U.S. FOOD & DRUG ADMINISTRATION

Deterministic Extraction vs. Probabilistic Extrapolation: A Pilot for R-enabled Augmentation of Information Retrieval by Humans

Selen Stromgren, PhD <selen.stromgren@fda.hhs.gov>
Danielle Larese, PhD <danielle.larese@fda.hhs.gov>
R Conference Government & Public Sector 2023
October 20, 2023

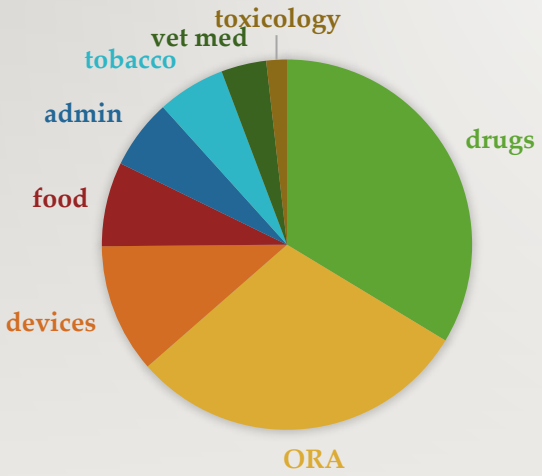
Disclaimer

This presentation reflects the views of the speakers and should not be construed to represent FDA's views or policies.

ORA within FDA

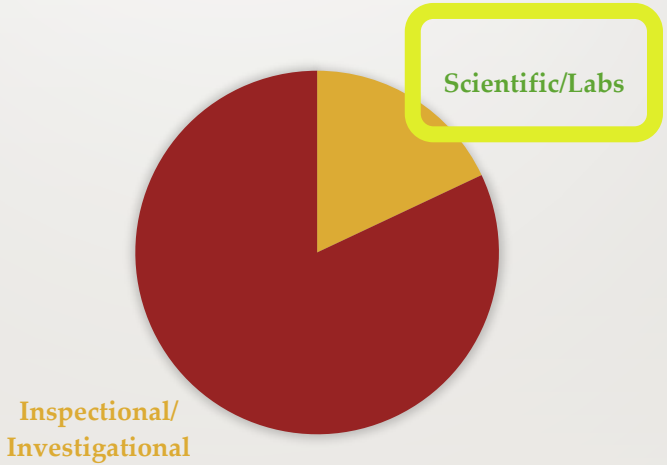
The Office of Regulatory Affairs (ORA) works with FDA Centers to provide oversight on FDA regulated consumer products

