Supervised vs. Unsupervised Automated Categorization

Joseph Busch, Principal Analyst

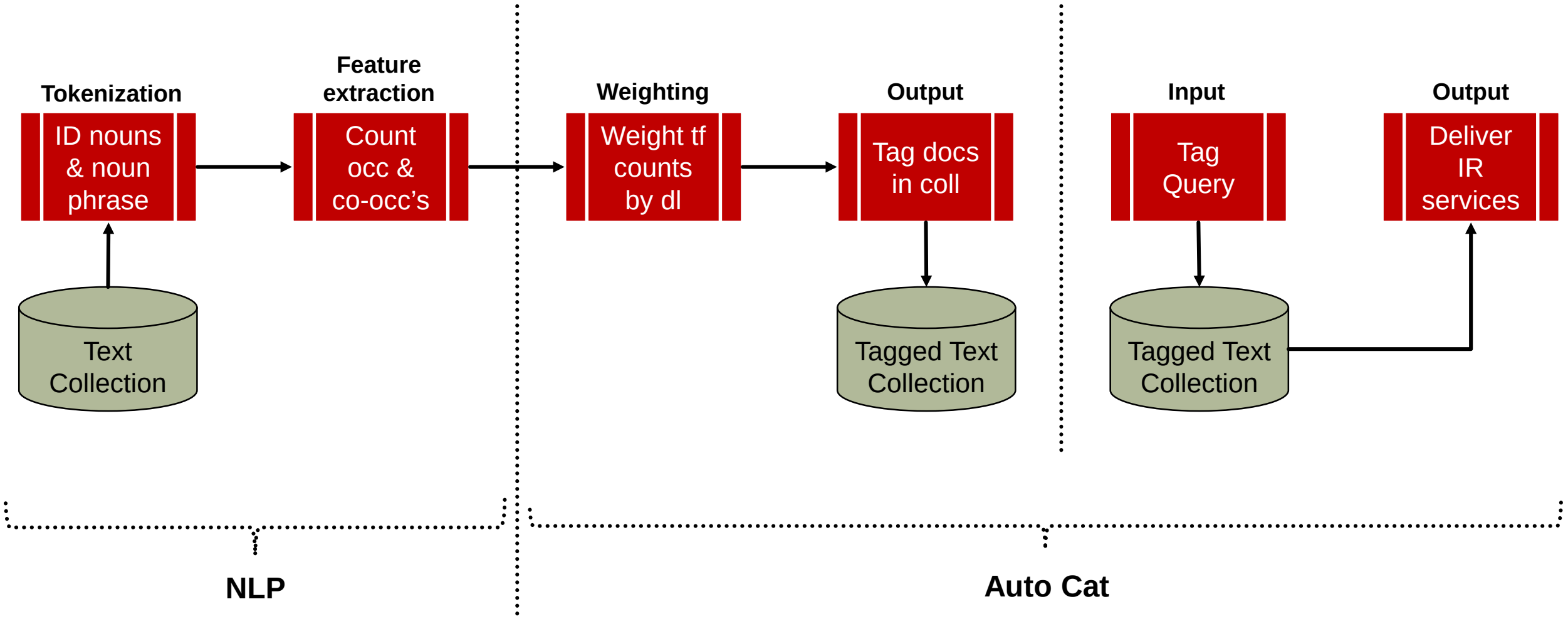
Agenda

- How to tag content using text analytics
- How to choose appropriate target collections
- How to evaluate text analytics accuracy
- RWJF case study

Categorization goals

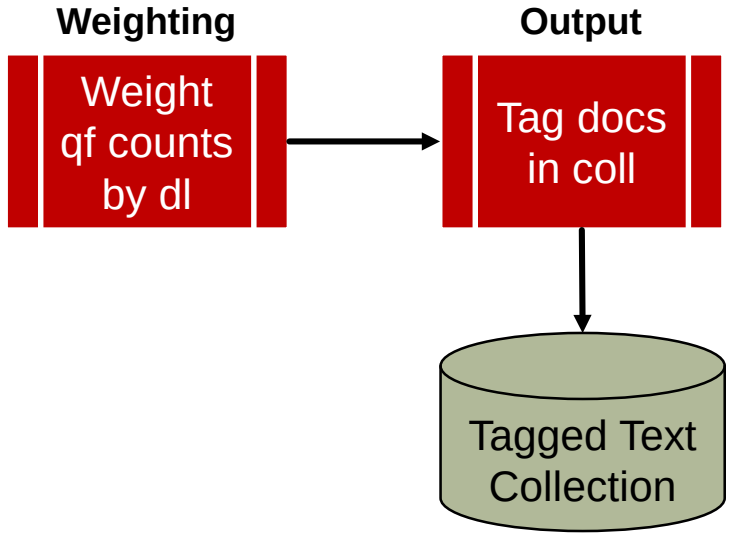
- Make sense
- Provide clear perception
- Is trustworthy and as unbiased as possible

