NATURAL LANGUAGE PROCESSING… 21st-CENTURY SURVEILLANCE

PHILIP HARBER MD MPH

Gondy Leroy PhD

Univ of Arizona

Medical Research Building- Room 112
University of Arizona- MEZCOPH
1656 E. Mabel St.
Tucson, AZ 85719
Telephone: 520-626-1263
pharber@email.Arizona.edu

Working Words: Real-Life Lexicon of
American Workers

Philip Harber, MD, MPH
Lori Crawford, BS
Katie Liu, BS
Levanto Schacter, MS

Work Coding: Beyor
and DOT Philip Harbe

Computer Algorithm for Automated Work
Group Classification From Free Text: The
DREAM Technique

Philip Harber, MD, MPH
Lori Crawford, BS
Amarpreet Cheema, BS
Levanto Schacter, MS

, BOC
, and Justine Smitherman, MS

Feasibility and Utility of Lexical Analysis for Occupational Health Text

Philip Harber, MD, MPH and Gondy Leroy, PhD

Assessing Work–Asthma Interaction With Amazon Mechanical Turk

Philip Harber, MD, MPH and Gondy Leroy, PhD

Social media use for occupational lung disease

Public sharing of medical advice using social media: an analysis of Twitter

Gondy Leroy*[1], PhD, Philip Harber[2], MD MPH, Debra Revere[3], MLIS, MA

- Thank you for the opportunity to speak here.

- “Hello there, my name is Phil Harber. I am currently speaking, but sometimes I see patients as a doctor in occupational medicine or in pulmonary medicine. I also play with computers and try to make them understand language sometimes. I am employed by a university and in a public health college but also go to the hospital sometimes. I sometimes do some administrative work and even traveled on aircraft yesterday. Last week, I coughed a little after I open my mouth doing a racing turn in the swimming pool. “

- “I was welding on the stainless storage tank for the apple juice but my foot slipped and I had when in so I began to cough. The doctor said I might have asthma or maybe just a bad cold. “

We talk and think in words, not numbers

Using codes loses most information of prior slide:

- Good morning 19-1041
- 51-4120 111339 J45
- 786.2 541511 611310

- Language is rich, natural, flexible, and cheap.

- Codes are narrow, inter-relate poorly, and are not spoken by “real workers”.

Today’s Tools, Tomorrow’s Needs

- Storage: ~ Unlimited
- Computing Power:
- Data Sources: Many, mainly unstructured
- Programs- Rapidly evolving

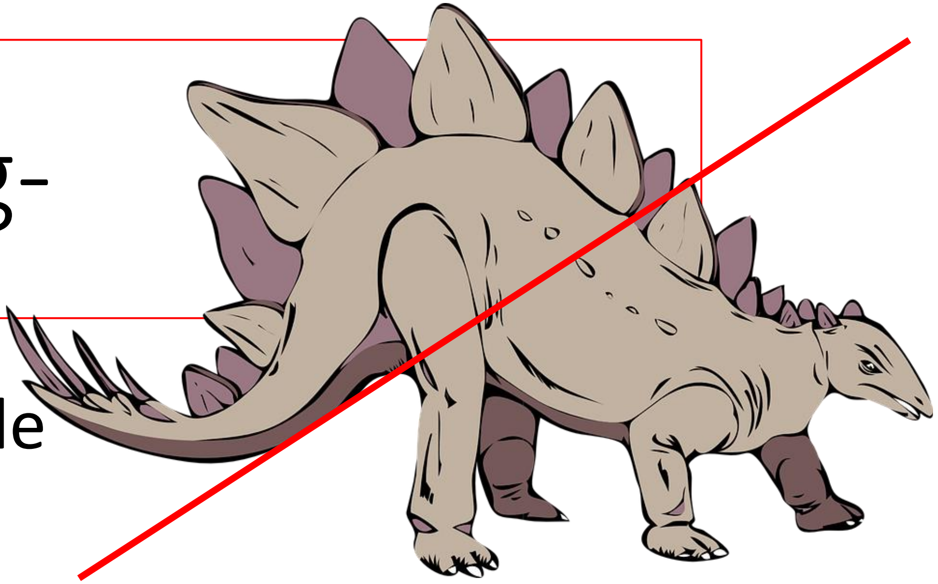
…But….

- Interoperability
- Motivation to use
- //pixabay.com/en/cartoon-cat-face-feline-unhappy-1296508/

P Harber Univ of Arizona

https://pixabay.com/en/cloud-cumulus-clouds-cumulus-8075/
https://pixabay.com/en/abacus-count-mathematics-485704/

Limitations of Numerical Coding-



- Many jobs, industries, situations - poorly codable
- Only predesignated questions
- Highly structured input and highly structured output
- Very simple questions
- Modern Luddites

- Codes vs real life

Use #1: Autocoding (Structured → Structured)
Code: Replace a structured field with a structured number:

