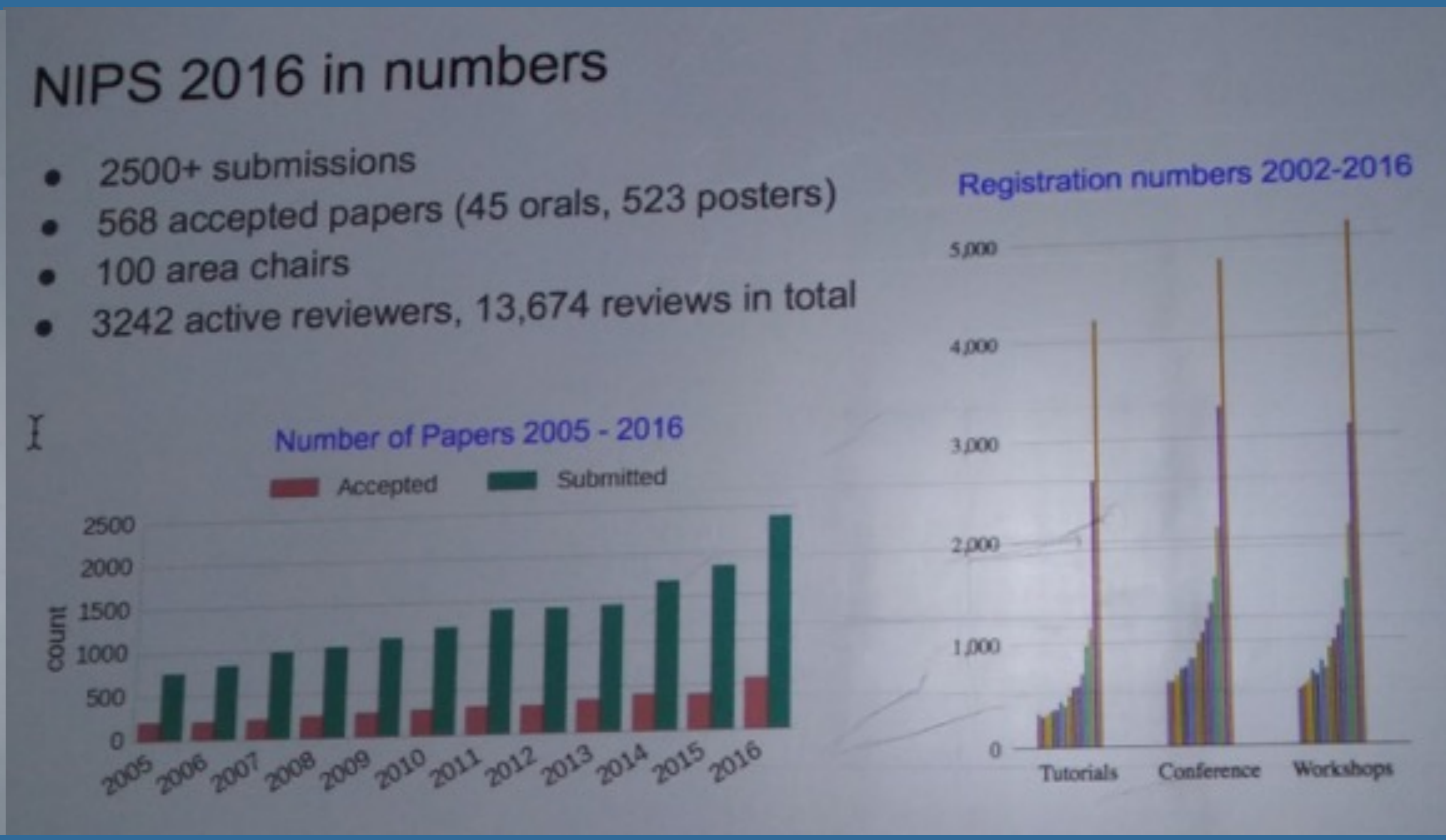
NIPS 2016 Highlights

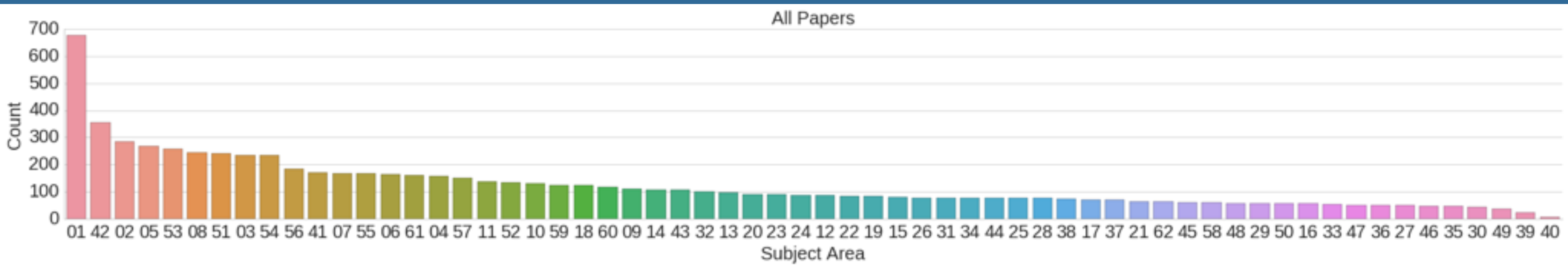
Sebastian Ruder
PhD Candidate, Insight Centre
Research Scientist, AYLIEN

Agenda

1. NIPS overview
2. Generative Adversarial Networks
3. Building applications with Deep Learning
4. RNNs
5. Improving classic algorithms
6. Reinforcement Learning
7. Learning-to-learn / Meta-learning
8. General AI
9. NLP
10. Miscellaneous

NIPS Overview





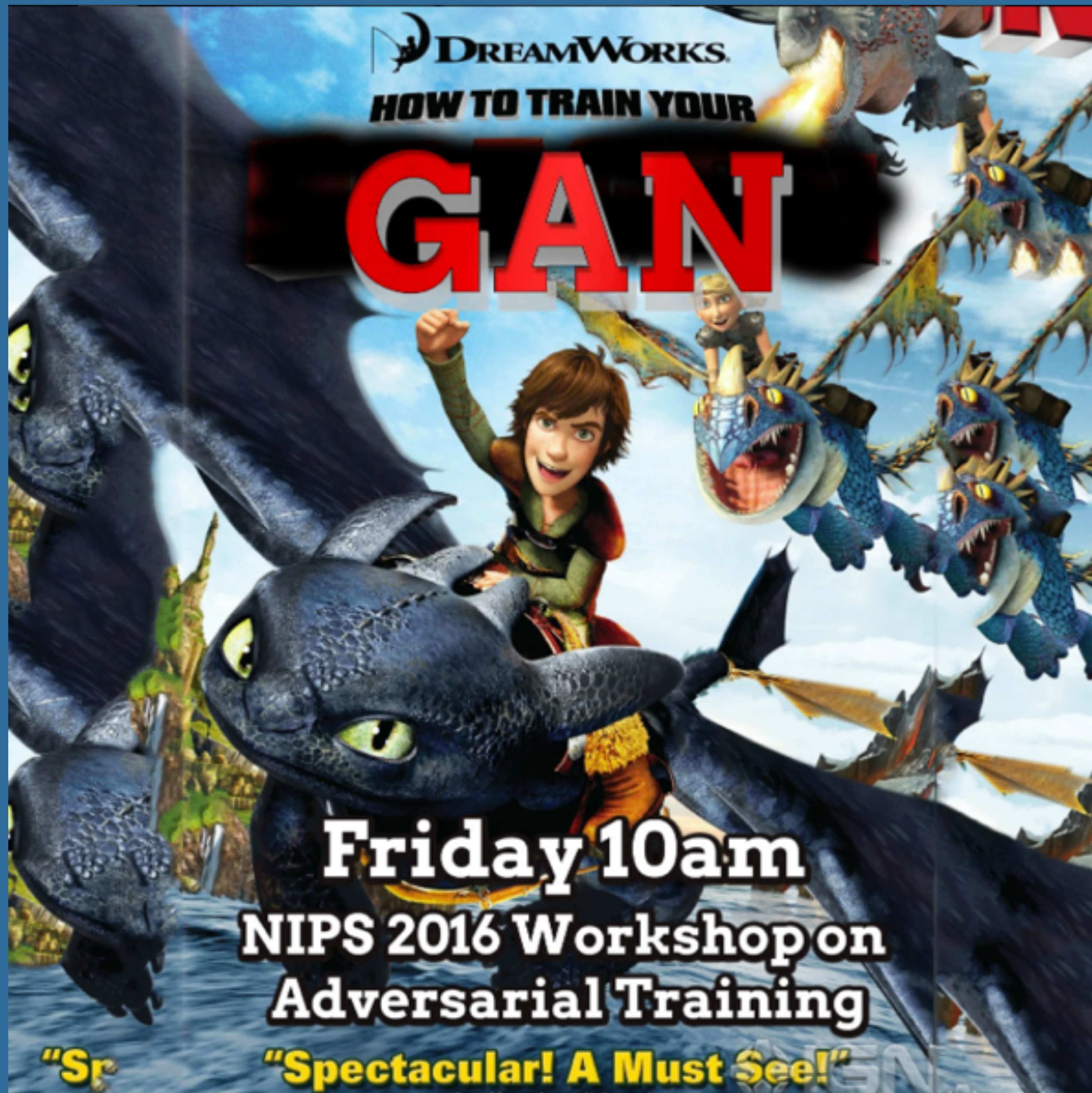
Source: http://www.tml.cs.uni-tuebingen.de/team/luxburg/misc/nips2016/index.php

