour of Birth.	(3.) Sex.	(4.) Name, Surname, and Rank or Profession of Father.	(5.) Name, and Maiden Surname of Mother.	(6.) Signature, Qualification, and Residence of Informant, if out of the House in which the Birth occurred.	(7.) When and Where Registered, and Signature of Registrar.
25	Johnson Francis	1856 May Eleventh 2h. A.M. at Greiting	M	Francis Johnson Fisherman	Helen Johnson Maiden name Laing	Francis Johnson his + mark, Father (Present) G. Williamson Registrar Witness	1856 May 23rd at Twatt G. Williamson Registrar
26	Johnston Grizell	1856 May Sixteenth 2h. 30m. A.M. at Troilhoulan	Fem	John Johnston Fisherman	Joan Johnston Maiden name Henry	John Johnston his + mark. Father (not Present) G. Williamson Registrar Witness	1856 May 25th at Twatt G. Williamson Registrar
27	Rankine Adam John	1856 May Seventeenth 1h. A.M. at Greiting	M	Hugh Mackenzie Rankine General Teacher	Elizabeth Rankine Maiden name Dickson	Hugh Mck. Rankine Father (Present)	1856 May 26th at Twatt G. Williamson Registrar

G. Williamson, Registrar

Digitising Scotland

- Approximately 29 million records
- Approximately 50 million occupation strings, 8 million causes of death
- Classify occupations to Historical International Standard Classification of Occupations (HISCO)
- Cause of death to ICD10



MARRIAGE | District No. 107 | Year 1972 | Entry No. 1

IN THE DISTRICT OF Nesting

1. When and where married
19.72 December ~~Eigth~~ Eighth
The Manse, Nesting

	BRIDEGROOM	BRIDE
Surname	Williamson	Pottinger
Name(s)	George Angus	Agnes Anne
	(Signed) George A Williamson	(Signed) Agnes Anne Pottinger
3. Occupation	Civil Servant	Farm Worker
4. Marital status	Bachelor	Spinster
5. Date of birth Year / Month / Day	~~1951~~ 1951 / 6 / 17	1952 / 7 / 26
6. Birthplace	~~Sellivoe Bridge of Walls~~ Lerwick	Benston
7. Usual residence	Selivoe, Bridge of Walls	Benston, Nesting
8. Father's name(s) surname and occupation	—	John Charles Pottinger Farmer

MARRIAGE	District No. 107	Year 1972	Entry No. 1

IN THE DISTRICT OF Nesting

1. When and where married
19.72 December ~~Eigth~~ Eighth
The Manse, Nesting

	BRIDEGROOM	BRIDE
2. Surname	Williamson	Pottinger
Name(s)	George Angus	Agnes Anne
(Signed)	George A Williamson	Agnes Anne Pottinger
3. Occupation	Civil Servant	Farm Worker
4. Marital status	Bachelor	Spinster
5. Date of birth — Year	~~1951~~ 1951	1942
Month	6	7
Day	17	26
6. Birthplace	~~[illegible]~~ ~~Bridge of Walls~~ Lerwick	Benston,
7. Usual residence	Bellevue Bridge-of-Walls	Benston, Nesting
8. Father's name(s) surname and occupation	— —	John Charles Pottinger Farmer

MARRIAGE	District No. 107	Year 1972	Entry No. 1

IN THE DISTRICT OF Nesting

1. When and where married
19.72 December ~~Eigth~~ Eighth
The Manse, Nesting

	BRIDEGROOM	BRIDE
Surname	Williamson	Pottinger
Name(s)	George Angus	Agnes Anne
	(Signed) George A Williamson	(Signed) Agnes Anne Pottinger
3. Occupation	Civil Servant	Farm Worker
4. Marital status	Bachelor	4. Spinster
5. Date of birth (Year / Month / Day)	~~1951~~ 1951 / 6 / 17	5. 1942 / 7 / 26
6. Birthplace	~~Lerwick~~ ~~Bridge of Walls~~ Zerwick	Benston,
7. Usual residence	Selivoe, Bridge-of-Walls	Benston, Nesting
8. Father's name(s) surname and occupation	— —	John Charles Pottinger Farmer

MARRIAGE | District No. 107 | Year 1972 | Entry No. 1

IN THE DISTRICT OF Nesting

1. When and where married
19.72 December ~~Eigth~~ Eighth
The Manse, Nesting

	BRIDEGROOM	BRIDE
Surname	Williamson	Pottinger
Name(s)	George Angus	Agnes Anne
	(Signed) George A Williamson	(Signed) Agnes Anne Pottinger
3. Occupation	Civil Servant	Farm Worker
4. Marital status	Bachelor	Spinster
5. Date of birth (Year / Month / Day)	~~1951~~ 1951 / 6 / 17	1942 / 7 / 26
6. Birthplace	~~[illegible] Bridge of Walls~~ Lerwick	Benston,
7. Usual residence	Belivoe, Bridge of Walls	Benston, Nesting
8. Father's name(s) surname and occupation	—	John Charles Pottinger Farmer