Natural language processing enables unsupervised automated categorization



Individual terms can be replaced by pre-defined queries chosen and constructed by editors

- tf becomes qf
- Simple pre-defined query for “Affordable Care Act”
  - ("Affordable Care Act" OR "ACA" OR Obamacare)
- Complex pre-defined query for “Medicare Navigator”
  - ((navigator* OR assistor* OR assister*) NEAR (("health insurance" OR Medicare OR Medicaid) NEAR enroll*))



Agenda

- How to tag content using text analytics
- How to choose appropriate target collections
- How to evaluate text analytics accuracy
- RWJF case study

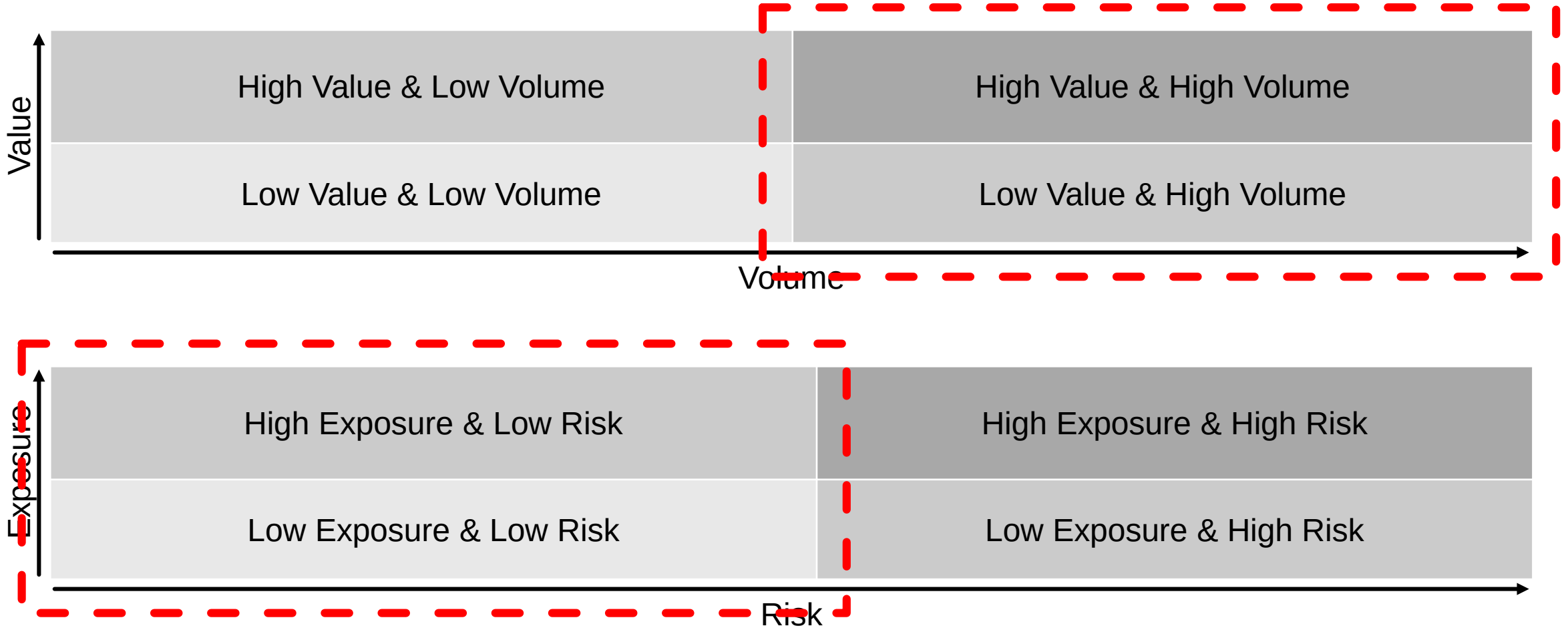
How reliable does a categorizer need to be?