Generative Adversarial Networks

- Ian Goodfellow's GAN tutorial[1]
- Soumith Chintala's "How to train your GAN" talk[2]

[1] http://www.iangoodfellow.com/slides/2016-12-04-NIPS.pdf
[2] https://github.com/soumith/ganhacks

Generative Adversarial Networks

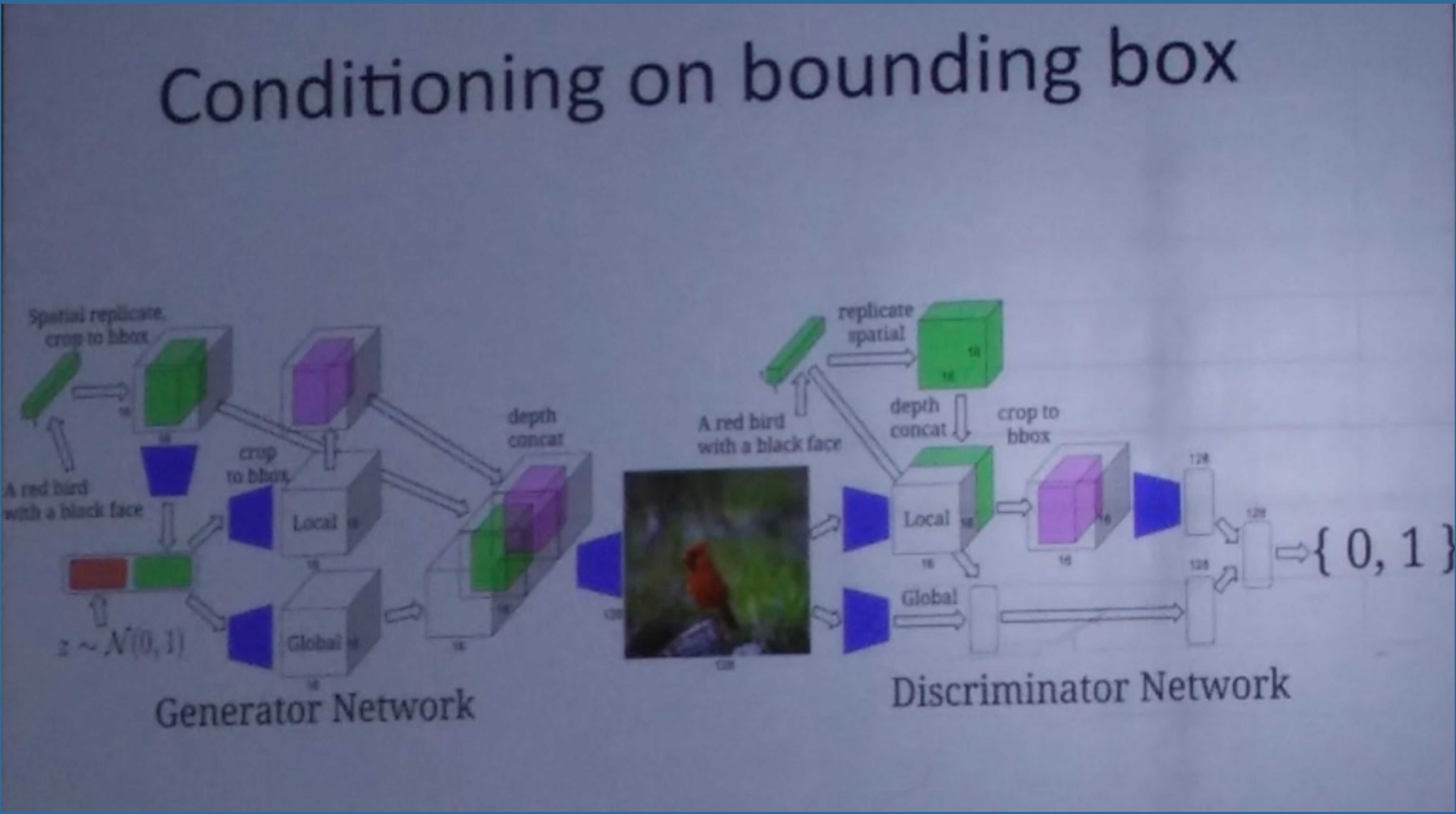


Generative Adversarial Networks

- Ian Goodfellow's GAN tutorial[1]
- Soumith Chintala's "How to train your GAN" talk[2]
- Yann LeCun is very bullish on GANs; lots of work on GANs from FAIR, etc.[3]
- Cool extensions, e.g. GA what-were network[4]

[1] http://www.iangoodfellow.com/slides/2016-12-04-NIPS.pdf
[2] https://github.com/soumith/ganhacks
[3] https://drive.google.com/file/d/0BxKBnD5y2M8NREZod0tVdW5FLTQ/view
[4] https://papers.nips.cc/paper/6111-learning-what-and-where-to-draw.pdf

