DAY 3 - NIPS 2018 - MAIN DAY 2

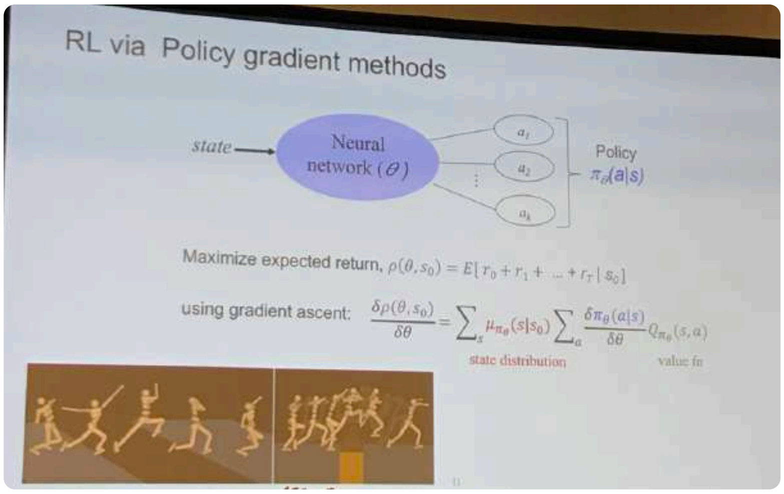
Talk 1 – Joelle Pineau (FAIR) – Reproducible, Reusable and Robust RL

- Reproducibility = ability to duplicate
  Reusability = using same materials
  Robustness = min. necessary addition

Nature (2016)
↳ 52% of researchers state that there is a crisis

- Deep RL that Matters (AAAI, 2017)

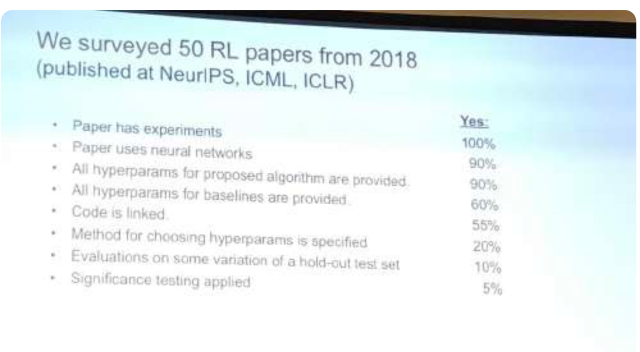
- Baseline comparison:
  - always logged
  - TRPO, PPO, DDPG, ACKTR
  ↳ different results for different domains / for different implementations / different network architectures / random seeds



- Underfitting the baseline! Overfitting your approach!

- Fair comparisons ⇒ robust conclusions
  → distinct hyperparams
  → sensitivity hyperparams
  → comp. budget

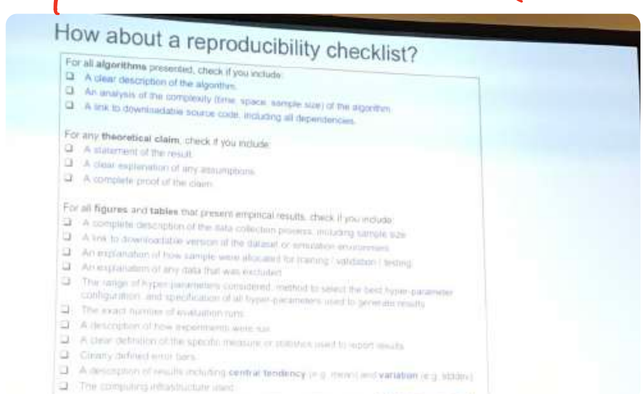
⟹ Careful thinking / reporting of experiments!


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%

→ add to submission process!

→ influence of hardware!
→ also specify!

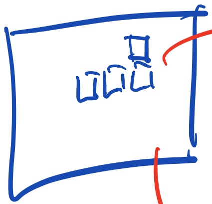

How about a reproducibility checklist?

