Some Success Stories in Bridging Theory and Practice

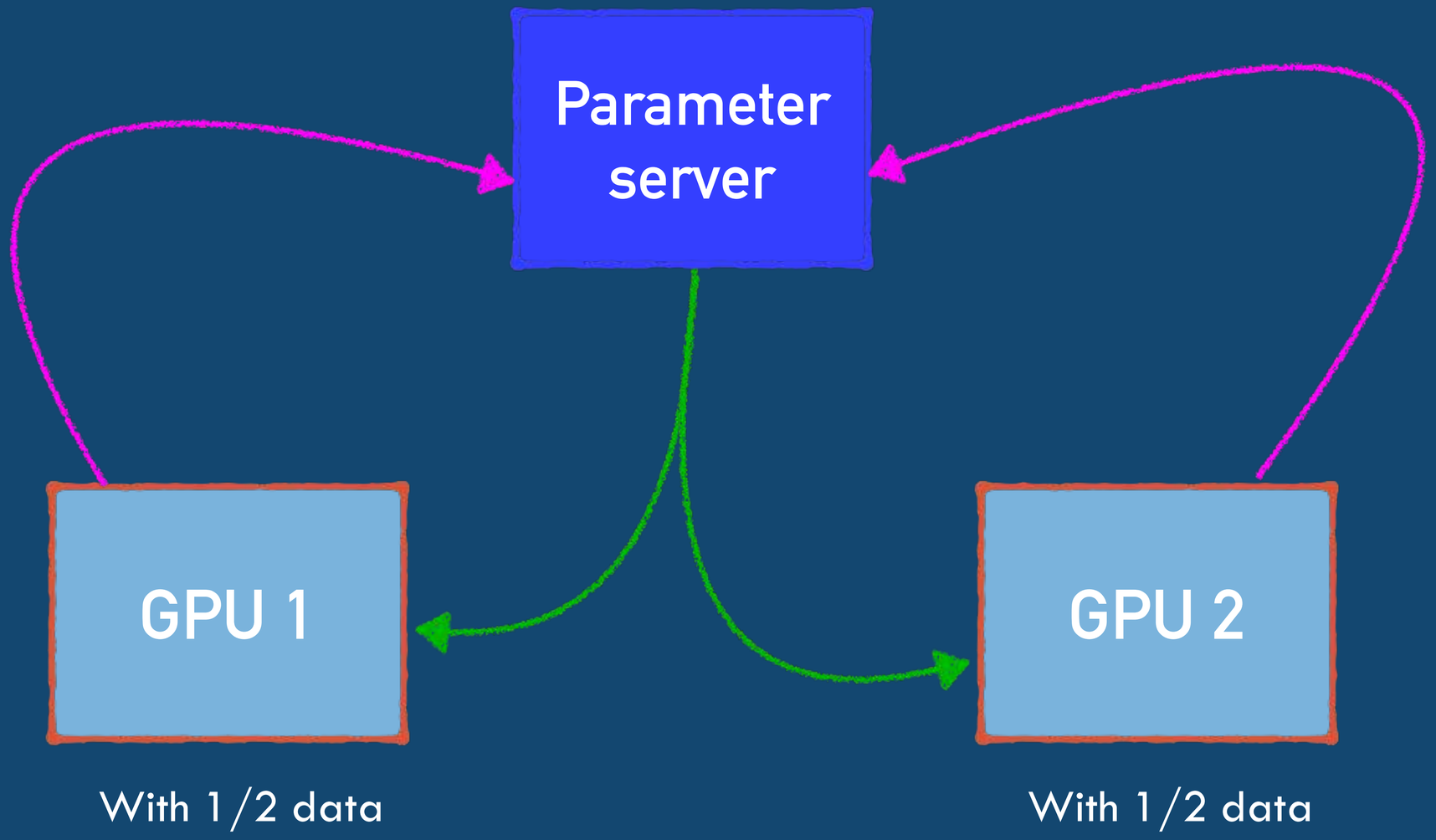
Anima Anandkumar

Bren Professor at Caltech
Director of ML Research at NVIDIA

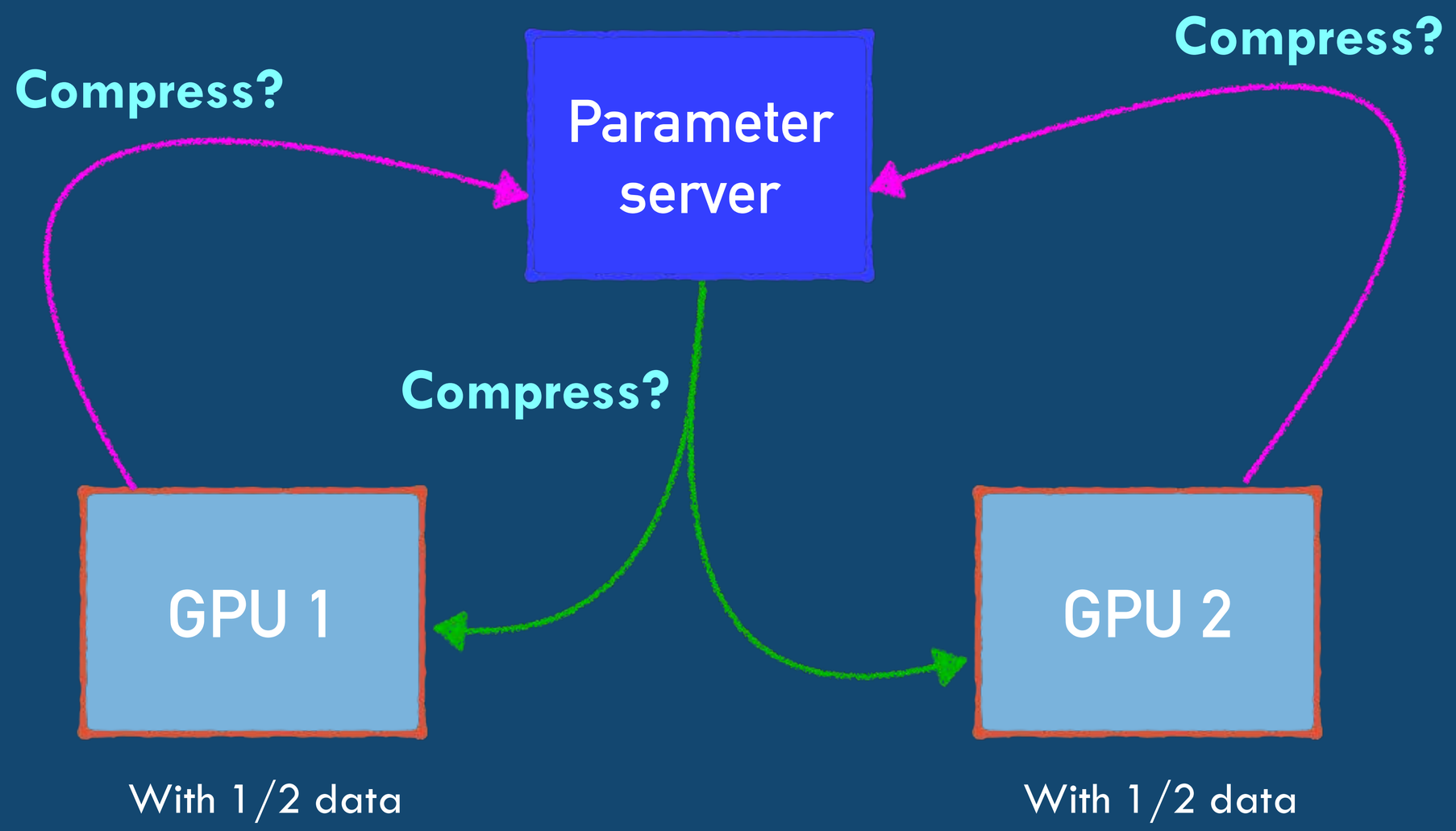
SIGNSGD: COMPRESSED OPTIMIZATION FOR NON-CONVEX PROBLEMS

JEREMY BERNSTEIN, JIAWEI ZHAO, KAMYAR AZZIZADENESHELI, YU-XIANG WANG, ANIMA ANANDKUMAR

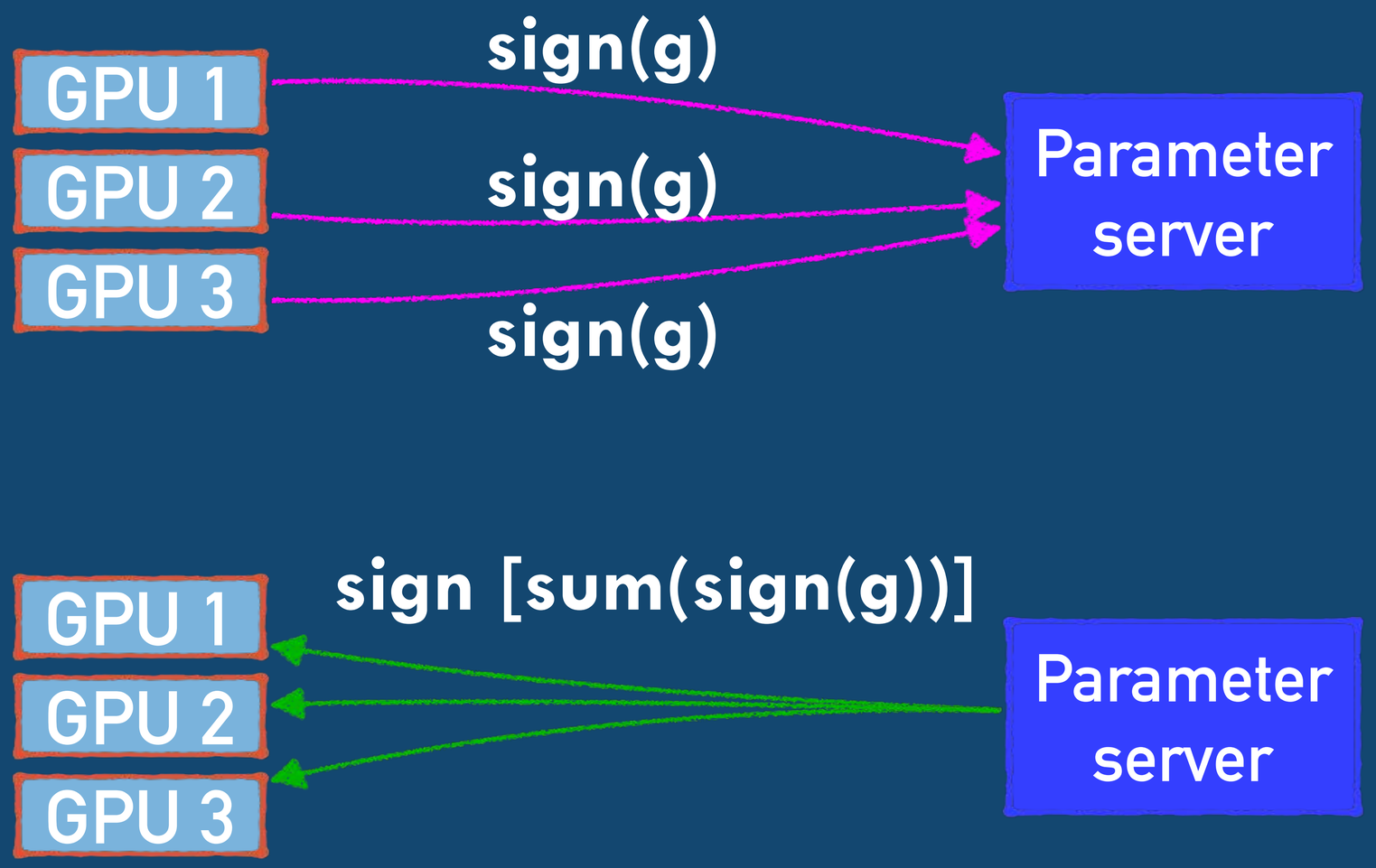
DISTRIBUTED TRAINING INVOLVES COMPUTATION & COMMUNICATION



DISTRIBUTED TRAINING INVOLVES COMPUTATION & COMMUNICATION



DISTRIBUTED TRAINING BY MAJORITY VOTE



LARGE-BATCH ANALYSIS

SINGLE WORKER RESULTS

Assumptions

- Objective function lower bound f_*
- Coordinate-wise variance bound $\vec{\sigma}$
- Coordinate-wise gradient Lipschitz $\vec{L}$