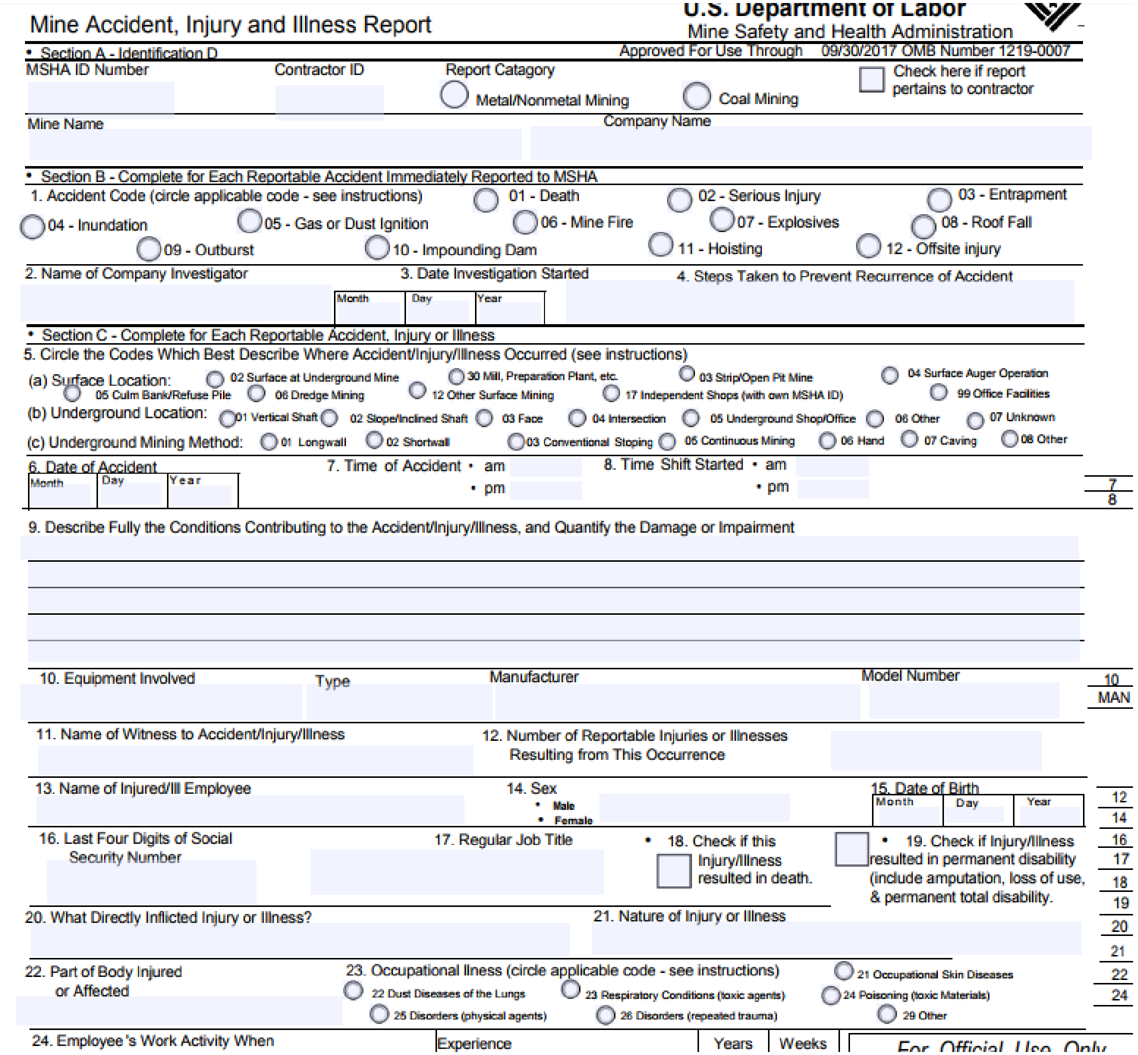
eg, Death certificate- Industry field → NAICS Code

Auto coding: Replace Nosologists with software

- Faster, cheaper, ? More accurate
- NIOSH NIOCCS, …

https://pixabay.com/en/clown-comedian-nose-circus-funny-362155/
https://pixabay.com/en/cat-relax-chill-out-camacho-98359/
https://pixabay.com/en/face-nose-eyes-plush-funny-2520636

#2: Abstraction



U.S. Department of Labor
Mine Safety and Health Administration

Mine Accident, Injury and Illness Report

• Section A - Identification D

Approved For Use Through 09/30/2017 OMB Number 1219-0007

MSHA ID Number

Contractor ID

Report Catagory ○ Metal/Nonmetal Mining ○ Coal Mining

□ Check here if report pertains to contractor

Mine Name

Company Name

• Section B - Complete for Each Reportable Accident Immediately Reported to MSHA

1. Accident Code (circle applicable code - see instructions) ○ 01 - Death ○ 02 - Serious Injury ○ 03 - Entrapment ○ 04 - Inundation ○ 05 - Gas or Dust Ignition ○ 06 - Mine Fire ○ 07 - Explosives ○ 08 - Roof Fall ○ 09 - Outburst ○ 10 - Impounding Dam ○ 11 - Hoisting ○ 12 - Offsite injury

2. Name of Company Investigator

3. Date Investigation Started Month Day Year

4. Steps Taken to Prevent Recurrence of Accident

• Section C - Complete for Each Reportable Accident, Injury or Illness

5. Circle the Codes Which Best Describe Where Accident/Injury/Illness Occurred (see instructions)

(a) Surface Location: ○ 02 Surface at Underground Mine ○ 30 Mill, Preparation Plant, etc. ○ 03 Strip/Open Pit Mine ○ 04 Surface Auger Operation ○ 05 Culm Bank/Refuse Pile ○ 06 Dredge Mining ○ 12 Other Surface Mining ○ 17 Independent Shops (with own MSHA ID) ○ 99 Office Facilities

(b) Underground Location: ○ 01 Vertical Shaft ○ 02 Slope/Inclined Shaft ○ 03 Face ○ 04 Intersection ○ 05 Underground Shop/Office ○ 06 Other ○ 07 Unknown

(c) Underground Mining Method: ○ 01 Longwall ○ 02 Shortwall ○ 03 Conventional Stoping ○ 05 Continuous Mining ○ 06 Hand ○ 07 Caving ○ 08 Other

6. Date of Accident Month Day Year

7. Time of Accident • am • pm

8. Time Shift Started • am • pm

7
8

9. Describe Fully the Conditions Contributing to the Accident/Injury/Illness, and Quantify the Damage or Impairment

10. Equipment Involved Type Manufacturer Model Number

10
MAN

11. Name of Witness to Accident/Injury/Illness

12. Number of Reportable Injuries or Illnesses Resulting from This Occurrence

13. Name of Injured/Ill Employee

14. Sex • Male • Female

15. Date of Birth Month Day Year

16. Last Four Digits of Social Security Number

17. Regular Job Title

• 18. Check if this □ Injury/Illness resulted in death.

□ • 19. Check if Injury/Illness resulted in permanent disability (include amputation, loss of use, & permanent total disability.

