Future of AI & Africa

Karim BEGUIR, Co-Founder & CEO, InstaDeep

kb@instadeep.com
@kbeguir

AI Today

AI Today

"AI = Reinforcement Learning + Deep Learning"

David Silver (AlphaZero's inventor) Google DeepMind

AI Today

MIT
Technology
Review

Reinforcement Learning

By experimenting, computers are figuring out how to do things that no programmer could teach them.

Deep Learning

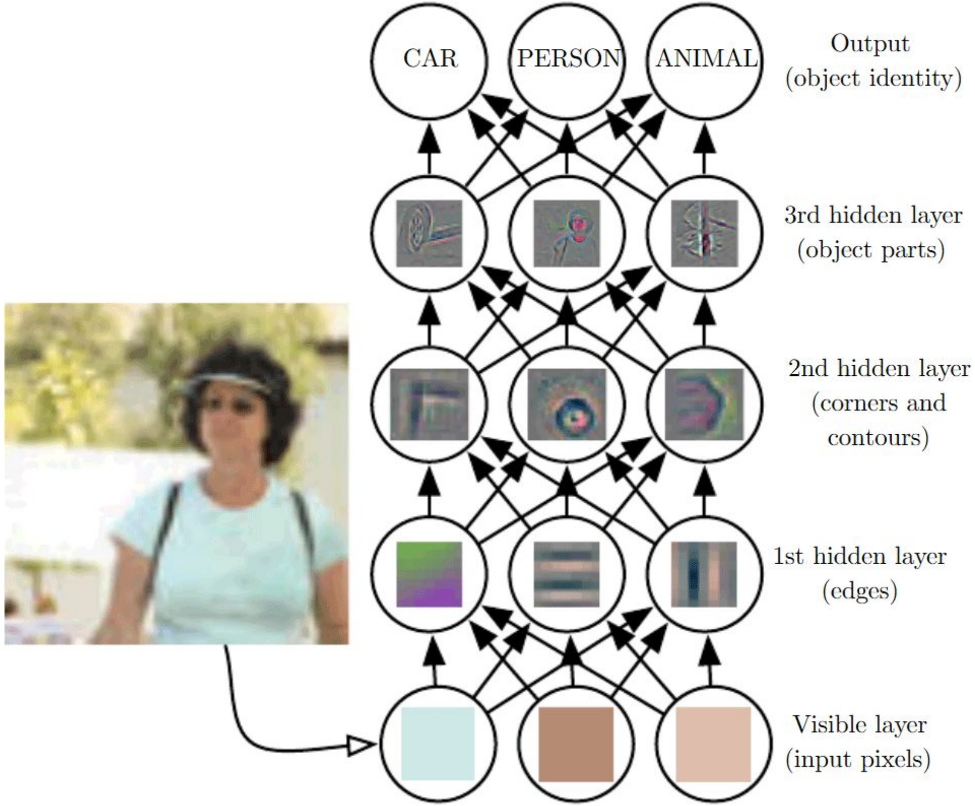
With massive amounts of computational power, machines can now recognize objects and translate speech in real time. Artificial intelligence is finally getting smart.

(MIT Tech Review)

Both in MIT TR 10 Breakthrough Techn (2017 and 2013)

InstaDeep™

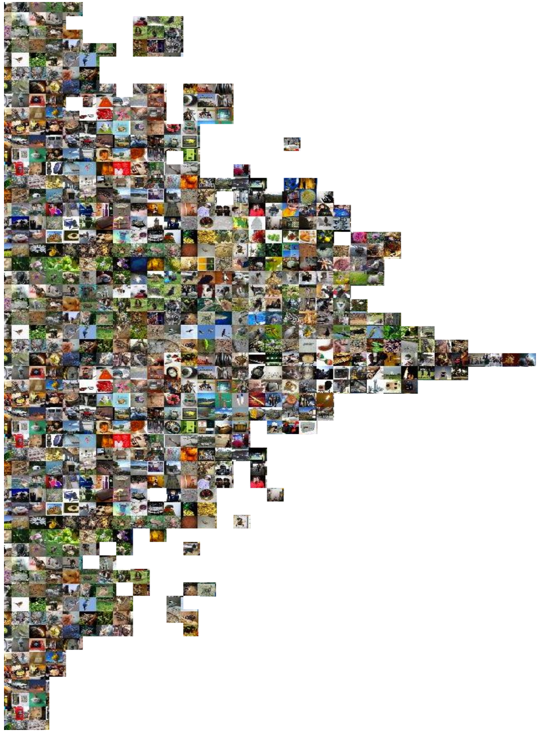
Deep Learning: Pattern Recognition



(Deep Learning, Goodfellow, Bengio & Courville 2016)

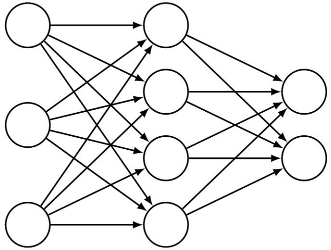
InstaDeep™

Deep Learning: 3 reasons for success

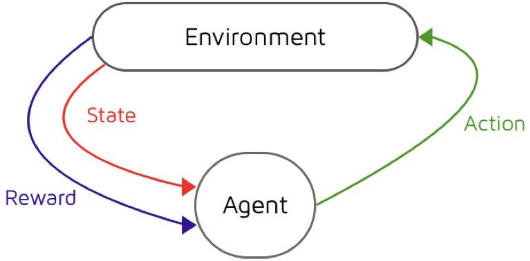


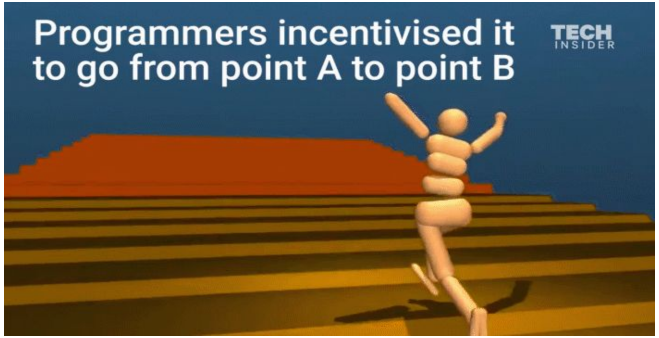
+

+

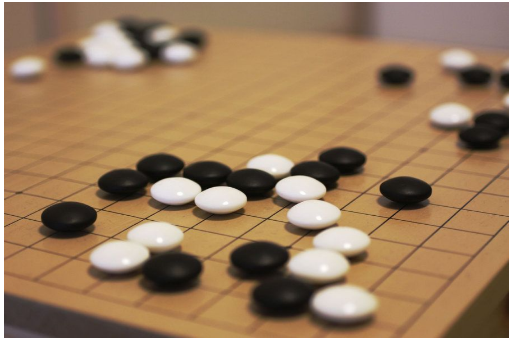


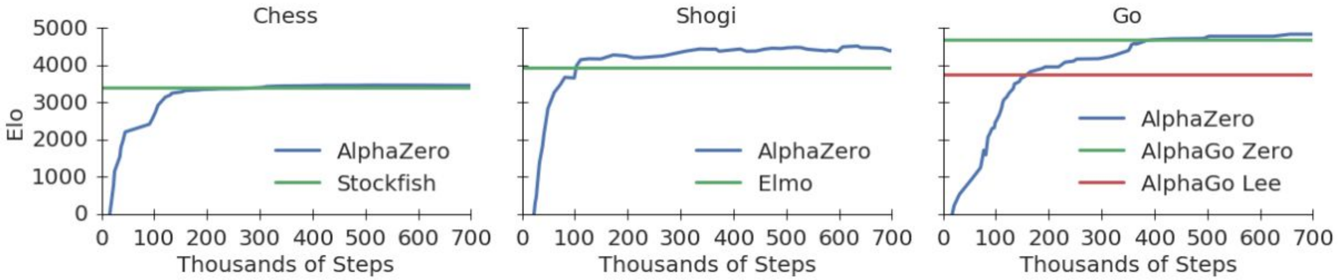
Reinforcement Learning (RL)





