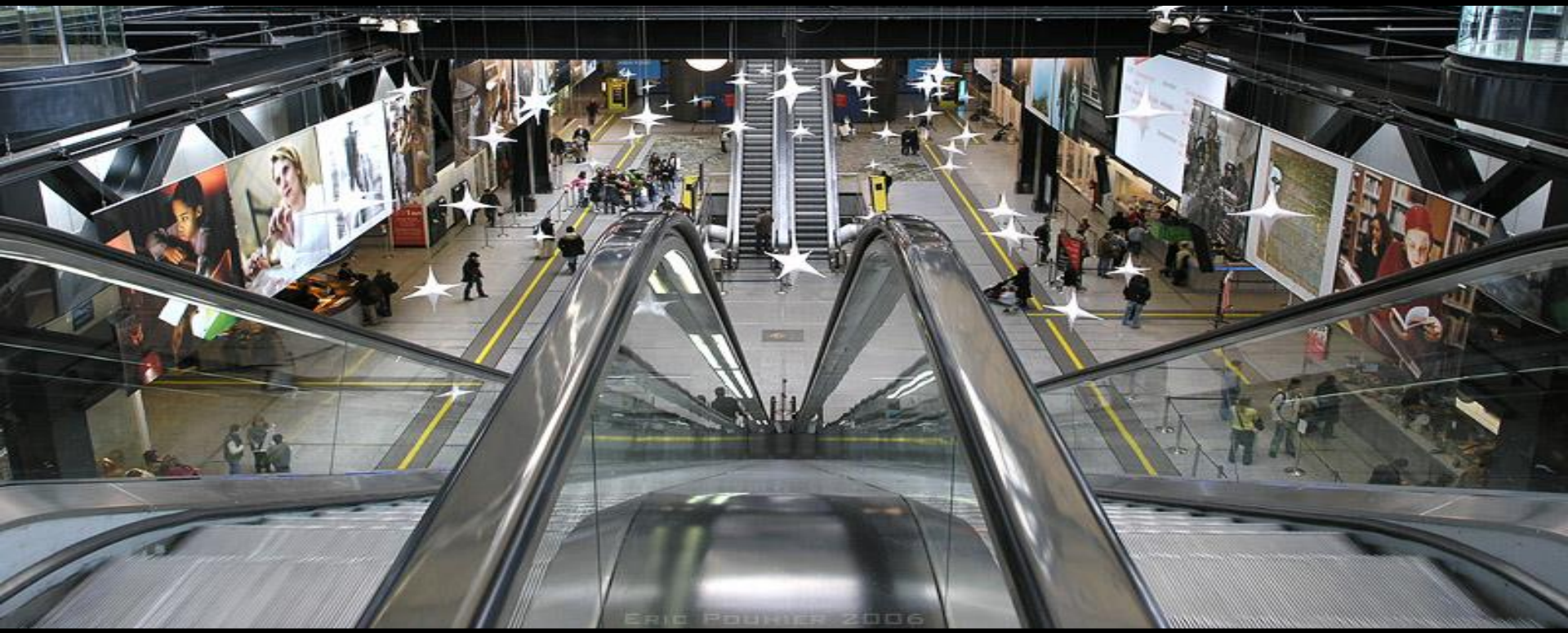
Statistical Significance Testing in Information Retrieval: An Empirical Analysis of Type I, Type II and Type III Errors

Julián Urbano, Harlley Lima, Alan Hanjalic @TU Delft



Picture by dalbera

SIGIR 2019 · July 23rd · Paris

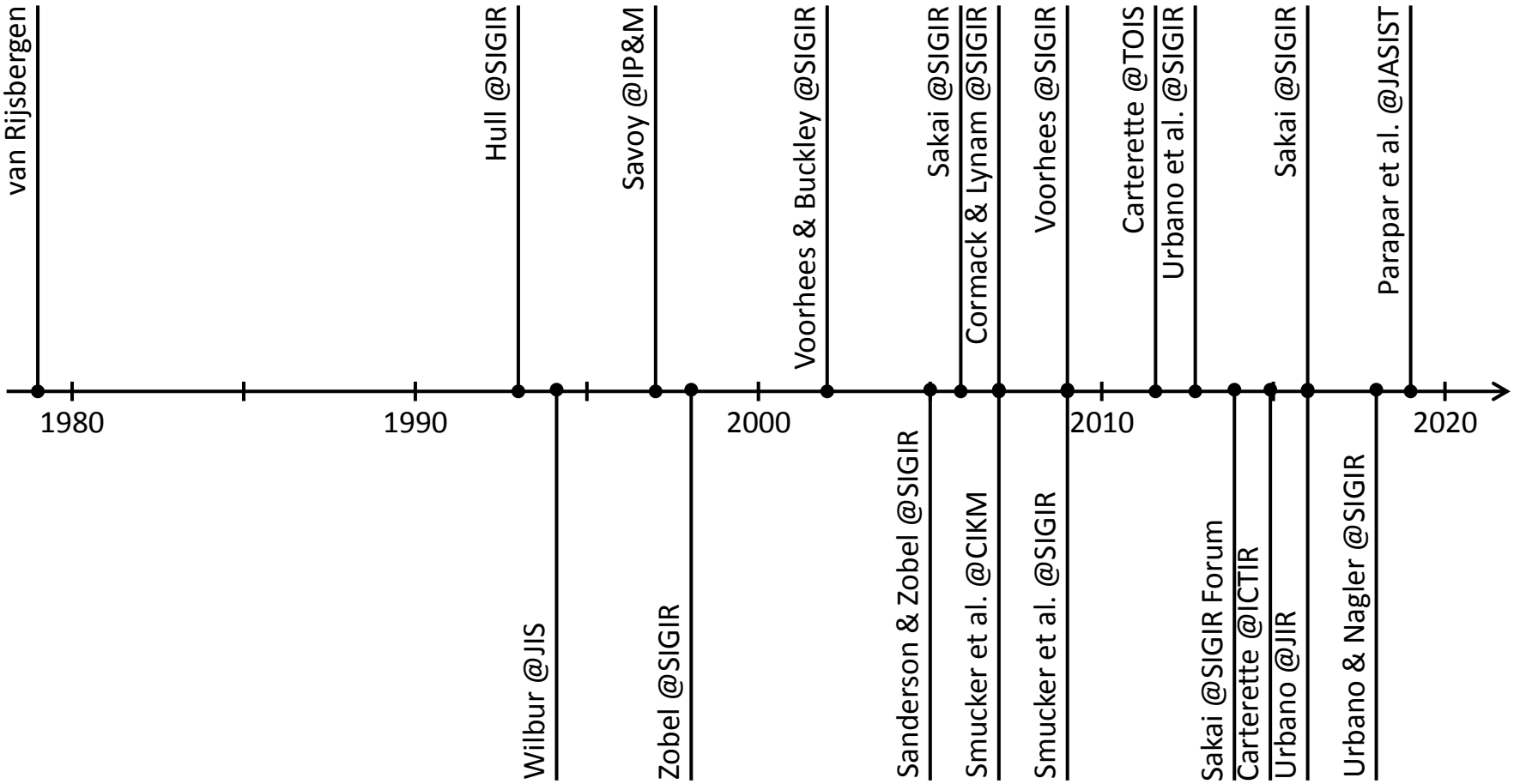
Current Statistical Testing Practice

- According to surveys by Sakai & Carterette
  - 60-75% of IR papers use significance testing
  - In the paired case (2 systems, same topics):
    - 65% use the paired t-test
    - 25% use the Wilcoxon test
    - 10% others, like Sign, Bootstrap & Permutation

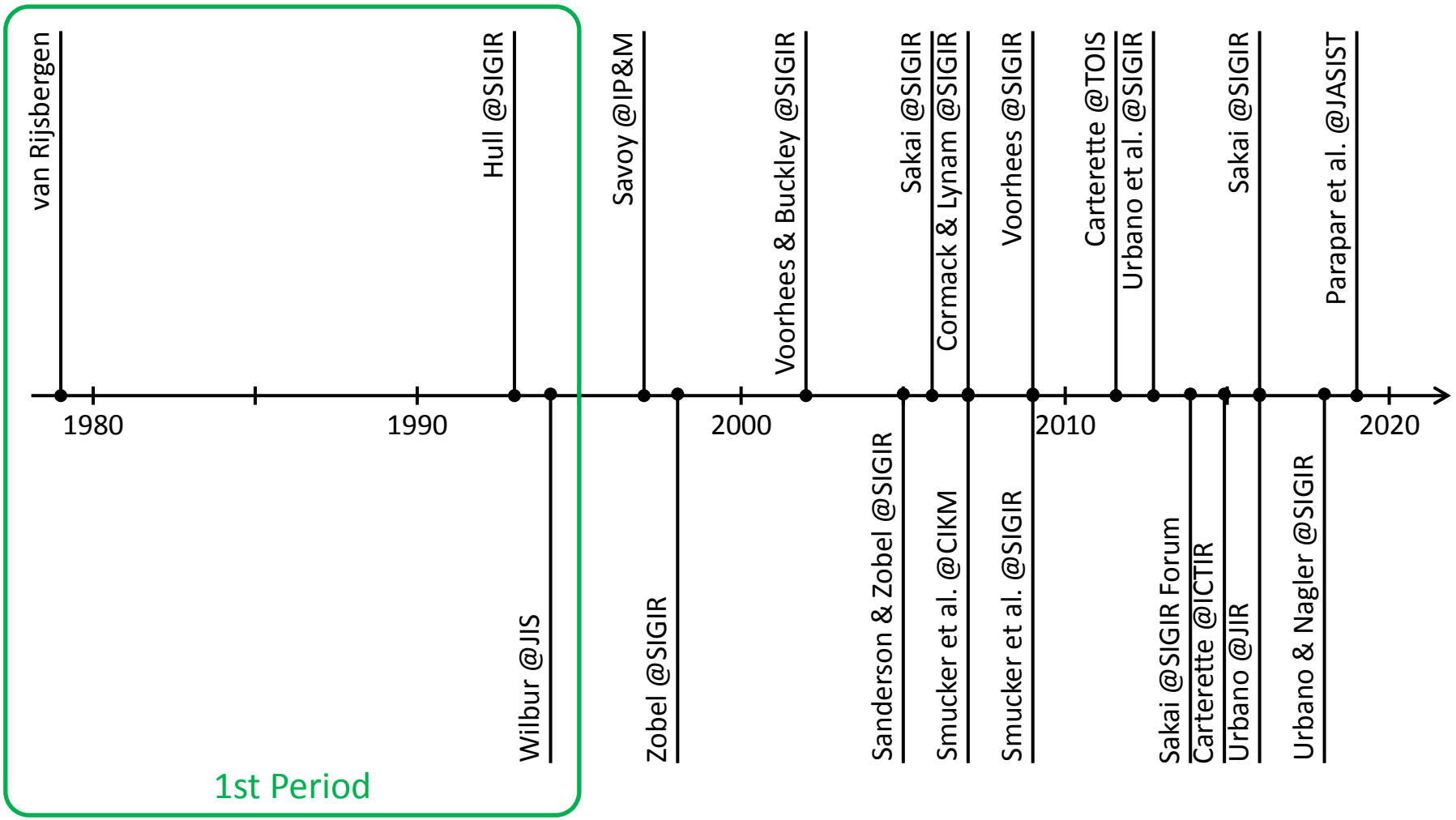
t-test and Wilcoxon are the de facto choice

Is this a *good* choice?

Our Journey


van Rijsbergen
Hull @SIGIR
Savoy @IP&M
Voorhees & Buckley @SIGIR
Sakai @SIGIR
Cormack & Lynam @SIGIR
Voorhees @SIGIR
Carterette @TOIS
Urbano et al. @SIGIR
Sakai @SIGIR
Parapar et al. @JASIST
1980
1990
2000
2010
2020
Wilbur @JIS
Zobel @SIGIR
Sanderson & Zobel @SIGIR
Smucker et al. @CIKM
Smucker et al. @SIGIR
Sakai @SIGIR Forum
Carterette @ICTIR
Urbano @JIR
Urbano & Nagler @SIGIR

