# Precision Public Health Surveillance and Machine Learning

**Better Tools to Tackle Old and New Challenges?**


**Matthew Maenner, PhD**

Chief, Child Development and Disability Branch


CDC

Oak Ridge Institute for Science and Education

# Confessions of a former "machine learning boy wonder"

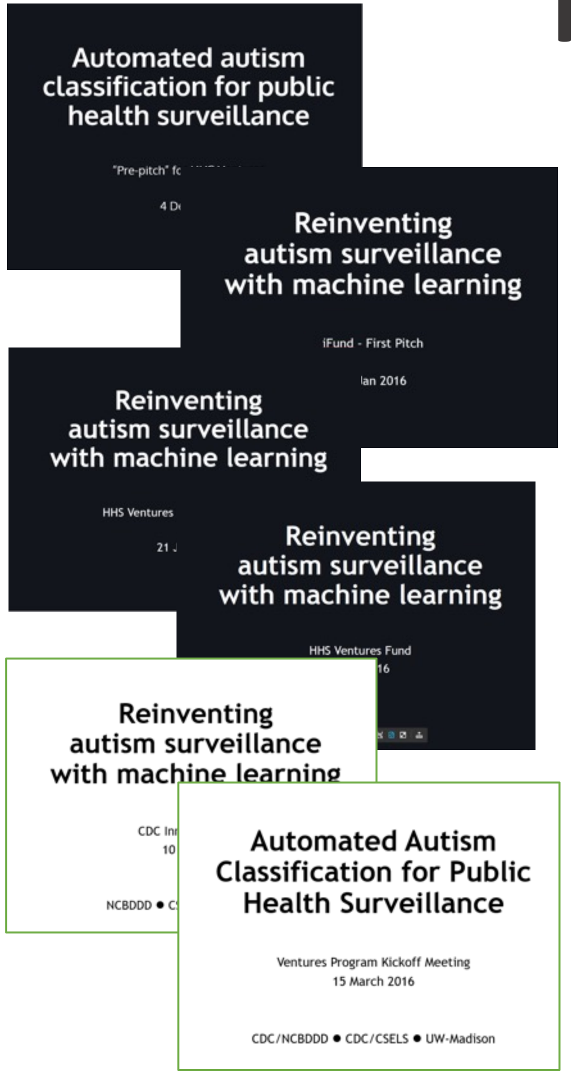
**Matthew Maenner, PhD**

Branch Chief
Child Development and Disability Branch

2023 February 8

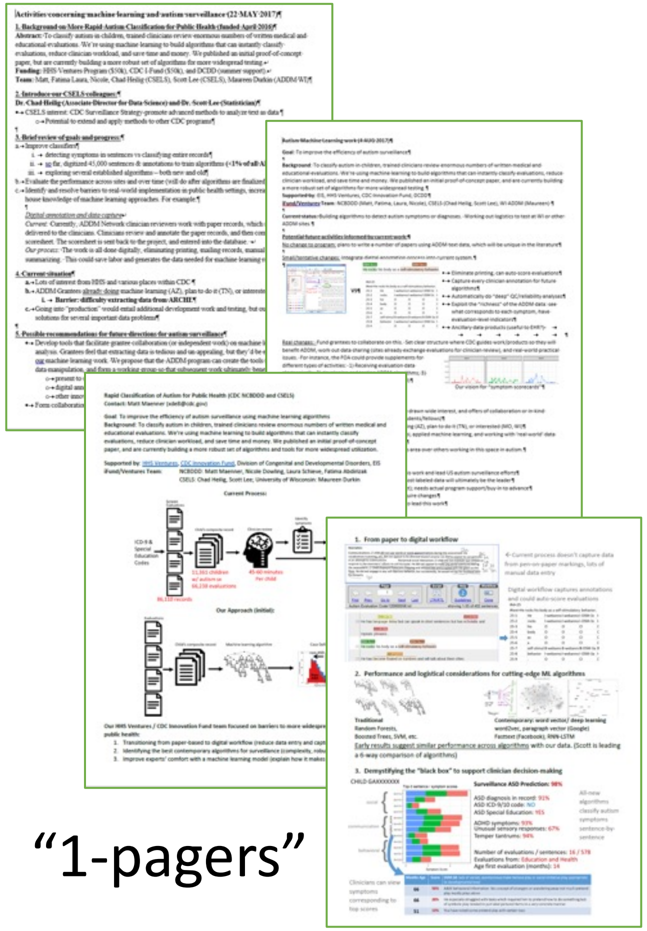
# I had a machine learning project

- Started as an EIS project
- HHS Ventures and CDC Innovation Fund project
- Influenced the updated autism surveillance system*




Pitches


Talks



“1-pagers”

*autism surveillance system does not use machine learning

# Do we have an "honest broker" for machine learning?*

"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."

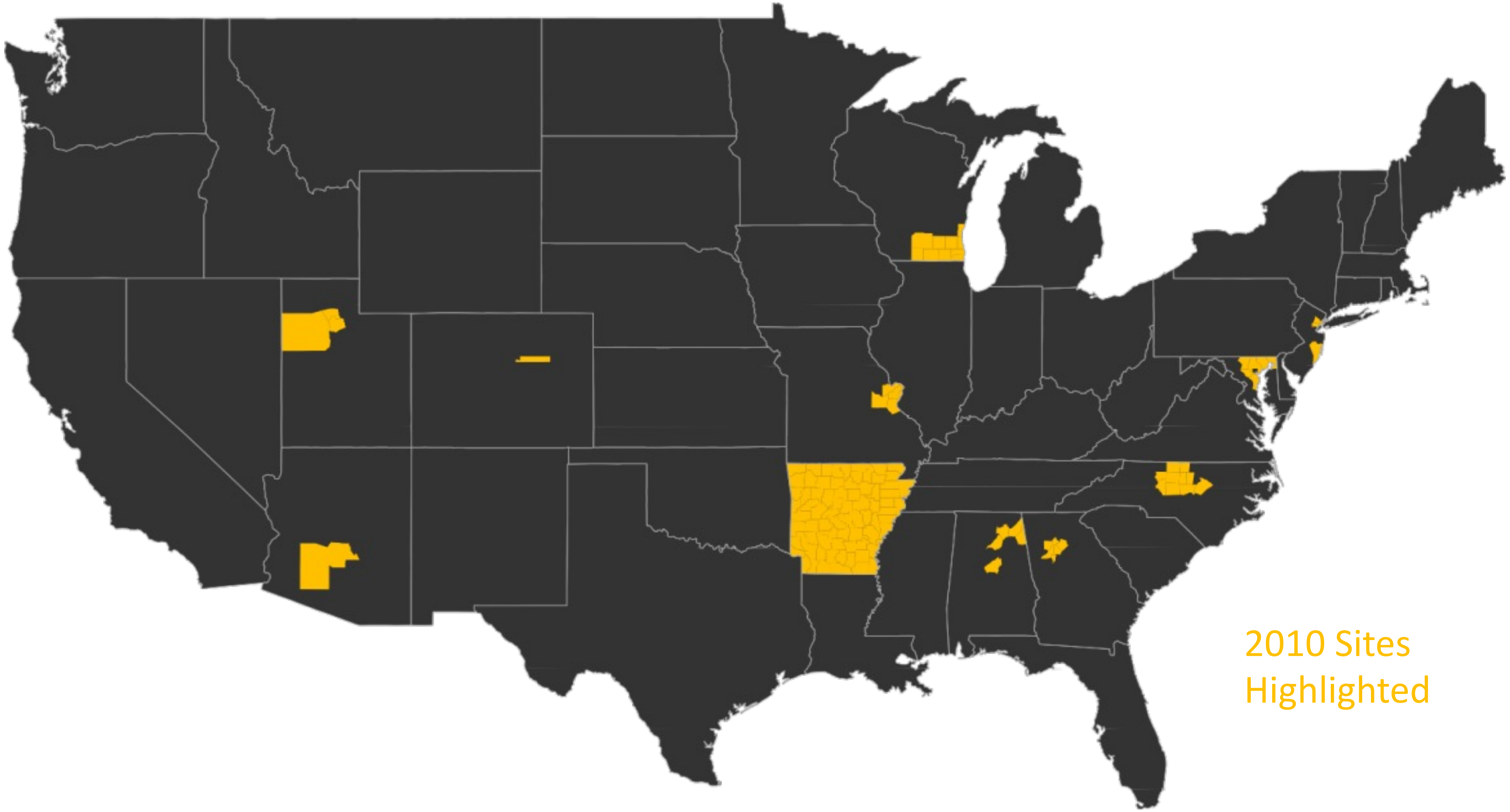
Khoury MJ, Bowen MS, Burke W, et al. Current priorities for public health practice in addressing the role of human genomics in improving population health. *Am J Prev Med*. 2011;40(4):486-93.

* see: https://en.wikipedia.org/wiki/Betteridge%27s_law_of_headlines

With the benefit of hindsight:

- Project aims
- How implementation is more complicated than research aims
- Judging utility of machine learning approaches in public health

# Autism and Developmental Disabilities Monitoring (ADDM) Network

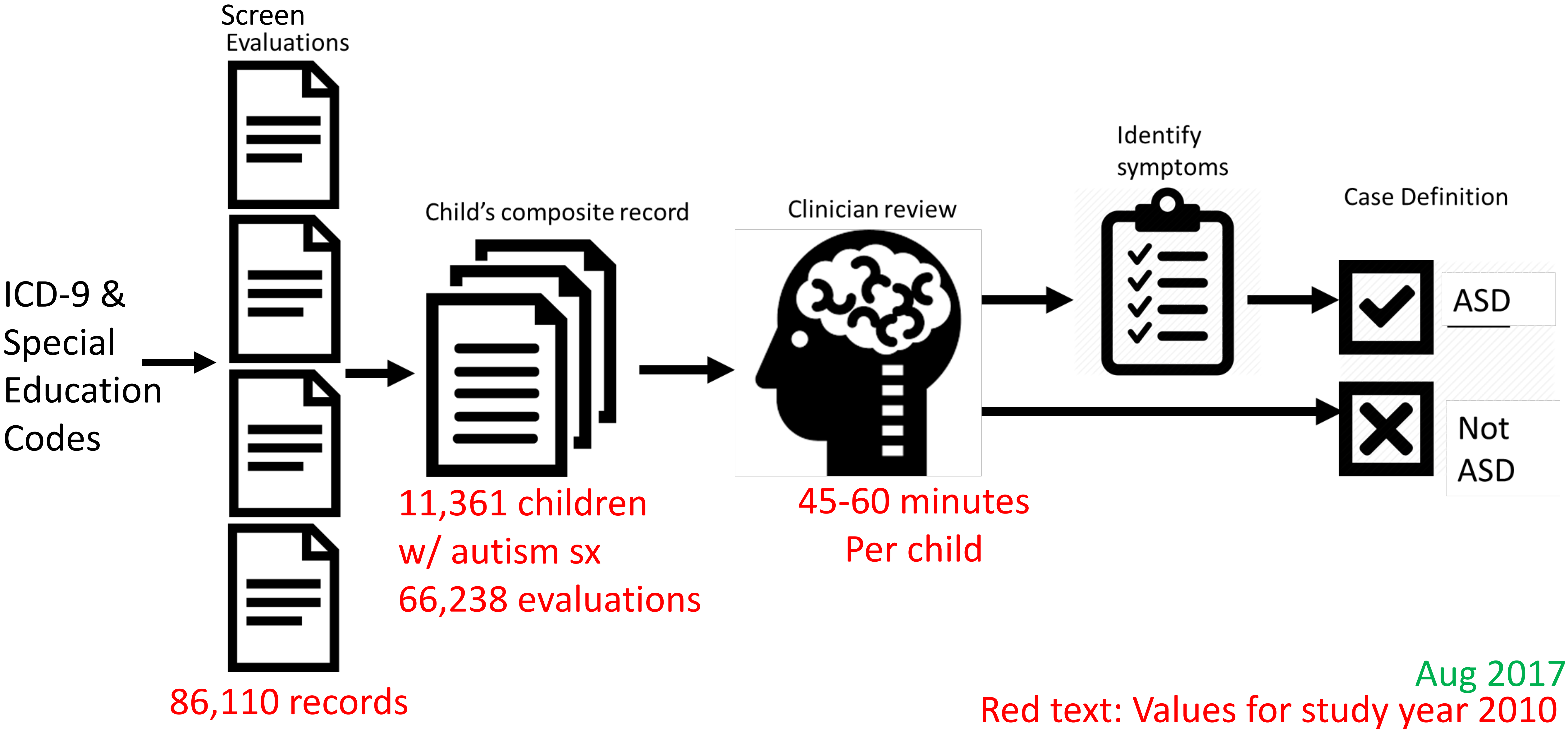


2010 Sites Highlighted

8-year-old children living in defined geographic areas
1-year period prevalence for even-numbered years beginning in 2000

May 2014

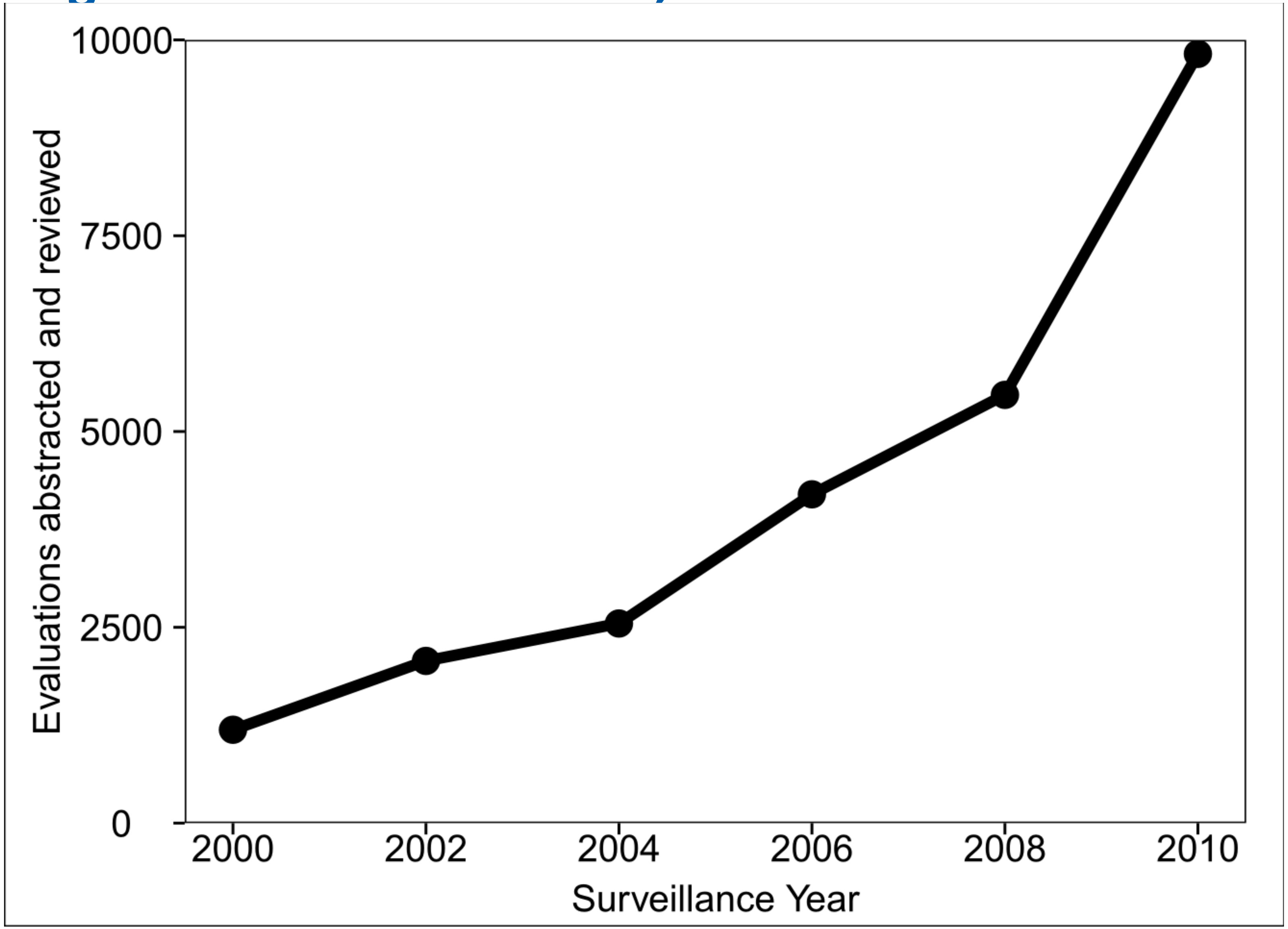
# CDC's population-based autism surveillance requires the manual review of ever-increasing numbers of records.