TOTAL FDA STAFF



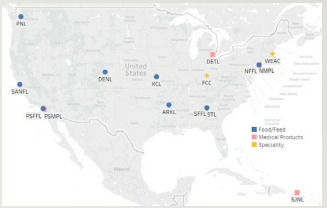
Total 18,062

ORA STAFF



Total 4,997

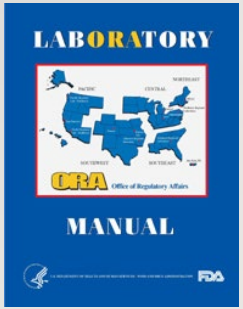
ORA is where the rubber hits the road

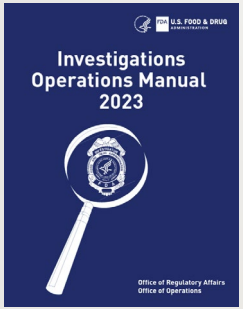


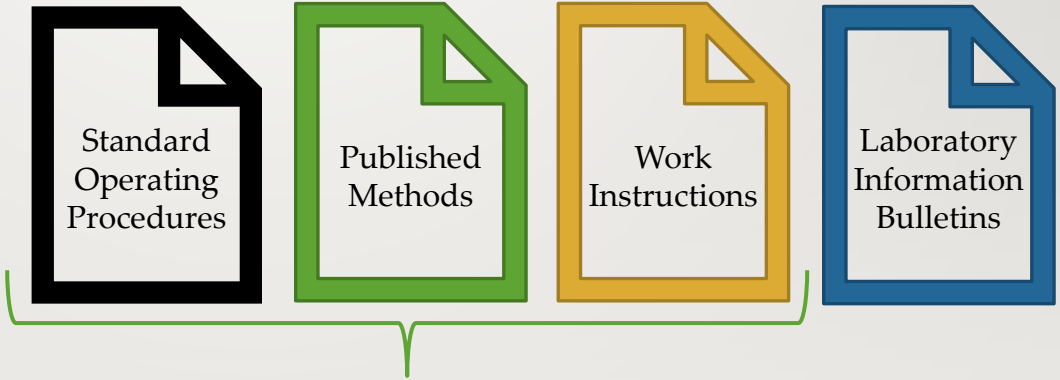
Scientific/Labs

Lots of technical PDFs at ORA

Core mission: regulatory laboratory network performs post-market analytical assessment of FDA-regulated products and provides legally defensible results