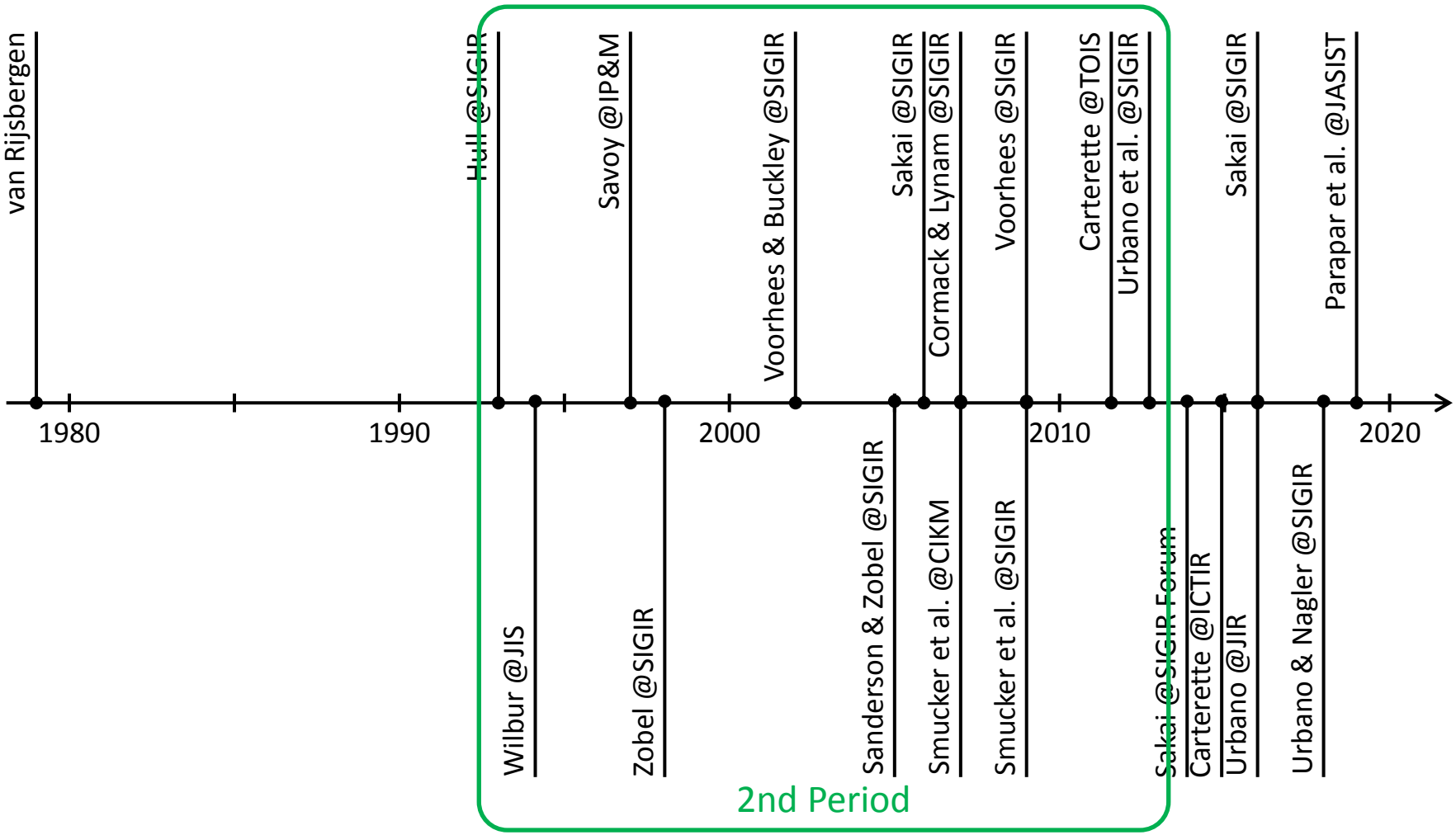
Our Journey



Statistical testing unpopular

Theoretical arguments around test assumptions

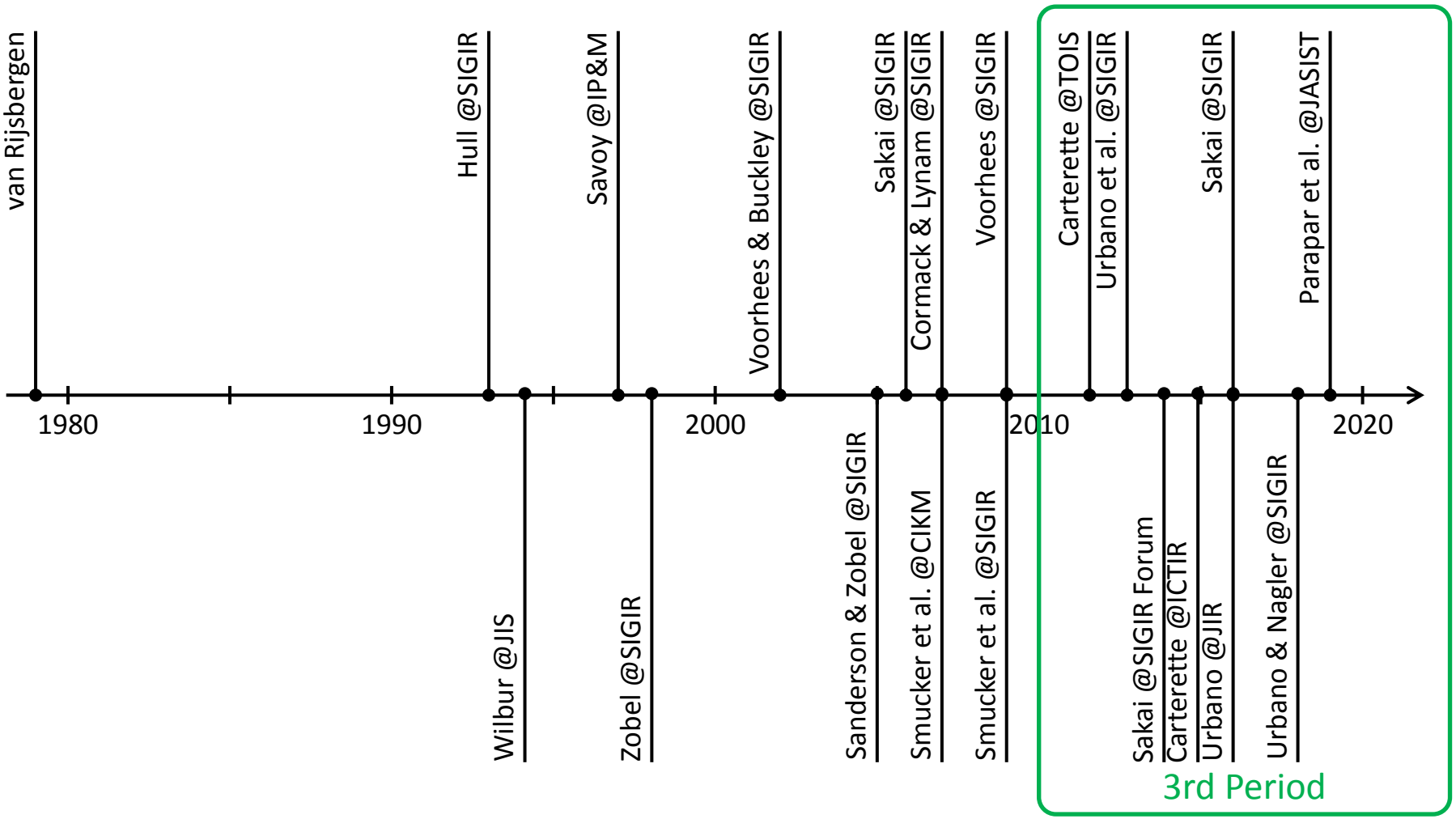
Our Journey



Empirical studies appear

Resampling-based tests and t-test

Our Journey



Wide adoption of statistical testing

Long-pending discussion about statistical practice

Our Journey

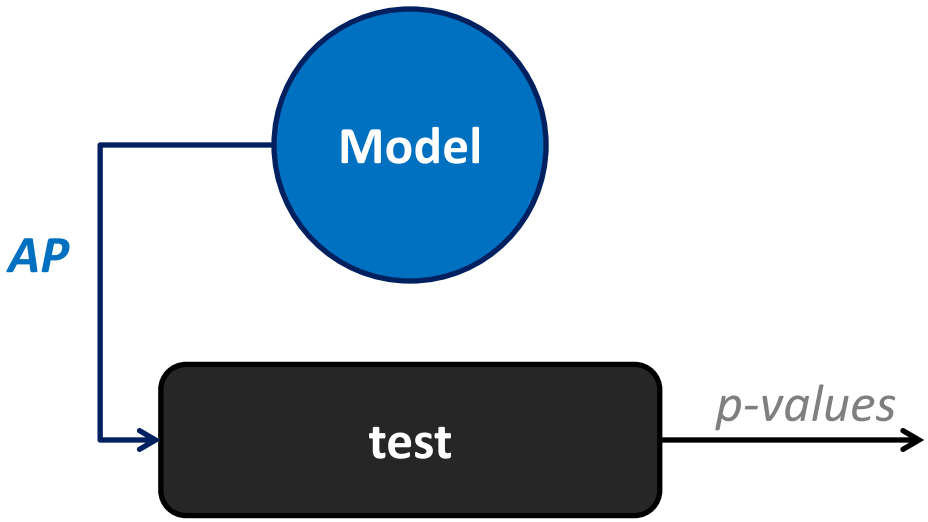
- **Theoretical and empirical** arguments **for and against** specific tests
- 2-tailed tests at α=.05 with AP and P@10, almost exclusively
- **Limited data**, resampling from the same topics
- **No control** over the null hypothesis
- Discordances or conflicts among tests, but **no actual error rates**

Main reason?

No control of the data generating process

PROPOSAL FROM SIGIR 2018

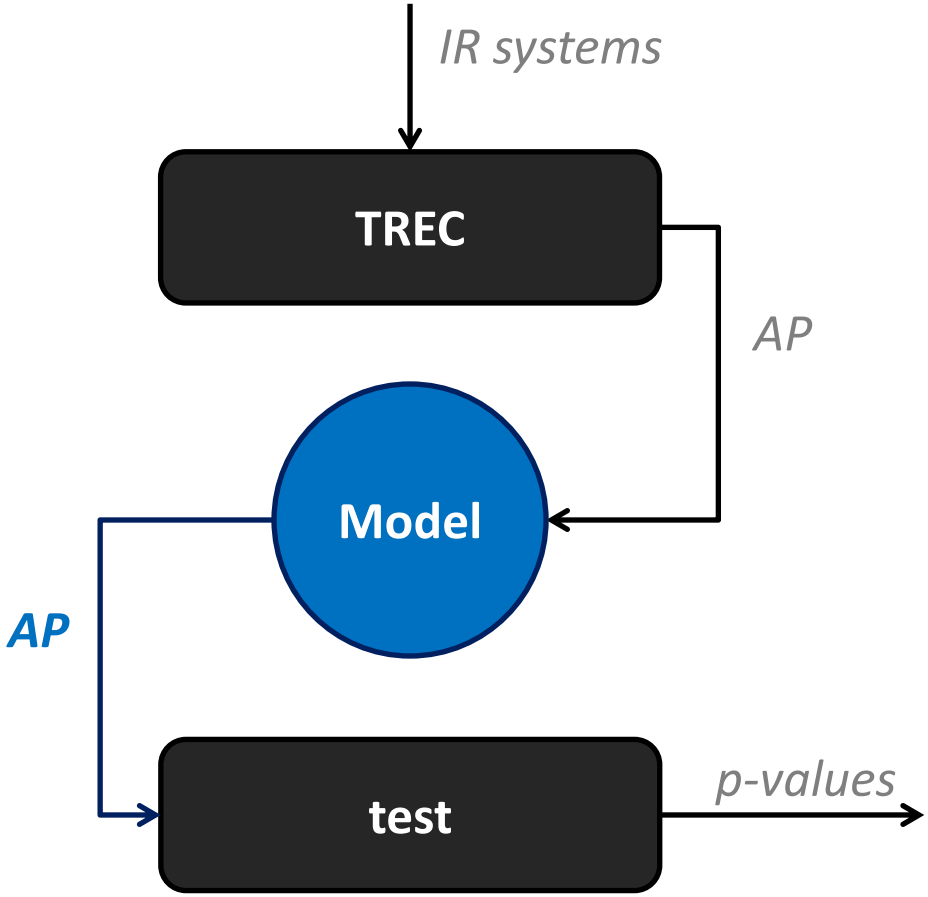
Stochastic Simulation



- Build a **generative model** of the **joint distribution** of system scores
- So that we can simulate scores on new, **random topics** (no content, only scores)
- **Unlimited data**
- **Full control over H_0**

Urbano & Nagler, SIGIR 2018

Stochastic Simulation

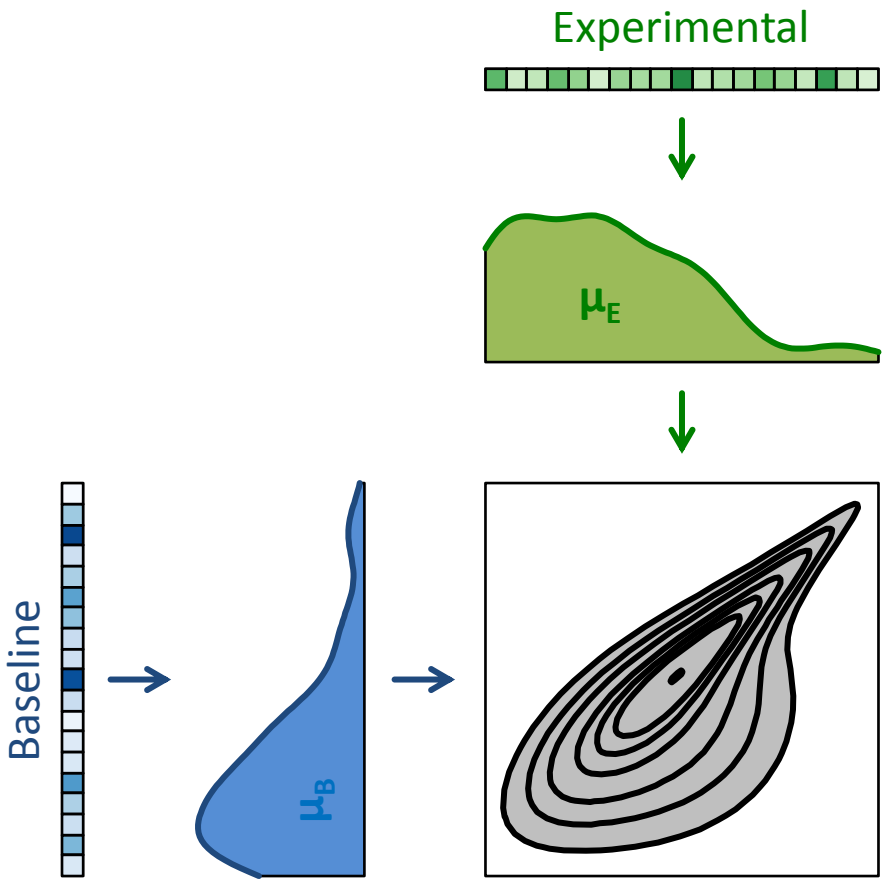


- Build a **generative model** of the **joint distribution** of system scores
- So that we can simulate scores on new, **random topics** (no content, only scores)
- **Unlimited data**
- **Full control over H_0**
- The model is **flexible**, and can be fit to existing data to make it **realistic**

