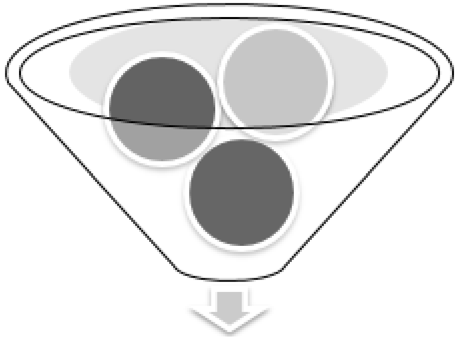
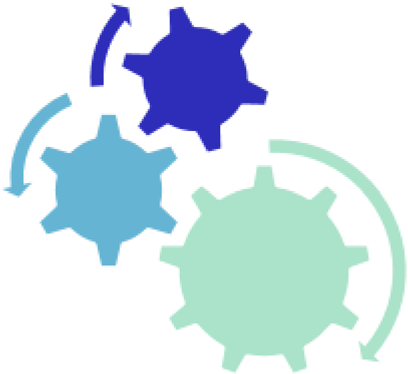




AI vs. Automation

Workshop

Agenda

Length	Format	Description	Outcomes
10 min 50 min	Activity Demos	▪ Demo some natural language processing, entity extraction, and complex Boolean query tagging tools (Lexalytics, Data Harmony, IBM Watson)	▪ Understand the types of features that differentiate automated tagging tools, and the criteria for evaluating them.
20 min 25 min 15 min	Lecture Activity Show/ Tell	▪ Participate in a query building exercise.	▪ Obtain a practical understanding of how to build an automated classifier.

Who’s in the room?

1) Pick the **one** profession you most strongly identify with.

1) Profession	No
Librarian/Archivists	9
Taxonomist	2
Info/Data Scientist	2
Researcher	1
Student	0
Entrepreneur	1
SW Engineer	1
Information Architect	1
Other Interested Party	1

Outline

- Tool demos
- What are Boolean queries, how to build them, and why

Tool Demos

Tool	Demo URL
Aylien	https://developer.aylien.com/text-api-demo?text=&language=en&tab=classify-taxonomy
Data Harmony *	http://demo.newsindexer.com/
IBM Watson	https://natural-language-understanding-demo.ng.bluemix.net/
Intellexer	http://demo.intellexer.com/
Lexalytics *	https://www.lexalytics.com/demo
Meaning Cloud	https://www.meaningcloud.com/demo
PoolParty PowerTagging	https://drupal.poolparty.biz/powertagging
Text Razor	https://www.textrazor.com/demo

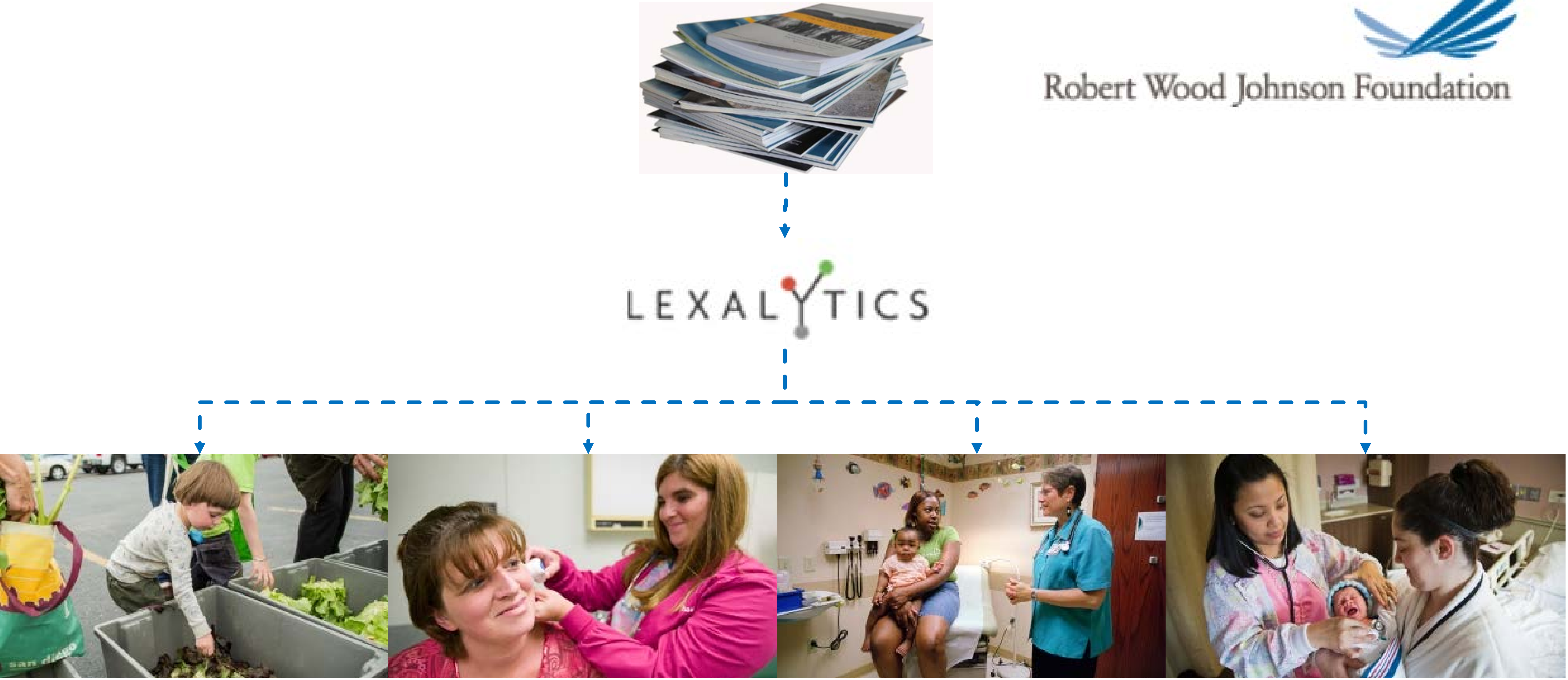
* In-depth demos using web client applications for these tools.

