NIPS 2016 Highlights

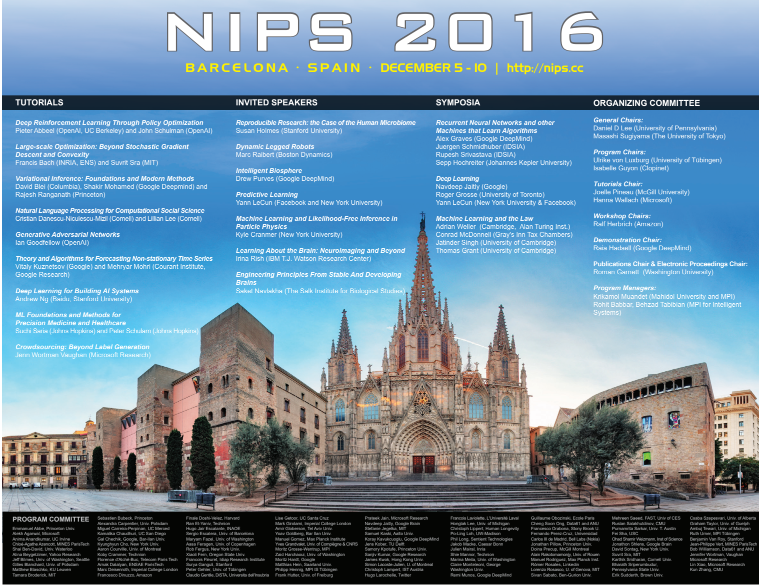
Norman Hendrich

University of Hamburg, Informatics Department
Vogt-Koelln-Str. 30, D-22527 Hamburg, Germany

17 Januar 2017

Outline

Neural Information Processing Systems 2016



Barcelona

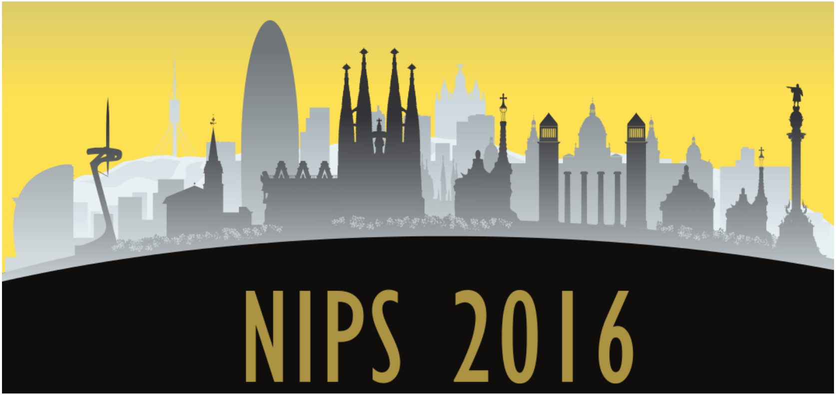
- capital city of Catalonia region
- 2nd largest city in Spain
- 1.6 M people (4.7 M urban area)

- gothic, art-nouveau/modernist and modern architecture icons
- hills, harbour, and beaches
- party all year round
- UNESCO world heritage
- Olympic games 1992
- FCB, Camp Nou, Lionel Messi

- NIPS-16 at CCIB conference center
- conveniently, 70 m from the beach

Las Ramblas, Parc Güell, Beach

Barcelona Landmarks



NIPS 2016

- 05-10 Dec 2016, Centre Convencions Internacional Barcelona
- organizers:
  - Daniel D. Lee, Masashi Sugiyama, General Chairs
  - Ulrike von Luxburg, Isabelle Guyon, Program Chairs

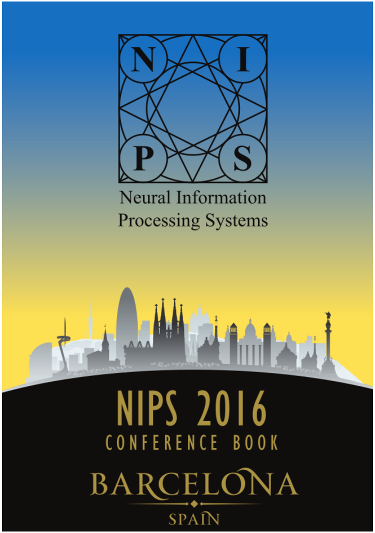
- 2500 submissions, 568 accepted (22%)
- 45 oral presentations (2%), 523 posters (20%)
- 13674 reviews

- two track conference, four poster sessions
- 1 plenary lecture: Yann LeCun „Predictive Learning"
- 9 tutorials, 3 symposia, 50 workshops
- ca. 40 exhibitors
- ca. 6000 participants

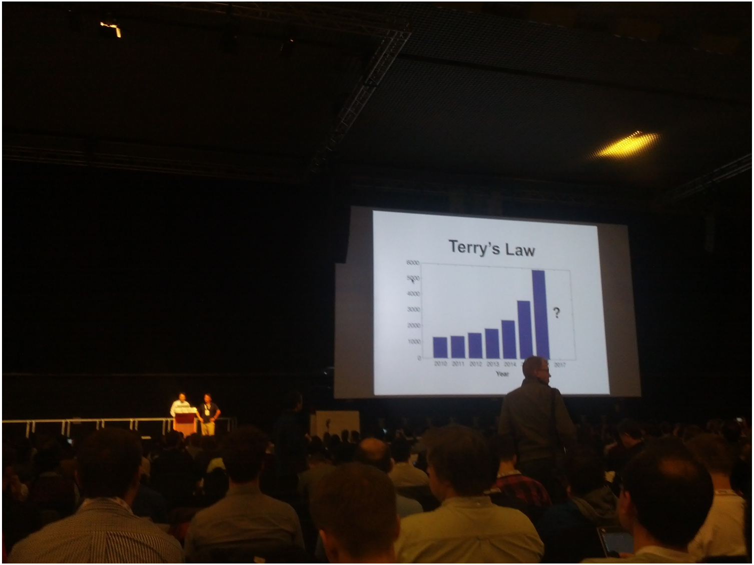
- open access: digest and all papers online

http://nips.cc/

- deep learning everywhere
  - image recognition
  - MNIST, CIFAR, …
  - recursive architectures
  - deep RL and A3C
- lots of Bayesian approaches
- few papers on „classical“ algorithms
- no SVMs at all

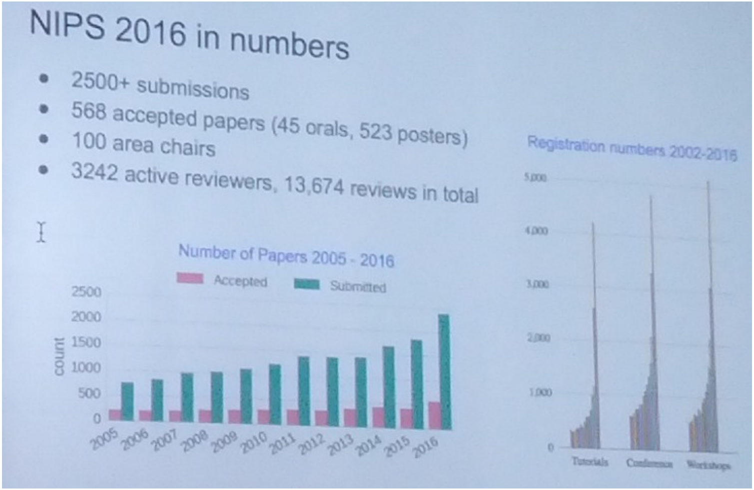


Opening Ceremony: Terry's Law



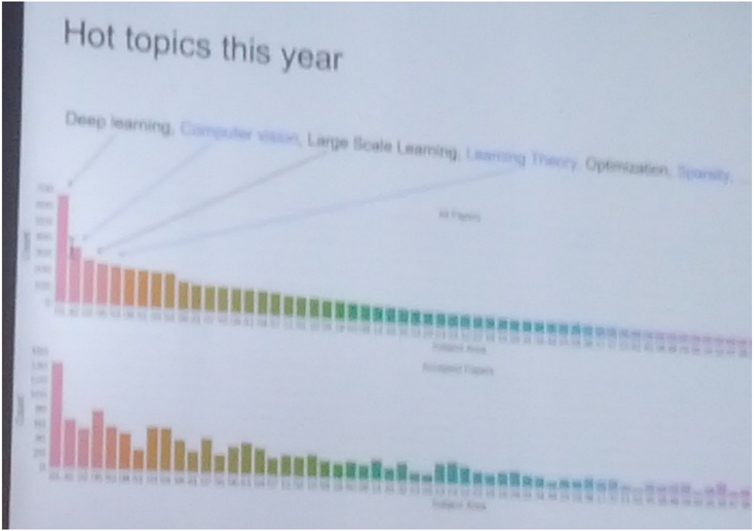
Opening Ceremony: Program Chairs

Submissions and Reviews

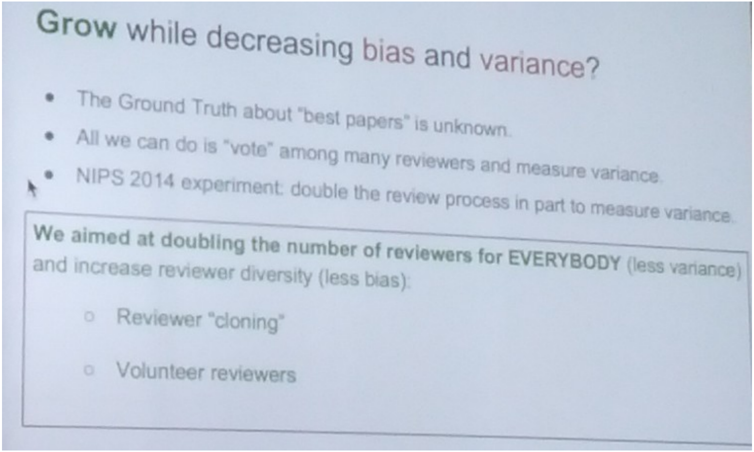


Hot Topics

Deep Learning, CV, Large-scale, Theory, plus a large tail



Reviews: Decrease Bias and Variance :-)



Reviews: Quality of Submissions

Quality of submissions

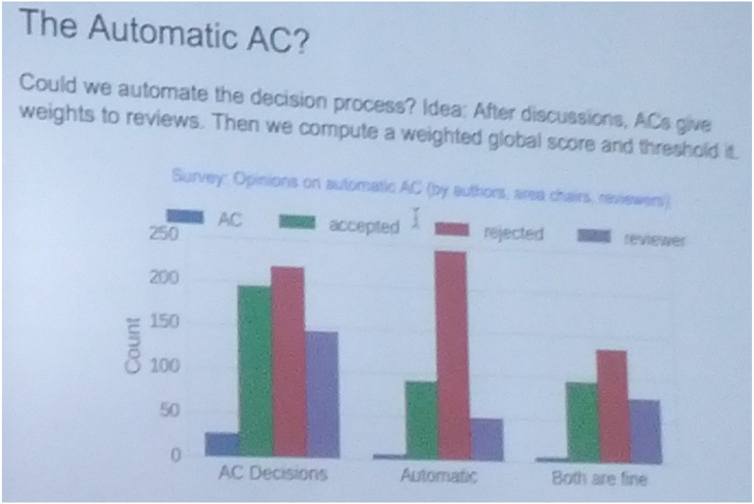
Distribution of Review Scores (in percent of submissions)

	1	2 (sub-standard)	3 (poster level: top 30%)	4 (oral level: top 3%)	5 (award level: top 0.1%)
Impact	6.5	36.1	45.7	10.5	1.1
Quality	6.7	38	44.7	9.5	1.1
Novelty	6.4	34.8	48.1	9.7	1.1
Clarity	7.1	28	48.6	14.6	1.8

More than expected!

Reviewers found: over 50% of the submissions of NIPS poster quality or better

Reviews: Automatic Area Chairs?



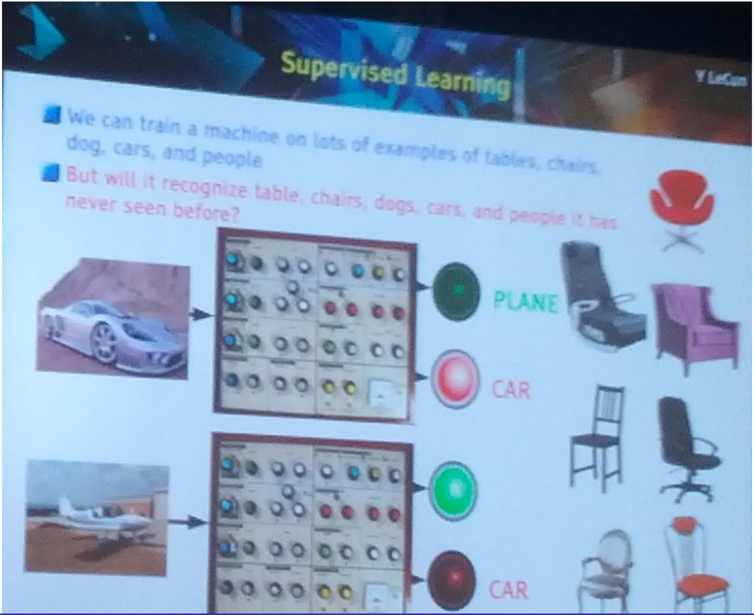
NIPS 2016
- Review Process Analysed

Plenary Speech and Tutorials
- Predictive Learning
- Nuts and Bolts of Deep Learning
- Deep RL Tutorial
- Generative Adversarial Networks

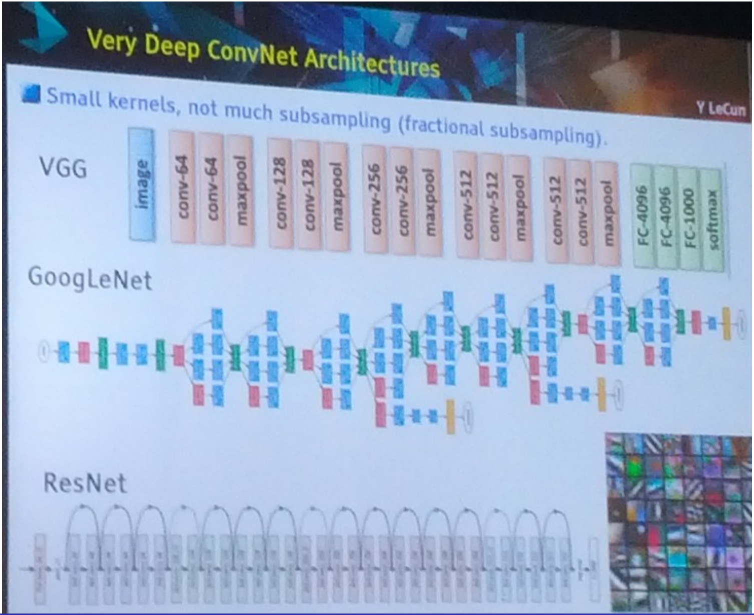
Workshops
- Efficient Methods for DNNs
- Continual Learning Workshop
- Neurobotics Workshop

