Forecasting ASD gene discovery

Jake Michaelson, PhD

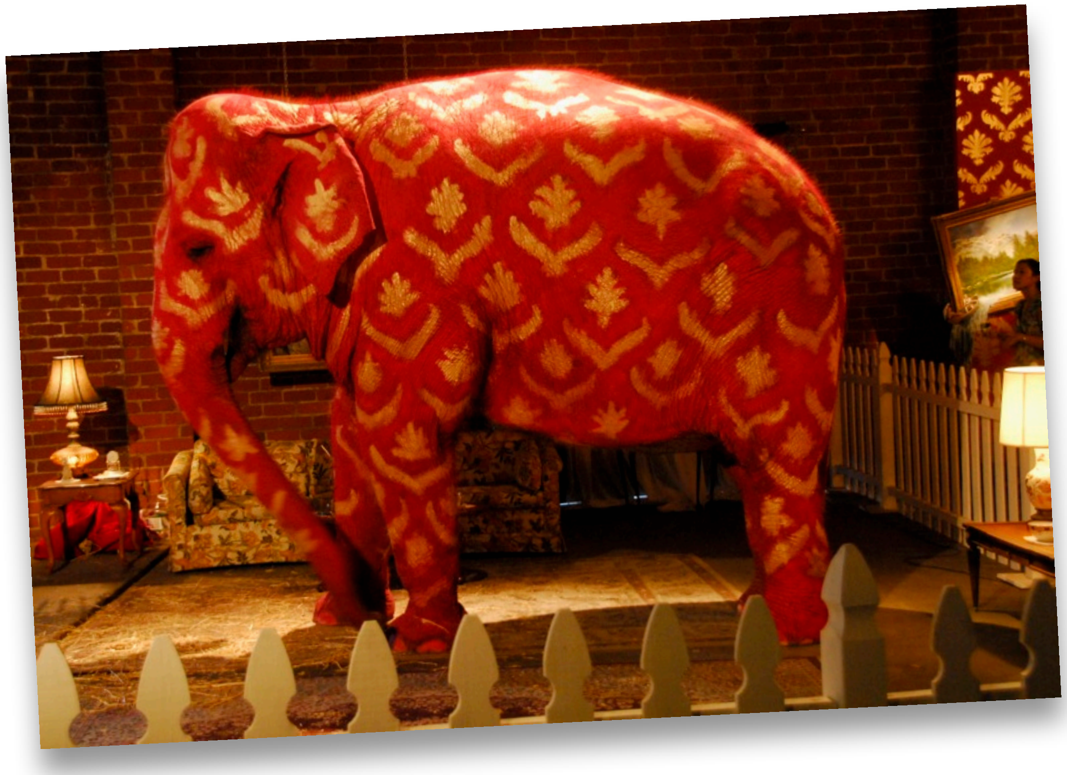
Department of Psychiatry
Department of Biomedical Engineering
Department of Communication Sciences & Disorders
University of Iowa

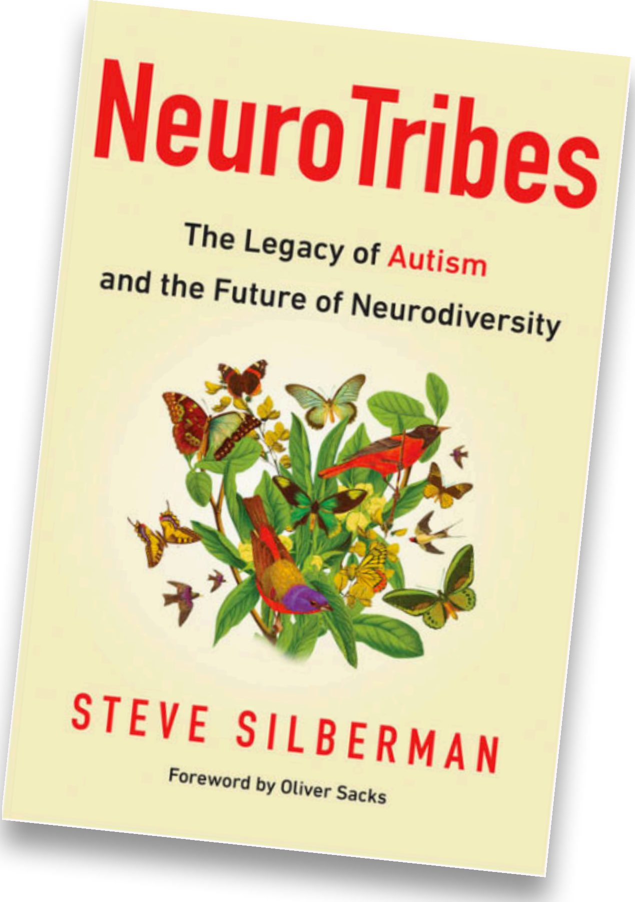
The University of Iowa

outline

- >> what is the point of genetic research in autism?
- >> how is ASD gene discovery pursued today?
  - >> SPARK
- >> how can we forecast gene discovery before the discovery happens?
  - >> machine learning

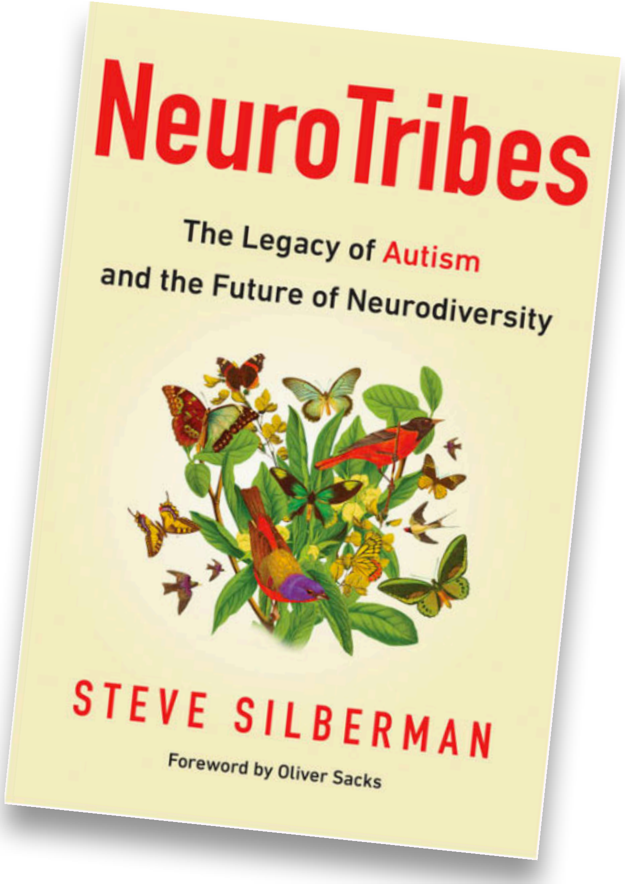
why research the causes of ASD?





why research the causes of ASD?

- “[Neurodiversity advocates] suggest that, instead of investing millions of dollars a year to uncover the causes of autism in the future, we should be helping autistic people and their families live happier, healthier, more productive, and more secure lives in the present” -- NeuroTribes, p. 470



why explore space?

- >> “In 1970, a Zambia-based nun named Sister Mary Jucunda wrote to Dr. Ernst Stuhlinger, then-associate director of science at NASA's Marshall Space Flight Center, in response to his ongoing research into a piloted mission to Mars.”

- >> “Specifically, she asked how he could suggest spending billions of dollars on such a project at a time when so many children were starving on Earth.”

  >> www.lettersofnote.com “Why Explore Space?”

why explore space?

“I even believe that by working for the space program I can make some contribution to the relief and eventual solution of such grave problems as poverty and hunger on Earth. Basic to the hunger problem are two functions: the production of food and the distribution of food. Food production by agriculture, cattle ranching, ocean fishing and other large-scale operations is efficient in some parts of the world, but drastically deficient in many others. For example, large areas of land could be utilized far better if efficient methods of watershed control, fertilizer use, weather forecasting, fertility assessment, plantation programming, field selection, planting habits, timing of cultivation, crop survey and harvest planning were applied.” (--> satellites)

advocacy AND research

- >> Questions research can’t address:
  - >> Why can’t I get the services my child needs?
  - >> What prospects does my child have once they reach adulthood? What will happen to my child when I’m gone?
  - >> How can I best ensure my child’s safety?
- >> Questions advocacy can’t address:
  - >> What treatment options are there for my child’s epilepsy?
  - >> Why won’t my child eat or sleep? What can I do?
  - >> What can I do about my child’s self-harm or aggressive behaviors?

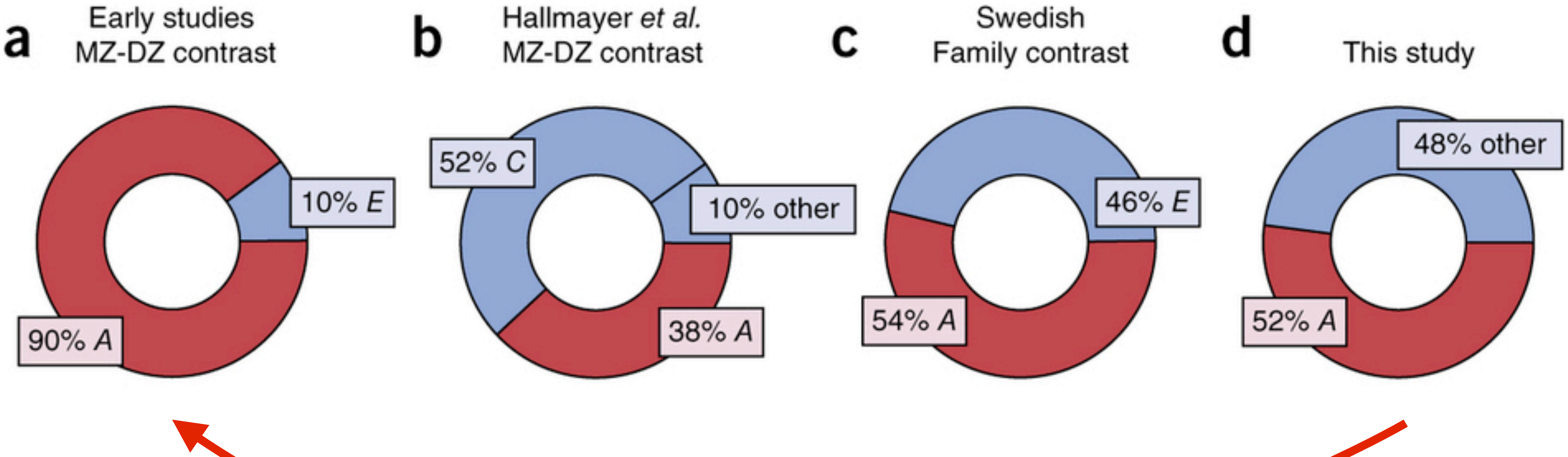
a holistic view

AUTISM SOCIETY

advocacy
(for today)

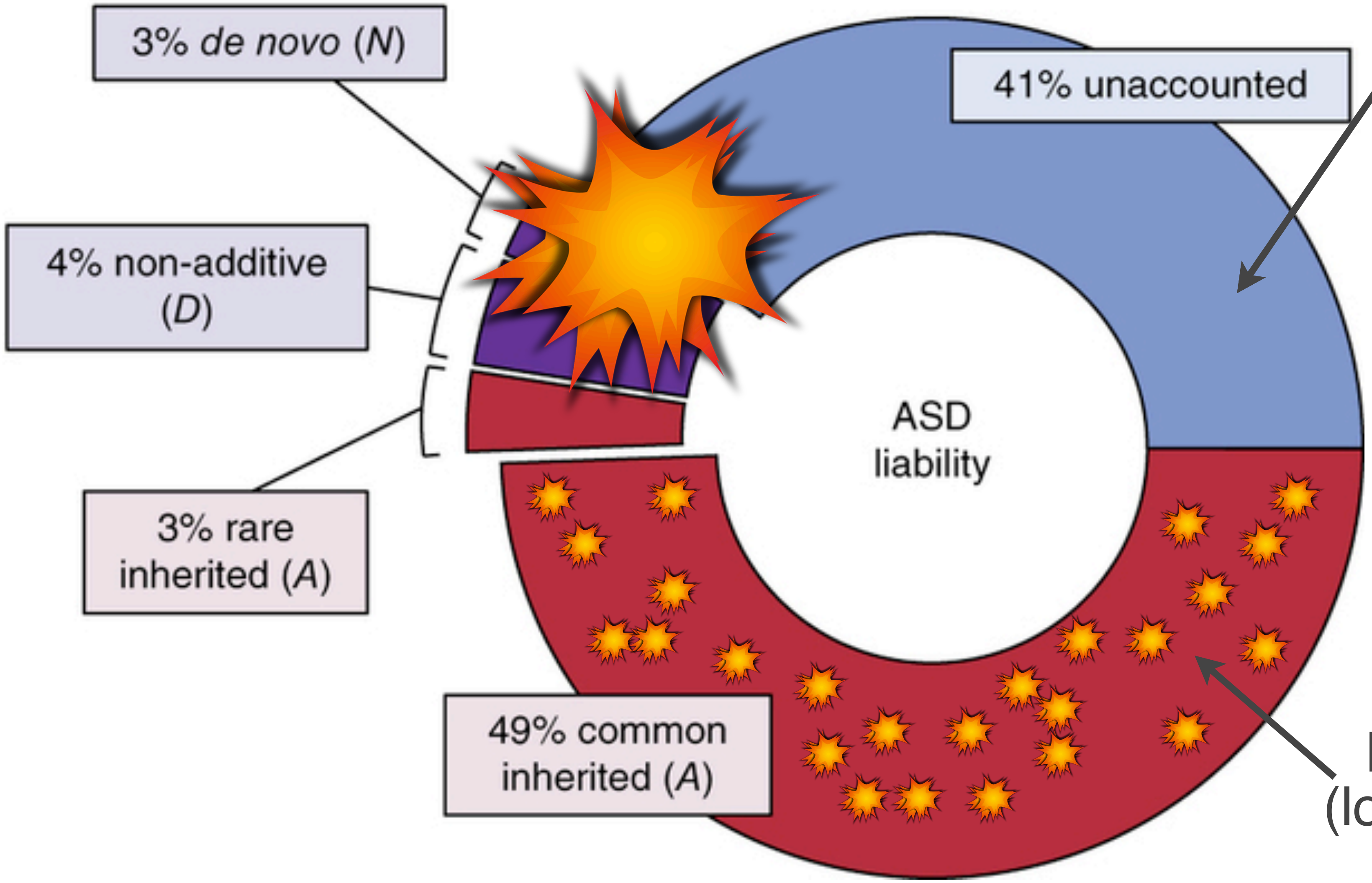
THE UNIVERSITY OF IOWA

research
(for tomorrow)



2017 JAMA update: autism is ~ **85% genetic**

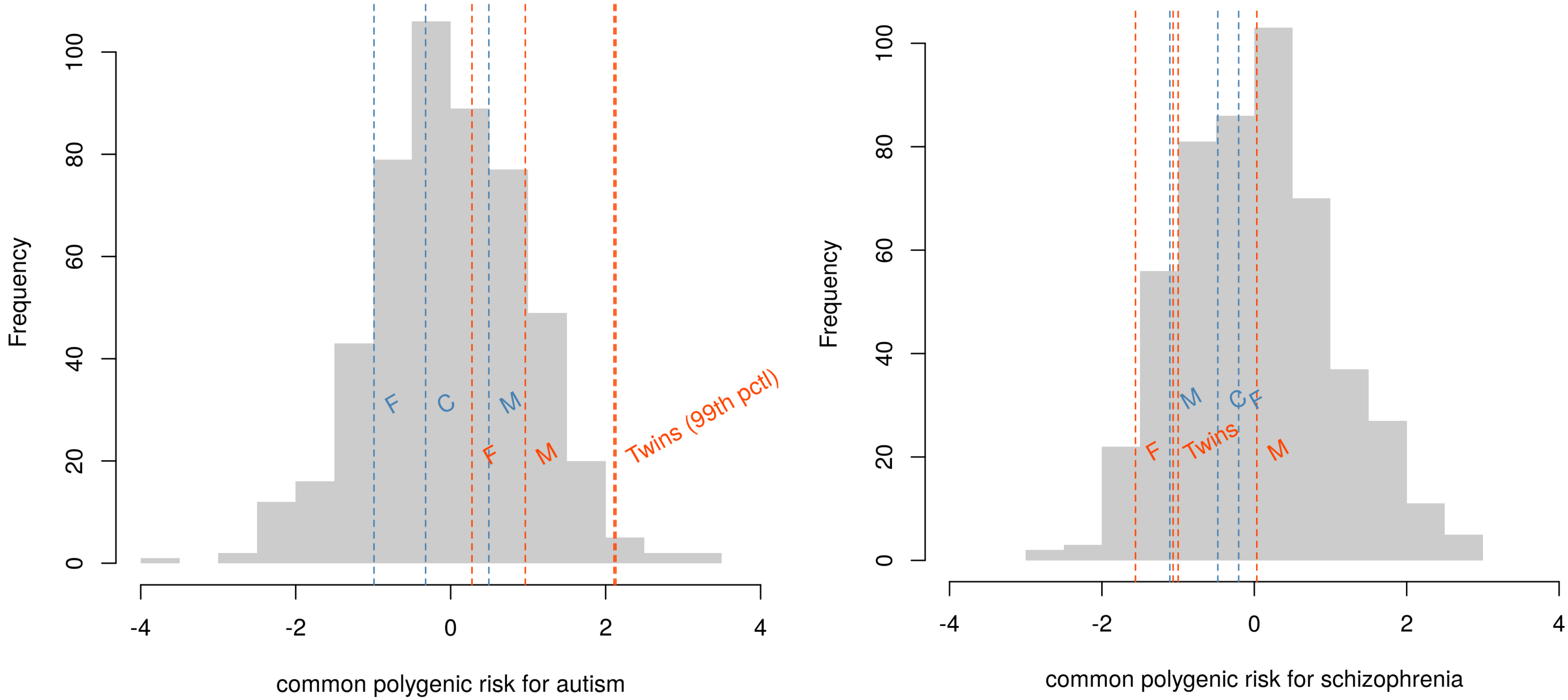
Gaugler, et al. *Nat Genetics* 2014



this has shrunk **substantially**

polygenic risk
(lots of tiny genetic risk factors)

Gaugler, et al. *Nat Genetics* 2014


Frequency
0
20
40
60
80
100
-4
-2
0
2
4
F
C
M
F
M
Twins (99th pctl)
common polygenic risk for autism
Frequency
0
20
40
60
80
100
-4
-2
0
2
4
M
C
F
F
Twins
M
common polygenic risk for schizophrenia

why genetics?

>> every day we’re learning more about the genes and biological processes that underlie autism

>> but why are we really doing this? to what end?

