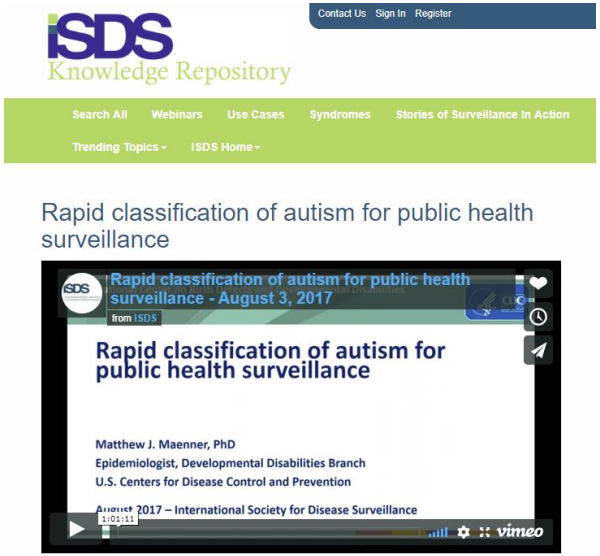
Human-learned lessons about machine learning in public health surveillance

Matthew J. Maenner, PhD
Surveillance Team Lead and Epidemiologist
Developmental Disabilities Branch
National Center on Birth Defects and Developmental Disabilities
Centers for Disease Control and Prevention

December 2018 ISDS webinar

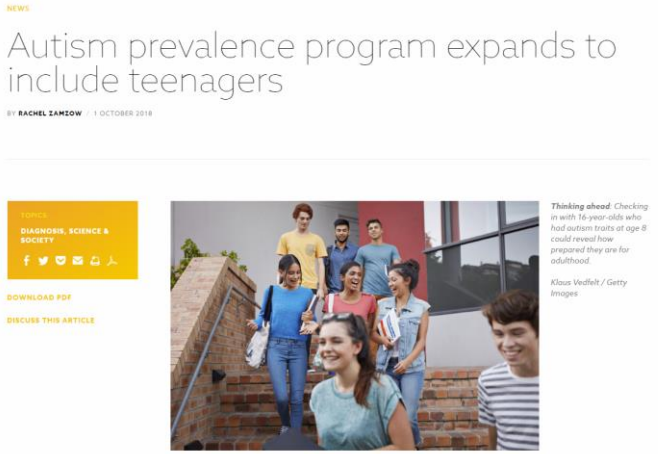
My perspective changed with my job

Last year…



I gave a talk about a collaborative project using machine learning to potentially speed up autism surveillance.

This year…



Getting ready to launch next round of autism surveillance activities

https://www.surveillancerepository.org/rapid-classification-autism-public-health-surveillance
https://www.spectrumnews.org/news/autism-prevalence-program-expands-include-teenagers/

"It worked great in the lab..."

"What we want are *new* weapons - weapons totally different from any that have been employed before. Such weapons can be made […] I have replaced some of the older scientists with young men and have directed research into several unexplored fields which show great promise. I believe, in fact, that a revolution in warfare may soon be upon us.“

-Professor-General Norden

[*Superiority* by Arthur C Clarke]

http://www.mayofamily.com/RLM/txt_Clarke_Superiority.html

The New York Times

Opinion

Artificial Intelligence Hits the Barrier of Meaning

Machine learning algorithms don't yet understand things the way humans do — with sometimes disastrous consequences.

By Melanie Mitchell
Ms. Mitchell is Professor of Computer Science at Portland State University.

Nov. 5, 2018

https://www.nytimes.com/2018/11/05/opinion/artificial-intelligence-machine-learning.html

"The most important priority for public health … genomics is to **be the honest broker to inform** providers, the public, and policymakers **whether** the deployment of a particular technology for a particular intended use can have a net positive health impact on the population."

Khoury MJ, Bowen MS, Burke W, et al. Current priorities for public health practice in addressing the role of human genomics in improving population health. *Am J Prev Med*. 2011;40(4):486-93.

Do we have an “honest broker” for machine learning?

A general answer to all the submitted questions:

- It depends

Will focus on

- Recommendations on what to do if you are just getting started
- Things that we learned are important, aside from machine learning

This is not a technical talk, but we will talk about tools

What are the goal(s) of using these methods (e.g., to develop a syndrome initially, to do ongoing surveillance, to improve existing keyword syndromes, other)?

- A collection of methods with different purposes, including
  - [**Supervised learning**] Classify case status from (labelled) data
    - E.g., using words in medical record, does child meet autism case def?
  - [**Information retrieval**] Extract data from unstructured text
    - E.g., what was the reported blood pressure during last office visit?
  - [**Unsupervised learning**] Represent patterns and relationships in data, without regard to a particular outcome
    - E.g., most similar patients, ICD codes, words…

What tools are used? CoT? R? Python? Proprietary SaaS?

- Recommendation: start with what you know – R and Python are good starting points
  - **SAS** … has SAS Textminer and **PROC HPFOREST** – I don't know a lot about it; not very popular for machine learning
  - **R**: **caret** tries to streamline process for multiple methods. Also, **xgboost**. **text2vec** package is fast and good at manipulating text.
  - **Python**: **scikit** has lots of classifiers available; **gensim** and **fasttext** offer easy ways to use word vector embeddings. Also, **spacy,** which is designed for production tasks. Many new methods will come to python before R. Python also has interface to tensorflow.
  - **Cutting edge/deep learning**: **tensorflow**, pytorch, etc.

What methods are relevant?

- Potentially, any of them.
  - Usually, cannot know prospectively what will be "best" on a given data set.

- Practical answer: for supervised problems, you will likely see similar results across a handful of well-known methods.

Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?

Manuel Fernández-Delgado MANUEL.FERNANDEZ.DELGADO@USC.ES
Eva Cernadas EVA.CERNADAS@USC.ES
Senén Barro SENEN.BARRO@USC.ES

This paper doesn't address deep learning methods, but the point is still relevant.

A COMPARISON OF MACHINE LEARNING ALGORITHMS FOR THE SURVEILLANCE OF AUTISM SPECTRUM DISORDER

Scott H Lee[1], Matthew J Maenner[1], Charles M Heilig[1]

Comparison of:

Random Forests
Latent Semantic Analysis
Multinomial Naïve Bayes
Support Vector Machine (linear kernel)
Naïve Bayes – Support Vector Machine
Neural Network based on Fasttext method

Key points:

- Newer methods performed about the same as RF – 86% accuracy
- Estimated Bayes error rate is ~88%, suggesting limited room for improvement.

https://arxiv.org/abs/1804.06223

What methods are relevant? [cont'd]

- **A good baseline:** Train a model using Random Forests, Boosted Trees, or Support Vector Machines.
  - These methods use "bag of words" as input where each word/phrase in a text field, or each code, are represented as features in the model.
  - Can apply weights to the words (binary, counts, TF-IDF)

Sent 1: He avoided eye contact.
Sent 2: He made good eye contact.

Sent#	he	avoided	eye	contact	made	good	he_avoided
0001	1	1	1	1	0	0	1
0002	1	0	1	1	1	1	0
…							

Outcome
1
0
…

What methods are relevant? [cont'd]

Chief complaint classification with recurrent neural networks

Scott H Lee, Drew Levin, Pat Finley, Charles M Heilig

(Submitted on 19 May 2018 (v1), last revised 12 Jul 2018 (this version, v2))

CCS	Description	GRU	LSTM	MNB_{bi}	MNB_{uni}	SVM_{bi}	SVM_{uni}
660	Alcohol-related disorders	78.91	**79.10**	69.58	72.68	74.41	65.45
128	Asthma	68.39	**68.49**	65.61	64.05	64.54	63.15
251	Abdominal pain	53.64	**53.98**	44.39	48.72	51.70	42.01
134	Other upper respiratory disease	51.72	**51.99**	42.50	45.73	45.44	45.22
250	Nausea and vomiting	**41.81**	41.76	19.31	33.74	36.39	26.71
133	Other lower respiratory disease	**35.57**	37.38	29.71	29.27	22.53	26.17
7	Viral infection	25.68	**33.30**	21.41	23.53	22.67	11.48
242	Poisoning by other medications and drugs	22.18	**25.50**	15.76	16.00	22.82	19.78
123	Influenza	14.80	13.41	13.11	13.27	**15.37**	15.18

Table 4. F_1 scores of each model for select conditions, with the highest score in bold.

