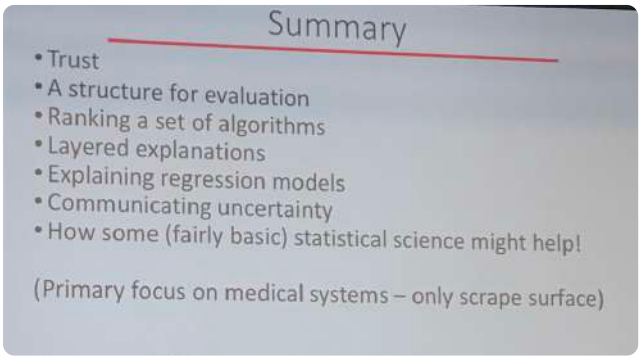
NIPS 2018 - DAY 4 - MAIN CONFERENCE 3

Breiman Lecture — D. Spiegelhalter — Trust: Transparency, Explanation, Validation

- Groups focus: communication of risk and uncertainty



→ Trust: Onora-O'Neill
to demonstrate trustworthiness
to do not aim to be trusted

→ Expect trustworthy claims
Cf toy and about the system

- Structure for eval:
  1. safety testing
  2. proof-of-concept
  3. Randomised controlled trials
  4. post-marketing surveillance

⟷ digital testing (mice/test)
⟷ lab testing (human/real)
⟷ field testing (put to practice)
⟷ routine use

1. Evaluation of different algorithms → problem of testing on same data
   → overfitting of test sets
   → solution: bootstrap sample from test set and get distribution of metric ⇒ compare distributions across algorithms!
   → but post CV estimate
2. Turing test → need to actually test outside of simulations
3. * simple randomised: A/B testing
   * cluster randomised: by team/user
   * stepped wedge: randomised rollout, when expect temporal effects (e.g. order of when/who gets testing)

- No transparency but "intelligent openness": accessible, intelligible, useable, assessable
  does not imply interpretability
  vs. explainability → global: trustworthiness of algorithm itself
  → local: current decision

- Communication – Visualisation ⇒ Tables, Curves, Charts, Texts, Icons
  - ↳ NHS Breast Cancer info
  - ↳ ease of use via configurables ⇒ click and drag
- Interpretability does need to trade-off with accuracy!
- Communicating uncertainty ⟷ trustworthiness ↓ ?
  - ↳ confident uncertainty does not reduce trust!

1st Oral/Spotlight session – RL and Neuroscience

Random Fourier Features (RFF)

[Rahimi and Recht, 2008] RFF mapping $\phi(\cdot)$:

$$k(\mathbf{x}, \mathbf{z}) = \mathbb{E}[\phi_{\mathbf{w}}(\mathbf{x})\phi_{\mathbf{w}}(\mathbf{z})]$$

$$\phi_{\mathbf{w}}(\mathbf{x}) = \left[\cos(\mathbf{w}^\top \mathbf{x}), \sin(\mathbf{w}^\top \mathbf{x})\right], \quad \mathbf{w} \sim p(\mathbf{w})$$

RFF ⟷ Monte Carlo approximation for $I(f)$

- Orthogonal points $\mathbf{w}$ ⟶ more accurate
- Structured $\mathbf{w}$ ⟶ faster
- Orthogonal + structured $\mathbf{w}$ ⟶ more accurate and faster

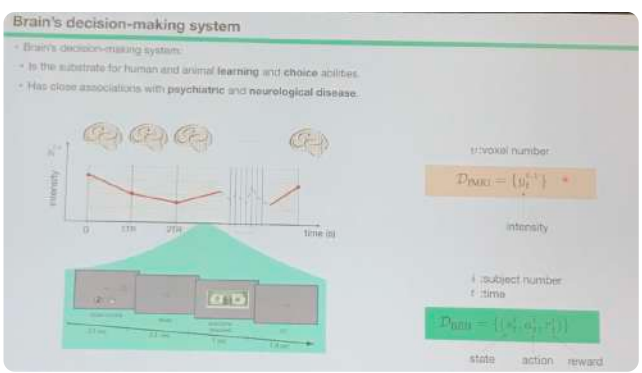
READ!

