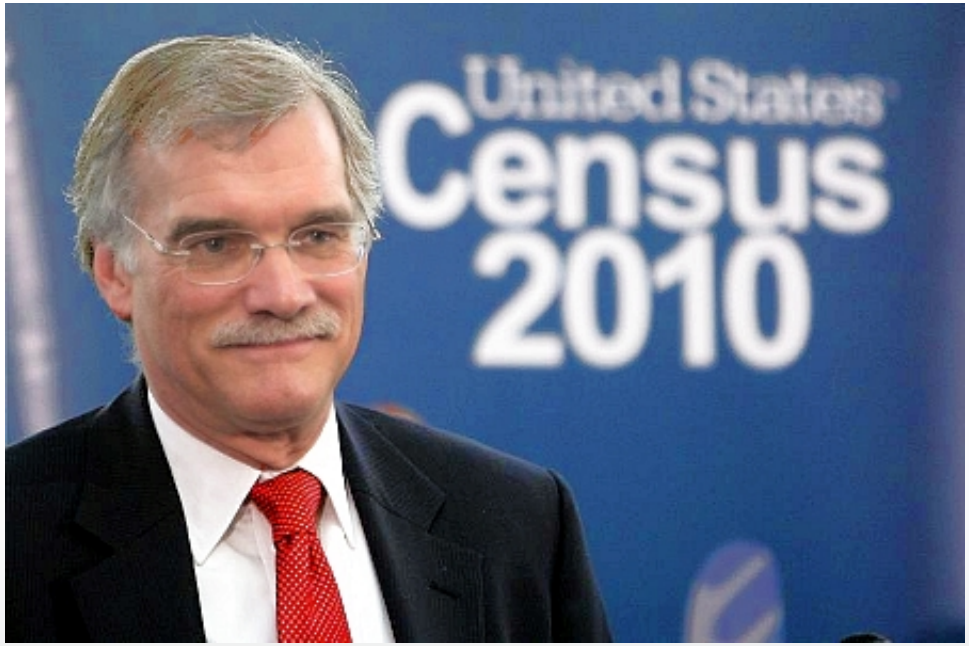
Using an Autocoder to Code Industry and Occupation in the American Community Survey

By Jennifer Cheeseman Day

Presentation for the Federal Economic Statistics Advisory Committee Meeting

June 13, 2014

Opportunity



In summer of 2010 Census Bureau Director Robert Groves announced the **Improving Operational Efficiency program,** to fund development of cost saving ideas suggested by Bureau staff.

Why Autocode Industry and Occupation?

- In 2010, workload of 2.48 million manually coded records per year
- Clerical - very labor intensive
  - Cost
  - Time
- Concern about accuracy and consistency
  - Keying and other errors
  - Variation in coding clerk's interpretations of the respondent's write-in
- Future increase in workload (e.g., sample expansion, bridge coding, competing surveys)

American Community Survey Variables Autocoded

Type and Method of Coding

Race Hispanic Origin Ancestry Language Health Insurance Field of Degree Computer Types Internet Service	Backcoding, automated with clerical follow-up
Industry	**Industry, automated with clerical follow-up**
Occupation	**Occupation, automated with clerical follow-up**
Place of Birth Migration Place of Work	Geocoding Automated with clerical follow-up