44 PDFs

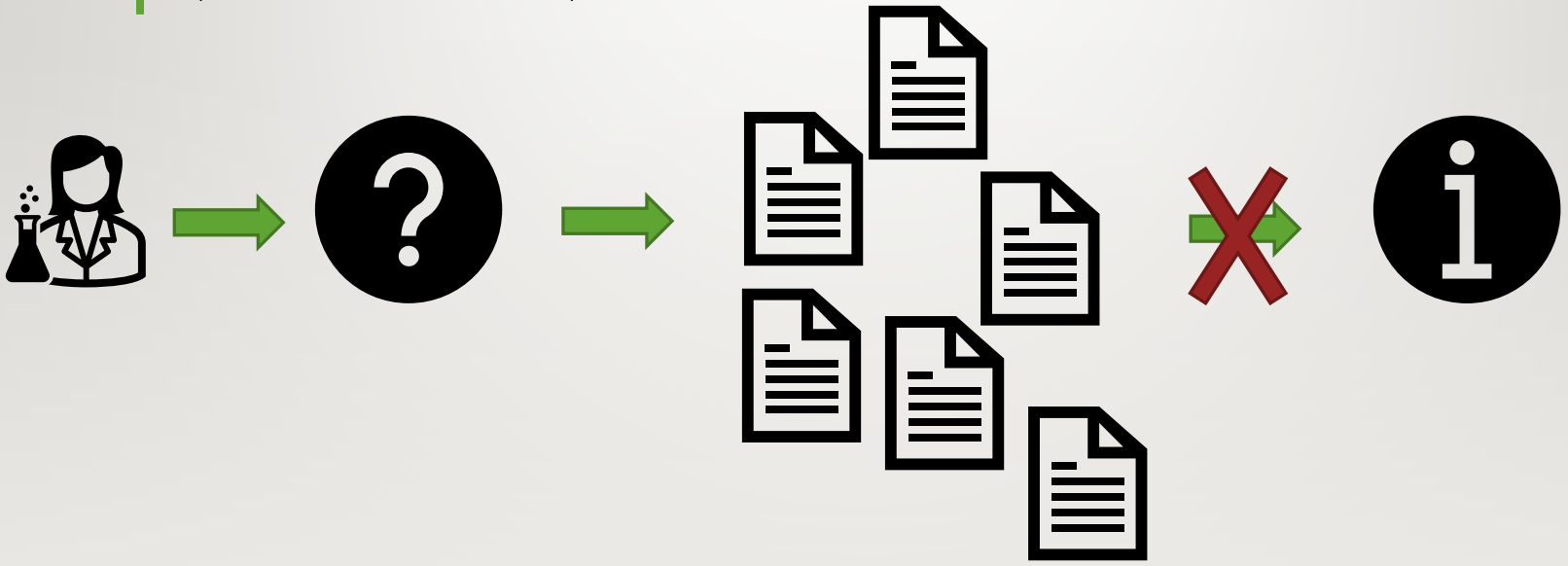
17 PDFs

~4000 PDFs

~4600 PDFs

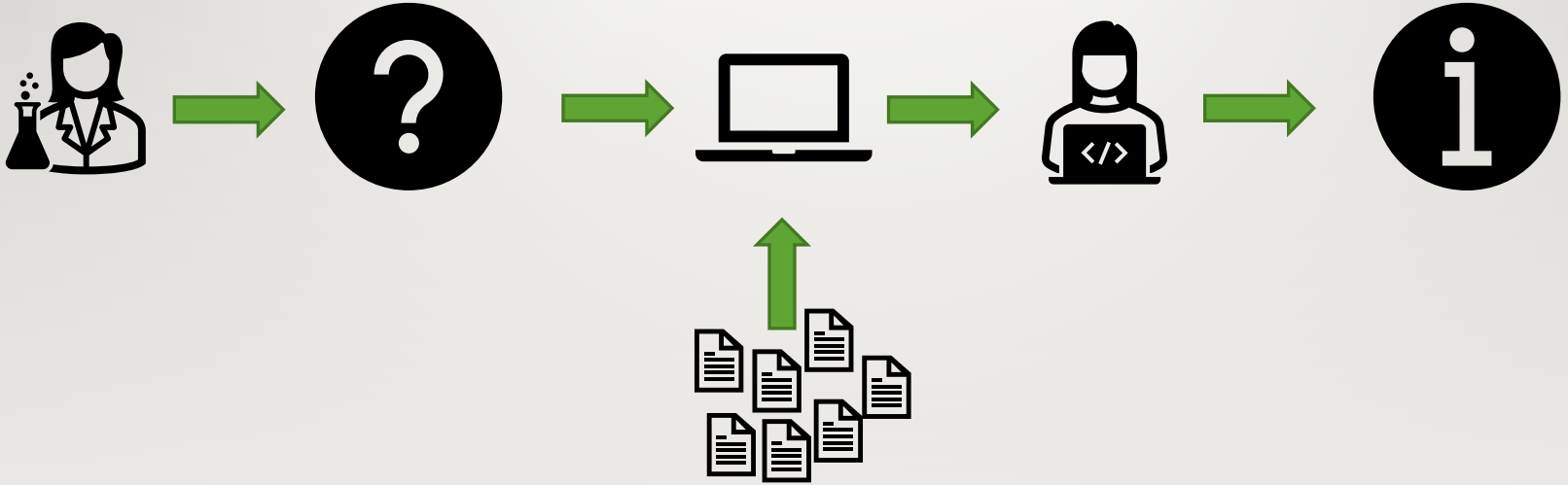
https://www.fda.gov/science-research/field-science-and-laboratories/field-science-laboratory-manual
https://www.fda.gov/inspections-compliance-enforcement-and-criminal-investigations/inspection-references/investigations-operations-manual

Factual information retrieval by humans (laboratorians)



100s or 1000s PDFs

R-enabled augmentation of information retrieval by humans (laboratorians)



Options for digesting large amounts of text



Probabilistic/AI

- Large Language Models (LLMs)
  - AzureAIStudio
  - LLaMA
  - chatGPT
  - BERT
- "Refined" LLMs
  - pharmBERT
  - pharmaGPT
  - bioBERT
  - bioGPT

Deterministic

- Human reading a document
- Basic search functions (like "find" in Word)
- Word frequency

Requirements framework for proof-of-concept project

DO want

- Simple interface for laboratorians to ask technical questions on the go within the lab
- Document security: documents are internal, often controlled in quality system
- Mathematically deterministic and accurate output

DON'T want

- Generative or "creative" responses
- False positives
- False negatives
- Data privacy or data leakage issues
- Non-reproducible answers
- "AI doesn't promise truthfulness"

Project motivation

- We're self-taught R-enthusiasts
- We're scientists who want to be responsible/informed users of AI-powered tools
- Concerns with explosion of AI-powered tools and their current appropriateness for use in regulatory lab work
- Want to turn a black box LLM into a glass box for ourselves
  - Not using LLM for ourselves, but gaining insight into how to make machines manipulate language
- Set up challenge for ourselves to see if we can produce a script in R that would perform accurate, deterministic text extraction to find answers to question inputs on a well-defined corpus

Proof of Concept Project: DIY, mathematically deterministic, simple way for us to explore the logic of Q&A across many sentences, with goal of extending script to many PDF documents.