It depends on the categorization target

Documents **VS.** **Projects**

Productivity vs risk models

Low vs high error tolerance



Agenda

- How to tag content using text analytics
- How to choose appropriate target collections
- How to evaluate text analytics accuracy
- RWJF case study

How to evaluate text analytics accuracy

- Build a prototype to test feasibility of auto-categorization.
  - Test selected categories individually
    - Is the recall and precision optimized on a category basis?
  - Test selected categories together
    - Can the categorizer appropriately distinguish between categories?
    - Are multiple categories identified when that is appropriate?
    - Are there relevance measures that can be meaningfully used to distinguish among multiple identified categories?
- Build production-quality categorizers.
  - Test all categories individually to optimize recall and precision
  - Test all categories together to ensure they can be appropriately distinguished

Agenda

- How to tag content using text analytics
- How to choose appropriate target collections
- How to evaluate text analytics accuracy
- RWJF case study

Former “A-Z” RWJF Taxonomy for their Project Information Management System (PIMS)

Access and barriers to care
Accountable care organizations
Accreditation
Addiction and substance abuse
Administration
Advanced practice nurses
Affordable Care Act (ACA)
Aging
Alcoholism
Allied health
Alternative and spiritual care
Ambulatory care
American Indian tribal government
Antibiotic resistance
Asthma
At-risk and vulnerable people
Barriers to care: cultural, gender and racial
Barriers to care: financial
Barriers to care: language and literacy
Barriers to care: logistics and transportation
Behavior change
Behavioral economics
Behavioral/mental health
Benchmarks and best practices
Built environment
Bundled payments
Business
Cancer
Care transitions
Careers
Case management
Child abuse and neglect
Child care centers
Child welfare
Childhood obesity
Children's health insurance program
Chronic disease management
Chronic illness
Clean air laws
Clinical care
Clinical research
Colleges and universities
Community benefit
Community development
Community outreach

Community violence
Community-based care
Competitions
Competitive foods
Complete streets
Conferences
Consumer engagement
Continuing education
Coordinated care
Cost of care
Crowd sourcing
Cultural competence
Data
Dental schools
Dentists
Developmental disabilities
Diabetes
Disparities
Disruptive innovations
Diverse communities
Diversity
Early childhood
Early intervention
Education and training
Education level
E-health
Eligibility and enrollment
Emergency care
Emergency preparedness and response
Employers
Employer-sponsored insurance
End-of-life care
Environmental health
Epidemiology
Episodes of care
Ethics/bioethics
Evaluation
Evidence-based
Families
Family caregiving
Federal government
Food access
Food safety
Foster care
Frontline workforce
Genetics
Health care delivery system

Health care markets
Health education
Health games
Health impact assessments
Health insurance
Health insurance exchanges
Health IT
Health plans
Health policy
Health promotion and disease prevention
Health records/electronic health records
Health reform
Healthy communities
Heart disease
Home health care
Home visiting
Homeless
Hospital readmissions
Hospital-acquired infections
Hospitals
Housing/public housing
Immigrants and refugees
Immunizations
Independent living or self-determination
Individual insurance
Industry regulation
Infectious diseases
Informatics
Information technology
Injury
Inmates and former inmates
Inpatient care
Institutional design
Insurance cost
Interpersonal violence
Interprofessional collaboration
Jails and prisons
Job satisfaction
Jobs
Joint use facilities
Journalists
Language services
Leadership development
Legal systems and issues
Lesbian/gay/bisexual/transgender
Local government
Long-term care

Low-birthweight infants
Managed care organizations
Marketing
Medicaid
Medical errors
Medical homes
Medical malpractice
Medical practices
Medical schools
Medical students and residents
Medical technology
Medical, dental and nursing workforce
Medically underserved areas
Medicare
Mentoring
Military/veterans
Mobile health/mhealth
Models
Networks
Nurse executives
Nurse midwives
Nurse practitioners
Nurses
Nursing homes
Nursing schools
Nutrition
Nutrition policy
Obesity policy
Open sourcing
Oral health
Out-of-school time
Overweight
Palliative care
Parks or playgrounds
Patient safety and outcomes
Patient satisfaction
Patient-centered care
Patients
Pay for performance
Payment reform
Pediatric care
Performance standards and measurement
Philanthropy
Physical activity
Physical activity policy
Physical education

Physician assistants
Physicians
Poor and economically disadvantaged
Practice guidelines
Pregnancy
Prenatal and neonatal care
Pre-schools
Prescription drugs
Prevention
Preventive care
Primary care
Primary care/generalist physicians
Prisoner reentry
Privacy and confidentiality
Provider incentives
Public health
Public health agencies
Public health law
Public health professionals
Public health schools
Public health system and finance
Public policy
Public-private partnerships
Quality of care
Recruitment and retention
Registered nurses
Religious congregations
Research
Research networks
Risky behavior
RWJF news
Safe routes to school
Safety nets
Scholars and fellows
School foods
School snacks
School/district policy
School-based health centers
Schools K-12
Scope of practice
Screening
Self-care
Shared decisionmaking
Shelters
Shortage of medical or nursing personnel
Small businesses
Social and emotional learning