Define

- Number of iterations K
- Number of backpropagations N

SGD gets rate

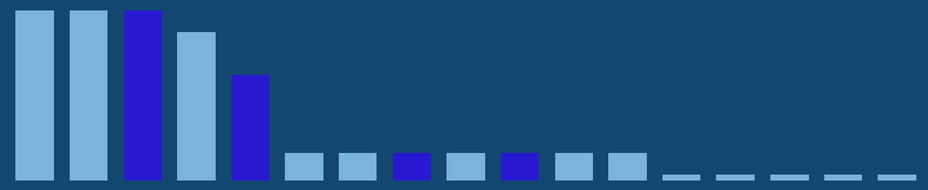
$$\mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1}\|g_k\|_2^2\right] \leq \frac{1}{\sqrt{N}}\left[2\|\vec{L}\|_\infty(f_0 - f_*) + \|\vec{\sigma}\|_2^2\right]$$

signSGD gets rate

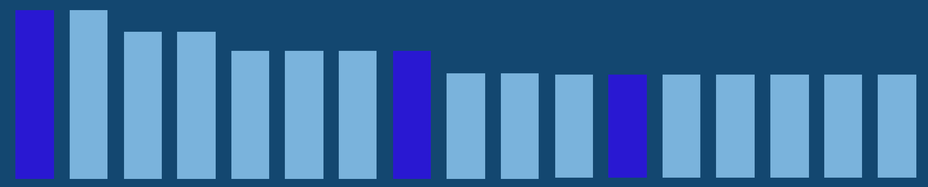
$$\mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1}\sqrt{d}\|\vec{g}_k\|_2\right]^2 \leq \frac{1}{\sqrt{N}}\left[\sqrt{d}\sqrt{\|\vec{L}\|_\infty}\left(f_0 - f_* + \frac{1}{2}\right) + 2\sqrt{d}\|\vec{\sigma}\|_2\right]^2$$

VECTOR DENSITY & ITS RELEVANCE IN DEEP LEARNING

A sparse vector



A dense vector

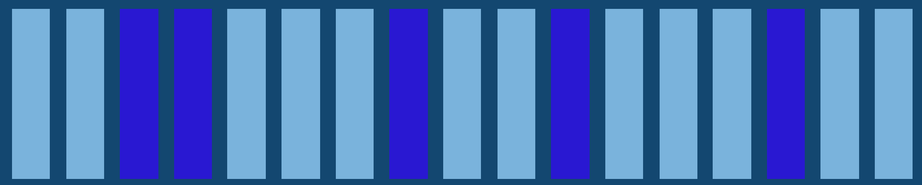


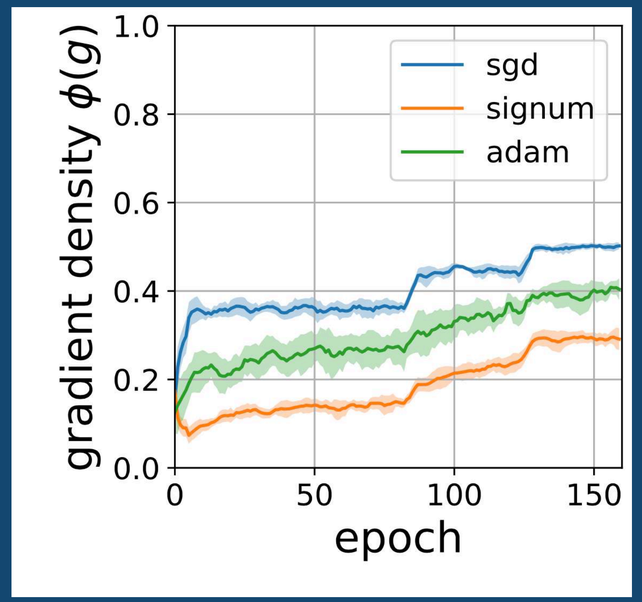
Natural measure of density

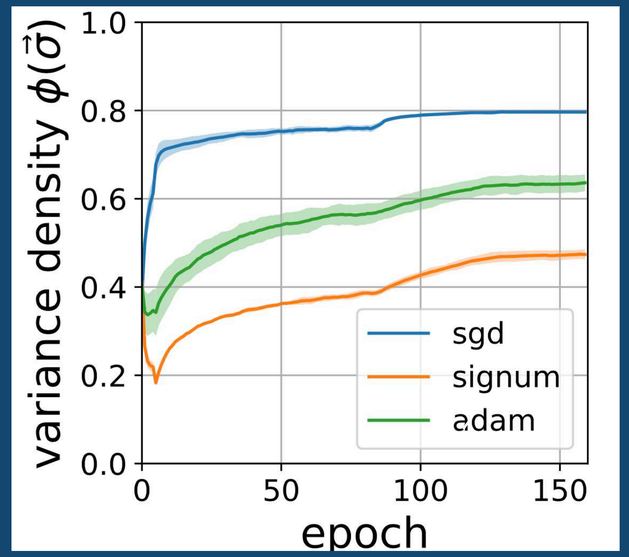
$$\phi(\vec{v}) = \frac{\|\vec{v}\|_1^2}{d\|\vec{v}\|_2^2}$$

=1 for fully dense v
≈0 for fully sparse v

Fully dense vector……a sign vector

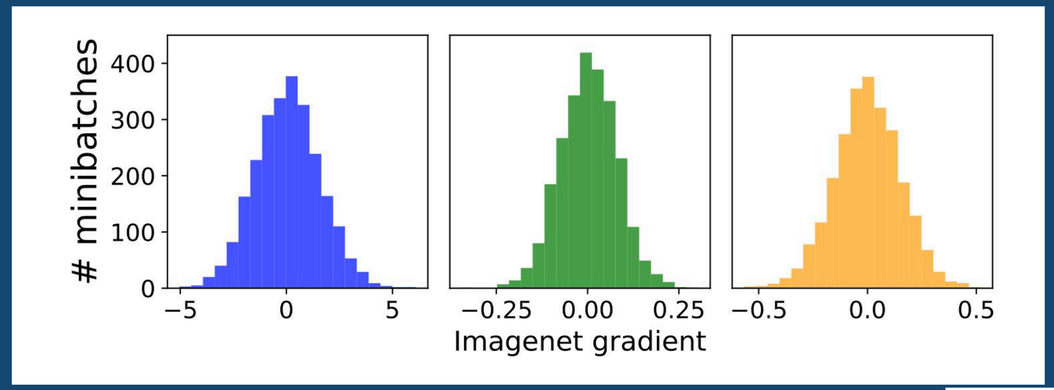






DISTRIBUTED SIGNSGD: MAJORITY VOTE THEORY

If gradients are unimodal and symmetric…



…reasonable by central limit theorem…

…majority vote with M workers converges at rate:

$$\mathbb{E}\left[\min_{0\le k\le K-1}\|g_k\|_1\right]^2$$

$$\le \frac{1}{\sqrt{N}}\left[\sqrt{\|\vec{L}\|_1}\left(f_0 - f_* + \frac{1}{2}\right) + \frac{2}{\sqrt{M}}\|\vec{\sigma}\|_1\right]^2$$

Same variance reduction as SGD

MINI-BATCH ANALYSIS

- Under symmetric noise assumption:

Theorem 1 (Non-convex convergence rate of small-batch SIGNSGD). *Run the following algorithm for K iterations under Assumptions 1 to 4: $x_{k+1} = x_k - \eta\,\mathrm{sign}(\tilde{g}_k)$. Set the learning rate, η, and mini-batch size, n, as*

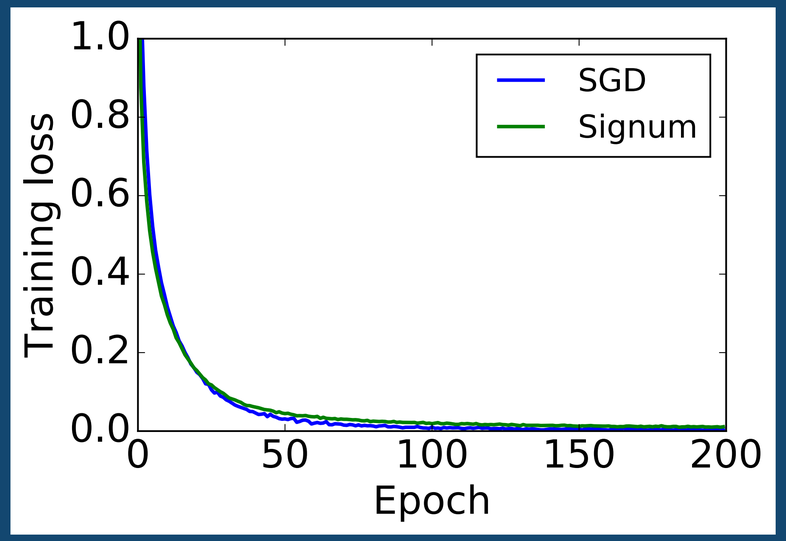
$$\eta = \sqrt{\frac{f_0 - f_*}{\|\vec{L}\|_1 K}}, \qquad n = 1.$$

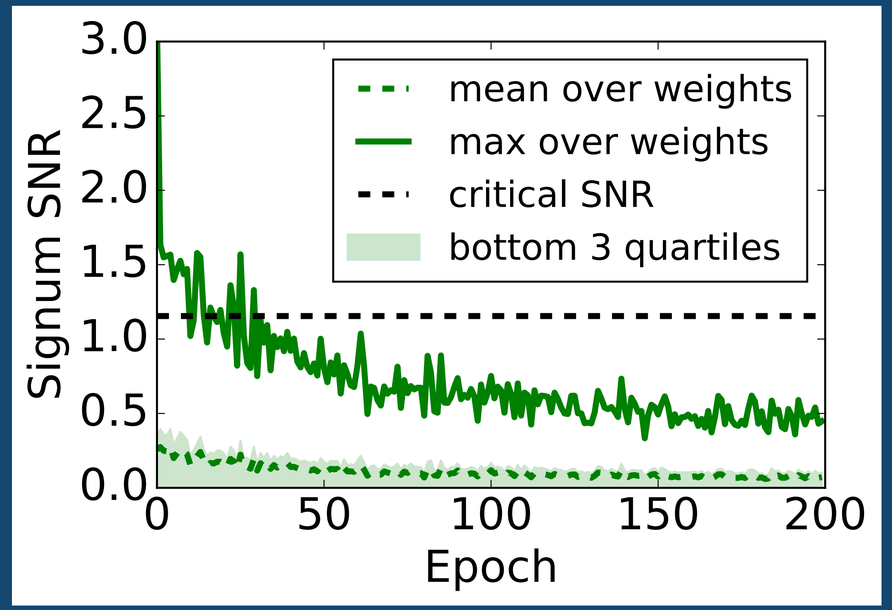
Let H_k be the set of gradient components at step k with large signal-to-noise ratio $S_i := \frac{|g_{k,i}|}{\sigma_i}$, i.e. $H_k := \left\{ i \middle| S_i > \frac{2}{\sqrt{3}} \right\}$. We refer to $\frac{2}{\sqrt{3}}$ as the 'critical SNR'. Then we have