Outline

- Tool demos
- What are Boolean queries, how to build them, and why

Case study

Robert Wood Johnson Foundation



Childhood Obesity

Disease Prevention and Health Promotion

Health Care Quality

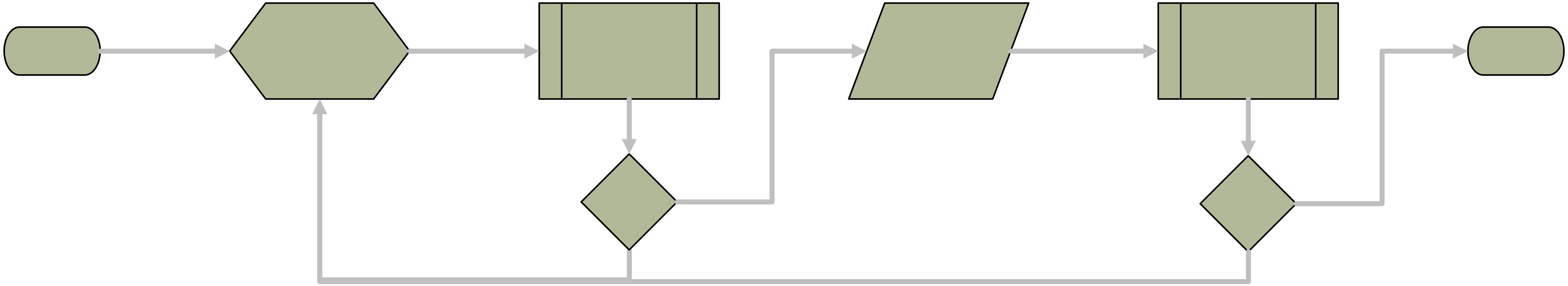
Health Coverage

Testing process: Text collections

User interfaces	API	Excel
Test collection	90 Repository long form assets	400 WCMS short form assets
Asset Types	Article, Book, Chart, Evaluation, Issue Brief, News Release, Newsletter, Proceedings, Promotion, Report, Speech, Survey, Testimony, Toolkit	Brief, Journal Article
Content	Full text	Title & summary only
Format	Clear text	Clear text, CSV
Topics	Childhood Obesity, Disease Prevention and Health Promotion, Health Care Quality, Health Coverage	Childhood Obesity, Disease Prevention and Health Promotion, Health Care Quality, Health Coverage

Test process: Categorization (to a pre-defined set of categories)

- Build and test a rule
  - A Boolean query with proximity operators to classify into a Topic (called a “configuration” in Lexalytics Semantria).
- Modify and test a rule.
- Obtain relevant classification
  - Identify the correct Topic, 80% or more of the time.
  - If an incorrect Topic is returned, why was it returned? Is an incorrect Topic potentially relevant?

