Functional Context Matters: Building and Applying a New Dictionary for Human Development

Donna K. Slonim

Dept. of Computer Science, Tufts University

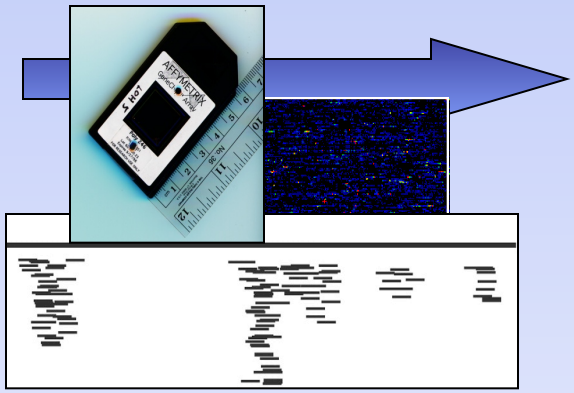
Dept. of Integrative Physiology and Pathobiology, Tufts University School of Medicine

Genetics Faculty, Sackler School of Graduate Biomedical Sciences

Tufts UNIVERSITY

September 1, 2015

Bioinformatics for Human Development



Goals:

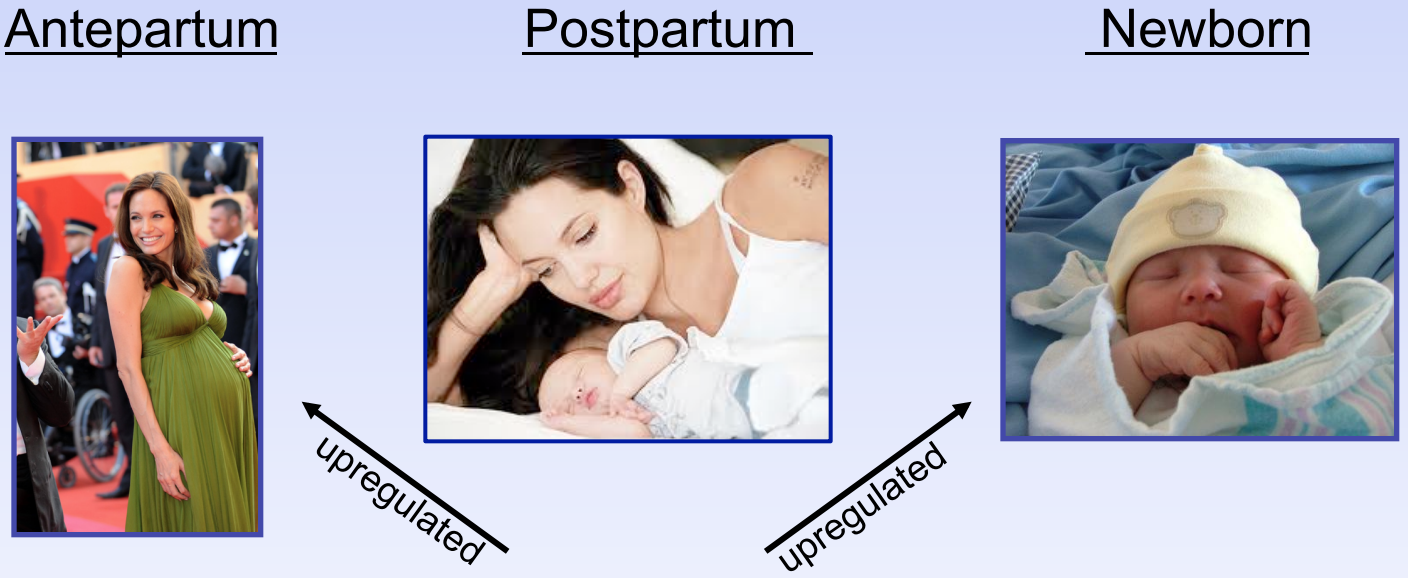
- Better understand normal development
- Diagnose developmental abnormalities
- Suggest novel therapies
- Discover developmental implications for adult disease

Tufts UNIVERSITY

Functional Analysis of Prenatal Gene Expression: Fetal RNA in Maternal Blood

Maron, et al., *J Clin Invest* 2007

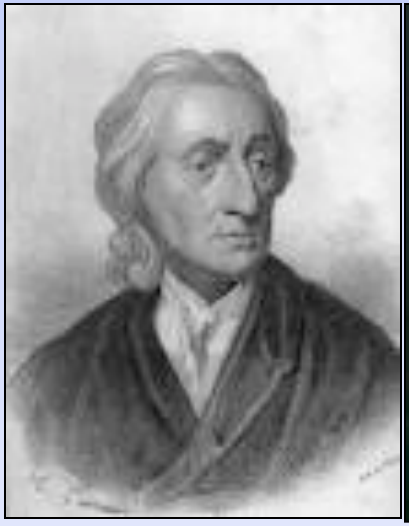
157 genes upregulated in pregnant moms and their babies, vs. postpartum moms.

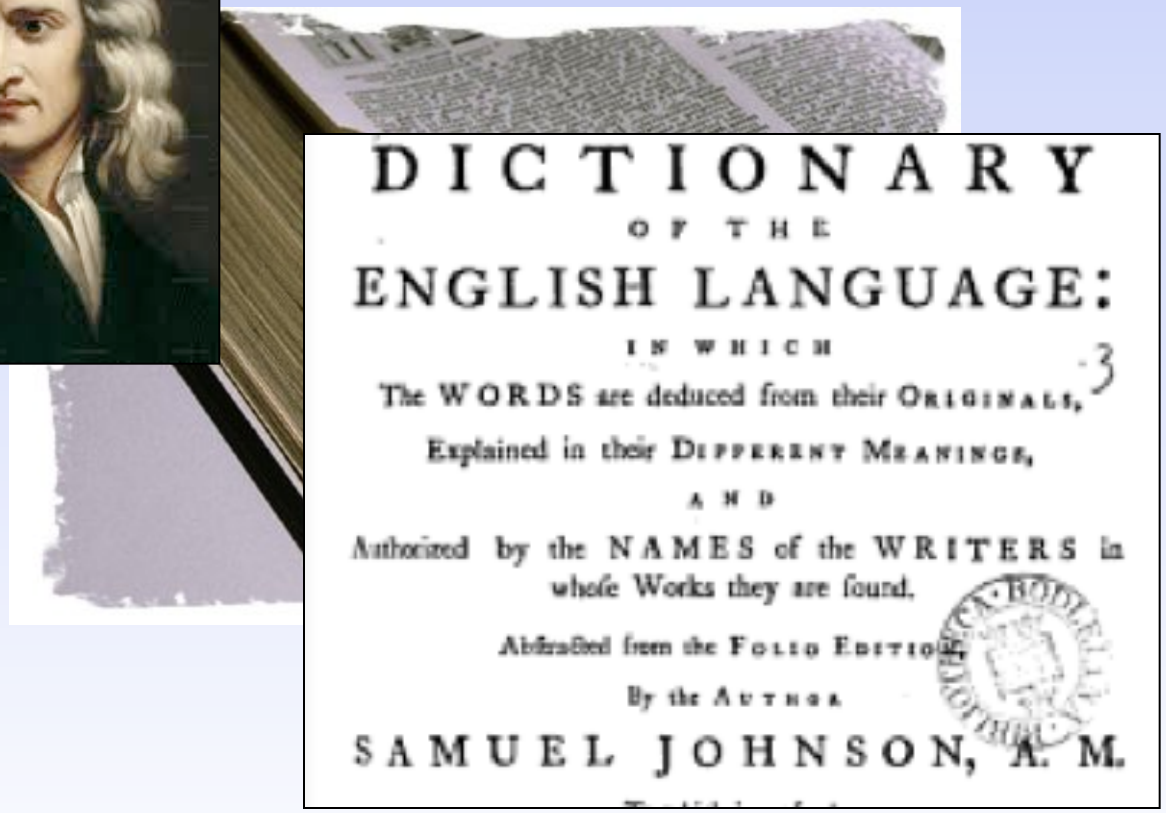


Automated functional analysis hinted at fetal origin

Evidence from manual review was much stronger. Why?

Using the Wrong Dictionary





We need to create our own dictionary.