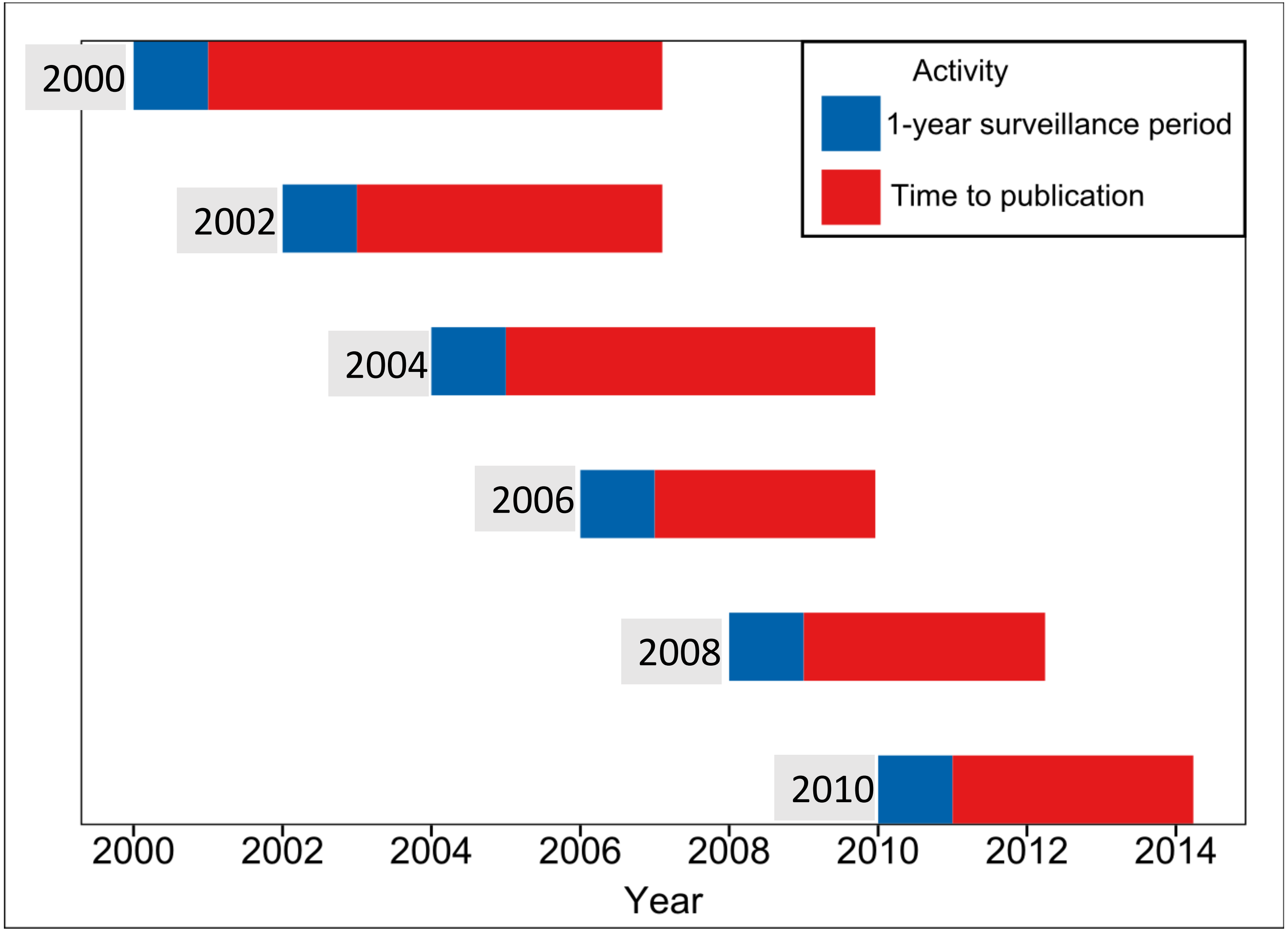

Aug 2017

Red text: Values for study year 2010

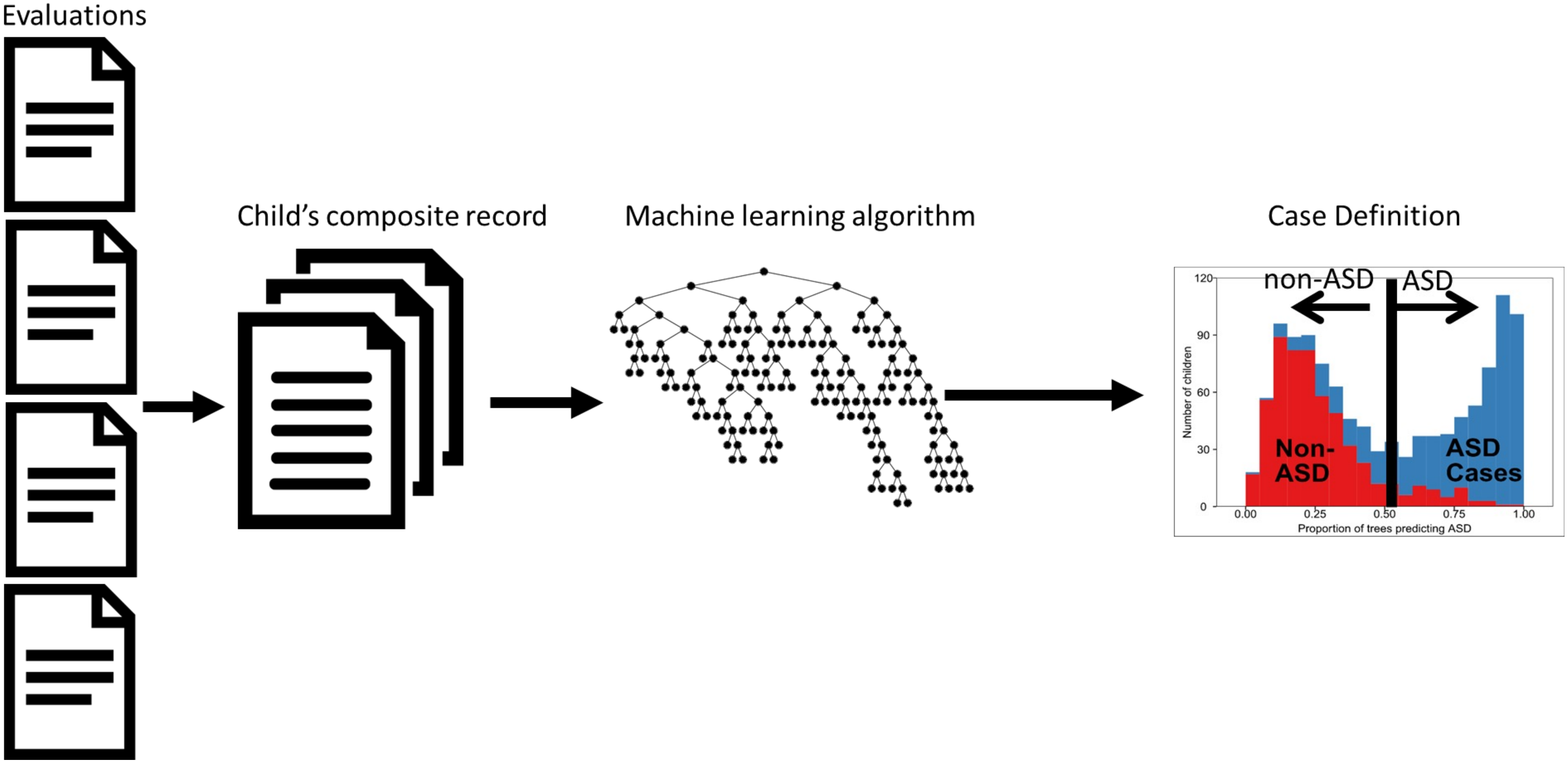
# Increasing number of ASD evaluations reviewed by Georgia ADDM Network site, 2000-2010



# Timeline of ADDM ASD surveillance reports



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



# Classification with random forests

Random Forests[1]

Training Data: 2008 Georgia ADDM site

- 1,162 children (601 met ASD case status)
- 5,396 evaluations
- 13,135 1-3 word phrases initially included
  - stemmed, tf-idf weighted, dropped rare words

Test Data: 2010 Georgia ADDM site

- 1,450 children (754 met ASD case status)
- 9,811 evaluations

Software: R (tm, RWeka, RandomForest)
Python (Scikit-learn, pandas)

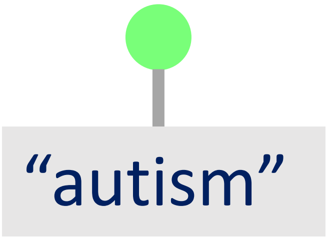
1. Breiman, 2001

# Random forests: training one tree

Sample N observations
<u>with replacement</u>
(leaves out ~37%, or 1/e)

2008 Training Data
N = 1,162 Children
K = 13,135 phrases

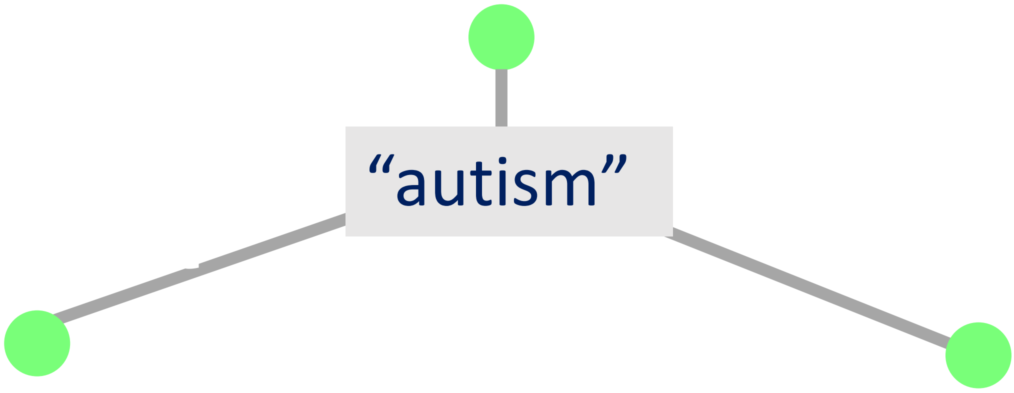
# Random forests: training one tree




2. Random subset of sqrt(K) words/phrases; choose term that best separates outcomes

2008 Training Data
N = 1,162 Children
K = 13,135 phrases

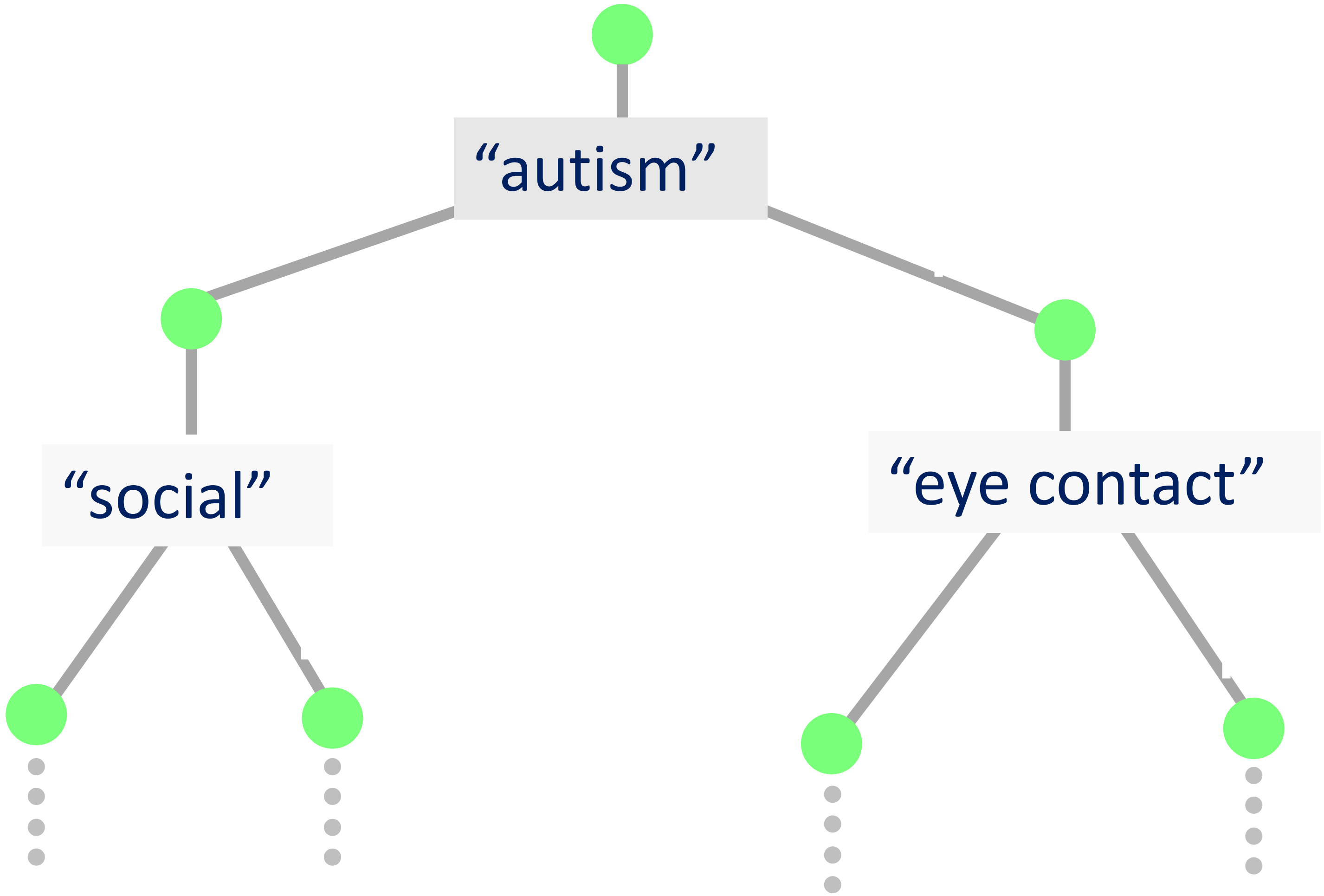
# Random forests: training one tree




3. **Split sample** using the values of the selected term

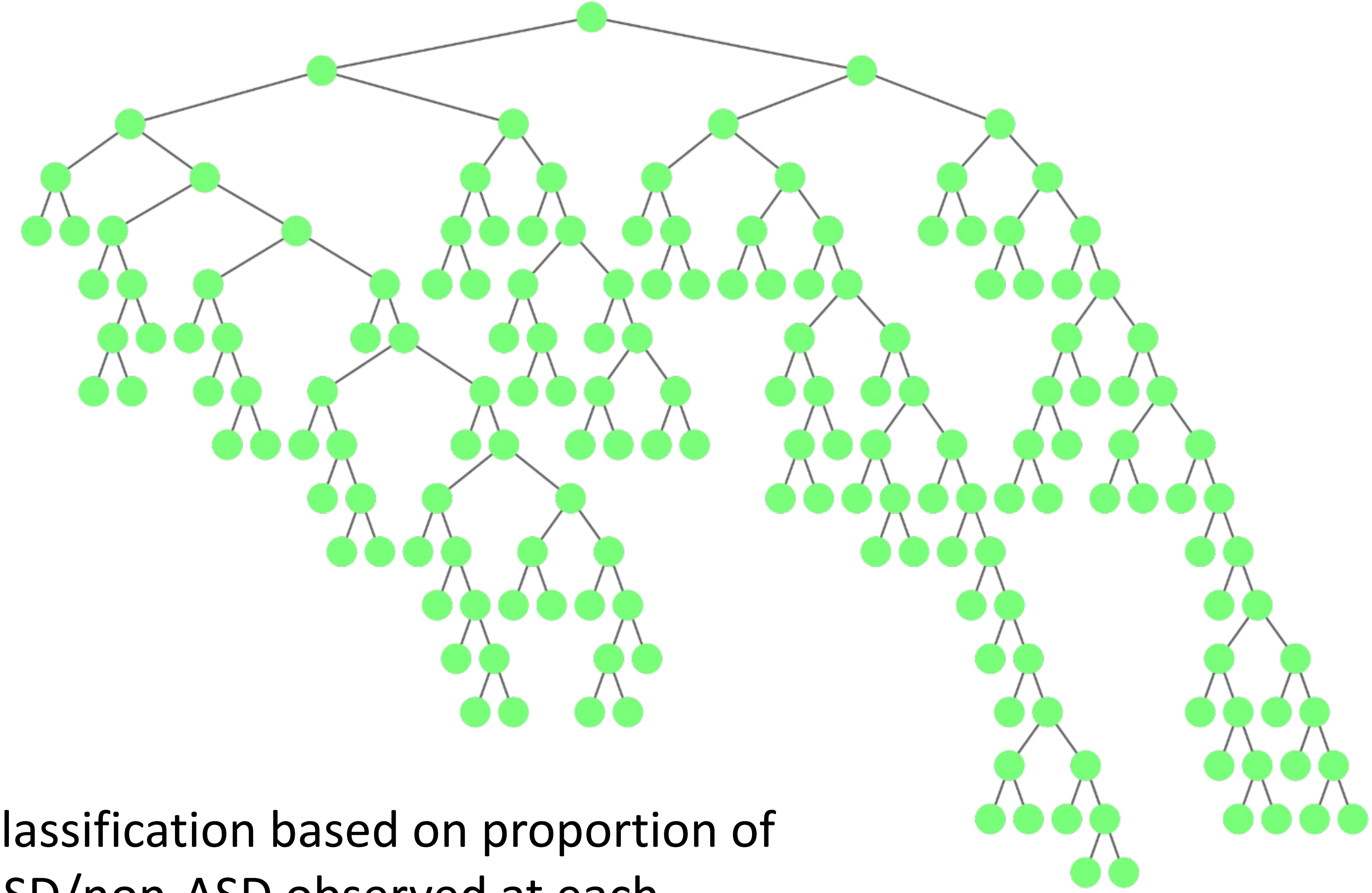
2008 Training Data
N = 1,162 Children
K = 13,135 phrases

# Random forests: training one tree



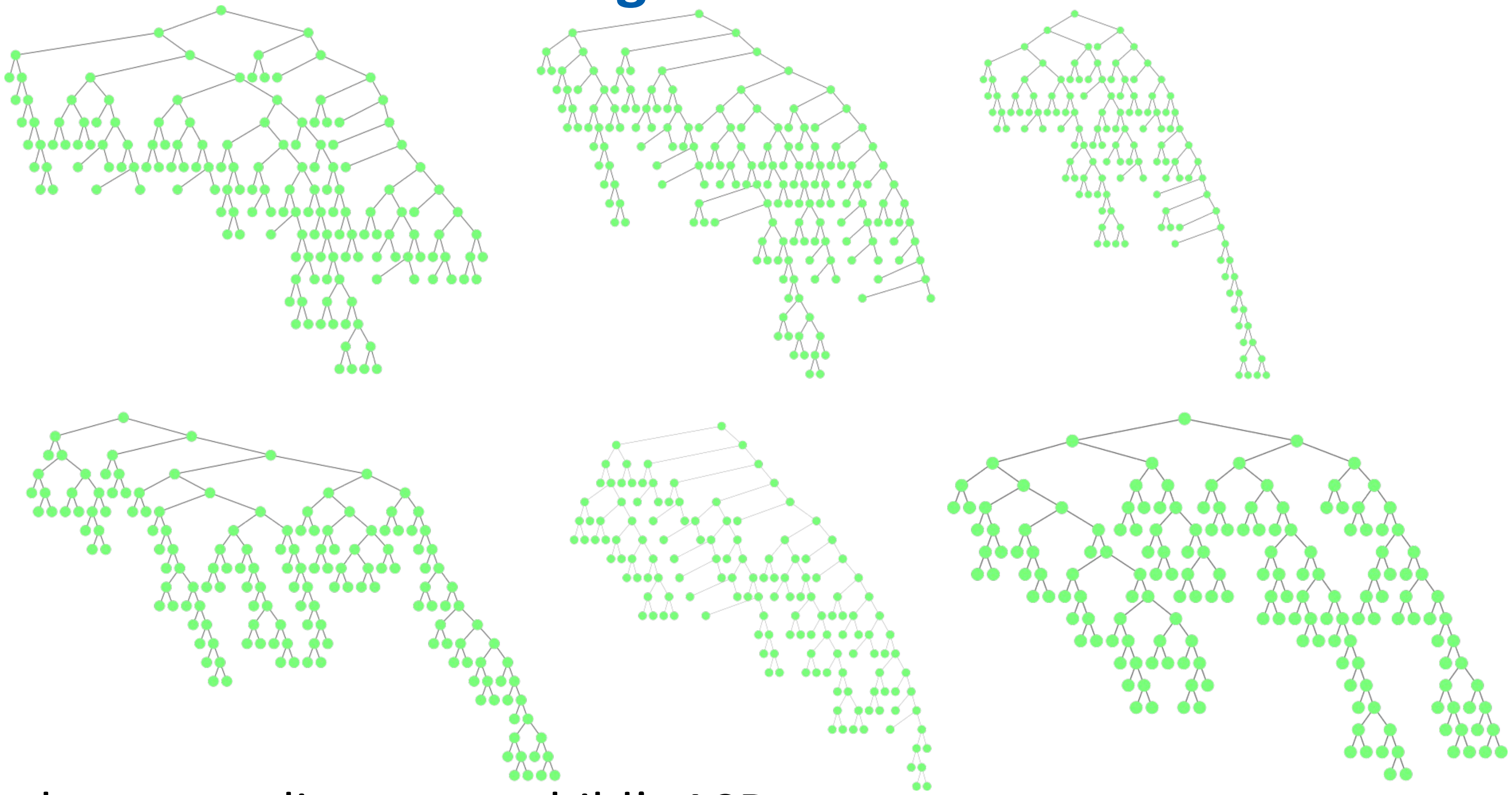

Repeat selection and splitting until tree is fully grown.