Social determinants of health
Social isolation
Social marketing
Social sciences
Social support services
Specialist physicians
Specialist physicians
State government
Substance abuse treatment
Substance use
Substance use
Sugary beverages
Sugary beverages
Supportive housing
Tax policy
Tax policy
Telemedicine and telehealth
Tobacco
Tobacco
Tobacco cessation
Tobacco cessation
Tobacco control
Transdisciplinary
Transparency/public reporting
Transportation
Transportation
Transportation policy
Transportation policy
Transportation policy
Transportation policy
Underinsured
Underserved populations
Underserved populations
Underserved populations
Uninsured
Value-based purchasing
Walking and biking
Walking and biking
Walking and biking
Work environment
Workflow
Workforce issues
Workforce supply and demand
Workforce supply and demand
Youth development

New RWJF Taxonomy model for their Project Information Management System (PIMS)

Topics	Activities	Geo Locations	Health Factors	Settings	Populations	Demographics
Built Environment & Health	Advocacy	+ Location Type	Access to Care	American Indian Tribal Government	Activist/Advocate	+ Age
Child & Family Well-being	Capacity & Organizational Development	+ Regions	Air & Water Quality	Business (General)	Administrative/Managerial	+ Gender
Childhood Obesity	Coalition & Network Building	+ States & Territories	Alcohol & Drug Use	Child Care	Caregiver	+ Education
Disease Prevention & Health Promotion	Communications & Media		Community Safety	Community (General)	Faith-based	+ Income Level
Early Childhood	Evaluation		Diet & Exercise	Federal Government	Health Professional	+ Race & Ethnicity
Health Care Coverage & Access	Idea Sourcing		Education	Health Care/Clinical	Immigrant or Migrant	
Health Care Quality & Value	Leadership, Education & Training		Employment & Income	Home/Household	Inmates or Former Inmates	
Health Disparities	Pilot Project		Family & Social Support	Local Government	LGBTQ Person	
Health Leadership Development	Research & Analysis		Genetics	Nonprofit/Philanthropy (General)	Non-Health Professional	
Nurses and Nursing			Housing, Transit & Infrastructure	School (General)	Parent	
Public & Community Health			Quality of Care	State Government	Person with Disability	
Social Determinants of Health			Tobacco Use		Person with Disease or Illness	
					Policymaker	
					Student	

Problem: How to convert legacy collection from as-is to the new taxonomy model?

- Option 1: Mapping from the as-is to the new.
- Option 2: Automated categorization
- Option 3: Combination of option 1 and 2

Option 1: Mapping

CODING_DEF_ID	PIMS Topics	New Facet	New Catgeory
350	Access and barriers to care	Topics	Health Care Coverage & Access
1707	Accountable care organizations	Topics	Health Care Quality & Value
386	Accreditation	Topics	Public & Community Health
508	Addiction and substance abuse	Health Fac	Alcohol & Drug Use
508	Addiction and substance abuse	Topics	Disease Prevention & Health Promotion
670	Administration	Population	Administrative/Managerial
670	Administration	Population	Health Professional
670	Administration	Topics	Health Leadership Development
470	Advanced practice nurses	Population	Health Professional
470	Advanced practice nurses	Topics	Nurses & Nursing
1709	Affordable Care Act (ACA)	Topics	Health Care Coverage & Access
1667	Aging	Demograp	Older Adults
509	Alcoholism	Health Fac	Alcohol & Drug Use
509	Alcoholism	Population	Person with Disease or Illness
509	Alcoholism	Topics	Disease Prevention & Health Promotion
478	Allied health	Population	Health Professional
478	Allied health	Topics	Health Leadership Development
406	Alternative and spiritual care	Settings	Health Care/Clinical
395	Ambulatory care	Settings	Health Care/Clinical
395	Ambulatory care	Topics	Health Care Quality & Value
306	American Indian tribal government	Settings	American Indian Tribal Government

Option 2-Step 1: Build a prototype to test feasibility of auto-categorization

Can we build some categorizers that accurately detect Topics?