Implementation

ACS data year 2012

	autocoded
Industry	**55%**
Occupation	**43%**
Joint I&O	**29%**

$avings

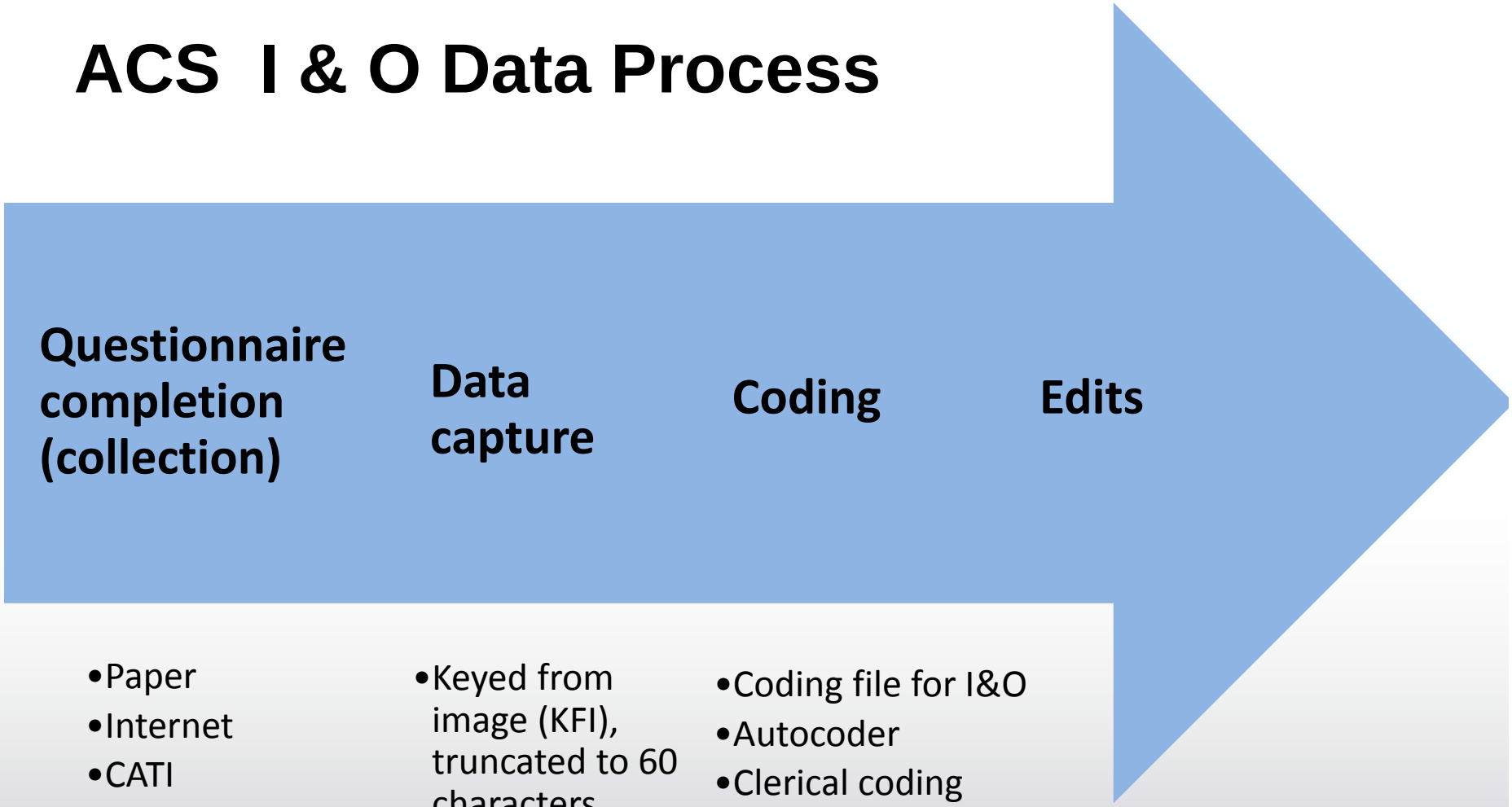
Year	Number of records	Autocoded Records	Records sent to clerical coding	Joint Coding Rate
2012	3,016,617	887,750	2,128,867	29.4
2013	2,823,924	824,122	1,999,802	29.2
2014 (1st qtr)	676,810	197,458	479,352	29.2

Autocoder

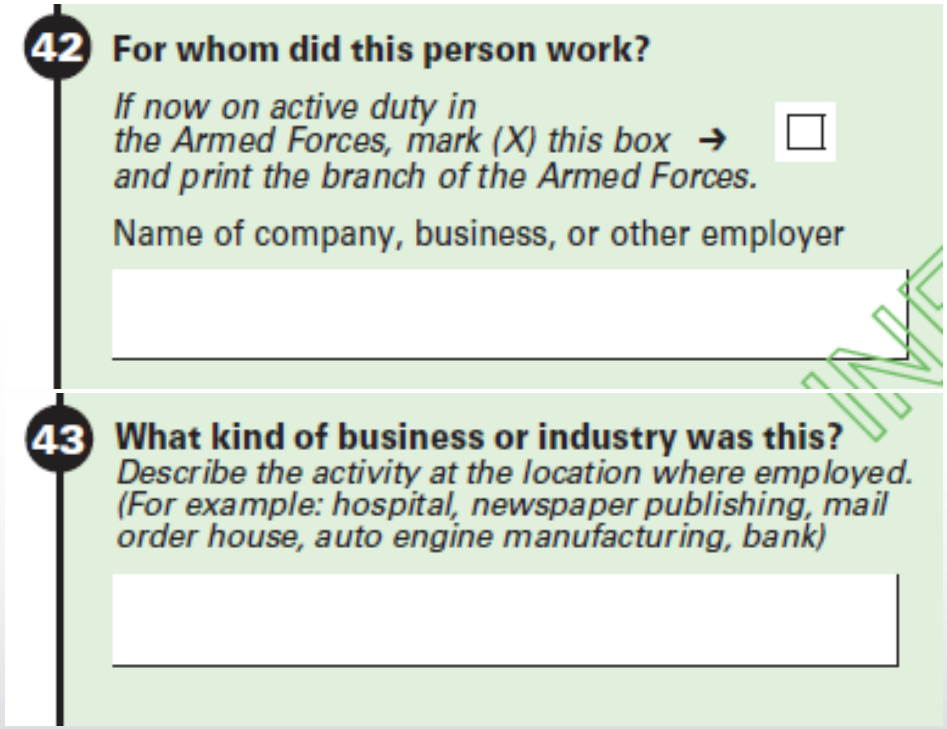
- All records with I&O write-in information go through autocoder
- Designed to replicate clerical coding
- Industry and occupation coded separately
- Assigned codes with quality scores below cutoff go to clerical coding
- Result:
  - 30% both codes assigned (no clerical coding)
  - 40% have one code assigned (partially coded)
  - 30% no codes assigned