12
14
16
17
18
19
20
21
22
24

20. What Directly Inflicted Injury or Illness?

21. Nature of Injury or Illness

22. Part of Body Injured or Affected

23. Occupational Ilness (circle applicable code - see instructions) ○ 21 Occupational Skin Diseases ○ 22 Dust Diseases of the Lungs ○ 23 Respiratory Conditions (toxic agents) ○ 24 Poisoning (toxic Materials) ○ 25 Disorders (physical agents) ○ 26 Disorders (repeated trauma) ○ 29 Other

24. Employee's Work Activity When

Experience Years Weeks

For Official Use Only

Abstraction

- Read the item
- Find the key term(s)
- Look up number
- Enter number in correct field
  - ☹ MSHA 7000 has *34* fields
- Needs domain knowledge

#3: Classifier

- Sorts items into one of a few alternative boxes
  - NIOSH trip/fall workers comp (Bertke)
  - CASCOT (UK)
  - DREAM (Harber)
  - SOCEYE (NCI)

- Numerous algorithms-identification, rule-based, probabilistic
- Most require **training set** and are **very specific**

- https://pixabay.com/en/boxes-drawers-mailboxes-1834406/

4: Search

Find the needle in the haystack

- Look for a specific term
- **Store raw words, not post- coded numbers**

- **Does the public care about work-related asthma?**
- Are ATS, AOEC, NIOSH outreach working ????

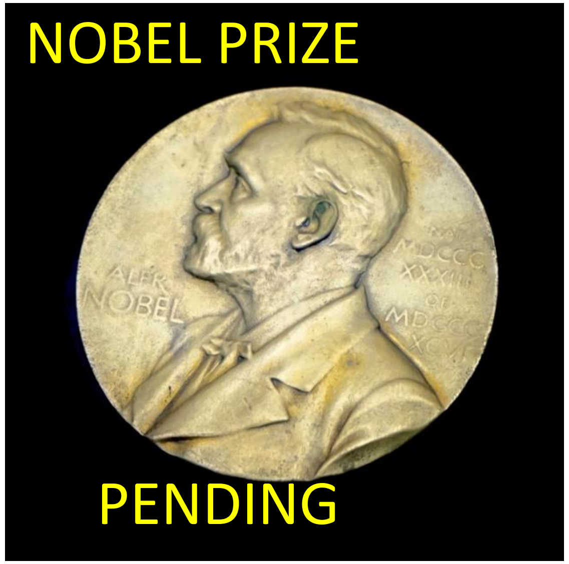
TWITTER:

- 40,000 tweets with “asthma”
- Very few involved a job, isocyanates,…
- Many, many, many included drugs, cam, children, air pollution, ozone,…

- https://pixabay.com/en/magnifying-glass-loupe-search-145942/ :: https://pixabay.com/en/needle-in-a-haystack-search-find-1706106/ https://pixabay.com/en/twitter-tweet-bird-funny-cute-117595/

What we did…

5. Classify ALL items
6. Knowledge Discovery



- 86,000 MSHA incident reports
- Tokenize, Count words
- Eliminate irrelevant words (stop terms, low relevance)
- Annotate : assign to a domain
  - (e.g., exposure, injury/illness, medical care, action, …)
- Count by domain
- Seek associations

What we did…

5. Classify **ALL** items
6. Knowledge **Discovery**

TOO POMPOUS

- 86,000 MSHA incident reports
- Tokenize, Count words
- Eliminate irrelevant words (stop terms, low relevance)
- Annotate : assign to a domain
  - (eg, exposure, injury/illness, medical care, action, …)
- Count by domain
- Seek associations

Probabilistic

Bayesian classifiers

Syntactic analyzers

Pattern recognition

Support Vector Machine

Training set

Parser

Normalizer

→ Lexical Analyzer

Negation

Temporal relationships

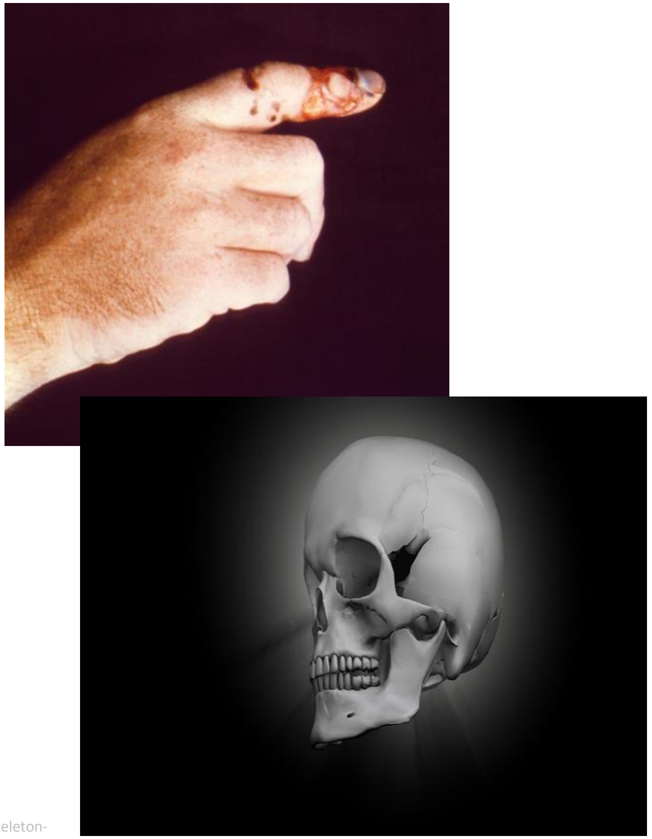
Sentiment

Summary:

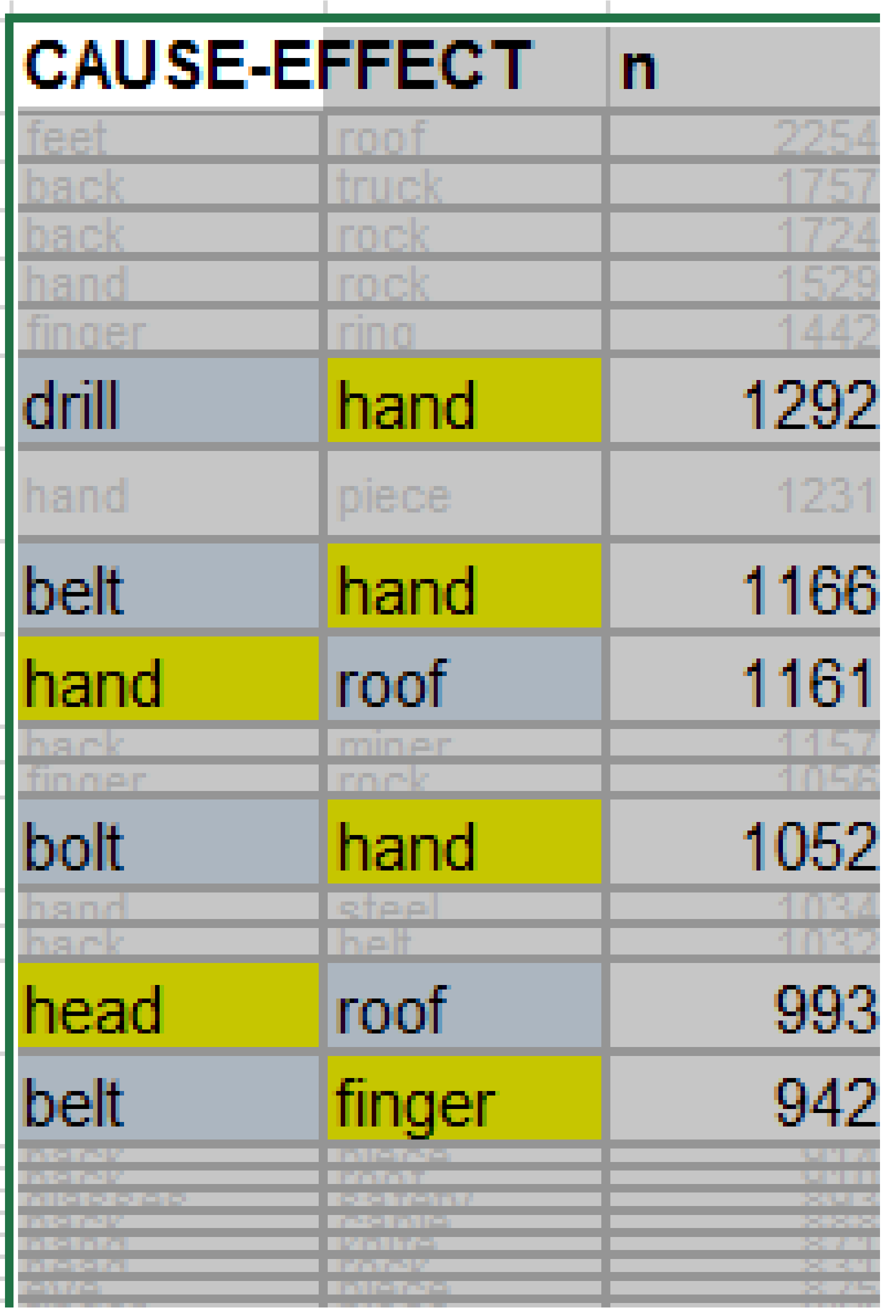
- Common Words (Injury/Illness Exposure/Hazard Treatment …)
- Common causes
- Common effects
- Common cause-effect relationships

Associations: Cause & Effect

- *Automatic* discovery of RELATIONS between HAZARDS and HEALTH EFFECTS



https://pixabay.com/en/skull-bone-head-skeleton-3d-model-1557446/ CDC PHIL

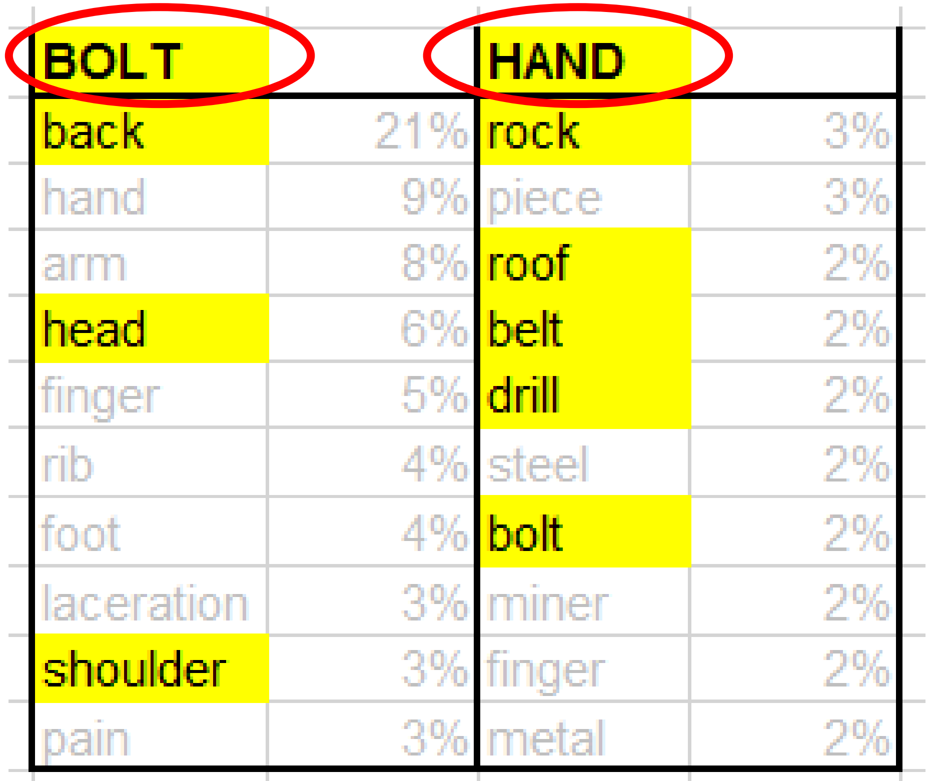


CAUSE-EFFECT		n
feet	roof	2254
back	truck	1757
back	rock	1724
hand	rock	1529
finger	ring	1442
drill	hand	1292
hand	piece	1231
belt	hand	1166
hand	roof	1161
back	miner	1157
finger	rock	1056
bolt	hand	1052
hand	steel	1034
back	belt	1032
head	roof	993
belt	finger	942
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]