# Random Forests: classification



Classification based on proportion of ASD/non-ASD observed at each terminal node

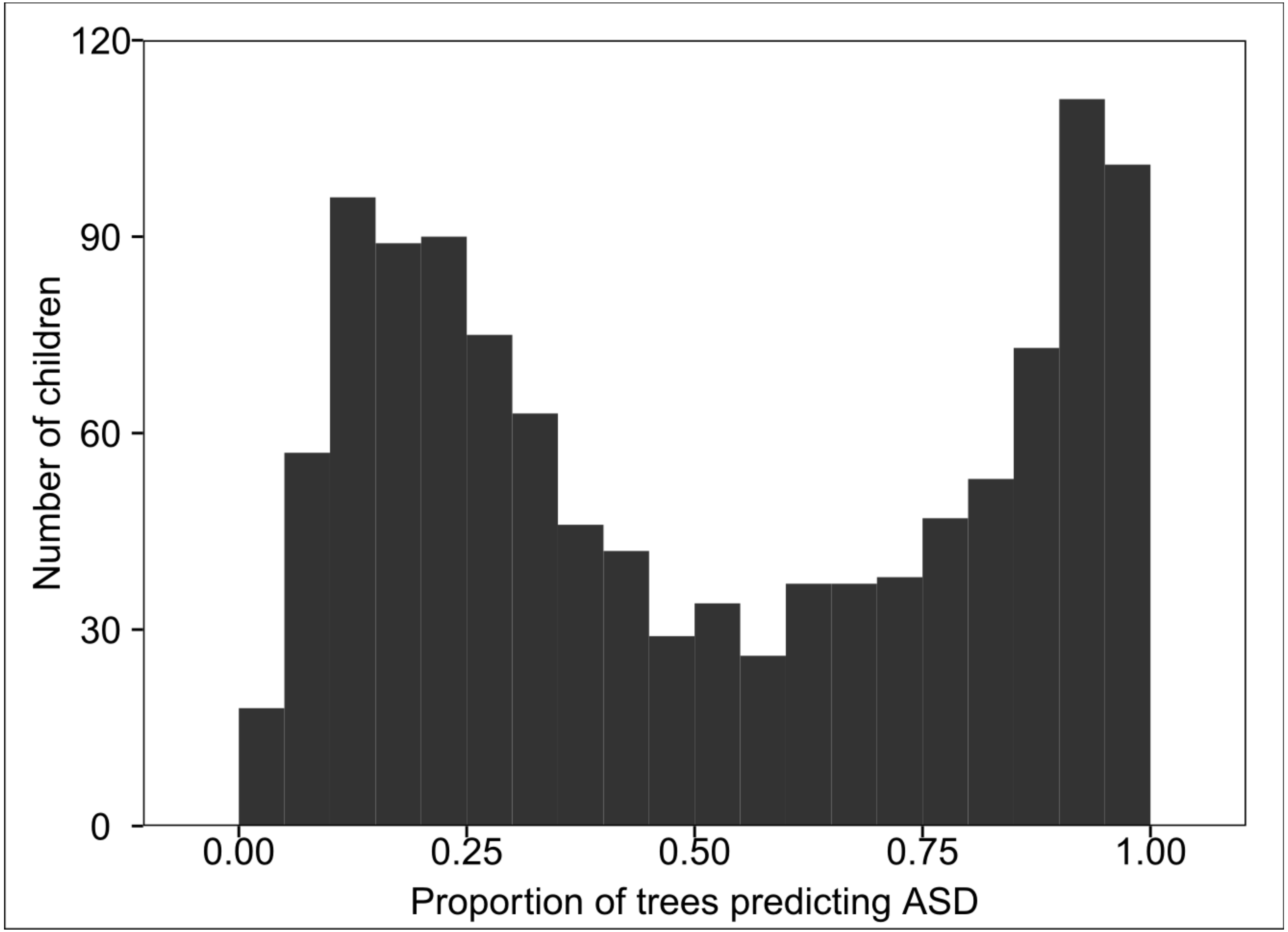
# Random Forests: voting on ASD case status



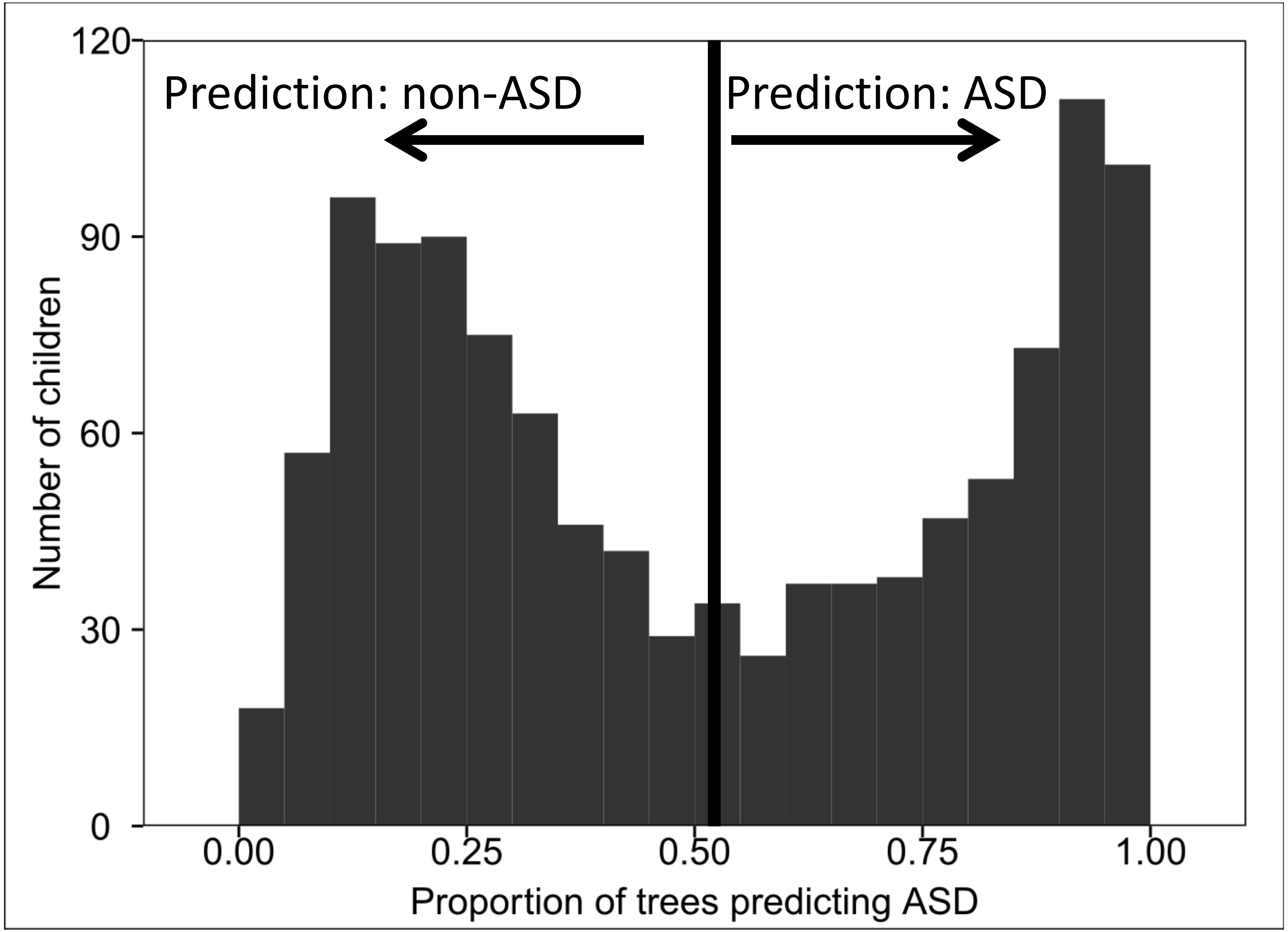
Each tree predicts every child's ASD case status.

$$\text{Child's classification score} = \frac{1}{nTree}\sum_{i=1}^{nTree}(Prediction_i)$$

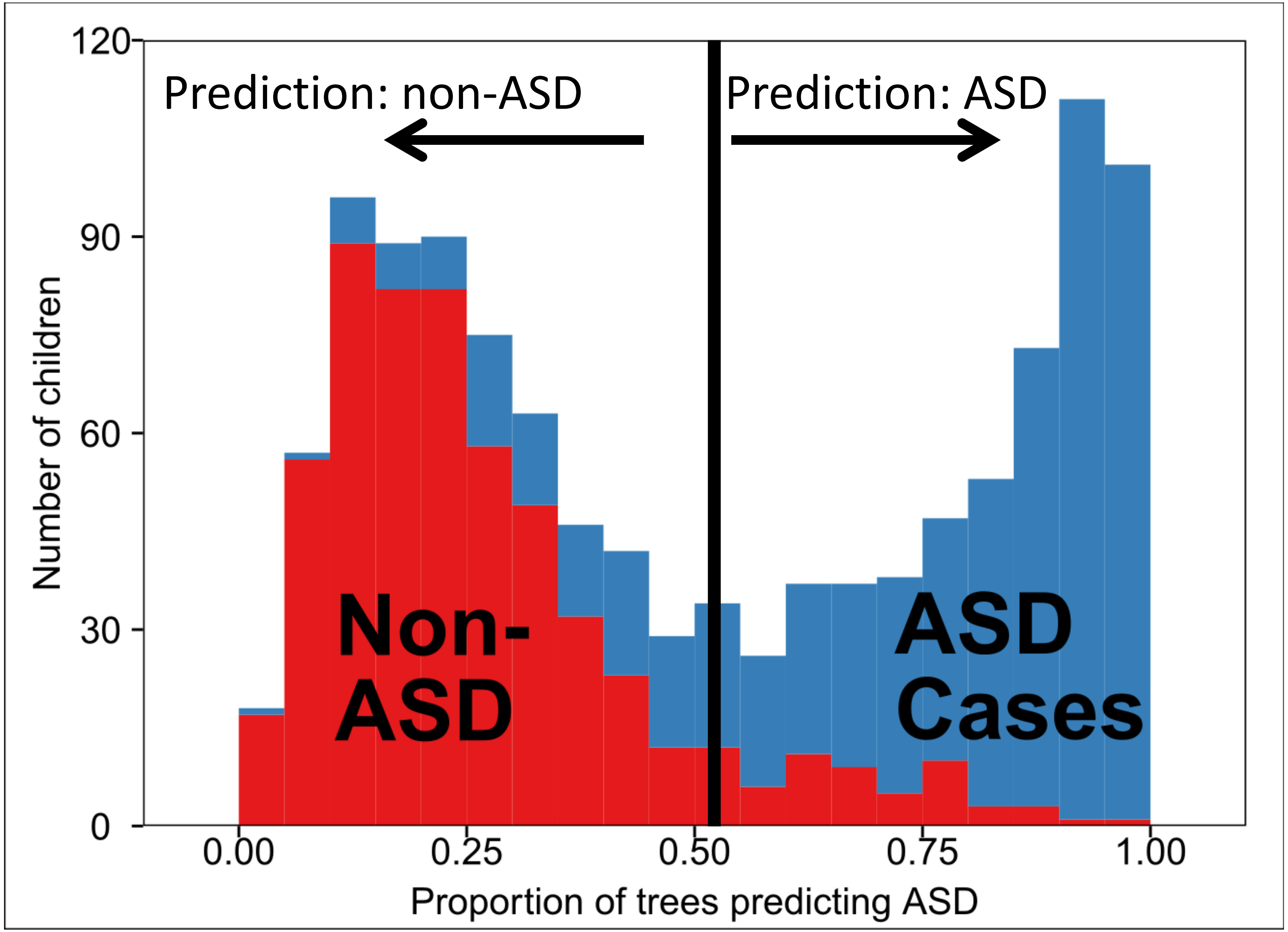
# Histogram of ASD prediction scores (N=1,165)

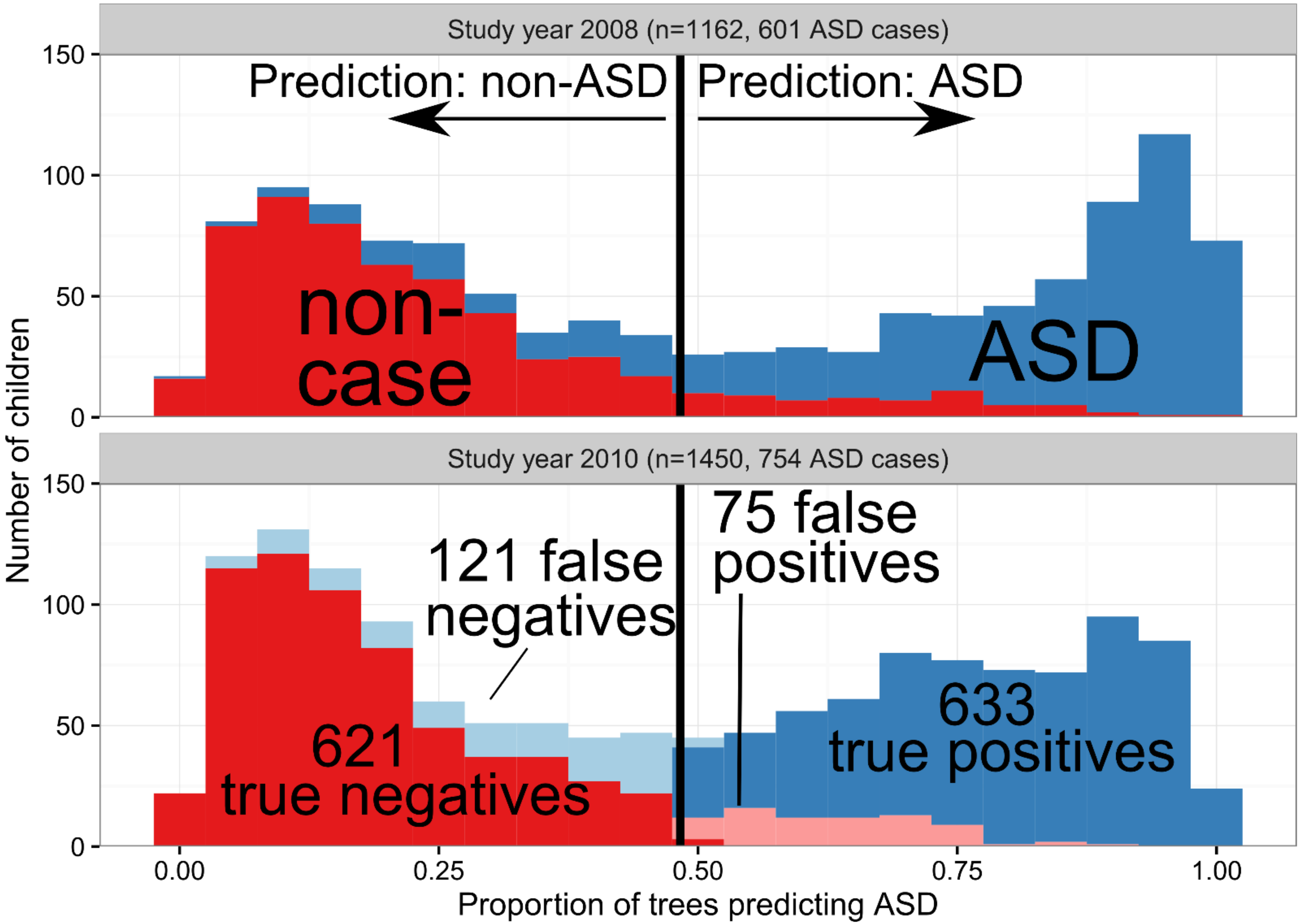


# Histogram of ASD prediction scores (N=1,165)



# Histogram of ASD prediction scores (N=1,165)





Study year 2008 (n=1162, 601 ASD cases)
Prediction: non-ASD
Prediction: ASD
non-case
ASD
Study year 2010 (n=1450, 754 ASD cases)
121 false negatives
75 false positives
621 true negatives
633 true positives
Number of children
150
100
50
0
0.00
0.25
0.50
0.75
1.00
Proportion of trees predicting ASD