Urbano & Nagler, SIGIR 2018

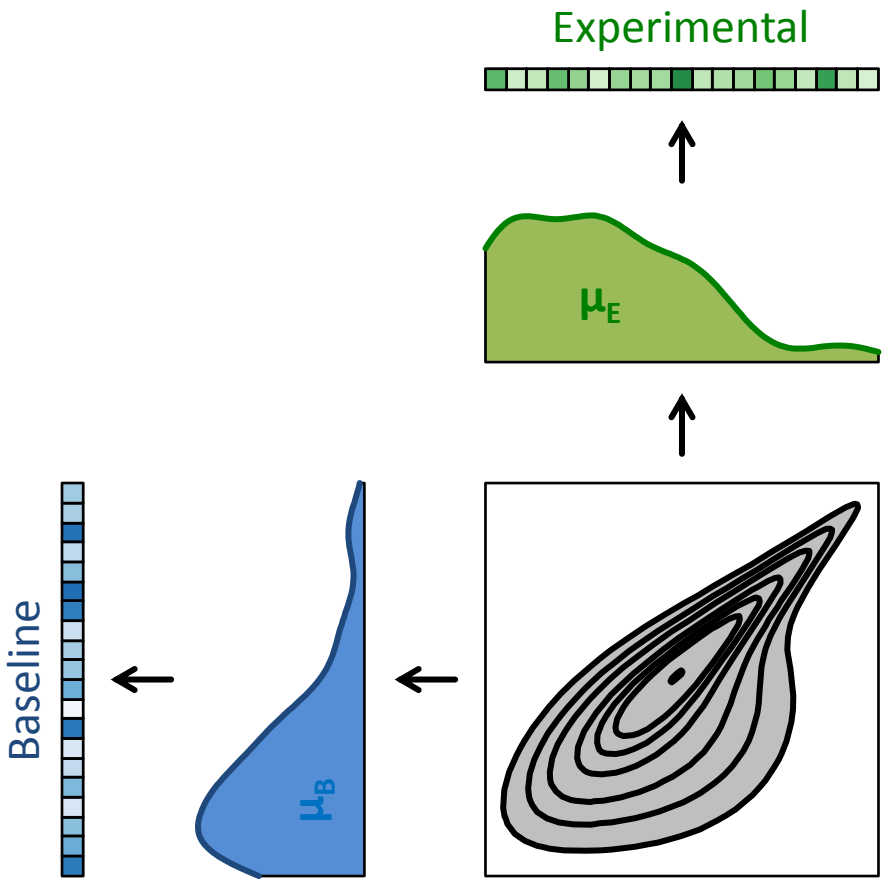
Stochastic Simulation

- We use copula models, which separate:
  1. **Marginal distributions**, of individual systems
     - Give us full knowledge and control over H_0
  2. **Dependence structure**, among systems



Stochastic Simulation

- We use copula models, which separate:
  1. **Marginal distributions**, of individual systems
     - Give us full knowledge and control over H_0
  2. **Dependence structure**, among systems



Research Question

- Which is the test that…
  1. Maintaining Type I errors at the α level,
  2. Has the highest statistical power,
  3. Across measures and sample sizes,
  4. **With IR-like data?**

Factors Under Study

- **Paired test:** Student's t, Wilcoxon, Sign, Bootstrap-shift, Permutation
- **Measure:** AP, nDCG@20, ERR@20, P@10, RR
- **Topic set size n:** 25, 50, 100
- **Effect size δ:** 0.01, 0.02, …, 0.1
- **Significance level α:** 0.001, …, 0.1
- **Tails:** 1 and 2

- Data to fit stochastic models: TREC 5-8 Ad Hoc and 2010-13 Web

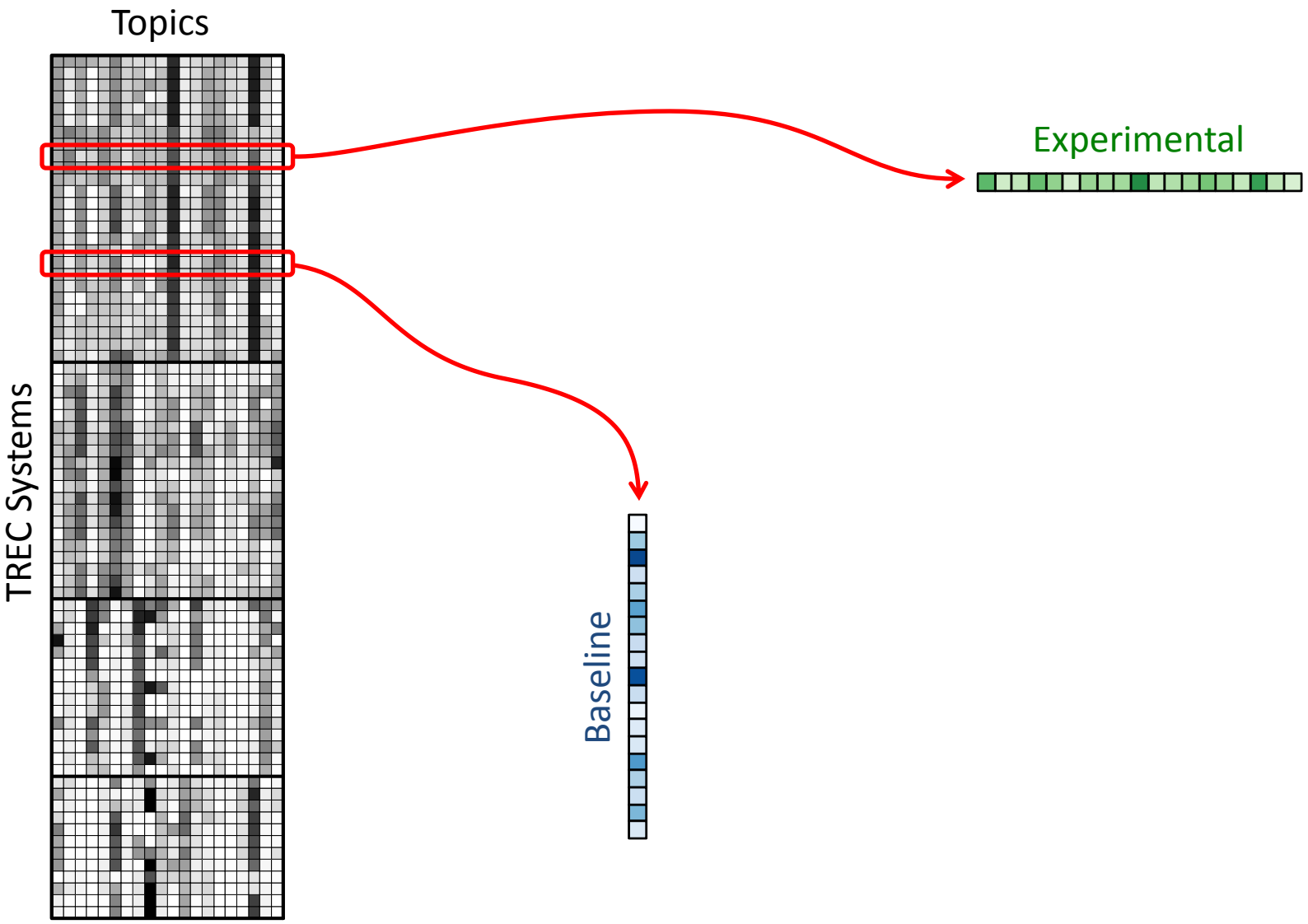
We report results on >500 million p-values