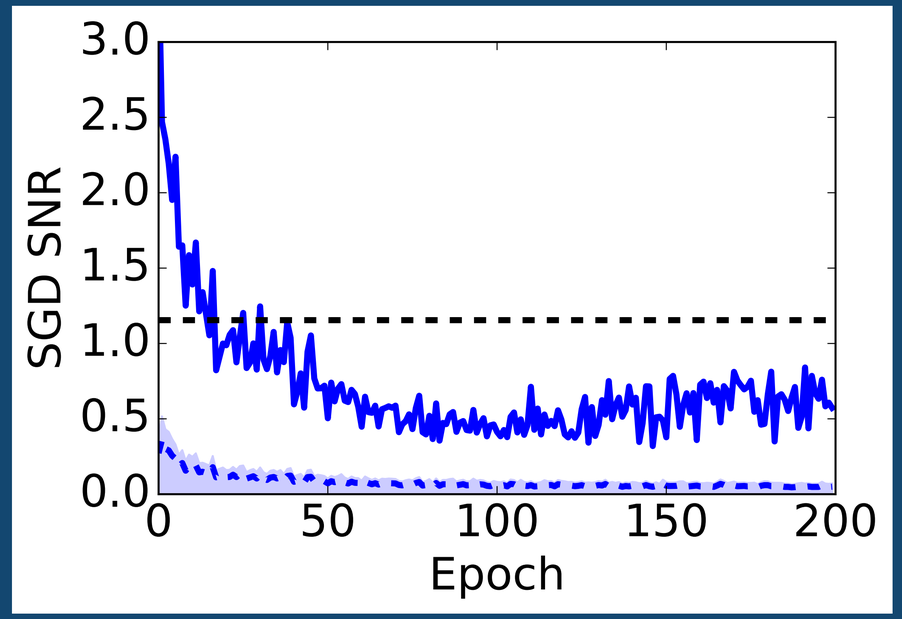
$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\left[\sum_{i \in H_k} |g_{k,i}| + \sum_{i \notin H_k} \frac{g_{k,i}^2}{\sigma_i} \right] \leq 3\sqrt{\frac{\|\vec{L}\|_1 (f_0 - f_*)}{N}}.$$

where $N = K$ is the total number of stochastic gradient calls up to step K.

CIFAR-10 SNR

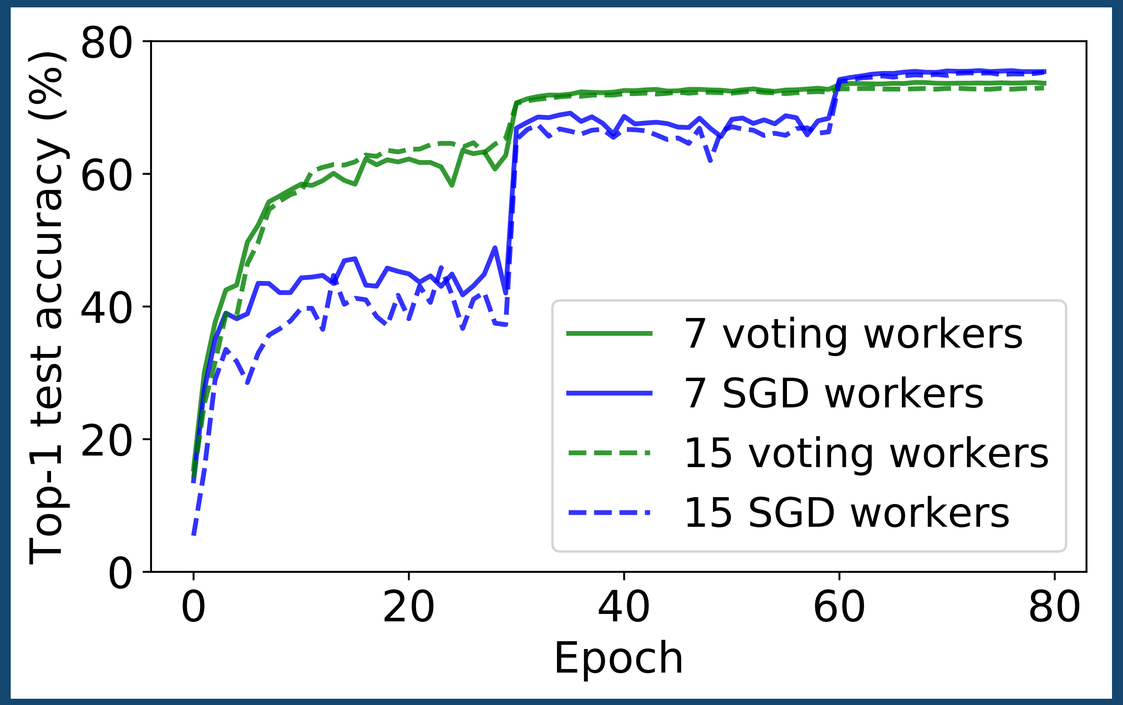

Training loss
1.0
0.8
0.6
0.4
0.2
0.0
0
50
100
150
200
Epoch
SGD
Signum

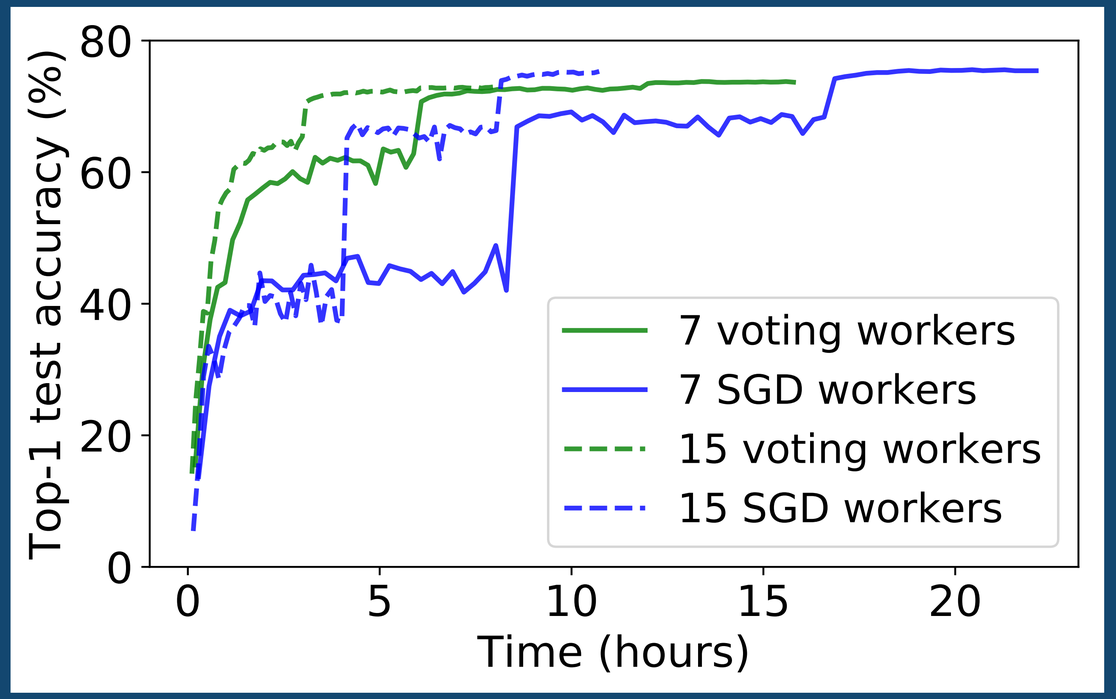

Signum SNR
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
50
100
150
200
Epoch
mean over weights
max over weights
critical SNR
bottom 3 quartiles


SGD SNR
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
50
100
150
200
Epoch

SIGNSGD PROVIDES “FREE LUNCH"

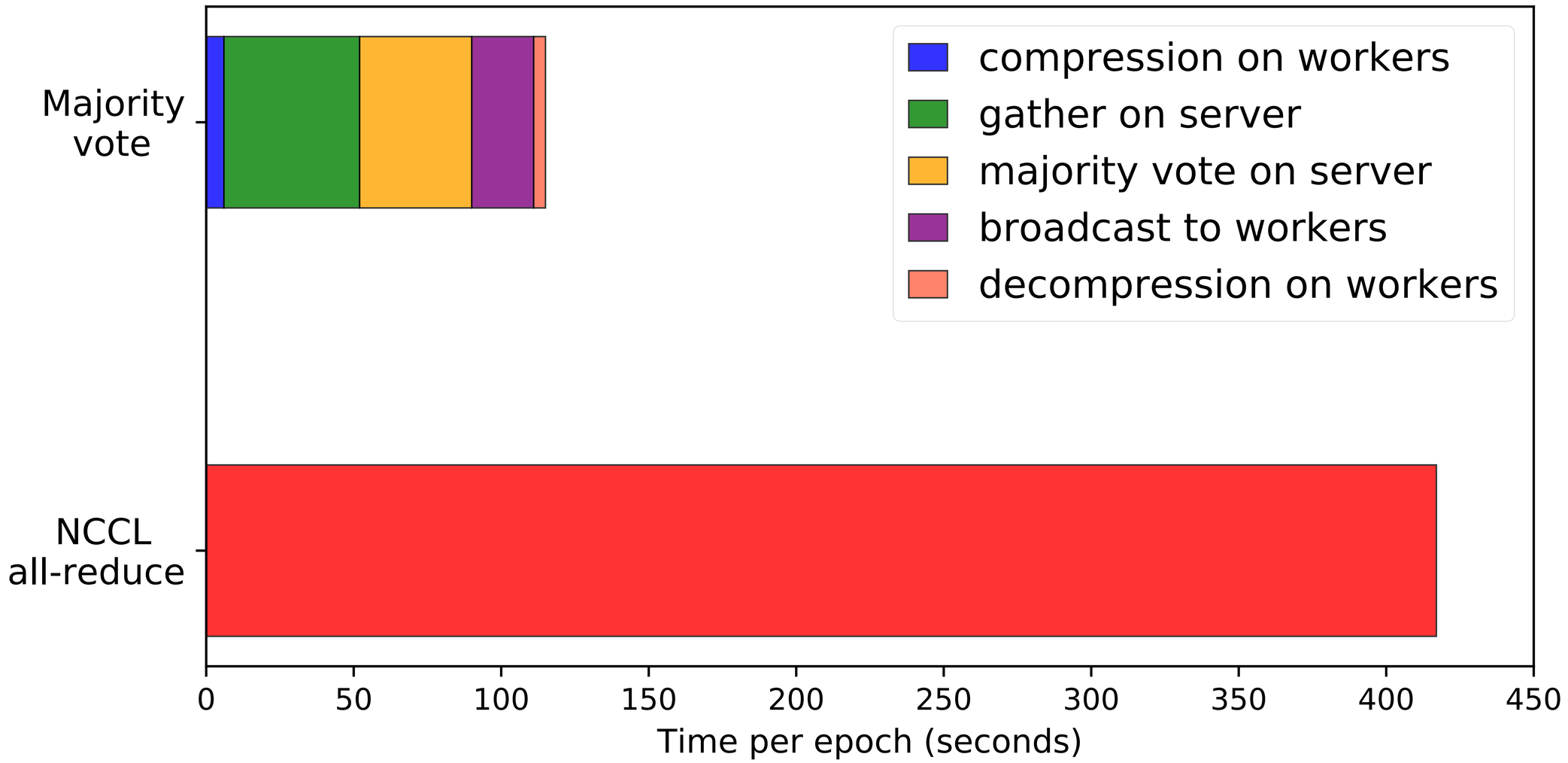
P3.2x machines on AWS, Resnet50 on imagenet