>> do we want to “cure” autism? Even if we did, is that possible?

early diagnosis

>> it has been demonstrated repeatedly that early intervention can lead to lasting improvements in core autistic traits (though not raw cognitive functioning/intelligence)

>> diagnosis is being pushed back earlier by some non-genetic approaches

>> ...but genetics will win out in the end as the “earliest” indicator

robingulon7@reddit

personalized treatment

- autism comes in many flavors and varieties
- "catch-all" treatments can be ineffective or even dangerous for some children
- treatment adapted to the individual's biology holds much promise:
- Joe Gleeson's lab demonstrated a case of severe autism that was due largely to a mutation that rendered patients unable to make a certain amino acid
  - dietary supplementation* was promising in this case (in mice)

sharing genetic findings

- because there are so many possible ways to arrive at autism, it's rare to find someone else who has the exact same kind of autism
- as genetic diagnostics become routine, it's important for researchers and clinicians to share knowledge of the mutations that are connected to a particular flavor of autism
  - otherwise it can be impossible for clinicians to know what the genetic results might mean

is research therapeutic?

- the mere participation in research might be therapeutic in and of itself
  - increased appreciation for condition being rooted in biology (i.e. guilt alleviated)
  - altruism, helping others
  - "I'm not alone"

differences vs. deficits

>> the ASD community is rightly sensitive to the idea of people crusading for a “cure” for autism

>> it is a condition that in many ways is linked to one’s identity

>> nevertheless there are some aspects of autism (specifically comorbidities) that cause a great deal of distress to children and their families

>> a better understanding of autism’s biologies can help cultivate an appreciation for what a difference is vs. what a deficit is

SPL

human uniqueness

- many of the biological processes that underlie autism are the same that underlie human uniqueness
  - brain size & growth
  - social cognition
  - trade-offs between social vs. other skills
- human evolution continues, and on some level individuals with autism are on the front lines of it, with selective pressure (both medical and societal) pushing back

putting a face to a gene

- >> SHANK3 (Phelan-McDermid)
- >> nearly 1% of ASD cases have mutations here
- >> affects neuron synapse connections

- >> FOXP1
- >> ASD + ID with language delay

- >> regulator of gene expression

- >> low muscle tone, eye, kidney, bladder problems

- >> SYNGAP1
- >> ID, epilepsy, ASD

- >> brain becomes hyperexcitable (seizures)

- >> sleep issues, constipation, low muscle tone

- >> CHD8
- >> ASD with macrocephaly

- >> chromatin remodeling

- >> sleep issues, gastrointestinal issues

SPARK

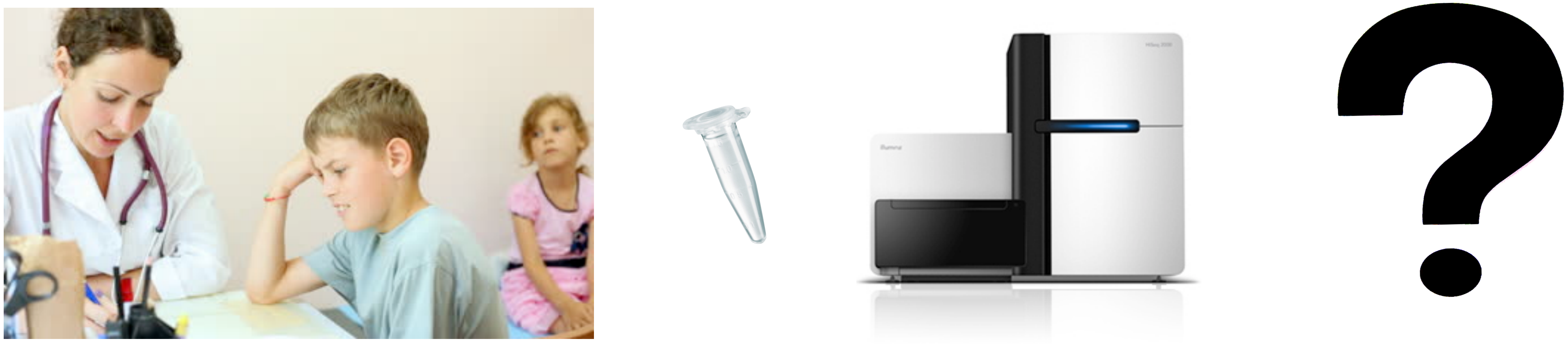
Igniting autism research
Improving lives

autism research

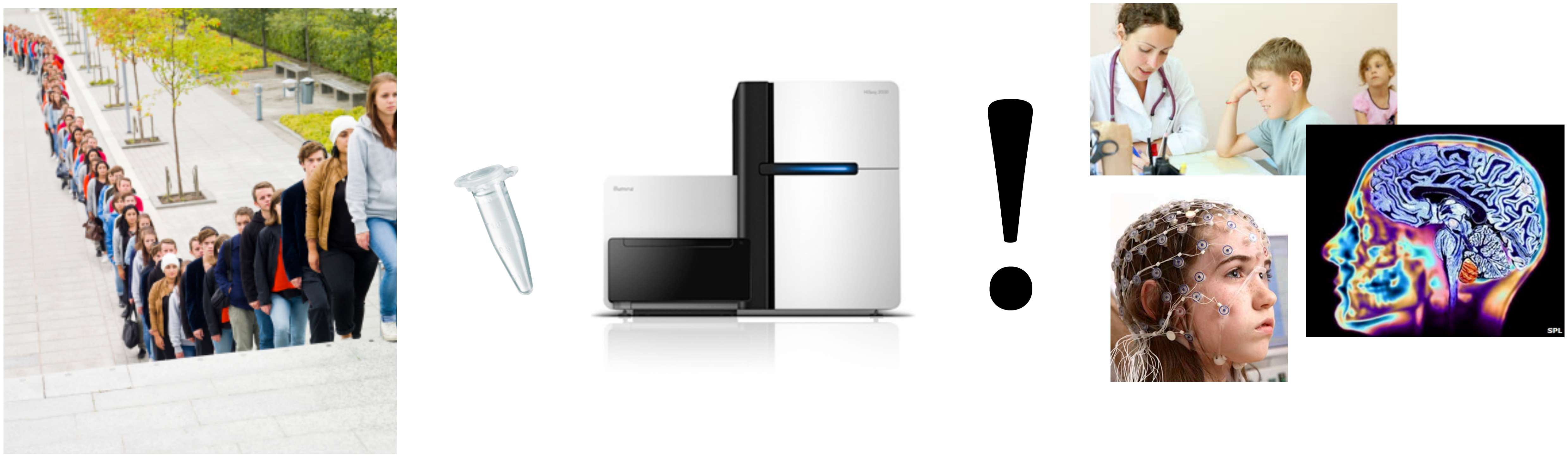
SPARK

genotype first approach

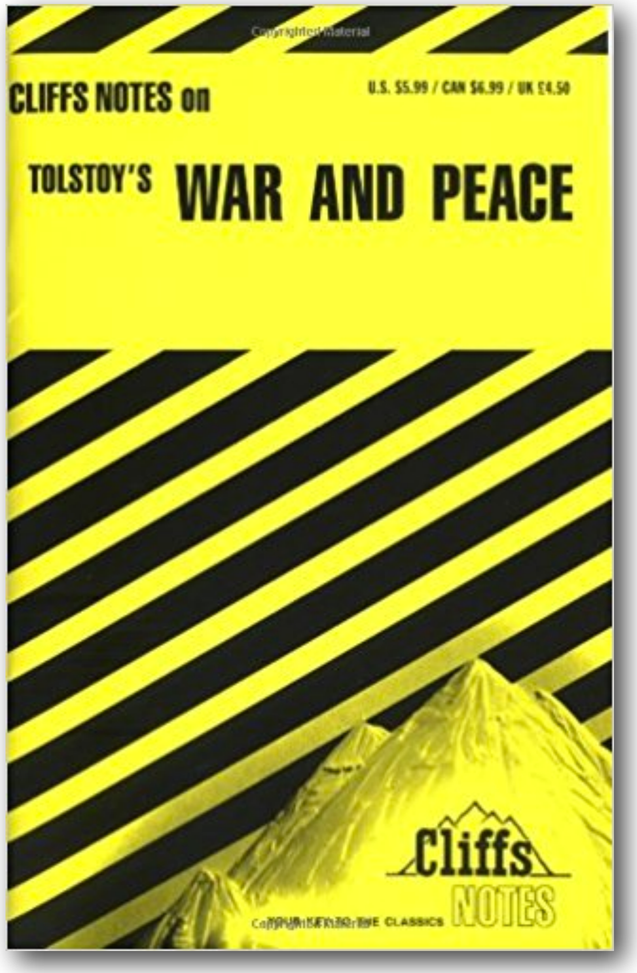
usually:



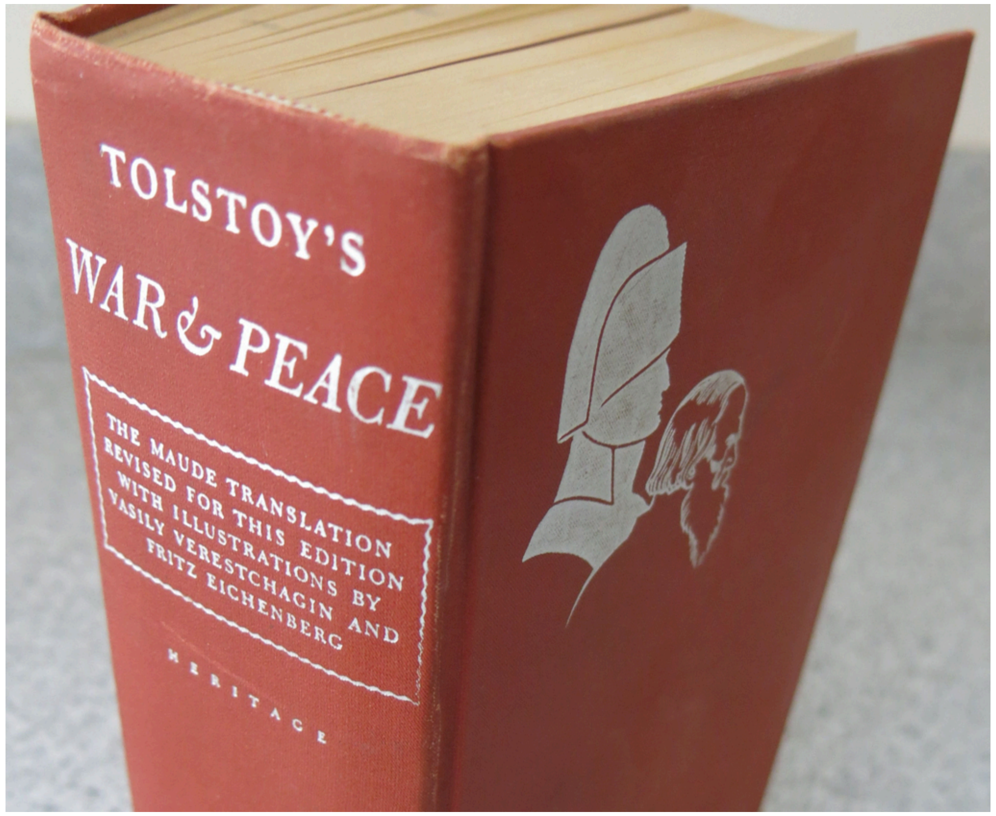
genotype-first approach:



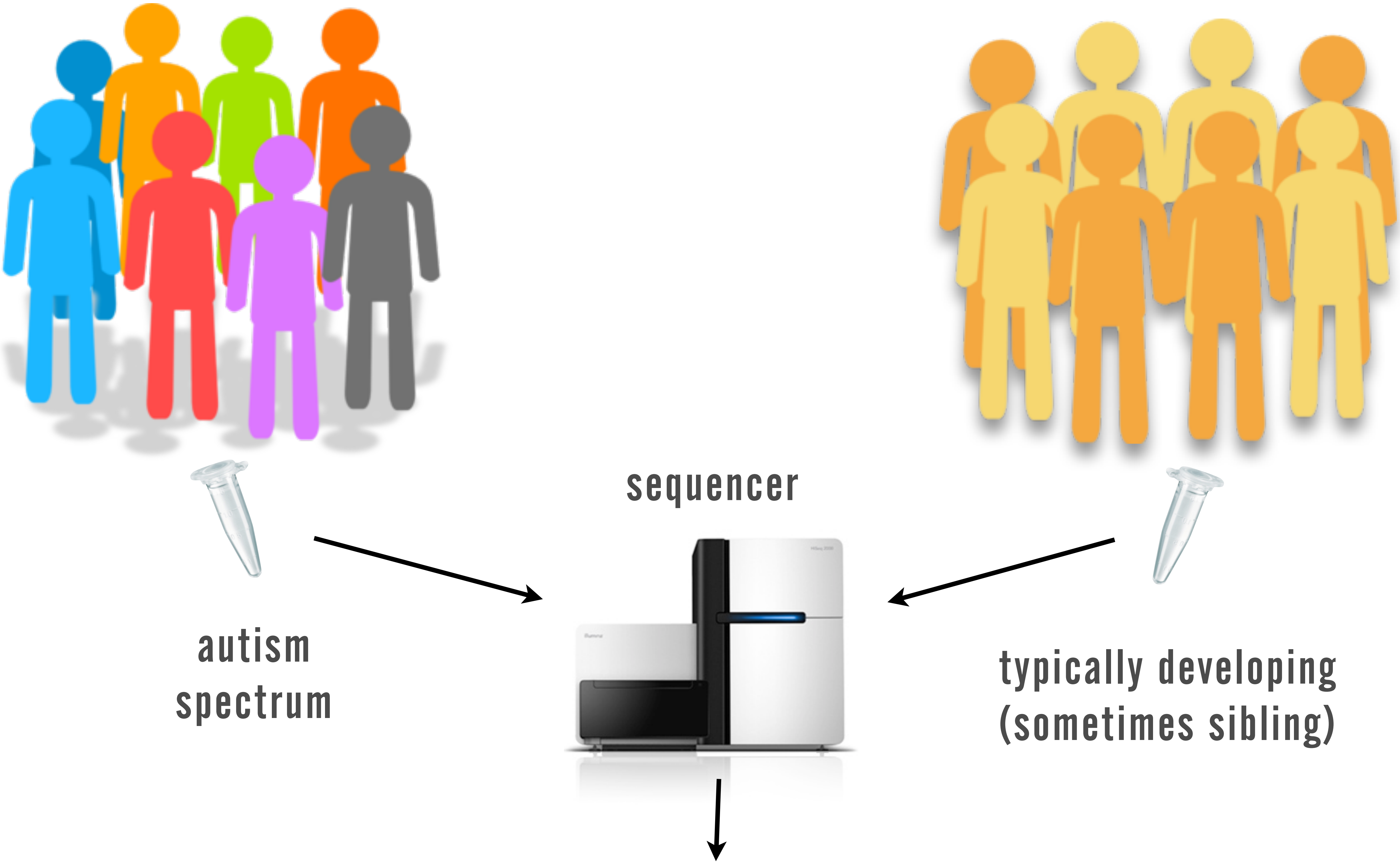
aside: exome vs. whole genome



exome
sequencing
(protein-coding parts only)


TOLSTOY'S
WAR & PEACE
THE MAUDE TRANSLATION REVISED FOR THIS EDITION WITH ILLUSTRATIONS BY VASILY VERESTCHAGIN AND FRITZ EICHENBERG
HERITAGE

whole genome
sequencing



sequencer

autism
spectrum

typically developing
(sometimes sibling)

what's
different?

SPARK

- SPARK aims to recruit 50,000 individuals with autism and their parents
  - “professional diagnosis of autism” --> self-report
- 25 participating sites nationwide (Iowa is one of them)
- an online community and resource for autism research for the next 20 years

SPARK

>> registration online: http://sparkforautism.org/uiowa

>> saliva collection kit sent to you (affected child, parents)

>> spit in tube, mail back in prepaid package

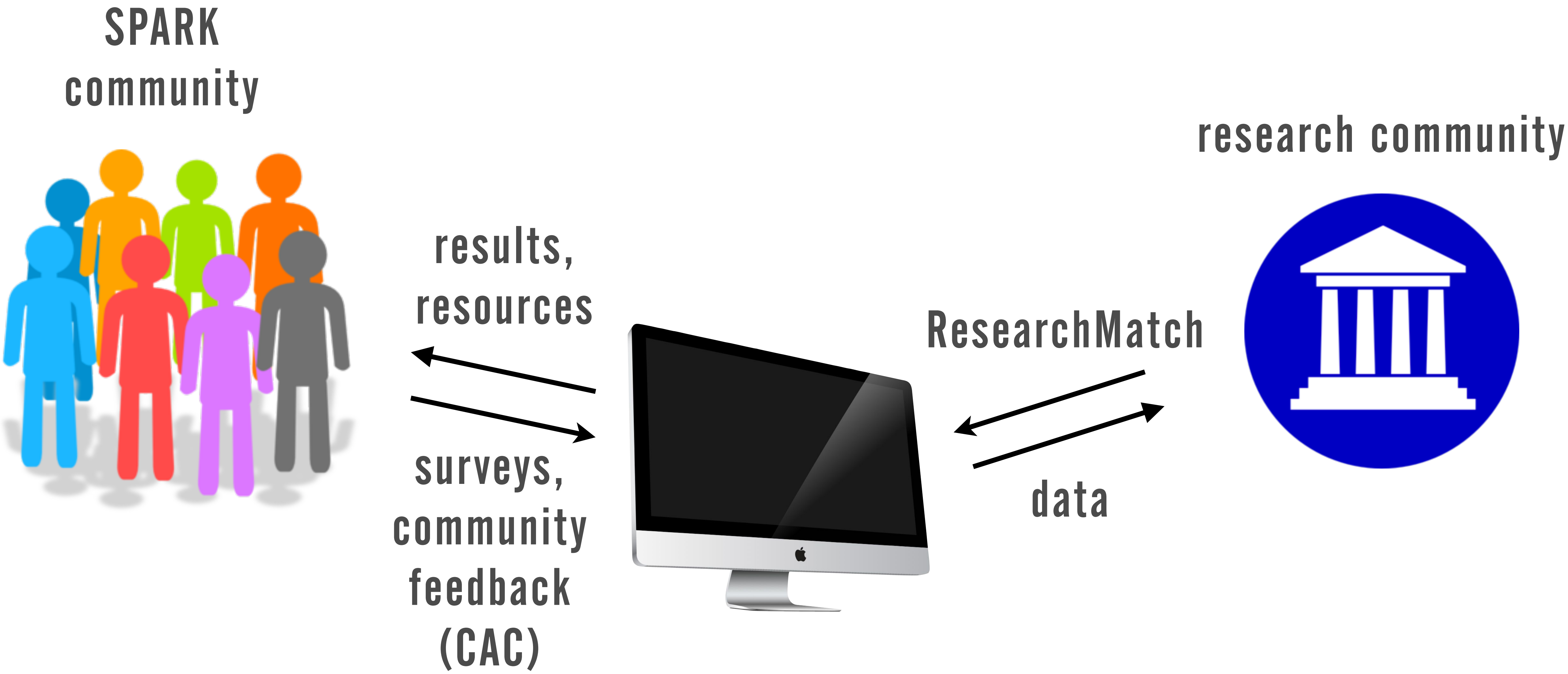
  >> $50 participation bonus

>> DNA extracted, analyzed (exome sequencing)

  >> data provided to qualified autism researchers

>> periodically participate in follow-up surveys at your discretion

  >> compensation on a per-survey basis



SPARK community

results, resources

surveys, community feedback (CAC)

ResearchMatch

data

research community

SPARK: why 50,000?

