University of Crete

Open BioMedical Data for Integrative Analysis

Ioannis Tsamardinos

Associate Professor, Computer Science Department, University of Crete

Founder & CEO, Gnosis Data Analysis PC

Affiliated Faculty, ICS-FORTH

γ Gnosis Data Analysis

Data Repositories in Biomedicine

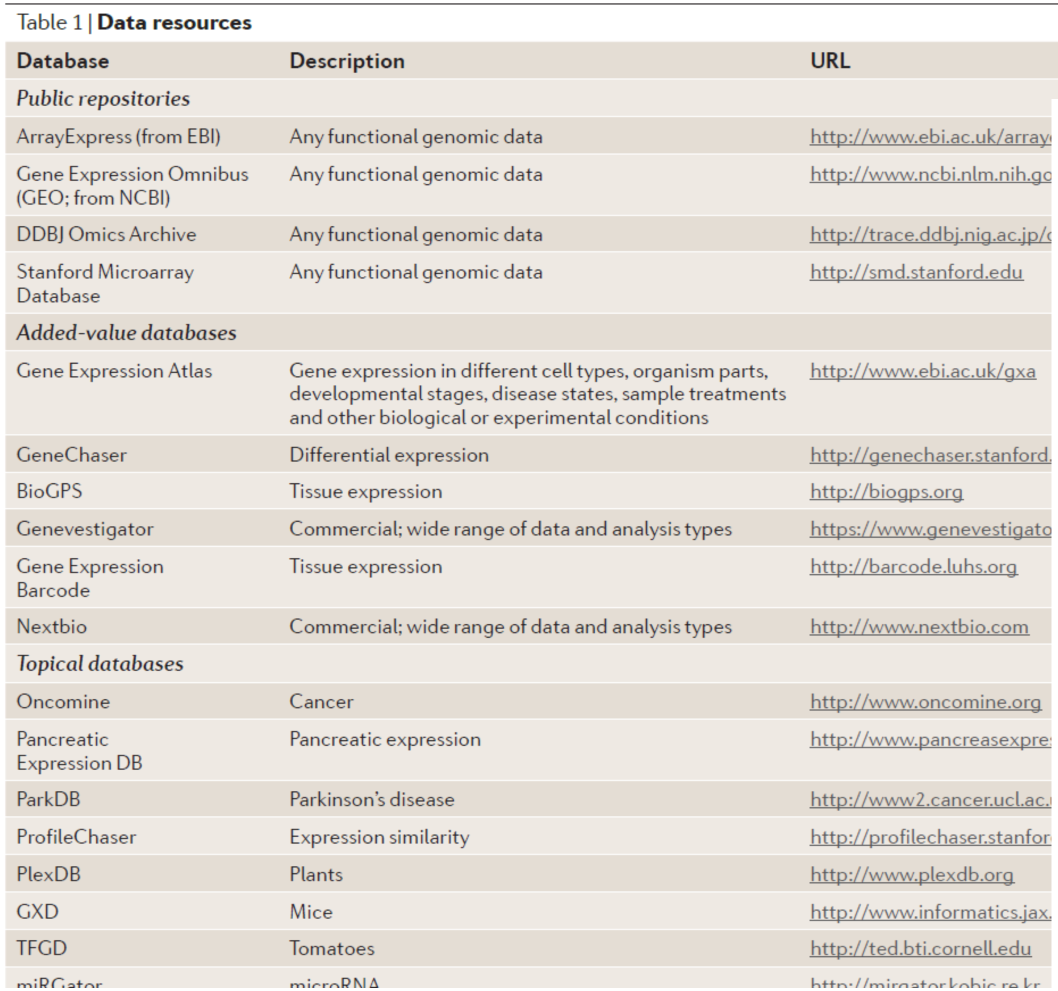


Table 1 | **Data resources**

Database	Description	URL	Refs
Public repositories			
ArrayExpress (from EBI)	Any functional genomic data	http://www.ebi.ac.uk/array	
Gene Expression Omnibus (GEO; from NCBI)	Any functional genomic data	http://www.ncbi.nlm.nih.go	
DDBJ Omics Archive	Any functional genomic data	http://trace.ddbj.nig.ac.jp/	
Stanford Microarray Database	Any functional genomic data	http://smd.stanford.edu	
Added-value databases			
Gene Expression Atlas	Gene expression in different cell types, organism parts, developmental stages, disease states, sample treatments and other biological or experimental conditions	http://www.ebi.ac.uk/gxa	
GeneChaser	Differential expression	http://genechaser.stanford.	
BioGPS	Tissue expression	http://biogps.org	
Genevestigator	Commercial; wide range of data and analysis types	https://www.genevestigato	
Gene Expression Barcode	Tissue expression	http://barcode.luhs.org	
Nextbio	Commercial; wide range of data and analysis types	http://www.nextbio.com	
Topical databases			
Oncomine	Cancer	http://www.oncomine.org	
Pancreatic Expression DB	Pancreatic expression	http://www.pancreasexpre	
ParkDB	Parkinson's disease	http://www2.cancer.ucl.ac.	
ProfileChaser	Expression similarity	http://profilechaser.stanfor	
PlexDB	Plants	http://www.plexdb.org	
GXD	Mice	http://www.informatics.jax.	
TFGD	Tomatoes	http://ted.bti.cornell.edu	
miRGator	microRNA	http://mirgator.kobic.re.kr	

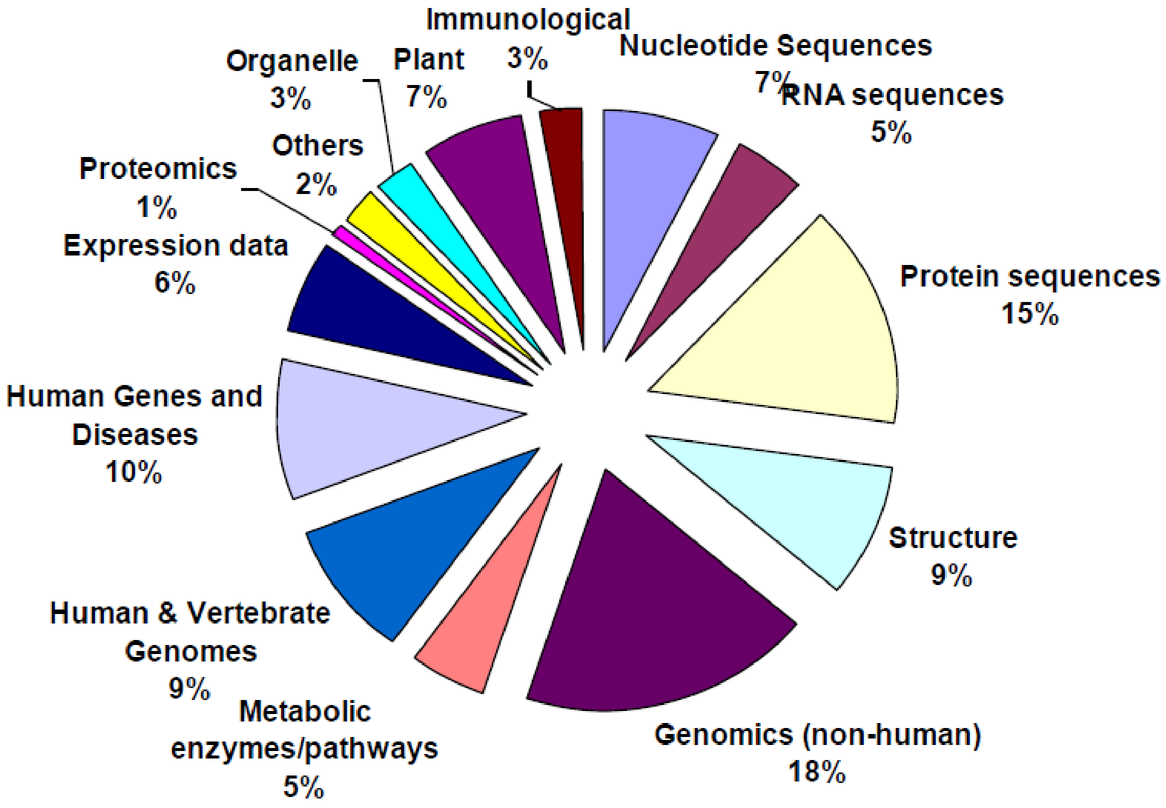


Figure 1. Subject matter of biological databases (2005)

Source: Elixir strategy for data resources report

Constructing a Statistical Model

Rule-Based Model (Decision Tree)

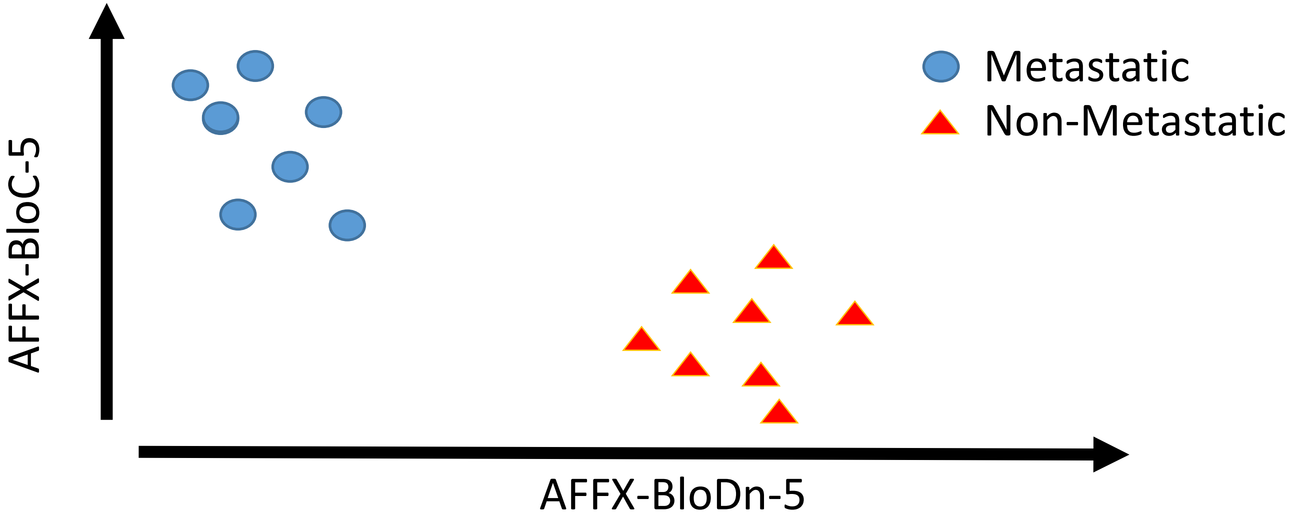
If AFFX-BloC-5 is Overexpressed and
AFFX-BloDn-5 is Underexpressed