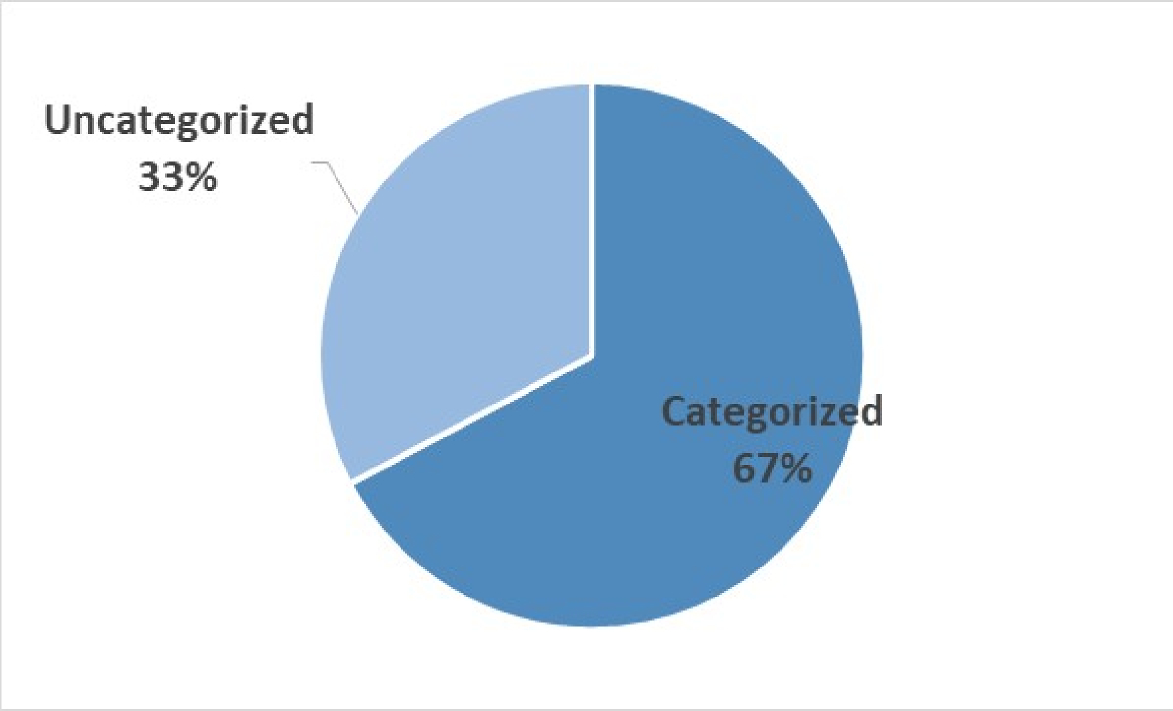


Sample pre-defined query: Health coverage

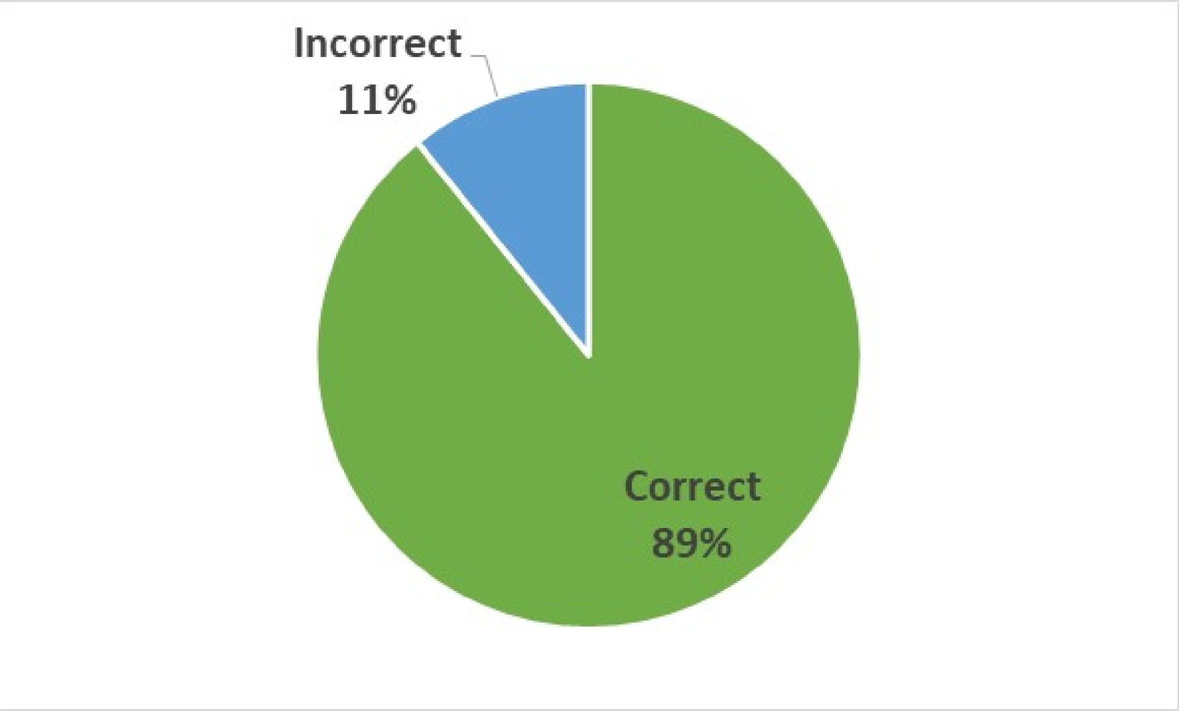
(("health insurance coverage" OR "health coverage") OR ("healthcare reform" OR "health care reform") OR ("Better Care Reconciliation Act" OR BCRA) OR ("American Health Care Act" OR AHCA) OR ("Affordable Care Act" OR "ACA" OR Obamacare) OR ((Medicare OR Medicaid) NEAR/5 (spend* OR cover* OR expan*)) OR ("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"))

Overall trial results

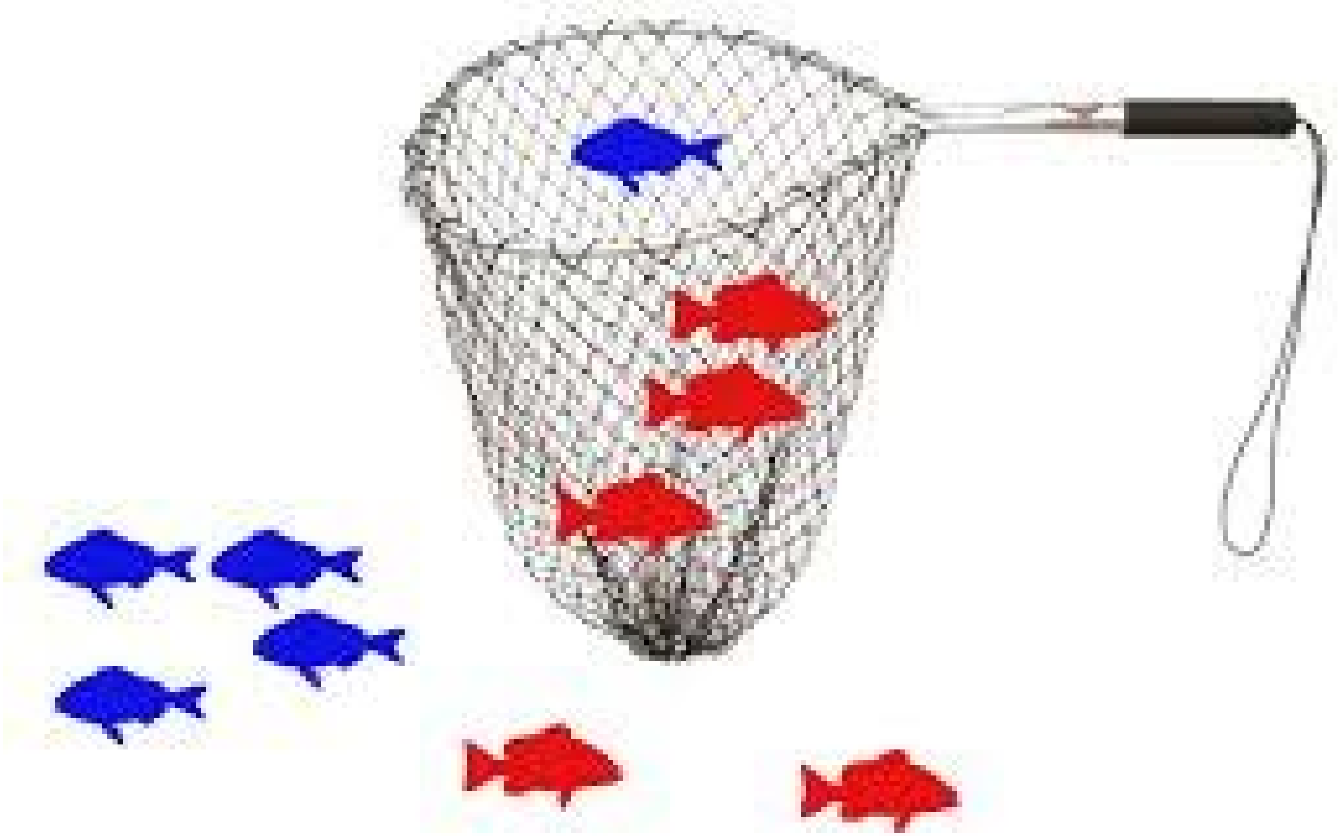
Categorized to a topic (Recall)



