NIH Intramural Research Program
Our Research Changes Lives

SOCcer: Automatic coding of free-text job descriptions to standardized occupation codes

Daniel E. Russ[1], Kwan-Yuet Ho[1],Calvin A. Johnson[1], and Melissa C. Friesen[2]

[1]Division of Computational Bioscience, Center for Information Technology, NIH
[2]Occupational and Environmental Epidemiology Branch, Division of Cancer Epidemiology and Genetics, NCI, NIH

Overview

- Occupation information
  - Use and collection
  - Standardized classification systems
- SOCcer
  - Framework
  - Performance measures
  - Future advancements and new directions

Occupation Information

Finance, credit, purchase preferences

Medical diagnosis

Surveillance to improve health policy and identify priorities

Primary or confounding factor in **epidemiologic studies**

What is your job title?

- Many ways of describing an occupation
- Different level of detail
- Changes depending on who is asking

Wile E. Coyote
GENIUS
HAVE BRAIN WILL TRAVEL

<u>Standardized Occupation Categories</u>

Mechanical Engineer
Misc. Engineer
Physical Scientist
Chief Executive

Standard Occupational Codes

US SOC 2010

Major group	19-0000	Life, Physical, and Social Science Occupations
Minor Group	19-1000	Life Scientists
Broad Group	19-1040	Medical Scientists
Detailed Occupation	19-1041	Epidemiologists
	19-1042	Medical Scientists, Except Epidemiologists

Major group	29-0000	Healthcare Practitioners and Technical Occupations
Minor Group	29-1000	Health Diagnosing and Treating Practitioners
Broad Group	29-1060	Physicians and Surgeons
Detailed Occupation	29-1062	Family and General Practitioners
	29-1063	General Internist
	29-1069	Physicians and Surgeons, All Other

Occupation in population-based studies

- Current job, longest job, usual job, all jobs
- Wide variety of occupations, industries
- Open-ended questions:
  - What was your job title?
  - What were your main tasks and activities in this job?
  - Who was your employer?
  - What services were provided or what products were made by your employer?
  - Start year/stop year
- Coded to standardized occupation and industry classification systems: SOC and SIC

Coding: time-consuming, modestly reliable

- Manual process
- Based on limited information
- No gold standard
- Agreement between 2 coders is poor/moderate (Koeman et al. 2013)
  - 5-digit level ISCO68 agreement: 36%
  - 3-digit level ISCO68 agreement: 55%
- Preferably independent assignments by 2 coders, resolve discordant assignments
- Costly in large-scale studies

Multiple recent efforts to automate

- NIOSH: http://wwwn.cdc.gov/niosh-nioccs/
- U. Montreal: www.caps-canada.ca
- Burstyn et al. (2014) Beyond crosswalks: reliability of exposure assessment following automated coding of free-text job descriptions for occupational epidemiology.
- Patel et al. (2012) Performance of automated and manual coding systems for occupational data: A case study of historical records

- Batch vs. job-by-job.
- Most require the user to make the final determination from multiple choices or do not provide an assignment when low confidence/no match.
- Different coding schemes.

Objective

- Develop a computer algorithm to assign standardized occupation classifications (SOC) based on free-text responses.
  - Reduce but not replace expert coding

- Cross-NIH institute collaboration with Division of Computational Bioscience
  - Expertise in natural language processing and classification

Our framework

- Adaptable system
  - [Initially] Code to US SOC 2010
    - Electronic knowledge base of job titles and tasks for each SOC (O*NET)
  - [Future] Other classification systems
- Assumption: Better matches by using multiple aspects of job description
  - Job title, tasks, coded industry

