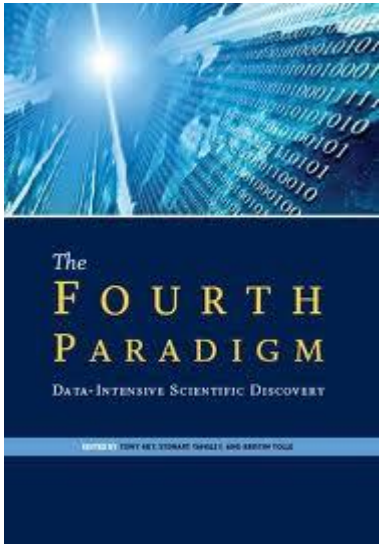
信頼できる科学のための統計検定手法（20min）

津田　宏治

東京大学大学院新領域創成科学研究科
メディカル情報生命専攻　教授

(兼任）
理研 革新知能統合研究センター
物材機構

第4のパラダイム：データ中心科学

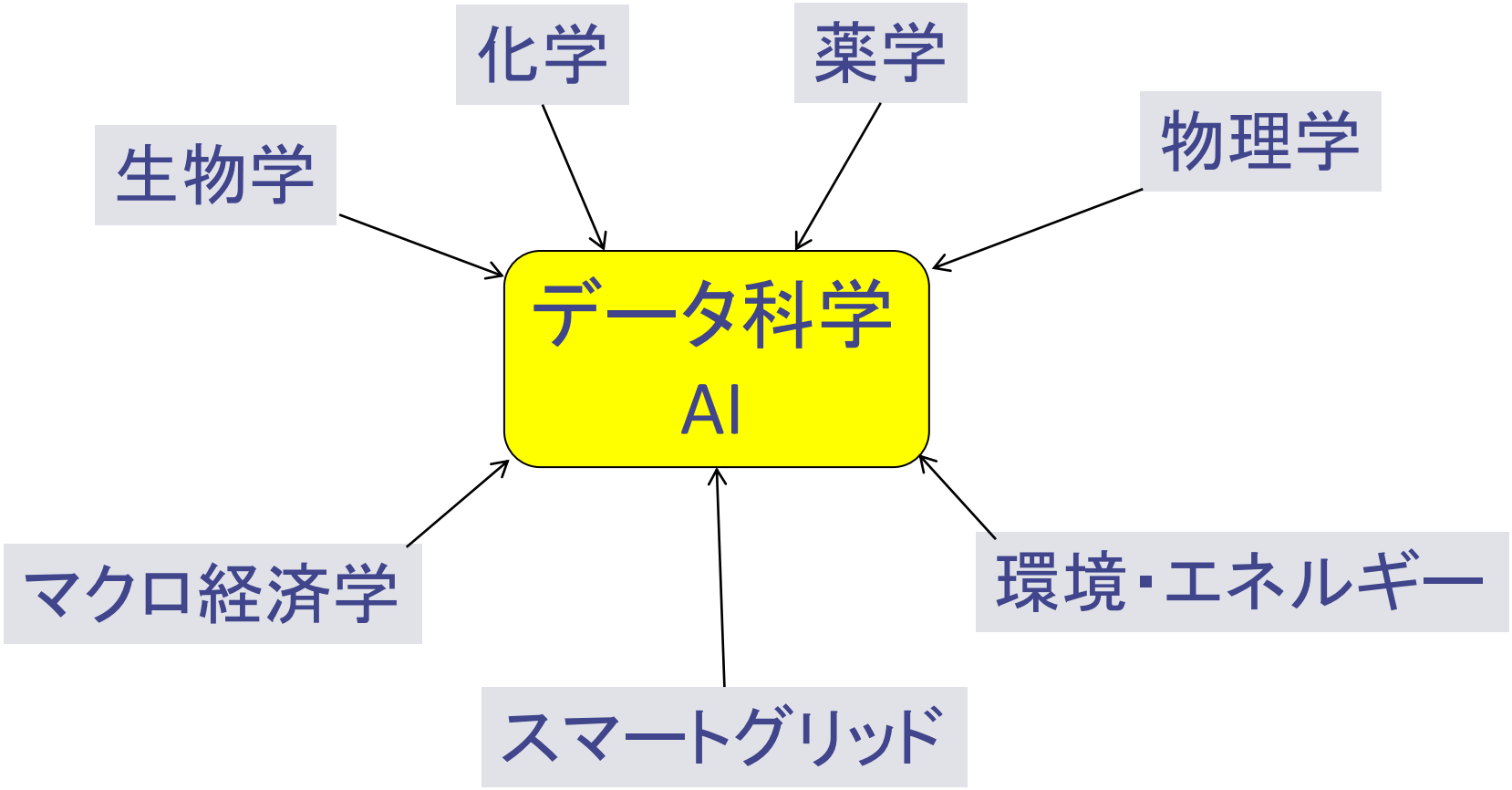


- 経験科学・理論科学・計算科学（シミュレーション）に代わる21世紀の科学

- 従来：　直感・経験に基づく仮説生成　⇒　実験に基づく検証

- 第4：　網羅的データからの**知識発見**による仮説生成 (AI)⇒実験に基づく検証

様々な分野からの要望が、データ科学に集中： 深刻な人材不足



データ解析結果の信頼性に関する思考実験

- ゲノムデータ解析で原因遺伝子を探る
- AさんとBさんが同じデータを解析：結果が違う
- どちらを信じるべきか？

- 有名な方を信じる？　ブランド

