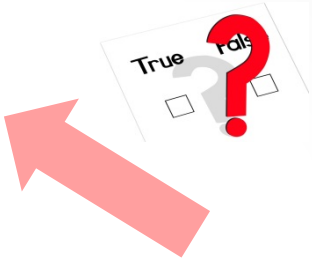
Preserving Validity in Adaptive Data Analysis

Vitaly Feldman

Accelerated Discovery Lab

IBM Research - Almaden

Distribution
P over X



Parameter estimate

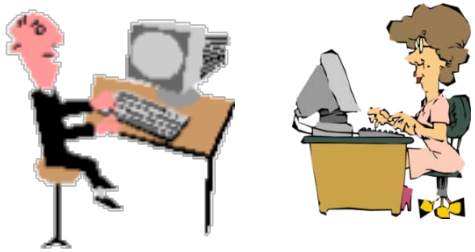
Classifier

Clustering

etc.

Results

$S = (x_1, \dots, x_n) \sim P^n$



Analysis

Data Analysis 101

Does student nutrition predict academic performance?

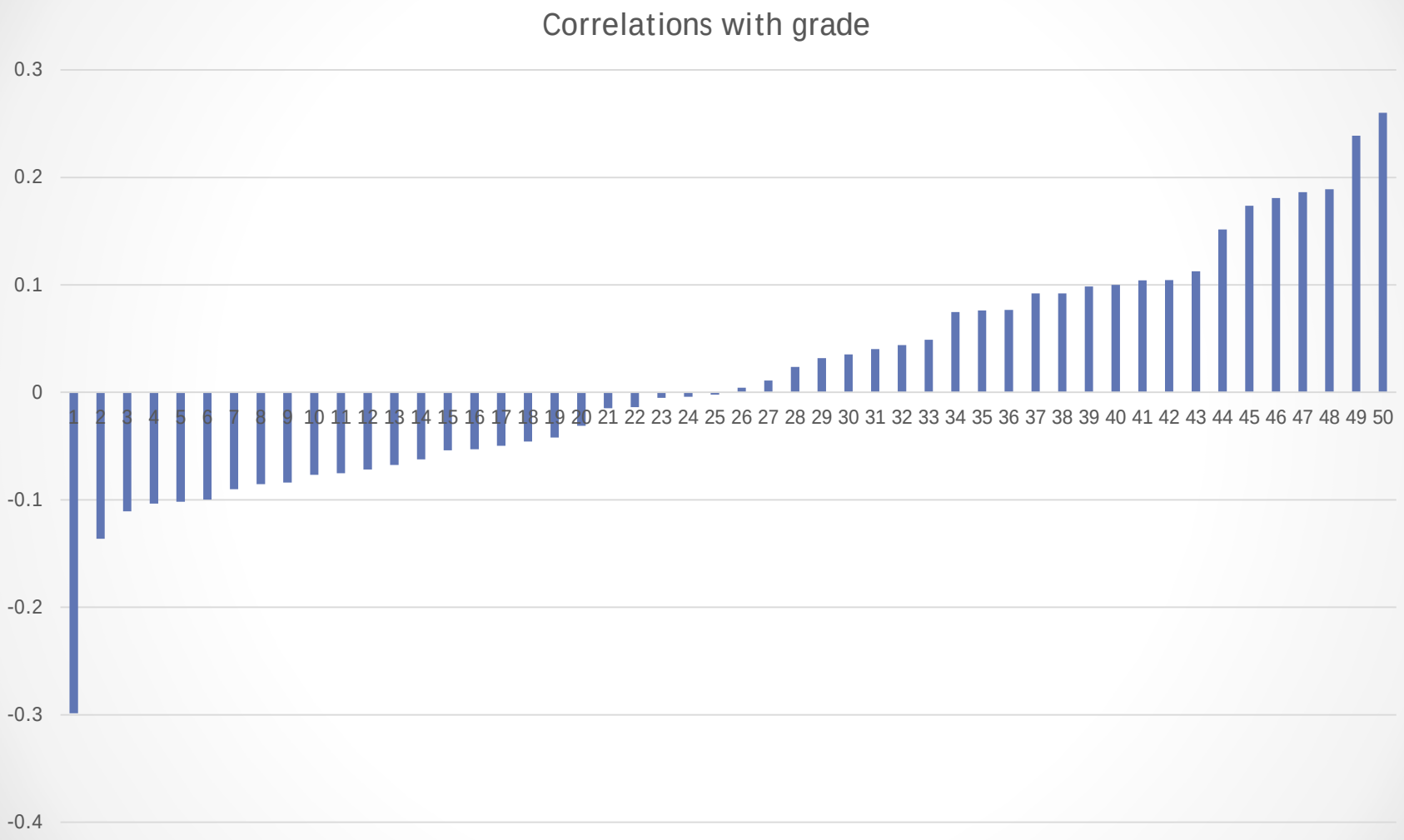
50

Normalized grade

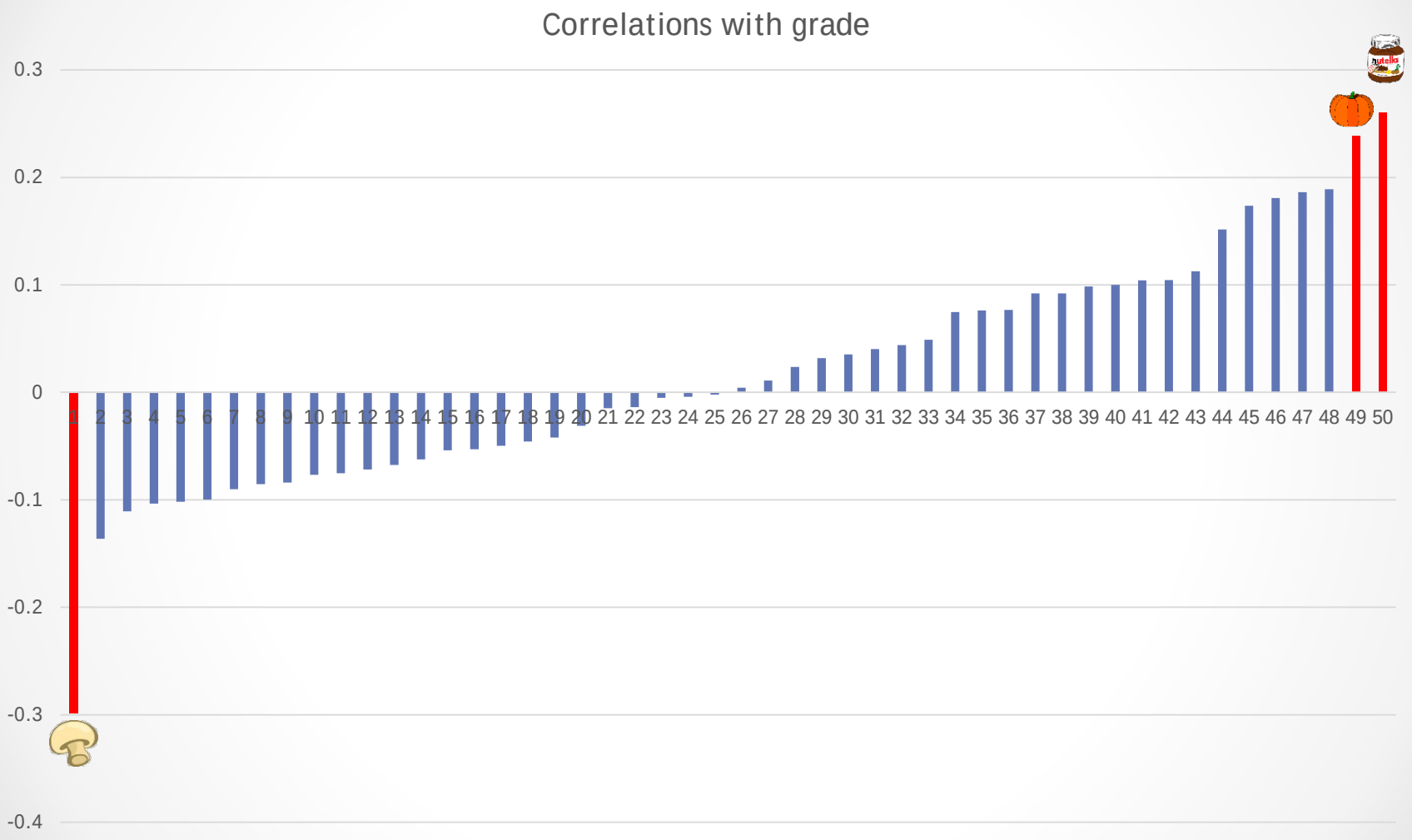
100

0.13	0.73	1.06	1.23	1.87	-0.97	-0.97	1.03	-0.24	-0.13	-0.63	-0.01	0.04	-0.54	0.14	-0.20	1.03	2.01	-1.16	1.40	-0.54	-0.56	0.06
-0.44	-0.72	0.30	0.13	-0.18	-2.13	0.30	1.94	1.88	0.03	-2.26	1.30	-0.27	-0.20	0.80	1.10	0.65	0.48	-1.07	1.11	0.98	0.34	3.01
-0.10	0.16	-0.31	-0.96	-0.40	0.93	1.86	-1.16	-0.13	-1.07	0.96	-0.57	-0.45	0.38	0.82	-0.77	0.48	-0.42	1.10	-0.76	1.64	1.18	1.15
-0.38	0.34	-1.61	0.28	-0.53	-2.07	1.87	-2.51	-0.28	-0.58	-0.40	-1.12	-0.56	0.91	-0.79	0.65	-1.83	0.52	2.61	0.81	0.87	0.38	0.73
-0.16	0.39	-1.42	-0.55	-0.45	-0.94	-1.99	0.17	-0.05	1.14	0.03	0.64	-0.94	0.31	2.01	-0.04	1.90	-0.68	-0.62	-1.21	0.36	-0.53	-2.63
1.31	-1.21	0.00	0.99	0.07	0.58	-1.30	0.08	0.43	-1.51	-1.20	0.05	-2.54	1.03	2.05	-0.50	1.60	-1.24	-0.37	-0.57	1.24	-1.38	1.94
1.19	-0.52	1.10	-1.04	-0.18	0.16	0.62	0.70	1.07	0.41	0.02	0.46	0.63	1.01	0.91	-0.75	0.70	0.48	-0.59	-1.76	1.23	1.30	0.71
0.81	-0.60	1.13	-2.30	1.23	-0.31	0.36	0.43	2.36	1.43	-1.97	1.30	-1.17	0.08	1.67	-0.14	-0.40	0.02	0.23	0.79	-1.34	1.50	1.77
1.51	-0.35	-0.49	-0.51	0.36	1.76	2.42	-0.31	-1.14	1.86	0.45	0.71	1.26	-1.34	-0.09	0.86	1.11	0.47	-0.39	0.69	-0.87	-2.33	0.31
1.46	0.97	0.44	0.03	0.11	-0.39	0.14	0.63	-0.76	0.22	0.28	-0.53	-0.05	0.81	1.22	1.02	-1.27	-0.83	0.65	1.31	1.56	-0.53	-1.23
0.31	-0.77	0.16	0.25	0.00	1.03	-0.53	-0.78	-0.64	-2.31	-1.59	0.20	-0.28	-0.74	0.32	0.05	-0.38	1.04	1.70	1.18	-1.54	0.18	0.09
-1.70	0.11	1.31	0.50	-0.71	0.71	2.43	0.27	-0.80	1.50	0.46	-1.20	-0.61	1.30	-0.48	-0.01	-0.82	-1.03	0.38	0.63	-1.11	1.33	1.70
-0.20	0.19	-0.41	-0.45	-0.82	-0.70	1.16	-1.79	-0.21	1.45	-0.29	-0.85	1.39	-2.55	0.71	-0.90	1.91	1.75	-0.38	-0.89	0.60	-0.47	1.08
-0.42	0.20	2.03	-0.03	1.81	0.87	1.21	0.15	-0.24	-2.07	1.04	-1.90	0.01	1.47	-0.33	1.69	0.00	0.44	0.02	-0.11	0.12	-1.28	-0.08