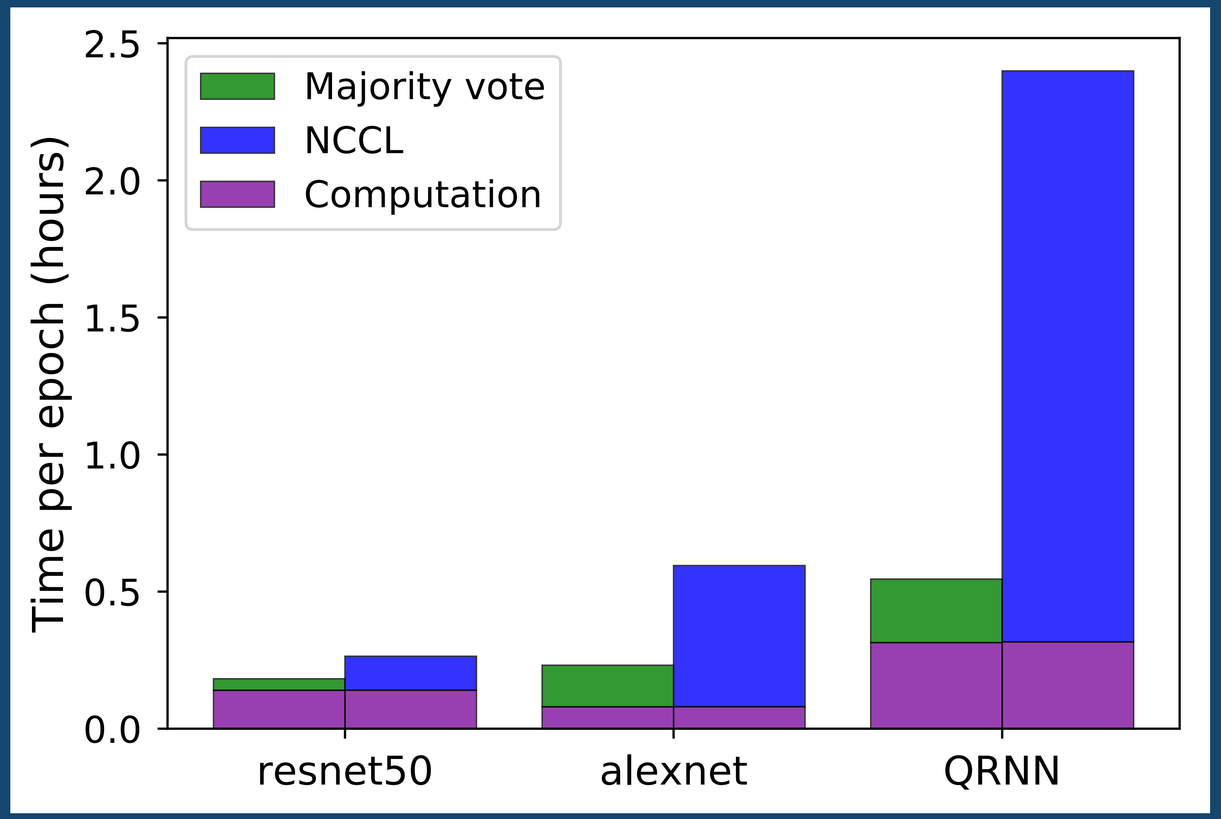


Throughput gain with only tiny accuracy loss

SIGNSGD: TIME PER EPOCH



SIGNSGD ACROSS DOMAINS AND ARCHITECTURES



Huge throughput gain!

BYZANTINE FAULT TOLERANCE

- Under symmetric noise assumption:

Theorem 2 (Non-convex convergence rate of majority vote with adversarial workers). *Run algorithm 1 for K iterations under Assumptions 1 to 4. Switch off momentum and weight decay ($\beta = \lambda = 0$). Set the learning rate, η, and mini-batch size, n, for each worker as*

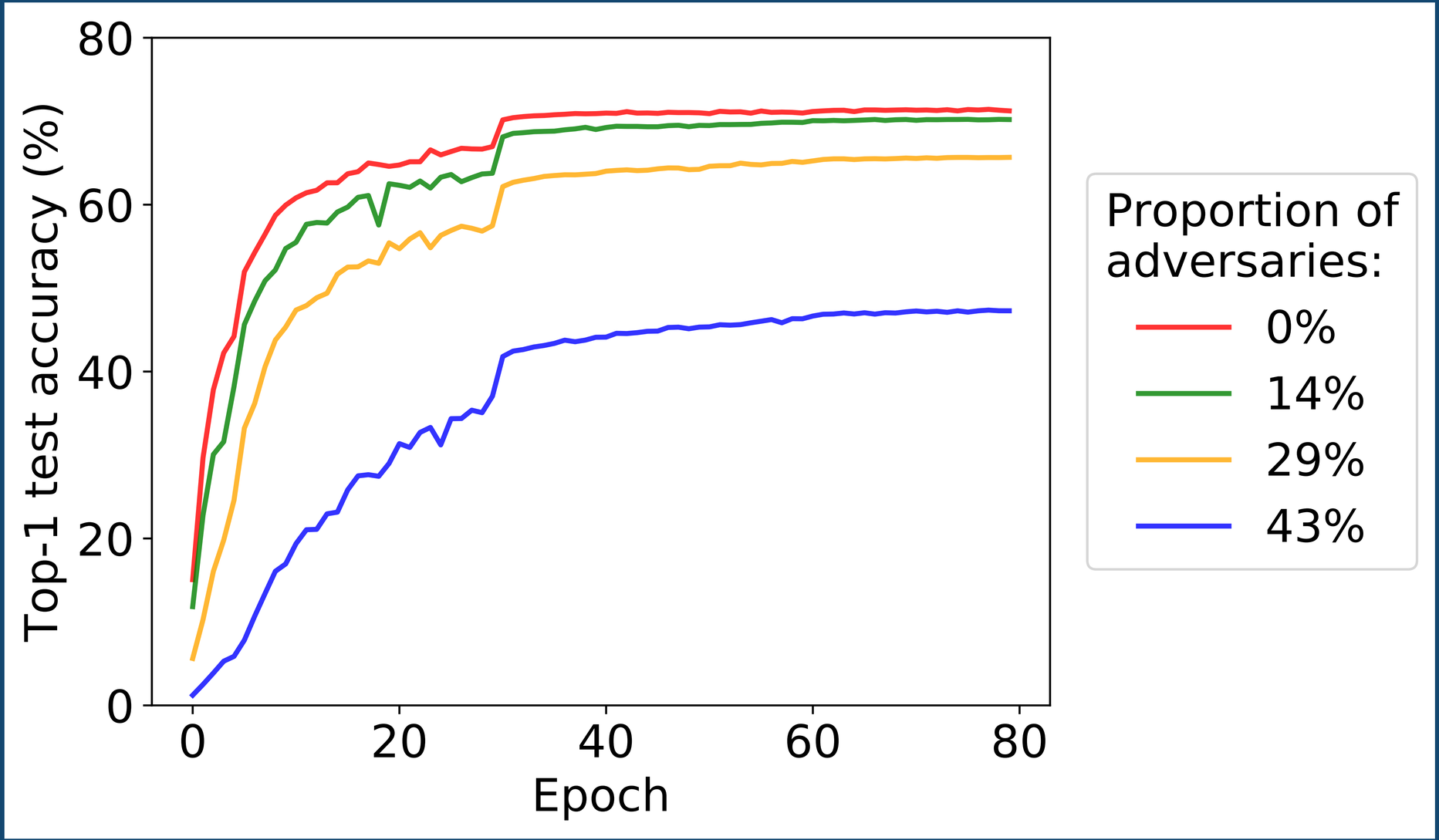
$$\eta = \sqrt{\frac{f_0 - f_*}{\|L\|_1 K}}, \qquad n = K.$$

Assume that a fraction $\alpha < \frac{1}{2}$ of the M workers behave adversarially by sending to the server the negation of their sign gradient estimate. Then majority vote converges at rate:

$$\left[\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\,\|g_k\|_1\right]^2 \leq \frac{4}{\sqrt{N}}\left[\frac{1}{1-2\alpha}\frac{\|\vec{\sigma}\|_1}{\sqrt{M}} + \sqrt{\|L\|_1(f_0 - f^*)}\right]^2$$

where $N = K^2$ is the total number of stochastic gradient calls per worker up to step K.

SIGNSGD IS ALSO BYZANTINE FAULT TOLERANT



TAKE-AWAYS FOR SIGN-SGD

- Convergence even under biased gradients and noise.
- **Faster than SGD** in theory and in practice.
- For distributed training, similar variance reduction as SGD.
- In practice, similar accuracy but with **far less communication.**