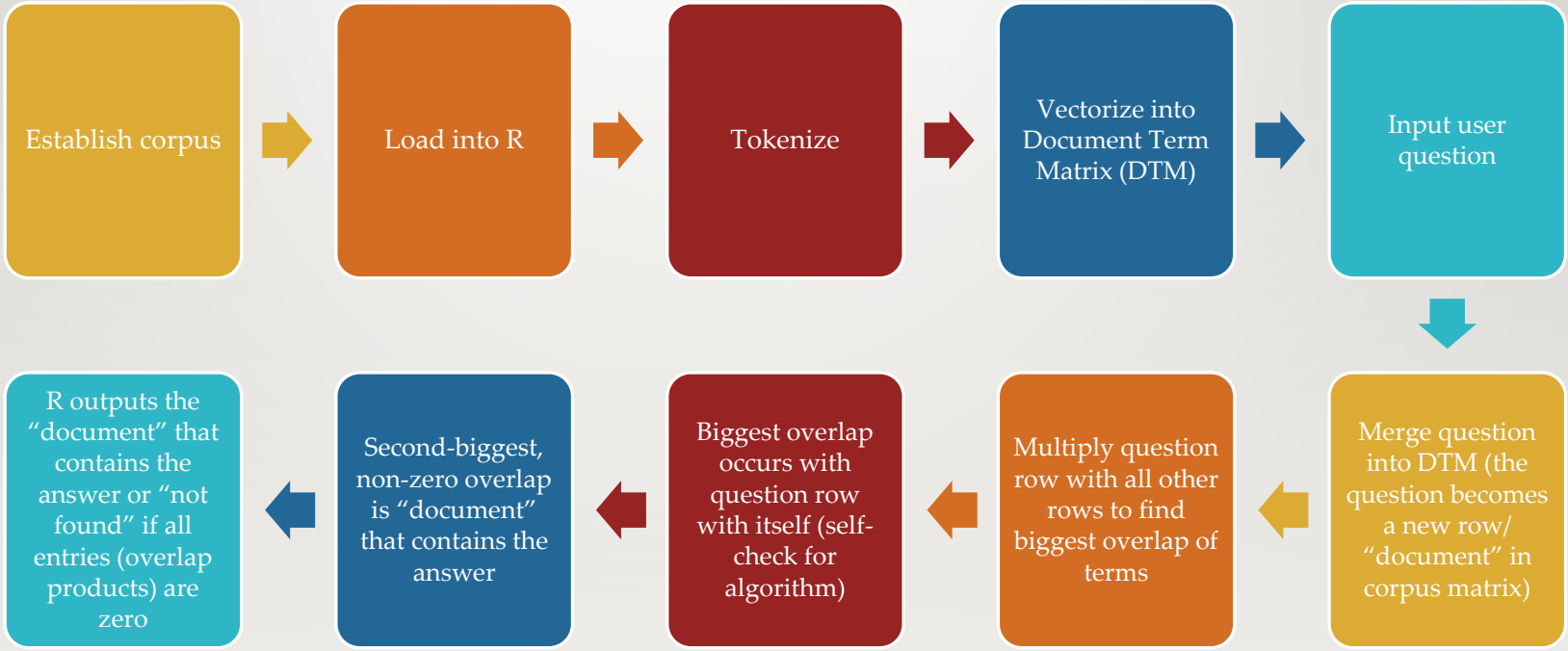
Proof-of-concept project description

- R script to extract high-accuracy answers from well-defined corpus in response to user questions
- Nano-scale corpus (5 sentences)
- Math will not include any extrapolation nor probability-based guessing
- Will produce an output only if the answer to the question posed is present in the reference corpus (as defined by 'overlap')
- If insufficient 'overlap' will return "not found"