DFLAT

Developmental FunctionaL Annotation at Tufts

Wick, et al., *BMC Bioinfo* 2014

keywords, genes, developmental processes

PubMed.gov

Manual curation

Literature-derived annotation

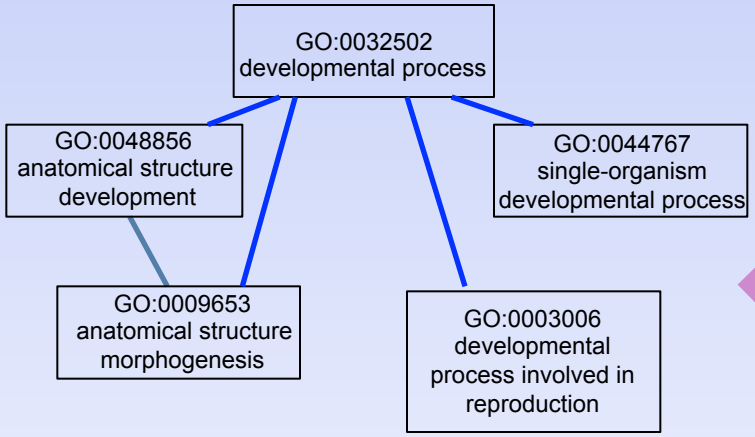
GENEONTOLOGY
Unifying Biology

GONE

DFLAT

Developmental FunctionaL Annotation at Tufts

Wick, et al., *BMC Bioinfo* 2014



Mouse developmental genes

Total: 13,344 new annotations

Human orthologs

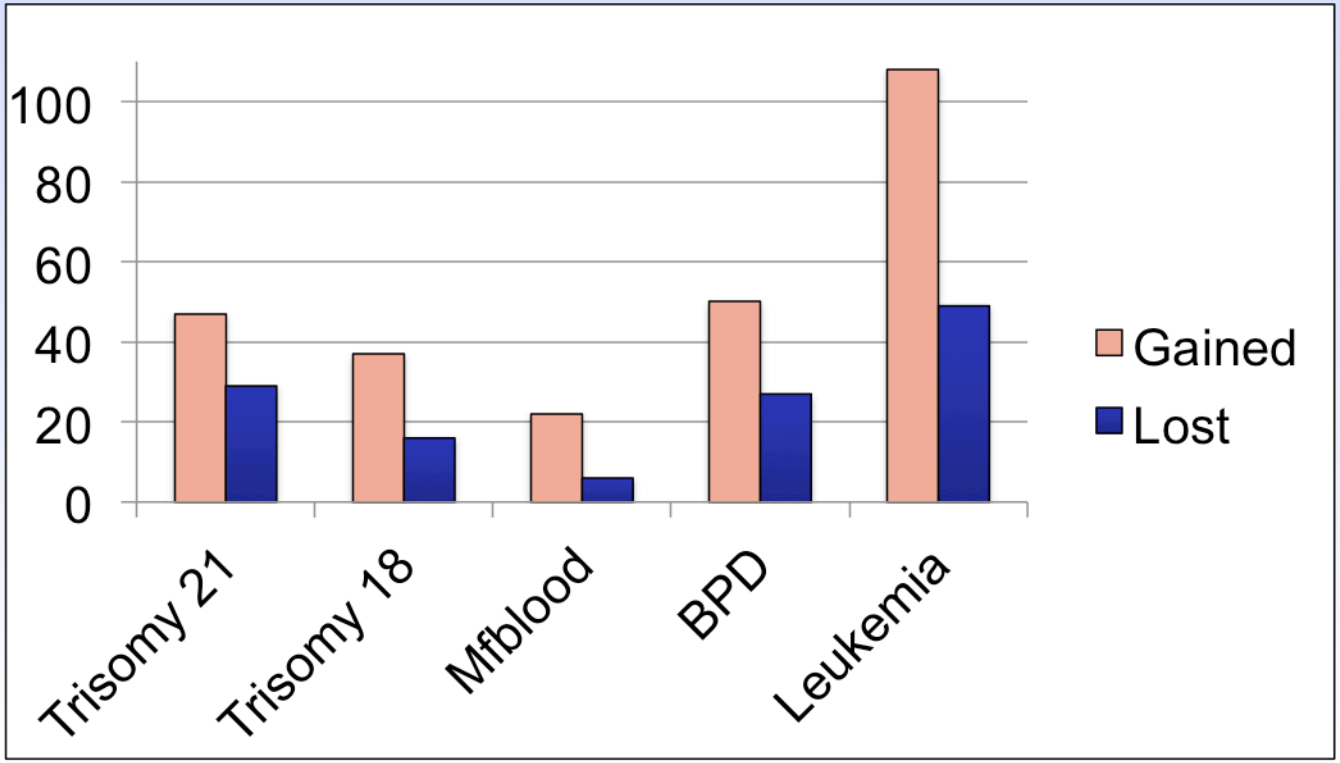
Ortholog-derived annotation

Tufts UNIVERSITY

DFLAT's impact on expression analysis

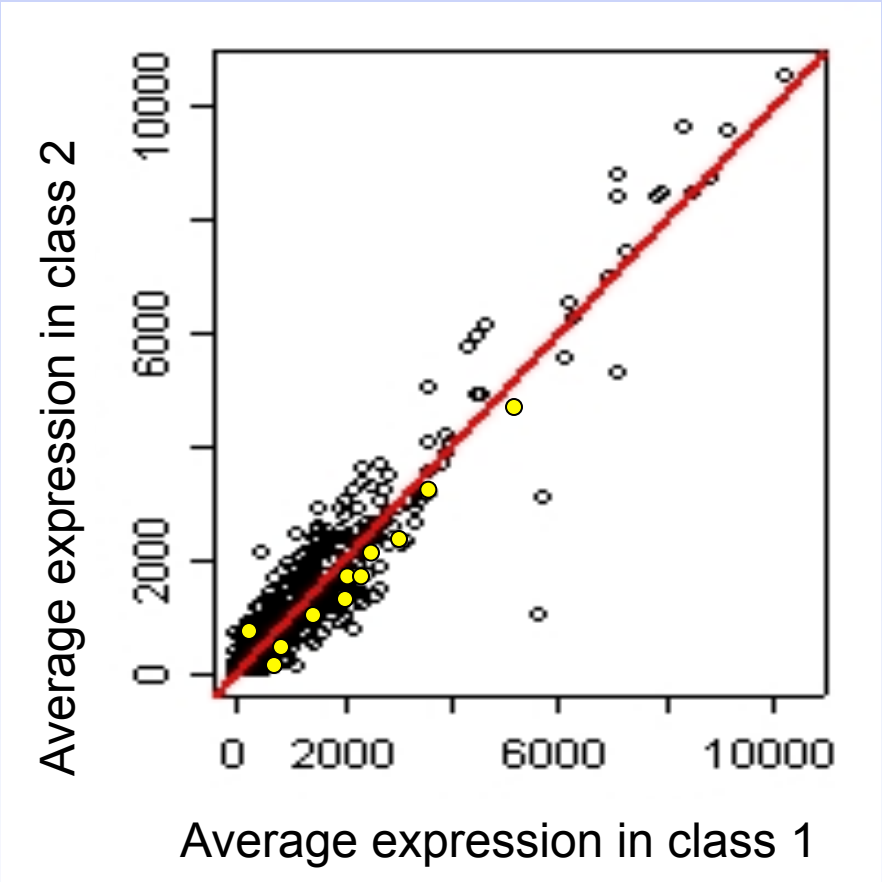
GSEA with GO Biological Process terms

Significant gene sets gained/lost with DFLAT:



Tufts UNIVERSITY

Differential expression in gene *sets*



Differences in expression may be subtle for each gene.

But, if entire set shows consistent changes, we can still detect this.

GSEA software from Broad Institute does this (as do others)

But, need to know gene sets ahead of time.

Tufts UNIVERSITY

Fetal RNA just before birth: New results with DFLAT

"I'm getting ready to feed"

"I am going to be challenged by new odors"

"I will see something soon"

"My bones and muscles are still developing"

"I am going to have to fight bacteria on my own"

"My nervous system is getting ready for me to face the world"