Aug 2017

# Algorithm vs clinician ASD classification
# Georgia ADDM Site

| Statistic | 2008 | 2010 |
|---|---|---|
| Simple Agreement | 86.3% | 86.5% |
| Sensitivity | 84.5% | 84.0% |
| Specificity | 88.2% | 89.2% |
| Predictive Value Positive | 88.5% | 89.4% |
| Predictive Value Negative | 84.2% | 83.7% |
| Kappa | 0.73 | 0.73 |
| Area Under Receiver-Operating Characteristic Curve | 0.932 | 0.932 |

# Pilot results: 2010 Georgia ADDM data

| | Algorithm | Official* |
|---|---|---|
| Agrees with clinicians | 86.5% | 90.7% |
| Autism prevalence | 1.5% | 1.6% |
| Time for clinician review | 1 second** | 1088-1450 hours |

Clinician costs for the entire 2010 ADDM network:
> $1 Million per surveillance year

*MMWR 2014

**for prediction after data processing

# "It worked great in the lab! "

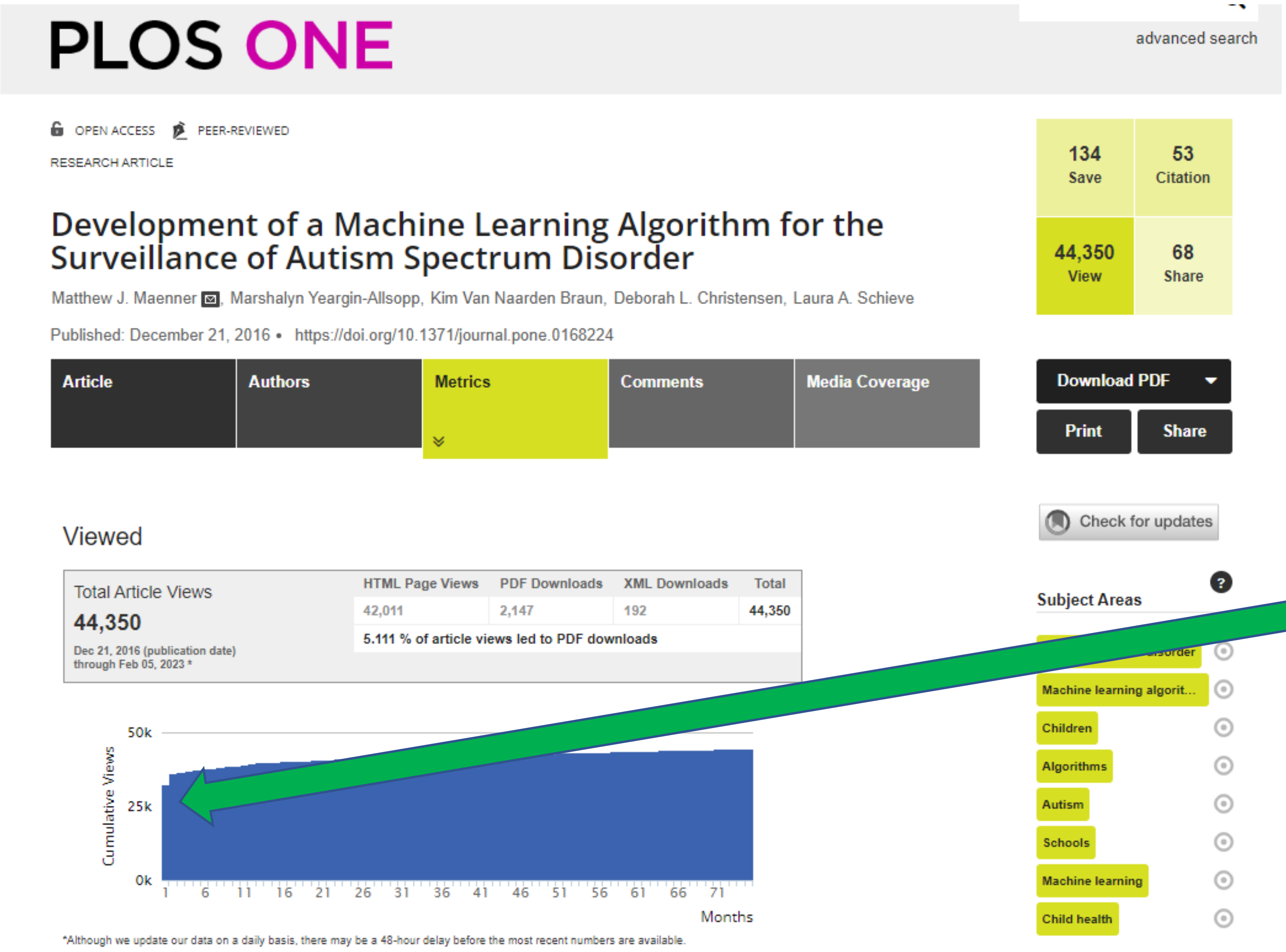
"What we want are *new* weapons - weapons totally different from any that have been employed before. Such weapons can be made [… ] and have directed research into several unexplored fields which show great promise. I believe, in fact, that a revolution in warfare may soon be upon us.“

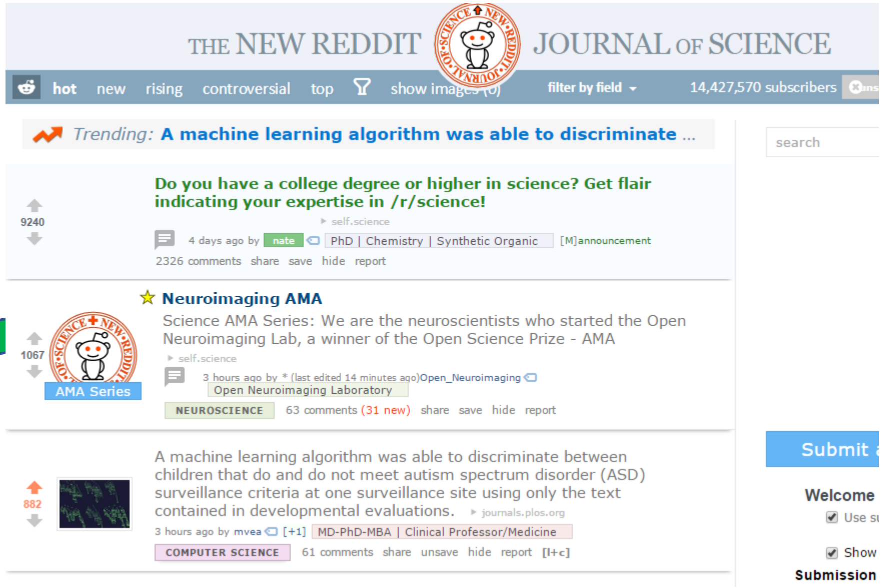
-Professor-General Norden

[from *Superiority* by Arthur C Clarke]

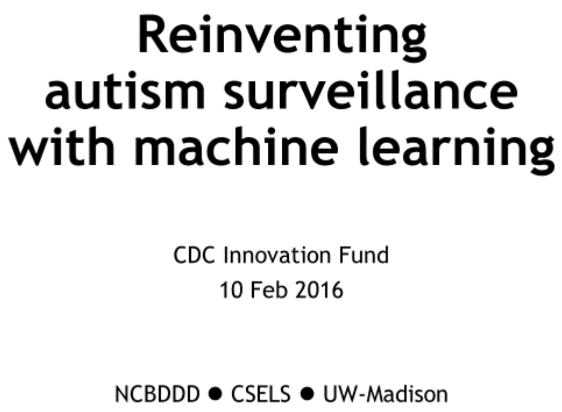
http://www.mayofamily.com/RLM/txt_Clarke_Superiority.html

# Publication




https://journals.plos.org/plosone/article/authors?id=10.1371/journal.pone.0168224





# Our Team

**Chad Heilig** (CSELS)

**Fatima Abdirizak** (NCBDDD)

**Nicole Dowling** (NCBDDD)

**Maureen Durkin** (U Wisc)

**Scott Lee** (CSELS)

**Laura Schieve** (NCBDDD)

Advisors

Juliana Cyril (CDC) & Bonny Harbinger (HHS)

Executive Sponsors

Coleen Boyle (NCBDDD) & Bill Mac Kenzie (CSELS)

# paragraph vectors (Le & Mikolov 2014)

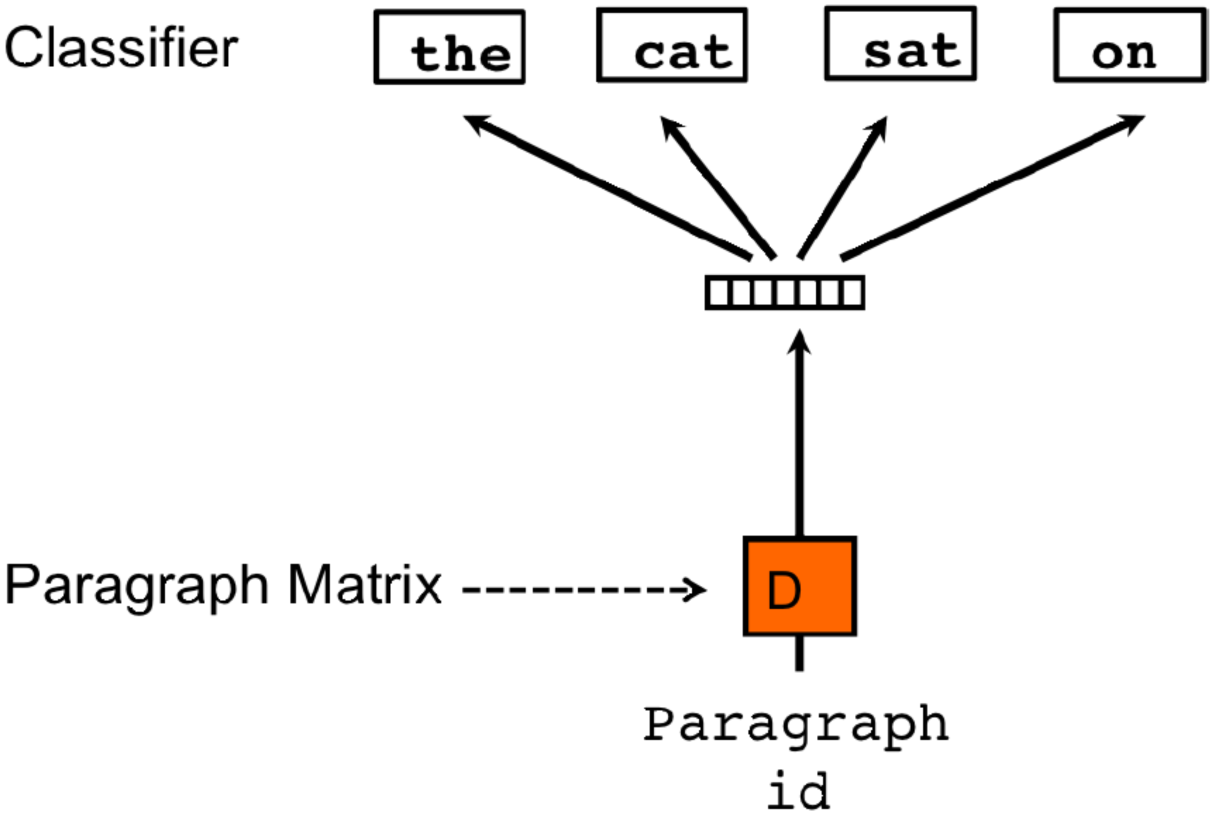



*Figure 3.* Distributed Bag of Words version of paragraph vectors. In this version, the paragraph vector is trained to predict the words in a small window.

# April 2015 – Langmuir Lecture


TF: What do 1000d vectors have to do with public health?

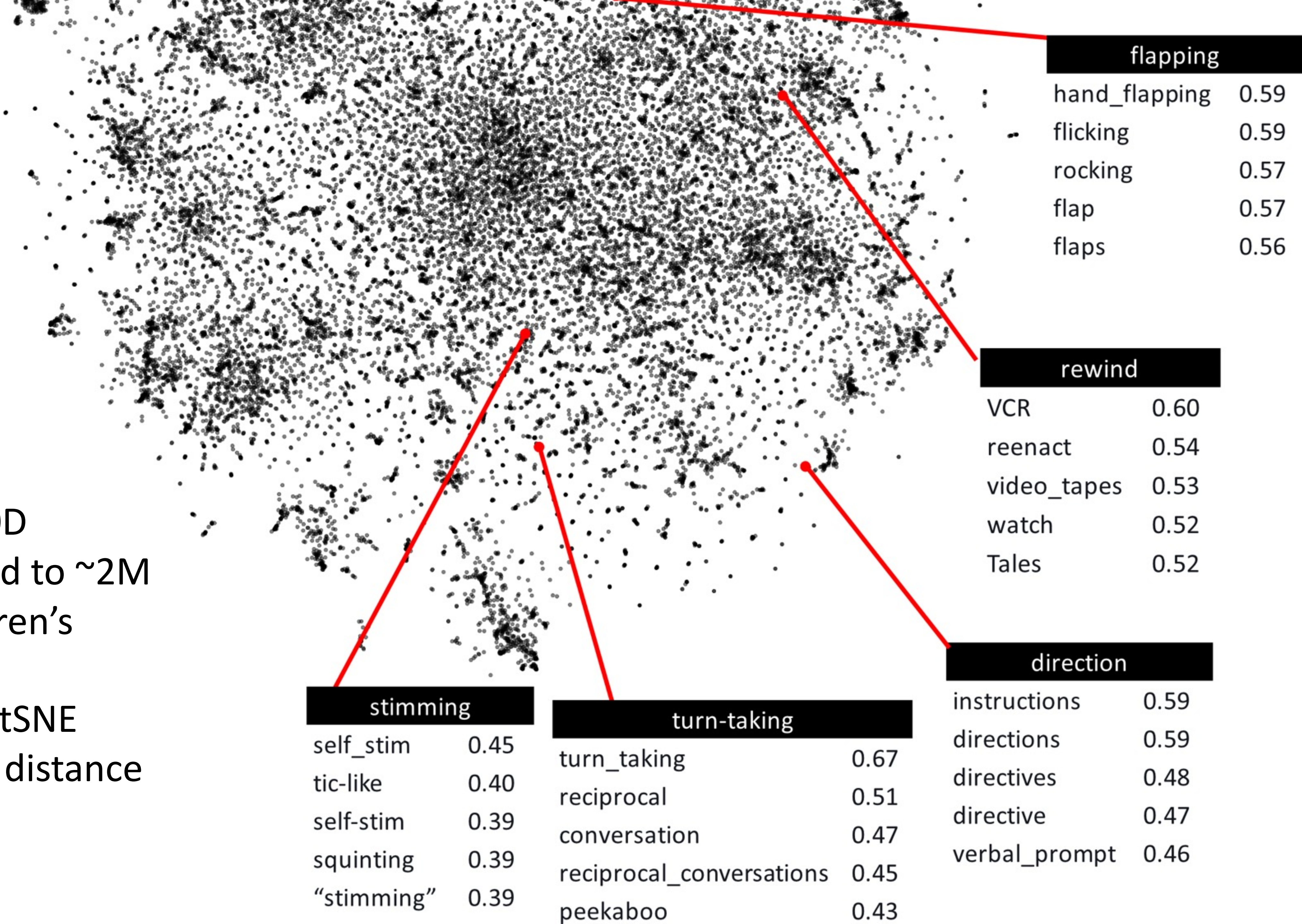
JD: If you have a new and rare disease, you can see how it is similar to other diseases, by looking at surrounding information and get new ideas.

https://www.youtube.com/watch?v=kO-Iw9xlxy4

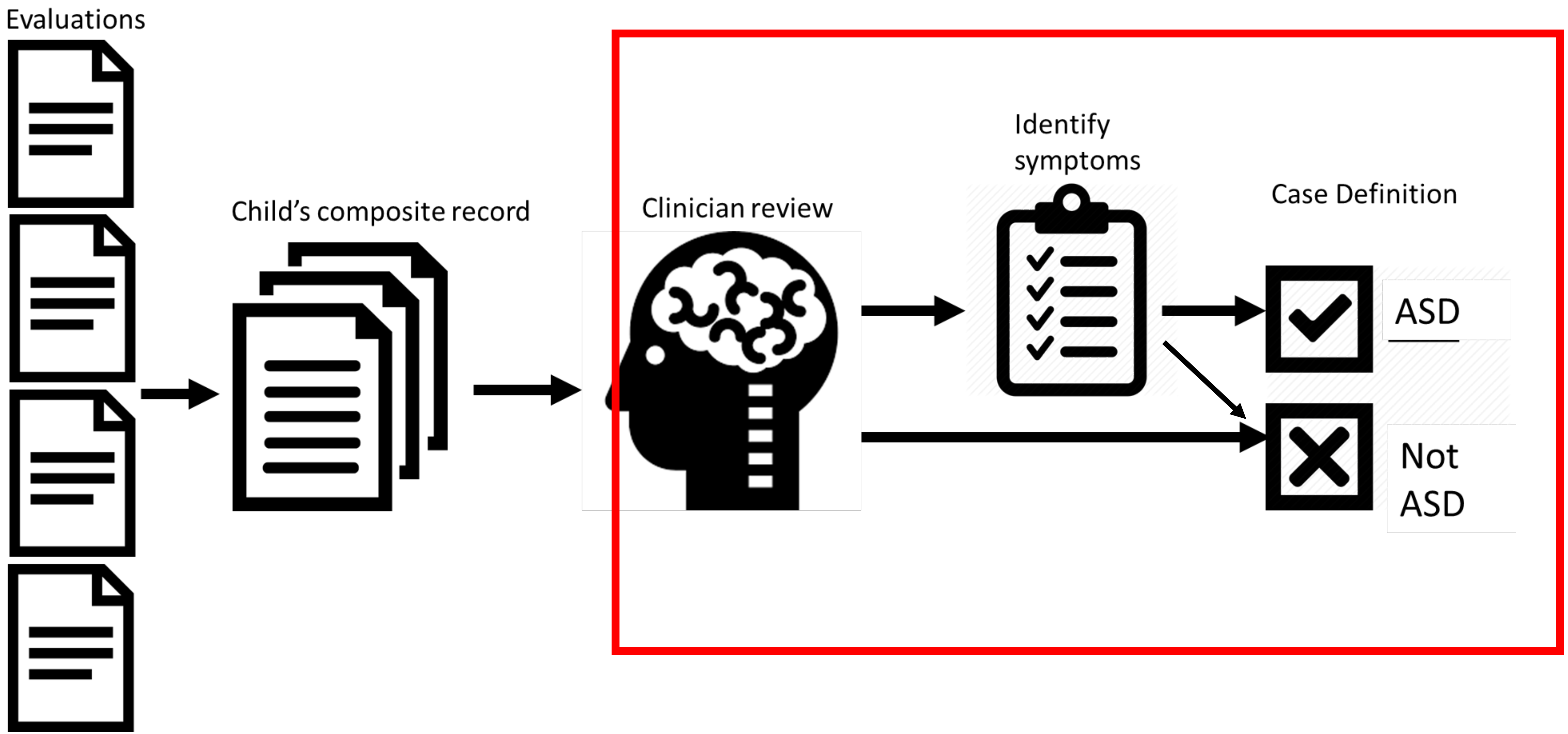
Distributed word embeddings (300D word2vec) applied to ~2M words from children’s evaluations.