-0.78	-0.69	-0.05	-1.17	-1.24	-0.34	-0.50	-0.09	-1.35	0.46	0.94	-1.69	0.96	-0.10	-0.30	-0.74	-0.63	0.17	0.03	1.19	-0.77	1.56	-0.43
0.15	0.45	0.34	0.20	-0.88	1.11	0.88	-0.47	0.87	-1.67	0.08	1.21	-1.82	-0.98	-0.09	0.42	-0.64	-0.02	2.03	-0.27	-0.89	-0.90	-2.02
1.20	1.18	0.69	0.04	-1.06	0.90	1.23	-0.48	-0.28	-1.23	-0.96	-0.72	0.45	0.36	0.65	0.57	1.22	0.28	-0.55	-0.43	0.07	2.50	-1.18
-0.76	-0.30	-1.25	1.07	0.55	1.35	-1.33	-1.04	-1.26	1.46	-0.73	0.16	0.84	-1.37	0.89	-0.74	-0.25	-0.66	2.13	-0.82	-0.90	0.40	-1.26
0.59	0.00	-0.28	-1.00	1.09	0.27	0.63	1.03	-1.19	0.88	1.89	1.29	-1.17	0.99	1.72	-1.69	-1.03	-2.03	-0.56	-1.12	-1.43	-1.10	-0.36
-0.50	0.98	0.40	-1.93	-1.73	-2.28	0.85	0.61	-0.53	-0.20	-0.97	0.74	-1.06	-0.31	-0.55	-0.44	1.17	-1.51	-0.32	0.57	-0.09	0.31	0.82
0.39	1.95	-0.39	-1.09	2.11	0.14	0.05	2.78	0.87	-0.59	0.14	0.49	-0.82	0.20	0.01	0.16	-0.33	-0.31	-0.81	-1.10	1.21	0.66	0.82
-0.82	0.10	-1.00	1.38	-0.40	0.29	-0.49	-0.62	-1.40	-0.90	0.87	-0.13	0.23	0.65	-1.42	-0.39	0.97	0.35	-1.55	0.16	-0.19	1.13	0.46
0.89	-1.83	-0.01	-1.02	-0.18	-0.47	0.17	-1.92	0.23	-0.98	0.12	0.58	-0.91	-1.31	0.71	-0.39	-0.17	-1.97	1.45	0.22	-0.87	-0.12	0.92
-0.69	-0.74	1.46	-0.87	-1.21	0.42	-1.18	-1.28	0.19	-0.14	-0.29	-1.04	-0.28	0.66	2.03	-0.76	0.12	2.81	-0.31	-1.40	-1.03	-0.31	0.07
0.08	-0.02	-0.28	-1.57	1.33	1.64	-0.20	-0.53	-0.74	0.39	-0.11	-0.34	-1.21	0.99	0.06	-0.46	-0.11	-0.53	-0.54	-0.23	0.28	0.79	0.36
-0.41	-1.53	1.42	-0.49	1.68	-0.22	0.89	-0.68	-1.83	-0.79	1.12	-2.09	-0.08	1.32	-0.50	-0.58	0.90	1.87	0.30	-0.79	0.62	0.17	0.45
0.06	-1.13	-0.64	0.68	0.43	0.06	-0.05	-0.67	-0.29	-1.73	-0.66	0.39	1.00	-0.03	0.27	-1.37	0.66	0.13	0.97	0.90	0.40	-0.01	0.05
-0.78	1.86	1.10	2.14	-0.73	-0.03	2.05	2.18	-1.76	-0.49	1.31	-0.38	-0.18	-0.54	0.19	-0.53	0.29	-0.74	-0.02	-0.13	0.49	-2.36	0.34
0.45	0.54	0.05	-0.13	-0.96	0.20	0.49	-1.35	-0.45	-0.74	-0.96	-0.89	0.17	-0.23	-1.20	-1.49	-0.45	-0.84	-0.20	-1.94	1.01	-0.80	-0.13
-1.40	-0.11	1.52	-0.31	0.01	0.35	1.50	0.13	-0.65	-0.74	0.36	-1.55	0.41	-1.47	0.38	-0.10	-1.45	-1.87	-0.40	0.78	-0.90	1.80	0.11
0.44	-0.21	0.42	0.31	-1.35	0.56	-0.68	-0.57	1.96	0.43	-1.50	-0.10	0.66	2.11	0.13	0.24	0.22	-0.40	-1.54	-0.94	0.00	-0.01	-0.97

