Limitless Arity Multiple Testing Procedures for Combinatorial Hypotheses

Koji Tsuda

Department of Computational Biology

Graduate School of Frontier Sciences

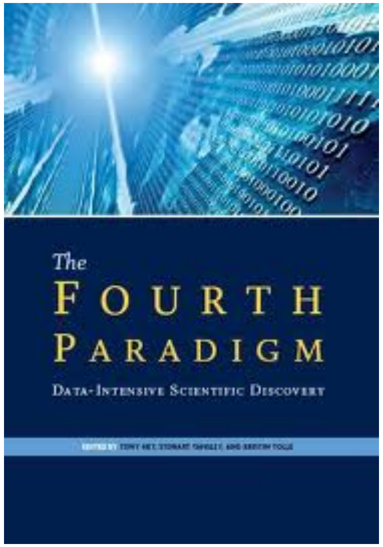
University of Tokyo

Joint work with Aika Terada, Mariko Okada-Hatakeyama and Jun Sese

Agenda

- Fourth Paradigm: Data-intensive Science
  - Efficiency and Reliability
- Itemset mining
- Novel multiple testing procedure for discovering combinatorial factors (LAMP)

Fourth Paradigm
Data-intensive Science

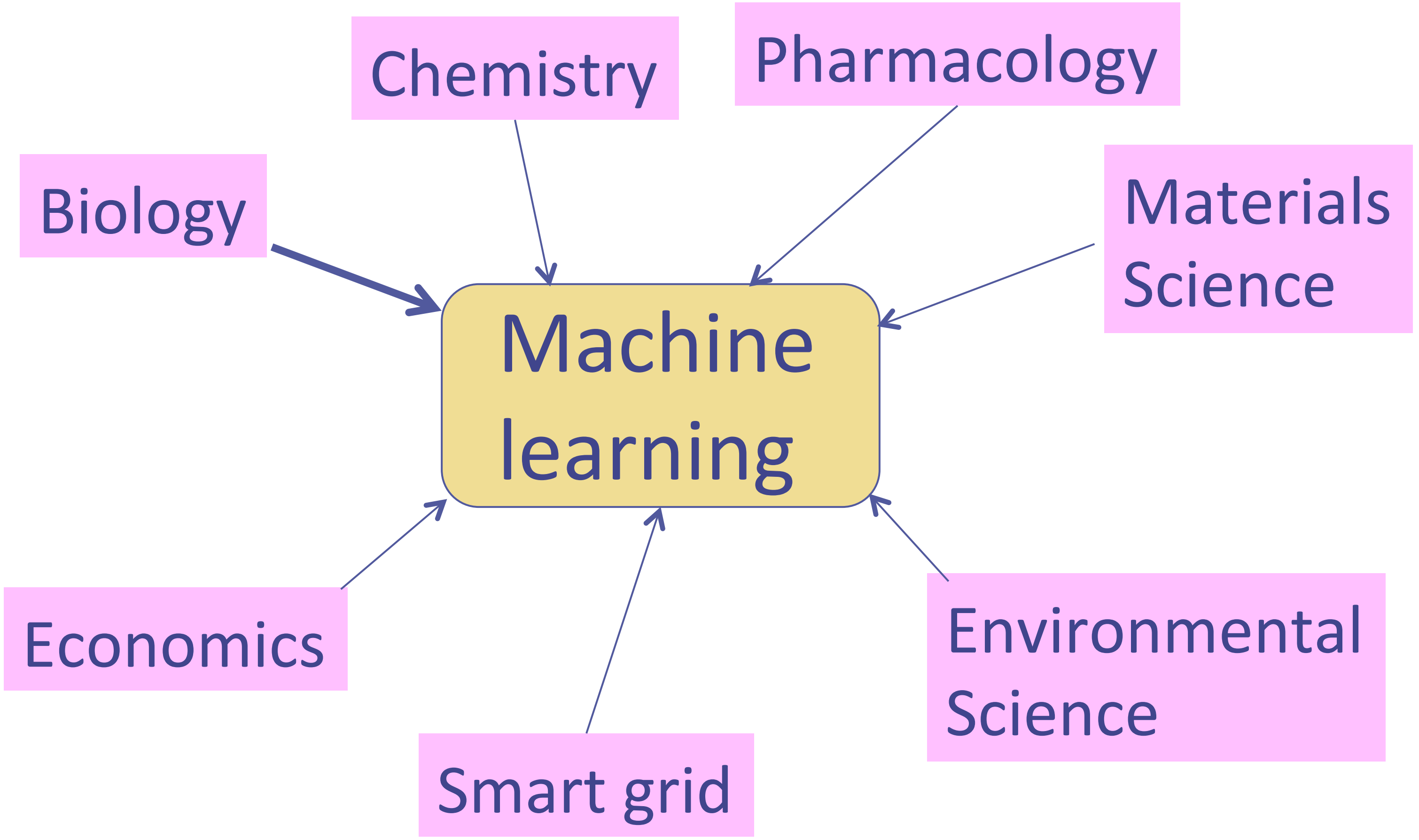


- 1st: Empirical Science
- 2nd: Theoretical Science
- 3rd: Computational Science (Simulation)
- 4th: Data-intensive Science

- Hypothesis determined by humans ⇒ Verification by data

- NEW : Hypothesis generated by data analysis ⇒Verification by data

Ever increasing demand for data scientists



Chemistry

Pharmacology

Biology

Materials Science

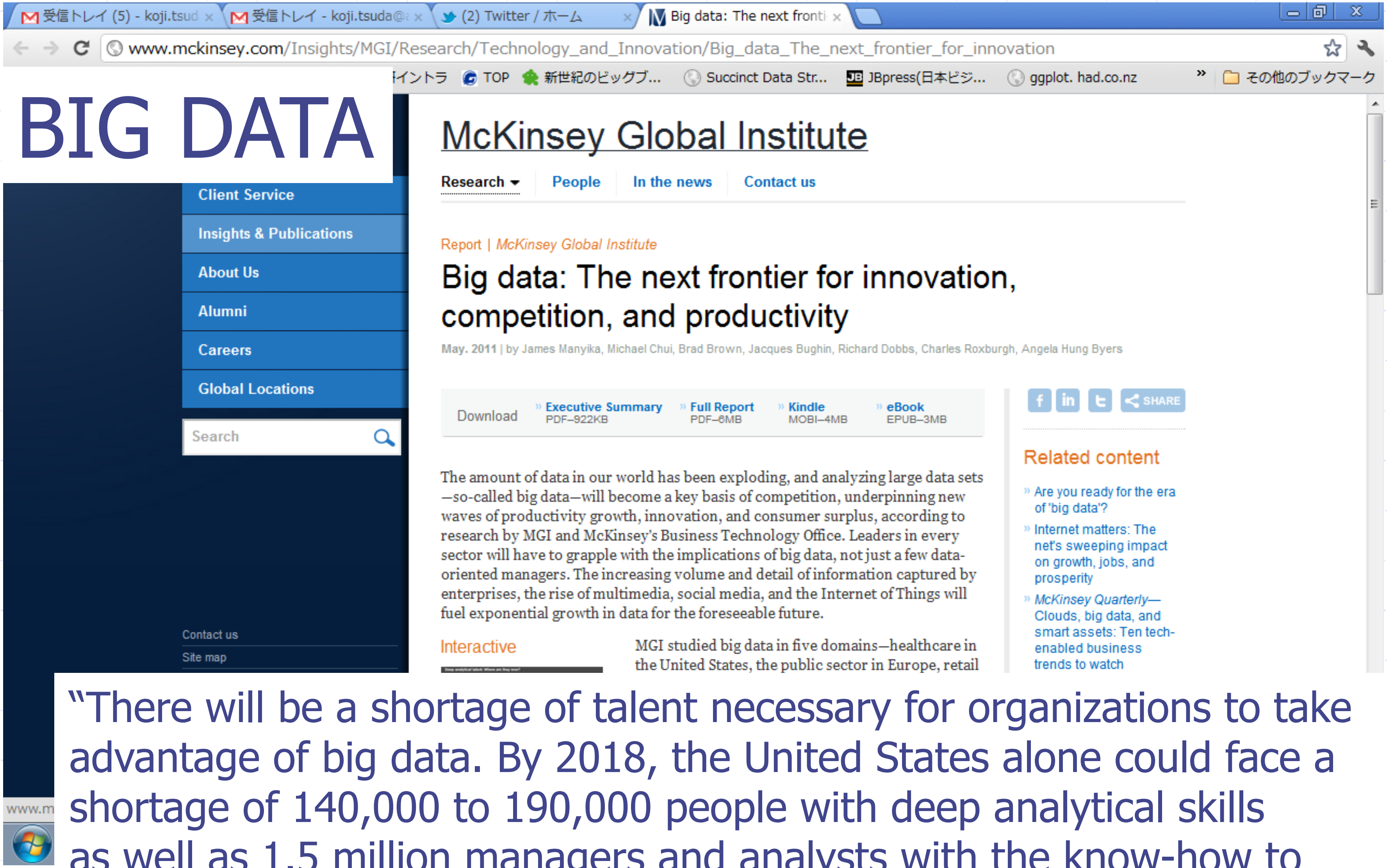
Machine learning

Economics

Environmental Science

Smart grid

BIG DATA



McKinsey Global Institute

Research People In the news Contact us

Report | *McKinsey Global Institute*

Big data: The next frontier for innovation, competition, and productivity

May. 2011 | by James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byers

Download » Executive Summary PDF–922KB » Full Report PDF–6MB » Kindle MOBI–4MB » eBook EPUB–3MB

The amount of data in our world has been exploding, and analyzing large data sets—so-called big data—will become a key basis of competition, underpinning new waves of productivity growth, innovation, and consumer surplus, according to research by MGI and McKinsey's Business Technology Office. Leaders in every sector will have to grapple with the implications of big data, not just a few data-oriented managers. The increasing volume and detail of information captured by enterprises, the rise of multimedia, social media, and the Internet of Things will fuel exponential growth in data for the foreseeable future.