Visualization: 2D tSNE

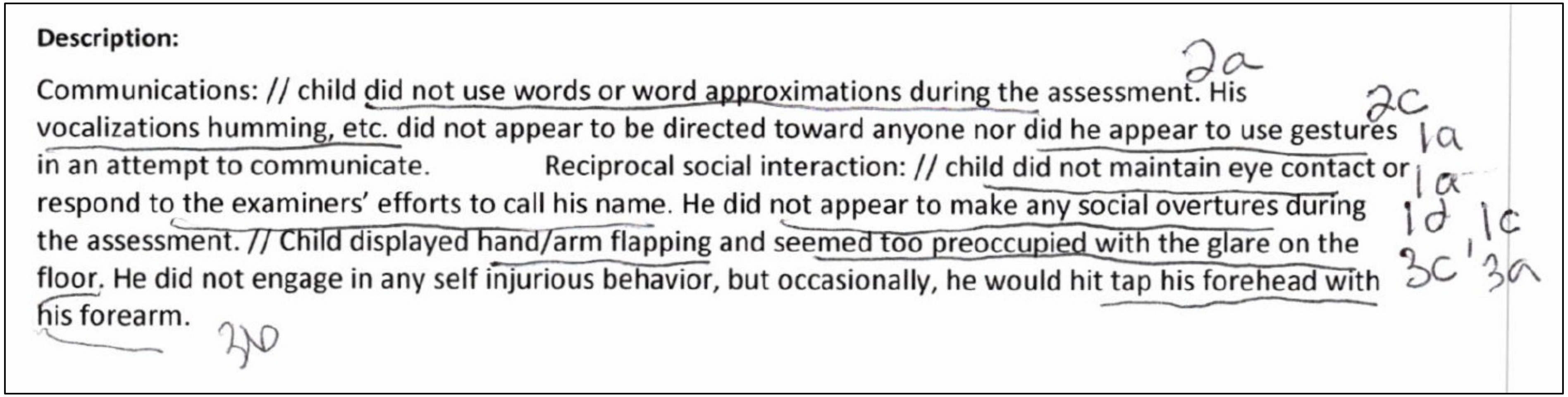
Similarity: cosine distance



# Training a classifier to detect autism symptoms

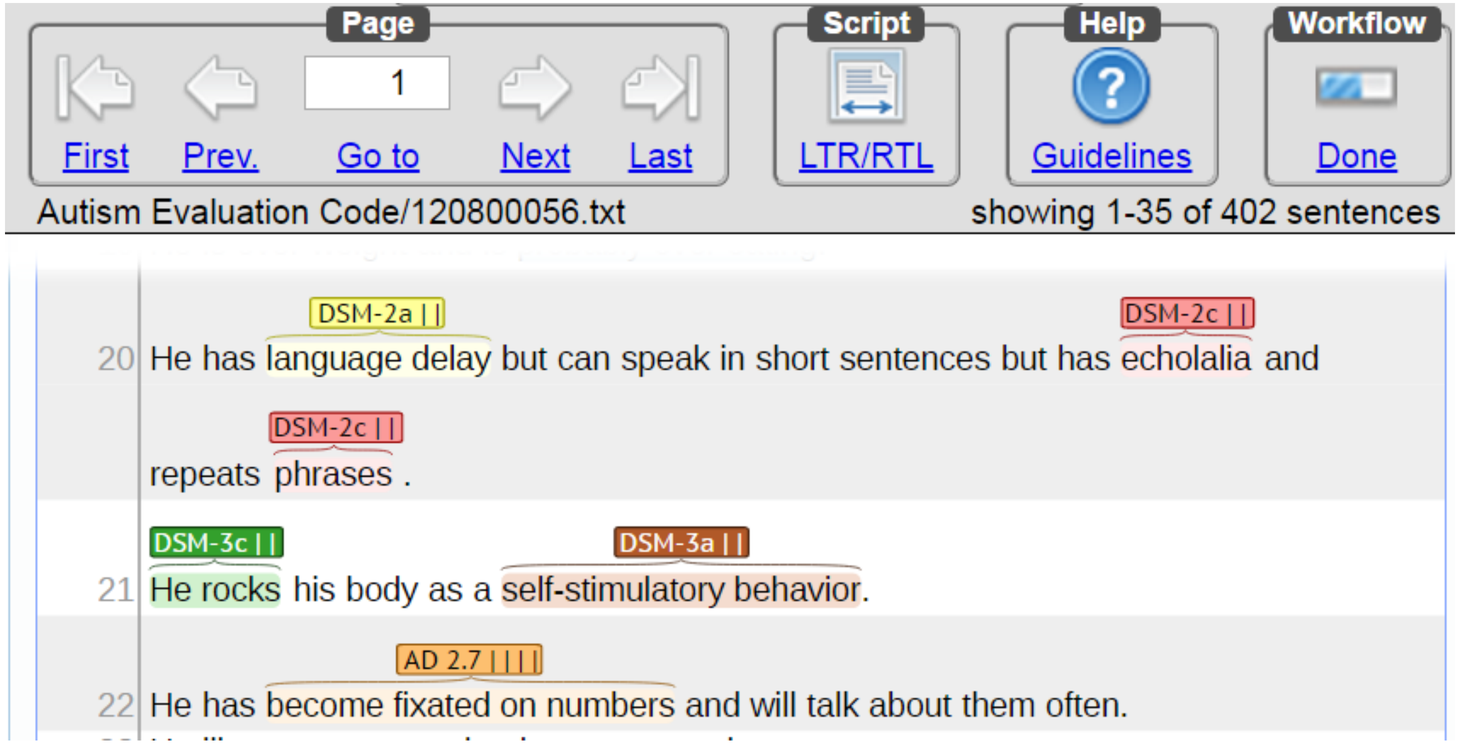



Dec 2018


**Description:**

Communications: // child did not use words or word approximations during the assessment. His vocalizations humming, etc. did not appear to be directed toward anyone nor did he appear to use gestures in an attempt to communicate. Reciprocal social interaction: // child did not maintain eye contact or respond to the examiners' efforts to call his name. He did not appear to make any social overtures during the assessment. // Child displayed hand/arm flapping and seemed too preoccupied with the glare on the floor. He did not engage in any self injurious behavior, but occasionally, he would hit tap his forehead with his forearm.

# Digitizing 50,000 sentences of paper-based annotations



*DSM-IV-TR* 1A: Marked impairment in the use of multiple nonverbal behaviors such as eye to-eye gaze, facial expression, body postures, and gestures to regulate social interaction.

Algorithm: Positive / Clinician: Negative

```
[1] "Sustained eye contact with people was fleeting, but present
     for short periods."
[2] "Makes eye contact with speakers. 2."
[3] "Behavior: calm, cooperative and poor eye contact."
```

Annotation data was "noisy" and didn't improve the overall models.

- Algorithms not as savvy as people
- Lots of complex coding rules make it difficult to score items out of context

PLOS | ONE

PUBLISH ABOUT BROWSE

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

# A comparison of machine learning algorithms for the surveillance of autism spectrum disorder

Scott H. Lee, Matthew J. Maenner, Charles M. Heilig

Published: September 25, 2019 • https://doi.org/10.1371/journal.pone.0222907

Compared:

- LDA
- LSA
- Random Forests
- MNB
- SVM
- NB-SVM*
- Neural network adapted from fastText*

Using Data From:

- Georgia ADDM Site
- 2008, 2010, 2012
- 3,739 children
- 59,660 unique words
- 7.8M total words
- Evaluations range from few words to >10,000

Feb 2018

https://journals.plos.org/plosone/article/comments?id=10.1371/journal.pone.0222907

*Scott's own implementation

# An algorithmic shoot-out.

| Model | Sens | Spec | PPV | NPV | $F_1$ | Acc (95% CI) | Diff acc (95% CI) |
|---|---|---|---|---|---|---|---|
| LDA | 44.2 | 72.4 | 60.6 | 57.5 | 51.1 | 58.6 (55.0, 62.2) | -29.0 (-32.4, -25.6) |
| MNB | 82.3 | 72.6 | 74.2 | 81.0 | 78.0 | 77.3 (73.9, 80.7) | -10.3 (-12.5, -8.1) |
| SVM | 83.5 | 84.5 | 83.8 | 84.2 | 83.6 | 84.0 (80.8, 87.2) | -3.7 (-6.6, -0.7) |
| LSA | 81.5 | 88.5 | 87.2 | 83.3 | 84.2 | 85.1 (83.1, 87.0) | -2.6 (-4.2, -0.9) |
| $NN_{sum}$ | 85.5 | 84.7 | 84.4 | 86.0 | 84.9 | 85.1 (81.9, 88.3) | -2.6 (-5.2, 0.1) |
| $NN_{avg}$ | 86.3 | 86.4 | 85.9 | 86.9 | 86.0 | 86.3 (84.4, 88.2) | -1.3 (-3.3, 0.6) |
| RF | **87.0** | 87.1 | 86.6 | **87.5** | 86.8 | 87.1 (83.8, 90.4) | -0.5 (-2.2, 1.1) |
| NB-SVM | 85.2 | **89.9** | **89.0** | 86.4 | **87.1** | **87.6** (85.2, 90.1) | * |

https://doi.org/10.1371/journal.pone.0222907.t002

- Several methods performed similarly well, but overall no great improvement over Random Forests.
- Likely limited benefit of using "deep learning" methods
- Able to replicate earlier results on a slightly broader dataset

Lee, Maenner, Heilig, 2019

# On choosing the “best” algorithm

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

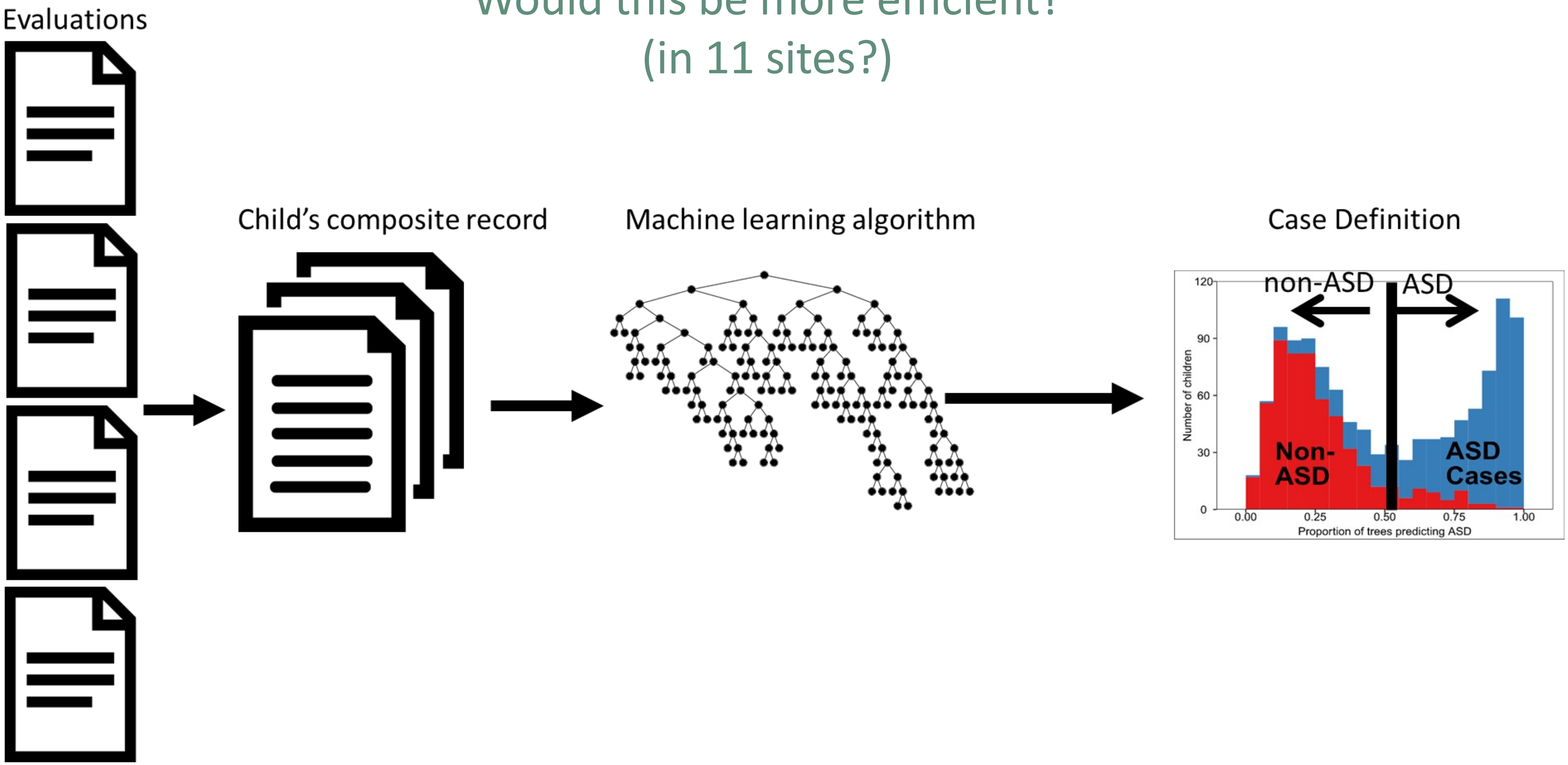
Manuel Fernández-Delgado MANUEL.FERNANDEZ.DELGADO@USC.ES
Eva Cernadas EVA.CERNADAS@USC.ES
Senén Barro SENEN.BARRO@USC.ES
*CITIUS: Centro de Investigación en Tecnoloxías da Información da USC*
*University of Santiago de Compostela*
*Campus Vida, 15872, Santiago de Compostela, Spain*

Dinani Amorim DINANIAMORIM@GMAIL.COM
*Departamento de Tecnologia e Ciências Sociais- DTCS*
*Universidade do Estado da Bahia*
*Av. Edgard Chastinet S/N - São Geraldo - Juazeiro-BA, CEP: 48.305-680, Brasil*

(hint: Betteridge’s Law)

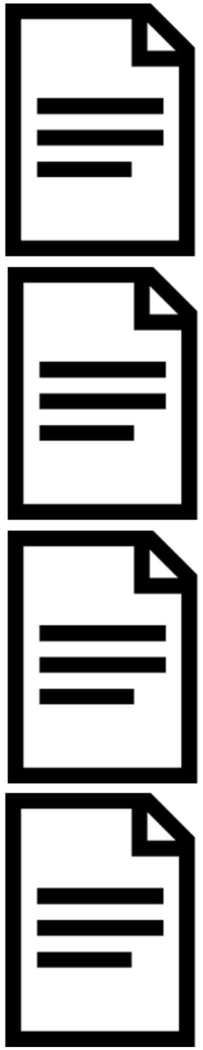
Aug 2017

# Would this be more efficient?
# (in 11 sites?)



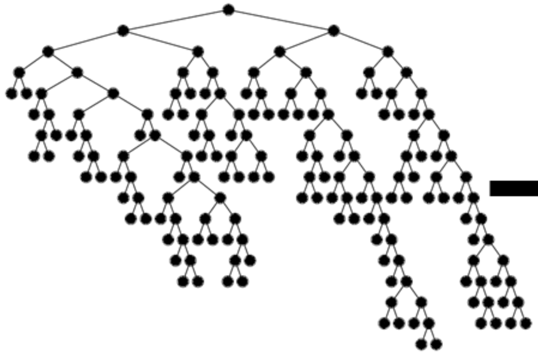
## Would this be more efficient?
## (in 11 sites?)

