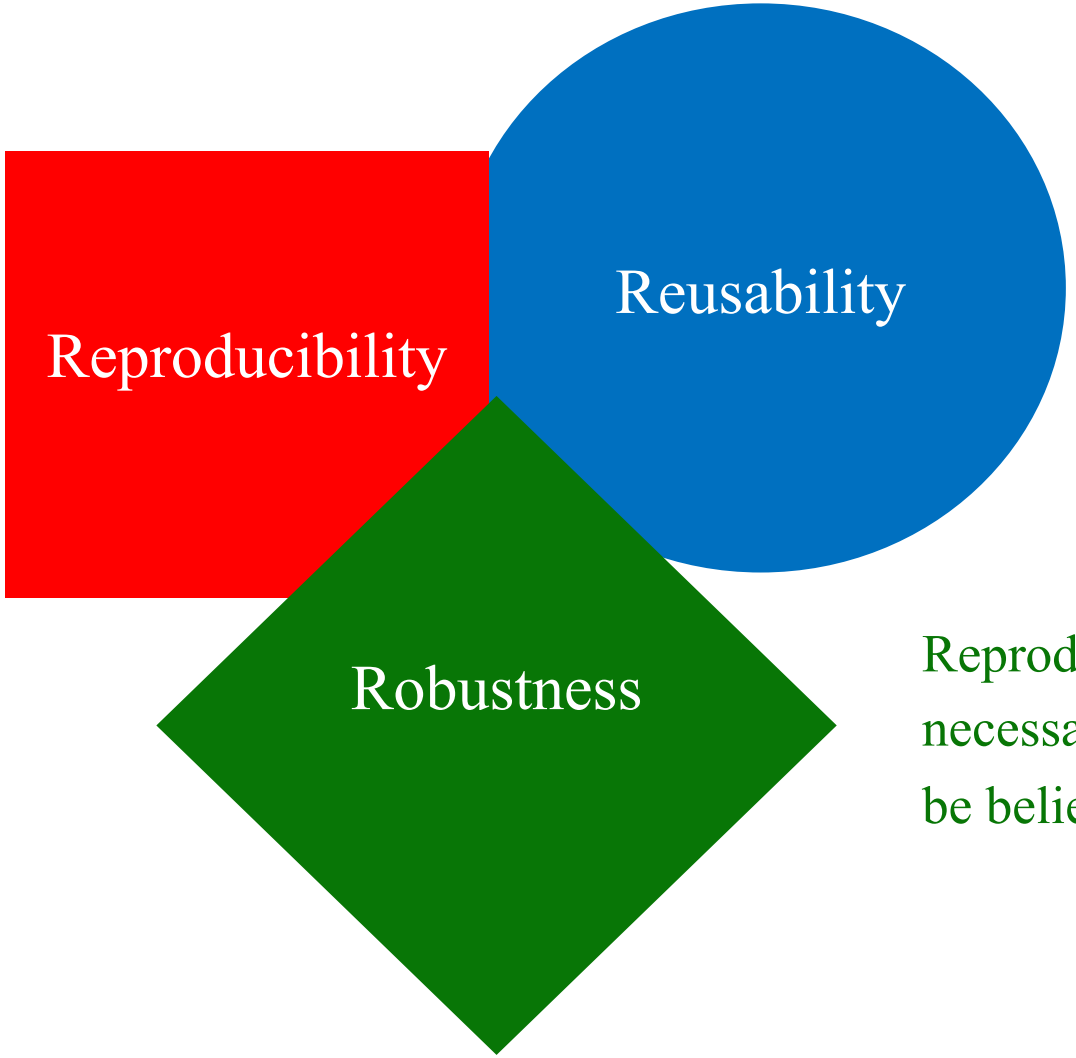
Can pre-registration lead to better reproducibility in ML research?

Joelle Pineau
Facebook AI Research (FAIR)
School of Computer Science, McGill University
Mila, CIFAR

NeurIPS 2020 workshop: The pre-registration experiment: an alternative publication model for machine learning research

December 2020

"**Reproducibility** refers to the ability of a researcher to duplicate the results of a prior study….

