Rapid classification of autism for public health surveillance

Matthew J. Maenner, PhD

Epidemiologist, Developmental Disabilities Branch

U.S. Centers for Disease Control and Prevention

August 2017 – International Society for Disease Surveillance

"Laws, like sausages, cease to inspire respect in proportion as we know how they are made."

-John Godfrey Saxe

http://en.wikipedia.org/wiki/John_Godfrey_Saxe

This presentation will address

- Diagnosing autism and tracking autism prevalence
- Automating autism surveillance with machine learning
- Practical considerations for real-world use

Autism Spectrum Disorder (ASD)

A group of neurodevelopmental disorders diagnosed based on observed behavior[1]

- Impairments in social communication
  - e.g., lack of eye contact, inability to hold a conversation
- Presence of repetitive behaviors or restricted interests
  - e.g., motor stereotypies, narrow interests, routines

No established biomarkers

First described in 1943; formal criteria in *DSM-III* (1980), revised *DSM-III-R, DSM-IV, DSM-5*

The "gold standard" is expert clinical judgment

		Truth		
Dx by DSM-III-R		AD	Not AD	*n*
	AD	19	32	51
	Not AD	1	148	149
		20	180	200

Clearly, there is no marker that can be used to diagnose autism without error (i.e., there is no true gold standard). It should be emphasized that this is a problem for the evaluation of any diagnostic criteria in psychiatry, not only for autism (Robins, 1985).

Szatmari, JADD 1992

Clinician reliability—*DSM-5* Field Trials

Target DSM-5 Diagnosis and Field Trial Site	Intraclass Kappa	95% CI	Interpretation
Autism spectrum disorder[b]			
Baystate	0.66	0.51–0.79	Very good
Stanford	0.72	0.54–0.86	Very good
Pooled	0.69	0.58–0.79	Very good

Regier et al. *Am J Psych* 2013

Subjective interpretations of behavior

"A given act such as hand flapping may be described as stereotypic, self-stimulatory, ritualistic, perseverative, gesturing, or posturing by different clinicians"

-Bodfish et al. 2000

Current preferred assessment tools

- Researchers often use two instruments, which lead to better reliability:
  - *Autism Diagnostic Interview – Revised (**ADI-R**)*
  - *Autism Diagnostic Observation Schedule (**ADOS**)*

- Expensive; ~3.5 hours to administer both
- Not uniformly used in community settings[1]

1. Rice et al. IMFAR 2014 https://imfar.confex.com/imfar/2014/webprogram/Paper17138.html

Current diagnostic practices in research

	ASD	Non-ASD
ADOS met	536	133
ADOS not met	48	205
ADI-R met	450	90
ADI-R not met	134	248
Concordant ADOS + ADI-R met	438	60
Concordant ADOS + ADI-R not met	146	278
SEED ASD criteria met	500	87
SEED ASD criteria not met	84	251

81.4% agree
kappa = 0.60

"Diagnostic instruments alone cannot replace informed clinical judgment when diagnosing children with ASD."

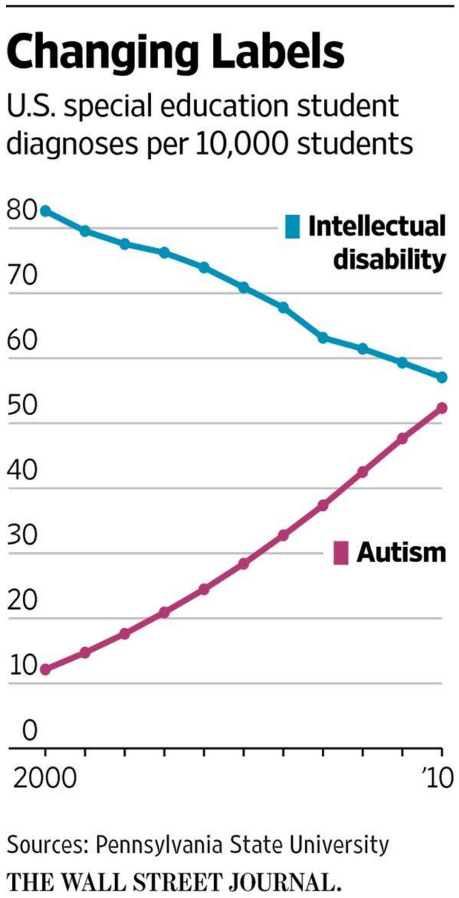
Wiggins et al 2015

Autism prevalence from administrative data

Often linked to education or services.

Autism Special Education Exceptionality

- not equivalent to a medical diagnosis
- introduced in 1992, number of children in category rapidly increased
- Accompanied by decrease in intellectual disability category ("diagnostic substitution", Shattuck 2006)



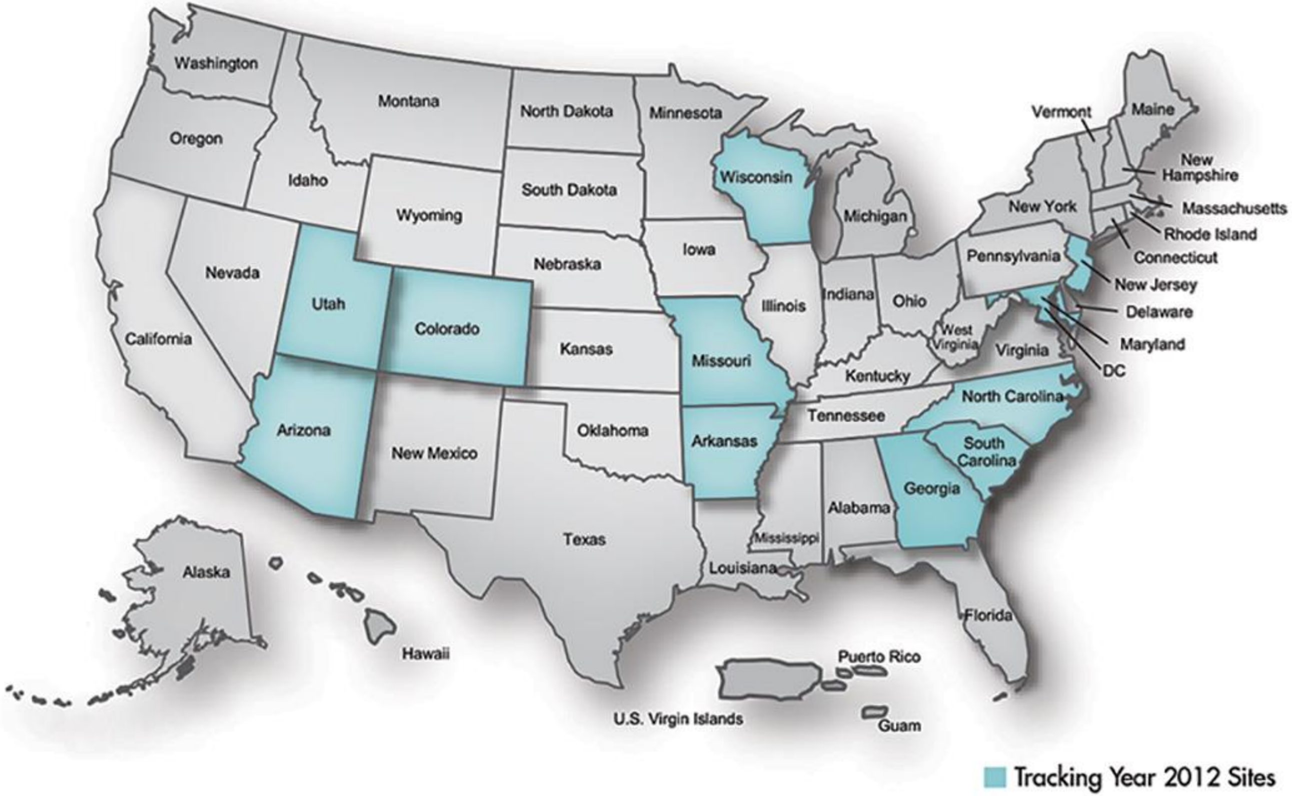
CDC's autism surveillance system

- Children's Health Act of 2000 authorized CDC to develop a program for autism surveillance

The Autism and Developmental Disabilities Monitoring (ADDM) Network

- uses a consistent case definition based on documented symptoms
- does not rely entirely on existing diagnoses

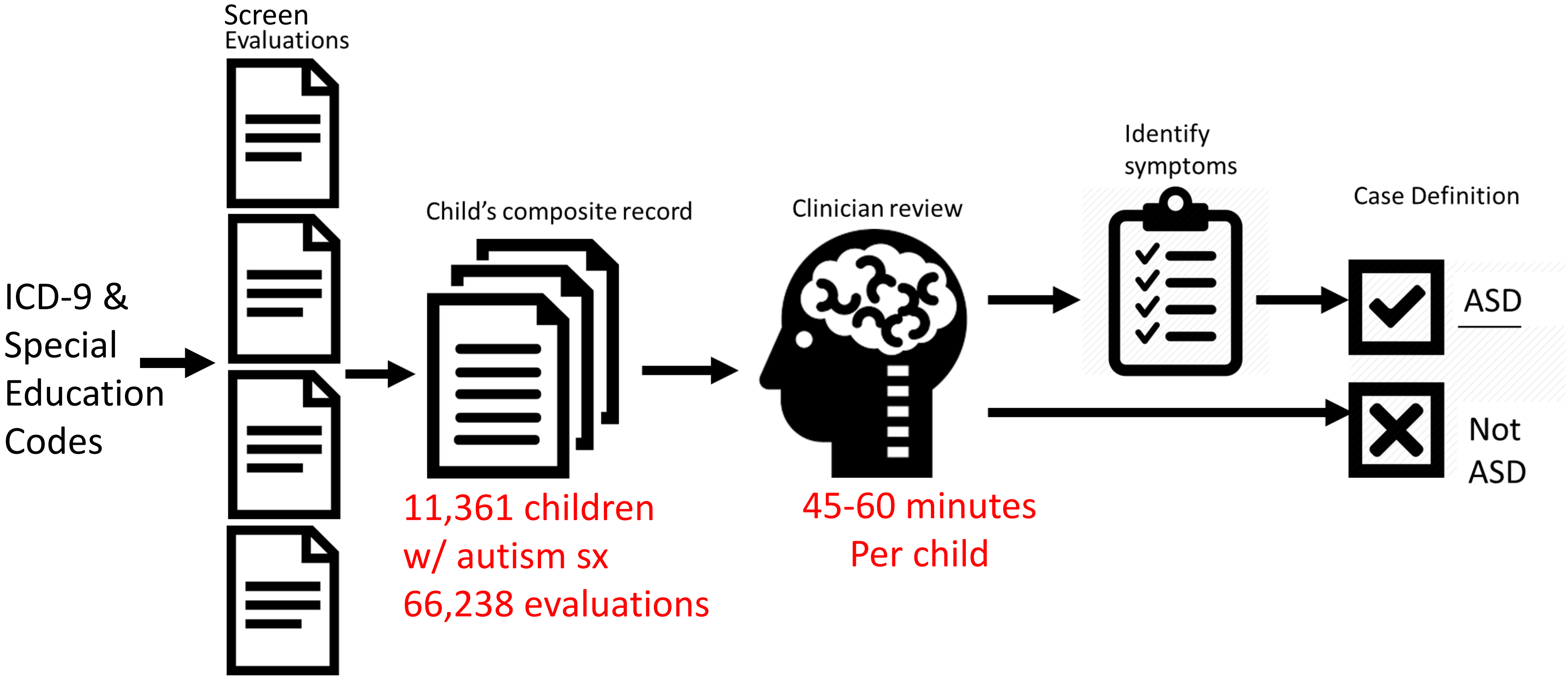
Autism and Developmental Disabilities Monitoring (ADDM) Network Sites