OUCH!!*#

Targeted a priori example searches-

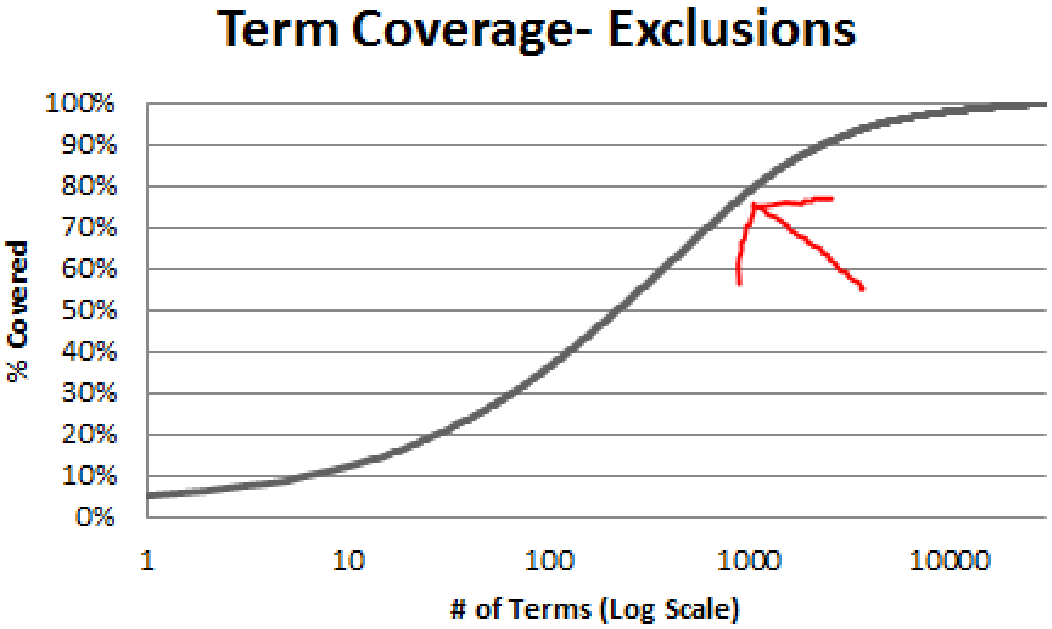
Start with either a hazard (Bolt) or an effect (Hand)



BOLT		HAND	
back	21%	rock	3%
hand	9%	piece	3%
arm	8%	roof	2%
head	6%	belt	2%
finger	5%	drill	2%
rib	4%	steel	2%
foot	4%	bolt	2%
laceration	3%	miner	2%
shoulder	3%	finger	2%
pain	3%	metal	2%

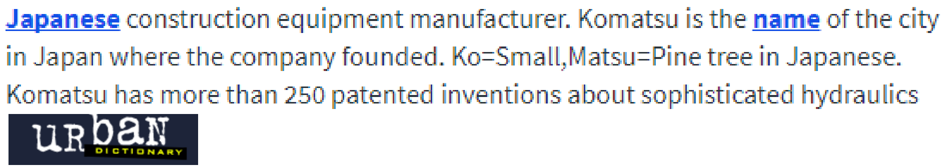
Annotation: Feasible & Specific

“How many words must I code to cover most info?”



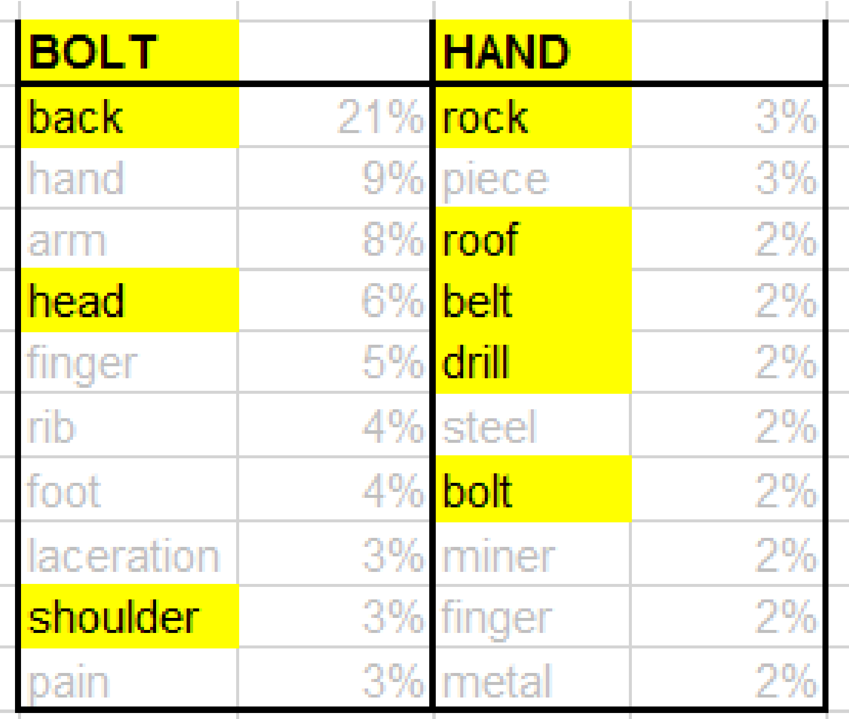
- 1000 terms gets you > 80% coverage

- Topic-Specific Annotation
  - Exposure, technical term, health effect, medical care, activity, …
- Even easier if stems/ morphemes
- “KOMATSU“


Japanese construction equipment manufacturer. Komatsu is the name of the city in Japan where the company founded. Ko=Small,Matsu=Pine tree in Japanese. Komatsu has more than 250 patented inventions about sophisticated hydraulics