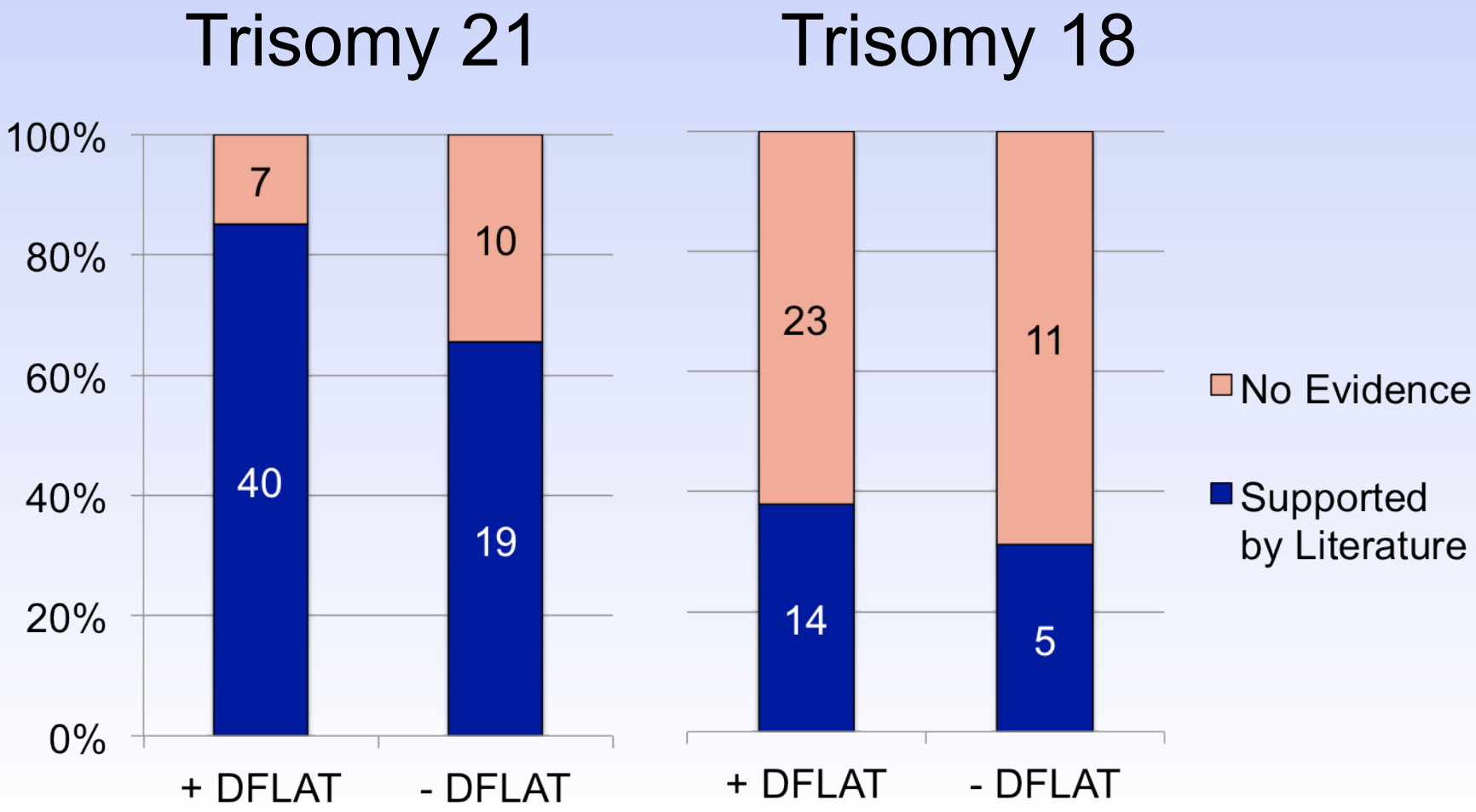
"I'm going to need to breathe for myself"

"Sounds and speech are going to be important to me"

Tufts UNIVERSITY

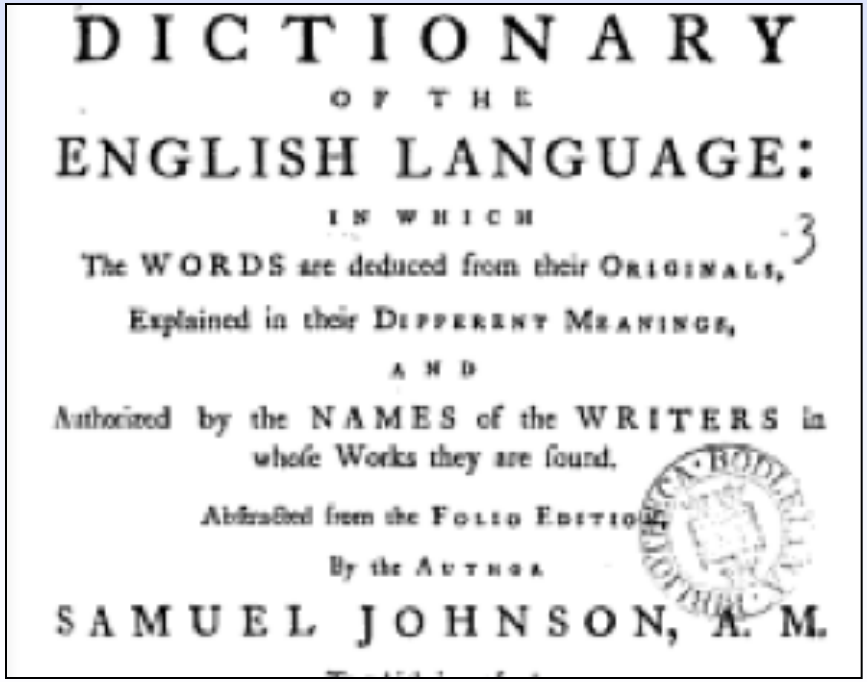
DFLAT's impact on expression analysis

Verified new functions significant ($p < 0.05$) in:



Applying the Dictionary

Characterizing Individual Developmental Anomalies



DICTIONARY

OF THE

ENGLISH LANGUAGE:

IN WHICH

The WORDS are deduced from their ORIGINALS,

Explained in their DIFFERENT MEANINGS,

AND

Authorized by the NAMES of the WRITERS in whose Works they are found.

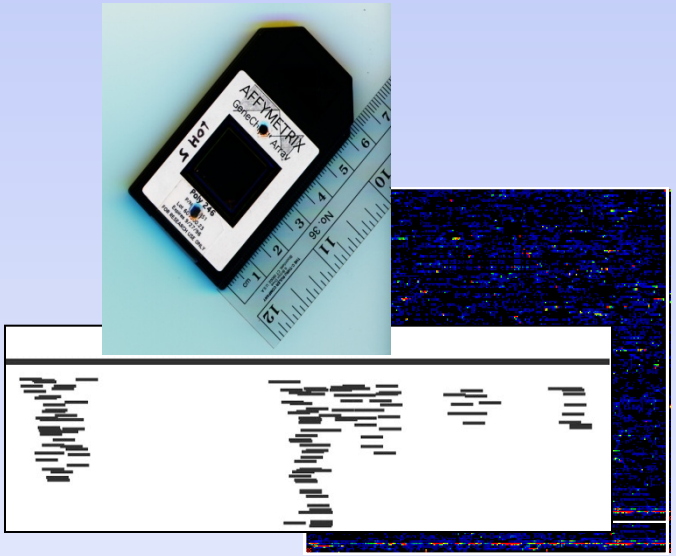
Abridged from the FOLIO EDITION

By the AUTHOR

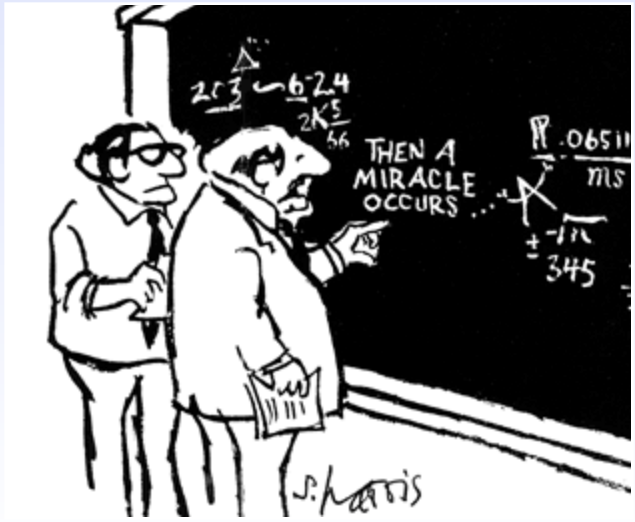
SAMUEL JOHNSON, A. M.

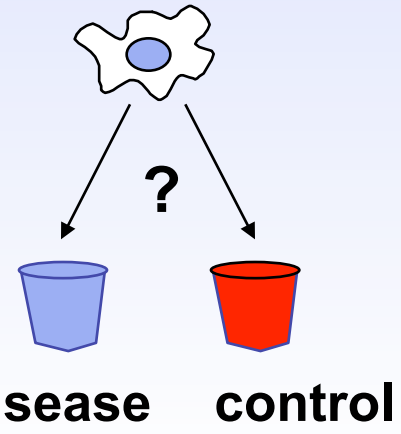
Tufts UNIVERSITY

Expression Data Analysis

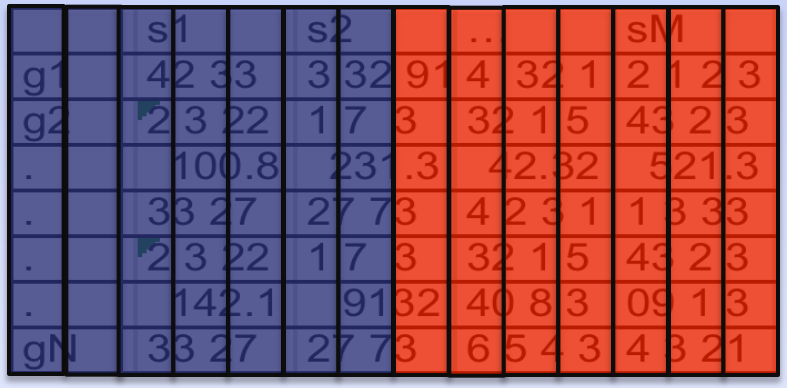


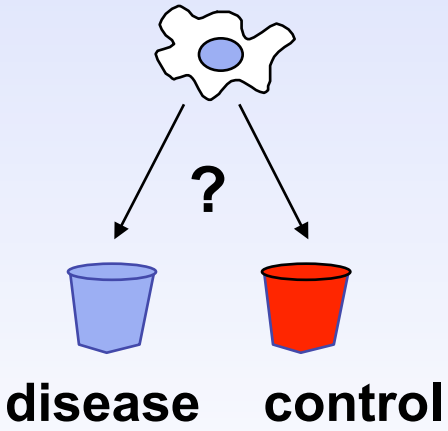
	s1	s2	…	sM
g1	3.2	23		8.2
g2	4.6	4		4.1
.	100.7	143		127
.	33	72		86
.	22.3	16		19
.	66	38		47
gN	5.3	7.2		3.7