→ Bayesian papers!

- Interpreted accounts of behavioural and neuroimaging data using RNNs

#

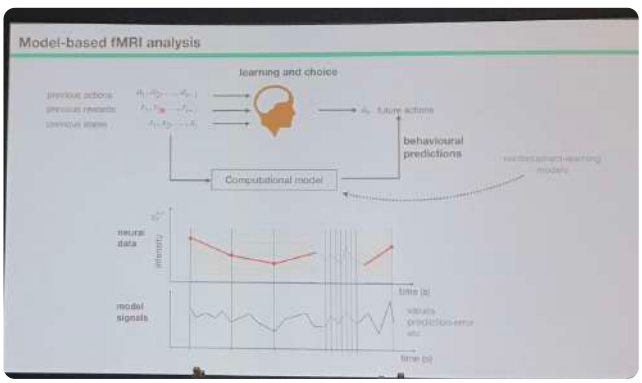
model-based fMRI



→ mechanisms

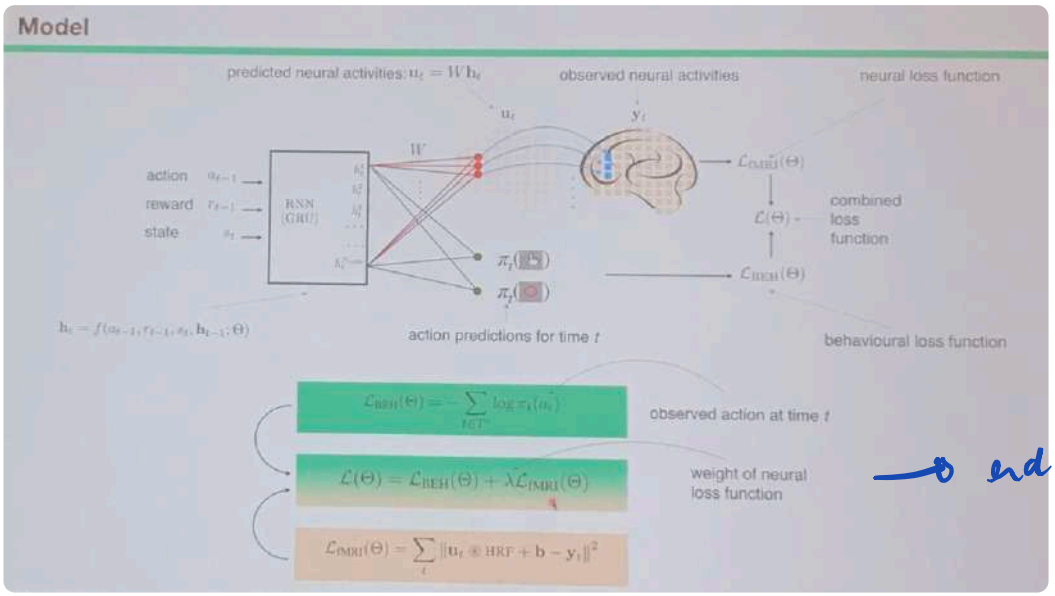
↓

→ outputs



→ computes behaviour ⇒ does not account for neural dynamics!