Yann LeCun Plenary Talk: Predictive Learning

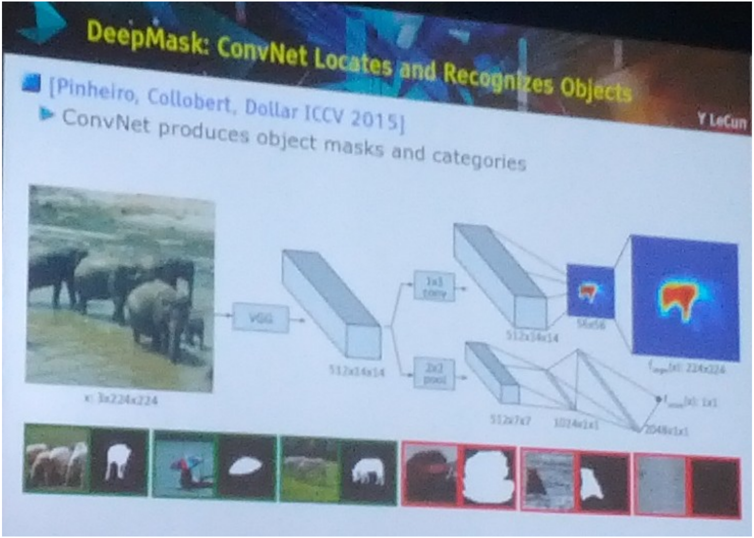
Supervised Learning



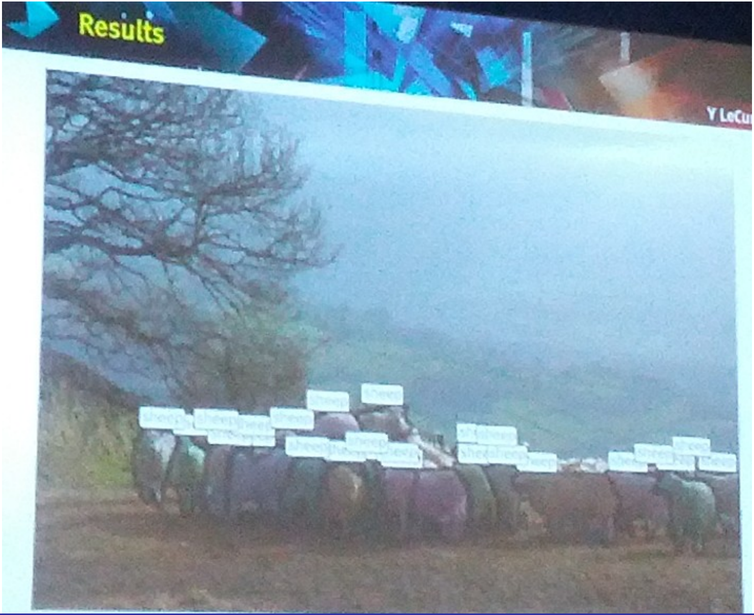
Very Deep Convolutional Nets



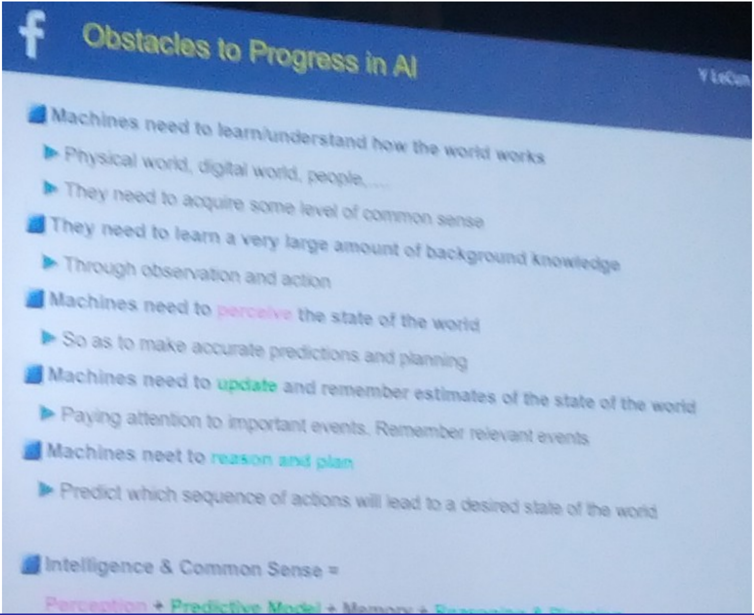
DeepMask


DeepMask: ConvNet Locates and Recognizes Objects
Y LeCun
[Pinheiro, Collobert, Dollar ICCV 2015]
ConvNet produces object masks and categories
VGG
512x14x14
x: 3x224x224
512x7x7
1024x1x1
2048x1x1
224x224

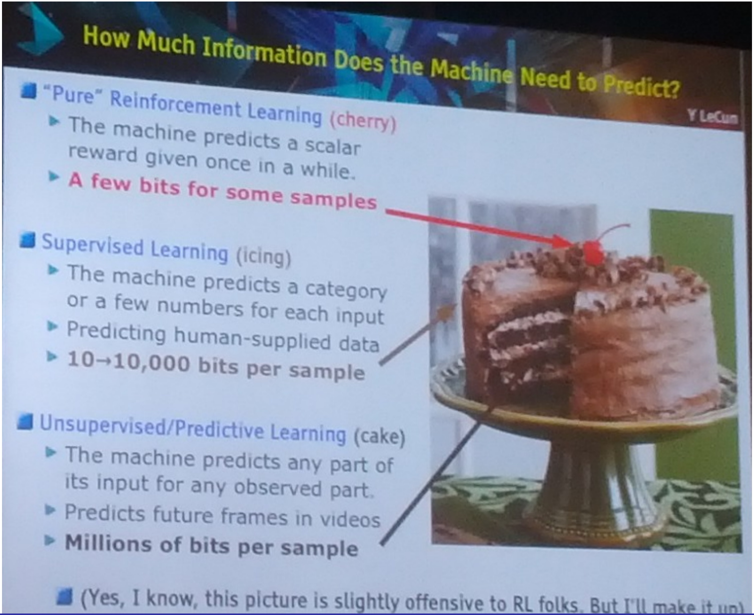
DeepMask: Sheep



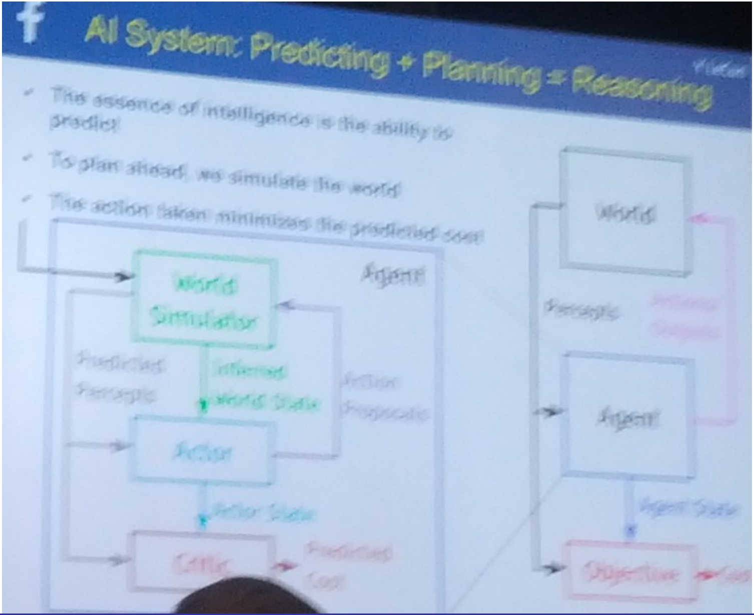
Obstacles to Progress in AI



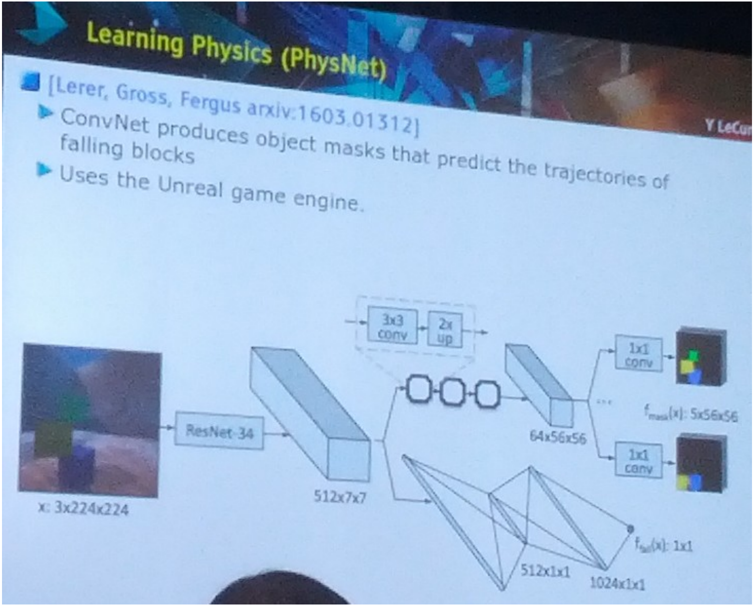
Learning Efficiency: the Cake Analogy



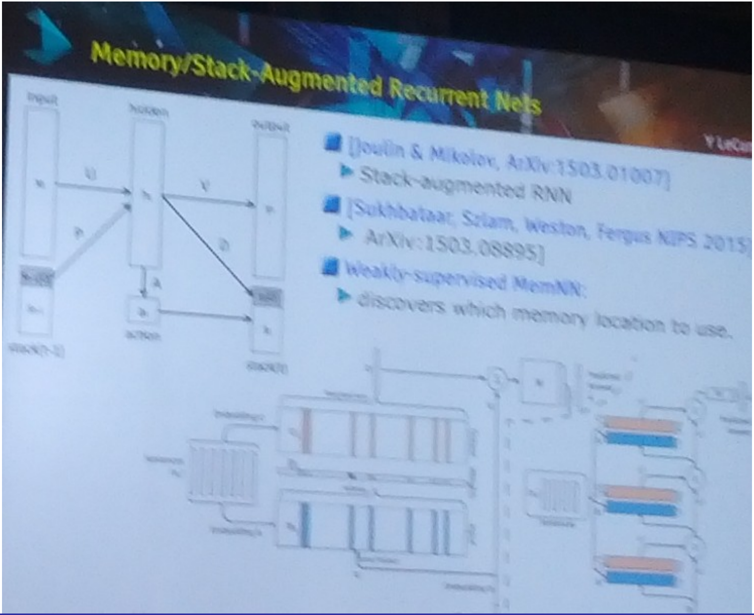
Predicting + Planning = Reasoning



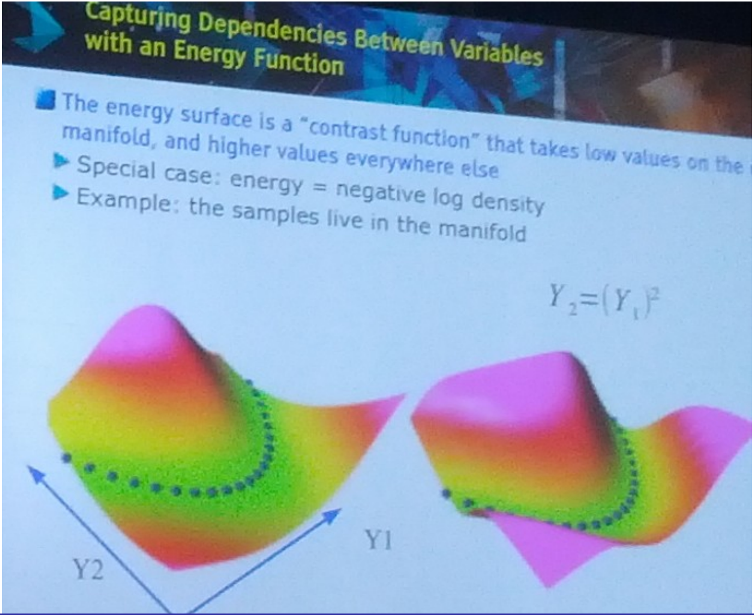
PhysNet: Learning Physics



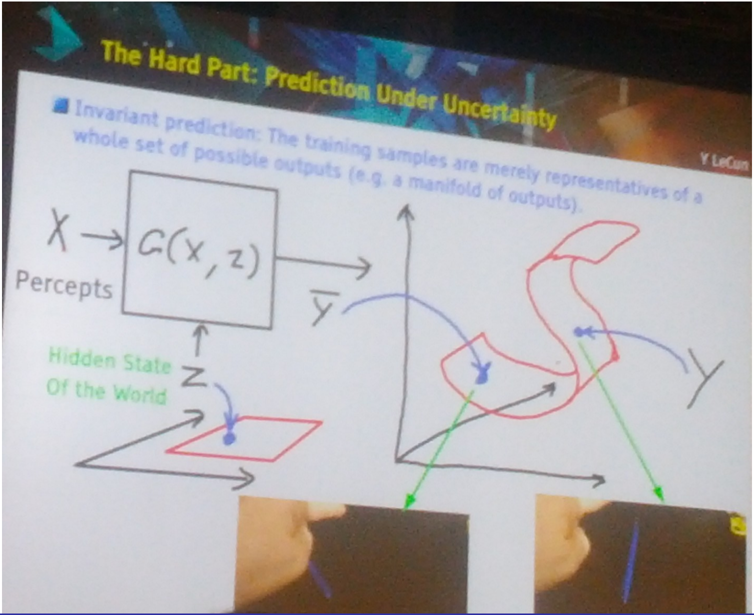
Memory/Stack-Augmented Recurrent Nets



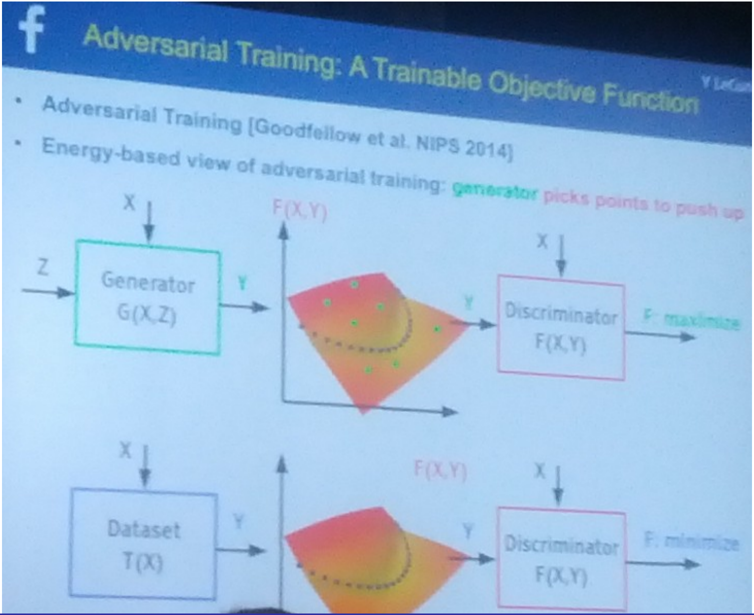
Capturing Dependencies with an Energy Function