Interactive

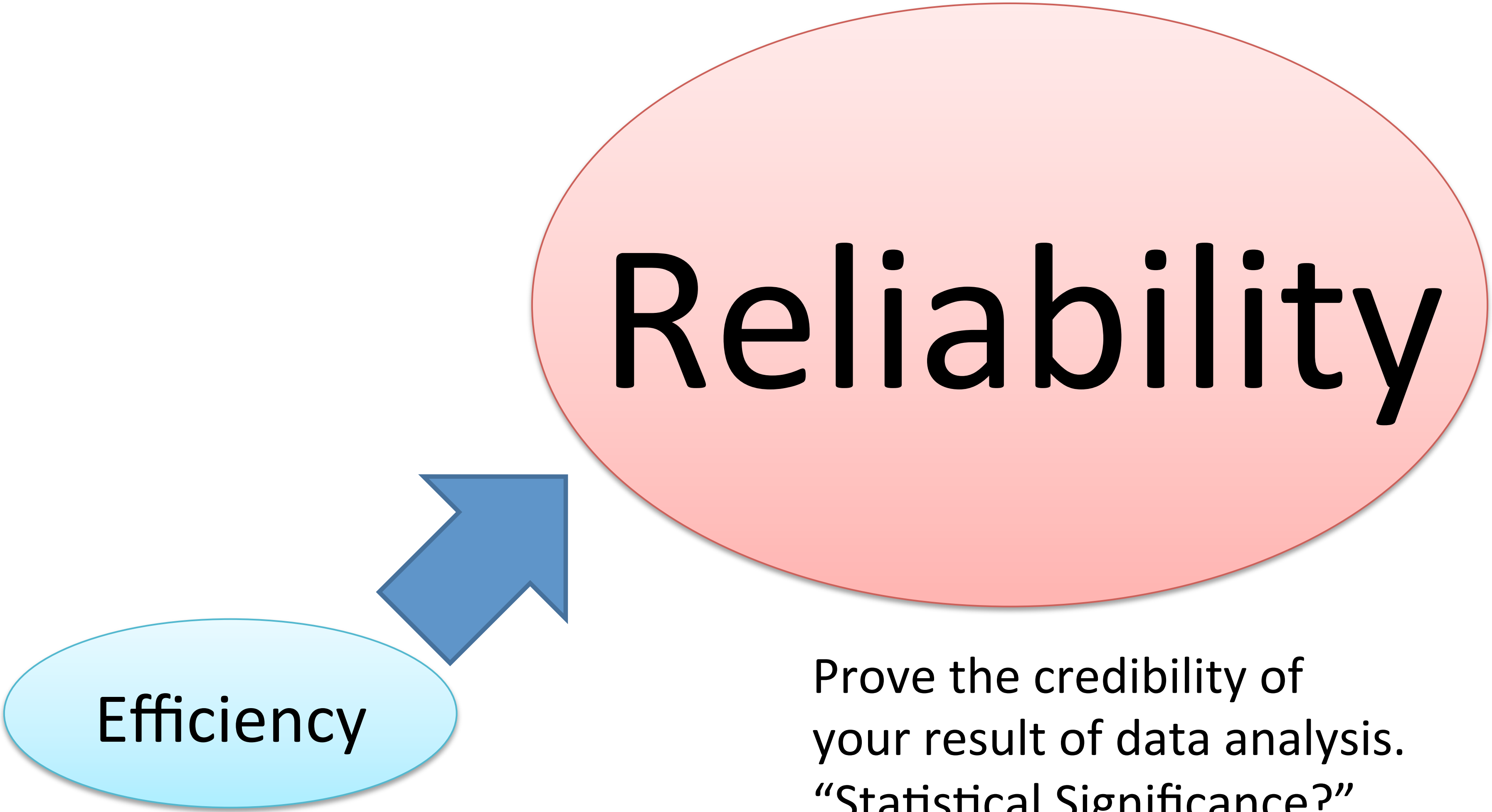
MGI studied big data in five domains—healthcare in the United States, the public sector in Europe, retail

Related content

- Are you ready for the era of 'big data'?
- Internet matters: The net's sweeping impact on growth, jobs, and prosperity
- *McKinsey Quarterly*—Clouds, big data, and smart assets: Ten tech-enabled business trends to watch

"There will be a shortage of talent necessary for organizations to take advantage of big data. By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions."

What’s important in data-intensive science

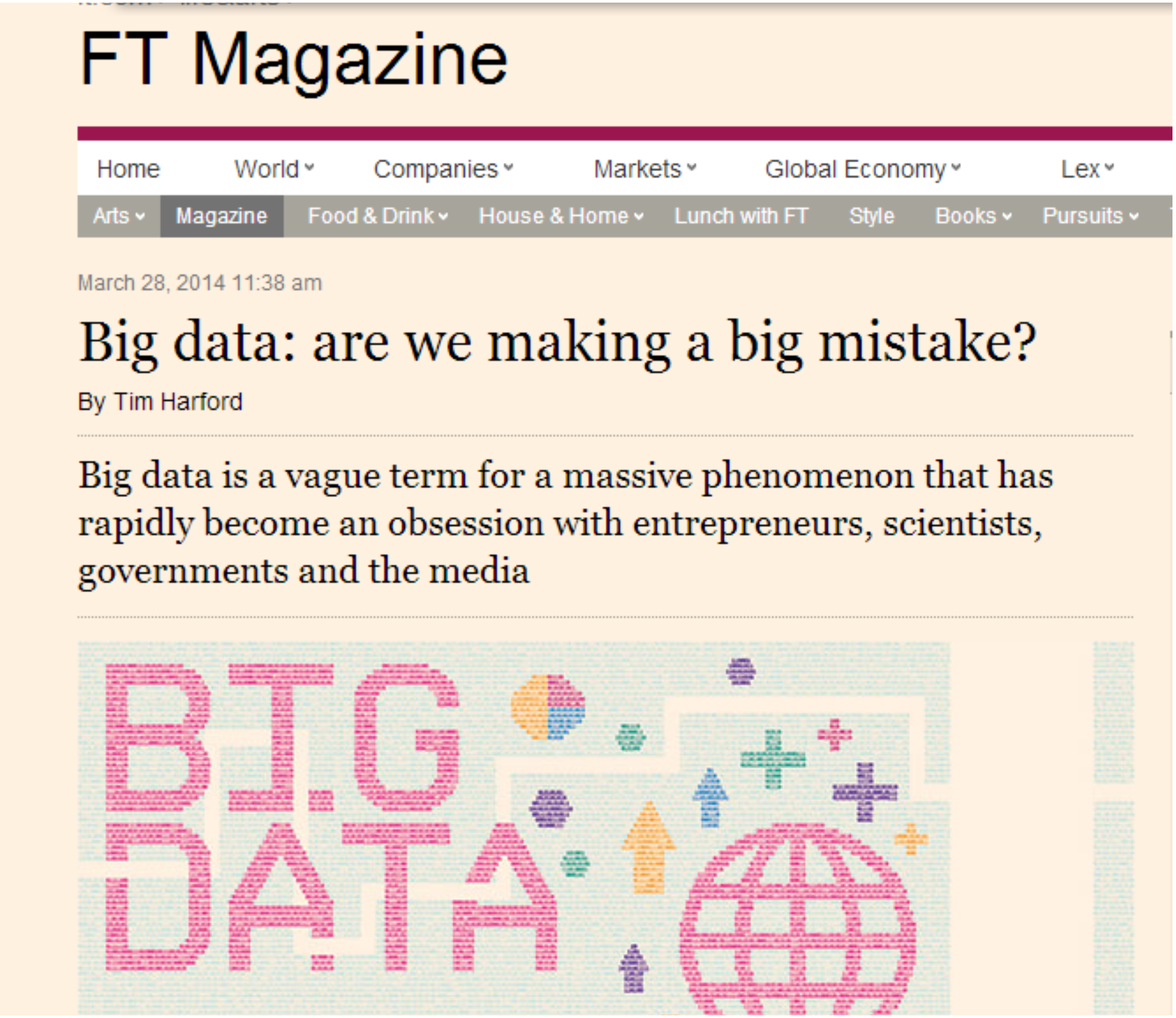


Reliability

Efficiency

Common Big-Data Studies

Prove the credibility of
your result of data analysis.
“Statistical Significance?”



March 28, 2014 11:38 am

Big data: are we making a big mistake?