347,000 8-year-old children living in defined geographic areas in 2012

1-year period prevalence for even-numbered years beginning in 2000

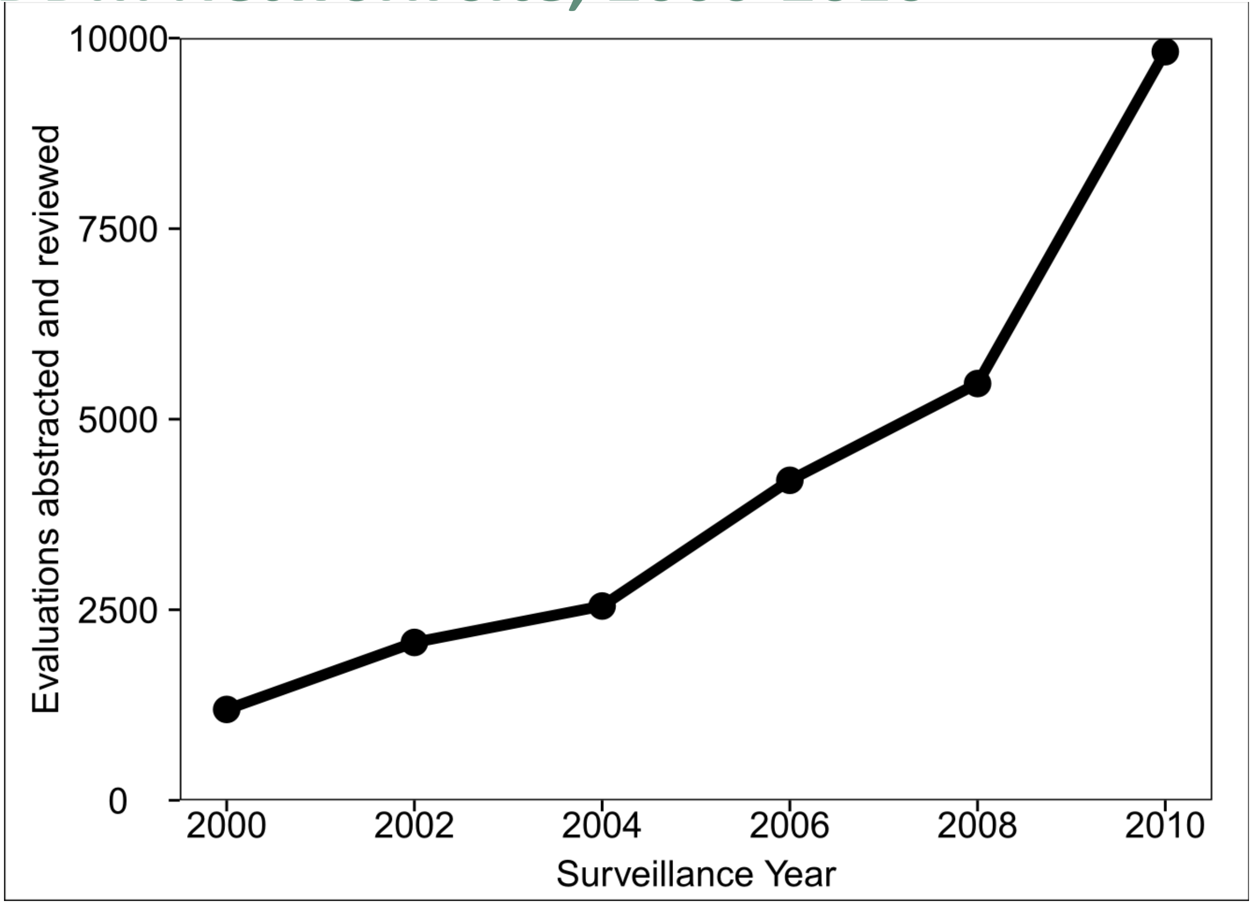
CDC's population-based autism surveillance requires the manual review of ever-increasing numbers of records.



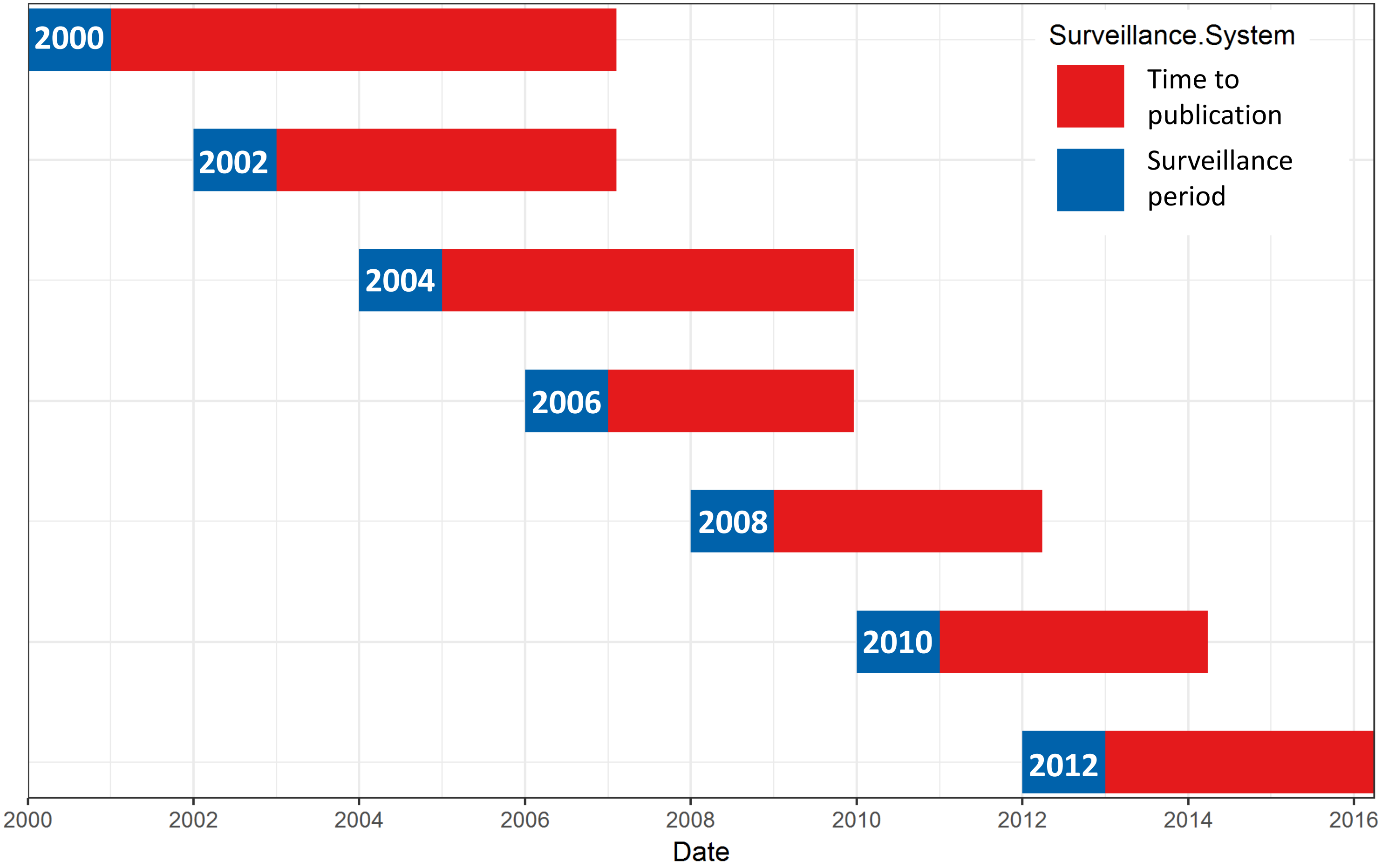
86,110 records

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

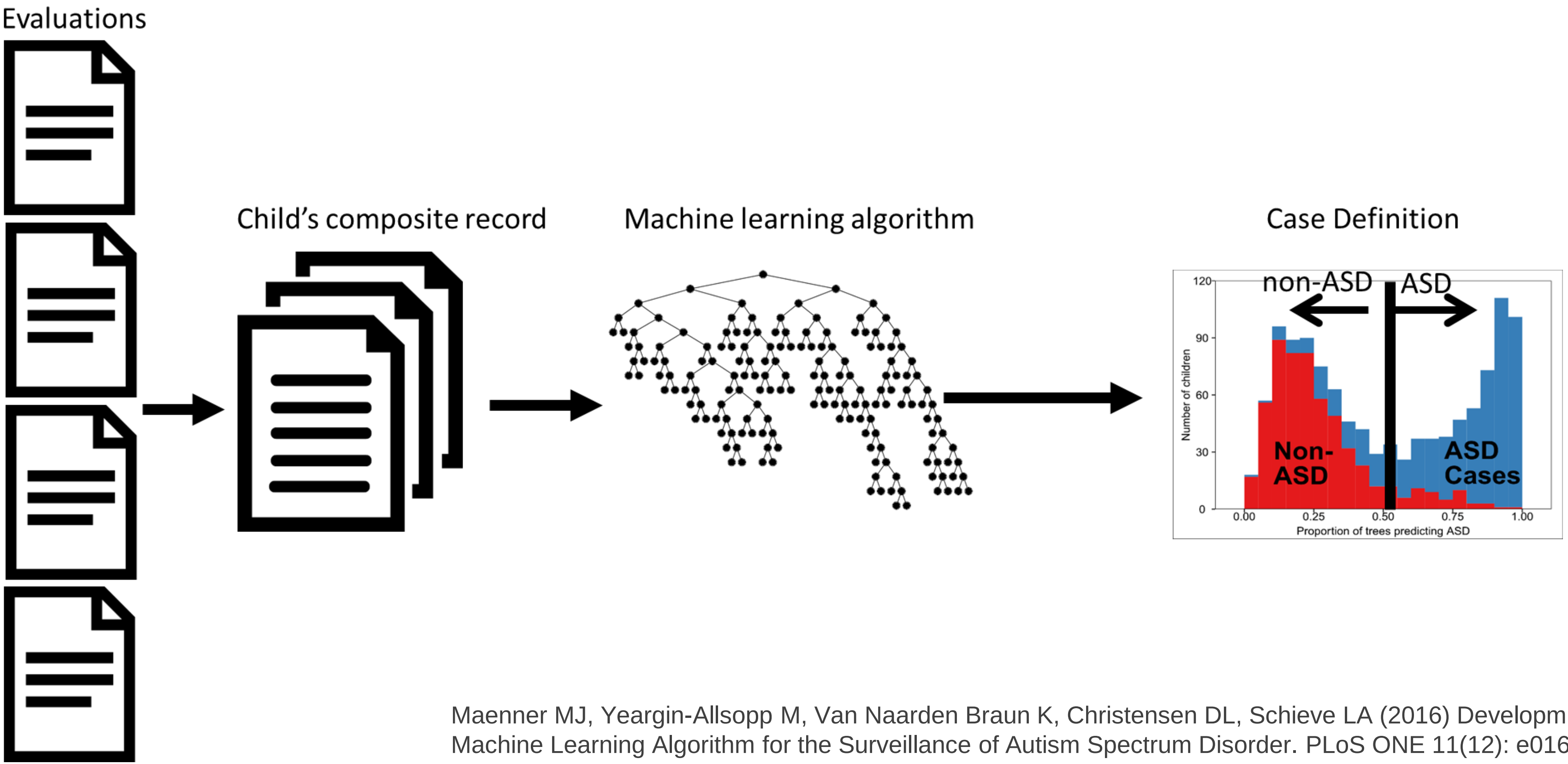


“[ADDM] is in many ways considered a gold-standard measure of autism prevalence, but **it takes a long time to compile that information**.”