MARRIAGE District No. 107 Year 1972 Entry No. 1

IN THE DISTRICT OF Nesting

1. When and where married

19.72 December ~~Eigth~~ Eighth

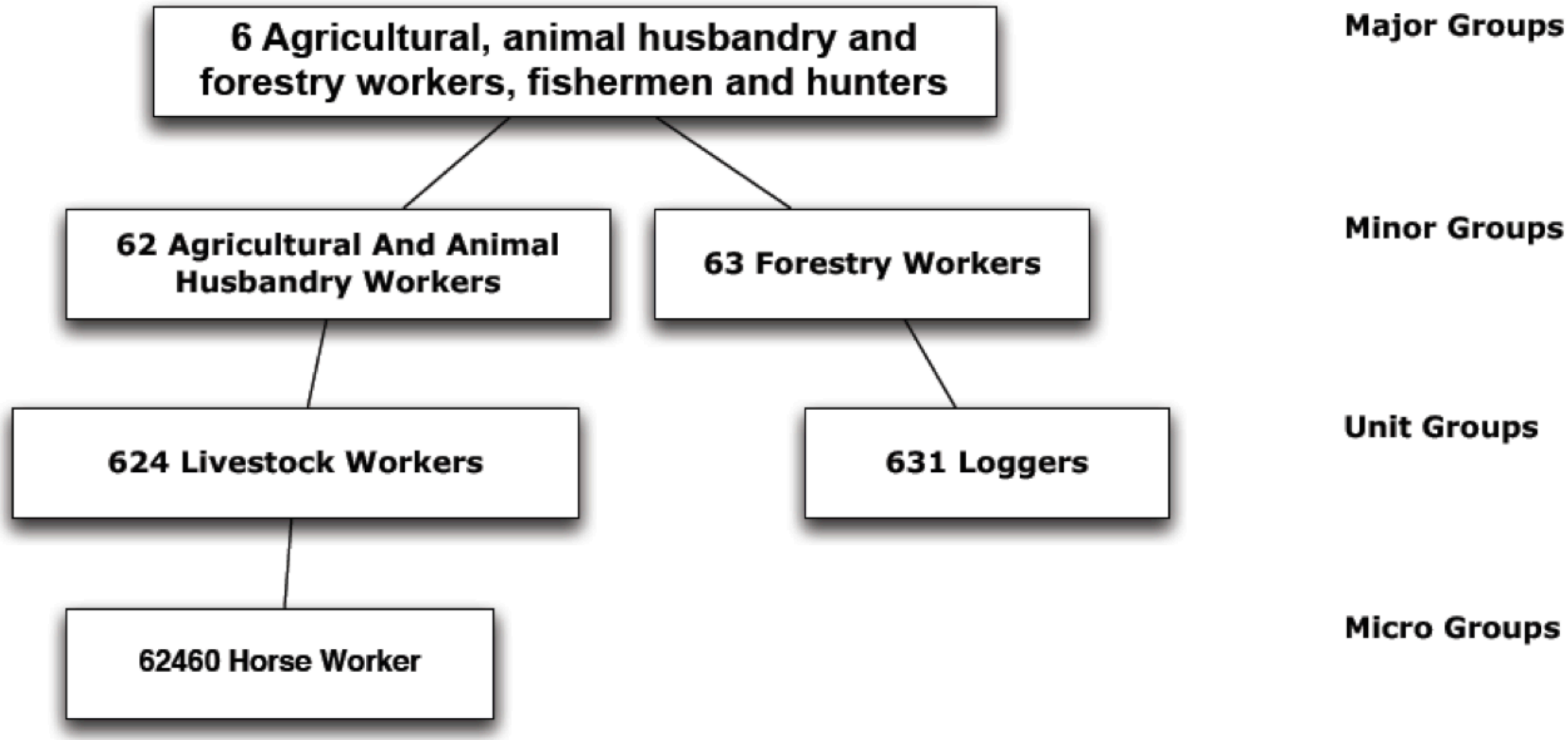
The Manse, Nesting

	BRIDEGROOM	BRIDE
Surname	Williamson	Pottinger
Name(s)	George Angus	Agnes Anne
(Signed)	George A Williamson	Agnes Anne Pottinger
3. Occupation	Civil Servant	Farm Worker
4. Marital status	Bachelor	Spinster
5. Date of birth (Year / Month / Day)	~~1951~~ 1951 / 6 / 17	1952 / 7 / 26
6. Birthplace	~~[illegible] Bridge of Walls~~ Zetrick	Benston,
7. Usual residence	Selivoe, Bridge of Walls	Benston, Nesting
8. Father's name(s) surname and occupation	—	John Charles Pottinger Farmer

Experimental Dataset

- Vital event records currently being transcribed
- Use a dataset with similar content for experiments
- 60,000 records from the Cambridge Family History Study (records from 1800-1990)
- Occupation descriptions and associated HISCO codes
- HISCO coding done by historians
- Dataset contains 330 different HISCO codes

HISCO Hierarchy Example



Classification Example

String from record	Gold Standard Classification	Automatic Classification Output
Farm horseman	62460	62460
Shoe maker	80110	80110
Fireman (railway)	98330	98330
Fireman	58100	58100
Stationer	41000	91000

Classification Example

String from record	Gold Standard Classification	Automatic Classification Output
Farm horseman	62460 Horse Worker	62460 Horse Worker
Shoe maker	80110 Shoemaker, General	80110 Shoemaker, General
Fireman (railway)	98330 Railway Steam-Engine Fireman	98330 Railway Steam-Engine Fireman
Fireman	58100 Fire-Fighter	58100 Fire-Fighter
Stationer	41000 Working Proprietors (Wholesale and Retail Trade)	91000 Paper and Paperboard product makers