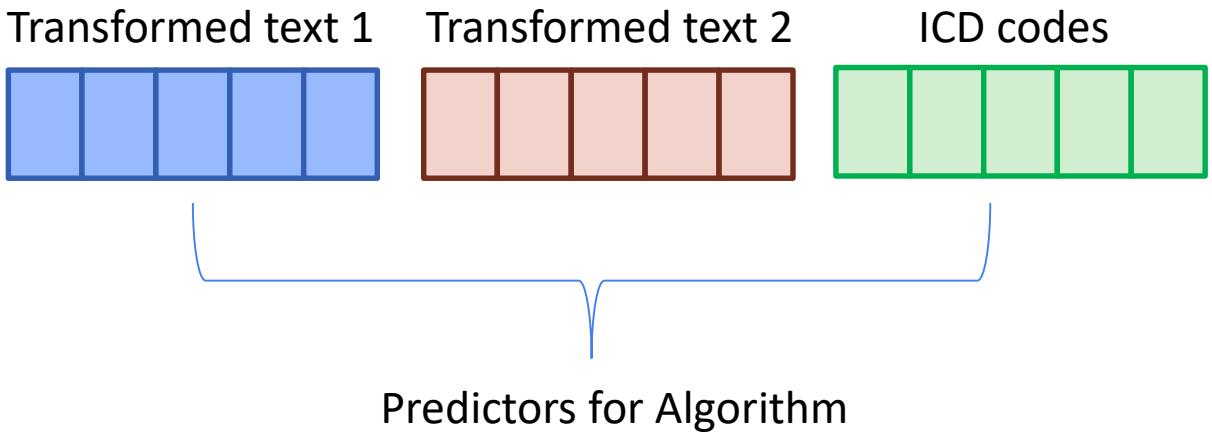
https://arxiv.org/abs/1805.07574

Some of the insights presented:

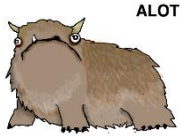
- Deep learning methods consistently perform ~5% better…
  - … so SVM could give a ballpark estimate of how more sophisticated methods may perform.
- A nice discussion of training a single model for multiple outcomes vs discrete models for each outcome

Can we take advantage of the contents of multiple fields?

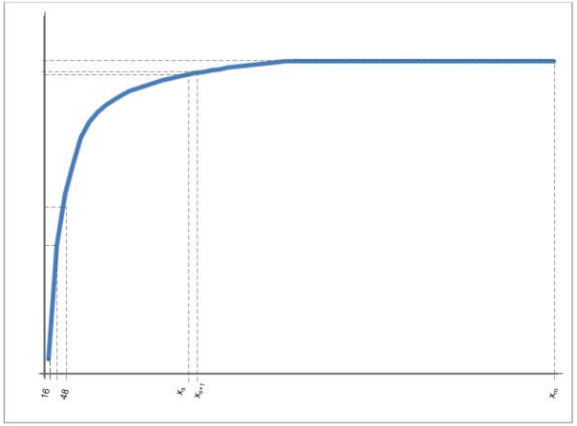
- YES!
- Combine the fields (or their transformed representations) as additional variables/features in the model:



How much data are required? I hear A LOT but how much is enough?



- More is better, but often diminishing returns



Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH. Predicting sample size required for classification performance. *BMC Med Inform Decis Mak*. 2012;12:8. Published 2012 Feb 15. doi:10.1186/1472-6947-12-8

- For autism classification, we got similar performance training on 1162 observations vs ~3000.
- More complex models likely need more data to reach full performance
  - Word vectors – millions of words (or more)
  - For sample sizes for "deep learning" models, look at benchmark datasets

https://hyperboleandahalf.blogspot.com/2010/04/alot-is-better-than-you-at-everything.html

What hardware is required? Do we need special servers/hardware?

- Can do a lot with a new-ish PC (multi-core CPU and 8-16GB RAM)
- Things that matter:
  - Efficiency of code
    - R < Python < compiled languages
    - R and Python can call external compiled libraries
  - Computational intensity of algorithm
    - Can use optimized algorithm (as with tSNE) or methods geared toward efficiency (such as fasttext)
  - Size of data
    - and whether it is read from disk or stored in RAM

What hardware is required? Do we need special servers/hardware? [cont'd]

- Usually several options to solve computational limits
  - E.g., R randomforest package was taking > 1 day to run and would run out of memory.
    - Option 1: (Python) scikit-learn random forest much more efficient
    - Option 2: use R, but use text2vec package and xgboost library (both are fast)
    - Option 3: reduce number of words/features that are in the model, by eliminating rare words/variables. Or, use a dense-vector representation to reduce dimensionality (paragraph vector, LDA).

What hardware is required? Do we need special servers/hardware? [cont'd]

- Consider hardware upgrades if
  - Dealing with lots of data in R [more RAM]
  - Want to use deep learning methods [powerful GPU]
  - You plan to put a model "in production" or will be doing intensive computations [workstation/server]

  … And you believe a more complex/intensive model will perform better

Does this process have to be done for each topic we want to address?

- Hopefully, you can re-use parts of the process
  - E.g., use data manipulation, cleaning, word vectors, etc., but run a new classifier for each condition
  - Will the algorithm classify mutually exclusive things?
    - E.g., a child has autism *and* ADHD
  - Will you ever want to "tweak" the algorithm for one topic without affecting the others? Or include data that is useful to only certain outcomes?

How much time is required to develop these methods?

- If you have the correct data in-hand, you can run an initial model very quickly.
  - E.g., let's train an algorithm to predict if Ira Glass is speaking using the first 596 transcripts from *This American Life*

text	speaker
When they came to V103, Max and Tony made some changes in personnel, they tinkered with the station's slogan	Ira Glass
A higher degree of discipline. And what, again, my experience has been, a short playlist always seems to fix the p	Tony Gray
When I went to visit V103, I had not been in a commercial radio station for 20 years. And I was curious about what	Ira Glass
Tony and Max and I are the same age-- 39 and 40. All three of us started in radio as teenagers Ask them what they l	Ira Glass
It happens every month. I can't sleep the night before the book comes out. I can't do it. I can't sleep. I can't do it. A	Max Myrick
Running a station the scientific way means, of course, that most radio stations sound the same. It's why most radic	Ira Glass
And one of the interesting things about V103 is that their format does include pockets of individuality, most notab	Ira Glass
So we kind of made a little noise, huh, Mr. Novak?	Tom Joyner
Indeed you did. You've got some power, I would say.	Robert Novak
No. No, no, the people, not me.	Tom Joyner
The people, OK.	Robert Novak
It's not me.	Tom Joyner
A few weeks ago, Joyner and his morning crew took on conservative columnist Robert Novak. This is while Preside	Ira Glass
Some of our really fine citizens are African-Americans-- in government, in business, athletics and show business. '	Robert Novak

~143k "utterances" – Ira was speaking 23% of the time

How much time does it take? [Cont'd]

Punctuation/ numbers	chose to remove both	gsub("[[:punct:]]") gsub("[[:digit:]]")
capitalization	chose lowercase	
stem (remove suffixes)	can help, especially with verbs	
ngrams	chose 1 and 2-word phrases	
remove sparse terms	minimum of 10 times	
word/ngram weights	choose: Binary (present/absent), Frequency counts, TF-IDF	