For all **algorithms** presented, check if you include:
- ❑ A clear description of the algorithm.
- ❑ An analysis of the complexity (time, space, sample size) of the algorithm.
- ❑ A link to downloadable source code, including all dependencies.

For any **theoretical claim**, check if you include:
- ❑ A statement of the result.
- ❑ A clear explanation of any assumptions.
- ❑ A complete proof of the claim.

For all **figures** and **tables** that present empirical results, check if you include:
- ❑ A complete description of the data collection process, including sample size.
- ❑ A link to downloadable version of the dataset or simulation environment.
- ❑ An explanation of how sample were allocated for training / validation / testing.
- ❑ An explanation of any data that was excluded.
- ❑ The range of hyper-parameters considered, method to select the best hyper-parameter configuration, and specification of all hyper-parameters used to generate results.
- ❑ The exact number of evaluation runs.
- ❑ A description of how experiments were run.
- ❑ A clear definition of the specific measure or statistics used to report results.
- ❑ Clearly defined error bars.
- ❑ A description of results including **central tendency** (e.g. mean) and **variation** (e.g. stddev).
- ❑ The computing infrastructure used.

- RL: train ⊕ test on same env/data set → generalization ⇒ AGI
  ↳ different tasks
  ↳ Somewhere in between: use different seeds for train/test!
  ↳ simulators never can depict complexity of real world! → descript. length < 100 kB
- → simulator is convenient for reproducibility!
- → Zhang, Ballas, Pineau (2018) ⇒ active inference!

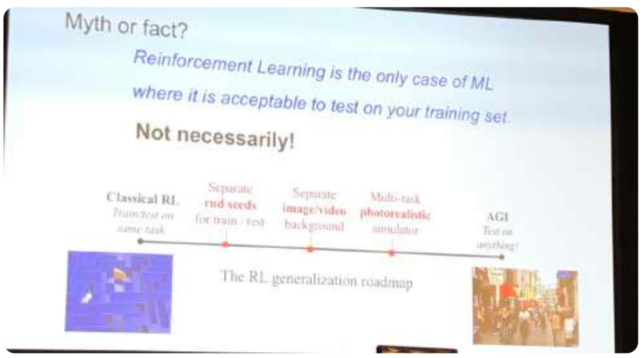


→ RL agent moves around image until sure, then reports classification → correct? : get reward

→ Natural

→ train on images / test on others!

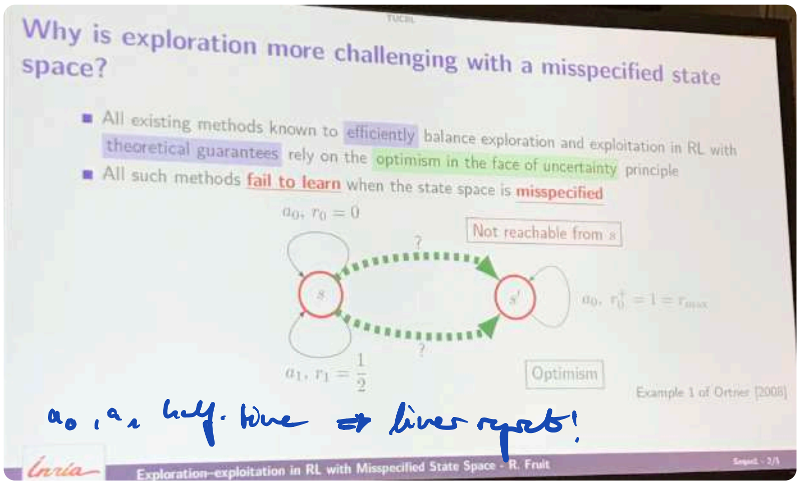
- → Video background for ALE envs, photorealistic simulators



- Pineau works on neural stim. with help of RL
  ↳ Go into real world!
  ↳ No exploration usually possible!
  ↳ very slow data acquisition