- 従来の知識に合っている方を信じる？　進歩は？

FT Magazine

Home World Companies Markets Global Economy Lex

Arts Magazine Food & Drink House & Home Lunch with FT Style Books Pursuits

March 28, 2014 11:38 am

Big data: are we making a big mistake?

By Tim Harford

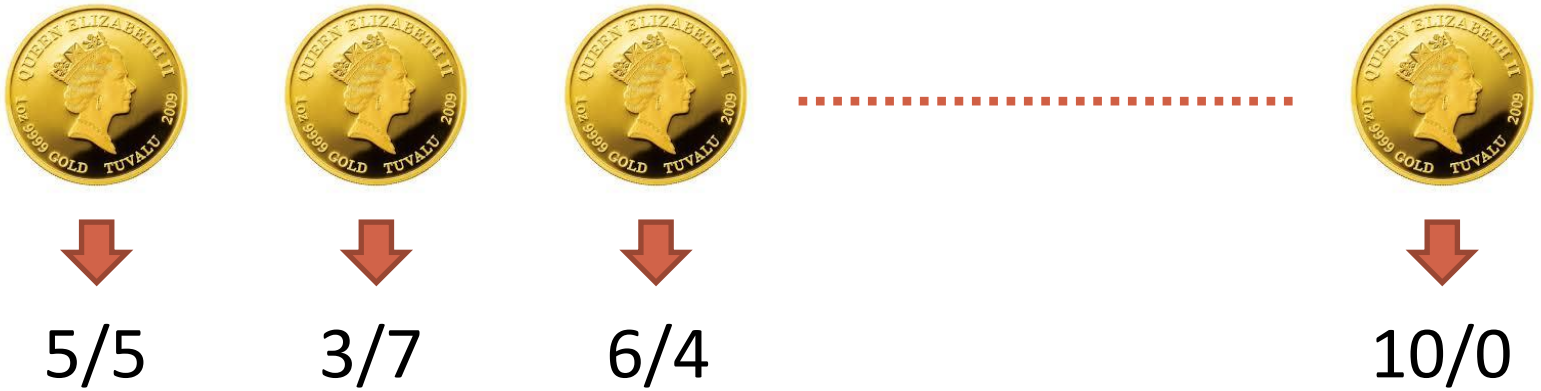
Big data is a vague term for a massive phenomenon that has rapidly become an obsession with entrepreneurs, scientists, governments and the media