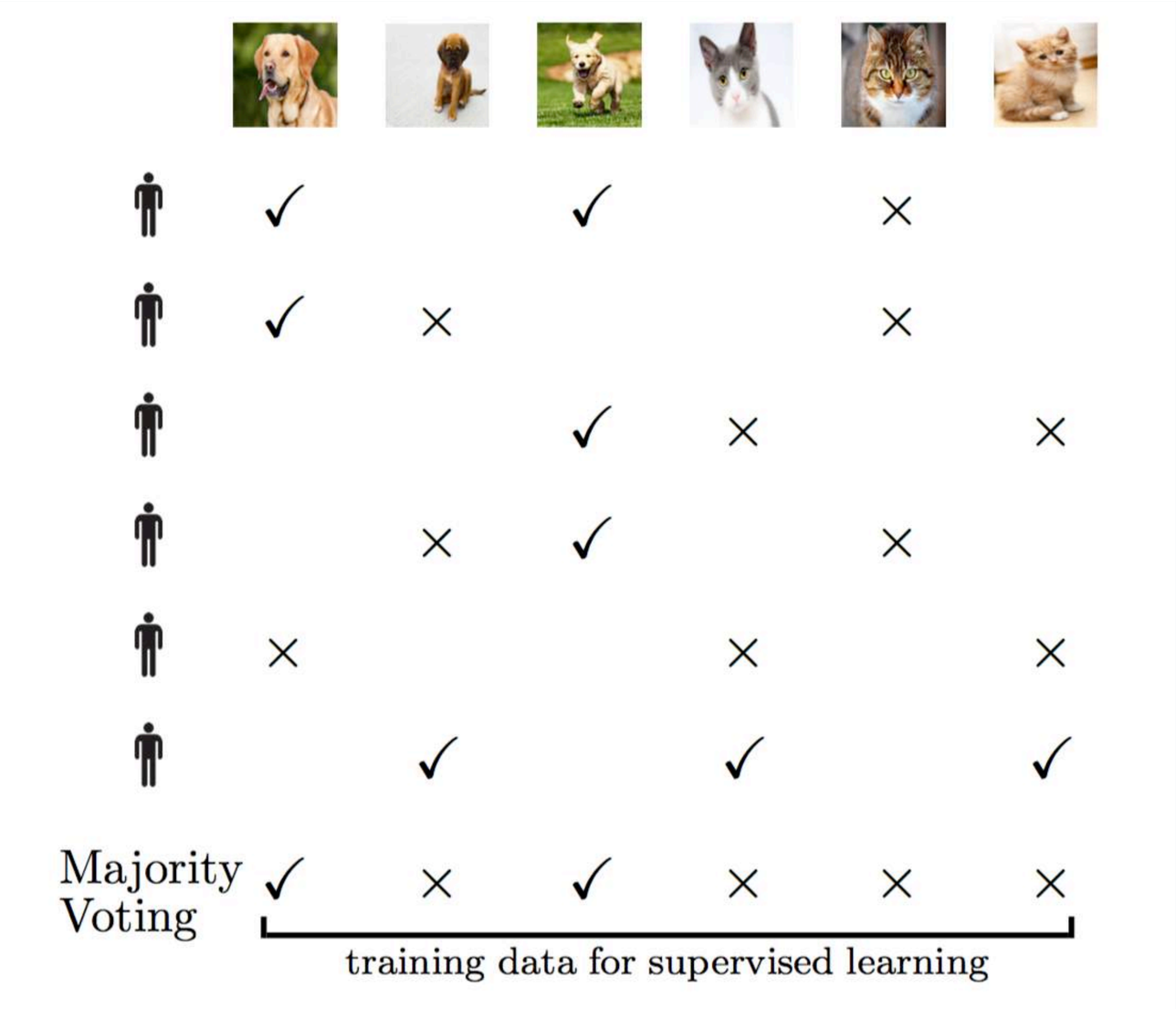
LEARNING FROM NOISY SINGLY-LABELED DATA

ASHISH KHETAN, ZACHARY C. LIPTON, ANIMA ANANDKUMAR

CROWDSOURCING: AGGREGATION OF CROWD ANNOTATIONS

Majority rule

- Simple and common.
- Wasteful: ignores annotator quality of different workers.



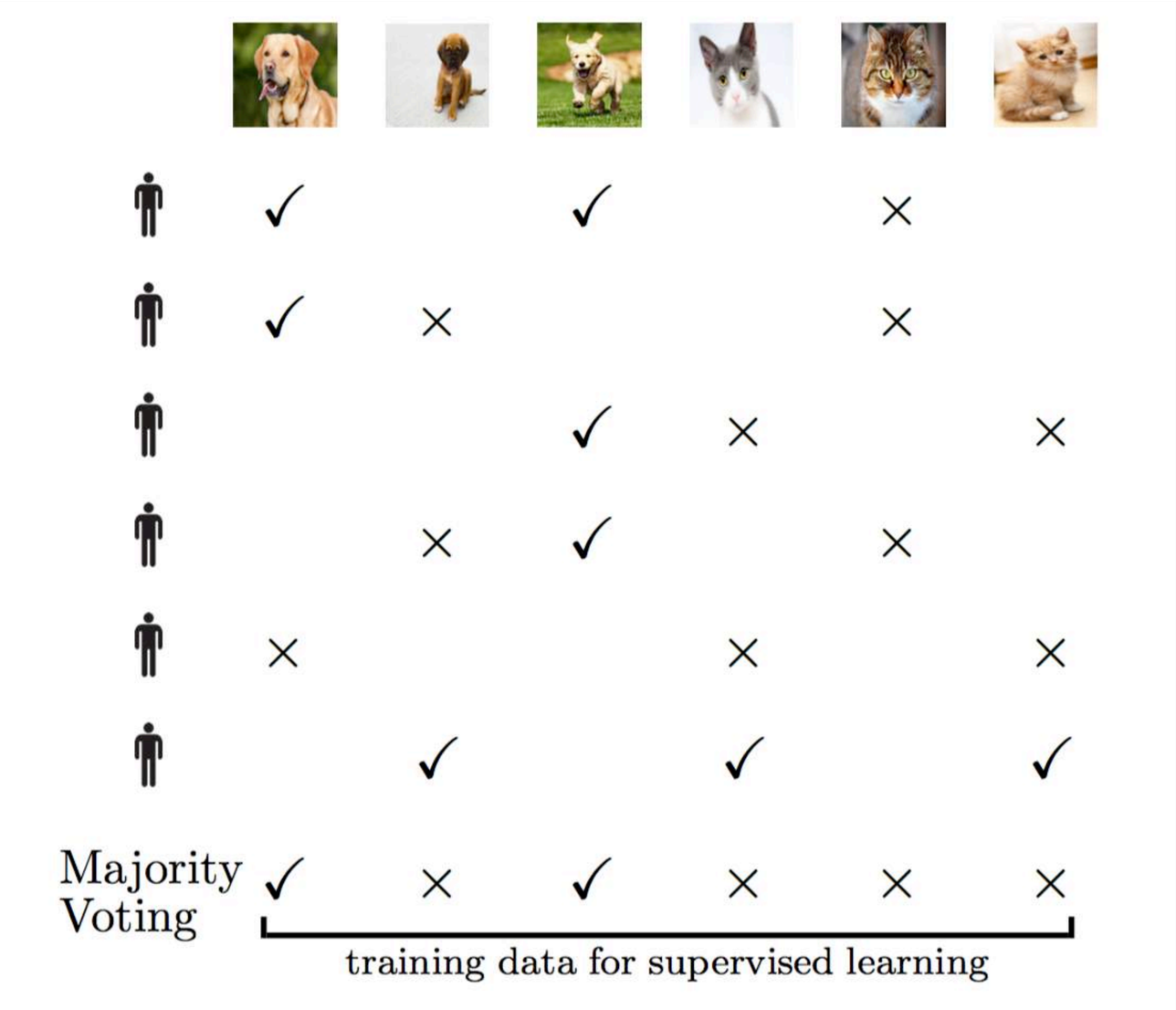
CROWDSOURCING: AGGREGATION OF CROWD ANNOTATIONS

Majority rule

- Simple and common.
- Wasteful: ignores annotator quality of different workers.

Annotator-quality models

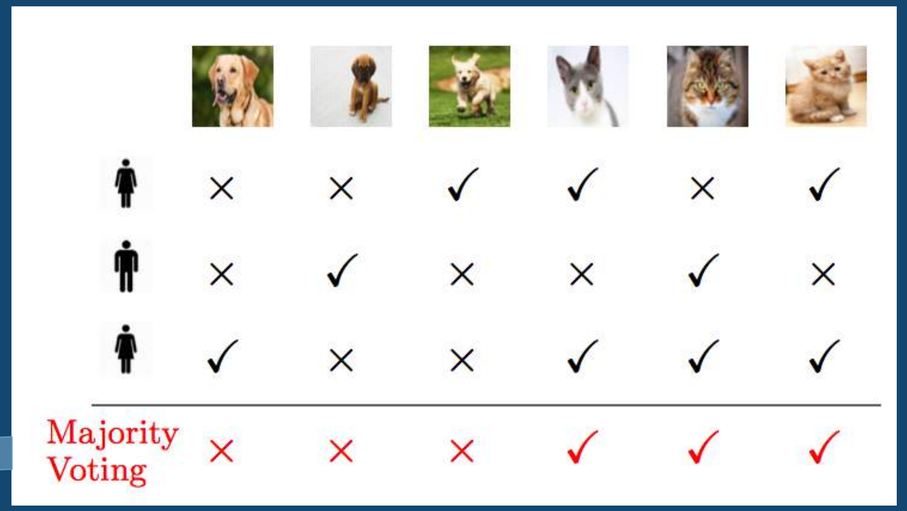
- Can improve accuracy.
- Hard: needs to be estimated without ground-truth.



SOME INTUITIONS

Majority rule to estimate annotator quality

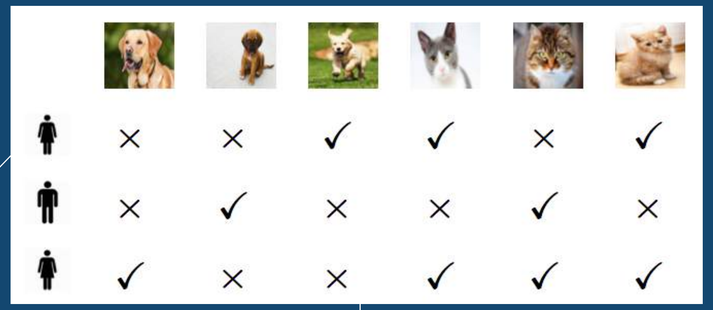
- **Justification:** Majority rule approaches ground-truth when enough workers.
- **Downside:** Requires large number of annotations for each example for majority rule to be correct.



Annotator quality model
(Prob. of correctness)

PROPOSED CROWDSOURCING ALGORITHM

Noisy crowdsourced annotations

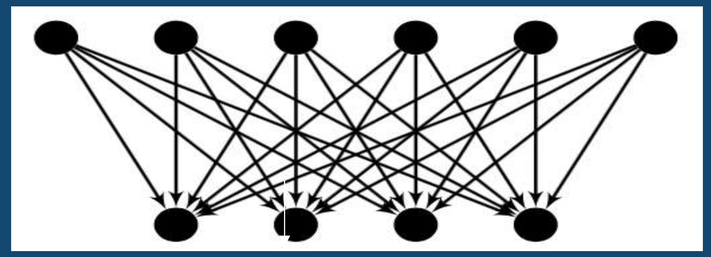


Repeat

Posterior of ground-truth labels given annotator quality model

cat	1/3	1/3	1/3	2/3	2/3	2/3
not cat	2/3	2/3	2/3	1/3	1/3	1/3

Training with weighted loss. Use posterior as weights



Use trained model to infer ground-truth labels

cat	0	0	1	1	1	0
not cat	1	1	0	0	0	1

MLE : update Annotator quality using inferred labels from model

0.7	0.3	0.5

LABELING ONCE IS OPTIMAL: THEORY

Theorem:

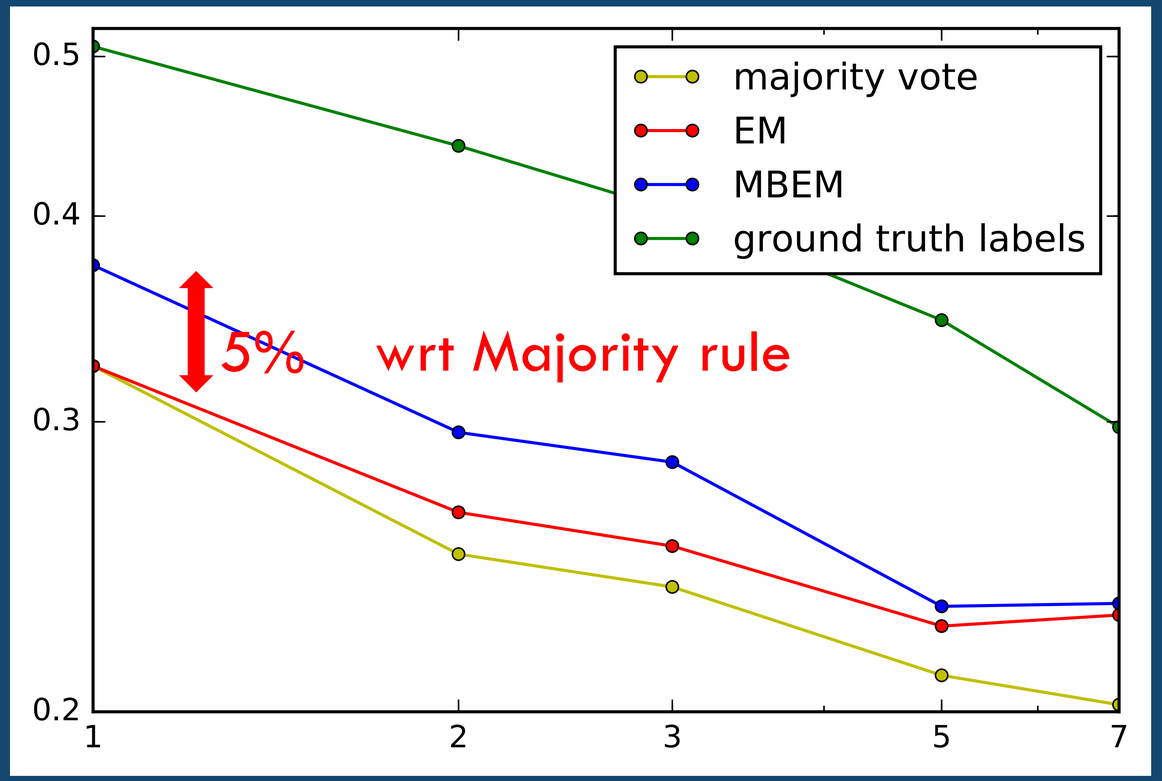
Under fixed budget, generalization error minimized with **single annotation** per sample.