Generative Adversarial Networks

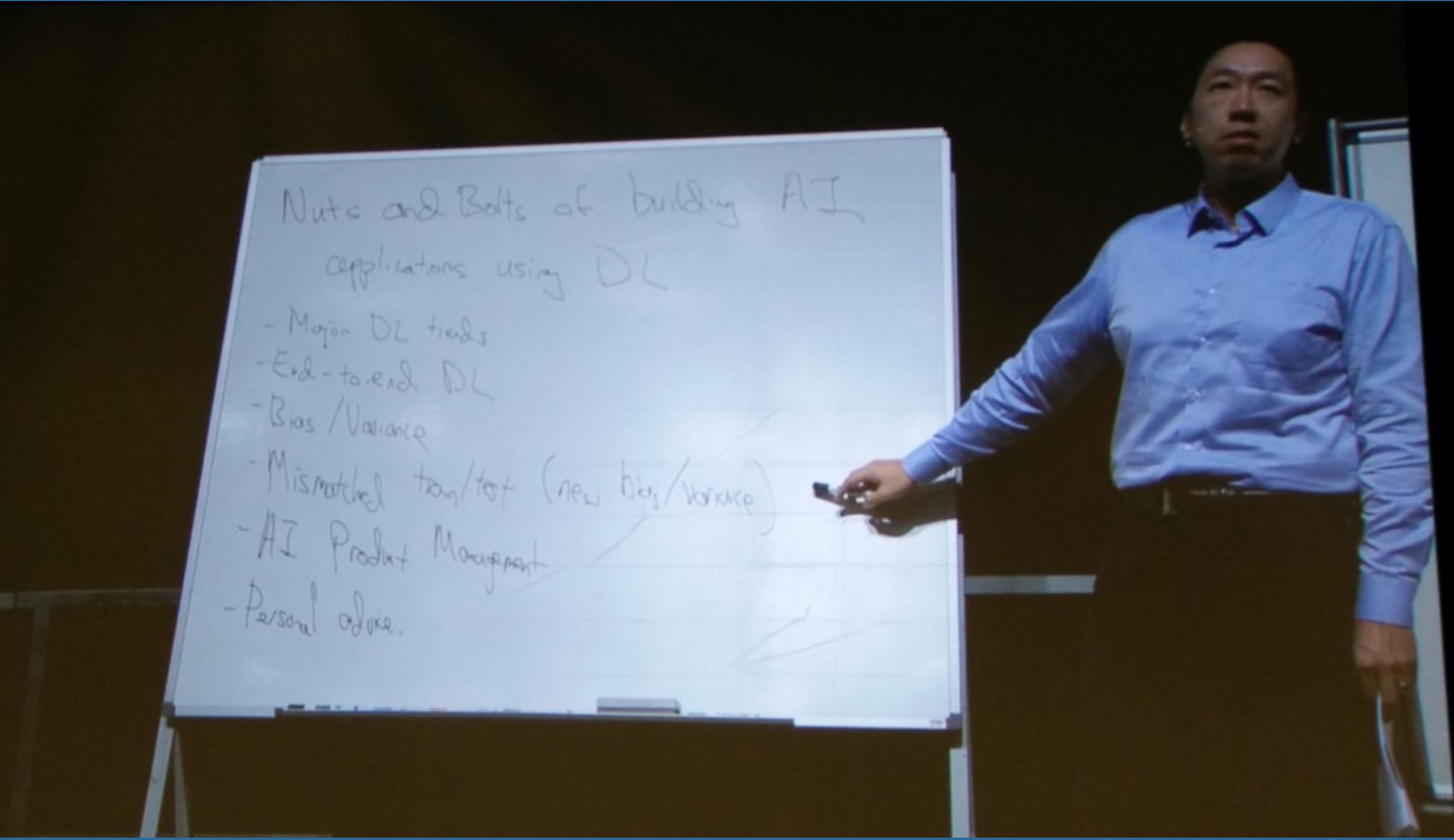


Generative Adversarial Networks

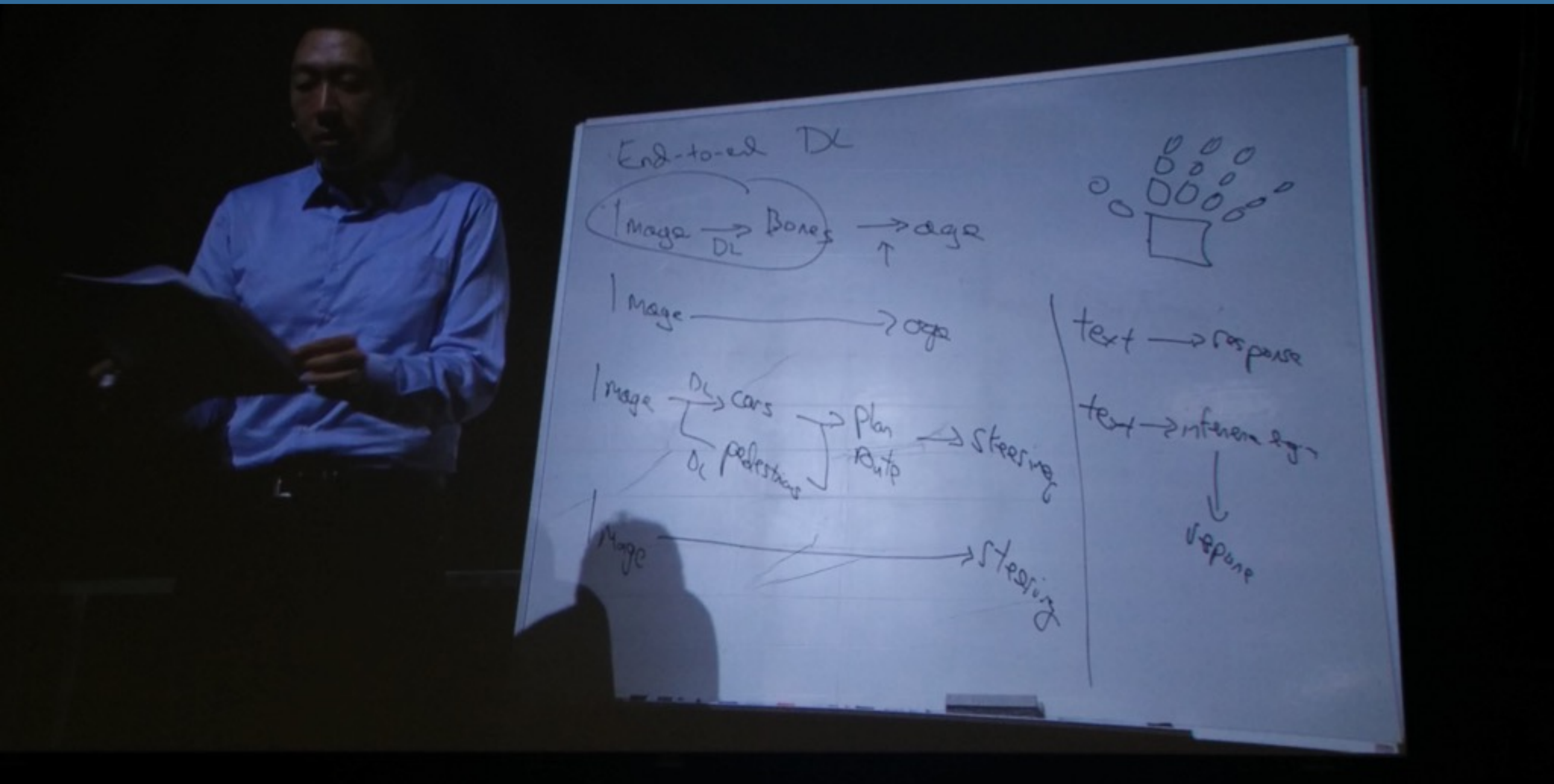
- Ian Goodfellow's GAN tutorial[1]
- Soumith Chintala's "How to train your GAN" talk[2]
- Yann LeCun is very bullish on GANs; lots of work on GANs from FAIR, etc.[3]
- Cool extensions, e.g. GA what-were network[4]
- Mostly used in CV for image generation; discriminator can be used as feature extractor
- Less work in NLP and other areas; promising directions at GAN workshop (1, 2, 3, 4)

[1]http://www.iangoodfellow.com/slides/2016-12-04-NIPS.pdf
[2]https://github.com/soumith/ganhacks
[3]https://drive.google.com/file/d/0BxKBnD5y2M8NREZod0tVdW5FLTQ/view
[4]https://papers.nips.cc/paper/6111-learning-what-and-where-to-draw.pdf
[5]https://arxiv.org/abs/1611.01144

Building applications with DL



Building applications with DL



RNNs

- 20 year anniversary of LSTM…

RNNs

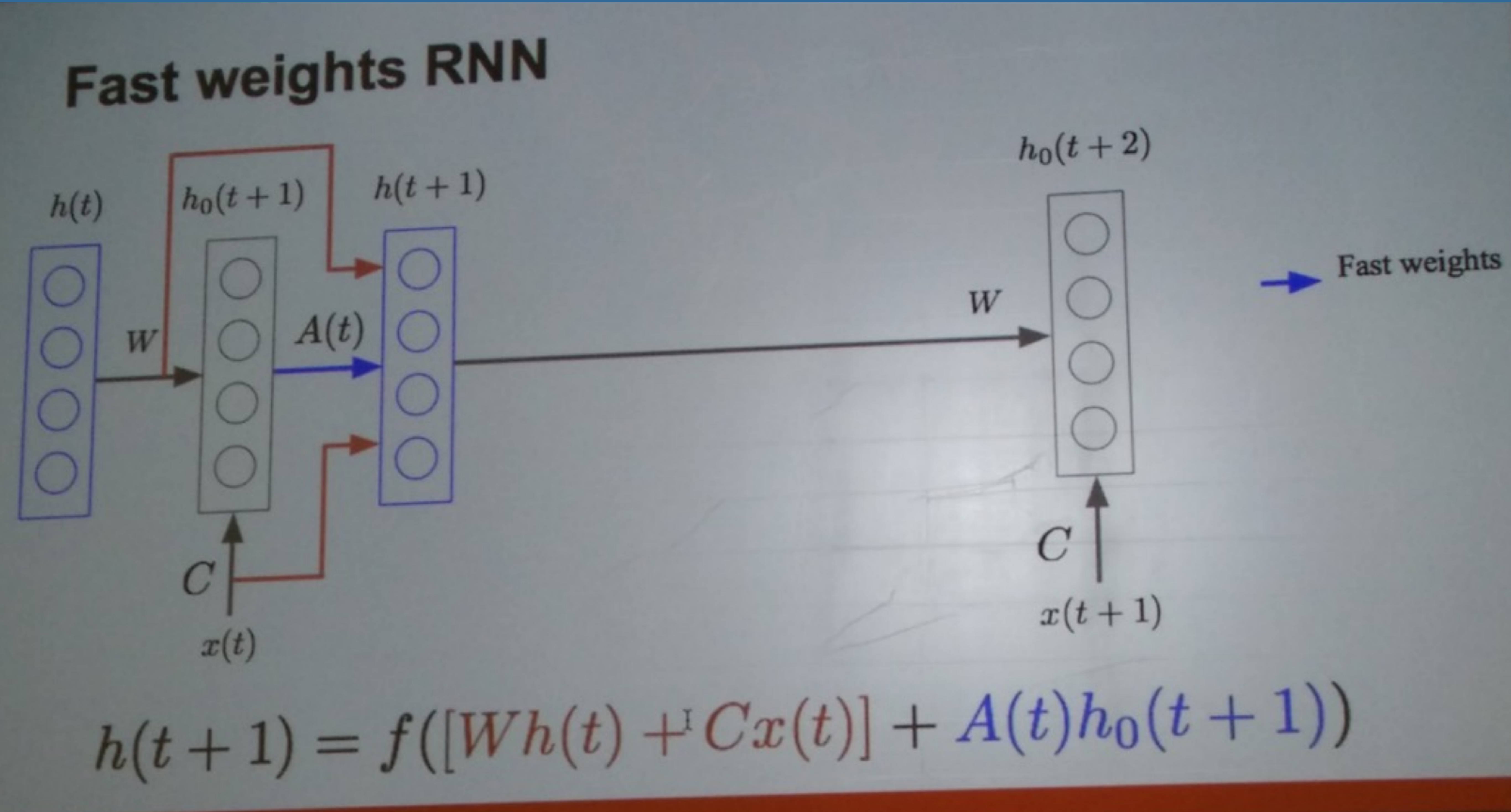
- 20 year anniversary of LSTM… being rejected from NIPS 1996 — perseverance pays off

RNNs

- RNN symposium
- Many papers on improving RNNs:
  - Handling different time scales[1,2]

[1] https://papers.nips.cc/paper/6057-using-fast-weights-to-attend-to-the-recent-past.pdf
[2] https://papers.nips.cc/paper/6310-phased-lstm-accelerating-recurrent-network-training-for-long-or-event-based-sequences.pdf