-Stephen Blumberg, NCHS

https://spectrumnews.org/opinion/q-and-a/questions-for-stephen-blumberg-tracking-autisms-transience/

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

Classification with random forests

Random Forests[1]

- Ensemble classifier, 10,000 trees initially

Training Data: 2008 Georgia ADDM site

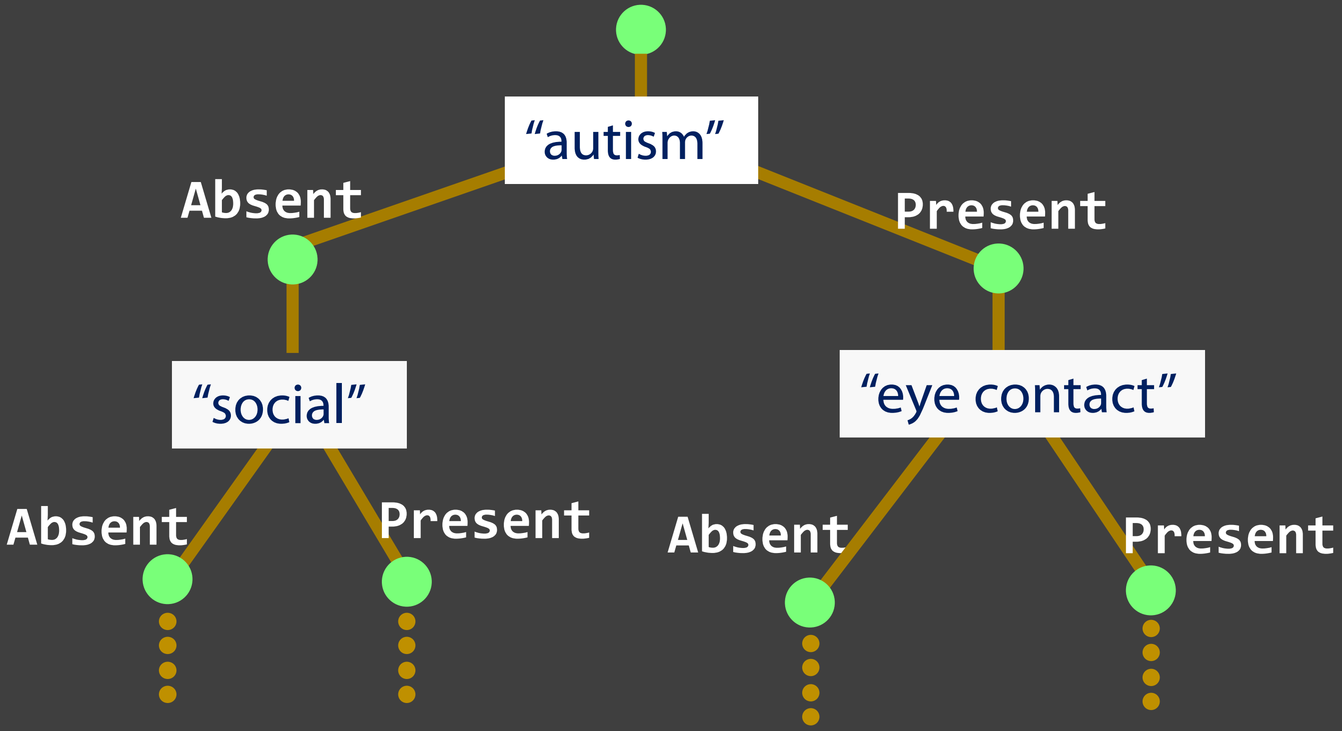
- 1,162 children (601 met ASD case status)
- 5,396 evaluations
- 13,135 1-3 word phrases initially included
    - Each child's evaluations concatenated, stemmed, and used Term Frequency – Inverse Document Frequency weights

Testing Data: 2010 Georgia ADDM site

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

1. Breiman, 2001

Random forests: training one tree



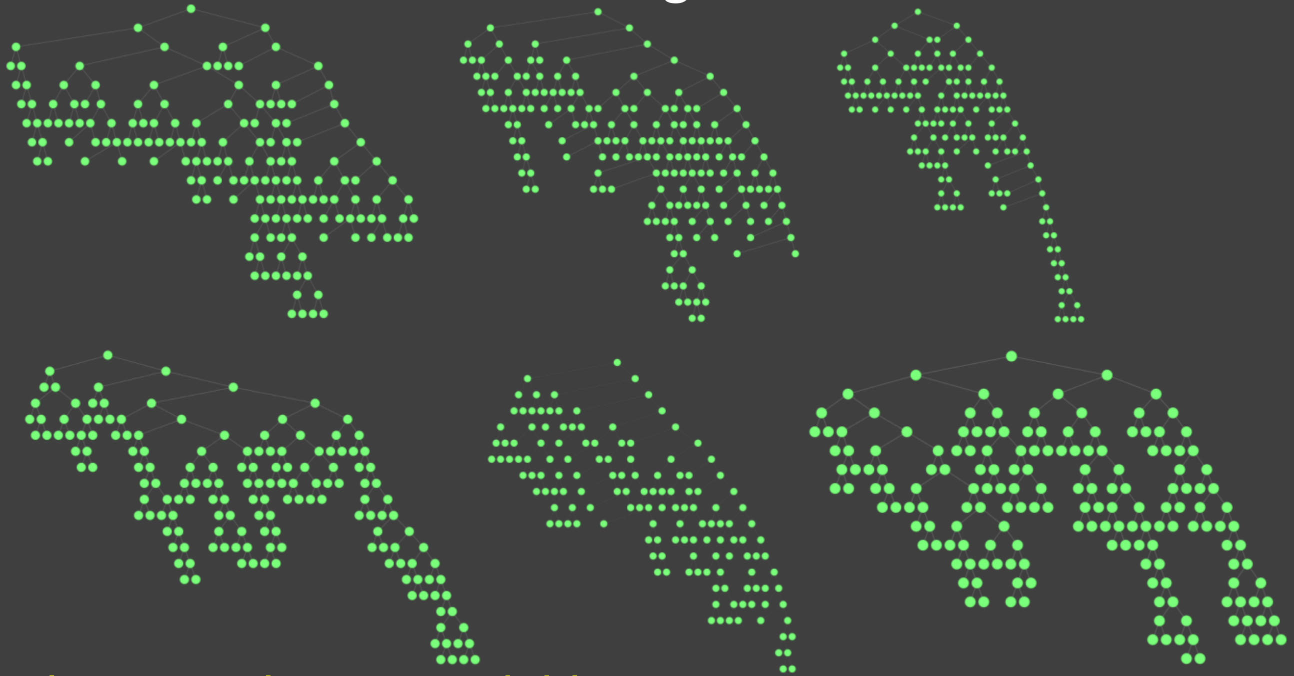
Repeat selection and splitting until tree is fully grown.

Random Forests: classification

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

RF Tree
1 of 10,000

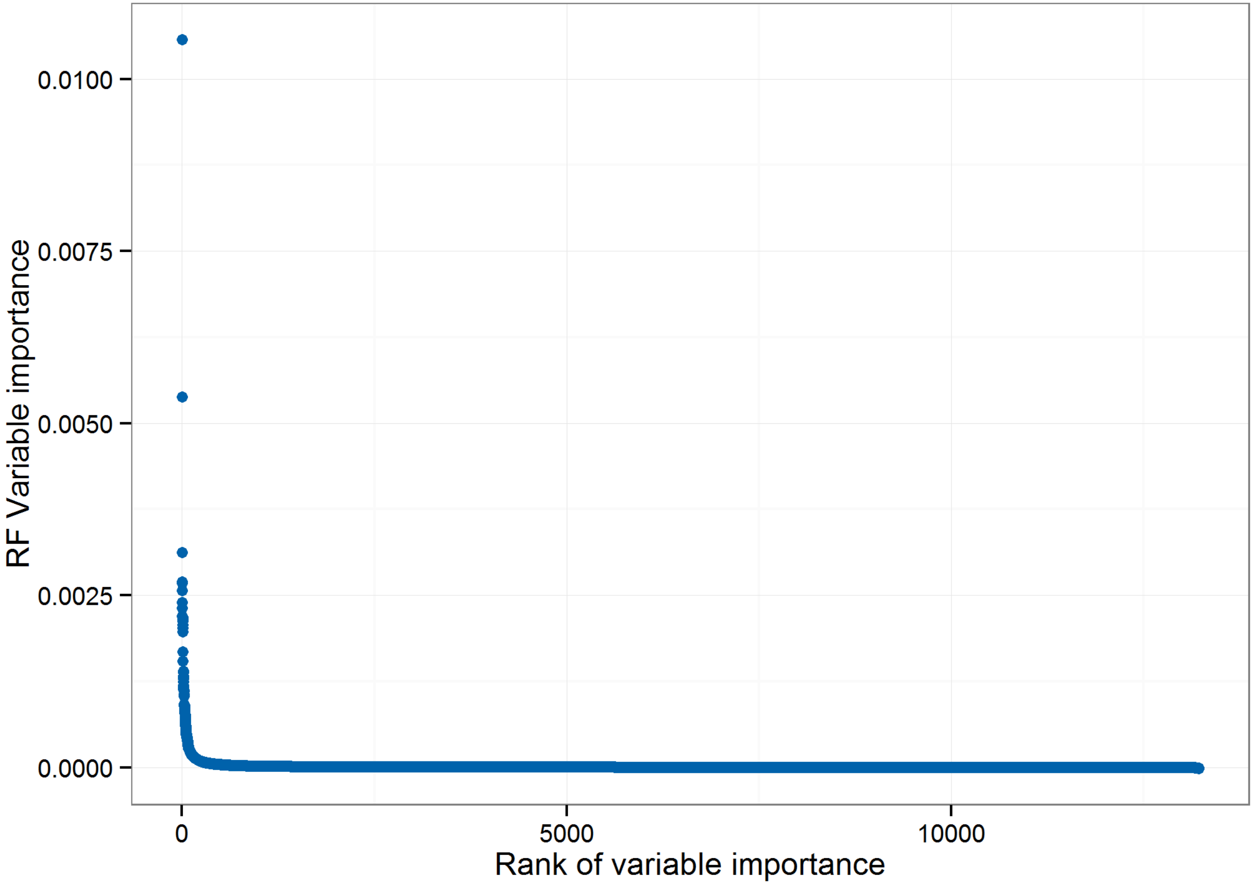
Random Forests: voting on ASD case status



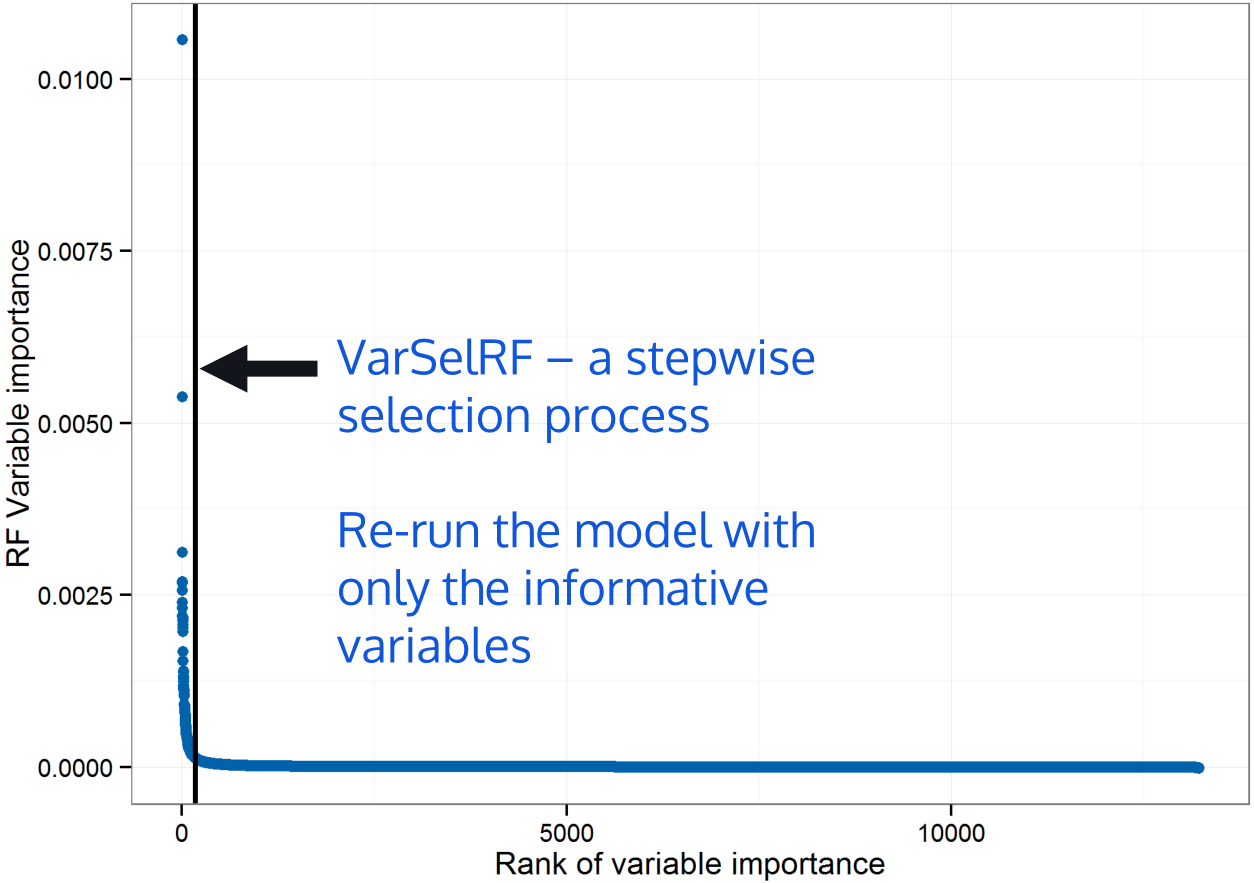
Each tree predicts every child's ASD case status.