Check correlations



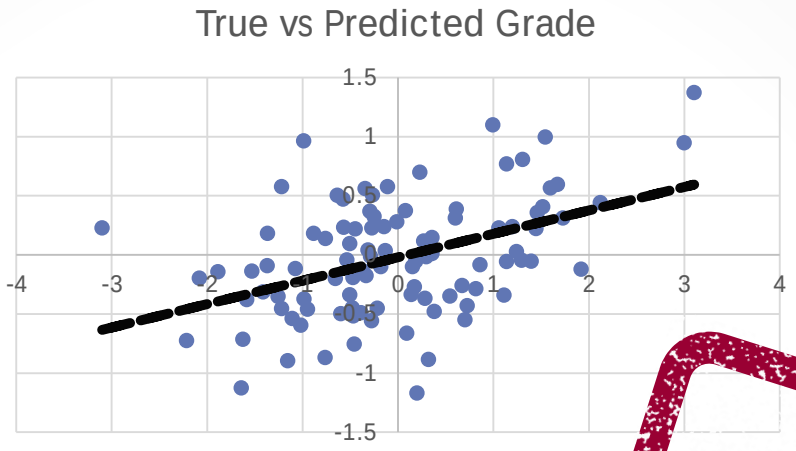
Pick candidate foods



Fit linear function of 3 selected foods

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.4453533
R Square	0.1983396
Adjusted R Square	0.1732877
Standard Error	1.0041891
Observations	100



ANOVA

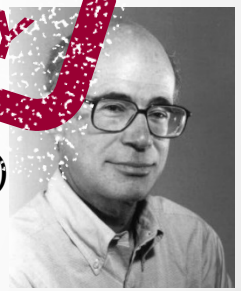
	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	23.95086544	7.983622	7.917151	8.987[illegible] ≈ 0.00001
Residual	96	96.80600126	1.008396		
Total	99	120.7568667			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-0.044248	0.100545016	-0.44008	0.660868
Mushroom	-0.296074	0.10193011	-2.90468	0.004563
Pumpkin	0.255769	0.108443069	2.358555	0.020373
Nutella	0.2671363	0.095186165	2.806462	0.006066

True P:

$x_i, y \sim N(0,1)$

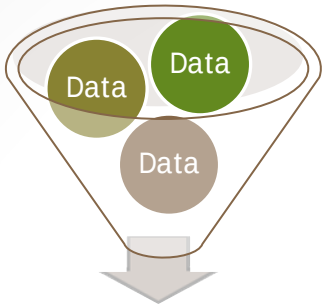
Uncorrelated

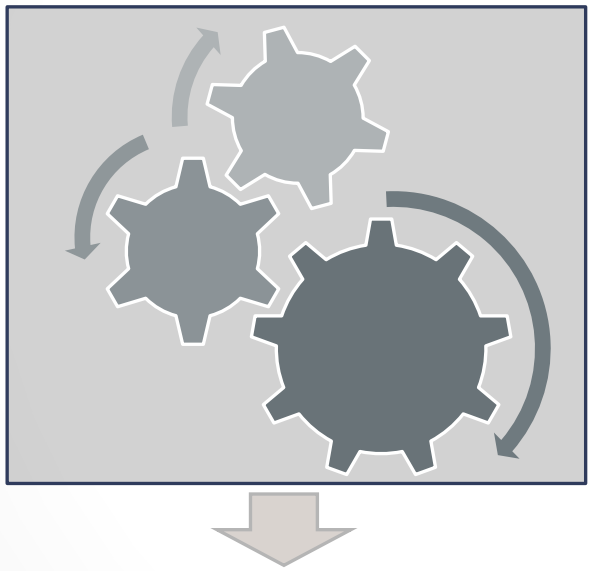


Freedman's Paradox [1983]