Evaluations



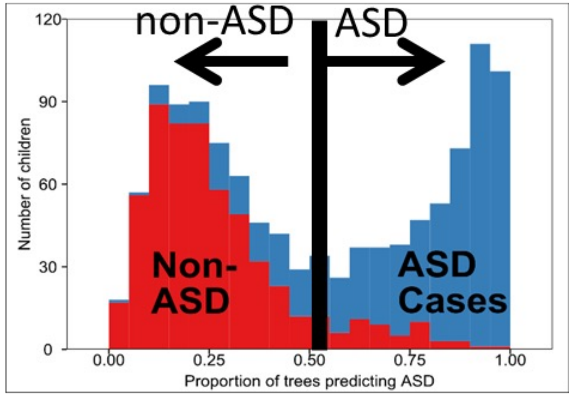
Child's composite record

Machine learning algorithm

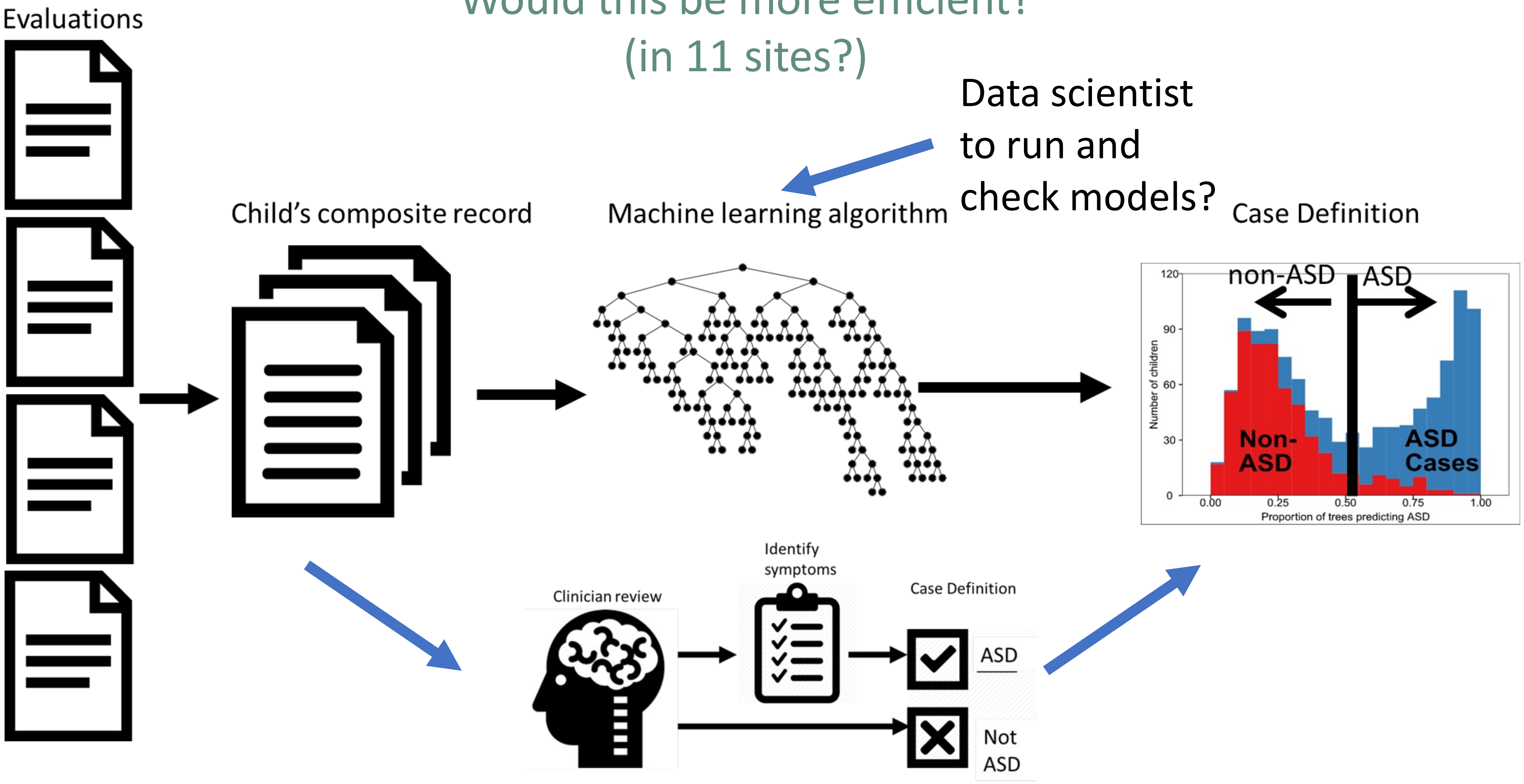


Data scientist
to run and
check models?

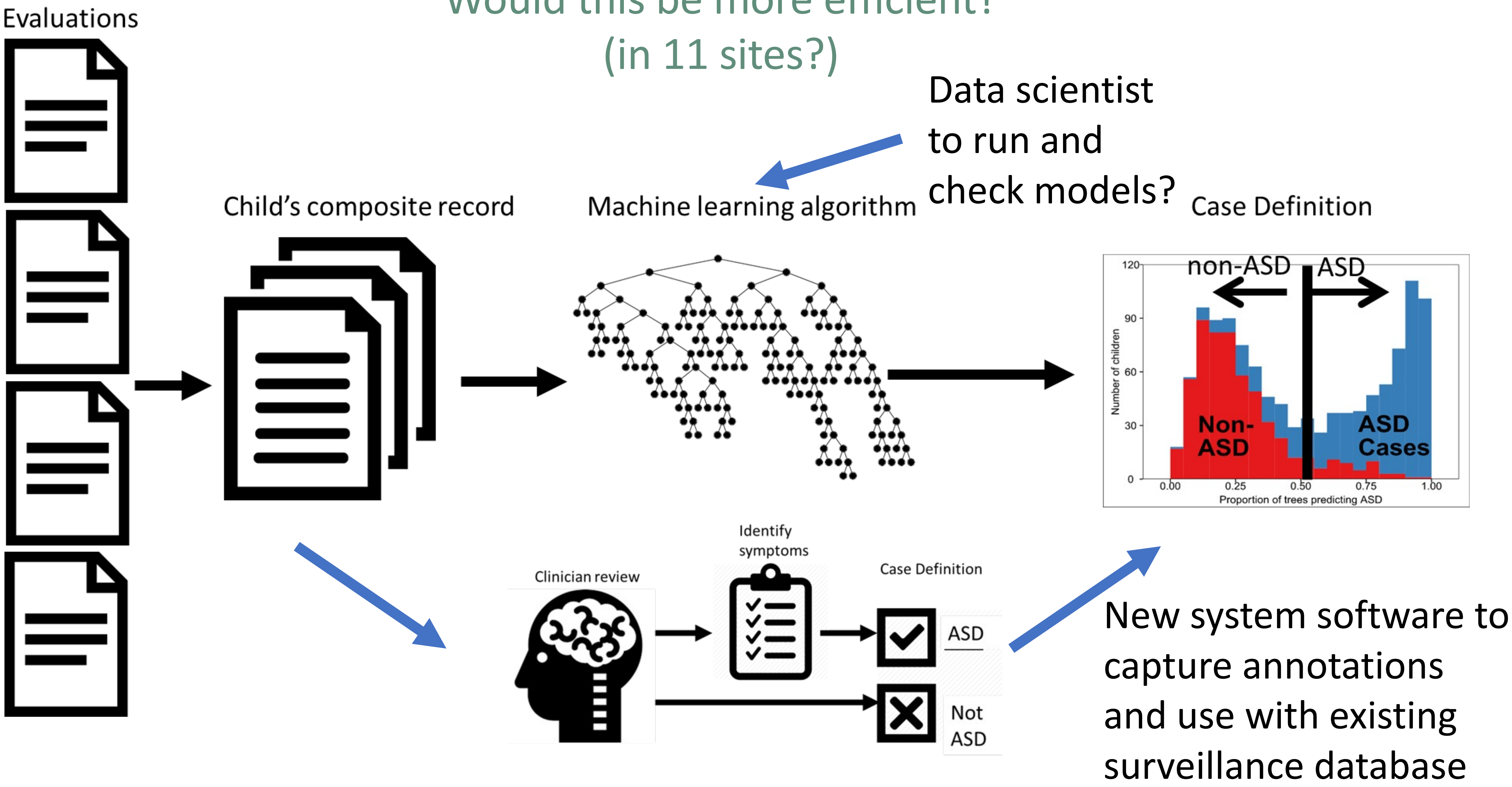
Case Definition



# Would this be more efficient?
# (in 11 sites?)



# Would this be more efficient?
# (in 11 sites?)




March 2021

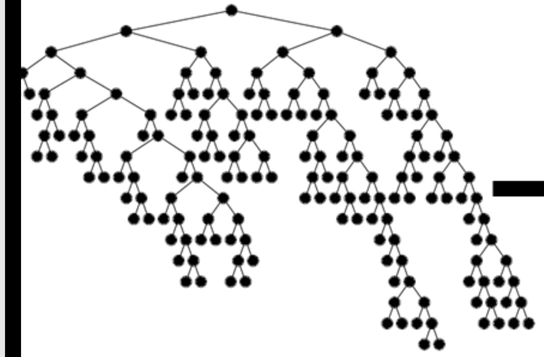
## Would this be more efficient?
## (in 11 sites?)

Evaluations

Child's composite record

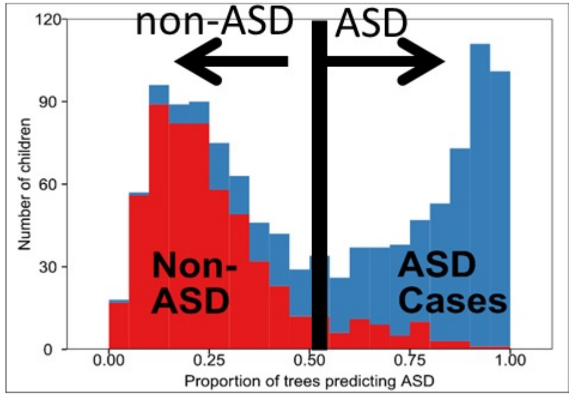
**Reclassify historical data with updated algorithms to consistently measure trends?**

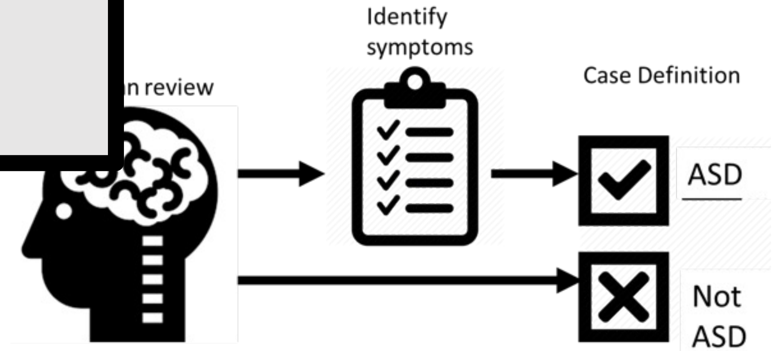
Machine learning algorithm



Data scientist to run and check models?

Case Definition



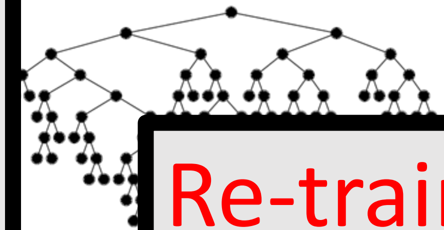

New system software to capture annotations and use with existing surveillance database

# Would this be more efficient?
# (in 11 sites?)

Evaluations

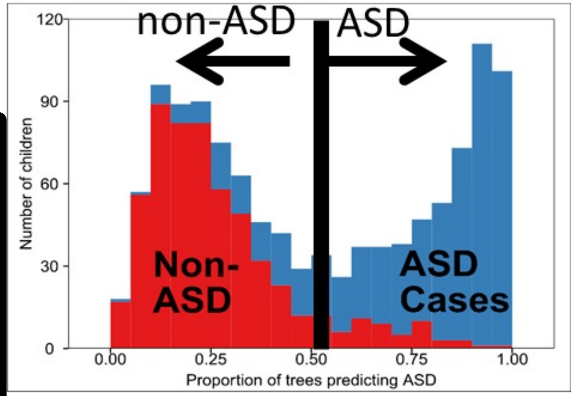
Child's composite record

Machine learning algorithm



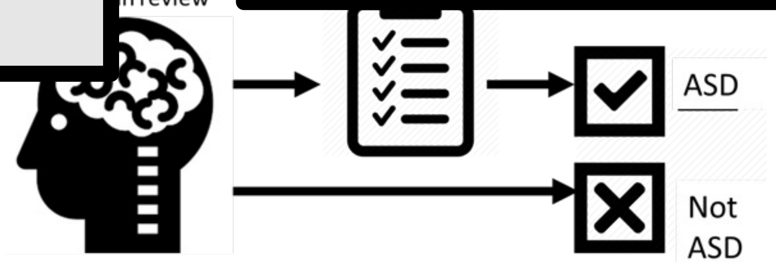
Data scientist to run and check models?

Case Definition




Reclassify historical data with updated algorithms to consistently measure trends?

Re-train all models for DSM-5 criteria




New system software to capture annotations and use with existing surveillance database

# TINSTAAFL

## Hidden Technical Debt in Machine Learning Systems

**D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips**
{dsculley,gholt,dgg,edavydov,toddphillips}@google.com
Google, Inc.

**Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François**
{ebner,vchaudhary,mwyoung,jfcrespo,dennison
Google, Inc.

### Abstract

Machine learning offers a fantastically powerful toolkit for building useful complex prediction systems quickly. This paper argues it is dangerous to think of these quick wins as coming for free. Using the software engineering framework of *technical debt*, we find it is common to incur massive ongoing maintenance costs in real-world ML systems. We explore several ML-specific risk factors to account for in system design. These include boundary erosion, entanglement, hidden feedback loops, undeclared consumers, data dependencies, configuration issues, changes in the external world, and a variety of system-level anti-patterns.

https://proceedings.neurips.cc/paper_files/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf

# Can we stand behind the algorithm-generated result?


WIRED BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY SIGN IN SUBSCRIBE

DAVID LAZER RYAN KENNEDY OPINION OCT 1, 2015 7:00 AM

## What We Can Learn From the Epic Failure of Google Flu Trends

GFT seemingly presented an eerily accurate projection of the prevalence of the flu, turning the digital refuse of people's searches into potentially life saving insights.

And then, GFT failed—and failed spectacularly—**missing at the peak of the 2013 flu season by 140 percent**. When Google quietly euthanized the program, called Google Flu Trends (GFT), it turned the poster child of big data into the poster child of the foibles of big data. […] what we like to call "big data hubris."

https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/

# Evaluating a machine-learning approach for autism surveillance

- Simplicity – more complex
- Flexibility – hypothetically, but training each model requires resources
- Data quality & Acceptability – not if algorithm produces odd results
- Sensitivity & PPV – still unknown and likely varies by site
- Representativeness – potentially apply algorithms to larger datasets
- Timeliness: maybe a little, but **large majority of effort is the data collection needed for clinician review**
- Stability –will system enhancements and revisions to algorithms lead to problems?

https://www.cdc.gov/mmwr/preview/mmwrhtml/rr5013a1.htm

Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead

*Cynthia Rudin*
*Duke University*
*cynthia@cs.duke.edu*

https://arxiv.org/abs/1811.10154

## The Rashomon set argument:

Consider that the data permit a large set of reasonably accurate predictive models to exist.

Because this set of accurate models is large, it often contains at least one model that is **interpretable.**

When models are inherently interpretable, they provide their own explanations, which are faithful to what the model actually computes.

https://www.youtube.com/watch?v=wL4X4lG20sM

Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead

*Cynthia Rudin*
*Duke University*
*cynthia@cs.duke.edu*

https://arxiv.org/abs/1811.10154

## The Rashomon set argument:

Consider that the data permit a large set of reasonably accurate predictive models to exist.

Because this set of accurate models is large, it often contains at least one model that is **interpretable.**

When models are inherently interpretable, they provide their own explanations, which are faithful to what the model actually computes.

PLOS ONE
PUBLISH ABOUT BROWSE

OPEN ACCESS PEER-REVIEWED
RESEARCH ARTICLE

A comparison of machine learning algorithms for the surveillance of autism spectrum disorder

Scott H. Lee, Matthew J. Maenner, Charles M. Heilig

Published: September 25, 2019 • https://doi.org/10.1371/journal.pone.0222907

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

Conditions met!

https://www.youtube.com/watch?v=wL4X4lG20sM

"….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"* "[data science is] a set of core activities for **asking good questions** and **lining up the tools** to **answer them rigorously** using data."