Then
Classify as Metastatic

Else
Classify as Non-Metastatic

Linear Model:

Metastatic = sign (0,5 × AFFX-BloC-5 – 0,5 × AFFX-BloDn-5 + 3)



Expression Values

		Genes / Probe Sets							**Metastatic ?**
		AFFX-BloB-5_at	**AFFX-BloB-M_at**	**AFFX-Blob-3_at**	**AFFX-BloC-5_at**	**…**	**Affx-Bloc-3_at**	**AFFX-BloDn-5_at**	
Sample	1	123.00	1.00	2,3	12.00		23.00	34.00	Yes
	2	323.00	23.00	4,54	2.00		21.00	65.00	No
									No
									No
	N	232.00	4,5	23.00	0,55		75.00	343.00	Yes

Constructing a Statistical Model

Rule-Based Model (Decision Tree)

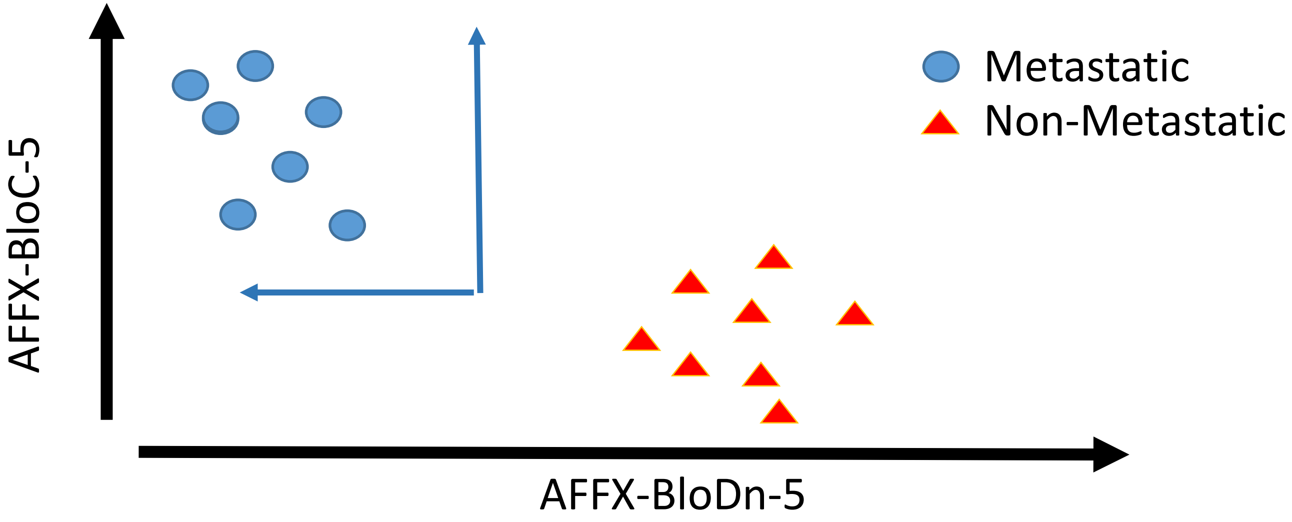
If AFFX-BloC-5 is Overexpressed and
AFFX-BloDn-5 is Underexpressed

Then
Classify as Metastatic

Else
Classify as Non-Metastatic

Linear Model:

Metastatic = sign (0,5 × AFFX-BloC-5 – 0,5 × AFFX-BloDn-5 + 3)



Expression Values

Genes / Probe Sets

		AFFX-BloB-5_at	**AFFX-BloB-M_at**	**AFFX-Blob-3_at**	**AFFX-BloC-5_at**	**...**	**Affx-Bloc-3_at**	**AFFX-BloDn-5_at**	**Metastatic ?**
Sample	1	123.00	1.00	2,3	12.00		23.00	34.00	Yes
	2	323.00	23.00	4,54	2.00		21.00	65.00	No
									No
									No
	N	232.00	4,5	23.00	0,55		75.00	343.00	Yes

Constructing a Statistical Model

Rule-Based Model (Decision Tree)

If AFFX-BloC-5 is Overexpressed and

AFFX-BloDn-5 is Underexpressed

Then

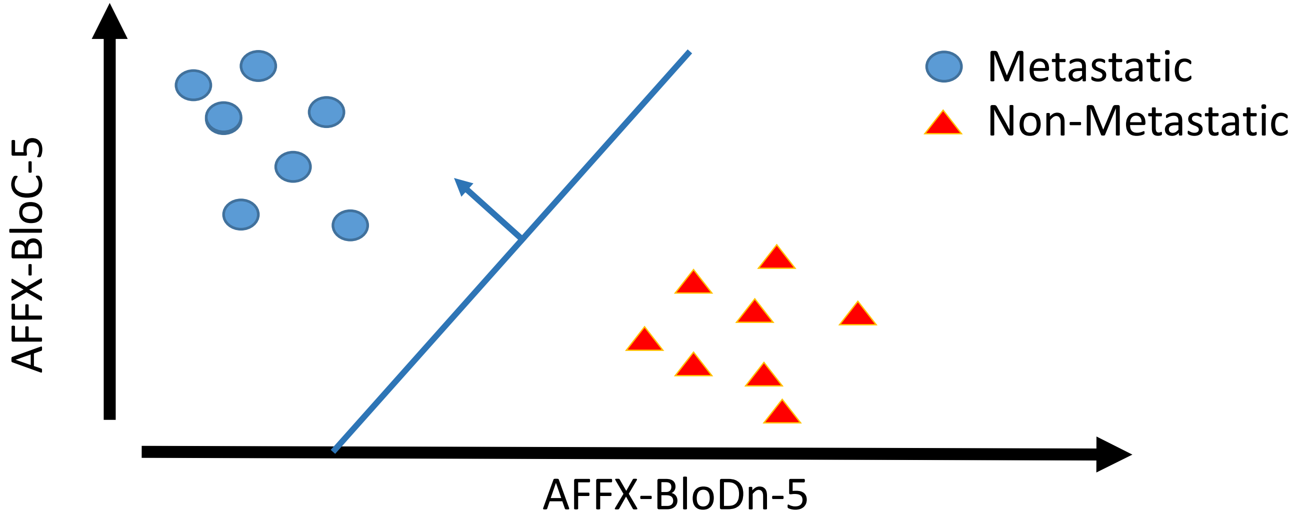
Classify as Metastatic

Else

Classify as Non-Metastatic

Linear Model:

Metastatic = sign (0,5 × AFFX-BloC-5 – 0,5 × AFFX-BloDn-5 + 3)

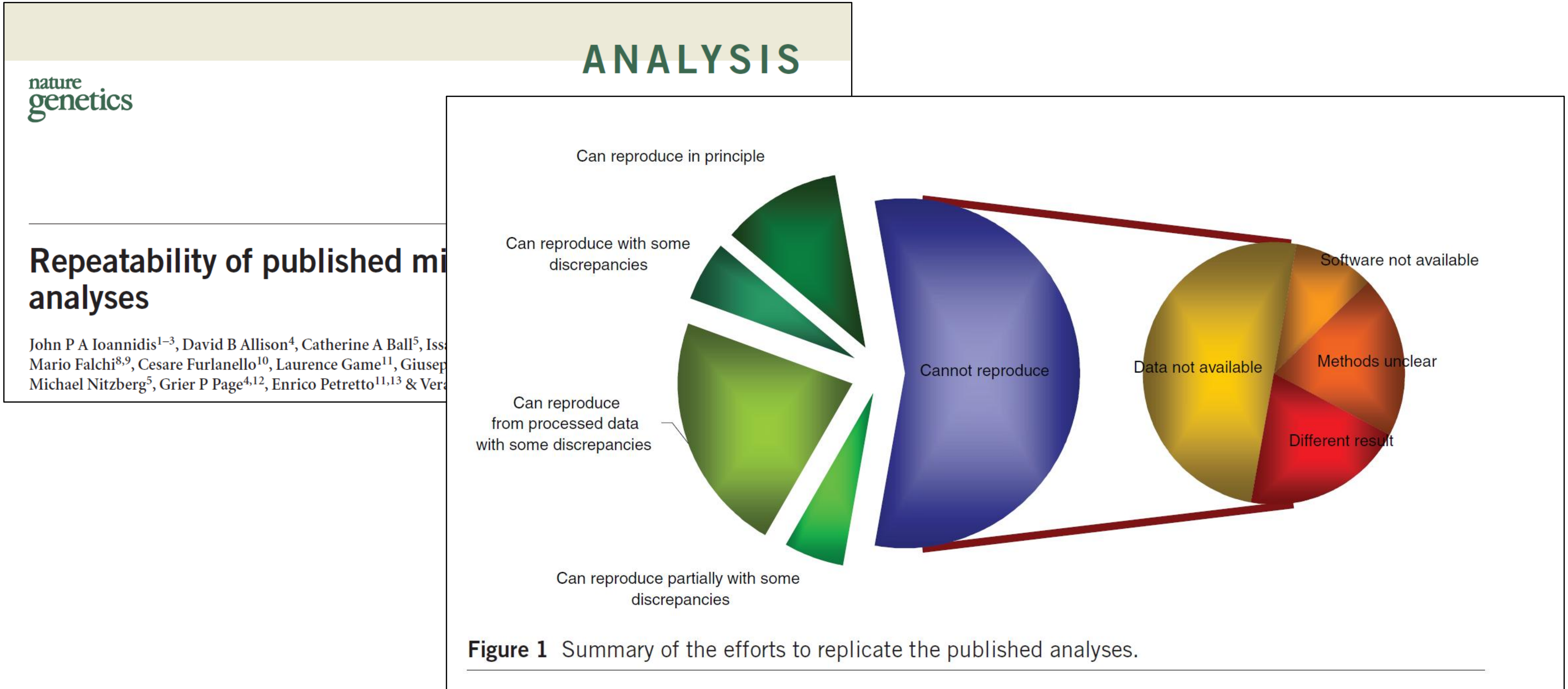


Expression Values

Genes / Probe Sets

		AFFX-BloB-5_at	**AFFX-BloB-M_at**	**AFFX-Blob-3_at**	**AFFX-BloC-5_at**	**...**	**Affx-Bloc-3_at**	**AFFX-BloDn-5_at**	**Metastatic ?**
Sample	1	123.00	1.00	2,3	12.00		23.00	34.00	Yes
	2	323.00	23.00	4,54	2.00		21.00	65.00	No
									No
									No
	N	232.00	4,5	23.00	0,55		75.00	343.00	Yes