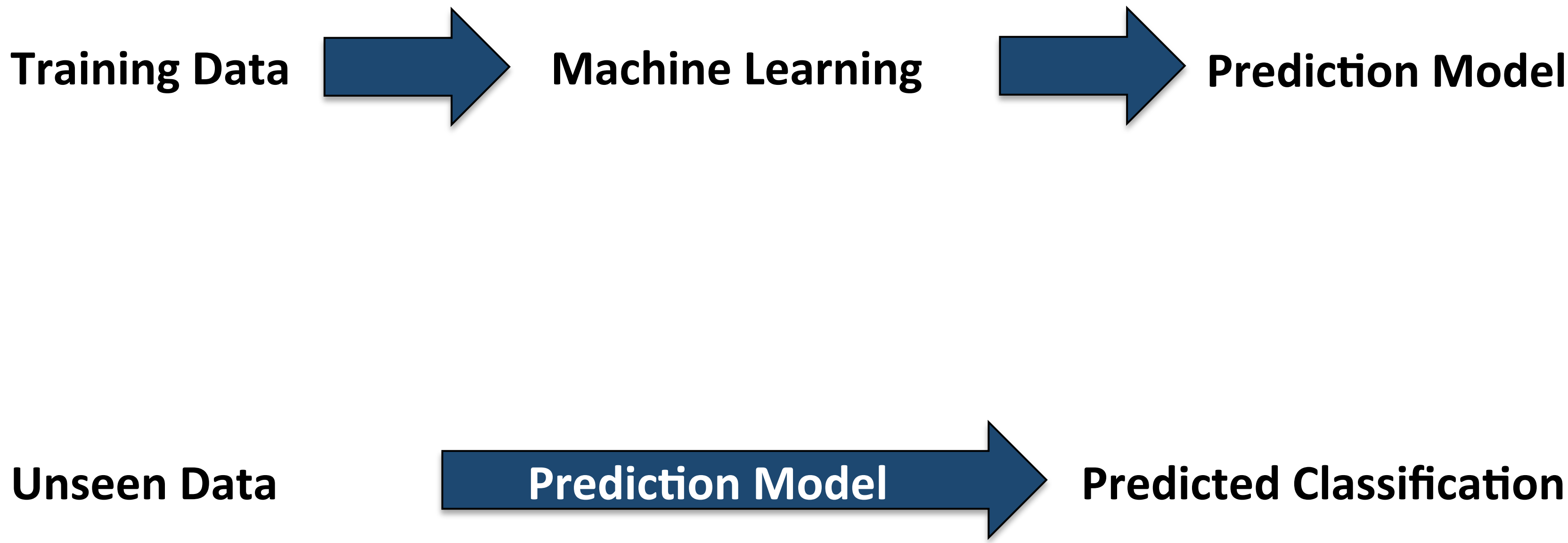
Classification Example

String from record	Gold Standard Classification	Automatic Classification Output
Farm horseman	62460 Horse Worker	62460 Horse Worker
Shoe maker	80110 Shoemaker, General	80110 Shoemaker, General
Fireman (railway)	98330 Railway Steam-Engine Fireman	98330 Railway Steam-Engine Fireman
Fireman	58100 Fire-Fighter	58100 Fire-Fighter
Stationer	41000 Working Proprietors (Wholesale and Retail Trade)	91000 Paper and Paperboard product makers

Approach

- Text analysis
- Supervised machine learning
  - Apache Mahout framework.
- Combination of these techniques.

Supervised Machine Learning



Supervised Machine Learning

Training Data → **Machine Learning** → **Prediction Model**

Farm horseman	62460
Shoe maker	80110
Fireman	58100
Stationer	41000

Unseen Data → **Prediction Model** → **Predicted Classification**

Supervised Machine Learning

Training Data → **Machine Learning** → **Prediction Model**

Farm horseman	62460
Shoe maker	80110
Fireman	58100
Stationer	41000

Unseen Data → **Prediction Model** → **Predicted Classification**

Farm horseman
Boot maker
Fireman
Painter

Supervised Machine Learning

Training Data → **Machine Learning** → **Prediction Model**

Farm horseman	62460
Shoe maker	80110
Fireman	58100
Stationer	41000

Unseen Data → **Prediction Model** → **Predicted Classification**

Farm horseman
Boot maker
Fireman
Painter

?

Machine Learning

- Inputs are split into features and converted to high dimension vectors

Record	Original Input	Cleaned input	Vector
A	Boot and shoe maker		
B	Boot and shoe dealer		
C	Fireman		
D	Cattle (& sheep) farmer		

Machine Learning

- Inputs are split into features and converted to high dimension vectors

Record	Original Input	Cleaned input	Vector
A	Boot **and** shoe maker	boot shoe maker	
B	Boot **and** shoe dealer	boot shoe dealer	
C	Fireman	fireman	
D	Cattle **(&** sheep**)** farmer	cattle sheep farmer	

Machine Learning

- Inputs are split into features and converted to high dimension vectors

Record	Original Input	Cleaned input	Vector
A	Boot **and** shoe maker	boot shoe maker	
B	Boot **and** shoe dealer	boot shoe dealer	
C	Fireman	fireman	
D	Cattle **(&** sheep**)** farmer	cattle sheep farmer	

	boot	cattle	dealer	farmer	fireman	horse	maker	sheep	shoe
A									
B									
C									
D									

Machine Learning

- Inputs are split into features and converted to high dimension vectors

Record	Original Input	Cleaned input	Vector
A	Boot **and** shoe maker	**boot** shoe maker	
B	Boot **and** shoe dealer	boot shoe dealer	
C	Fireman	fireman	
D	Cattle **(&** sheep**)** farmer	cattle sheep farmer	

	boot	cattle	dealer	farmer	fireman	horse	maker	sheep	shoe
A	**1**								
B									
C									
D									

Machine Learning

- Inputs are split into features and converted to high dimension vectors

Record	Original Input	Cleaned input	Vector
A	Boot **and** shoe maker	boot shoe maker	
B	Boot **and** shoe dealer	boot shoe dealer	
C	Fireman	fireman	
D	Cattle **(&** sheep**)** farmer	cattle sheep farmer	

	boot	cattle	dealer	farmer	fireman	horse	maker	sheep	shoe
A	1	0							
B									
C									
D									

Machine Learning

- Inputs are split into features and converted to high dimension vectors

Record	Original Input	Cleaned input	Vector
A	Boot **and** shoe maker	boot shoe maker	
B	Boot **and** shoe dealer	boot shoe dealer	
C	Fireman	fireman	
D	Cattle **(**& sheep**)** farmer	cattle sheep farmer	

	boot	cattle	dealer	farmer	fireman	horse	maker	sheep	shoe
A	1	0	0	0	0	0			
B									
C									
D									

Machine Learning

- Inputs are split into features and converted to high dimension vectors

Record	Original Input	Cleaned input	Vector
A	Boot **and** shoe maker	boot shoe **maker**	
B	Boot **and** shoe dealer	boot shoe dealer	
C	Fireman	fireman	
D	Cattle **(**& sheep**)** farmer	cattle sheep farmer	

	boot	cattle	dealer	farmer	fireman	horse	maker	sheep	shoe
A	1	0	0	0	0	0	**1**		
B									
C									
D									

Machine Learning

- Inputs are split into features and converted to high dimension vectors

Record	Original Input	Cleaned input	Vector
A	Boot **and** shoe maker	boot **shoe** maker	
B	Boot **and** shoe dealer	boot shoe dealer	
C	Fireman	fireman	
D	Cattle **(**& sheep**)** farmer	cattle sheep farmer	

	boot	cattle	dealer	farmer	fireman	horse	maker	sheep	shoe
A	1	0	0	0	0	0	1	0	**1**
B									
C									
D									

Machine Learning

- Inputs are split into features and converted to high dimension vectors

Record	Original Input	Cleaned input	Vector
A	Boot **and** shoe maker	boot shoe maker	
B	Boot **and** shoe dealer	boot shoe dealer	
C	Fireman	fireman	
D	Cattle (& sheep) farmer	cattle sheep farmer	

	boot	cattle	dealer	farmer	fireman	horse	maker	sheep	shoe
A	1	0	0	0	0	0	1	0	1
B	1	0	1	0	0	0	0	0	1
C	0	0	0	0	1	0	0	0	0
D	0	1	0	1	0	0	0	1	0