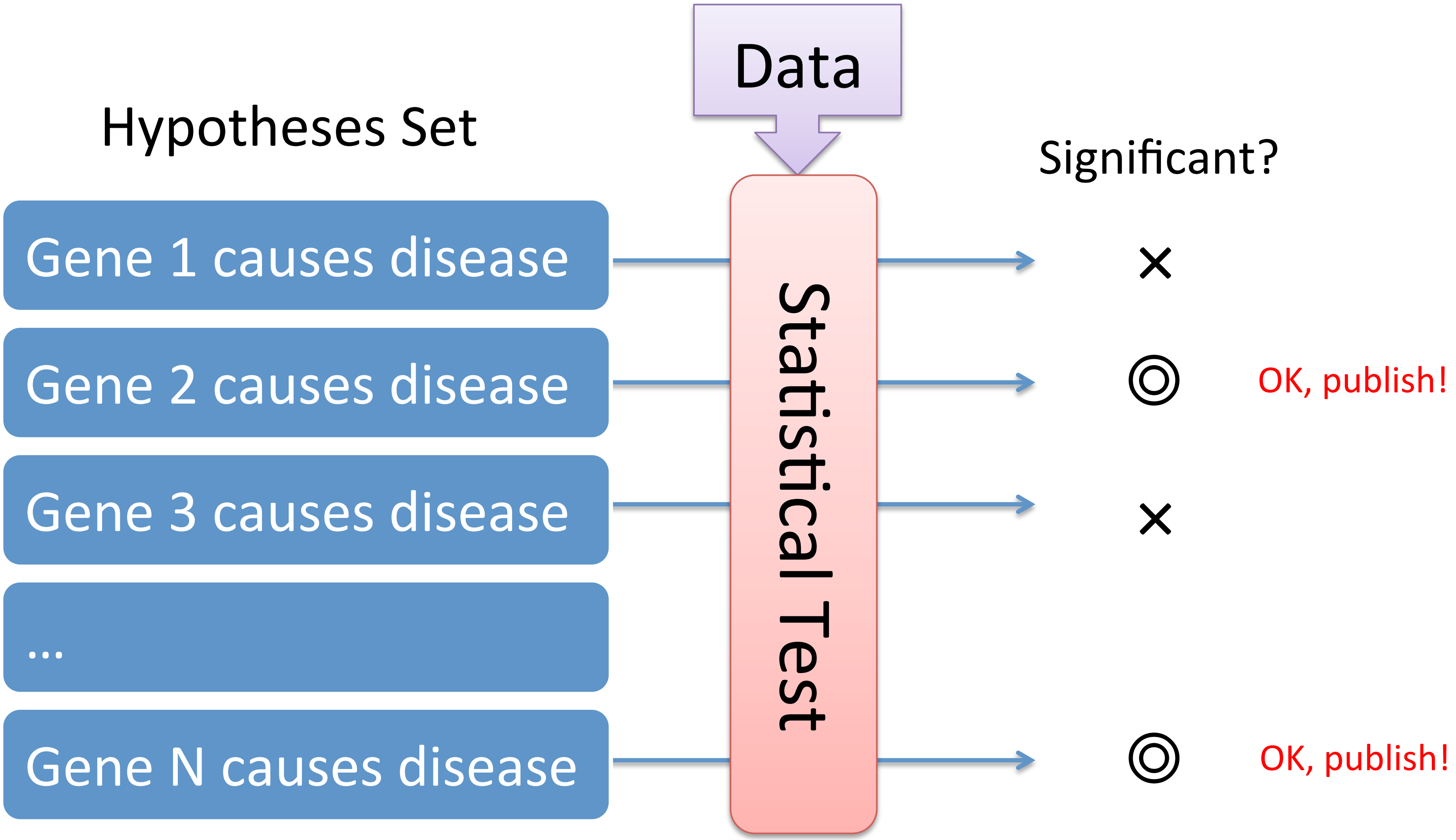
By Tim Harford

Big data is a vague term for a massive phenomenon that has rapidly become an obsession with entrepreneurs, scientists, governments and the media

Quotes

- In 2005, John Ioannidis, an epidemiologist, published a research paper with the self-explanatory title, “Why Most Published Research Findings Are False”. The paper became famous as a provocative diagnosis of a serious issue. One of the key ideas behind Ioannidis’s work is what statisticians call the “multiple-comparisons problem”.

Publishing system in life sciences



Hypotheses Set

Data

Significant?

Gene 1 causes disease

Statistical Test

×

Gene 2 causes disease

◎

OK, publish!

Gene 3 causes disease

×

…

Gene N causes disease

◎

OK, publish!

Reproducibility Crisis!
(and statistics is to blame)

- Biological results reported in journals cannot be reproduced
  - Bayer: could not reproduce 43 of 67 studies
  - Amgen: could not reproduce 47 of 53 studies

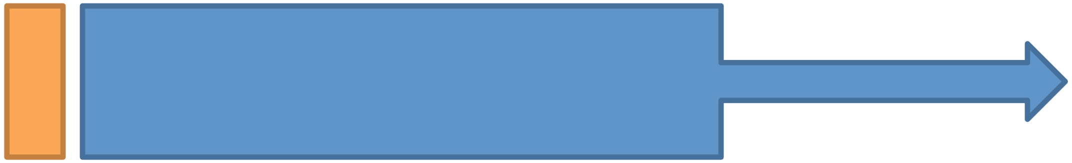
J. Ioannidis (Stanford)

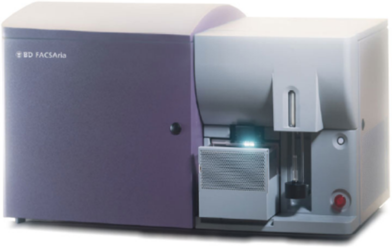
Data that are multidimensional (ie contain many features) are particularly at risk of false positives and overfitting, particularly when analyzed by inexperienced or untrained analysts.
(Lancet, 2014)

Curse of Dimensionality in Testing

Huge increase in explanatory variables

No increase in examples



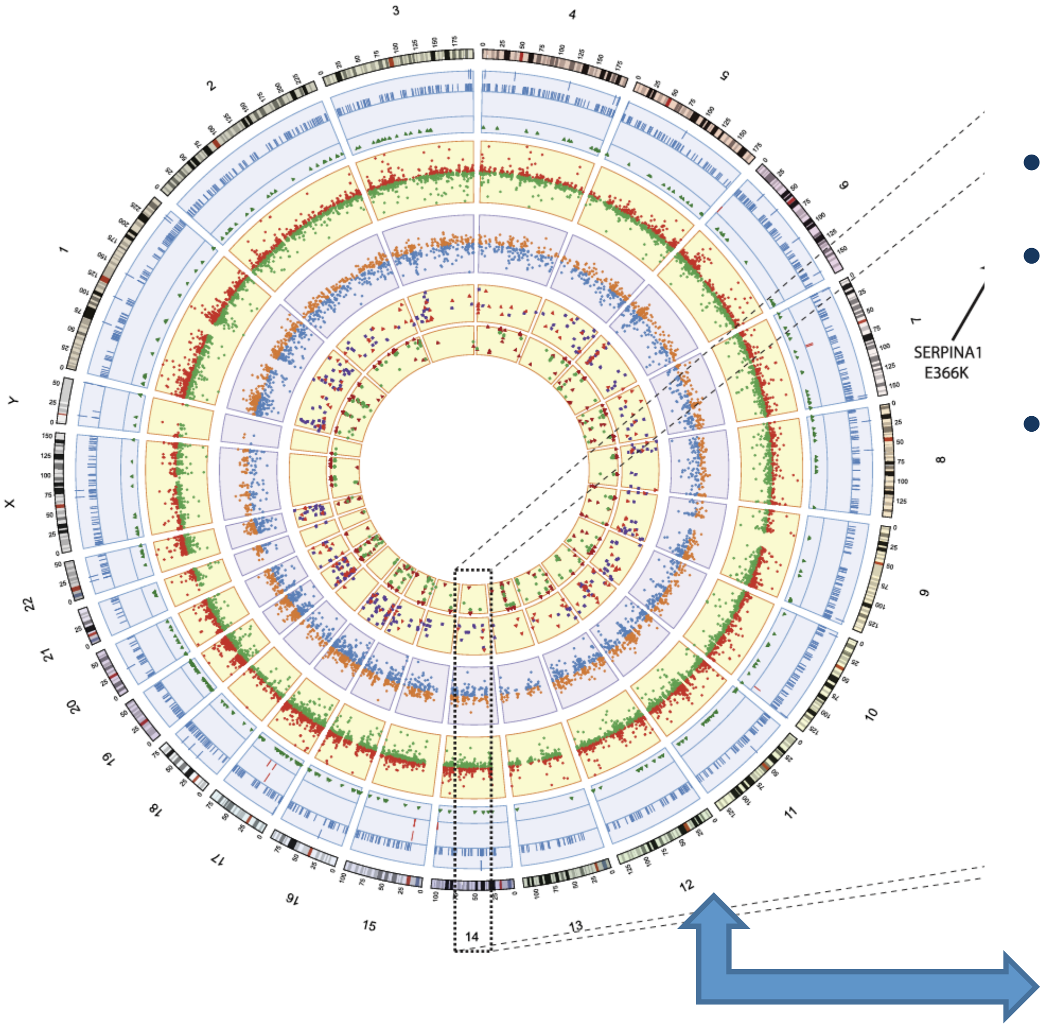


False positive more likely

→Have to apply stricter criterion

→Fewer discovery (!)

Trans-omics Data



- DNA（mutation, insertion, deletion, CNV etc）
- DNA methylation, Histon modification
- mRNA expression, ncRNA
- Protein expression, modification
- Metabolite（Sugar, Amino acids, Nucleotides, lipids)

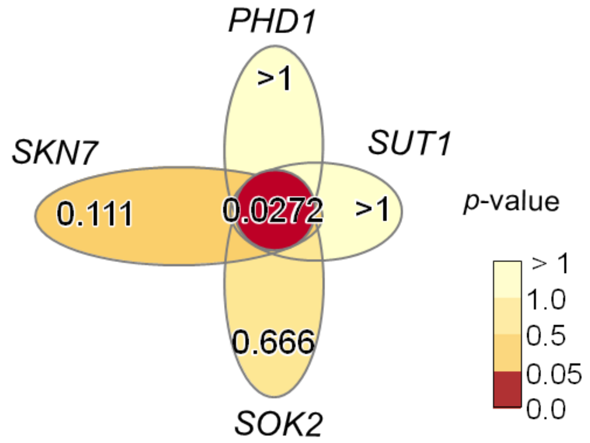
Clinical Data

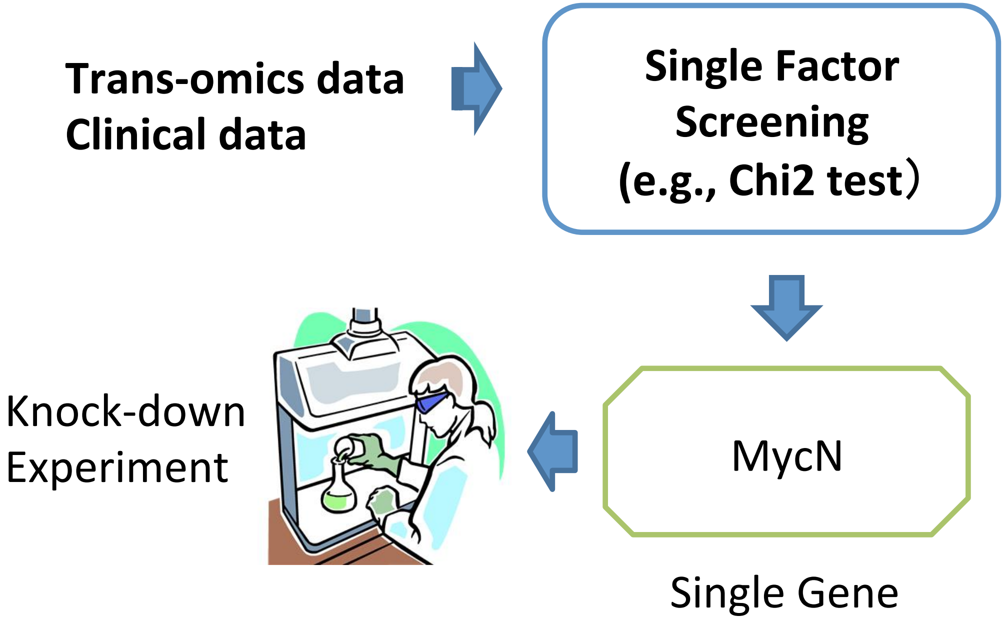
Survival rate, Drug resistance, Relapse, Family history

