Machine Learning Algorithms to Code State Public Health Spending Accounts

Keeneland Conference

K15: Session 3C: April 22nd 2015

Johns Hopkins Bloomberg School of Public Health

de Beaumont Foundation

Acknowledgements

This work was generously supported by a grant from the de Beaumont Foundation

Research Team

YN Alfonso
D Bishai
E Brady
N Kish
J Le
JP Leider
B Resnick
A Sensenig

Overview

Motivation for this project
- Project aims
- Current estimates of public health spending

Data sources
- Expenditure data
- Re-classification of public health spending

Machine Learning application
- What it is, how apply to this context
- Results
- Conclusions and potential applications

Motivation and Aims

To refine existing public health spending estimates to ascertain what we actually spend on public health

Knowing what we spend on public health is fundamental to demonstrating public health value, and effectiveness

The Problem

Estimating the value of public health spending is difficult

- Lack of consistent reporting and coding in public health activities and definitions
- No systematic dataset on how much in total we spend on public health
- Current public health spending estimates exclude non-health agencies that do some public health work (e.g., agriculture, environment, etc.)

2008 and 2011 State Public Health Spending Estimates

(in billions)

Year	ASTHO*	TFAH**	Census
2011	**$26.5**	**$10.4**	**$55**
2008	**$24**	**$12**	**$60**
Notes	State health agency spending. This estimate does not include behavioral health or Medicaid	Does not include federal funds or some “non comparable” programs (e.g., behavioral health)	Comprises all state agencies (not only health). Includes $39 (2008) $36 (2011) for current operations and $20 (2008) $18 (2011) in state to local transfers

*Association of State and Territorial Health Officials
**Trust for America’s Health

The Data

Census of Governments

Census of Governments is a US Census Bureau program to collect county expenditure data every 5 years

Multiple categories, sub-categories of spending

Examples: Hospital spending, Police, Sewerage, Solid Waste Management, Environmental, Education, Housing

Code 32 is "Current Operations – Health – Other" contains much public health spending