1.5 years of CPU time

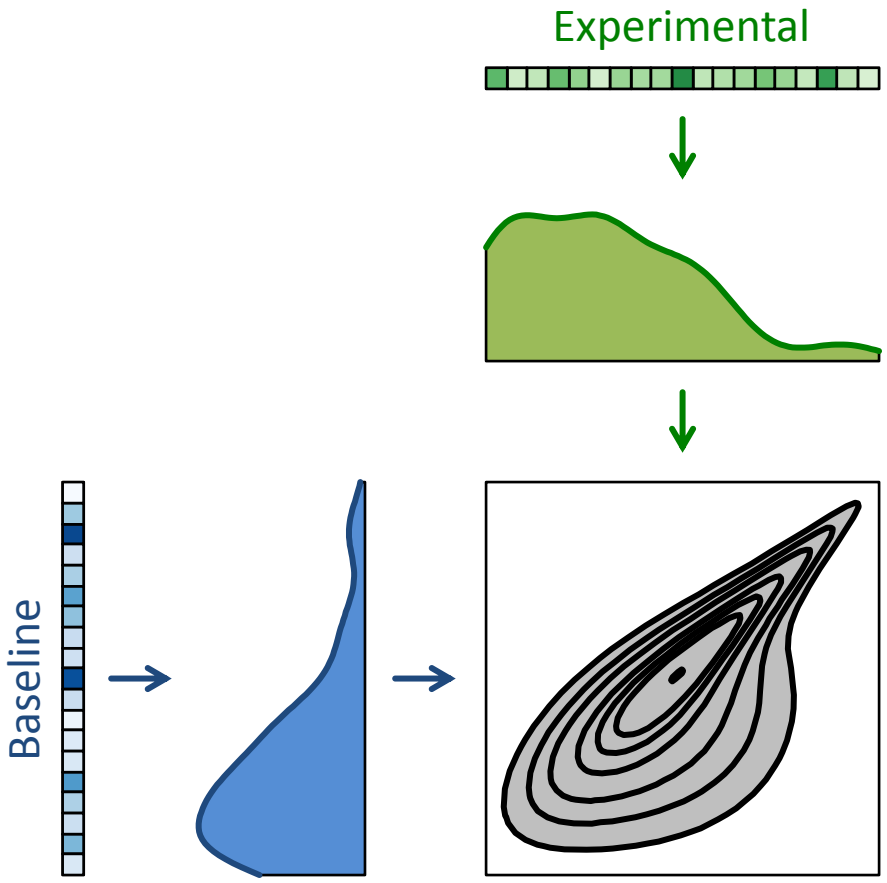
¯_(ツ)_/¯

TYPE I ERRORS

Simulation such that $\mu_E = \mu_B$

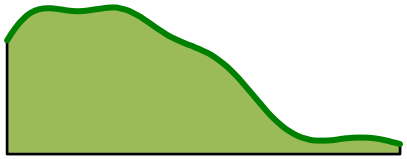


Simulation such that $\mu_E = \mu_B$

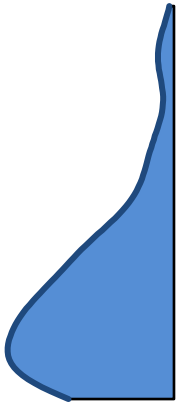


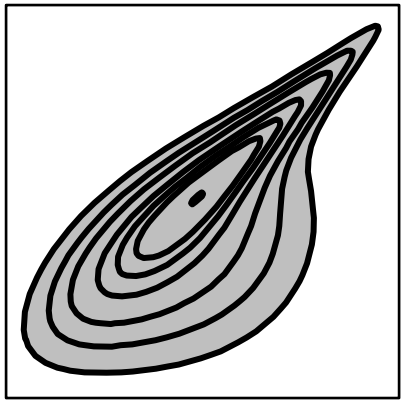
Simulation such that $\mu_E = \mu_B$

Experimental



Baseline

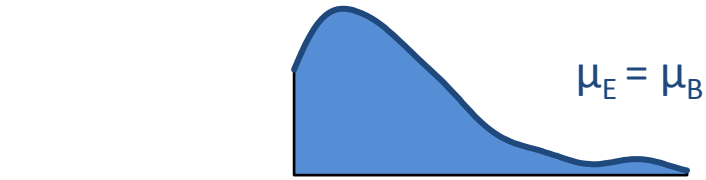




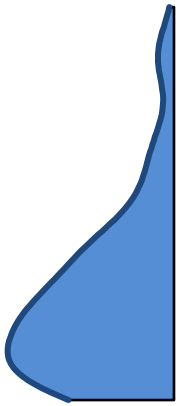
Simulation such that $\mu_E = \mu_B$

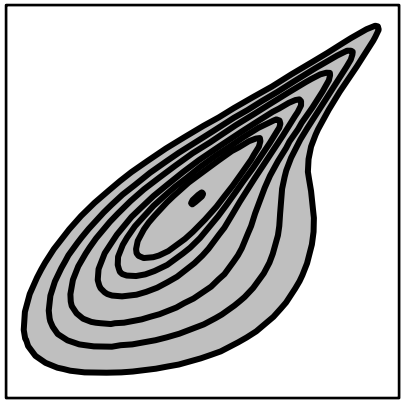
Experimental

$\mu_E = \mu_B$

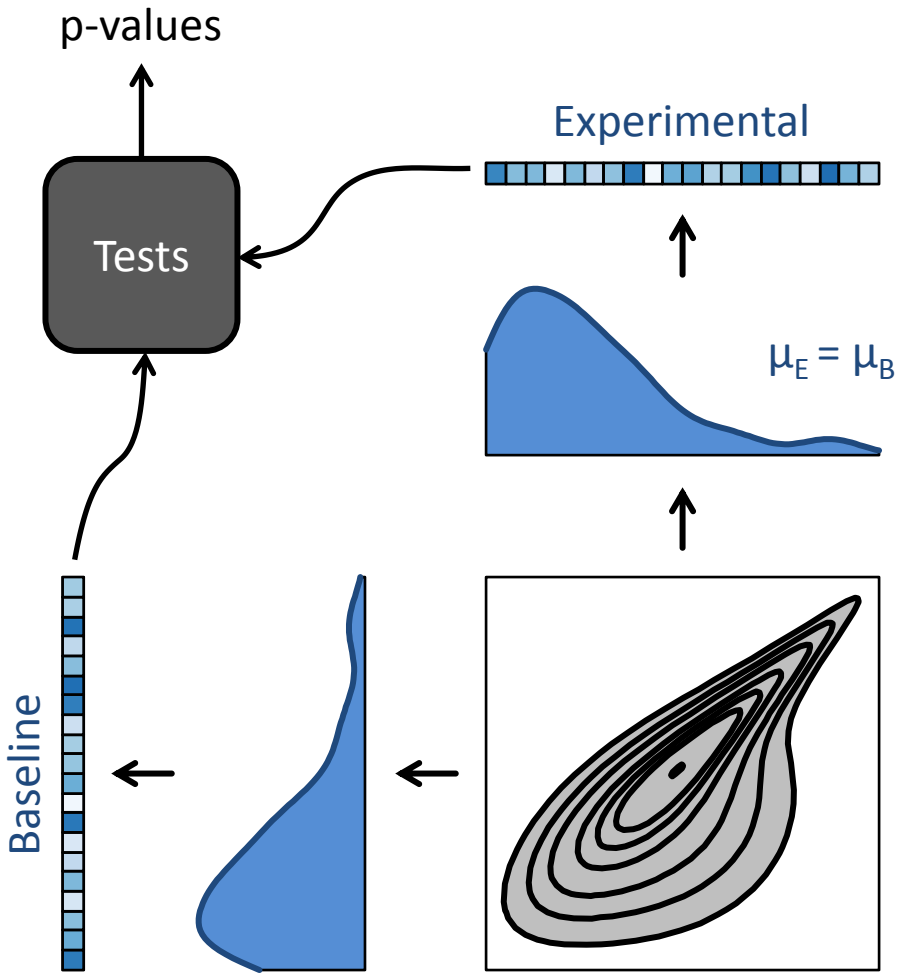


Baseline





Simulation such that $\mu_E = \mu_B$

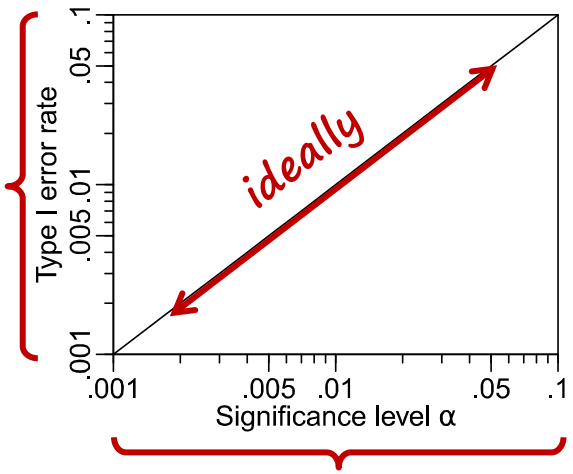


Simulation such that $\mu_E = \mu_B$

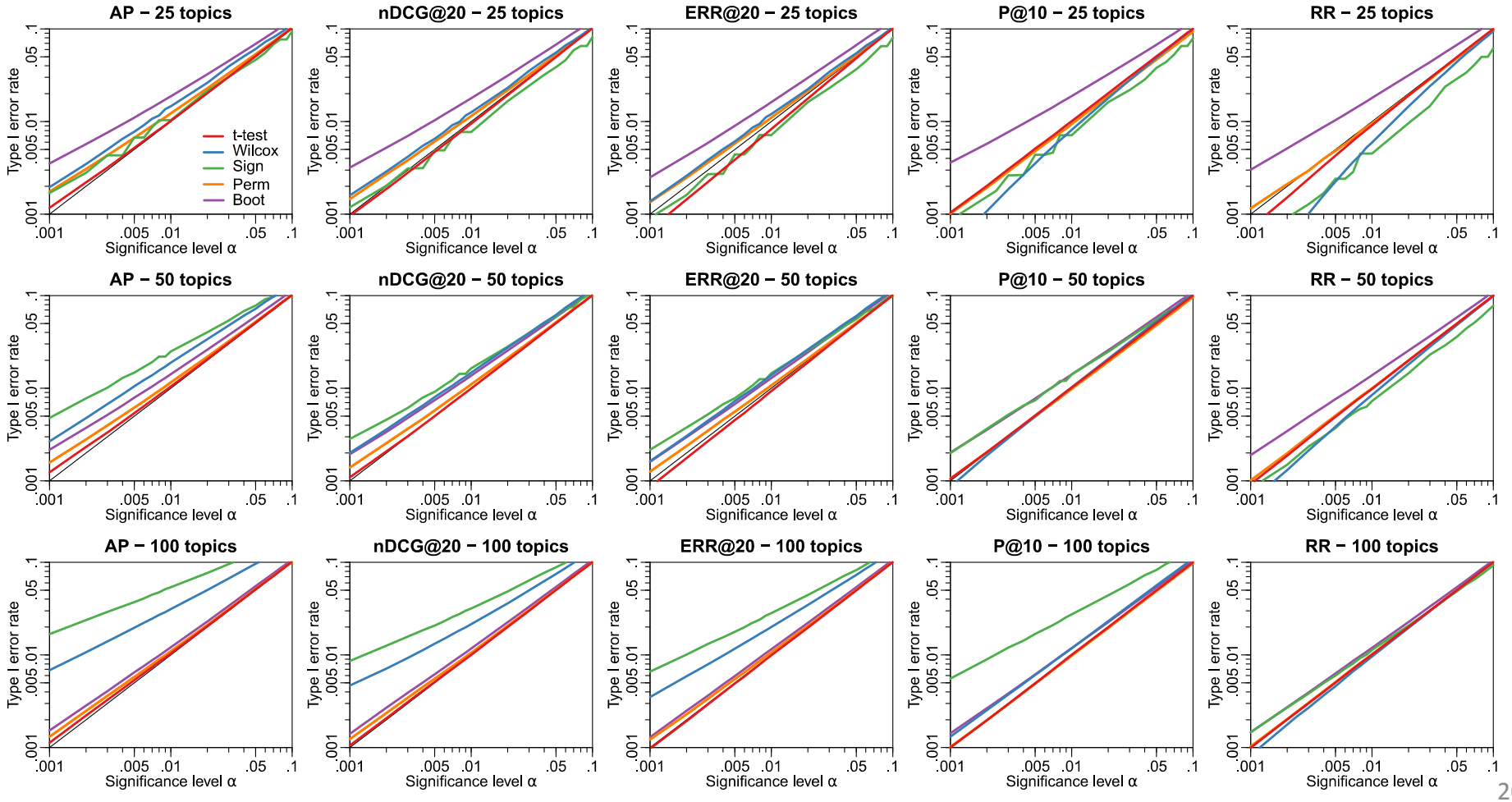
- Repeat for each measure and topic set size n
  - 1,667,000 times
  - ≈8.3 million 2-tailed p-values
  - ≈8.3 million 1-tailed p-values
- Grand total of >250 million p-values
- **Any $p<\alpha$ corresponds to a Type I error**

Type I Errors by α | n 2-tailed

Not so interested in specific points but in trends

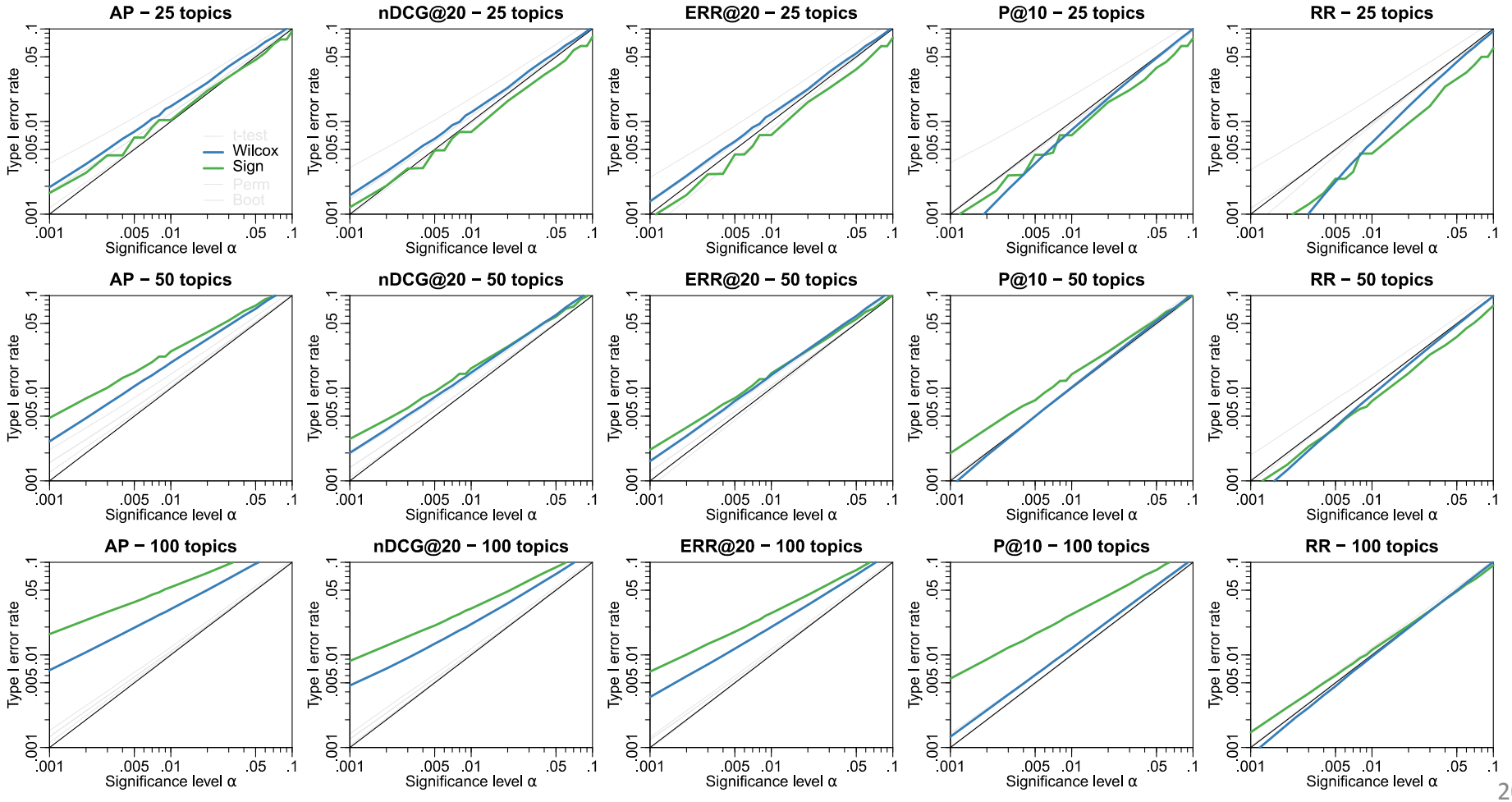


Type I Errors by α | n 2-tailed



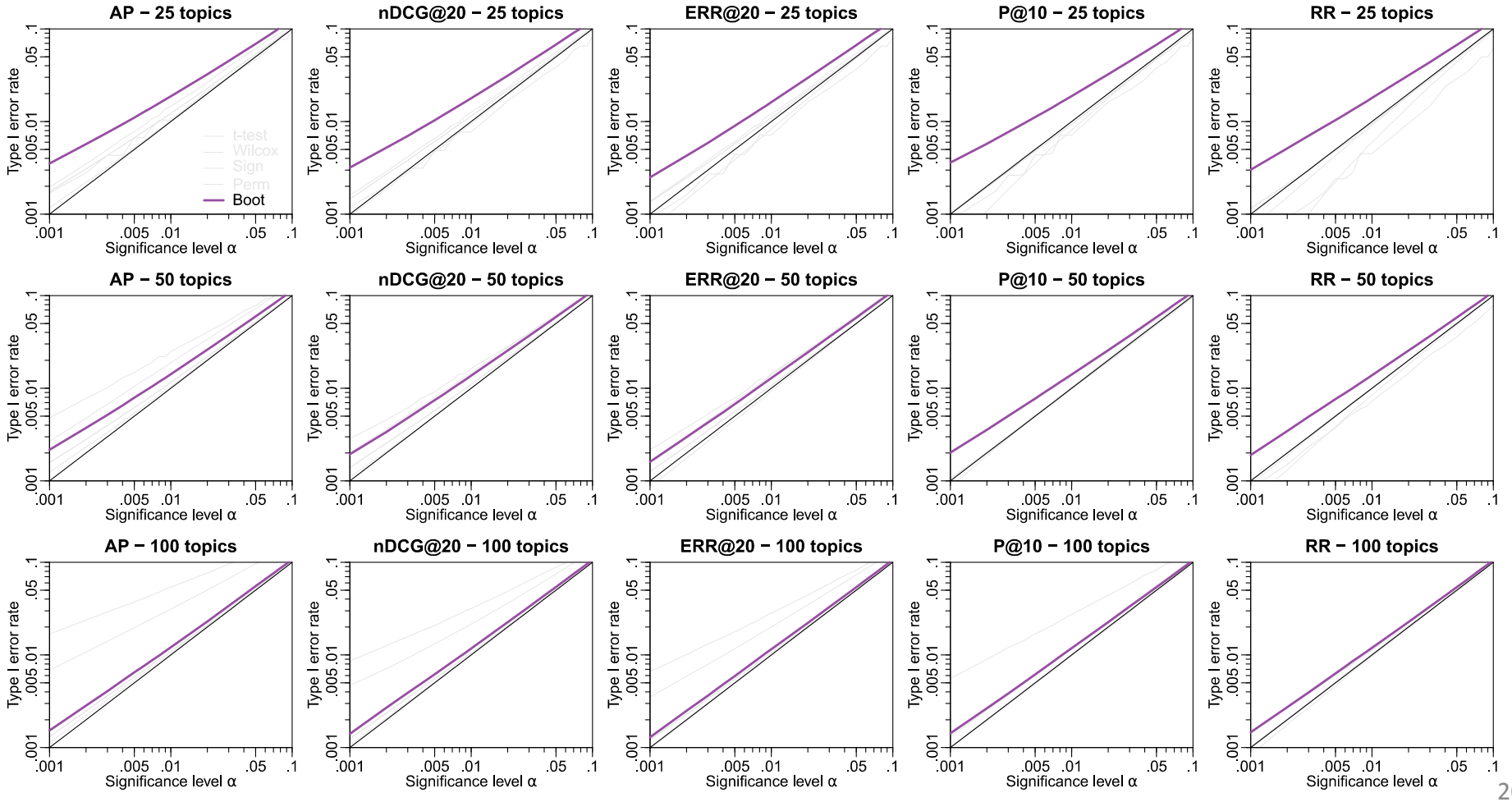
Type I Errors by α | n 2-tailed

- Wilcoxon and Sign have higher error rates than expected
- Wilcoxon better in P@10 and RR because of symmetricity
- **Even worse as sample size increases (with RR too)**