Looking on a gene-by-gene basis:

this is where studies have been
in the past few years

N	ASD w/ mutation	typical dev. w/ mutation
100	1	0
1000	2	0
10,000	3	1
50,000	15	1

SPARK pilot findings

- currently about 4-5% of families have a “returnable result” (provided with genetic counseling)
  - i.e., a clear loss of function variant in one of 78 known ASD genes
  - this is very conservative
    - about 25% of families in the pilot study had highly suggestive findings; very useful from a research perspective
  - the list is growing (which is one of the main purposes of SPARK)
- see SPARK Snapshot for more details on initial findings

accelerating gene discovery

gene discovery

- large-scale genetic studies are being funded to enumerate ASD risk genes
  - expensive and relatively slow
- what if we could fast forward and know those genes now?
  - given the closed set of ∼20,000 genes, can we make educated guesses about which are most likely the ∼1000 autism genes?

The Weather

Hill Valley Telegraph

Vol. XVII. No. 22

THURSDAY, OCTOBER 15, 1959

Index

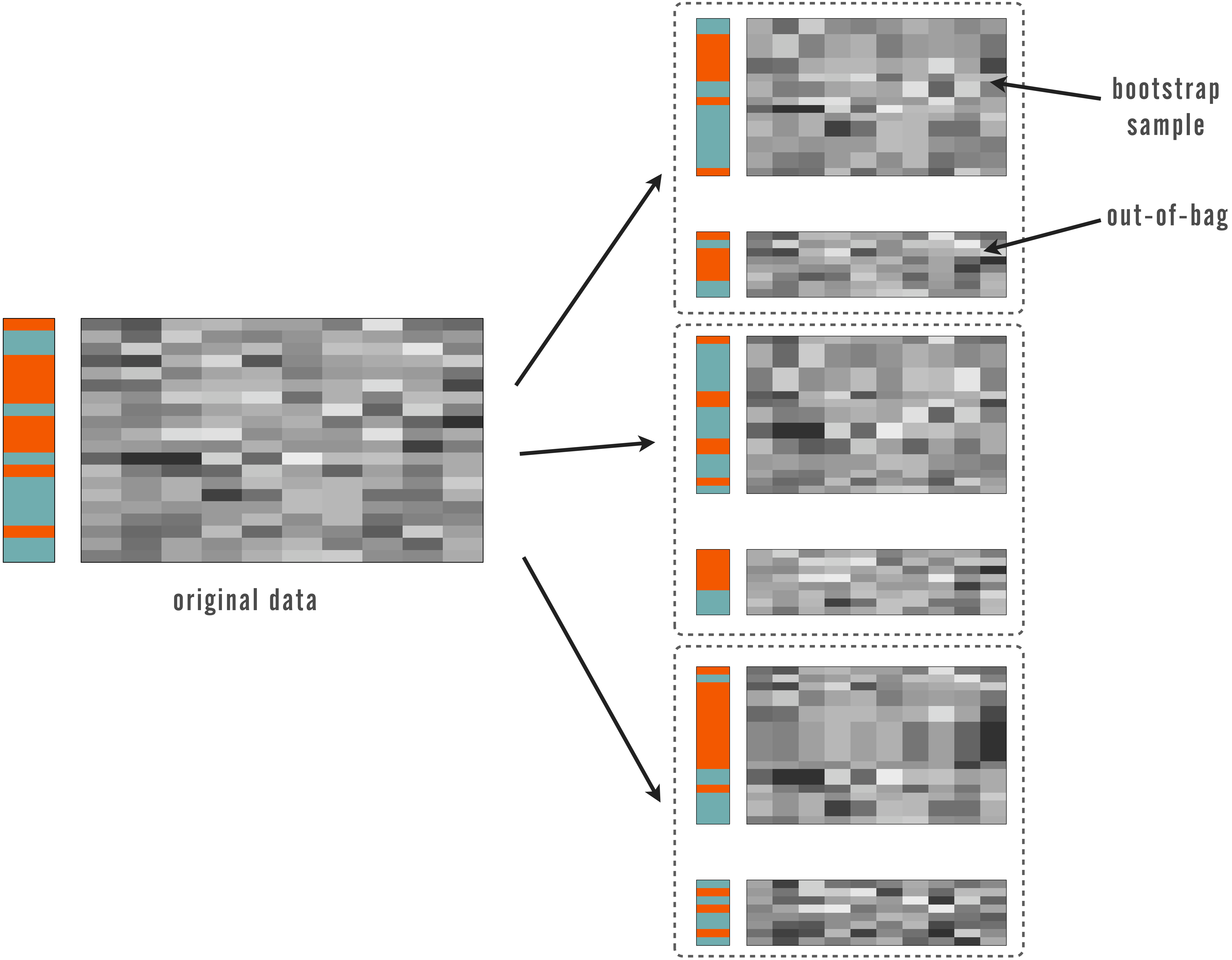
PRICE: 10 CENTS

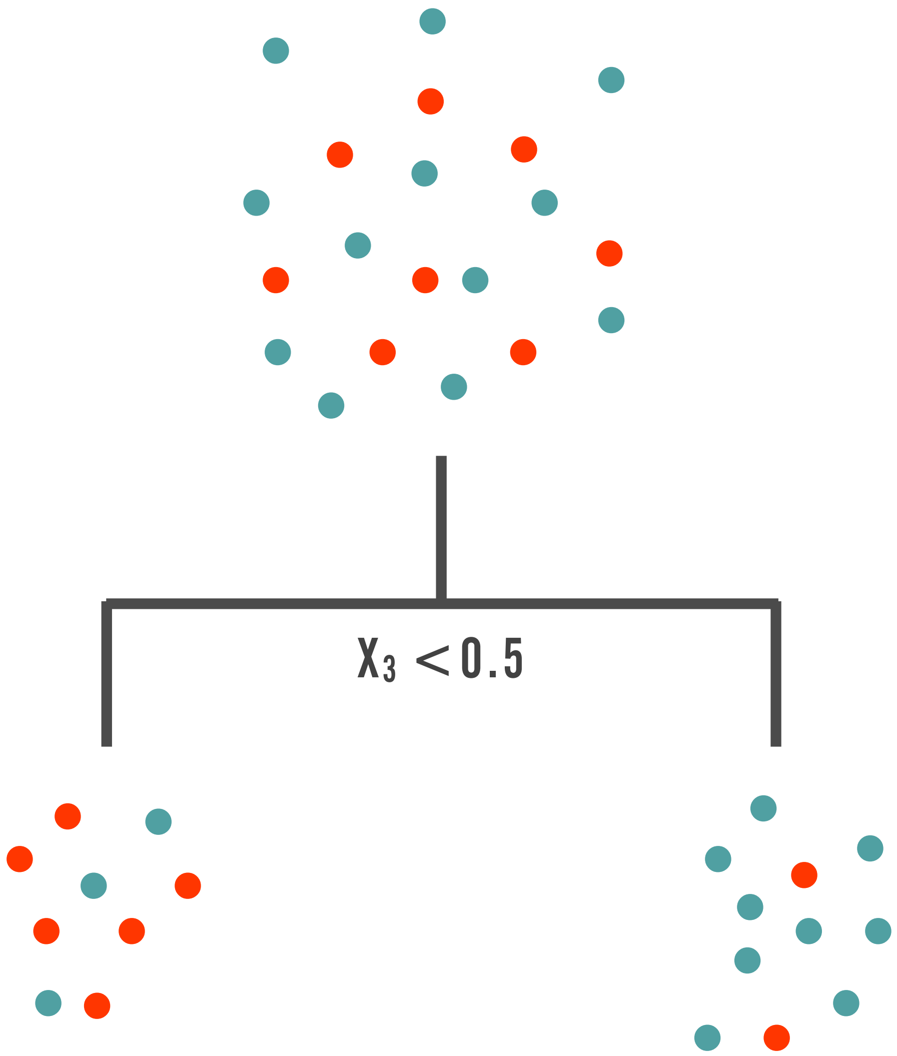
BIFF WINS AGAIN

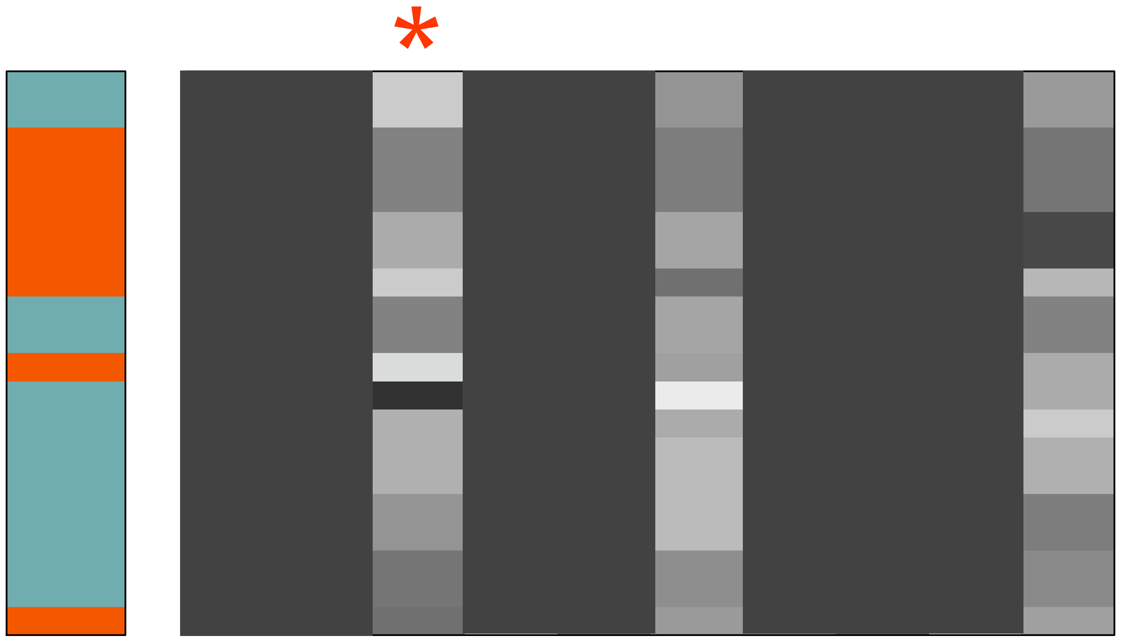
Khruschev Offers Dates for Summit

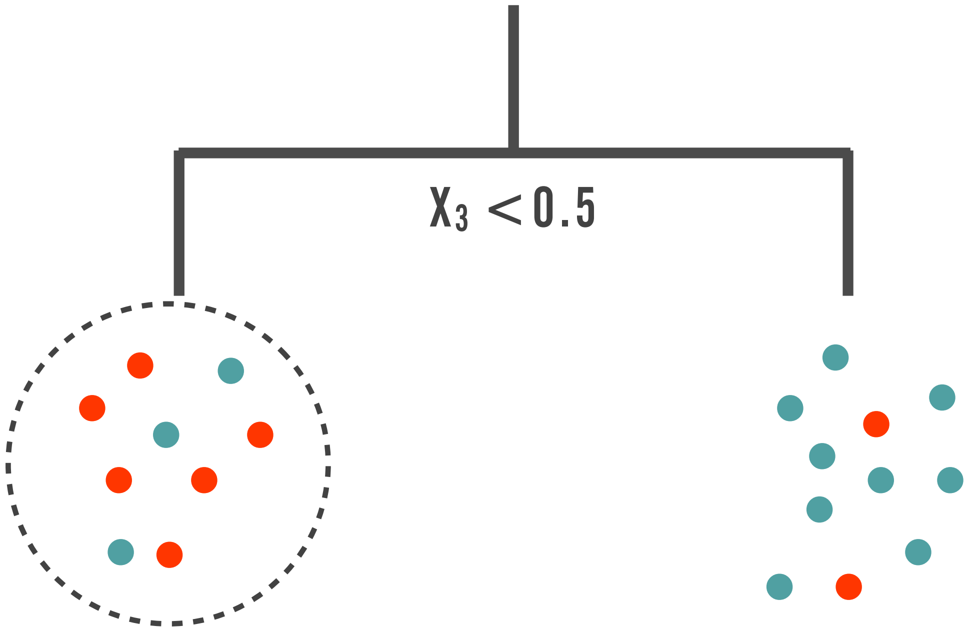
random forests

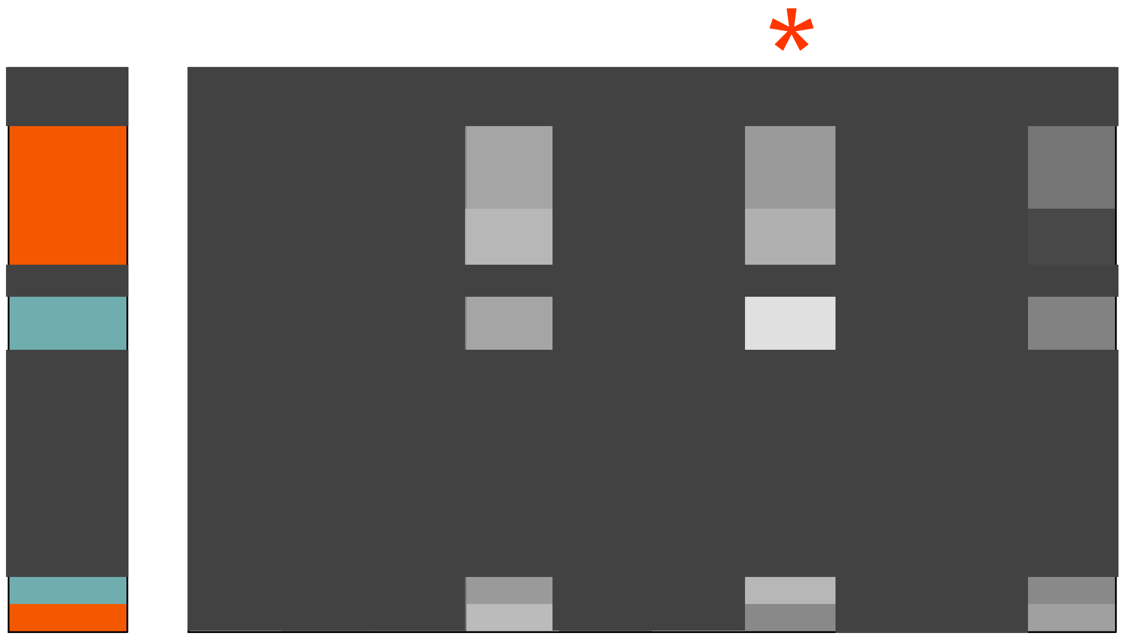