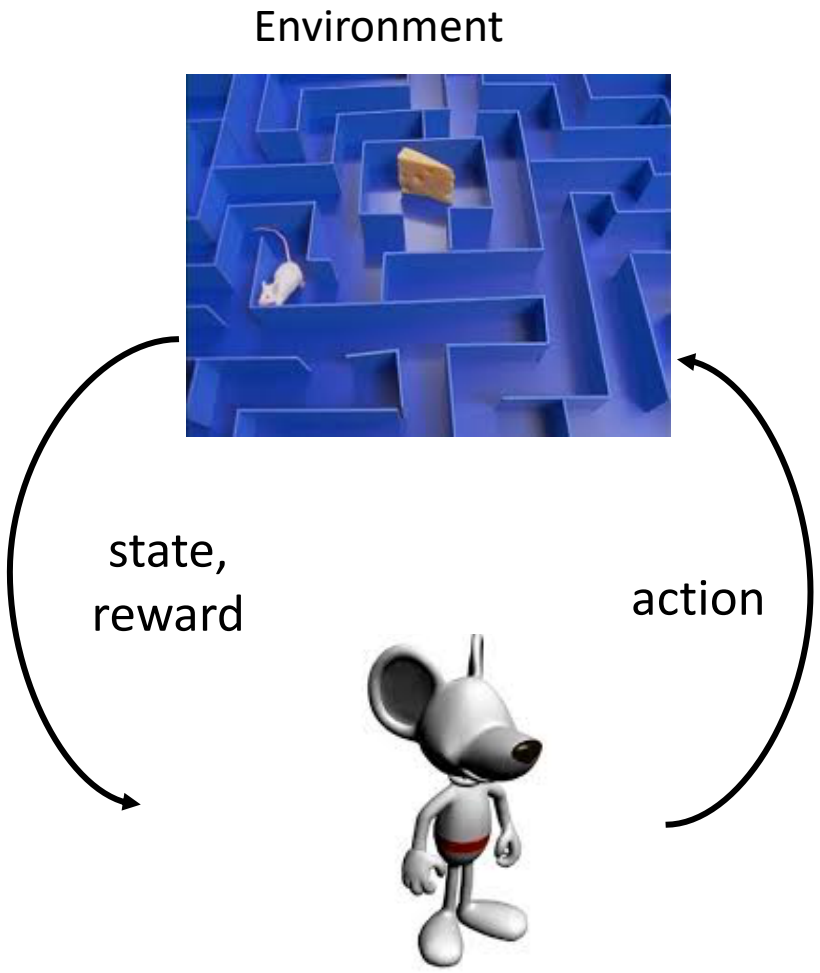


… using the same materials as were used by the original investigator.

Reproducibility is a minimum necessary condition for a finding to be believable and informative."

Bollen et al.
National Science Foundation, 2015.

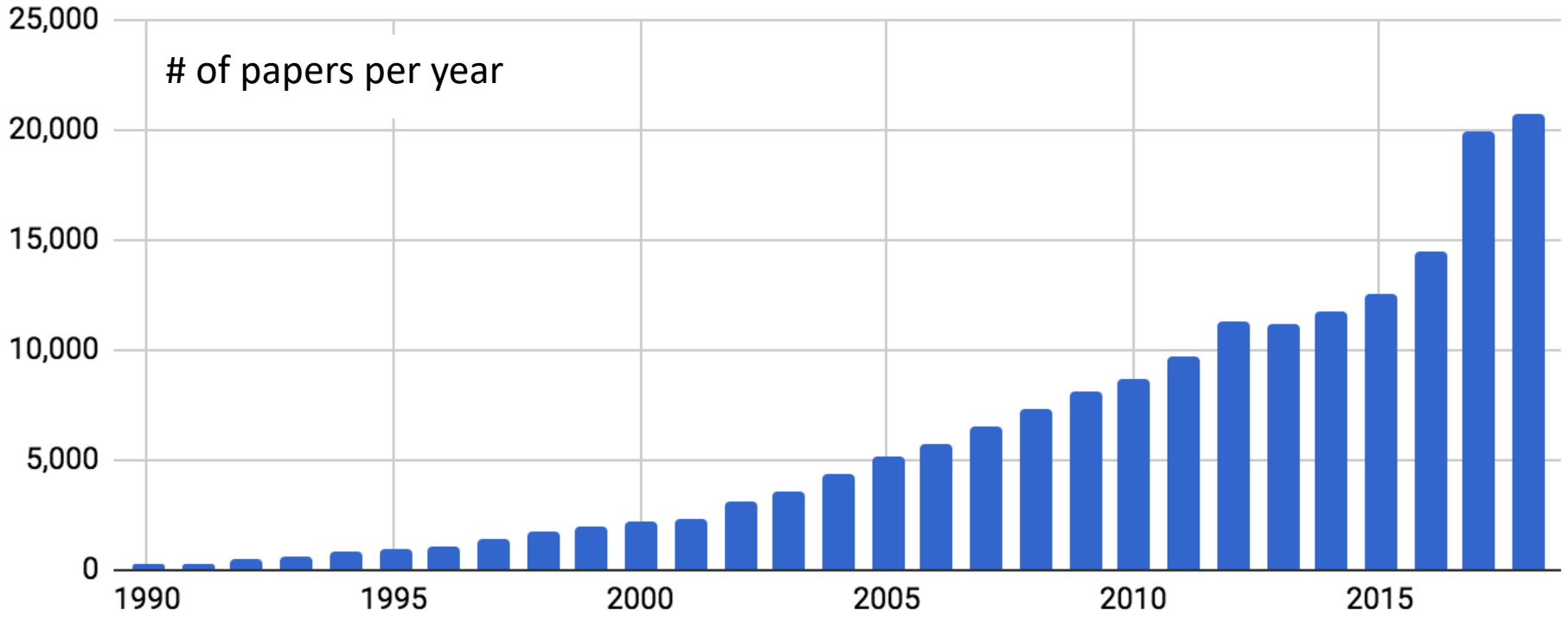
Reinforcement learning (RL)



- Very general framework for sequential decision-making!
- Learning by trial-and-error, from sparse feedback.
- Improves with experience, in real-time.