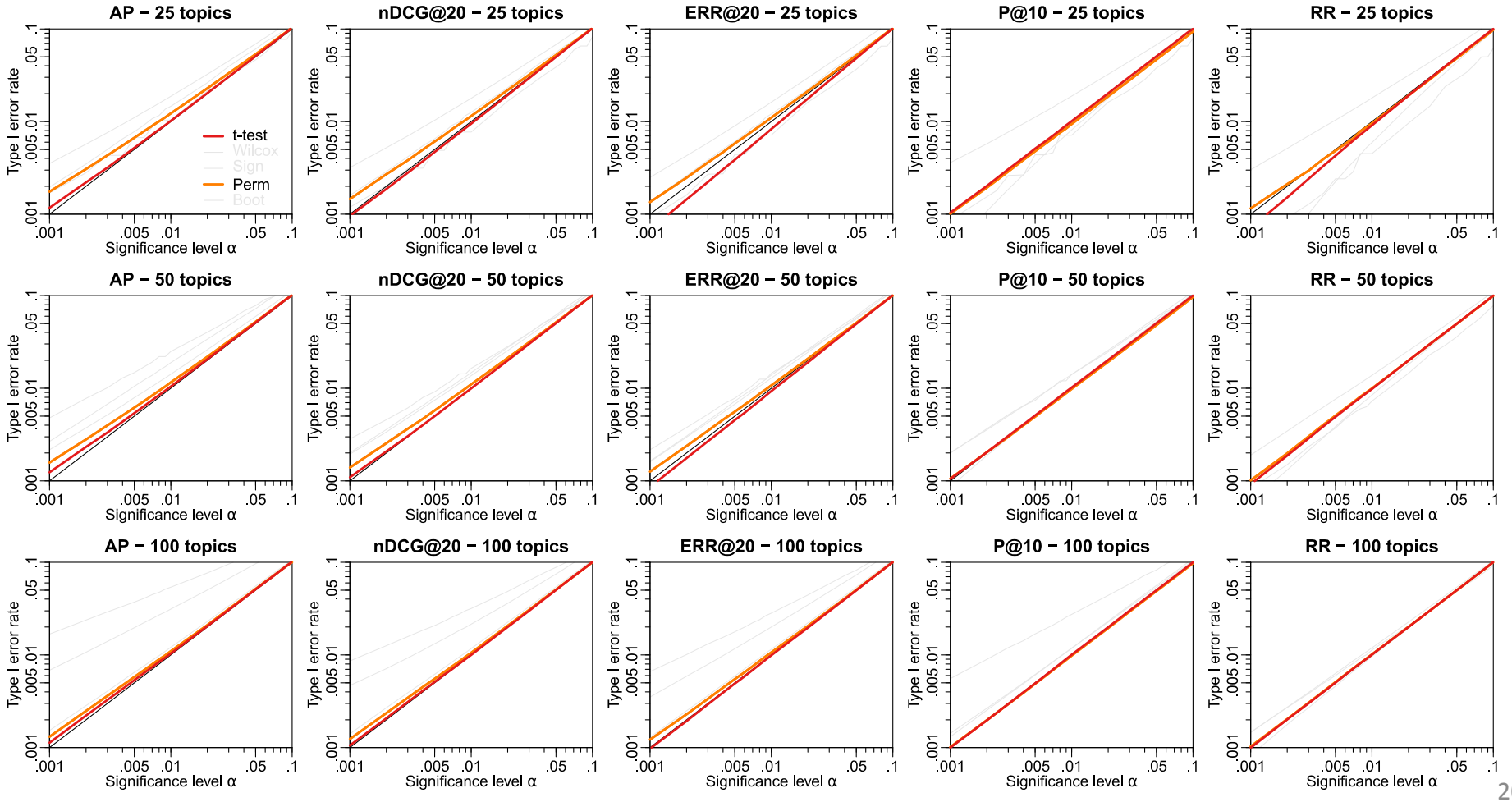
Type I Errors by α | n 2-tailed

- **Bootstrap has high error rates too**
- Tends to correct with sample size because it estimates the sampling distribution better



Type I Errors by α | n 2-tailed

- **Permutation and t-test have nearly ideal behavior**
- Permutation very slightly sensitive to sample size
- t-test remarkably robust to it



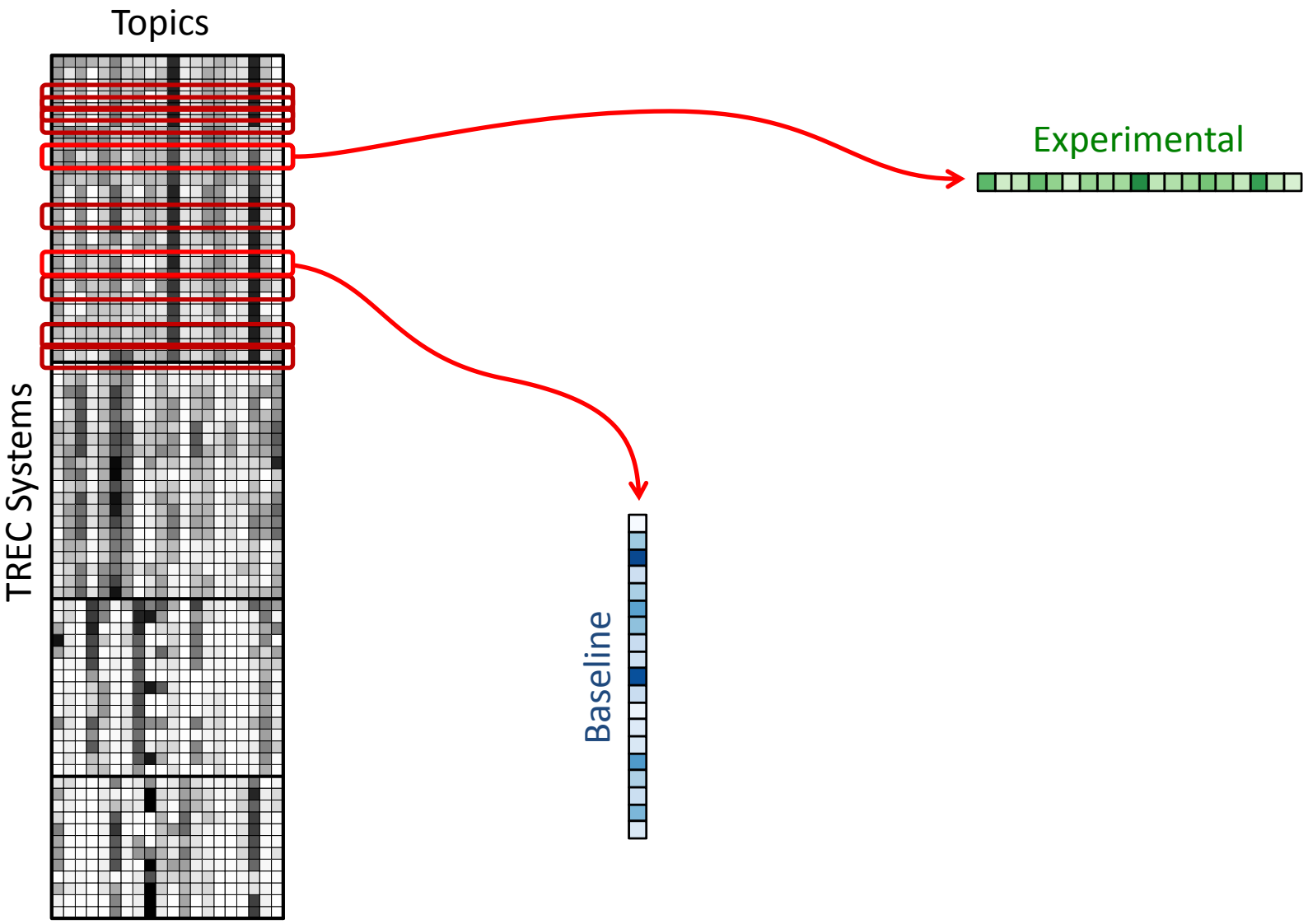
Type I Errors - Summary

- Wilcoxon, Sign and Bootstrap test tend to make more errors than expected
- Increasing sample size helps Bootstrap, but hurts Wilcoxon and Sign even more
- Permutation and t-test have nearly ideal behavior across measures, even with small sample size
- **t-test is remarkably robust**

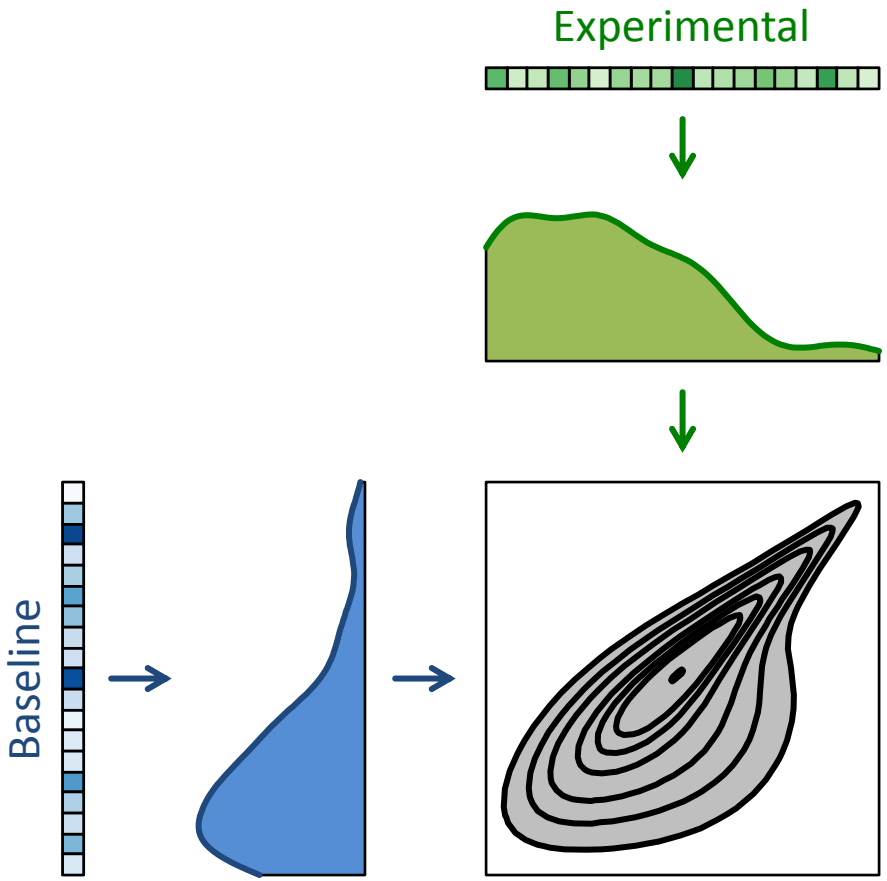
- Same conclusions with 1-tailed tests

TYPE II ERRORS

Simulation such that $\mu_E = \mu_B + \delta$

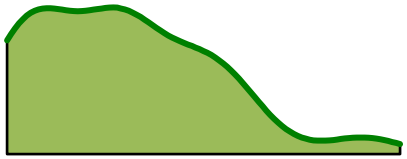


Simulation such that $\mu_E = \mu_B + \delta$

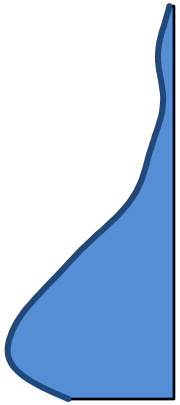


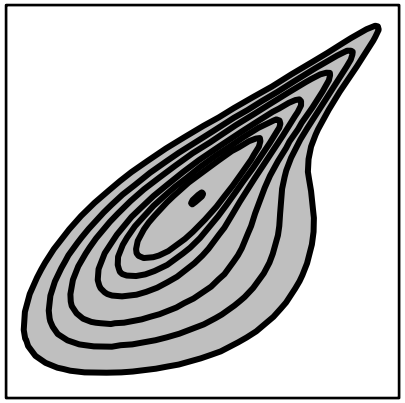
Simulation such that $\mu_E = \mu_B + \delta$

Experimental



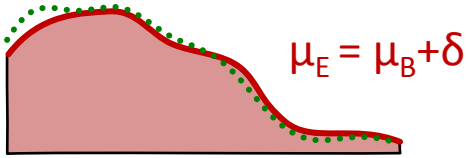
Baseline



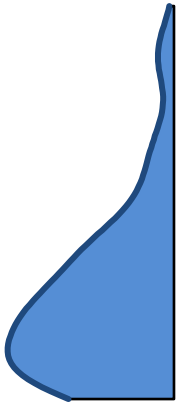


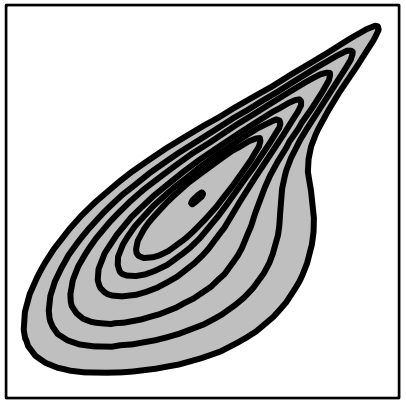
Simulation such that $\mu_E = \mu_B + \delta$