Statistical inference

"Fresh" i.i.d. samples

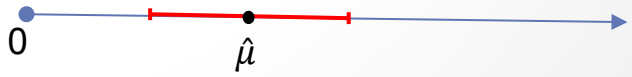




Result + generalization guarantees

Procedure

- Hypothesis tests
- Regression
- Learning



Data analysis is adaptive

Steps depend on previous analyses of the same dataset

Data analyst(s)

A_1

v_1

A_2

v_2

$\vdots$

A_m

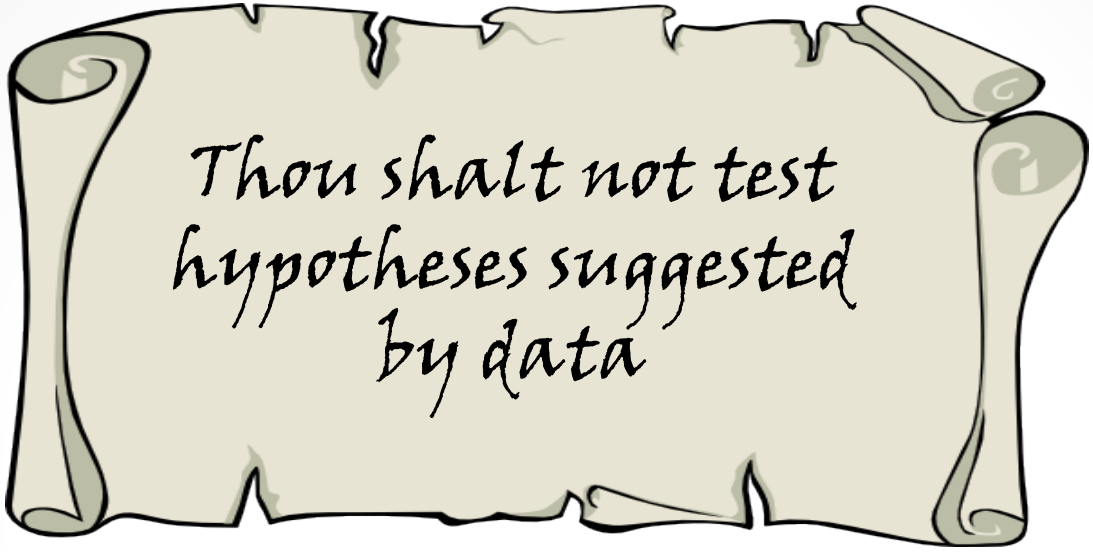
v_m

S

Data cleaning
Exploratory data analysis
Variable selection
Hyper-parameter tuning
Shared datasets
….

$$A_i : X^n \rightarrow Y_i$$
$$v_i = A_i(S)$$

It's an old problem



"Quiet scandal of statistics"
[Leo Breiman, 1992]

Is this a real problem?

"Why Most Published Research Findings Are False" [Ioannidis 2005]

"Irreproducible preclinical research exceeds 50%, resulting in approximately US$28B/year loss" [Freedman,Cockburn,Simcoe 2015]

Adaptive data analysis is one of the causes

- p-hacking
- Researcher degrees of freedom [Simmons, Nelson, Simonsohn 2011]
- Garden of forking paths [Gelman, Loken 2015]