Drawbacks of “Single Factor Screening”

- Discover single factor causing phenotype (e.g., disease)
- BUT cellular processes are highly combinatorial

Single factor screening misses combinatorial causes





Trans-omics data
Clinical data

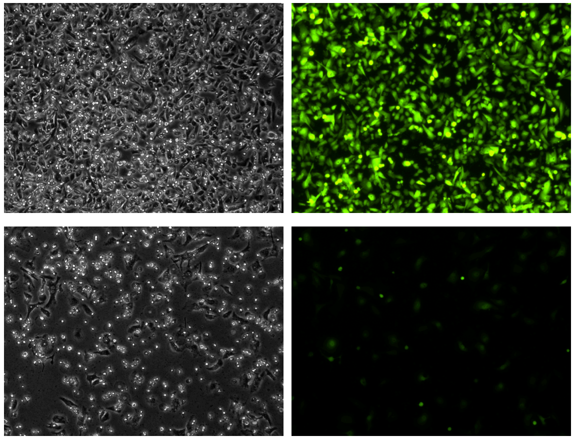
Single Factor Screening (e.g., Chi2 test)

MycN

Single Gene

Knock-down Experiment

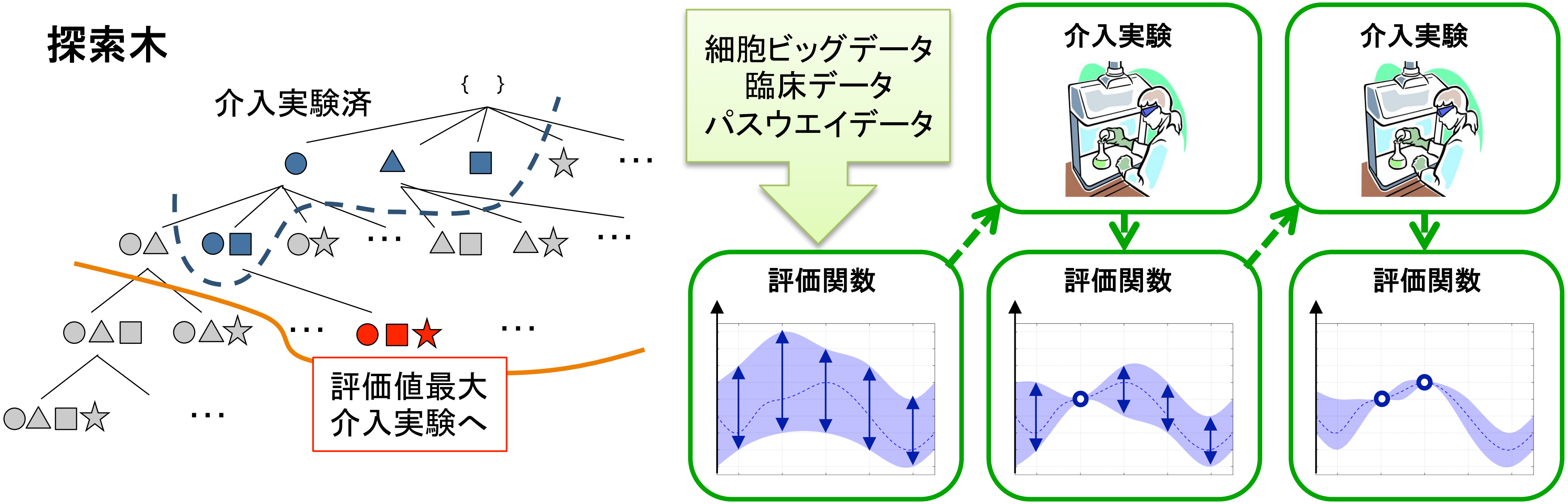
Knock down Experiments



Challenge:

Discovering Combinatorial Factors Associated with Biological Phenomena

- **Combinatorial Explosion**
  - 100m SNP x 10,000 Expression x 10,000 CNV = 100 trillion scores
- Search tree, Evaluation scores, Pruning, Ranking branches
- Develop sophisticated algorithms including itemset mining



探索木

介入実験済

{ }

評価値最大
介入実験へ

細胞ビッグデータ
臨床データ
パスウエイデータ

介入実験

介入実験

評価関数

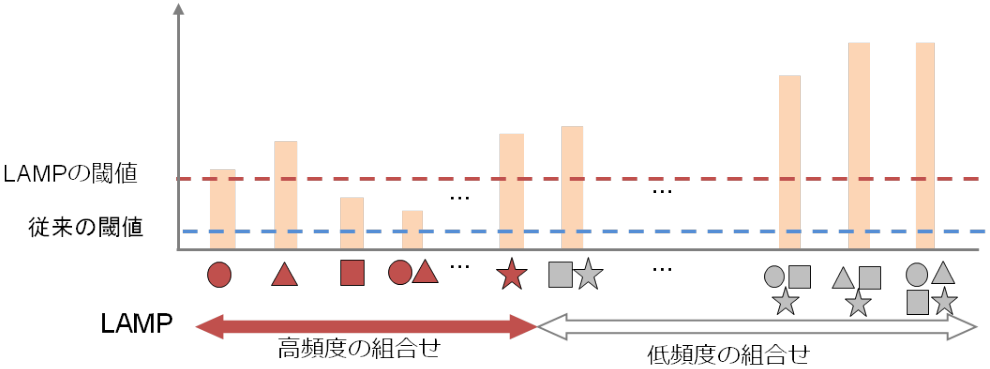
評価関数

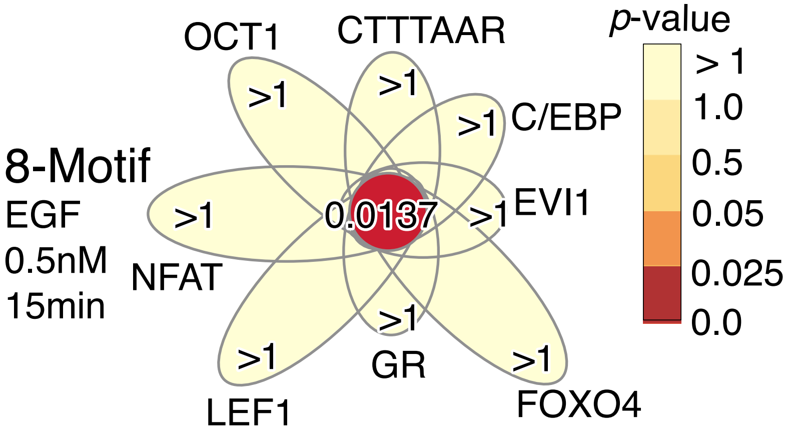
評価関数

Limitless Arity Multiple testing Procedure (LAMP)

Terada, Okada-Hatakeyama, Tsuda and Sese, PNAS, published (July 23)

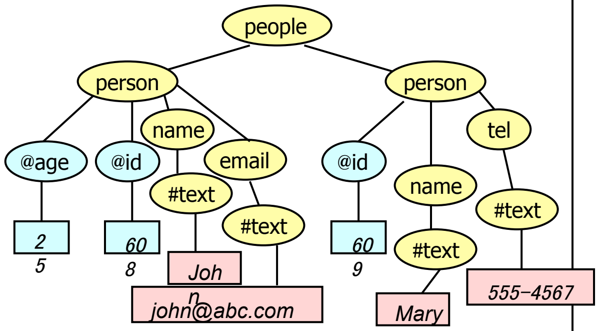
- **Reliability of scientific discovery is assessed by P-values**
- **Multiple test (Bonferroni): If n candidate factors are available, use 0.05/n as significance level**
- **Number of combinatorial factors is huge: No chance of discovery**
- **Reduce the Bonferroni factor dramatically by itemset mining-based algorithm**





Itemset mining

Data Mining



- A formal study of efficient methods for extracting interesting rules and patterns from massive data

- Frequent itemset mining (Agrawal and Srikant 1994)
- Closed pattern mining
- Structured data mining (Sequence, Trees, and Graphs)

Frequent Itemset Mining

[Agrawal, Srikant, VLDB'94]

- Finding all "frequent" sets of elements (items) appearing σ times or more in a database

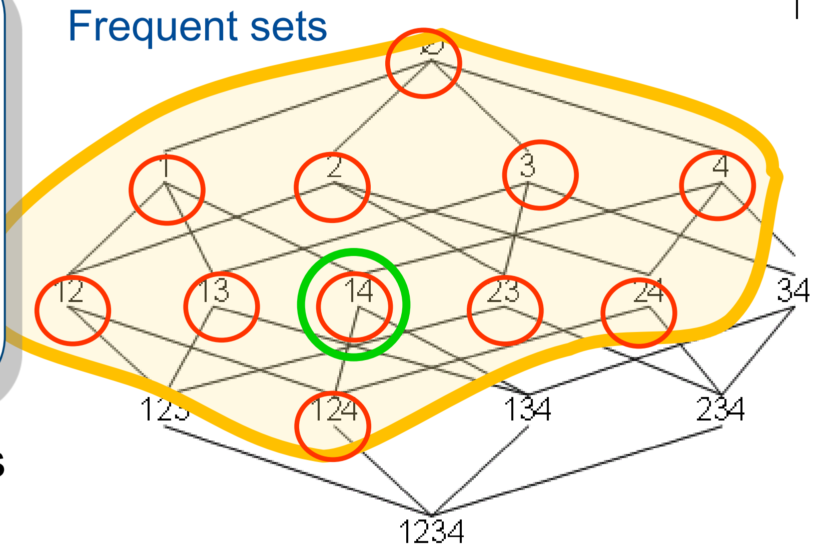
Minsup σ= **2**

	1	2	3	4	5
t1	○		○		
t2		○		○	
t3	○	○	○	○	
t4		○	○		○
t5	○	○		○	

database

Frequent sets
∅,
1, 2, 3, 4,
12, 13, 14,
23, 24, 124

X = {2, 4} appears **three times**, thus **frequent**



The itemset lattice (2^{Σ}, ⊆)

Market Basket Data

- Popular application of itemset mining
- Business and Market data analysis

- Transaction Data of purchase

- Item

- a transaction or a "basket"