chriswenger@flickr

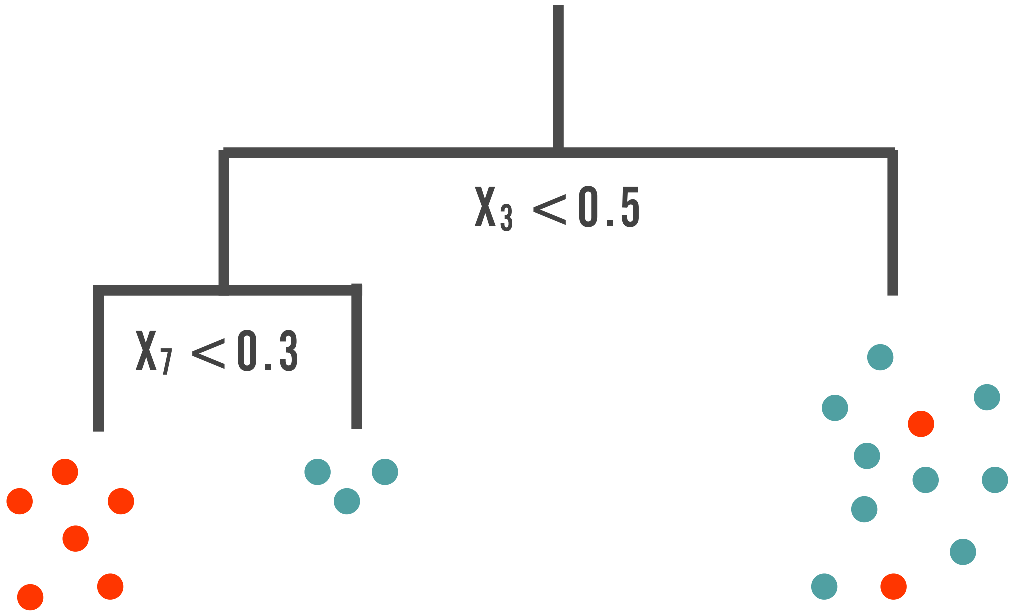

bootstrap sample
out-of-bag
original data

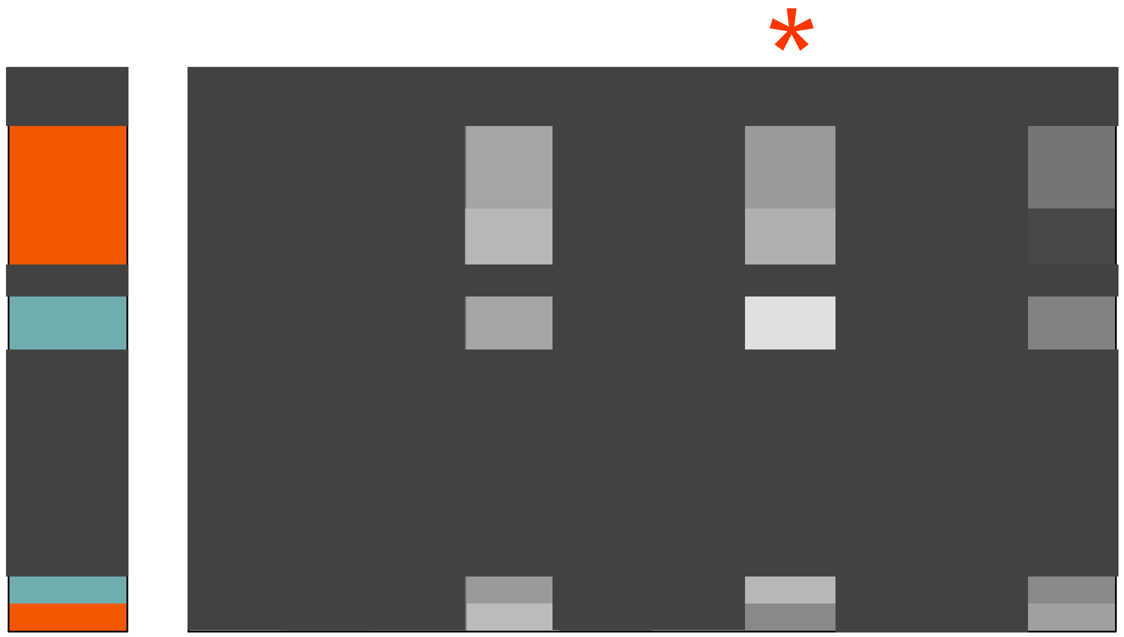

$X_3 < 0.5$

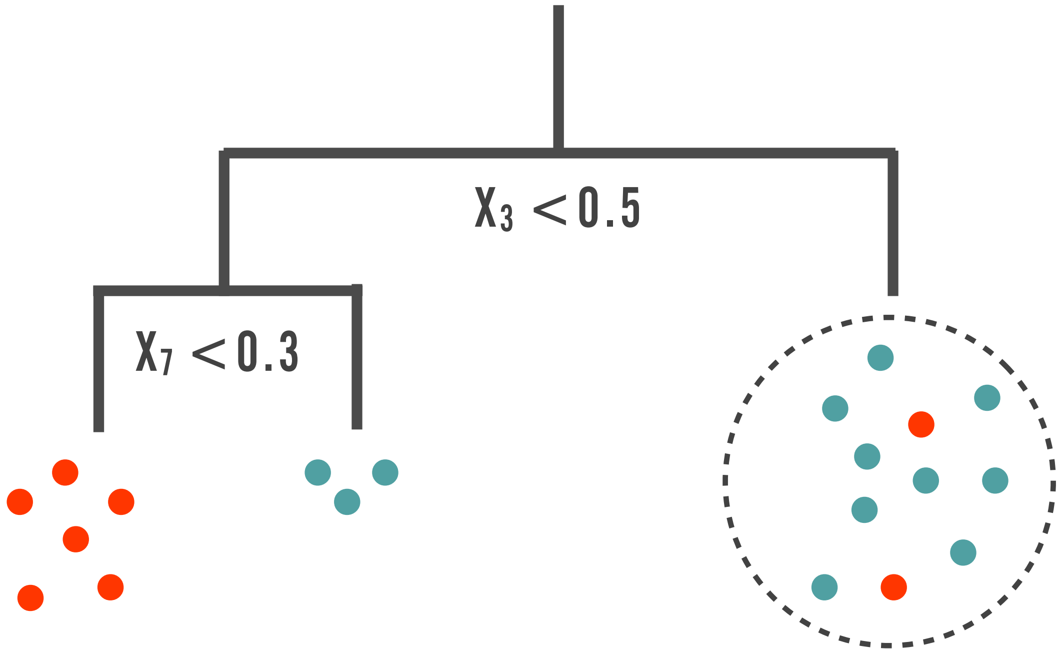

*

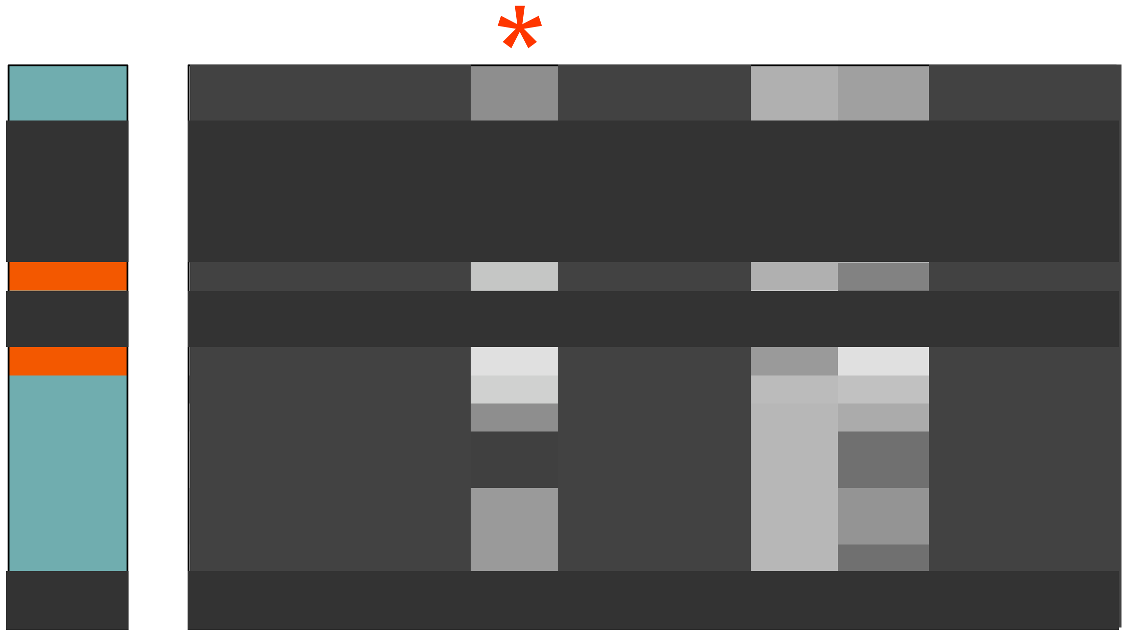

$X_3 < 0.5$

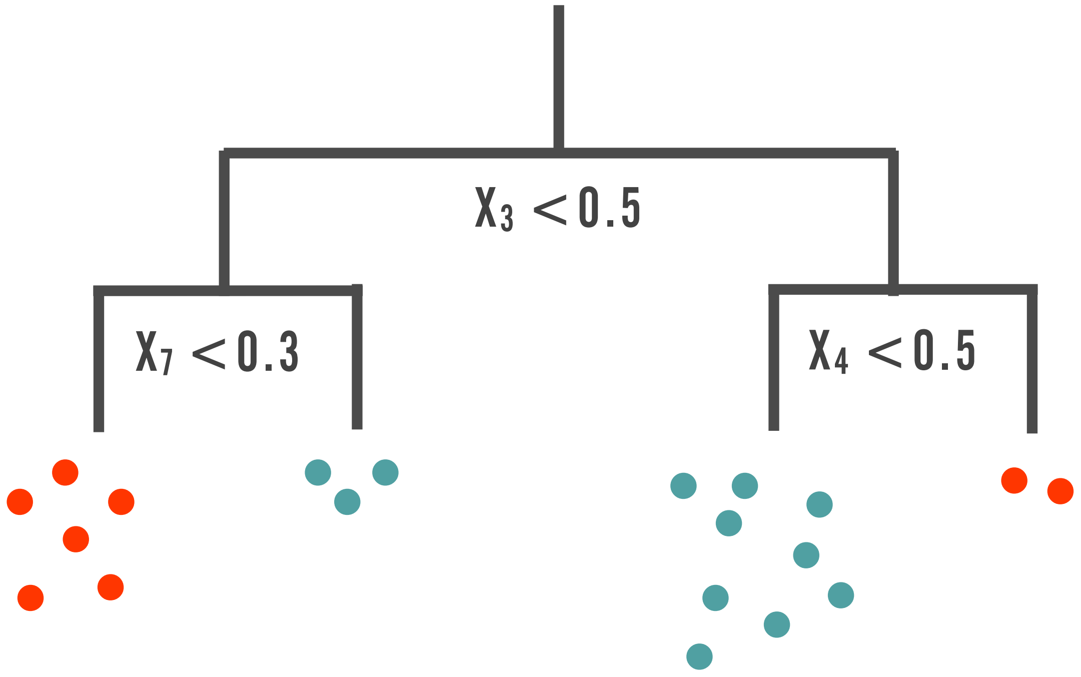

*

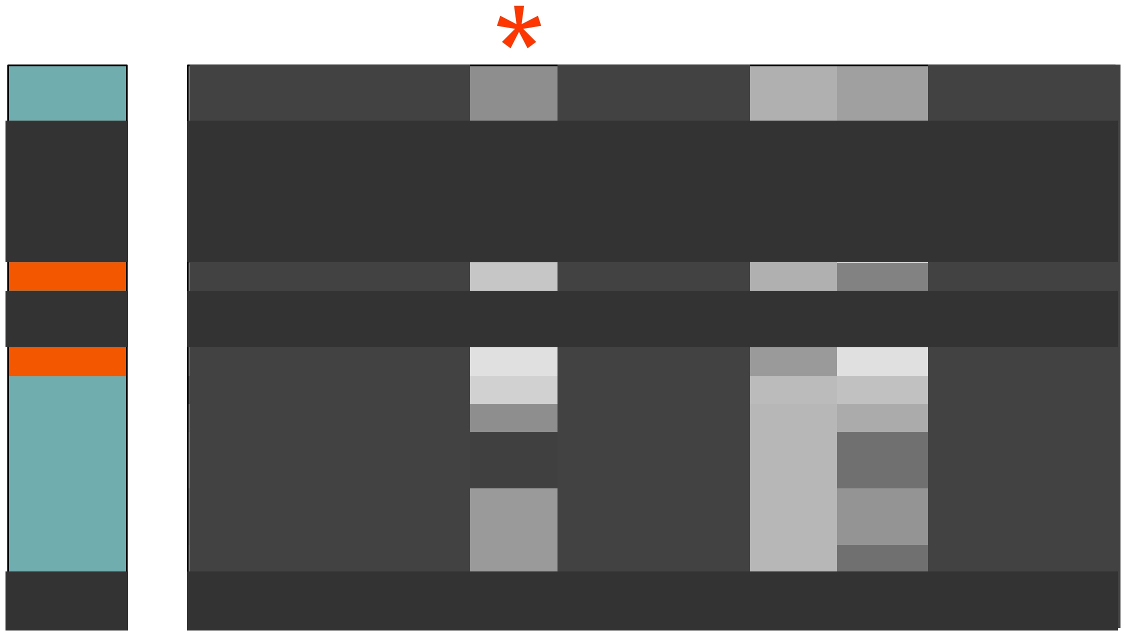

$X_3 < 0.5$
$X_7 < 0.3$

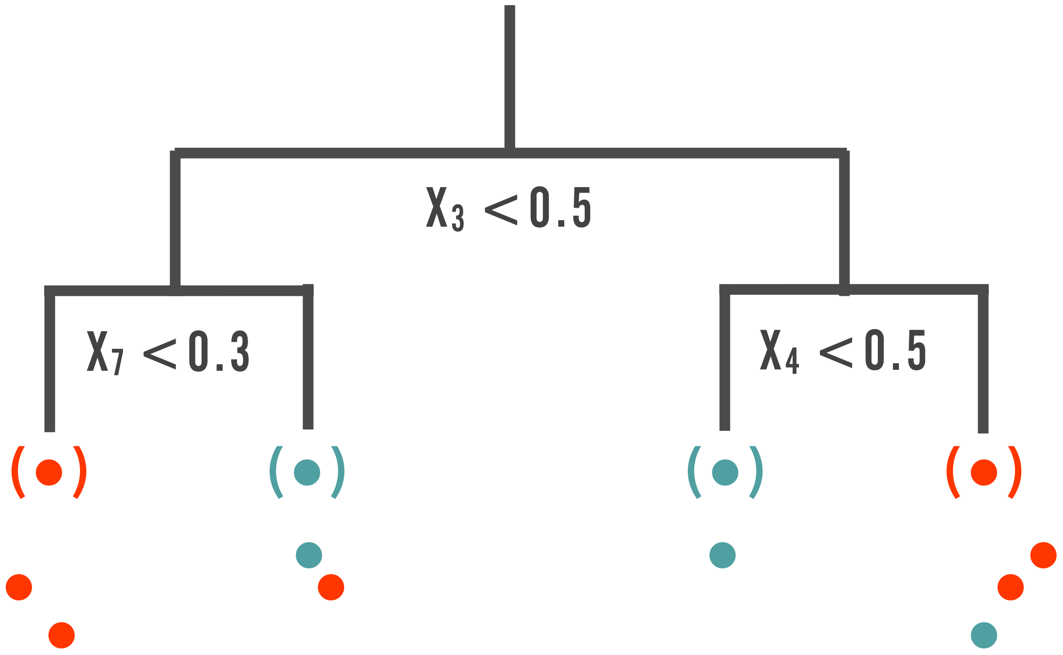

*


$X_3 < 0.5$
$X_7 < 0.3$

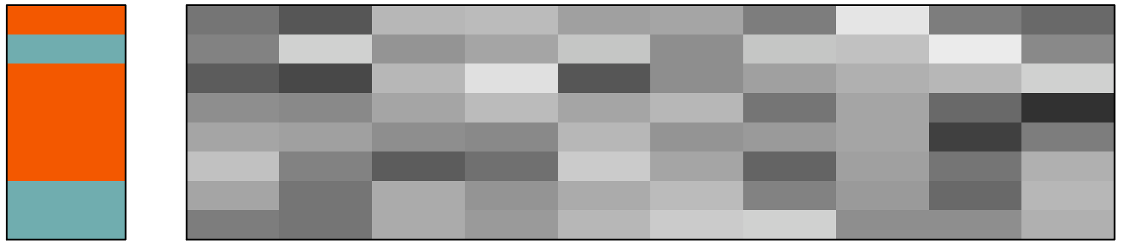

*


$X_3 < 0.5$
$X_7 < 0.3$
$X_4 < 0.5$


*



OOB data (test set)



trees to forests

- >> each tree tries to solve the same problem under slightly different conditions
  - >> different samples, different predictors available at each decision split
- >> prediction emerges from the forest (an ensemble)
  - >> not the output of a single tree
    - >> average (regression), majority vote (classification)

assessment

pro

- >> open-ended, model-free approach
- >> can detect interactions
- >> works with categorical and continuous responses and predictors
- >> not prone to overfitting
- >> easily parallelized
- >> applicable to a wide variety of data and problems
  - >> prediction, feature selection, clustering, etc.

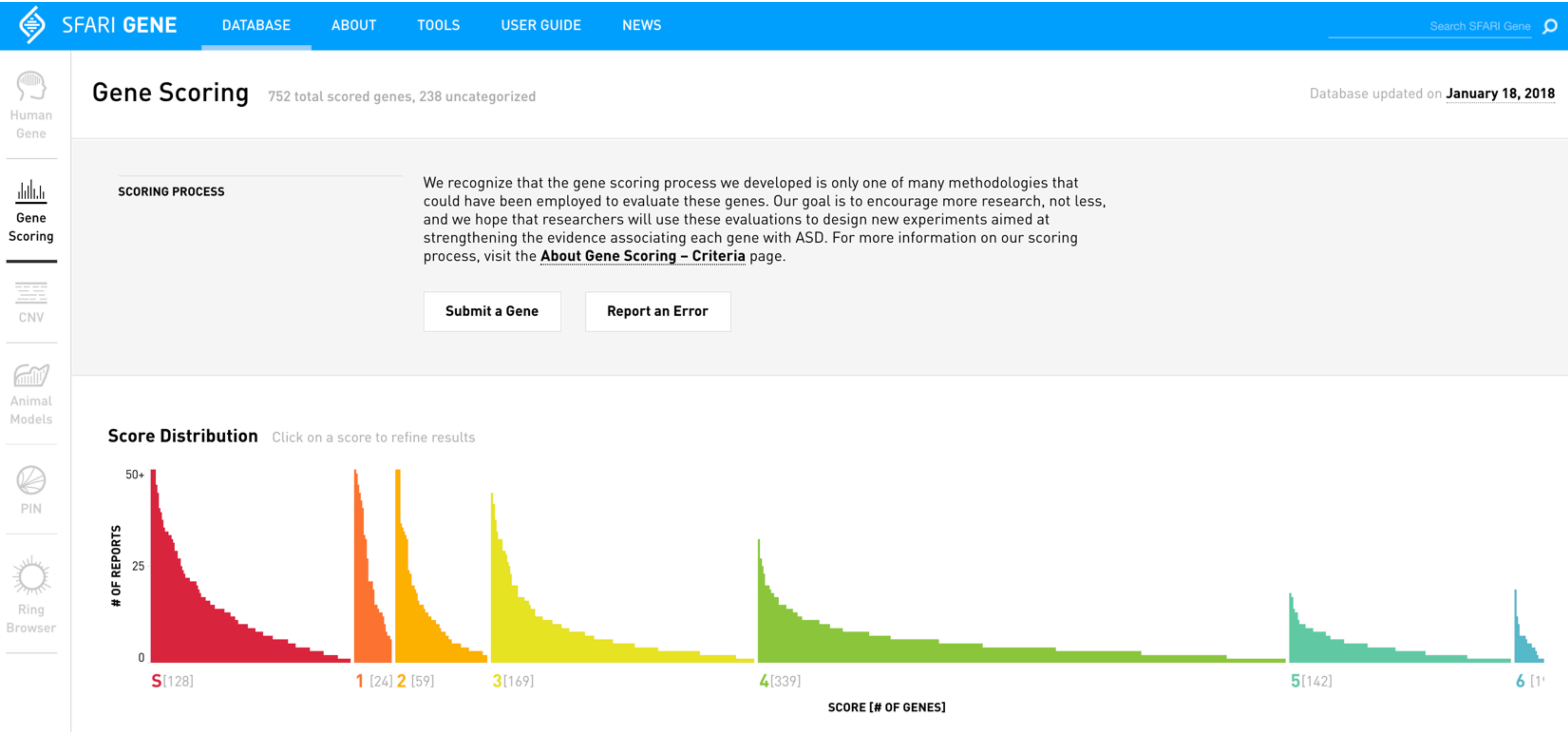
con

- >> “model” interpretation
- >> it’s stochastic
- >> memory, time efficiency

back to the task at hand...