Reproducibility of Statistical Analysis



ANALYSIS

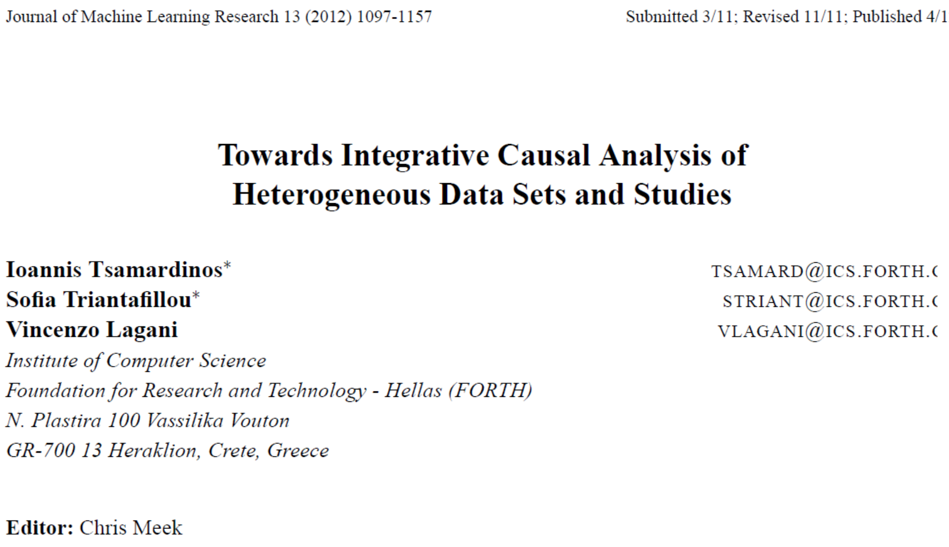
nature genetics

Repeatability of published mi analyses

John P A Ioannidis[1–3], David B Allison[4], Catherine A Ball[5], Iss
Mario Falchi[8,9], Cesare Furlanello[10], Laurence Game[11], Giusep
Michael Nitzberg[5], Grier P Page[4,12], Enrico Petretto[11,13] & Vera

Figure 1 Summary of the efforts to replicate the published analyses.

Development of Novel Methods

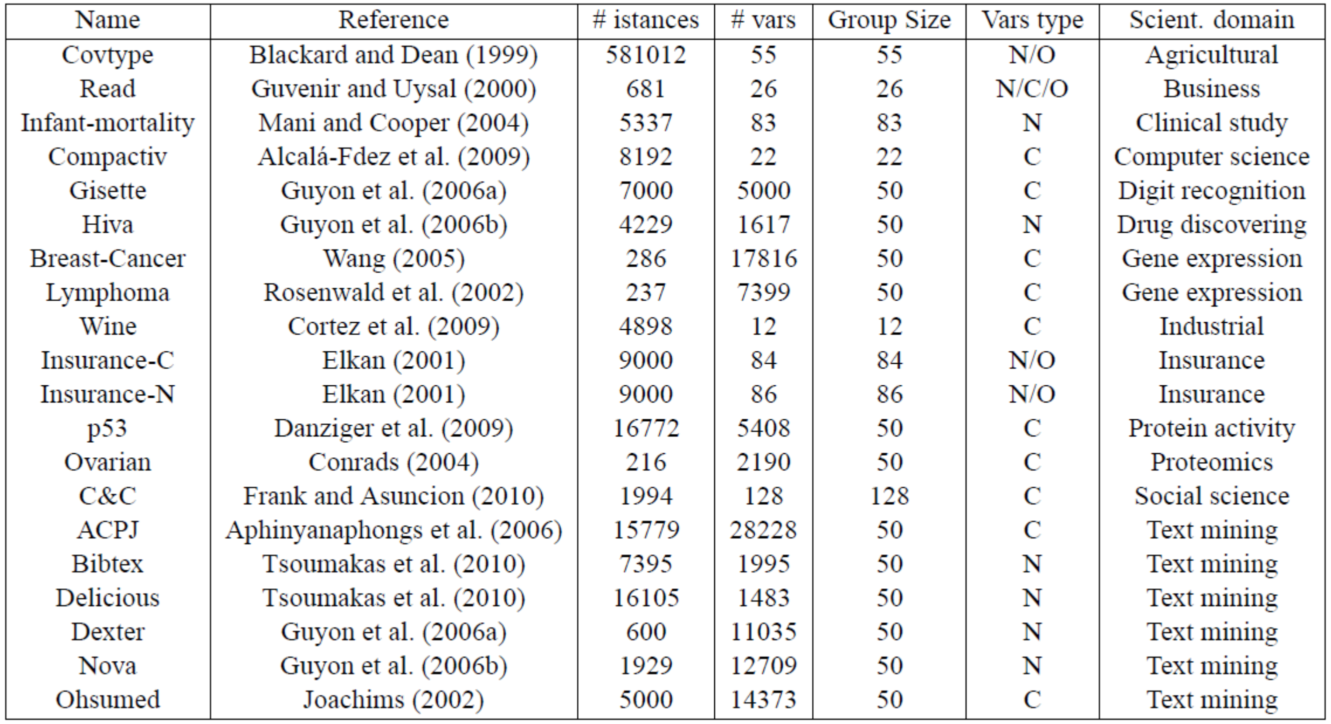


Journal of Machine Learning Research 13 (2012) 1097-1157 Submitted 3/11; Revised 11/11; Published 4/12

Towards Integrative Causal Analysis of Heterogeneous Data Sets and Studies

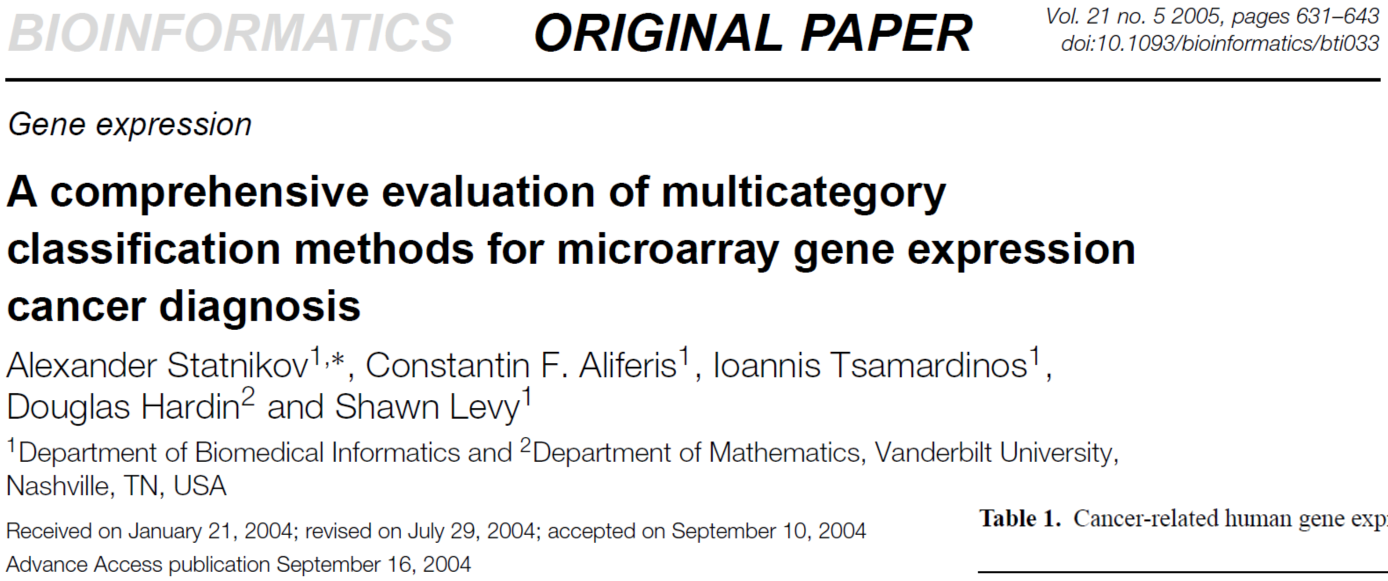
Ioannis Tsamardinos* TSAMARD@ICS.FORTH.(
Sofia Triantafillou* STRIANT@ICS.FORTH.(
Vincenzo Lagani VLAGANI@ICS.FORTH.(
Institute of Computer Science
Foundation for Research and Technology - Hellas (FORTH)
N. Plastira 100 Vassilika Vouton
GR-700 13 Heraklion, Crete, Greece

Editor: Chris Meek



Name	Reference	# istances	# vars	Group Size	Vars type	Scient. domain
Covtype	Blackard and Dean (1999)	581012	55	55	N/O	Agricultural
Read	Guvenir and Uysal (2000)	681	26	26	N/C/O	Business
Infant-mortality	Mani and Cooper (2004)	5337	83	83	N	Clinical study
Compactiv	Alcalá-Fdez et al. (2009)	8192	22	22	C	Computer science
Gisette	Guyon et al. (2006a)	7000	5000	50	C	Digit recognition
Hiva	Guyon et al. (2006b)	4229	1617	50	N	Drug discovering
Breast-Cancer	Wang (2005)	286	17816	50	C	Gene expression
Lymphoma	Rosenwald et al. (2002)	237	7399	50	C	Gene expression
Wine	Cortez et al. (2009)	4898	12	12	C	Industrial
Insurance-C	Elkan (2001)	9000	84	84	N/O	Insurance
Insurance-N	Elkan (2001)	9000	86	86	N/O	Insurance
p53	Danziger et al. (2009)	16772	5408	50	C	Protein activity
Ovarian	Conrads (2004)	216	2190	50	C	Proteomics
C&C	Frank and Asuncion (2010)	1994	128	128	C	Social science
ACPJ	Aphinyanaphongs et al. (2006)	15779	28228	50	C	Text mining
Bibtex	Tsoumakas et al. (2010)	7395	1995	50	N	Text mining
Delicious	Tsoumakas et al. (2010)	16105	1483	50	N	Text mining
Dexter	Guyon et al. (2006a)	600	11035	50	N	Text mining
Nova	Guyon et al. (2006b)	1929	12709	50	N	Text mining
Ohsumed	Joachims (2002)	5000	14373	50	C	Text mining

Table 1: Data Sets included in empirical evaluation of Section 6.3. N- Nominal, O - Ordinal, C - Continuous.



BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 5 2005, pages 631–643
doi:10.1093/bioinformatics/bti033

Gene expression

A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis

Alexander Statnikov[1,*], Constantin F. Aliferis[1], Ioannis Tsamardinos[1],
Douglas Hardin[2] and Shawn Levy[1]

[1]Department of Biomedical Informatics and [2]Department of Mathematics, Vanderbilt University,
Nashville, TN, USA

Received on January 21, 2004; revised on July 29, 2004; accepted on September 10, 2004
Advance Access publication September 16, 2004

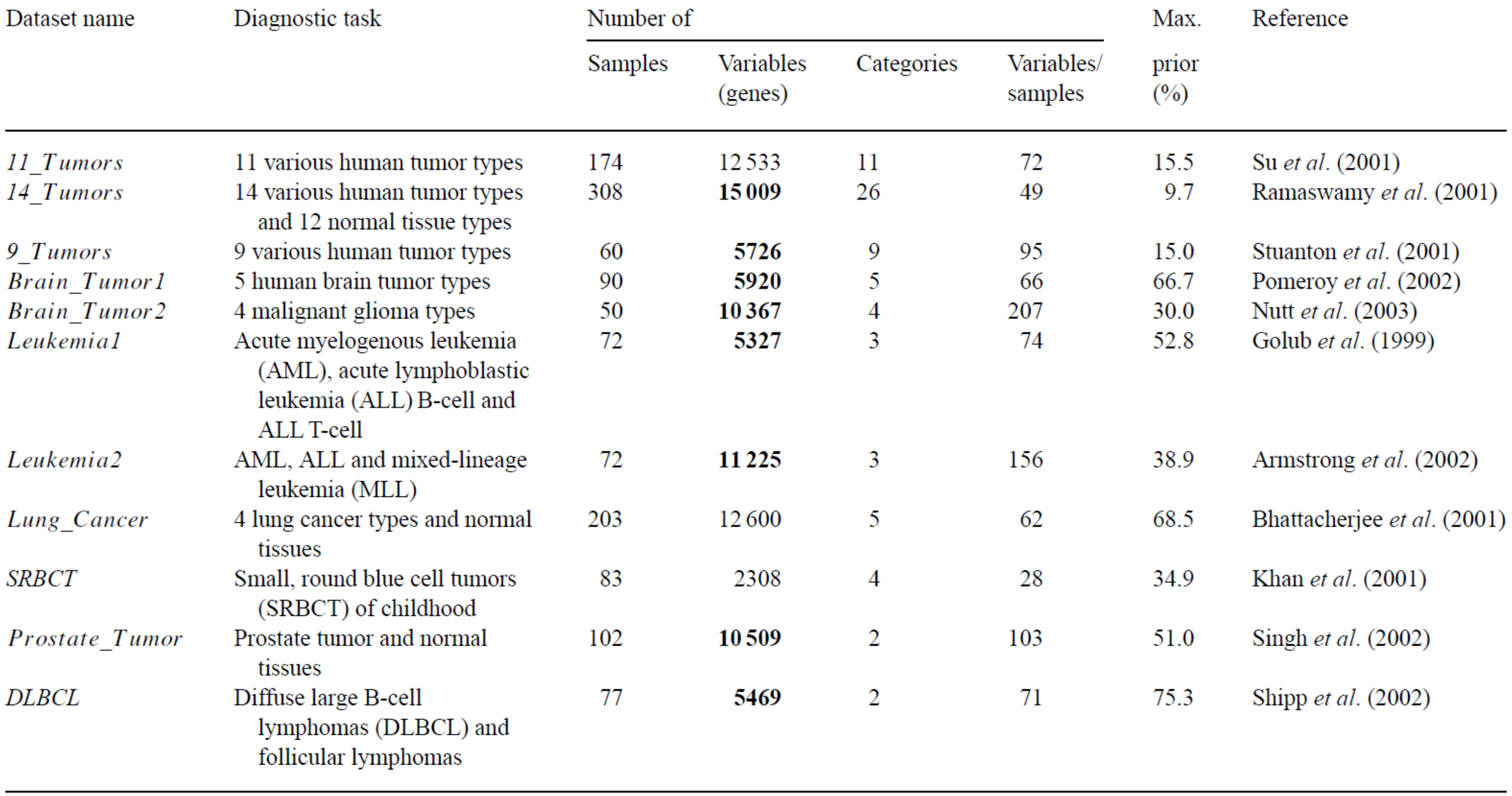
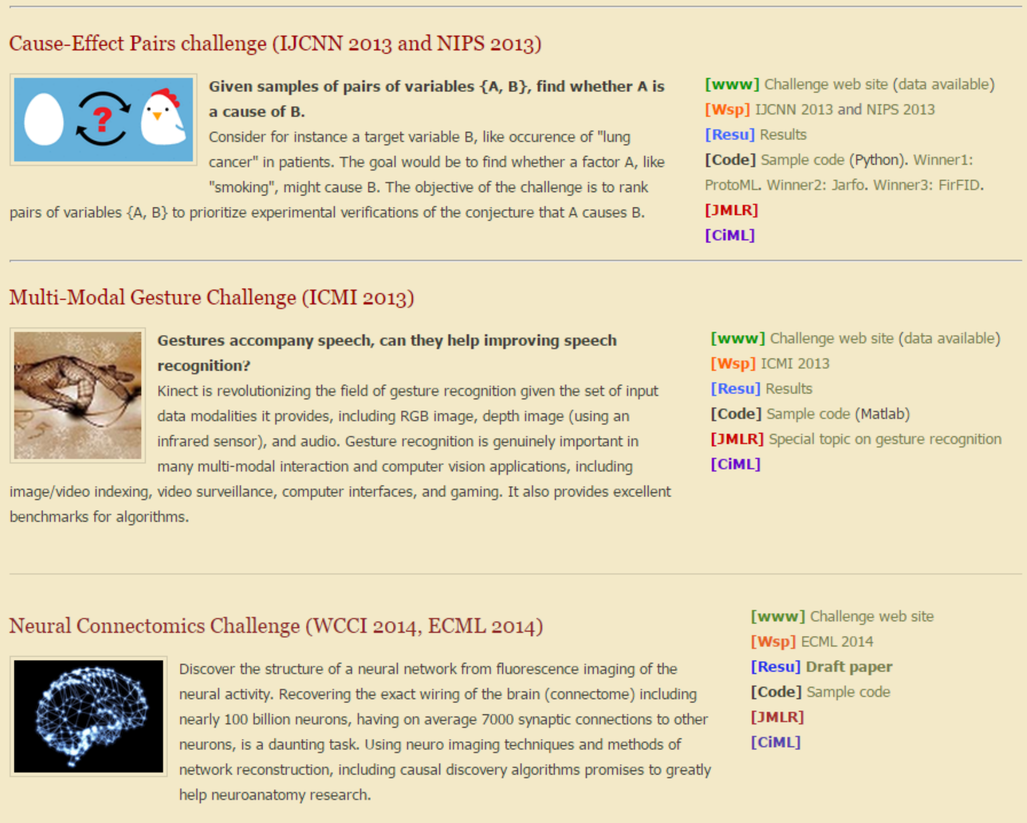


Table 1. Cancer-related human gene expression datasets used in this study

Dataset name	Diagnostic task	Number of				Max.	Reference
		Samples	Variables (genes)	Categories	Variables/ samples	prior (%)	
11_Tumors	11 various human tumor types	174	12 533	11	72	15.5	Su *et al*. (2001)
14_Tumors	14 various human tumor types and 12 normal tissue types	308	**15 009**	26	49	9.7	Ramaswamy *et al*. (2001)
9_Tumors	9 various human tumor types	60	**5726**	9	95	15.0	Stuanton *et al*. (2001)
Brain_Tumor1	5 human brain tumor types	90	**5920**	5	66	66.7	Pomeroy *et al*. (2002)
Brain_Tumor2	4 malignant glioma types	50	**10 367**	4	207	30.0	Nutt *et al*. (2003)
Leukemia1	Acute myelogenous leukemia (AML), acute lymphoblastic leukemia (ALL) B-cell and ALL T-cell	72	**5327**	3	74	52.8	Golub *et al*. (1999)
Leukemia2	AML, ALL and mixed-lineage leukemia (MLL)	72	**11 225**	3	156	38.9	Armstrong *et al*. (2002)
Lung_Cancer	4 lung cancer types and normal tissues	203	12 600	5	62	68.5	Bhattacherjee *et al*. (2001)
SRBCT	Small, round blue cell tumors (SRBCT) of childhood	83	2308	4	28	34.9	Khan *et al*. (2001)
Prostate_Tumor	Prostate tumor and normal tissues	102	**10 509**	2	103	51.0	Singh *et al*. (2002)
DLBCL	Diffuse large B-cell lymphomas (DLBCL) and follicular lymphomas	77	**5469**	2	71	75.3	Shipp *et al*. (2002)

Challenges and Competitions



Cause-Effect Pairs challenge (IJCNN 2013 and NIPS 2013)

Given samples of pairs of variables {A, B}, find whether A is a cause of B.

Consider for instance a target variable B, like occurence of "lung cancer" in patients. The goal would be to find whether a factor A, like "smoking", might cause B. The objective of the challenge is to rank pairs of variables {A, B} to prioritize experimental verifications of the conjecture that A causes B.

[www] Challenge web site (data available)
[Wsp] IJCNN 2013 and NIPS 2013
[Resu] Results
[Code] Sample code (Python). Winner1: ProtoML. Winner2: Jarfo. Winner3: FirFID.
[JMLR]
[CiML]

Multi-Modal Gesture Challenge (ICMI 2013)

Gestures accompany speech, can they help improving speech recognition?