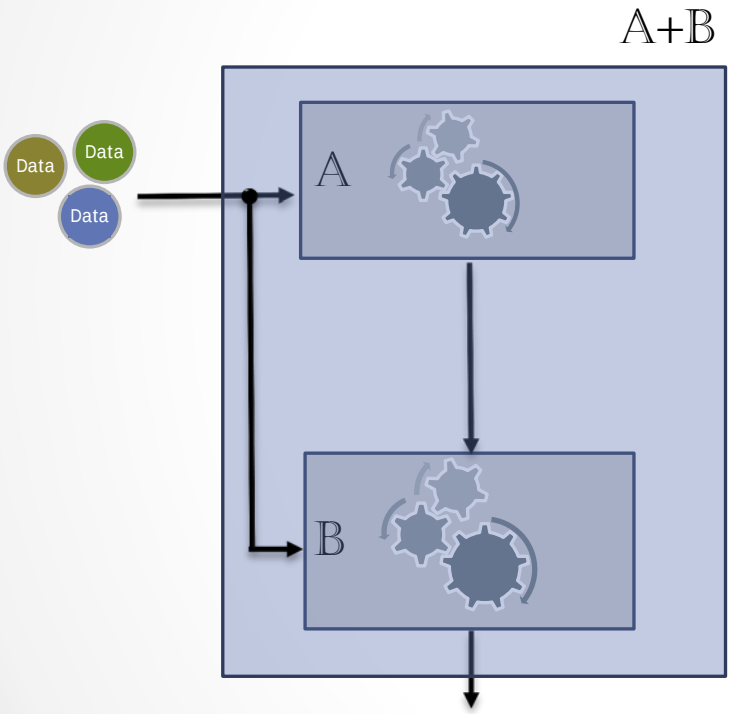
Existing approaches I

Abstinence

Pre-registration

Existing approaches II

Selective inference



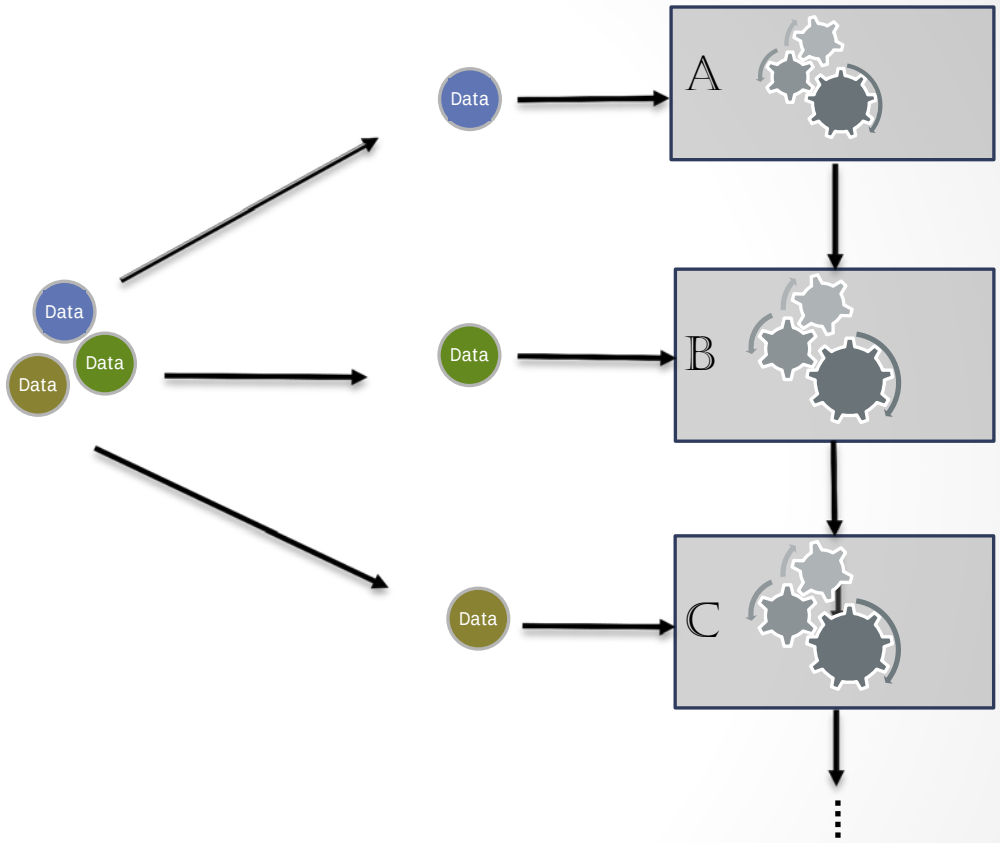
Examples:

- Model selection + parameter inference
- Variable selection + regression

Survey: [Taylor, Tibshirani 2015]