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

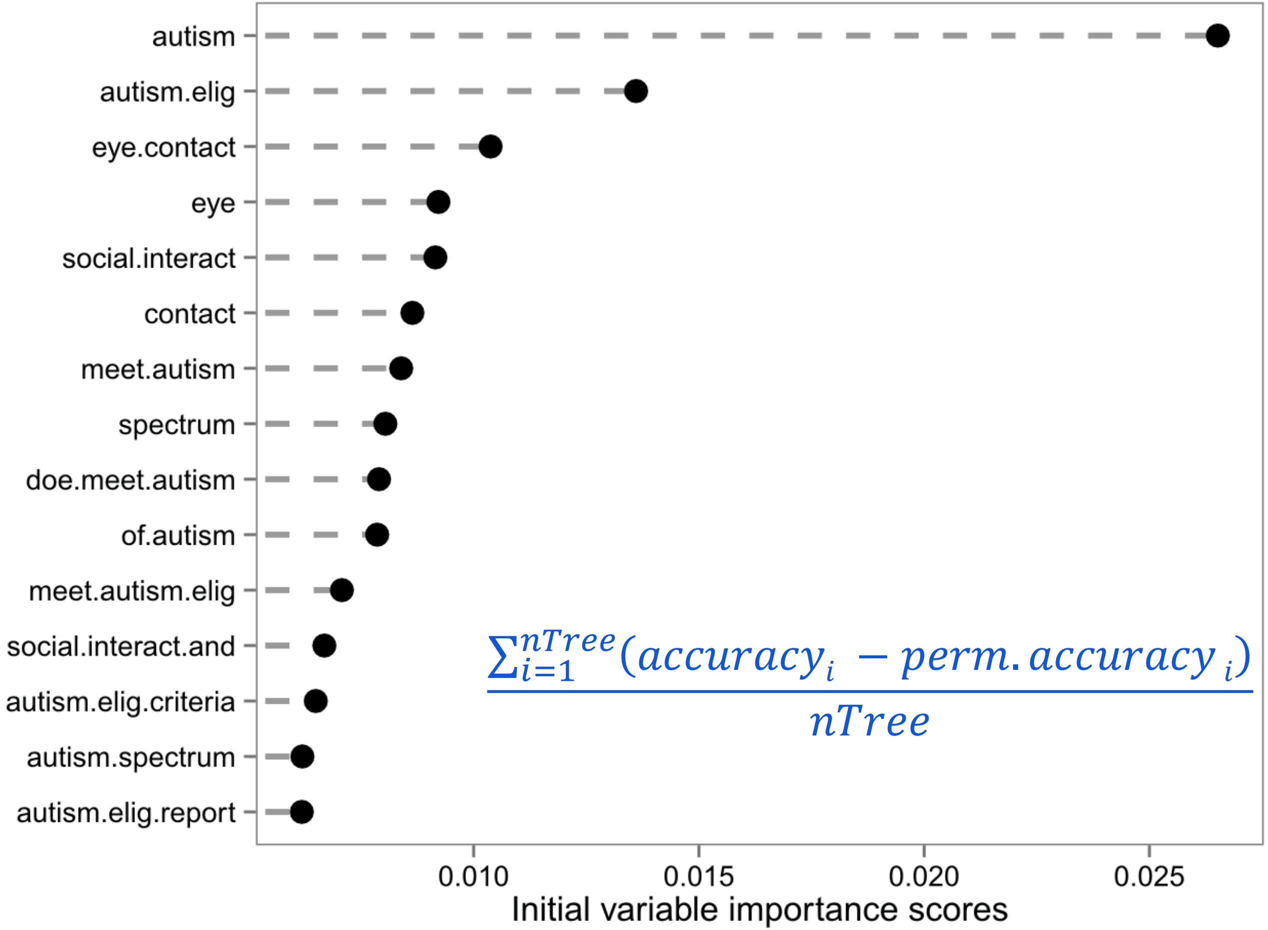
Word / phrase **un**importance:



Word / phrase **un**importance:

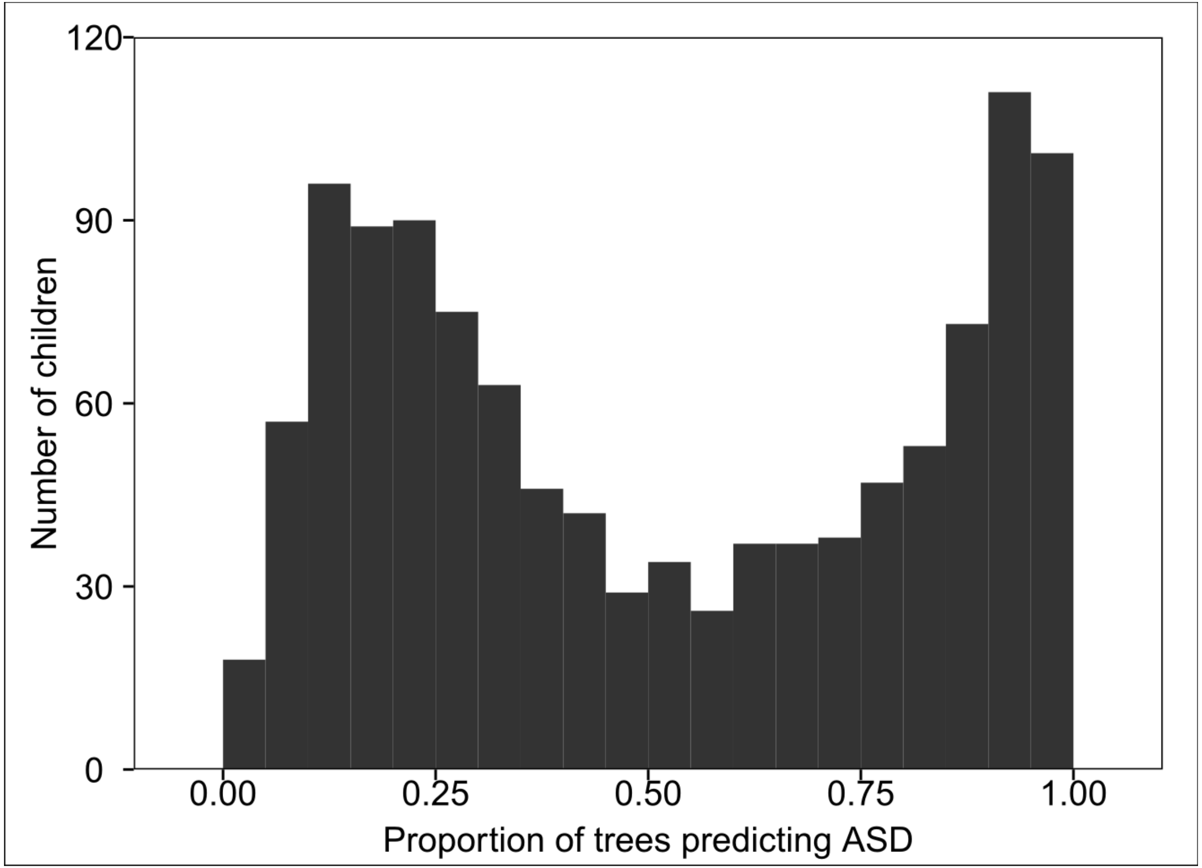


Word/phrase importance scores

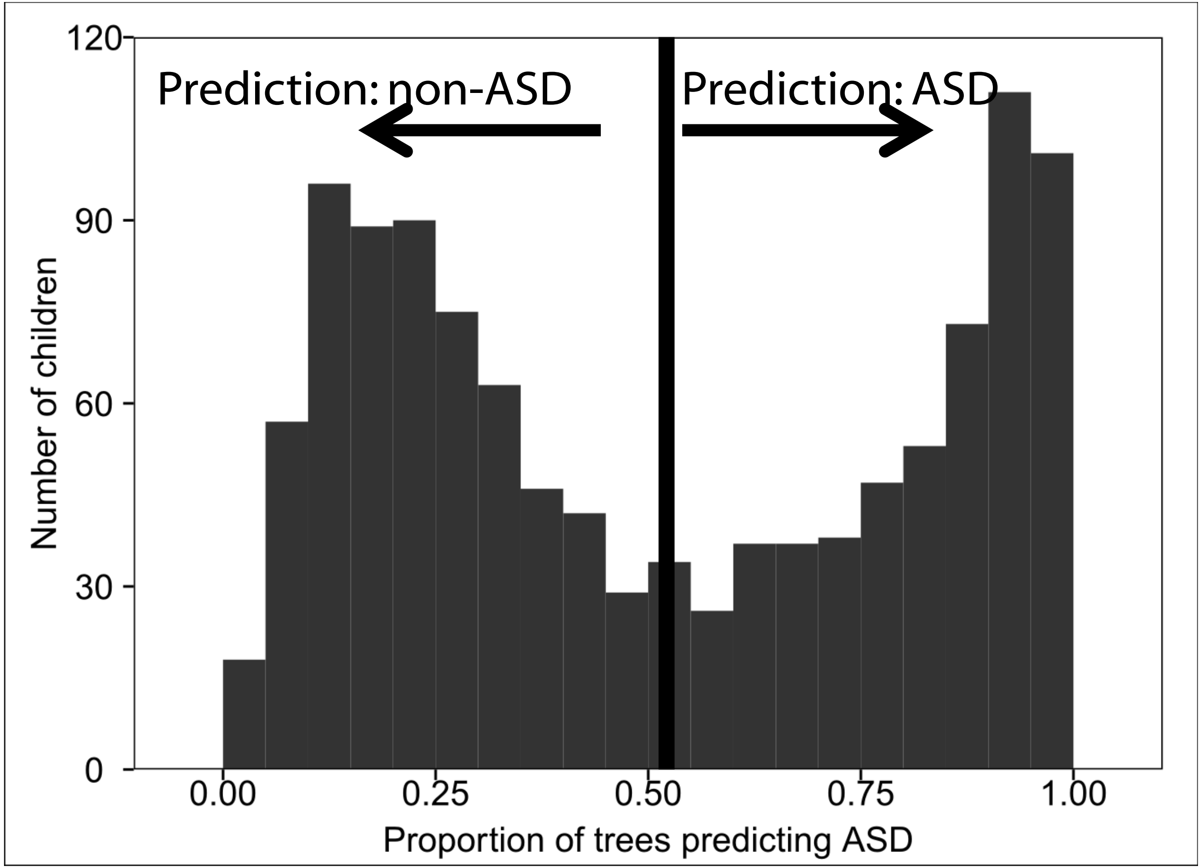


$$\frac{\sum_{i=1}^{nTree}(accuracy_i - perm.accuracy_i)}{nTree}$$

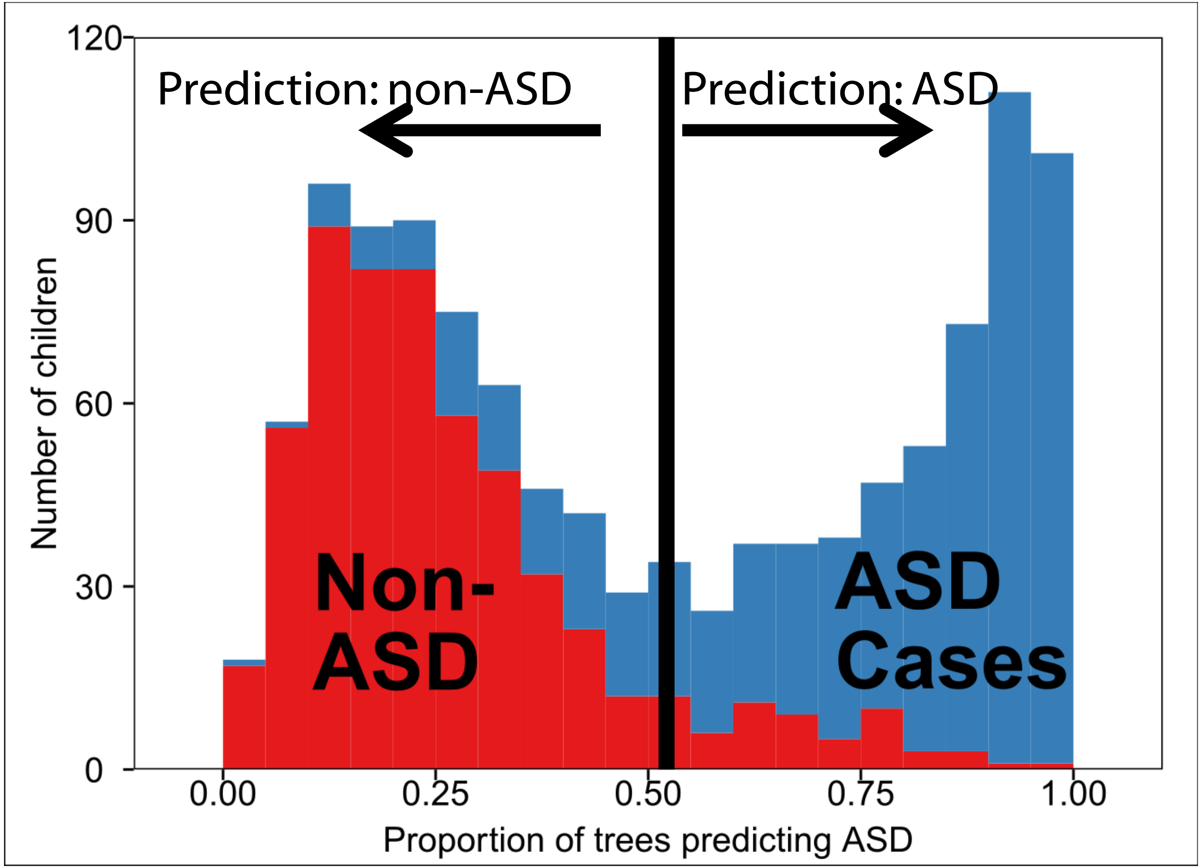
Histogram of ASD prediction scores (N=1,162)

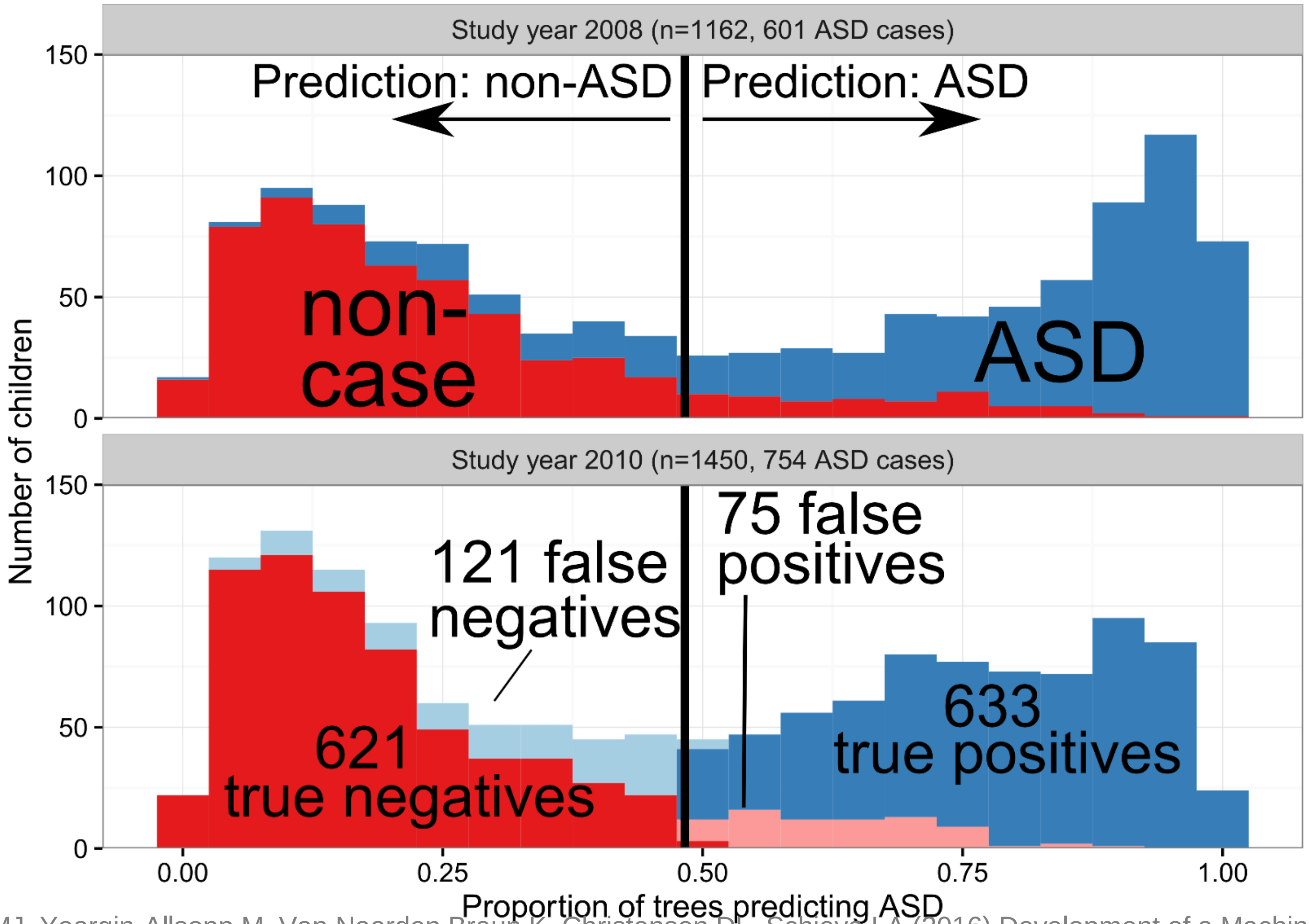


Histogram of ASD prediction scores (N=1,162)



Histogram of ASD prediction scores (N=1,162)





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 vs clinician ASD classification
Georgia ADDM Site

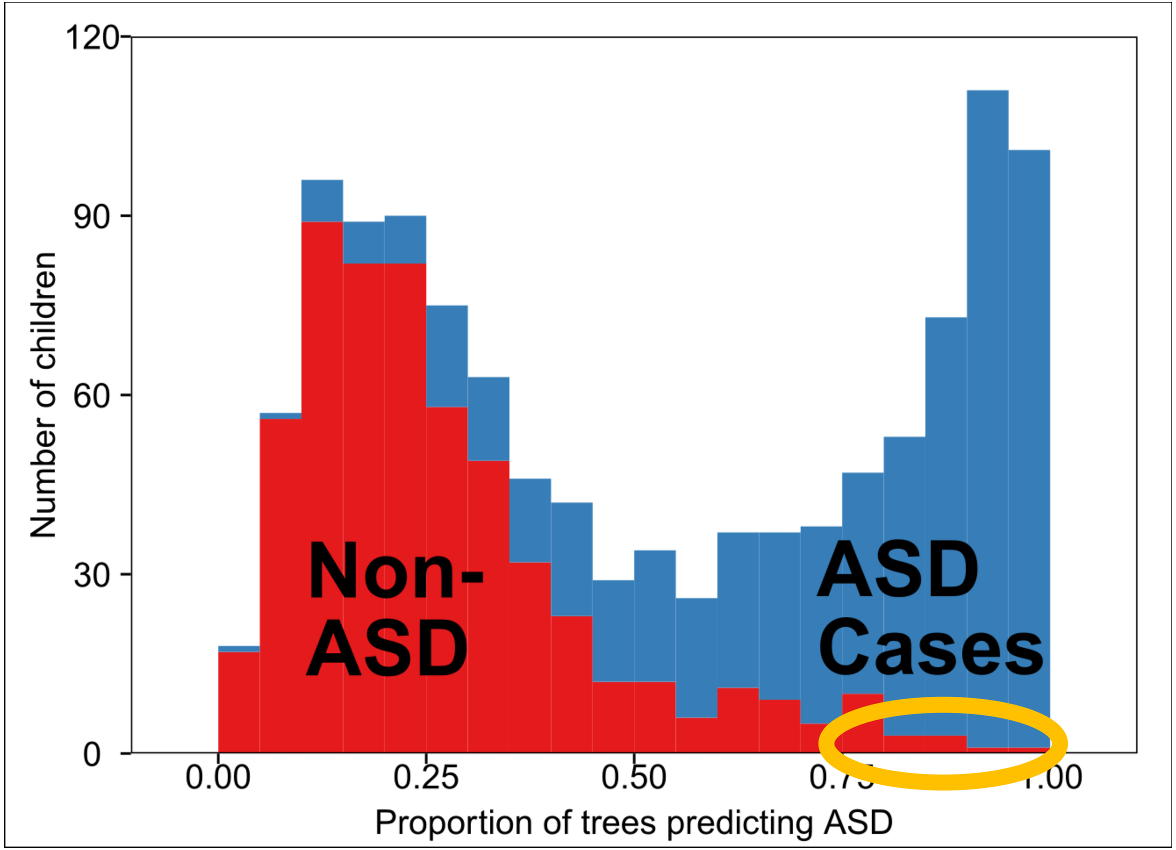
Statistic	2008	2010
Simple Agreement	86.3%	86.5%
Sensitivity	84.5%	84.0%
Specificity	88.2%	89.2%
Predictive Value Positive (PVP)	88.5%	89.4%
Predictive Value Negative (PVN)	84.2%	83.7%
Kappa	0.73	0.73
Area Under Receiver-Operating Characteristic Curve	0.932	0.932

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
Boys	25.4	(23.5-27)	24.1	(22.3-26.1)	0.95
Girls	5.5	(4.6-6.5)	4.9	(4.1-5.9)	0.89
Non-Hispanic White	18.2	(16.2-20.4)	17.4	(15.5-19.5)	0.95
Non-Hispanic Black	14.0	(12.5-15.7)	13.0	(11.5-14.6)	0.93
Hispanic	10.7	(8.7-13.1)	10.1	(8.2-12.5)	0.94

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

Disagreements and uncertainty



CDC Innovation Fund and HHS Ventures Program Project

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)

Project goals

1. Create more refined, **symptom-specific** algorithms
2. Test across surveillance sites and years
3. Make tools and processes scalable and more accessible across the agency (Aligns w/ CDC Surveillance Strategy)

Traditional method: Bag of words

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
…							

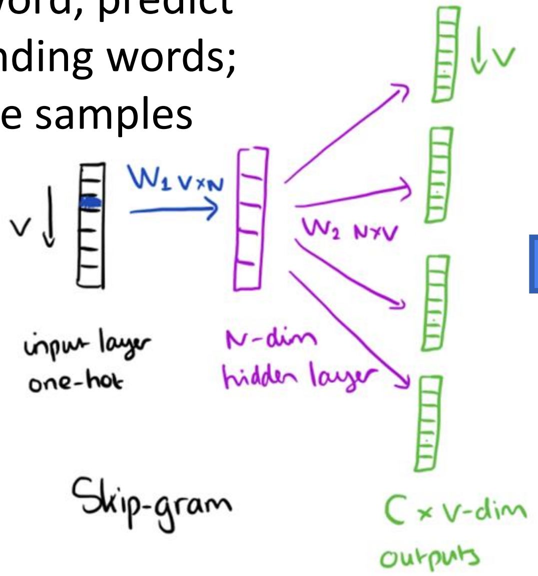
Each word or phrase is a column (variable) in the dataset

Pros: easy to use, variety of established classifiers

Cons: could lead to very "wide" datasets; sensitive to vocabulary changes

Newer methods: Distributed representations

given word, predict surrounding words; negative samples



>>> model['friendly']
array([-0.14460348, 0.22440973, -0.00493282, -0.08833114, 0.0131678 , -0.19822162 ....