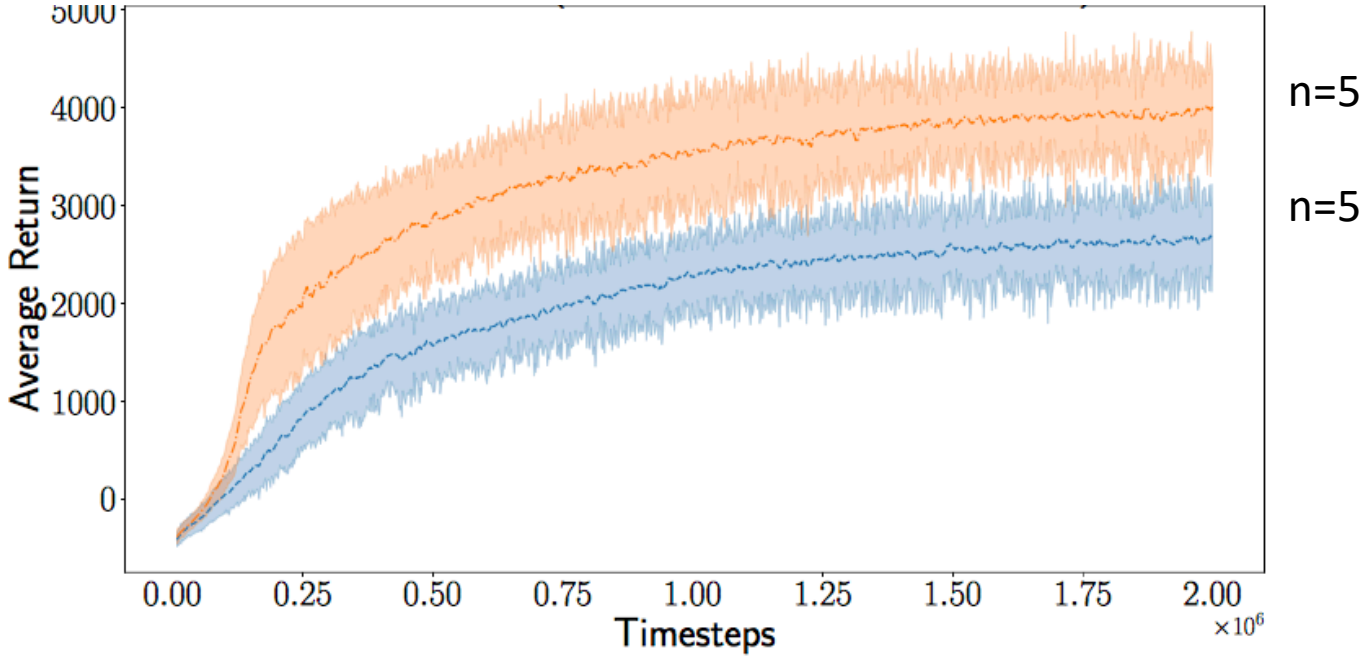
Learn $\boldsymbol{\pi}$ = *strategy to find this cheese!*

25+ years of RL papers



P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, D. Meger. *Deep Reinforcement Learning that Matters*. AAAI 2017 *(+updates)*.

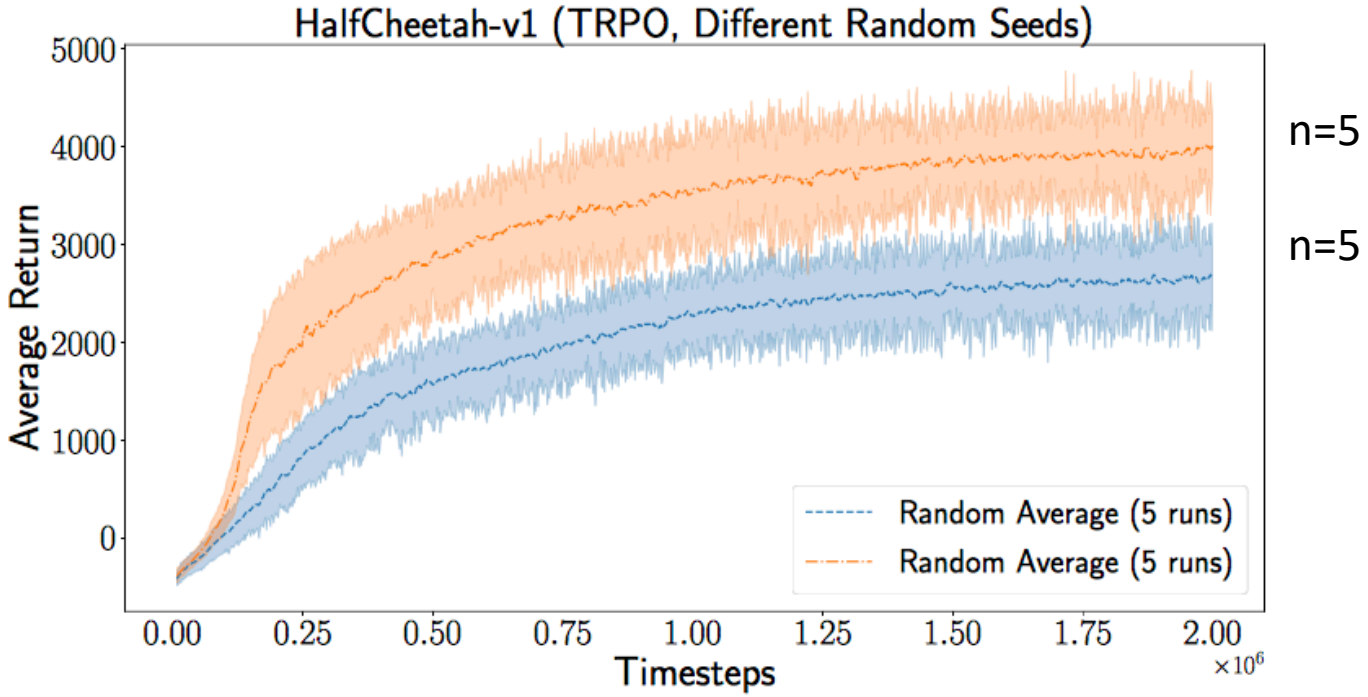
One particular experiment



P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, D. Meger. *Deep Reinforcement Learning that Matters*. AAAI 2017 *(+updates)*.