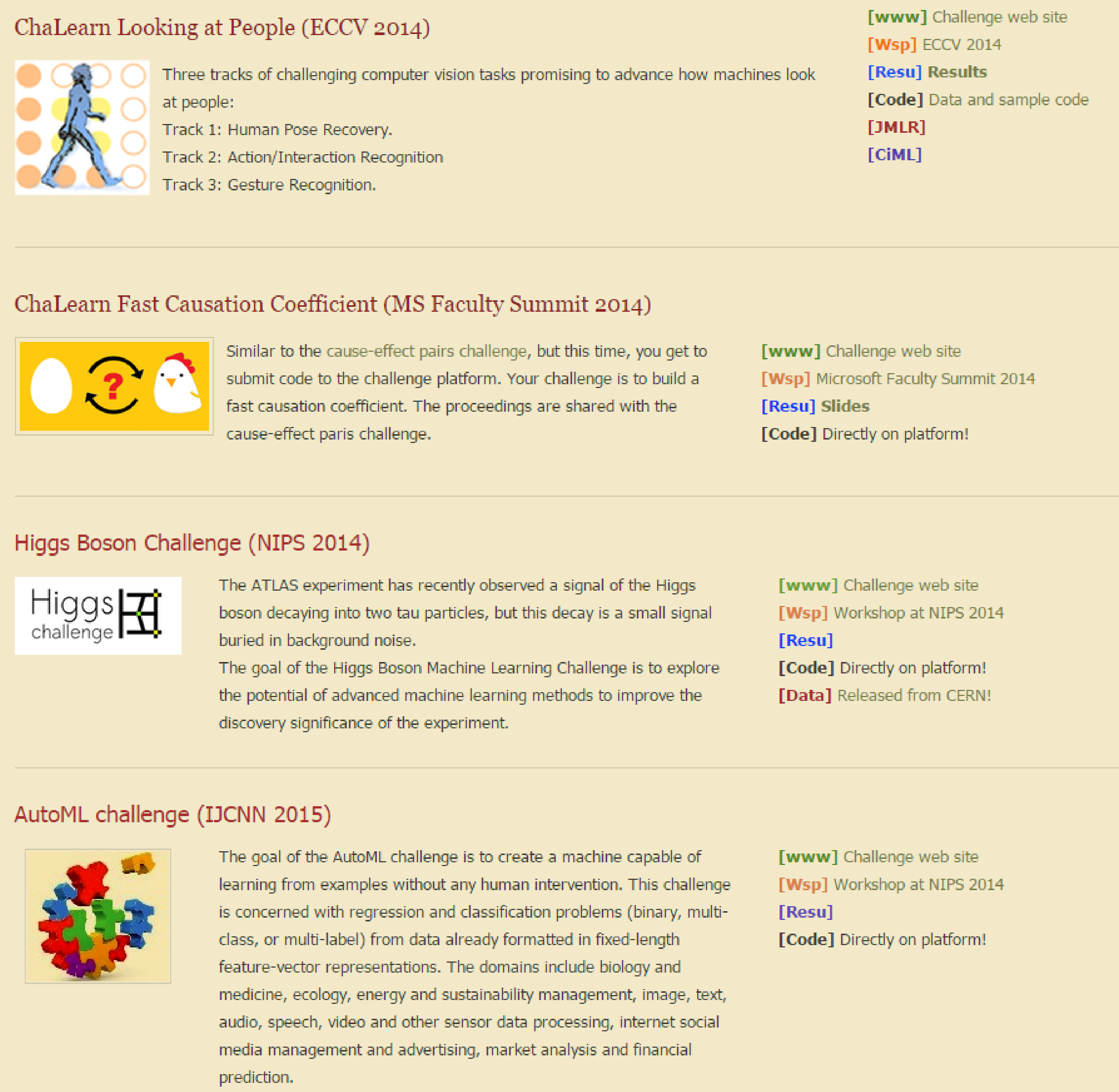
Kinect is revolutionizing the field of gesture recognition given the set of input data modalities it provides, including RGB image, depth image (using an infrared sensor), and audio. Gesture recognition is genuinely important in many multi-modal interaction and computer vision applications, including image/video indexing, video surveillance, computer interfaces, and gaming. It also provides excellent benchmarks for algorithms.

[www] Challenge web site (data available)
[Wsp] ICMI 2013
[Resu] Results
[Code] Sample code (Matlab)
[JMLR] Special topic on gesture recognition
[CiML]

Neural Connectomics Challenge (WCCI 2014, ECML 2014)

Discover the structure of a neural network from fluorescence imaging of the neural activity. Recovering the exact wiring of the brain (connectome) including nearly 100 billion neurons, having on average 7000 synaptic connections to other neurons, is a daunting task. Using neuro imaging techniques and methods of network reconstruction, including causal discovery algorithms promises to greatly help neuroanatomy research.

[www] Challenge web site
[Wsp] ECML 2014
[Resu] **Draft paper**
[Code] Sample code
[JMLR]
[CiML]



ChaLearn Looking at People (ECCV 2014)

Three tracks of challenging computer vision tasks promising to advance how machines look at people:
Track 1: Human Pose Recovery.
Track 2: Action/Interaction Recognition
Track 3: Gesture Recognition.

[www] Challenge web site
[Wsp] ECCV 2014
[Resu] **Results**
[Code] Data and sample code
[JMLR]
[CiML]

ChaLearn Fast Causation Coefficient (MS Faculty Summit 2014)

Similar to the cause-effect pairs challenge, but this time, you get to submit code to the challenge platform. Your challenge is to build a fast causation coefficient. The proceedings are shared with the cause-effect paris challenge.

[www] Challenge web site
[Wsp] Microsoft Faculty Summit 2014
[Resu] **Slides**
[Code] Directly on platform!

Higgs Boson Challenge (NIPS 2014)

The ATLAS experiment has recently observed a signal of the Higgs boson decaying into two tau particles, but this decay is a small signal buried in background noise.

The goal of the Higgs Boson Machine Learning Challenge is to explore the potential of advanced machine learning methods to improve the discovery significance of the experiment.

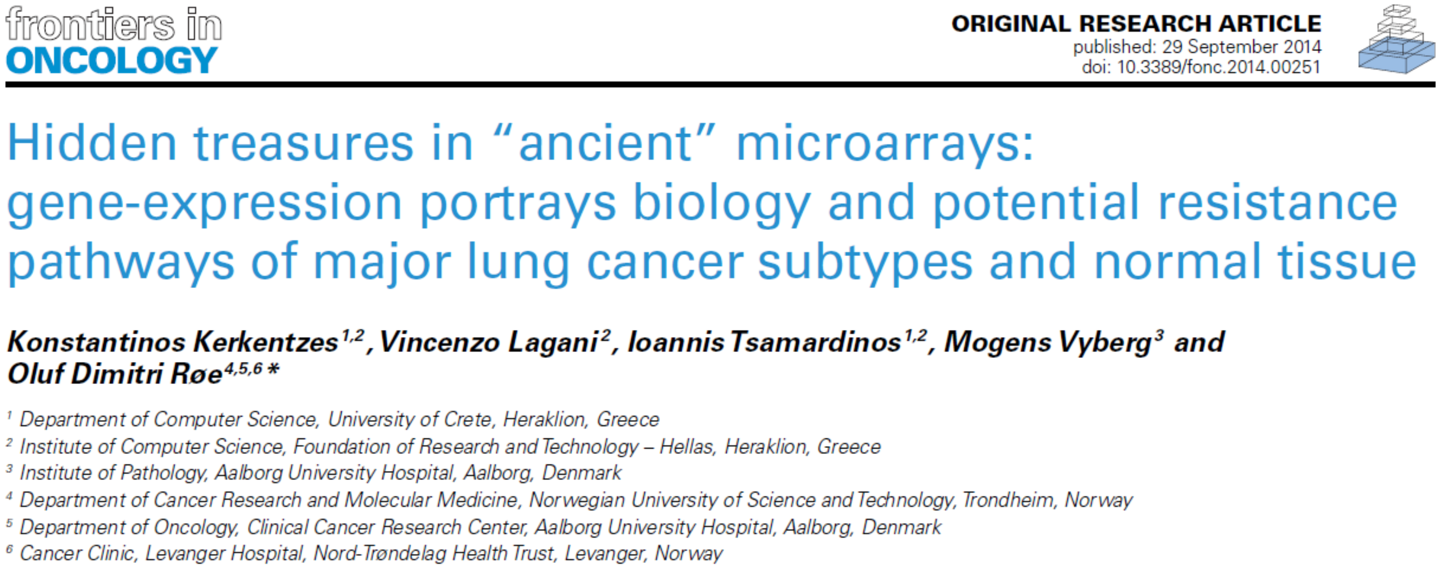
[www] Challenge web site
[Wsp] Workshop at NIPS 2014
[Resu]
[Code] Directly on platform!
[Data] Released from CERN!

AutoML challenge (IJCNN 2015)

The goal of the AutoML challenge is to create a machine capable of learning from examples without any human intervention. This challenge is concerned with regression and classification problems (binary, multi-class, or multi-label) from data already formatted in fixed-length feature-vector representations. The domains include biology and medicine, ecology, energy and sustainability management, image, text, audio, speech, video and other sensor data processing, internet social media management and advertising, market analysis and financial prediction.

[www] Challenge web site
[Wsp] Workshop at NIPS 2014
[Resu]
[Code] Directly on platform!