Assumptions:

- Best predictor is accurate enough (under no label noise).
- Simplified case: All workers have same quality.
- Prob. of being correct > 83%

LABELING ONCE IS OPTIMAL: PRACTICE

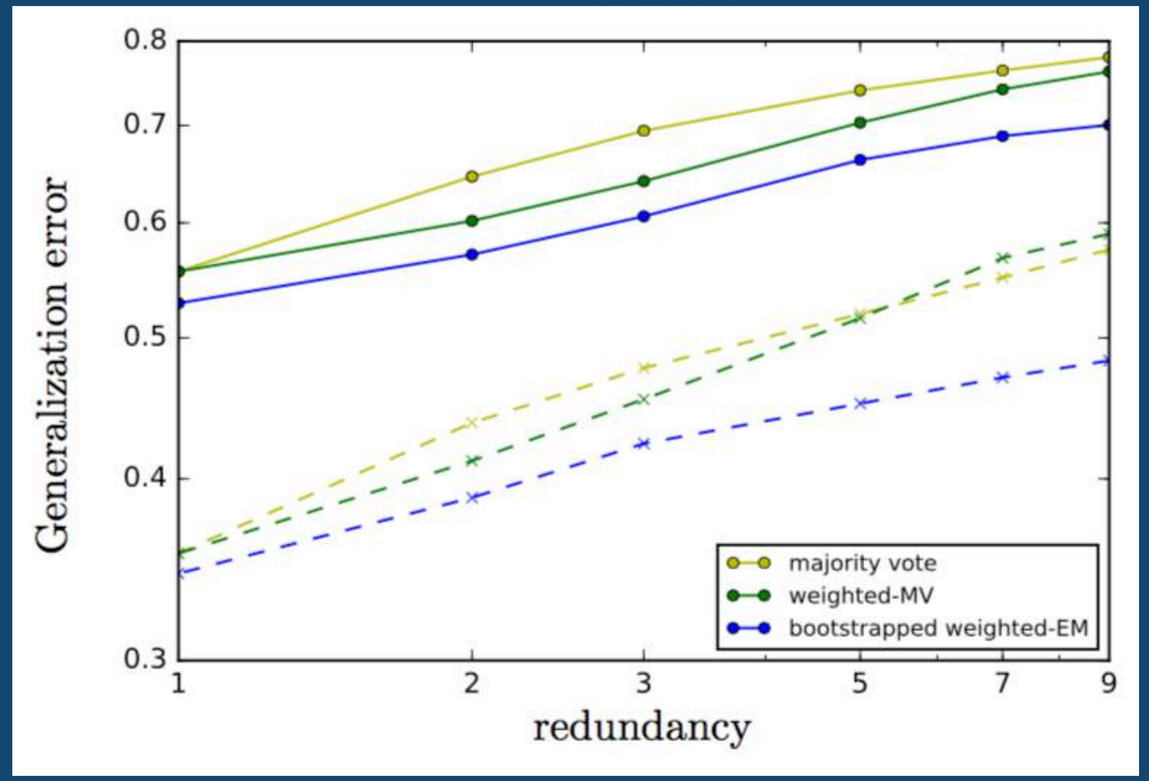
MS-COCO dataset.
Fixed budget: 35k annotations



F1 score

No. of workers

Imagenet dataset.
Simulated workers and fixed budget

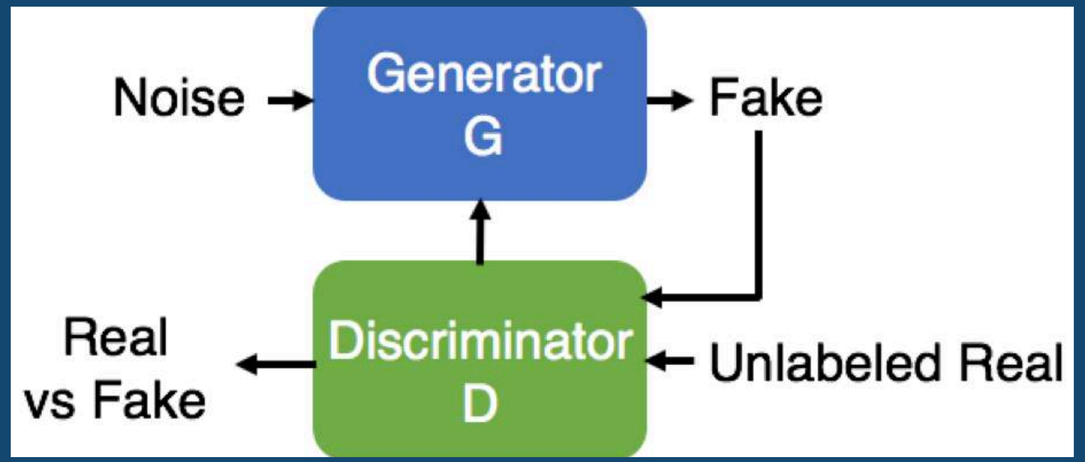


NEURAL RENDERING MODEL (NRM): JOINT GENERATION AND PREDICTION FOR SEMI-SUPERVISED LEARNING

Nhat Ho, Tan Nguyen, Ankit Patel, **A.** , Michael Jordan, Richard Baraniuk

SEMI-SUPERVISED LEARNING WITH GENERATIVE MODELS?

GAN



Merits

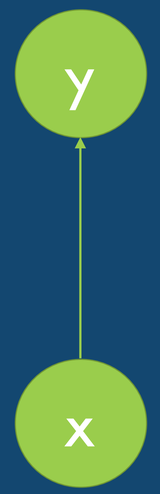
- Captures statistics of natural images
- Learnable

Peril

- Feedback is real vs. fake: different from prediction.
- Introduces artifacts

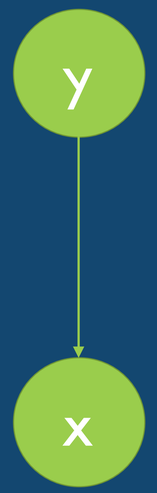
PREDICTIVE VS GENERATIVE MODELS

$P(y \mid x)$



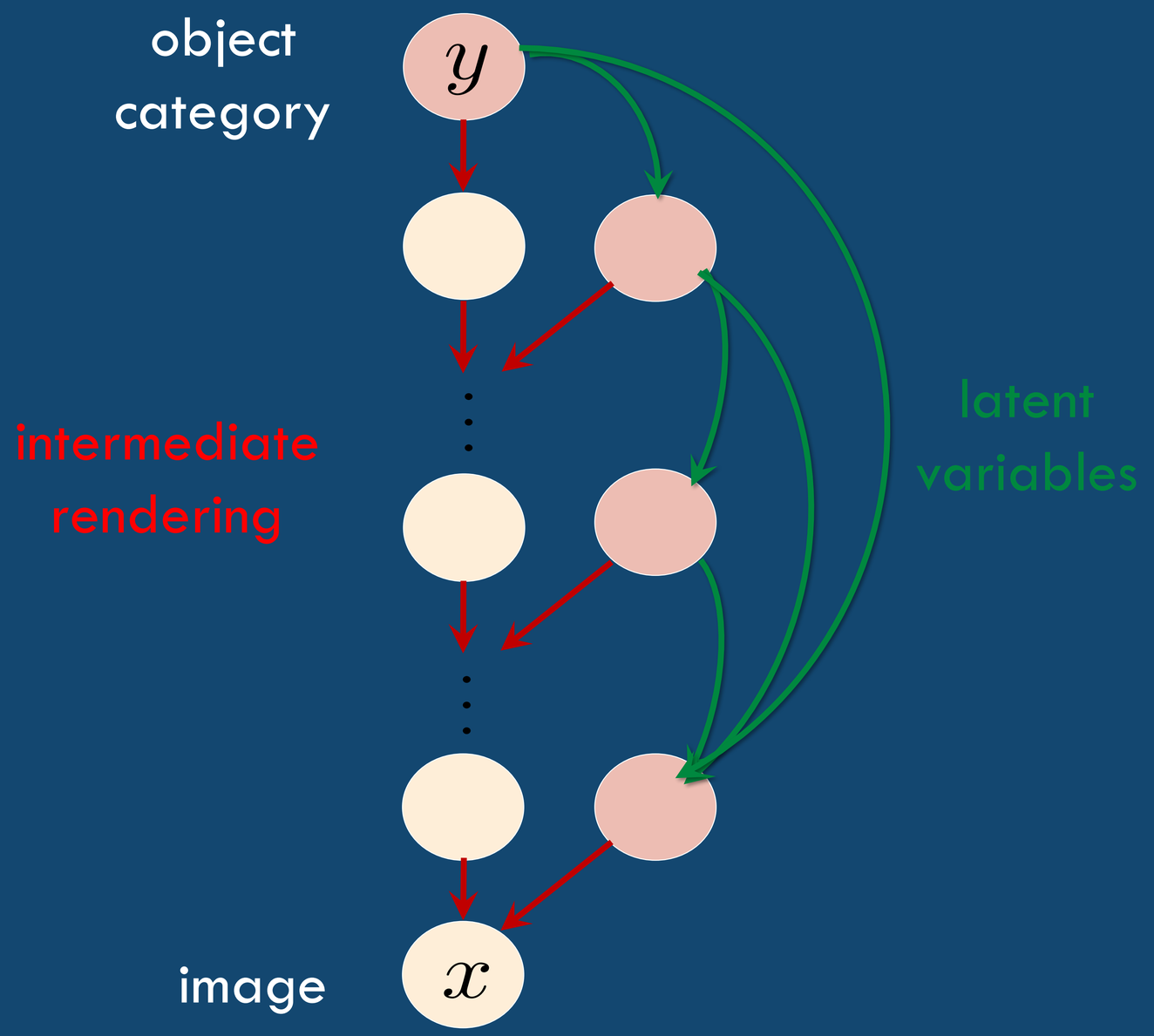
One model to do both?