One particular experiment

Both are same RL code with best hyperparameter configuration!

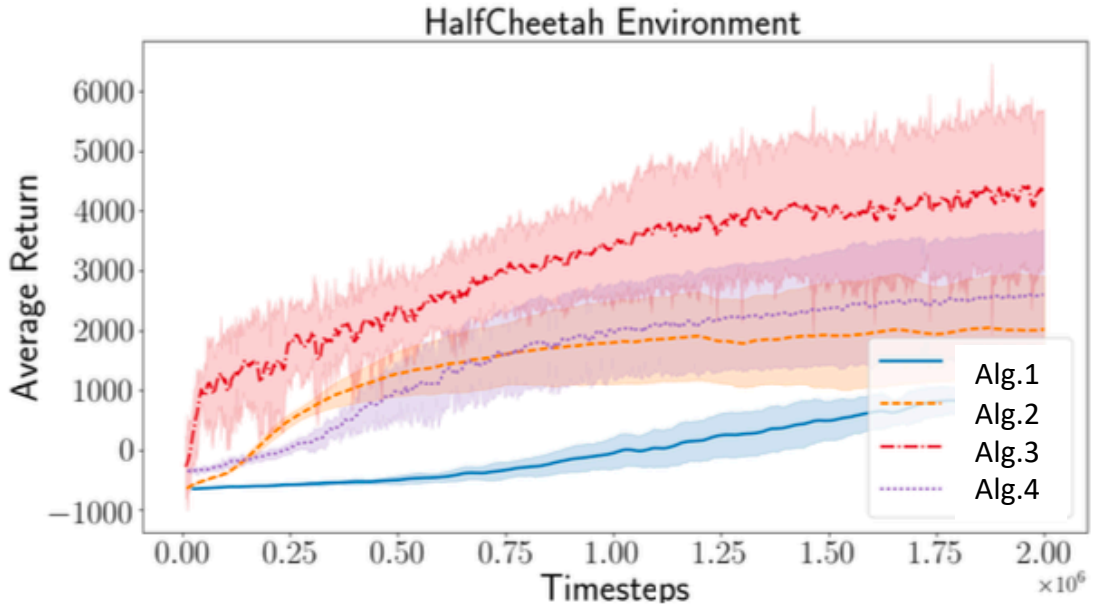


n=5

n=5

P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, D. Meger. *Deep Reinforcement Learning that Matters*. AAAI 2017 *(+updates)*.

How should we measure performance of the learned policy?



Average return over test trials? $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Confidence interval? $\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$

How do we pick *n*?

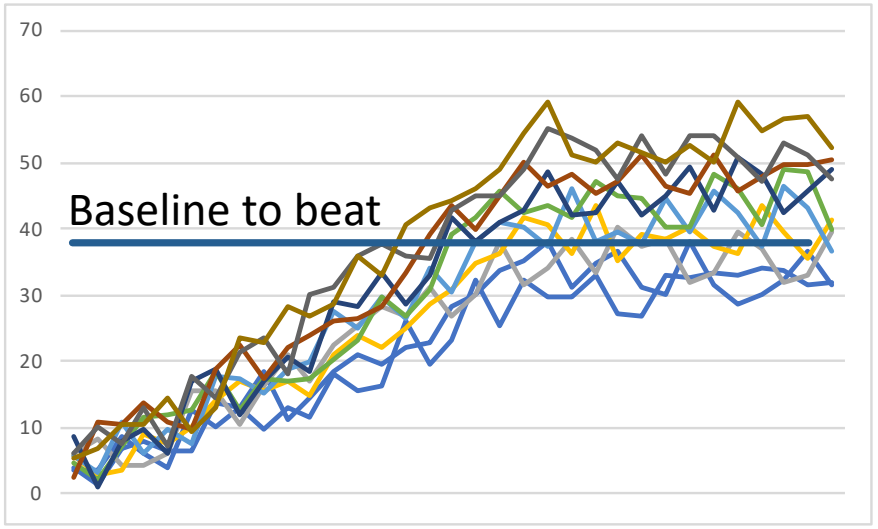
P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, D. Meger. *Deep Reinforcement Learning that Matters*. AAAI 2017 *(+updates)*.

How many trials?

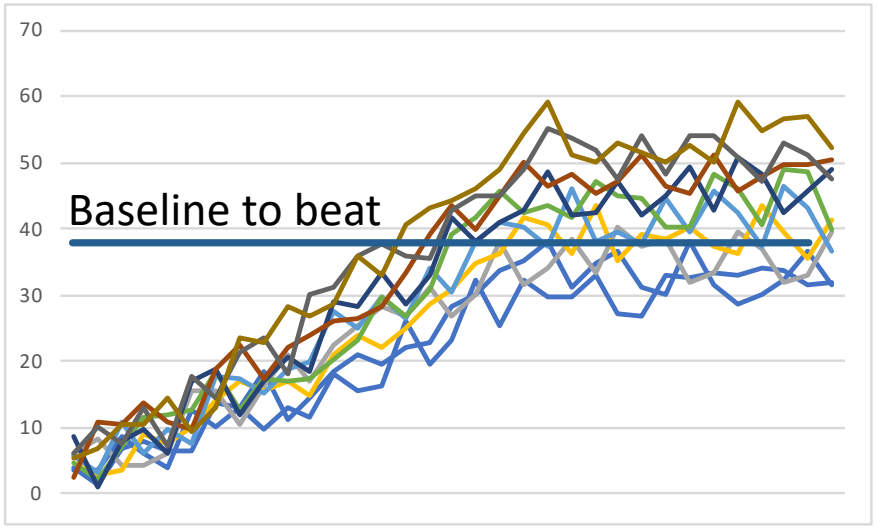
Work	Number of Trials
(et al. 2016)	top-5
(et al. 2017)	3-9
(et al. 2016)	5 (5)
(et al. 2017)	3
(et al. 2015b)	5
(et al. 2015a)	5
(et al. 2017)	top-2, top-3

P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, D. Meger. *Deep Reinforcement Learning that Matters*. AAAI 2017 *(+updates)*.

