

# Extracting Information from Unstructured Data

**Neil Alexander Kattampallil**

Biocomplexity Institute, University of Virginia

**Gary Anderson**, National Center for Science and Engineering Statistics (NCSES), National  
Science Foundation



# Unstructured Text Data : An Overview

Any information that doesn't follow conventional data models, making it difficult to store and manage in a Relational Database.

- Word Documents
- Emails
- Social Media posts
- Medical notes (Healthcare NLP)

Most of this exists as 'dark data', data that is collected by an organization which is not used for analytics or insight generation, and is usually stored only for compliance purposes.

# Unstructured Data as an information asset

Taken from Healthcatalyst.com, [5 Reasons Healthcare Data is Difficult to Measure](#)

"...a recent initiative to reduce unnecessary C-sections at a large health system in the Northwest. The first task for the team was to understand how the indications for C-section were documented in the Electronic Medical Record. It turned out that there were only two options to choose from: 1) fetal indication and 2) maternal indication. Because these were the only two options, delivering clinicians would often **choose to document the true indication for C-section in a free text form**, while others did not document it at all. "

# Methods: BERT



A pre-trained unsupervised Natural Language Processing model  
developed by Google in 2018

# BERT Benefit

*Pre-trained models solve challenge of lack of enough training data*

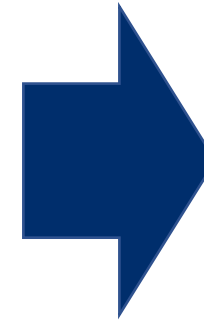


Book Corpus:  
2,500 million words



WIKIPEDIA  
The Free Encyclopedia

Wikipedia:  
800 million words




Can use pre-trained models to fine tune on smaller-task specific datasets and improves accuracy drastically

# How is BERT Trained?

One of the first of these models, BERT, is trained by taking sentences, splitting them into individual words, randomly hiding some of them, and predicting what the hidden words are. After doing this millions of times, BERT has “read” enough Shakespeare to predict how this phrase usually ends:

to be or not to be , that is the   ;



56.987% question	3.610% difference	3.004% answer
2.691% problem	2.623% key	0.954% challenge
0.899% truth	0.743% game	0.719% point
0.678% definition	0.618% riddle	0.584% idea
0.576% dilemma	0.568% message	0.555% phrase

BERT's predictions for what should fill in the hidden word

# Semantic Similarity

Language models like BERT are based on training with large data sets, such as:

- full text of Wikipedia
- Open Books corpus

The goal is to develop a vector representation of words, in a given sentence that represent a given idea.

These Language models contain vector representations of words, clustered by similarity of ideas in an n-dimensional vector space.

# Goals for this presentation:

- **The problem:** Extracting information from text articles
- **Why it is a problem:** Lot of information is stored in the form of human readable text articles, which may not be easily accessible to automated systems, but this information is vital for research and analysis tasks.
- **What has been done about the problem:**  
Currently, to measure Innovation, NCSES uses Surveys to businesses to obtain information about what their new products are.
- **What you are doing (or have done) about the problem:**  
We are using news articles and press releases as a dataset, as companies often try to announce new product launches, and we can extract product name, company name and product features from articles.



# Current Measure of Innovation



## ***EuroStat: Community Innovation Survey***

- Conducted every 2 years by EU member states and ESS member countries
- Survey of innovation activity by enterprises

## ***Annual Business Survey (US Census Bureau and NSF/NCSES)***

- Collects data on R&D, innovation, technology, intellectual property, and business owner characteristics
- Annual; initial year: 2018
- Previously, Business R&D Innovation Survey (BRDIS) launched in 2009

### **C.1 New or Improved Goods**

During the three years 2017 to 2019, did this business introduce to the market any new or improved goods that differed significantly from this business's previous goods? *(This includes the addition of new functions or improvements to existing functions or user utility. Functions include quality, technical specifications, reliability, durability, economic efficiency during use, affordability, convenience, usability, and user friendliness. User utility includes attributes such as affordability and financial convenience.)*

**Goods:** usually a tangible object such as a smartphone, furniture, or packaged software, but also includes digital goods such as downloadable software, music, and film. *(Exclude the simple resale of new goods or changes of a solely aesthetic nature.)*