A new type of AI Champion





(Silver et Al. 2017)

Starting random, **beats world's best Chess and Go programs in hours** using distributed ML

Future of AI

AI boosts Data

(*Creative Adversarial Nets*, Elgammal & al. 2017)

Painting set	Q1 (std)
CAN	53% (18%)†

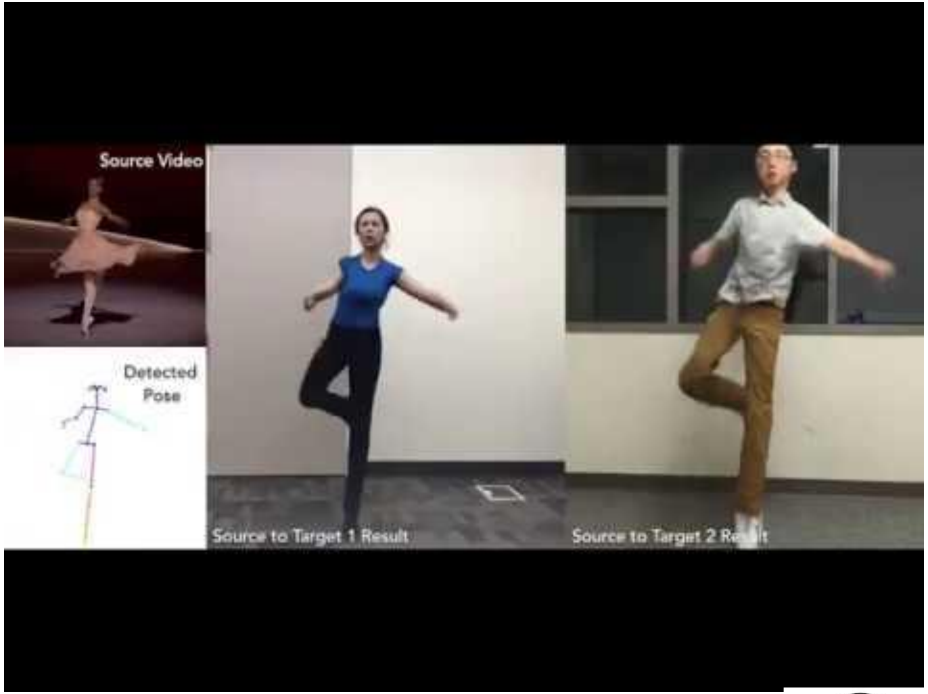
(*Towards Automatic Anime Characters Creation*, Jin & al. 2017)

(*Progressive Growing of GANs*, Karras & al. 2017)

InstaDeep™

AI boosts Data

(Video-to-Video Synthesis, Wang & al. 2018)



(Everybody Dance Now, Chan & al. 2018)

DEEP LEARNING
INDABA

AI boosts GPUs

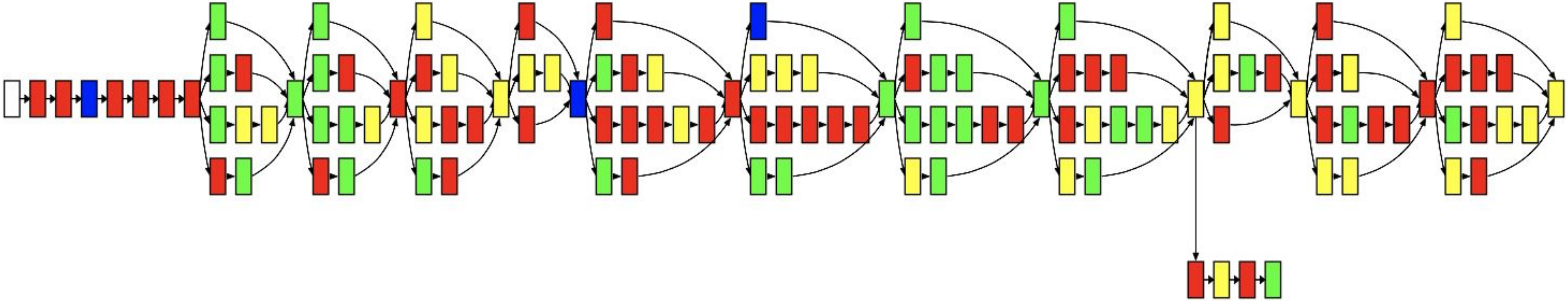


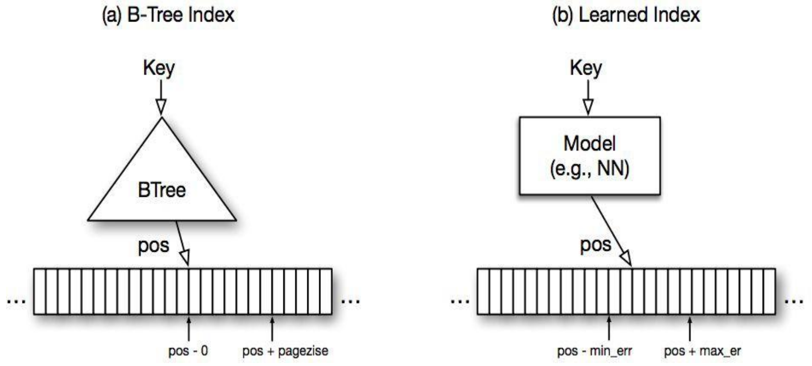
Figure 5. RL-based placement of Inception-V3. Devices are denoted by colors, where the transparent color represents an operation on a CPU and each other unique color represents a different GPU. RL-based placement achieves the improvement of 19.7% in running time

(Device Placement Optimization with Reinforcement Learning, Mirhoseini & al. 2017)

19.7% faster than GPU experts for Deep Learning tasks

AI boosts CS

AI Data Index **60% faster and 20X lighter**



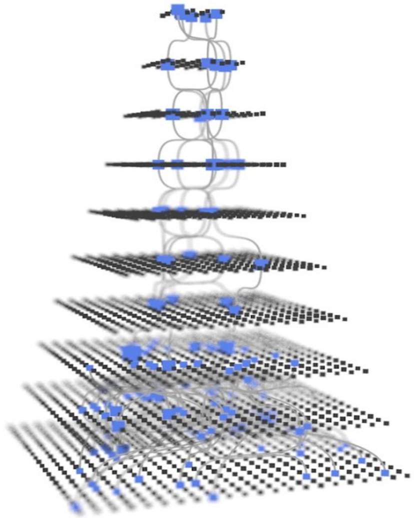
Type	Config	Lookup time	Speedup vs. Btree	Size (MB)	Size vs. Btree
BTree	page size: 128	260 ns	1.0X	12.98 MB	1.0X
Learned index	2nd stage size: 10000	222 ns	1.17X	0.15 MB	0.01X
Learned index	2nd stage size: 50000	**162 ns**	**1.60X**	**0.76 MB**	**0.05X**
Learned index	2nd stage size: 100000	144 ns	1.67X	1.53 MB	0.12X
Learned index	2nd stage size: 200000	126 ns	2.06X	3.05 MB	0.23X