Categorized to the correct topic (Precision)

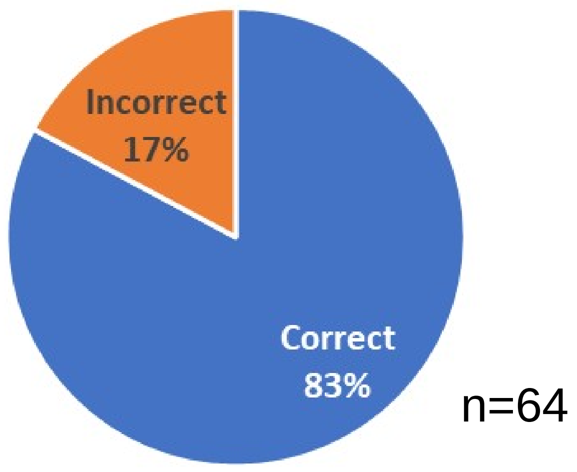


Precision and recall tradeoff

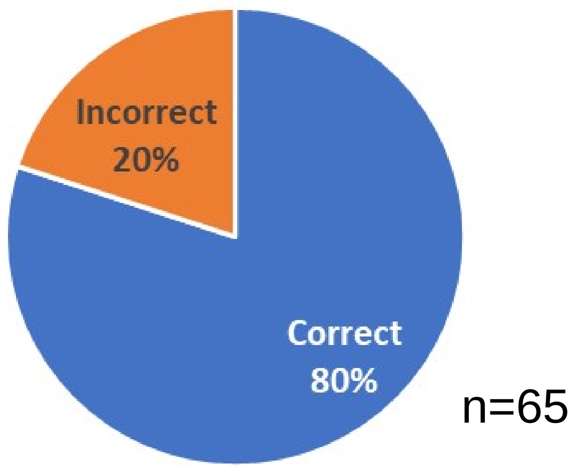


Trial results for each topic

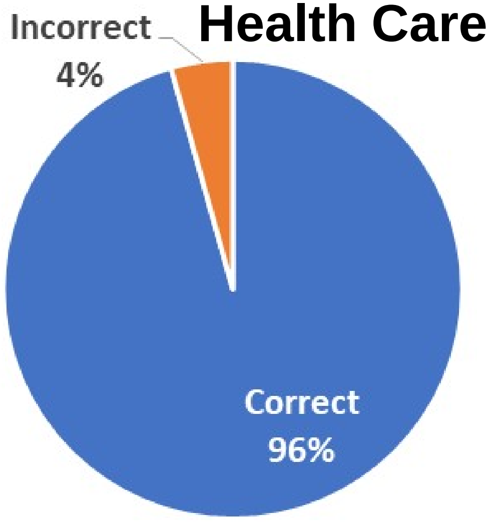
Childhood Obesity



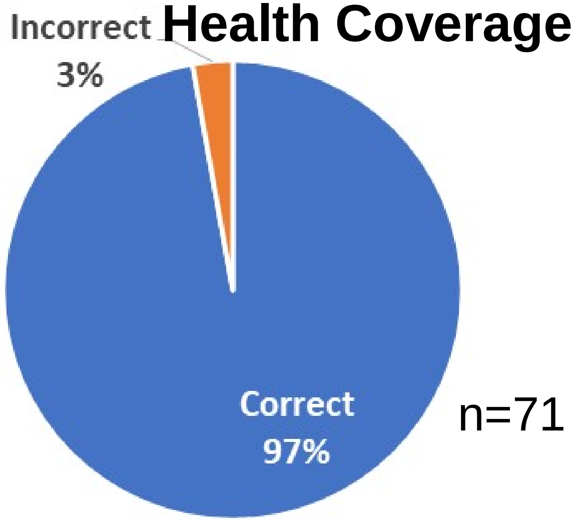
Disease Prevention and Health Promotion



Health Care Quality



Health Coverage



Boolean search: How hard is it do?

Machine learning

- Machine learning is opaque.
  - Benefit: No preparation. Content just gets processed.
  - Problem: Categories are generic, may be irrelevant, can be biased, and are difficult to change or tune.

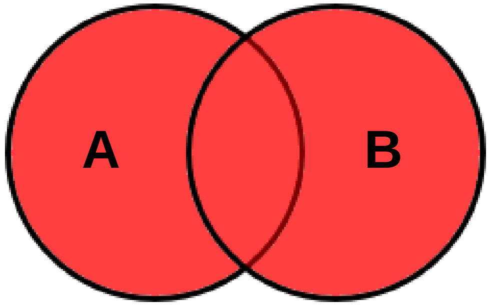
Boolean queries

- Pre-defined categories (aka a taxonomy) plus Boolean queries to scope the context for categories are transparent.
  - Benefit: Relevant categories.
  - Problem: Requires work to set up, and specialized skills.

