Centre for Molecular Medicine
and Therapeutics

Multiple Testing

Bernard Ng

Department of Statistics, UBC

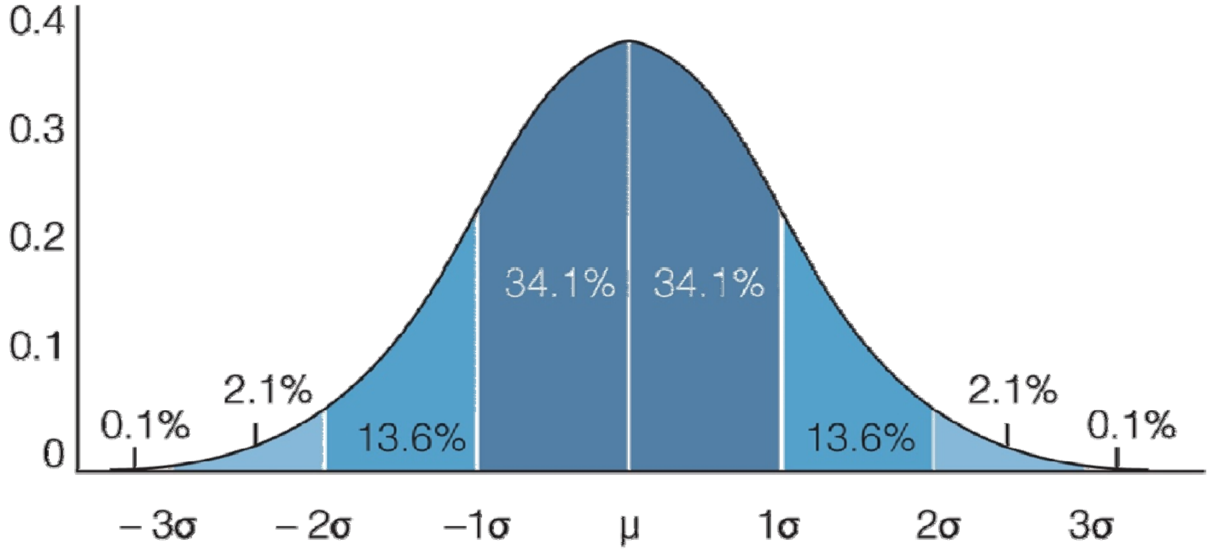
www.cmmt.ubc.ca

Outline

- Single Hypothesis Testing
- Multiple Hypothesis Testing
- Bonferroni Correction
- Step-up Procedure
- False Discovery Rate Correction
- Max-t Permutation Test
- Recent Topics
- Neuroimaging Applications

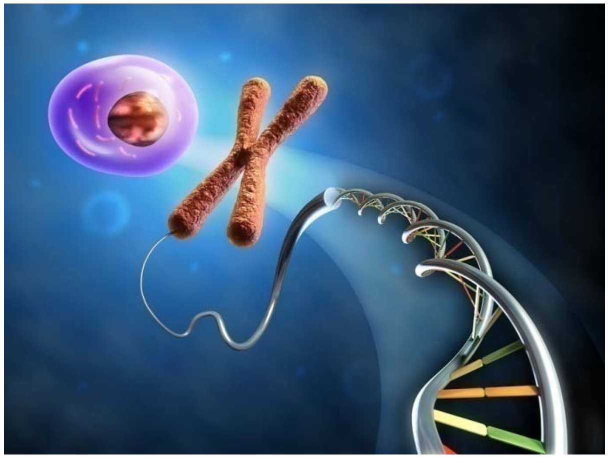
Single Hypothesis Testing

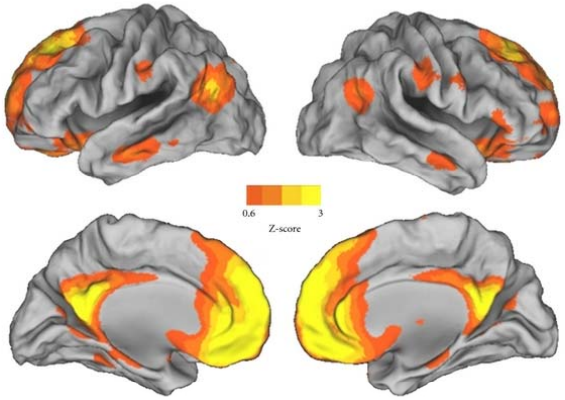
- In the past, a handful of hypotheses with a lot of samples, e.g. census data.
- H_0: $x = \mu$ vs. H_A: $x \neq \mu$
  - Are girls smarter than guys? => two sample t-test
  - Do last minute studying affect scores? => regression
- Generate statistics e.g. z, t, F, …
- p < 0.05



Multiple Hypothesis Testing

- Nowadays, a lot of (unplanned) hypotheses but not enough samples (for medical research)





10^6

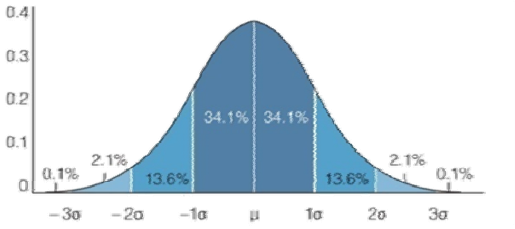
10^3 genomics

10^5

10^2 Neuroimaging

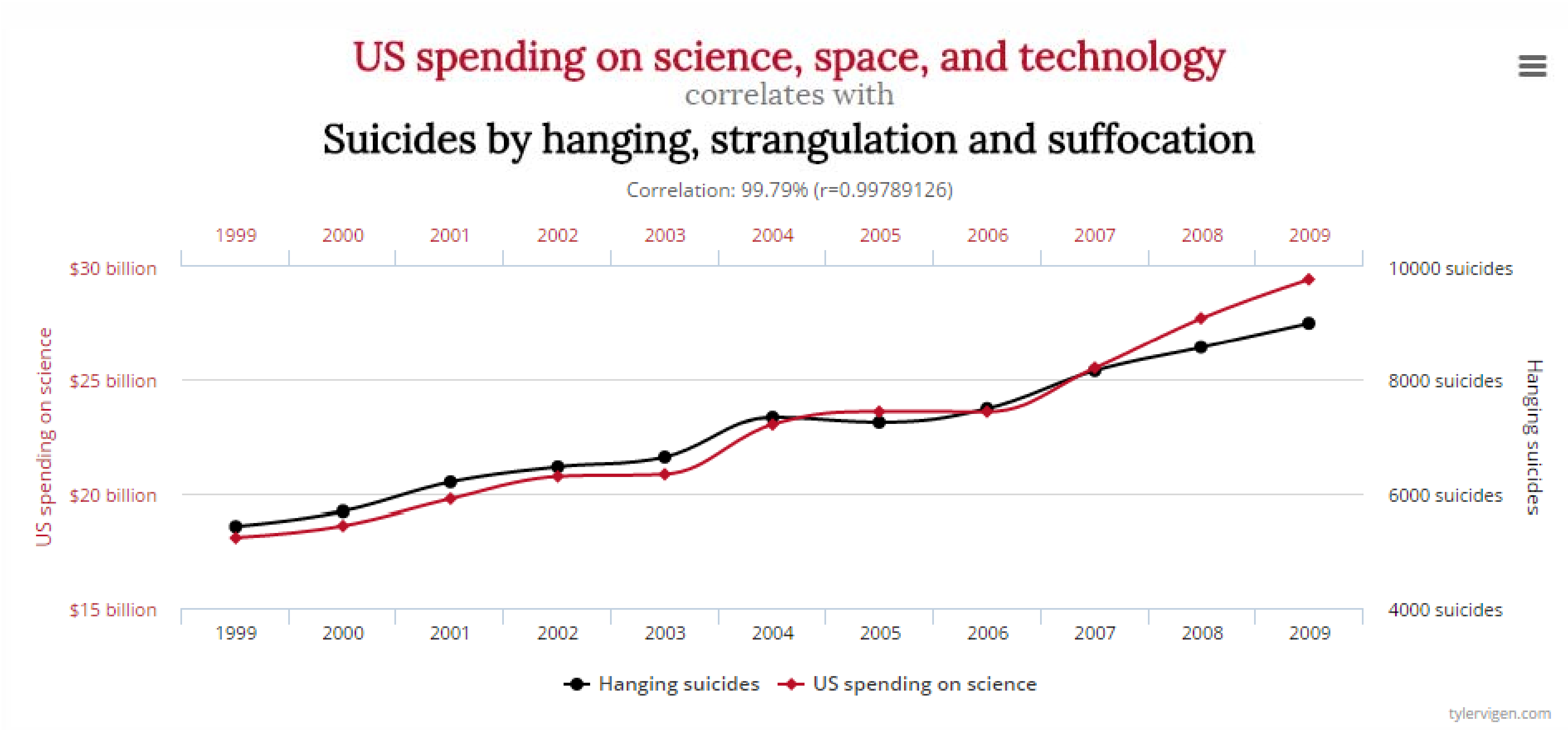
Multiple Testing Problem

- P(not rejecting 1 hypothesis) = 1- α
- P(not rejecting all *n* hypotheses) = (1- α)^n
- $\alpha_{\mathrm{FWER}} = 1\text{-}(1\text{-}\alpha)^n$
  - – $\alpha = 0.05$, $n = 10$: $\alpha_{\mathrm{FWER}} = 0.4013$
  - – $\alpha = 0.05$, $n = 10^2$: $\alpha_{\mathrm{FWER}} \approx 1$
- So if e.g. run 100 experiments, then $\alpha_{\mathrm{FWER}}\cdot100$ of them would have ≥1 hypothesis falsely rejected.
- Intuition from ML perspective is that the more we sample the variable space, the more "likely" we will get some "extreme" samples.