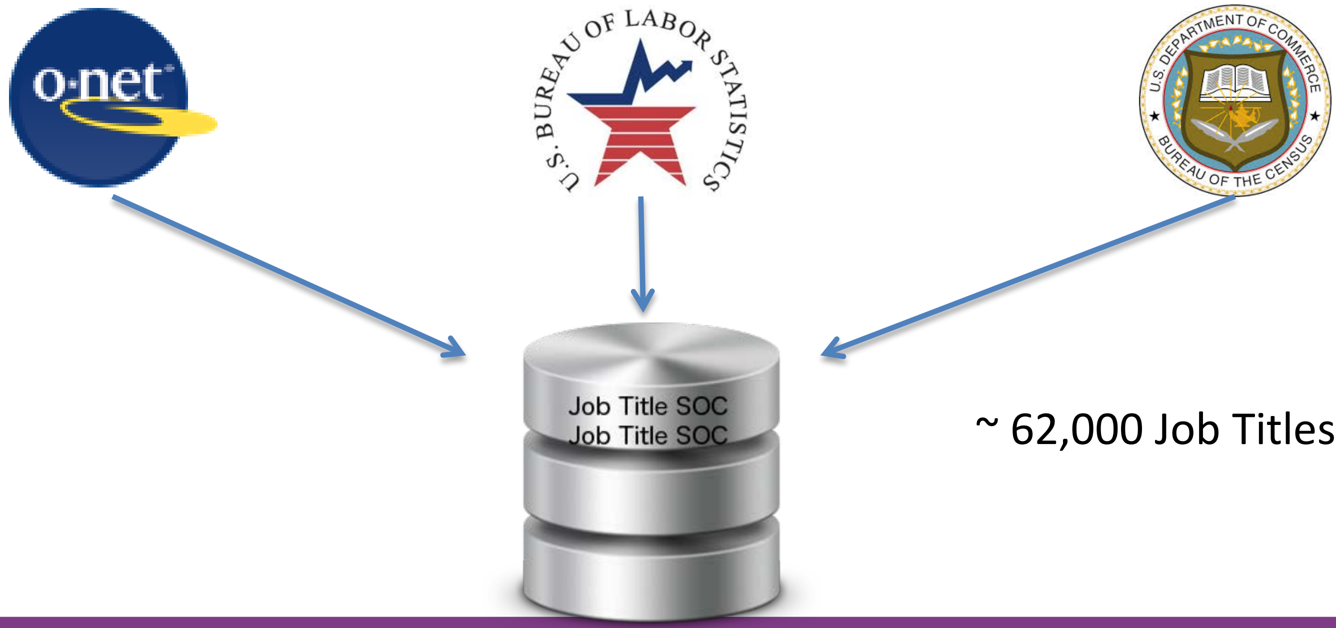
SOCcer

Standardized Occupation Coding for Computer-assisted
Epidemiological Research

http://soccer.nci.nih.gov

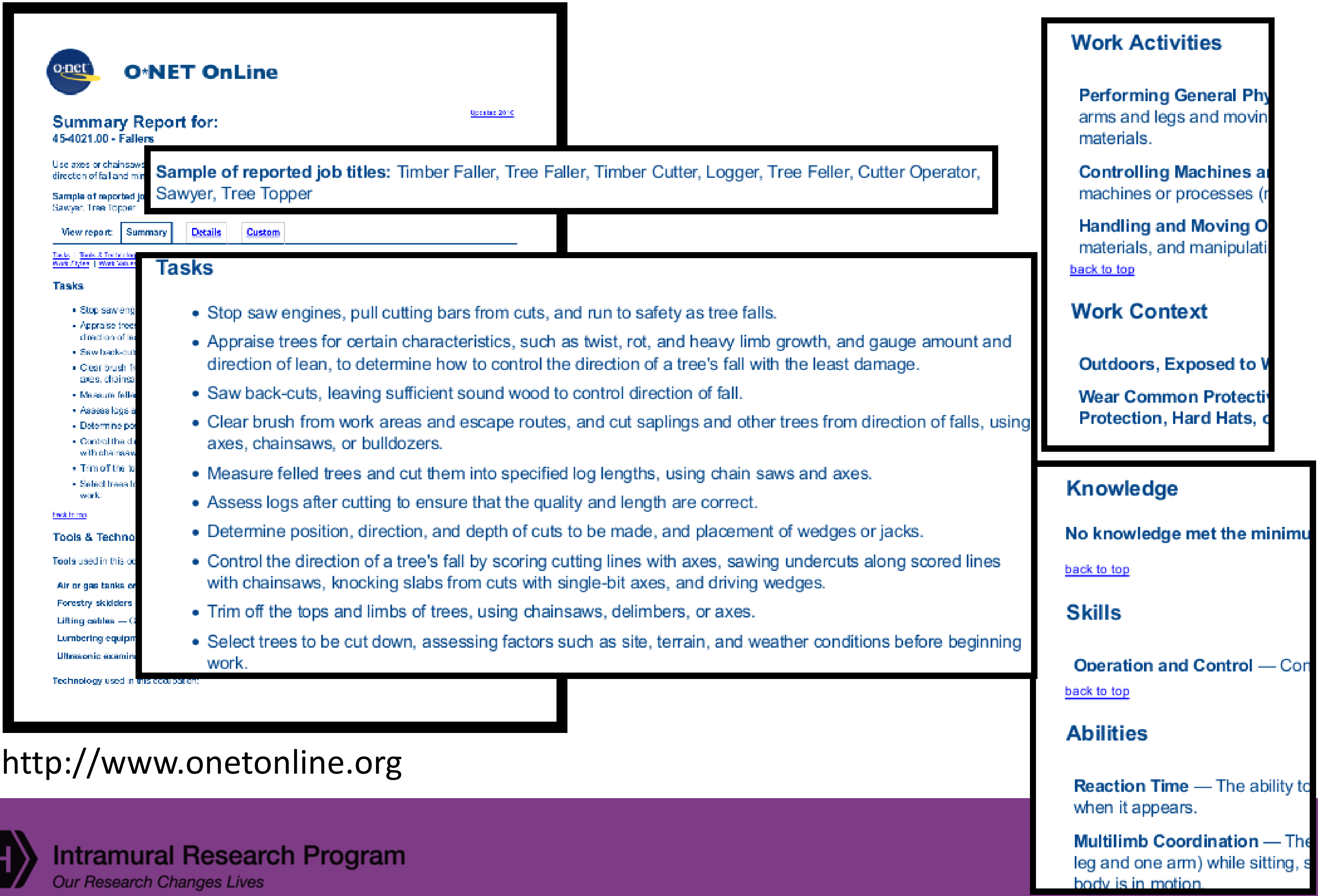
Knowledge Base Development

Data Sources		
Production database of job titles Crosswalk information	Direct Match Title File	U.S. Census Occupational Index



~ 62,000 Job Titles

O*NET for US SOC 2010



O*NET OnLine

Updated 2010

Summary Report for:
45-4021.00 - Fallers

Use axes or chainsaws … direction of fall and mi…

Sample of reported jo… Sawyer, Tree Topper

View report: Summary | Details | Custom

Tasks | Tools & Technology … Work Styles | Work Values …

Sample of reported job titles: Timber Faller, Tree Faller, Timber Cutter, Logger, Tree Feller, Cutter Operator, Sawyer, Tree Topper

Tasks

- Stop saw engines, pull cutting bars from cuts, and run to safety as tree falls.
- Appraise trees for certain characteristics, such as twist, rot, and heavy limb growth, and gauge amount and direction of lean, to determine how to control the direction of a tree's fall with the least damage.
- Saw back-cuts, leaving sufficient sound wood to control direction of fall.
- Clear brush from work areas and escape routes, and cut saplings and other trees from direction of falls, using axes, chainsaws, or bulldozers.
- Measure felled trees and cut them into specified log lengths, using chain saws and axes.
- Assess logs after cutting to ensure that the quality and length are correct.
- Determine position, direction, and depth of cuts to be made, and placement of wedges or jacks.
- Control the direction of a tree's fall by scoring cutting lines with axes, sawing undercuts along scored lines with chainsaws, knocking slabs from cuts with single-bit axes, and driving wedges.
- Trim off the tops and limbs of trees, using chainsaws, delimbers, or axes.
- Select trees to be cut down, assessing factors such as site, terrain, and weather conditions before beginning work.

back to top

Tools & Techno…

Tools used in this oc…

Air or gas tanks or…
Forestry skidders …
Lifting cables — C…
Lumbering equipm…
Ultrasonic examin…

Technology used in this occupation:

Work Activities

Performing General Phy… arms and legs and movin… materials.