ASD gene predictors

- >> Krishnan, et al. (Nature Neuroscience 2016)
  - >> machine learning approach, uses coexpression and a few other metrics, class labels drawn from literature
- >> DAMAGES (Zhang & Shen, Hum Mut 2017)
  - >> uses pLI & cell-type specific expression profiles
- >> TADA (Sanders, et al. Neuron 2015)
  - >> Bayesian approach to estimate excess of functional/pathogenic variation; gives gene-level score



Gene Scoring

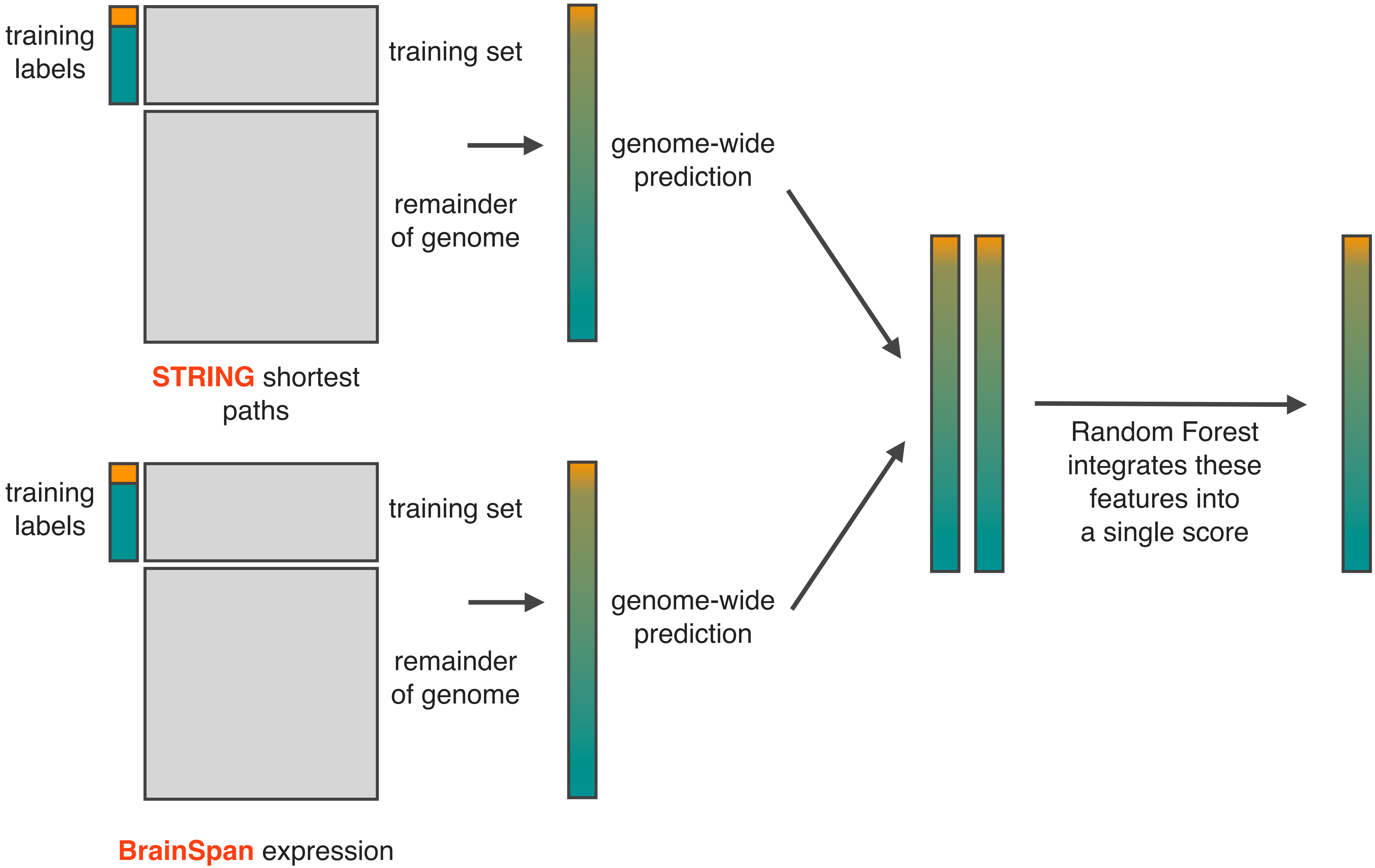
752 total scored genes, 238 uncategorized

Database updated on **January 18, 2018**

SCORING PROCESS

We recognize that the gene scoring process we developed is only one of many methodologies that could have been employed to evaluate these genes. Our goal is to encourage more research, not less, and we hope that researchers will use these evaluations to design new experiments aimed at strengthening the evidence associating each gene with ASD. For more information on our scoring process, visit the **About Gene Scoring – Criteria** page.

Score Distribution Click on a score to refine results



training labels

training set

remainder of genome

genome-wide prediction

STRING shortest paths

training labels

training set

remainder of genome

genome-wide prediction

BrainSpan expression

Random Forest integrates these features into a single score

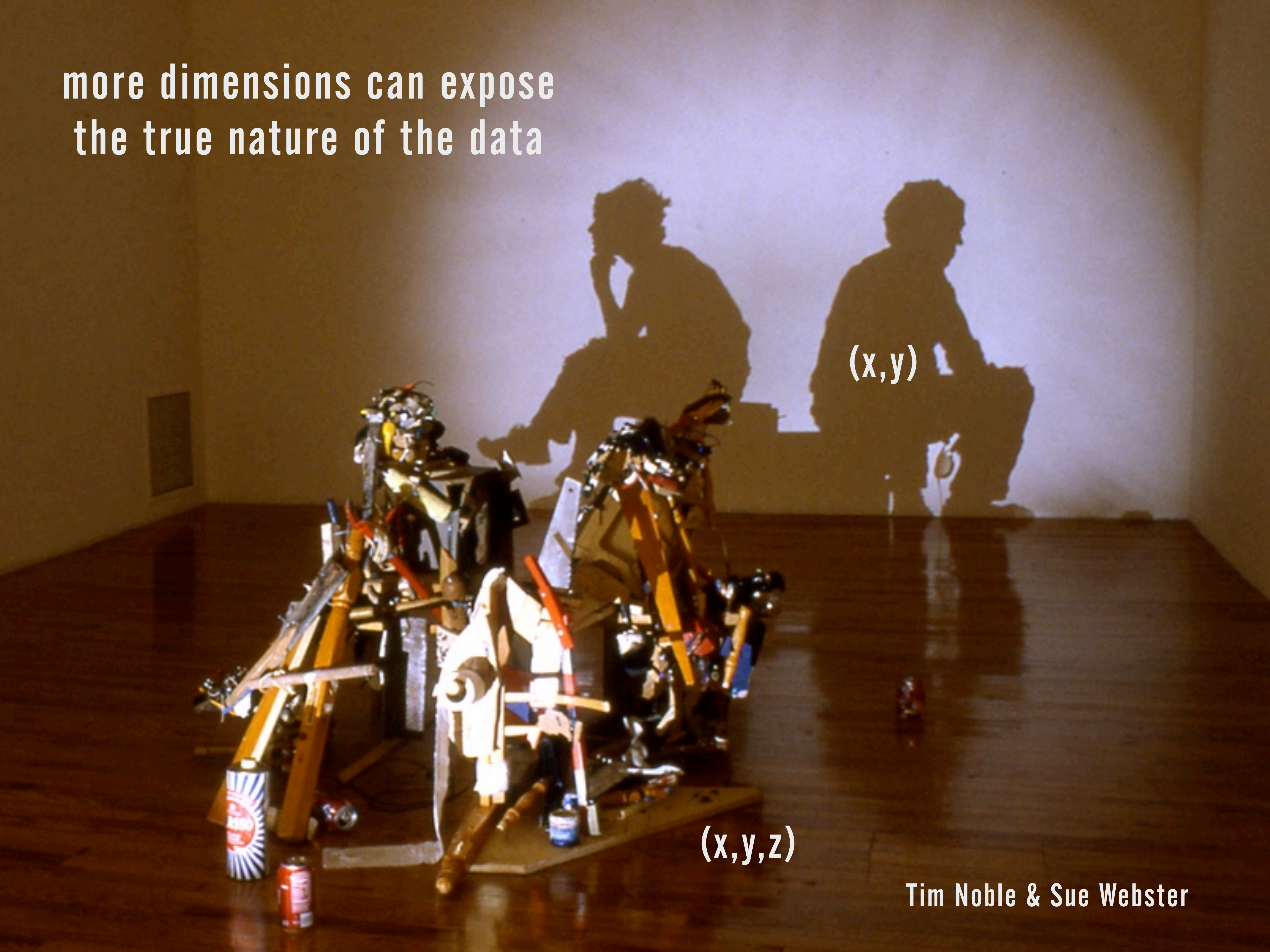
initial performance

>> initial tests showed improvement over both Krishnan and DAMAGES, and TADA in some settings

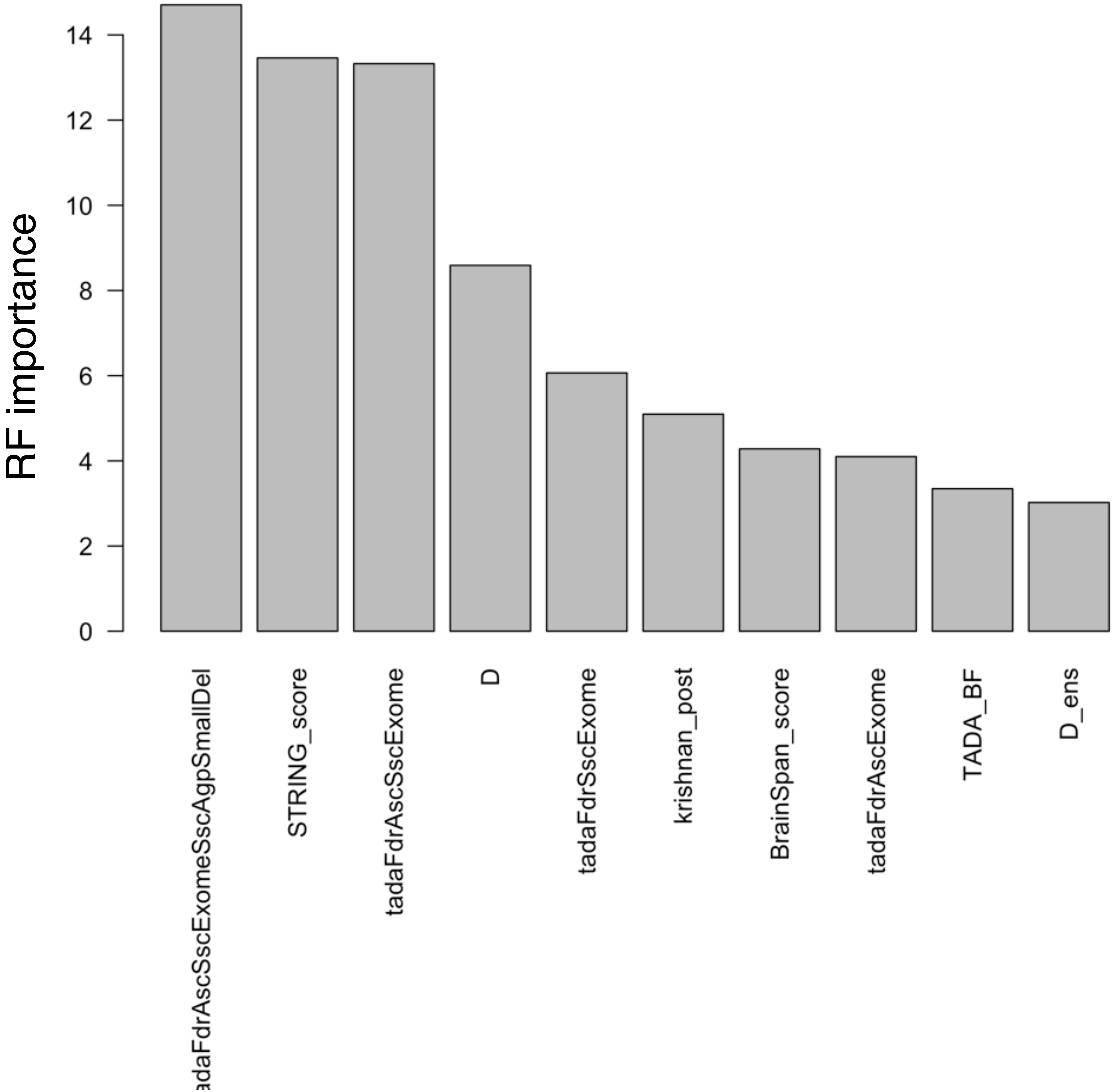
>> why not combine all of these scores into a single score?

YOU WILL BE ASSIMILATED

more dimensions can expose the true nature of the data

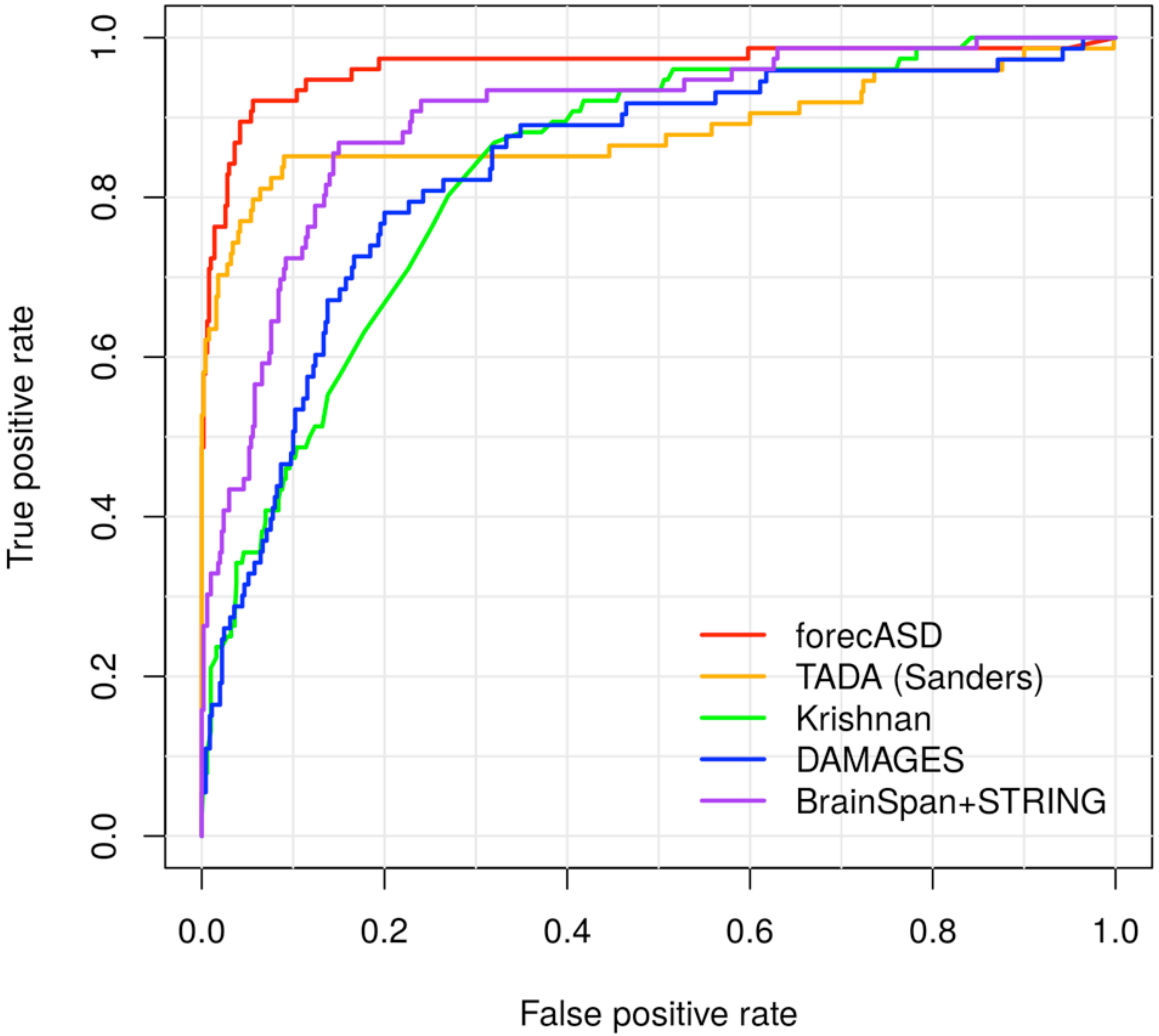


Tim Noble & Sue Webster


RF importance
0
2
4
6
8
10
12
14
adaFdrAscSscExomeSscAgpSmallDel
STRING_score
tadaFdrAscSscExome
D
tadaFdrSscExome
krishnan_post
BrainSpan_score
tadaFdrAscExome
TADA_BF
D_ens

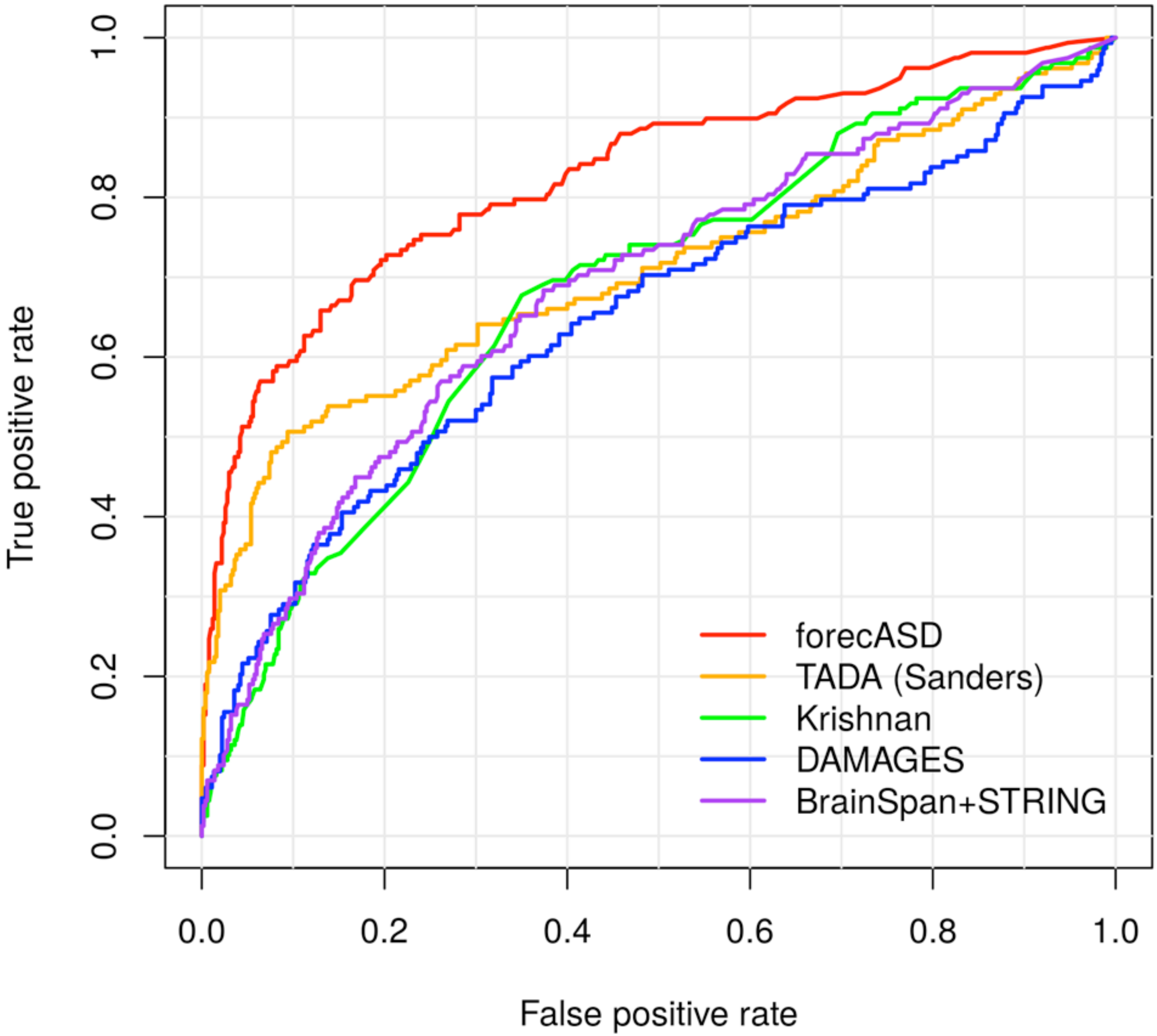
performance

- >> we **compared*** these methods based on their ability to discriminate SFARI-scored genes from a random sample of genes **not** listed in SFARI Gene
  - >> “negative” genes were matched to SFARI genes based on mutation rate (no distributional difference between pos/neg)
- >> first, we looked at well-accepted ASD genes (SFARI Gene scores 1 and 2):