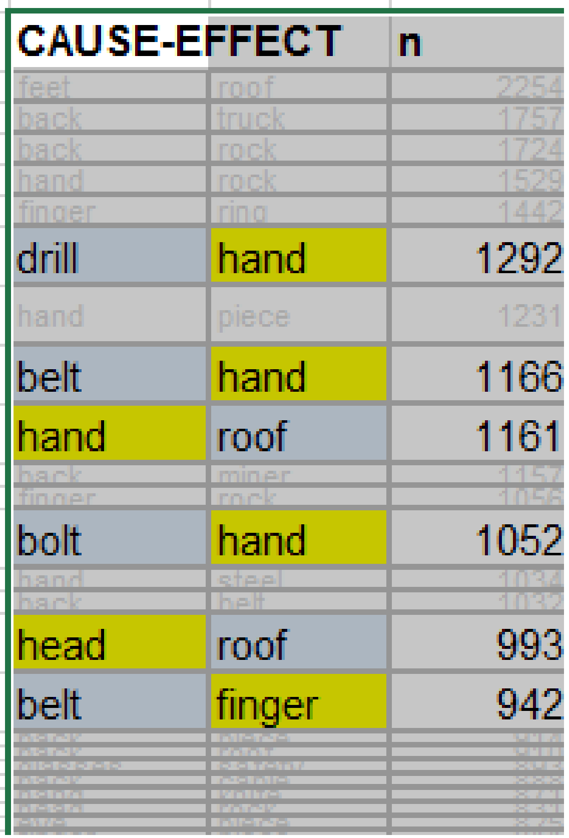
DISCOVERY vs HYPOTHESIS DRIVEN

HYPOTHESIS DRIVEN



BOLT		HAND	
back	21%	rock	3%
hand	9%	piece	3%
arm	8%	roof	2%
head	6%	belt	2%
finger	5%	drill	2%
rib	4%	steel	2%
foot	4%	bolt	2%
laceration	3%	miner	2%
shoulder	3%	finger	2%
pain	3%	metal	2%

AUTOMATED KNOWLEDGE DISCOVERY



CAUSE-EFFECT		n
feet	roof	2254
back	truck	1757
back	rock	1724
hand	rock	1529
finger	ring	1442
drill	hand	1292
hand	piece	1231
belt	hand	1166
hand	roof	1161
back	miner	1157
finger	rock	1056
bolt	hand	1052
hand	steel	1034
back	belt	1032
head	roof	993
belt	finger	942
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]

abay.com/en/dice-red-cubes-die-shapes-two-575658/

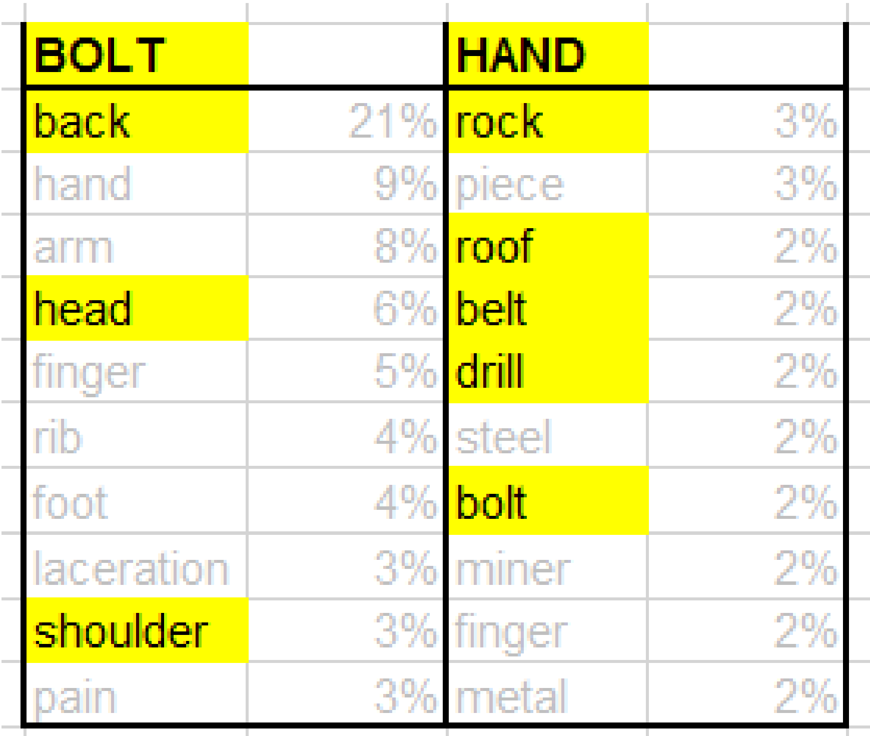
https://pixabay.com/en/ancient-blind-boys-brain