MATLAB Demo

Why Important?

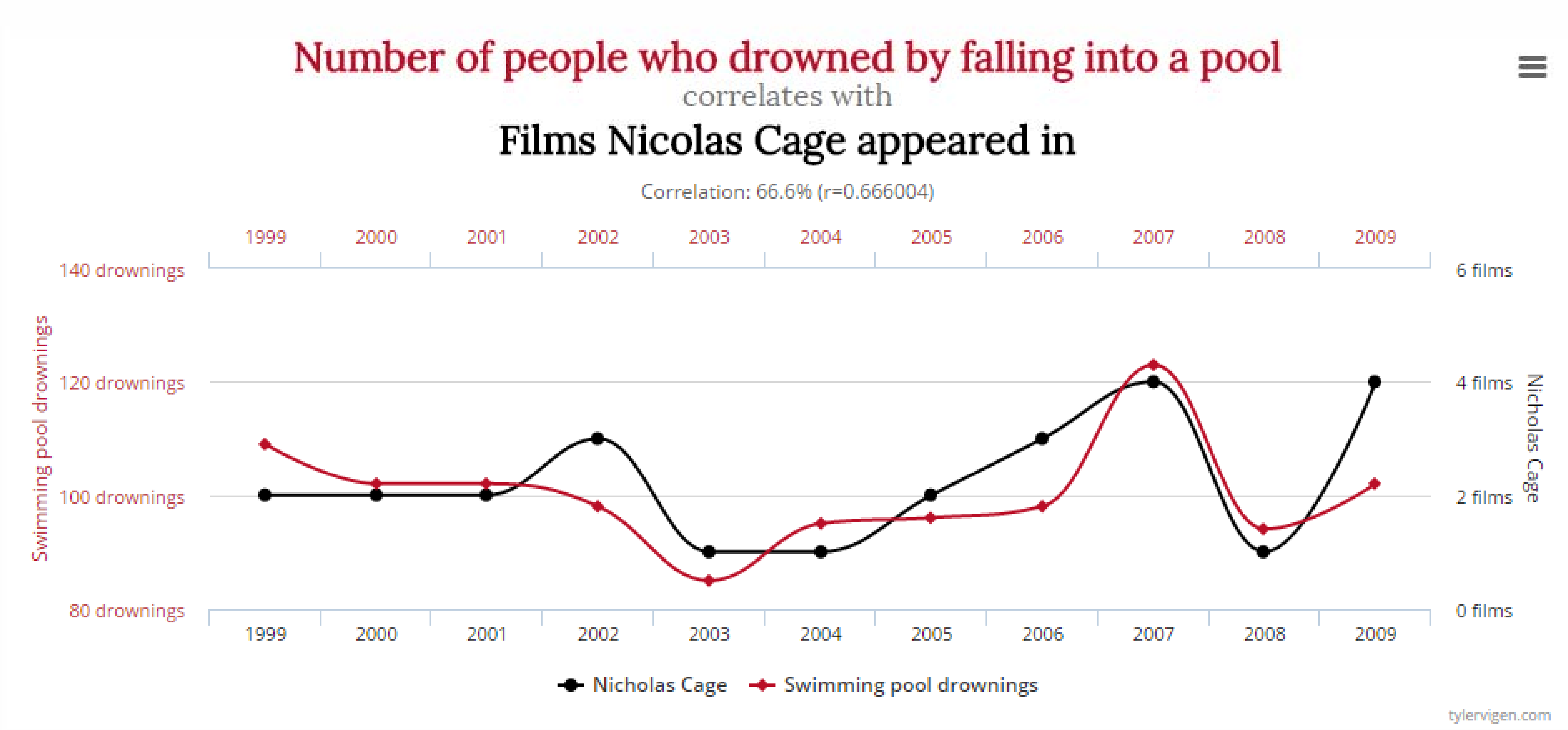


US spending on science, space, and technology
correlates with
Suicides by hanging, strangulation and suffocation

Correlation: 99.79% (r=0.99789126)

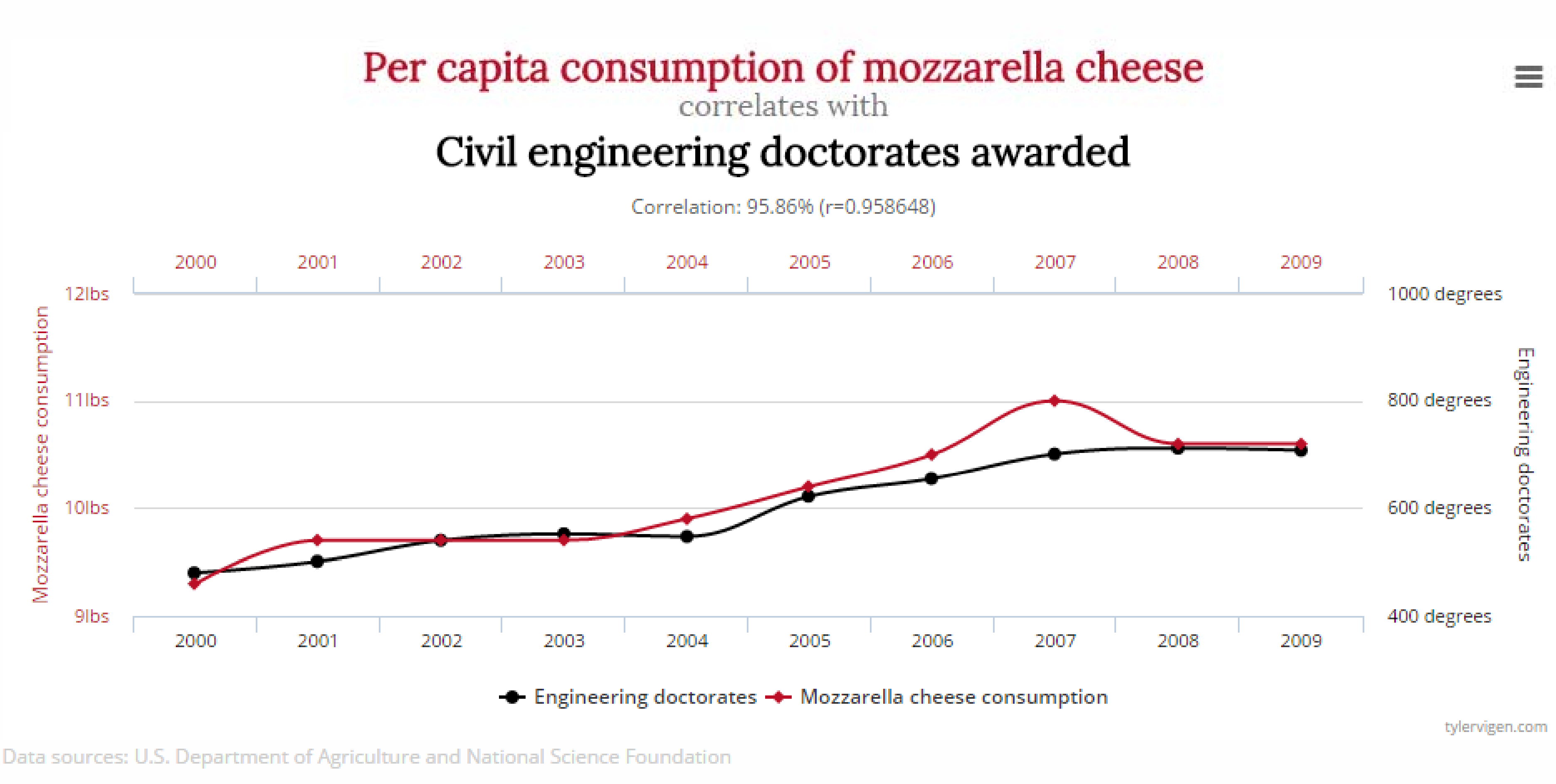
Data sources: U.S. Office of Management and Budget and Centers for Disease Control & Prevention

Why Important?



Data sources: Centers for Disease Control & Prevention and Internet Movie Database

Why Important?



Per capita consumption of mozzarella cheese

correlates with

Civil engineering doctorates awarded

Correlation: 95.86% (r=0.958648)

Data sources: U.S. Department of Agriculture and National Science Foundation

Why Important?

tylervigen.com

about | twitter | email | subscr

Spurious correlations

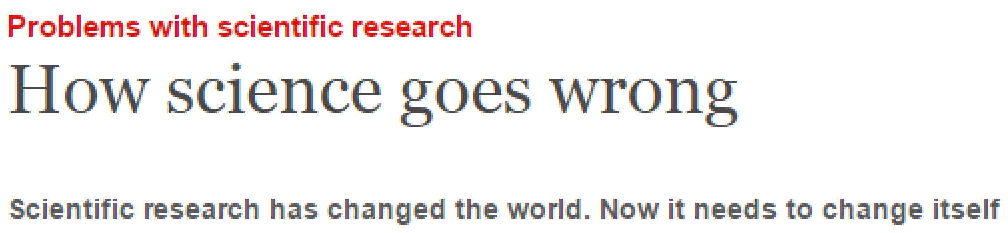
Now a ridiculous book!