一部の訳

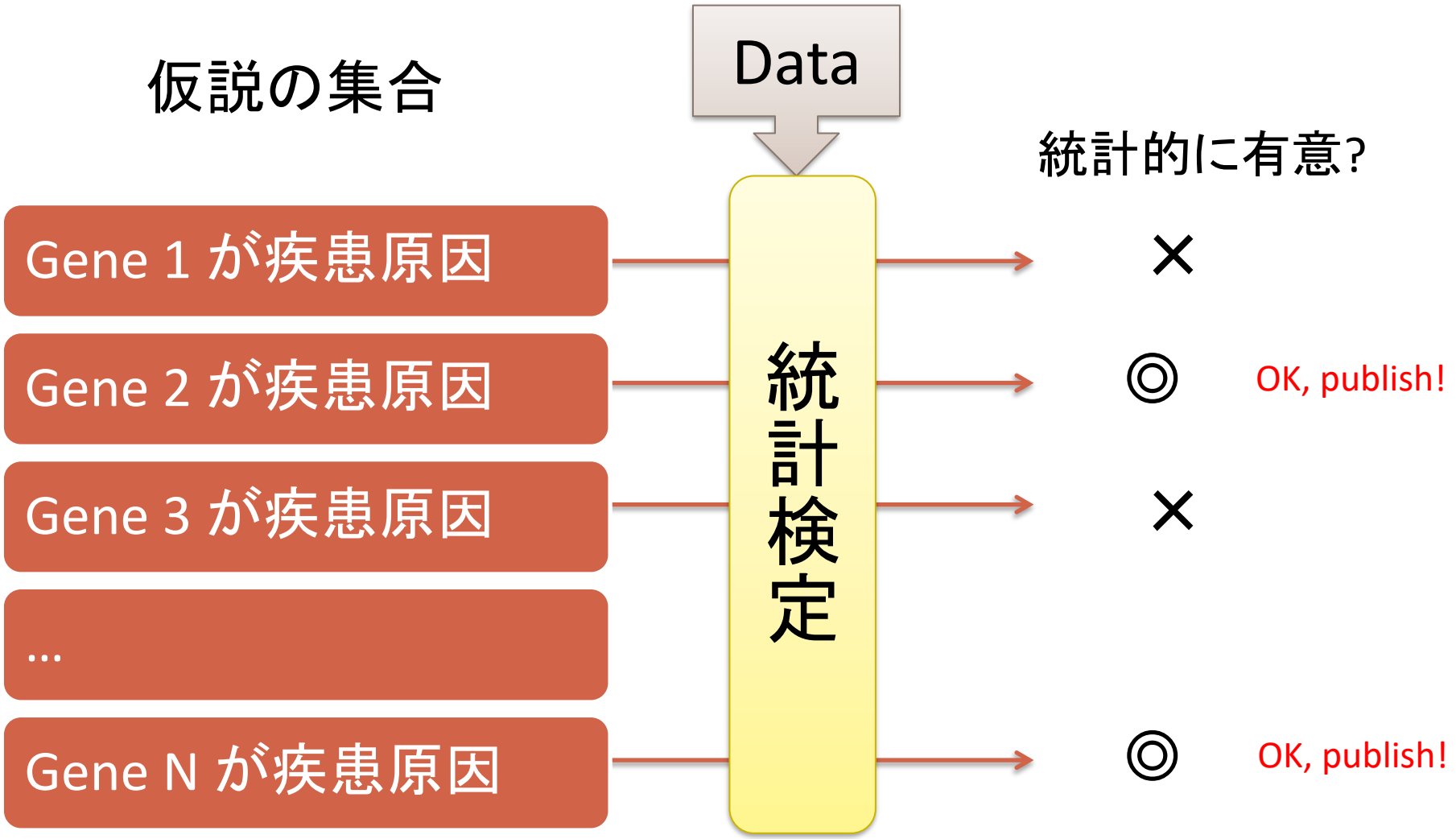
- 2005年、疫学者J. Ioannidisは、“Why Most Published Research Findings Are False”という論文を発表した。この論文は、深刻な問題を刺激的に取り上げたことで有名になった。この論文の基礎は、統計学者が「多重検定」と呼ぶ問題である

異常なコインを見つけよう

- 1000個の正常なコインを10回ずつ振る
- 10回全部表がでるものがある確率は、62.4%
- 全くのランダムでも、「それっぽい」結果が出る
- 多くの仮説を同時に扱うには注意が必要