Machine Learning

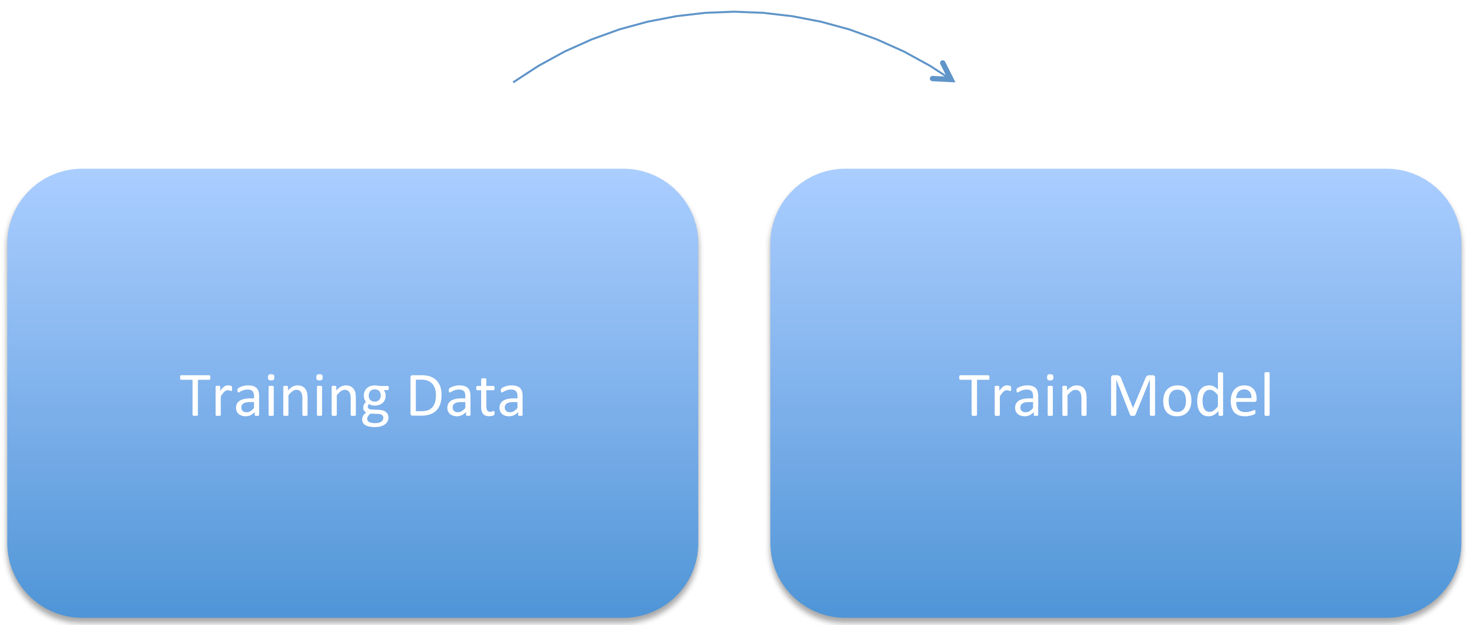
- Inputs are split into features and converted to high dimension vectors

Record	Original Input	Cleaned input	Vector
A	Boot **and** shoe maker	boot shoe maker	100000101
B	Boot **and** shoe dealer	boot shoe dealer	101000001
C	Fireman	fireman	000010000
D	Cattle (& sheep) farmer	cattle sheep farmer	010100010

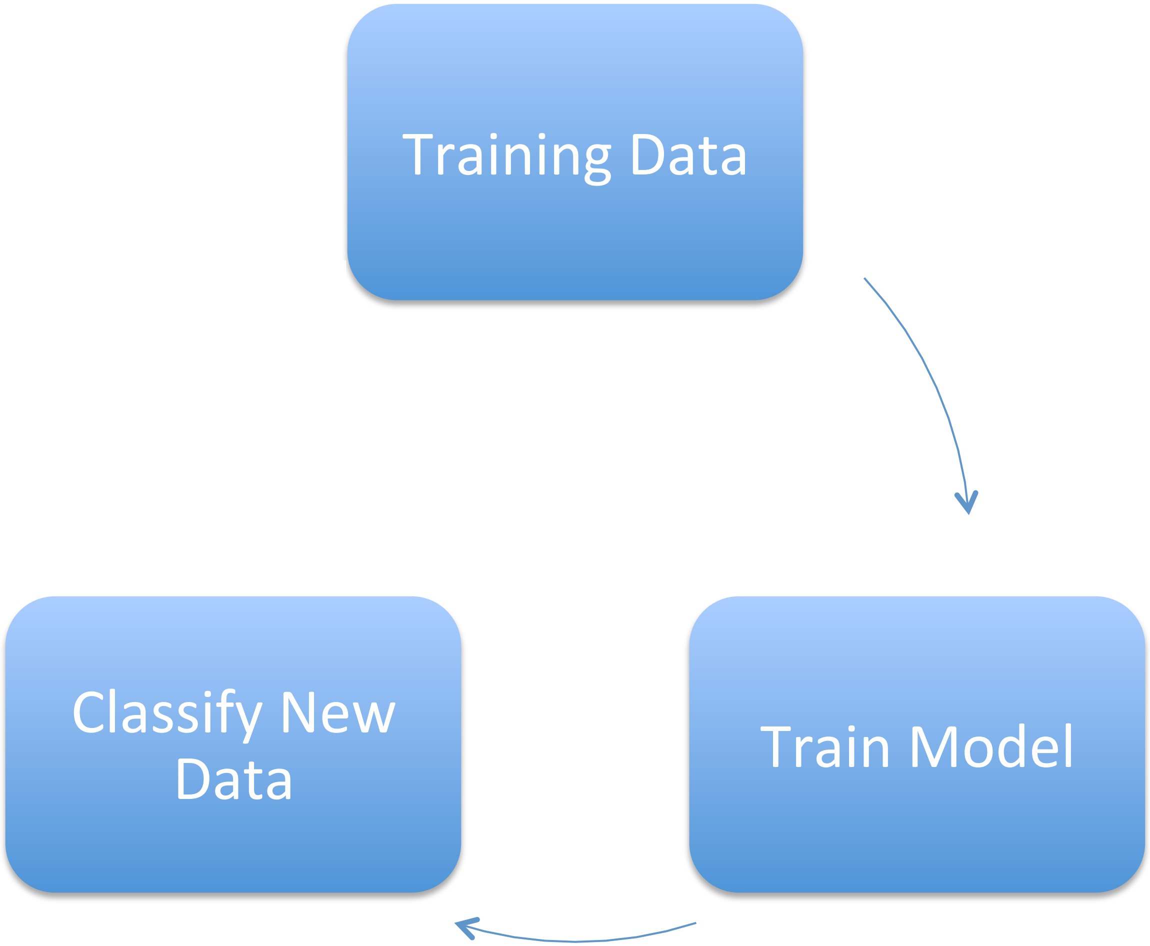
	boot	cattle	dealer	farmer	fireman	horse	maker	sheep	shoe
A	1	0	0	0	0	0	1	0	1
B	1	0	1	0	0	0	0	0	1
C	0	0	0	0	1	0	0	0	0
D	0	1	0	1	0	0	0	1	0