Train GAN to Predict World Behaviour



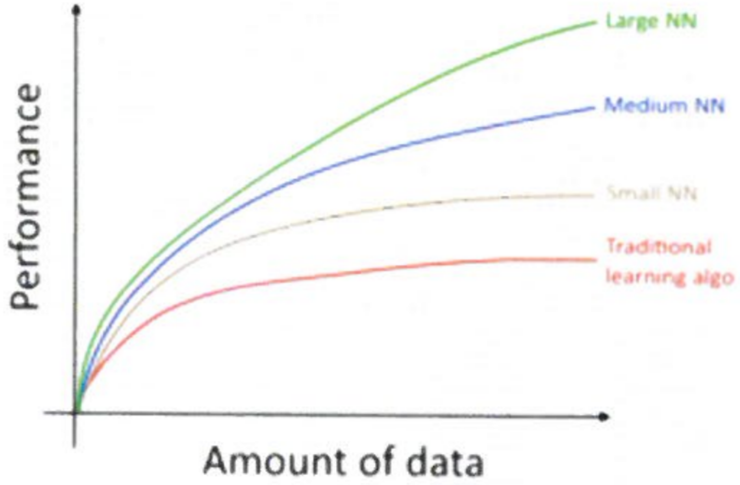
Adversarial Training



Poster-Session

Giveaways: New Tux Mutations

Tencent Labs



Andrew Ng, handout slides

The Rise of End-to-End Learning

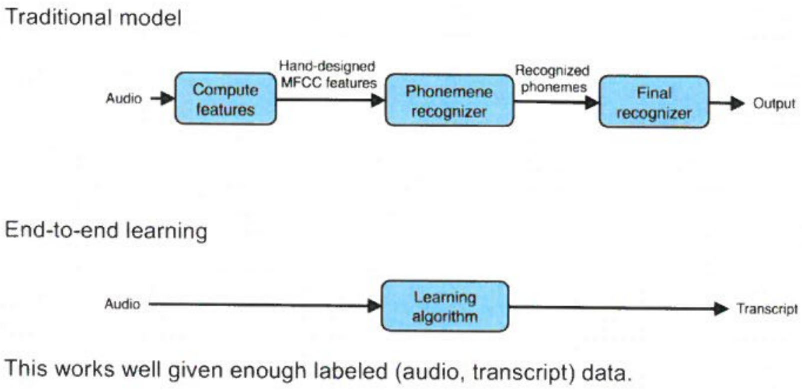
Learning with integer or real-valued outputs:

Problem	X	Y
Spam classification	Email	Spam/Not spam (0/1)
Image recognition	Image	Integer label
Housing price prediction	Features of house	Price in dollars
Product recommendation	Product & user features	Chance of purchase

Learning with complex (e.g., string valued) outputs:

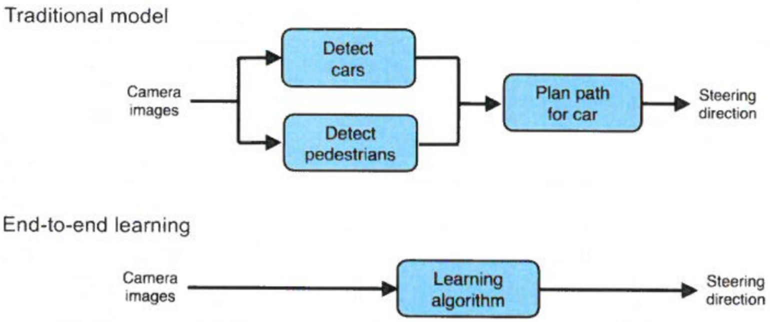
Problem	X	Y	Example
Image captioning	Image	Text	Mao et al., 2014
Machine translation	English text	French text	Suskever et al., 2014
Question answering	(Text,Question) pair	Answer text	Bordes et al., 2015
Speech recognition	Audio	Transcription	Hannun et al., 2015
TTS	Text features	Audio	van der Oord et al., 2016

End-to-End Learning: Speech Recognition



- the era of hand-coded features is over…

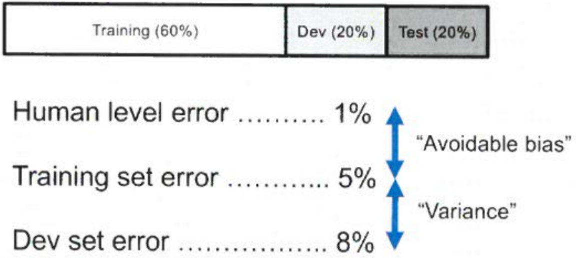
End-to-End Learning: Autonomous Driving



Given the safety-critical requirement of autonomous driving and thus the need for extremely high levels of accuracy, a pure end-to-end approach is still challenging to get to work. End-to-end works only when you have enough (x,y) data to learn function of needed level of complexity.

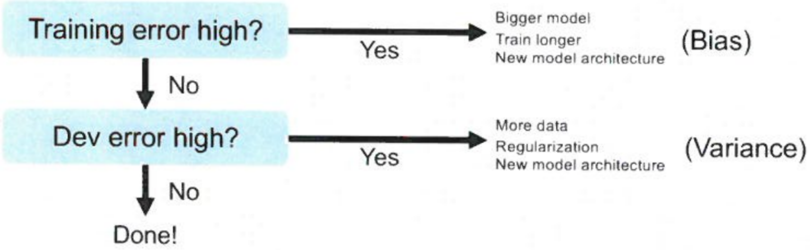
Traditional Training Data Split

Say you want to build a human-level speech recognition system. You split your data into train/dev/test:



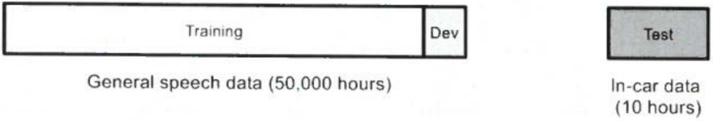
Compared to earlier eras, we still talk about bias and variance, but somewhat less about the "tradeoff" between them.

Basic Machine Learning Recipe



Check your Training Data Distributions

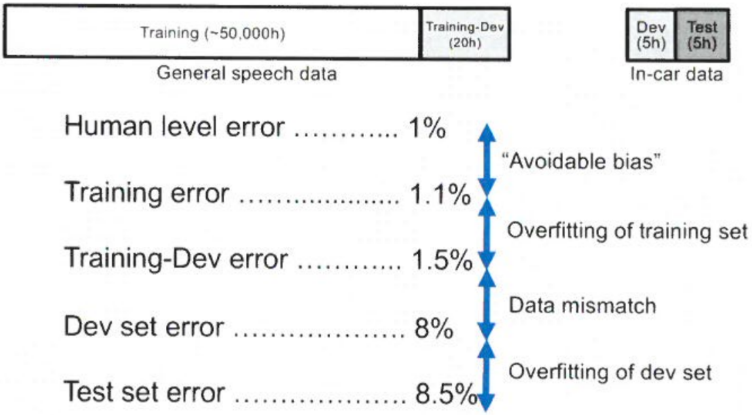
Say you want to build a speech recognition system for a new in-car rearview mirror product. You have 50,000 hours of general speech data, and 10 hours of in-car data. How do you split your data? This is a **bad** way to do it:



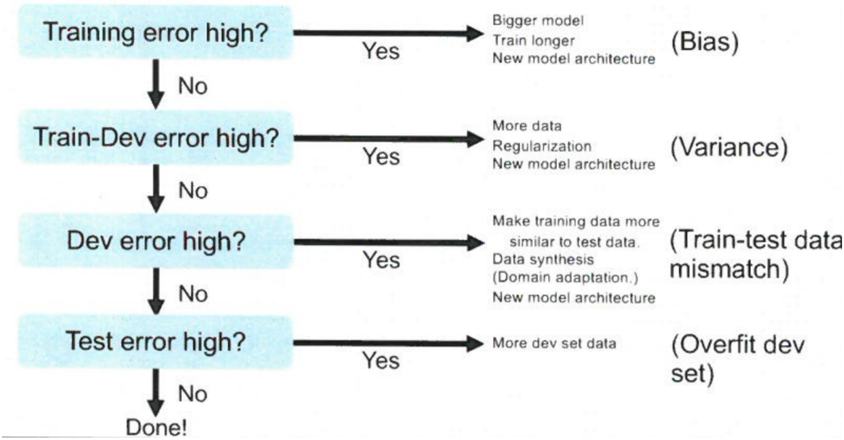
Having mismatched dev and test distributions is not a good idea. Your team may spend months optimizing for dev set performance only to find it doesn't work well on the test set.

Better Training Data Split

Better way: Make the dev and test sets come from the same distribution.



Improved Machine Learning Recipe



General Human/Bias/Variance Analysis

	General speech data (50,000 hours)	In-car speech data (10 hours)
Performance of humans	Human-level error	(Carry out human evaluation to measure.)
Performance on examples you've trained on	Training error	(Insert some in-car data into training set to measure.)
Performance on examples you haven't trained on	Training-Dev error	Dev/Test error

"Avoidable bias"

"Variance"/degree of overfitting

Data mismatch

Human Level Performance

You'll often see the fastest performance improvements on a task while the ML is performing worse than humans.

- Human-level performance is a proxy for Bayes optimal error, which we can never surpass.
- Can rely on human intuition: (i) Have humans provide labeled data. (ii) Error analysis to understand how humans got examples right. (iii) Estimate bias/variance. E.g., On an image recognition task, training error = 8%, dev error = 10%. What do you do? Two cases:

Human level error 1%
"Avoidable bias"
Training set error 8%
"Variance"
Dev set error 10%

Focus on bias.

Human level error 7.5%
"Avoidable bias"
Training set error 8%
"Variance"
Dev set error 10%

Focus on variance.

Quiz: Medical Imaging

Suppose that on an image labeling task:

Typical human .... 3% error
Typical doctor .... 1% error
Experienced doctor .... 0.7% error
Team of experienced doctors .... 0.5% error

What is "human-level error"?

Answer: For purpose of driving ML progress, 0.5% is best answer since it's closest to Bayes error.

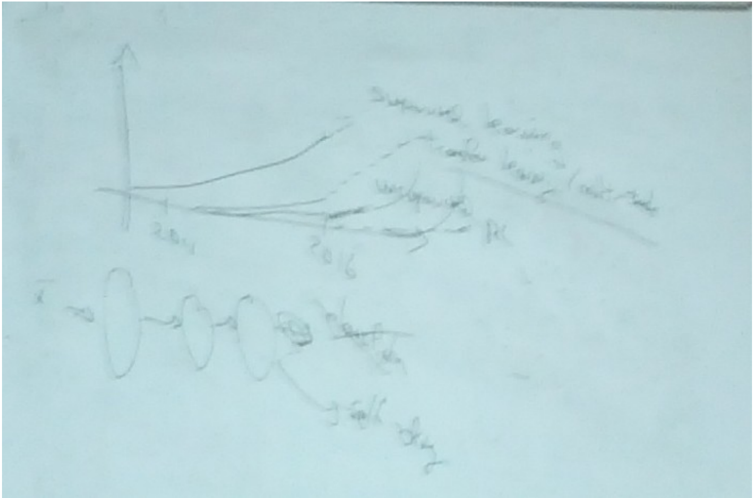
OT: Stanford PhD Recipe

proven Stanford Univ. approach to PhD research:

- read (lots of) papers
- try to reproduce their results
- understand what works and what doesn't

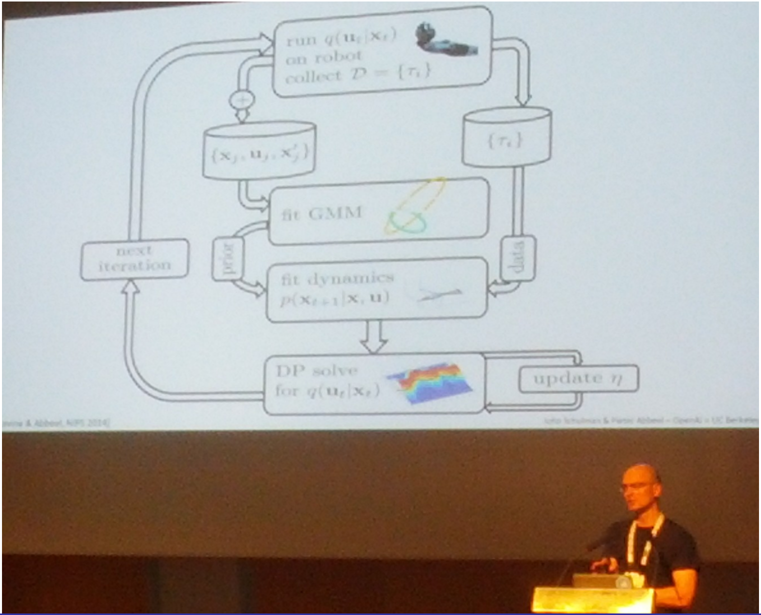
- get your hands dirty (real data, real systems)

Deep-Learning Outlook…

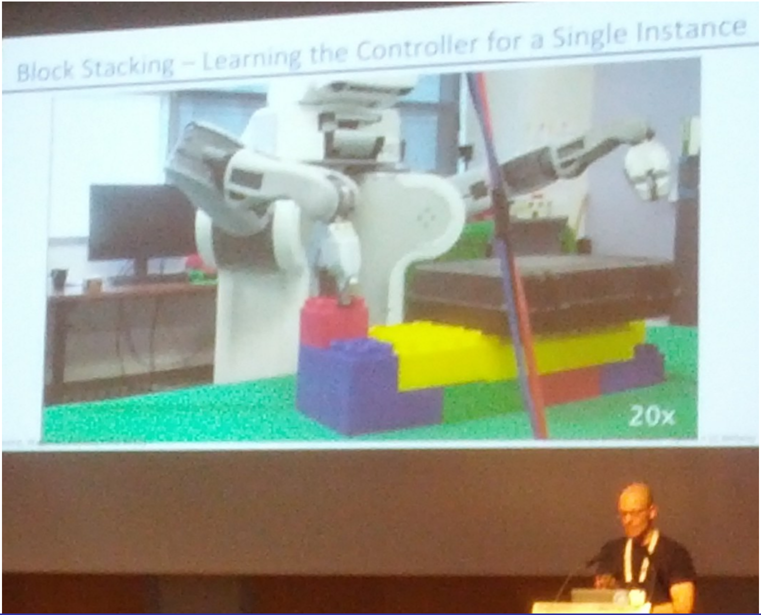


- low-hanging fruit: supervised, larger datasets, transfer learning
- mid-term: unsupervised learning
- far in the future: real robust RL applications