P Harber Univ of Arizona

HYPOTHESIS DRIVEN vs DISCOVERY

HYPOTHESIS DRIVEN

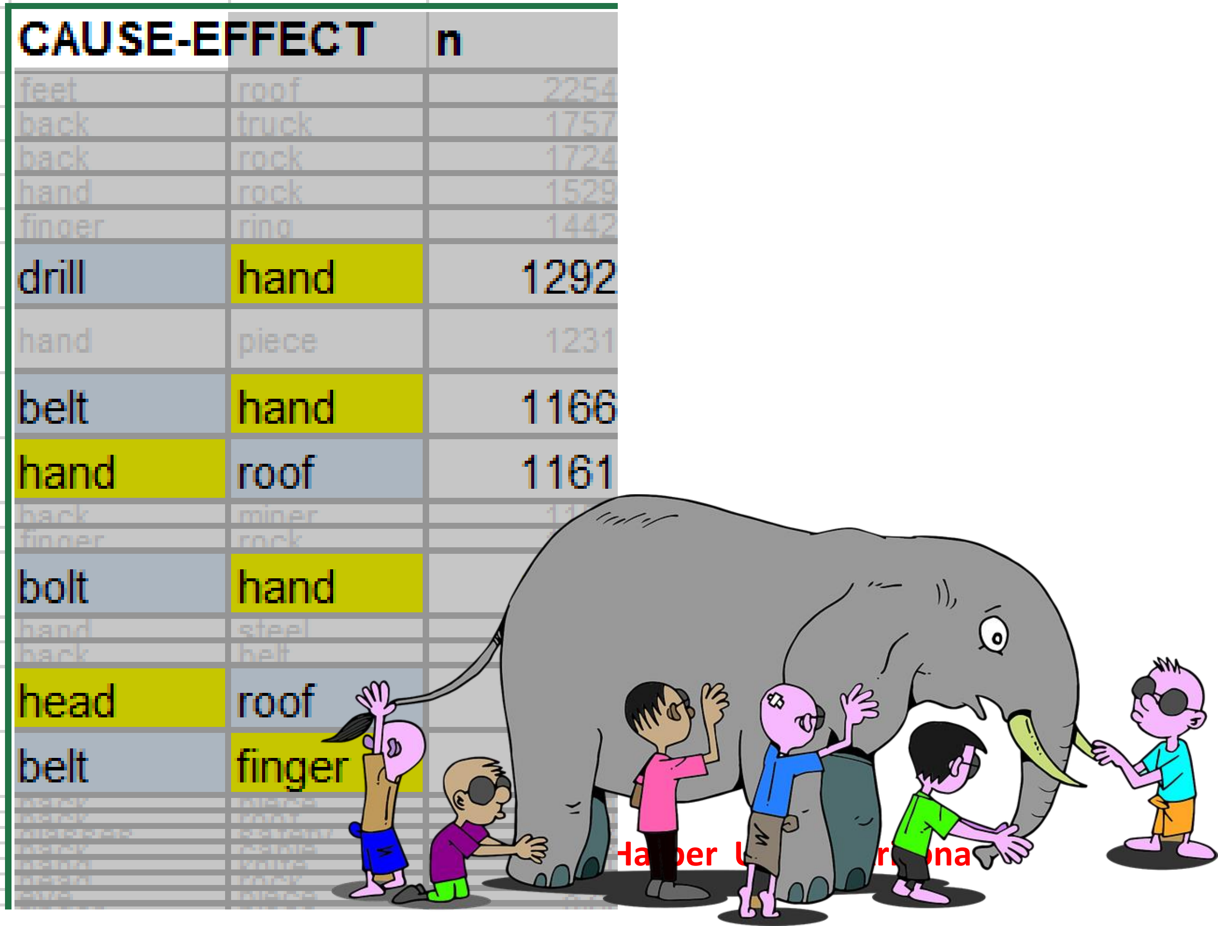
Serendipity



BOLT		HAND	
back	21%	rock	3%
hand	9%	piece	3%
arm	8%	roof	2%
head	6%	belt	2%
finger	5%	drill	2%
rib	4%	steel	2%
foot	4%	bolt	2%
laceration	3%	miner	2%
shoulder	3%	finger	2%
pain	3%	metal	2%

AUTOMATED KNOWLEDGE DISCOVERY

Less biased

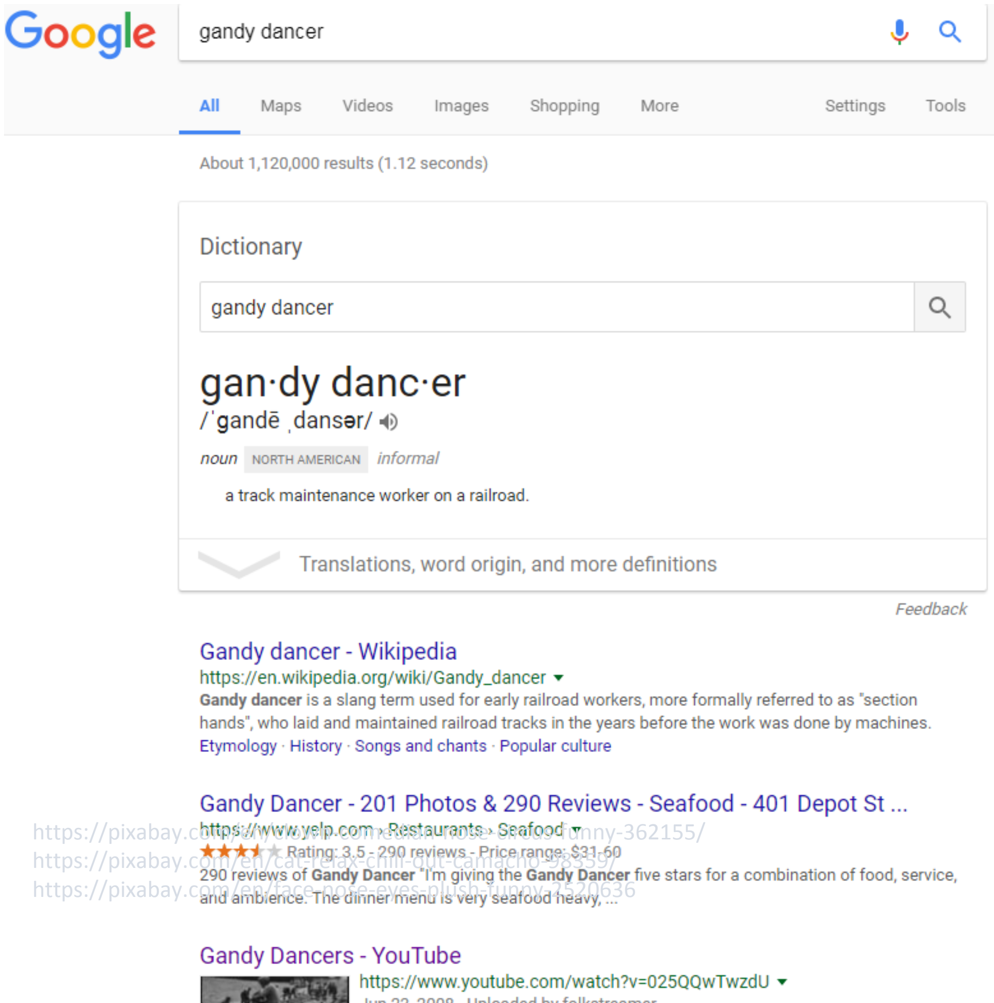


CAUSE-EFFECT		n
feet	roof	2254
back	truck	1757
back	rock	1724
hand	rock	1529
finger	ring	1442
drill	hand	1292
hand	piece	1231
belt	hand	1166
hand	roof	1161
back	miner	[illegible]
finger	rock	[illegible]
bolt	hand	[illegible]
hand	steel	[illegible]
back	belt	[illegible]
head	roof	[illegible]
belt	finger	[illegible]

“Real language of work, workers…”

Gandy dancers, Santa Claus, Occupational physicians

https://pixabay.com/en/santa-claus-st-nicholas-funny-x-mas-153309/



gandy dancer

About 1,120,000 results (1.12 seconds)

Dictionary

gandy dancer

gan·dy danc·er
/ˈgandē ˌdansər/

noun NORTH AMERICAN *informal*

a track maintenance worker on a railroad.