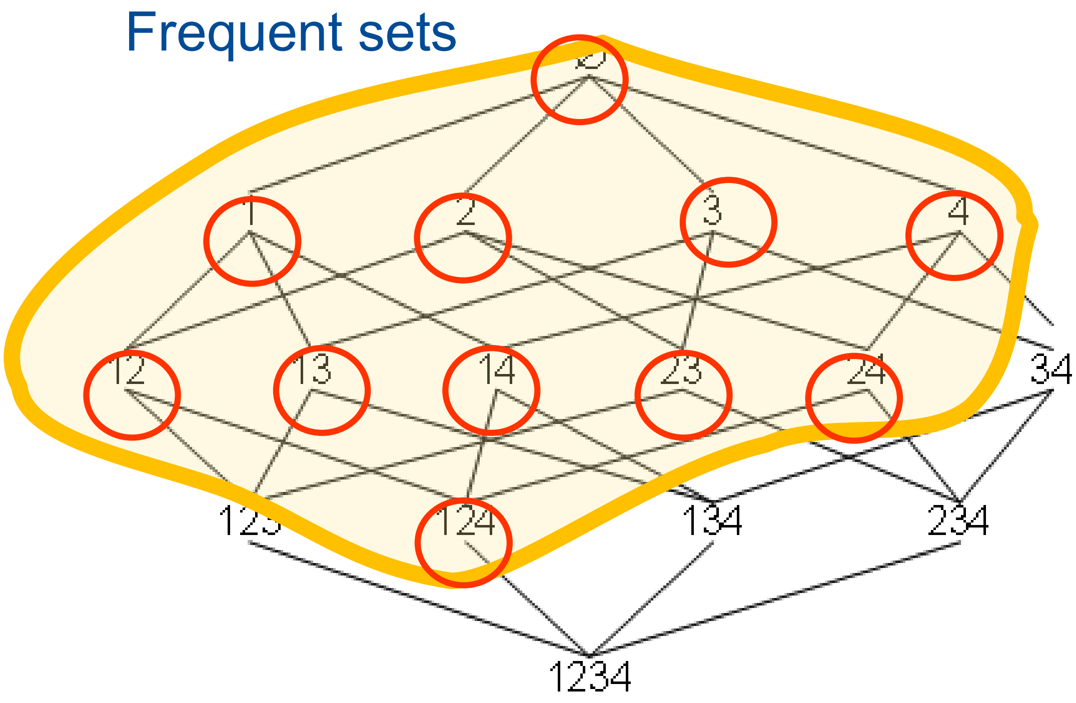
ID	Chips	Mustard	Sausage	Softdrink	Beer
001	1	0	0	0	1
002	1	1	1	1	1
003	1	0	1	0	0
004	0	0	1	0	1
005	0	1	1	1	1
006	1	1	1	0	1
007	1	0	1	1	1
008	1	1	1	0	0
009	1	0	0	1	0

- **Meaning of the transaction 003**

  "Custmer 003 bought Chips and Sausage together in his basket"

Backtracking Algorithm: FP Growth etc.

- Monotonicity: Support only decreases
- Depth First Traversal, Prune if support < σ

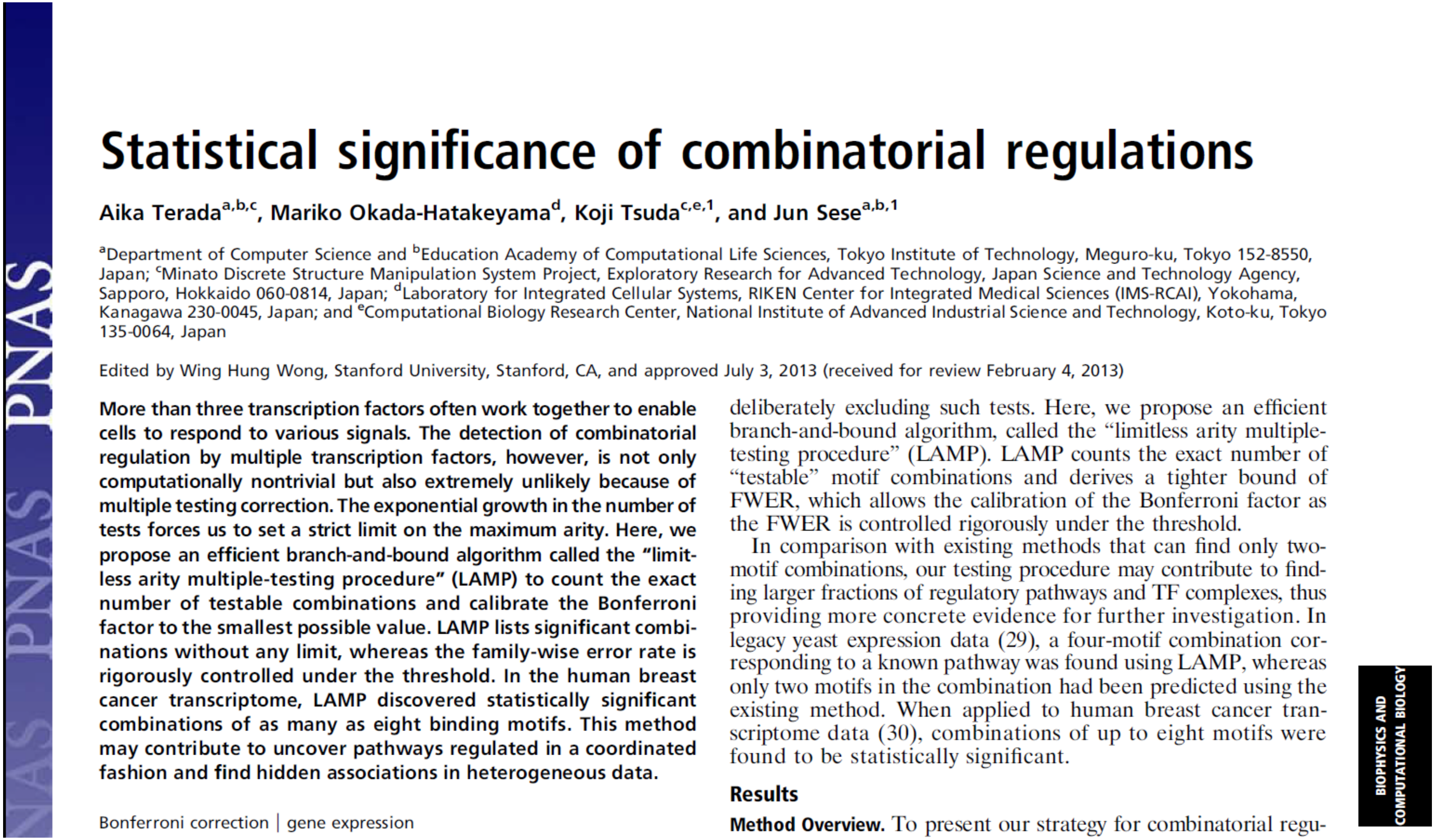


Summary: Itemset mining

- Itemset mining is the simplest of all mining algorithms
- Need to maintain occurrence of each pattern in database
- Tree by lexicographical order is (implicitly) used

Novel multiple testing procedure for discovering combinatorial factors (LAMP)

PNAS, 2013



Statistical significance of combinatorial regulations

Aika Terada[a,b,c], Mariko Okada-Hatakeyama[d], Koji Tsuda[c,e,1], and Jun Sese[a,b,1]

[a]Department of Computer Science and [b]Education Academy of Computational Life Sciences, Tokyo Institute of Technology, Meguro-ku, Tokyo 152-8550, Japan; [c]Minato Discrete Structure Manipulation System Project, Exploratory Research for Advanced Technology, Japan Science and Technology Agency, Sapporo, Hokkaido 060-0814, Japan; [d]Laboratory for Integrated Cellular Systems, RIKEN Center for Integrated Medical Sciences (IMS-RCAI), Yokohama, Kanagawa 230-0045, Japan; and [e]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Koto-ku, Tokyo 135-0064, Japan

Edited by Wing Hung Wong, Stanford University, Stanford, CA, and approved July 3, 2013 (received for review February 4, 2013)

More than three transcription factors often work together to enable cells to respond to various signals. The detection of combinatorial regulation by multiple transcription factors, however, is not only computationally nontrivial but also extremely unlikely because of multiple testing correction. The exponential growth in the number of tests forces us to set a strict limit on the maximum arity. Here, we propose an efficient branch-and-bound algorithm called the "limitless arity multiple-testing procedure" (LAMP) to count the exact number of testable combinations and calibrate the Bonferroni factor to the smallest possible value. LAMP lists significant combinations without any limit, whereas the family-wise error rate is rigorously controlled under the threshold. In the human breast cancer transcriptome, LAMP discovered statistically significant combinations of as many as eight binding motifs. This method may contribute to uncover pathways regulated in a coordinated fashion and find hidden associations in heterogeneous data.

Bonferroni correction | gene expression

deliberately excluding such tests. Here, we propose an efficient branch-and-bound algorithm, called the "limitless arity multiple-testing procedure" (LAMP). LAMP counts the exact number of "testable" motif combinations and derives a tighter bound of FWER, which allows the calibration of the Bonferroni factor as the FWER is controlled rigorously under the threshold.

In comparison with existing methods that can find only two-motif combinations, our testing procedure may contribute to finding larger fractions of regulatory pathways and TF complexes, thus providing more concrete evidence for further investigation. In legacy yeast expression data (29), a four-motif combination corresponding to a known pathway was found using LAMP, whereas only two motifs in the combination had been predicted using the existing method. When applied to human breast cancer transcriptome data (30), combinations of up to eight motifs were found to be statistically significant.