(The Case for Learned Index Structures, Kraska, Jeff Dean & al. 2017)

InstaDeep™

Neural Architecture Search to find a model

Controller: proposes ML models



Iterate to find the most accurate model

Train & evaluate models

Keynote, Jeff Dean, TensorFlow Dev Summit March 2018

Obtained NN:

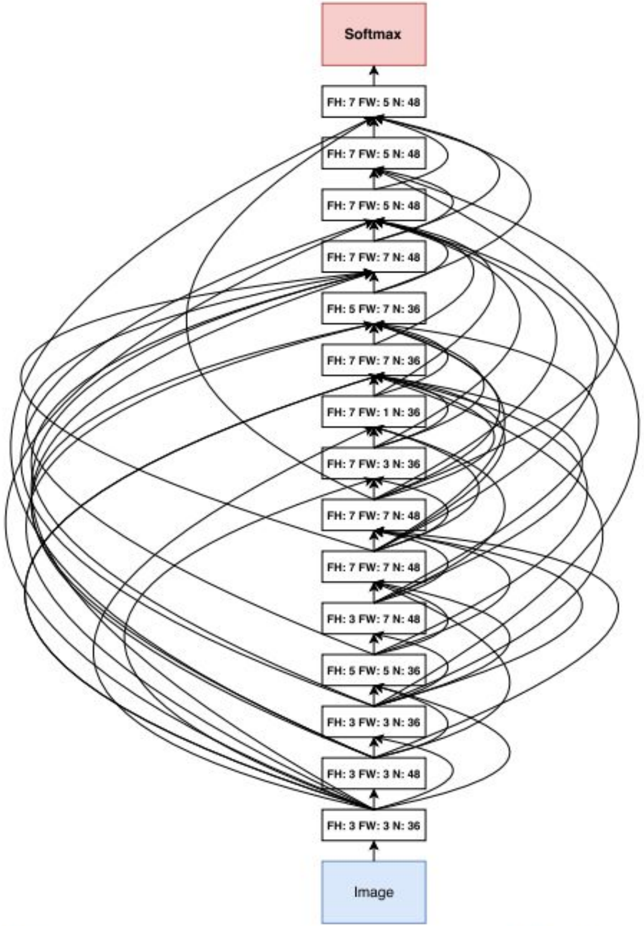
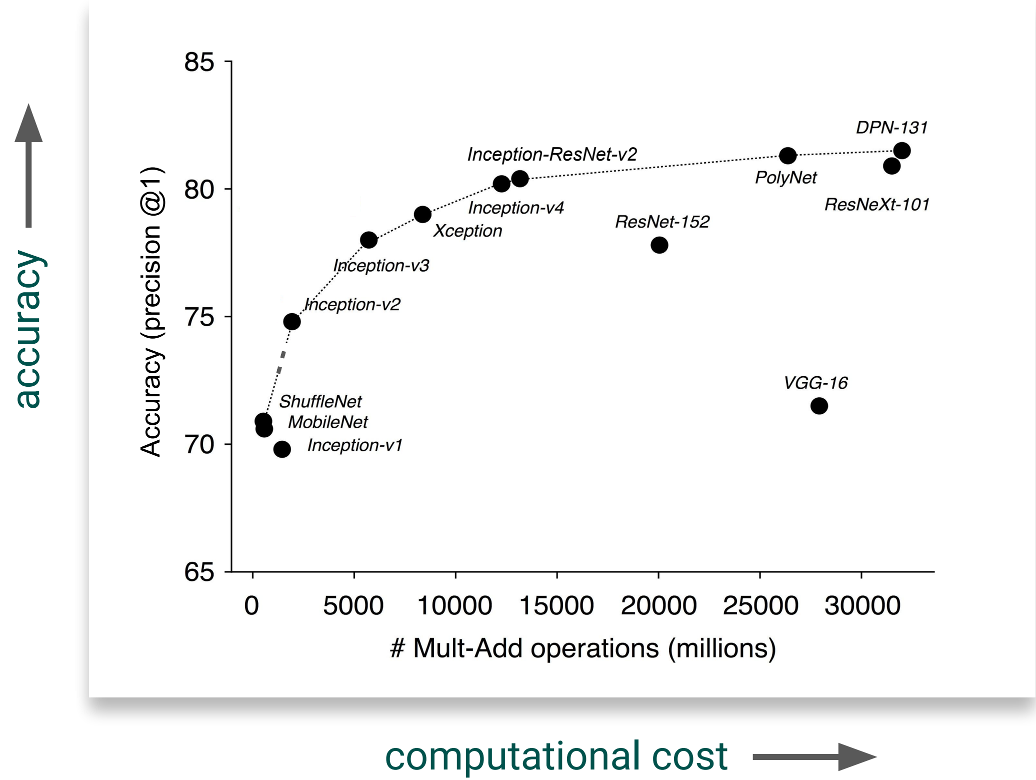


Figure 7: Convolutional architecture discovered by our method, when the search space does not have strides or pooling layers. FH is filter height, FW is filter width and N is number of filters.

Keynote, Jeff Dean, TensorFlow Dev Summit March 2018

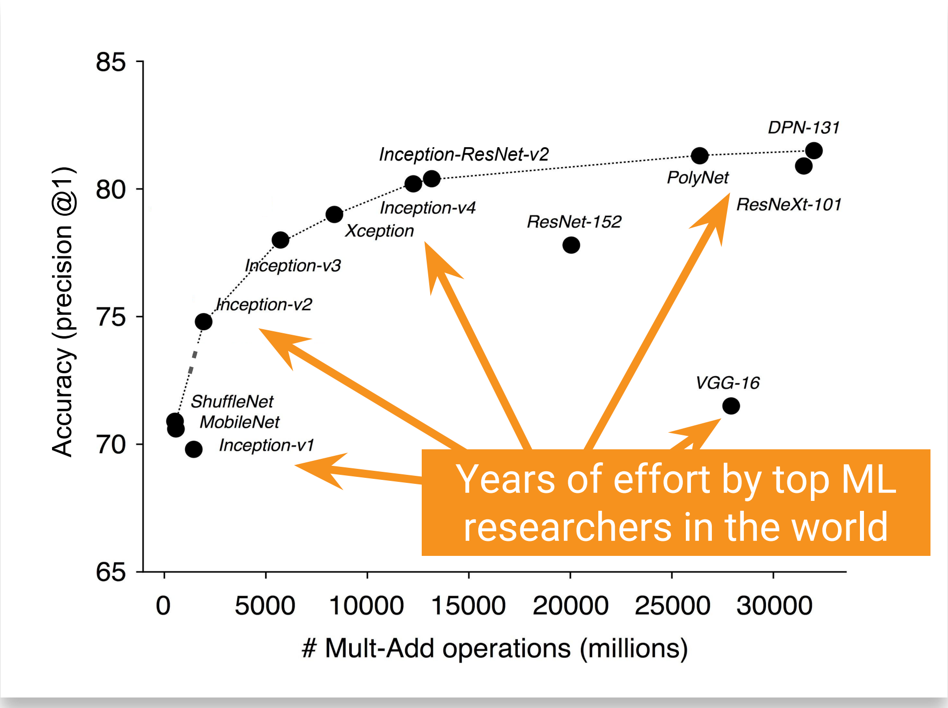
AutoML outperforms handcrafted models



Learning Transferable Architectures for Scalable Image Recognition, Zoph et al. 2017, https://arxiv.org/abs/1707.07012