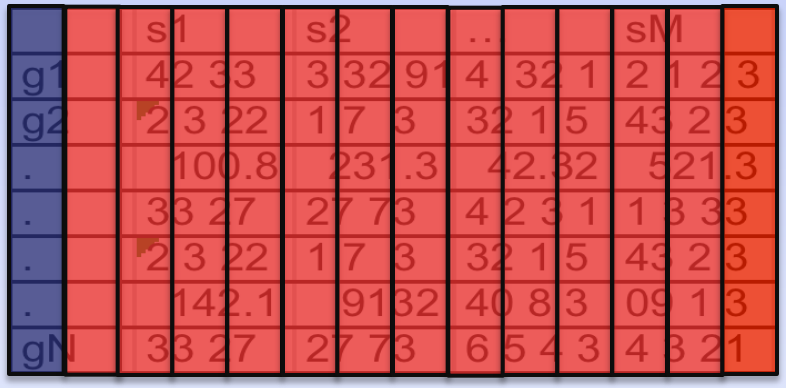
Training Data

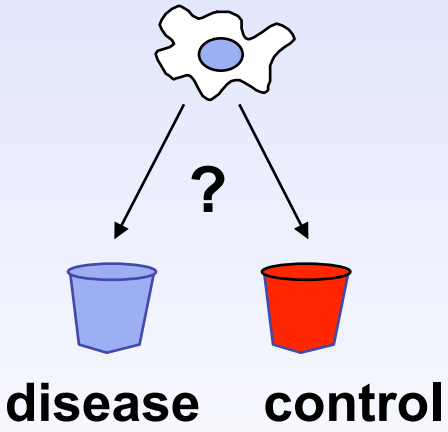




Tufts UNIVERSITY

The Challenge of Rare Samples





Tufts UNIVERSITY

Anomaly Detection

Computational field of identifying unusual data points in multi-dimensional space

Training data: normal examples only

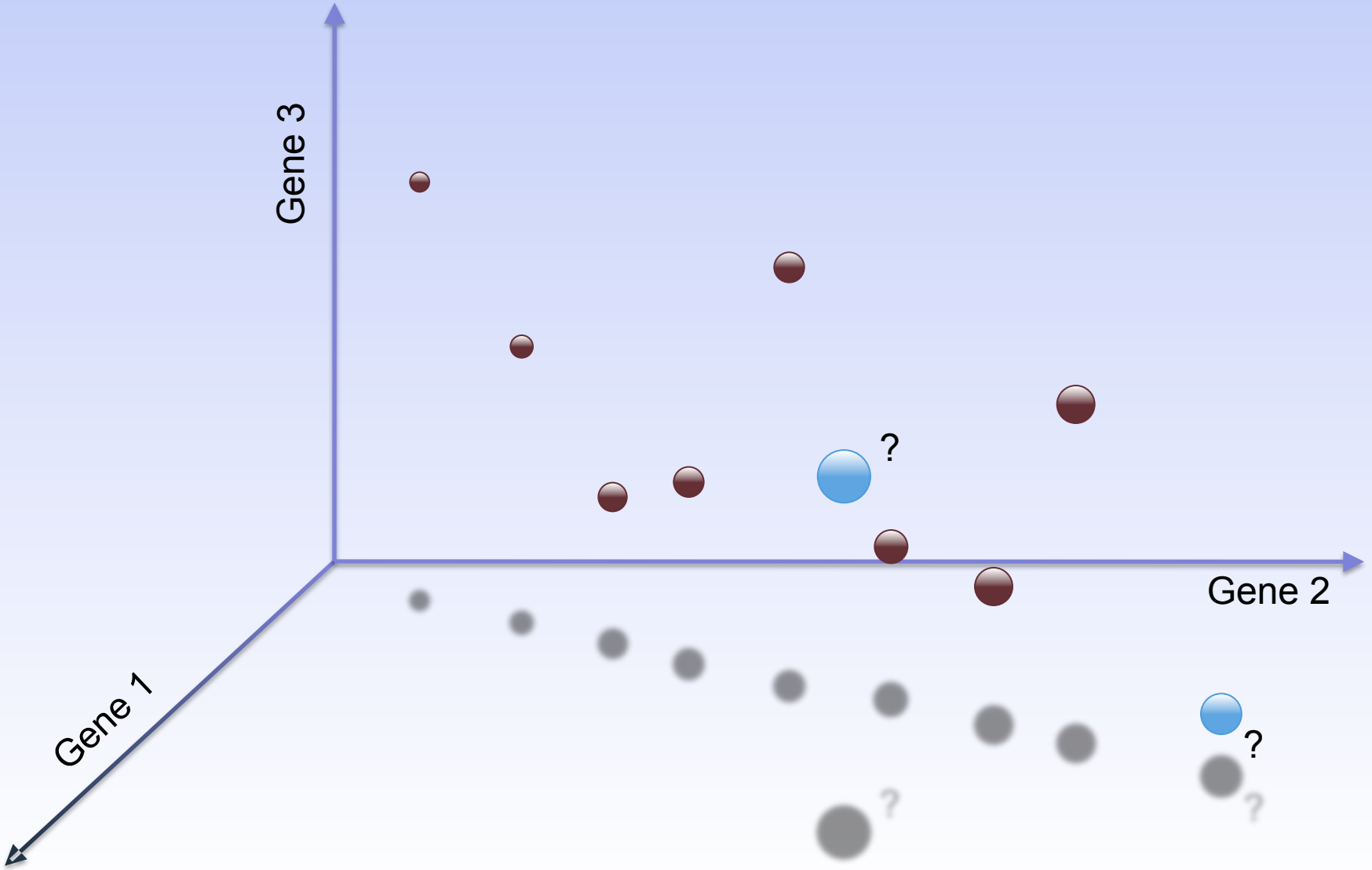
Many methods; best include:

- 1-class SVMs
- Local Outlier Factor (LOF)

But distance-based methods struggle in high-dimensional spaces

Feature Regression and Classification (FRaC)

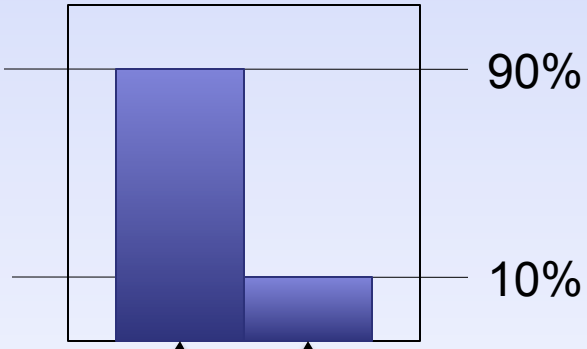
Noto, et al., *Intl Conf Data Min 2010*



Combine Predictors via Information Theory

Features (binary)	Reliability (on Training Data)	Error (on Query Data)
1	50% accuracy	misclassified
2	90% accuracy	correct
3	90% accuracy	misclassified