- Spurious charts
- Fascinating factoids
- Commentary in the footnotes

Amazon | Barnes & Noble | Indie Bound

http://tylervigen.com/spurious-correlations

Why Important?



Problems with scientific research

How science goes wrong

Scientific research has changed the world. Now it needs to change itself

Last year researchers at one biotech firm, Amgen, found they could reproduce just **six of 53** “landmark” studies in cancer research. Earlier, a group at Bayer, a drug company, managed to repeat just a **quarter of 67** similarly important papers. A leading computer scientist frets that **three-quarters** of papers in his subfield are bunk. In 2000-10 roughly **80,000 patients** took part in clinical trials based on research that was **later retracted** because of mistakes or improprieties.

Notations and Terminologies

🟦 unobserved

		Predicted		
		True	**False**	
Ground Truth	**True**	U True Negative	V False Positive	n_0
	False	T False Negative	S True Positive	$n-n_0$
		n-R	R	n

Sensitivity = $S/(n-n_0)$
Specificity = U/n_0

Note: All terms are defined wrt null, e.g. if the ground truth is that the null hypothesis is True, but we claimed it is false, then we have a false positive.

Bonferroni Correction

Procedures

- Recall $\alpha_{\mathrm{FWER}} = 1\text{-}(1\text{-}\alpha)^n$
- Set $\alpha = 1\text{-}(1\text{-}\ \alpha_{\mathrm{FWER}})^{1/n} \approx 1\text{-}(1\text{-}\alpha_{\mathrm{FWER}}/n) = \alpha_{\mathrm{FWER}}/n$

Examples

- $\alpha_{\mathrm{FWER}} = 0.05$ and $n = 10$, needs $\alpha = 0.05/10 = 0.005$
- $\alpha_{\mathrm{FWER}} = 0.05$ and $n = 10^6$, needs $\alpha = 0.05/10^6 = 5\times10^{-8}$

Properties

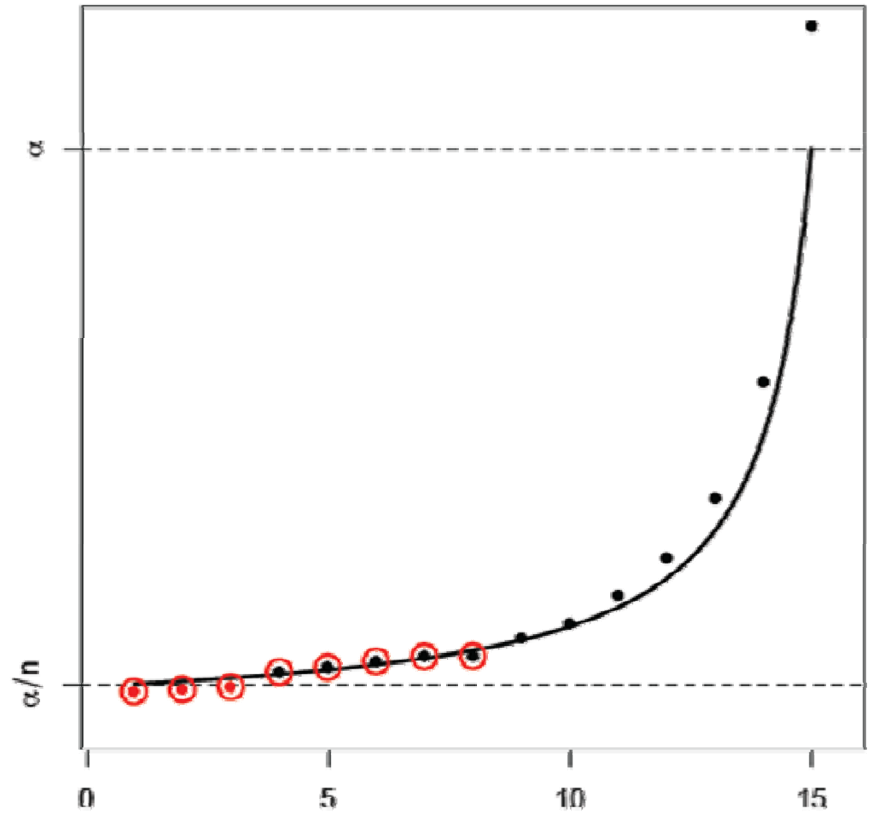
- Ctrls FWER = P(V≥1) in *strong* sense.
- Can handle correlated hypotheses.
- Very stringent

		Predicted		
		True	**False**	
Ground Truth	**True**	U	V	n_0
	False	T	S	$n\text{-}n_0$
		n-R	R	n

MATLAB Demo

Step-up Procedure

- aka Hochberg’s procedure
- Sort p in descending order
- $p(i) \le \alpha/(n-i+1)$
  1. $p(n) \le \alpha/(n-n+1)$
  2. $p(n-1) \le \alpha/(n-(n-1)+1)$
  3. …
- Controls FWER in strong sense
- Holm’s Step-down procedure uses same threshold but less sensitive.



MATLAB Demo

False Discovery Rate

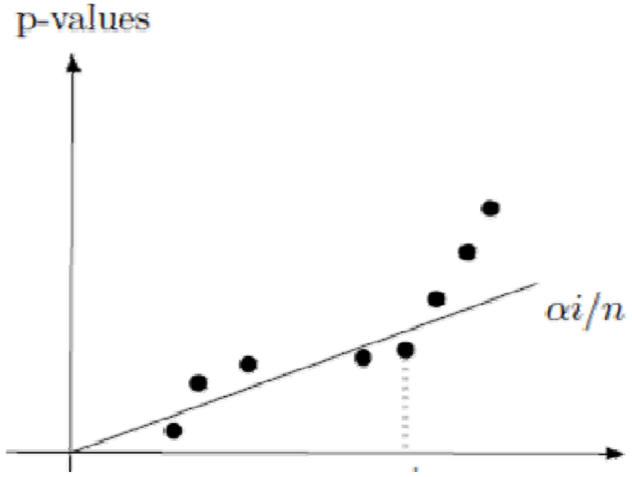
Idea

- Benjamini & Hochberg, 1995
- Recall FWER = $P(V \geq 1)$
- Fdp = $V/\max(R,1)$
- But V unobserved, so: FDR = E(Fdp)

		Predicted		
		True	**False**	
Ground Truth	**True**	U	V	n_0
	False	T	S	$n-n_0$
		n-R	R	n

Procedures

- Sort p in ascending order.
- Find $i_0 = \max i$ s.t. $p(i) \leq i \cdot q/n$



False Discovery Rate

Properties

- If hypotheses are independent, then FDR < q for *all* configurations of hypotheses.
- If data are Gaussian and hypotheses are positively correlated, i.e. $\Sigma_{ij} \geq 0$, then FDR < q.
- If hypotheses are correlated,

  $FDR < q \cdot (\log(n)+0.577)$

  $\Rightarrow p(i) < i \cdot q/n \,/\, (\log(n)+0.577)$

  BUT $i = 1$, $p(i) < q/n \,/\, (\log(n)+0.577) < q/n$

		Predicted		
		True	**False**	
Ground Truth	**True**	U	V	n_0
	False	T	S	$n-n_0$
		n-R	R	n

False Discovery Rate

Properties

- $n=n_0$, then FDR = FWER since:

  $V=R$, so $V=0$ iff Fdp = 0 and $V\geq 1$ iff Fdp=1,

  i.e. Fdp = indicator variable,

  thus $P(V\geq 1)$ = E(Fdp) = FDR => ctrls FWER *weakly*
- $n<n_0$, controlling FWER controls FDR
- Adaptive: 5/100, 50/1000, …
- More sensitive than Hochberg