RNNs

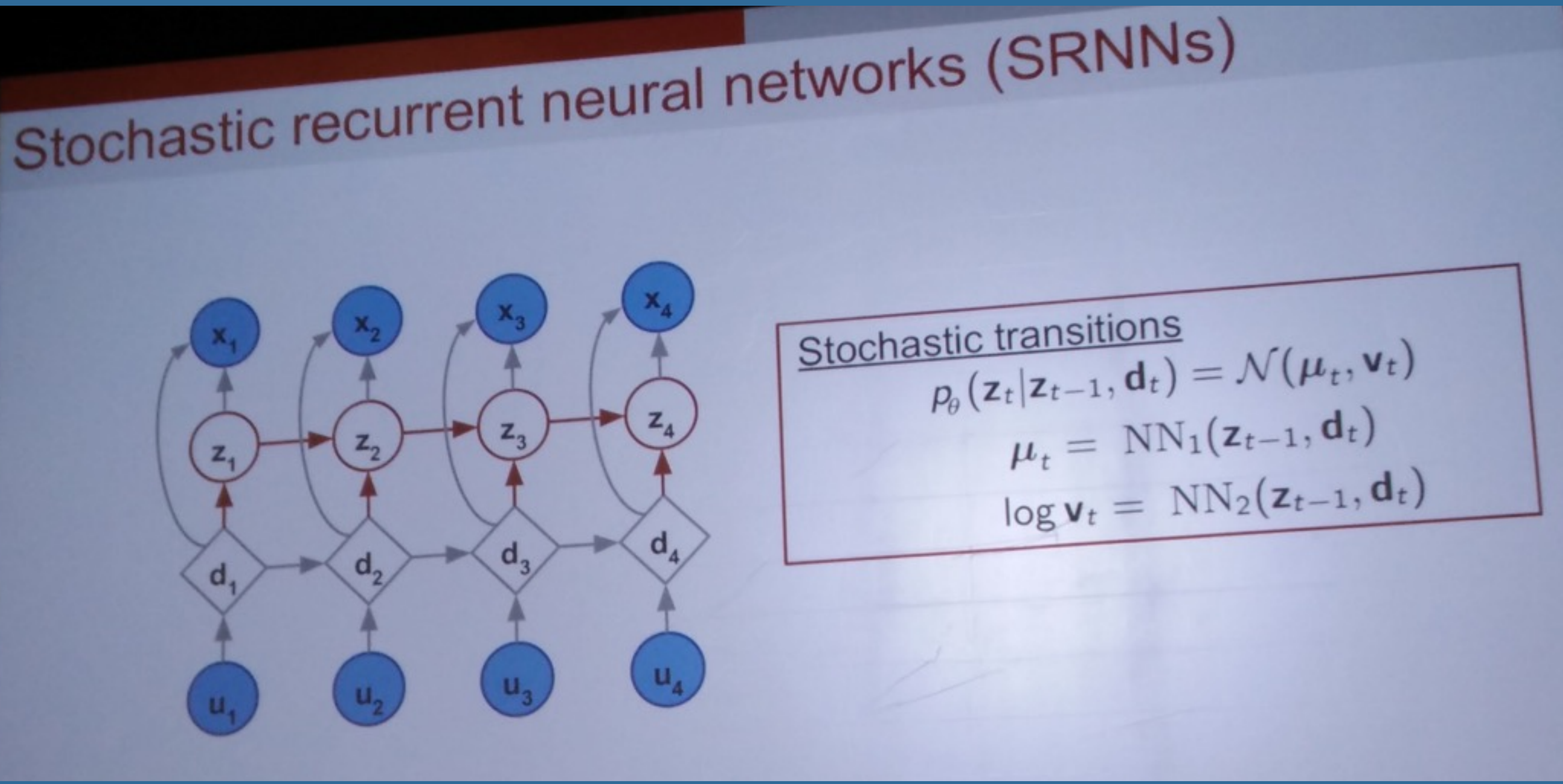

Fast weights RNN
$h(t)$
$h_0(t+1)$
$h(t+1)$
$h_0(t+2)$
W
$A(t)$
W
C
$x(t)$
C
$x(t+1)$
Fast weights
$h(t+1) = f([Wh(t) + Cx(t)] + A(t)h_0(t+1))$

RNNs

- RNN symposium
- Many papers on improving RNNs:
  - Handling different time scales[1,2]
  - Modelling uncertainty[3]

[1]https://papers.nips.cc/paper/6057-using-fast-weights-to-attend-to-the-recent-past.pdf
[2]https://papers.nips.cc/paper/6310-phased-lstm-accelerating-recurrent-network-training-for-long-or-event-based-sequences.pdf
[3]https://papers.nips.cc/paper/6039-sequential-neural-models-with-stochastic-layers.pdf

RNNs


Stochastic recurrent neural networks (SRNNs)
x1 x2 x3 x4
z1 z2 z3 z4
d1 d2 d3 d4
u1 u2 u3 u4
Stochastic transitions
$p_\theta(\mathbf{z}_t | \mathbf{z}_{t-1}, \mathbf{d}_t) = \mathcal{N}(\boldsymbol{\mu}_t, \mathbf{v}_t)$
$\boldsymbol{\mu}_t = \mathrm{NN}_1(\mathbf{z}_{t-1}, \mathbf{d}_t)$
$\log \mathbf{v}_t = \mathrm{NN}_2(\mathbf{z}_{t-1}, \mathbf{d}_t)$

RNNs

- RNN symposium
- Many papers on improving RNNs:
  - Handling different time scales[1,2]
  - Modelling uncertainty[3]
- General DL improvements
  - Weight normalisation[4]
  - Multinomial dropout[5]
- Edward Grefenstette's talk at NAMPI workshop[6]

[1] https://papers.nips.cc/paper/6057-using-fast-weights-to-attend-to-the-recent-past.pdf
[2] https://papers.nips.cc/paper/6310-phased-lstm-accelerating-recurrent-network-training-for-long-or-event-based-sequences.pdf
[3] https://papers.nips.cc/paper/6039-sequential-neural-models-with-stochastic-layers.pdf
[4] https://papers.nips.cc/paper/6114-weight-normalization-a-simple-reparameterization-to-accelerate-training-of-deep-neural-networks.pdf
[5] https://papers.nips.cc/paper/6561-improved-dropout-for-shallow-and-deep-learning.pdf
[6] https://uclmr.github.io/nampi/talk_slides/grefenstette_limitations_of_rnns.pdf

Improving classic algorithms

- Matrix completion: Non-convex objective for matrix completion has no spurious local minima[1]
- Clustering: Seedings for large-scale k-means++ clustering[2]

[1] https://papers.nips.cc/paper/6048-matrix-completion-has-no-spurious-local-minimum.pdf
[2] https://papers.nips.cc/paper/6478-fast-and-provably-good-seedings-for-k-means.pdf

k-Means algorithms

SEEDING
Find initial cluster centers

→

FINE-TUNING
Iteratively improve solution

k-Means++ seeding SLOW

??

Lloyd's algorithm SLOW

Mini-batch k-Means FAST

NEED FOR FAST AND GOOD SEEDINGS

Improving classic algorithms

- Matrix completion: Non-convex objective for matrix completion has no spurious local minima[1]
- Clustering: Seedings for large-scale k-means++ clustering[2]
- Clustering: Using queries to convey domain information and reduce computational complexity[3]

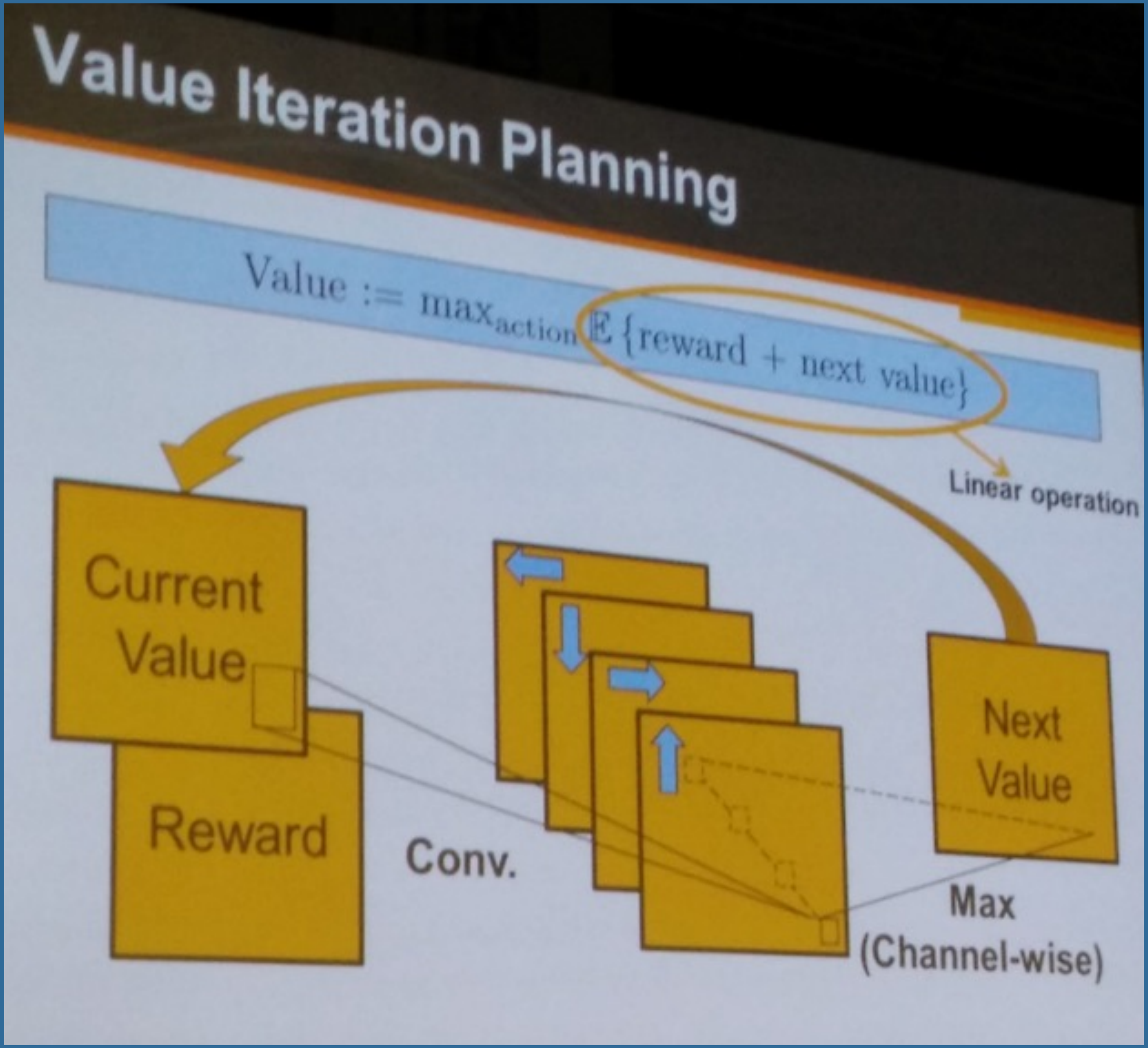
[1] https://papers.nips.cc/paper/6048-matrix-completion-has-no-spurious-local-minimum.pdf
[2] https://papers.nips.cc/paper/6478-fast-and-provably-good-seedings-for-k-means.pdf
[3] https://papers.nips.cc/paper/6449-clustering-with-same-cluster-queries.pdf