Goes to clerical coding

ACS I & O Data Process



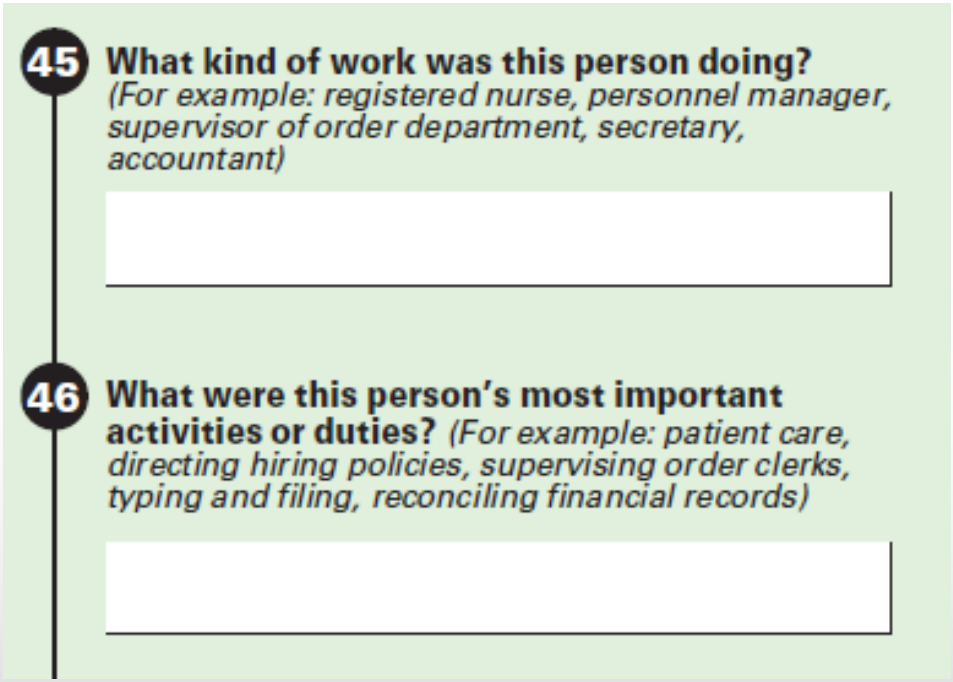
ACS Questionnaire Industry Items



- Industry data describe the kind of business conducted by a person's employing organization

- 3 industry questions
  - 2 write-ins
  - 1 checkbox

ACS Questionnaire Occupation Items


45 **What kind of work was this person doing?** *(For example: registered nurse, personnel manager, supervisor of order department, secretary, accountant)*