☐ Yes

☐ No

### **C.2 New or Improved Services**

During the three years 2017 to 2019, did this business introduce to the market any new or improved services that differed significantly from this business's previous services? *(This includes the addition of new functions or improvements to existing functions or user utility. Functions include quality, technical specifications, reliability, durability, economic efficiency during use, affordability, convenience, usability, and user friendliness. User utility includes attributes such as affordability and financial convenience.)*

**Services:** Intangible activities, such as retailing, insurance, educational courses, air travel, consulting, etc., also includes digital services. *(Exclude the simple resale of new services.)*

☐ Yes

☐ No

Annual Business Survey (ABS) question on innovation



# Background and Goals

- Can we use alternative (non-survey) data sources to measure business innovation?
  - While **ABS** measures innovation incidence, i.e., the number of innovating firms, we aim to test the feasibility of developing methods using non-traditional data to obtain richer and complementary innovation measures.
- We develop *natural language processing and machine learning methods* using non-survey data to obtain richer and complementary innovation measures.
- Focus on *opportunity and administrative data* (e.g., product announcements, press releases, financial filings).
- What *type(s) of innovation* or improvement (i.e., greater efficacy, resource efficiency, reliability and resilience, affordability, convenience and usability) could be captured using opportunity data sources?
- How does it vary across different *companies and sectors*?

# Current Focus

- **Innovation Type:** Product innovation as *defined by* the OSLO Manual
- **Sectors:**
  - Pharmaceutical and Medicine Manufacturing (NAICS 3254)
  - Food Processing & Beverage Manufacturing (NAICS 311 & 3121)
  - Computer Systems Design and Related Services (NAICS 5415)
- **Data Sources:** News articles
- **Innovation Metrics:** Number of new products/launches by company & measures of “novelty of innovation”
- **Validation:** Comparison of extracted metrics to ‘ground-truth’ data and computation of performance metrics to evaluate methods
- **Repeatability:** Applying these data and methods to other sectors?



# Data Sources

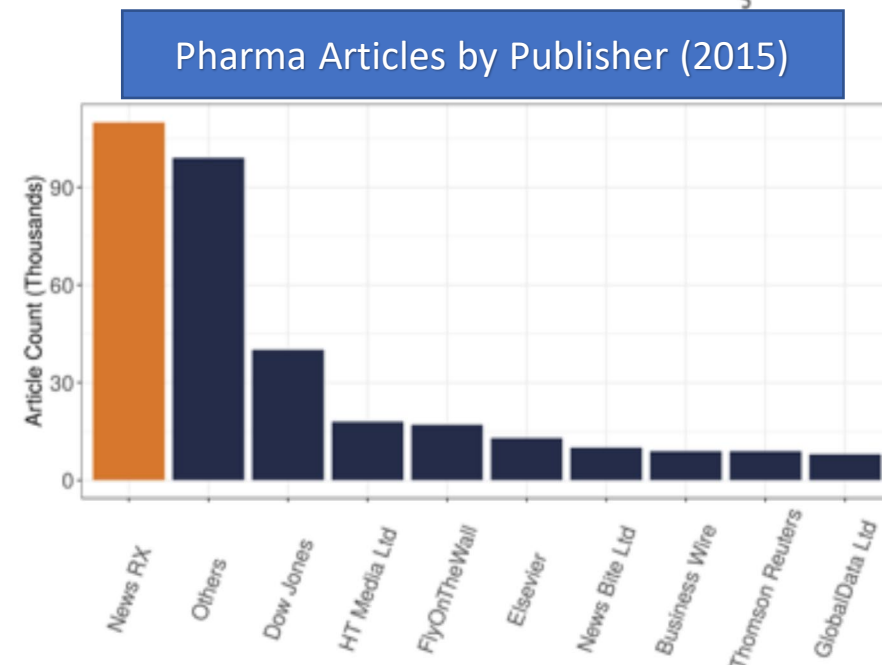
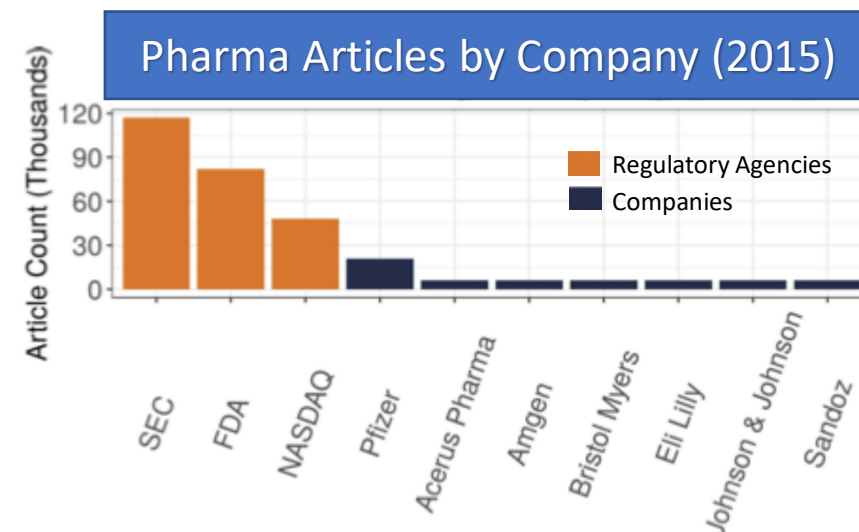