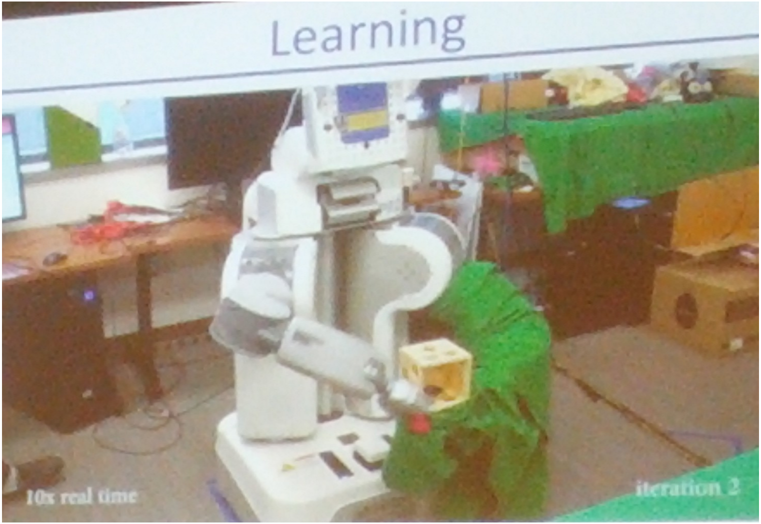
Abbeel & Shulman: Deep RL Tutorial



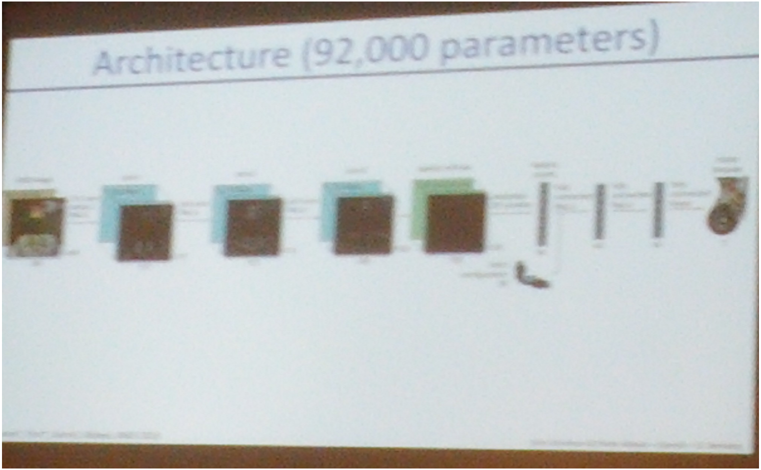
PR2 Manipulation Learning: Duplo



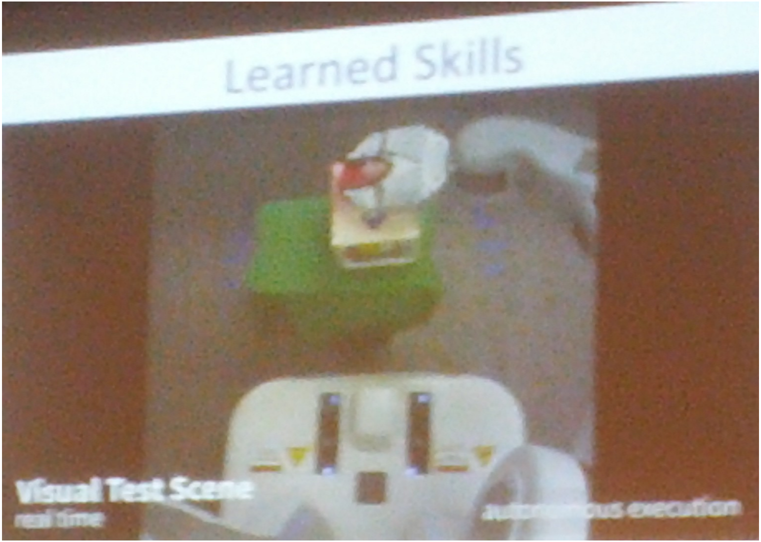
PR2 Manipulation Learning: Shape-sorting Box



PR2 Manipulation Learning: Network Archictecture



PR2 Manipulation learning: Autonomous Execution



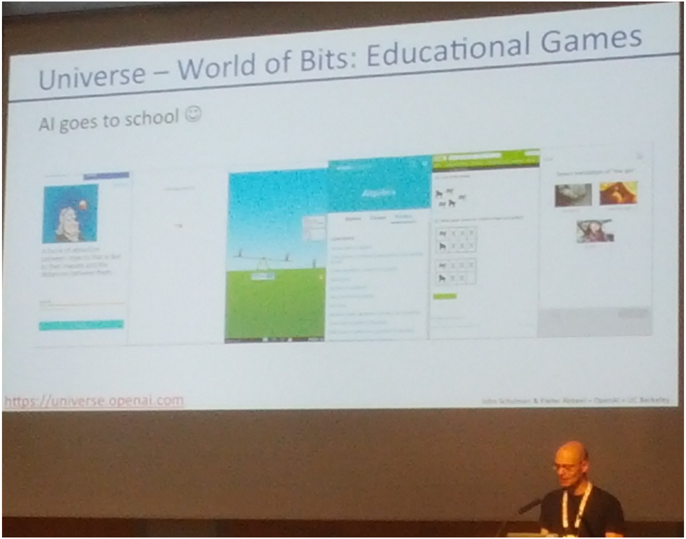
Abbeel: Current Frontiers

- off-policy policy gradients / Off-policy actor-critic
- exploration
- auxiliary objectives
- multi-task and transfer (incl. sim2real)
- meta-RL
- 24/7 data collection
- safety
- architectures
- inverse RL, model-based RL, hierarchical RL

How to get started?

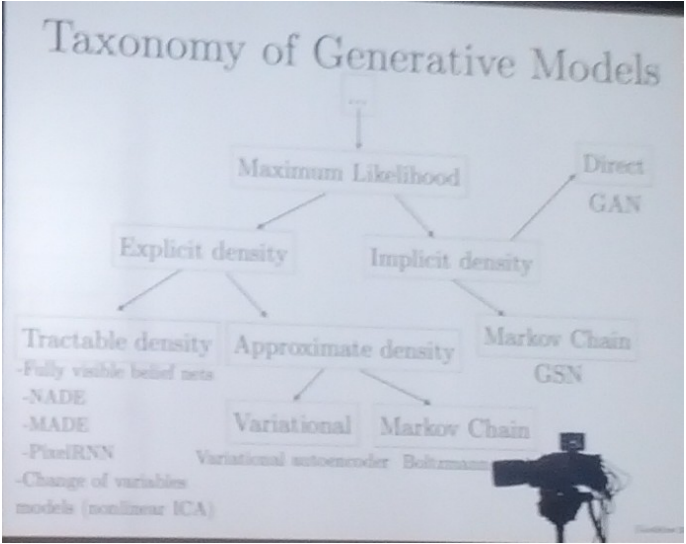
- deep RL courses on the web
- deel RL code bases
- environments, e.g. OpenAI Universe

Universe — World of Bits: Educational Games

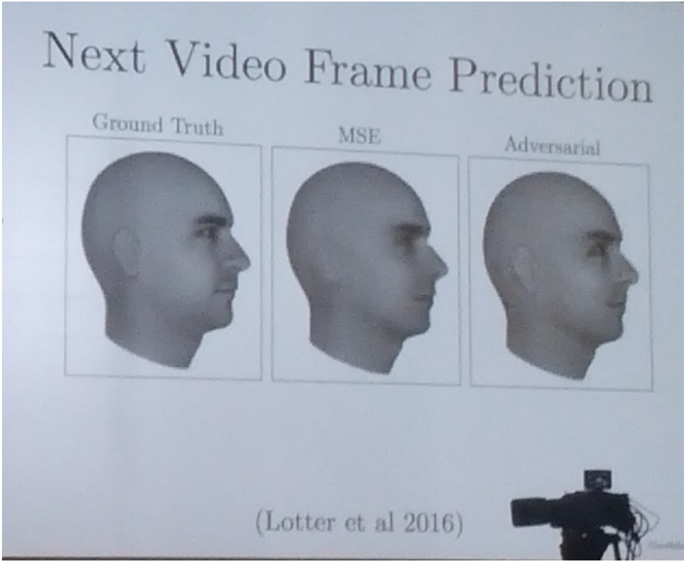


https://universe.openai.com

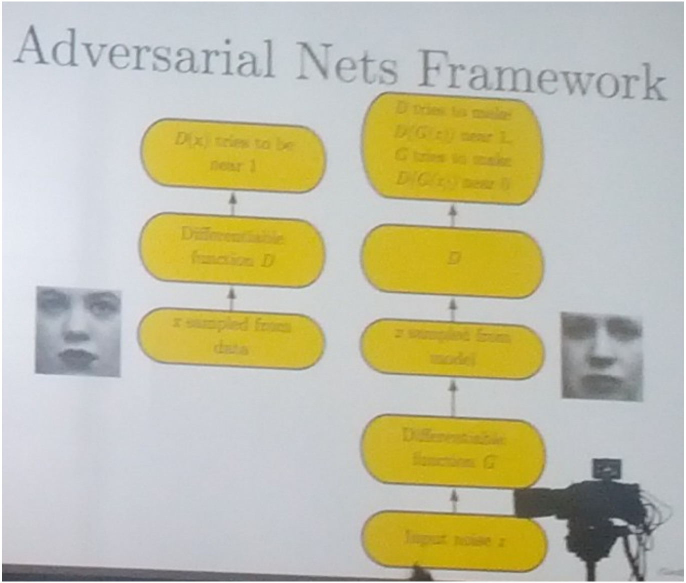
Ian Goodfellow: Generative Adversarial Networks



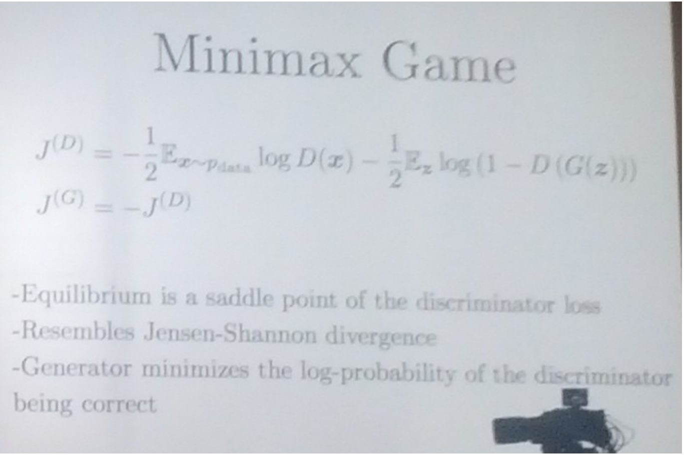
GAN: Video Frame Prediction

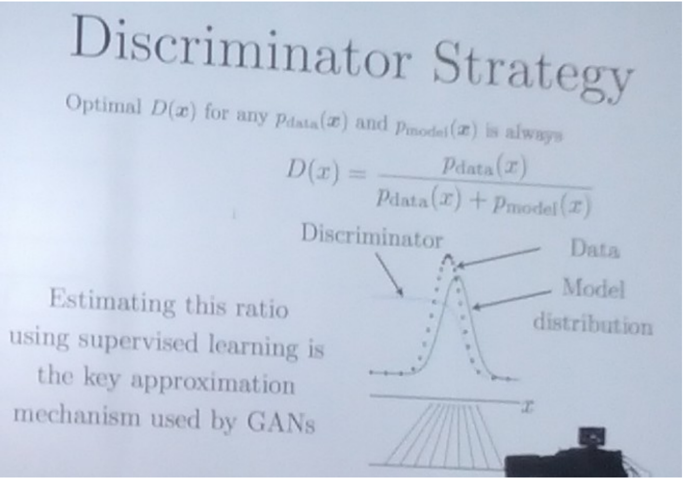


GAN: General Framework



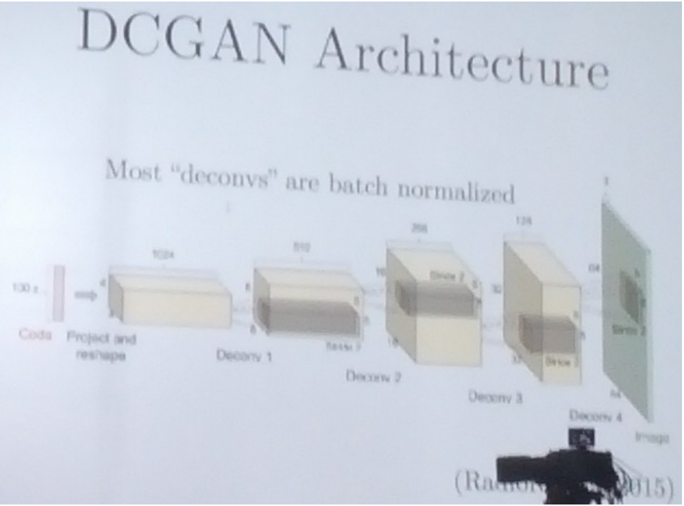
GAN: Minimax Game



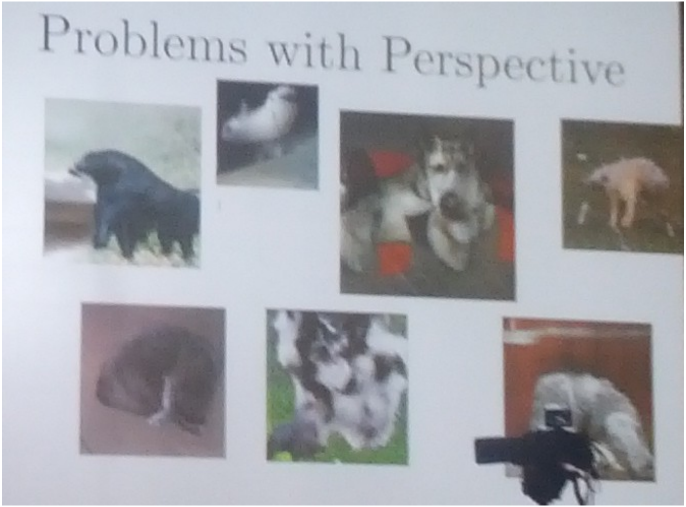


Discriminator Strategy
Optimal $D(x)$ for any $p_{data}(x)$ and $p_{model}(x)$ is always
$$D(x) = \frac{p_{data}(x)}{p_{data}(x) + p_{model}(x)}$$
Discriminator
Data
Model
distribution
Estimating this ratio
using supervised learning is
the key approximation
mechanism used by GANs
x