Approach to Classification

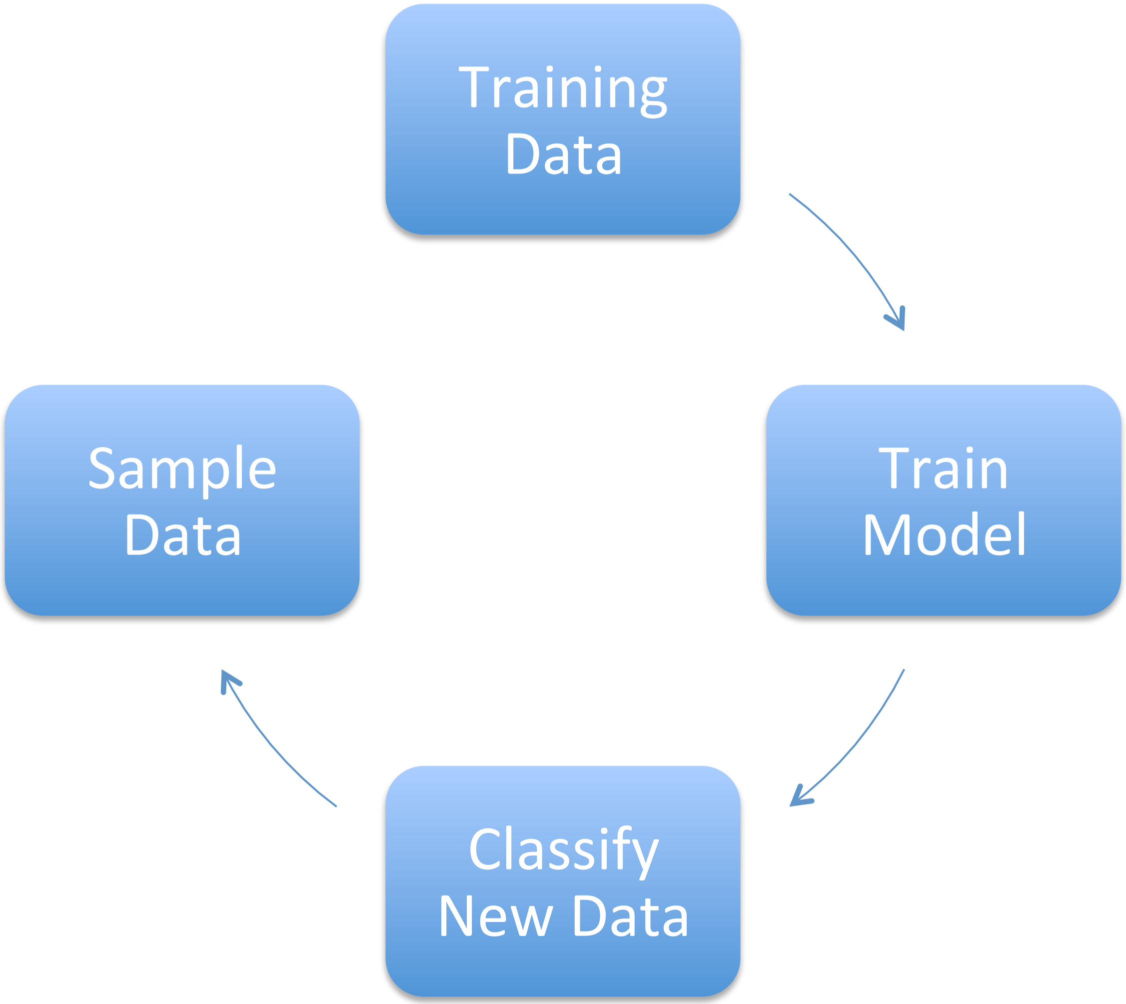
Approach to Classification



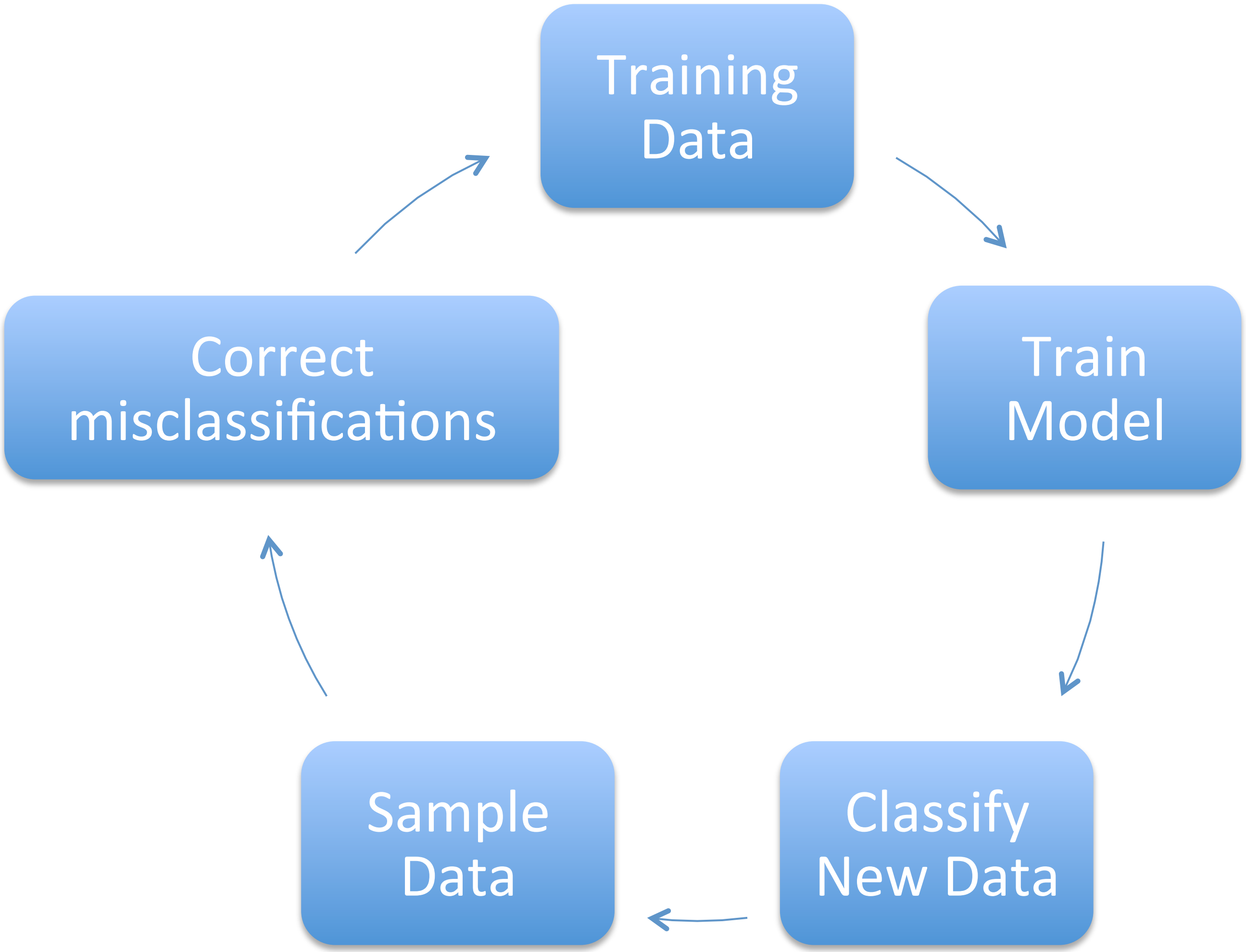
Approach to Classification



Approach to Classification



Approach to Classification



Feature Selection

- Changes to the data input to the classification system
- Feature Selection
  - Selecting the most appropriate features to use in the training data
- Analyse the input to identify features are likely to harm the quality of the classification
- Common words that appear in lots of different output codes.
- Example: “farmers daughter”, “Butchers daughter”…
  - Remove the word daughter.

Gold Standard Misclassification

- Variations in coding of unique strings make it harder to calculate a good model
- Different coders, extra data, mistakes
- Try removing strings coded to multiple codes
- Try changing less common codes to most common

Edit Distance Classifier

- Relatively simple string similarity classifier
- HISCO uses numerical codes, so compare with code description
- Assume similar inputs have similar descriptions
- Similarity measured using edit distance
  - Number of single-character insertions, deletions or replacements needed to transform
- Look for highest number of exact matches between words, fall back to similarity if equal number of matches.

Edit Distance Example

Occupation	Gold Standard Output	Edit Distance Output
Hotel proprietor	Working Proprietor (Hotel and Restaurant)	Working Proprietor (Hotel and Restaurant)
Taxi driver	Taxi driver	Taxi driver
Tax clerk	Tax collector	Tax collector