## News articles from Dow Jones Data, News, and Analytics (DNA) database

- English language
- Articles between 2013 and 2021
- Region Code USA
- Variables: company\_codes, subject\_codes, region\_codes, word\_count
- Additional DNA Subject Codes that can be used to further improve data selection, e.g., c22 – New Product/Service

Sector	Number of articles
Pharmaceutical and Medicine Manufacturing (NAICS 3254)	1.8 M (2013 - 2018)
Food Processing & Beverage Manufacturing Sectors (NAICS 311 & 3121)	600 K (2013 - 2021)
Computer Systems Design and Related Services (NAICS 5415)	1.2 M (2013 - 2021)

Data Obtained



# Named Entity Recognition

Named-entity recognition (NER), also known as

- (named) entity identification
- entity chunking, and entity extraction)

This is a subtask of information extraction that seeks to locate and classify named entities mentioned in unstructured text into pre-defined categories such as

- person names
- organizations, locations
- medical codes
- time expressions
- quantities
- monetary values
- percentages, etc.

# Categories for NER

Normally, **NER uses eight categories**—location, person, organization, date, time, percentage, monetary value, and “none-of-the-above”. NER first finds named entities in sentences and declares the category of the entity. In the sentence:

“**Apple** [organization] CEO **Tim Cook** [Person] Introduces 2 New, Lager iPhones, Smart Watch at **Cupertino** [Location] **Flint Center** [Organization] Event.”

Note that “Apple” is recognized as an organization name instead of a fruit name in terms of its context.

# BERT NER models

bert-base-NER is a fine-tuned BERT model that is ready to use for Named Entity Recognition and achieves state-of-the-art performance for the NER task. It has been trained to recognize four types of entities: location (LOC), organizations (ORG), person (PER) and Miscellaneous (MISC).

Specifically, this model is a bert-base-cased model that was fine-tuned on the English version of the standard CoNLL-2003 Named Entity Recognition dataset.

# BERT - Named Entity Recognition (NER)

- Identifies and categorizes entities into 4 categories based on the context that the words are used in
  - Location, Person, Organization, Miscellaneous

My name is **Wolfgang** **PER** and I live in **Berlin** **LOC**. I work at **Google** **ORG** and drink **Starbucks** **MISC**

Tagging by huggingface model `dlssim/bert-base-NER`.

- Tokenizes words
  - Craftsbury becomes Crafts## and ##bury
- Uses surrounding words (Sentence context)
  - John works at \_\_\_\_\_.
- Is Case Sensitive
- Can be further fine-tuned by training on a corpus



# BERT - Named Entity Recognition (NER)

- Use NER to extract organization names from the articles to identify company names

Year_y	Lead Paragraph_512	parsed_sentence	labels_for_sentences	Orgs from NER	Misc from NER
2013	VP COPELAND Sells 5,000 Of EPIZYME INC >EPZM(E...	[copeland, epizyme, inc, dow, jones, gmt]	[B-PER, B-ORG, I-ORG, B-ORG, I-ORG, B-MISC]	[epizyme, inc, dow, jones]	[gmt]
2013	Millennium Research Group; US Market for Denta...	[millennium, research, group, us, zimmer, keys...]	[B-ORG, I-ORG, I-ORG, B-MISC, B-ORG, B-ORG, I-...	[millennium, research, group, zimmer, keystone...]	[us, us]
2013	PRESS RELEASE: Synergy Pharmaceuticals Present...	[synergy, pharmaceuticals, plecanatide, synerg...]	[B-ORG, I-ORG, B-MISC, B-ORG, I-ORG, B-LOC, I-...	[synergy, pharmaceuticals, synergy, pharmaceut...]	[plecanatide, nasdaq, cic, ibs, -, c, guanylat...]