Controlling Machines a… machines or processes (r…

Handling and Moving O… materials, and manipulati…

back to top

Work Context

Outdoors, Exposed to W…

Wear Common Protecti… Protection, Hard Hats, o…

Knowledge

No knowledge met the minimu…

back to top

Skills

Operation and Control — Co…

back to top

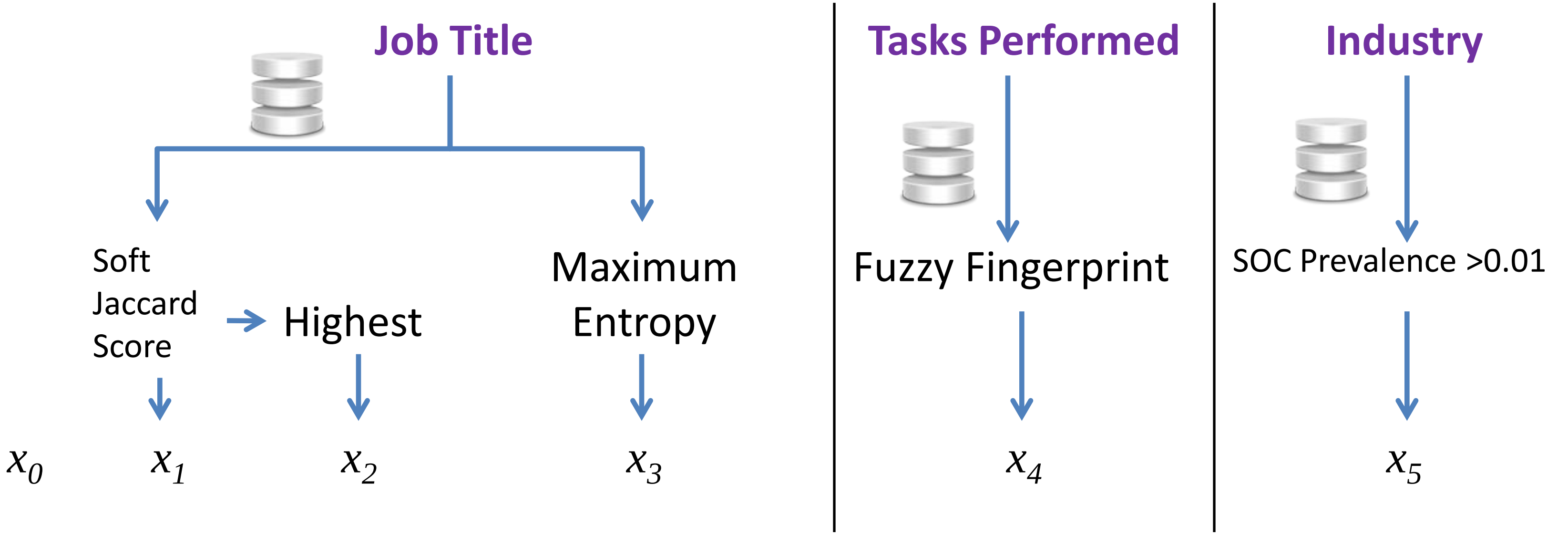
Abilities

Reaction Time — The ability to… when it appears.

Multilimb Coordination — The… leg and one arm) while sitting, s… body is in motion.

http://www.onetonline.org

Overview of SOCcer



Job Title

Soft Jaccard Score → Highest

Maximum Entropy

Tasks Performed

Fuzzy Fingerprint

Industry

SOC Prevalence >0.01

x_0 x_1 x_2 x_3 x_4 x_5

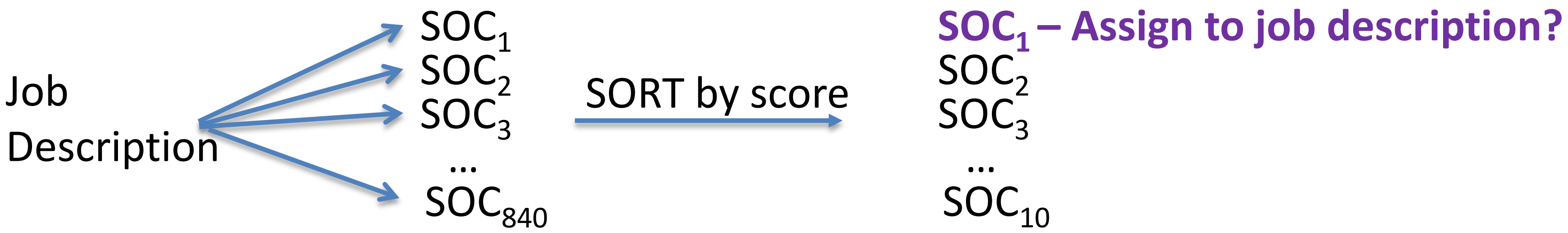
Logistic Regression to obtain score

$$\ln\left(\frac{p}{1-p}\right) = f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

SOCcer score and SOC code assignment

Algorithm score:
probability that an expert
coder would assign that code

SOCcer output:
Top 10 scoring SOCs



Job Description

SOC_1
SOC_2
SOC_3
…
SOC_{840}

SORT by score

SOC_1 – Assign to job description?
SOC_2
SOC_3
…
SOC_{10}

SOCcer's performance

Validity assessments at 6-digit level:

“Gold standard”: Consensus **expert** SOC assignment

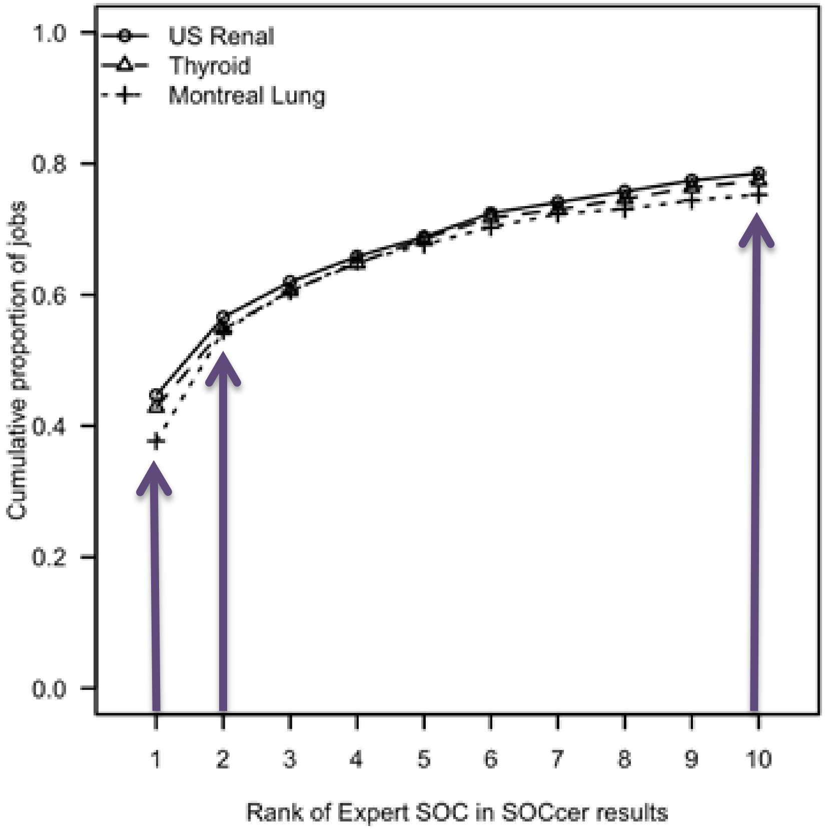
Vs. **Highest scoring SOC-2010 code** from SOCcer

Overall agreement

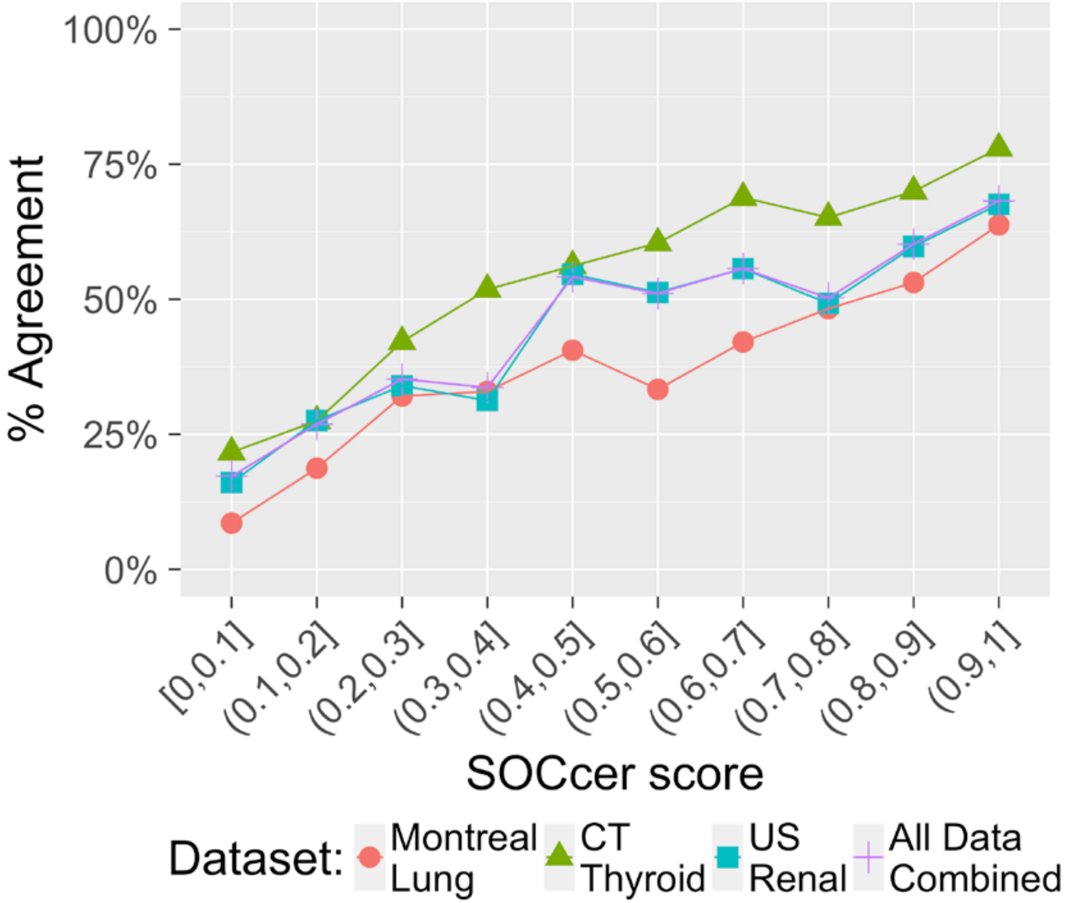
Study	# Jobs	Percent Agreement at SOC Level				Median SOCcer score (IQR)
		2-Digit	3-Digit	5-Digit	6-Digit	
US Renal	11,943	77	64	52	45	0.46 (0.24-0.77)
CT Thyroid	1,942	73	59	50	44	0.22 (0.08-0.51)
Montreal Lung	829	74	56	46	38	0.45 (0.22-0.79)
Combined	**14,519**	**76**	**63**	**51**	**41**	**0.44 (0.23-0.75)**

	Coder vs. Coder	Computer vs. Coders		
	Koeman	Burstyn	Patel (NIOSH)	US Renal
- 3-digit	55%	31-85%	63%	63%
- 5-digit	36%	9-72%	52%	51%

Did the expert assigned code appear in any of the top 10 codes from SOCcer?