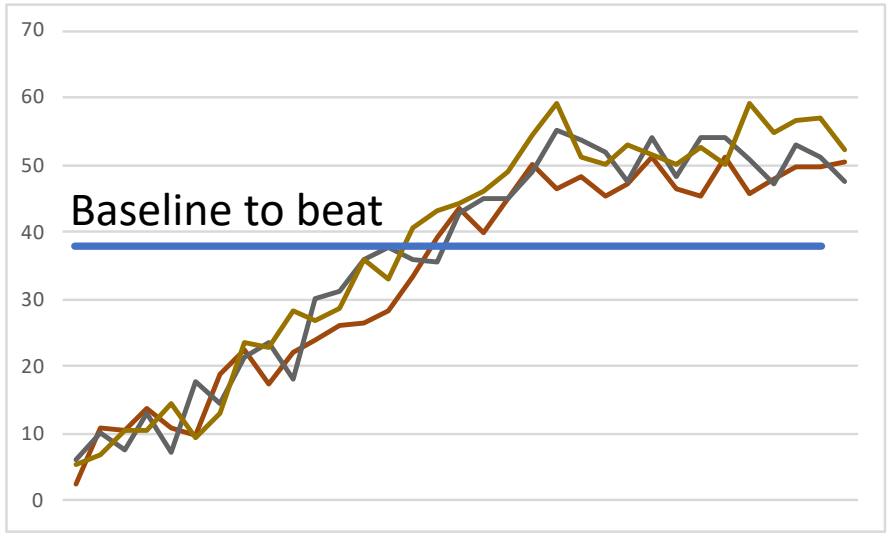
Consider the case of *n=10*



Consider the case of *n=10*



Top-3 results



- Strong positive bias: seems to beat the baseline!
- Variance appears much smaller.

We surveyed 50 RL papers from 2018
(published at NeurIPS, ICML, ICLR)

	Yes:
• Paper has experiments	100%
• Paper uses neural networks	90%
• All hyperparams for proposed algorithm are provided.	90%
• All hyperparams for baselines are provided.	60%
• Code is linked.	55%
• Method for choosing hyperparams is specified	20%
• Evaluations on some variation of a hold-out test set	10%
• Significance testing applied	5%

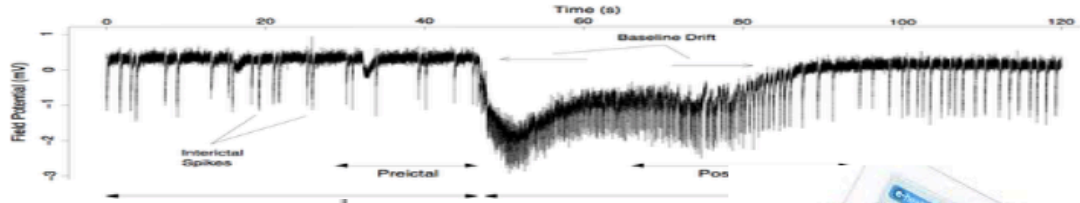
Should we do more pre-registration?

Pros:

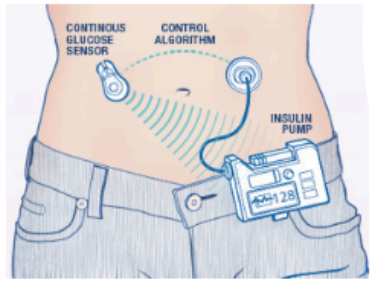
- Explicit and detailed record of experimental methodology.
- Peer feedback on methodology alone, earlier in the process.
- More reliable measurement of performance and generalization.

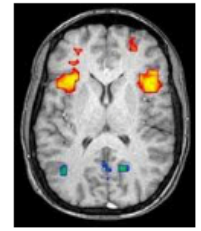
Cons:

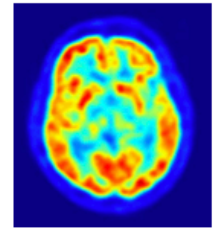
- Two-stage review (pre-/post-results). Slower to produce results.
- Difficulty ensuring results were not generated previously.
- Enabling more exploratory (“understand”) research work.