Re-Using and Revisiting



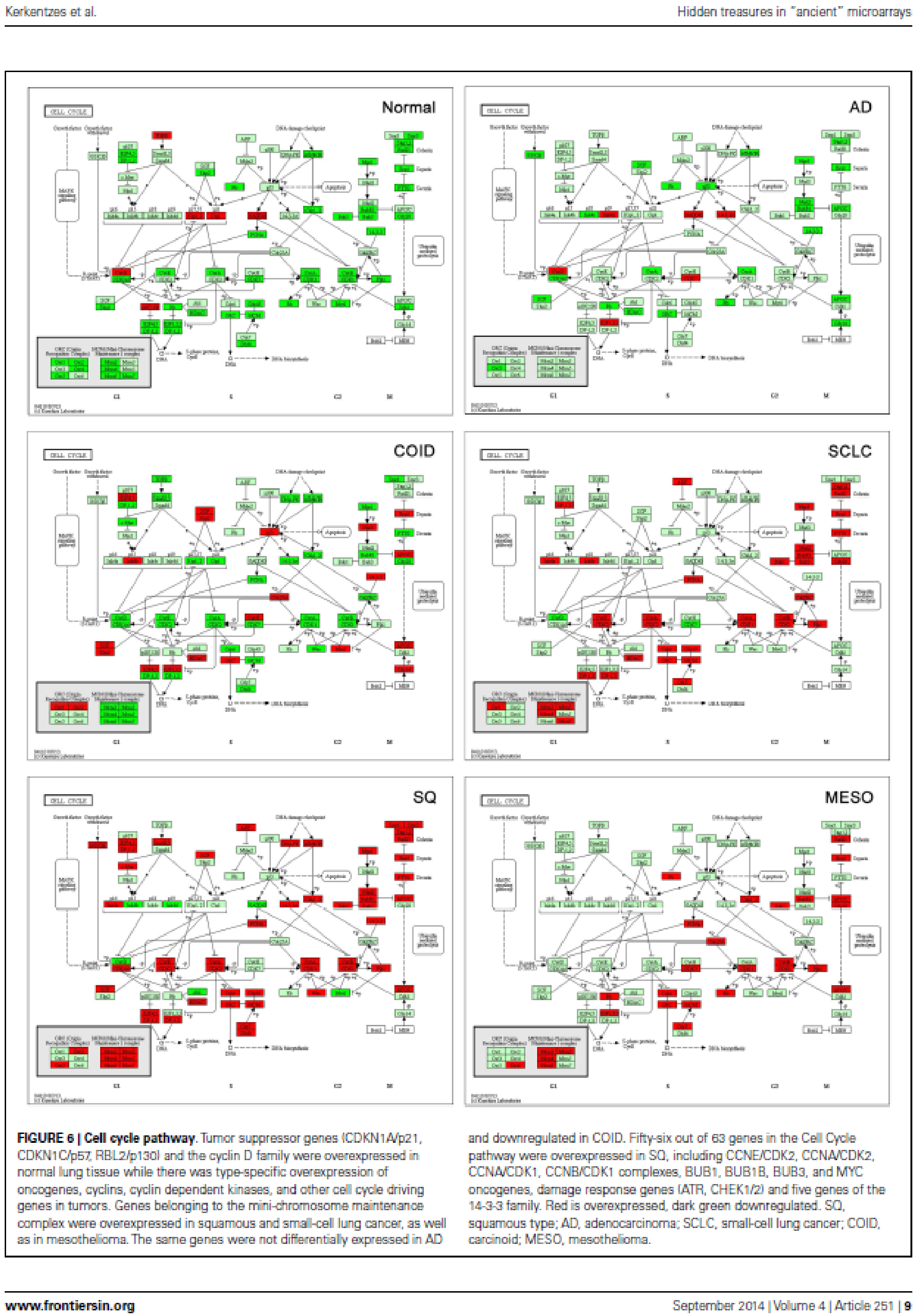
frontiers in
ONCOLOGY

ORIGINAL RESEARCH ARTICLE
published: 29 September 2014
doi: 10.3389/fonc.2014.00251

Hidden treasures in "ancient" microarrays: gene-expression portrays biology and potential resistance pathways of major lung cancer subtypes and normal tissue

Konstantinos Kerkentzes*[1,2], *Vincenzo Lagani*[2], *Ioannis Tsamardinos*[1,2], *Mogens Vyberg*[3] and *Oluf Dimitri Røe*[4,5,6]

[1] *Department of Computer Science, University of Crete, Heraklion, Greece*
[2] *Institute of Computer Science, Foundation of Research and Technology – Hellas, Heraklion, Greece*
[3] *Institute of Pathology, Aalborg University Hospital, Aalborg, Denmark*
[4] *Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway*
[5] *Department of Oncology, Clinical Cancer Research Center, Aalborg University Hospital, Aalborg, Denmark*
[6] *Cancer Clinic, Levanger Hospital, Nord-Trøndelag Health Trust, Levanger, Norway*



Kerkentzes et al. Hidden treasures in "ancient" microarrays

FIGURE 6 | Cell cycle pathway. Tumor suppressor genes (CDKN1A/p21, CDKN1C/p57, RBL2/p130) and the cyclin D family were overexpressed in normal lung tissue while there was type-specific overexpression of oncogenes, cyclins, cyclin dependent kinases, and other cell cycle driving genes in tumors. Genes belonging to the mini-chromosome maintenance complex were overexpressed in squamous and small-cell lung cancer, as well as in mesothelioma. The same genes were not differentially expressed in AD and downregulated in COID. Fifty-six out of 63 genes in the Cell Cycle pathway were overexpressed in SQ, including CCNE/CDK2, CCNA/CDK2, CCNA/CDK1, CCNB/CDK1 complexes, BUB1, BUB1B, BUB3, and MYC oncogenes, damage response genes (ATR, CHEK1/2) and five genes of the 14-3-3 family. Red is overexpressed, dark green downregulated. SQ, squamous type; AD, adenocarcinoma; SCLC, small-cell lung cancer; COID, carcinoid; MESO, mesothelioma.

www.frontiersin.org September 2014 | Volume 4 | Article 251 | 9

Meta-Analysis

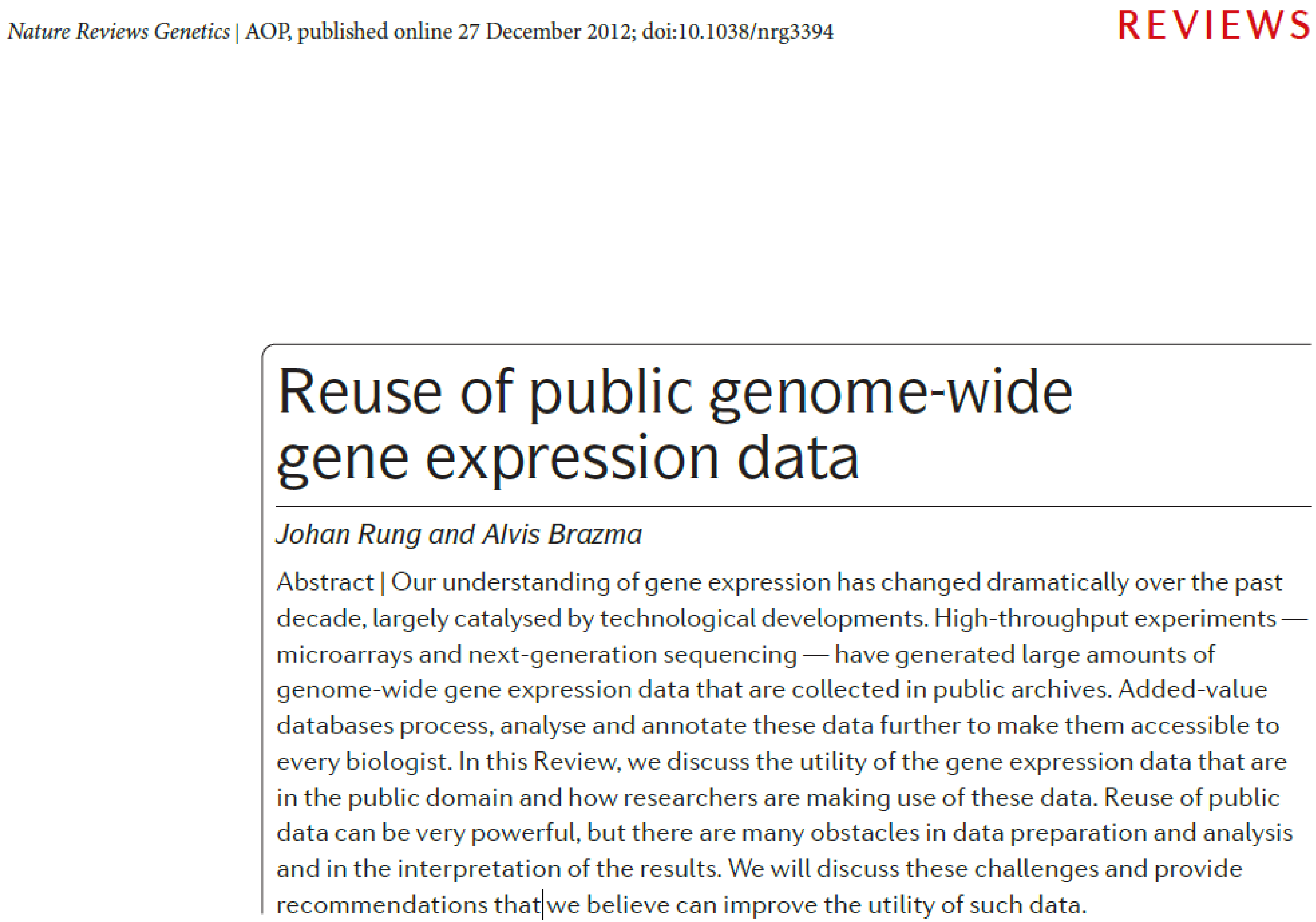
- Does psychotherapy reduce depression?
- 375 studies were included
- 2 years of collecting the **papers (!)**

- Combine p-values, regression coefficients, statistics found in the **papers (not the raw data)**

Gene V Glass

Meta-Analysis on the Raw Data

- One research hypothesis for each gene!
- Raw data allow more sophisticated statistical methods
  - Batch-Effect Removal



Nature Reviews Genetics | AOP, published online 27 December 2012; doi:10.1038/nrg3394

REVIEWS

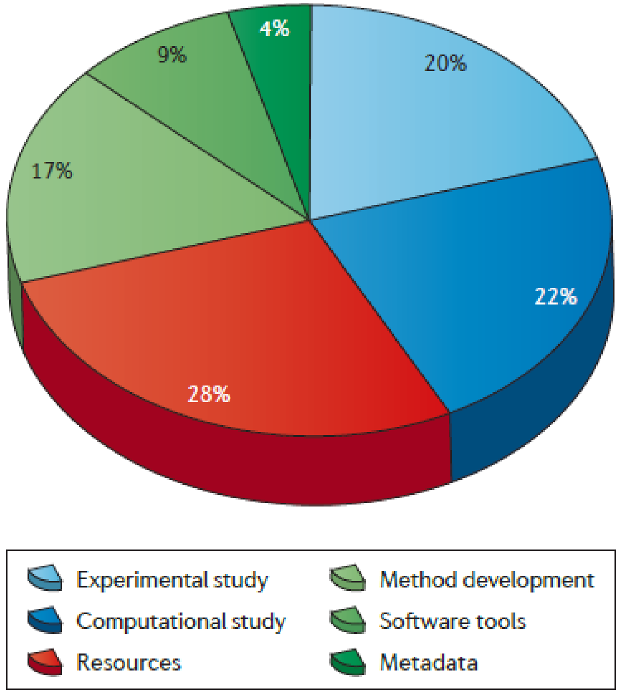
Reuse of public genome-wide gene expression data