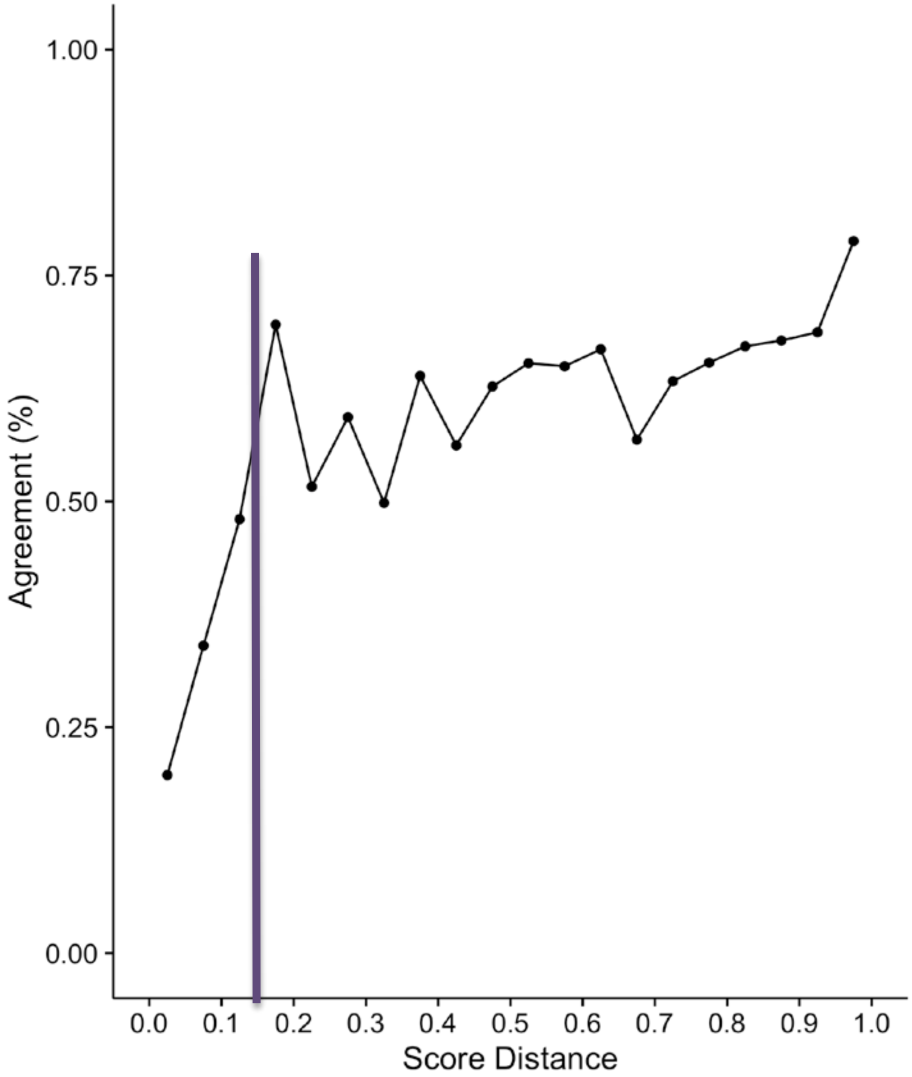


Agreement by score



Agreement by score distance to 2nd ranked SOC code:

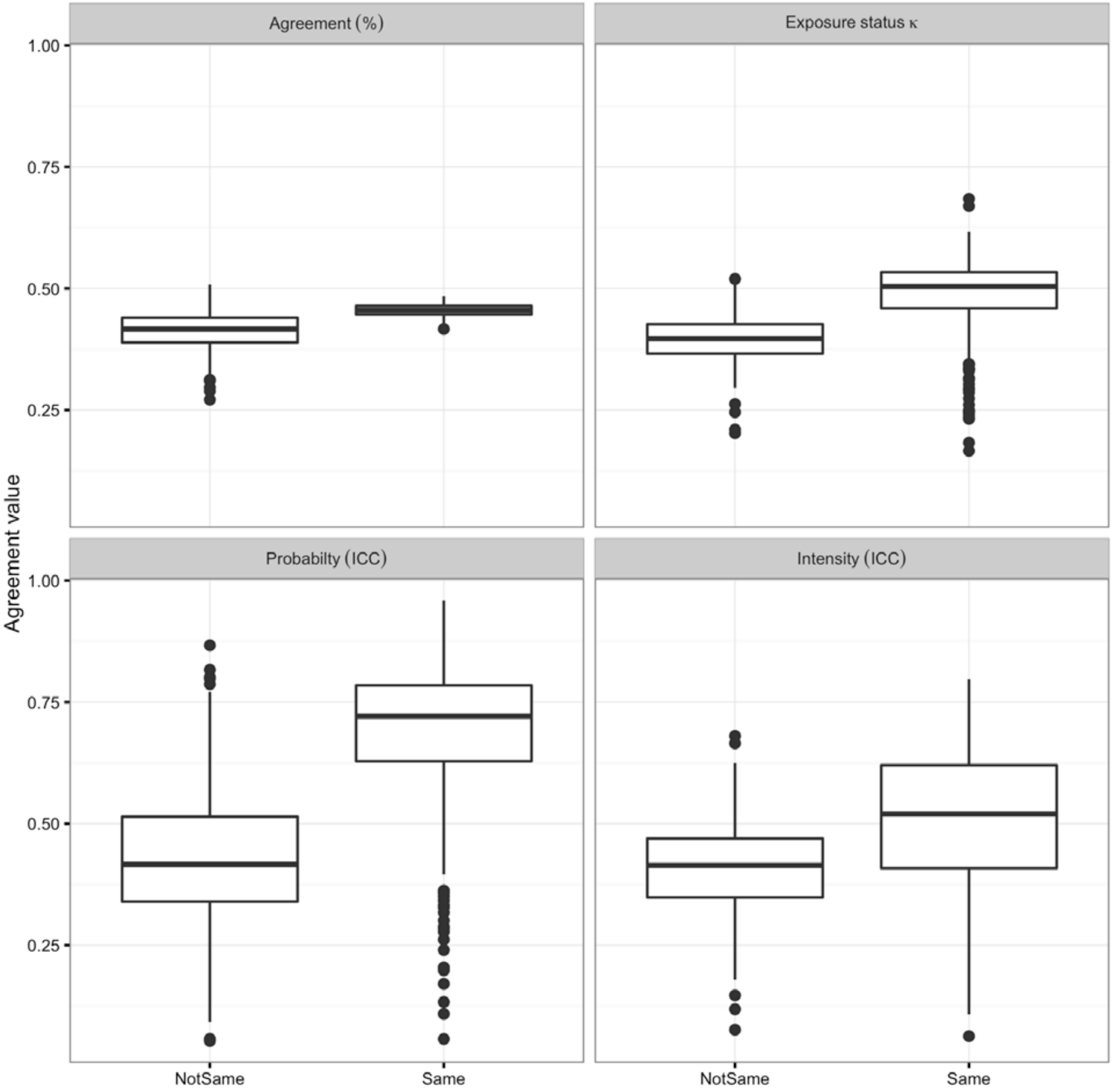
Score distance = $\text{score}_1 - \text{score}_2$



Do the mis-assignments matter?

