IT'S EVALUATION'S WORLD, WE JUST LIVE IN IT

Asad Sayeed

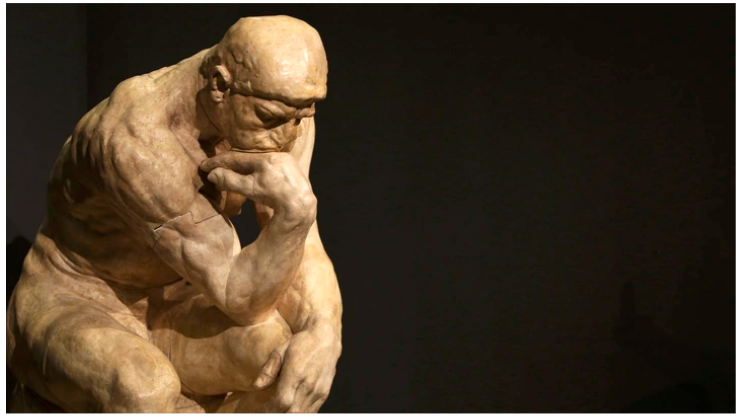
University of Gothenburg Ph. D. course – Introductory session

SCIENCE

Psychology's Replication Crisis Is Running Out of Excuses

Another big project has found that only half of studies can be repeated. And this time, the usual explanations fall flat.

ED YONG NOVEMBER 19, 2018



The Thinker, by Auguste Rodin (JASON LEE / REUTERS)

Over the past few years, an international team of almost 200 psychologists has been trying to repeat a set of previously published experiments from its field, to see if it can get the same results. Despite its best efforts, the project, called Many Labs 2, has only succeeded in 14 out of 28 cases. Six years ago, that might have been shocking. Now it comes as expected (if still somewhat disturbing) news.

MORE STORIES

Science Is Getting Less Bang for Its Buck

PATRICK COLLISON AND MICHAEL NIELSEN

Why Rich Kids Are So Good at the Marshmallow Test

JESSICA MCCRORY CALARCO

REPLICATION CRISIS

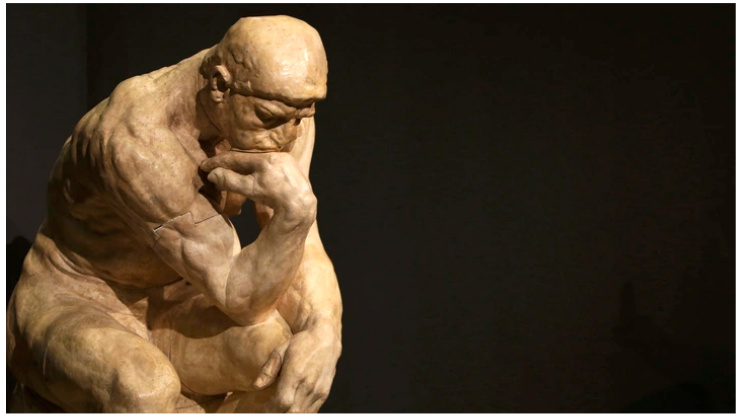
- Cohen et al (2018) LREC:
  - Replicability/repeatability: ability to run the experiment again
  - Reproducibility: ability to reach the same conclusions

SCIENCE

Psychology's Replication Crisis Is Running Out of Excuses

Another big project has found that only half of studies can be repeated. And this time, the usual explanations fall flat.

ED YONG NOVEMBER 19, 2018



The Thinker, by Auguste Rodin (JASON LEE / REUTERS)

Over the past few years, an international team of almost 200 psychologists has been trying to repeat a set of previously published experiments from its field, to see if it can get the same results. Despite its best efforts, the project, called Many Labs 2, has only succeeded in 14 out of 28 cases. Six years ago, that might have been shocking. Now it comes as expected (if still somewhat disturbing) news.