AutoML outperforms handcrafted models

accuracy

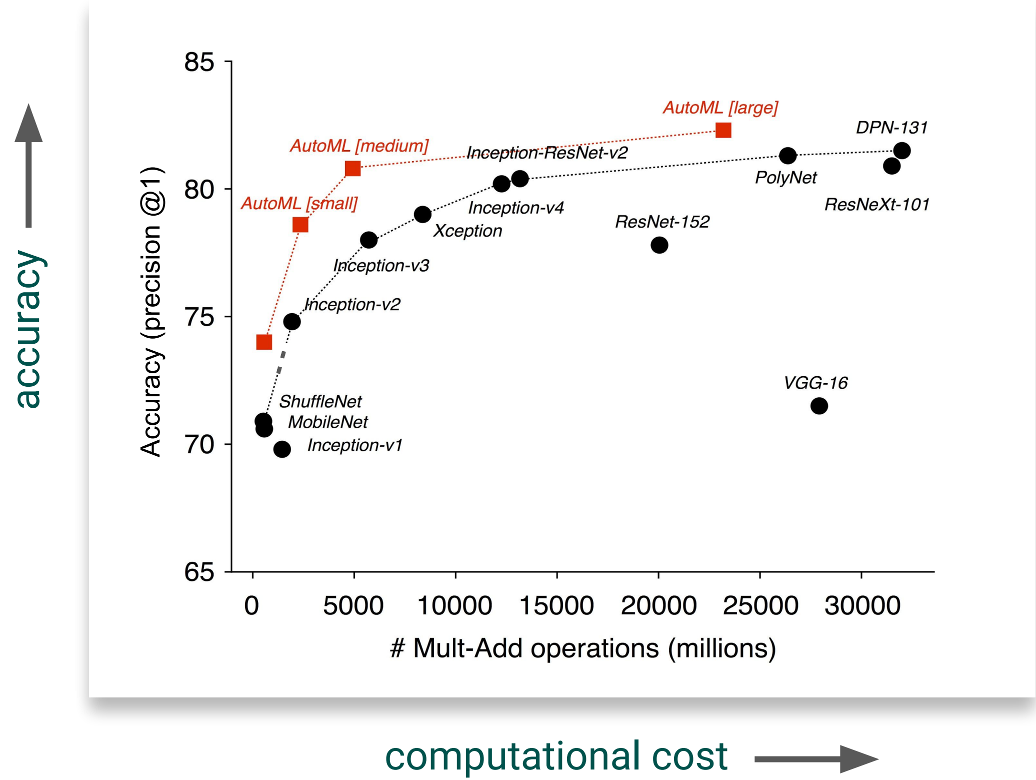


computational cost

Learning Transferable Architectures for Scalable Image Recognition, Zoph et al. 2017, https://arxiv.org/abs/1707.07012

InstaDeep™

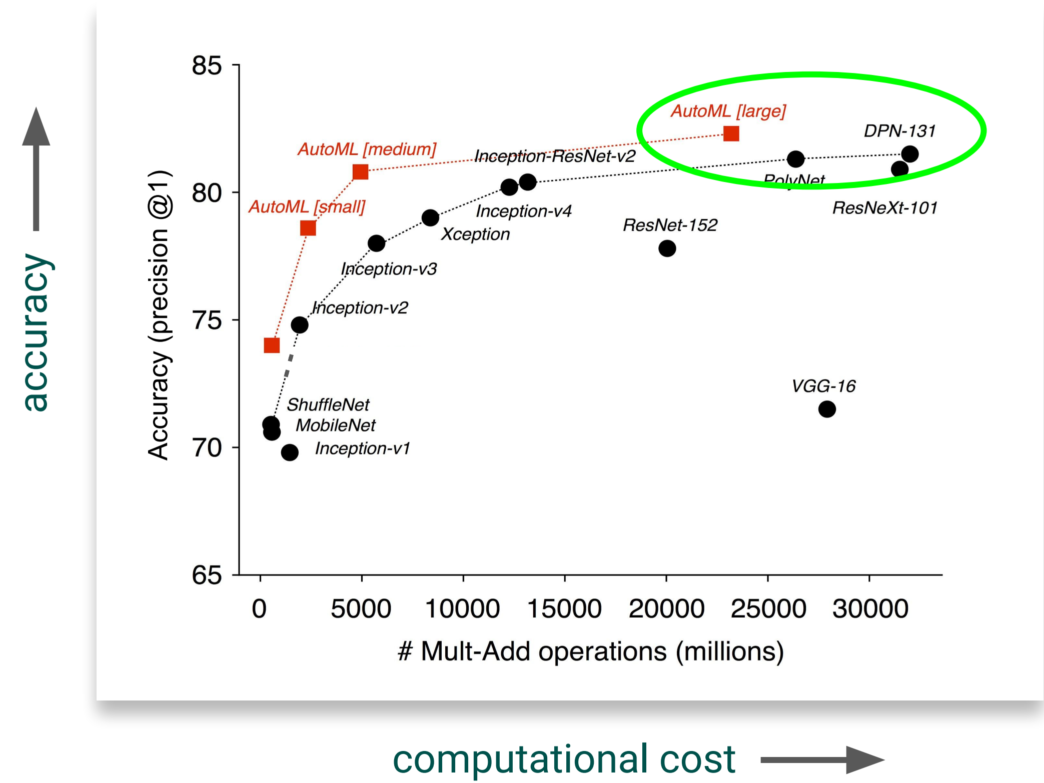
AutoML outperforms handcrafted models



Learning Transferable Architectures for Scalable Image Recognition, Zoph et al. 2017, https://arxiv.org/abs/1707.07012

InstaDeep™

AutoML outperforms handcrafted models



Learning Transferable Architectures for Scalable Image Recognition, Zoph et al. 2017, https://arxiv.org/abs/1707.07012

NAS itself is improving!

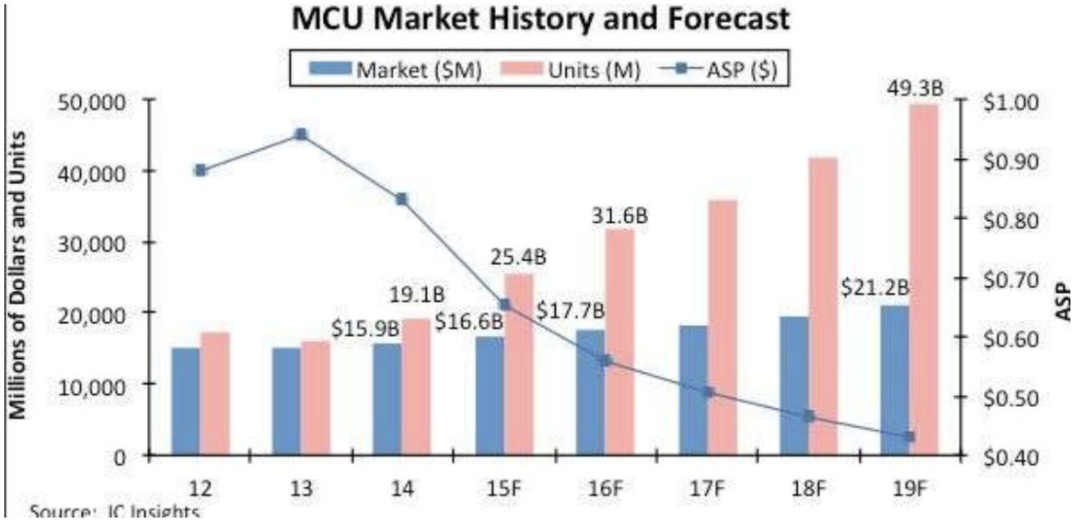
Method	GPUs	Times (days)	Params (million)	Error (%)
NAS (Zoph & Le, 2017)	800	21-28	7.1	4.47
NAS + more filters (Zoph & Le, 2017)	800	21-28	37.4	**3.65**
ENAS + macro search space	1	0.32	21.3	4.23
ENAS + macro search space + more channels	1	0.32	38.0	**3.87**