生命科学の出版システム



生物学における再現性危機

- 論文に載っている統計的に有意な結果が再現できない
- Bayer: 67の結果のうち、43が再現不能
- Amgen: 53の結果のうち、47が再現不能

J. Ioannidis (Stanford)

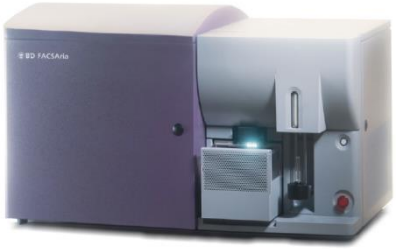
高次元データを扱う際には、偽陽性の確率が大幅に増加する (Lancet, 2014)

統計検定における次元数の呪い

New Machines

説明変数の飛躍的増加

サンプル数は、あまり増えない



偽陽性の確率増

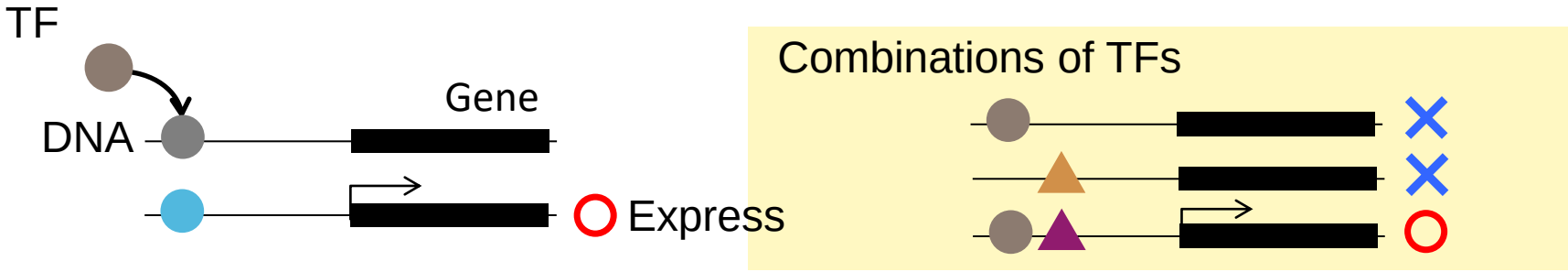
→従来の統計検定では、

有意性基準が過剰に厳しくなる

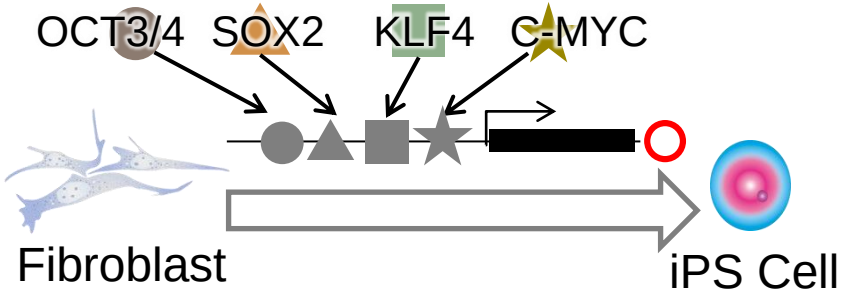
→科学的発見の減少

転写因子の組合せ効果