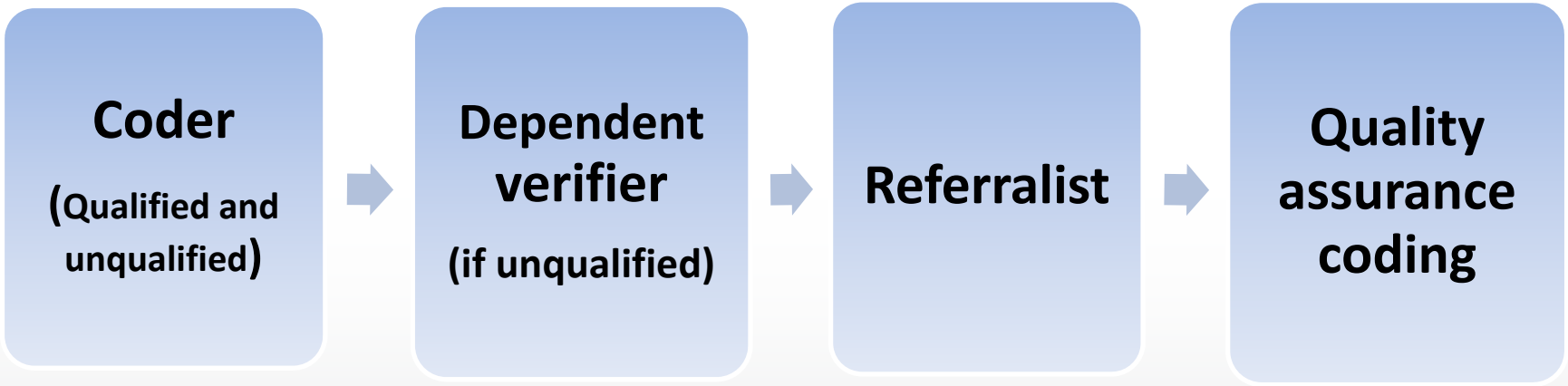
46 **What were this person's most important activities or duties?** *(For example: patient care, directing hiring policies, supervising order clerks, typing and filing, reconciling financial records)*

- Occupation describes the kind of work a person does on the job
- 2 occupation questions

Example of Write-ins

Industry Write-in (INW3)	Occupation Write-in (OCW1)
BANK	ASSISTANT VP LOAN DEPARTMENT
CONSTRUCTION COMPANY	PROJECT MANAGER
CONSTRUCTION CONTRACTER	BRIDGE CARPENTER
CROP CATTLE FARM	ALL JOBS THAT NEEDED DOING
UNIVESITY HIGHER EDUCATION	ASSISTANT DEAN/ADMINISTRATION
CHILED CARE	BBOSS
PROPERTY MANAGEMENT	PROPERTY MANAGER
MEAT PROCESSING	MEAT PROCESSING
BOWLING ALLEY- SNACK BAR	NIGHT SUPERVISOR FOR SNACK BAR
STEEL BRIDGE FABRICATOR	MACHINE PROGRAMMER-LAYOUT
HIGH SCHOOL	COUNSELOR
NON-PROFIT LAW FIRM	ATTORNEY
UNIVERSIDAD	AYUDANTE DE PROGRAMA P E A N
EDUCATION	INSTRUTOR OF ENGLISH
RESTAURANT	COOK
PIZZA	HE MAKES PIZZAS
BURYING PEOPLE	I BURY THE DEAD
MILITARY	RADIO MAINTENANCE
GOVERNMENT	D

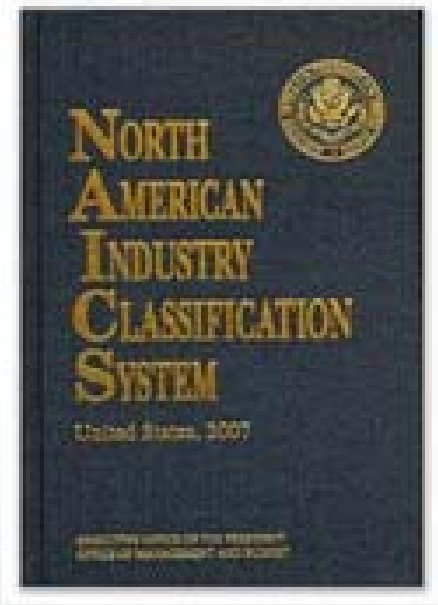
Clerical Coding Process



Variables Coders Use to Code I&O

- Age
- Sex
- Date of birth
- Educational attainment
- Residence county, state
- Active duty Armed Forces checkbox
- Class of worker checkbox
- Employer name write-in
- Kind of business write-in
- Industry type checkbox
- Kind of work write-in
- Job duties write-in

Industry and Occupation Codes

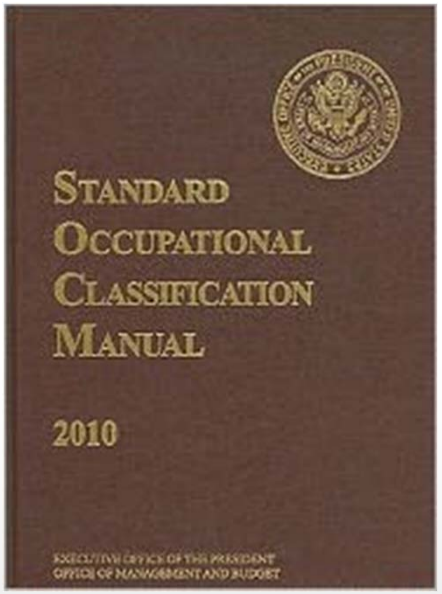


Census INDUSTRY Codes

- Covers all 20 sectors
- Classified based on NAICS two-digit through six-digit codes
- 269 Census industry codes (4 digits)