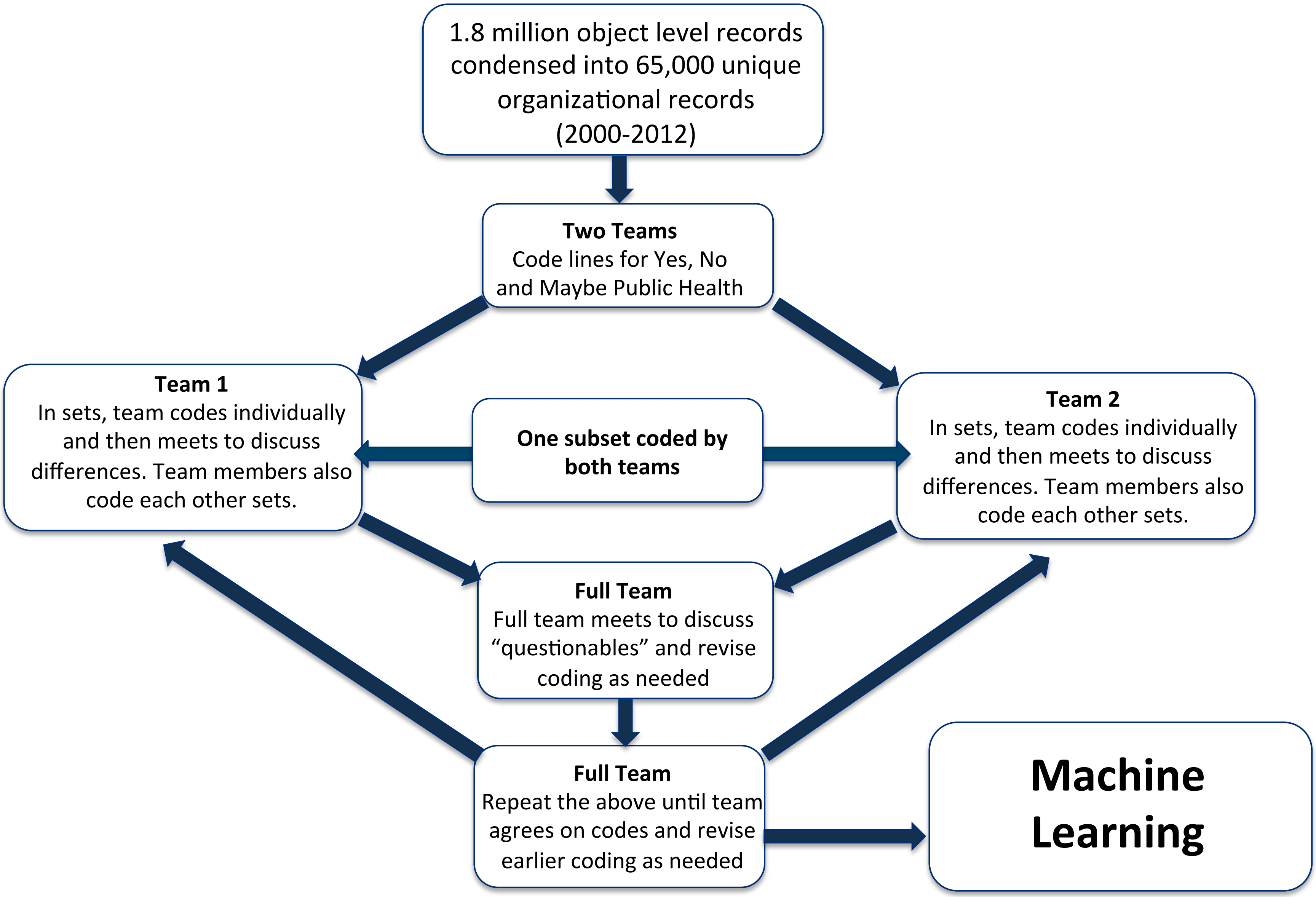
State level data 2000-2012

Source: U.S. Census of Governments http://www.census.gov/govs/cog/

Individual records

	B	C	D	E	G	I	K	M	O	Q	S	T	U
1	ID	SURVEY		CD	LVL1_DESC	LVL2_DESC	LVL3_DESC	LVL4_D	LVL5_D	LVL6_D	OBJ_DESC	AMT	Amoun
2	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	FAMILY HEALTH SERVICES	PUBLIC HEALTH SERVICES			GROUP HEALTH INSURANCE	7	7000
3	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	FAMILY HEALTH SERVICES	PUBLIC HEALTH SERVICES			SALARIES, REGULAR	17	17000
4	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	FAMILY HEALTH SERVICES	PUBLIC HEALTH SERVICES			TRAIN/REG-INDIVIDUAL/GOVERNMT	1	1000
5	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	FAMILY HEALTH SERVICES	PUBLIC HEALTH SERVICES			OFFICE OPERATION	1	1000
6	01000[illegible]0	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	FAMILY HEALTH SERVICES	PUBLIC HEALTH SERVICES			ANSWERING SERVICES	17	17000
7	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	FAMILY HEALTH SERVICES	PUBLIC HEALTH SERVICES			MEDICAL SERVICES-PROFESSIONAL	282	282000
8	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	FAMILY HEALTH SERVICES	PUBLIC HEALTH SERVICES			ADVERTISING-PROFESSIONAL	15	15000
9	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	DISEASE CONTROL	PUBLIC HEALTH SERVICES			BOOKS, SUBSCRIPTIONS & PERIODI	1	1000
10	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	DISEASE CONTROL	PUBLIC HEALTH SERVICES			FICA	218	218000
11	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	DISEASE CONTROL	PUBLIC HEALTH SERVICES			SALARIES, REGULAR	2448	2448000
12	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	DISEASE CONTROL	PUBLIC HEALTH SERVICES			TERMINATION COST, ANNUAL LEAVE	72	72000
13	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	DISEASE CONTROL	PUBLIC HEALTH SERVICES			SICK LEAVE	133	133000
14	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	DISEASE CONTROL	PUBLIC HEALTH SERVICES			LONGEVITY ALLOWANCES	35	35000
15	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	DISEASE CONTROL	PUBLIC HEALTH SERVICES			TERMINATION COSTS, SICK LEAVE	63	63000
16	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	DISEASE CONTROL	PUBLIC HEALTH SERVICES			COMPENSATORY LEAVE	2	2000
17	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	DISEASE CONTROL	PUBLIC HEALTH SERVICES			ANNUAL LEAVE	231	231000
18	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	DISEASE CONTROL	PUBLIC HEALTH SERVICES			JURY DUTY	2	2000
19	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	DISEASE CONTROL	PUBLIC HEALTH SERVICES			ASSOCIATION DUES	11	11000
20	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	DISEASE CONTROL	PUBLIC HEALTH SERVICES			MEDICAL	4820	4820000
21	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	DISEASE CONTROL	PUBLIC HEALTH SERVICES			OFFICE OPERATION	1	1000
22	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	DISEASE CONTROL	PUBLIC HEALTH SERVICES			MEDICAL SERVICES-PROFESSIONAL	112	112000
23	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	DISEASE CONTROL	PUBLIC HEALTH SERVICES			PRINTNG/REPRODUCTN/PHOTO EQUIP	2	2000
24	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	DISEASE CONTROL	PUBLIC HEALTH SERVICES			GROUP HEALTH INSURANCE	549	549000
25	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	DISEASE CONTROL	PUBLIC HEALTH SERVICES			TRAIN/REG-INDIVIDUAL/GOVERNMT	2	2000
26	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	COUNTY OPERATIONS	PUBLIC HEALTH SERVICES			ANNUAL LEAVE	318	318000
27	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	COUNTY OPERATIONS	PUBLIC HEALTH SERVICES			JURY DUTY	2	2000
28	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	COUNTY OPERATIONS	PUBLIC HEALTH SERVICES			FICA	902	902000
29	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	COUNTY OPERATIONS	PUBLIC HEALTH SERVICES			GROUP HEALTH INSURANCE	1198	1198000
30	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	COUNTY OPERATIONS	PUBLIC HEALTH SERVICES			MILEAGE	1	1000
31	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	COUNTY OPERATIONS	PUBLIC HEALTH SERVICES			LEGAL- PROFESSIONAL	1	1000
32	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	COUNTY OPERATIONS	PUBLIC HEALTH SERVICES			TERMINATION COSTS, SICK LEAVE	50	50000
33	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	COUNTY OPERATIONS	PUBLIC HEALTH SERVICES			OFFICE OPERATION	1	1000
34	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	COUNTY OPERATIONS	PUBLIC HEALTH SERVICES			STATE & FED-TAXES/LICENSES	1	1000
35	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	COUNTY OPERATIONS	PUBLIC HEALTH SERVICES			SALARIES, REGULAR	7323	7323000
36	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	COUNTY OPERATIONS	PUBLIC HEALTH SERVICES			TERMINATION COST, ANNUAL LEAVE	56	56000
37	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	COUNTY OPERATIONS	PUBLIC HEALTH SERVICES			SICK LEAVE	138	138000
38	010000000	2011	E	32	PUBLIC HEALTH	HEALTH-GENERAL FUND	COUNTY OPERATIONS	PUBLIC HEALTH SERVICES			LONGEVITY ALLOWANCES	37	37000