  $i/n \,/\, (1/(n-i+1)) = i\cdot(1-(i-1)/n)$

  e.g. $i = n/2$ => ~n/4 gain

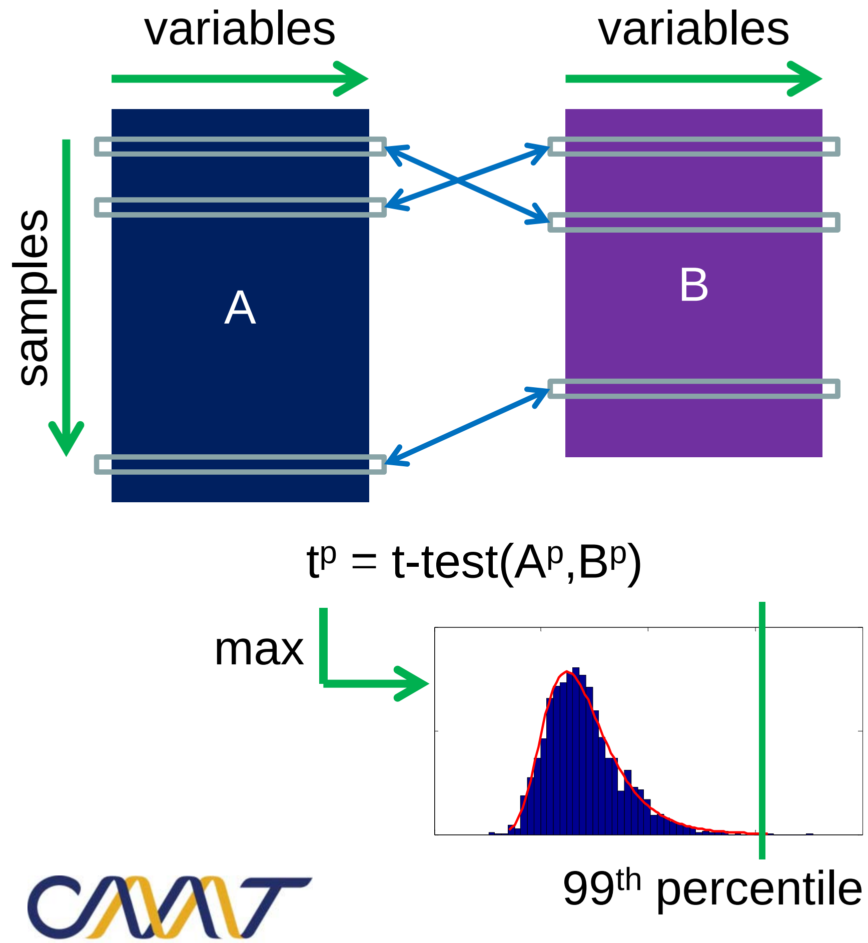
		Predicted		
		True	**False**	
Ground Truth	**True**	U	V	n_0
	False	T	S	$n-n_0$
		n-R	R	n

MATLAB Demo

Max-t Permutation Test

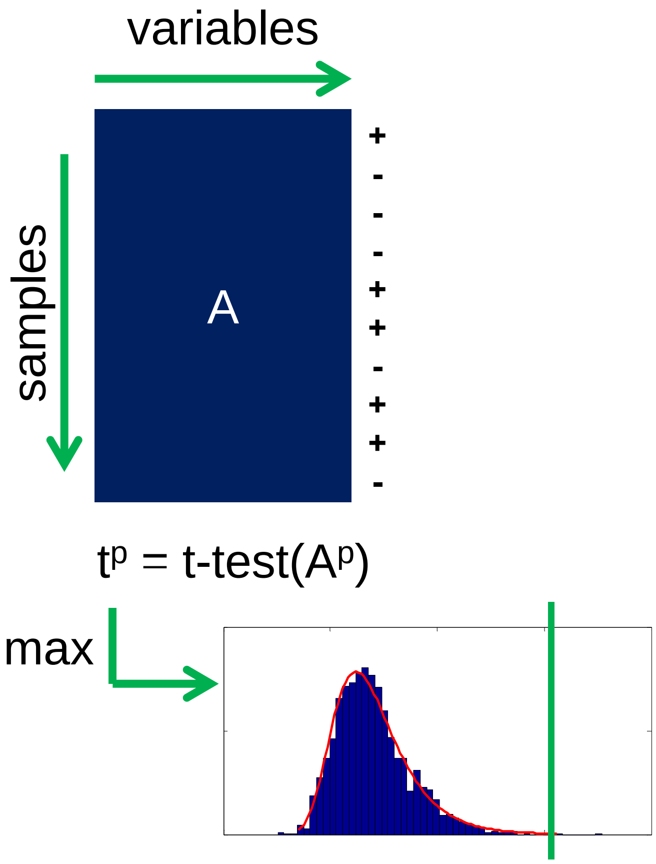
- Strongly controls FWER under any kind of dependence structure under certain assumptions.

Two Sample



t^p = t-test(A^p,B^p)

One Sample

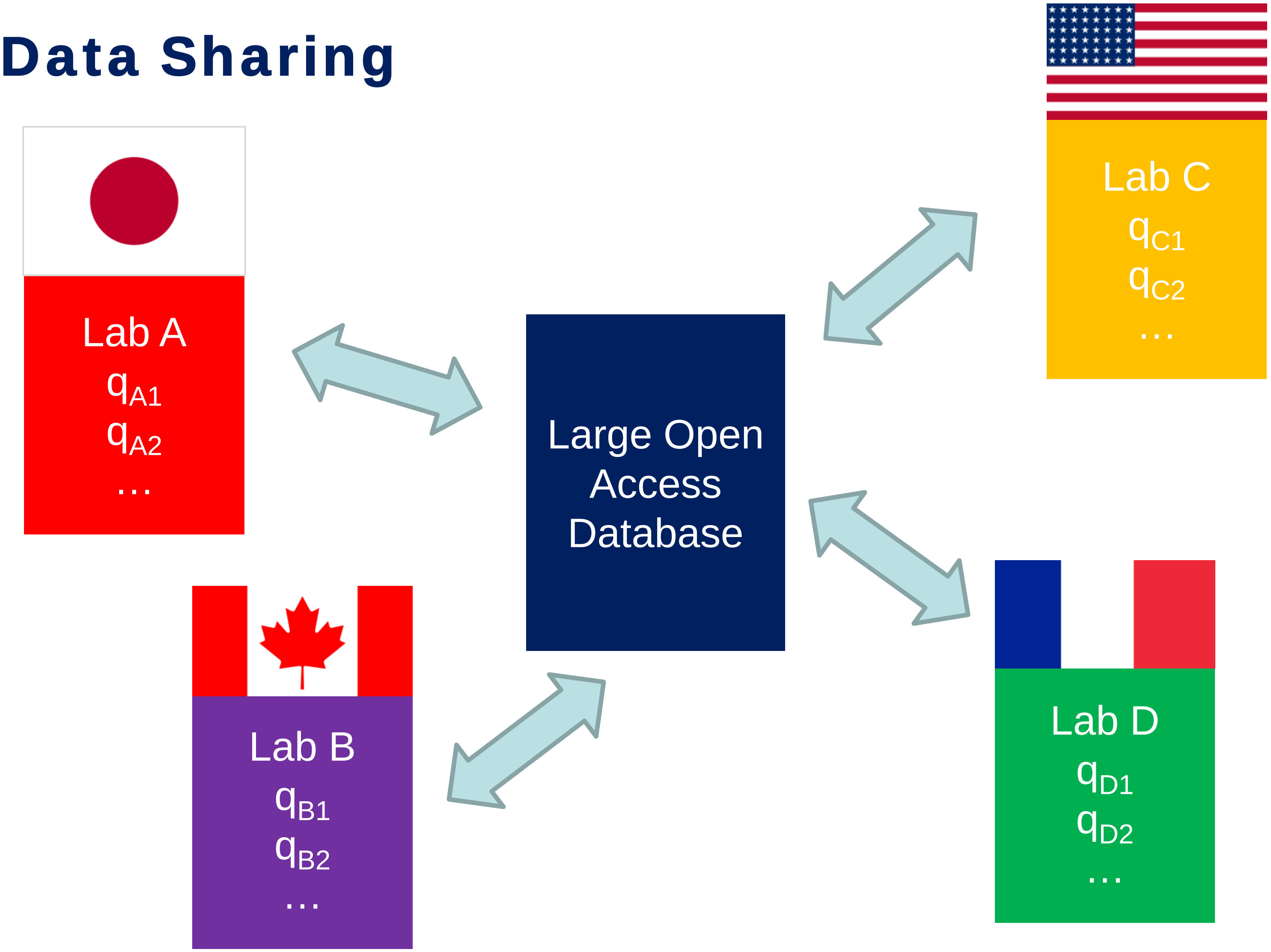


t^p = t-test(A^p)