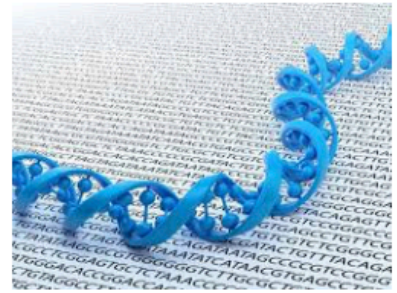

Time (s)
0
20
40
60
80
100
120
Baseline Drift
Field Potential (mV)
1
0
-1
-2
-3
Interictal
Spikes
Preictal
Pos

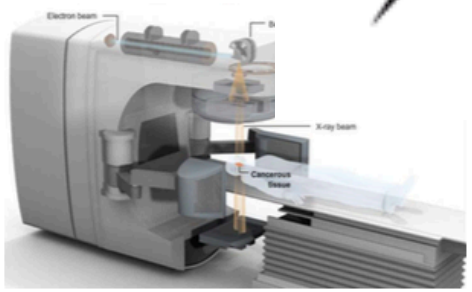

Medical Record
Search
A-F
G-L
M-R
T-Z
Home
Schedule
Health plan
Appointment

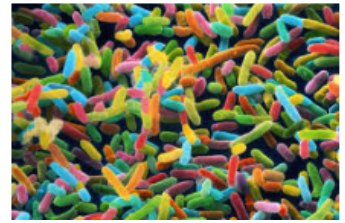

CONTINOUS
GLUCOSE
SENSOR
CONTROL
ALGORITHM
INSULIN
PUMP







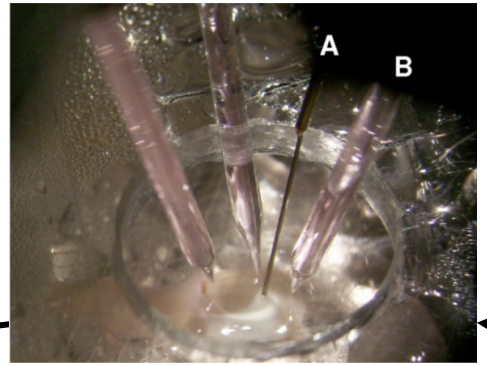

Electron beam
X-ray beam
Cancerous
tissue



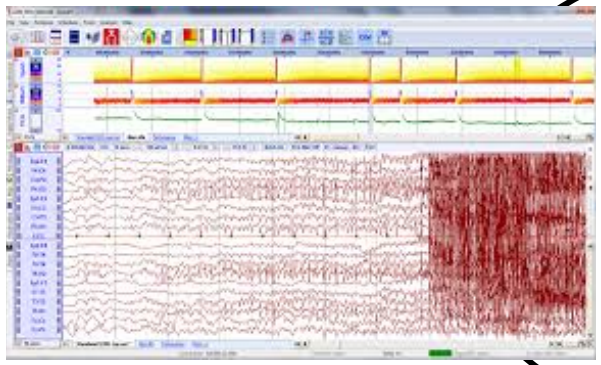
A case study of RL for adaptive neurostimulation (2008-2013)

Panuccio, Guez, Vincent, Avoli, Pineau, Exp Neurol, 2013

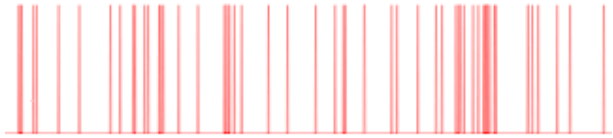
A case study of RL for adaptive neurostimulation (2008-2013)



state, reward



action

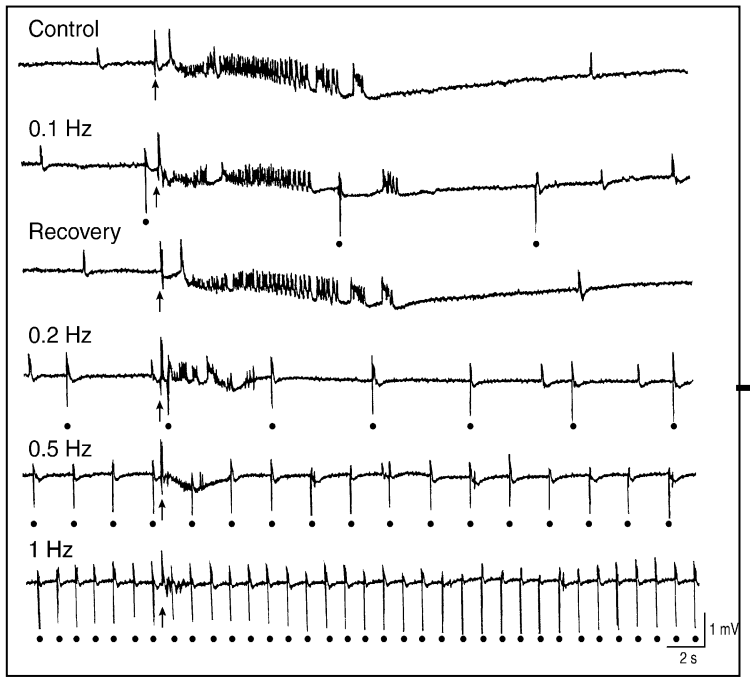


Goal = minimize seizures and minimize stimulation

Panuccio, Guez, Vincent, Avoli, Pineau, Exp Neurol, 2013

A case study of RL for adaptive neurostimulation (2008-2013)

Training data (N=8) from periodic strategies.