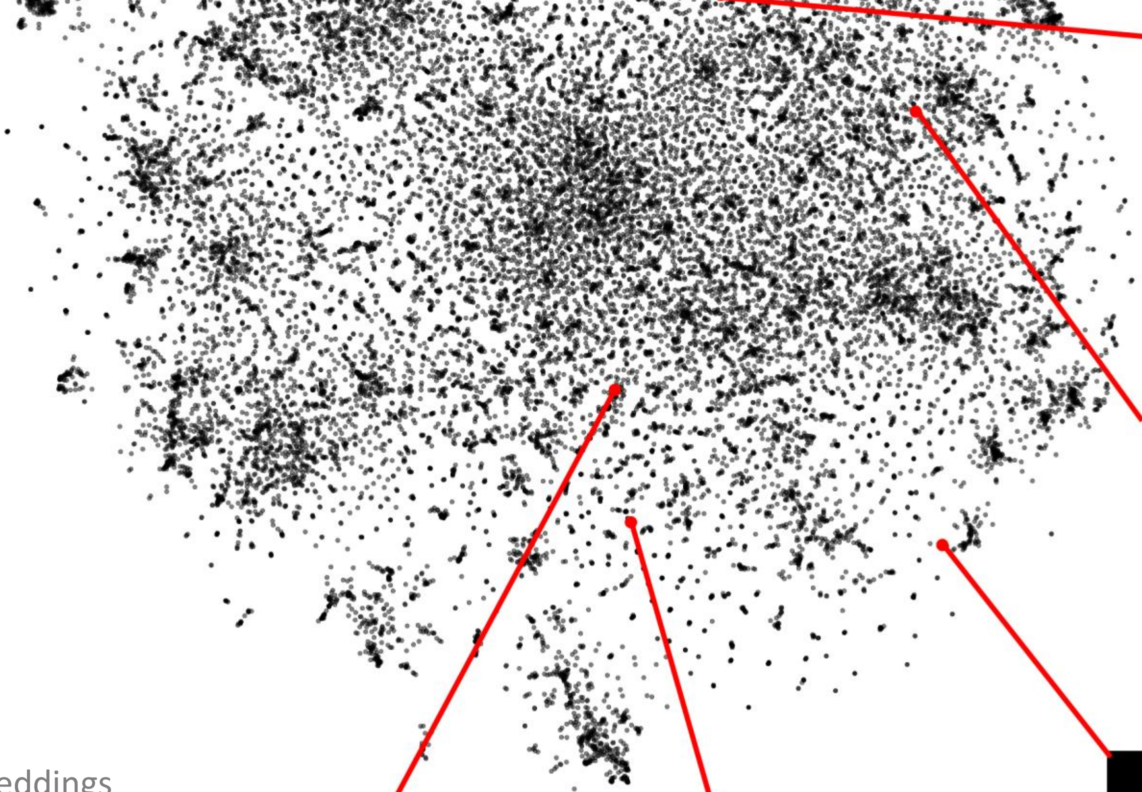
Classifiers:
Facebook Fasttext, and RNN-LSTM or CNN models
(i.e., "deep learning")

Distributed word representations (word2vec, fasttext)

Pros: learn word relationships from larger corpus; use that information in classification task

Cons: new methods; "data hungry"

Quantifying relationships between words



flapping	
hand_flapping	0.59
flicking	0.59
rocking	0.57
flap	0.57
flaps	0.56

rewind	
VCR	0.60
reenact	0.54
video_tapes	0.53
watch	0.52
Tales	0.52

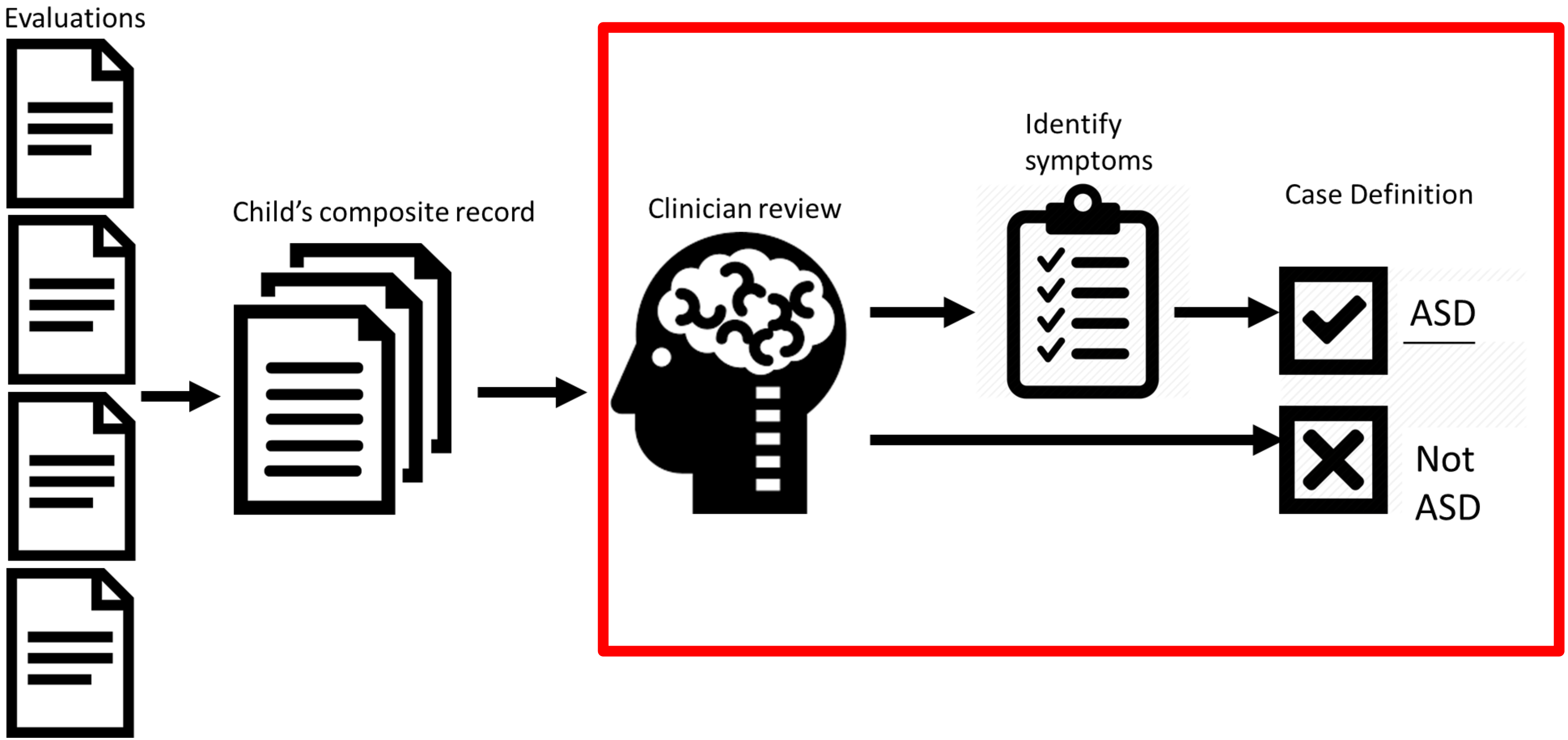
stimming	
self_stim	0.45
tic-like	0.40
self-stim	0.39
squinting	0.39
"stimming"	0.39

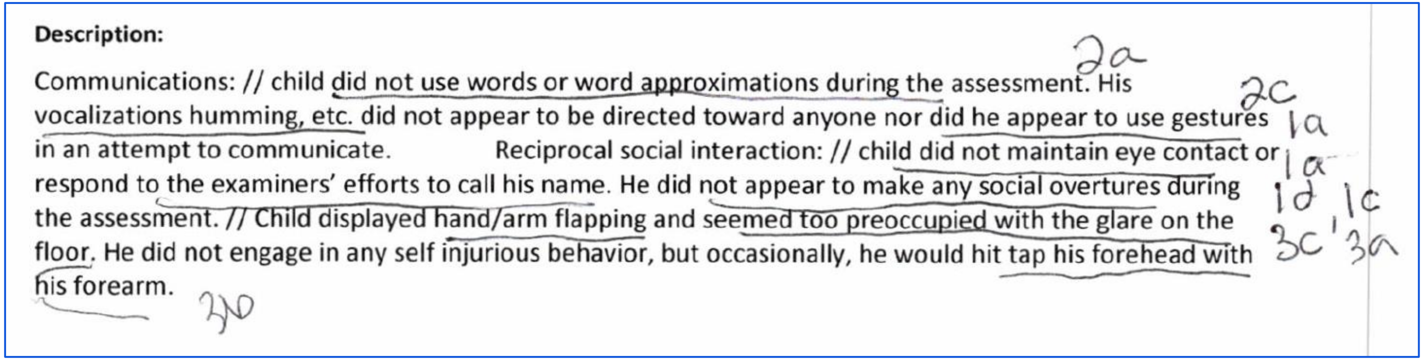
turn-taking	
turn_taking	0.67
reciprocal	0.51
conversation	0.47
reciprocal_conversations	0.45
peekaboo	0.43

direction	
instructions	0.59
directions	0.59
directives	0.48
directive	0.47
verbal_prompt	0.46

Distributed word embeddings (300D word2vec) applied to ~2M words from children's evaluations.
Visualization: 2D tSNE
Similarity: cosine distance

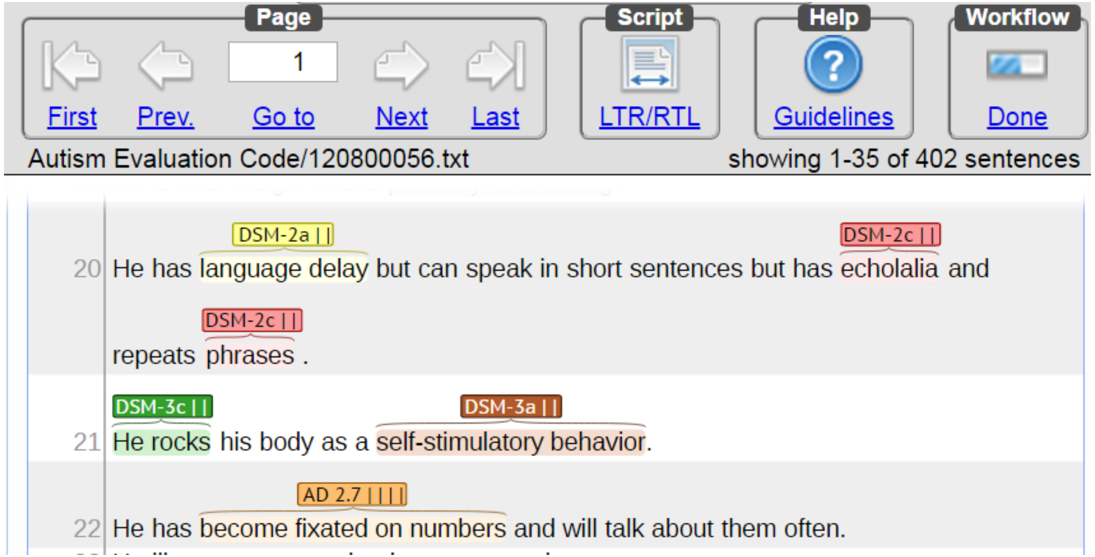
Training a classifier to detect autism symptoms