Census OCCUPATION Codes

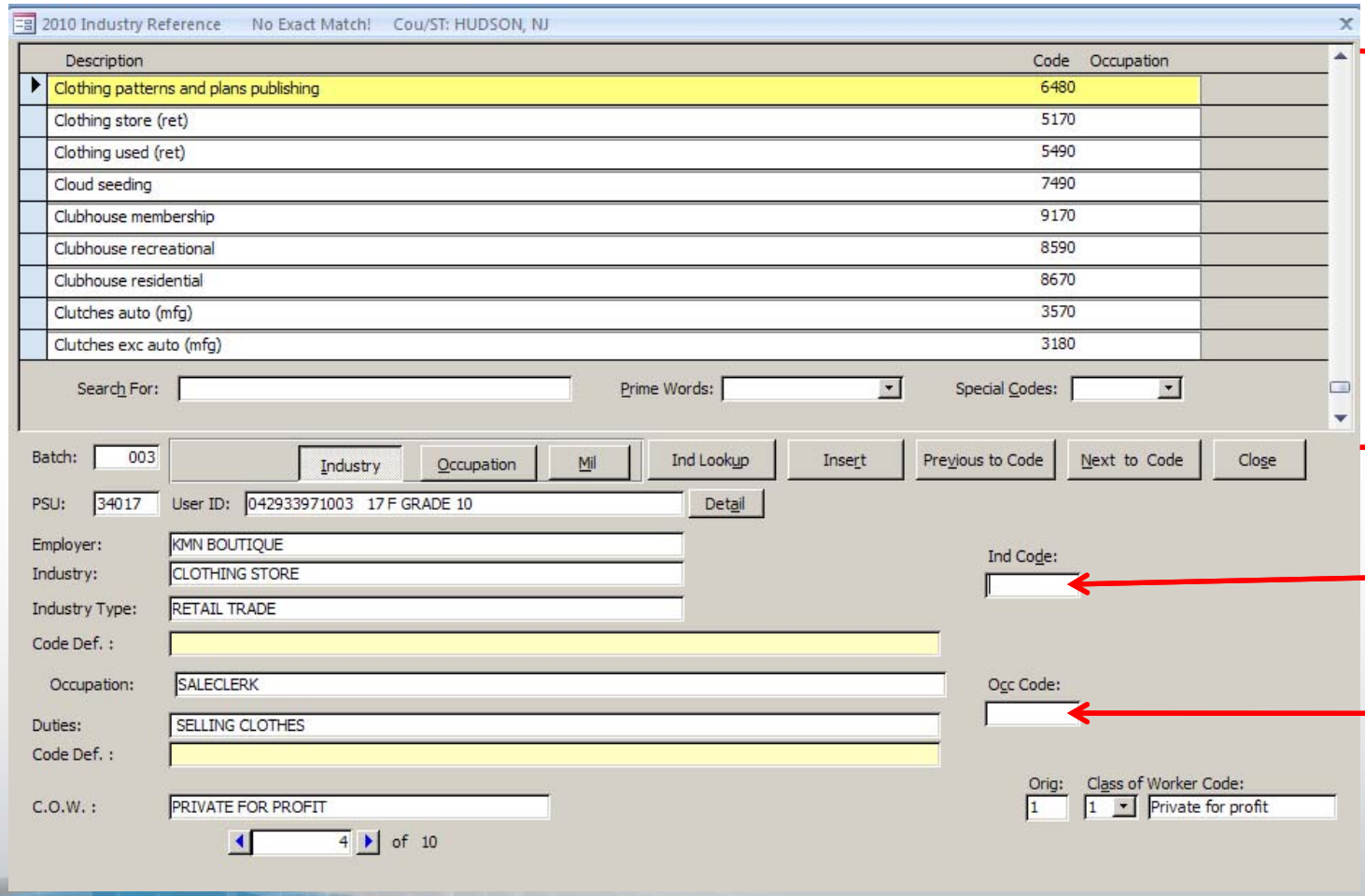
- Covers all 23 major occupation groups
- Based on SOC two-digit through six-digit codes
- 539 Census occupation codes (4 digits)