MORE STORIES

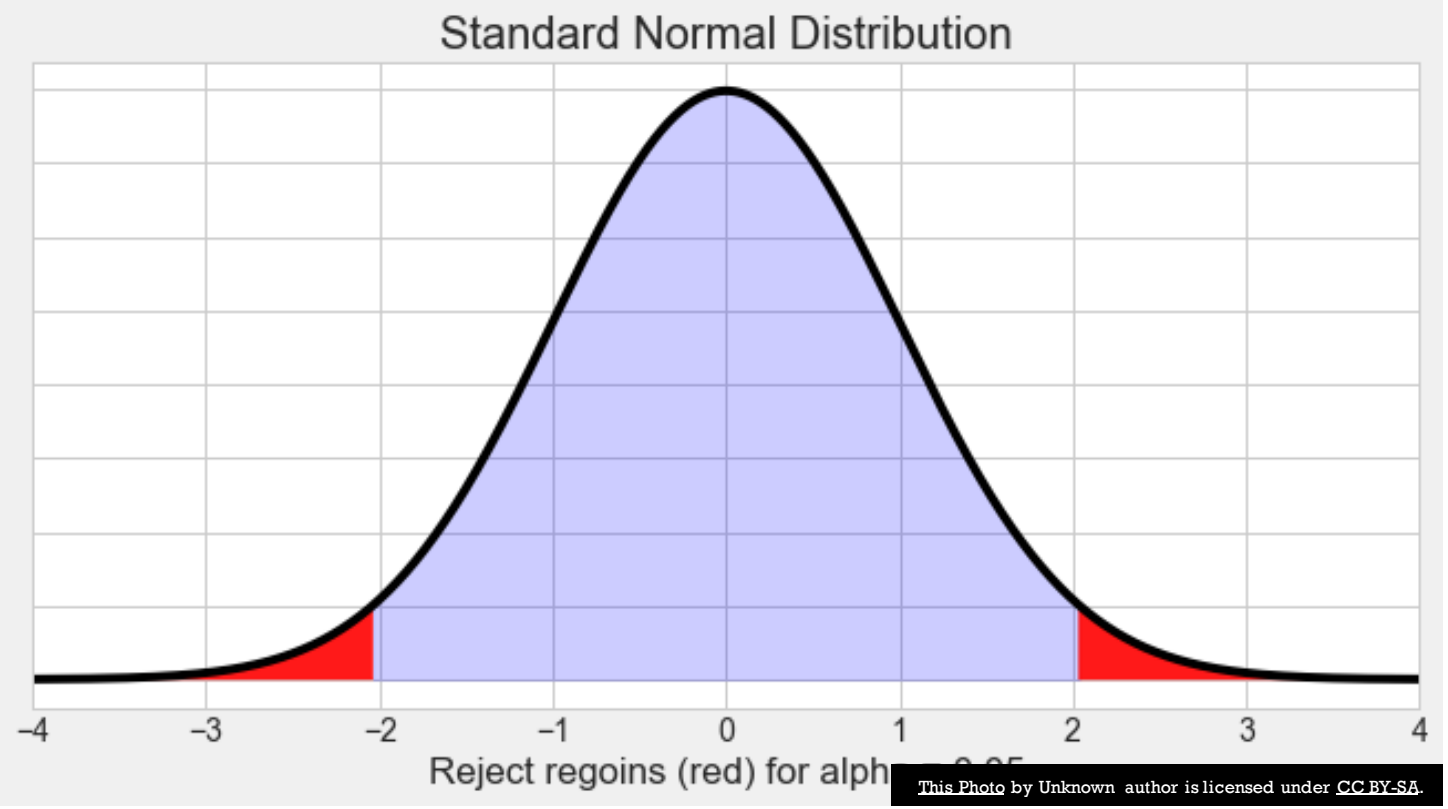
Science Is Getting Less Bang for Its Buck