Manual inter-rater Coding Process



Manual inter-rater Coding Process

Two Teams
Code lines for Yes, No
and Maybe Public Health

1=Not Public Health
2= Maybe Public Health
3=Public Health

Machine Learning

Automatic Coding using Machine Learning

Aims to replicate 'gold standard' classifications generated manually

Automation will save time and should improve consistency of classification

Manual codes are considered the 'truth', used to train machine algorithms in classification decisions

65,000+ organizational records split up, majority used to train algorithms, two subsets set aside for testing and validation of predictions

Agreement unlikely to be perfect, 90% inter-rater (machine/human) agreement considered acceptable

Steps in training and testing models

Data formatted as corpus (large, structured set of text objects) split into training, testing & validation sets: 3/5 1/5 1/5

Pre-processing includes text mining, condensing the data, removing unnecessary features, can include re-weighting, manual adjustments

Algorithms selected to fit models to the data
(eg. Random Forests, Tree, Bootstrap aggregation, Support Vector Machine, Maximum Entropy)

Training set used to fit parameters with true classifiers as ‘dependent variable’

Steps in training and testing models

Based on these parameters, for each line of testing set, a class is predicted and compared with true class (1/2/3)

Differences between prediction and true class may arise due to model structure, heterogeneity in data.

In this case, another source of error could be inconsistencies in manual coding

Risk of over-fitting to training data, use k-fold cross-validation for out-of-sample accuracy

Results 1: Confusion Matrix

Initial look at how each specific algorithm compares with true classification

Helps to identify sources of error (off diagonal), classes to investigate e.g. more ‘maybe’s being predicted as ‘not PH’

Sum of diagonal as a % of total: 85.4% (matches)

Public health as a % of total: ‘True’ =52%, Predicted=49%

		Predicted class		
		1	2	3
True class	1	5081	51	679
	2	215	298	358
	3	568	32	5764

Results 2: Algorithm performance

Non-parametric models seem to perform best overall – Random Forests, Aggregate Bootstrapping

Algorithm	Precision	Recall	F-score
Forests	0.86	0.85	0.86
Bagging	0.85	0.85	0.85
SVM	0.85	0.84	0.85
SLDA	0.85	0.84	0.85
GLM net	0.84	0.84	0.84
Max Entropy	0.84	0.84	0.84
Boosting	0.75	0.85	0.8
Tree	0.74	0.74	0.74
Neural net	0.56	0.91	0.69