$$Surprisal = -log_2(p)$$



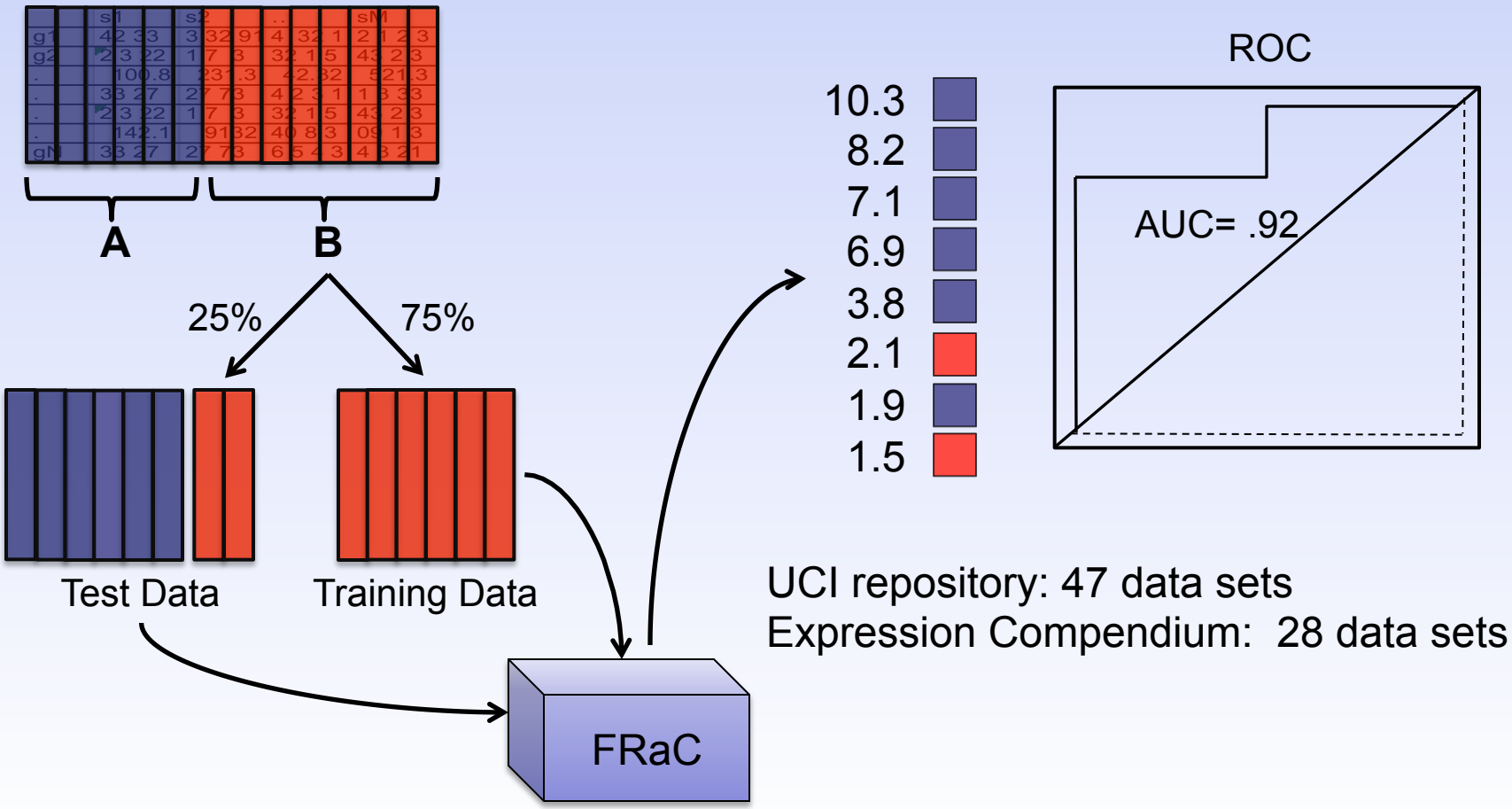
If I see this, it tells me 0.152 bits of information

If I see this, it tells me 3.32 bits of information

Tufts UNIVERSITY

Evaluating Anomaly Detection

Create data sets from classification tasks:

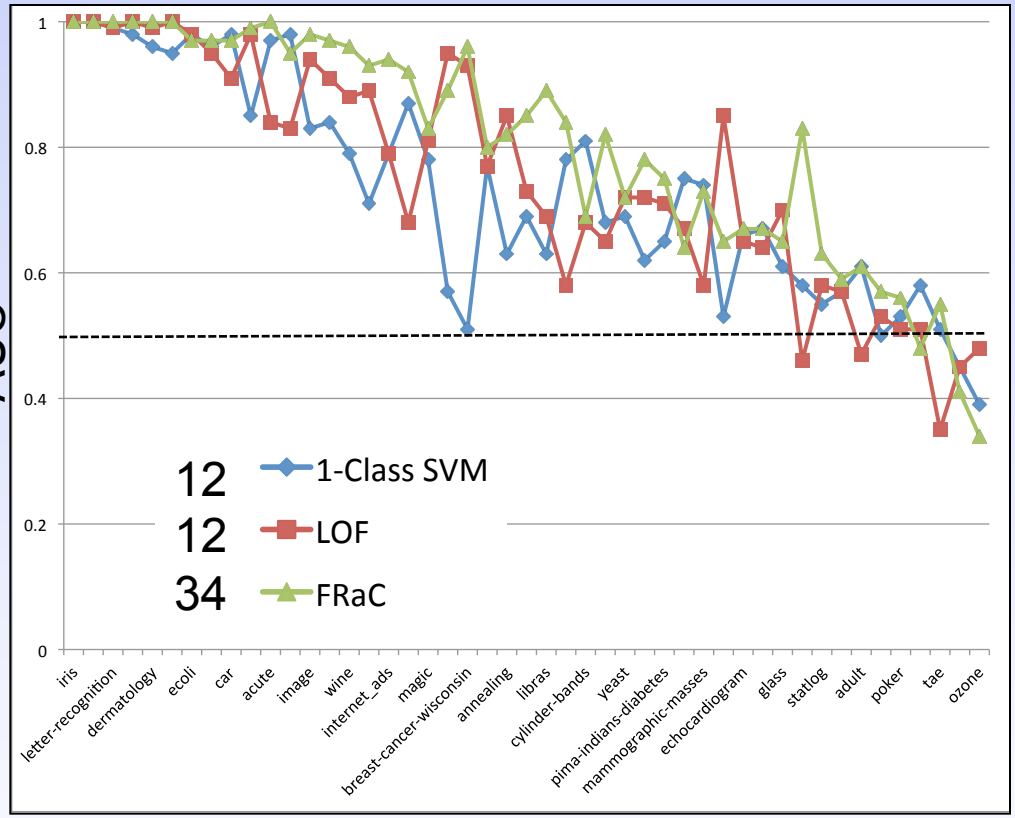


UCI repository: 47 data sets
Expression Compendium: 28 data sets

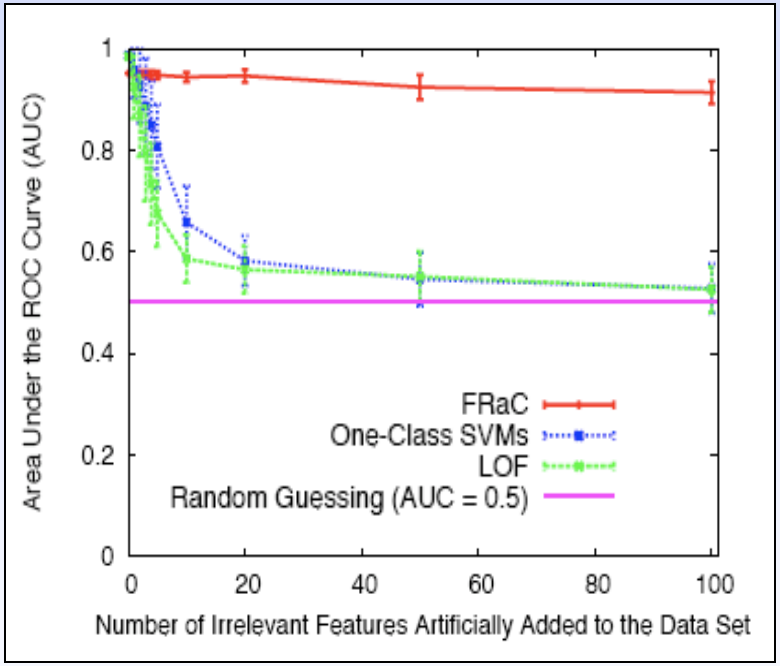
FRaC Wins on Machine Learning Data

47 UCI data sets

Noto, et al., *Data Min Knowl Disc 2012*

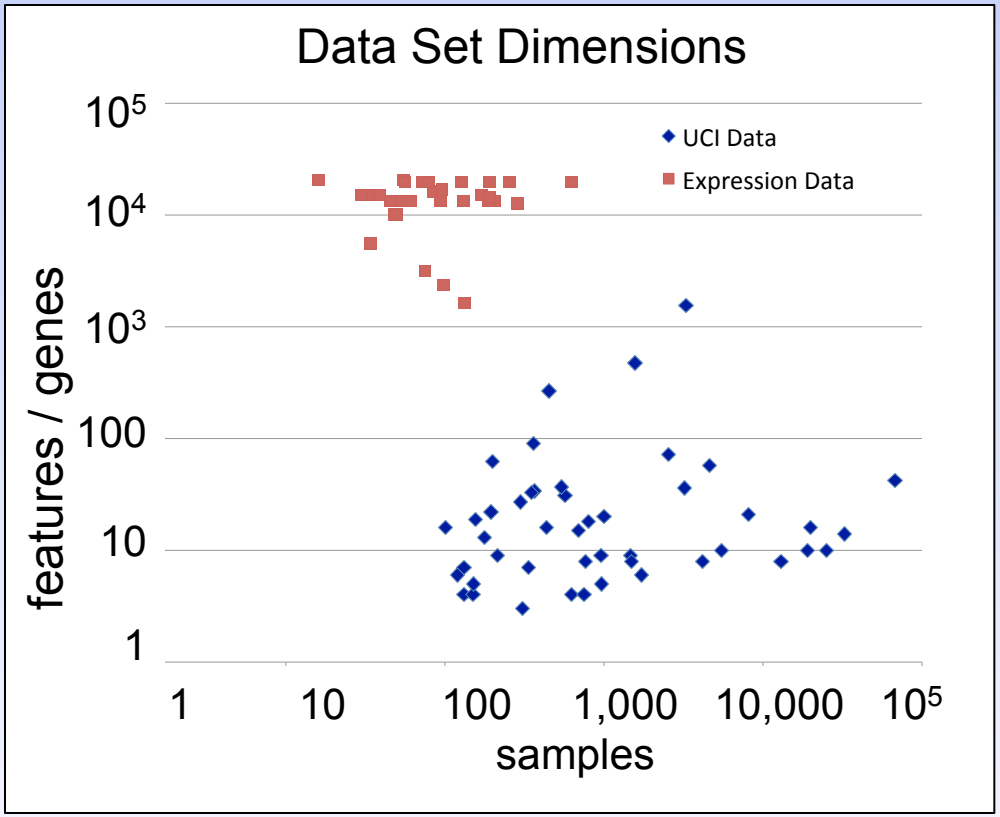


Much more robust to irrelevant features



Tufts UNIVERSITY

Why Genomic Problems are Harder



Dimensions

Irrelevant genes

Interpretation?

…which gene sets are anomalous?

Tufts UNIVERSITY

CSAX: Characterizing Systematic Anomalies in eXpression Data

Main Idea:

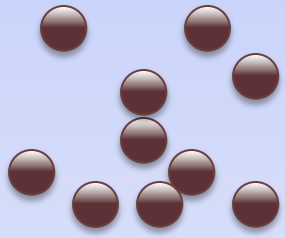
- Rank genes by FRaC surprisal
- Find enriched functions in ranked list of genes (GSEA)

But this “FRaC + enrichment” is not enough…

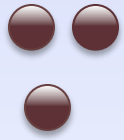
Tufts UNIVERSITY

Limited Normal Training Data

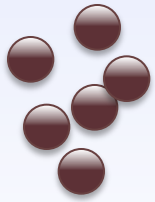
Bare minimum numbers of "normal" examples



Complex, heterogeneous high-dimensional domains



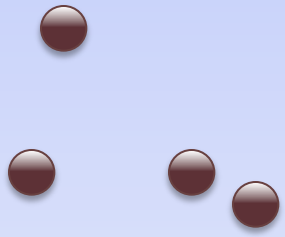
Add cross-validation:



Tufts UNIVERSITY

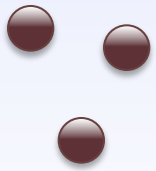
Limited Normal Training Data

Bare minimum numbers of "normal" examples



Complex, heterogeneous high-dimensional domains

Add cross-validation:



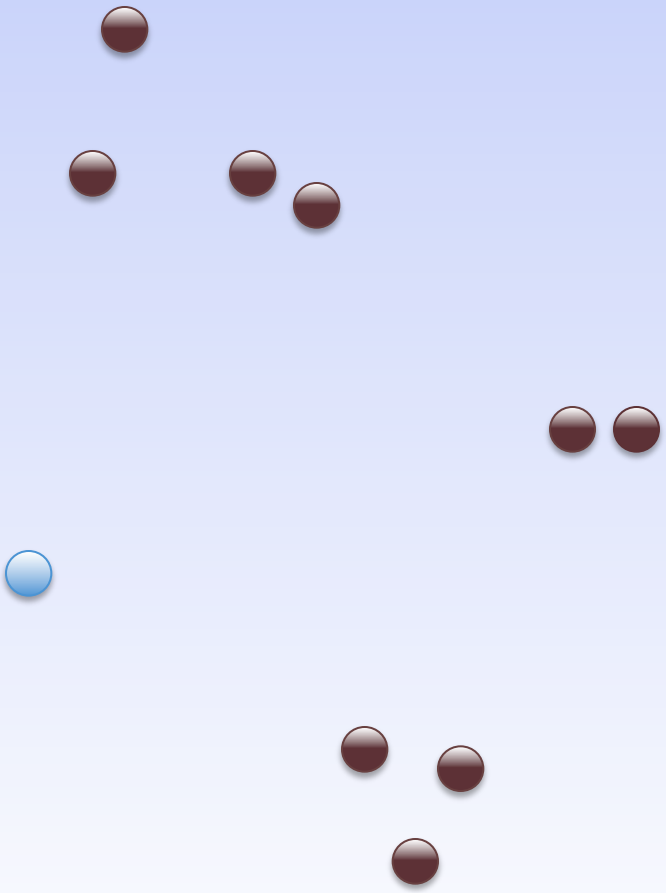
Tufts UNIVERSITY

Limited Normal Training Data

Bare minimum numbers of "normal" examples

Complex, heterogeneous high-dimensional domains

Add cross-validation:



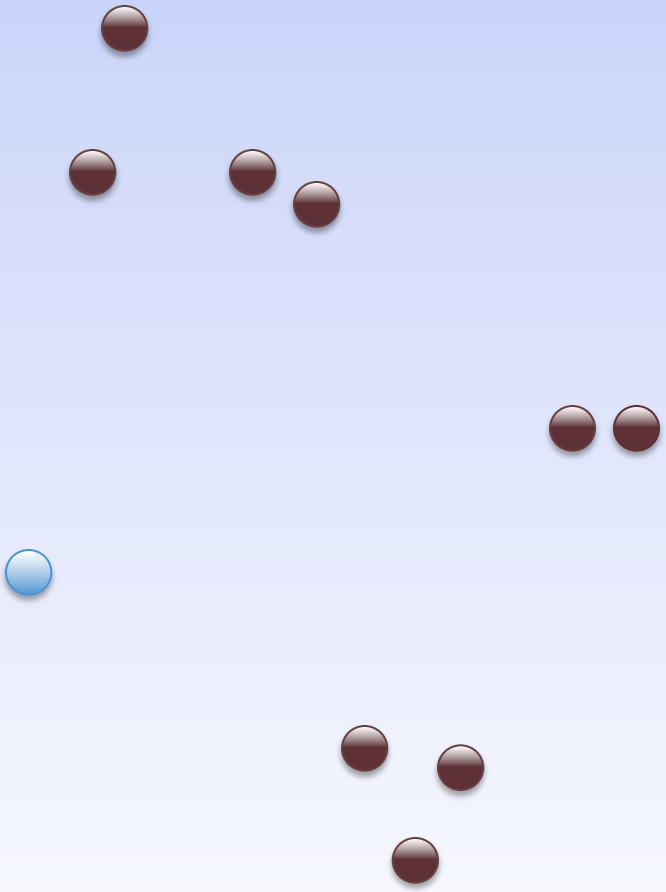
Tufts UNIVERSITY

Limited Normal Training Data

Bare minimum numbers of "normal" examples

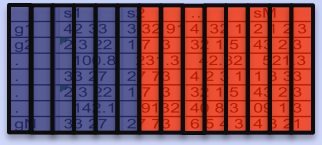
Complex, heterogeneous high-dimensional domains