Painter & decorator	Painters, Construction	Sign Painter
File Cutter	Machinery Fitters, Machine Assemblers and Precision Instrument Makers (except Electrical) NEC	Stock Clerks

Edit Distance Example

Occupation	Gold Standard Output	Edit Distance Output
Hotel proprietor	Working Proprietor (Hotel and Restaurant)	Working Proprietor (Hotel and Restaurant)
Taxi driver	Taxi driver	Taxi driver
Tax clerk	Tax collector	Tax collector
Painter & decorator	Painters, Construction	Sign Painter
File Cutter	Machinery Fitters, Machine Assemblers and Precision Instrument Makers (except Electrical) NEC	Stock Clerks

Edit Distance Example

Occupation	Gold Standard Output	Edit Distance Output
Hotel proprietor	Working Proprietor (Hotel and Restaurant)	Working Proprietor (Hotel and Restaurant)
Taxi driver	Taxi driver	Taxi driver
Tax clerk	Tax collector	Tax collector
Painter & decorator	Painters, Construction	Sign Painter
File Cutter	Machinery Fitters, Machine Assemblers and Precision Instrument Makers (except Electrical) NEC	Stock Clerks

Individual Machine Learning Classifiers

- Naïve Bayes
  - Probabilistic classifier
  - Co-occurrence of features
- Stochastic Gradient Descent
  - Optimisation of logistic regression

Ensemble Approaches

- Majority voting
  - Pick the most frequent classification
- Confidence threshold technique
  - Pick the SGD classification unless its likelihood value is below a given threshold
- Pseudo confidence threshold
  - Produce a pseudo measure of likelihood for the Naive Naïve classifier. Pick the best classification from Naïve Bayes and SGD.