Experimental

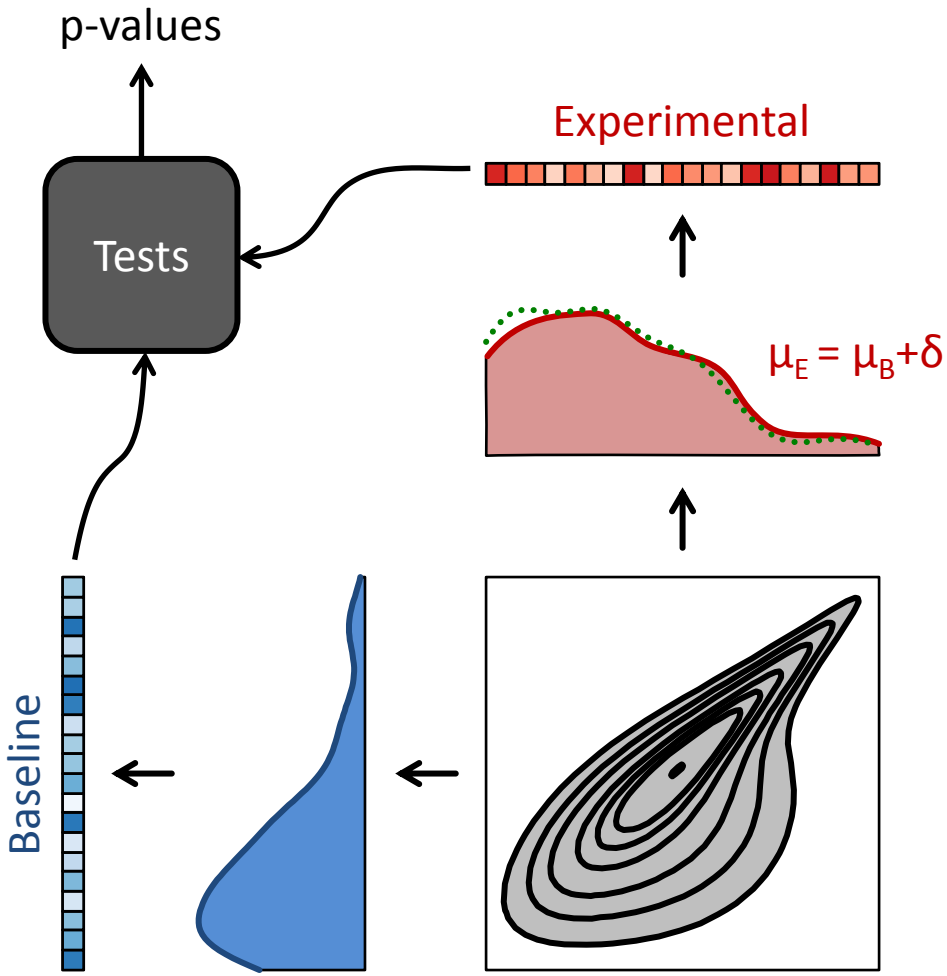


Baseline





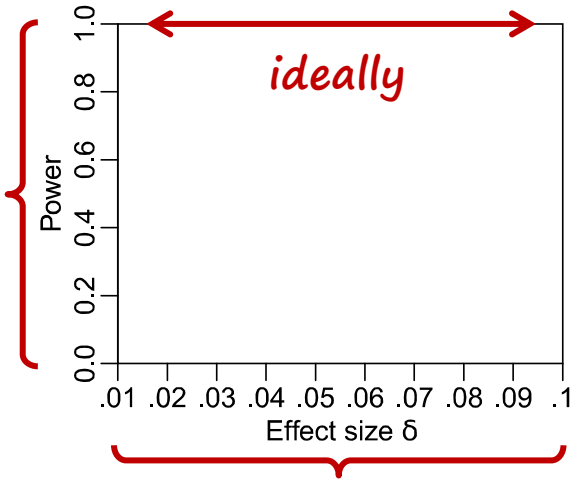
Simulation such that $\mu_E = \mu_B + \delta$



Simulation such that $\mu_E = \mu_B + \delta$

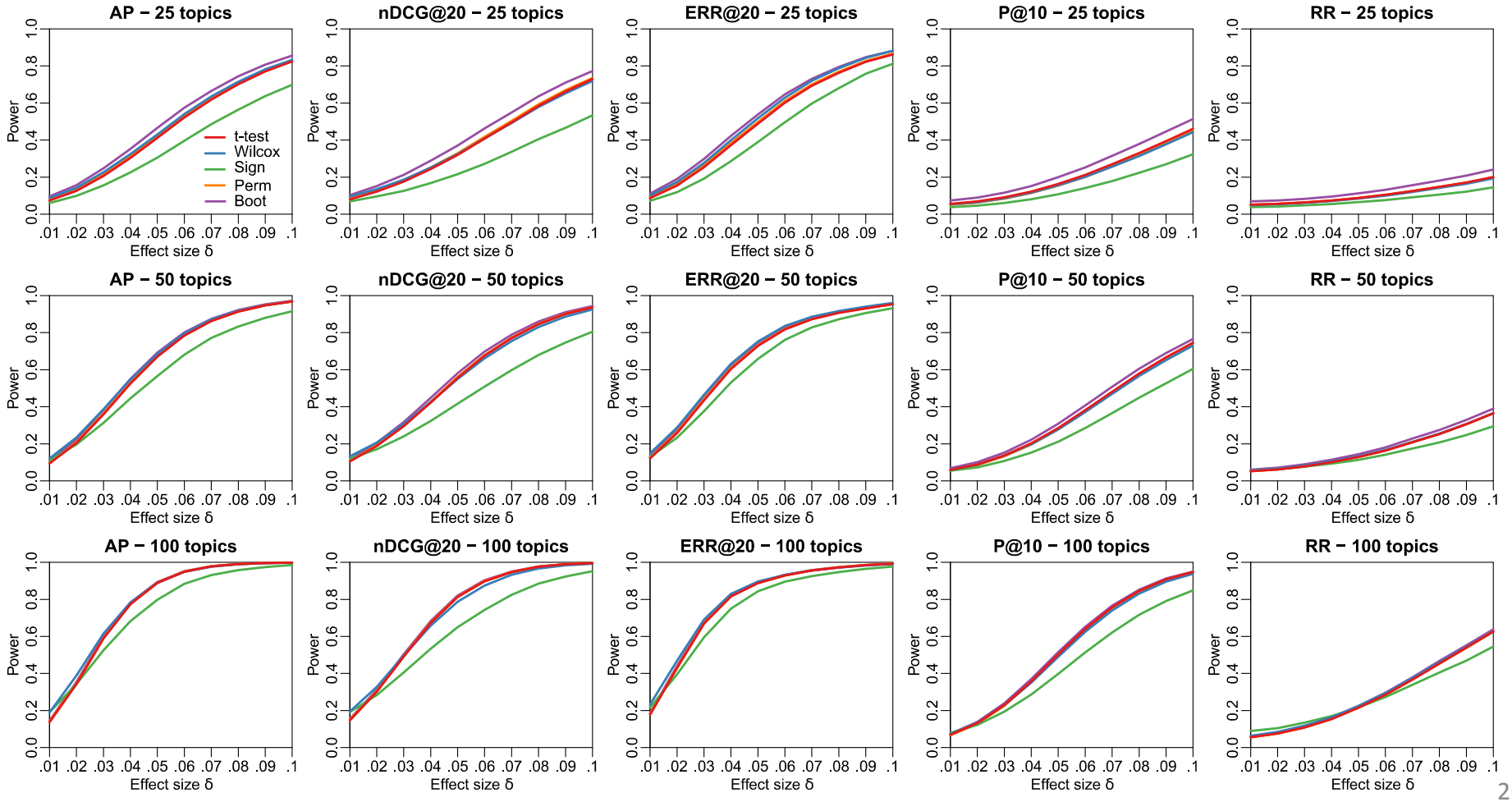
- Repeat for each measure, topic set size n and effect size δ
  - 167,000 times
  - ≈8.3 million 2-tailed p-values
  - ≈8.3 million 1-tailed p-values
- Grand total of >250 million p-values
- **Any $p>\alpha$ corresponds to a Type II error**

Power by δ | n α=.05, 2-tailed



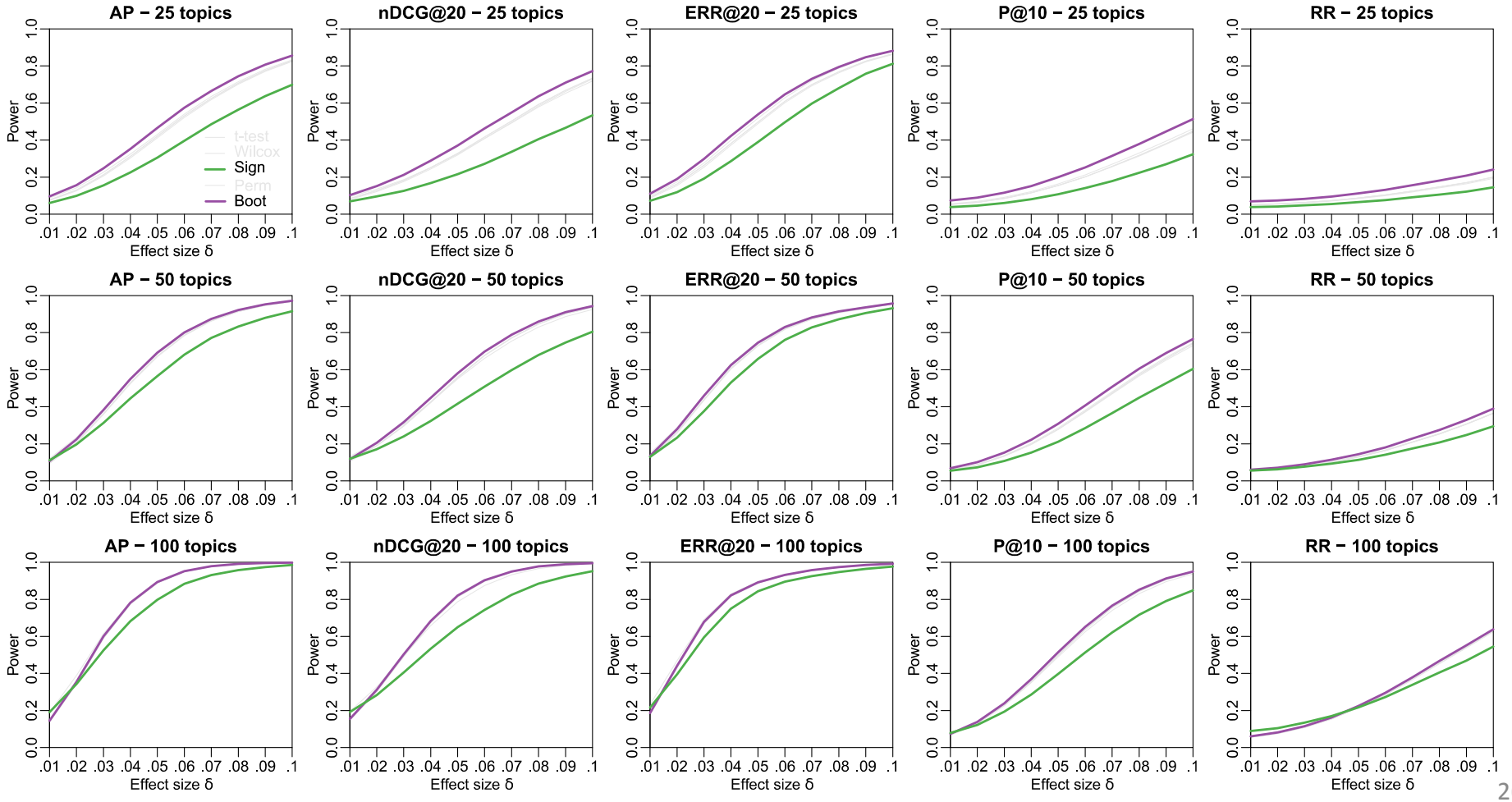
Power by δ | n α=.05, 2-tailed

- Clear effect of effect size δ
- Clear effect of sample size n
- Clear effect of measure (via σ)



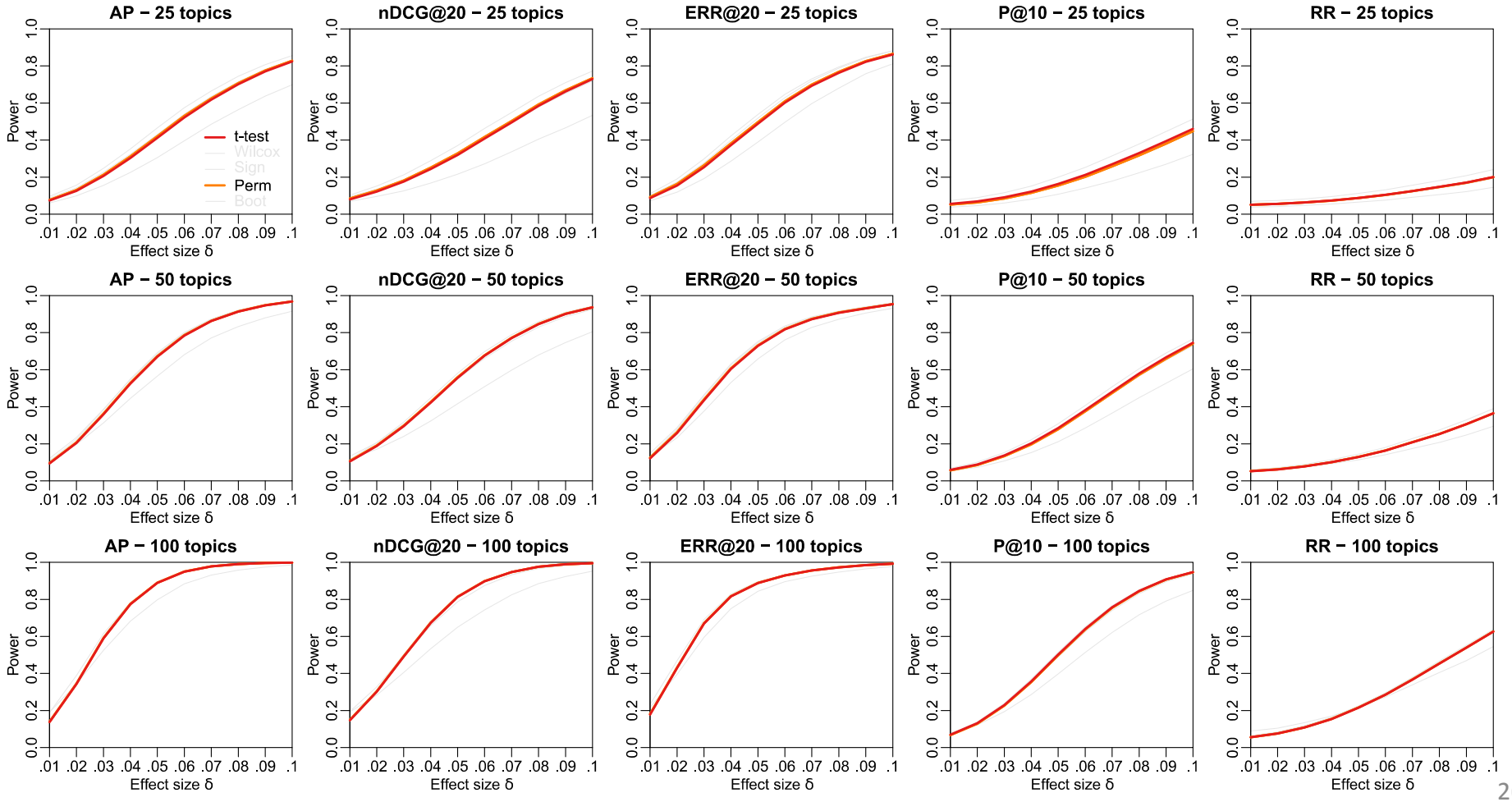
Power by δ | n α=.05, 2-tailed

- Sign test consistently the least powerful (disregards magnitudes)
- **Bootstrap test consistently the most powerful, specially for small n**