Solution: bagging and weighting

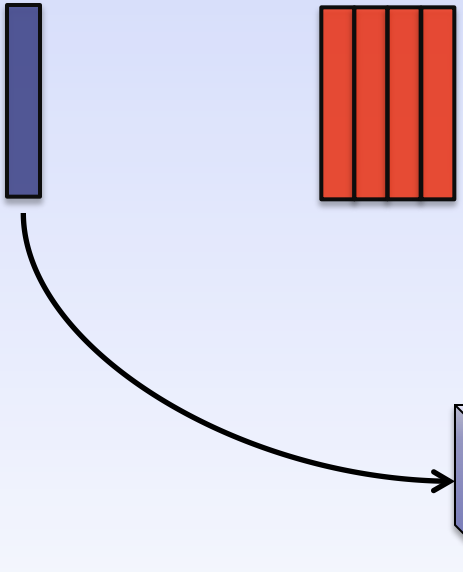


Tufts UNIVERSITY

CSAX: Characterizing Systematic Anomalies in eXpression Data

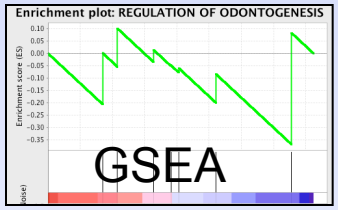


Random 50%



FRaC

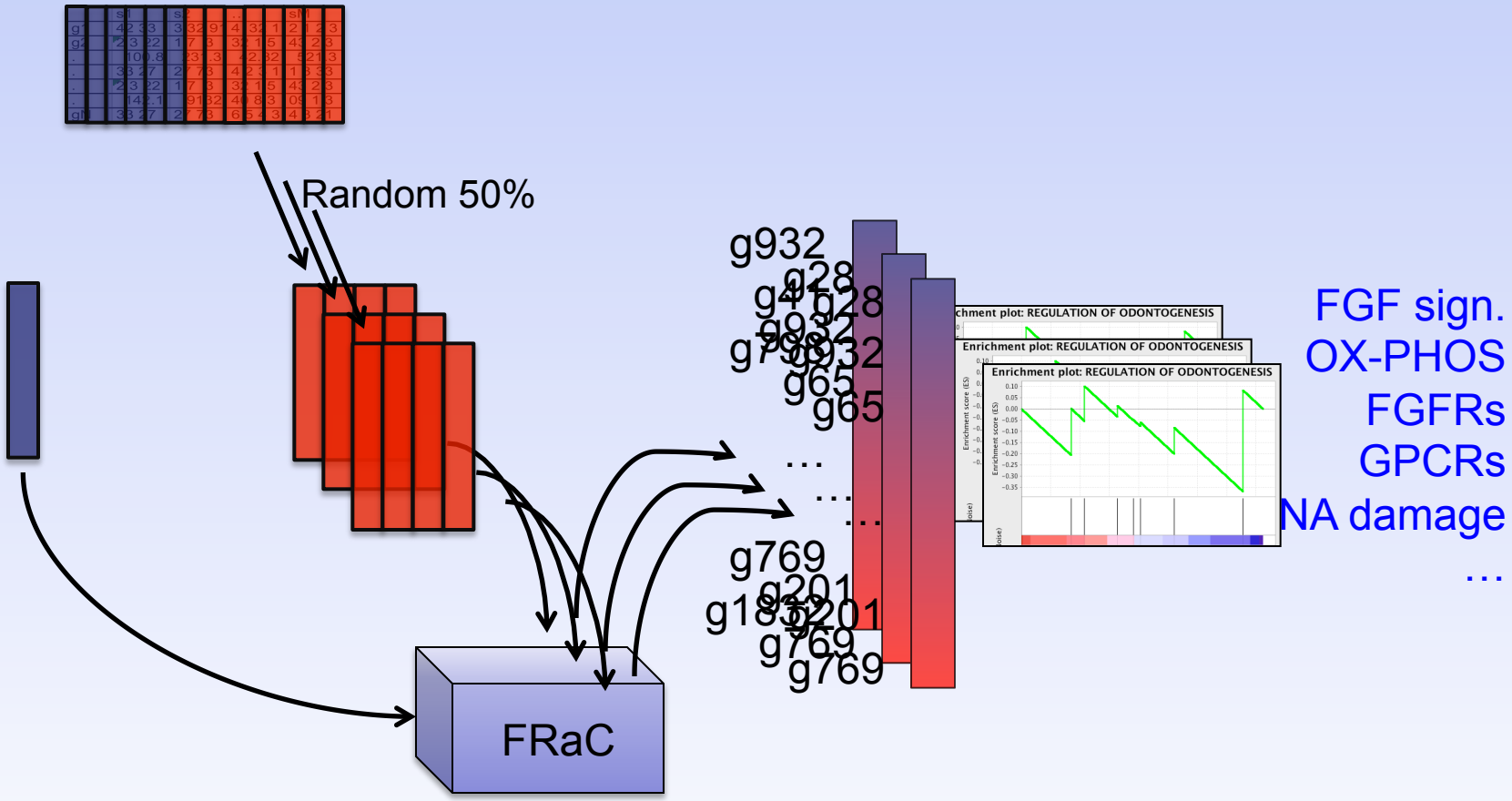
g932
g41
g798
…
g769
g1832



OX-PHOS
FGF sign.
GPCRs
MAPK path
DNA damage
…

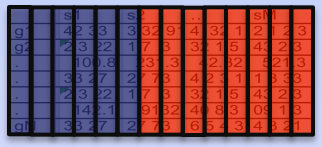
Noto, et al., *JCB 2015*

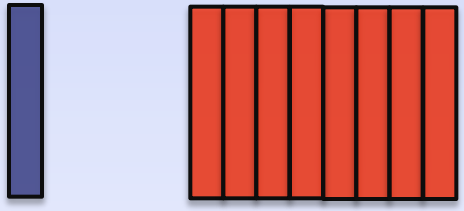
CSAX: Characterizing Systematic Anomalies in eXpression Data



Noto, et al., *JCB 2015*

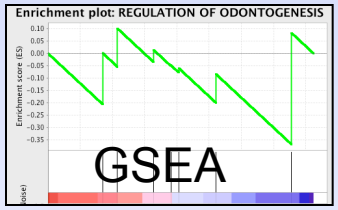
CSAX: Characterizing Systematic Anomalies in eXpression Data





FRaC

g932
g41
g798
…
g769
g1832

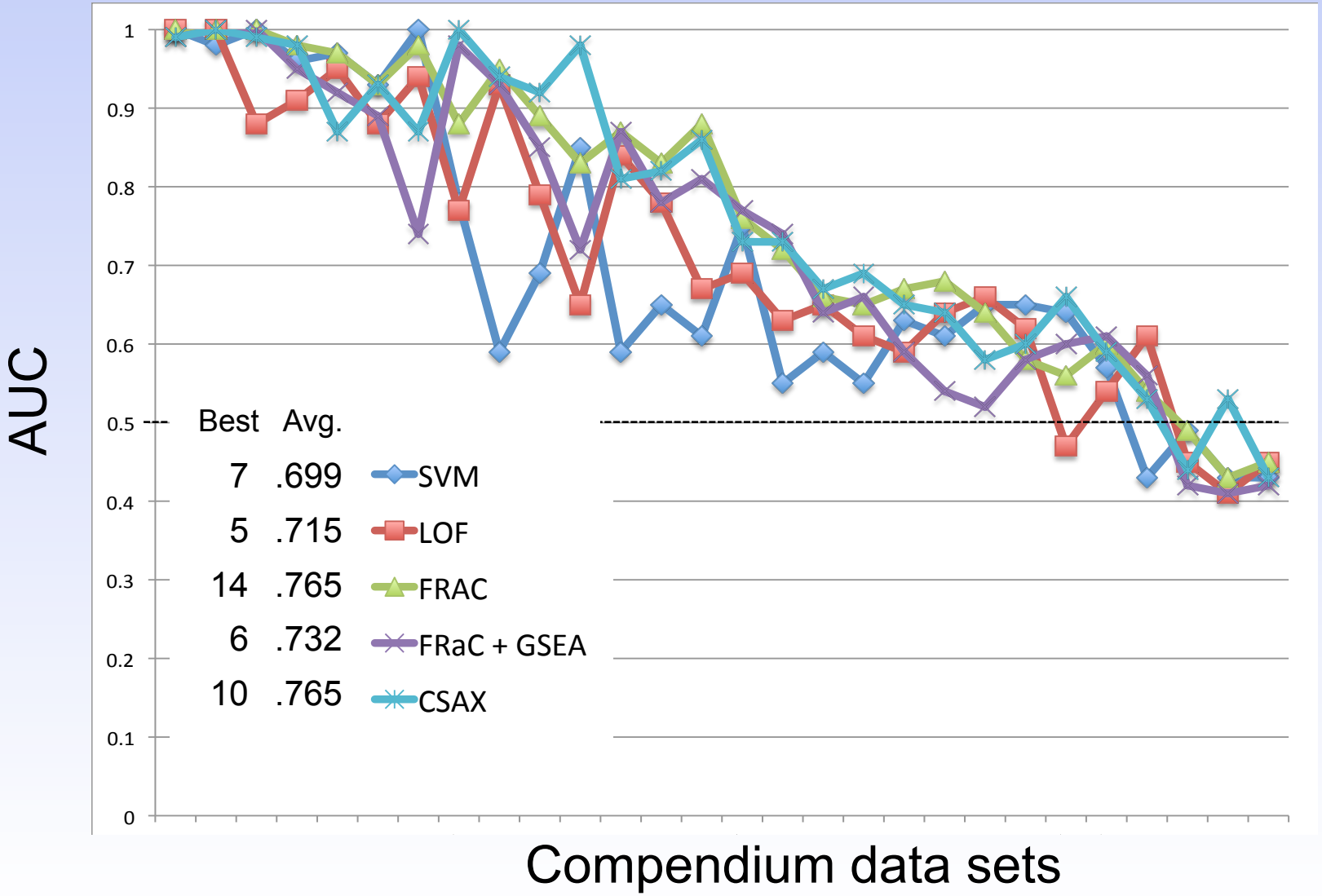


OX-PHOS
GPCRs
MAPK path
DNA damage
WNT sign.
…

Outlier score: combined enrichment scores, weighted by func. of ranks for gene sets