- 複数の転写因子の結合によって、下流の遺伝子の発現が引き起こされる

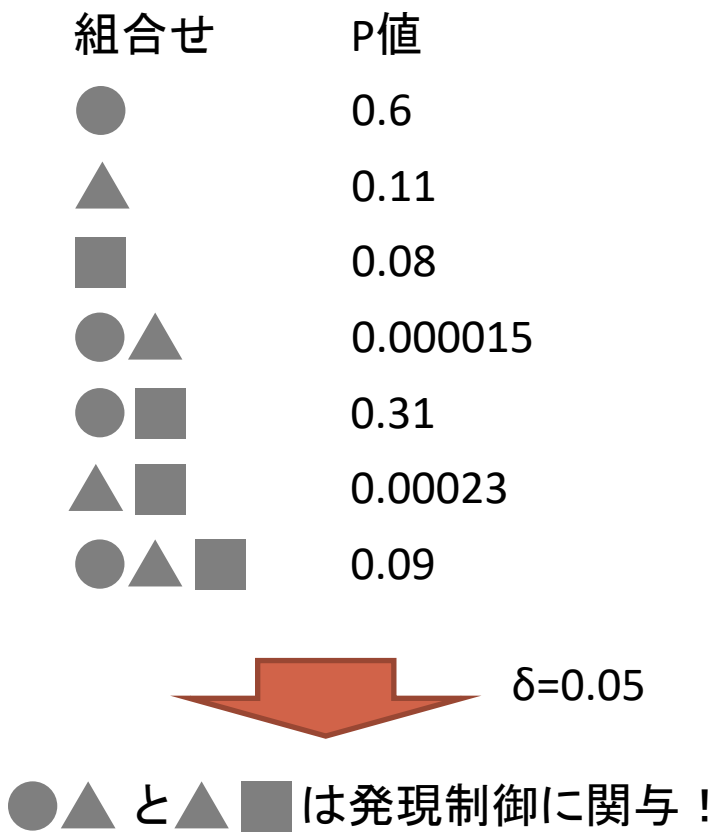


Example: Yamanaka Factor（K. Okita *et al.*, Nature, 2007）



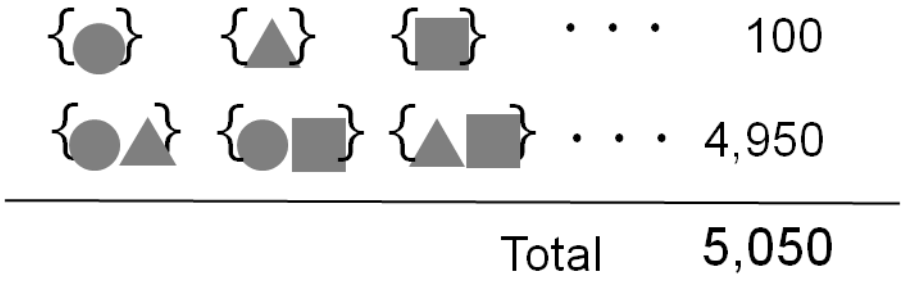
多重検定 (Multiple Testing)

- 二つまでの転写因子の組合せを考え、各々のP値を計算
- 閾値より小さいものを発見として採用
- P値＝誤発見の確率を評価したもの



Bonferroni補正

- 100個転写因子があるとし、二個の組合せまで見た場合の検定数



- Bonferroni: P値に検定値を掛けて閾値より低ければ有意
- 補正された閾値　$\delta=0.05/5050 = 9.9 \times 10^{-6}$
- Bonferroniは保守的すぎる