(Efficient Neural Architecture Search via Parameter Sharing, Pham & al. 2018)

State of the art results with a **1000X speed-up in a year**

AI is getting small

Price of powerful hardware is collapsing. AI models are getting smarter
-> You can run **visual AI on a camera** for **one year** on a **coin battery**!



AI is getting big

OpenAI Five: **~1,000** actions to choose from (Chess **~35**, Go **~250**)
State space: **20,000 Dimensions** (Chess **70D**, Go **400D**)

	OPENAI 1V1 BOT	OPENAI FIVE
CPUs	60,000 CPU cores on Azure	128,000 preemptible CPU cores on GCP
GPUs	256 K80 GPUs on Azure	256 P100 GPUs on GCP
Experience collected	~300 years per day	~180 years per day (~900 years per day counting each hero separately)

Compute: **190 petaflops/s-days** to reach **99.95%** playing rank

DEEP LEARNING
INDABA

AI at InstaDeep

Classic Approaches to Logistics Optimization

Optimization Solvers

- ✓ Guaranteed optimality
- ✗ Poor scalability for large instances

Heuristics

- ✓ Fast and scalable
- ✗ Tailor-made for a specific problem
- ✗ No guarantee of optimality

Monte-Carlo Search Algorithms

- ✓ Fast and scalable
- ✓ No specific heuristic or evaluation function
- ✗ No guarantee of short-term optimality

Common Flaw: No Learning!

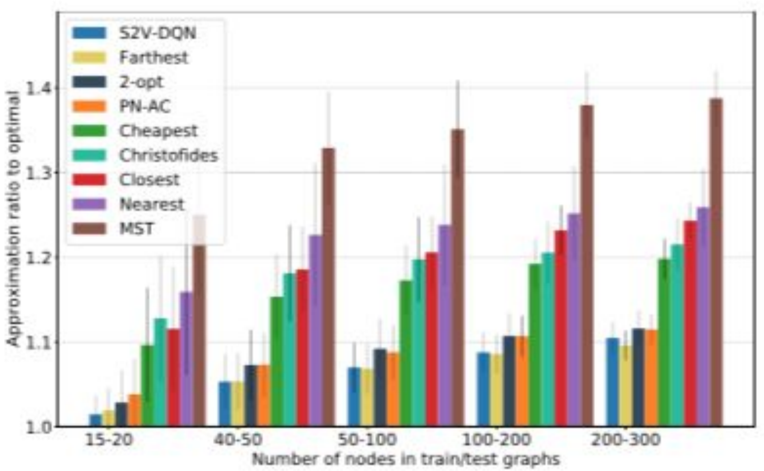
- **Do not** leverage past computations.
- **Unable** to generalize to **unseen** instances
- Might have to be **redesigned** if constraints **change**

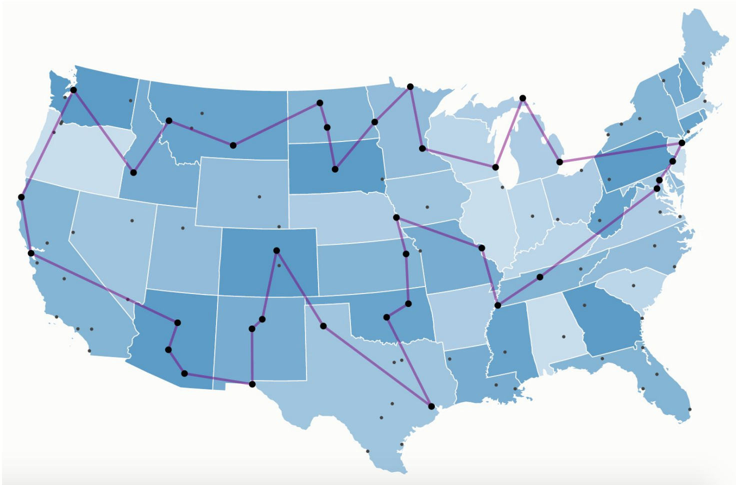
Learning the Traveling Salesman Problem (TSP)

Bello et al. (2017) show how it is possible to train a seq2seq PointerNet to **learn** to minimize trip length **from scratch** using REINFORCE (Williams, 1992):

$$\nabla_\theta J(\theta \mid s) = \mathbb{E}_{\pi \sim p_\theta(.|s)} \left[\left(L(\pi \mid s) - b(s) \right) \nabla_\theta \log p_\theta(\pi \mid s) \right]$$

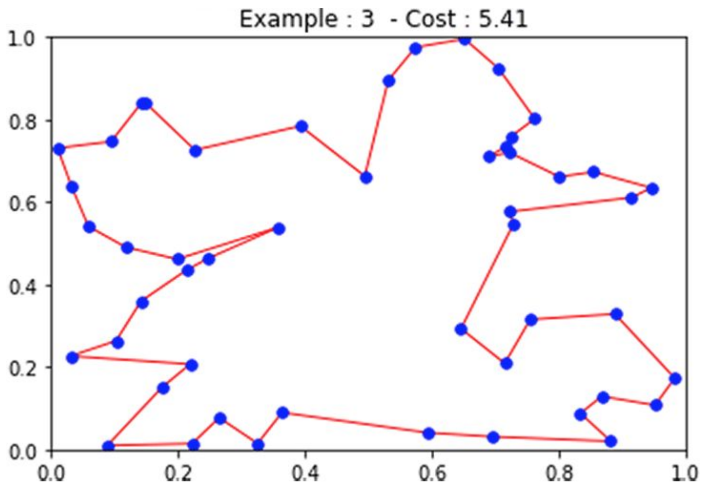
Dai et al. (2017): **data-driven heuristics** using structure2vec and DQN (Mnih, 2015)





Learning the Traveling Salesman Problem (TSP)

Starting from scratch, AI algorithm learns to find better paths progressively:



When compared to Google OR Tools, 2.9% improvement on the TSP-30 Algorithmic Complexity: Exact solver, Held Karp $O(n^2\ 2^n)$ not useful if n>50
Actor-Critic $O(n^2)$ vs heuristics Christophides $O(n^3)$, LKH $O(n^2)$.

InstaDeep™

Learning to Binpack

Packing Items efficiently is another NP-hard logistics problem.