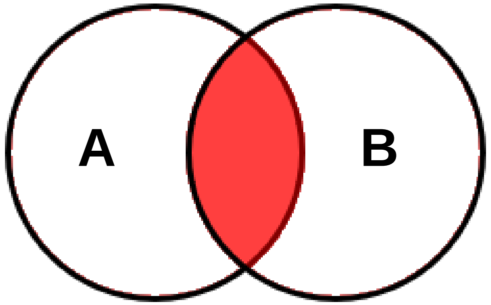
Boolean queries

- Basic operators
  - AND (conjunctive)
  - OR (disjunctive)
  - NOT (negation)

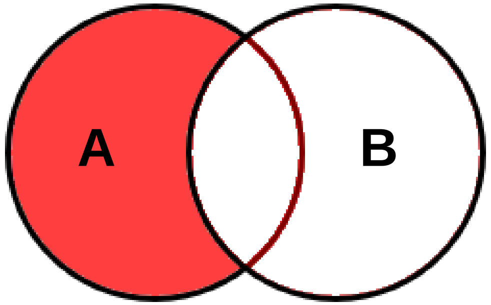
- Venn diagrams



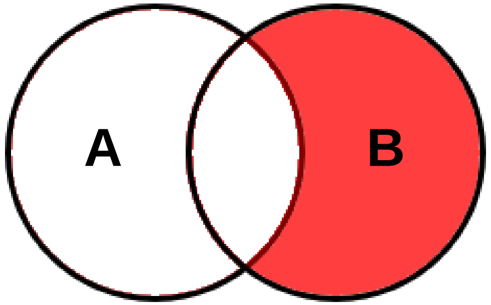
A OR B



A AND B



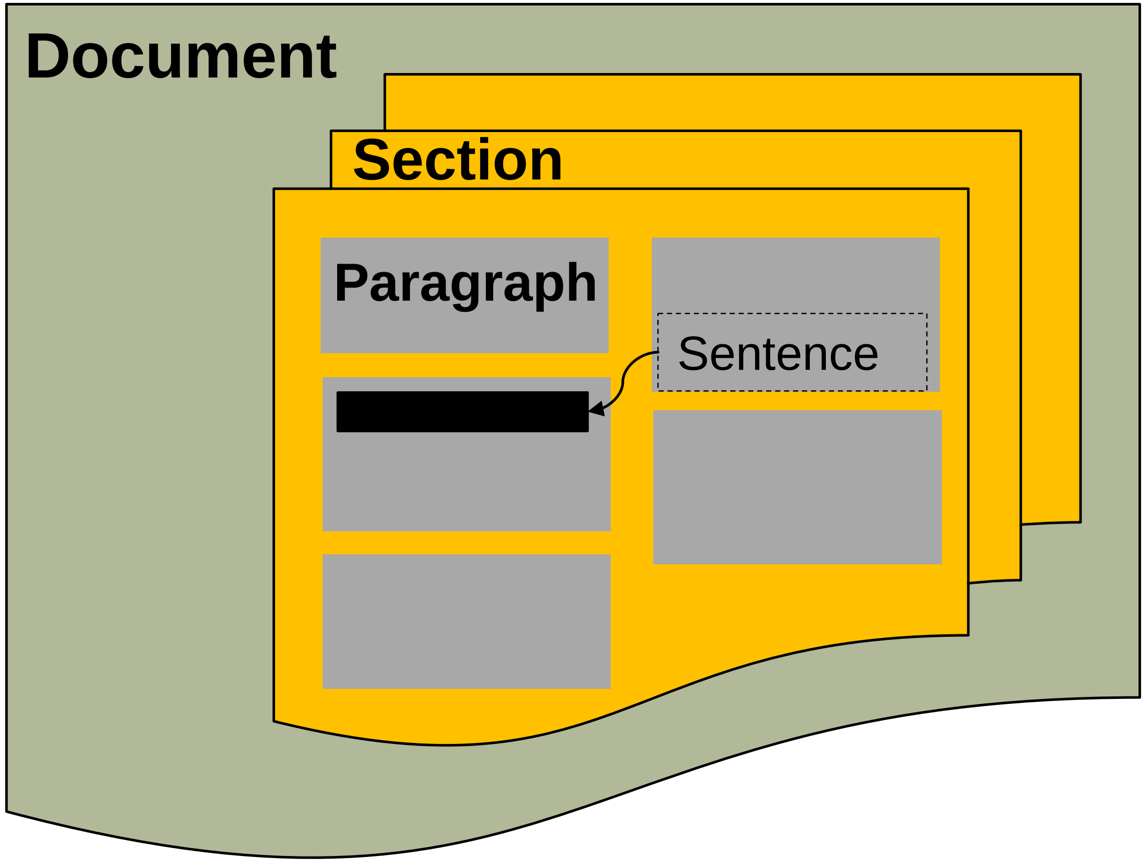
A NOT B



B NOT A

Proximity operators

- Proximity search (specified distance). Hint: Proximity operators and syntax are not standardized.
  - NEAR
  - NOT NEAR
  - FOLLOWED BY
  - NOT FOLLOWED BY
  - SENTENCE
  - FAR



Query syntax

- Bounded phrase
  - Usually quotation marks, e.g.

    “health insurance”
- Truncation (right, left, internal)
  - Usually an asterisk, e.g.

    child*

    “pre-existing condition*”
- Nested statements
  - Parentheses (that must match up)

    (“health insurance” AND (children* OR “pre-existing condition*”))

How to create a Boolean query (1)

1) Brainstorm a list of 10 relevant words and phrases.
2) Use that list to identify 10 relevant items (articles, videos, websites, etc.)
   - E.g., do a Google search, search Google Scholar, search the NYT (or any other newspaper that you subscribe to), search Library of Congress Chronicling America (1789-1963), etc.
3) Review 10 relevant items and write down the words and phrases that provide a context for the theme/topic/concept.
   - Titles, headings, summaries, introductions (at the beginning) and conclusions (at the end) are good areas to focus on without having to read the whole thing.
4) Note any named entities (people, organizations, events, laws, etc.) that are closely associated with the theme/topic/concept.
   - E.g., for gun violence Gabrielle Giffords, Michael Bloomberg, Doctors Against Gun Violence, March for our Lives, etc.