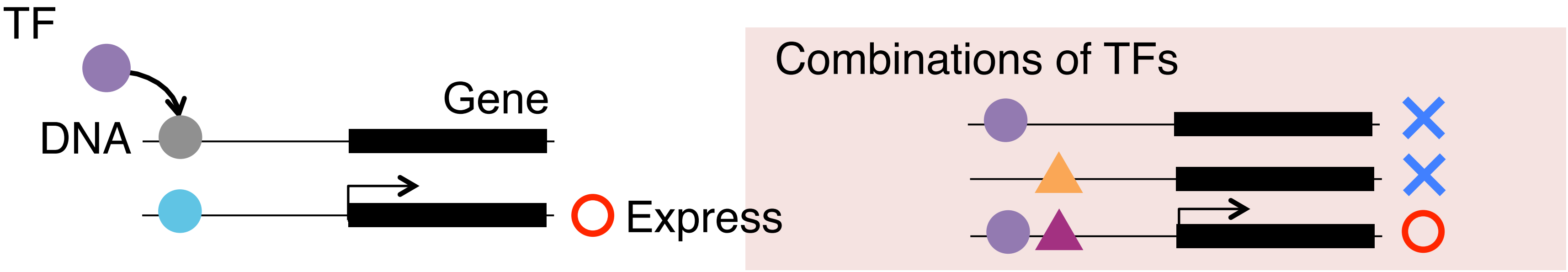
Results

Method Overview. To present our strategy for combinatorial regu-

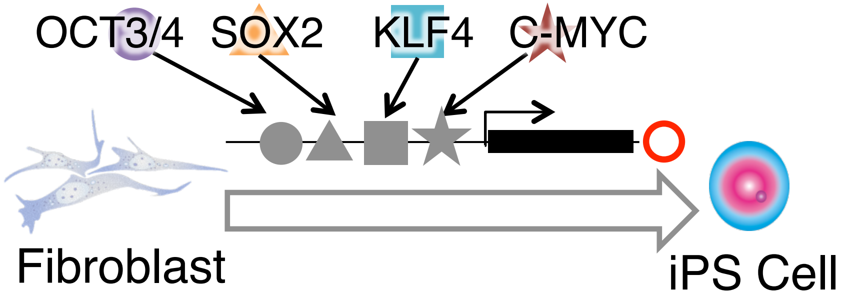
BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Transcription factors (TFs) work in combination

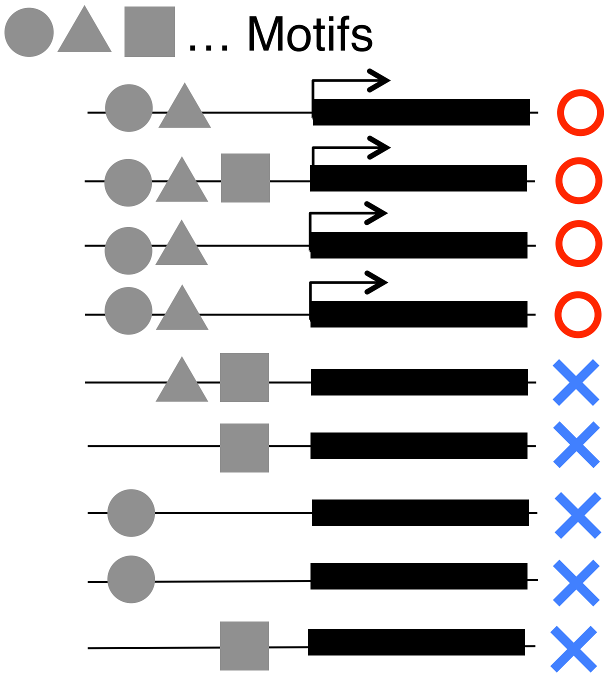
- Often several TFs are necessary to induce the expression of downstream genes



Example: Yamanaka Factor（K. Okita *et al.*, Nature, 2007）



Find statistically significant combinations of TF binding motifs



Contingency table for ●▲

	Up-regulated	No-regulated
With Motif Combination	4	0
Without	0	5

P-value by Fisher exact test 0.0079

Significant?

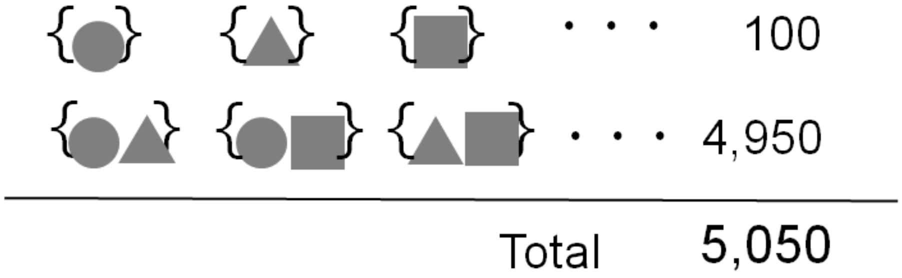
No – You have to apply multiple testing procedure

Bonferroni Correction

- Family-wise error rate(FWER)
  - At least one false discovery occurs
- P-value threshold δ is determined such that FWER is below α
- For m tests,

$$\delta = \frac{\alpha}{m}$$

- 100 motifs in total
- Number of tests



- Corrected threshold

  $\delta = 0.05/5050$

  $= 9.9\times10^{-6}$
- Bonferroni is too conservative!

New Proposal:
Limitless Arity Multiple testing Procedure (LAMP)

- Count the exact number of “testable” combinations
  - Infrequent combinations do not affect family-wise error rate
  - Stepwise procedure involving itemset mining
- Calibrate the correction factor to the smallest possible value
- Discovered statistically significant motif combinations in yeast and breast cancer expression data

Raw p-value

	Up regulated	No regulated
With Motif Combination	a	b
Without	c	d

- Null Hypothesis H
  - Two variables are independent
- P-value: p(a,b,c,d)
  - Probability of observing stronger table than observed
  - If smaller than α, reject H (discovery!)
- Type-I error: reject H when it is true
- Probability of type-I error must satisfy

$$P(p < \alpha \mid H) \leq \alpha$$

Multiple Tests

- m null hypotheses $H_1,…,H_m$
- V: Number of rejections in m tests
- Probability that more than one type-I error occurs: Family-wise error rate (FWER)

$$P(V > 0 \mid \bigcap_{i=1}^{m} H_i)$$

- Multiple testing procedures aim to control FWER under α

Bonferroni Correction

- Given threshold δ、FWER is bounded as

$$P(V > 0 \mid \bigcap_{i=1}^{m} H_i) \le \sum_{i=1}^{m} P(p_i \le \delta \mid H_i)$$

Union bound

$$\le m\delta$$

Definition of p-value

- Thus, setting δ=α/m calibrate FWER bound to α

	Up-regulated	Not regulated	
With Motif Combination	a	b	x
Without	c	d	N-x
	n_u	N-n_u	N

Occurrence Frequency

- P-value by Fisher exact test cannot be smaller than

$$f(x) = \binom{n_u}{x} \Big/ \binom{N}{x}$$

- No chance of false discovery, if $f(x) \geq \delta$

$$P(p < \delta \mid H) = 0$$

Tarone Correction (Biometrics, 1990)

- Considering minimum p-value, FWER is bounded as follows

$$P(V > 0 \mid \bigcap_{i=1}^{m} H_i) \le \sum_{i=1}^{m} P(p_i \le \delta \mid H_i)$$

Union bound

$$= \sum_{\{i \mid f(x_i) \ge \delta\}} P(p_i \le \delta \mid H_i)$$

Use minimum p-value to remove hypotheses

$$\le \mid \{i \mid f(x_i) \ge \delta\} \mid \delta$$

Definition of p-value

- Take maximum δ that keeps FWER bound below α

- FWER is represented as

$$g(\delta) = |\{i \mid f(x_i) \geq \delta\}| \delta$$

- Identify all motif combinations that satisfy

$$f(x) \geq \delta$$

- Inverse function

$$f^{-1}(\delta) = \lambda \; s.t. \; f(\lambda) \leq \delta \leq f(\lambda - 1)$$

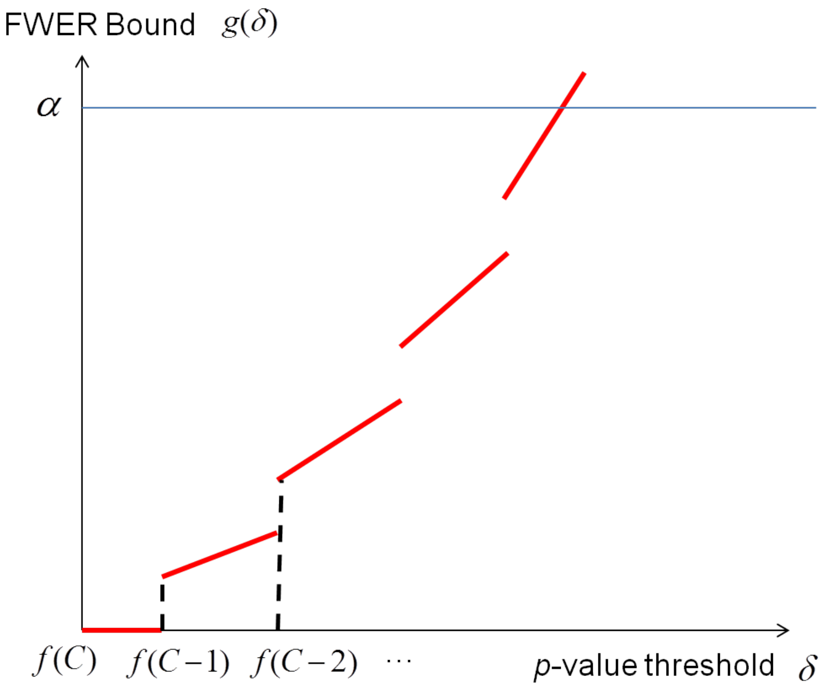
- Find all combinations whose frequency is λ or more by itemset mining
- FWER bound is computed as