**-Chad Heilig**
(Formerly) Associate Director for Data Science, CSELS, CDC

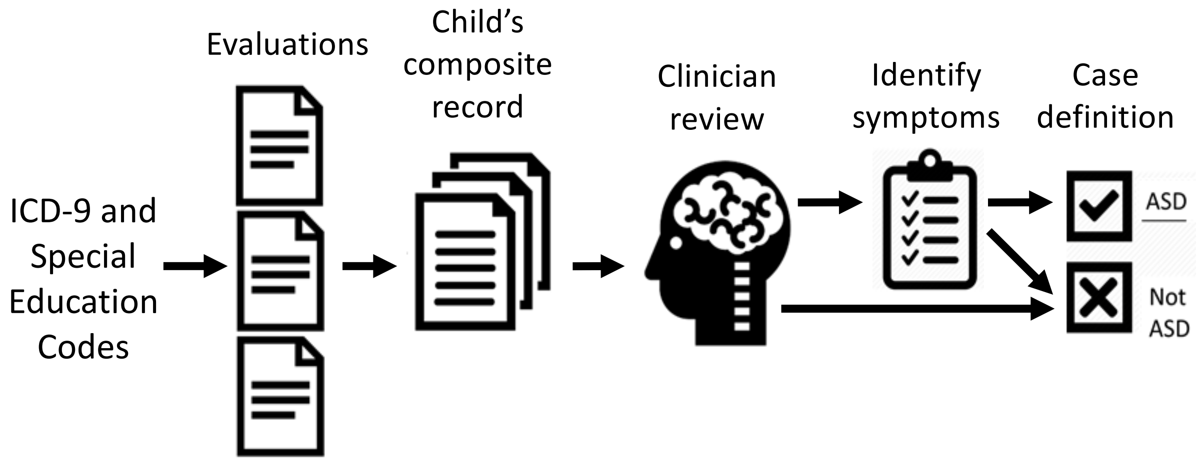
http://intranet.cdc.gov/expression-data-science/2016/05/17/welcome-to-expression-of-data-science/

I put this in all my talks

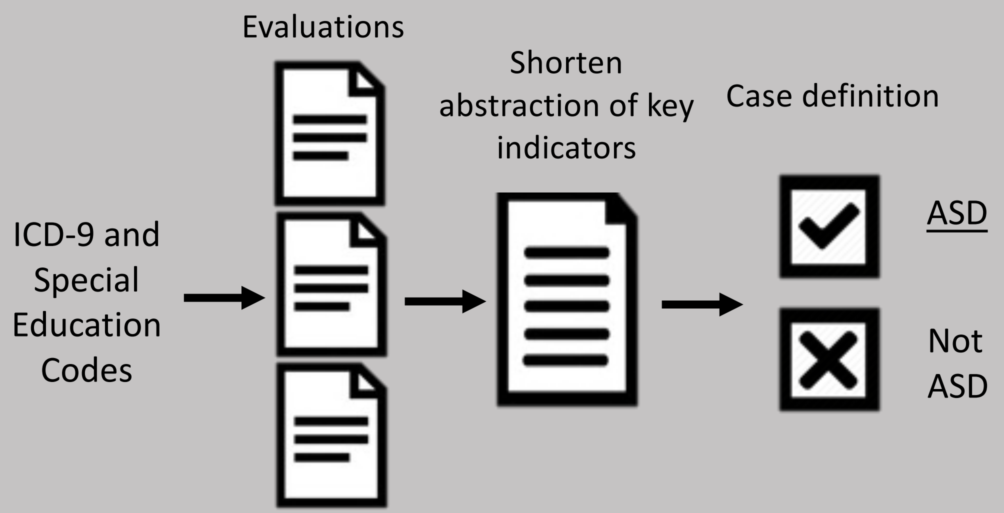
# Postscript: interpretable solution from our set of models

**Very similar results!**

**Previous approach**




**Current approach (1 year faster)**




**Case Definition: child has at least one of the following**

ASD diagnosis (ever)

ASD special education classification (ever)

ASD ICD Code (299.XX, F84.X)

Maenner et al, AJE, 2021
https://pubmed.ncbi.nlm.nih.gov/33847734/

Adapted from March 2021

# Judging the potential of an ML project?

| Issue | Go for it! | Investigate further… |
|---|---|---|
| Where is the data? | in a database you can access | In multiple places and needs manual data entry |
| How much data? | A lot | Not enough for automation savings to offset ML investment. \| Not enough to train algorithm |
| "honest broker" ? | uninvolved experts | Anyone who stands to benefit from contract or CoAg, or staff keen to do an "ML project" |
| High stakes results? | It's ok if algorithm gives crazy results | CDC must stand behind all results; analyst is treating it like a "black box" |
| How much time / resources saved? | Significant proportion | Small amount relative to entire project and/or costs of ML project team |
| Purpose | one-off paper | Integrating ML into ongoing system, anything requiring modifications to infrastructure or permanent resources |
| Is ML the simplest effective approach? | Yes | No, but it is the most interesting. |

# Acknowledgements

**Machine Learning project collaborators/co-authors**

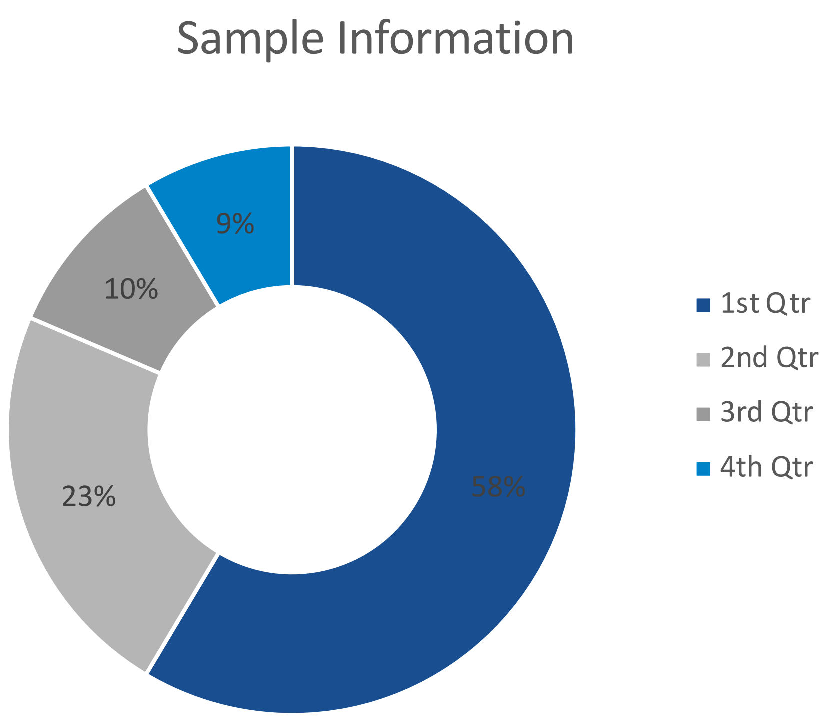
Chad Heilig (Assoc Dir of Data Science?, CSELS?)
Fatima Abdirizak (NCBDDD)
Nicole Dowling (Former Branch Chief, DDB)
Maureen Durkin (UW-Madison)
Scott Lee (CSELS?)
Laura Schieve (EIS supervisor)
Daisy Christensen (CDDB)
Kim Van Naarden Braun (CDDB)
Marshalyn Yeargin-Allsopp (Former Branch Chief, DDB)
Patty Dietz (Former Branch Chief, CDDB)

**CDDB Surveillance team (2019-2022)**

Esther Amoakohene
Monica DiRienzo
Michelle Hughes
Dedria McArthur
Mary Patrick
Ashley Robinson Williams
Corshae Robinson
Kelly Shaw
Anita Washington
Susan Williams

Matt: xde8@cdc.gov

# Presentation Header



# Presentation Header

Sample

- Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.
- Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.
- Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

Sample

- Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.
- Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.
- Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

# Presentation Header

### Sample

- Soluptate a adicil magnam ne quia cum, conempos doloris magnisimi, aspelibero modit pre perum que etur si sitatiorum eum comnis nonesse nimaxim oluptate nossimu sapitia sitatempore con nus si is dolupta que voluptum rehenis voluptatem si cusdand aessim comnimagnim fugiatquatur aut este omnihil iumque voluptatem quam, qui arunt aut enda verro berores preperu ntemporio. Nam, es


### Sample

- Soluptate a adicil magnam ne quia cum, conempos doloris magnisimi, aspelibero modit pre perum que etur si sitatiorum eum comnis nonesse nimaxim oluptate nossimu sapitia sitatempore con nus si is dolupta que voluptum rehenis voluptatem si cusdand aessim comnimagnim fugiatquatur aut este omnihil iumque voluptatem quam, qui arunt aut enda verro berores preperu ntemporio. Nam, es

### Sample

- Soluptate a adicil magnam ne quia cum, conempos doloris magnisimi, aspelibero modit pre perum que etur si sitatiorum eum comnis nonesse nimaxim oluptate nossimu sapitia sitatempore con nus si is dolupta que voluptum rehenis voluptatem si cusdand aessim comnimagnim fugiatquatur aut este omnihil iumque voluptatem quam, qui arunt aut enda verro berores preperu ntemporio. Nam, es

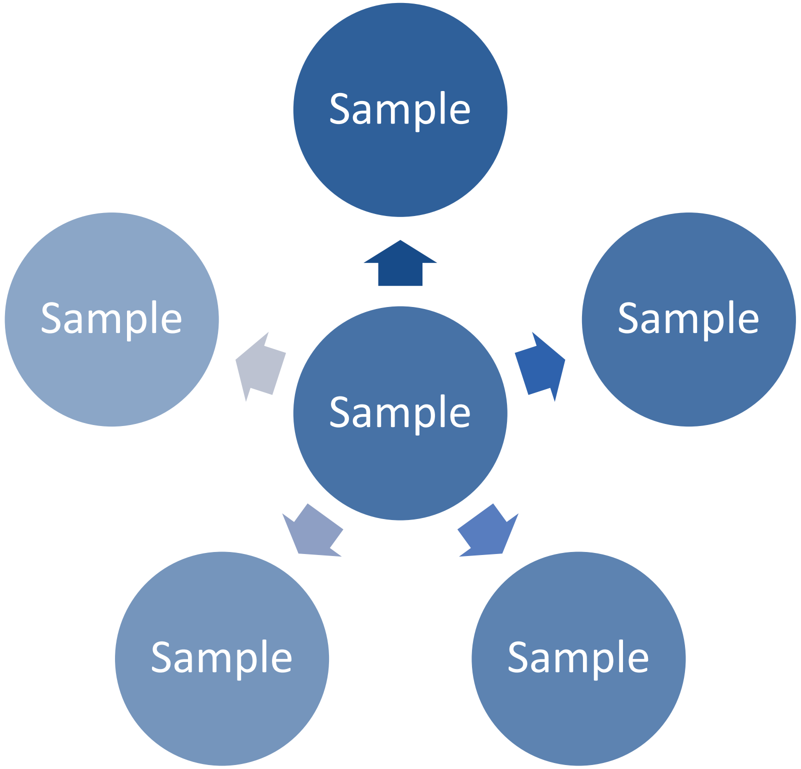
### Sample

- Soluptate a adicil magnam ne quia cum, conempos doloris magnisimi, aspelibero modit pre perum que etur si sitatiorum eum comnis nonesse nimaxim oluptate nossimu sapitia sitatempore con nus si is dolupta que voluptum rehenis voluptatem si cusdand aessim comnimagnim fugiatquatur aut este omnihil iumque voluptatem quam, qui arunt aut enda verro berores preperu ntemporio. Nam, es

### Sample

- Soluptate a adicil magnam ne quia cum, conempos doloris magnisimi, aspelibero modit pre perum que etur si sitatiorum eum comnis nonesse nimaxim oluptate nossimu sapitia sitatempore con nus si is dolupta que voluptum rehenis voluptatem si cusdand aessim comnimagnim fugiatquatur aut este omnihil iumque voluptatem quam, qui arunt aut enda verro berores preperu ntemporio. Nam, es

# Presentation Header




## Subheader

Soluptate a adicil magnam ne quia cum, conempos doloris magnisimi, aspelibero modit pre perum que etur si sitatiorum eum comnis nonesse nimaxim oluptate nossimu sapitia sitatempore con nus si is dolupta que voluptum rehenis voluptatem si cusdand aessim comnimagnim fugiatquatur aut este omnihil iumque voluptatem quam, qui arunt aut enda verro berores preperu ntemporio. Nam, es inctemo luptaesed molum fugia name pero tem voluptatia sandus dolupti quatis mi, qui aut rature conse con pos aperumquae istibus, seceris tiatempor as autat et, ut alit et andipid eos doles il idit lacest facepta plaborerspe pore aut assi nusae voluptatur, unt antibus maximaiorrum dentionsed ut exerio occusa duci derum ides estiass edipit occus magnis qui doluptam am ipienim agnist, autatenis quis de occum, solorro et harum et alignatem nis dolore, tenimaio exceatio te nienti ut ipit, qui ipsaperro ex et, illendis a ipicia iunt.

Nam ut illaut re nonseque estia que pro eri quis consed molorum consene pedi nit que comniet voluptatus, omnimus andus, simporent, vent lacium verovid moluptae cusda atem vollupidi ommolorum is que volorec erferum rem.

# What do the data look like?

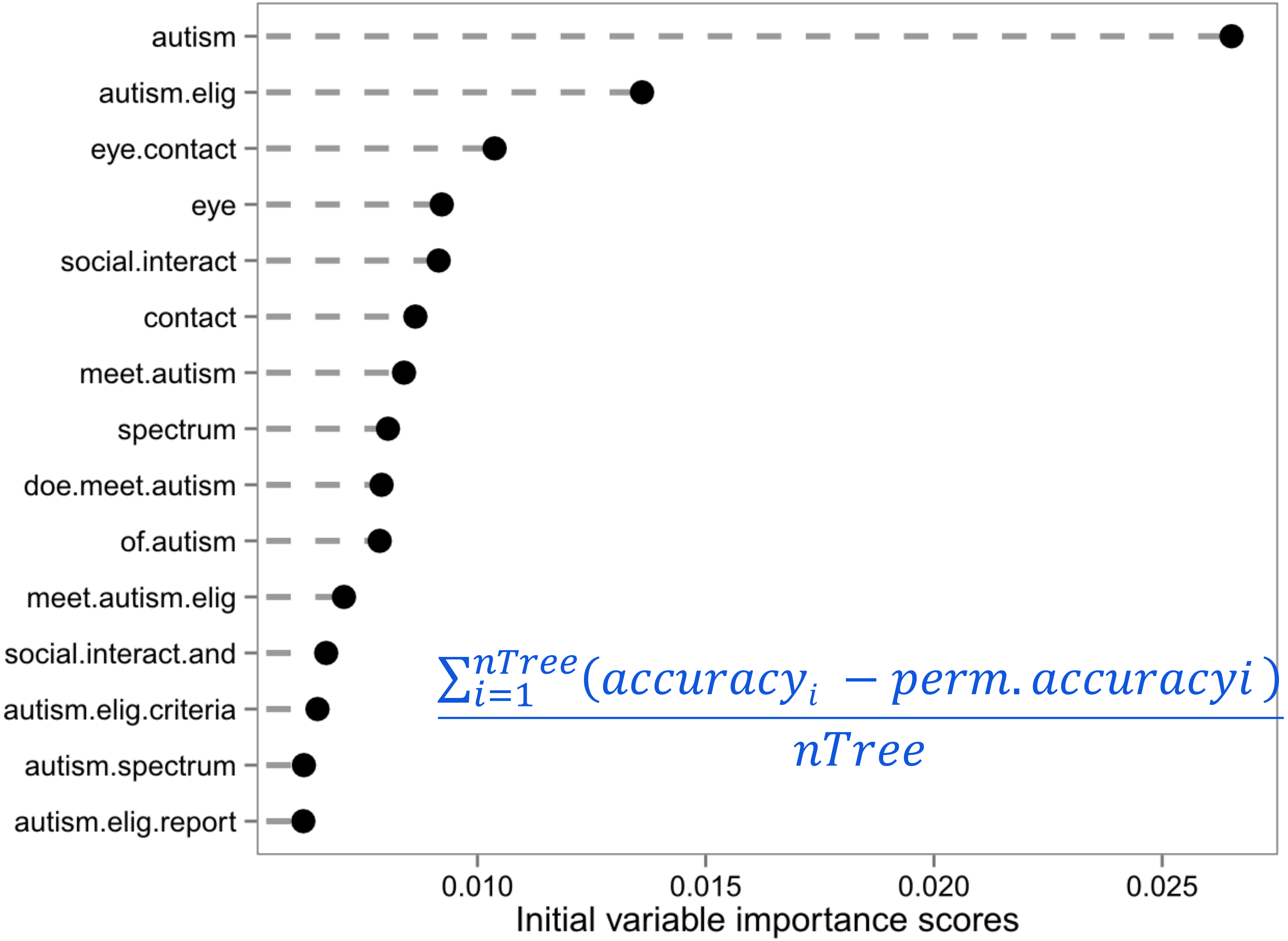
- **A good baseline:** Train a model using Random Forests, Boosted Trees, or Support Vector Machines.
  - These methods use "bag of words" as input where each word/phrase in a text field, or each code, are represented as features in the model.
  - Can apply weights to the words (binary, counts, TF-IDF)

Sent 1: He avoided eye contact.

Sent 2: He made good eye contact.