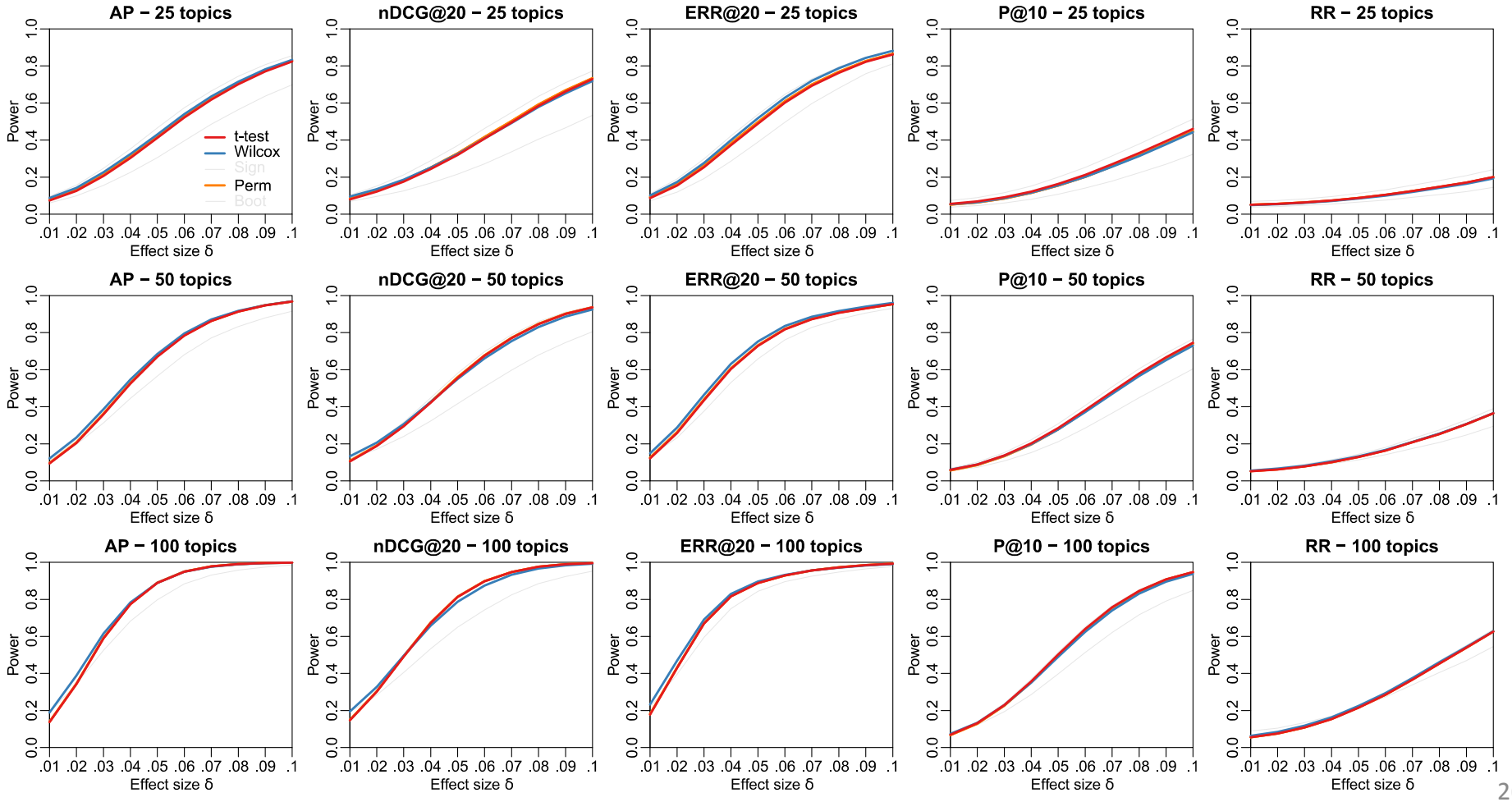
Power by δ | n α=.05, 2-tailed

- Permutation and t-test are almost identical again
- Very close to Bootstrap as sample size increases

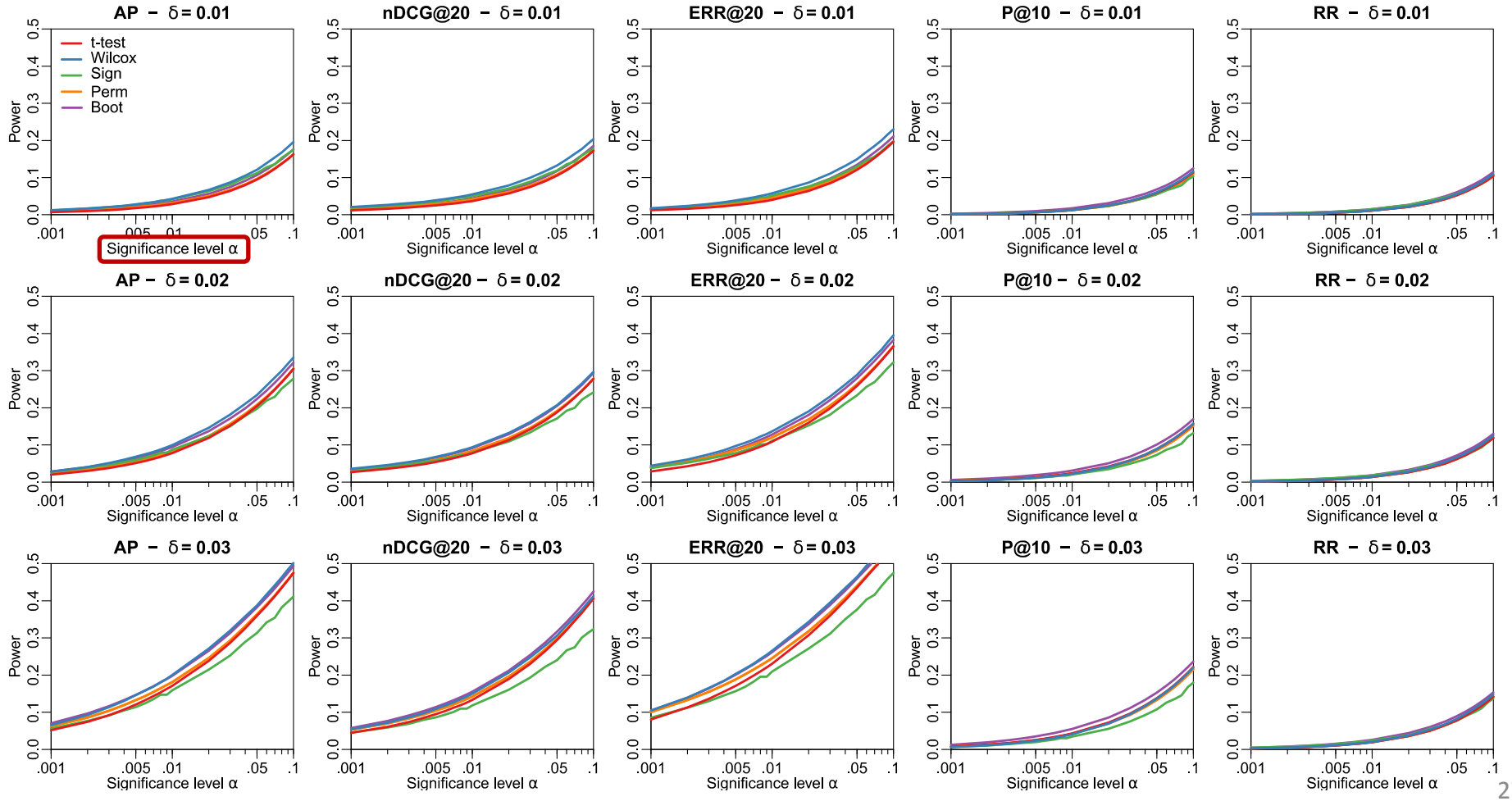


Power by δ | n α=.05, 2-tailed

- Wilcoxon is very similar to Permutation and t-test
- Even slightly better with small n or δ, specially for AP, nDCG and ERR (it's indeed more efficient with some asymmetric distributions)

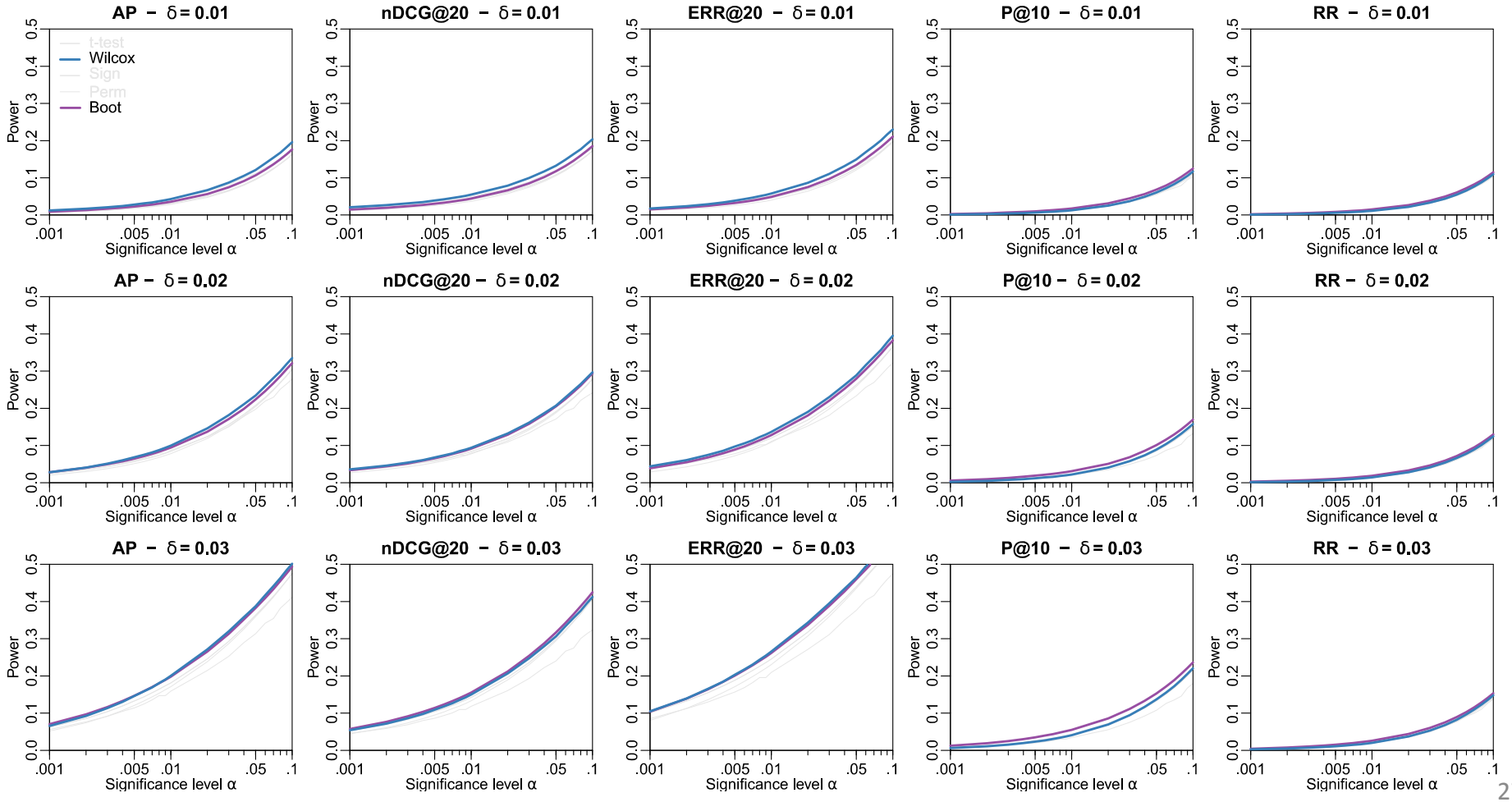


Power by α | δ n=50, 2-tailed



Power by α | δ n=50, 2-tailed

- With small δ, Wilcoxon and Bootstrap consistently the most powerful
- With large δ, Permutation and t-test catch up with Wilcoxon



Type II Errors - Summary

- All tests, except Sign, behave very similarly

- Bootstrap and Wilcoxon are consistently a bit more powerful across significance levels
  - But more Type I errors!
- With larger effect sizes and sample sizes, Permutation and t-test catch up with Wilcoxon, but not with Bootstrap

- Same conclusions with 1-tailed tests

TYPE III ERRORS

Type III what?