MATLAB Demo

Recent Topics

- **Data Sharing**



Lab A

q_{A1}

q_{A2}

…

Lab B

q_{B1}

q_{B2}

…

Large Open Access Database

Lab C

q_{C1}

q_{C2}

…

Lab D

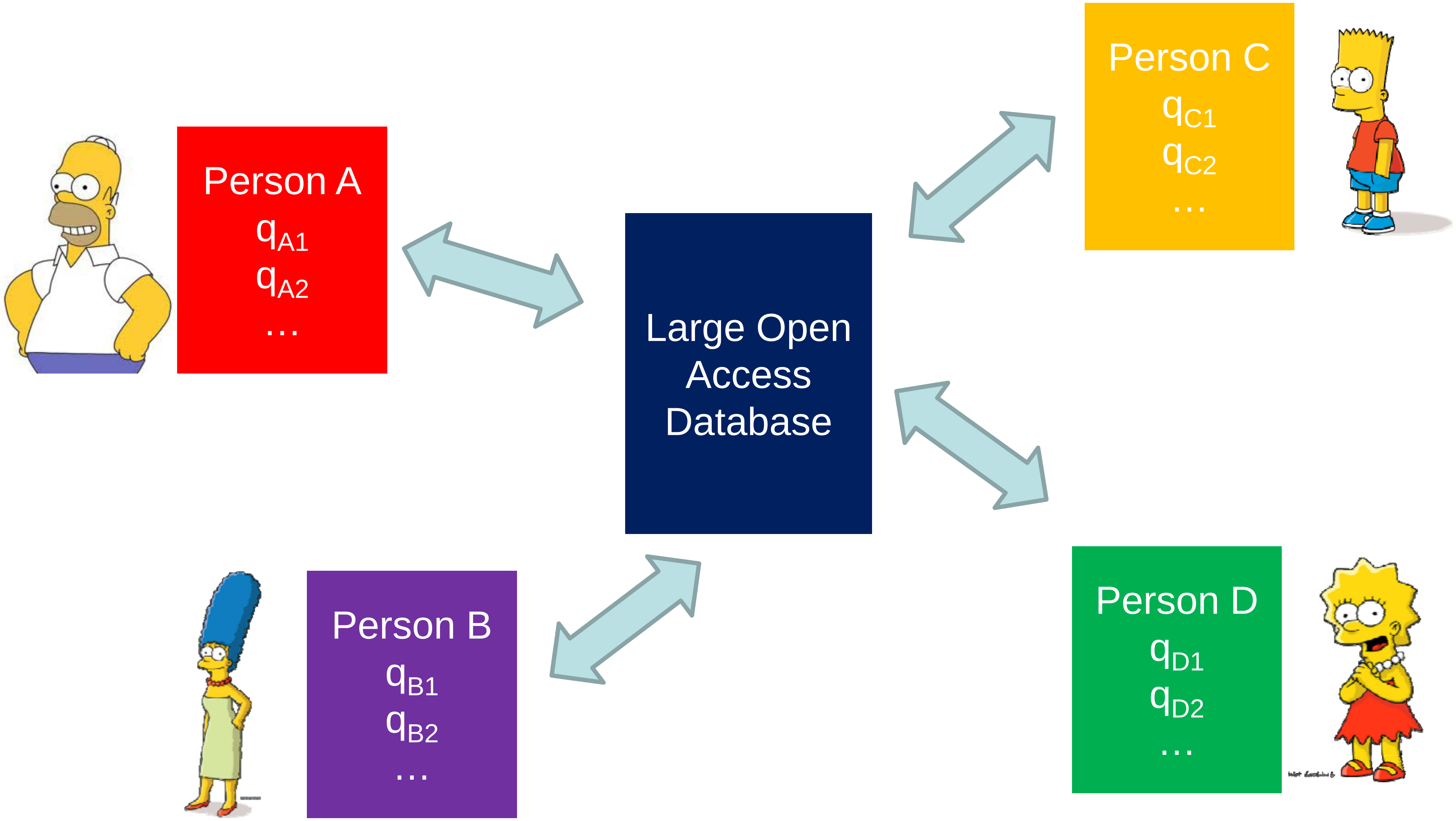
q_{D1}

q_{D2}

…

Recent Topics

- **Data Sharing**



Person A
q_{A1}
q_{A2}
…

Person B
q_{B1}
q_{B2}
…

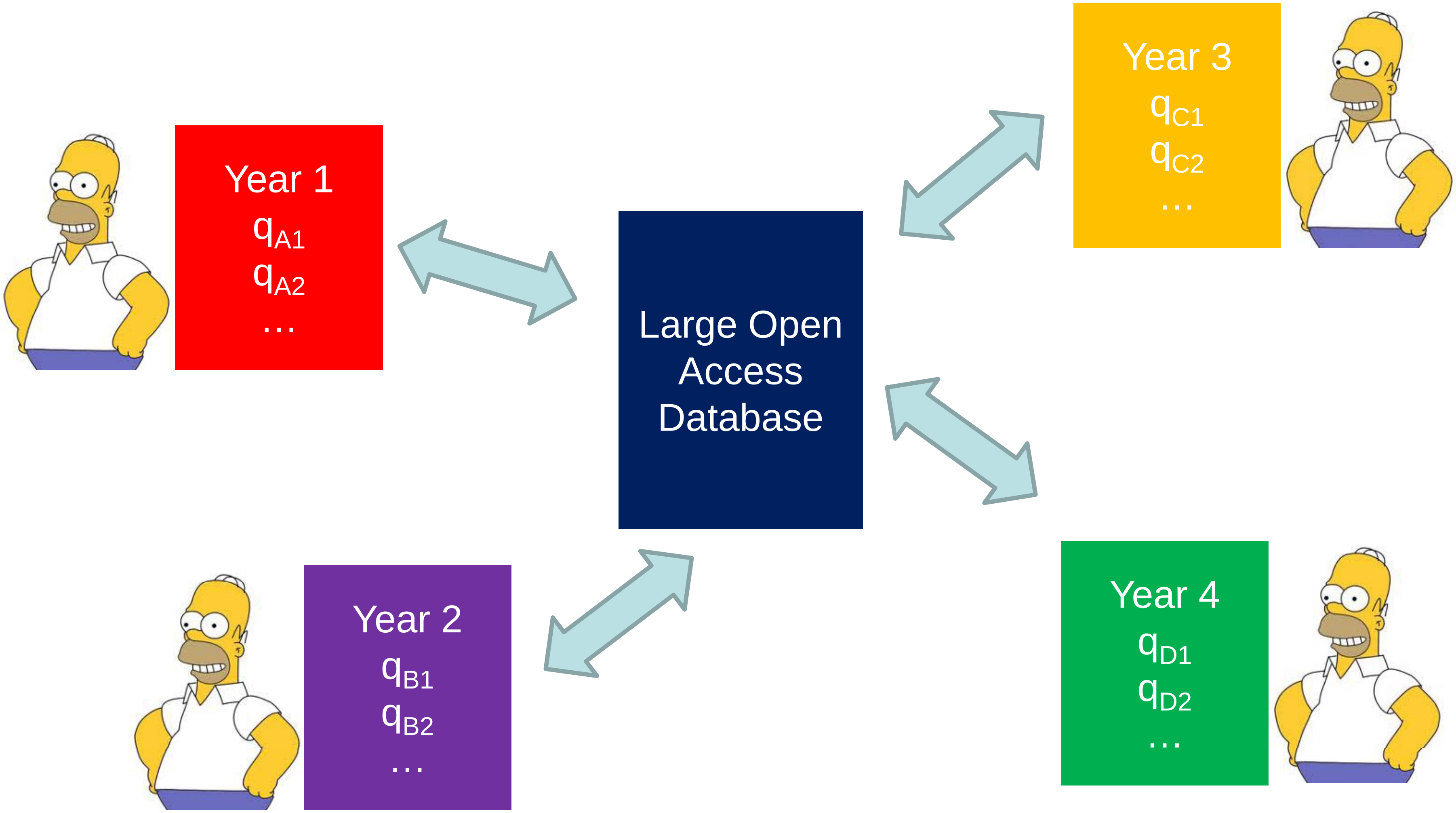
Large Open Access Database

Person C
q_{C1}
q_{C2}
…

Person D
q_{D1}
q_{D2}
…

Recent Topics

- **Data Reuse**



Year 1
q_{A1}
q_{A2}
…

Large Open Access Database

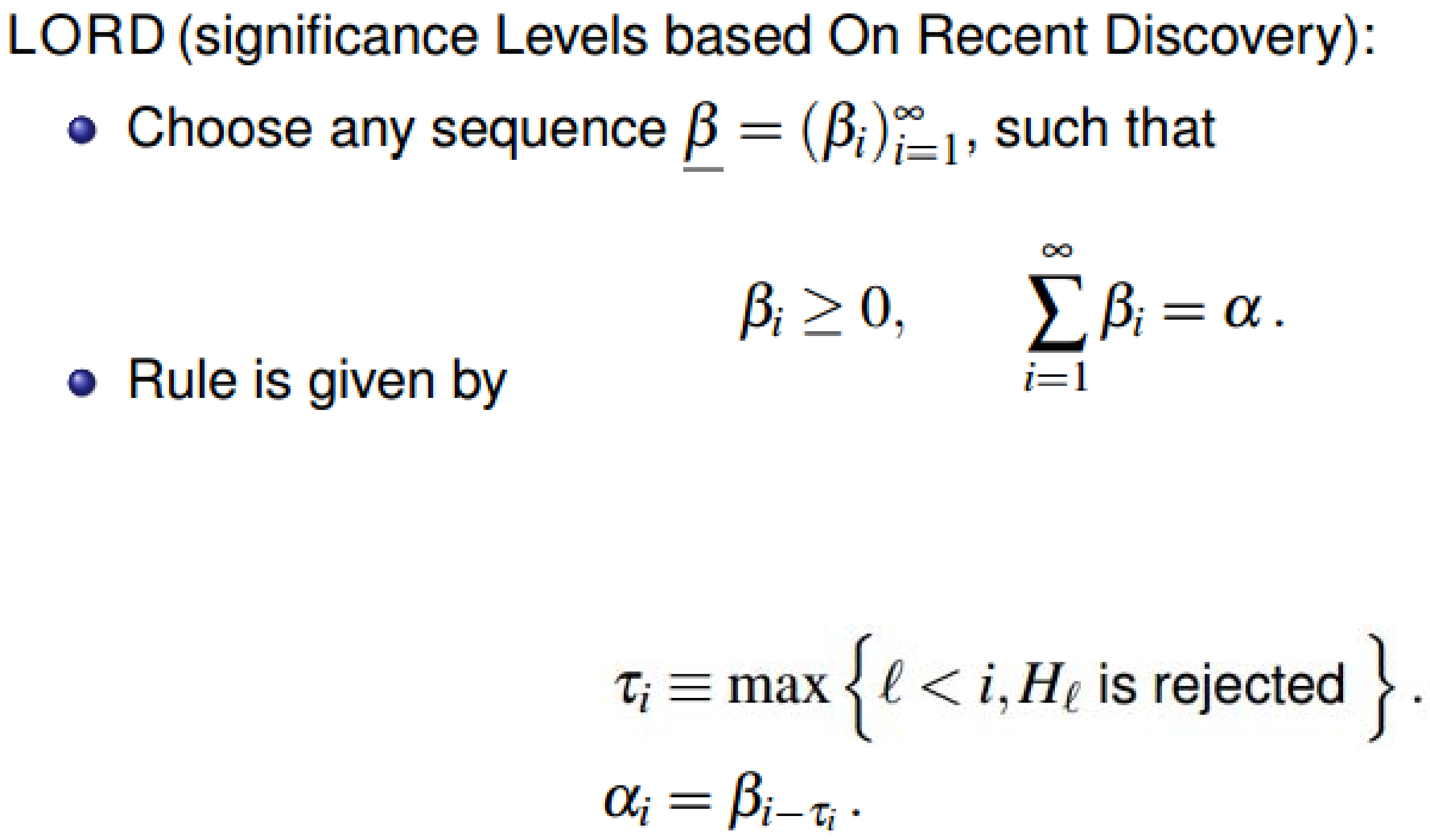
Year 2
q_{B1}
q_{B2}
…

Year 3
q_{C1}
q_{C2}
…

Year 4
q_{D1}
q_{D2}
…

Online FDR

Javanmard and Montanari, On Online Control of False Discovery Rate, 2015: arXiv:1502.06197