GAN: DCGAN Architecture



GAN: Problems with Perspective



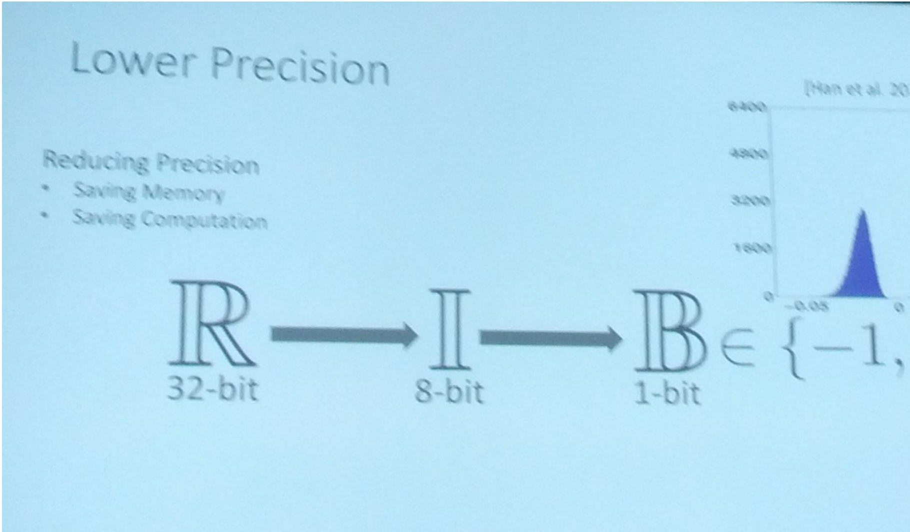
Workshop on Efficient Methods for DNNs

M. Rastegari & M. Courbariaux, http://allenai.org/plato/emdnn/

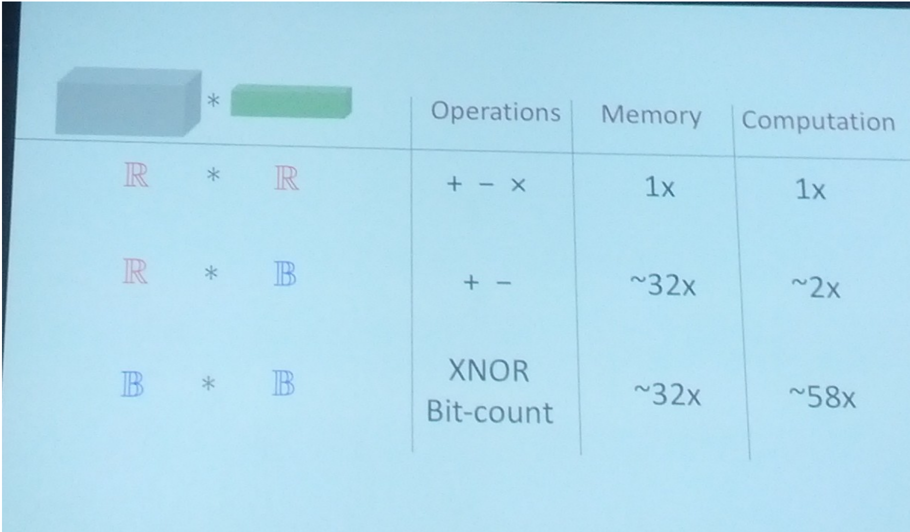
- nVidia
- Intel „Nervana“ (hardware+compiler+cloud)
- binarized networks, real-time on iPhone/Raspberry
- FPGAs and ASICs
- good poster session

- fun discussions
  - Intel: we have 10x more multipliers
  - nVidia: what you have is vaporware

Ali Farhadi: DL on Resource-Constrainted Platforms

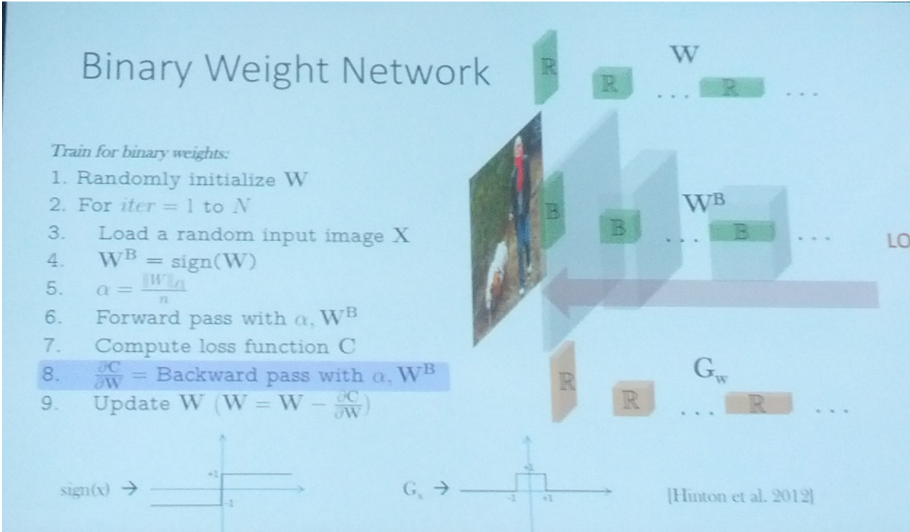


Float vs. Binary

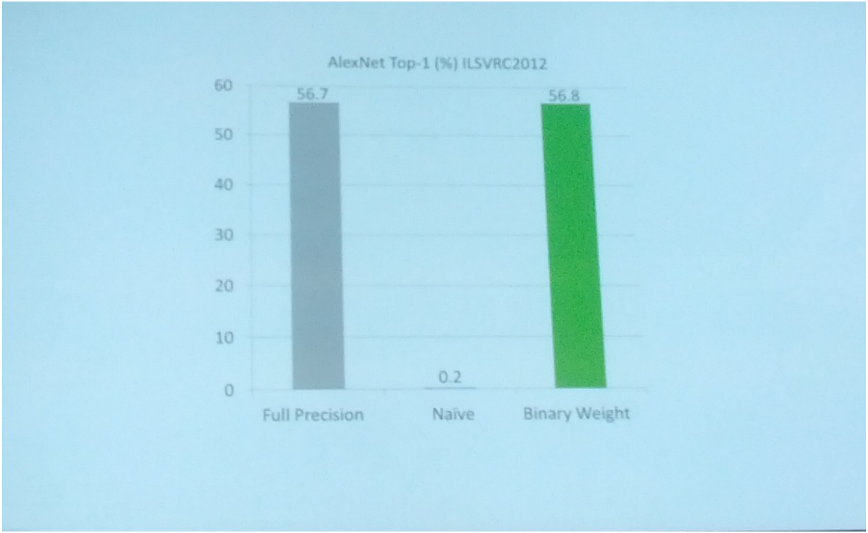


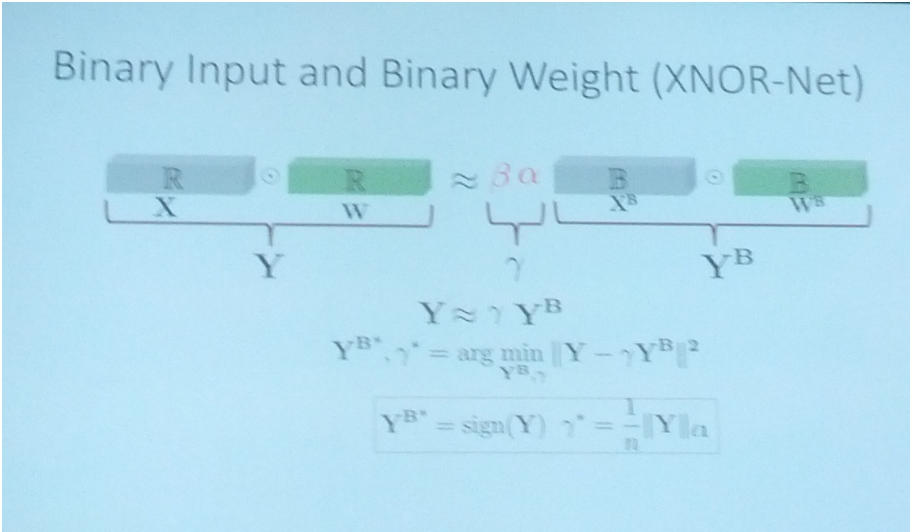
*	Operations	Memory	Computation
ℝ * ℝ	+ − ×	1x	1x
ℝ * 𝔹	+ −	~32x	~2x
𝔹 * 𝔹	XNOR Bit-count	~32x	~58x

Deep Binary Weight Network

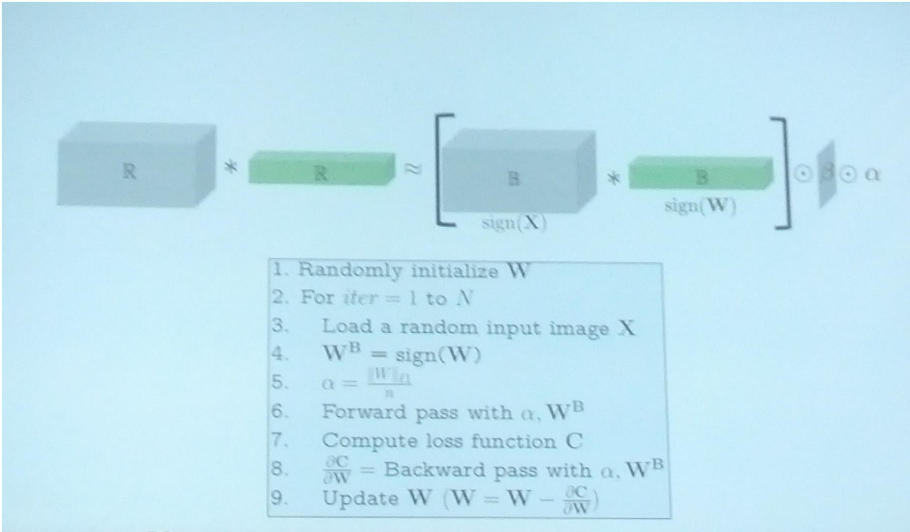


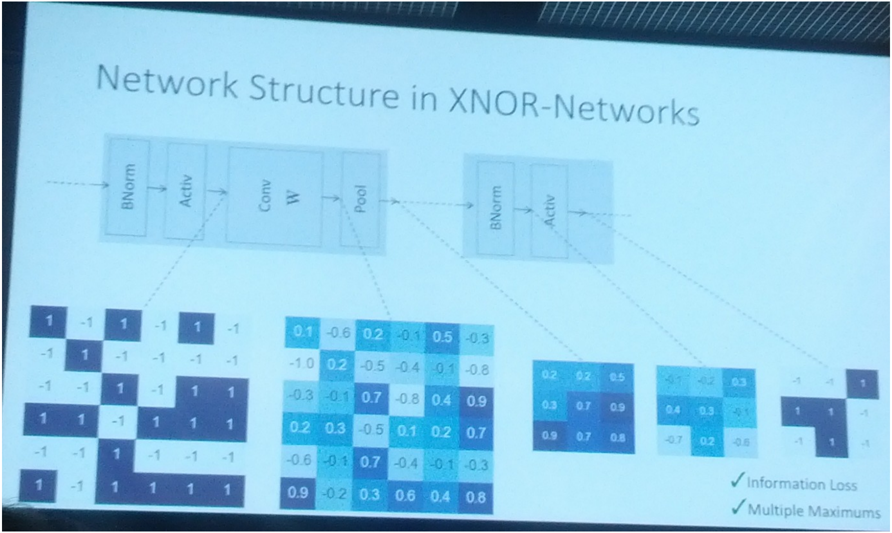
AlexNet vs. Binary AlexNet




Binary Input and Binary Weight (XNOR-Net)
$\mathbb{R}$ ⊙ $\mathbb{R}$ ≈ $\beta\alpha$ $\mathbb{B}$ ⊙ $\mathbb{B}$
X W
X^{B} W^{B}
Y
γ
Y^{B}
$Y \approx \gamma Y^{B}$
$Y^{B^*}, \gamma^* = \arg\min_{Y^{B},\gamma} \|Y - \gamma Y^{B}\|^2$
$Y^{B^*} = \mathrm{sign}(Y) \quad \gamma^* = \frac{1}{n}\|Y\|_{\ell 1}$

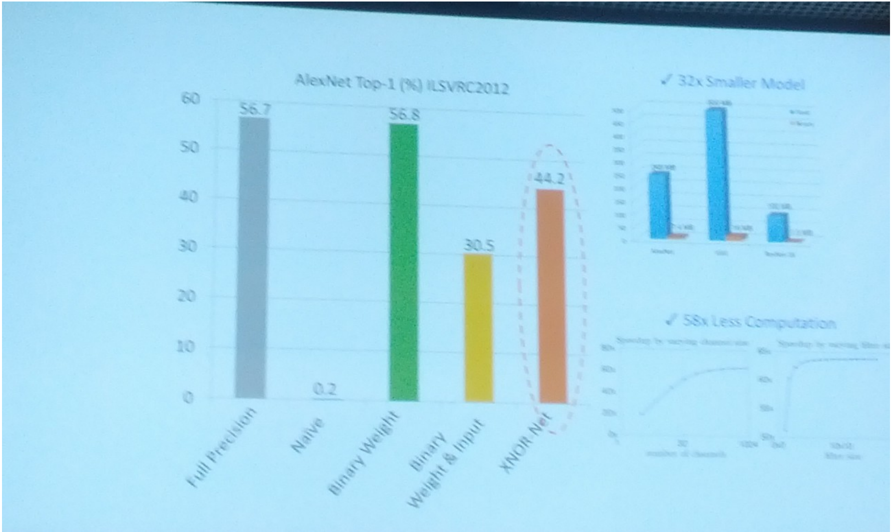
XNOR-Net


R
*
R
≈
B
sign(X)
*
B
sign(W)
⊙ β ⊙ α
1. Randomly initialize W
2. For $iter = 1$ to N
3. Load a random input image X
4. $\mathbf{W}^{\mathbf{B}} = \mathrm{sign}(\mathbf{W})$
5. $\alpha = \frac{\|\mathbf{W}\|_{\ell 1}}{n}$
6. Forward pass with α, $\mathbf{W}^{\mathbf{B}}$
7. Compute loss function C
8. $\frac{\partial C}{\partial \mathbf{W}}$ = Backward pass with α, $\mathbf{W}^{\mathbf{B}}$
9. Update W $(\mathbf{W} = \mathbf{W} - \frac{\partial C}{\partial \mathbf{W}})$