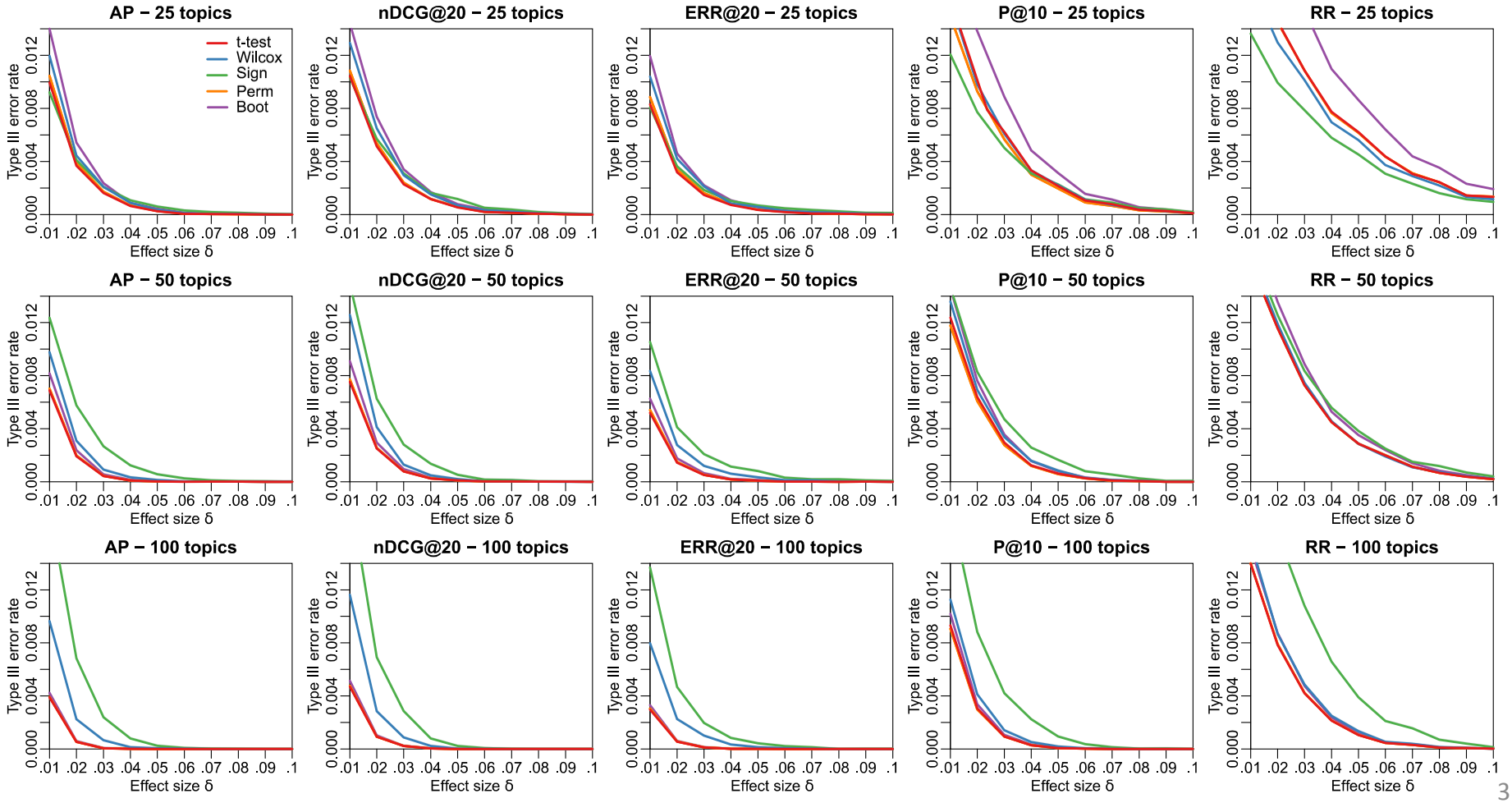
- A wrong directional decision based on the correct rejection of a non-directional hypothesis

- Example:
  - We observe a positive result, $\bar{E} > \bar{B}$
  - We run a 2-tailed test, $H_0: \mu_E = \mu_B$
  - Find $p < \alpha$, so we reject and conclude $\mu_E > \mu_B$

  - **But H_0 is non-directional**
  - **What if we just got lucky, and really $\mu_E < \mu_B$?**

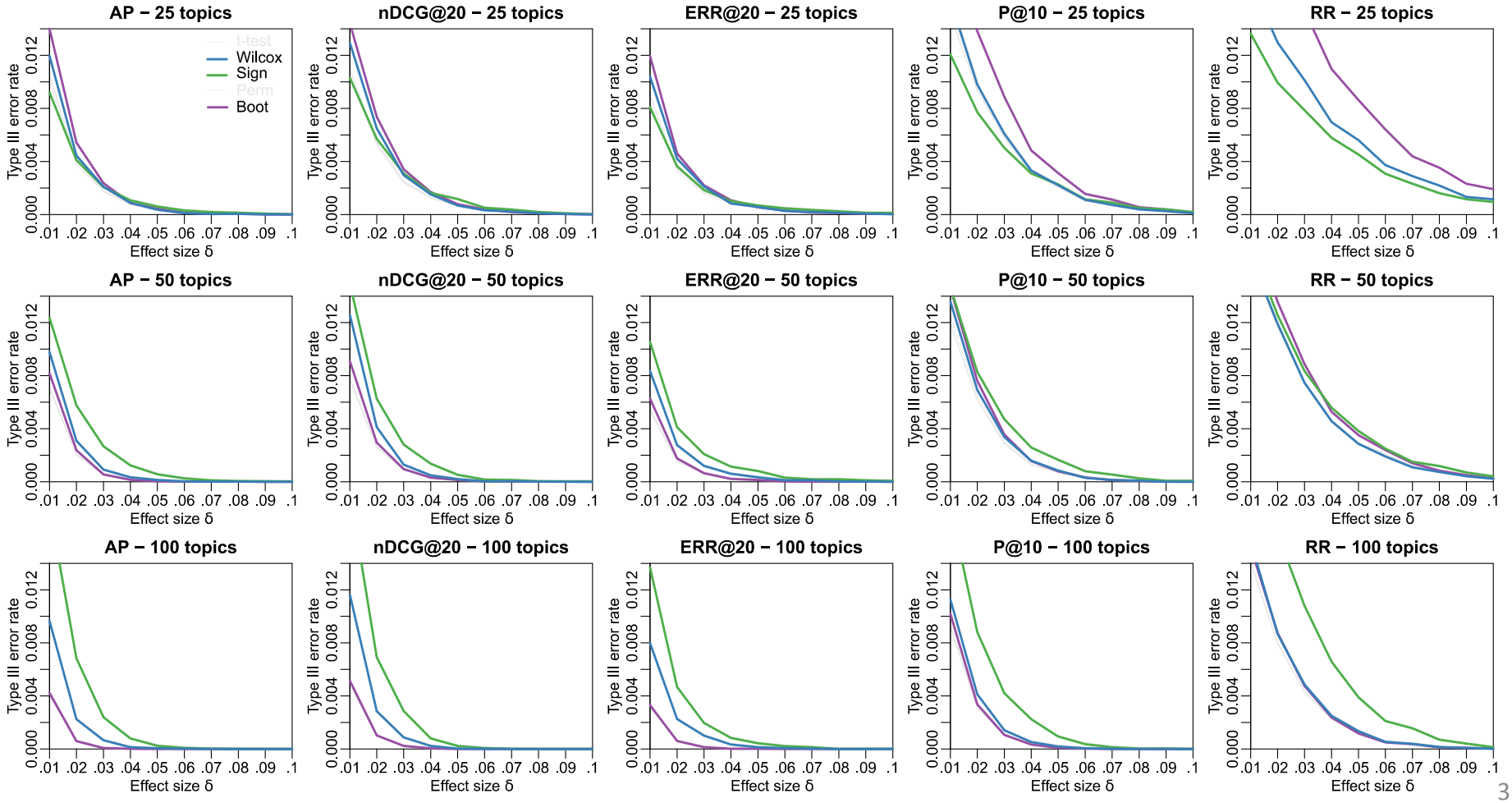
Type III Errors by δ | n α=.05

- Clear effect of δ and n
- **P@10 and RR substantially more problematic because of higher σ**



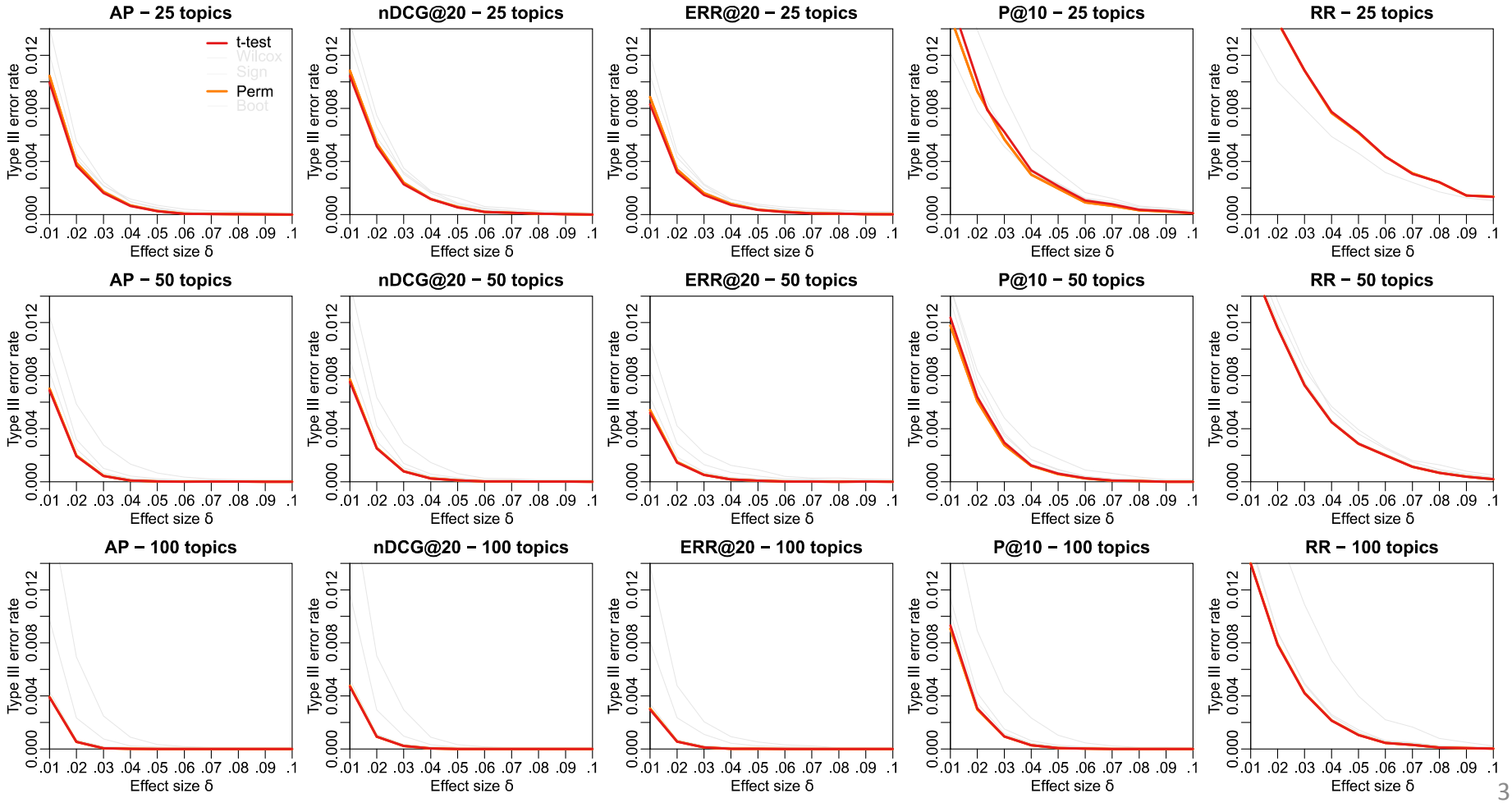
Type III Errors by δ | n α=.05

- Bootstrap tends to correct with sample size
- Wilcoxon stays the same, and Sign test gets even worse



Type III Errors by δ | n α=.05

- Bootstrap tends to correct with sample size
- Wilcoxon stays the same, and Sign test gets even worse



Type III Errors in Practice

- How much of a problem could this be?

- Example: AP and n=50 topics
  - Improvement of +0.01 over the baseline
  - 2-tailed t-test comes up significant
  - 7.3% probability that it is a Type III error and your system is actually worse

  - Is that too high?

CONCLUSIONS

What We Did

- First empirical study of actual error rates with IR-like data
- Comprehensive
  - Paired test: Student's t, Wilcoxon, Sign, Bootstrap-shift, Permutation
  - Measure: AP, nDCG@20, ERR@20, P@10, RR
  - Topic set size: 25, 50, 100
  - Effect size: 0.01, 0.02, …, 0.1
  - Significance level: 0.001, …, 0.1
  - Tails: 1 and 2
- More than 500 million p-values
- All data and many more plots are available online
  https://github.com/julian-urbano/sigir2019-statistical

Recommendations

- Don’t use Wilcoxon or Sign tests anymore
- For statistics other than the mean, use permutation, and bootstrap only if you have many topics
- For typical tests about mean scores, the **t-test** is simple, the most robust, behaves as expected w.r.t. Type I errors, and is nearly as powerful as the Bootstrap. **Keep using it**