LORD (significance Levels based On Recent Discovery):

- Choose any sequence $\underline{\beta} = (\beta_i)_{i=1}^{\infty}$, such that

$$\beta_i \geq 0, \qquad \sum_{i=1}^{\infty} \beta_i = \alpha \, .$$

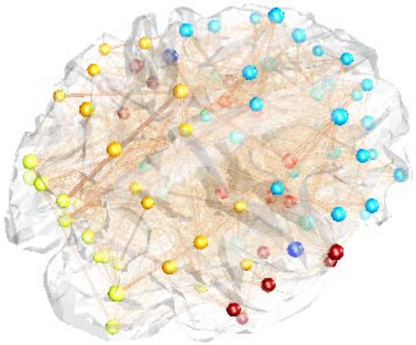
- Rule is given by

$$\tau_i \equiv \max \Big\{ \ell < i, H_\ell \text{ is rejected} \Big\} \, .$$

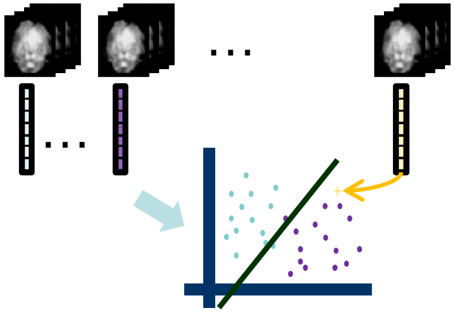
$$\alpha_i = \beta_{i-\tau_i} \, .$$

Neuroimaging Applications

Activation Detection



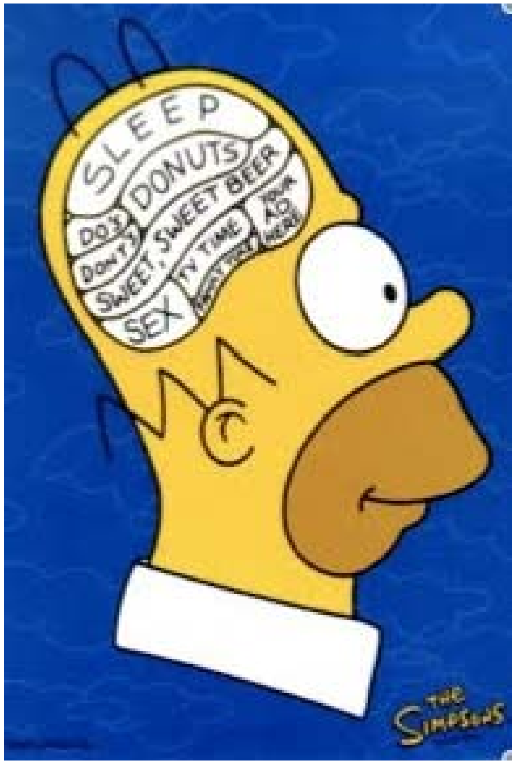
Connectivity Estimation



Brain Decoding

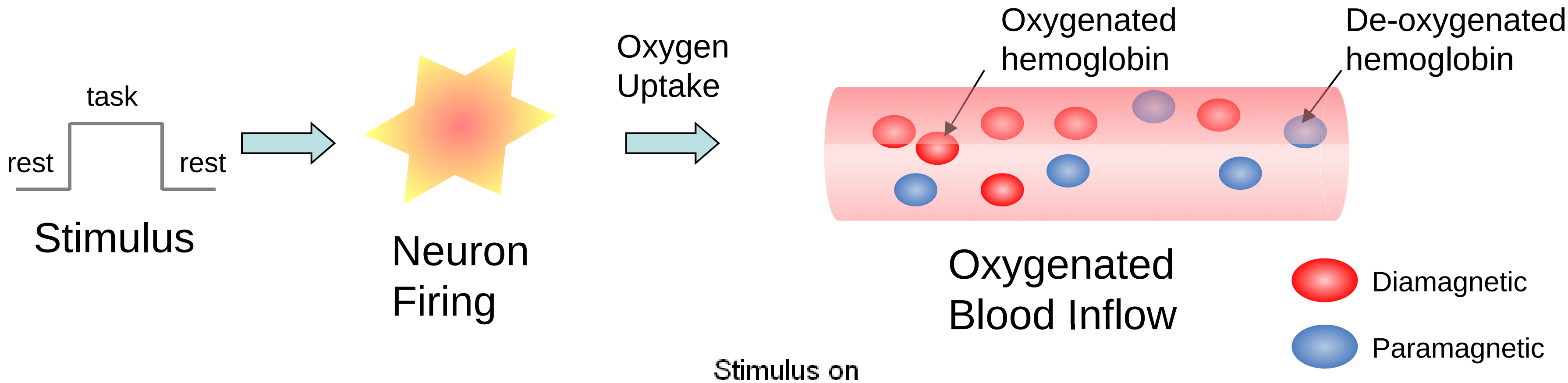
Shape vs. Function

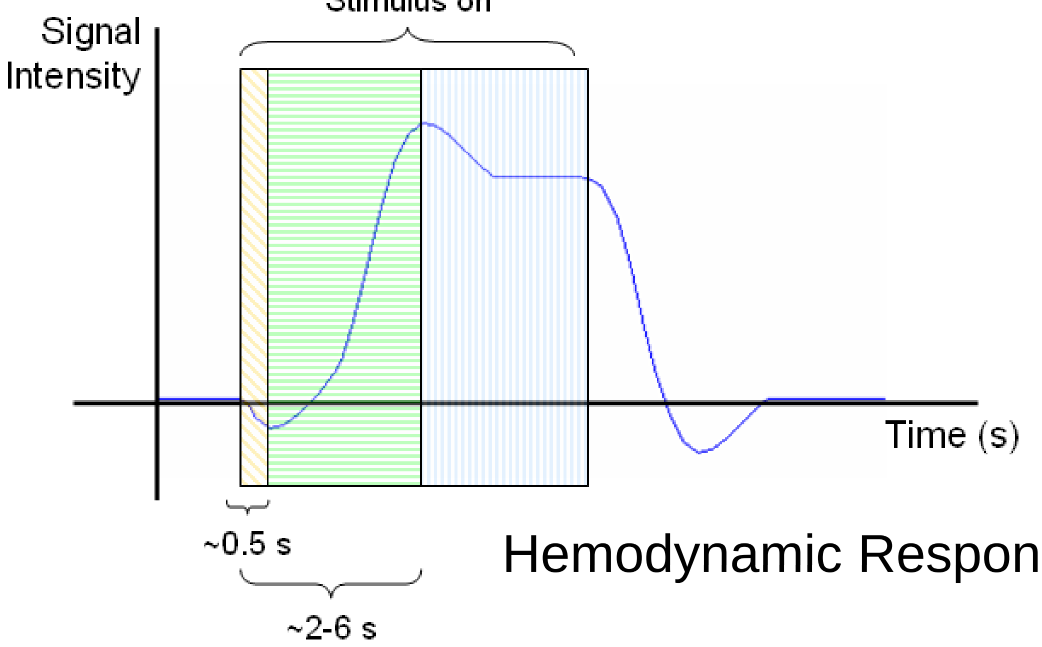
Magnetic Resonance
Imaging (MRI)



Functional
MRI (fMRI)