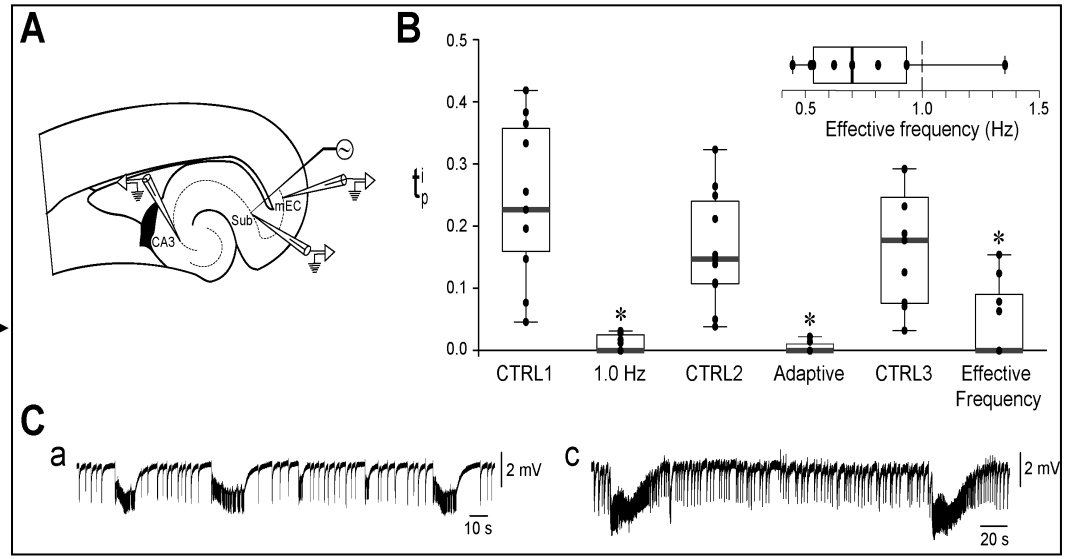


6 months

Learning algorithm
$Q_t(s,a)$

6 months

Testing data (N=11) from RL strategies.

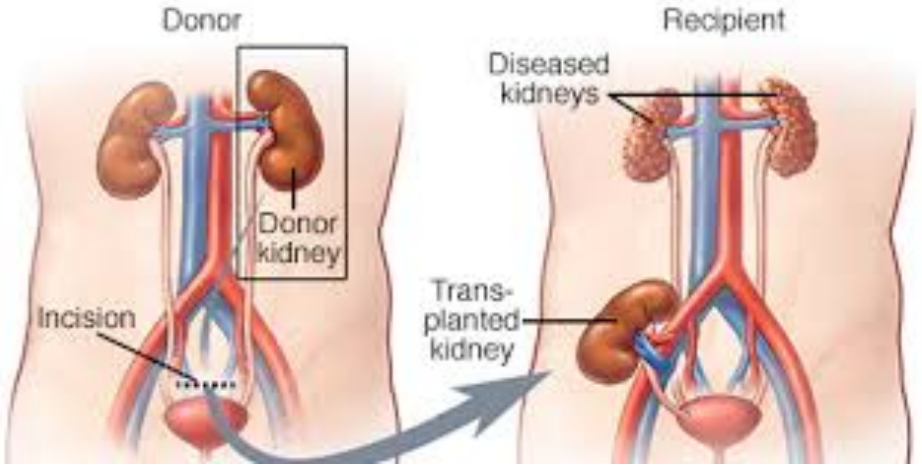


24 months

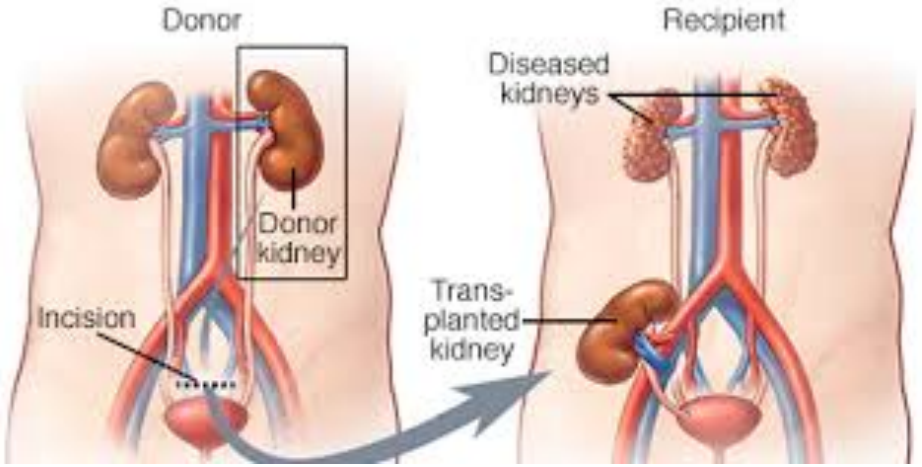
Panuccio, Guez, Vincent, Avoli, Pineau, Exp Neurol, 2013

Self-registration: A case study of ML for Organ Transplantation (2019)

- N=75 patients, split into 50 train / 25 test.
  Lock-up the test data!



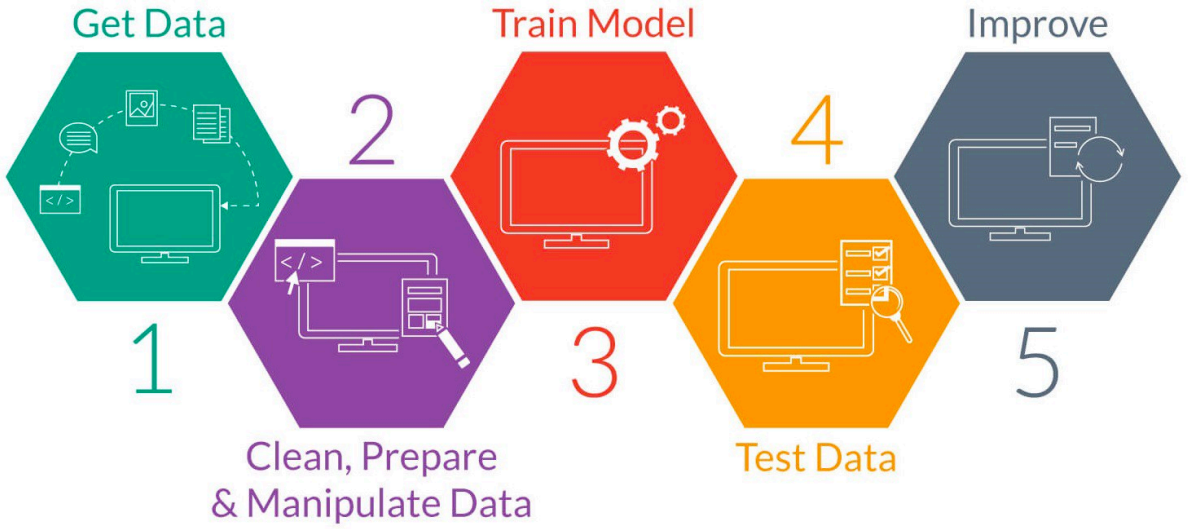
Self-registration: A case study of ML for Organ Transplantation (2019)