- ICLR reproducibility challenge

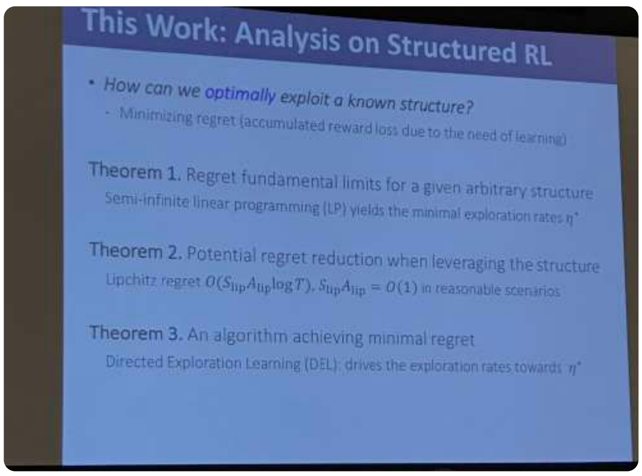
1st Oral / Spotlight Session – RL and Neuroscience



a_0, a_1 half. time ⇒ linear regret!

Oral 1: Exploration in Structured RL

→ structure via fct. approx ⇒ imposed on Q-fct. → optimally exploit!



→ structured MDP — combining reward / transitions

↳ $\phi \in \Phi$

MDP — structure ⇒ E.g. Lipschitz MDP

↳ How to exploit the structure?!

↳ covering number

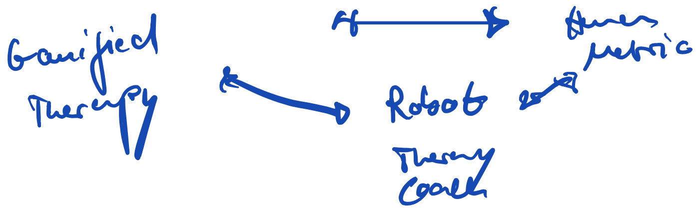
Talk 2 - A. Howard - Human-AI Trust Phenomenon

- Trust → belief truster that trustee mitigates truster's risk

→ Human-robot interaction : proximity ⇒ physiotherapy

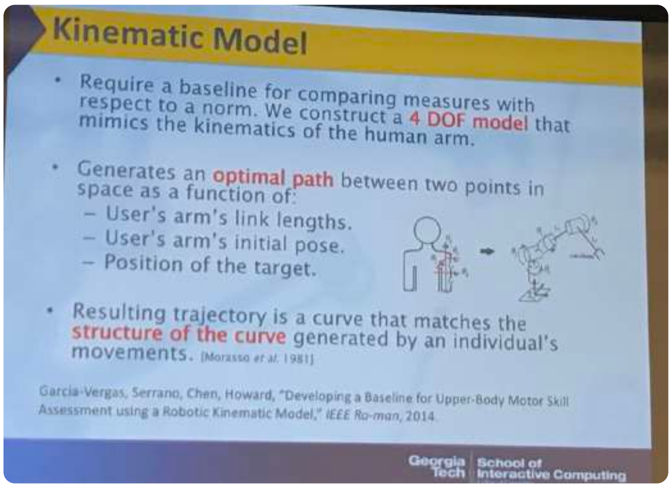
→ Cognitive attention of human/child ⇒ feather in!

↳ bonding / emotional engagement ⇒ not anger but sadness



- Goal : Build robot for effective physiotherapy sessions

↳ Humans trust robots often more than humans!