Logic for constructing R script

- Principle: turn words into numbers, turn sentences into matrix
- To achieve high accuracy, want output to be relevant sentence or “none”
- Deterministic: no probabilistic extrapolation (such as predicting next word based on previous ones)
- We return answer with biggest overlap with question words
  - Translated language problem to a math problem by looking for maximum overlap between vectorized question sentence and vectorized corpus sentences

R script workflow



DTM = machine-readable sparse matrix which holds information about "documents" in the corpus and unique terms in the corpus

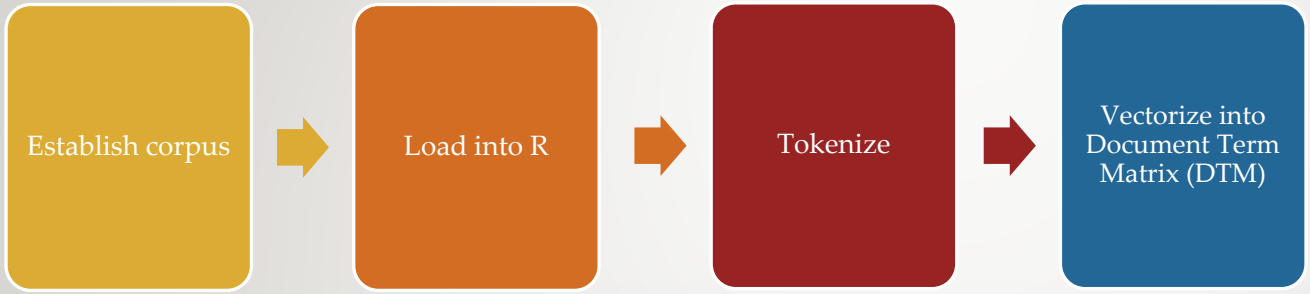
Establish corpus → Load into R

```
Last summer my family went on an amazing picnic in a park.
We brought ham sandwiches, apple pie and ice cold lemonade.
While we enjoyed our picnic lunch under a tree, a small gray squirrel crawled close to us.
Suddenly this small squirrel grabbed my delicious piece of pie off my plate and scurried off.
Hard to believe such a small animal could grab such a big piece of food.
```

five_sentences.txt - Notepad

```
1  library(data.table)
2  txt_in <- readLines("five_sentences.txt")
3  table_in <- data.table(txt_in)
4  id <- 1:5
5  table_corpus <- cbind(table_in, id)
```

	txt_in	id
1	Last summer my family went on an amazing picnic in a park.	1
2	We brought ham sandwiches, apple pie and ice cold lemona...	2
3	While we enjoyed our picnic lunch under a tree, a small gray...	3
4	Suddenly this small squirrel grabbed my delicious piece of p...	4
5	Hard to believe such a small animal could grab such a big pi...	5




```
library(text2vec)
prep_fun = tolower
tok_fun = word_tokenizer
iter = itoken(table_corpus$txt_in,
              preprocessor = prep_fun,
              tokenizer = tok_fun,
              ids = table_corpus$id,
              progressbar = FALSE)
```

```
vocabiter = create_vocabulary(iter)
vectorizer = vocab_vectorizer(vocabiter)
dtm_corpus = create_dtm(iter, vocabiter)
```

DTM is the sparse matrix that R uses to store all the content of the corpus (not human comprehensible)



Because we're working at nano-scale, we can 'densify' this 'sparse' matrix to visually inspect

	amazing	an	animal	apple	believe	big	brought	close	cold	could	crawled	delicious	enj
1	1	1	0	0	0	0	0	0	0	0	0	0	
2	0	0	0	1	0	0	1	0	1	0	0	0	
3	0	0	0	0	0	0	0	1	0	0	1	0	
4	0	0	0	0	0	0	0	0	0	0	0	1	
5	0	0	1	0	1	1	0	0	0	1	0	0	