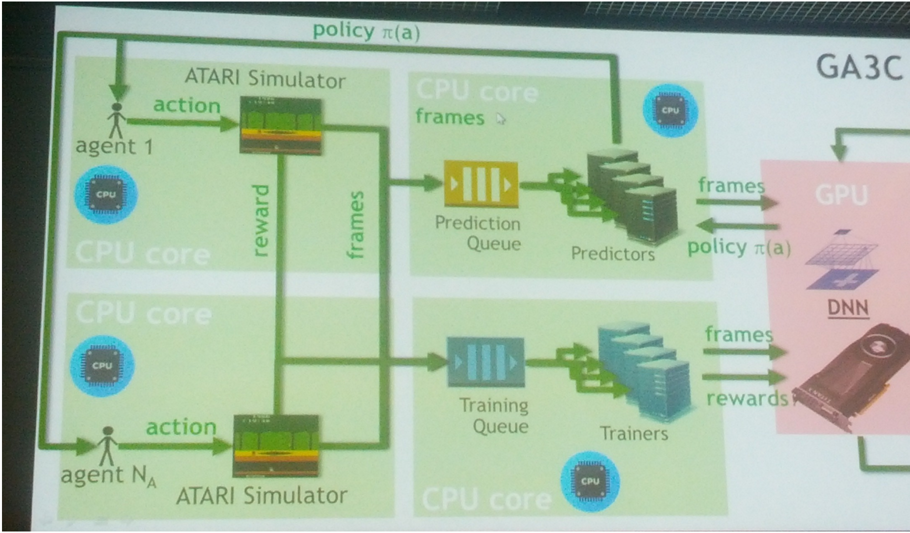
Network Structure in XNOR-Networks
BNorm
Activ
Conv
W
Pool
BNorm
Activ
✓Information Loss
✓Multiple Maximums

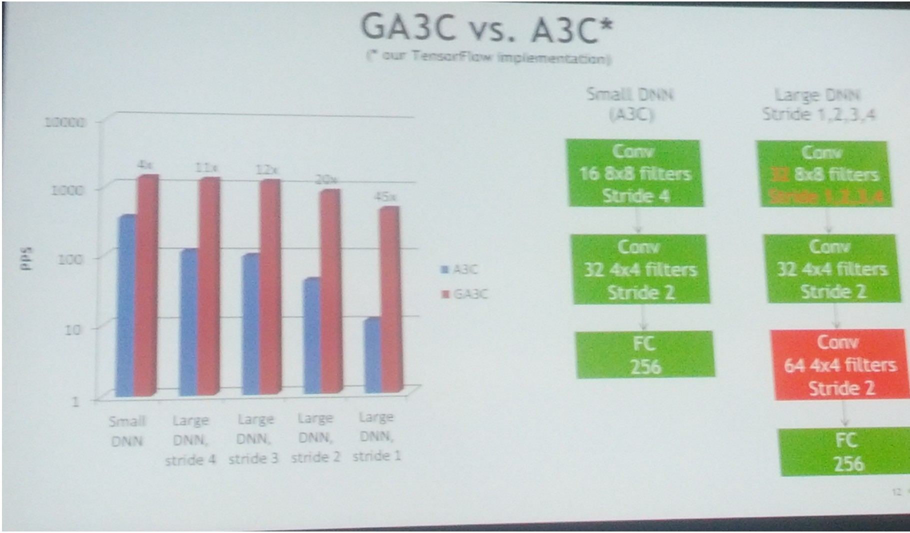
XNOR-Net Performance



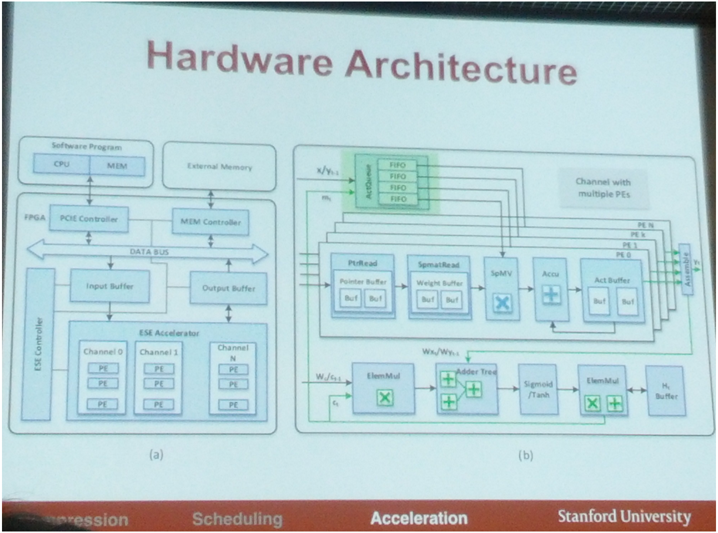
Live Demo: AlexNet on iPhone and Raspberry Zero

nVidia: A3C on GPU

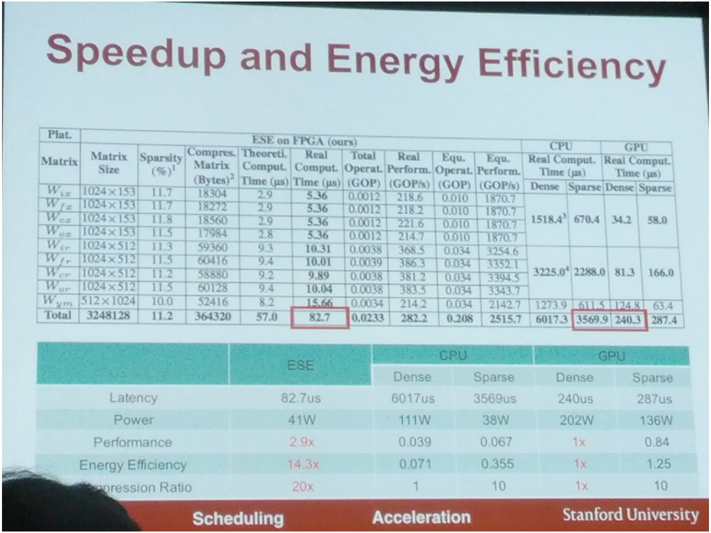



GA3C vs. A3C*
(* our TensorFlow implementation)
PPS
10000
1000
100
10
1
4x
11x
12x
20x
45x
Small DNN
Large DNN, stride 4
Large DNN, stride 3
Large DNN, stride 2
Large DNN, stride 1
A3C
GA3C
Small DNN (A3C)
Conv 16 8x8 filters Stride 4
Conv 32 4x4 filters Stride 2
FC 256
Large DNN Stride 1,2,3,4
Conv 32 8x8 filters Stride 1,2,3,4
Conv 32 4x4 filters Stride 2
Conv 64 4x4 filters Stride 2
FC 256

Song Han (?), Stanford: DNN FPGA Architecture



DNN FPGA Speedup and Efficiency



Speedup and Energy Efficiency

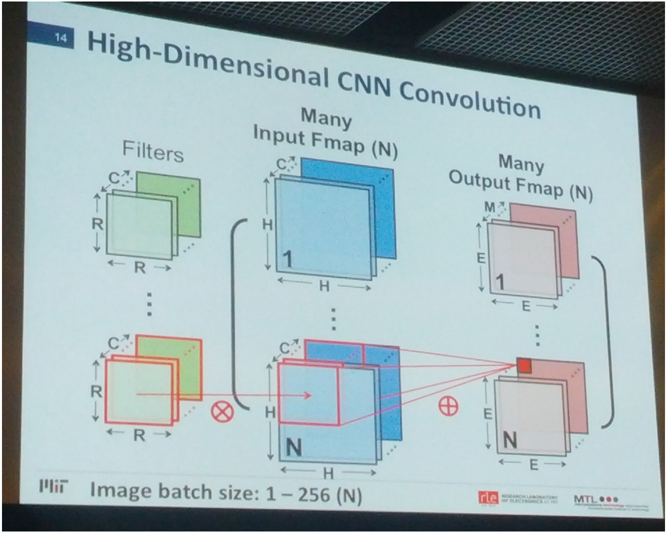
Plat.	ESE on FPGA (ours)									CPU		GPU	
Matrix	Matrix Size	Sparsity (%)[1]	Compres. Matrix (Bytes)[2]	Theoreti. Comput. Time (μs)	Real Comput. Time (μs)	Total Operat. (GOP)	Real Perform. (GOP/s)	Equ. Operat. (GOP)	Equ. Perform. (GOP/s)	Real Comput. Time (μs) Dense	Real Comput. Time (μs) Sparse	Real Comput. Time (μs) Dense	Real Comput. Time (μs) Sparse
W_{ix}	1024×153	11.7	18304	2.9	**5.36**	0.0012	218.6	0.010	1870.7	**1518.4**[3]	**670.4**	**34.2**	**58.0**
W_{fx}	1024×153	11.7	18272	2.9	**5.36**	0.0012	218.2	0.010	1870.7				
W_{cx}	1024×153	11.8	18560	2.9	**5.36**	0.0012	221.6	0.010	1870.7				
W_{ox}	1024×153	11.5	17984	2.8	**5.36**	0.0012	214.7	0.010	1870.7				
W_{ir}	1024×512	11.3	59360	9.3	**10.31**	0.0038	368.5	0.034	3254.6	**3225.0**[4]	**2288.0**	**81.3**	**166.0**
W_{fr}	1024×512	11.5	60416	9.4	**10.01**	0.0039	386.3	0.034	3352.1				
W_{cr}	1024×512	11.2	58880	9.2	**9.89**	0.0038	381.2	0.034	3394.5				
W_{or}	1024×512	11.5	60128	9.4	**10.04**	0.0038	383.5	0.034	3343.7				
W_{ym}	512×1024	10.0	52416	8.2	**15.66**	0.0034	214.2	0.034	2142.7	1273.9	611.5	124.8	63.4
Total	**3248128**	**11.2**	**364320**	**57.0**	**82.7**	**0.0233**	**282.2**	**0.208**	**2515.7**	**6017.3**	**3569.9**	**240.3**	**287.4**

	ESE	CPU		GPU	
		Dense	Sparse	Dense	Sparse
Latency	82.7us	6017us	3569us	240us	287us
Power	41W	111W	38W	202W	136W
Performance	2.9x	0.039	0.067	1x	0.84
Energy Efficiency	14.3x	0.071	0.355	1x	1.25
...pression Ratio	20x	1	10	1x	10

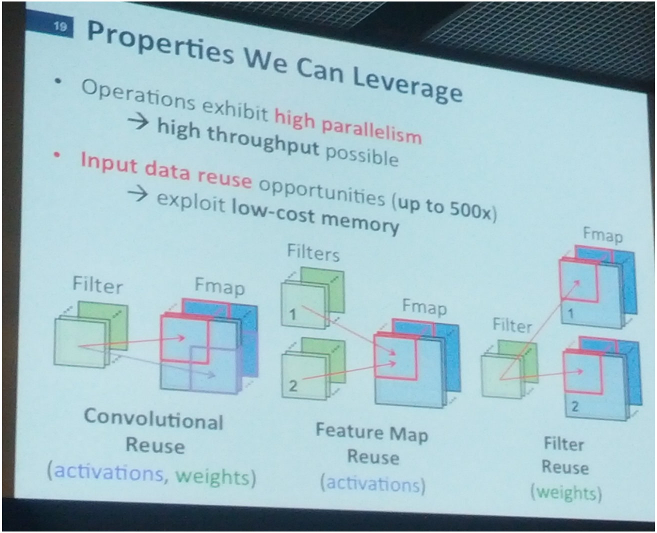
Scheduling Acceleration Stanford University