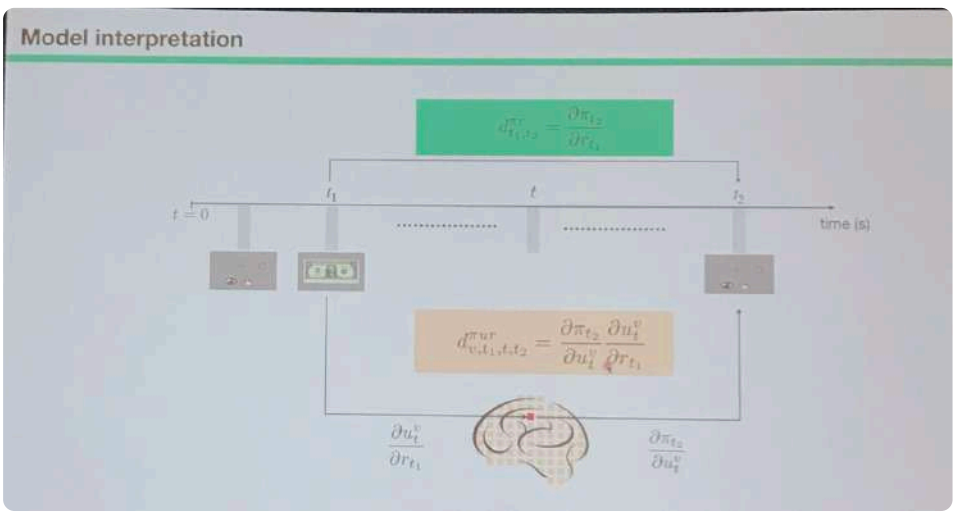
Meta-learning RNN



→ feed RNN layer with voxel activity!

→ end to end optimization of 55 GRU cells

→ Decoding brain activity ⊕ behavior ⊕ reward!



⇒ reward temporal activity of reward processing!

Posters 102!

2nd talk – Uwe Oluktun – Designing Computer Systems for Softw. 2.0?

- Moore's law slowing down → computation is limited by power
  - ↳ Dennard scaling: 2x transistors ⇔ 2x power
- Software 2.0 ⇒ weights of NN ⇔ optimization
  - ↳ predictable in terms of runtime / memory usage
- Data dependent / Programming → Snorkel ⇒ A. Ratner, C. Ré
  - ↳ semi-supervised weak labelling

Power =

Performance

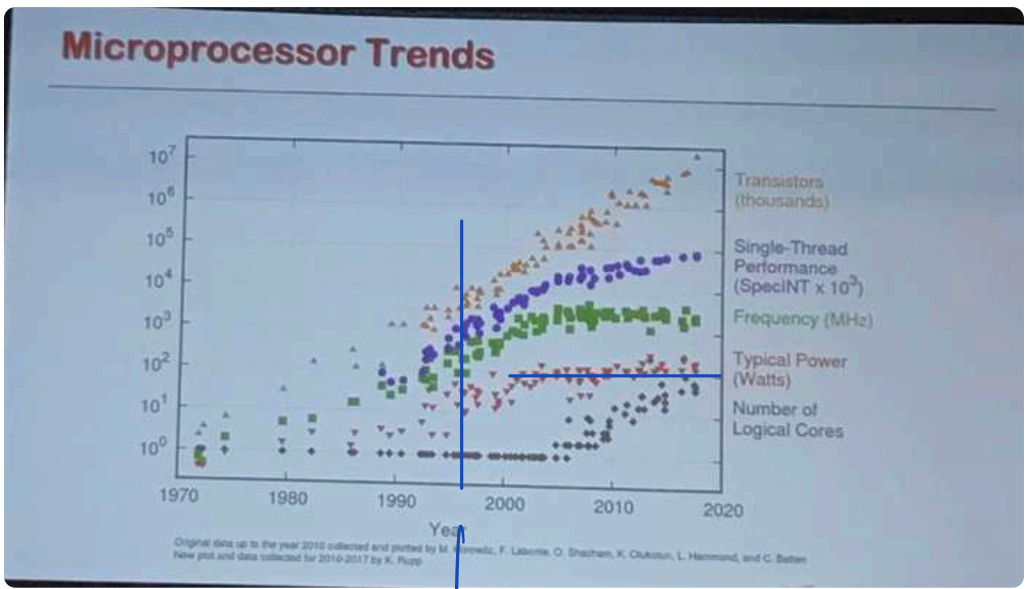
&

Energy

Efficiency

↓

more efficient via specialization



① Algos
② Langs / Compilers
③ Hardware

Multicore ⇒ first time microprocessors were designed in despair!