Coding Indexes

OCCUPATION TITLE	NAICS RESTRICTION	SOC CODE	INDUSTRY RESTRICTION	OCCUPATION CODE
Teacher, elementary school	6111	25-2021	7860	2310
Teacher, french	6112, 6113	25-1124	7870	2200
Teacher \ n.s.	51331	43-2011	6680	5010
Teacher \ n.s.	6112, 6113	25-1199	7870	2200
Teacher \ n.s.	524, 8121, 6114, 6115	25-1194	6990, 7880, 8870, 8880, 8970	2200
Teacher \ n.s.	Bible school 611699	21-2099	Bible school 7890	2060
Teacher \ n.s.	Elementary school 6111	25-2021	Elementary school 7860	2310

I&O Coding System

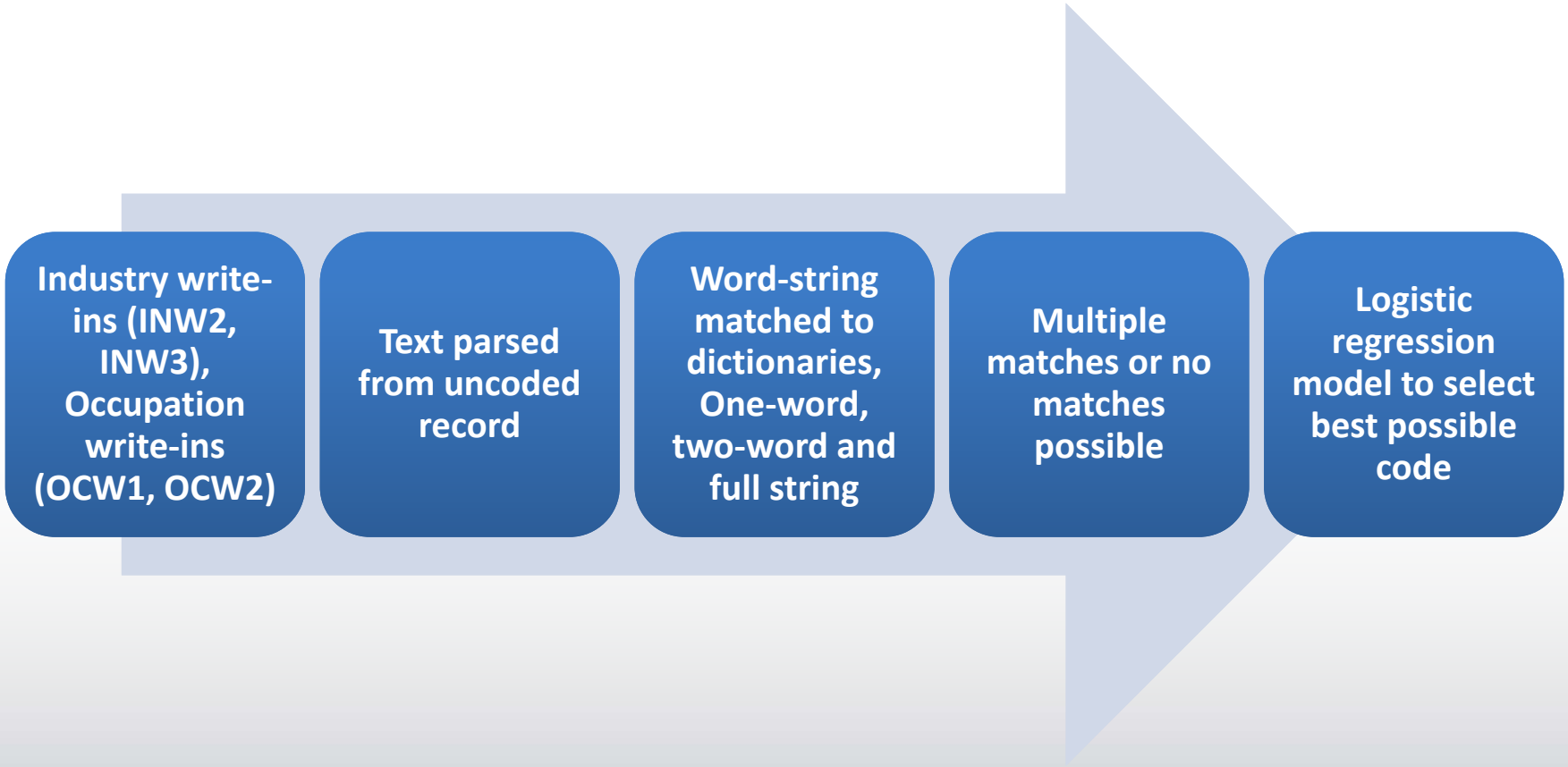


Industry, occupation, or military Index

Industry code

Occupation code

Autocoder Process



12 Autocoder Dictionaries

- 1.5 million coded records from 2010 ACS used to build dictionaries – random selection
- 3 dictionaries for each write-in variable – 6 for industry and 6 for occupation
- Cross-coding in dictionaries to assign industry from occupation entries and vice versa