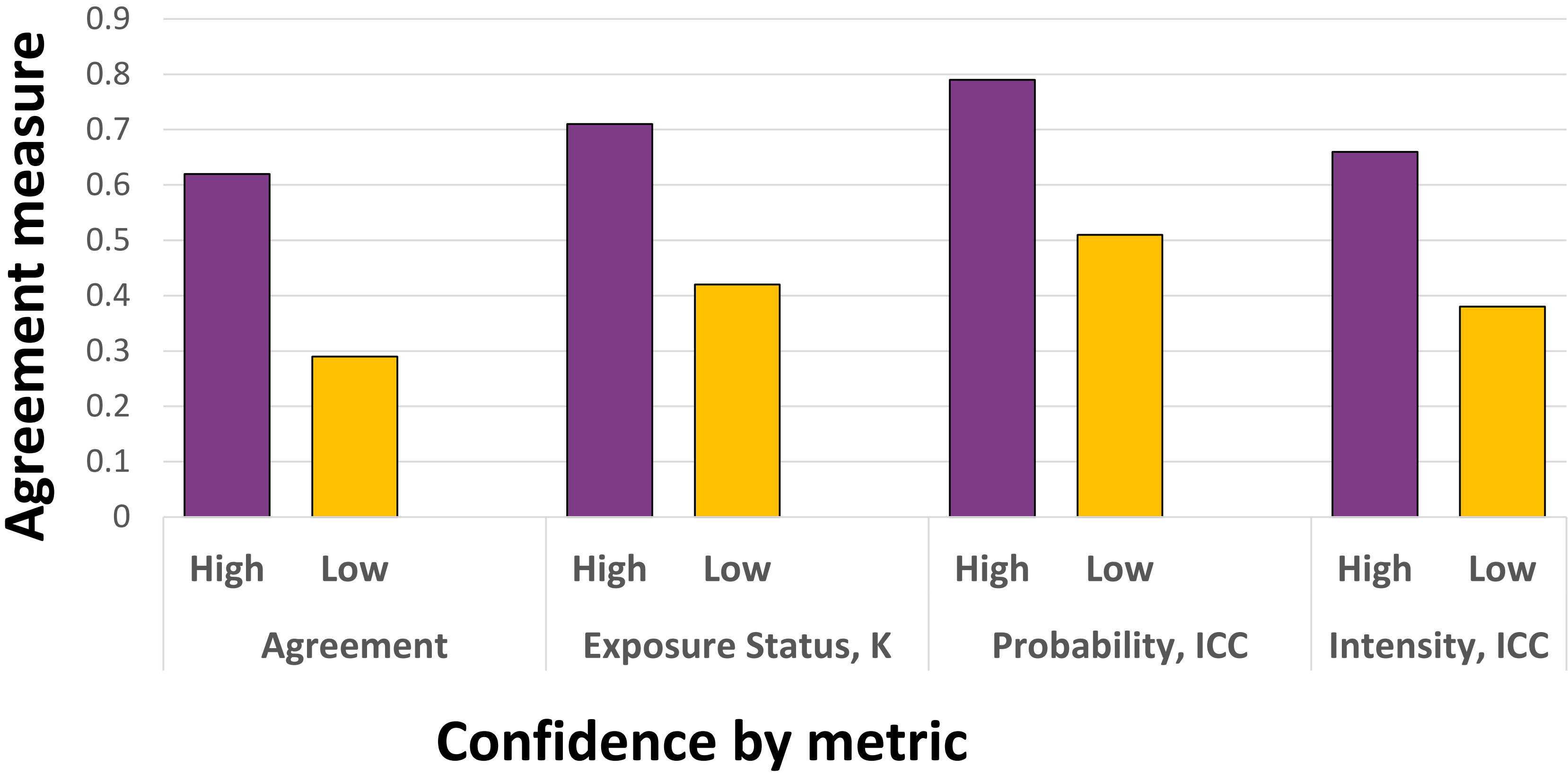
- Generic codes may be difficult to correctly code, but may result in same exposure estimate (e.g., welder)
- Linked expert & highest scoring SOC code to CANJEM
- Compared agreement in exposure estimates
  - Generally similar patterns to SOC code
  - Median kappa on exposed/not exposed: 0.56 (IQR 0.52-0.58)
  - Median ICC on continuous probability metric: 0.66 (IQR 0.58-0.73)
  - Median ICC on continuous probability metric: 0.50 (IQR 0.44-0.56)

Did the top 2 ranked SOC codes assign the same exposure?