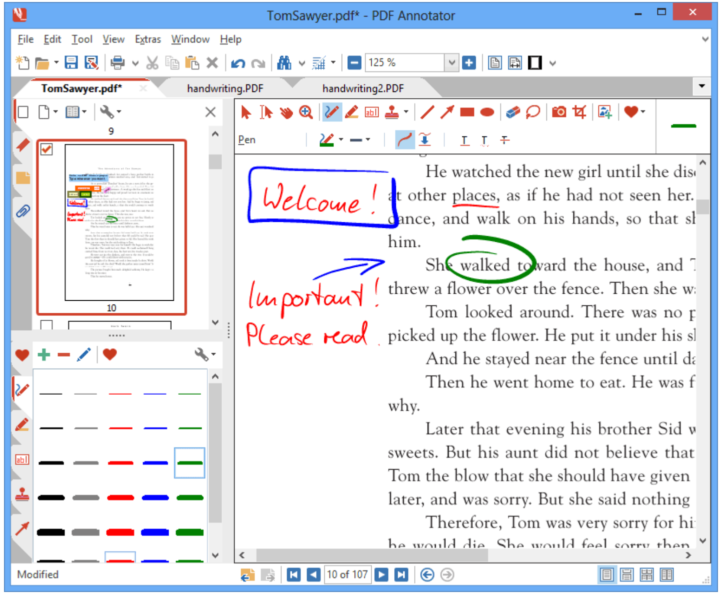

Description:

Communications: // child did not use words or word approximations during the assessment. His 2a
vocalizations humming, etc. did not appear to be directed toward anyone nor did he appear to use gestures 2c 1a
in an attempt to communicate. Reciprocal social interaction: // child did not maintain eye contact or 1a
respond to the examiners' efforts to call his name. He did not appear to make any social overtures during 1d 1c
the assessment. // Child displayed hand/arm flapping and seemed too preoccupied with the glare on the 3c 3a
floor. He did not engage in any self injurious behavior, but occasionally, he would hit tap his forehead with
his forearm.

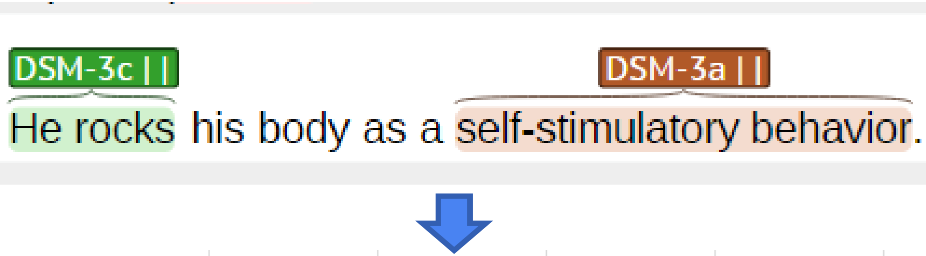
Our first major obstacle was digitizing paper-based annotations



The difference is the underlying data…



VS

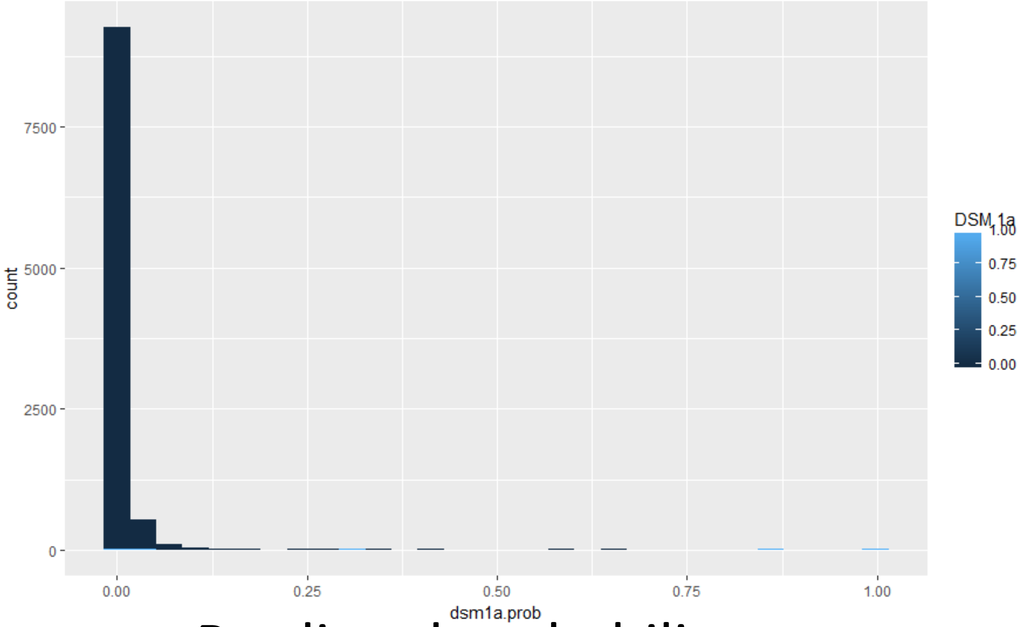


#id=25

#text=He rocks his body as a self-stimulatory behavior.

25-1	He	I-webanno	I-webanno	I-DSM-3c	I-
25-2	rocks	I-webanno	I-webanno	I-DSM-3c	I-
25-3	his	O	O	O	C
25-4	body	O	O	O	C
25-5	as	O	O	O	C
25-6	a	O	O	O	C
25-7	self-stimul	B-webann	B-webann	B-DSM-3a	B
25-8	behavior	I-webanno	I-webanno	I-DSM-3a	I-
25-9	.	O	O	O	C

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.

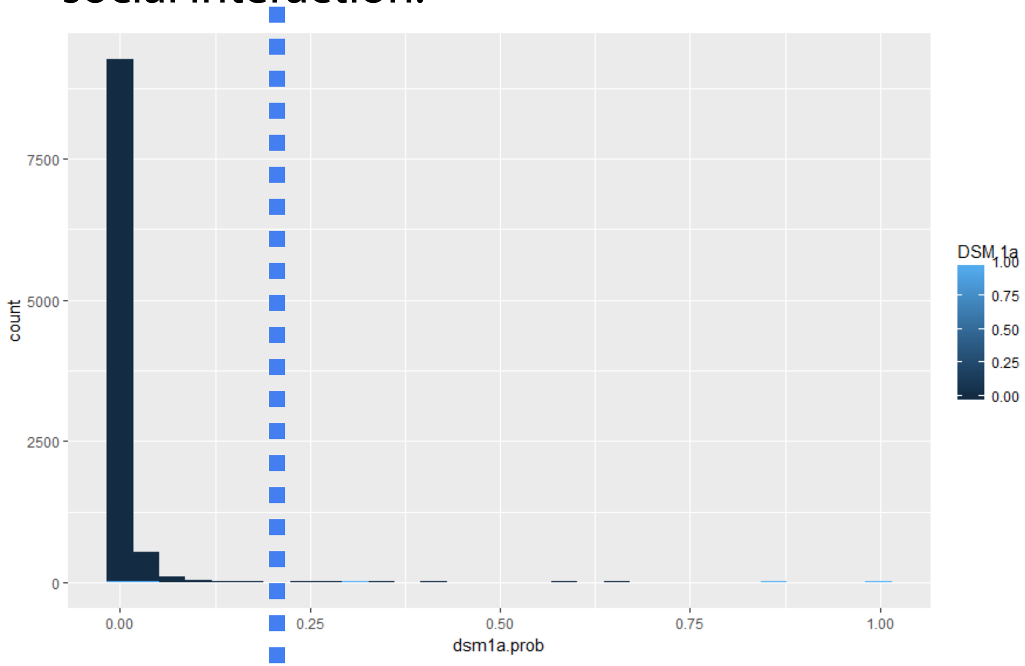


Predicted probability
(1 = symptom present)

(any symptom occurs in a small percentage of sentences – very 'unbalanced' data)

Software: Fasttext
Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759

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.



Fasttext, 100D

	Predict+	Predict-
DSM-1A +	120	49
DSM-1A -	51	9946

Sensitivity: 70.2%

PPV: 71.0%

Cohen kappa: 0.71

ROC AUC: 0.962

Examining the disagreements…

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."
[4] "With regard to behavioral characteristics consistent with
     Autism Spectrum Disorder, child's father indicated that
     child has difficulty using verbal and nonverbal
     communication appropriately to initiate, engage in and
     maintain social contact."
```

Examining the disagreements…

Algorithm: Negative / Clinician: Positive

```
[1] “Patient did not gesture or point to obtain a desired
     object .”
[2] "His expression of affect has been. reportedly restricted,
     but mother also noted that child displays behaviors.
     consistent with empathy as well as a sense of humor."
[3] "Child also appears to have limited visual tracking and
     visual awareness."
[4] "He established fleeting eye contact and often appeared
     disengaged or disconnected from the testing session."
```

Detecting abstract concepts

- For the DSM 1-A example, the phrase “eye contact”—by itself—has a sensitivity of 0.65 and a PPV of 0.61
- Other symptoms will be difficult:
  - *(c) a lack of spontaneous seeking to share enjoyment, interests, or achievements with other people (e.g., by a lack of showing, bringing, or pointing out objects of interest*