Autocoder Dictionaries

		1 word bit	2 word bits	full write in
Industry	INW1	1	2	3
	INW2	4	5	6
Occupation	OCW1	7	8	9
	OCW2	10	11	12

Autocoder Dictionaries Criteria

Criteria for inclusion:

- 30 occurrences for 1-word, 2-word
- 15 occurrences for full-string
- 50% map to one code
- 75% map to one code for cross-codes

Dictionaries include word-bit, code, total frequency and frequency percentage

Examples from Dictionaries for Occupation Write-in

OCW1	wordcnt	occ	occcnt	freqpct	ind	indcnt	indpct
PICKING	43	6050	25	58.1%		0	0.0%
PICKING FRUIT	22	6050	20	90.9%		0	0.0%
PICKING GRAPES	31	6050	31	100.0%	170	24	77.4%
PICKING GRAPES	33	6050	33	100.0%	170	26	78.8%
PICKING ORANGES	15	6050	15	100.0%	170	13	86.7%
PICKING UP TRASH	17	9720	11	64.7%	7790	14	82.4%

Selecting the Best Code: Logistic Regression Model

- 2 models: industry and occupation
- SAS Proc Logistic with Stepwise
- Independent variables:
  - Variables used in clerical coding
  - Plus Coding dictionary (frequency percentage, total frequency)
  - 111 variables in industry model; 91 variables in occupation model
- Dependent variable
  - 1 if dictionary match agrees with assigned code
  - 0 if code does not agree
- Model estimates the probability the code agrees with what a clerk would assign

Hardcodes

- Industry or occupation code reassignment based on additional text information
- Example: elementary and secondary education reassigned to colleges and universities based on ‘University’
- Corrects most common errors

Ensuring Quality

Disagreement Rates Between Autocoder and Clerical Coding*

Coding Rate	Industry	Occupation
30%	1.01%	4.49%
40%	2.12%	5.30%
50%	3.96%	8.35%
60%	5.39%	13.43%

*From expert coding test with 2,000 records

Cutoff Level

- Cutoffs based on model score
  - Keep code if score>cutoff
  - Separate cutoffs for industry and occupation
  - Set cutoffs to generate similar quality as 100% clerical coding
- Records with scores below cutoff go to clerical coding

	Industry	Occupation
Error rate	4.5%	5.9%
Acceptable Codes	56%	43%

Production QC Process

- Sample by code categories
- 3 times annually
- Clerical coding done by referralists
- Track error rates
- Evaluate problem wordbits
- Test change in autocoder

Updating Codes

- Constructed using 2010 I&O coding
- Updates based on I&O Indexes, QC, and coding changes
  - 2010 – occupation updates consistent with 2010 SOC changes
  - 2012 – industry updates consistent with 2012 NAICS update

Industries with Highest Frequency of Autocoding

Industry	% autocoded	Industry	% autocoded
Postal service	95	Legal services	83
Elementary and secondary schools	94	Veterinary services	82
Beauty salons	92	Funeral homes	81
Offices of dentists	90	Rail transportation	80
Restaurants	90	Landscaping services	77
Religious organizations	86	Libraries	77
Hospitals	86	Barber shops	76
Banking	84	Grocery stores	76
Insurance carriers	84	Accounting services	75
Coast guard	84	Bowling centers	75
Justice, public order, and safety	84	Nursing care	74

Half of autocoded records to these industries

Occupations with Highest Frequency of Autocoding

Occupation	% autocoded	Occupation	% autocoded
Hairdressers	95	Clergy	85
Speech-lang pathologists	91	Dentists	84
Massage therapists	88	Dishwashers	83
Postal service mail carriers	88	Bartenders	82
Dental hygienists	88	Hosts	81
Flight attendants	87	Truck drivers	81
Elementary and middle school teachers	86	Pilots	80
Firefighters	86	Cashiers	79
Lawyers	86	Dental assistants	78
Waiters	85	Nurse practitioners	78
Respiratory therapists	85	Misc personal appearance	78

One-quarter of autocoded records to these occupations

Autocoding Rates for Largest Industries

Industry	% autocoded
Elementary and secondary schools	94
Construction	72
Restaurants	90
Hospitals	86
Colleges	67
Grocery stores	76
Justice, public order, and safety	84
Department stores and discount stores	42
Insurance carriers	84

These industries total one-third of I&O records across all industries.