Reinforcement Learning

- RL tutorial by Pieter Abbeel and John Schulman: Slides[1]; practical advice[2]
- Best paper: Value Iteration Networks[3]
  - Learns to plan; contribution: differentiable approximation to classic algorithm via CNN

[1] http://people.eecs.berkeley.edu/~pabbeel/nips-tutorial-policy-optimization-Schulman-Abbeel.pdf
[2] http://rll.berkeley.edu/deeprlcourse/docs/nuts-and-bolts.pdf
[3] https://papers.nips.cc/paper/6046-value-iteration-networks.pdf

Reinforcement Learning



Reinforcement Learning

- RL tutorial by Pieter Abbeel and John Schulman: Slides[1]; practical advice[2]
- Best paper: Value Iteration Networks[3]
  - Learns to plan; contribution: differentiable approximation to classic algorithm via CNN
- New research environments:
  - OpenAI's Universe
  - Deep Mind Lab
  - FAIR's Torchcraft

[1] http://people.eecs.berkeley.edu/~pabbeel/nips-tutorial-policy-optimization-Schulman-Abbeel.pdf
[2] http://rll.berkeley.edu/deeprlcourse/docs/nuts-and-bolts.pdf
[3] https://papers.nips.cc/paper/6046-value-iteration-networks.pdf

Learning-to-learn / Meta-learning

- Several papers, e.g.
  Learning to learn by gradient descent by gradient descent[1]
- Meta-learning panel at RNN symposium
  - Sutskever: No good meta-learning models so far
  - Schmidhuber: Focus on universal model
- Focus at Neural Abstract Machines workshop

[1] https://papers.nips.cc/paper/6461-learning-to-learn-by-gradient-descent-by-gradient-descent.pdf

General Artificial Intelligence

- Topic in keynote talks
  - Yann LeCun: Focus on unsupervised learning

Common Sense is the ability to fill in the blanks

Y LeCun

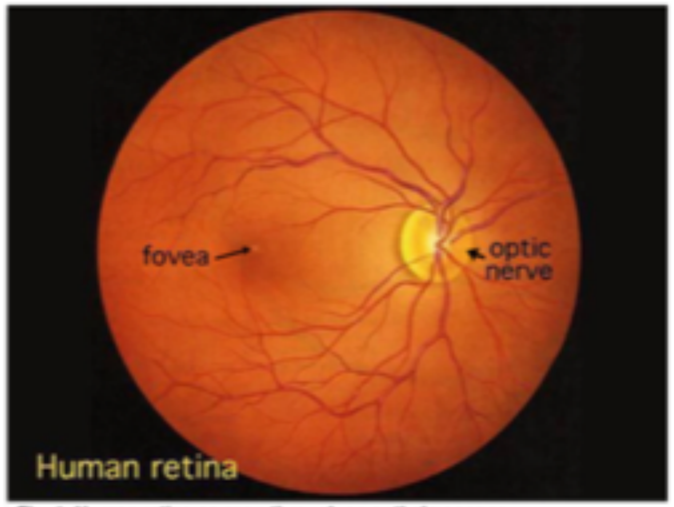


Fig. 1. Human retina as seen through an opthalmoscope.

- Infer the state of the world from partial information
- Infer the future from the past and present
- Infer past events from the present state

- Filling in the visual field at the retinal blind spot
- Filling in occluded images
- Fillling in missing segments in text, missing words in speech.
- Predicting the consequences of our actions
- Predicting the sequence of actions leading to a result

- Predicting any part of the past, present or future percepts from whatever information is available.

- That's what predictive learning is
- But really, that's what many people mean by unsupervised learning

How Much Information Does the Machine Need to Predict?

Y LeCun

- **"Pure" Reinforcement Learning (cherry)**
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- **Supervised Learning (icing)**
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- **Unsupervised/Predictive Learning (cake)**
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

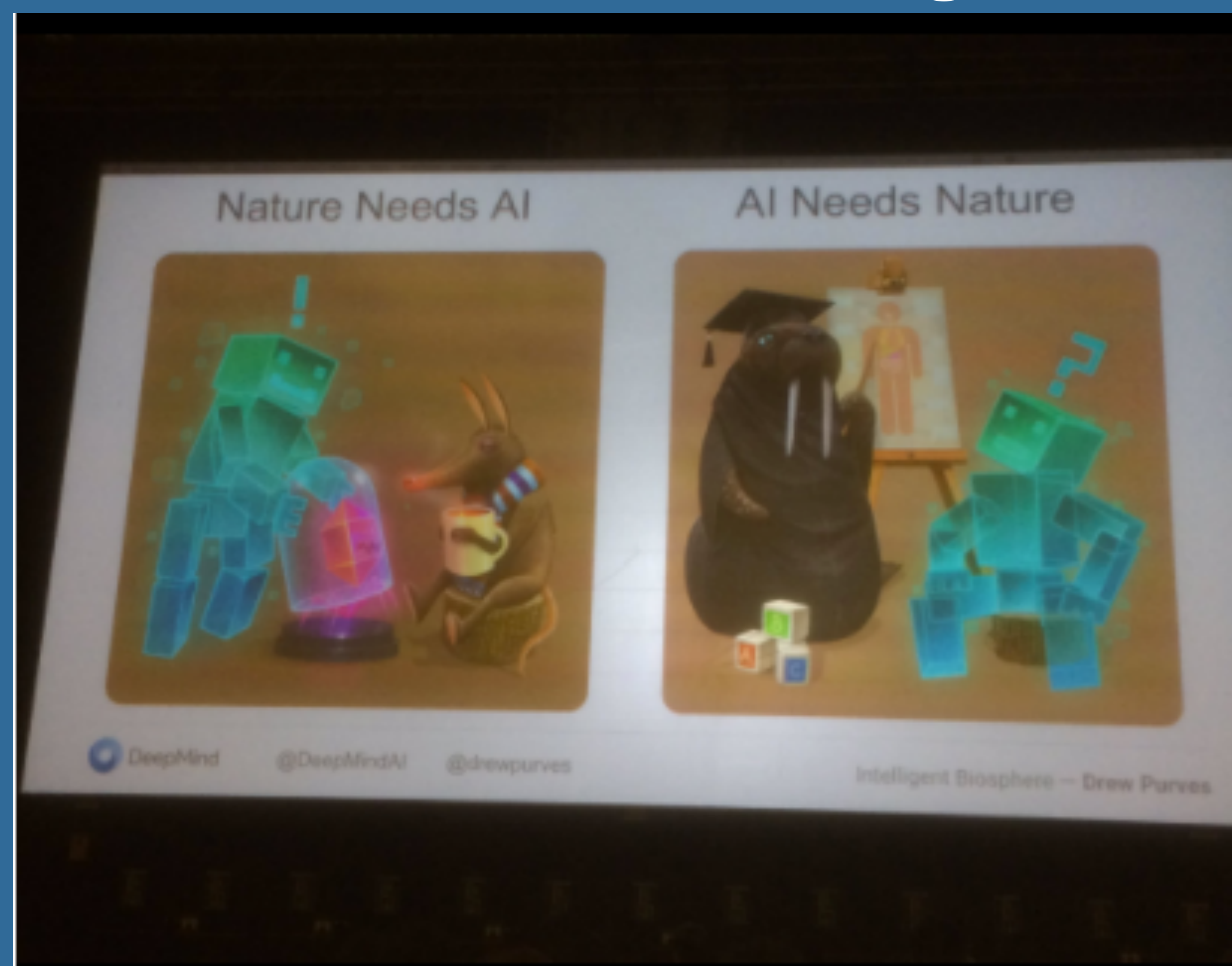
- (Yes, I know, this picture is slightly offensive to RL folks. But I'll make it up)