How to create a Boolean query (2)

5) Consolidate the terms.
   - Identify duplicates, synonyms, as well as any concepts that you want to combine even if they are not synonyms.
   - Re-label the term as needed to reflect the concept/category. Also consider and note any other relationships between terms. Prioritize the terms. Rank from 1-N, most relevant to least relevant.
   - Hint: Rank each term by higher, medium, lower relevance, then sort and rank from 1-N.
6) Write a query for each term.
   - Note that regular plurals (-s, -es, -ies) are usually (but not always) included automatically, but you always need to specify irregular plurals, e.g., “mice”.
7) Qualify the scope for each term.
   - Does the term require any qualification of the scope, e.g., by population, setting, geography, etc.?
   - Validate that the term is disjunctive, distinct, and requires no further qualification.
8) Combine the terms into a single nested query with an OR operator.

Activity: Create a Boolean search statement to scope one of the following concepts – Choose your level

Your Level	Your Concept
Basic	school shootings
Intermediate	public health
Advanced	what can public health do to curb school shootings

Questions, summary & evaluation

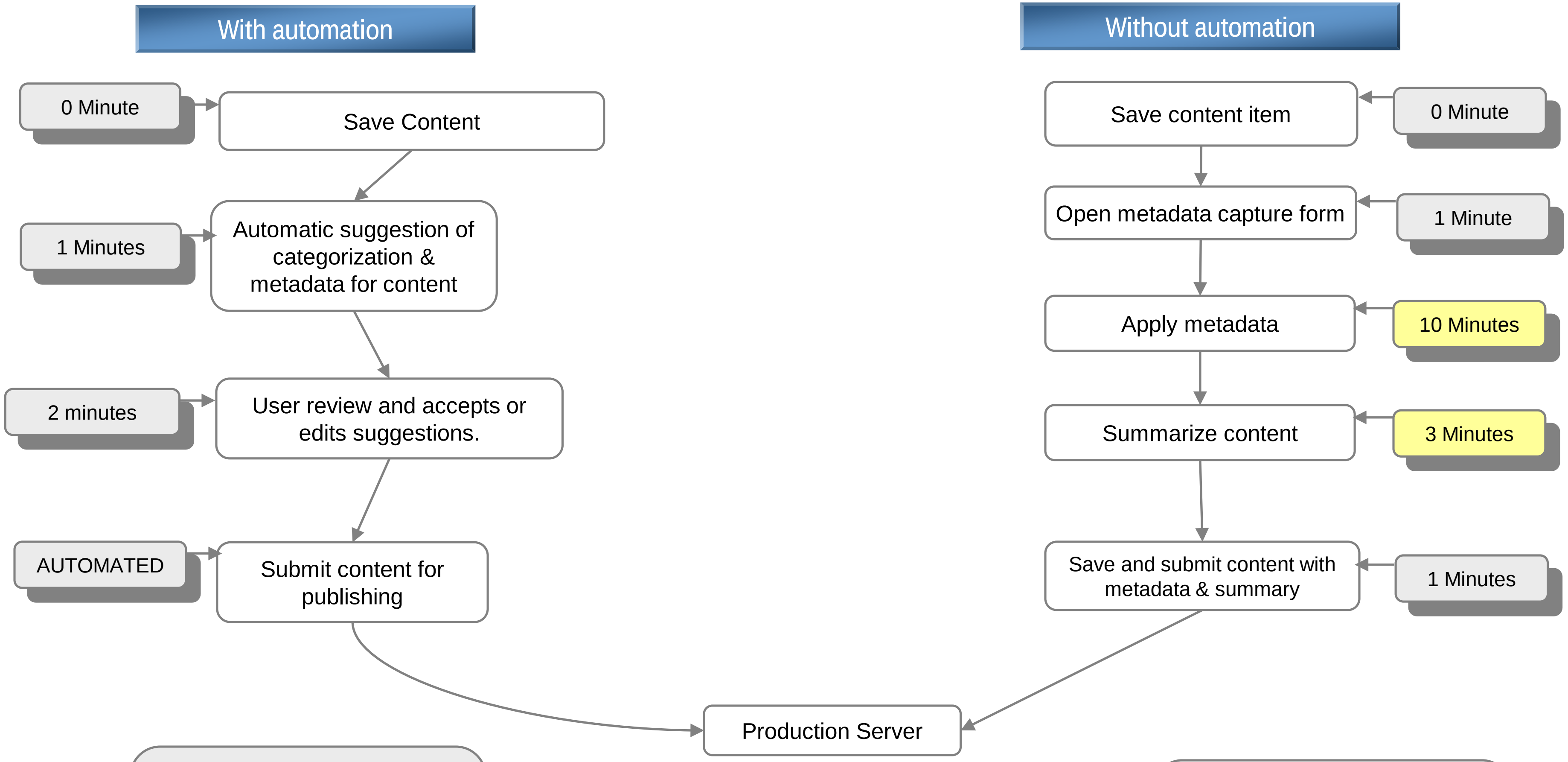
10 minutes

Joseph Busch,
jbusch@taxonomystrategies.com
joseph@semanticstaffing.com
m 415-377-7912

Appendix

- More information
- Cost-benefit analysis
- Pricing models

Cost benefit analysis: With and without automatic tagging



With automation

0 Minute → Save Content

1 Minutes → Automatic suggestion of categorization & metadata for content

2 minutes → User review and accepts or edits suggestions.