Autocoding Rates for Largest Occupations

Occupation	% autocoded
Secretaries	66
Cashiers	79
Elementary and middle school teachers	86
Retail salespersons	24
Truck drivers	81
Managers, all other	3
Janitors and building cleaners	38
Registered nurses	72
First-line supervisors of retail sales workers	13
Nursing aides	57
Customer service	52
Cooks	77
Laborers and freight, stock, and material movers, hand	24
Waiters	85
Accountants	52
Construction laborers	8
Stock clerks	54

These occupations total one-third of records for all occupations

Codes Not Autocoded

Industry ~26 codes

- 10 percent of codes
- 1.3 percent of records

Occupation ~ 100 codes

- 20 percent of codes
- 1.6 percent of records

Partial Coding

Clerical coders can change the autocoded code

Industry:	1.8 percent of industry partial codes
Occupation:	1.5 percent of occupation partial codes

Evaluating the 2012 Autocoder
(in progress)

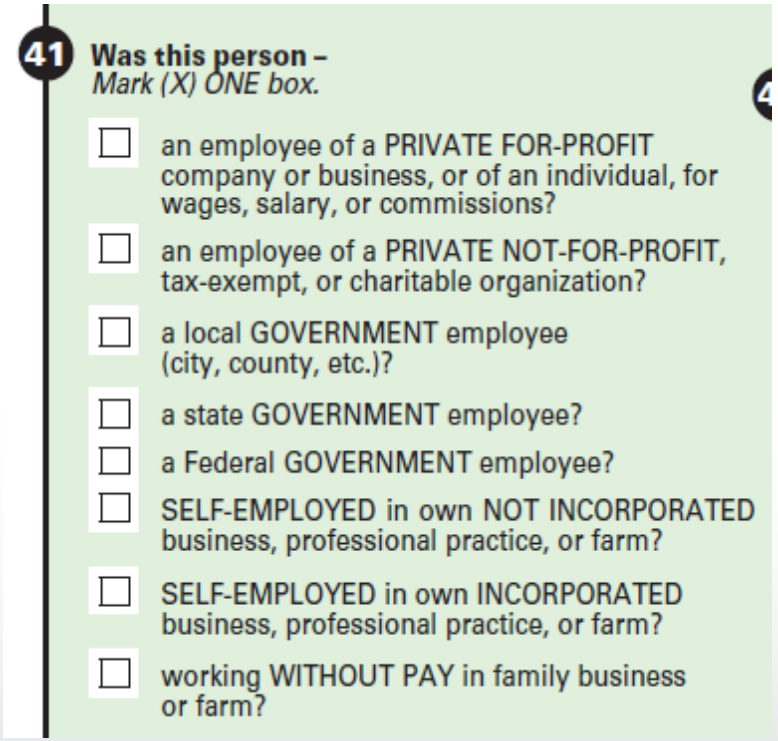
- **Multiple meaning for key words**
  - Manager, dealer, editor
- **Too much weight given to particular words**
  - E.g., “X-ray installer”
- **Phrasing of activities**
  - “engineering manager” vs. “engineer, managing a team”
- **Spelling - not all variations included**
- **Occupation types difficult to autocode -** “all other”, managers, engineers, teachers, laborers, designers, entertainers, editors, cafeteria workers

Unexpected Consequences

Class of worker:

- Can be changed by ACS clerical coders
- Not changed by autocoder
- Not changed in CPS clerical coding

Distributions vary, particularly when crossed by occupation



41 **Was this person –**
Mark (X) ONE box.

- ☐ an employee of a PRIVATE FOR-PROFIT company or business, or of an individual, for wages, salary, or commissions?
- ☐ an employee of a PRIVATE NOT-FOR-PROFIT, tax-exempt, or charitable organization?
- ☐ a local GOVERNMENT employee (city, county, etc.)?
- ☐ a state GOVERNMENT employee?
- ☐ a Federal GOVERNMENT employee?
- ☐ SELF-EMPLOYED in own NOT INCORPORATED business, professional practice, or farm?
- ☐ SELF-EMPLOYED in own INCORPORATED business, professional practice, or farm?
- ☐ working WITHOUT PAY in family business or farm?

Can we do better?

- How can we improve our autocoders with minimal additional resources?

- How can we best utilize the autocoders, particularly during real-time data collection?

More Information

Census Industry and Occupation website:
http://www.census.gov/people/io/

Industry and Occupation Statistics Branch
Telephone: 301-763-3239

Jennifer.Cheeseman.Day@Census.gov
Telephone: 301-763-3399