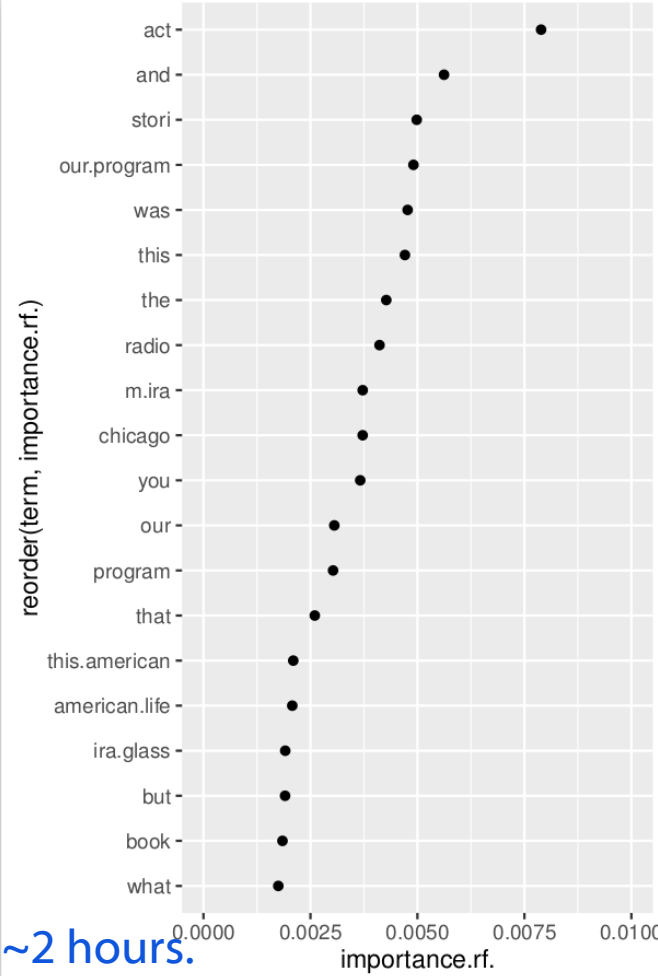
```
# all of this done with text2vec

sent_train = itoken(docs$sentence,
                    preprocessor = tolower,
                    tokenizer = stem_tokenizer,
                    progressbar = TRUE)
vocab = create_vocabulary(sent_train,
ngram=c(1L,2L))
pruned <- prune_vocabulary(vocab,
                           term_count_min = 10)
vectorizer = vocab_vectorizer(pruned)

tfidf= TfIdf$new()

dtm_train_tfidf <- fit_transform(dtm_train, tfidf)
```

^ not a lot of code to do all this

How much time does it take? [Cont'd]

```
rf <- ranger(formula = is_ira~., data=tdm_token_df,
                    num.trees=500, classification = TRUE,
                    importance="permutation",replace = TRUE,
                    seed = 123456)
```

Predictions	Reality: Ira	Reality: Not Ira
That's SO Ira!	8803	778
That's NO Ira!	17995	115720



Using 12 cores, trees took 20 minutes; permutation test took ~2 hours.

How much time does it take? [Cont'd]

All of these could affect the data and/or accuracy:

- Sparsity threshold (minimum of 10 words)
- Stem? If so, what algorithm?
- Ngrams? (uni, bi, tri)
- Lowercase?
- **Choice of classifier (RF, NB, SVM, etc.)**
- Use feature weights?
- Optimize classification rules for unbalanced classes?
  - Currently, just used a 50/50 split
- Add "handcrafted rules" using knowledge to help the algorithm?

Would different choices lead to even better accuracy?*

*this can lead to madness, and this is where you might spend all your time

Can ML* (or other methods) help to streamline/automate some of the iterative process required to initially develop a keyword-based syndrome?

- Yes. These can be great exploratory data tools and will likely show you something interesting, even if they don't outperform current methods:
  - Importance of features / variables
  - Relationships between data elements or observations
  - Ideas for rules/keywords

*machine learning

Unsupervised methods for exploratory data analysis

Start with all MMWRs (1982-2016, weekly and SS)

- 9,576 articles (through Apr 2016),~80MB of text
- 491k sentences
- 22M tokens (words and punctuation symbols)
- python / gensim

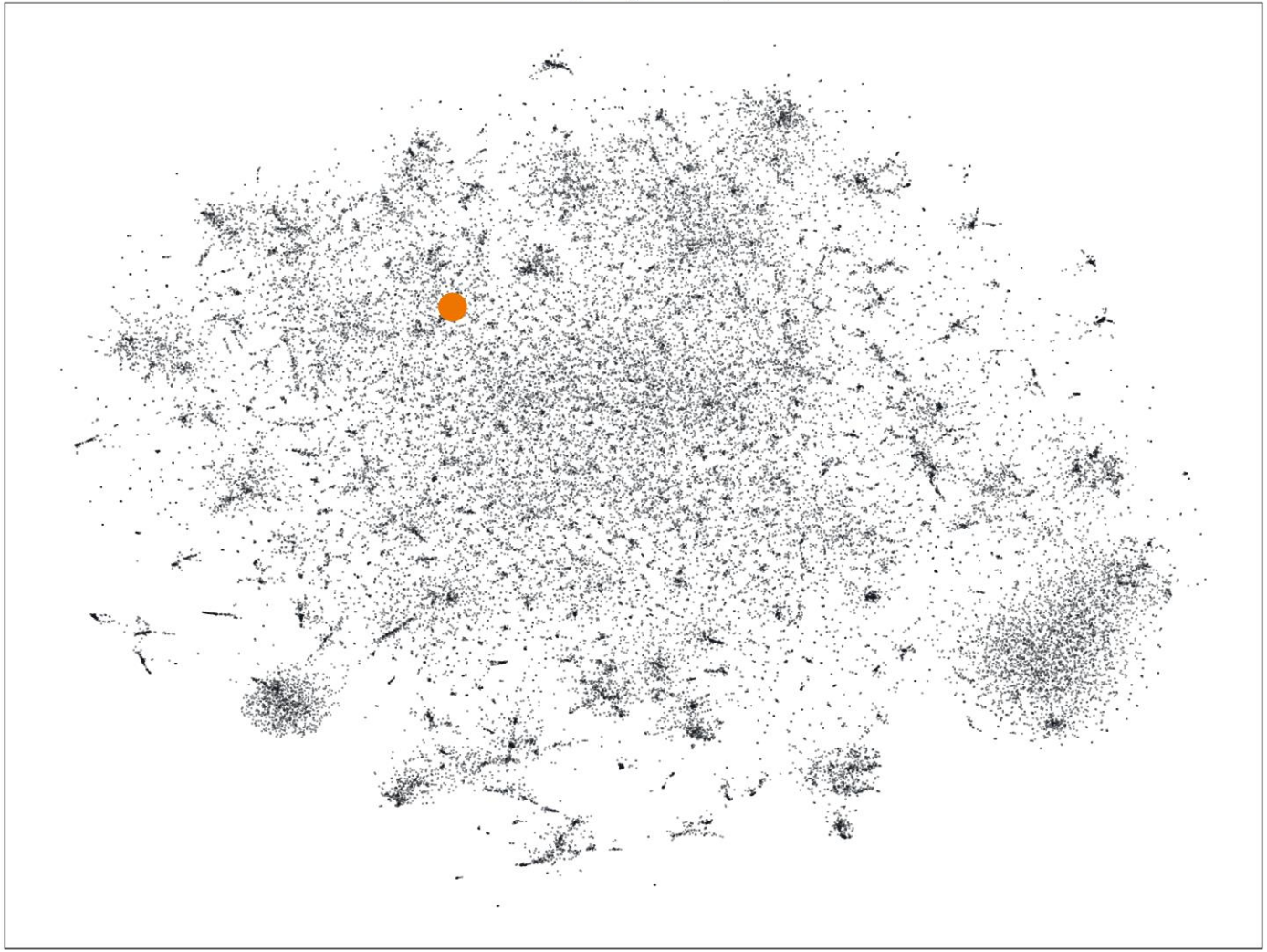
- **train** word vectors on sentences. [word2vec / gensim / python]
  - tokenized, keep punctuation, pad begin/end
  - 300d, SGNS (10 neg samples), 24 epochs
  - ~ 45 minutes (200k words per second)