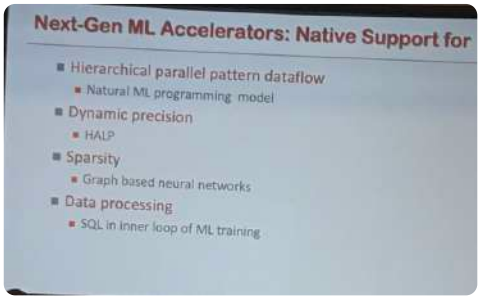
① • Statistical correctness of computation ⇔ accuracy in terms of performance
- Example: SGD ⇒ batch size – hardware / acc. tradeoff
- Example: precision ⇒ floating to fixed point → less energy ⊕ memory!
  - ↳ lower accuracy
  - ↳ also throughput advantages
  - ↳ HALP SGD → bit centering ⇒ implicit regularization

SIMD

- Relaxation cache, synchronizable, etc.

② • Domain specific languages (Matlab, SQL) → DSL
• Explicit graph construction vs. explicit such as in TF
• Integration of multiple DSLs → Parallel patterns
• Low vs. high level couplers
→ Optimizing locality ⇒ tiling / fusing
→ Exploit parallelism ⇒ hierarchical pipelining

③ TPU — ML unit + SW cache
— ISA → Instruction Set Architecture bottleneck
CUDA → PTX → GPU



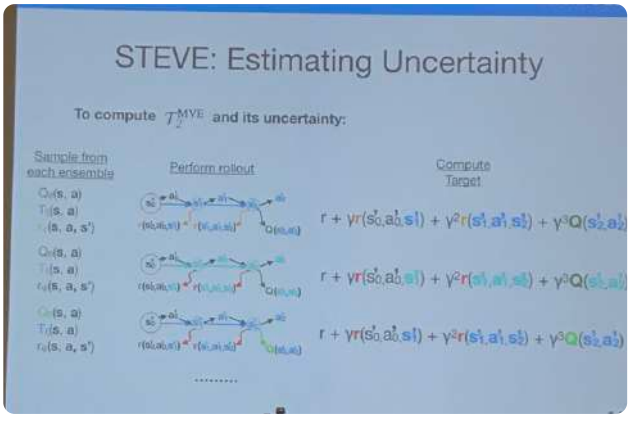
⇒ HIERARCHICAL PARALLEL PATTERNS

pattern → Compute Units
pattern → Memory Units

Best Oral/Spotlight Stats — RL and Neuroscience

• Sample-Efficient RL with Stochastic Ensemble Value Expansion
→ Ensemble NNs ⇒ uncertainty est. → use model-based of action!
→ Model value expansion ⇒ Feinberg et al 2018 → rollout the model
to extra discounting → faster convergence! (MVE)
to exploration → adversarial to model-based control
→ STEVE ⇒ different kinds of rollout

⇒ Target: mean
to Variance: uncertainty est.
↳ inverse variance weights
→ more flexible in TD(λ)
→ computationally very slow!



- Non-delusional Q-Learning and Value Iteration

→ Non-robust ⇒ delusional bias: ① restrictions due to approx.
② indep. Bellman backups

↳ incompatibility of non-jointly realizable updates!

→ information set ⇒ test consistency/feasibility! → prune

→ develop policy-constrained versions!

↳ VC dim. improves!

→ Efficient feasibility check!

Theoretical Guarantees

PCVI / PCQL Theorem

- ***Convergence & Correctness:*** Information sets converge and Q-values converge and are correct
- ***Optimality & Non-delusion:*** Optimal greedy policy and non-delusional values can be extracted
- ***Runtime:*** PCVI converges in polytime (w.r.t. VC-dimension of greedy policy class)

PCVI polytime for linear function approximation