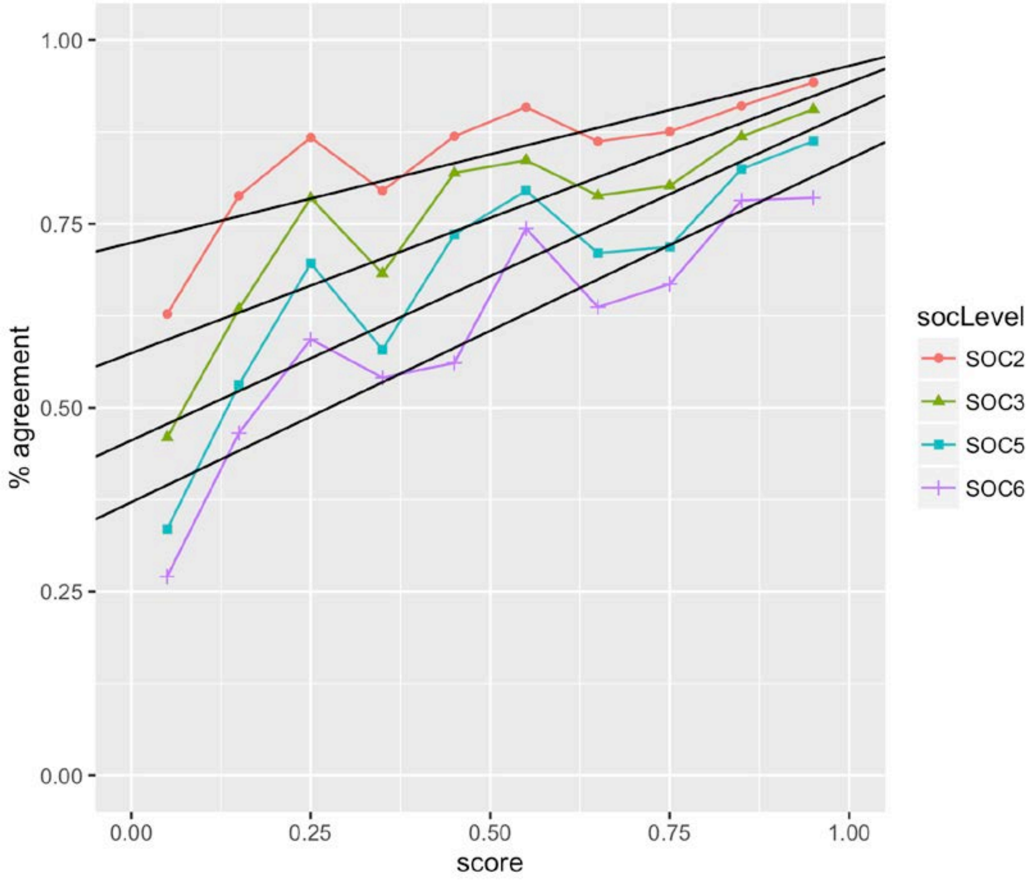


Agreement by confidence in assignment

HIGH: score ≥0.3 *and* score$_{1\text{-}2}$ ≥0.15



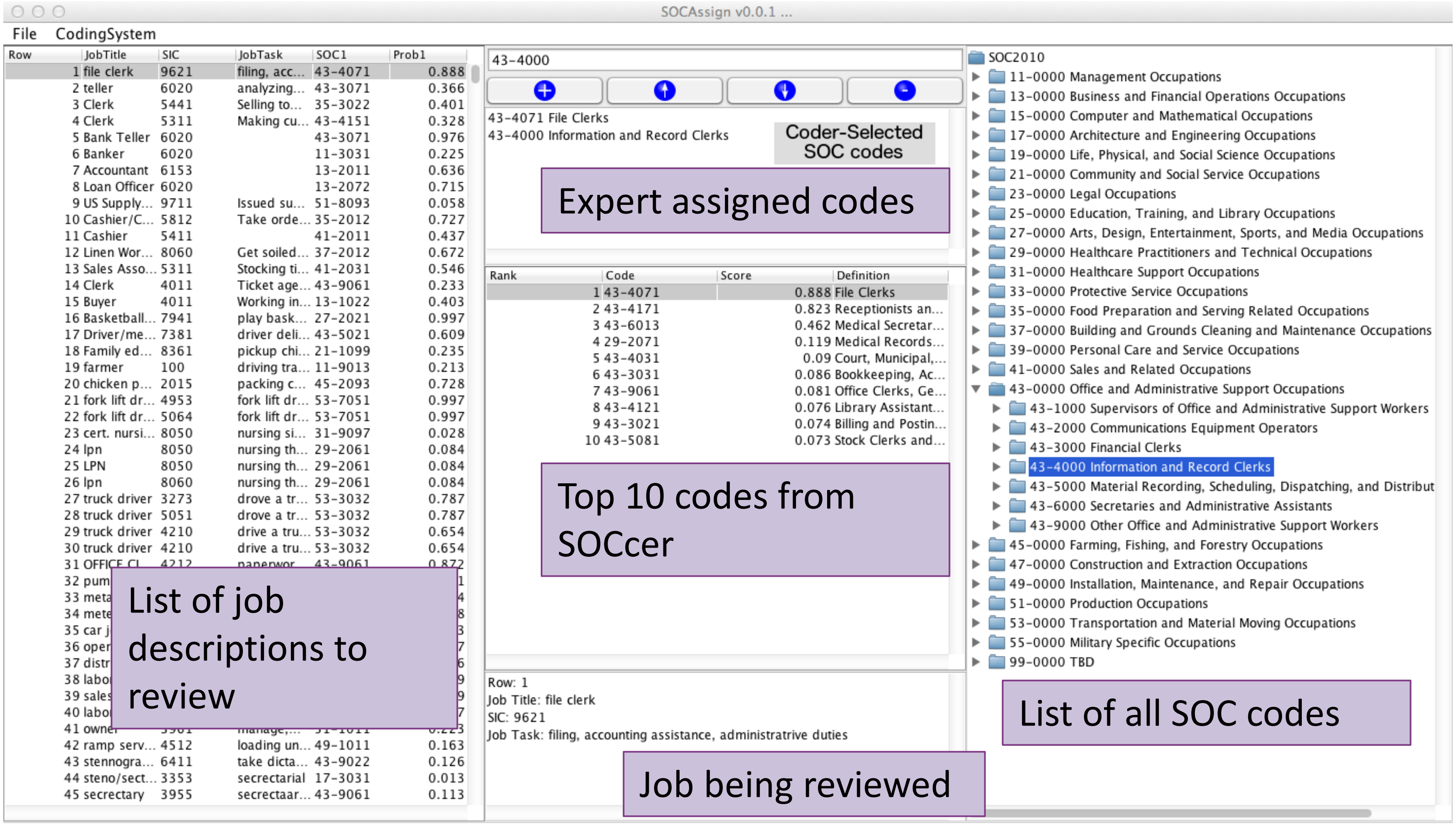
Agreement at varying hierarchy levels



Main findings

- Can reduce expert coding task
- Detailed coding not always possible (data quality)
- 6-digit level
  - Manual coding necessary for nontrivial # of jobs
  - Score and JEM-based metrics to prioritize expert assessment
- 3-digit level
  - Excellent overall agreement and at scores ≥0.25