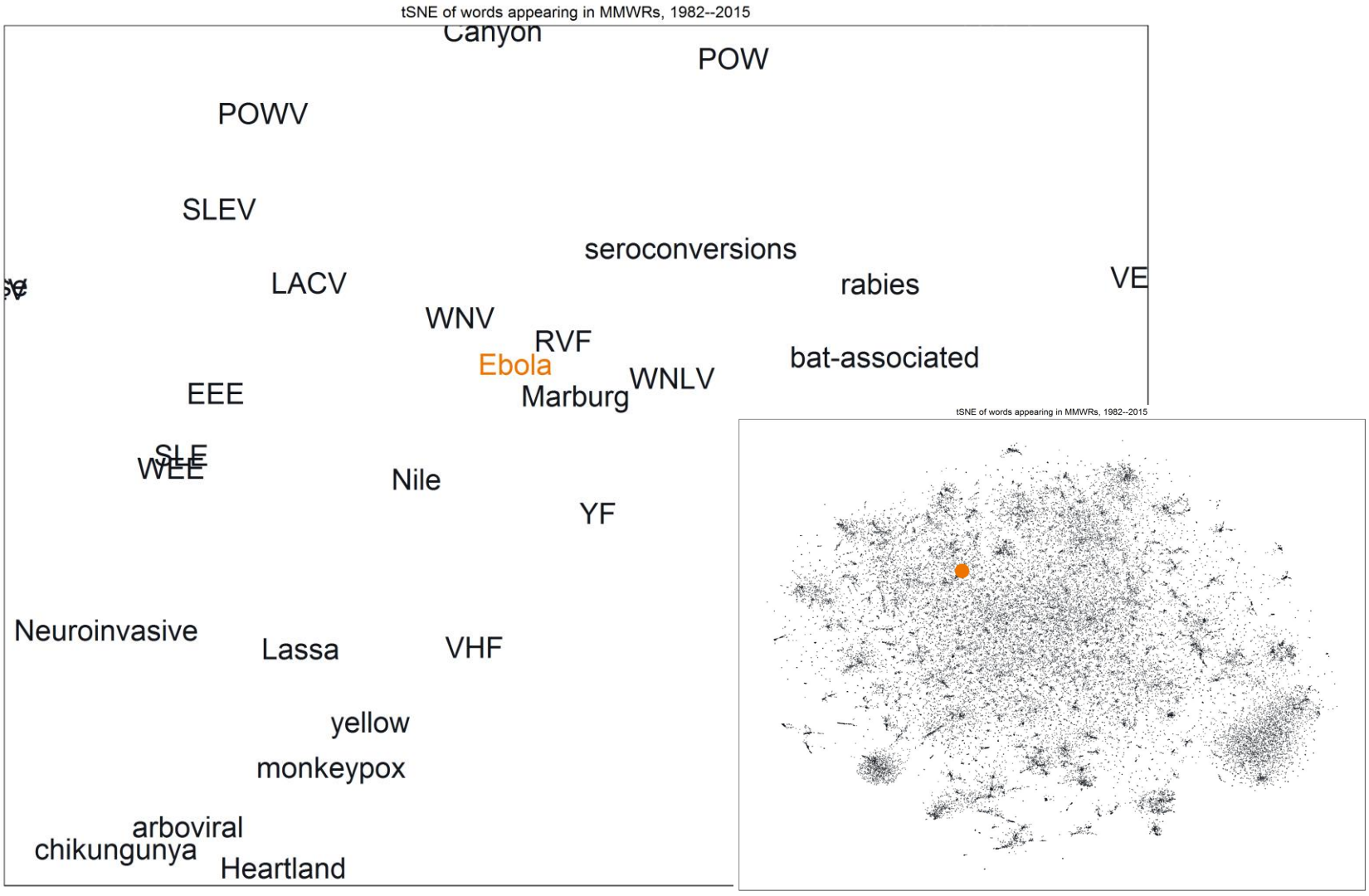
Unsupervised methods for exploratory data analysis

```
>>> model['Ebola']
array([-0.14460348,  0.22440973, -0.00493282, -0.08833114,  0.0131678 ,
       -0.19822162,  0.02534309, -0.0668368 ,  0.03818744, -0.18109842,
       -0.41992345,  0.32917005,  0.01639159, -0.07962775, -0.28792953,
       -0.08893112, -0.15143138,  0.05854373,  0.06473679,  0.19741157,
        0.21824922,  0.0939331 ,  0.05588859,  0.11730633, -0.16790462,
       -0.10254665, -0.18680874,  0.1206033 ,  0.10025534, -0.04680029,
        0.01593073,  0.51376379,  0.45748442,  0.40389478,  0.21005039,
        0.35991672,  0.04849171,  0.1887261 ,  0.11572687,  0.19430879,
        0.13049534, -0.29265872,  0.14755349, -0.4181219 ,  0.11367325,
        0.03661683,  0.23251511, -0.32505658, -0.02226758, -0.21435446,
       -0.12629834, -0.24388364,  0.38983586,  0.12900235, -0.0615279 ,
        0.01633649,  0.26458219, -0.20829263,  0.08852377, -0.26943061,
        0.11430427, -0.18097813, -0.30034429, -0.34245819, -0.03131076,
       -0.19752358,  0.36804938,  0.11156882,  0.16858512,  0.07501094,
        0.34935117,  0.17049626, -0.26456356,  0.16339448, -0.11351007,
       -0.29486787,  0.08876017,  0.01470964,  0.09294634,  0.1701287 ,
        0.23453704,  0.07594802, -0.3196139 ,  0.26534384, -0.44254655,
        0.00987968, -0.1357805 ,  0.12750803, -0.02711842, -0.00833873,
        0.3136344 ,  0.13502555,  0.16862279, -0.1486944 , -0.0024121 ,
        0.23119669,  0.07498727, -0.10351934,  0.19400899,  0.07872754,
       -0.06845266, -0.12752971, -0.15465078,  0.23827283,  0.07808877,
       -0.10813881, -0.13198243,  0.03134936, -0.31827468,  0.04945973,
        0.07361871,  0.55537355,  0.10771009, -0.43416414,  0.12181319,
       -0.41696385,  0.07888453, -0.55023706, -0.25540403, -0.14259575,
        0.17780991,  0.13589111,  0.16771944, -0.05802483,  0.16589975,
        0.3033174 , -0.14670326, -0.1092925 , -0.1804069 , -0.15395285,
        0.11851311, -0.12359003,  0.00323989, -0.32766417, -0.03804096,
       -0.16791679,  0.13432926, -0.04799558, -0.01985986,  0.03814695,
        0.02397371,  0.08304751, -0.00849702,  0.08110802,  0.26945314,
       -0.18982057,  0.11722723,  0.08636708,  0.28872421,  0.10633576,
       -0.04036408,  0.22189879, -0.28164443, -0.062053 ,  0.1569656 ,
        0.16907023, -0.43023589,  0.03781117, -0.1042143 ,  0.08678224,
       -0.03095428, -0.04490658,  0.09076487, -0.01483379,  0.1673985 ,
       -0.15804893, -0.10696812,  0.28772974, -0.18464214, -0.14182086,
        0.15284358,  0.13060325, -0.07196767, -0.18290038,  0.06439415,
       -0.18806683,  0.0175377 ,  0.22852072,  0.0434635 , -0.0535018 ,
        0.36121598,  0.31143379, -0.1191148 ,  0.17147142,  0.13983424,
        0.00119175, -0.01417412, -0.11025971, -0.13450806,  0.10632839,
        0.06920454, -0.00469677,  0.26147556,  0.09659825, -0.07432177,
        0.15252161, -0.30333135,  0.06334595,  0.11064886,  0.2156457 ,
        0.10056917,  0.53568804,  0.11798401,  0.49144864, -0.40535948,
        0.00301202, -0.08856452,  0.1318565 , -0.0977622 ,  0.07621381,
       -0.05712342,  0.07297543,  0.14582643, -0.32511228,  0.16298187,
        0.33866856, -0.04855051,  0.20587687,  0.27307385, -0.26193365,
       -0.01421137, -0.29414845, -0.30652016,  0.03621313, -0.11553205,
       -0.0331264 , -0.07681023,  0.02867636,  0.28418851,  0.26406464,
       -0.07401082, -0.09791714,  0.00257847,  0.04077858, -0.07154669,
       -0.12416645, -0.42704177, -0.04683092, -0.17901744, -0.26264372,
       -0.07537364,  0.16165955,  0.0011476 ,  0.01473377, -0.34305459,
        0.07362207,  0.02701855,  0.19833933,  0.04008583, -0.13152441,
        0.01620991,  0.04406546,  0.08561434, -0.04365898, -0.01782265,
        0.05294642,  0.00978605,  0.30910796,  0.09726601, -0.13310005,
        0.19718421,  0.10914054,  0.02821998,  0.24906589, -0.10399321,
       -0.33096734, -0.07929723, -0.02719171,  0.04258193, -0.41672957,
       -0.29224831,  0.32875165,  0.00236925,  0.19716987,  0.27938706,
        0.27162772,  0.14347738,  0.11451535,  0.36522067, -0.1076664 ,
       -0.30282694, -0.23203145, -0.11529484, -0.1467851 , -0.26459244,
        0.09815317,  0.04802833, -0.39088652, -0.38216135, -0.43946764,
       -0.12977573, -0.38795078, -0.1800321 ,  0.48116568,  0.08991092,
        0.01640161,  0.00849182,  0.02570708, -0.03093085,  0.04272429], dtype=float32)
```