- N=75 patients, split into 50 train / 25 test.
  Lock-up the test data!

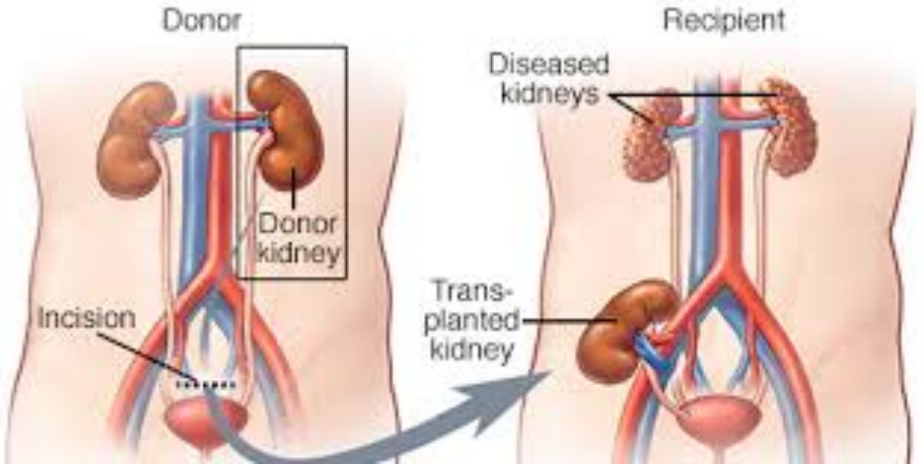
- Apply ML methods. Write the paper.

- When all is done, pull out test data and repeat the testing procedure.



https://houseofbots.com/news-detail/2449-Ranking-Popular-Deep-Learning-Libraries-for-Data-Science

Self-registration: A case study of ML for Organ Transplantation (2019)

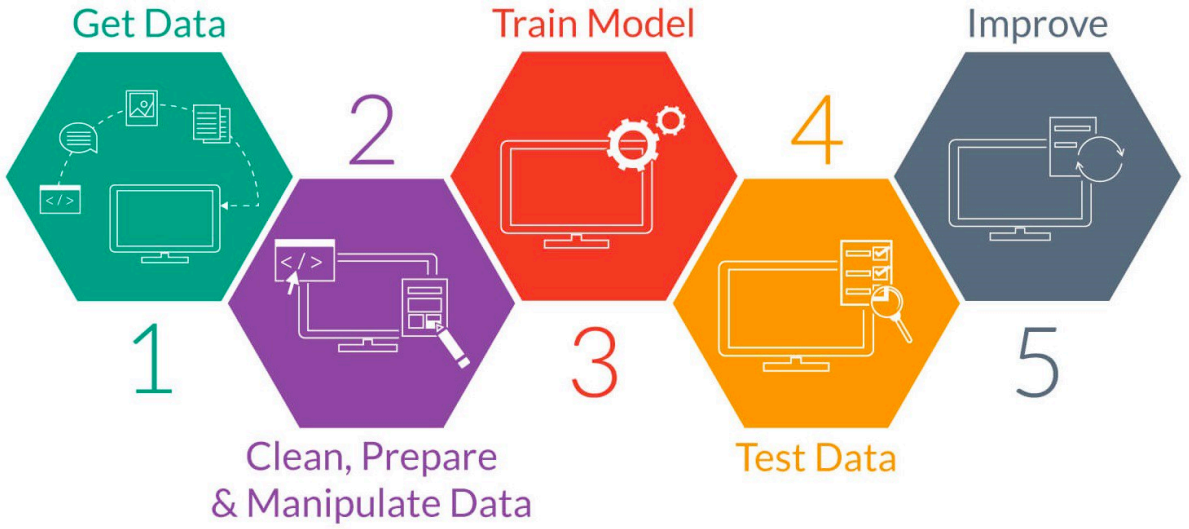


- N=75 patients, split into 50 train / 25 test.
  Lock-up the test data!

- Apply ML methods. Write the paper.

- When all is done, pull out test data and repeat the testing procedure.

FAIL!



https://houseofbots.com/news-detail/2449-Ranking-Popular-Deep-Learning-Libraries-for-Data-Science

What then?

- Get more data.
- Get much more data!
- Try a different model? (That requires more data too.)
- Move on to another project…

Discussion

- Can we narrow down the criteria for what type of study should be pre-registered?
- Does pre-registration make sense when experiments are so cheap?
- Can you re-pre-register? (aka *What is the "resubmission" policy?*)
- Should we have two separate research teams, one registering a methodology, and one executing the methodology to produce results?
- What is the threshold for a negative result to be informative / interesting? How do we weed out trivial / uninformative results?

Thank you!