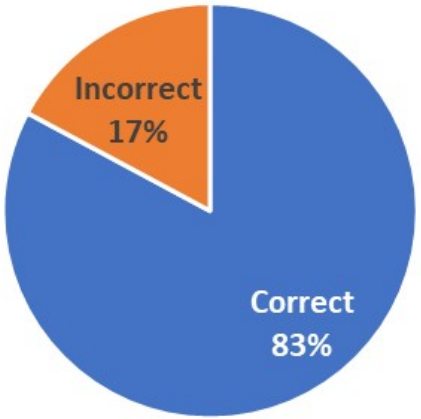
- Collected 200 categorized examples for each Topic to be tested.
- Divided test collection in half – 100 to use to build the categorizer, and 100 to use in a blind test.
- Reviewed 100 relevant items and identify the words and phrases that provide a context for the topic.
- Noted any named entities (people, organizations, events, laws, etc.) that are closely associated with the topic.
- Consolidated the terms – identify duplicates, synonyms, as well as any concepts that you want to combine even if they are not synonyms.
- Wrote a query for each term.
- Combined the terms into a single nested query.

Pre-defined query for “Health Care Coverage & Access” topic

(("health insurance coverage" OR "health coverage") OR ("healthcare reform" OR "health care reform") OR ("Better Care Reconciliation Act" OR BCRA) OR ("American Health Care Act" OR AHCA) OR ("Affordable Care Act" OR "ACA" OR Obamacare) OR ((Medicare OR Medicaid) NEAR/5 (spend* OR cover* OR expan*)) OR ("health insurance exchange" OR "HIE") OR ("health insurance" NEAR/5 marketplace*) OR ("federal* facilitated marketplace*" NEAR/10 "health insurance") OR ("federal* run marketplace*" NEAR/10 "health insurance") OR ((state NEAR/5 marketplace*) NEAR/10 "health insurance") OR ("small business marketplace*" NEAR/10 "health insurance") OR ("small-business marketplace*" NEAR/10 "health insurance") OR (("small business" NEAR/5 exchange*) NEAR/10 "health insurance") OR (("high-risk" OR "high risk") NEAR/10 "health insurance") OR (uninsured NEAR/5 (veteran* OR child* OR adult* OR people OR kid* OR citizen*)) OR (("pre-existing condition*" OR "preexisting condition*") NEAR/10 "health insurance") OR "health insurance rate*" OR ((cost* OR rate* OR payment*) NEAR/10 "health insurance") OR ("health insurance" NEAR/10 "tax credit*") OR ((healthcare OR "health care") NEAR/5 spending) OR ((healthcare OR "health care") NEAR/5 utilization) OR (("high-deductible" OR "high deductible") NEAR/10 "health insurance") OR (("mental health" OR "substance abuse") NEAR/10 "health insurance") OR ("provider network*" NEAR/10 "health insurance") OR (("in-network" OR "out-of-network") NEAR/10 "health insurance") OR ((PPO* OR HMO*) NEAR/5 (marketplace* OR plan* OR provider*)) OR ("health insurance" NEAR/10 (enroll* OR "re-enroll*" OR renew* OR "open-enrollment" OR "open enrollment")) OR ((navigator* OR assistor* OR assister*) NEAR/10 (("health insurance" OR Medicare OR Medicaid) NEAR/5 enroll*)) OR ("CHIP" OR "Children’s Health Insurance Program") OR ("individual mandate" NEAR/10 "health insurance") OR "employer-sponsored insurance" OR ((employer OR employee) NEAR/10 "health insurance"))

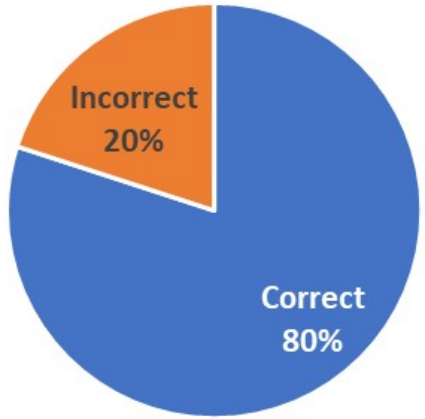
Trial results for each pre-defined category