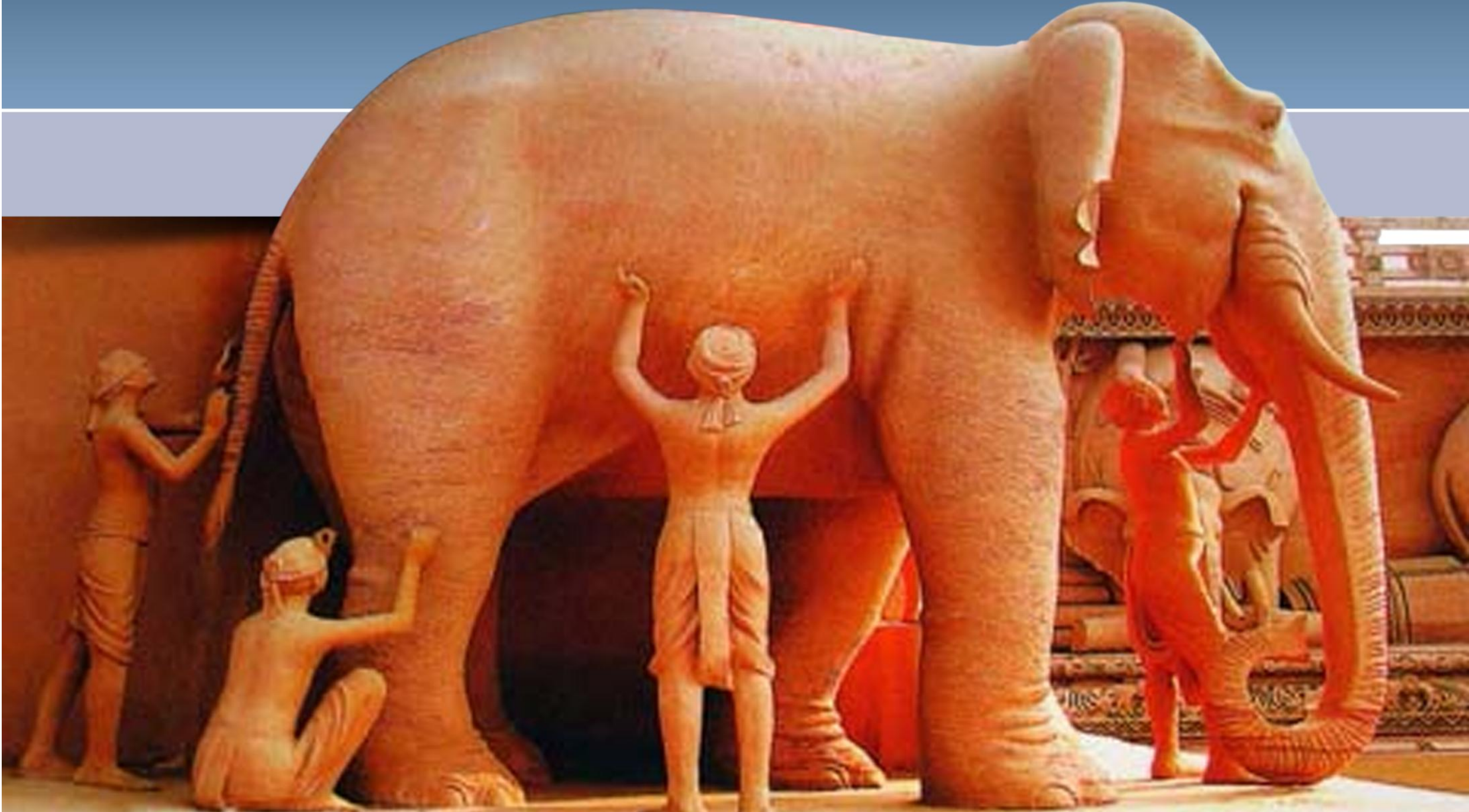
Johan Rung and Alvis Brazma

Abstract | Our understanding of gene expression has changed dramatically over the past decade, largely catalysed by technological developments. High-throughput experiments — microarrays and next-generation sequencing — have generated large amounts of genome-wide gene expression data that are collected in public archives. Added-value databases process, analyse and annotate these data further to make them accessible to every biologist. In this Review, we discuss the utility of the gene expression data that are in the public domain and how researchers are making use of these data. Reuse of public data can be very powerful, but there are many obstacles in data preparation and analysis and in the interpretation of the results. We will discuss these challenges and provide recommendations that we believe can improve the utility of such data.

Type	Use	Count
Biological	Experimental study	18
Biological	Computational study	20
Technical	Method development	15
Technical	Software tools	8
Technical	Metadata	4
Databases	Resources	25
Other	Reviews	20
Other	Data submission	8
Other	Other context	10

'Other' not included in graph because these do not reanalyse or process data

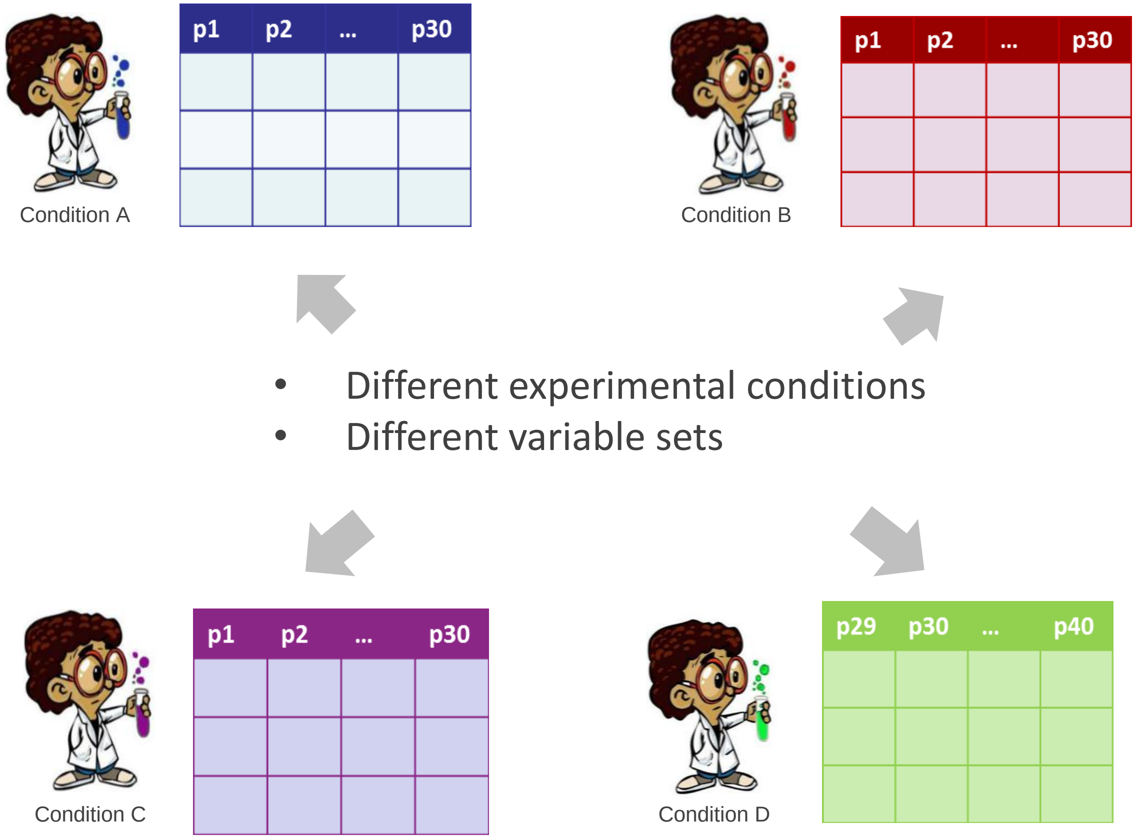




Motivation: Heterogeneous data sets measuring the same system under study

Variables Study	**Thrombosis** (Yes/No)	**Contraceptives** (Yes/No)	**Protein C** (numerical)	**Cancer** (Yes/No)	**Protein Y** (numerical)	**Protein Z** (numerical)
1 **observational data**	Yes	No	10.5	Yes	-	-
	No	Yes	5.3	No	-	-
					-	-
	No	Yes	0.01	No	-	-
2 **experimental data**	No	No	0 (Control)	No	-	-
	Yes	No	0 (Control)	Yes	-	-
					-	-
	Yes	Yes	5.0 (Treat.)	Yes	-	-
	No	Yes	5.0 (Treat.)	No	-	-
3 **different variables**	-	-	-	Yes	0.03	9.3
	-	-	-			
	-	-	-	No	3.4	22.2
4 **prior knowledge**	Use of contraceptives cause thrombosis: Contraceptives→→ Thrombosis					