SOCAssign: Companion Software



Row	JobTitle	SIC	JobTask	SOC1	Prob1
1	file clerk	9621	filing, acc...	43-4071	0.888
2	teller	6020	analyzing...	43-3071	0.366
3	Clerk	5441	Selling to...	35-3022	0.401
4	Clerk	5311	Making cu...	43-4151	0.328
5	Bank Teller	6020		43-3071	0.976
6	Banker	6020		11-3031	0.225
7	Accountant	6153		13-2011	0.636
8	Loan Officer	6020		13-2072	0.715
9	US Supply...	9711	Issued su...	51-8093	0.058
10	Cashier/C...	5812	Take orde...	35-2012	0.727
11	Cashier	5411		41-2011	0.437
12	Linen Wor...	8060	Get soiled...	37-2012	0.672
13	Sales Asso...	5311	Stocking ti...	41-2031	0.546
14	Clerk	4011	Ticket age...	43-9061	0.233
15	Buyer	4011	Working in...	13-1022	0.403
16	Basketball...	7941	play bask...	27-2021	0.997
17	Driver/me...	7381	driver deli...	43-5021	0.609
18	Family ed...	8361	pickup chi...	21-1099	0.235
19	farmer	100	driving tra...	11-9013	0.213
20	chicken p...	2015	packing c...	45-2093	0.728
21	fork lift dr...	4953	fork lift dr...	53-7051	0.997
22	fork lift dr...	5064	fork lift dr...	53-7051	0.997
23	cert. nursi...	8050	nursing si...	31-9097	0.028
24	lpn	8050	nursing th...	29-2061	0.084
25	LPN	8050	nursing th...	29-2061	0.084
26	lpn	8060	nursing th...	29-2061	0.084
27	truck driver	3273	drove a tr...	53-3032	0.787
28	truck driver	5051	drove a tr...	53-3032	0.787
29	truck driver	4210	drive a tru...	53-3032	0.654
30	truck driver	4210	drive a tru...	53-3032	0.654
31	OFFICE CL...	4212	paperwor...	43-9061	0.872
32	pum				1
33	meta				4
34	mete				8
35	car j				3
36	oper				7
37	distr				6
38	labo				9
39	sales				9
40	labo				7
41	owner		manage,...		0.223
42	ramp serv...	4512	loading un...	49-1011	0.163
43	stennogra...	6411	take dicta...	43-9022	0.126
44	steno/sect...	3353	secrectarial	17-3031	0.013
45	secrectary	3955	secrectaar...	43-9061	0.113

List of job descriptions to review

43-4000

43-4071 File Clerks
43-4000 Information and Record Clerks

Coder-Selected SOC codes

Expert assigned codes

Rank	Code	Score	Definition
1	43-4071	0.888	File Clerks
2	43-4171	0.823	Receptionists an...
3	43-6013	0.462	Medical Secretar...
4	29-2071	0.119	Medical Records...
5	43-4031	0.09	Court, Municipal,...
6	43-3031	0.086	Bookkeeping, Ac...
7	43-9061	0.081	Office Clerks, Ge...
8	43-4121	0.076	Library Assistant...
9	43-3021	0.074	Billing and Postin...
10	43-5081	0.073	Stock Clerks and...

Top 10 codes from SOCcer

Row: 1
Job Title: file clerk
SIC: 9621
Job Task: filing, accounting assistance, administratrive duties

Job being reviewed

- SOC2010
  - 11-0000 Management Occupations
  - 13-0000 Business and Financial Operations Occupations
  - 15-0000 Computer and Mathematical Occupations
  - 17-0000 Architecture and Engineering Occupations
  - 19-0000 Life, Physical, and Social Science Occupations
  - 21-0000 Community and Social Service Occupations
  - 23-0000 Legal Occupations
  - 25-0000 Education, Training, and Library Occupations
  - 27-0000 Arts, Design, Entertainment, Sports, and Media Occupations
  - 29-0000 Healthcare Practitioners and Technical Occupations
  - 31-0000 Healthcare Support Occupations
  - 33-0000 Protective Service Occupations
  - 35-0000 Food Preparation and Serving Related Occupations
  - 37-0000 Building and Grounds Cleaning and Maintenance Occupations
  - 39-0000 Personal Care and Service Occupations
  - 41-0000 Sales and Related Occupations
  - 43-0000 Office and Administrative Support Occupations
    - 43-1000 Supervisors of Office and Administrative Support Workers
    - 43-2000 Communications Equipment Operators
    - 43-3000 Financial Clerks
    - 43-4000 Information and Record Clerks
    - 43-5000 Material Recording, Scheduling, Dispatching, and Distribut
    - 43-6000 Secretaries and Administrative Assistants
    - 43-9000 Other Office and Administrative Support Workers
  - 45-0000 Farming, Fishing, and Forestry Occupations
  - 47-0000 Construction and Extraction Occupations
  - 49-0000 Installation, Maintenance, and Repair Occupations
  - 51-0000 Production Occupations
  - 53-0000 Transportation and Material Moving Occupations
  - 55-0000 Military Specific Occupations
  - 99-0000 TBD

List of all SOC codes

Preview of future advancements

- **SOCcer 1.1**
  - Expand training data to include job descriptions from epidemiologic studies and retrain algorithm
  - Increases **overall** 6-digit agreement to **>50%**
  - March-April 2017
- **SOCcer 2.0**
  - Add and refine classifiers and add training data
  - Increases **overall** 6-digit agreement to **~60%**
  - Late 2017
- **Consider more than one plausible code?**

Expansion to other systems?

- **Requires training data in target system**
  - Previously coded job descriptions
  - O*NET equivalent data source
  - Crosswalk to US SOC 2010
- **Canadian system?**
  - Current and previous versions
  - English & French?

Thank you

- *Expert coders*
  - Susan Viet, Pabitra Josse, Sarah Locke
- *CANJEM*
  - Jerome Lavoue, Thomas Remen
- *Epidemiologic studies*
  - Connecticut Thyroid Study
  - Montreal Lung Cancer Study
  - NCI-SEER NHL Study
  - New England Bladder Cancer Study
  - US Renal Cell Cancer Study