- **estimate** similarity with cosine distance
- **visualize** w/ t-SNE
  - rtsne; graphs made in ggplot2

tSNE of words appearing in MMWRs, 1982--2015




tSNE of words appearing in MMWRs, 1982--2015
Canyon
POW
POWV
SLEV
seroconversions
LACV
rabies
VE
WNV
RVF
Ebola
bat-associated
WNLV
EEE
Marburg
SLE
WEE
Nile
YF
Neuroinvasive
Lassa
VHF
yellow
monkeypox
arboviral
chikungunya
Heartland
tSNE of words appearing in MMWRs, 1982--2015

Cosine distance - similarity scores

"Ebola"

Word	Similarity score
RVG	.803
Jamestown_Canyon	.798
Marburg_virus	.793
VHF	.790
CCHF	.786
POW	.786
BHF	.786
Zaire	.783
hemorrhagic_fever	.782
Zika	.782

"investigation"

Word	Similarity score
epidemiologic_investigation	.842
investigations	.831
inquiry	.789
traceback_investigations	.788
evaluation	.788
entomologic	.778
epidemiologic_investigations	.775
assessment	.771
audit	.769
outbreak-control_measures	.768

Similarity scores are unsupervised

"obesity"

Similar Words
overweight
physical_inactivity
hypertension
diabetes
sedentary_lifestyle
cardiovascular_disease
cigarette_smoking
blood_pressure
cholesterol
chronic_conditions

- synonyms / similar conditions
- causes of obesity
- co-occurring health conditions
- other risk factors / behaviors

Could be difficult to distinguish important nuances

Possible to focus on specific relationships:

```
>>> model.doesnt_match(
["physical_inactivity", "sedentary_lifestyle",
“overweight"])

‘overweight'
```

Can ML (or other methods) help to streamline/automate some of the iterative process required to initially develop a keyword-based syndrome? [Cont'd]

- CDC does cerebral palsy surveillance -- reviews text from medical evaluations and ICD codes.
  - Trained a random forest model to find out which ICD codes were most useful in predicting cerebral palsy (CP).
    - The only informative codes were those normally associated with CP
    - Tried to simplify from a random forest to a rule, evaluated options:
      - number of relevant CP ICD codes
      - Presence of any CP ICD code
  - Lead to 2019 ISDS abstract: Comparing Cerebral Palsy Surveillance Definition to ICD Codes and Written Diagnoses – Sarabeth Mathis et al.

Are there certain topics where ML might perform better than manual syndromes than others?

- Hard to predict, but some thoughts:
  - How complex is the manual syndrome? (e.g., are there 2 rules or 1000)?
  - If topic uses a variety of spellings / synonyms that are difficult to define manually, some ML methods could pick this up.
  - If outcome can be predicted from contextual information that was learned on another dataset:
    - 2019 ISDS abstract: A machine learning algorithm to identify persons with chronic hepatitis C infection in health insurance claims data. Khan et al.

Are there certain topics where ML might perform better than manual syndromes than others? [cont'd]

Algorithm using data from case reports outperformed standard measure using ICD-9 codes.

But doesn't say how it compares to simple keywords for "VTE" or "venous thromboembolism"?

Med Care. 2018 Sep;56(9):e54-e60. doi: 10.1097/MLR.0000000000000831.

Improved Identification of Venous Thromboembolism From Electronic Medical Records Using a Novel Information Extraction Software Platform.

Dantes RB[1], Zheng S[2], Lu JJ[3], Beckman MG[4], Krishnaswamy A[5], Richardson LC[6], Chernetsky-Tejedor S[1,2], Wang F[2].

⊕ Author information

Abstract

INTRODUCTION: The United States federally mandated reporting of venous thromboembolism (VTE), defined by Agency for Healthcare Research & Quality Patient Safety Indicator 12 (AHRQ PSI-12), is based on administrative data, the accuracy of which has not been consistently demonstrated. We used IDEAL-X, a novel information extraction software system, to identify VTE from electronic medical records and evaluated its accuracy.

METHODS: Medical records for 13,248 patients admitted to an orthopedic specialty hospital from 2009 to 2014 were reviewed. Patient encounters were defined as a hospital admission where both surgery (of the spine, hip, or knee) and a radiology diagnostic study that could detect VTE was performed. Radiology reports were both manually reviewed by a physician and analyzed by IDEAL-X.

RESULTS: Among 2083 radiology reports, IDEAL-X correctly identified 176/181 VTE events, achieving a sensitivity of 97.2% [95% confidence interval (CI), 93.7%-99.1%] and specificity of 99.3% (95% CI, 98.9%-99.7%) when compared with manual review. Among 422 surgical encounters with diagnostic radiographic studies for VTE, IDEAL-X correctly identified 41 of 42 VTE events, achieving a sensitivity of 97.6% (95% CI, 87.4%-99.6%) and specificity of 99.8% (95% CI, 98.7%-100.0%). The performance surpassed that of AHRQ PSI-12, which had a sensitivity of 92.9% (95% CI, 80.5%-98.4%) and specificity of 92.9% (95% CI, 89.8%-95.3%), though only the difference in specificity was statistically significant (P<0.01).

CONCLUSION: IDEAL-X, a novel information extraction software system, identified VTE from radiology reports with high accuracy, with specificity surpassing AHRQ PSI-12. IDEAL-X could potentially improve detection and surveillance of many medical conditions from free text of electronic medical records.

PMID: 29087984 PMCID: PMC5927846 [Available on 2019-09-01] DOI: 10.1097/MLR.0000000000000831

https://www.ncbi.nlm.nih.gov/pubmed/29087984

If ongoing surveillance is the goal, how would we mechanize this to be done in an automated fashion?