Co-analyzing data sets from different experimental conditions with overlapping variable sets



p1	p2	…	p30

p1	p2	…	p30

p1	p2	…	p30

p29	p30	…	p40

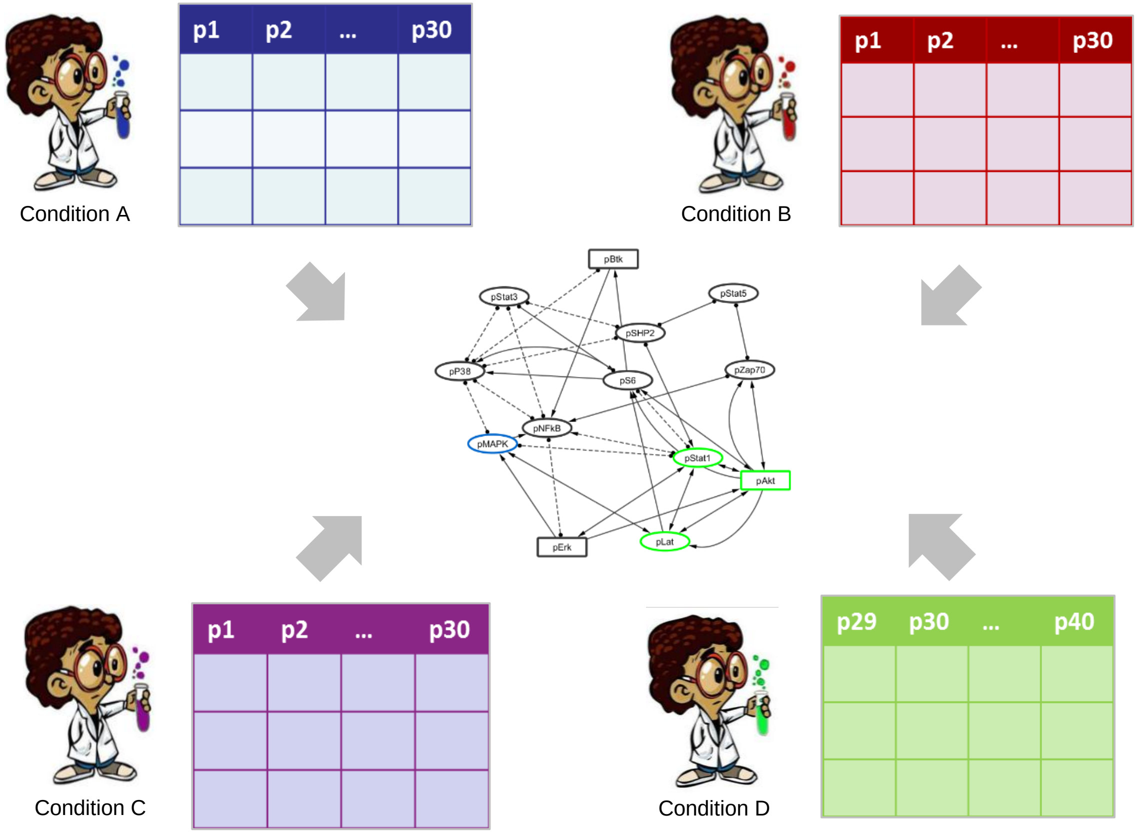
- Different experimental conditions
- Different variable sets

- Data can not be pulled together because they come from **different distributions**

Key point:

- Data come from the **same causal mechanism**
- Principles of causality links them to the underlying causal graph

Co-analyzing data sets from different experimental conditions with overlapping variable sets



Condition A

p1	p2	…	p30

Condition B

p1	p2	…	p30

Condition C

p1	p2	…	p30

Condition D

p29	p30	…	p40

Key point:
Identify the **causal graphs** that **simultaneously fit** all data

Predicting Correlations between Quantities Never Jointly Measured [JMLR 2012]

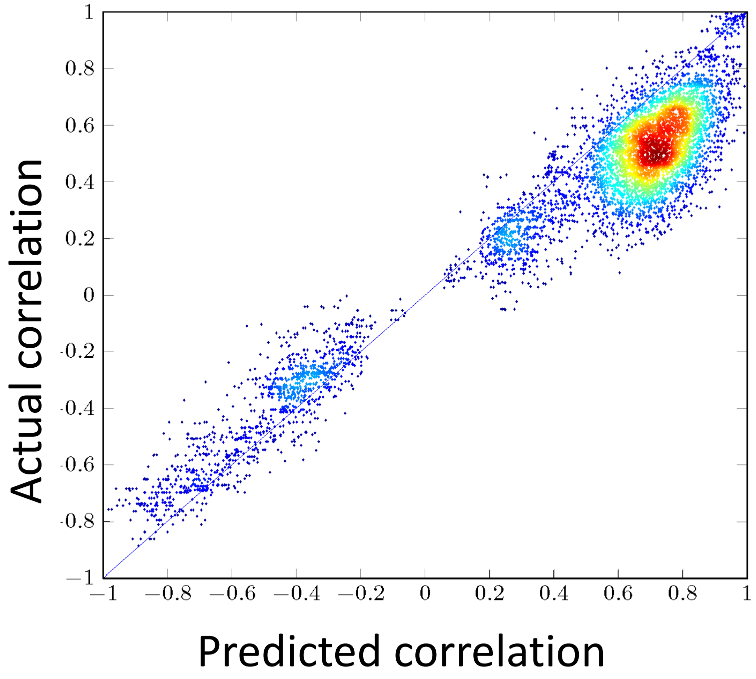
20 datasets

- **Biological, Financial, Text, Medical, Social**

698897 predictions

98% accuracy

- **vs. 16% for random guessing**



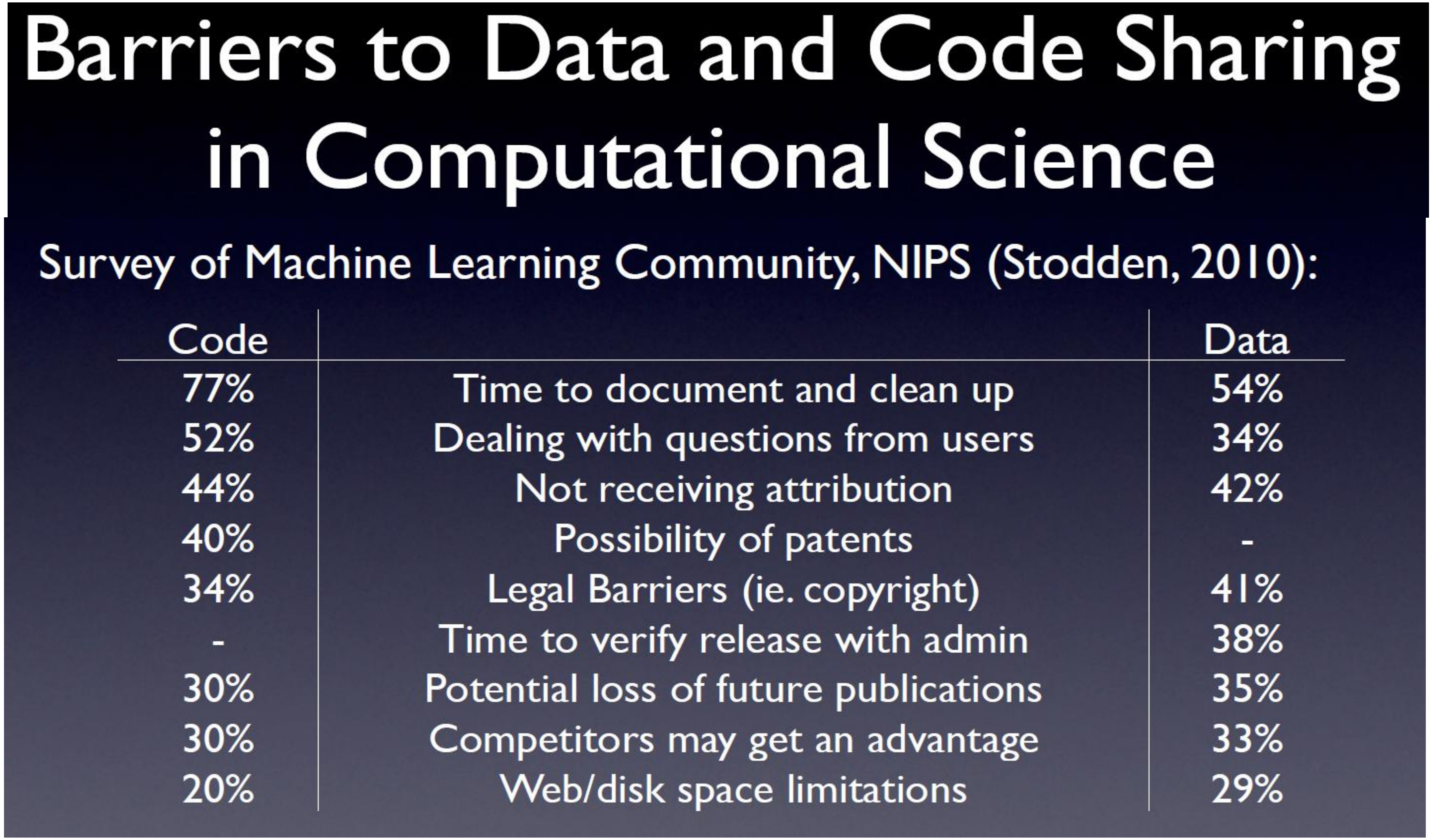
0.79 R^2 between predicted and sample correlation

ERC Consolidator Grant started 1/1/2015

Being Scientific About Science

- NASA lost the original Moon Landing Tapes!
- Beliefs and Knowledge based on Evidence
- Which evidence (data)?
- Which analysis methods?
- Which implementation of software?
- Combine with which knowledge?

Challenges: Culture



Barriers to Data and Code Sharing in Computational Science

Survey of Machine Learning Community, NIPS (Stodden, 2010):

Code		Data
77%	Time to document and clean up	54%
52%	Dealing with questions from users	34%
44%	Not receiving attribution	42%
40%	Possibility of patents	-
34%	Legal Barriers (ie. copyright)	41%
-	Time to verify release with admin	38%
30%	Potential loss of future publications	35%
30%	Competitors may get an advantage	33%
20%	Web/disk space limitations	29%

Challenge: Lack of Policies and Regulations

- Suppose the cure for cancer is found as follows
  - Multiple pharmaceuticals shared their data
  - With several data analysts
  - Who employed several data analysis software products
  - That were run on several computing infrastructures and clouds
  - Initiated by the request of an oncologist

- Who owns the IPR? What's the liability of each one for security leaks? What is the liability on the quality of service? What if entities involved are sited in different countries?

Challenge : Anonymization

- Dropping the name, tax id, etc. from patient data does not ensure anonymity due to cross correlation with other public data (de-identification)
- Solution 1: Distort the original data in order to reduce the re-identification risks while also preserving the statistical utility of the data
- **Share the analysis code not the data**; Distributed analysis

Conclusions

- Public data invaluable in biomedicine
  - Reproducibility
  - Testing new methods
  - External validation of results
  - Meta-Analysis
  - Integrative Analysis
- Open Data necessary for being Scientific About Science
- Still great cultural but also technical challenges