Childhood Obesity



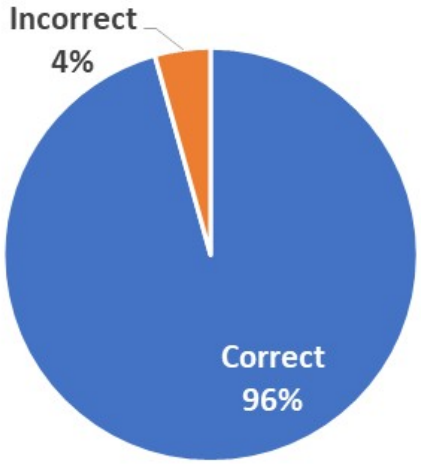
n=64

Disease Prevention & Health Promotion



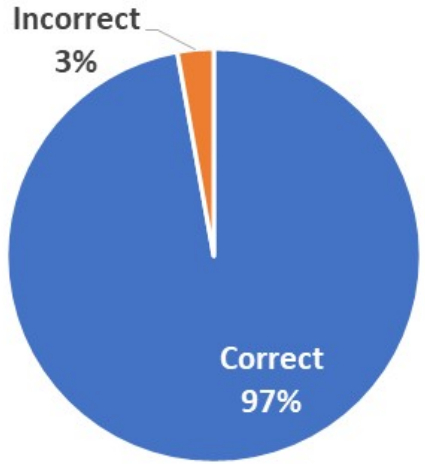
n=65

Health Care Quality & Value



n=69

Health Care Coverage & Access



n=71

Option 2-Step 2: Build production-quality categorizers for all 12 topics

- For each of the 12 RWJF Topics:
  - Gather a test collection of grant summaries
  - Build and test a rule – a Boolean query with proximity operators to classify into an RWJF Topic
    - Start with the RWJF Topic string.
    - Add synonyms.
    - Add legacy PIMS Topics that were mapped to the 12 RWJF Topics and synonyms.
    - Review the test collection of grant titles & summaries tagged for that RWJF Topic. (A sample size of 100 should suffice). Identify distinctive words and phrases.
    - Group Boolean search strings together by each discrete topic as much as possible.
    - Review and revise.
  - Test a categorizer for an RWJF Topic. (12 categorizers; 12 tests)
    - Test **recall** – Are **all** the relevant grants being identified by the query, 70-80% of the time? If not, revise query and re-test.
    - Test **precision** – Are **only** the relevant grants being identified by the query, 70-80% of the time? If not, revise query and re-test.
- Consolidate 12 RWJF Topics into a master categorizer.
  - Test master categorizer against NEW test collection.

Health Care Coverage > Health Insurance Characteristics sub-queries

(Note: we tested at the whole Topic level, not the sub-queries)

- Health Care Access
  - Access to Health Care
  - Barriers to Health Care
  - Various Health Outcomes Based on Access
- Health Care Coverage
  - Health Care Reform – Legislation
  - Health Insurance Characteristics
    - Health Insurance Cost
    - High Deductible Insurance
    - High Risk
    - Mental Health-Substance Abuse
    - PPO-HMO
    - Pre-existing Condition
    - Provider Network-In-Out
  - Health Insurance Enrollment
  - Insurance Marketplace

- ("health care" OR healthcare OR "health insurance") AND (premium* OR payment* OR rate* OR cost*)
- (high-deductible OR "high deductible") NEAR/5 "health insurance“
- (high-risk OR "high risk") NEAR/5 "health insurance“
- ("mental health" OR "substance abuse") NEAR/5 "health insurance“
- (PPO* OR HMO*) NEAR/5 (marketplace* OR plan* OR provider*)
- ("preexisting condition*" OR "pre-existing condition*") NEAR/5 "health insurance“
- (("provider network*" NEAR/5 "health insurance") OR (("in-network" OR "out-of-network") NEAR/5 "health insurance"))