Noto, et al., *JCB 2015*

Outlier Detection for Expression Data



Example: impact of maternal obesity

Amniotic fluid from 17-week fetuses

Normal = mom has healthy BMI

Maternal obesity known to affect neuronal development

Top CSAX pathways (# individuals):

-axonogenesis (2)

-oxidative stress and inflammation (2)

-DNA damage response (1)

Tufts UNIVERSITY

Example: Blood from Preterm Infants

Blood from 100 infants born preterm

“Normal” = no complications during hospital stay

Complications include:

- BPD (pulmonary)
- ROP (vision)
- PVL (brain)

Pietrzky, et al., *PLoS One* 2013

Example: Blood from Preterm Infants

CSAX:

14 with periventricular leukomalacia (PVL)

calcium signaling / homeostasis (in 6 of them)

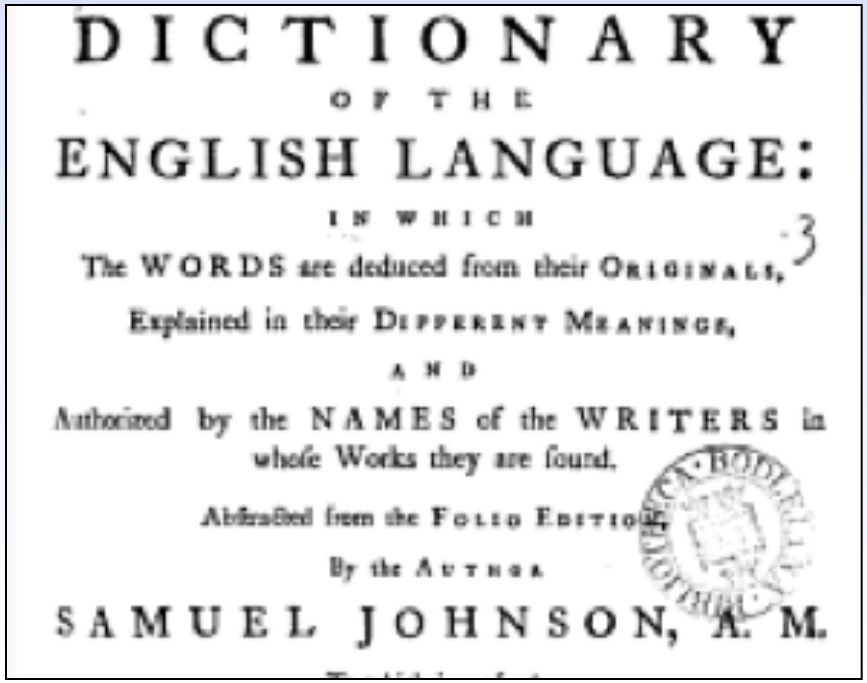
Tufts UNIVERSITY

Conclusions

Gene annotation focused on relevant context helps!

Using gene sets derived from this annotation, we can characterize individual anomalies.

Progress toward precision medicine: what does it mean when n=1?


DICTIONARY
OF THE
ENGLISH LANGUAGE:
IN WHICH
The WORDS are deduced from their ORIGINALS,
Explained in their DIFFERENT MEANINGS,
AND
Authorized by the NAMES of the WRITERS in whose Works they are found.
Abridged from the FOLIO EDITION
By the AUTHOR
SAMUEL JOHNSON, A. M.

Tufts UNIVERSITY

Acknowledgements

Slonim Lab
Jisoo Park
Michael Pietras
Heather Wick
Keith Noto
Saeed Majidi
Craig Fournier
Michael Sackman
Huy Ngu

Tufts CS
Carla Brodley

Women and Infants' Hospital:
Umadevi Tantravahi

Jackson Laboratory
Judy Blake
David Hill
Harold Drabkin

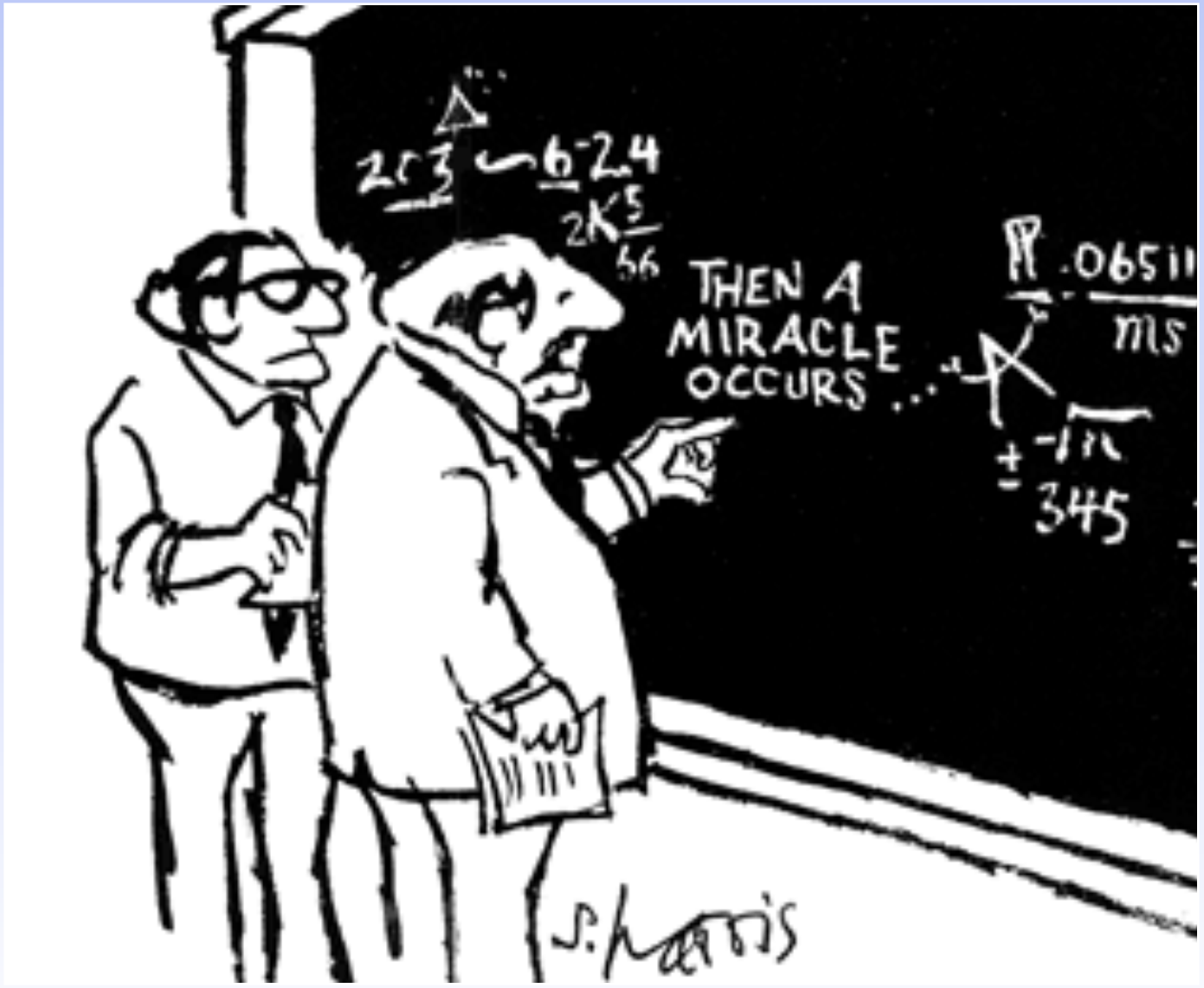
Tufts Medical Center
Diana Bianchi
Vidya Iyer
Jaclyn Ruggiero
Betsabee Castillo
Andrea Edlow
Jill Maron
Faycal Guedj
Janet Cowan
Zina Jarrah
Phil Hinds

Broad Institute
Jill Mesirov
Arthur Liberzon
Pablo Tamayo
George Steinhardt
Aravind Subramanian

EBI
Emily Dimmer
Rachael Huntley
Rolf Apweiler

Funding:
NIH R01HD058880
NIH R01HD076140

Tufts UNIVERSITY


THEN A
MIRACLE
OCCURS
.0651
ms
345
S. Harris