PATRICK COLLISON AND MICHAEL NIELSEN

Why Rich Kids Are So Good at the Marshmallow Test

JESSICA MCCRORY CALARCO

~~REPLICATION~~ REPRODUCIBILITY CRISIS

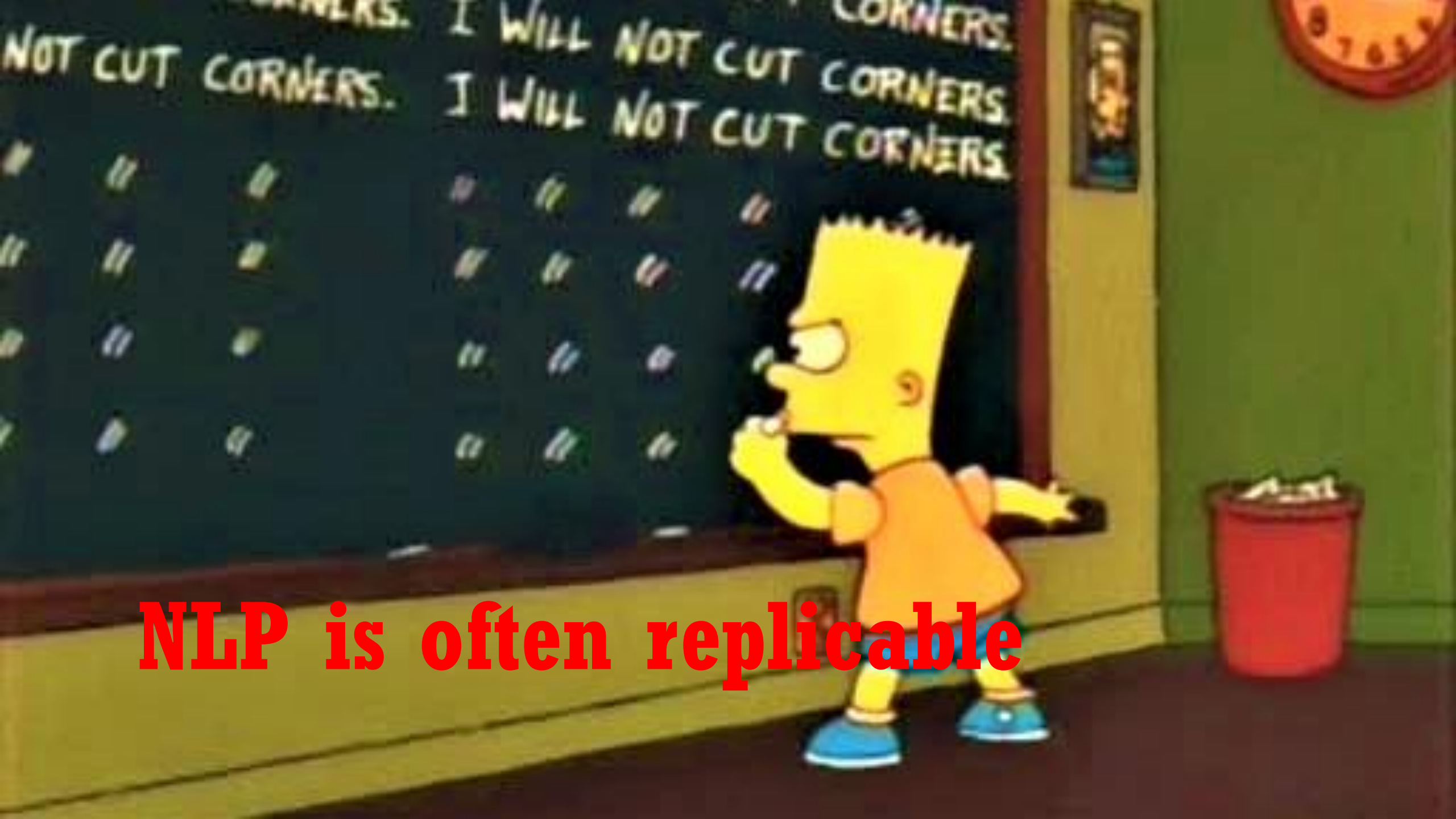


WHY A CRISIS?

- Null hypothesis testing
  - "P-hacking"
  - Insufficient power

BAYESIANITY DOESN'T SAVE US

SIGNIFICANCE: NEEDS TAILORED APPROACHES


I WILL NOT CUT CORNERS.
I WILL NOT CUT CORNERS.
NOT CUT CORNERS.
NLP is often replicable

But reproducibility?
No one knows…

THE NLP "HYPOTHESIS"

My system predicts human linguistic behaviour/performs a language task better than some previous system.

NLP EVALUATION: ROOTED IN INFORMATION RETRIEVAL

Precision, recall, etc.