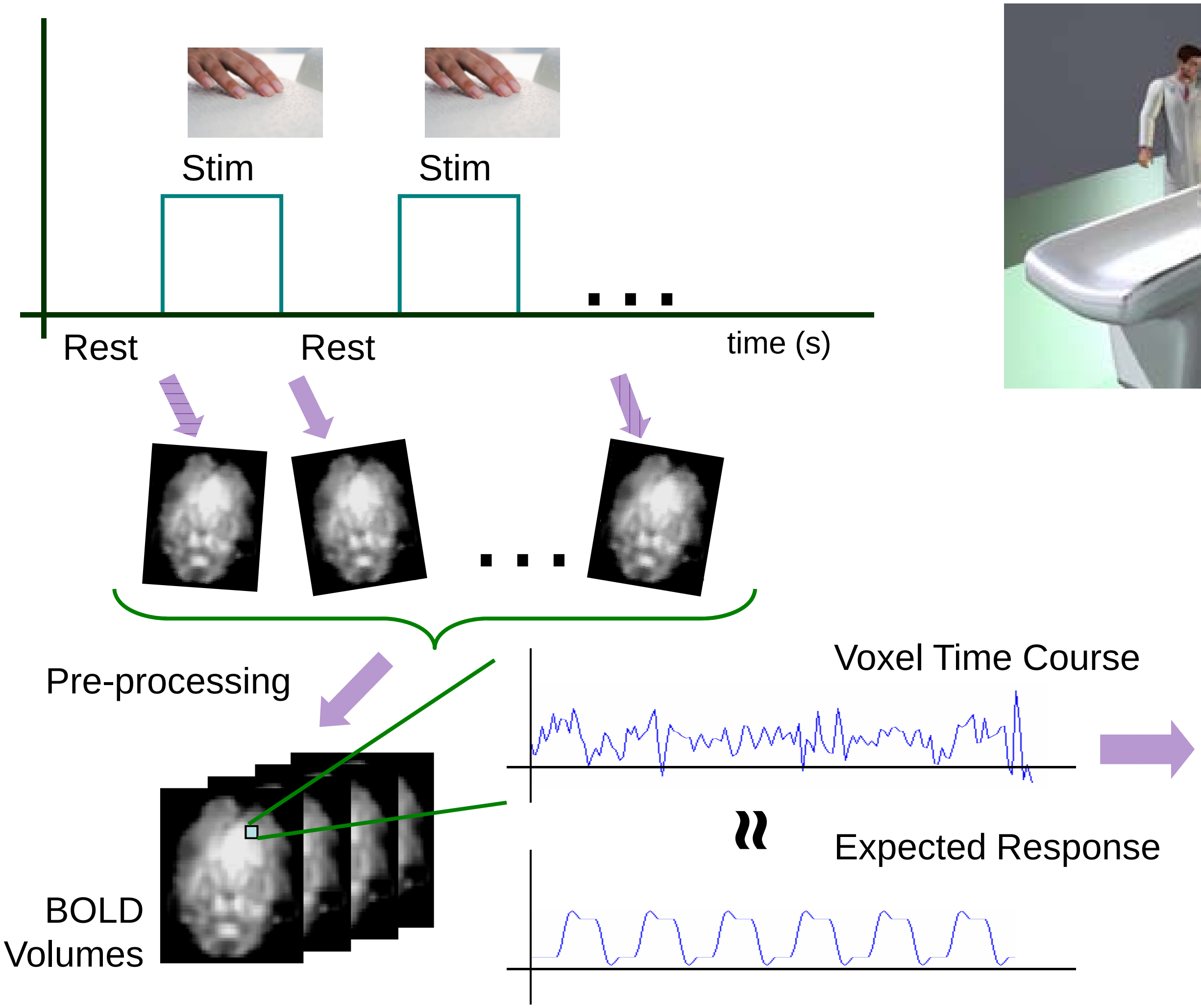
BOLD Effect





Hemodynamic Response

Task-based fMRI



Stim

Stim

Rest

Rest

time (s)