Vivienne Sze: EyeRiss Energy-Efficient DNNs

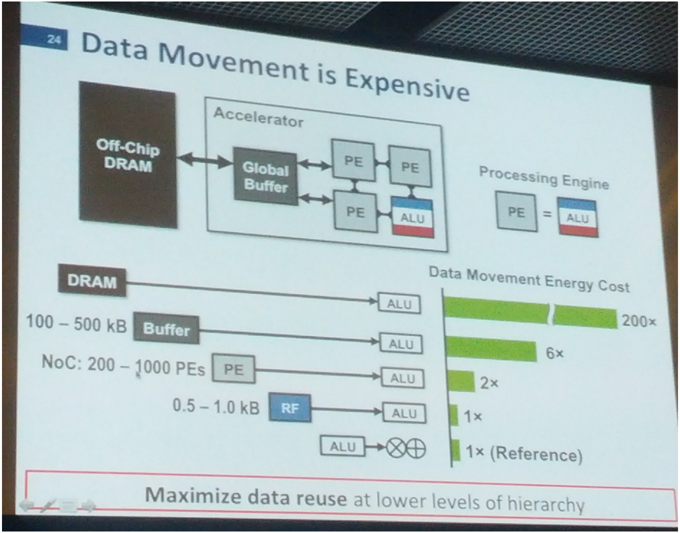
High-Dimensional CNN Convolution



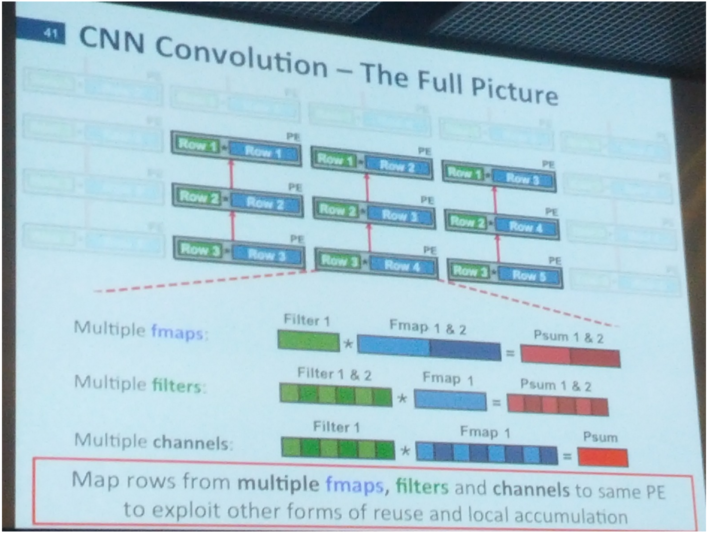
Properties we can leverage



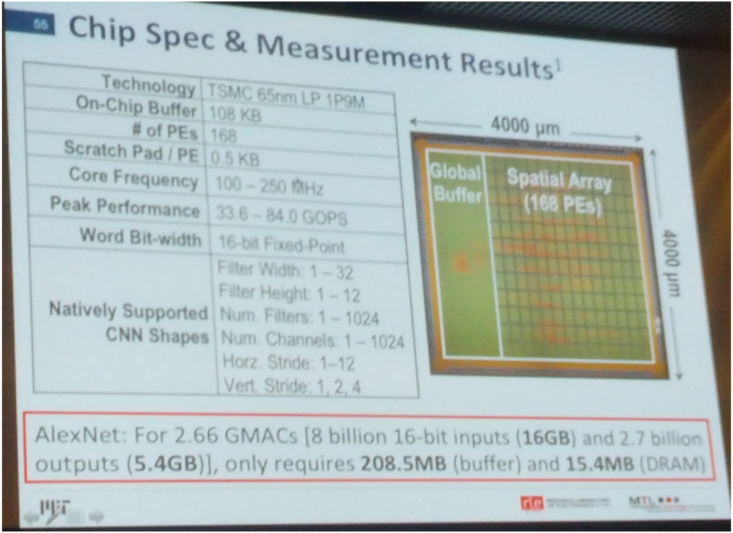
Data Movement is Expensive



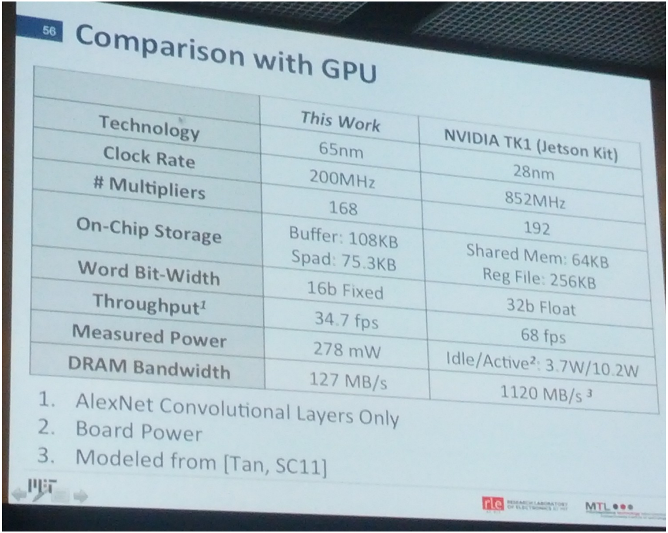
CNN Convolution Options



Chip Spec


Chip Spec & Measurement Results[1]
Technology
TSMC 65nm LP 1P9M
On-Chip Buffer
108 KB
# of PEs
168
Scratch Pad / PE
0.5 KB
Core Frequency
100 – 250 MHz
Peak Performance
33.6 – 84.0 GOPS
Word Bit-width
16-bit Fixed-Point
Natively Supported CNN Shapes
Filter Width: 1 – 32
Filter Height: 1 – 12
Num. Filters: 1 – 1024
Num. Channels: 1 – 1024
Horz. Stride: 1–12
Vert. Stride: 1, 2, 4
4000 µm
4000 µm
Global Buffer
Spatial Array (168 PEs)
AlexNet: For 2.66 GMACs [8 billion 16-bit inputs (16GB) and 2.7 billion outputs (5.4GB)], only requires 208.5MB (buffer) and 15.4MB (DRAM)

Comparison with (mobile) GPU



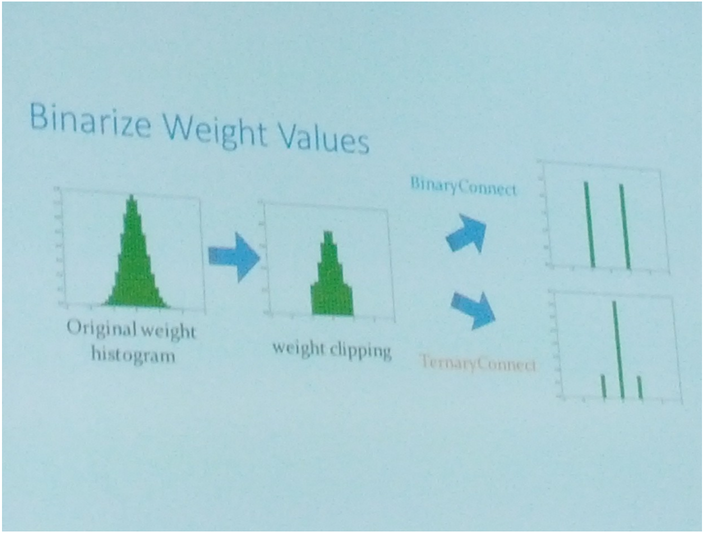
56

Comparison with GPU

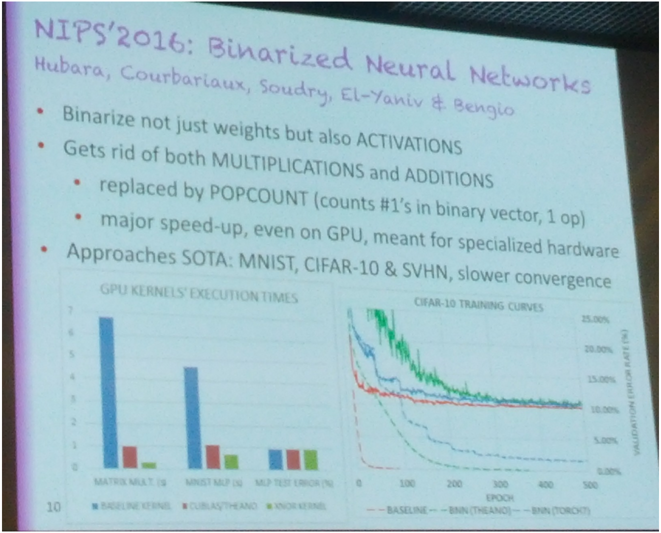
	This Work	NVIDIA TK1 (Jetson Kit)
Technology	65nm	28nm
Clock Rate	200MHz	852MHz
# Multipliers	168	192
On-Chip Storage	Buffer: 108KB Spad: 75.3KB	Shared Mem: 64KB Reg File: 256KB
Word Bit-Width	16b Fixed	32b Float
Throughput[1]	34.7 fps	68 fps
Measured Power	278 mW	Idle/Active[2]: 3.7W/10.2W
DRAM Bandwidth	127 MB/s	1120 MB/s [3]

1. AlexNet Convolutional Layers Only
2. Board Power
3. Modeled from [Tan, SC11]

Yoshua Bengio: Low Precision Neural Nets



Binarized Neural Nets

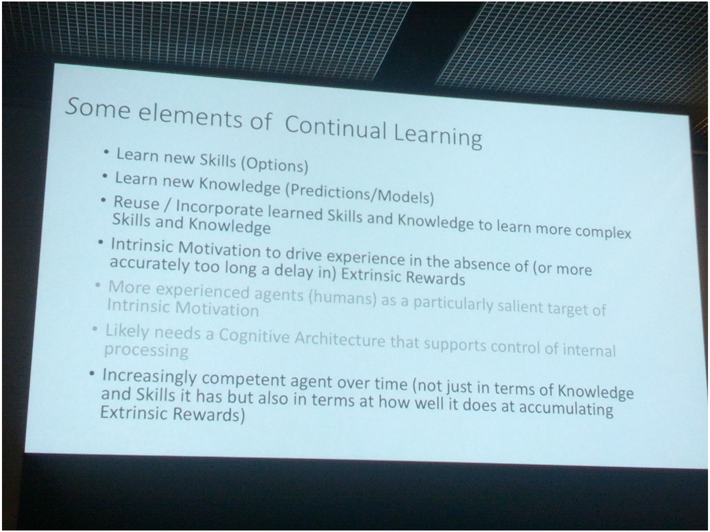


Continual Learning Workshop

Mark Ring et.al., https://sites.google.com/site/cldlnips2016/

Time	*Title*	*Speaker*
8:30 AM	Welcoming remarks and introduction	Mark Ring (Cogitai)
8:50 AM	Invited talk	Rich Sutton (U of Alberta)
9:20 AM	Spotlight 1	B. Lüders, M. Schläger and S. Risi
9:30 AM	Spotlight 2	T. Miconi
9:40 AM	Invited talk	Claudia Clopath (Imperial College)
10:10 AM	*Coffee break (+ poster set-up)*	
10:40 AM	Invited talk	Satinder Singh Bajeva (U of Michigan/Cogitai)
11:10 AM	Spotlight 3	E. de Jong
11:20 AM	Poster session	(see submission page for full list)
12:20 PM	*Lunch*	
2:00 PM	Invited talk	Honglak Lee (U of Michigan)
2:30 PM	Spotlight 4	M. Pickett, R. Al-Rfou, L. Shao and C. Tar
2:40 PM	Spotlight 5	H. Kasaei
2:50 PM	Invited talk	Eric Eaton (U of Pennsylvania)
3:20 PM	Systems that are alert to context and surprise	Hava Siegelmann (DARPA)
3:30 PM	*Coffee break*	
4:00 PM	Invited talk	Raia Hasdell (DeepMind)
4:30 PM	Invited talk	Doina Precup (McGill University)
5:00 PM	Panel discussion	(all invited speakers)
6:00 PM	*Workshop ends*	

Satinder Singh: Simple Continual Learning Demonstration



Neurobotics Workshop

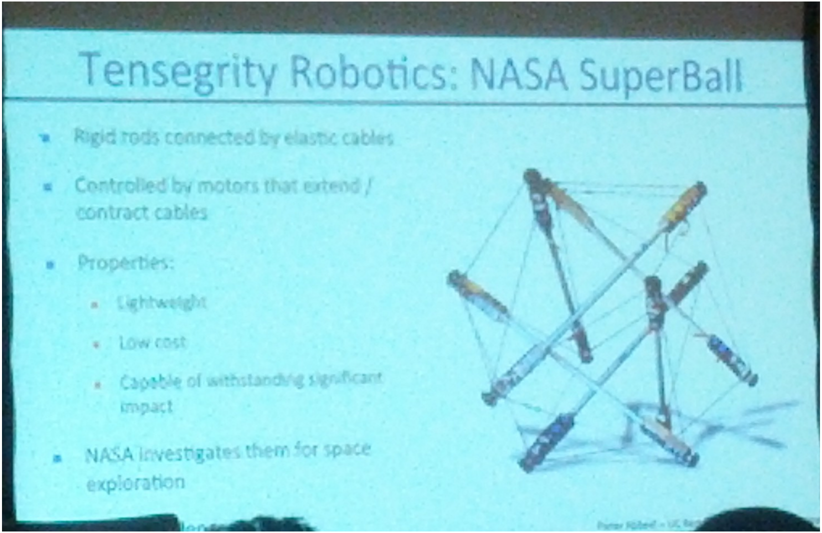
https://rueckert.lima-city.de/NIPSWS2016/WebContent

Day One, Neurorobotics WS, Fri Dec 9th 2016		
	14.20-14.30	Introduction by **Elmar Rueckert** and **Martin Riedmiller**
Session One: Reinforcement Learning, Imitation, and Active Learning		
1	14.30-15.00	**Juergen Schmidhuber** (Scientific Director of the Swiss AI Lab IDSIA
	15.00-15.30	Posters and Coffee
2	15.30-16.00	**Sergey Levine** (University of California, Berkeley)
3	16.00-16.30	**Pieter Abbeel** (University of California, Berkeley)
4	16.30-17.00	**Johanni Brea** (École polytechnique fédérale de Lausanne, EPFL)
	17.00-17.20	Posters and Coffee
5	17.20-17.45	**Paul Schrater** (University of Minnesota)
6	17.45-18.10	**Frank Hutter** (University Freiburg)
7	18.10-18.35	**Raia Hadsell** (Google DeepMind)
	18.35-19.00	Panel Discussion, Session One

day two

Day Two, Neurorobotics WS, Sat Dec 10th 2016		
Session One: Reinforcement Learning, Imitation, and Active Learning		
	08.30-08.35	Introduction by **Elmar Rueckert** and **Martin Riedmiller**
8	08.35-09.05	**Robert Legenstein** (Graz University of Technology)
9	09.05-09.35	**Sylvain Calinon**(Idiap Research Institute, EPFL Lausanne)
10	09.35-10.05	**Chelsea Finn** (University of California, Berkeley)
11	10.05-10.35	**Peter Stone** (University of Texas at Austin)
	10.35-11.00	Posters and Coffee
12	11.00-11.30	**Paul Verschure** (Catalan Institute of Advanced Research)
Session Two: Model Representations and Features		
13	11.30-12.00	**Tobi Delbrück** (University of Zurich and ETH Zurich)
14	12.00-12.30	**Moritz Grosse-Wentrup** (Max Planck Institute Tuebingen)
15	12.30-13.00	**Kristian Kersting** (Technische Universität Dortmund)
	13.00-14.00	Lunch break
Session Three: Feedback and Control		
16	14.00-14.30	**Emo Todorov** (University of Washington)
17	14.30-15.00	**Richard Sutton** (University of Alberta)
	15.00-15.30	Posters and Coffee
18	15.30-16.00	**Bert Kappen** (Radboud University)
19	16.00-16.30	**Jean-Pascal Pfister** (University of Zurich and ETH Zurich)
	16.30-17.00	Posters and Coffee
20	17.00-17.30	**Jan Babic** (Josef Stefan Institute Ljubijana)
21	17.30-18.00	**Martin Giese** (University Clinic Tübingen)
	18.00-18.30	Panel Discussion, Session Two and Session Three

Tensegrity Robotics: NASA SuperBall



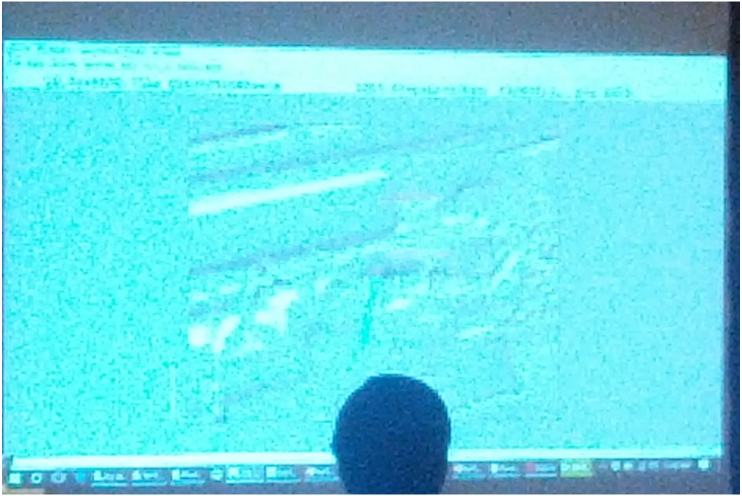
Tobi Delbrück, Zürich: DVS Event Camera

- custom event-based image sensor, CMOS prototype
- per pixel brightness ($\log I$) change detection
- only active pixels are read out (image gradients)

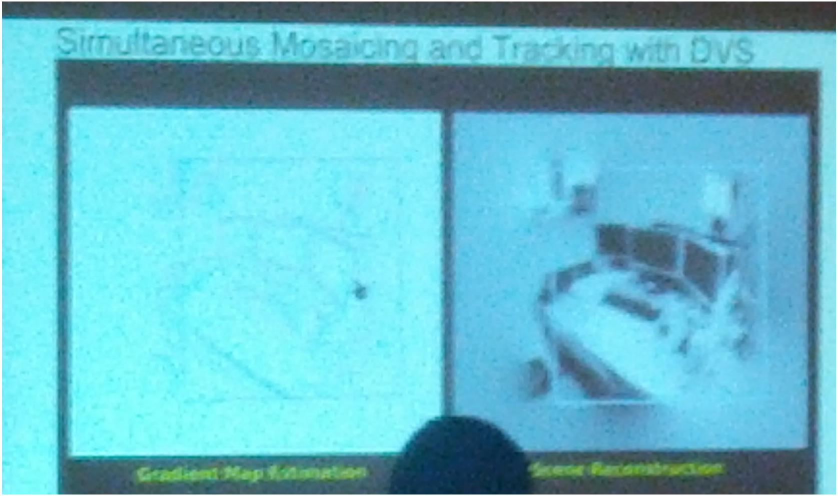
- static scenes: no image readout at all
- requires custom software to reconstruct full image
- fast tracking of moving objects
- ego-motion scene-reconstruction
- rock-paper-scissors live demo