Another area that could consume all your time

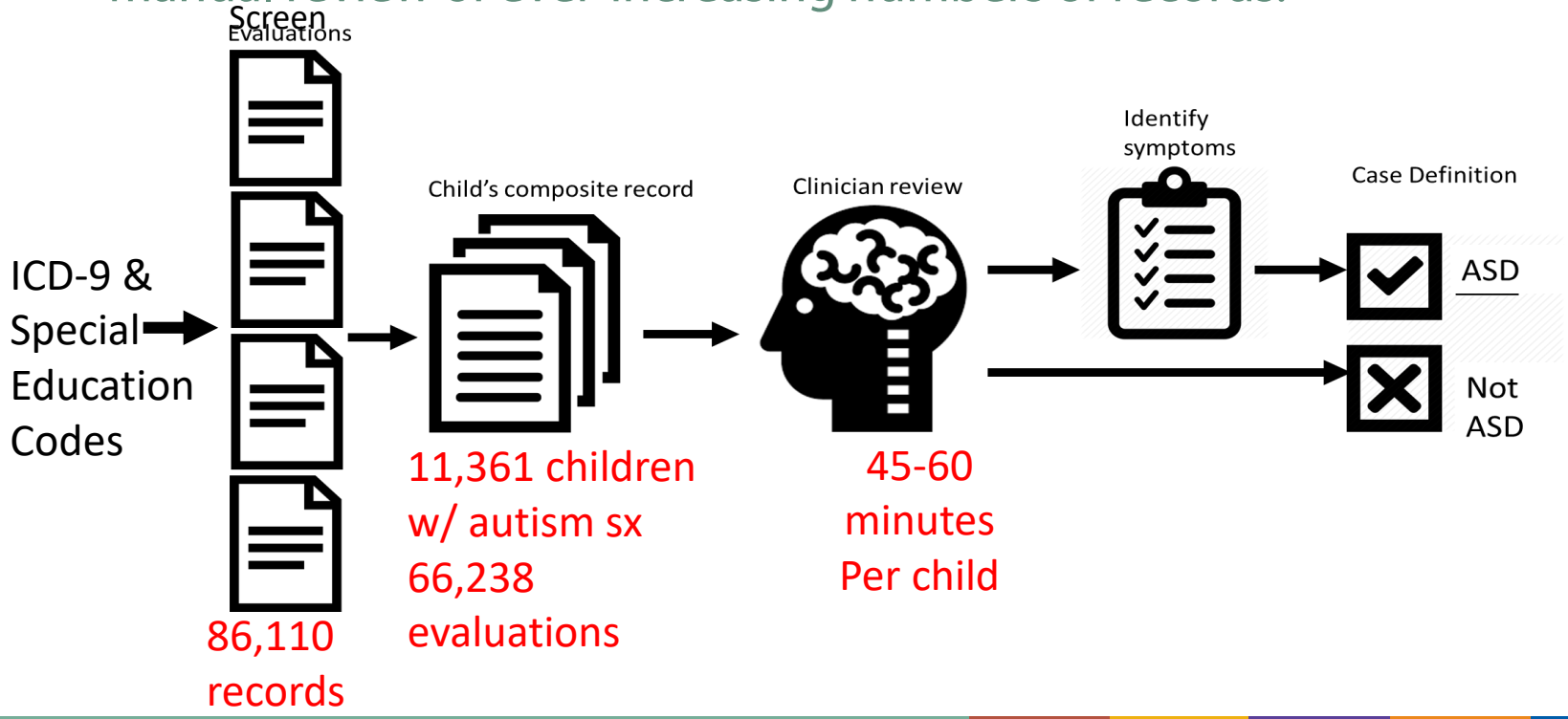
- Entire process needs to be standardized
  - Data transformation, manipulation, inputs into algorithm
- Need to make decisions about how it will work:
  - If it will run on data from multiple sites/states, will the training data reflect all the sites?
  - If each site/state has to run on their own data, they need to do everything exactly the same as training data.
  - How often will the algorithm need to be validated against a "gold standard" or retrained on new data? Does it matter if the algorithm changes over time?

Is the juice worth the squeeze? How much might it improve a syndrome over what we can develop using traditional methods? If there is improvement, how much improvement do we need to make the initial set-up work worth it?

- The most important question(s)
- Gather information to compare methods on accuracy, efficiency, timeliness, etc.
- Would you lose anything by switching to ML? Will people accept it?
- Evaluate the amount of resources / infrastructure needed to maintain this model

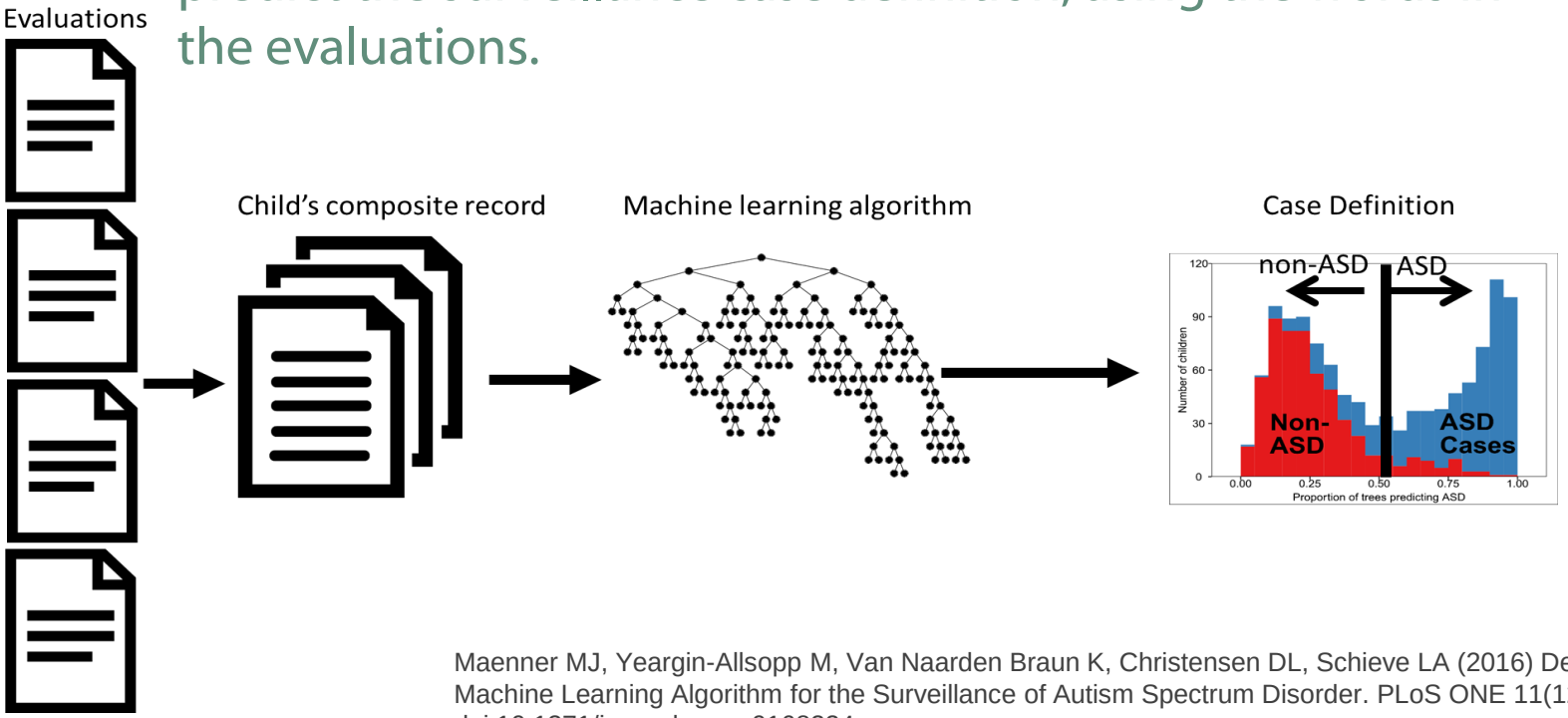
Is ML worth it? [Cont'd]

CDC's population-based autism surveillance requires the manual review of ever-increasing numbers of records.



Is ML worth it? [Cont'd]

To potentially improve efficiency, we had an algorithm predict the surveillance case definition, using the words in the evaluations.