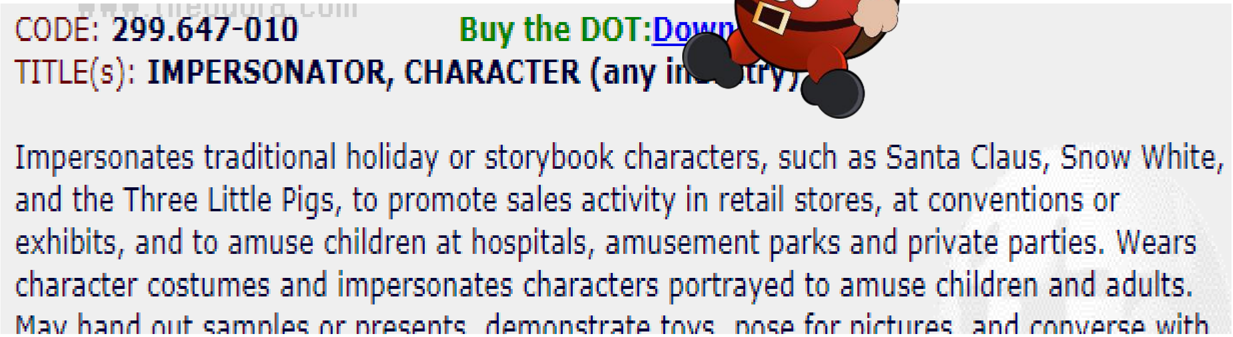
Translations, word origin, and more definitions

Gandy dancer - Wikipedia
https://en.wikipedia.org/wiki/Gandy_dancer ▾
Gandy dancer is a slang term used for early railroad workers, more formally referred to as "section hands", who laid and maintained railroad tracks in the years before the work was done by machines.
Etymology · History · Songs and chants · Popular culture

Gandy Dancer - 201 Photos & 290 Reviews - Seafood - 401 Depot St ...
https://www.yelp.com › Restaurants › Seafood ▾
★★★★☆ Rating: 3.5 - 290 reviews - Price range: $31-60
290 reviews of **Gandy Dancer** "I'm giving the **Gandy Dancer** five stars for a combination of food, service, and ambience. The dinner menu is very seafood heavy, ...

Gandy Dancers - YouTube
https://www.youtube.com/watch?v=025QQwTwzdU ▾

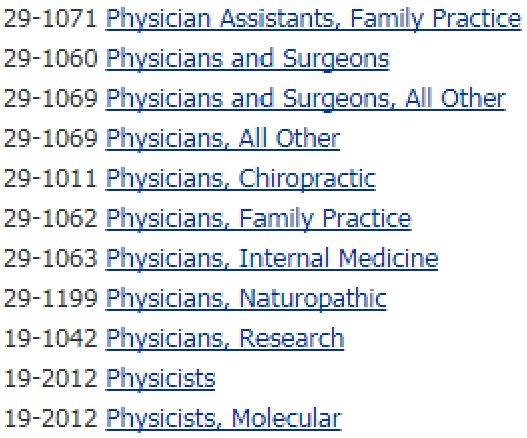
https://pixabay.com/en/clown-comedian-nose-circus-funny-362155/
https://pixabay.com/en/cat-relax-chill-out-camacho-98359/
https://pixabay.com/en/face-nose-eyes-plush-funny-2520636



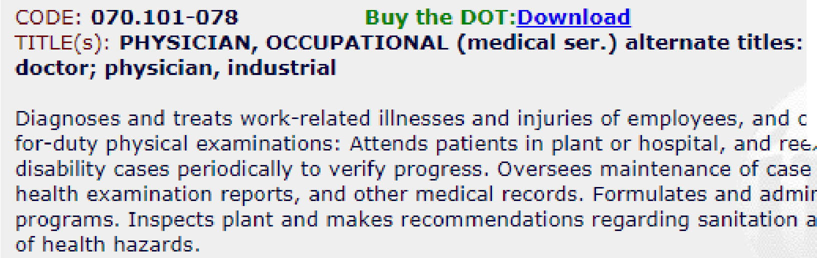
CODE: **299.647-010** Buy the DOT:Down

TITLE(s): **IMPERSONATOR, CHARACTER (any in try)**

Impersonates traditional holiday or storybook characters, such as Santa Claus, Snow White, and the Three Little Pigs, to promote sales activity in retail stores, at conventions or exhibits, and to amuse children at hospitals, amusement parks and private parties. Wears character costumes and impersonates characters portrayed to amuse children and adults. May hand out samples or presents, demonstrate toys, pose for pictures, and converse with



29-1071	Physician Assistants, Family Practice
29-1060	Physicians and Surgeons
29-1069	Physicians and Surgeons, All Other
29-1069	Physicians, All Other
29-1011	Physicians, Chiropractic
29-1062	Physicians, Family Practice
29-1063	Physicians, Internal Medicine
29-1199	Physicians, Naturopathic
19-1042	Physicians, Research
19-2012	Physicists
19-2012	Physicists, Molecular



CODE: **070.101-078** Buy the DOT:Download

TITLE(s): **PHYSICIAN, OCCUPATIONAL (medical ser.) alternate titles: doctor; physician, industrial**

Diagnoses and treats work-related illnesses and injuries of employees, and c for-duty physical examinations: Attends patients in plant or hospital, and reexamines disability cases periodically to verify progress. Oversees maintenance of case histories, health examination reports, and other medical records. Formulates and administers health programs. Inspects plant and makes recommendations regarding sanitation and elimination of health hazards.

Occupational health has its own language
(“Ontology”-Vocabulary, relations)

NIH U.S. National Library of Medicine

Unified Medical Language System® (UMLS®)

- Unified Medical Language System (UMLS) does not meet occupational health needs
- Medical terms→ frequently in UMLS concepts, generally accurate
- Workplace terms→ rarely present, usually wrong

WORDS- NOT NUMBER CODES
FULL MEANING AND FLEXIBLE
PEOPLE NOT PUNCH CARDS

HAIKU 5-7-5

THANK YOU!!

- ✓ Worker/patient talks to Automated Sir Val Ence (knighted epidemiologist)
- ✓ Who discovers new knowledge
- ✓ Provided by Marion (the automated librarian) to other workers

P Harber Univ of Arizona

https://pixabay.com/en/image-woodblock-printing-woodcut-1247354/