Experiments

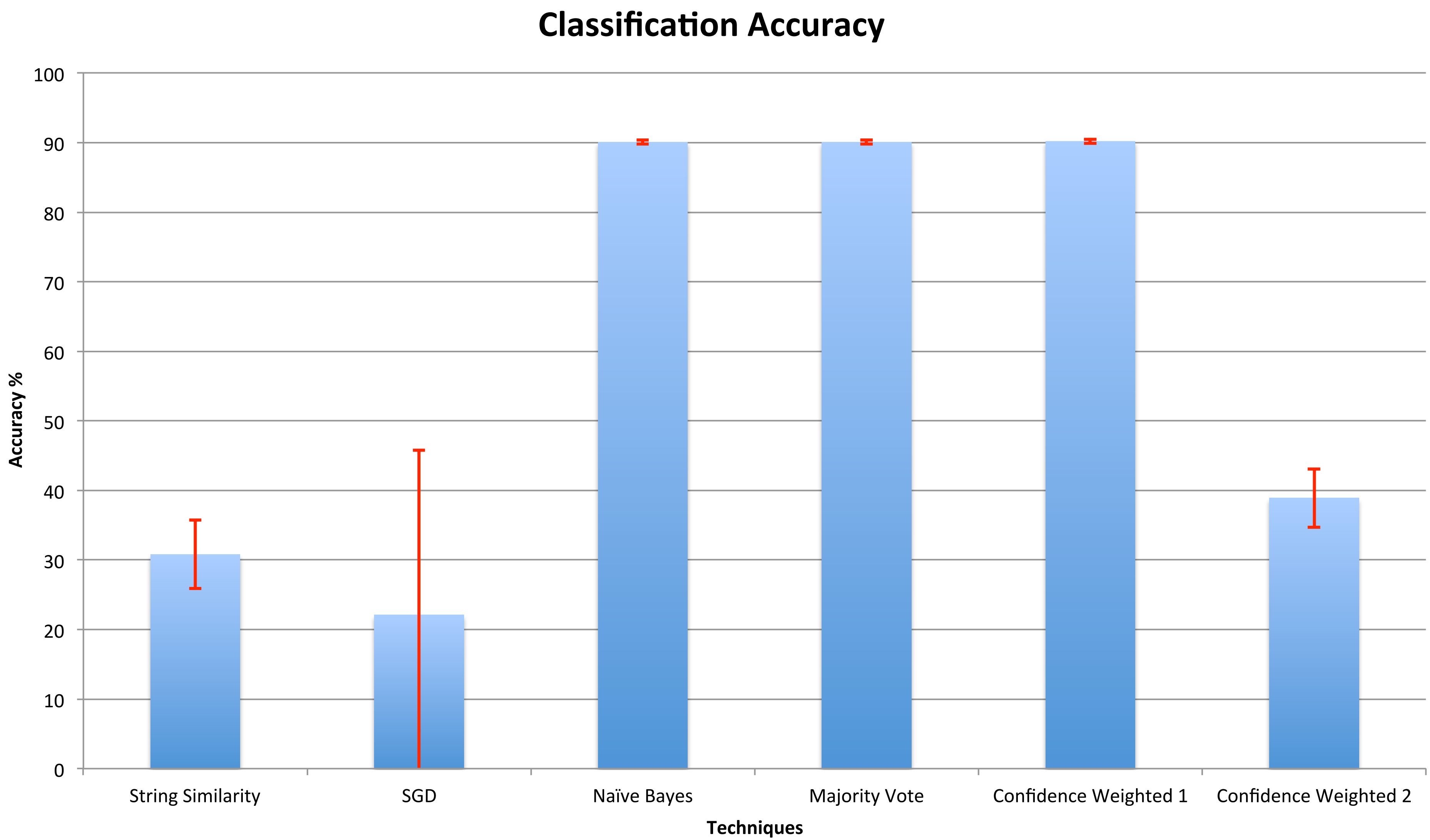
- Which single classification technique produces the highest accuracy when classifying occupations to HISCO?
- Which ensemble technique produces the highest accuracy?
- What difference, if any, does using feature selection make?
- What difference does fixing or removing multiple codings make?
- What effect does classifying to different HISCO levels make?