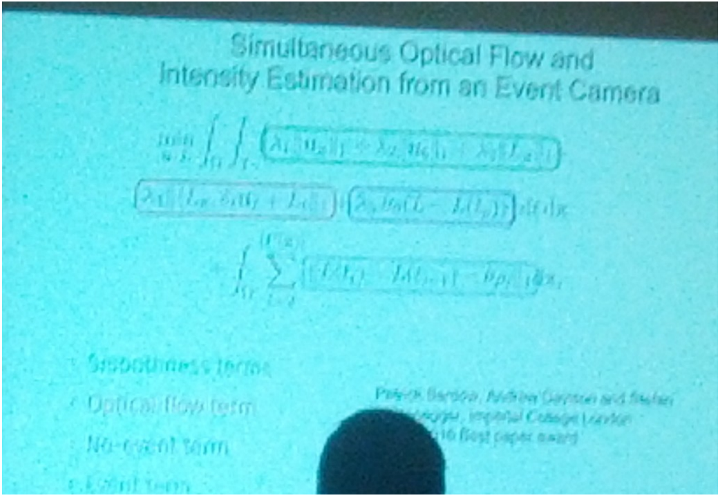
- insight: spiking neurons a means to save energy

DVS Camera: Live Demo

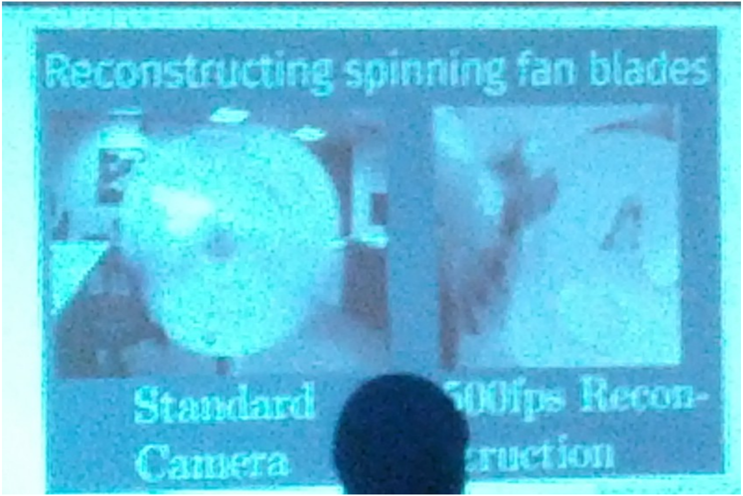


DVS Camera: Tracking

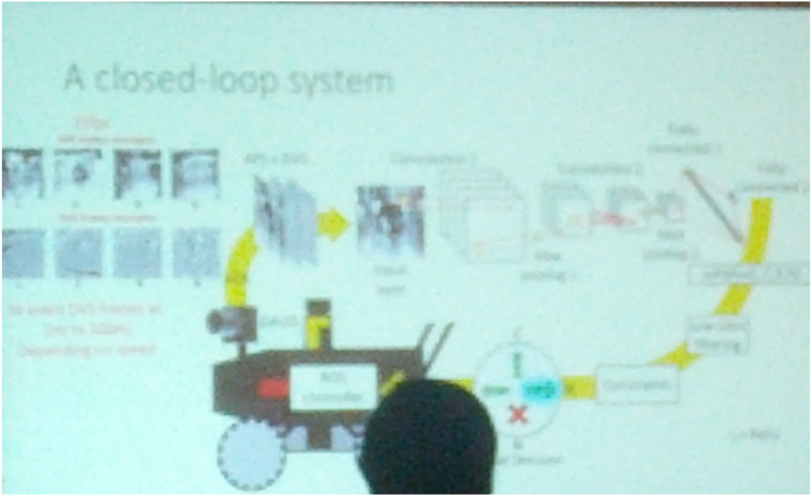



Simultaneous Optical Flow and
Intensity Estimation from an Event Camera

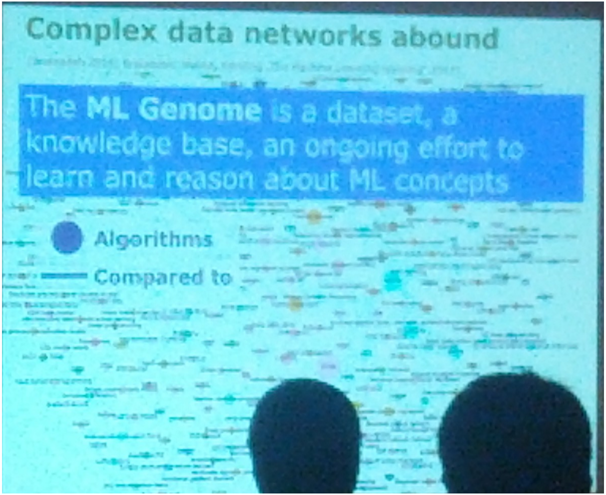
DVS Camera: 500fps Spinning Fan Blades



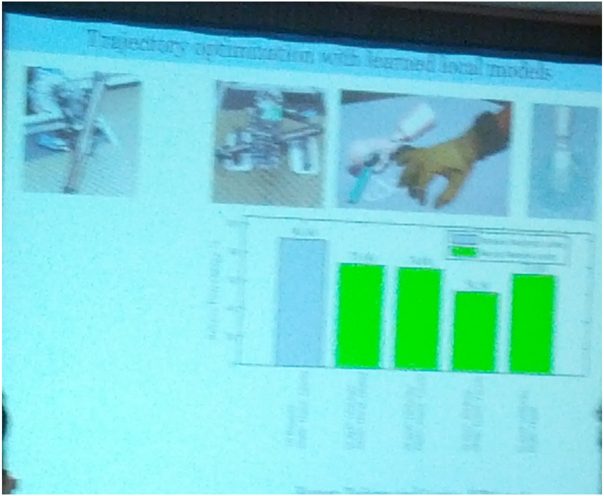
DVS Camera: Autonomous Vehicle with DNN



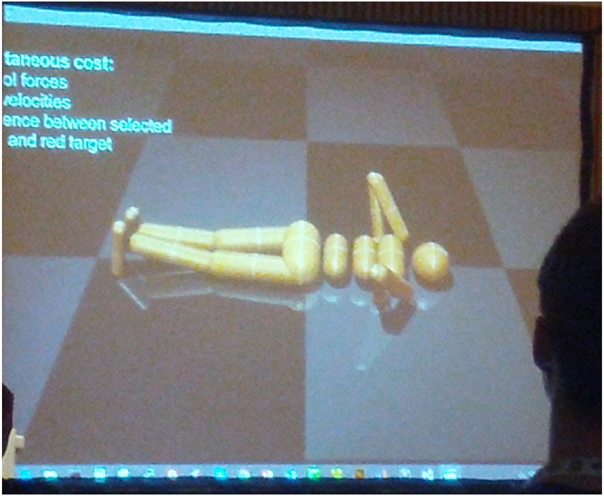
Kristian Kersting: ML Genome: Learning about Learning Algorithms



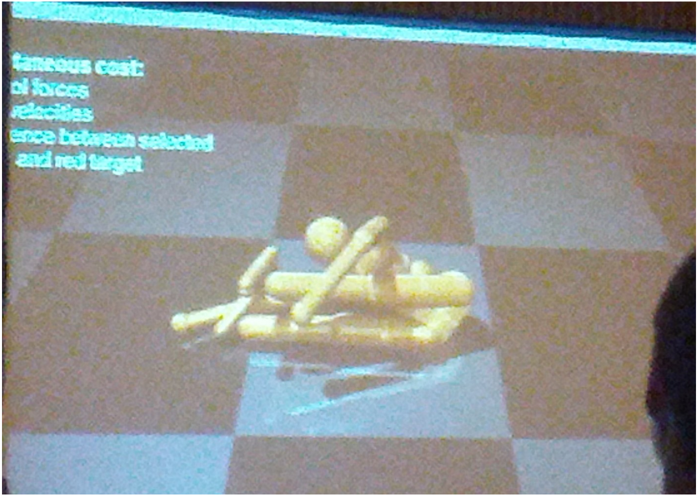
Emo Todorov: Optimal Control



Goal Directed Dynamics



Keeping your head up while falling



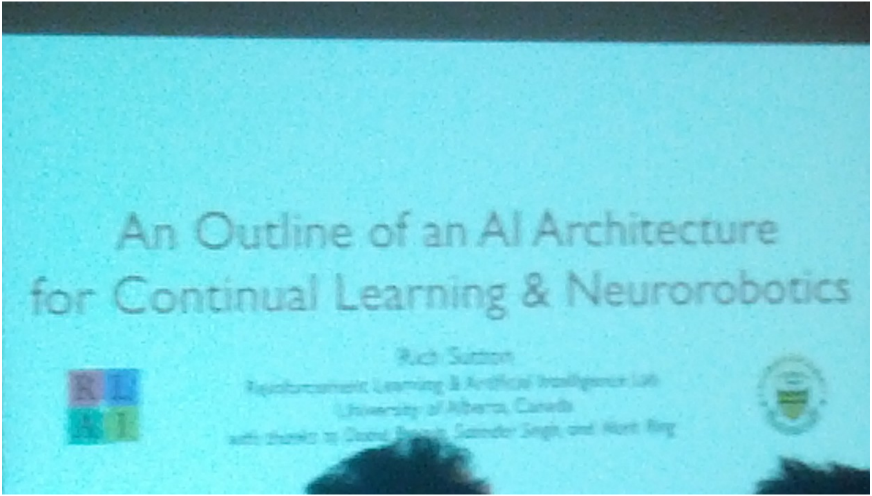
Q&A

(disclaimer: from memory, not accurate transsscript...)

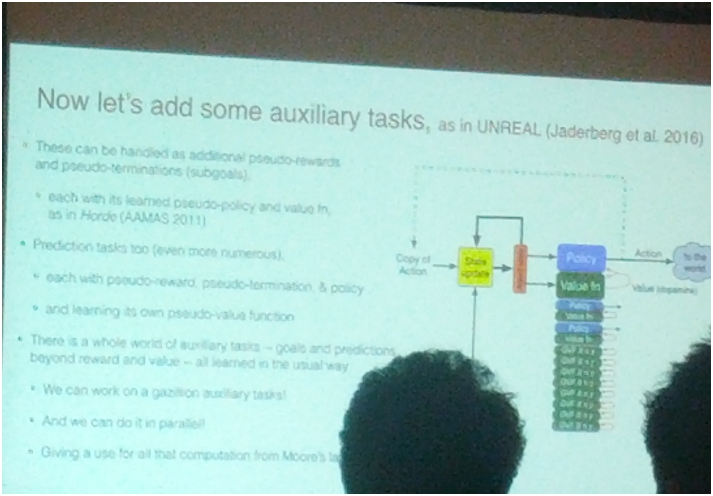
Q: you are using dynamics, are you also going to try DNNs?

A: Well, we know that NNs are universal function approximators.
So, there exists this network where I write any title and it outputs a complete paper that is then accepted at Nature.
Am I going to chase this network? No!

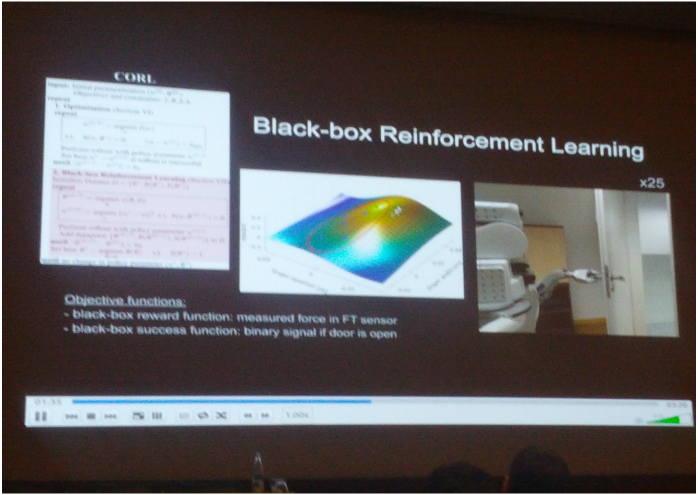
Rich Sutton: General AI Architecture



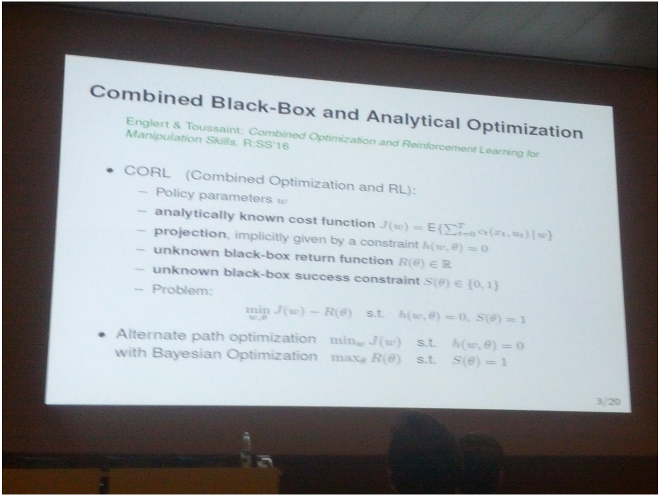
RL with Auxiliary Tasks



Bayesian Workshop: Toussaint: Learning Manipulation



CORL


Combined Black-Box and Analytical Optimization
Englert & Toussaint: Combined Optimization and Reinforcement Learning for Manipulation Skills, R:SS'16
• CORL (Combined Optimization and RL):
– Policy parameters w
– analytically known cost function $J(w) = \mathrm{E}\{\sum_{t=0}^{T} c_t(x_t, u_t) \mid w\}$
– projection, implicitly given by a constraint $h(w, \theta) = 0$
– unknown black-box return function $R(\theta) \in \mathbb{R}$
– unknown black-box success constraint $S(\theta) \in \{0, 1\}$
– Problem:
$\min_{w,\theta} J(w) - R(\theta)$ s.t. $h(w, \theta) = 0$, $S(\theta) = 1$
• Alternate path optimization $\min_w J(w)$ s.t. $h(w, \theta) = 0$
with Bayesian Optimization $\max_\theta R(\theta)$ s.t. $S(\theta) = 1$
3/20

Bayesian Extimation of Success Region