```
"Child did not respond to the examiners social smiles or social
overtures."
"He required numerous prompts to participate in the reciprocal activity
of throwing the ball back and forth with the examiner."
"Child reportedly does not greet people unless they are. extremely
familiar."
```

We just need to capture *enough* of these signals to make good predictions

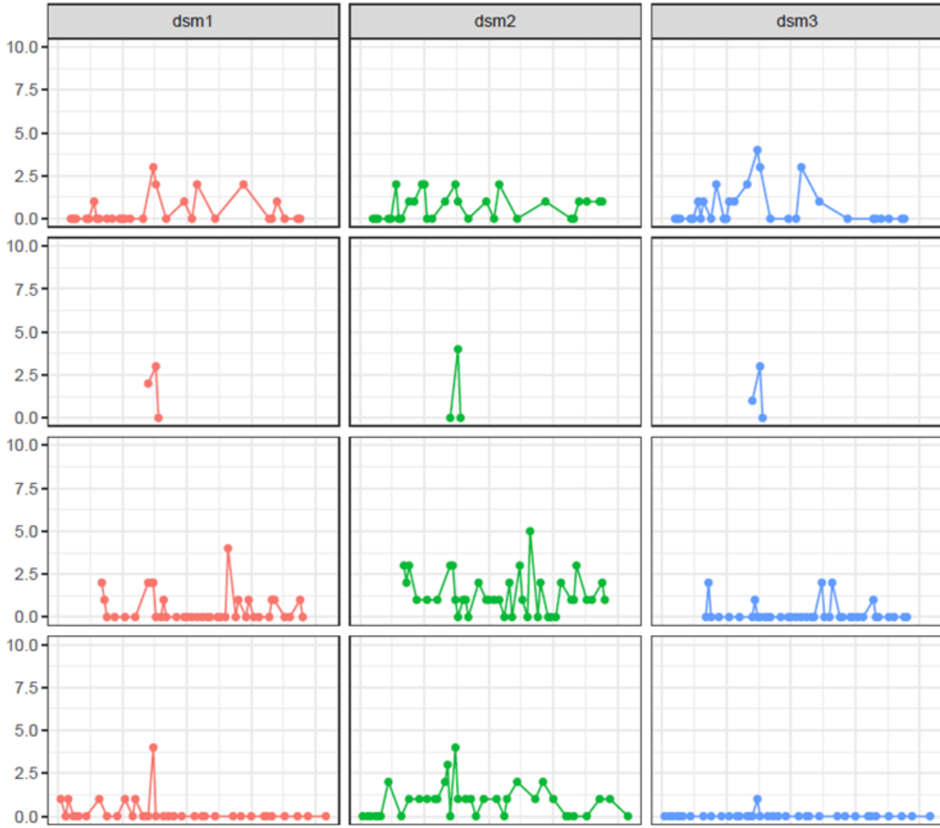
Caveats

- Human inter-rater sentence level reliability is unknown
- Annotations were recorded as on paper
  - Not always precise
  - Some “unknown” or illegible
  - Very complex coding schemes– depends on whether it is the first or subsequent occurrence
- Hunch: better to lump symptoms into groups that are useful for prediction vs studying individual symptoms

On performance…

Now building models / ensembles

- Already observed 1-2% improvement on initial bag-of words models using more years and different algorithms
- Looking at several levels (child, evaluation, sentence)
- Not ignoring non-text information (ICD codes)



(Example of symptom “scorecards”)

On whether we have "big data" or just alot

Data considerations for choosing a method:

- amount
    - 10s of 1000s of annotations
    - 10s of 1000s of evaluations
    - ~5M-10M words
    - 1000s of children in GA ADDM
        - over 1k / year
- Data augmentation/pre-training needs to be relevant to context
- expected performance VS simpler methods, **given the data size**

- ML experts might have different goals and priorities than scientists

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

On choosing the "best" algorithm

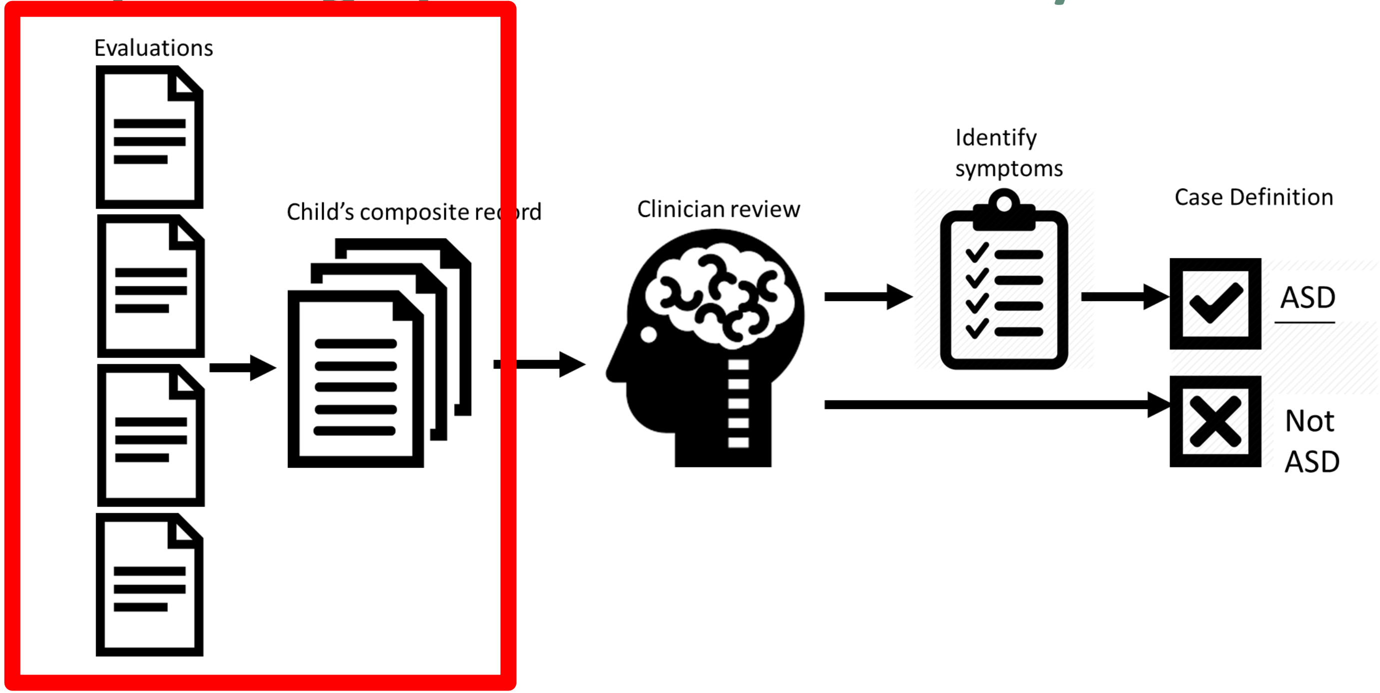
Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?

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)

On speeding up record abstraction / initial screening

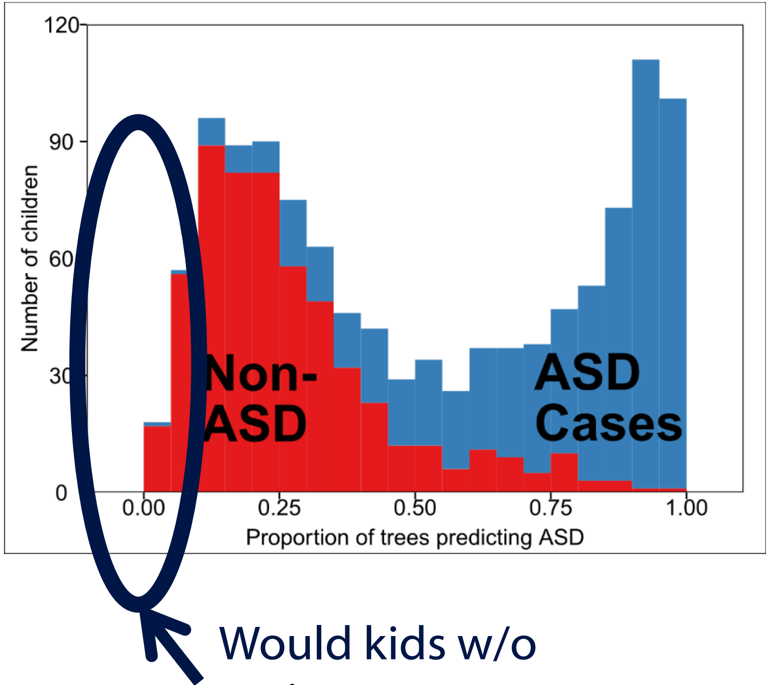


- The initial review of records is done manually, and takes a lot of time
- Potentially seen as less controversial than automating clinician review

On speeding up record abstraction / screening (cont'd)

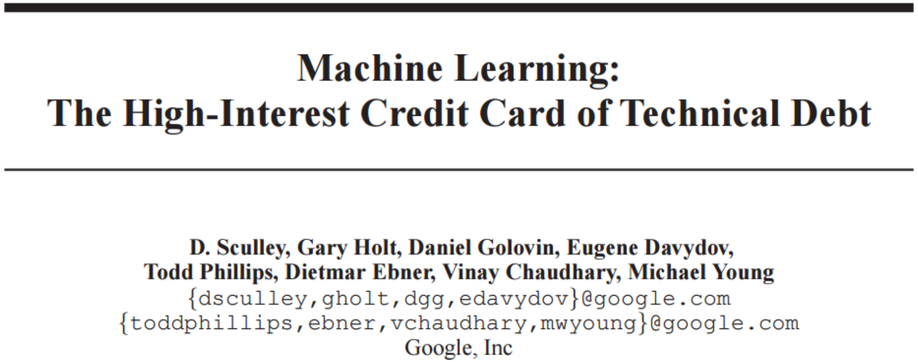
Needs two things:

- Receive evaluation data digitally
  - (people filter records *and* copy text into database)
- A classifier to identify which children likely have ASD



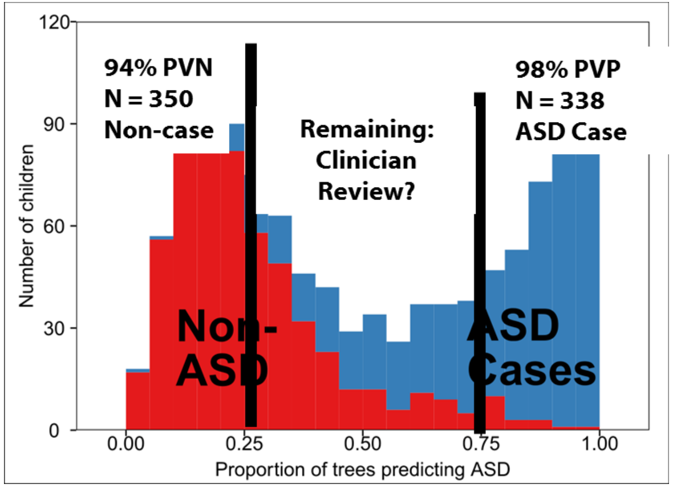
Would kids w/o autism symptoms score here?
I think most **would**.

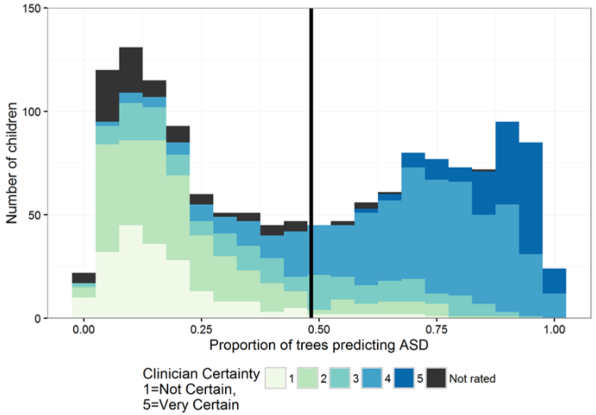
On “replacing clinicians”



Machine Learning:
The High-Interest Credit Card of Technical Debt

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov,
Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young
{dsculley,gholt,dgg,edavydov}@google.com
{toddphillips,ebner,vchaudhary,mwyoung}@google.com
Google, Inc





- Still need people!!1
  - ML could allow clinicians to focus on challenging records
  - Ongoing QC, could adjust based on subsample agreement (two-phase design)

"[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**
Associate Director for Data Science
CSELS, CDC

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

Project Team:

Chad Heilig (Assoc Dir of Data Science, CSELS)
Fatima Abdirizak (NCBDDD)
Nicole Dowling (Branch Chief, DDB)
Maureen Durkin (UW-Madison)
Scott Lee (CSELS)
Laura Schieve (Epi Team Lead, DDB)

Supported by:

HHS Secretary's Ventures Program
CDC Innovation Fund
NCBDDD Division of Congenital and Developmental Disorders

Thanks to our colleagues:

Juliana Cyril (CDC OADS), Bonny Harbinger (HHS OCTO)

Daisy Christensen, Kim van Naarden Braun,
Marshalyn Yeargin-Allsopp

mmaenner@cdc.gov

Word meaning depends on context

"Stimming"

Wikipedia	ADDM
Stimm	Stimulatory
Stimmt	Flapping
Stimme	Stimulating
Stimmel	Flicking
Stimmung	Stimulator
Stimmet	Rocking
Stimmen	Stimulations

"flapping"

Wikipedia	ADDM
Flappie	Stimulatory
Flapped	Spinning
Fluttering	Flicking
Flappet	Rocking
Wingbeats	Posturing
Flutters	Repetitive
Flappy	Excited

(top 7 by cosine distance)