Existing approaches III

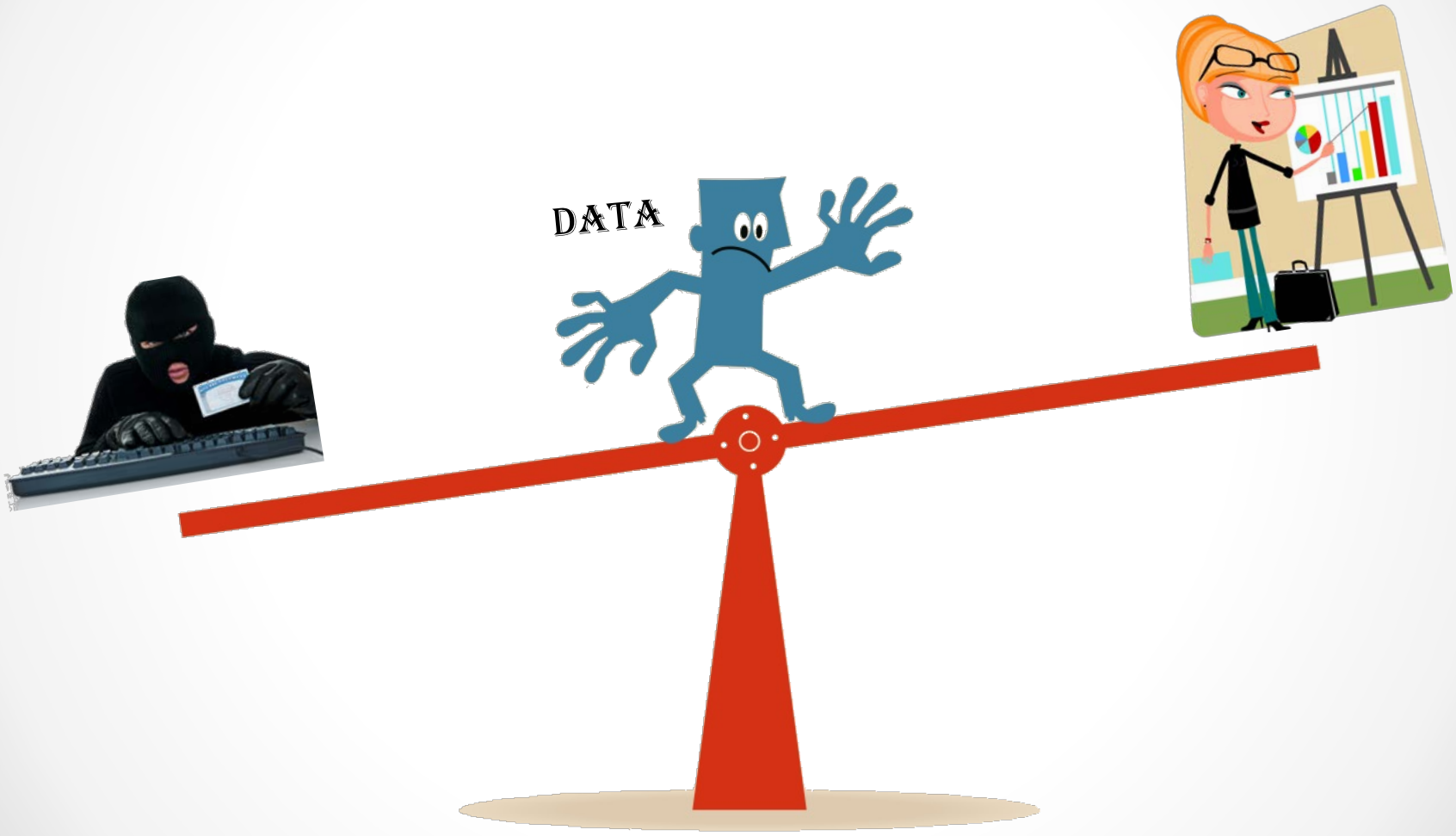
Sample splitting



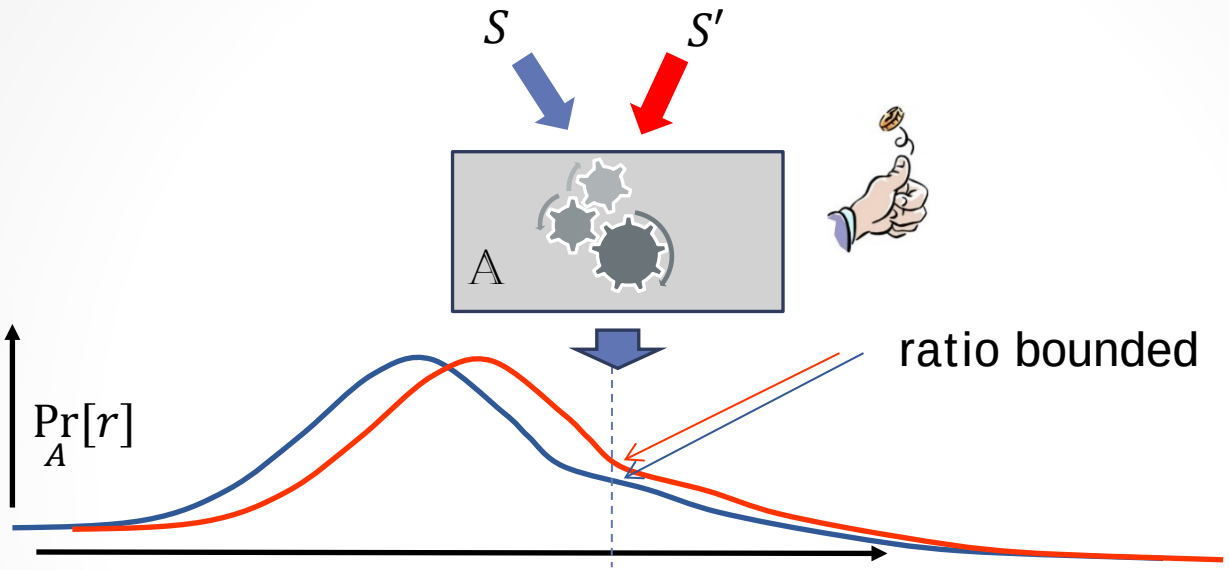
Might be necessary for standard approaches

New approach: differential privacy

[Dwork,McSherry,Nissim,Smith 06]

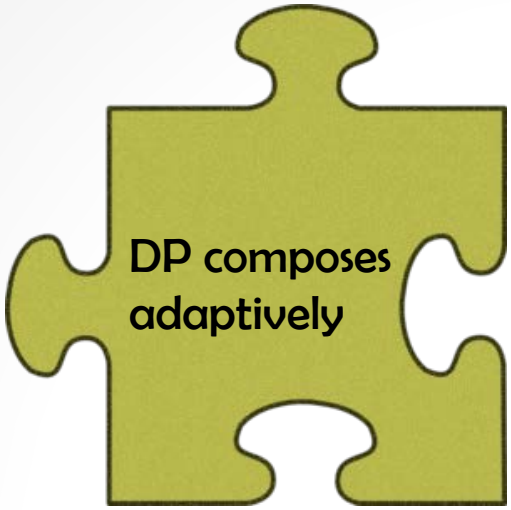


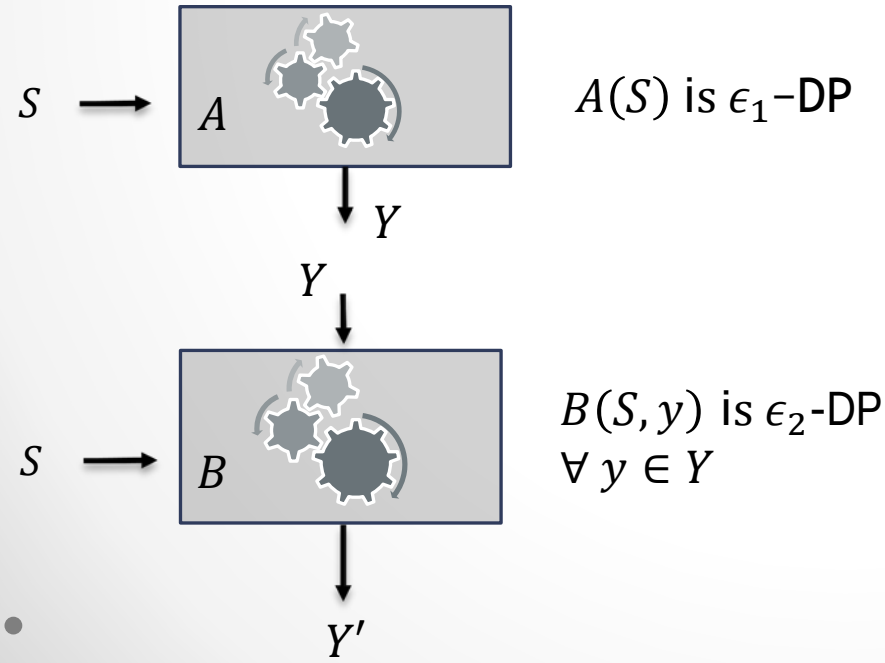
Differential privacy as outcome stability



Randomized algorithm A is ϵ-**differentially private (DP)** if for any two data sets S, S' such that $\mathrm{dist}(S, S') = 1$:

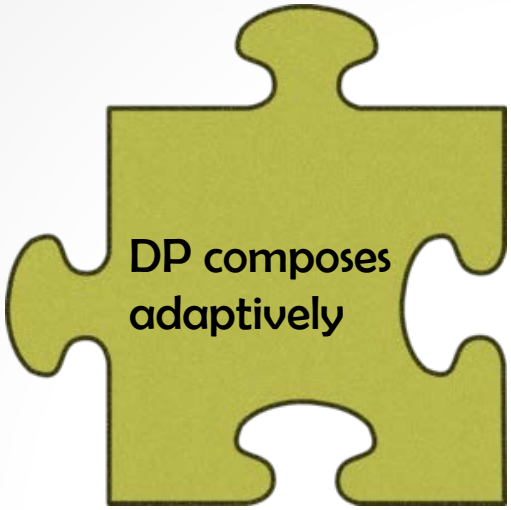
$$\forall Z \subseteq \mathrm{range}(A), \qquad \Pr_{\boldsymbol{A}}[\boldsymbol{A}(S) \in Z] \leq e^{\epsilon} \cdot \Pr_{\boldsymbol{A}}[\boldsymbol{A}(S') \in Z]$$