AUTOMATED → Submit content for publishing

Without automation

Save content item ← 0 Minute

Open metadata capture form ← 1 Minute

Apply metadata ← 10 Minutes

Summarize content ← 3 Minutes

Save and submit content with metadata & summary ← 1 Minutes

Production Server

Cost of Labor = $100/hr
Total Time = 3:00 minutes

Cost per page = $5

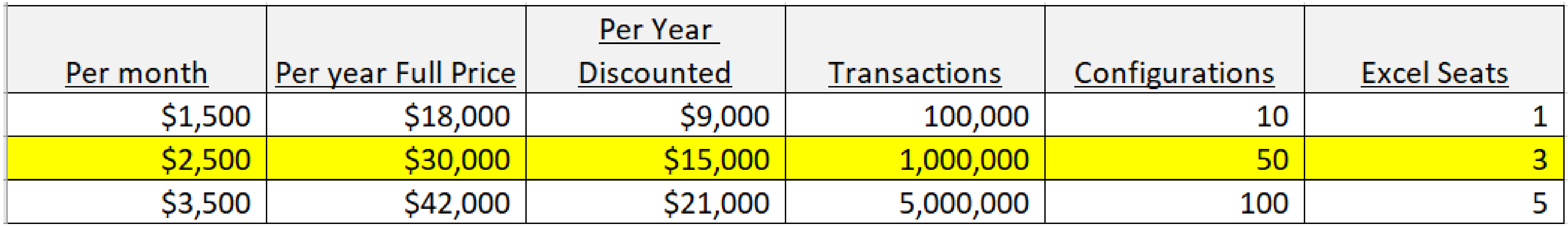
Savings Per Page = $20
Pages = 100,000

Total Savings = $2,000,000

Cost of Labor = $100/hr
Total Time = 15:00 minutes

Cost per page = $25

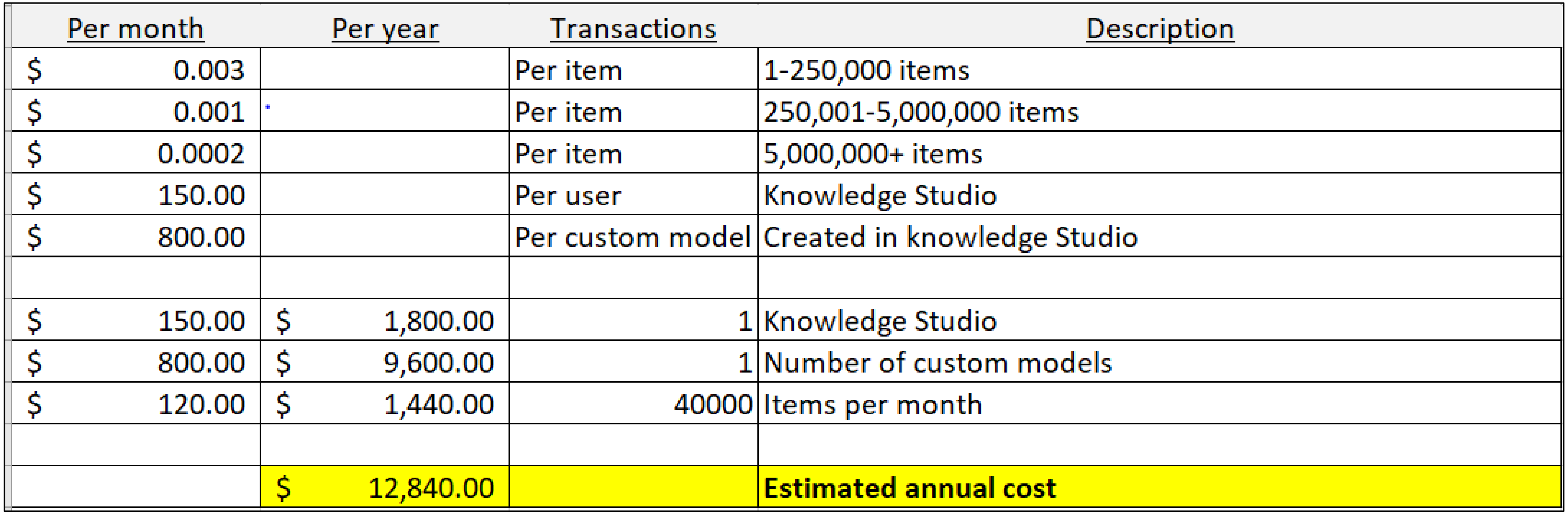
Pricing example: Lexalytics
Pricing by documents processed



Per month	Per year Full Price	Per Year Discounted	Transactions	Configurations	Excel Seats
$1,500	$18,000	$9,000	100,000	10	1
$2,500	$30,000	$15,000	1,000,000	50	3
$3,500	$42,000	$21,000	5,000,000	100	5

- Offers educational and non-profit pricing. "Pricing for educational institutions is 50% off the original package price. This can also be offered to certain non-profit organizations. https://www.lexalytics.com/prices.
- This model only considers the Semantria API topic queries method. Lexalytics offers other categorization methods including model based classifiers and concept matrix. We don't know what these other services might cost. https://www.lexalytics.com/technology/categorization

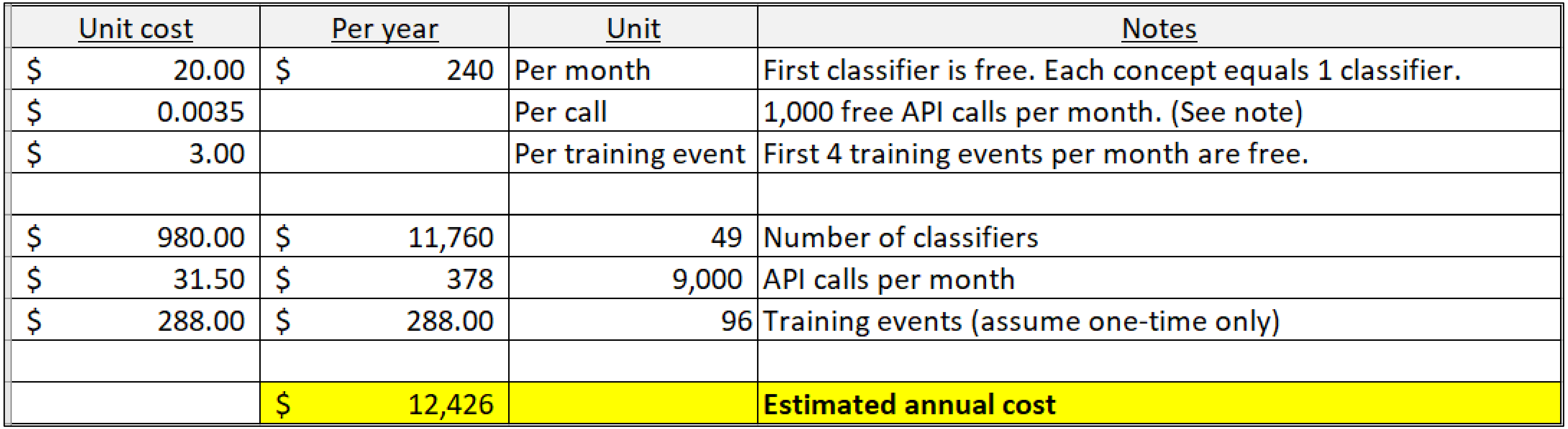
Pricing example: IBM
Pricing by classification method calls - Natural Language Understanding