True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
forecASD
TADA (Sanders)
Krishnan
DAMAGES
BrainSpan+STRING

performance

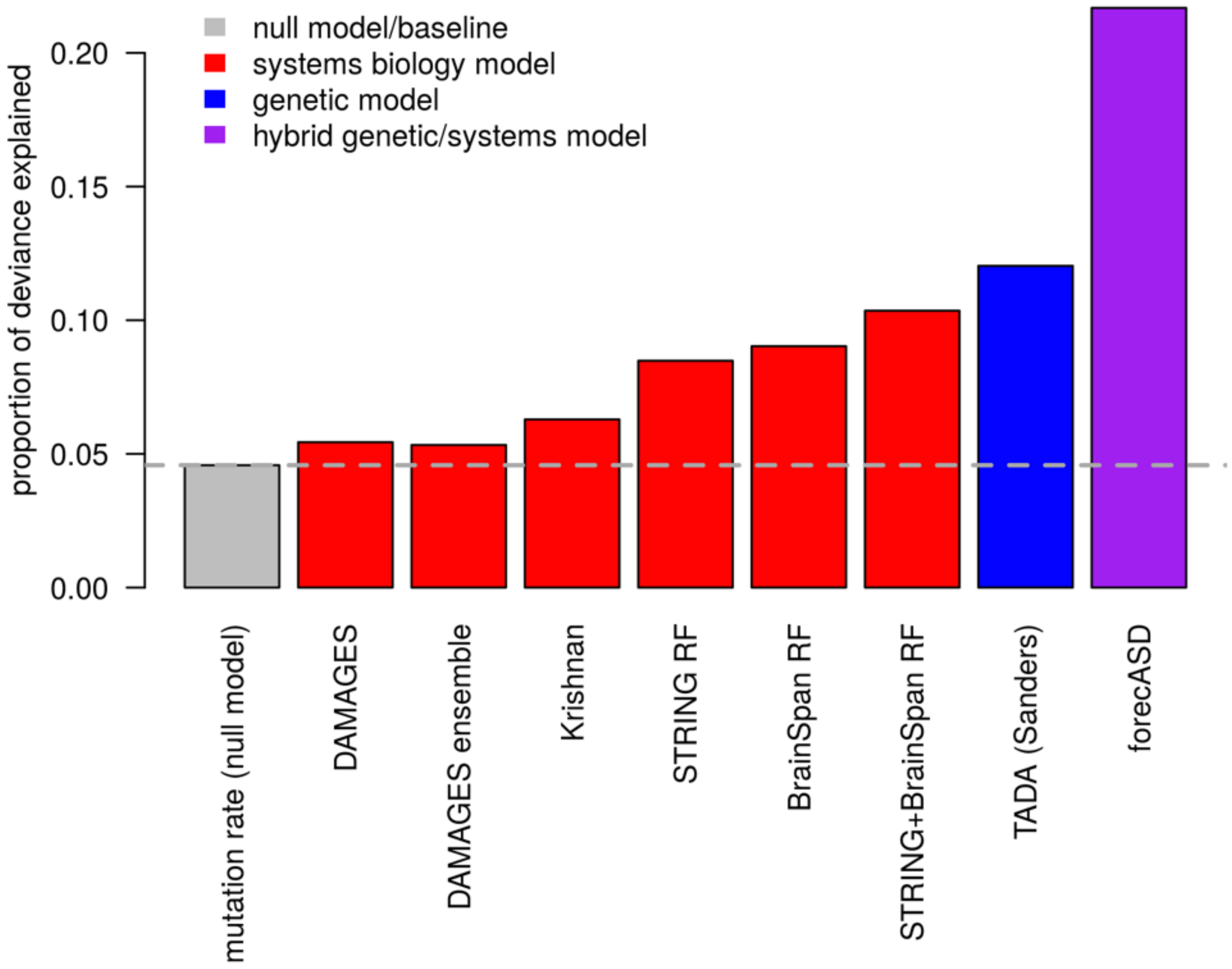
- >> next, we looked at “on the cusp” ASD genes (SFARI Gene score of 3)
  - >> this is where things get more interesting because it speaks to each method’s capacity for sensitivity to new discoveries:


True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
forecASD
TADA (Sanders)
Krishnan
DAMAGES
BrainSpan+STRING

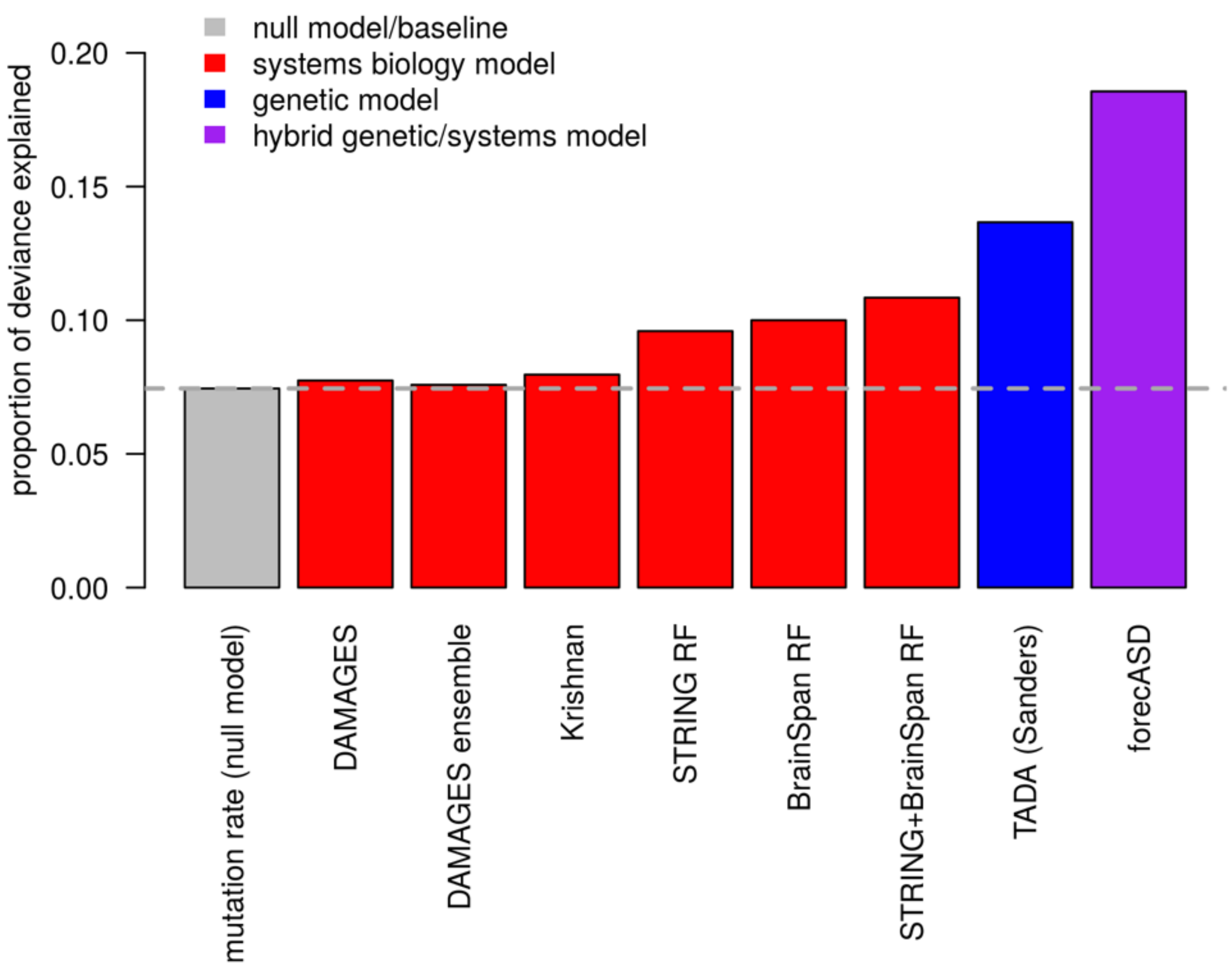
performance

- >> next, we took a slightly different approach: effectiveness as a predictor of mutation excess
- >> a good ASD gene score will predict mutation counts above and beyond what mutation rate alone will predict
  - >> adding the score to a statistical model of mutation counts (already including mutation rate) should result in increased deviance explained
- >> we pulled mutations from denovo-db:

denovo-db:
excess DNMs
in SFARI genes



denovo-db:
excess DNMs
in **non**-SFARI genes



performance

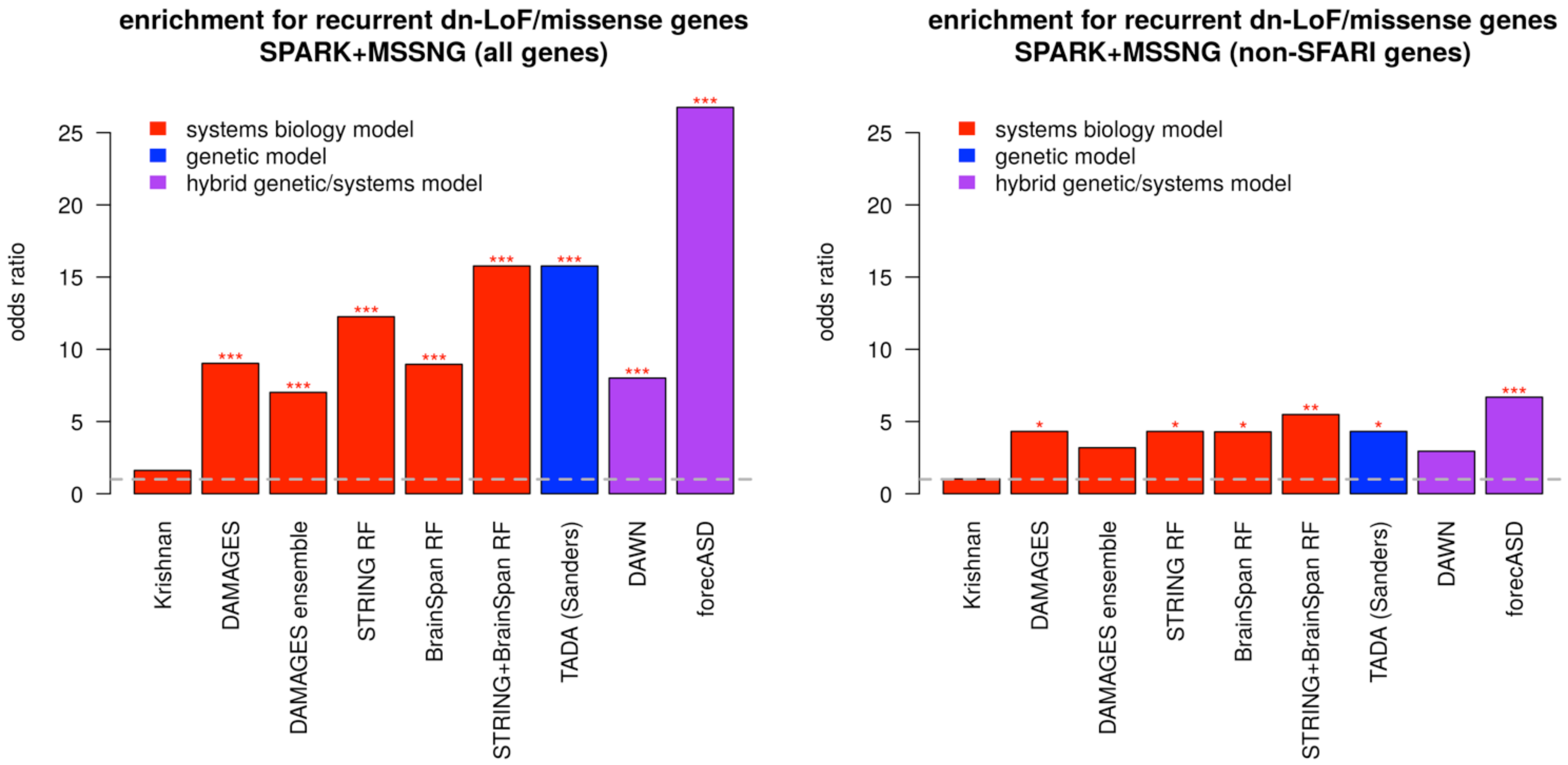
>> to see whether this trend holds, we looked at mutations in the SPARK pilot data (which is not yet in denovo-db)

>> note that the effect is much less pronounced because the SPARK pilot has far fewer mutations than denovo-db

SPARK

Igniting autism research
Improving lives

de novo mutations



functional characteristics

>> higher forecASD scores:

   >> higher pLI (loss-of-function-intolerant genes)

   >> increasing gene size

   >> more citation in ASD literature

   >> more protein-protein interactions (network hubs)

   >> functional enrichment

applications

- where is a genome-wide, gene-wise score useful?
  - parsing risk genes in CNV loci
  - focusing burden/association tests
    - non-coding elements
  - understanding the “big picture”

generalization: new ASD studies

>> I chatted with a researcher from a prominent group; they're about to submit a new "big" paper in ASD gene discovery

>> they proposed 17 "new" ASD genes

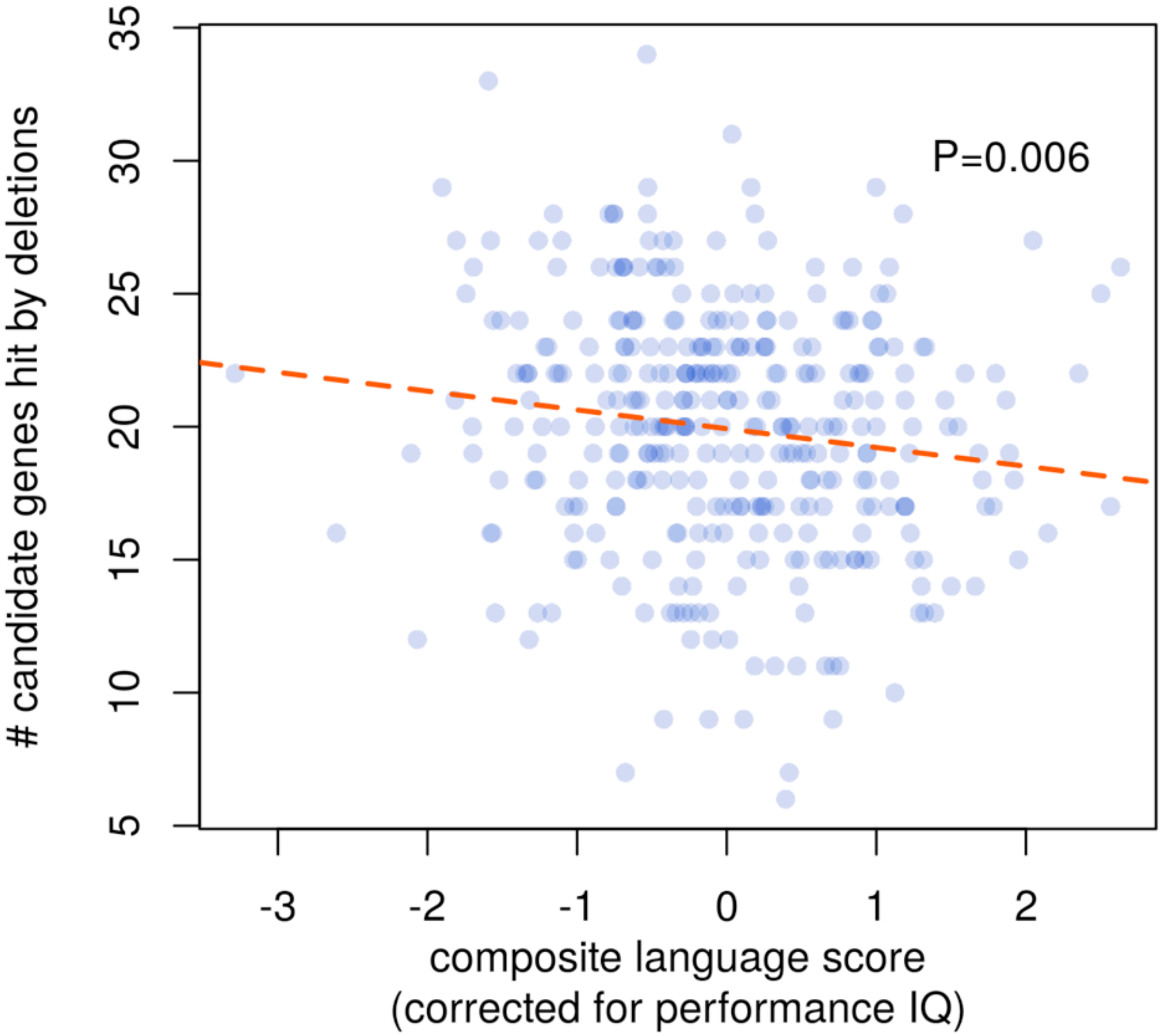
  >> our score predicted 13 of them as probable ASD risk genes