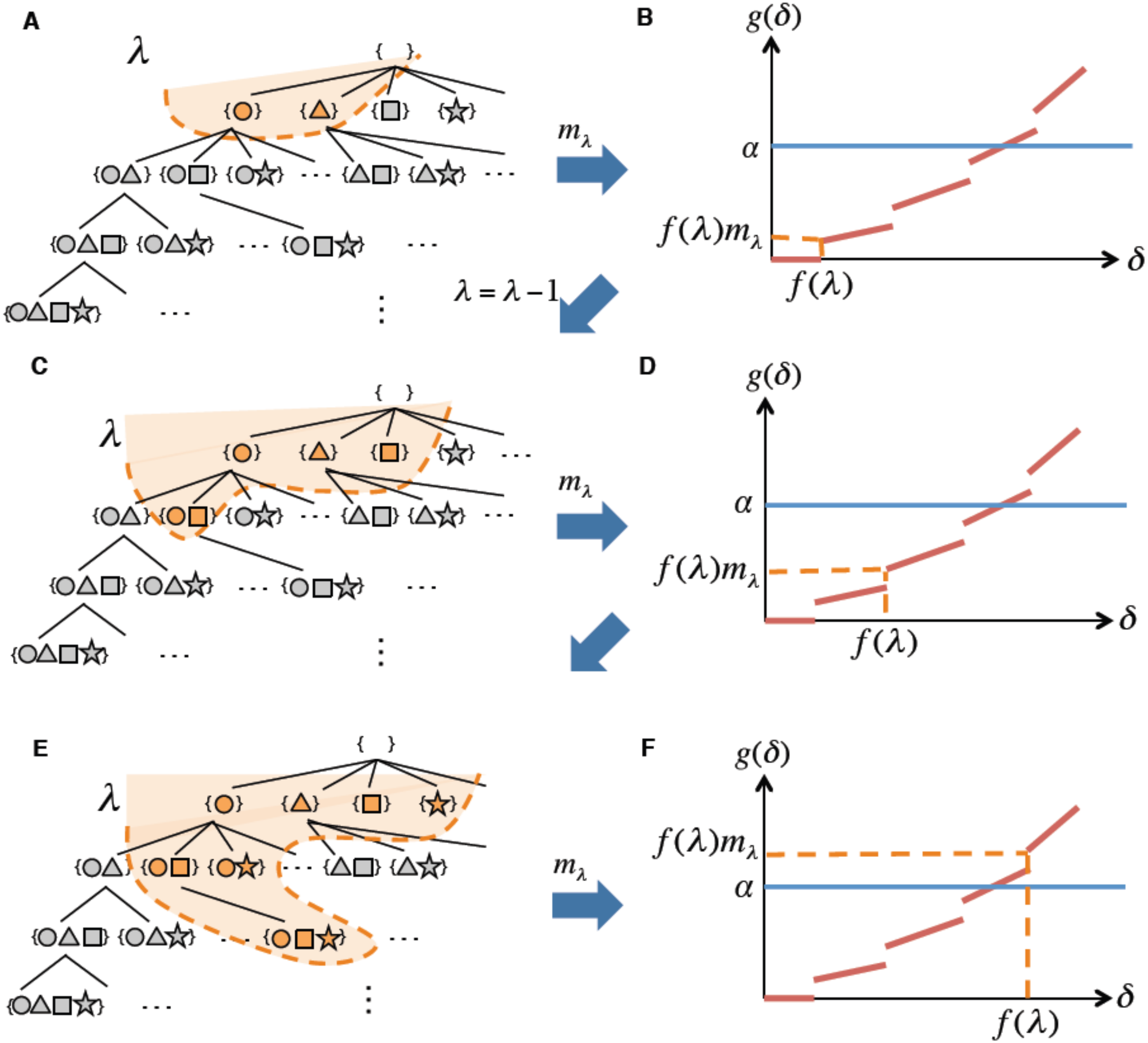
$$g(\delta) = m'\delta$$

m’: Number of motif combintions whose
frequency is λ or more

Finding optimal δ that calibrates FWER bound to α

- FWER bound is piecewise linear
- Repeat itemset mining with decrementing the frequency parameter
- A line segment drawn by a mining call
- Finish if line segment reaches α




A
λ
{ }
{○} {△} {□} {☆}
{○△} {○□} {○☆} ··· {△□} {△☆} ···
{○△□} {○△☆} ··· {○□☆} ···
{○△□☆} ···
m_λ
$\lambda = \lambda - 1$
B
$g(\delta)$
α
$f(\lambda)m_\lambda$
$f(\lambda)$
δ
C
λ
{ }
{○} {△} {□} {☆} ···
{○△} {○□} {○☆} ··· {△□} {△☆} ···
{○△□} {○△☆} ··· {○□☆} ···
{○△□☆} ···
m_λ
D
$g(\delta)$
α
$f(\lambda)m_\lambda$
$f(\lambda)$
δ
E
λ
{ }
{○} {△} {□} {☆}
{○△} {○□} {○☆} ··· {△□} {△☆}
{○△□} {○△☆} ··· {○□☆} ···
{○△□☆} ···
m_λ
F
$g(\delta)$
$f(\lambda)m_\lambda$
α
$f(\lambda)$
δ

Applications to Yeast Transcriptome

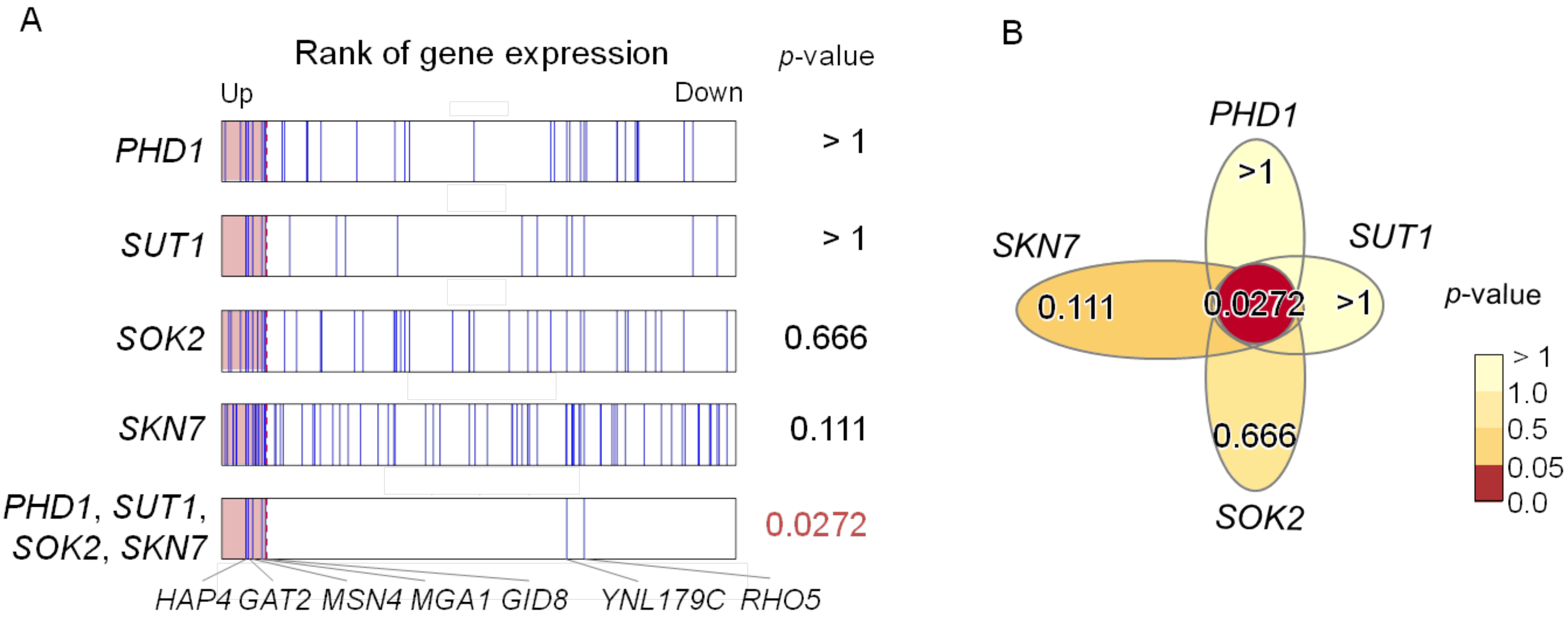
- Microarray data by Gasch et al
- Binding motif data by SGD Database
- 102 motifs, each binding to 30.1 genes on average
- Expressions of about 6000 genes measured on 173 different conditions

Statistically significant TF combinations under a heat shock condition

Corrected p-value（p-value*K）

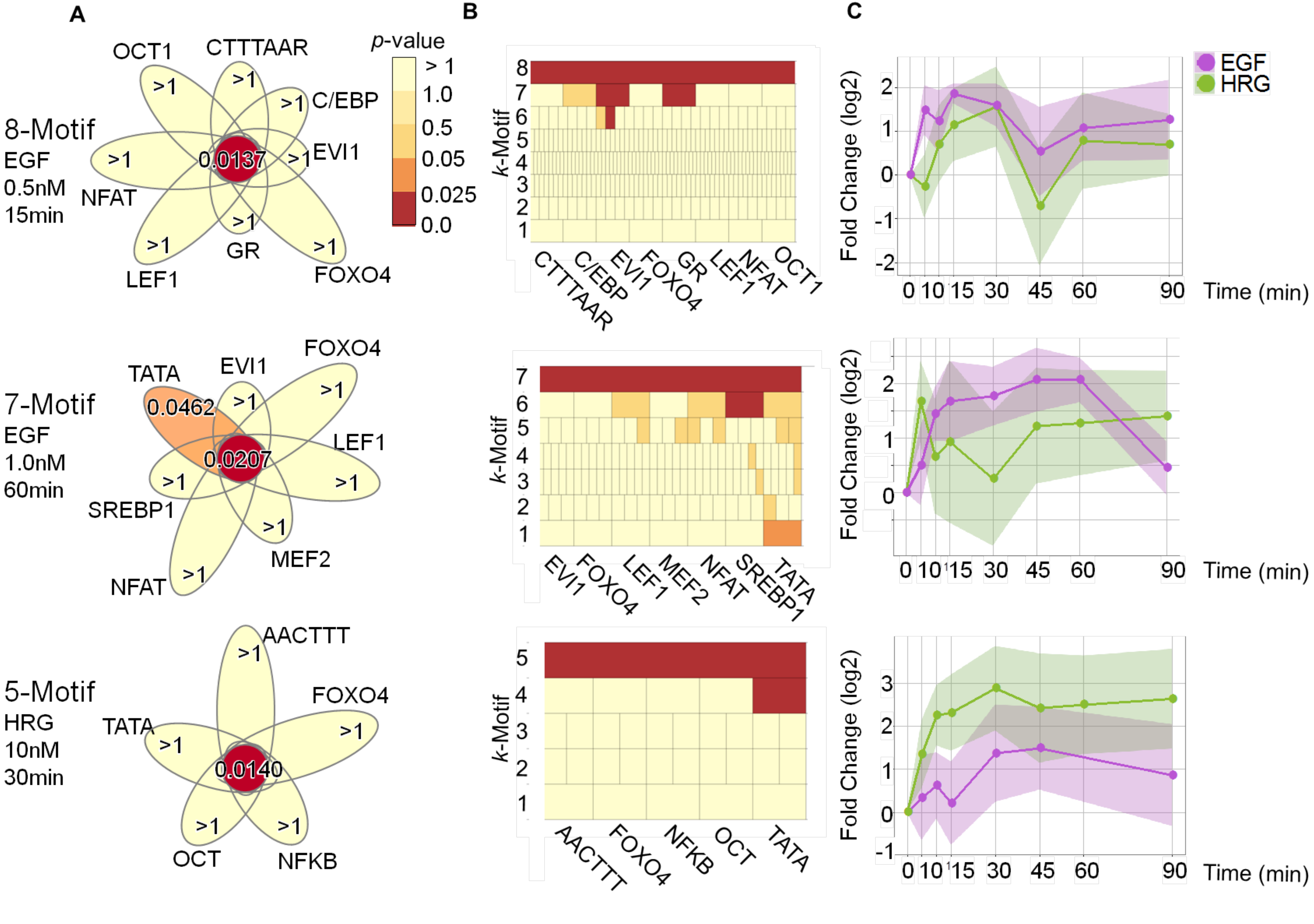
Combination	LAMP (≦102)	Bonferroni (≦4)
	K = 303	K = 4,426,528
HSF1	4.41E-24	6.44E-20
MSN2	3.73E-11	5.45E-07
MSN4	0.00053	> 1
SKO1	0.00839	> 1
SNT2	0.0192	> 1
PHD1, SUT1, SOK2, SKN7	0.0272	> 1