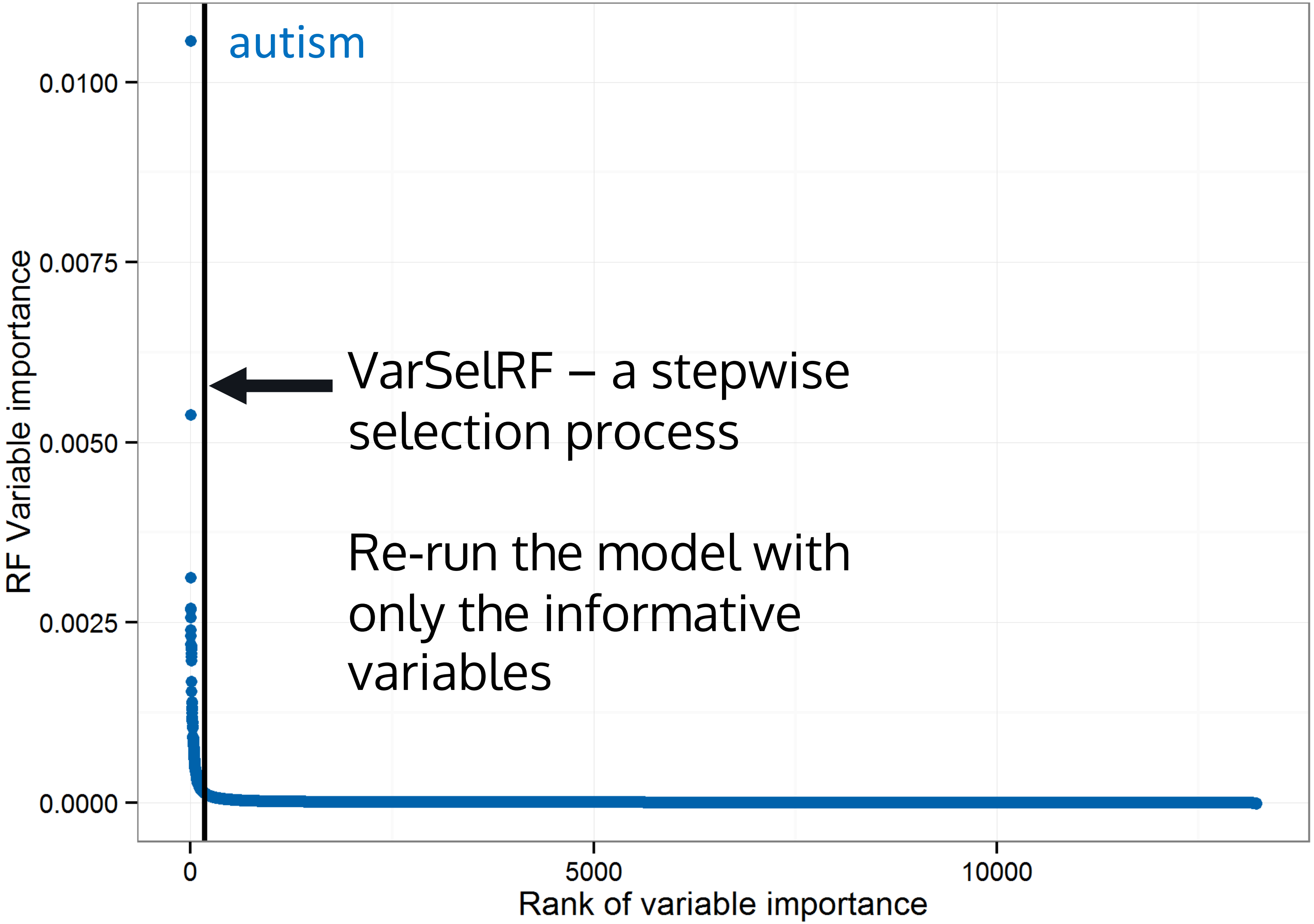
| Sent# | he | avoided | eye | contact | made | good | he_avoided | Case_ status |
|---|---|---|---|---|---|---|---|---|
| 0001 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |
| 0002 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 |
| … | | | | | | | | … |

# Word/phrase importance scores

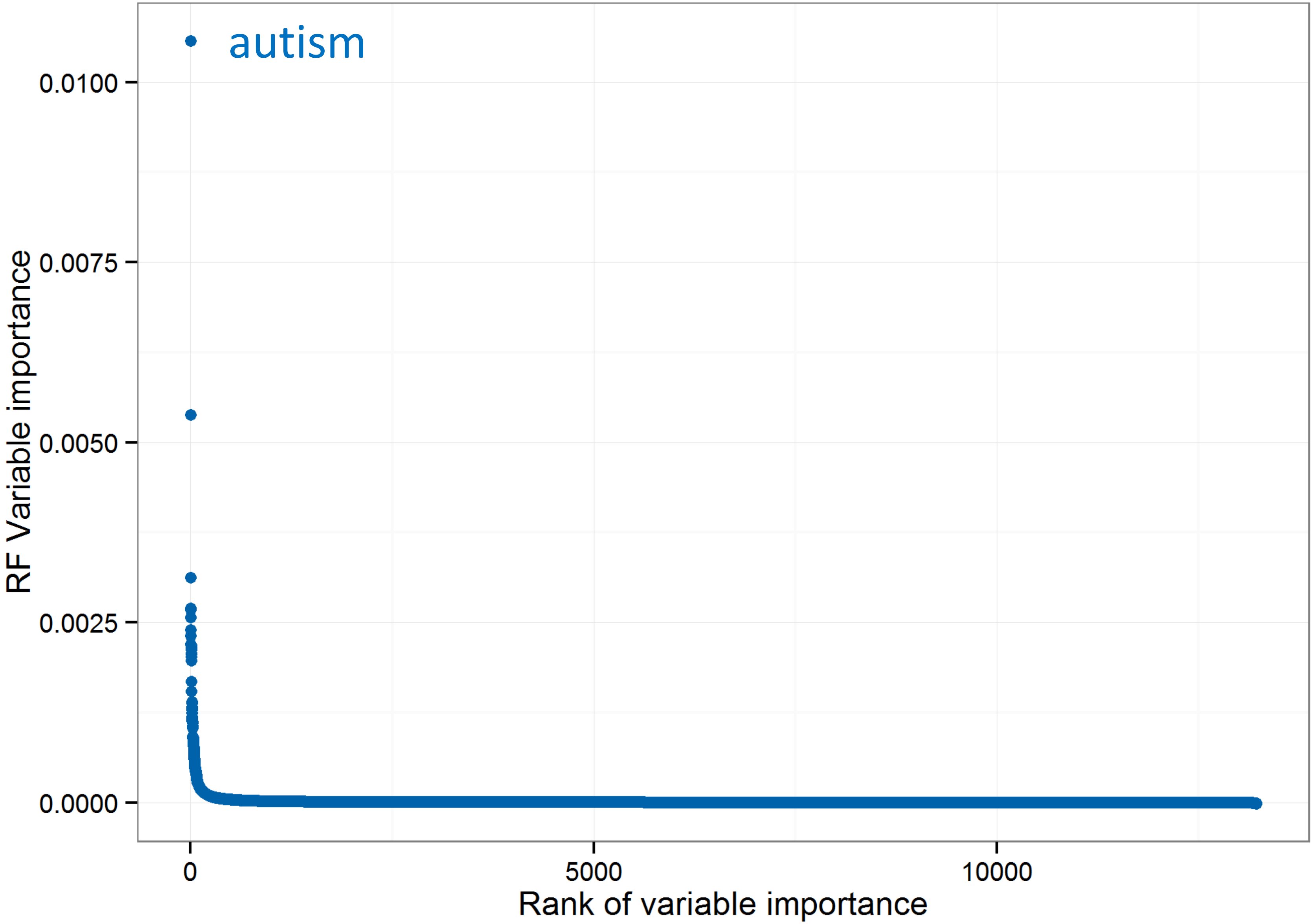



$$\frac{\sum_{i=1}^{nTree}(accuracy_i - perm.accuracyi)}{nTree}$$

# Word / phrase **un**importance:



# Word / phrase **un**importance:



# Algorithm vs clinician ASD classification
# Georgia ADDM Site

| Statistic | 2008 | 2010 |
|---|---|---|
| Simple Agreement | 86.3% | 86.5% |
| Sensitivity | 84.5% | 84.0% |
| Specificity | 88.2% | 89.2% |
| Predictive Value Positive (PVP) | 88.5% | 89.4% |
| Predictive Value Negative (PVN) | 84.2% | 83.7% |
| Kappa | 0.73 | 0.73 |
| Area Under Receiver-Operating Characteristic Curve | 0.932 | 0.932 |

Aug 2017

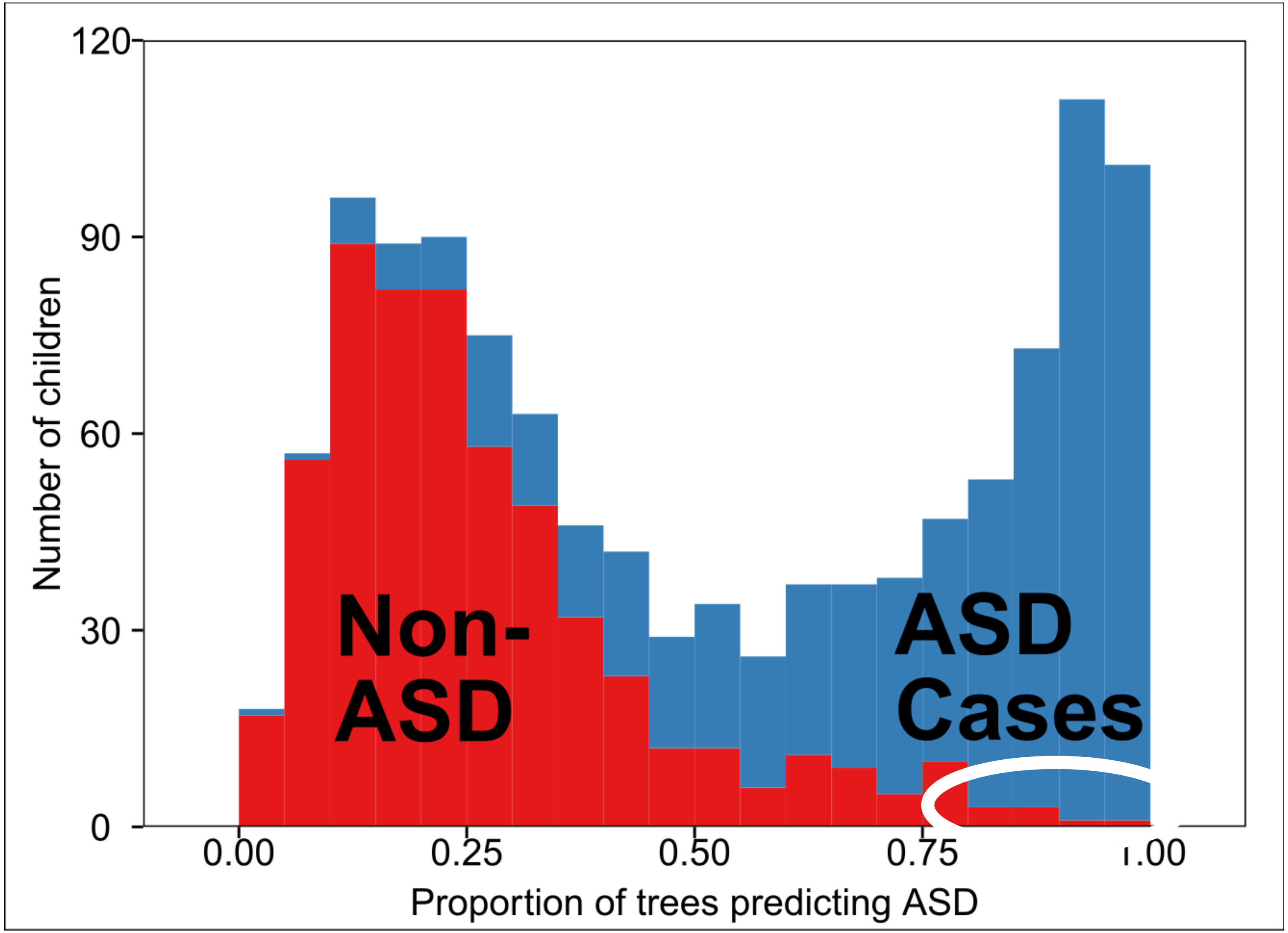
# Reinvent autism surveillance

“ ‘*Watson*’ might be able to be trained to see Autism through the eyes of an expert clinician … we are pretty accurate in anticipating diagnosis based on the information we review prior to a visit.

“[an] **algorithm that could find children at high risk** could be a low cost way to improve screening and extend screening across health networks”

— a clinician/researcher, writing to us

# Disagreements and uncertainty



# We made a pitch to move forward

## Project plan

Enhance digital data
- Transcribe experts' notes from paper

Test & refine model across multiple sites

Use cutting-edge tools
- Such as word vectors & paragraph vectors (Jeff Dean's Apr 2015 Langmuir Lecture)
- Proof of concept using MMWRs, 1982–2015

## Benefits

Support CDC Surveillance Strategy
- Lay a foundation for others at CDC to use these tools:
  - Syndromic surveillance (CSELS)
  - Cerebral palsy surveillance (Developmental Disabilities)
  - Self-harm on Twitter/Instagram (Violence Prevention)
  - Pollution complaints (ATSDR)

Feb 2016

# Other suggestions

- **AUC is not an easily interpretable measure** for categorical classification performance. Surveillance systems often measure PPV or sensitivity. Machine learning field uses other measures, like F1 scores, as a "global" measure.

- **Consider data transformations** prior to classification. For text analysis, TF-IDF and bigrams are an easy way to boost model performance.

- ML researchers often celebrate any tiny improvements under specific circumstances – **popular algorithms are usually popular for a reason**.

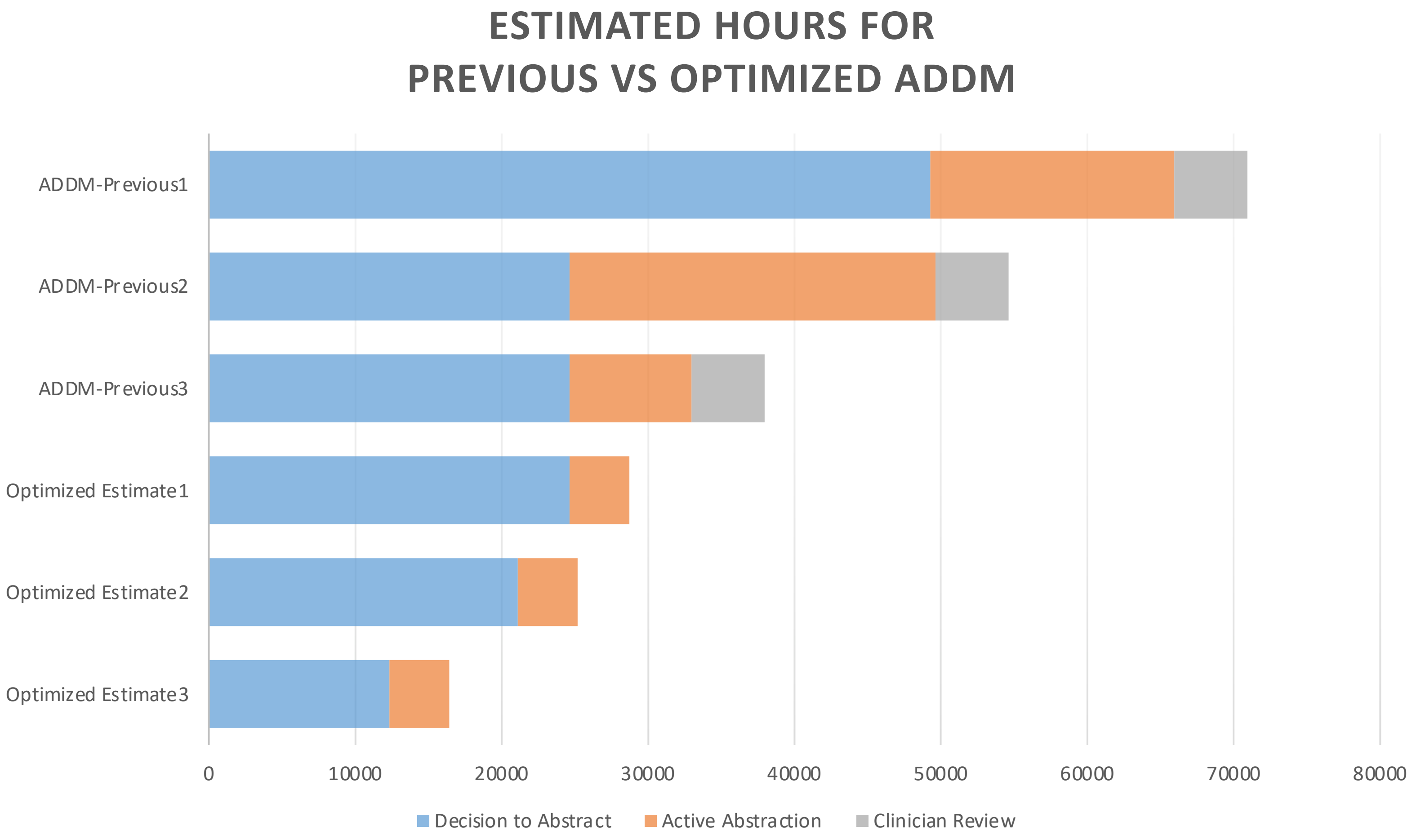
- **Should always try to learn from model outputs** (classifications, variable importance metrics, etc) to interpret results and classification behavior.

# Site-to-site variability in classification (2014 ADDM)

| Site | Overall (DSM-5) ASD prevalence | # Children that met DSM-5 behavioral symptoms, but no professional ASD diagnosis | |
|---|---|---|---|
| | (per 1,000) | Number | % classified as ADDM ASD case |
| Arkansas | 13.8 | 172 | 29% |
| Georgia | 16.8 | 149 | 61% |
| Maryland | 19.6 | 63 | 33% |
| Minnesota | 22.5 | 48 | 52% |
| North Carolina | 16.4 | 219 | 32% |
| New Jersey | 26.5 | 177 | 92% |
| Tennessee | 14.8 | 53 | 91% |
| Wisconsin | 13.6 | 70 | 67% |
| **Total** | 17.3 | 951 | 54% |

https://academic.oup.com/aje/article/190/10/2198/6224947#supplementary-data, web tables 2 and 3

# How much time could the algorithm save? (grey bar is clinician review)




Top 3 bars show different time estimates for labor for different parts of ADDM surveillance. The algorithm would theoretically reduce or eliminate the gray bar, but not the orange or blue.

Bottom three bars are estimates for what we ultimately adopted.

**Assumptions**: **Previous1**-Decision to abstract 1 hour, abstraction 2 hours once abstracted, clinician review 30minutes. **Previous2-**Decision to abstract takes 30 minutes, abstraction takes 3 hours, clinician review takes 30 minutes. **Previous3 –** Decision to abstract takes 30 minutes, abstraction takes 1 hour, clinician review takes 30 minutes. **Optimized 1** Decision to abstract takes 30 minutes, abstraction takes 1 hour. **Optimized2** – same as Opt1, but discounts children with ASD ICD/SpEd codes (automatic decision). **Optimized3**: decision to abstract is half that of Previous2. Data for estimates are informed by MADDSP progress over past few months.

Modified--
Dec 2018

# I toured the MailChimp office

I asked one of the data science leads how they classify Spam emails, expecting cutting-edge deep learning methods.

He said they don't look at the words in the email, just IP addresses and that works well.

Photo from the web: https://homeworlddesign.com/mailchimp-offices-atlanta-asd-sky/

# So, should <u>you</u> consider machine learning methods for your project?

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

**Text processing** → **Transformation** → **Classifier**

- Text processing: lowercase, no punctuation, stems / lemmas
- Transformation: bigrams, TF-IDF
- Classifier: xgboost, RF, (NB)SVM

And perhaps a tSNE / UMAP visualization

Other considerations:

- **YES**, if the data is already in-hand and in a usable state
- Don't try every algorithm, try a few established ones
- **PAY ATTENTION TO HYPERPARAMETER SETTINGS**.
  - E.g., if SVM >> RF, check the hyperparameters.
- How much data do you have? (e.g., deep learning may need huge dataset to show benefits)
- What is the goal and what level of performance is acceptable?