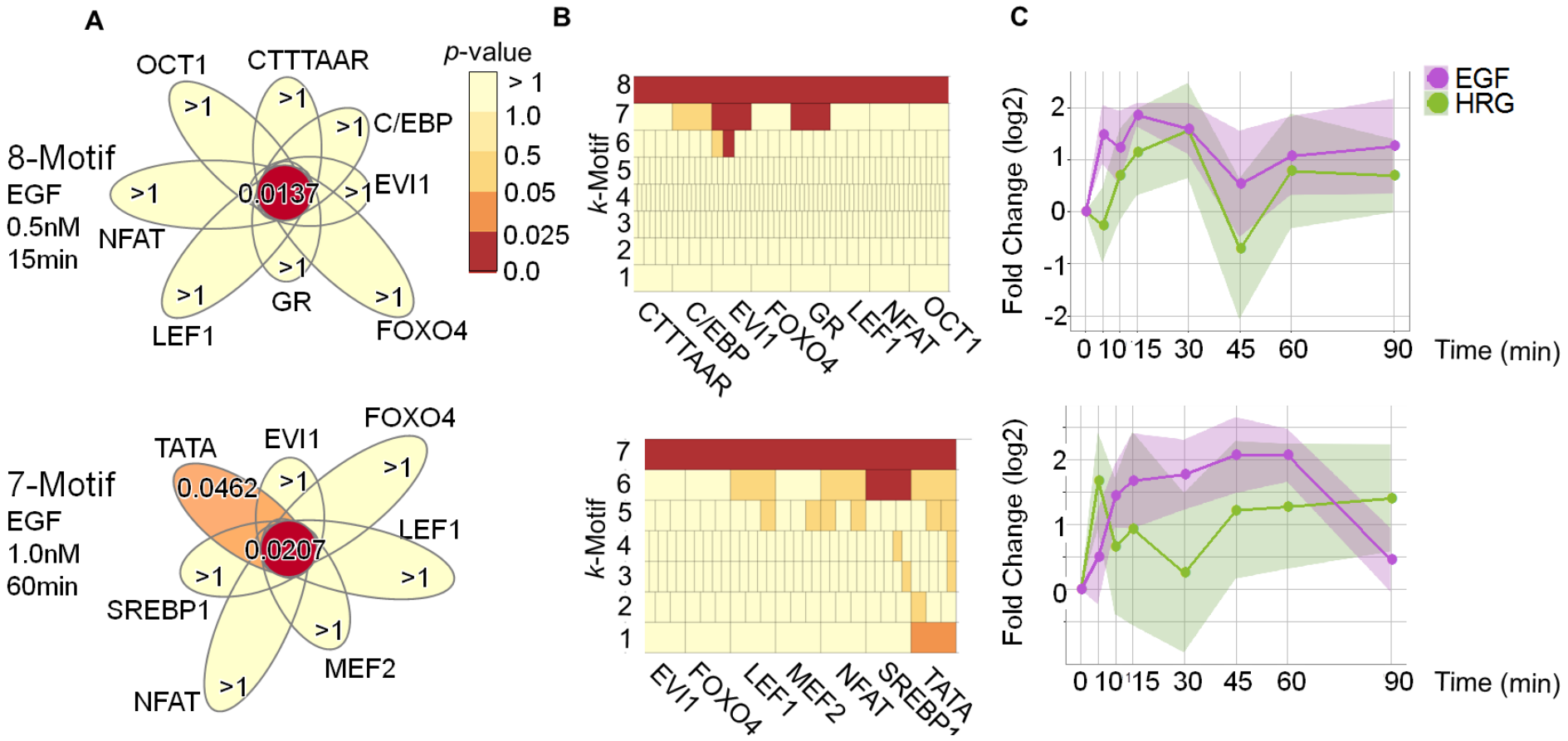
新手法

Limitless Arity Multiple testing Procedure (LAMP)

- データに表れる頻度が低い組合せは、理論的に補正係数に含める必要がない
- 有効な組合せを高速に数え上げることで、Bonferroni補正に比べて、補正係数を格段に低く抑えられる
- イースト菌、乳がん細胞株(MCF7)において、新規の転写因子組合せを発見