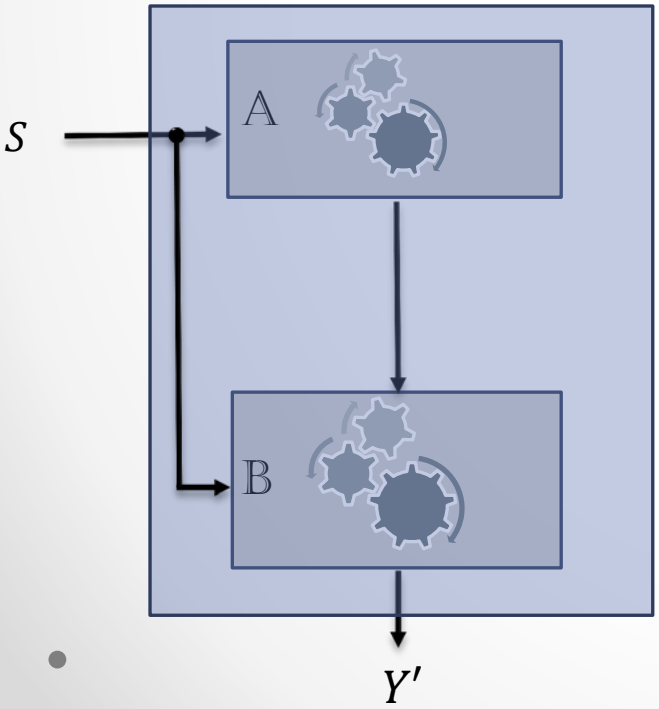




$A(S)$ is ϵ_1–DP

$B(S, y)$ is ϵ_2-DP
$\forall\, y \in Y$





$\epsilon_1 + \epsilon_2$-DP

DP is a strong form of algorithmic stability

Approaches

(Approximate) max-information

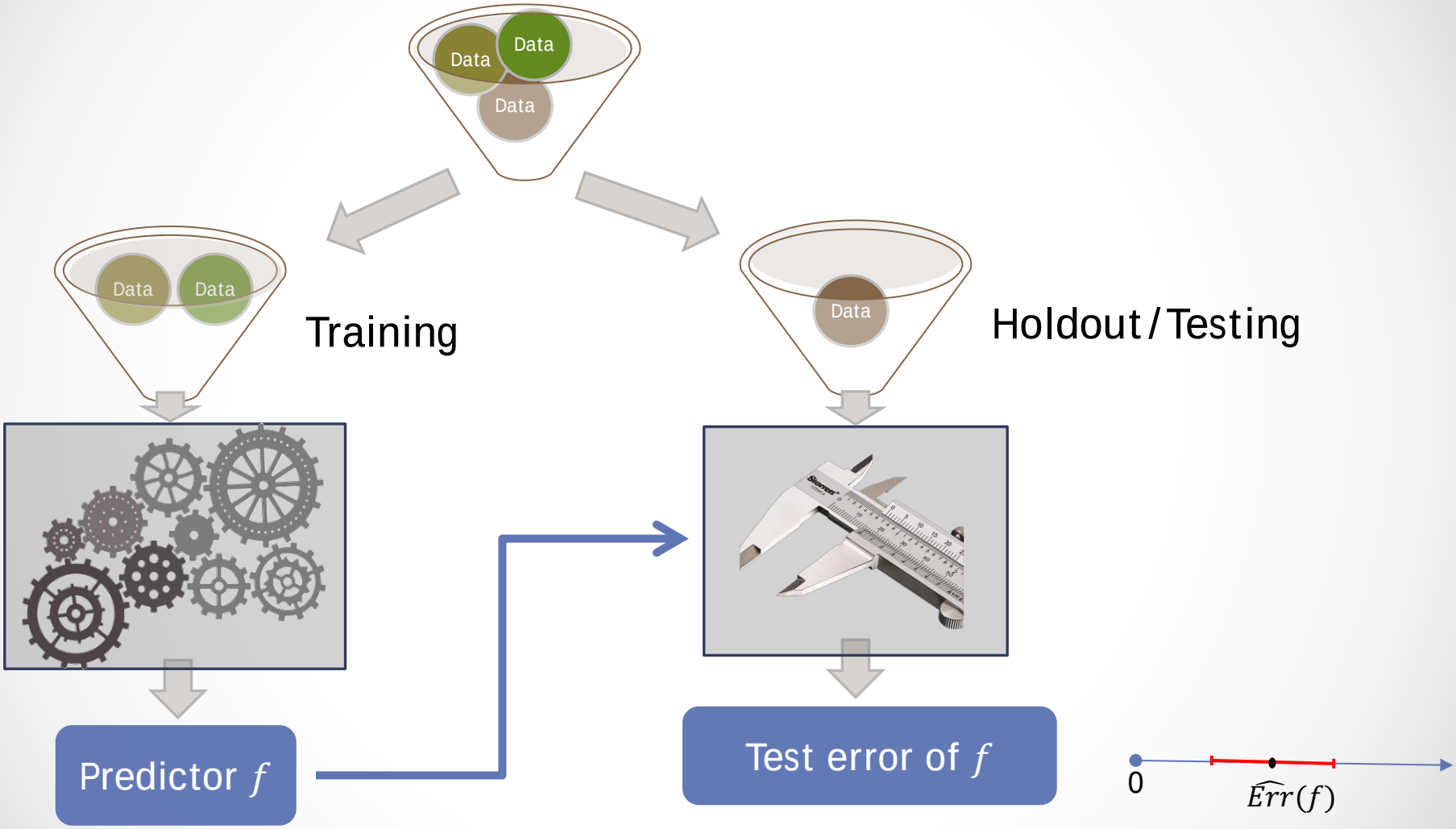
Differential privacy

Description length

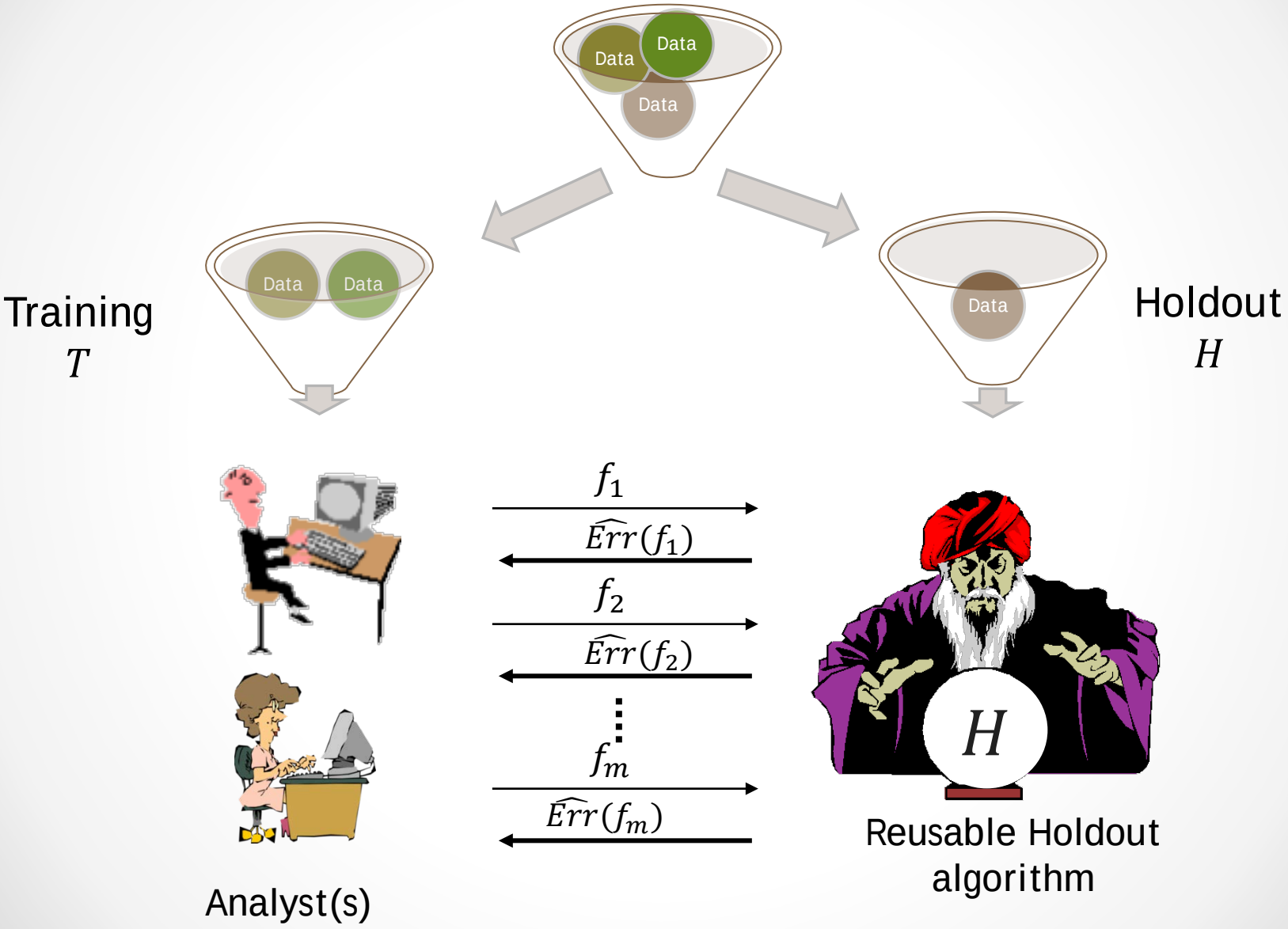
Additional approaches:

- Mutual information [Russo, Zhou 2016]
- KL-stability [Bassily,Nissim,Smith,Steinke,Stemmer,Ullman 2016]
- Typical stability [Bassily,Freund 2016]

Application: holdout validation



Reusable holdout



Thresholdout

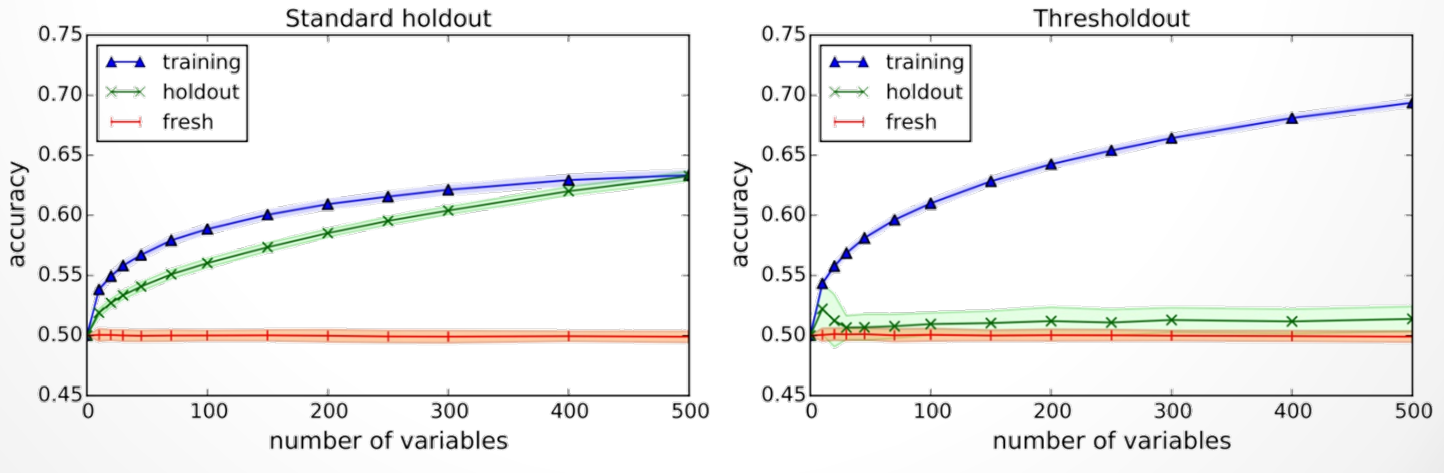
Theorem: Given a holdout set of n i.i.d. samples Thresholdout can estimate the true error of a "large" number of **adaptively chosen** predictors if "few" overfit to the training set

Can be applied to maintaining accurate leaderboard in machine learning competitions [Blum, Hardt 2015]

Illustration

- Given points in $\mathbf{R}^d \times \{-1, +1\}$
- Pick variables correlated with the label on the training set. **Validate on the holdout set**
- Check prediction error of the linear classifier given by $\text{sign}(\sum_{i \in V} s_i x_i)$ where s_i is the sign of the correlation

Data distribution: 10,000 points from $N(0,1)^{10,000}$ randomly labeled



Conclusions

Adaptive data analysis:

- Ubiquitous and useful
- Leads to false discovery and overfitting
- Possible to model and improve on standard approaches

Further work:

New and stronger theory

- Tune to specific applications

Practical applications

The cast

Cynthia Dwork
Harvard

Moritz Hardt
Google Res.

Toni Pitassi
U. of Toronto

Omer Reingold
Stanford

Aaron Roth
U. Penn

Reusable holdout: Preserving validity in adaptive data analysis.
Science, 2015

Preserving Statistical Validity in Adaptive Data Analysis,
Symposium on the Theory of Computing (**STOC**), 2015.

Generalization in Adaptive Data Analysis and Holdout Reuse
Neural Information Processing Systems (**NIPS**), 2015.