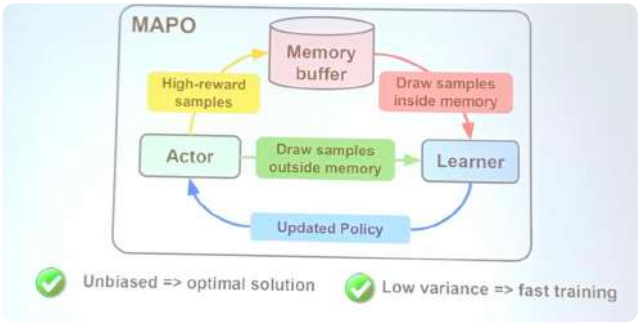
2nd oral / spotlight session — RL and Neuroscience

- Off-policy Evaluation ⇒ importance sampling
- Goal-dir. Molecular Graph generation ⇒ graph representation + RL
  - ↳ Graph Conv. Net
  - ↳ State: incomplete graph
  - ↳ Optimization via PPO

  validity check — optimization — + adv. ex. ↳ realistic

  TASKS: – Generation (full)
  – Completion (partially)

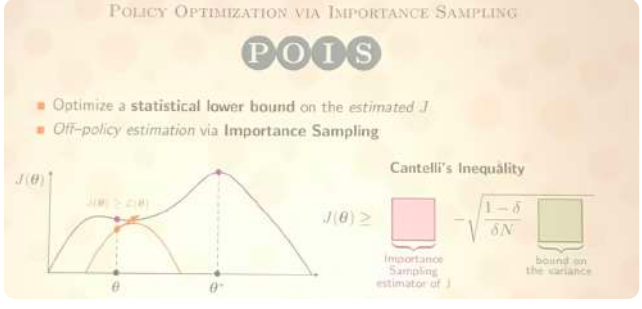
- Memory augmented PO ⇒ Semantic Parsing / program synthesis → treat as latent

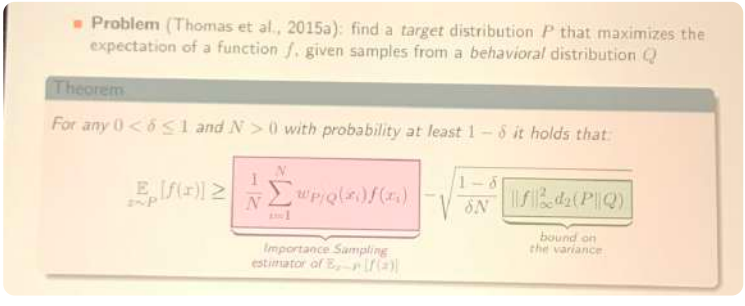


→ Variance reduction:
  exploration (inside memory)
  ⊕ sampling (outside memory)
→ slower convergence but better result

- Meta-RL for Structured Exploration Strategies
  → prior experience should guide exploration ⇒ structured stochasticity!
  → latent conditioned policy
    ↳ capture prior task distr.
  ⇒ Meta-learning in latent space
    ↳ more conditioning ⇒!

- Policy optimization via importance sampling (POIS)
  → policy search ⇒ θ^* ⇒ TRPO – reasonable neighborhood
  → PPO ⇔ surrogate loss
    ↳ statistical lower bound off-policy eval





→ action vs. parameter-based → downsides in experiments → variance reduction

→ offline hard optimization of neural policy!

→ hard performance in sparse rewards envs ⇒ shape trajectory based!

- Bayesian Generative Adversarial Imitation Learning

Generative Adversarial Imitation Learning (GAIL)

- Use **generative adversarial networks (GANs)** for imitation learning:

$$\min_{\pi} \max_{D} \mathbb{E}_{\pi}\left[\sum_{t=1}^{T} \log D(s_t, a_t)\right] + \mathbb{E}_{\pi_E}\left[\sum_{t=1}^{T} \log(1 - D(s_t, a_t))\right]$$

1. Sample trajectories by using $\pi(a|s)$ and $\pi_E(a|s)$ (expert demonstrations).
2. Train discriminator.
3. Update policy $\pi(a|s)$ by using reinforcement learning (RL), e.g., TRPO, PPO.

→ requires model-free inner loops ⇒ simple inefficient

→ train discriminator with Bayesian classification!