InstaDeep™

Method used (seen here in 2D):

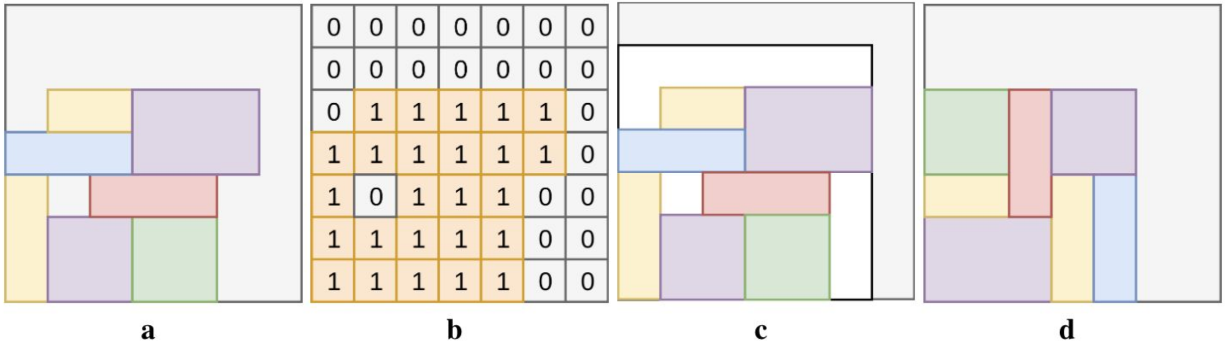
- Neural Nets (“thinking fast”)
- Search using MCTS (“thinking slow”)
- NN computes actions but also leaf values

Beyond the original AlphaZero algorithm:

- **Single-player** game rather than two-player
- Introduction of a **ranked reward** approach
- Environment with **physical constraints**

Learning to Binpack: Setup

Inputs: raw board representation ***s*** and a deep neural network with parameters θ



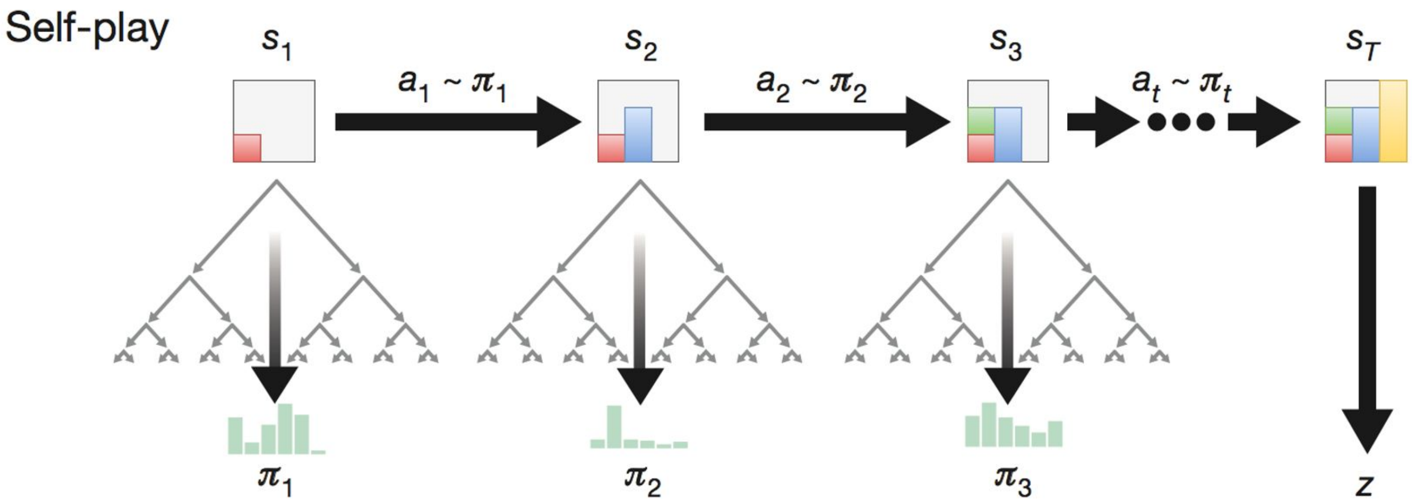
Outputs:

- a vector ***p*** which is a probability distribution over all legally allowed actions in state ***s***. This is the **policy component**
- a number ***v*** that depends on the expected reward ***r*** of packing (max 1.0 and min 0.0) from the current state ***s***. This is the **board evaluation component**

$$(\boldsymbol{p}, v) = f_\theta(s)$$

Learning to Binpack: Ranked Rewards

- Don't use policy ***p*** directly: use MCTS to **build a more accurate policy *π***
- Algorithm plays multiple games, ranks game outcomes ***r*** to obtain a **ranked reward *z*** defined as 1.0 if ***r*** above a set quantile (e.g. 75%) and 0.0 below.
- Reintroduces **relative** rather than **absolute** performance in one-player games.



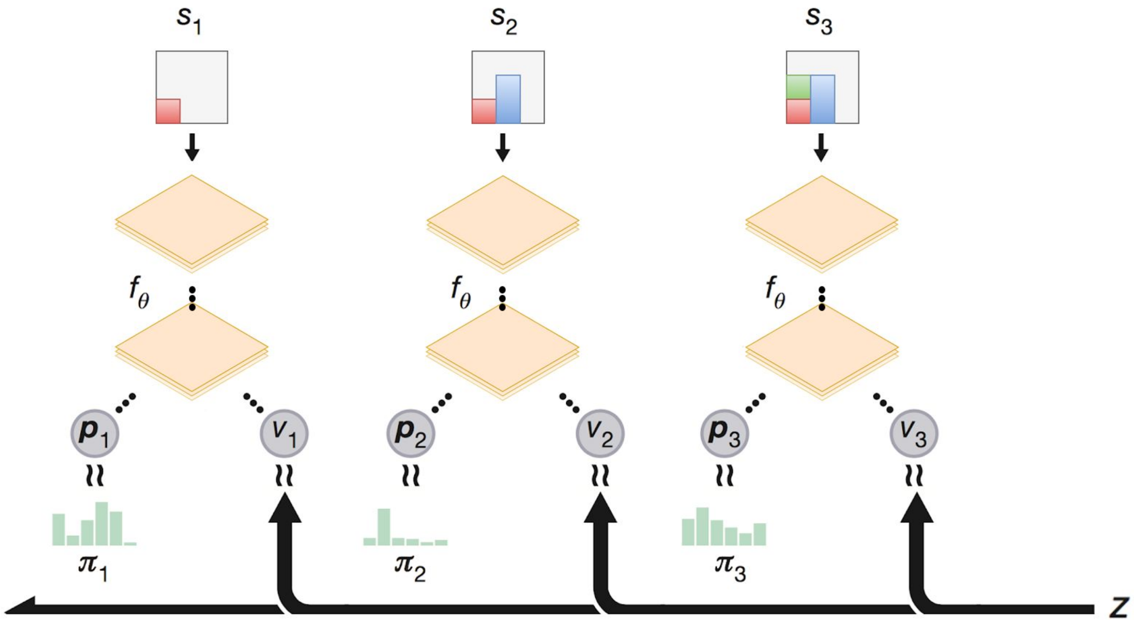
(Graphs from Silver et Al. 2017)

InstaDeep™

Learning to Binpack: Training

- Once a ranked game result z is known, parameters θ of the neural network are re-trained by gradient descent to minimize the loss function l below:

$$l = (z - v)^2 - \boldsymbol{\pi}^{\mathrm{T}} \log \boldsymbol{p} + c\|\theta\|^2 \qquad (\boldsymbol{p}, v) = f_\theta(s)$$



(Graphs from Silver et Al. 2017)