General Artificial Intelligence

- Topic in keynote talks
  - Yann LeCun: Focus on unsupervised learning
  - Drew Purves: AI for environment, ground for General AI

General Artificial Intelligence

- Topic in keynote talks
  - Yann LeCun: Focus on unsupervised learning
  - Drew Purves: AI for environment, ground for General AI



General Artificial Intelligence

Fuzzy

Scaleless

Granular

Emergent

Resources

Energy

Reproduction

Multi-agent

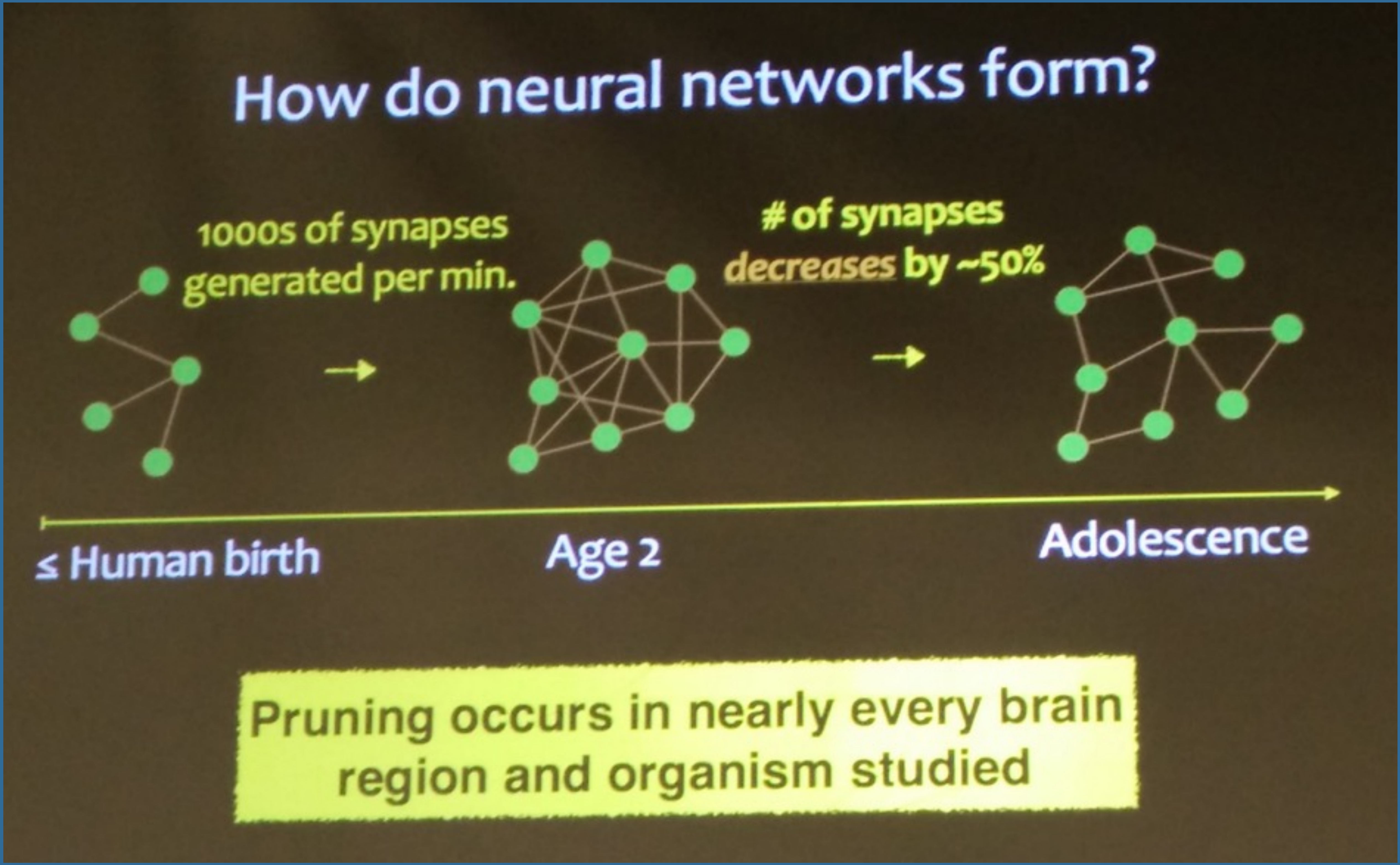
Diversity

Source: Drew Purves

General Artificial Intelligence

- Topic in keynote talks
  - Yann LeCun: Focus on unsupervised learning
  - Drew Purves: AI for environment, ground for General AI
  - Saket Navlakha: Borrowing engineering principles from the brain

General Artificial Intelligence



General Artificial Intelligence

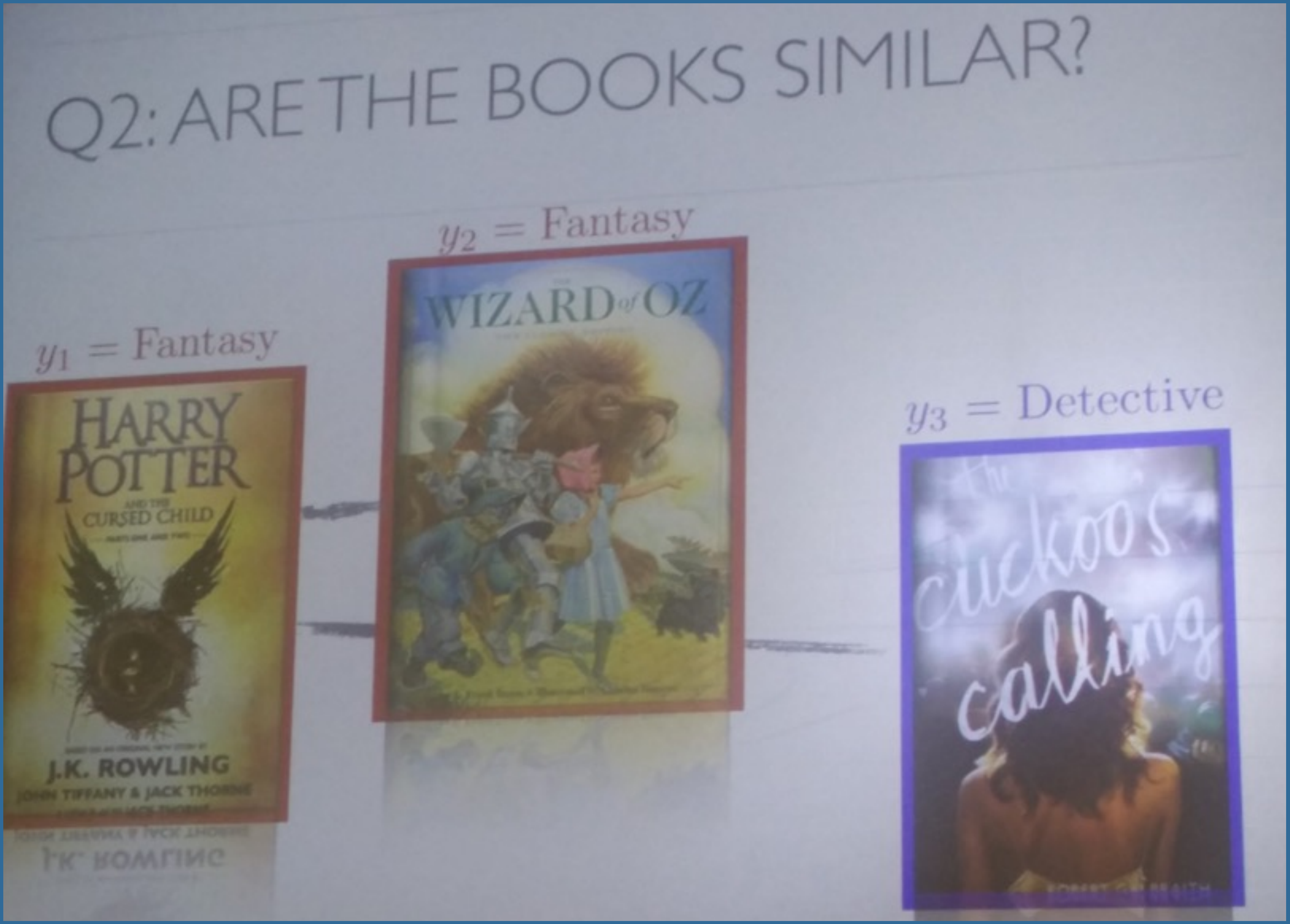
- Topic in keynote talks
  - Yann LeCun: Focus on unsupervised learning
  - Drew Purves: AI for environment, ground for General AI
  - Saket Navlakha: Borrowing engineering principles from the brain
- Machine Intelligence workshop:
  - FAIR's CommAI-env
  - Hierarchical representations, identifying patterns
  - Panel: Unclear if focus on specific tasks brings us closer to General AI.