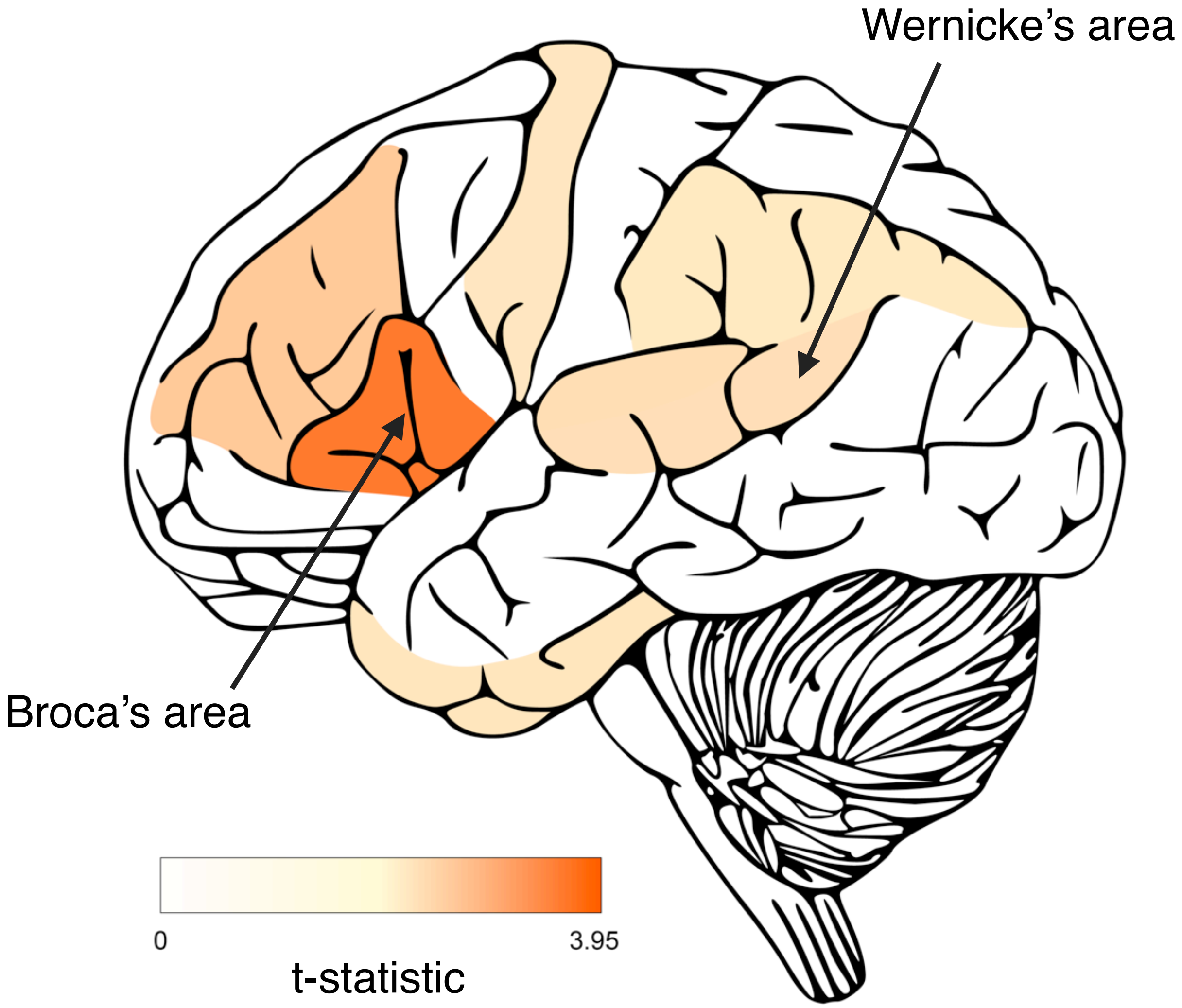
generalization: SLI

>> we are curious whether ASD genes are enriched for genes involved in language (seems reasonable)

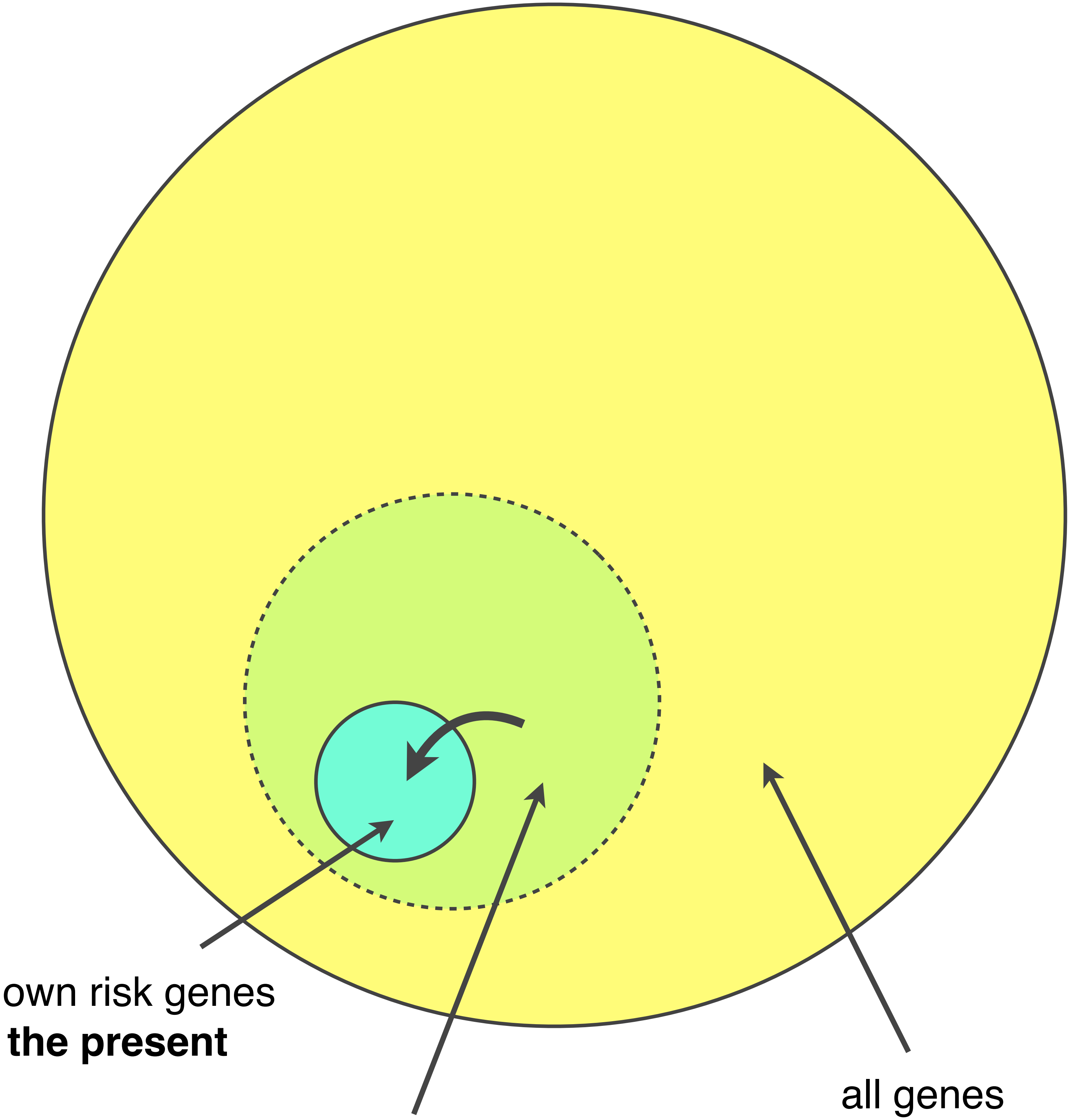
>> we have a N∼400 cohort of WGS on specific language impairment (SLI)

>> we counted the number of forecASD genes hit by a deletion (per individual) and then looked for an association with language ability:


P=0.006
# candidate genes hit by deletions
35
30
25
20
15
10
5
-3
-2
-1
0
1
2
composite language score
(corrected for performance IQ)


Wernicke’s area
Broca’s area
0
3.95
t-statistic

predicting the future



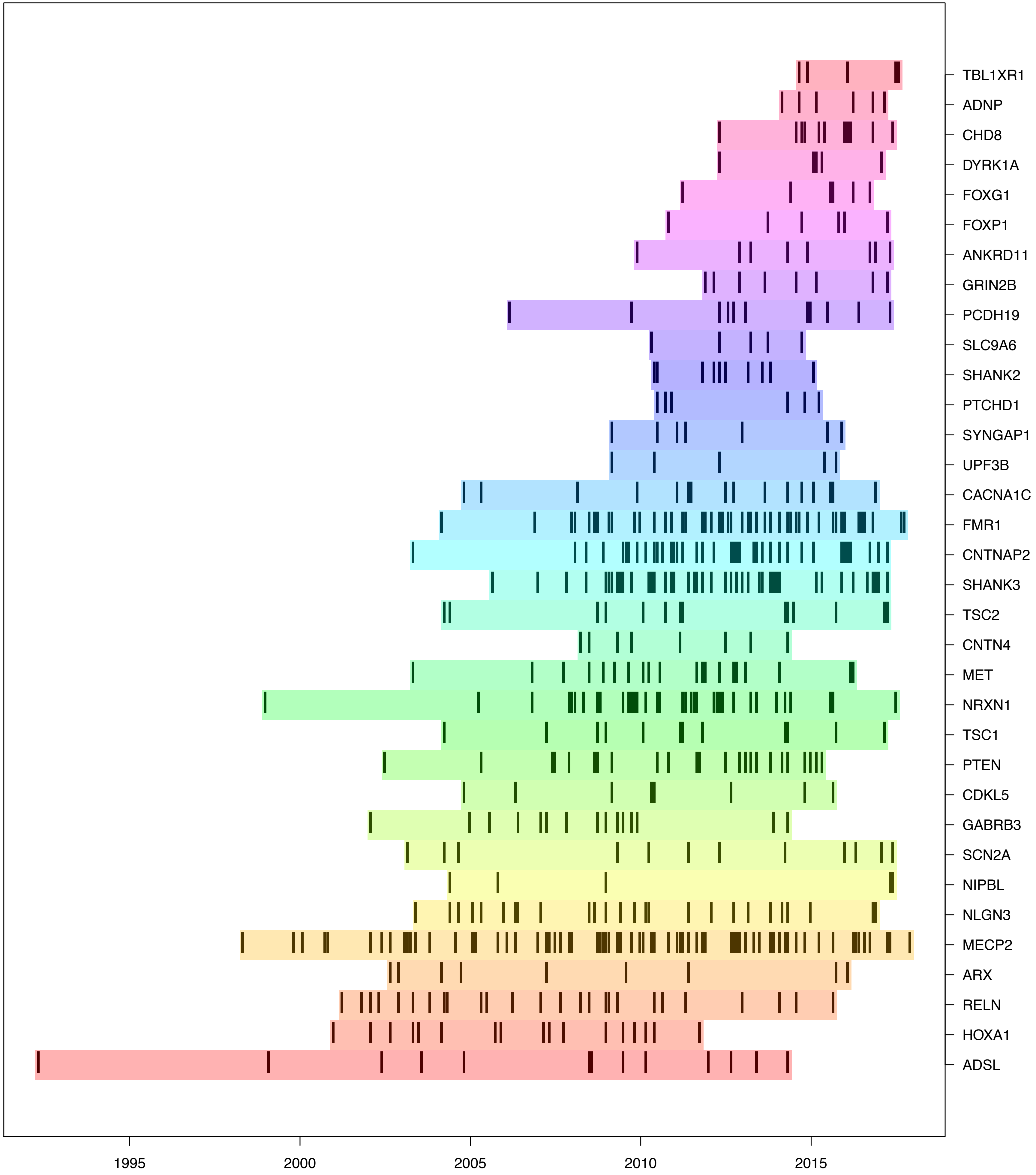
known risk genes
the present

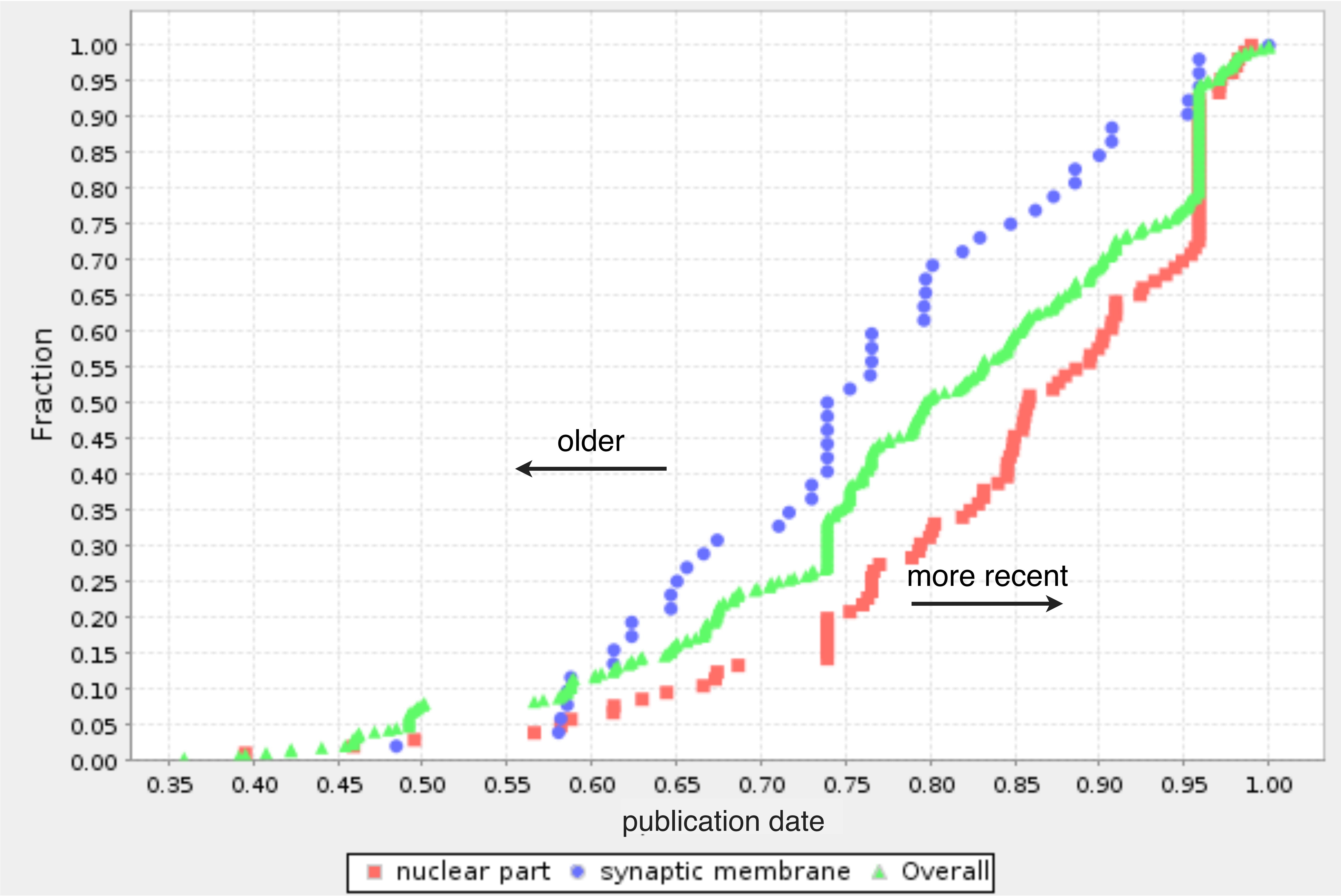
all genes

less established (or unknown) risk genes
the future

GSEA on time

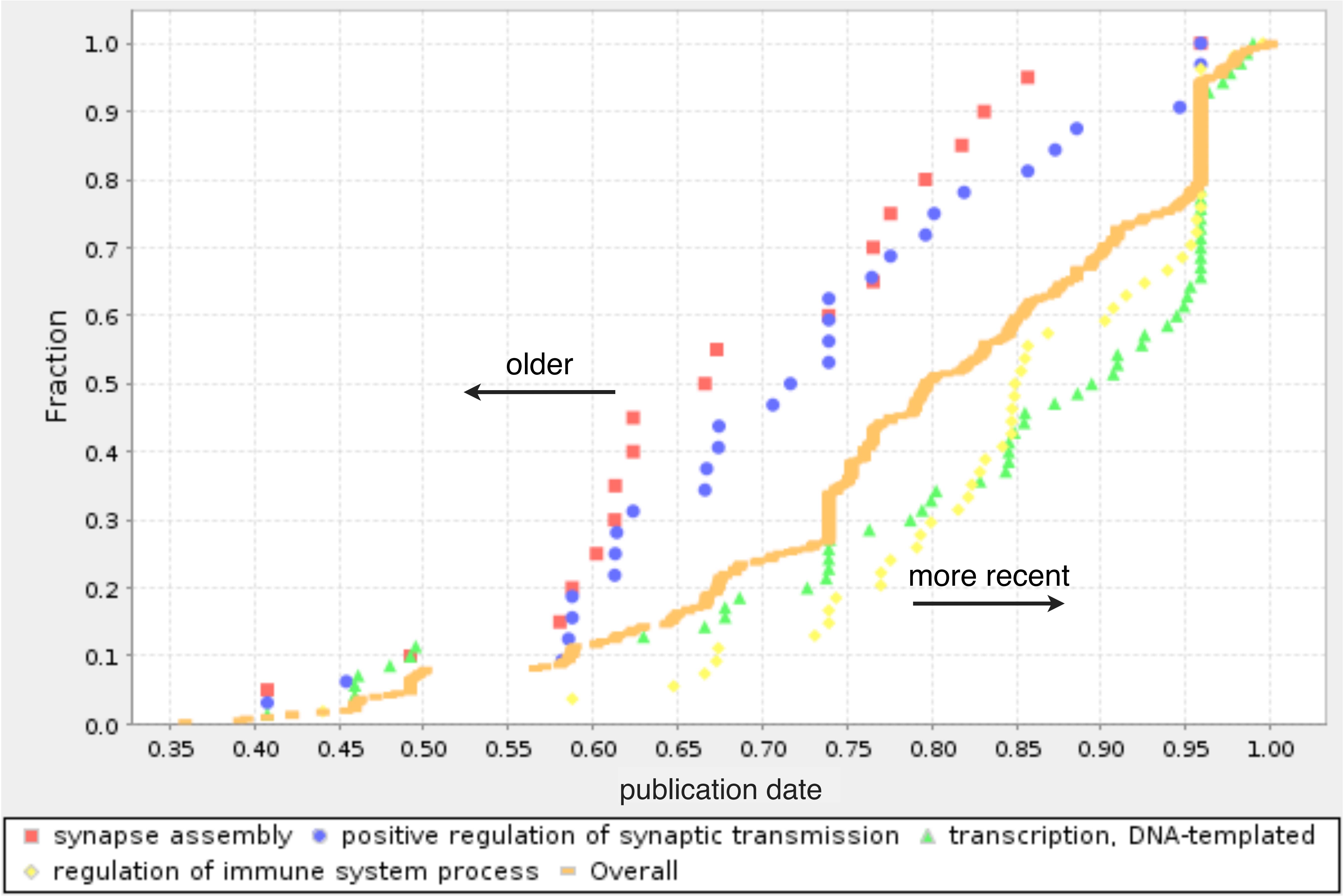
- gene set enrichment analysis (GSEA) allows one to test if biological annotations are associated with a quantitative gene-wise score (e.g. p-value or fold change)
- what if we used the date that a gene “became an ASD risk gene”* as that quantitative measure?
    - a GSEA in this case would tell us which biological themes were significantly earlier, vs. significantly more recent in the literature