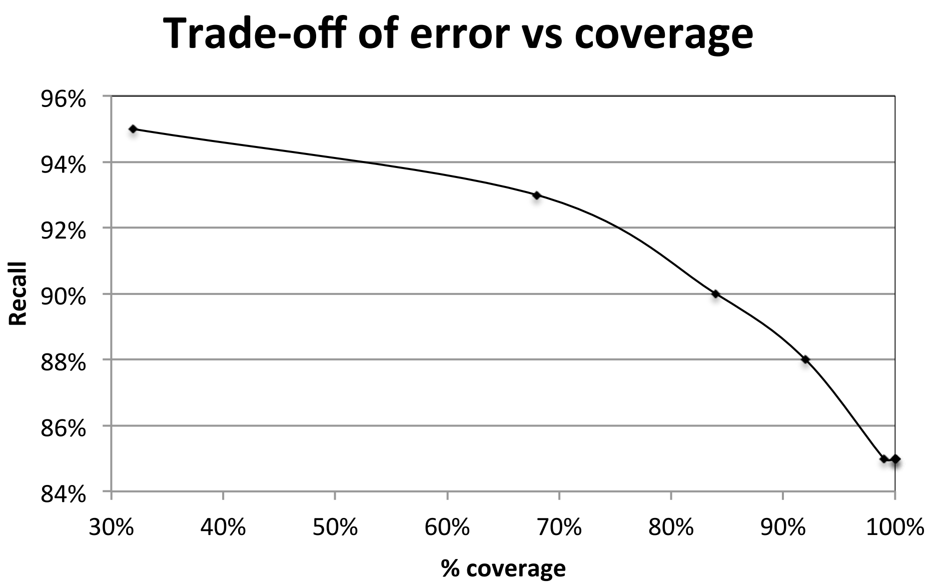
Notice several with good performance, not necessarily overlapping, can we take advantage of less well-performing algorithms?

Results 3: Ensemble agreement

'Ensembling' combines individual algorithm predictions to generate a more accurate 'ensemble' prediction

Trade-off coverage for accuracy

Number of algorithms	n-ENSEMBLE COVERAGE	n-ENSEMBLE RECALL
n >= 1	1	0.85
n >= 2	1	0.85
n >= 3	1	0.85
n >= 4	1	0.85
n >= 5	0.99	0.85
n >= 6	0.92	0.88
n >= 7	0.84	0.9
n >= 8	0.68	0.93



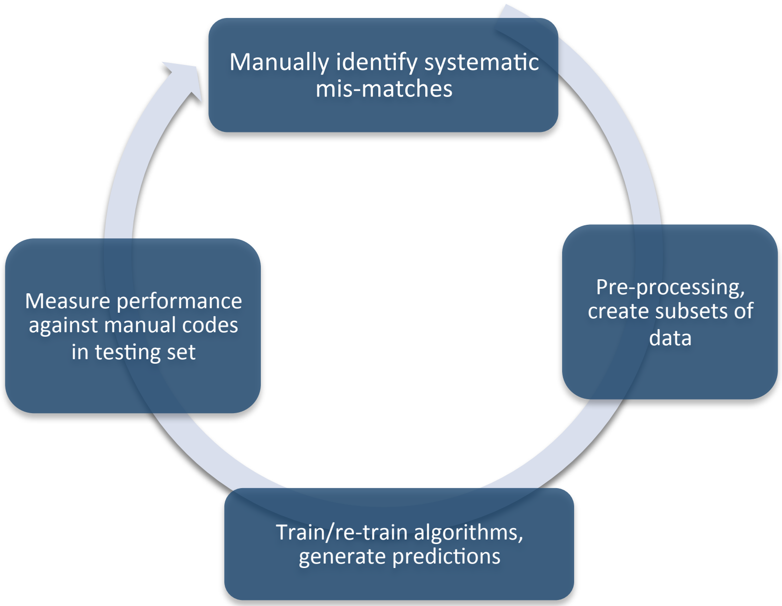
Summary results

Initial results are good against testing subset: 74-91% recall individual algorithms, up to 93% ensemble recall, ~88% out of sample error in cross-validation

Iterative process: Improve matching through pre-processing and model selection

Identification of inconsistent 'true' codes to be adjusted manually

Conclude that machines can classify this type of data to a high degree of accuracy



Implications for public health practice

In 2015 Census Bureau will have another million records of state spending on public health.

Human coding of local government spending on public health is expensive

Plan A) Census spends new federal money to code it using humans

Plan B) Foundations spend new money to code it

Plan C) Machines take over coding state public health spending and humans do a small sample as a cross check

With our work, we hope to lay a foundation for Plan C

Thank you

Johns Hopkins Bloomberg School of Public Health

de Beaumont Foundation