NLP

- Variations of Neural MT
  - Dual learning for MT (with RL)[1]
  - Structured prediction with bandit feedback[2]
- Improvements to existing methods
  - Supervised Word Mover's Distance[3]

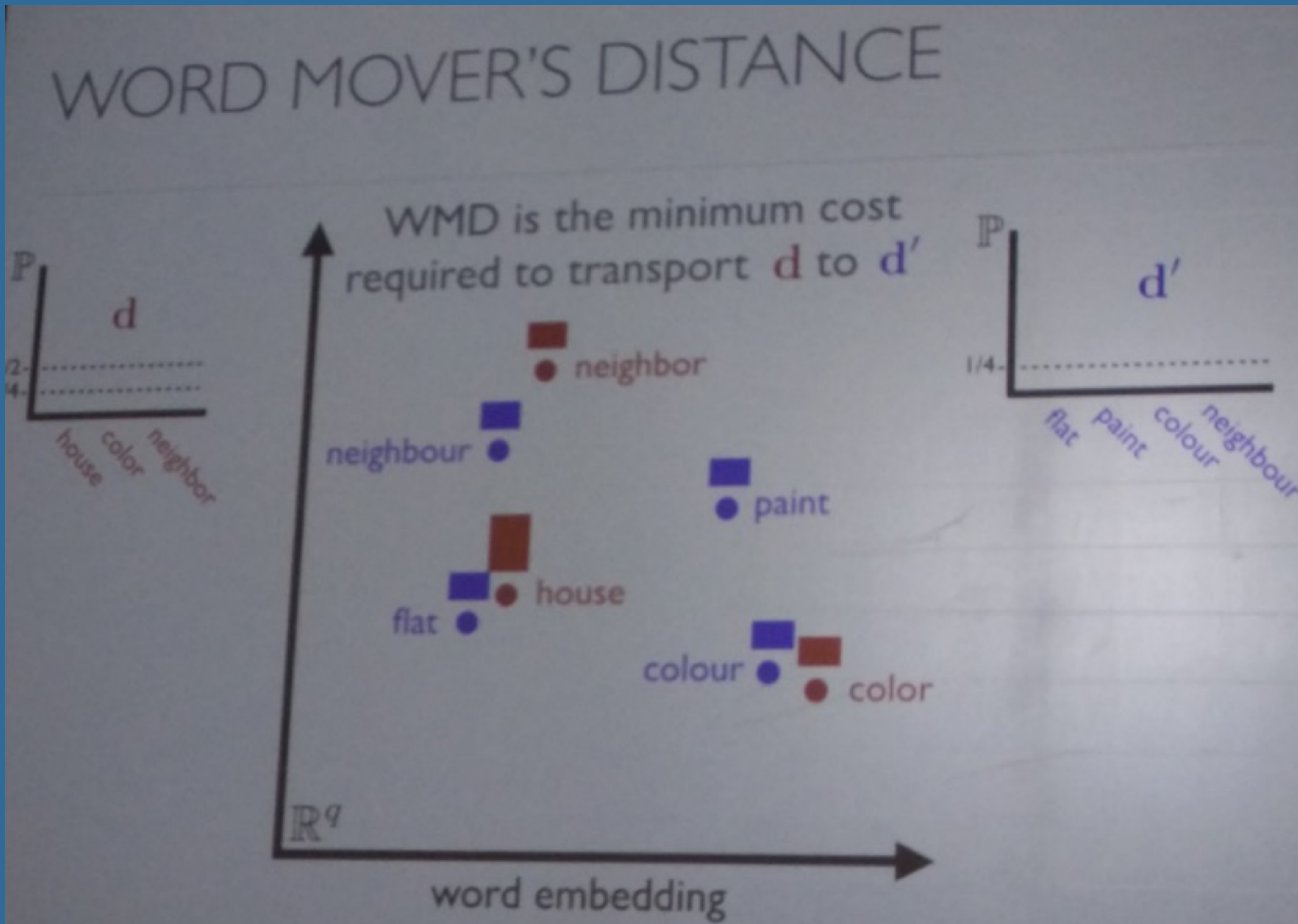
[1] https://papers.nips.cc/paper/6469-dual-learning-for-machine-translation.pdf
[2] https://papers.nips.cc/paper/6134-stochastic-structured-prediction-under-bandit-feedback.pdf
[3] https://papers.nips.cc/paper/6139-supervised-word-movers-distance.pdf

NLP



Q2: ARE THE BOOKS SIMILAR?
y_1 = Fantasy
y_2 = Fantasy
y_3 = Detective
HARRY POTTER
AND THE
CURSED CHILD
J.K. ROWLING
JOHN TIFFANY & JACK THORNE
WIZARD of OZ
The Cuckoo's calling

NLP

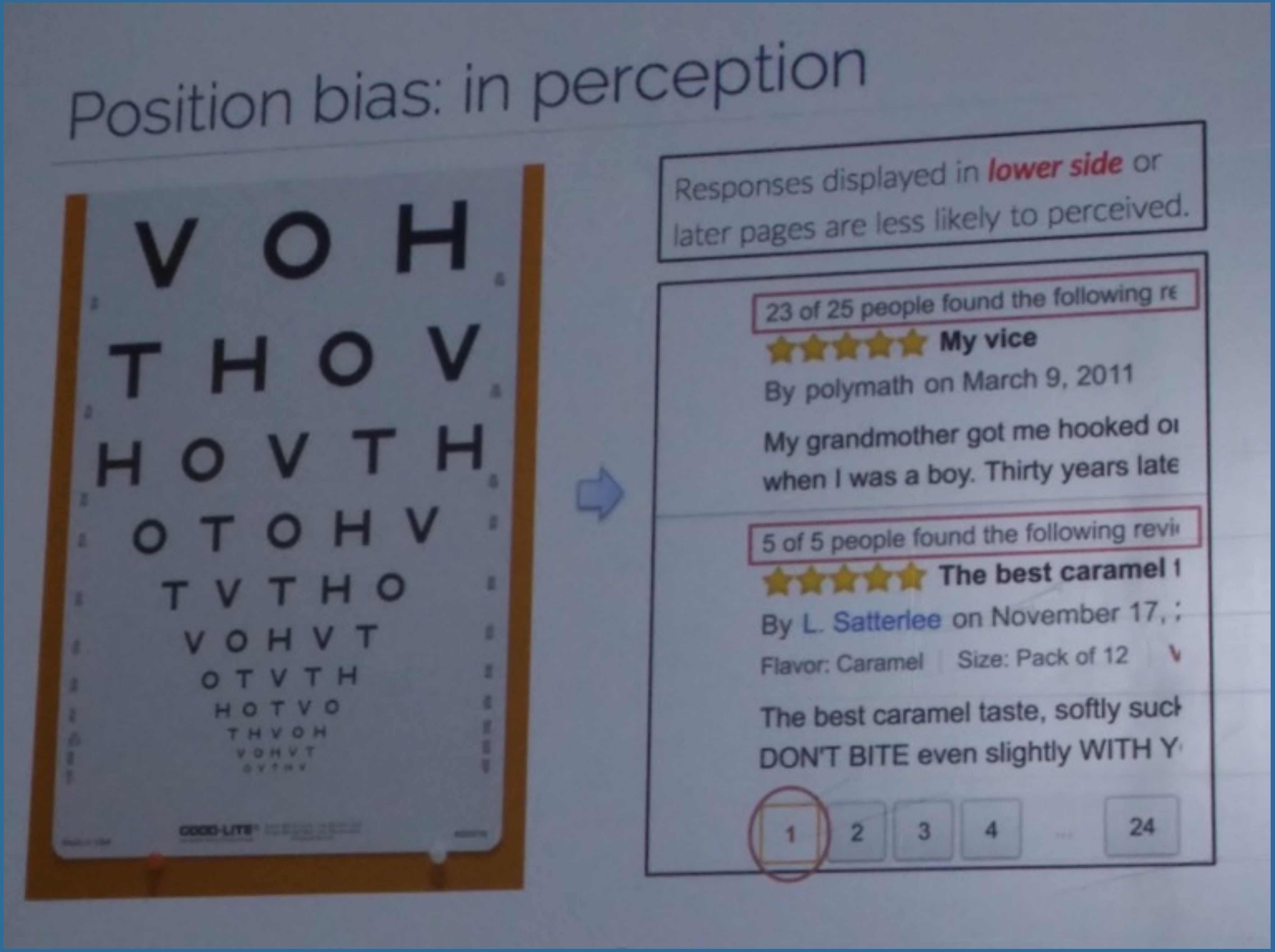

WORD MOVER'S DISTANCE
WMD is the minimum cost
required to transport d to d′
ℙ
d
house
color
neighbor
ℙ
d′
1/4
flat
paint
colour
neighbour
neighbor
neighbour
paint
house
flat
colour
color
ℝ^q
word embedding