InstaDeep™

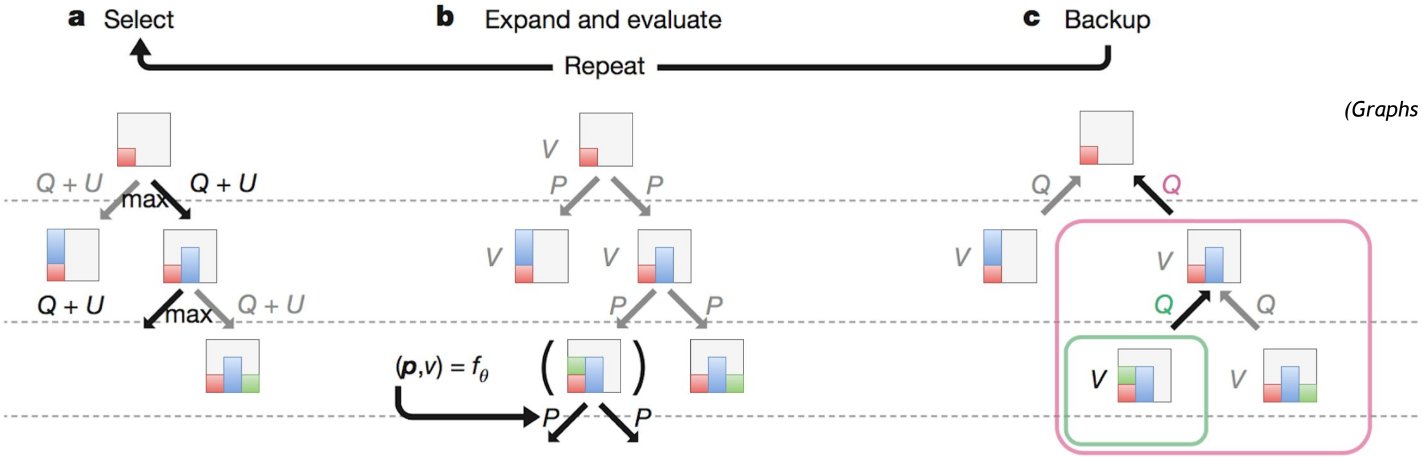
Learning to Binpack: MCTS Detailed

The MCTS decision tree gets built with **300** iterations of 3 steps below:

a) go through the tree till a leaf node by selecting max *Q+U* each time

b) expand that leaf node by computing its ***p*** and *v* values

c) back-propagate from leaf nodes values *v* to estimate values *Q+U*

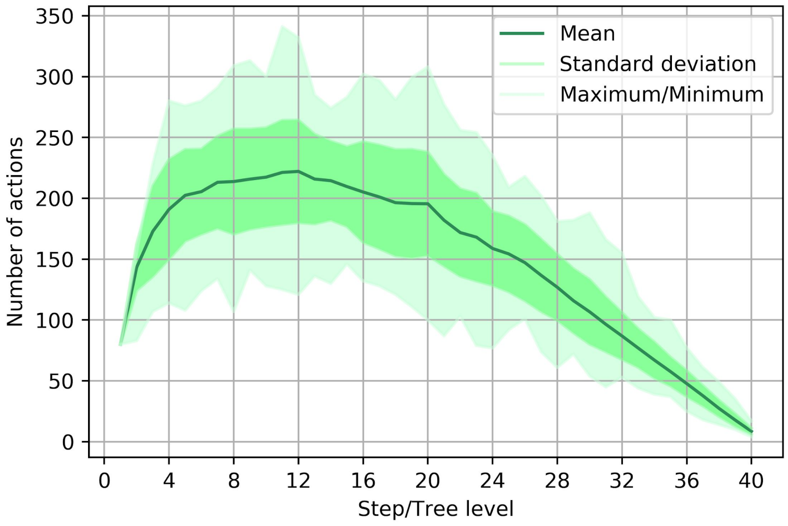


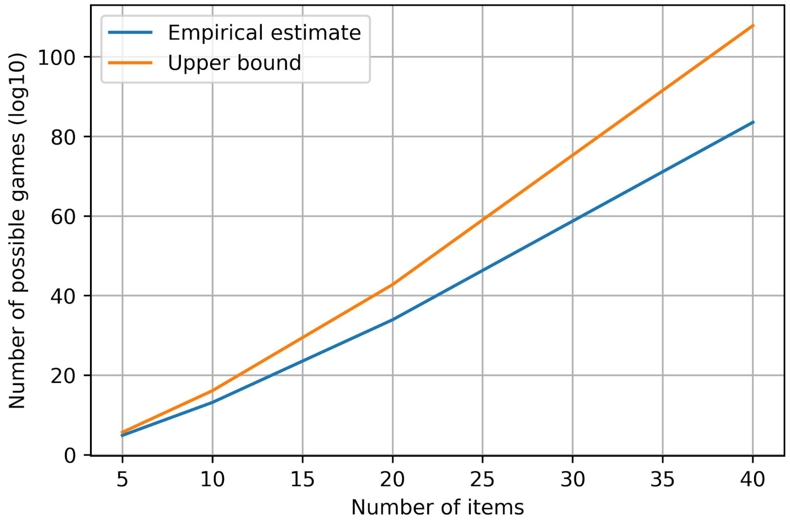
(Graphs from Silver et Al. 2017)

- *Q* is simply the **mean of leaf nodes values *v*** encountered below
- *U* is an extra **upper confidence** term to **favour exploring** new paths

InstaDeep™

Learning to Binpack: Exponential Complexity





- Chess = 10^{120}
- Go = 10^{174}
- Bin packing (40 items) ≈ 10^{84}
- Number of atoms in the known universe = 10^{82}

InstaDeep™

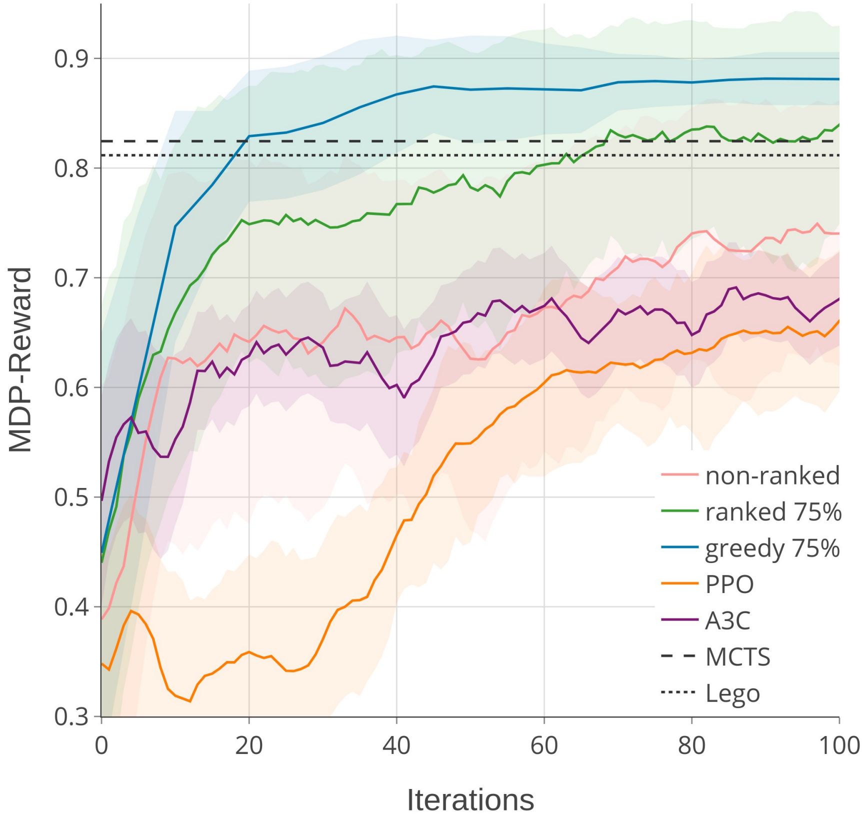
Learning to Binpack: 10 ITEMS RESULTS

- **Ranked** NN+MCTS **outperforms MCTS** and non-ranked NN+MCTS

- **Outperforms Deep RL** (A3C, PPO) and the Lego heuristic

- **Faster inference** than MCTS (no rollouts)

Combining Neural Nets (“thinking Fast”) and MCTS (“thinking slow) unlocks **better performance than taking each separately** and introduces **learnability**.