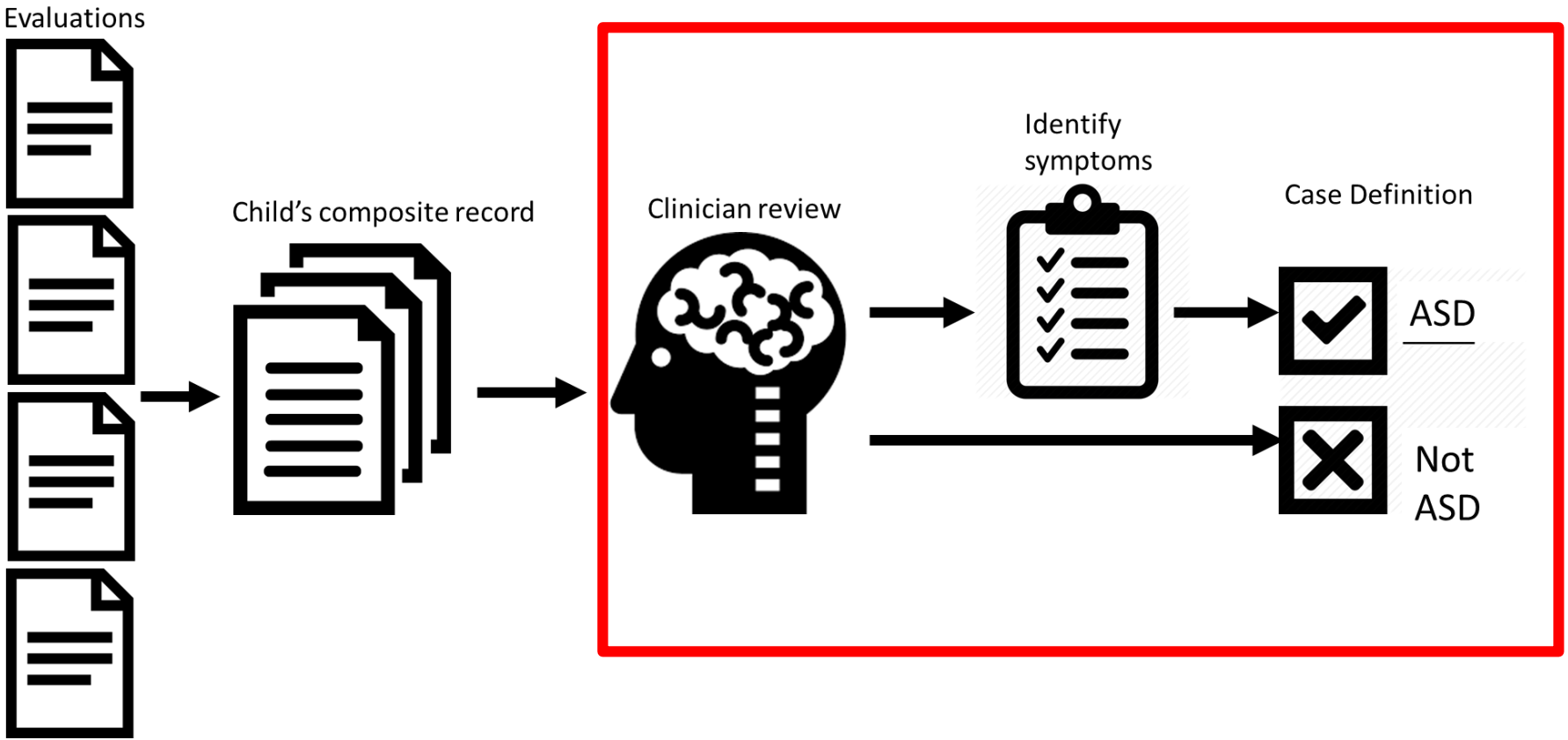


Maenner MJ, Yeargin-Allsopp M, Van Naarden Braun K, Christensen DL, Schieve LA (2016) Development of a Machine Learning Algorithm for the Surveillance of Autism Spectrum Disorder. PLoS ONE 11(12): e0168224. doi:10.1371/journal.pone.0168224

Algorithm-derived ASD “prevalence” per 1,000 kids

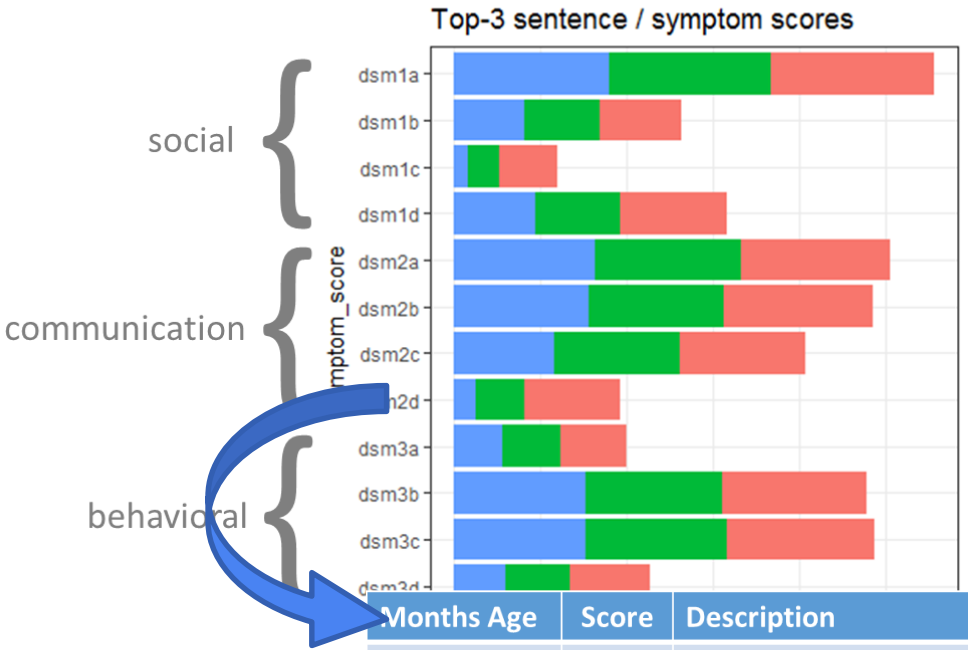
Group	Published		Algorithm-based		Ratio
Overall	15.5	(14.5-16.7)	14.6	(13.6-15.7)	0.94

Agrees w/ clinician	91%	87%
Time needed to review	Approx 1200 hours	Approx 1 second


Evaluations
Child's composite record
Clinician review
Identify symptoms
Case Definition
ASD
Not ASD