NLP

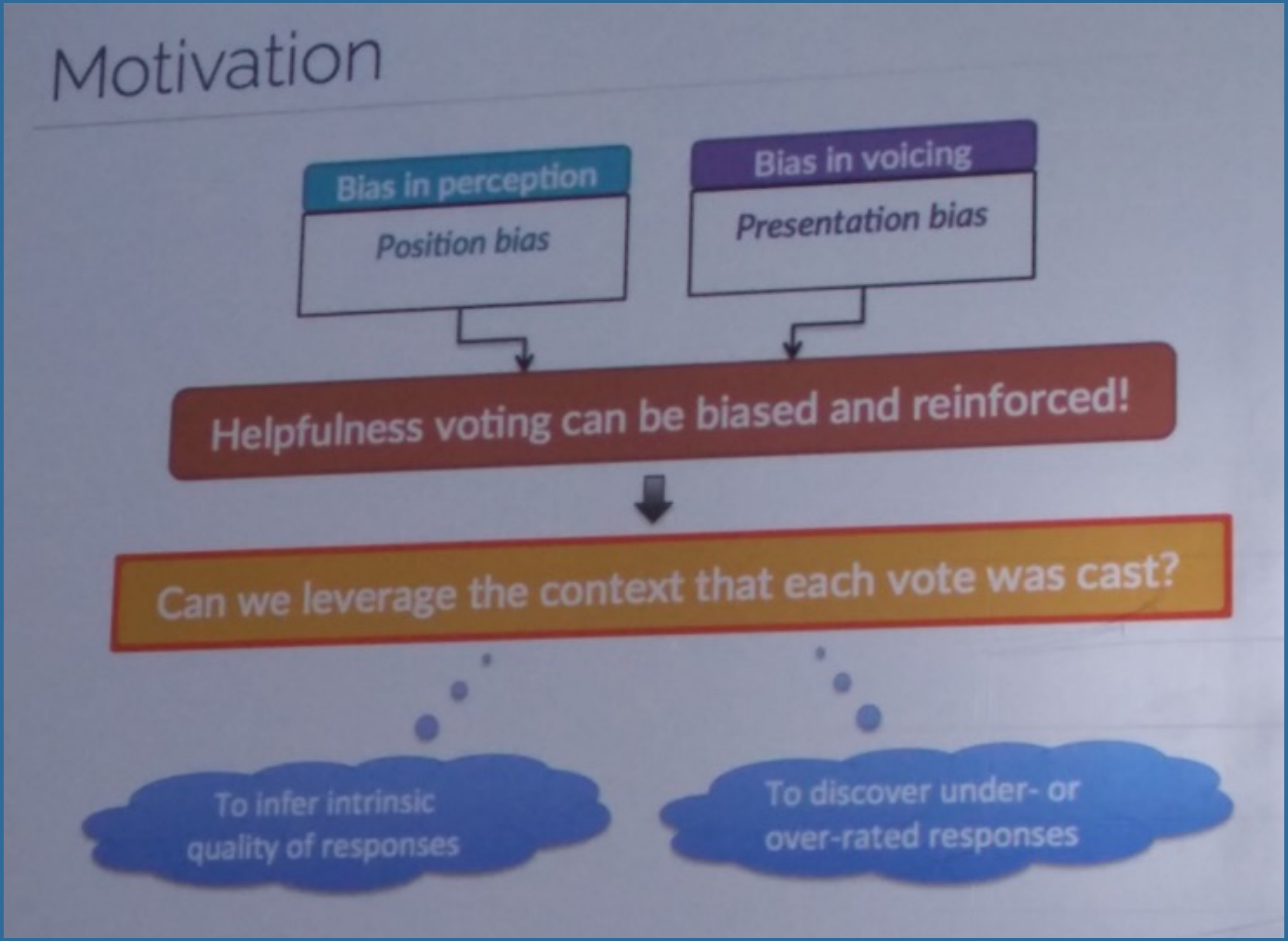
- Variations of Neural MT
  - Dual learning for MT (with RL)[1]
  - Structured prediction with bandit feedback[2]
- Improvements to existing methods
  - Supervised Word Mover's Distance[3]
  - Modelling user reviews[4]

[1] https://papers.nips.cc/paper/6469-dual-learning-for-machine-translation.pdf
[2] https://papers.nips.cc/paper/6134-stochastic-structured-prediction-under-bandit-feedback.pdf
[3] https://papers.nips.cc/paper/6139-supervised-word-movers-distance.pdf
[4] https://papers.nips.cc/paper/6362-beyond-exchangeability-the-chinese-voting-process.pdf

NLP


Position bias: in perception
V O H
T H O V
H O V T H
O T O H V
T V T H O
V O H V T
O T V T H
H O T V O
Responses displayed in lower side or later pages are less likely to perceived.
23 of 25 people found the following re
My vice
By polymath on March 9, 2011
My grandmother got me hooked o
when I was a boy. Thirty years late
5 of 5 people found the following revi
The best caramel
By L. Satterlee on November 17,
Flavor: Caramel | Size: Pack of 12
The best caramel taste, softly suc
DON'T BITE even slightly WITH Y
1
2
3
4
24

NLP


Motivation
Bias in perception
Position bias
Bias in voicing
Presentation bias
Helpfulness voting can be biased and reinforced!
Can we leverage the context that each vote was cast?
To infer intrinsic quality of responses
To discover under- or over-rated responses

NLP

- Variations of Neural MT
  - Dual learning for MT (with RL)[1]
  - Structured prediction with bandit feedback[2]
- Improvements to existing methods
  - Supervised Word Mover's Distance[3]
  - Modelling user reviews[4]
- Dialogue modelling workshop:
  - End-to-end systems, linguistics, and ML methods

[1] https://papers.nips.cc/paper/6469-dual-learning-for-machine-translation.pdf
[2] https://papers.nips.cc/paper/6134-stochastic-structured-prediction-under-bandit-feedback.pdf
[3] https://papers.nips.cc/paper/6139-supervised-word-movers-distance.pdf
[4] https://papers.nips.cc/paper/6362-beyond-exchangeability-the-chinese-voting-process.pdf

Miscellaneous

- Schmidhuber is everywhere!

hardmaru @hardmaru · 5. Dez.
NIPS2016 Day 1: Poor @Goodfellow_Ian gets Schmidhuber'ed during educational GAN Tutorial session.
9 30 144

Miscellaneous

- Schmidhuber is everywhere!
- Robotics

Miscellaneous

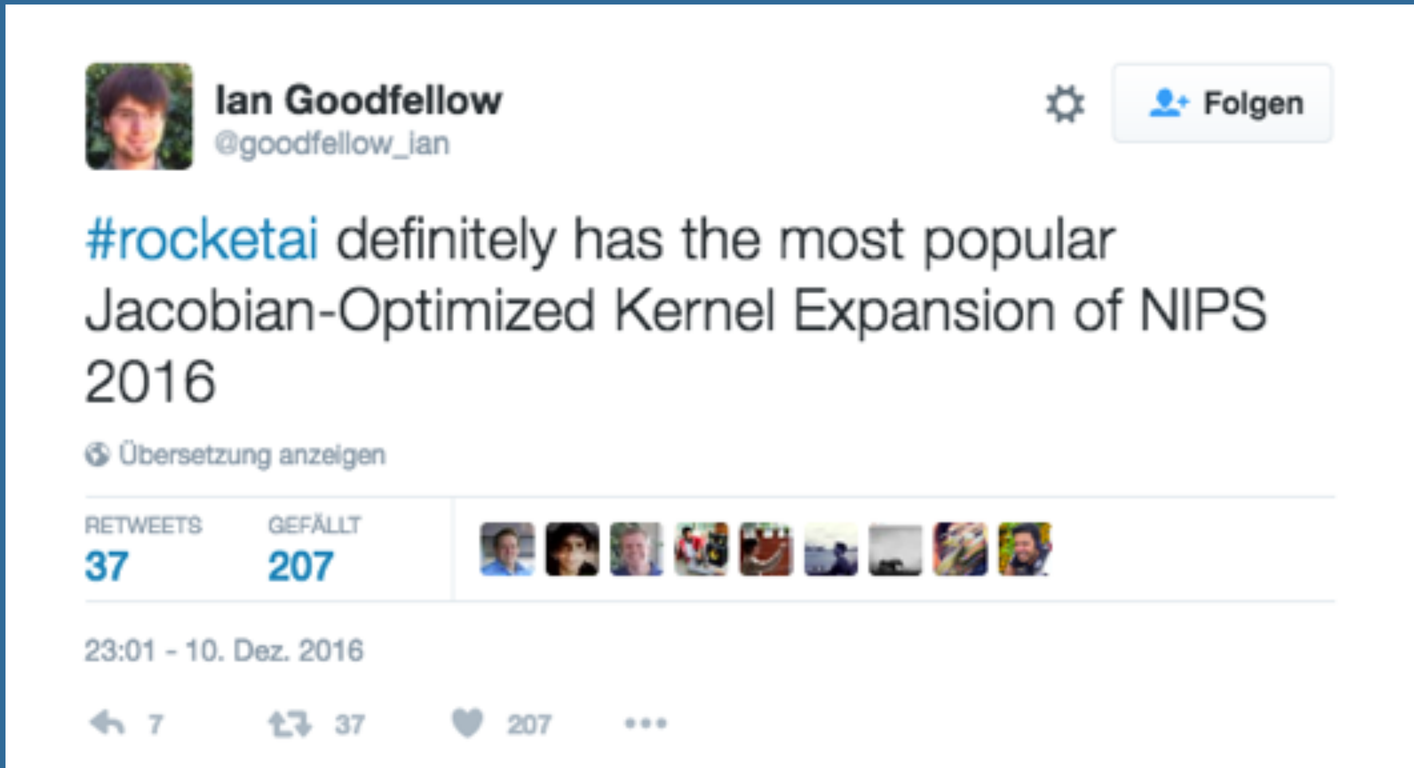
- Schmidhuber is everywhere!
- Robotics
- Apple starts publishing

Miscellaneous

- Schmidhuber is everywhere!
- Robotics
- Apple starts publishing
- AI startups:
  - Geometric Intelligence

Miscellaneous

- Schmidhuber is everywhere!
- Robotics
- Apple starts publishing
- AI startups:
    - Geometric Intelligence
    - RocketAI "launch" party



Thanks for your attention!

Questions?