$P(x \mid y)$



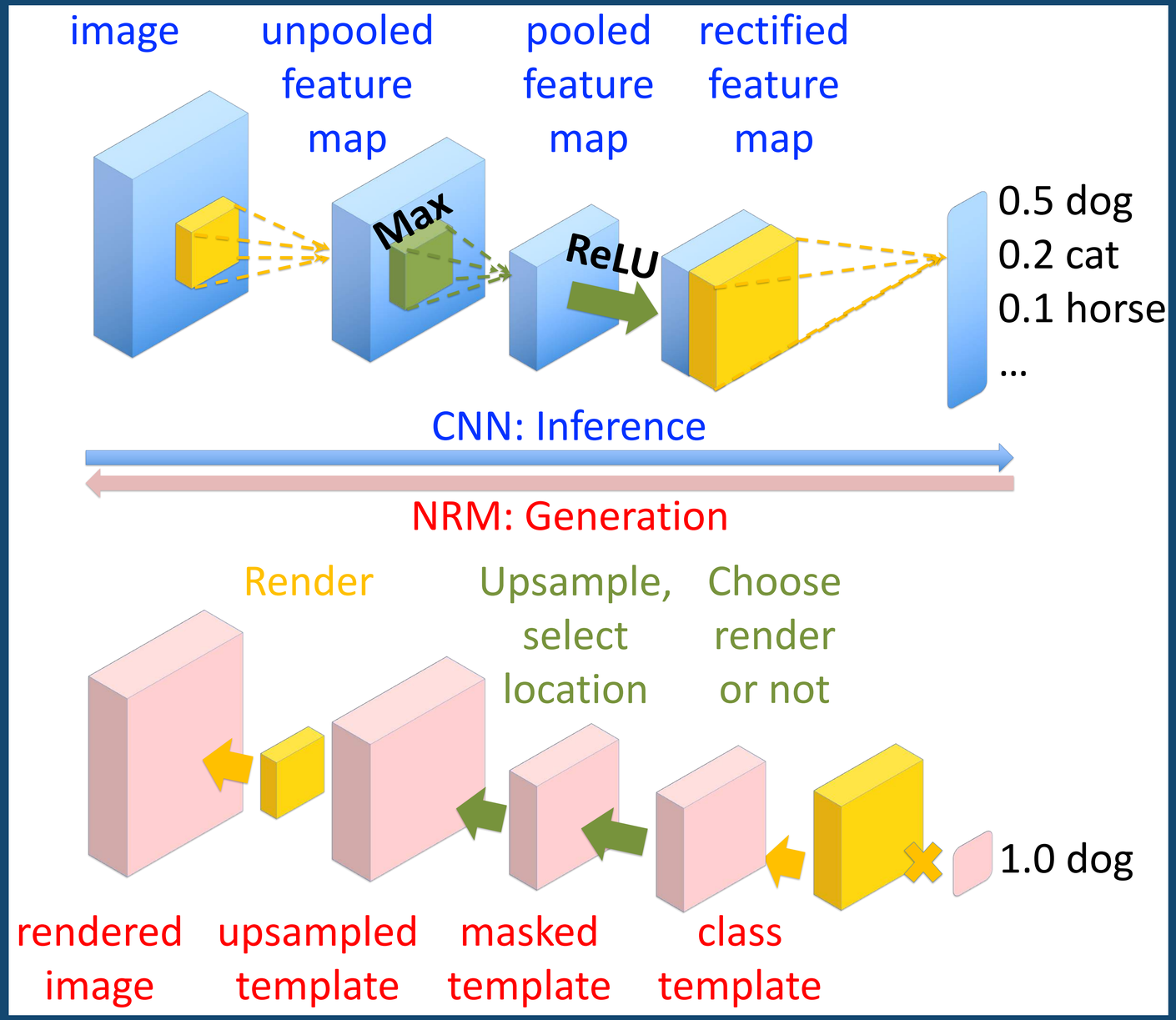
- SOTA prediction from CNN models.
- What class of $p(x|y)$ yield CNN models for $p(y|x)$?

NEURAL DEEP RENDERING MODEL (NRM)



Design joint priors for latent variables based on reverse-engineering CNN predictive architectures

NEURAL RENDERING MODEL (NRM)



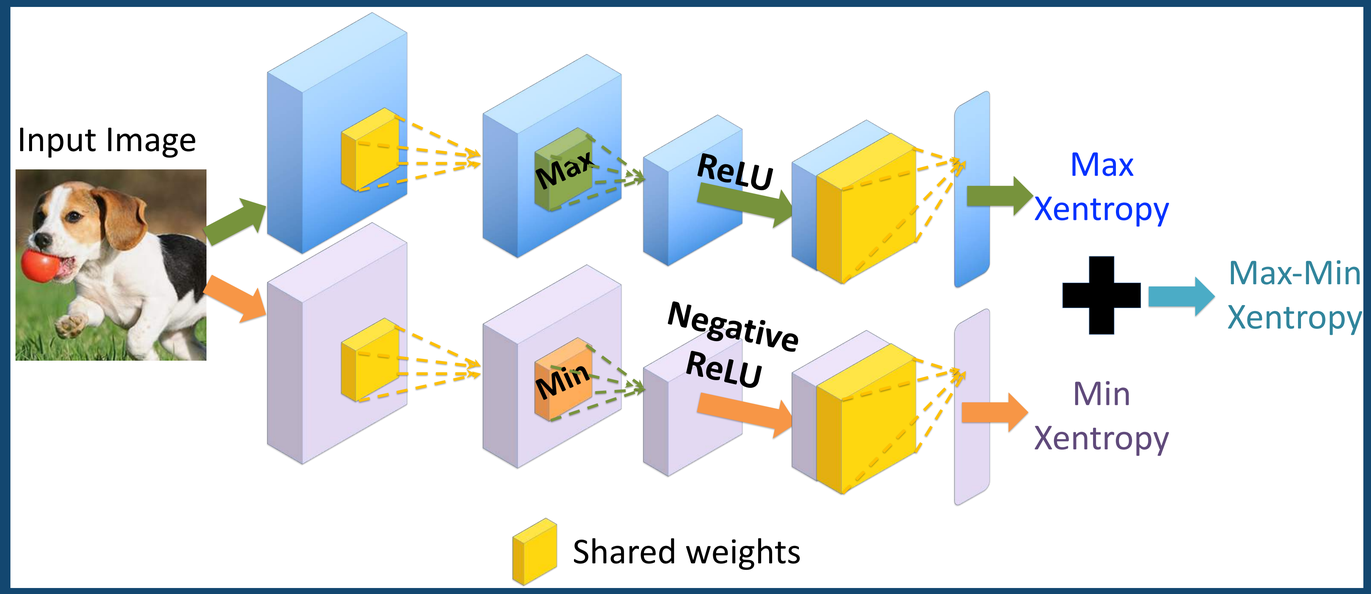
MAX-MIN CROSS-ENTROPY ➡ MAX-MIN NETWORKS

Cross-Entropy Loss for Training the CNNs with Labeled Data

$$\min_{\theta \in \mathcal{A}_\gamma} H_{p,q}(y|x, z^{\max}) \geq \min_{(z_i)_{i=1}^n, \theta} \frac{1}{n} \sum_{i=1}^{n} -\log p(y_i|x_i, z_i; \theta)$$

Max-Min Loss for Training the CNNs with Labeled Data

$$\alpha^{\max} H_{p,q}(y|x, z^{\max}) + \alpha^{\min} H_{p,q}(y|x, z^{\min})$$



- Max cross-entropy maximizes the posteriors of correct labels.
  Min cross-entropy minimizes the posteriors of incorrect labels.
- Co-learning: Max and Min networks try to learn from each other

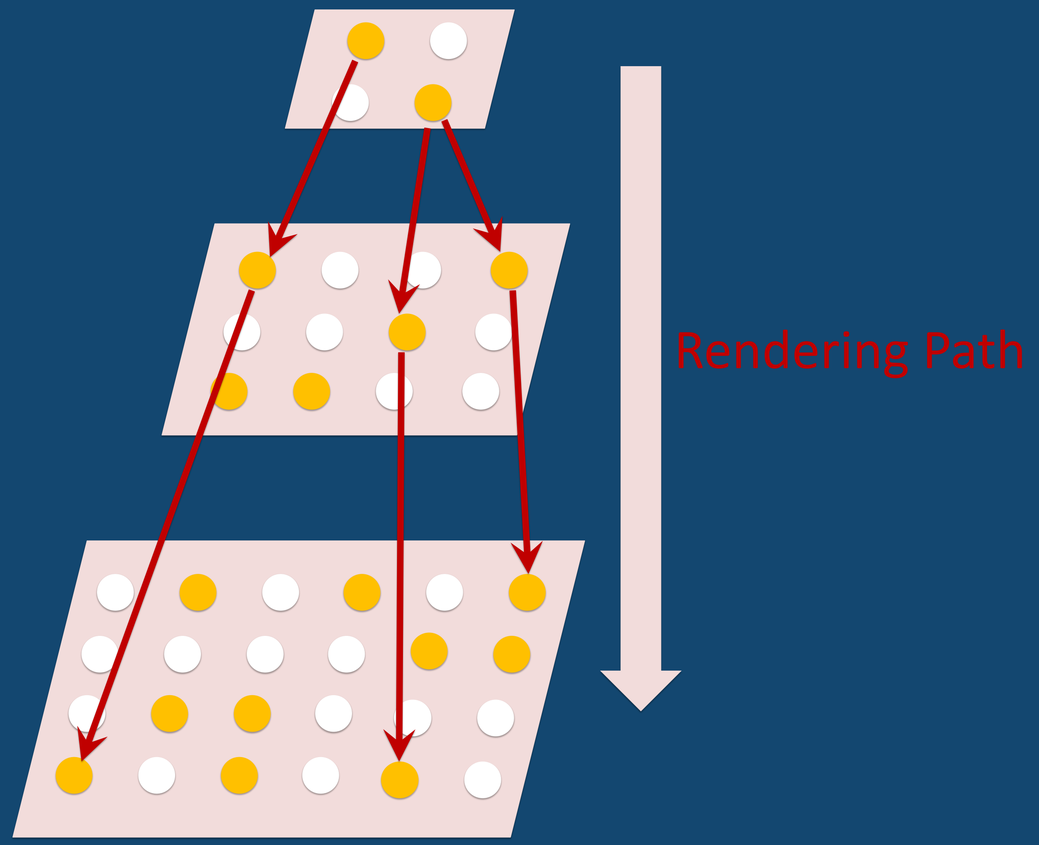
STATISTICAL GUARANTEES FOR THE NRM

Training loss in the CNNs equivalent to likelihood in NRM

Bound on the generalization error

Risk $\leq$

$$\frac{Number\ of\ active\ rendering\ \ paths}{n^{1/2}}$$



- Rendering path normalization:
- new form of regularization

Max-Min NRM with RPN achieves SOTA on benchmarks

EMPIRICAL RESULTS

Semi-Supervised Learning

- NRM + Max-Min improves SOTA by 0.7%-1.8% on CIFAR10, CIFAR100, SVHN.
- Especially in low labeled data setting.