Optimize recall and precision on a category-by-category basis

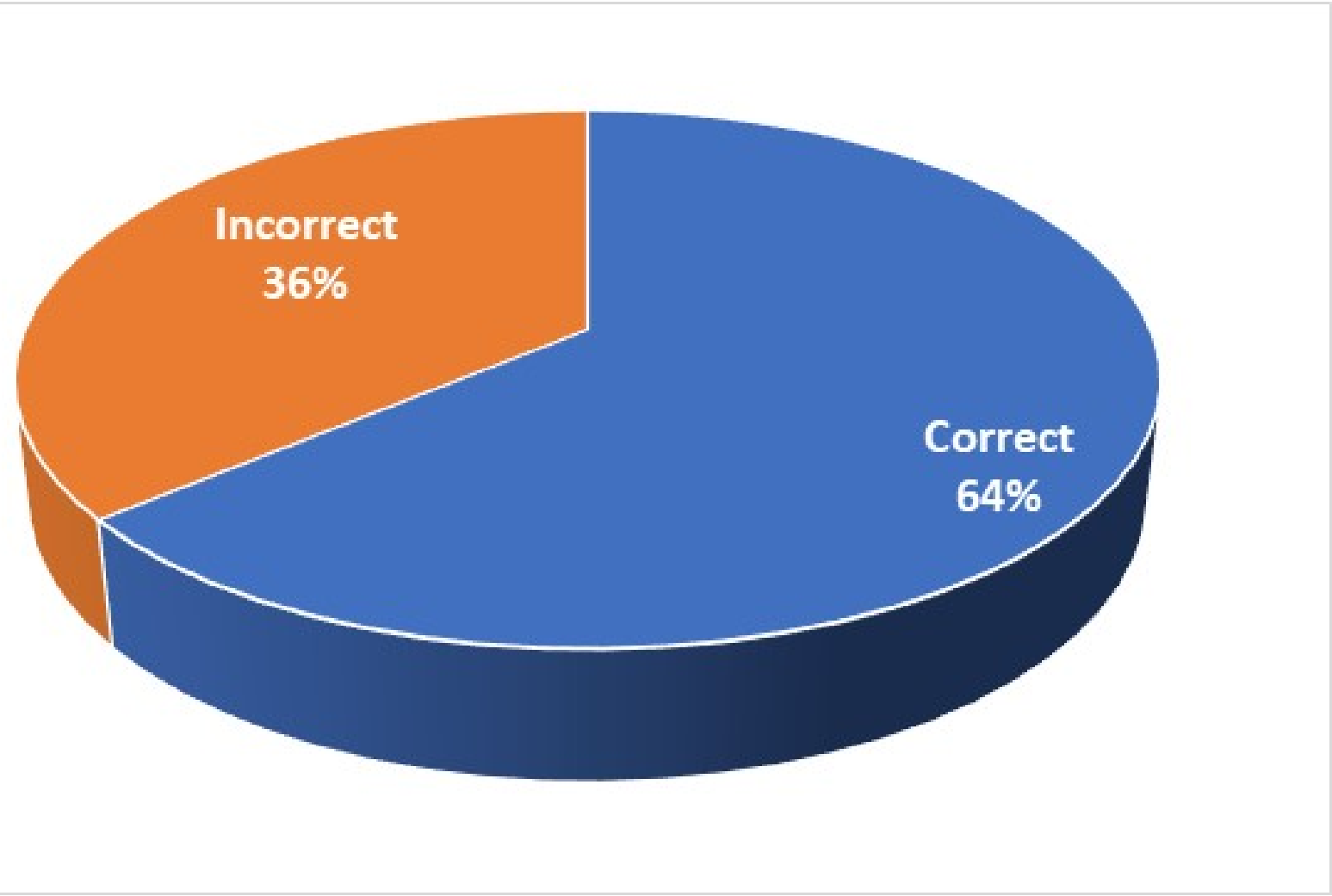
Topic	Recall	Precision
Built Environment & Health	78%	70%
Child & Family Well-being	77%	74%
Childhood Obesity	86%	97%
Disease Prevention & Health Promotion	57%	64%
Early Childhood Development	76%	82%
Health Care Coverage & Access	72%	79%
Health Care Quality & Value	69%	75%
Health Disparities	73%	81%
Health Leadership Development	71%	80%
Nurses and Nursing	94%	94%
Public & Community Health	51%	69%
Social Determinants of Health	71%	69%

Option 2-Step 3: Consolidate category-by-category queries into a master categorizer

- Combine all of the category queries into one master categorizer
  - Make changes to queries as needed to meet the parameters of a single categorizer
  - Add named entities where appropriate
- Run analysis and calculate precision and recall to discover if any unintended problems were introduced due to consolidating into one categorizer
  - Use same test set of assets (Step 2 test set) for Step 3