Per month	Per year	Transactions	Description
$ 0.003		Per item	1-250,000 items
$ 0.001		Per item	250,001-5,000,000 items
$ 0.0002		Per item	5,000,000+ items
$ 150.00		Per user	Knowledge Studio
$ 800.00		Per custom model	Created in knowledge Studio
$ 150.00	$ 1,800.00	1	Knowledge Studio
$ 800.00	$ 9,600.00	1	Number of custom models
$ 120.00	$ 1,440.00	40000	Items per month
	$ 12,840.00		**Estimated annual cost**

- This is a text analytics web service (comparable to Lexalytics Semantria)
- An item is 10,000 characters. Documents with greater than 10,000 characters are split into multiple 10K character items. Features include Categories, Concepts, Emotion, Entities, Keywords, Metadata, Relations, Semantic Roles, and Sentiment. Each feature in the API call is counted as a separate item.
- Only one custom model is required for RWJF Topics. These would be a specialization of the Categories feature. Documents are limited to 10,000 characters. Only 4 features are extracted per item. I.e., each document equals 4 items.

Pricing example: IBM
Pricing by classification method calls - Natural Language Classifier



Unit cost	Per year	Unit	Notes
$ 20.00	$ 240	Per month	First classifier is free. Each concept equals 1 classifier.
$ 0.0035		Per call	1,000 free API calls per month. (See note)
$ 3.00		Per training event	First 4 training events per month are free.
$ 980.00	$ 11,760	49	Number of classifiers
$ 31.50	$ 378	9,000	API calls per month
$ 288.00	$ 288.00	96	Training events (assume one-time only)
	$ 12,426		**Estimated annual cost**

- This is a statistical categorizer.
- Assume 1 API equals 1 document, but could be per classifer per document, i.e., x 50. https://www.ibm.com/watson/developercloud/nl-classifier.html#pricing-block

More information: Overview

- Performance Comparison of 10 Linguistic APIs for Entity Recognition. https://www.programmableweb.com/news/performance-comparison-10-linguistic-apis-entity-recognition/elsewhere-web/2016/11/03.
- Top 27 Free Software for Text Analysis, Text Mining, Text Analytics. http://www.predictiveanalyticstoday.com/top-free-software-for-text-analysis-text-mining-text-analytics/
- Is there any free tool available for text classification? https://www.quora.com/Is-there-any-free-tool-available-for-text-classification
- Satnam Alag. *Collective Intelligence in Action*. https://www.manning.com/books/collective-intelligence-in-action
- Haralambos Marmanis and Dmitry Babenko. *Algorithms of Intelligent Web*. https://www.manning.com/books/algorithms-of-the-intelligent-web

More information: Demos

- Amit Agarwal. “Perform Text Analysis with IBM Watson and Google Docs.” (Feb 19, 2018) https://www.labnol.org/internet/ibm-watson-google-docs-nlp/31481/.
- Andreas Blumauer. “PoolParty Semantic Classifier: Webinar. (Feb 22, 2018) https://www.slideshare.net/semwebcompany/poolparty-semantic-classifier.
- UNSILO Classify - Package Manager & Upcoming Features.(Oct 3, 2017) https://www.youtube.com/watch?v=ZPoVU_Jn4iw&feature=youtu.be.

More information: Challenges

- Jeff Catlin. “The Role of Artificial Intelligence in Ethical Decision Making.” Forbes Technology Council. (Dec 21, 2017) https://www.forbes.com/sites/forbestechcouncil/2017/12/21/the-role-of-artificial-intelligence-in-ethical-decision-making/#7d94a54f21dc.
- ProPublica. “Breaking the Black Box” series.
  - Julia Angwin, Terry Parris Jr. and Surya Mattu. “What Facebook Knows About You.” (September 28, 2016) https://www.propublica.org/article/breaking-the-black-box-what-facebook-knows-about-you.
  - Julia Angwin, Terry Parris Jr. and Surya Mattu. “When Algorithms Decide What You Pay.” (October 5, 2016) https://www.propublica.org/article/breaking-the-black-box-when-algorithms-decide-what-you-pay.
  - Julia Angwin, Terry Parris Jr., Surya Mattu and Seongtaek Lim. “When Machines Learn by Experimenting on Us.” (October 12, 2016) https://www.propublica.org/article/breaking-the-black-box-when-machines-learn-by-experimenting-on-us.
  - Jeff Larson, Julia Angwin and Terry Parris Jr. “How Machines Learn to Be Racist.” (October 19, 2016) https://www.propublica.org/article/breaking-the-black-box-how-machines-learn-to-be-racist?word=Trump.
- Seth Earley. “The Problem with AI.” 19 *IT Professional* 04 (July-Aug 2017) pp 63-67. https://www.computer.org/csdl/mags/it/2017/04/mit2017040063.html.