TBL1XR1
ADNP
CHD8
DYRK1A
FOXG1
FOXP1
ANKRD11
GRIN2B
PCDH19
SLC9A6
SHANK2
PTCHD1
SYNGAP1
UPF3B
CACNA1C
FMR1
CNTNAP2
SHANK3
TSC2
CNTN4
MET
NRXN1
TSC1
PTEN
CDKL5
GABRB3
SCN2A
NIPBL
NLGN3
MECP2
ARX
RELN
HOXA1
ADSL
1995
2000
2005
2010
2015

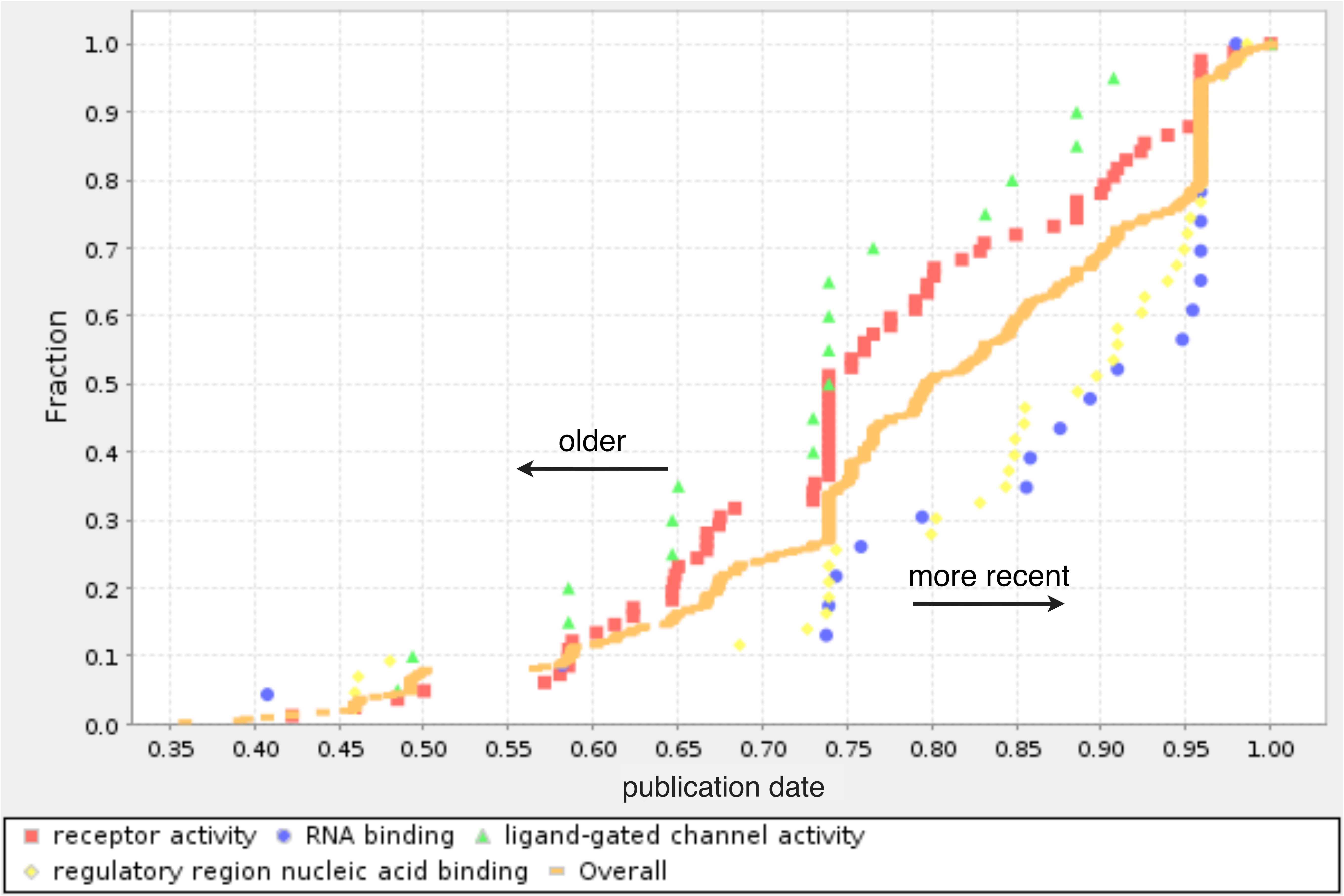


older

more recent

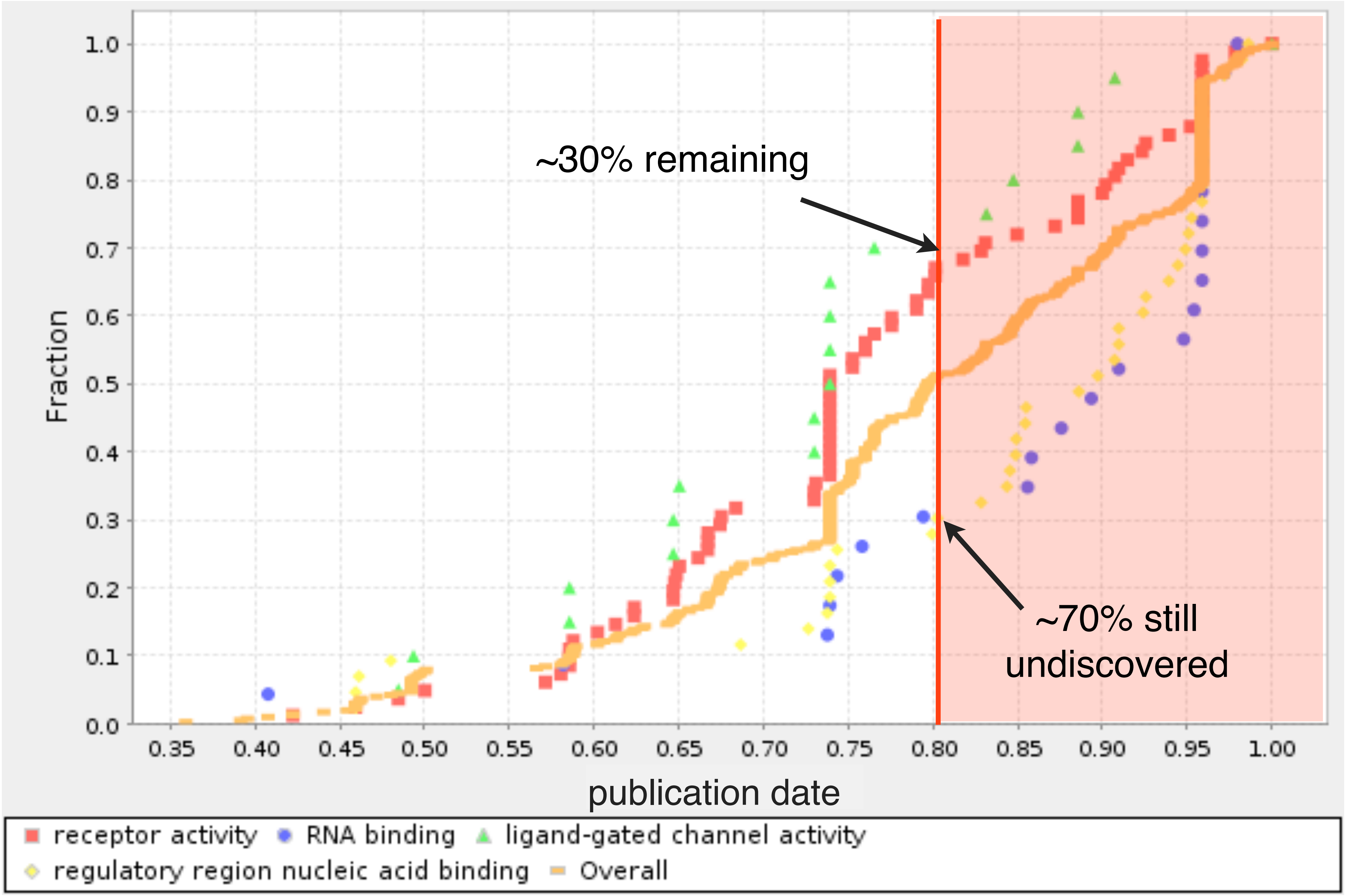


Fraction
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1.00
publication date
older
more recent
🟥 synapse assembly
● positive regulation of synaptic transmission
▲ transcription, DNA-templated
◇ regulation of immune system process
▬ Overall


Fraction
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1.00
publication date
older
more recent
🟥 receptor activity
● RNA binding
▲ ligand-gated channel activity
◇ regulatory region nucleic acid binding
▬ Overall

GSEA on time

- so a basic conclusion that we can draw from this is that synaptic risk is a long-established topic in the literature, while nuclear risk is a relatively new(er) development
    - this is probably intuitive for anyone working in ASD genetics
- how can we use this approach to anticipate currently under-represented molecular themes in ASD?



~30% remaining

~70% still undiscovered

predicting what's next

>> two comparisons:

>> 1) what are currently known ASD genes enriched for (compared to all genes)?

>> 2) what are novel forecASD genes enriched for?

>> what is the difference between the lists? This may give clues as to where potential new discoveries may come from.

predicting what’s next

Reactome pathway enrichment (a sampling)

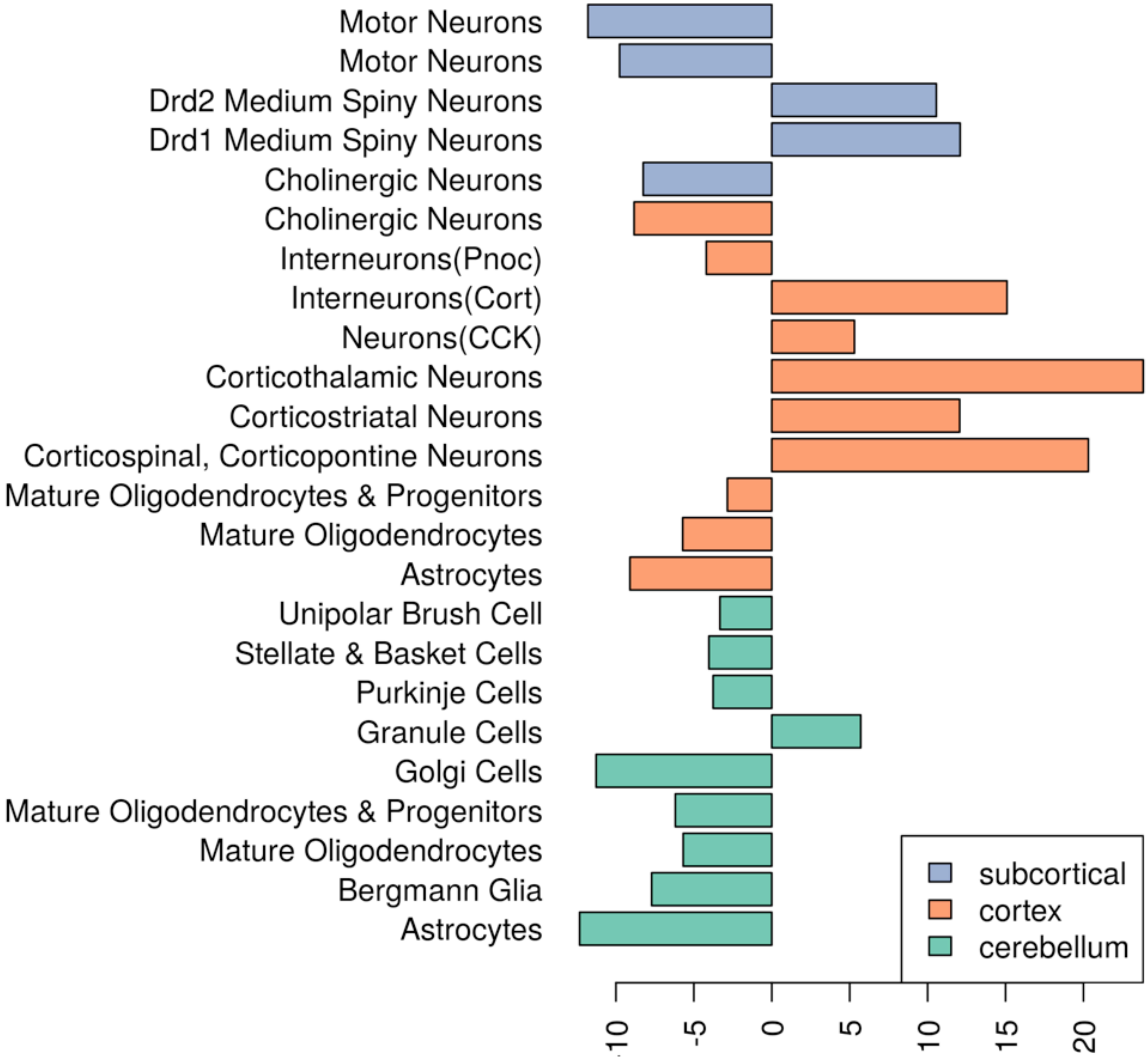
annotation	known (enrichment)	novel (enrichment)
mRNA Splicing	1.4x	**4.1x****
circadian clock	3.9x	**7.4x****
epigenetic regulation of gene expression	2.2x	**4.3x****
PPARA activates gene expression	3.3x	**4.6x****

what cell types in the brain?

>> there are a zillion kinds of cell in the brain

>> ASD has some characteristic behaviors, so it would make sense that there are specific cell types and circuits that are more related to ASD than not

>> with our “master” list of ASD genes, we can predict which cell types are most likely to be impacted by genetic risk


Motor Neurons
Motor Neurons
Drd2 Medium Spiny Neurons
Drd1 Medium Spiny Neurons
Cholinergic Neurons
Cholinergic Neurons
Interneurons(Pnoc)
Interneurons(Cort)
Neurons(CCK)
Corticothalamic Neurons
Corticostriatal Neurons
Corticospinal, Corticopontine Neurons
Mature Oligodendrocytes & Progenitors
Mature Oligodendrocytes
Astrocytes
Unipolar Brush Cell
Stellate & Basket Cells
Purkinje Cells
Granule Cells
Golgi Cells
Mature Oligodendrocytes & Progenitors
Mature Oligodendrocytes
Bergmann Glia
Astrocytes
subcortical
cortex
cerebellum
-10
-5
0
5
10
15
20

summary

>> forecASD combines both new and existing predictors of ASD involvement

  >> showed superior performance in recovering known genes and identifying probable genes

  >> a holistic, “living” approach to identifying genes on an individual basis

  >> ...and currently under-represented themes/pathways that suggest new directions

take-home messages

take-home messages

- studying the genetics of autism holds promise for the community
    - early diagnosis
    - individualized therapy
    - increased understanding and coping
    - increased understanding of human uniqueness

take-home messages

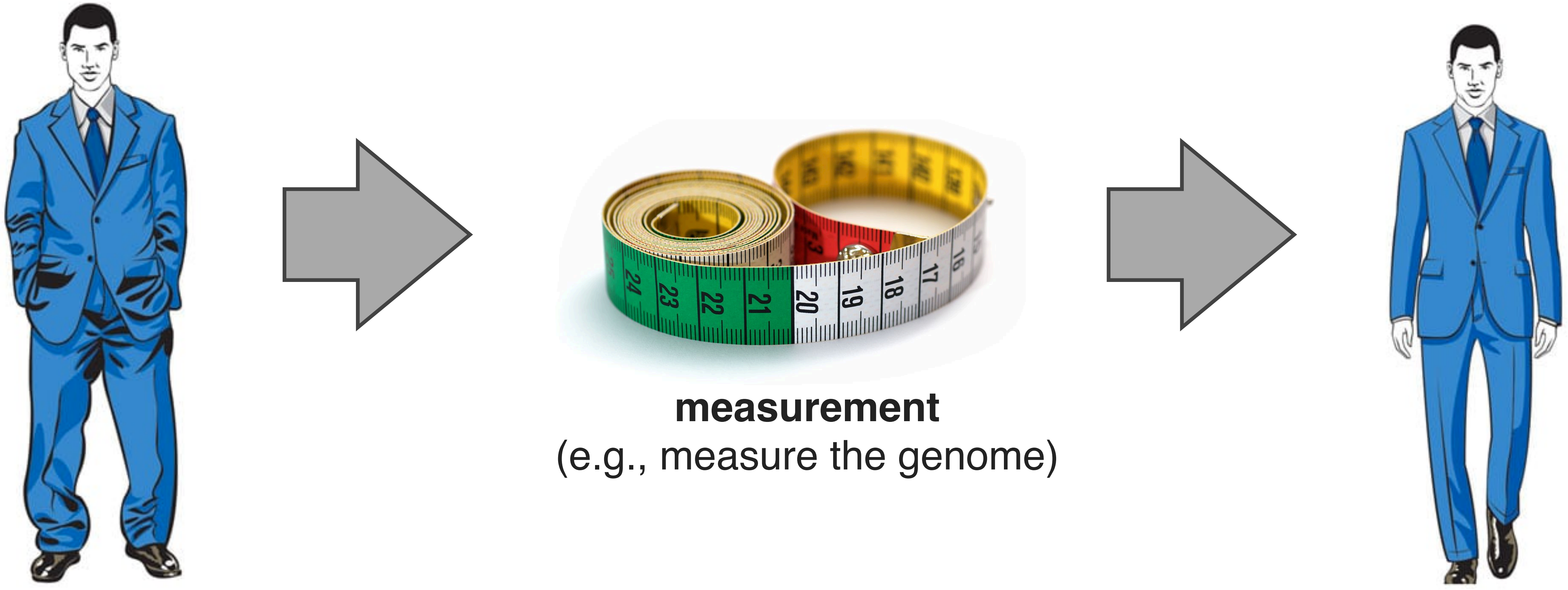
- >> autism has myriad risk factors, but genetics dominates all others
    - >> the sum of lots of tiny genetic risk factors accounts for most of the risk
        - >> this kind of risk is a two-edged sword: correlated with educational attainment
    - >> we expect that about 1000 genes give rise to the many types of autism
        - >> rare, damaging variation in these genes contribute an additional kind of risk (best at identifying specific genes)

take-home messages

>> community building (like SPARK) represents the next generation of autism research

>> more interaction, more back and forth, more partnership

>> more personalized medicine



measurement
(e.g., measure the genome)

acknowledgments

- >> Funding support:
  - >> NIMH, NIDCD, Simons Foundation, BBRF
- >> key lab members:
  - >> Leo Brueggeman, MSTP student
  - >> Natalie Pottschmidt, project coordinator
- >> SPARK participants
- >> SPARK Network Analysis WG

- >> questions?
  - >> jacob-michaelson@uiowa.edu