Optimize recall and precision for consolidated categorizer

Correct/Incorrect	DocNumFID	Bully Environment	Child and Family Well Being	Childhood Obesity	Disease Prevention and Health Promotion	Early Childhood Development	Health Care Coverage and Access	Health Care Quality and Value	Health Disparities	Health Leadership Development	Public and Community Health	Social Determinants of Health	(blank)	Grand Total
1	74742	10	2	0	0	0	0	0	0	0	0	0	0	12
0	74955	3	0	0	0	0	3	2	2	0	2	0	8	20
1	74968	0	2	0	2	2	0	0	0	0	0	0	0	6
1	75042	0	0	0	5	0	4	0	41	0	0	2	0	52
1	75059	1	3	0	4	0	5	2	9	0	3	4	6	37
1	75083	0	5	0	2	0	23	14	0	0	5	0	0	49
0	75089	0	2	0	8	0	3	2	0	0	1	0	0	22
0	75094	0	0	0	0	0	2	0	0	0	0	0	0	2
0	75095	4	0	0	3	0	10	11	13	0	11	4	0	56
1	75097	0	0	0	2	0	0	0	0	0	2	0	0	4
0	75103	0	0	0	0	0	4	5	0	0	3	0	3	21
0	75109	0	0	0	0	2	0	1	0	3	0	0	0	6
1	75116	0	5	0	2	5	0	0	0	0	0	0	0	12
0	75131	1	4	0	2	2	13	3	5	0	3	1	0	34
1	75148	0	0	0	4	0	3	0	0	0	0	0	0	7
1	75160	1	0	0	5	1	0	0	3	0	0	0	0	10
0	75163	10	2	0	0	0	0	0	6	2	3	0	0	23
1	75173	0	3	0	0	0	0	0	0	0	1	0	0	4
1	75184	0	0	0	0	0	5	4	0	0	0	0	0	9
1	75194	0	0	0	3	1	0	0	0	0	0	0	0	4
1	75196	15	0	0	2	0	0	0	6	0	4	0	6	33
1	75198	2	4	0	13	0	15	25	10	2	10	0	0	81
0	75199	1	0	0	2	0	4	0	3	3	0	5	0	18
1	75201	1	0	0	10	0	7	0	3	3	0	4	0	28
1	75202	2	0	0	2	0	6	0	3	3	0	5	0	21
0	75209	1	9	0	5	12	3	6	9	3	6	5	0	59
0	75213	0	2	0	0	5	0	0	0	0	0	0	0	7
0	75214	0	0	0	0	0	0	1	0	0	0	0	0	1
1	75215	9	8	5	0	2	0	0	0	0	0	0	4	28
1	75223	0	0	0	0	0	0	3	0	0	0	0	0	3
0	75319	0	0	0	0	0	0	0	2	0	0	0	0	2
1	75332	0	2	0	0	0	0	0	0	0	0	0	4	6
0	75333	0	5	0	0	0	0	0	0	0	0	0	0	5
1	75362	8	0	0	0	0	0	0	3	0	4	0	0	15
1	75373	0	0	0	3	0	4	9	0	0	0	0	0	16
1	75374	0	0	0	0	0	7	0	0	0	0	0	0	7
0	75386	2	3	0	0	0	0	3	3	0	0	2	0	13
1	75403	0	0	0	0	0	21	14	0	0	0	0	2	37
1	75417	0	11	0	2	3	0	0	0	0	0	5	0	21
1	75432	7	6	4	0	2	0	0	0	0	0	0	4	23
0	75443	0	6	0	5	8	0	0	8	0	0	0	5	32
1	75475	0	0	0	0	5	0	0	0	0	0	0	0	5
1	75493	0	2	0	4	2	0	1	3	0	3	4	0	19
1	75520	2	0	0	2	0	0	0	2	0	0	0	3	15
1	75523	0	0	0	0	0	0	0	3	0	4	0	0	7
1	75531	11	0	0	3	0	3	0	4	0	0	0	5	32
0	75541	2	0	0	0	0	2	0	3	0	0	5	0	12



… is this performance reliable enough for unsupervised categorization of high risk content?

Option 3: Combination of mapping and automated methods

- Used mapping to provide initial conversion of legacy content from A-Z to new facetted taxonomy.
- Identified 1) categories with no mapping to any grants, and 2) grants with too many Topics; then processed these grants with a combination of human review, and automated categorization to improve their indexing.
- Worked to build a critical mass of new grants indexed by editors using the new taxonomy.
- **Future goal:** Automation-assisted indexing.

Summary

- Natural language processing enables unsupervised automated categorization.
- Individual terms can be replaced by pre-defined queries chosen and constructed by editors.
- Good results on individual pre-defined queries/topics may not be as good when distinguishing which is most relevant among a group of pre-defined queries/topics.
- Automated categorization results may be good enough for some collections of content (high volume or low risk), but not for others (low value or high risk).

Questions

Joseph Busch
jbusch@taxonomystrategies.com or joseph@semanticstaffing.com
@joebusch
mobile +1 415-377-7912

Supervised vs. Unsupervised Automated Categorization: Summary

- AI promises to categorize all types of content with reliable results, but the reality is much more complex. Most applications won't work with a meat grinder approach where you pour a huge amount of content in one end and a perfectly organized collection comes out the other end. Effective automated categorization depends on defining a process workflow and assembling a stack of methods to process different types of content in different ways. Designing and validating a content processing workflow requires human judgements. So good quality categorization applications often rely on how to make the best use of people. This presentation provides a reality check on unsupervised automated categorization, and discusses a case study where the performance was suitable for editorial review and approval, but not for unsupervised processing of a large collection. The project is the Robert Wood John Foundation grants information system using Lexalytics configurations developed by Taxonomy Strategies.