Pre-processing

BOLD Volumes

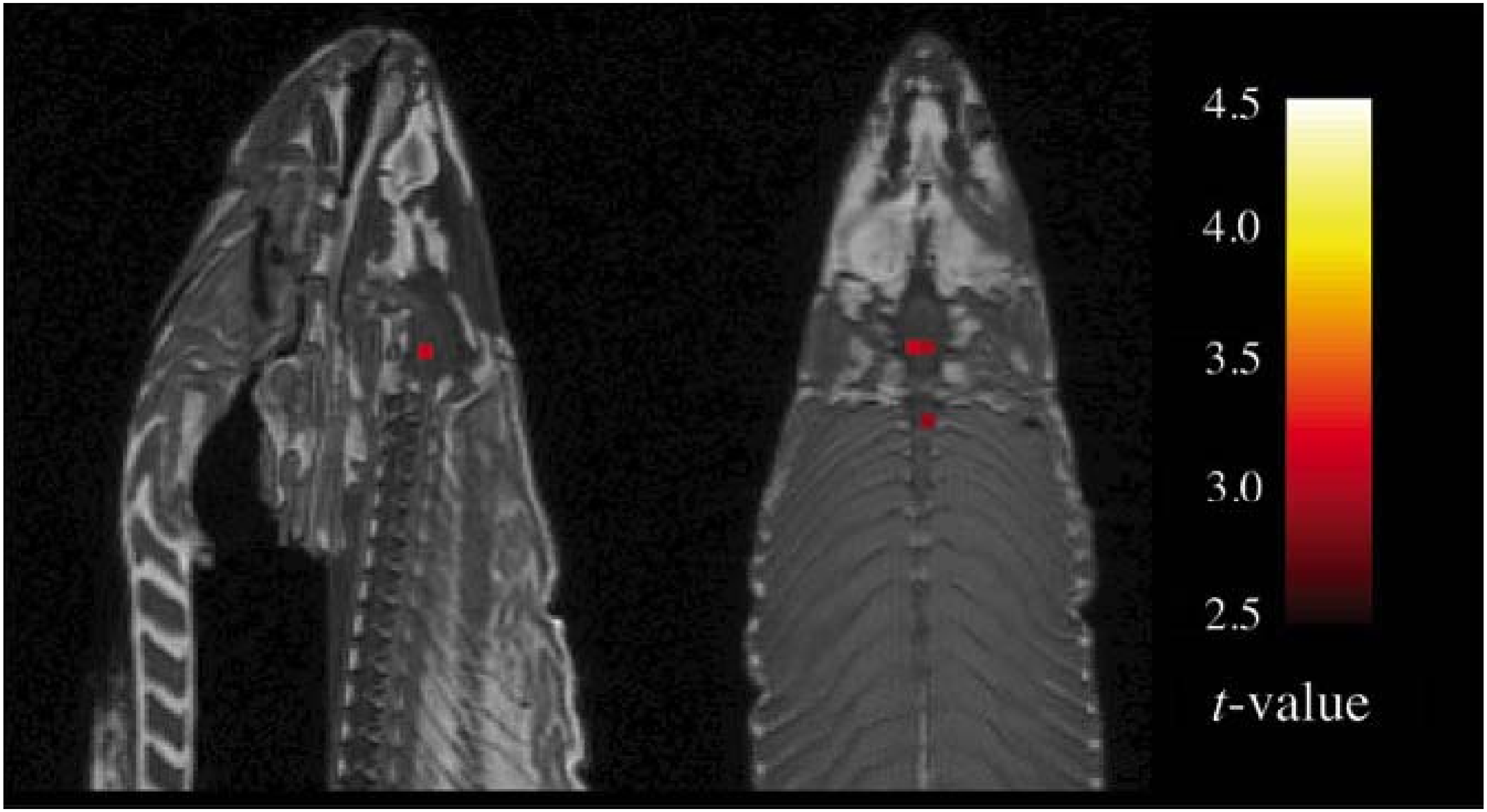
Voxel Time Course

≈

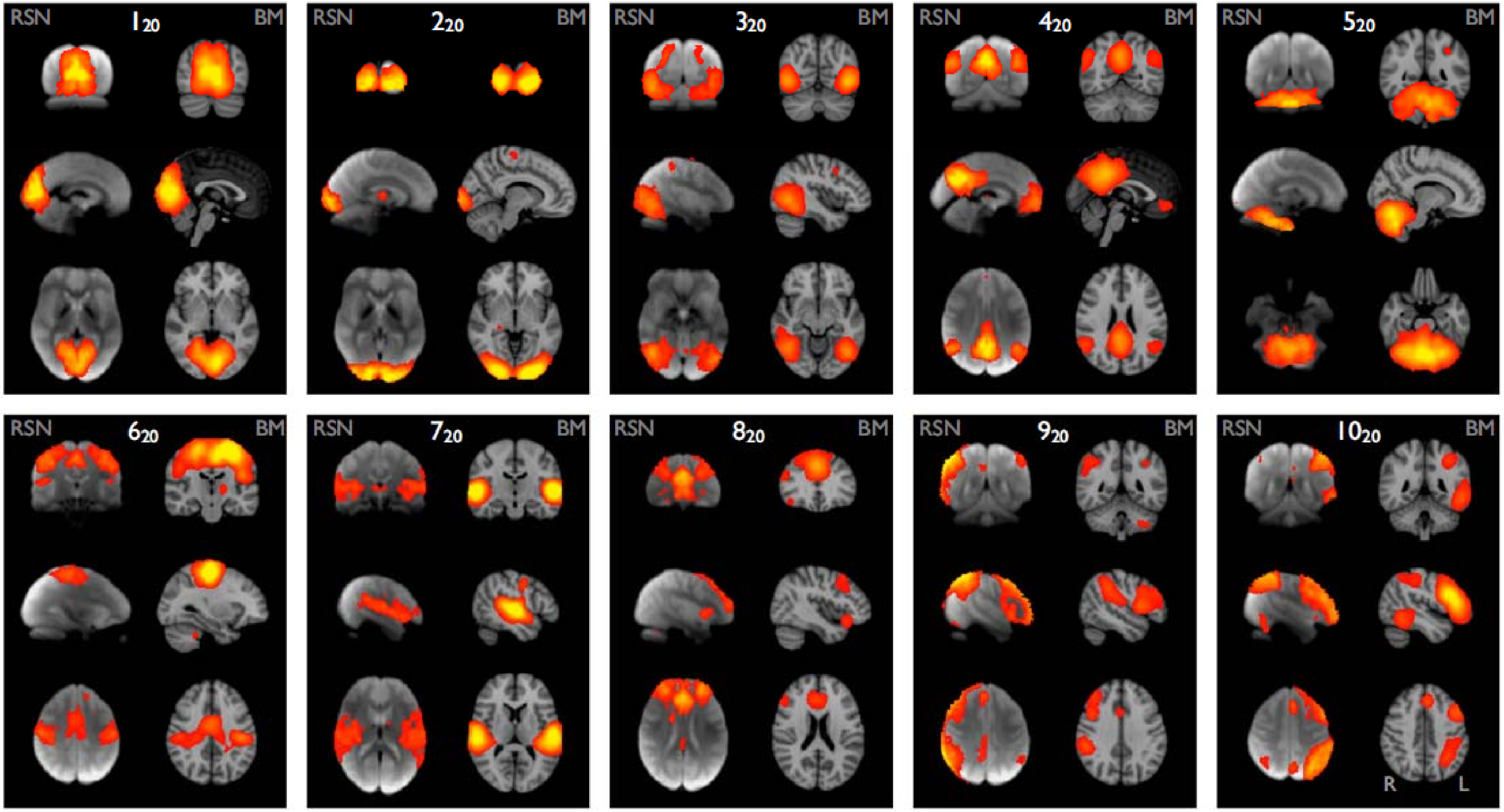
Expected Response

Activation Statistics Maps

Famous Dead Salmon Study

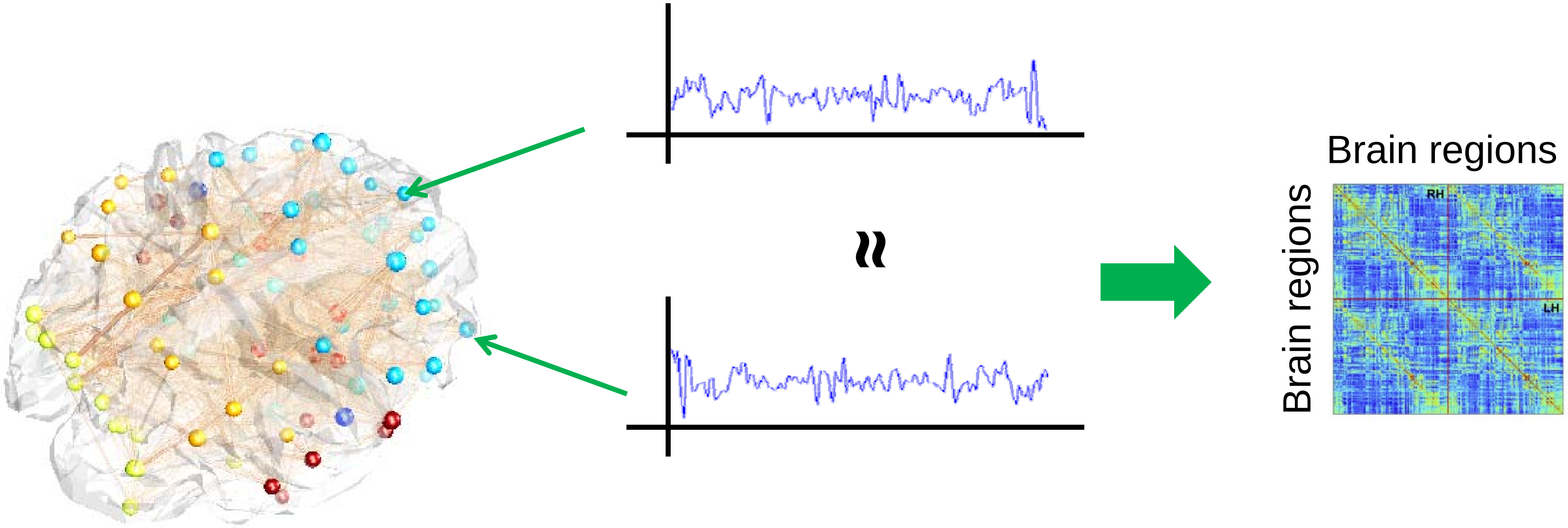


Resting State fMRI



SM. Smith, P.T. Fox, K.L. Miller, D.C. Glahn, P.M. Fox, C.E. Mackay, N. Filippini, K.E. Watkins, R. Toro, A.R. Laird, and C.F. Beckmann, “Correspondence of the Brain’s Functional Architecture During Activation and Rest,” Proc. Natl. Acad. Sci., vol.106, pp.13040-13045, 2009

Connectivity Estimation



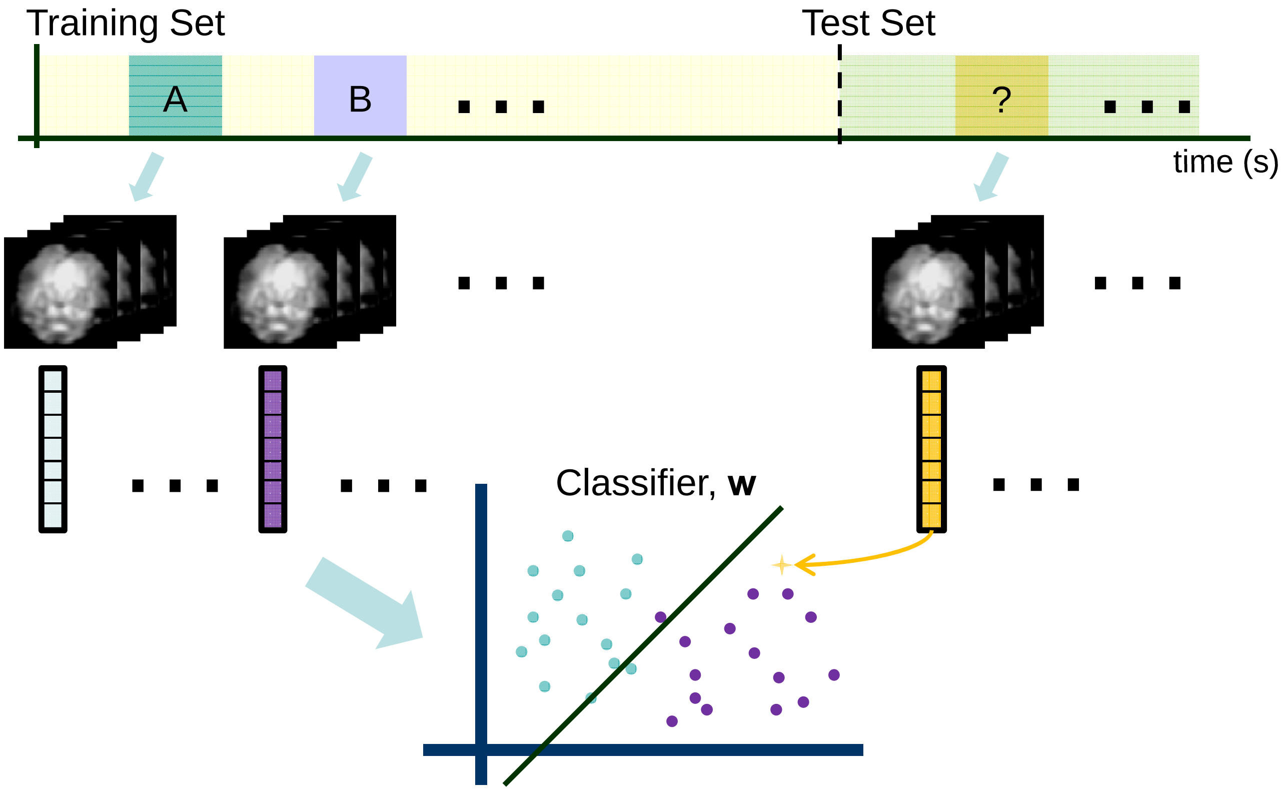
≈

Brain regions

Brain regions

Which connection significant?

Brain Decoding



From **w**, which variable significantly drives classification?

Summary

- Multiple testing can result in many false findings if the number of tests is not accounted for.
- Bonferroni correction is too stringent.
- FDR correction is a good compromise.
- Data sharing is creating new problems.