- Fuzzy Match (fuzzywuzzy) extracted names with DNA company names to evaluate NER

Innovators	Other_Companies	parsed_sentence	labels_for_sentences	Orgs from NER	Misc from NER	match ratios	highest match
23andMe	NaN	[gattaca, -, style, california, 23andme, new, ...]	[B-MISC, B-MISC, B-MISC, B-MISC, B-ORG, B-LOC, ...]	[23andme, genepeeks, bbc]	[gattaca, -, style, california, -, based, gatt...]	[(23andme, 100), (genepeeks, 25), (bbc, 0)]	(23andme, 100)
Innovus Pharmaceuticals	NaN	[bassam, damaj, innovus, pharmaceuticals, apea...]	[B-PER, I-PER, B-ORG, I-ORG, B-MISC, B-MISC, B-...	[innovus, pharmaceuticals]	[apeaz™, apeaz™, apeaz™]	[(innovus, 90), (pharmaceuticals, 90)]	(innovus, 90)

# BERT - Named Entity Recognition (NER)

- Fuzzy Match (fuzzywuzzy) extracted names with DNA company names to evaluate NER

Innovators	Other_Companies	parsed_sentence	labels_for_sentences	Orgs from NER	Misc from NER	match ratios	highest match
23andMe	NaN	[gattaca, -, style, california, 23andme, new, ...]	[B-MISC, B-MISC, B-MISC, B-MISC, B-ORG, B-LOC, ...]	[23andme, genepeaks, bbc]	[gattaca, -, style, california, -, based, gatt...]	[(23andme, 100), (genepeaks, 25), (bbc, 0)]	(23andme, 100)
Innovus Pharmaceuticals	NaN	[bassam, damaj, innovus, pharmaceuticals, apea...]	[B-PER, I-PER, B-ORG, I-ORG, B-MISC, B-MISC, B...]	[innovus, pharmaceuticals]	[apeaz™, apeaz™, apeaz™]	[(innovus, 90), (pharmaceuticals, 90)]	(innovus, 90)

# NER – Company Name Extraction

- Performance Results for labeled data from 3 Sectors:

	Pharma	Food & Beverage	Software
Total Number of Labeled Company Names:	284	129	232
Exact Matches:	110	36	112
Fuzzy Matches:	159	73	80
Total Matches:	269	109	192
Total Accuracy (%):	<b>95%</b>	<b>84%</b>	<b>83%</b>

Labeled data is a random sample of articles across all available years, as mentioned in slide 10

# Training for Additional Categories

If there are custom categories that need to be extracted, we can use transfer learning to build on the existing categories that the model is trained for, and then add training data for new categories.

# Applications of BERT:

## *Named Entity Recognition*

### Results of NER Annotation Pipeline

A 28-year-old female with a history of gestational diabetes mellitus diagnosed eight years prior to presentation and subsequent type two diabetes mellitus ( T2DM ), one prior episode of HTG-induced pancreatitis three years prior to presentation , associated with an acute hepatitis , and obesity with a body mass index ( BMI ) of 33.5 kg/m2 , presented with a one-week history of polyuria , polydipsia , poor appetite , and vomiting . Two weeks prior to presentation , she was treated with a five-day course of amoxicillin for a respiratory tract infection . She was on metformin , glipizide , and dapagliflozin for T2DM and atorvastatin and gemfibrozil for HTG . She had been on dapagliflozin for six months at the time of presentation . Physical examination on presentation was significant for dry oral mucosa ; significantly , her abdominal examination was benign with no tenderness , guarding , or rigidity . Pertinent laboratory findings on admission were : serum glucose 111 mg/dl , bicarbonate 18 mmol/l , anion gap 20 , creatinine 0.4 mg/dL , triglycerides 508 mg/dL , total cholesterol 122 mg/dL , glycated hemoglobin ( HbA1c ) 10% , and venous pH 7.27 . Serum lipase was normal at 43 U/L . Serum acetone levels could not be assessed as blood samples kept hemolyzing due to significant lipemia . The patient was initially admitted for starvation ketosis , as she reported poor oral intake for three days prior to admission . However , serum chemistry obtained six hours after presentation revealed her glucose was 186 mg/dL , the anion gap was still elevated at 21 , serum bicarbonate was 16 mmol/L , triglyceride level peaked at 2050 mg/dL , and lipase was 52 U/L . The  $\beta$ -hydroxybutyrate level was obtained and found to be elevated at 5.29 mmol/L - the original sample was centrifuged and the chylomicron layer removed prior to analysis due to interference from turbidity caused by lipemia again . The patient was treated with an insulin drip for euDKA and HTG with a reduction in the anion gap to 13 and triglycerides to 1400 mg/dL , within 24 hours . Her euDKA was thought to be precipitated by her respiratory tract infection in the setting of SGLT2 inhibitor use . The patient was seen by the endocrinology service and she was discharged on 40 units of insulin glargine at night , 12 units of insulin lispro with meals , and metformin 1000 mg two times a day . It was determined that all SGLT2 inhibitors should be discontinued indefinitely . She had close follow-up with endocrinology post discharge .