DTM: 5 rows, 52 columns
(5 sentences, 52 unique words)

```
17  question <- data.table("Where did the family go?","6")
```

Second-biggest, non-zero overlap is "document" that contains the answer

Biggest overlap occurs with question row with itself (self-check for algorithm)

Multiply question row with all other rows to find biggest overlap of terms

Merge question into DTM (the question becomes a new row/ "document" in corpus matrix)

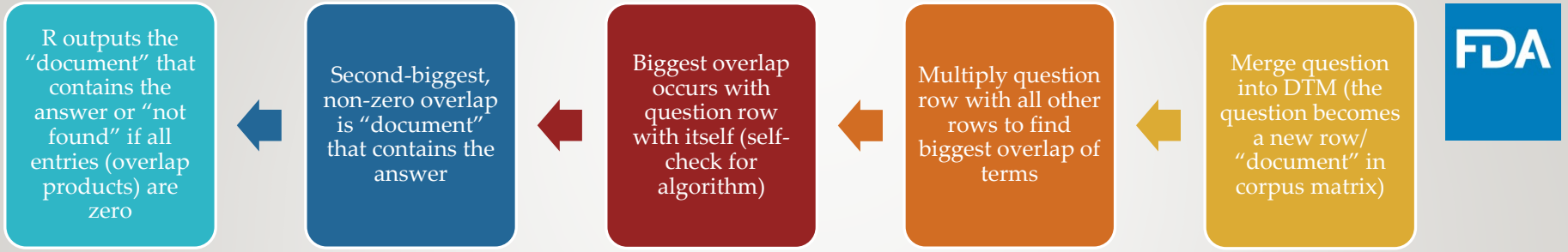
FDA

```
18 table_q <- rbind(table_corpus, question, use.names=FALSE)
19 iter = itoken(table_q$txt_in,
20                preprocessor = prep_fun,
21                tokenizer = tok_fun,
22                ids = table_q$id,
23                progressbar = FALSE)
24 vocabiter = create_vocabulary(iter)
25 vectorizer = vocab_vectorizer(vocabiter)
26 dtm_q = create_dtm(iter, vectorizer)
27 dtm_q@Dim
```

DTM: 5 rows, 52 columns
DTM+Q: 6 rows, 56 columns

```
28 dense_q <- as.matrix(dtm_q)
29 ans <- t(dense_q[6,]*t(dense_q))
30 rowSums(ans)
```

```
> rowSums(ans)
1 2 3 4 5 6
1 0 0 0 0 5
```




```
Last summer my family went on an amazing picnic in a park.
We brought ham sandwiches, apple pie and ice cold lemonade.
While we enjoyed our picnic lunch under a tree, a small gray squirrel crawled close to us.
Suddenly this small squirrel grabbed my delicious piece of pie off my plate and scurried off.
Hard to believe such a small animal could grab such a big piece of food.
```

```
"Where did the family go?"
```

```
1: Last summer my family went on an amazing picnic in a park.
```

Calculating question and answer 'overlap' is just left-multiplying the sentence row to the corpus matrix

Future steps: expanded use cases

- Possible expanded use to internal FDA standard operating procedure (SOPs)
  - Investigators during inspections
  - Lab analysts during assay
  - Helps not flip pages during job
- Limitation to overcome to achieve a larger size corpus: currently our script uses a “dense” matrix representation of corpus, need to work with sparse matrix for memory management
- Need to determine how to return null/‘not found’ threshold and correlate confidence in answer to degree of overlap
- What this project promises is:
  - Complete transparency into how the code works
  - Complete ownership of the code
  - Protection of internal confidential data
  - Accurate answers or null answers if the info is not contained in the corpus
  - This is especially important for scientific and regulatory work where the operations have to be by the book

Thank you!

Selen.Stromgren@fda.hhs.gov

Danielle.Larese@fda.hhs.gov

Connect with us on LinkedIn