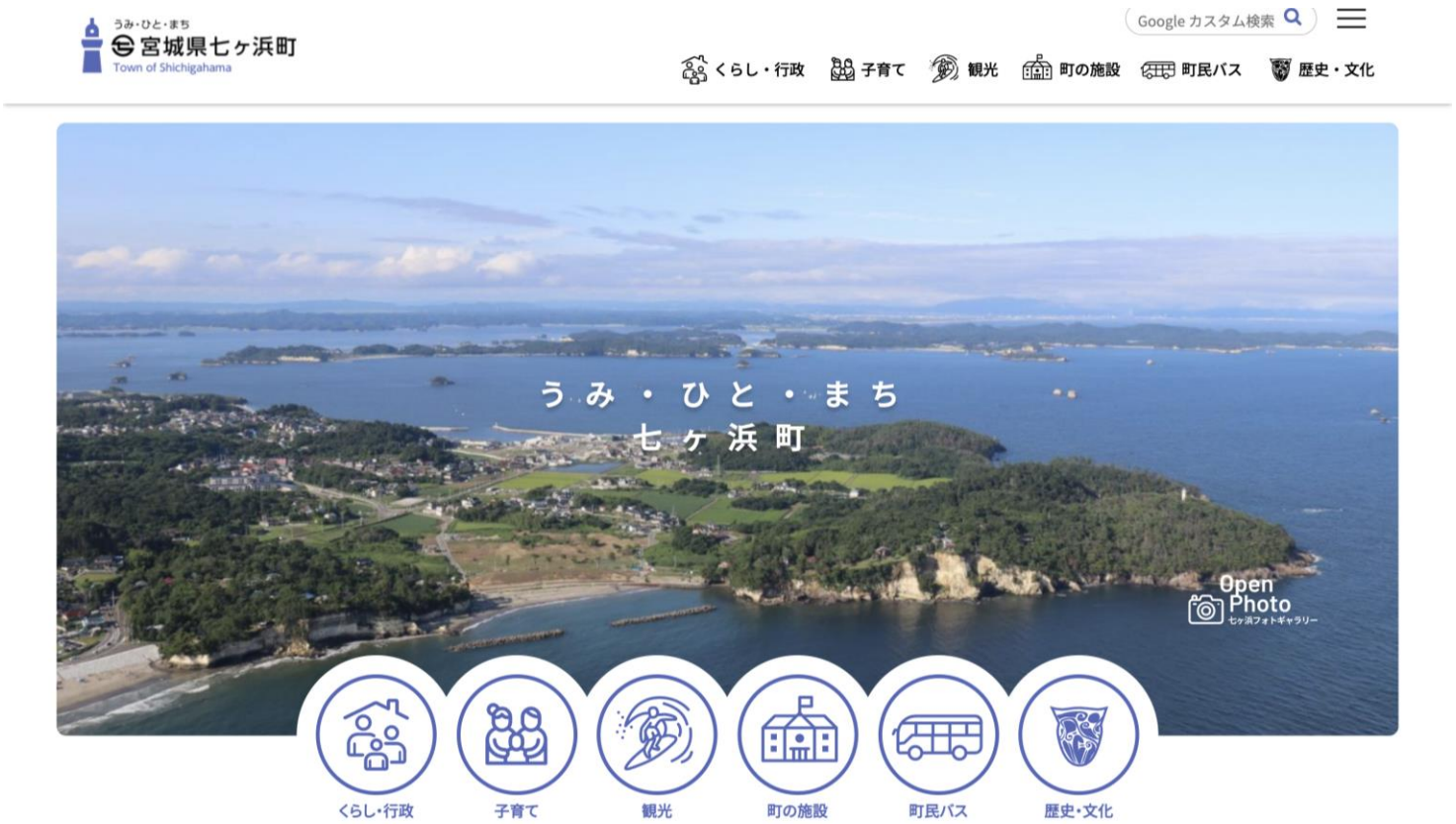
Terada et al., PNAS, 2013.

Application to MCF7 human breast cancer cells (GSE6462)