Color codes: Patient problem, Test, Treatment

Clinical NER model in Spark NLP library

<https://towardsdatascience.com/named-entity-recognition-ner-with-bert-in-spark-nlp-874df20d1d77>

# Bert - Question and Answering (QnA)

- BERT uses token embedding similarities to retrieve answers to a given question, from a given reference corpus of information.
- We use QnA to extract company and product names

- Input Question:

```
Where do water droplets collide with ice  
crystals to form precipitation?
```

- Input Paragraph:

```
... Precipitation forms as smaller droplets  
coalesce via collision with other rain drops  
or ice crystals within a cloud. ...
```

- Output Answer:

```
within a cloud
```



# BERT - Question and Answering (QnA)

100%|██████████| 1/1 [00:00<00:00, 1.66it/s]  
['ROSA(TM) BRAIN SYSTEM']

MEDTECH : MEDTECH ANNOUNCES NEW SALES OF THE ROSA(TM) BRAIN SYSTEM IN THE UNITED STATES AND CHINA

what's the new product? question

ROSA(TM) BRAIN SYSTEM answer

0.5903190581759756 probability

100%|██████████| 1/1 [00:02<00:00, 2.74s/it]  
['ROSA™ Spine']

The ROSA™\_Brain system at the Yale Comprehensive Epilepsy Center will be used for minimally invasive procedures such as stereoelectroencephalography (SEEG), deep brain stimulation, and biopsies. The program will be led by Dennis Spencer, MD, Surgical Director and Harvey and Kate Cushing Professor of Neurosurgery; and Jason Gerrard, MD, PhD, Assistant Professor of Neurosurgery and of Neuroscience and Section Chief, Stereotactic and Functional Neurosurgery. The other two sales of ROSA™ Brain were made in China, a fast-growing market representing a strong opportunity for Medtech. With these two new sales, Medtech now has nine ROSA™ robots in the Chinese territory. "We are pleased to continue to expand our ROSA™ technology in these large neurosurgical markets," said Bertin Nahum, CEO and Founder of Medtech. "The three new sales in China and the United States testify to the positive momentum of our commercial activity." CONTACT MEDTECH Christophe Sibillin Chief Financial Officer +33 (0)4 67 10 77 40 INVESTORS Corinne Puissant +33 (0)1 53 67 36 77 cpuissant@actus.fr PRESS Alexandra Prisa (EU) +33(0)1 53 67 36 90 aprisa@actus.fr Joanna Zimmerman (US) +1 646-536-7006 jzimmerman@theruthgroup.com About MEDTECH Founded in 2002 by Bertin NAHUM and based in Montpellier, MEDTECH is a European specialist in the design, development and marketing of innovative robotic appliances to assist surgeons during their medico-surgical interventions, thus contributing to the implementation of safer, more efficient, less-invasive treatment. In 2007, MEDTECH developed ROSA™, an innovative technological device devoted to brain surgery procedures. ROSA™ has been approved in Europe, the United States and Canada. In 2013 MEDTECH received the "European Company of the Year Award" in the "robotic neurosurgery" category from Frost & Sullivan. In July 2014, MEDTECH obtained the CE marking for its new product ROSA™ Spine, a robotic-assistive device for minimally invasive surgery of the spine. In October 2014, MEDTECH won the "Révélation" prize in the Mediterranean Deloitte Technology Fast 50 Awards. In 2015 MEDTECH received the "2016 Company of the Year Award" in the "robotic neurosurgery" category from Frost & Sullivan. In November 2015, MEDTECH was honored by Deloitte In Extensio for its excellent performance in the Technology Fast 50 Mediterranean Awards, in the "listed company" category. In January 2016, MEDTECH obtained the FDA clearance for its new product ROSA™ Spine, a robotic-assistive device for minimally invasive surgery of the spine. Information réglementée Communiqués au titre de l'obligation d'information permanente : - Autres communiqués Full and original press release in PDF: [http://www.actusnews.com/documents\\_communiqués/ACTU-0-43418-Medtech-3-ventes-1Yale-et-2Chine-ENG-TRG-FINAL-3-13-16.pdf](http://www.actusnews.com/documents_communiqués/ACTU-0-43418-Medtech-3-ventes-1Yale-et-2Chine-ENG-TRG-FINAL-3-13-16.pdf) [http://www.actusnews.com/documents\\_communiqués/ACTU-0-43418-Medtech-3-ventes-1Yale-et-2Chine-ENG-TRG-FINAL-3-13-16.pdf](http://www.actusnews.com/documents_communiqués/ACTU-0-43418-Medtech-3-ventes-1Yale-et-2Chine-ENG-TRG-FINAL-3-13-16.pdf)