Red：significant


A
Rank of gene expression
p-value
Up
Down
PHD1
> 1
SUT1
> 1
SOK2
0.666
SKN7
0.111
PHD1, SUT1, SOK2, SKN7
0.0272
HAP4
GAT2
MSN4
MGA1
GID8
YNL179C
RHO5
B
PHD1
>1
SKN7
0.111
0.0272
SUT1
>1
0.666
SOK2
p-value
> 1
1.0
0.5
0.05
0.0

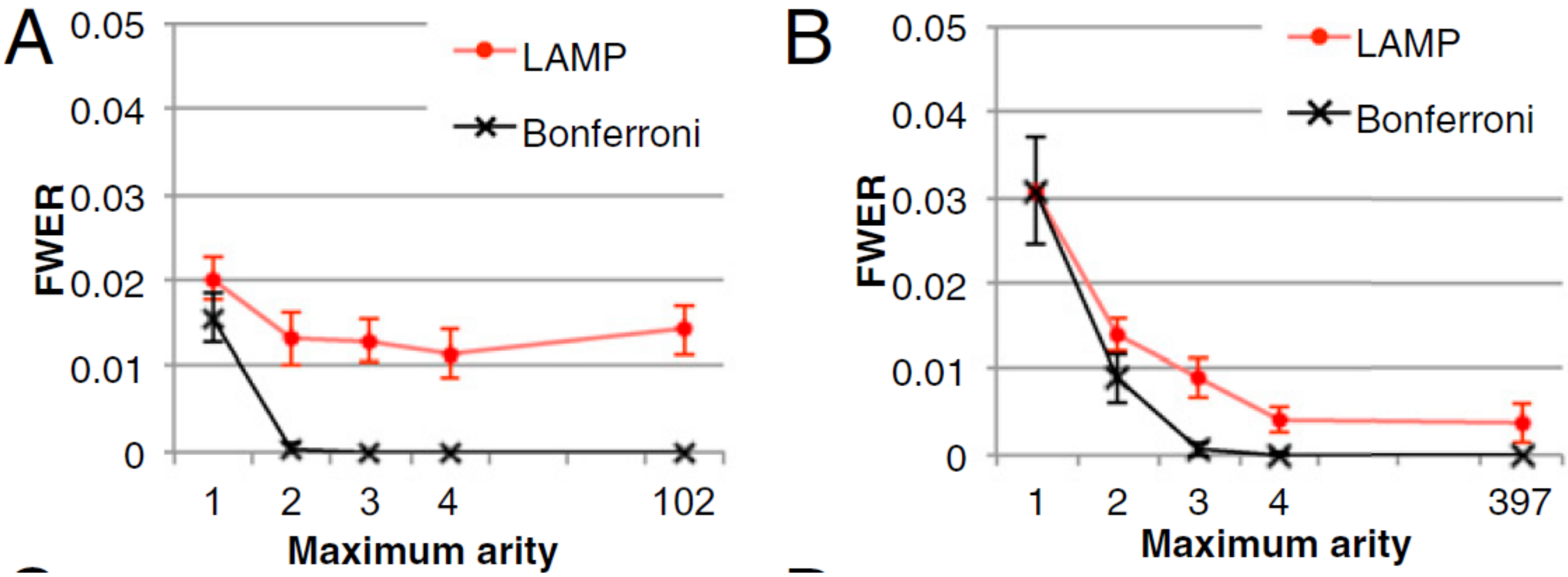
Application to MCF7 human breast cancer cells (GSE6462)

- Treated with epidermal growth factor (EGF) or heregulin (HRG)
  - 0.1, 0.5, 1, 10 nM
- Expression measured 5, 10, 15, 30, 45, 60 mins after
- Motifs taken from MSigDB
- 397 motifs, Approx. 12000 genes
- LAMP K= 1,174,108 ～ 3,750,336
- Bonferroni K=1.4 x 10^{16} (maximum arity =8)

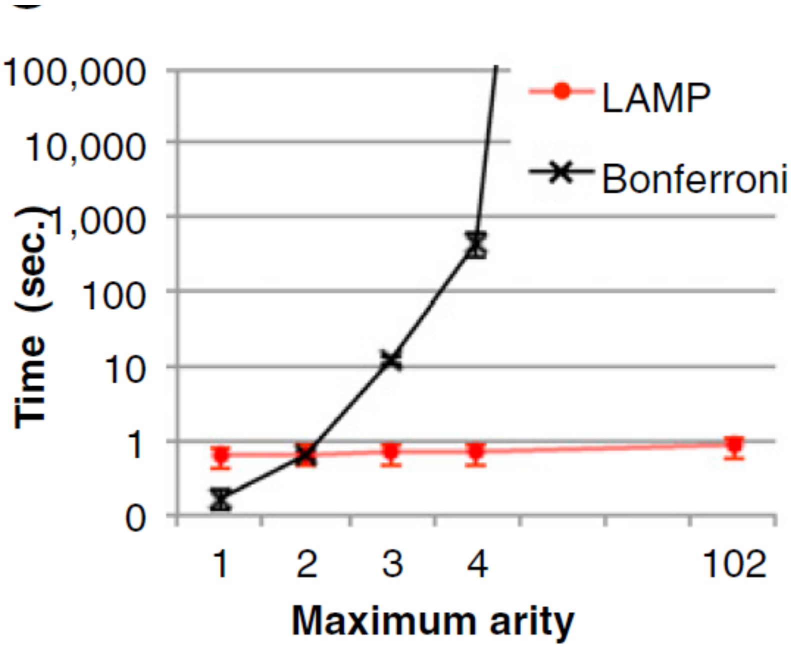

A
B
C
p-value
> 1
1.0
0.5
0.05
0.025
0.0
8-Motif
EGF
0.5nM
15min
OCT1
CTTTAAR
C/EBP
EVI1
NFAT
GR
LEF1
FOXO4
0.0137
>1
7-Motif
EGF
1.0nM
60min
TATA
EVI1
FOXO4
LEF1
SREBP1
MEF2
NFAT
0.0462
0.0207
5-Motif
HRG
10nM
30min
AACTTT
FOXO4
TATA
OCT
NFKB
0.0140
k-Motif
8
7
6
5
4
3
2
1
CTTTAAR
C/EBP
EVI1
FOXO4
GR
LEF1
NFAT
OCT1
EVI1
FOXO4
LEF1
MEF2
NFAT
SREBP1
TATA
AACTTT
FOXO4
NFKB
OCT
TATA
EGF
HRG
Fold Change (log2)
Time (min)
0
10
15
30
45
60
90
-2
-1
0
1
2
3

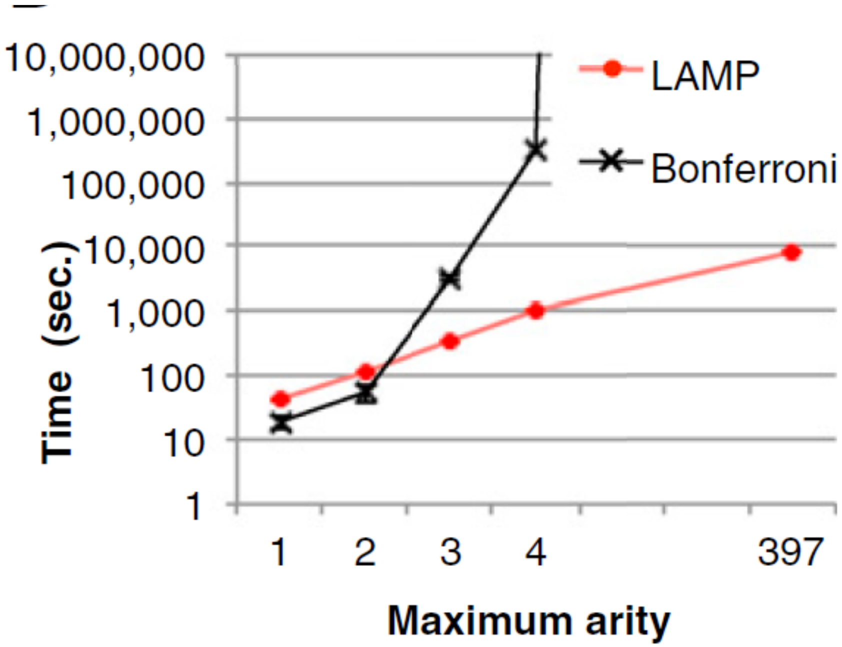
Empirical FWER

- LAMP’s FWER is much closer to the designated value 0.05



Computational Time





Concluding Remarks (LAMP)

- LAMP is much more sensitive than Bonferroni, whereas FWER is strictly kept under threshold
- Immediately applicable to sequences, trees and graphs
- Minimum p-value must be strictly positive
  - LAMP cannot be applied to t-test
  - Statistical tests with "robustness" can be combined with LAMP

Everything goes “Personal”

- Reference genome → Personal genomes
- Cell population → Single cell measurement
- Similar to current status of data mining
  - Analyze average behavior of customers （Obsolete!）
  - Focus on difference among customers
- You do not need sophisticated algorithms for studying averages
- Knowledge discovery tools to the center stage of sciences