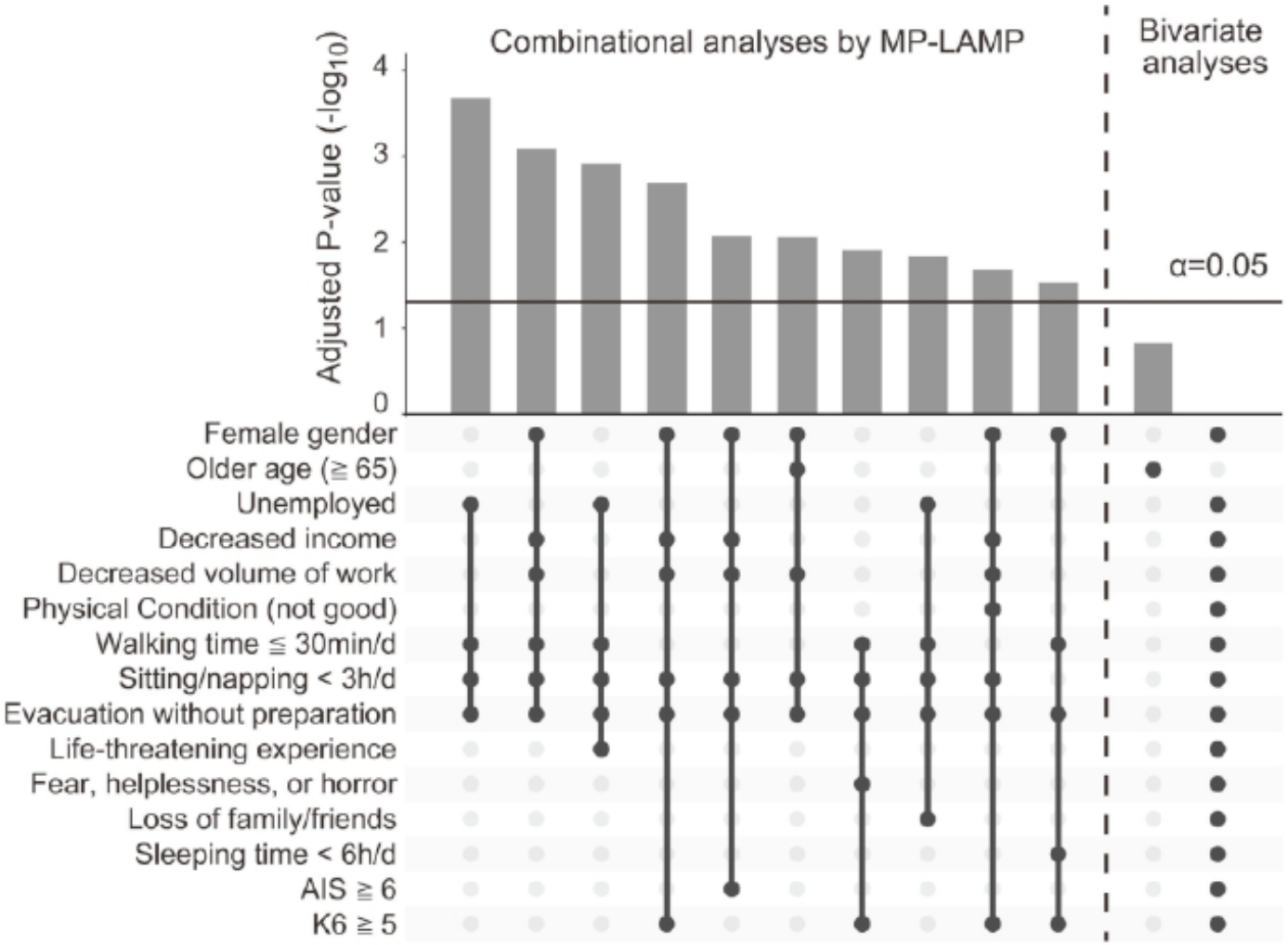
震災後PTSDのLAMPによる要因解析
(東北大との共同研究)

被災者の状態・行動が、PTSDからの回復に及ぼす影響の研究



Takahashi et al., Scientific Reports, 2020.

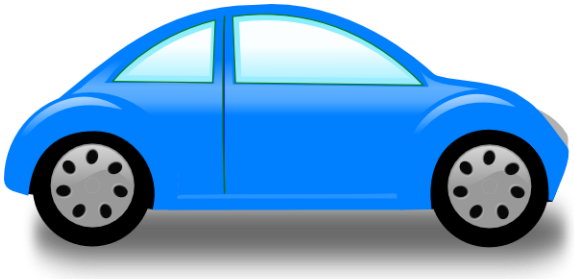
- 従来法では、年齢のみが有意
- LAMPでは、それに加え、歩く時間、失業、健康状態なども有意



おわりに

- 機械学習の目的「推論の自動化」「原因究明」のうち、後者の進展は遅い
- 統計学の強化による、信頼できるデータ科学の開発は最重要課題

ENGINE:
大量のデータを産生する実験機器



BRAKE:
統計学

LAMPによる組み合わせ効果の検定

- 10^5個のSNP（遺伝子変異)を扱う
- SNPの三つの組み合わせ
- Bonferroni factor ~ 10^15
- 実質的に、Bonferroni検定では、統計的有意にならない

- **疑問**：一度もデータに出てこない組み合わせが大半。これもBon-factorに入れる？

LAMPでは、Tarone検定を利用する

- 頻度が小さいイベントは、Bon-factorから除去可能 (Biometrika, 1990)
- 頻度閾値の最適値を、itemset mining-likeなアルゴリズムで求める