what's the new product?

ROSA™ Spine

0.978813461342744

Article title

leading paragraph

# Bert - Question and Answering (QnA)

	title	Lead Paragraph_512	names_company_about_list	Innovators	What's the new product?	What's the new product? **probability**	what's the company name?	what's the company name? **probability**	what's the new drug?	what's the new drug? **probability**	...	Which company announced the product? **probability**	When will the company announce the product?	When will the company announce the product? **probability**
0	MEDTECH : MEDTECH ANNOUNCES NEW SALES OF THE ROSA™ Brain system at the Yale Comprehensive...	The ROSA™ Brain system at the Yale Comprehensive...	[[Medtech SAS]]	NaN	ROSA™ Spine	0.978814	Medtech	0.798178	ROSA™ Spine	0.673567	...	0.737372	January 2016	0
1	CUBA SEEKS MALAYSIAN COLLABORATION VIA PHARMAC...	She said Cuba viewed Malaysia as a significant...	[]	NaN	lung cancer vaccine	0.580466	Cimavax	0.732677	lung cancer	0.334009	...	0.737267	next year	0
2	Stents appear to increase stroke patients' rec...	Currently, standard stroke care in the United ...	[]	Medtronic PLC	stent-based clot removal	0.293876	Medtronic	0.848025	tPA	0.865578	...	0.975851	Bloomberg Businessweek	0
3	PSIVIDA CORP. pSivida to Present At Two Invest...	pSivida will also present at the Stifel Nicola...	[[EyePoint Pharmaceuticals Inc'], ['Direct Ma...	EyePoint Pharmaceuticals Inc	injectable, sustained release micro-insert ILU...	0.298959	pSivida Corp	0.317354	Latanoprost	0.962495	...	0.439559	Thursday, September 12	0
4	Strides Shasun receives US FDA approval	Strides is launching the product immediately. ...	[[U.S. Food and Drug Administration'], ['Stri...	Strides Pharma Science Ltd	Tenofovir Disoproxil Fumarate	0.928353	Strides Pharma Inc	0.451422	Tenofovir Disoproxil Fumarate	0.881826	...	0.946747	immediately	0
5	Gilead Sciences Inc Files Patent Application f	The abstract of the patent published by the Co	[[Gilead Sciences Inc]]	Gilead Sciences Inc	experimental drug candidate	0.425972	Gilead Sciences, Inc.	0.505565	HIV/AIDS	0.851420	...	0.527109	July 31, 2015	0



# QnA - Company and Product Extraction

Pharma	Product	Company
Total Number of Labeled Names:	284	284
Exact Matches:	140	88
Fuzzy Matches:	111	131
Total Matches:	251	219
Total Accuracy (%):	<b>88.3%</b>	<b>77.1%</b>

# Thank you!

- Questions & Comments?

## Contact:

**Thomas Neil Alexander Kattampallil** ([nak3t@virginia.edu](mailto:nak3t@virginia.edu))

Research Scientist

Biocomplexity Institute, University of Virginia

**Acknowledgments:** This material is based on work supported by U.S. Department of Agriculture (58-3AEU-7-0074) and the National Science Foundation (Contract #49100420C0015)

**Disclaimer:** The views expressed in this paper are those of the authors and not necessarily those of their respective institutions.