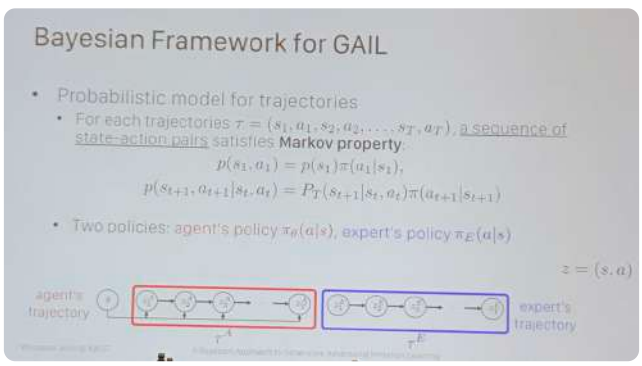
Bayesian Framework for GAIL

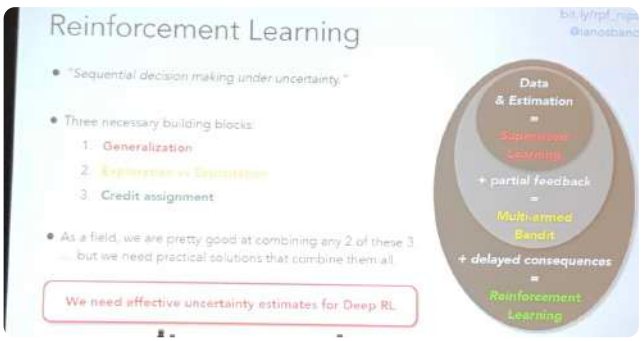
- Probabilistic model for trajectories
  - For each trajectories $\tau = (s_1, a_1, s_2, a_2, \ldots, s_T, a_T)$, a sequence of state-action pairs satisfies **Markov property**:

$$p(s_1, a_1) = p(s_1)\pi(a_1|s_1),$$
$$p(s_{t+1}, a_{t+1}|s_t, a_t) = P_T(s_{t+1}|s_t, a_t)\pi(a_{t+1}|s_{t+1})$$

- Two policies: agent's policy $\pi_\theta(a|s)$, expert's policy $\pi_E(a|s)$

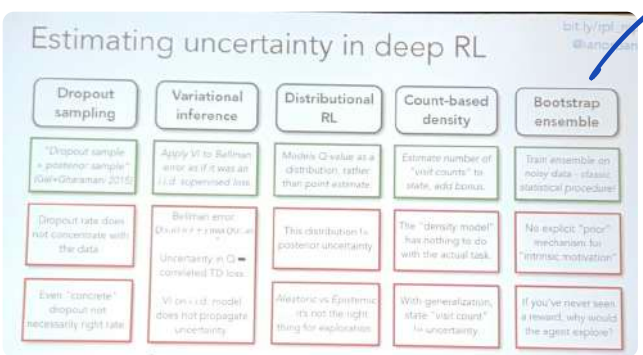
$z = (s, a)$





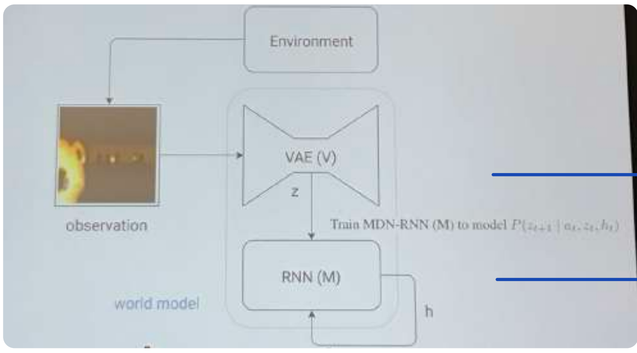
- Visual RL with Disagreement Goals
  - autonomous goal setting → only visual input
  - latent representation ⇒ VAE → sample new latents used as goals
    - reward fct.: distance in latent space!

- Randomized prior fct. for DRL → address shortcomings



- Playing hard exploration games by watching YouTube
  → Sound in ATARI games! → important signal!
- Recurrent World Models Facilitate Policy Evolution
  → VAE : compression → Density Estimation via RNN

→ No end-to-end but separate training!



z → Feed into controller module → output action
h