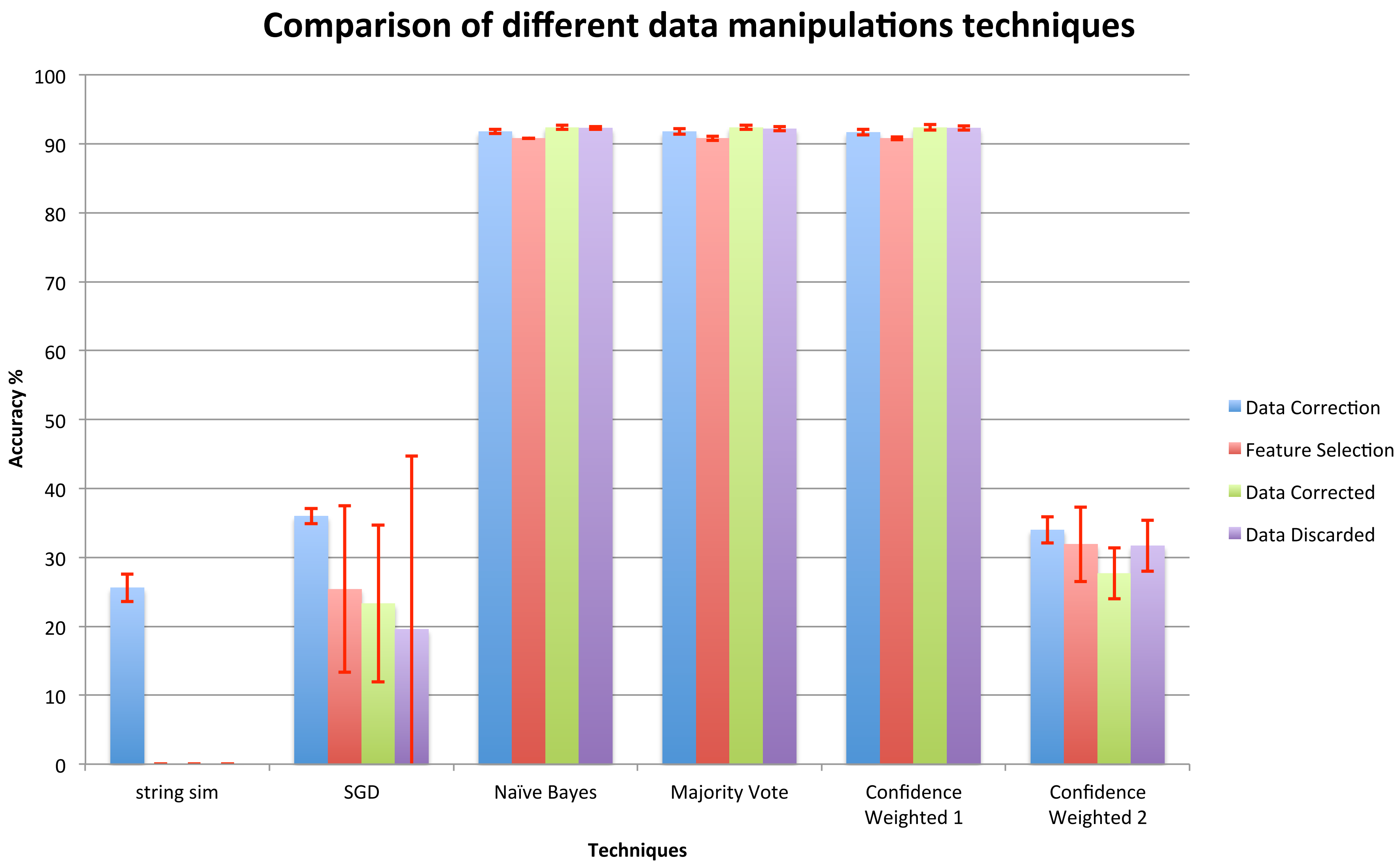
Evaluation

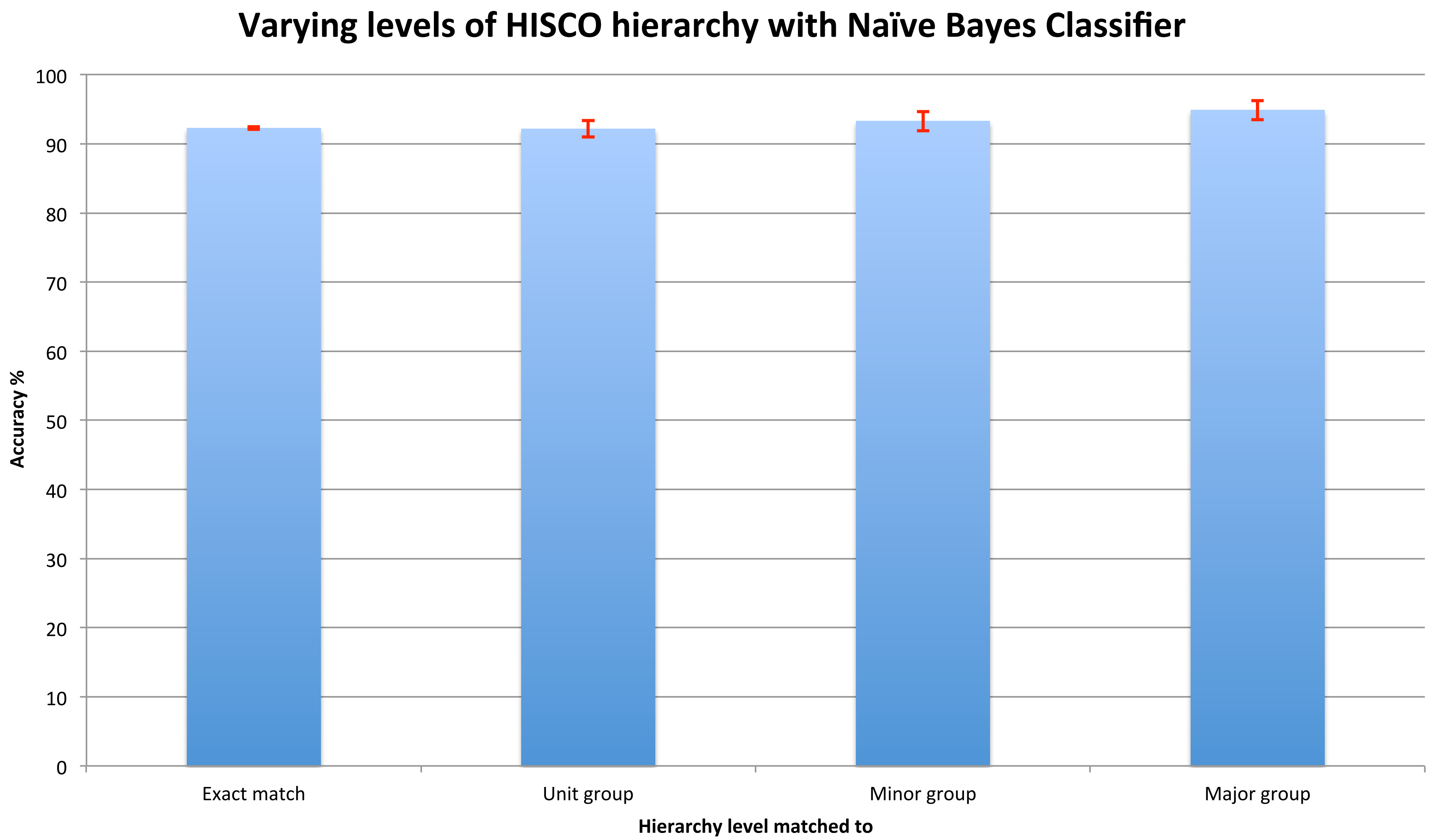
- Need to assess the quality of the automatic coding
- Hold out method
- Split data into two sets, a training set and a validation set
- 80% chosen for training, 20% for validation
- Pick a new training/test set each repetition
- Correct classification is gold standard code matches output code

Accuracy Measures

- HISCO employs a hierarchical structure
- If we are only interested in coarse classifications we can relax the closeness of the match required
- Match unit group
- Match minor group
- Match major group


Classification Accuracy
100
90
80
70
60
50
40
30
20
10
0
Accuracy %
String Similarity
SGD
Naïve Bayes
Majority Vote
Confidence Weighted 1
Confidence Weighted 2
Techniques


Comparison of different data manipulations techniques
Accuracy %
100
90
80
70
60
50
40
30
20
10
0
string sim
SGD
Naïve Bayes
Majority Vote
Confidence Weighted 1
Confidence Weighted 2
Techniques
Data Correction
Feature Selection
Data Corrected
Data Discarded


Varying levels of HISCO hierarchy with Naïve Bayes Classifier
100
90
80
70
60
50
40
30
20
10
0
Accuracy %
Exact match
Unit group
Minor group
Major group
Hierarchy level matched to

Summary

- Highest accuracy: Naïve Bayes classifier with feature selection and correction of multiply coded descriptions.
- Exact match accuracy using this technique was 92.3 ± 0.2%
- Considering only major group matching 94.9 ± 1.4% was achieved.
- Although the ensemble did not improve performance, addition of another high performance algorithm should yield gains.

Discussion

- Previous results classifying cause of death using ensemble methods showed improvement of 2-3%
- Run times:
  - String Similarity: a few minutes
  - Naïve Bayes: a few minutes
  - SGD: 3-4 hours depending on learning parameters
- SGD has been reworked, preliminary results: 88-94%

Future Work

- Continue machine learning and string matching development to classify cause of death and occupations
- Continue to examine behaviour of SGD algorithm to try and achieve better performance.
- Add further machine learning models, such as support vector machines into the ensemble