CHECKING SIGNIFICANCE IS RARE

With good reason …

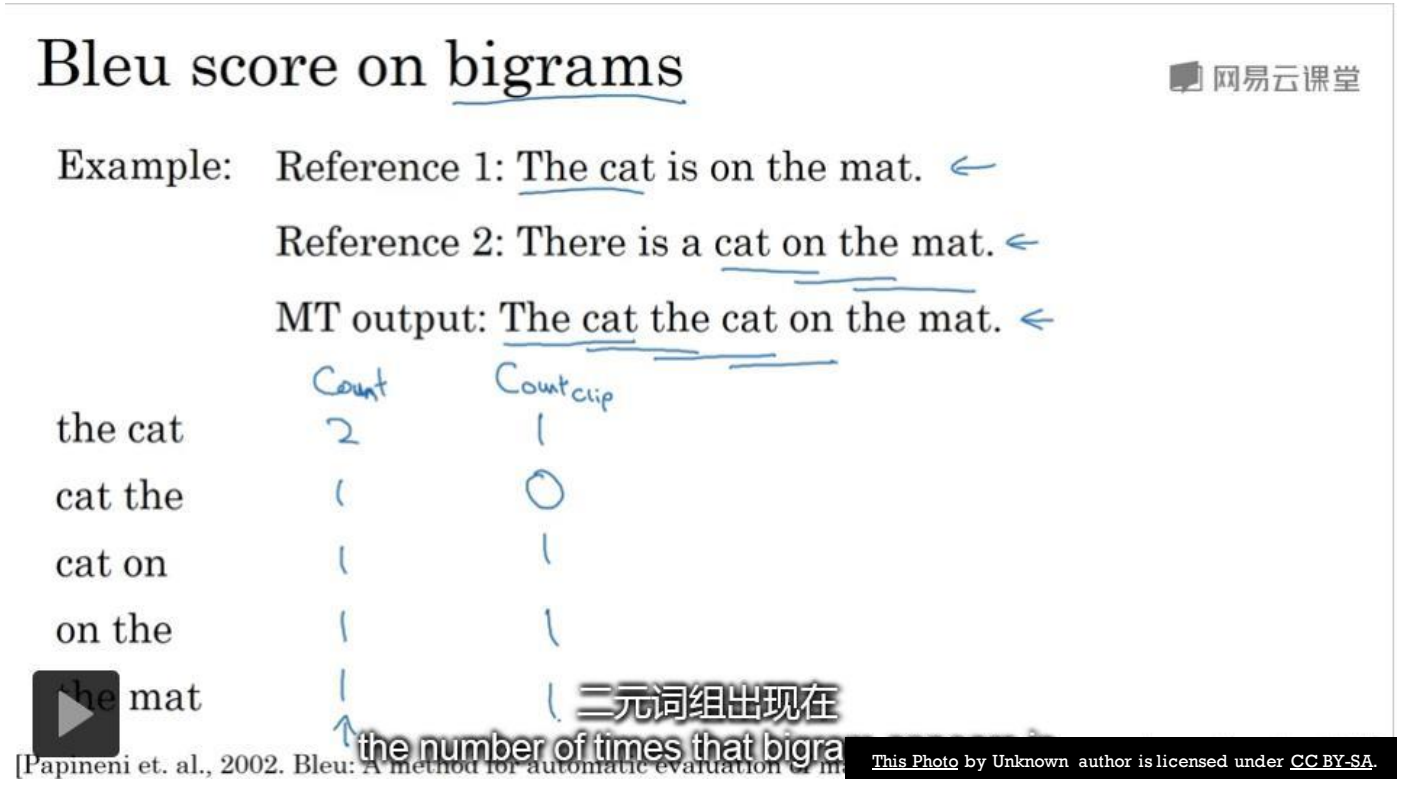
SIGNIFICANCE TESTING IN NLP

"Cookbook" null hypothesis testing can be misleading.

Assumptions of "standard" tests possibly not respected.

"Subject" variable either over or underpowered.

DO THE EVALUATION METRICS EVEN MAKE SENSE?



CONSIDER MACHINE TRANSLATION AND BLEU

EVALUATION AND SIGNIFICANCE

If the evaluations themselves are badly-suited to the task, how can we even talk about significance and reproducibility?

Need to rethink the "cookbook" -- for both evaluation and significance.

PURPOSE OF THE COURSE

- Explore/discuss/rethink various aspects of evaluation in language science and technology.
- Technical and philosophical perspectives.

PH.D. CREDIT

- 7.5 hpe
- Examination:
  - Leading discussion on a paper or topic.
  - Reasonably-sized research project.

COURSE MECHANICS

Weekly-to-biweekly reading group.

Meeting time negotiable.

Most meetings: led by student or other participant.

Use Microsoft Teams to coordinate?

NEXT WEEK'S READING

Dror et al. (2018). The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. LREC.

https://www.aclweb.org/anthology/P18-1128/

QUESTIONS FOR THE REST OF TODAY

- What are your expectations for the course?
- What misgivings do you have about evaluation in lang tech and science?
- Is there anything that works well, any advantages "we" have?