InstaDeep™

AI & Africa

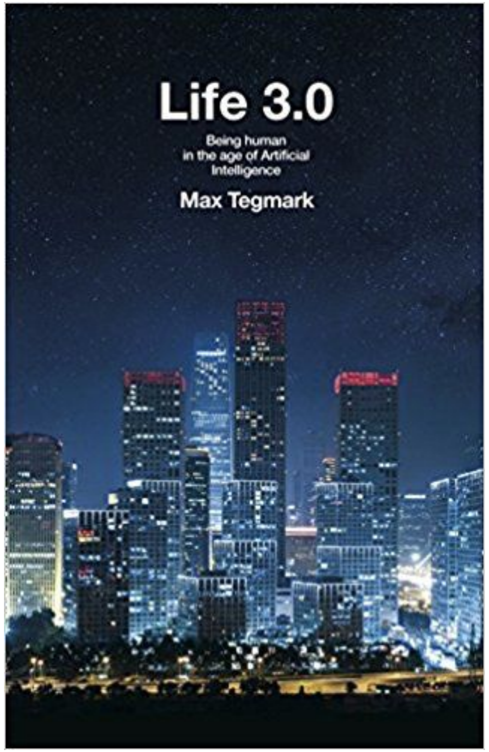
Fundamental Limits of Physics:

Energy: Given an amount of matter,
10 Billion Times more energy

Memory: **10^12 to 10^18** more **information** in a given space

Computation: **10^30 to 10^40** faster than today (!!!)

Be ready AI is **just getting started!**



(Life 3.0, Max Tegmark. 2017)

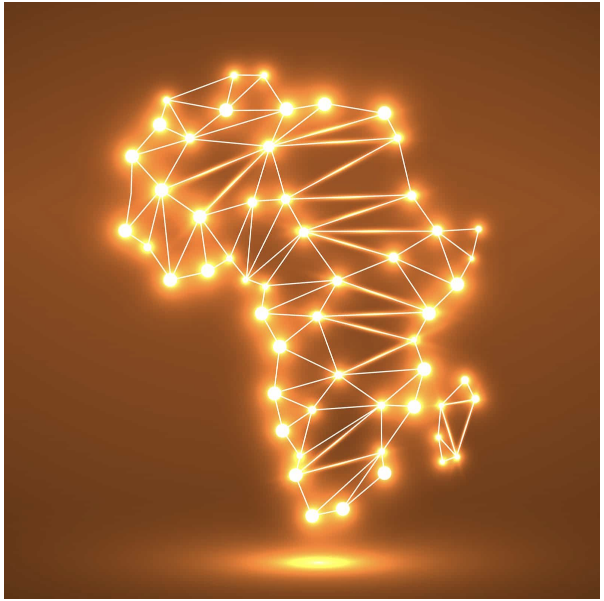
Why AI is REALLY important:

AI is a **threat** for the developing world:

- A lot less value on **energy** resources going fwd
- A lot less of value on **unskilled human work**

AI is an **opportunity:**

- **Low barrier** to entry: just curiosity and willpower
- Few of the limits in physical goods trading apply



InstaDeep's goal: **unlock the AI opportunity** for our communities

Tech Giants believe in Africa

Large pool of talent:

- **60%** of Africans are **under 25** years old
- **Talent** and **resourcefulness** in front of adversity

Initiatives are happening**:**

- Google Brain opening its own **AI Research** center in Ghana
- Facebook-Google **Masters of Machine Intelligence** in Rwanda

Question: can an **African AI startup** have an impact too?

InstaDeep Milestones

Founded in April 2014

Named in **top 20 startups** at Mobile World Congress (March 2017)

Invited to meet **Mark Zuckerberg** at table event (April 2017)

Joined **Nvidia's Inception AI** program (May 2017)

Selected by **Google** to become an **AI/ML Dev Expert** (Nov 2017)

Mentored at **Google Launchpad Accelerator Africa** (Mar 2018)

Published 1st **AI Research** article (July 2018)

African ML is getting organized now:

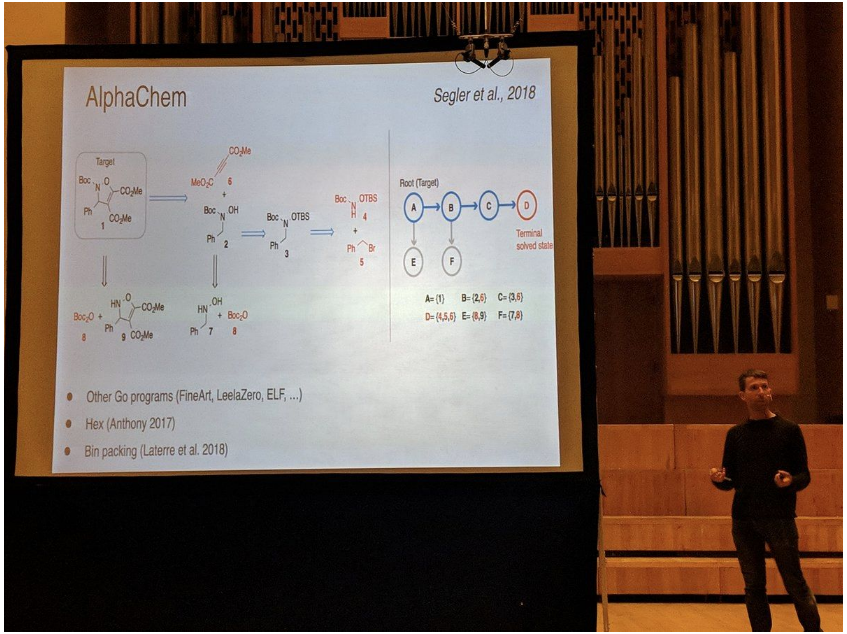
Major AI/ML event: Deep Learning Indaba 2018

550 students attended. Partners included:

IBM amazon.com DeepMind Google Microsoft InstaDeep™

DL Indaba 2018:

We won top **AI Research poster prize** out of 250 submissions!
David Silver mentioned our AI work in his lecture as a "Deep RL success" (!!!)



Conclusion

- ML is delivering **results** in speech, vision, NLP, chess, design, healthcare etc.

- **AI accelerating AI** through creative **data**, better **hardware** and better **models**. **The gap** between **AI-led** organizations and others will **increase**.We are very far from fundamental physical limits, **AI is just getting started.**

- The AI opportunity in Africa is **NOW**. If we believe in ourselves, we can prove to the world that Africans can contribute to the **AI revolution**.

Contact: **kb@instadeep.com** **@kbeguir**

DSN Data Science Nigeria

InstaDeep™

Google Developers Experts