Providing additional information

CHILD GAXXXXXXX



Surveillance ASD Prediction: 0.98

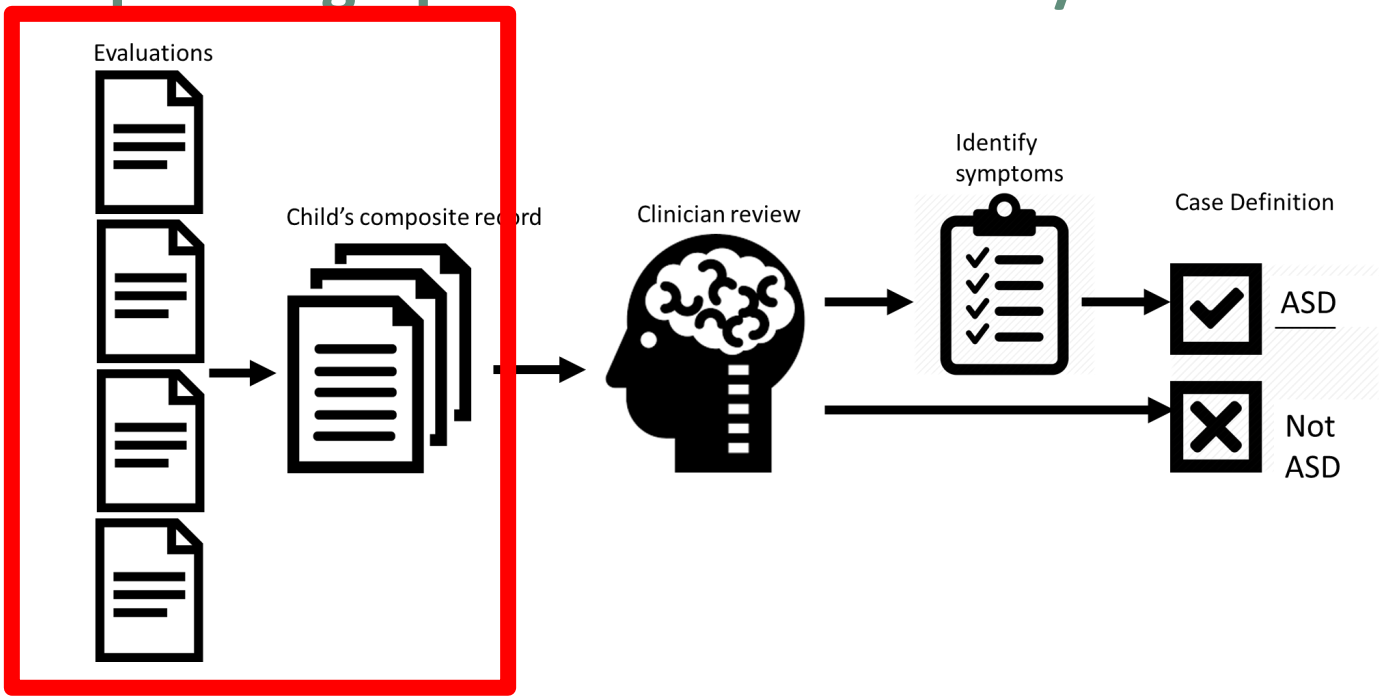
ASD diagnosis in record: 0.91
ASD ICD-9/10 code: NO
ASD Special Education: YES

ADHD symptoms: 0.93
unusual sensory responses: 0.67
Temper tantrums: 0.94

Number of evaluations / sentences: 16 / 578
Evaluations from: Education and Health

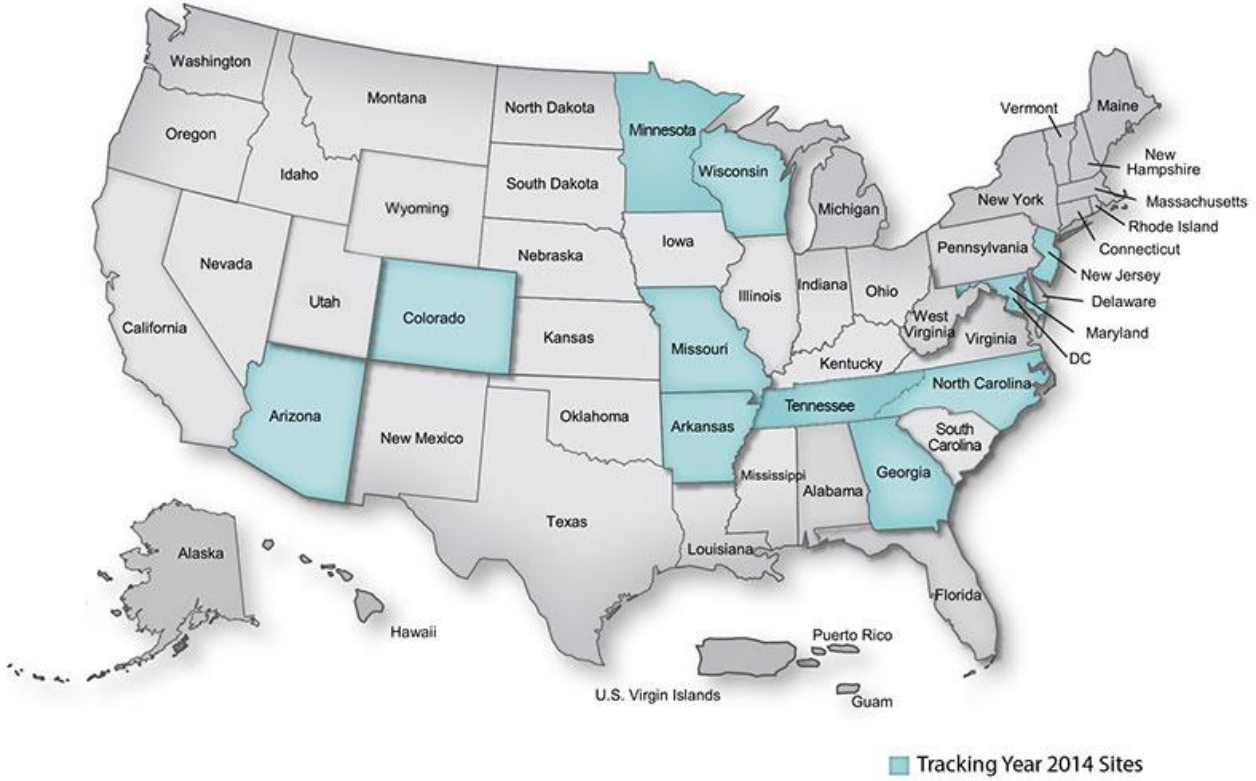
Months Age	Score	Description
66	56%	Addtl behavioral information: No concept of strangers or wandering away not much pretend play mostly plays alone
66	28%	He especially struggled with tasks which required him to pretend how to do something lack of symbolic play tended to just label pictured items in a very concrete manner
51	13%	You have noted some pretend play with certain toys

On speeding up record abstraction / initial screening



- The initial review of records is done manually, and takes a lot of time
- Records exist in various electronic and paper-based systems

Autism and Developmental Disabilities Monitoring (ADDM) Network Sites



"….taking your methods and looking for a problem is not the way to go about making a serious contribution to health in populations, which is what we as epidemiologists should be about."

…

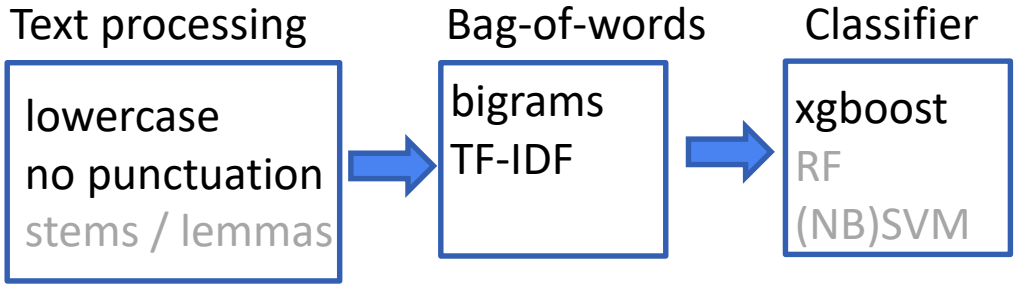
"Do not be governed entirely by your armamentarium, although one must stay within one's capacities. Choose the problem, a health problem of some sort."

-Mervyn Susser

Paneth, *"A conversation with Mervyn Susser"*

Overall: worth exploring

Possible to quickly get a sense of performance using basic tools.



Decide next steps based on

- what is needed for system to accomplish its job
- a reasonable expectation of cost/benefit for more advanced methods
- non-ML factors: personnel, infrastructure, support, ongoing quality assurance, prediction vs interpretation

Acknowledgments
Chad Heilig (CSELS)
Scott Lee (CSELS)
Fatima Abdirizak (NCBDDD)
Nicole Dowling (DDB)
Maureen Durkin (UW-Madison)
Laura Schieve (DDB)
Daisy Christensen (DDB)
Kim Van Naarden Braun (DDB)
Marshalyn Yeargin-Allsopp (DDB)
Sarabeth Mathis

For more information, contact CDC
1-800-CDC-INFO (232-4636)
TTY: 1-888-232-6348 www.cdc.gov

Autism machine learning project was supported by:
HHS Secretary's Ventures Program
CDC Innovation Fund
NCBDDD Division of Congenital and Developmental Disorders

mmaenner@cdc.gov