Supervised Learning

- Max-Min improves SOTA on CIFAR10 by 0.26% and on ImageNet by 0.17% (top 5 error)

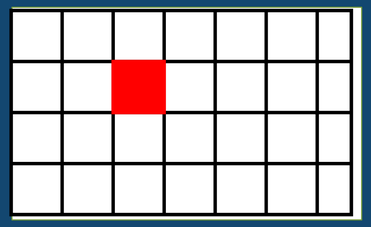
Max-Min NRM achieves SOTA on semi-supervised & supervised learning

TENSOR METHODS

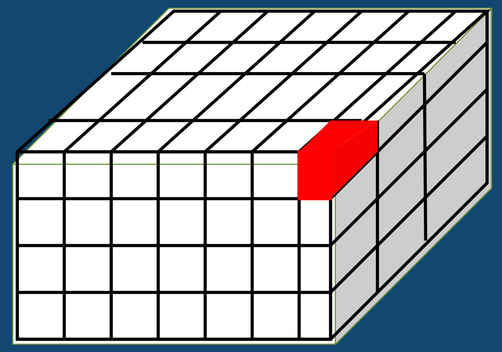
TENSORS FOR MULTI-DIMENSIONAL DATA AND HIGHER ORDER MOMENTS

Images: 3 dimensions

Videos: 4 dimensions



Pairwise correlations



Triplet correlations

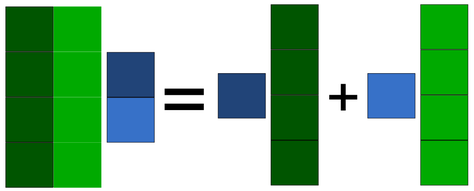
OPERATIONS ON TENSORS: TENSOR CONTRACTION

Tensor Contraction

Extends the notion of matrix product

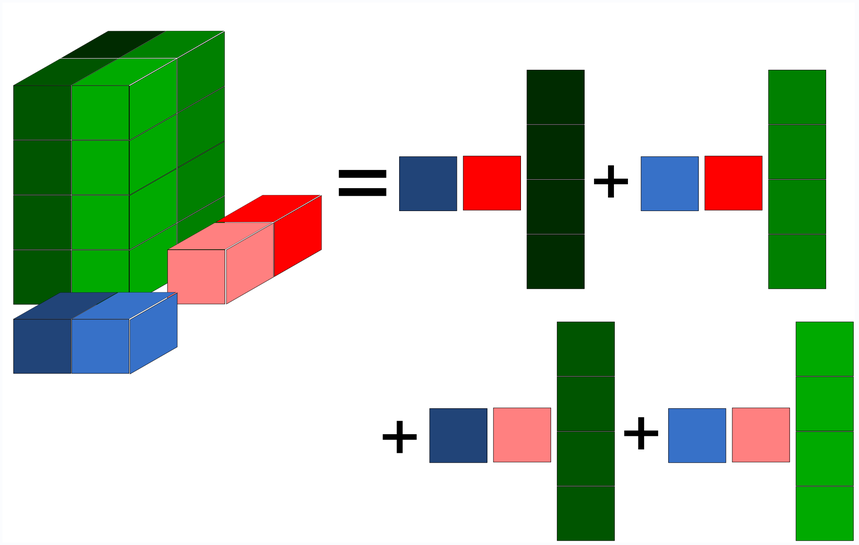
Matrix product

$$Mv = \sum_j v_j M_j$$



Tensor Contraction

$$T(u, v, \cdot) = \sum_{i,j} u_i v_j T_{i,j,:}$$



UNSUPERVISED LEARNING OF TOPIC MODELS THROUGH TENSOR METHODS

Topics

- Justice
- Education
- Sports

SECTIONS HOME SEARCH The New York Times

COLLEGE FOOTBALL

At Florida State, Football Clouds Justice

By MIKE McINTIRE and WALT BOGDANICH OCT. 10, 2014

Now, an examination by The New York Times of police and court records, along with interviews with crime witnesses, has found that, far from an aberration, the treatment of the Winston complaint was in keeping with the way the police on numerous occasions have soft-pedaled allegations of wrongdoing by Seminoles football players. From criminal mischief and motor-vehicle theft to domestic violence, arrests have been avoided, investigations have stalled and players have escaped serious consequences.

In a community whose self-image and economic well-being are so tightly bound to the fortunes of the nation's top-ranked college football team, law enforcement officers are finely attuned to a suspect's football connections. Those ties are cited repeatedly in police reports examined by The Times. What's more, dozens of officers work second jobs directing traffic and providing security at home football games, and many express their devotion to the Seminoles on social media.

TMZ, the gossip website, also requested the police report and later asked the school's deputy police chief, Jim L. Russell, if the campus police had interviewed Mr. Winston about the rape report. Mr. Russell responded by saying his officers were not investigating the case, omitting any reference to the city police, even though the campus police knew of their involvement. "Thank you for contacting me regarding this rumor — I am glad I can dispel that one!" Mr. Russell told TMZ in an email. The university said Mr. Russell was unaware of any other police investigation at the time of the inquiry. Soon after, the Tallahassee police belatedly sent their files to the news media and to the prosecutor, William N. Meggs. By then critical evidence had been lost and Mr. Meggs, who criticized the police's handling of the case, declined to

lson after the Seminoles' first game; five
am's second-leading receiver.

On Jan. 10, 2013, a female student at Florida State spotted the man she believed had raped her the previous month. After learning his name, Jameis Winston, she reported him to the Tallahassee police.

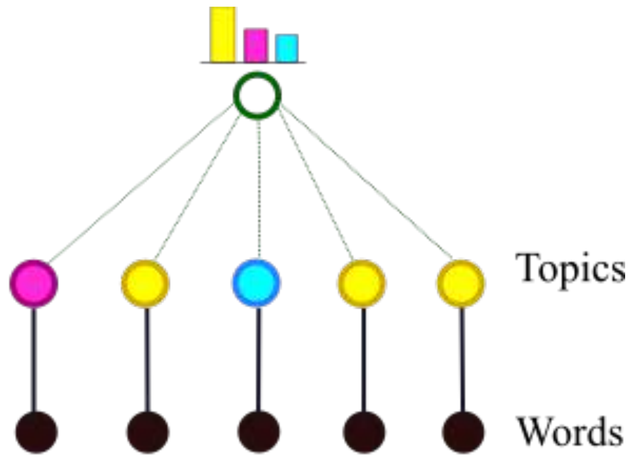
In the 21 months since, Florida State officials have said little about how they handled the case, which is no investigated by the federal Depart

Most recently, university officials suspended Mr. Winston for one game after he stood in a public place on campus and, playing off a running Internet gag, shouted a crude reference to a sex act. In a news conference afterward, his coach, Jimbo Fisher, said, "Our hope and belief is Jameis will learn from this and use better judgment and language and decision-making."

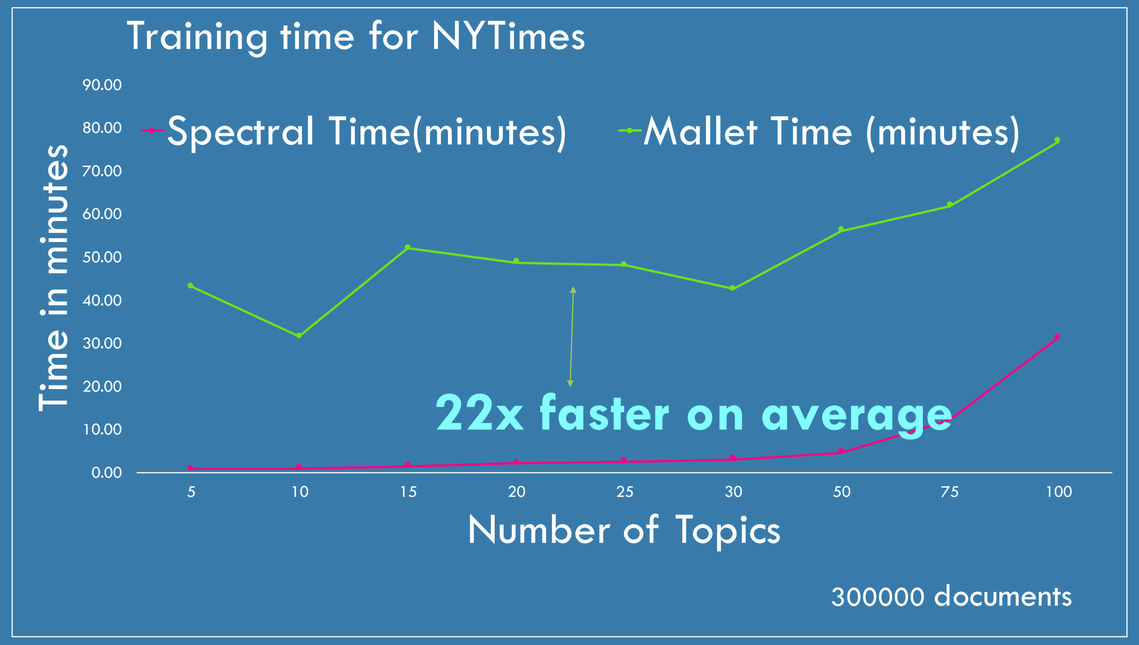
As The Times reported last April, the Tallahassee police also failed to aggressively investigate the rape accusation. It did not become public until November, when a Tampa reporter, Matt Baker, acting on a tip, sought records of the police investigation.

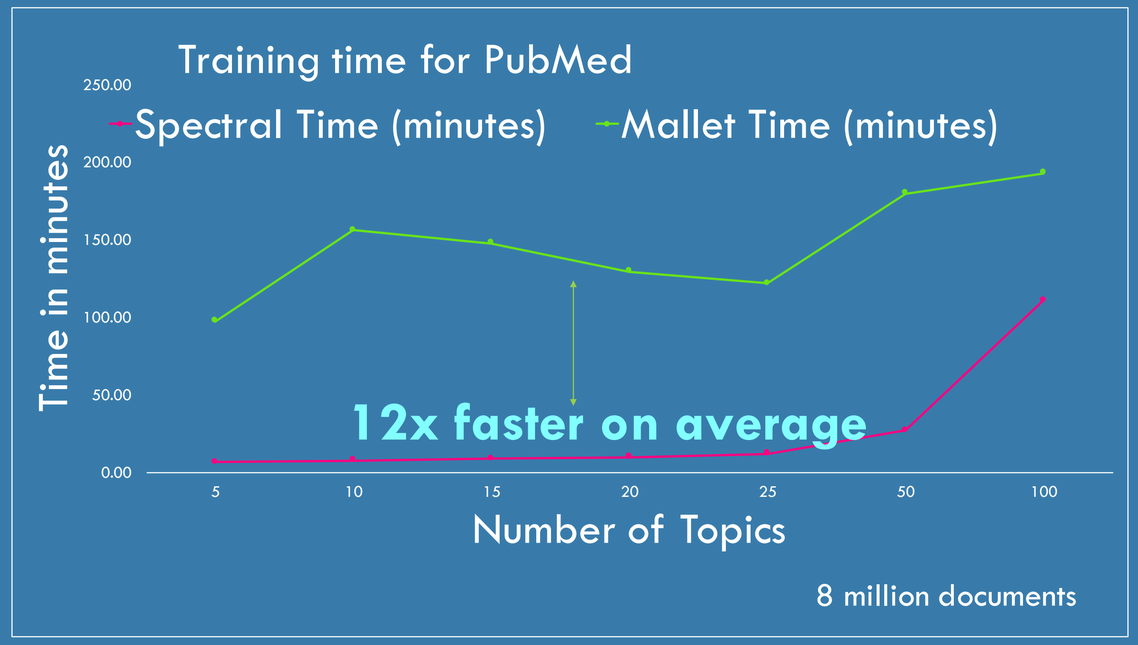
Upon learning of Mr. Baker's inquiry, Florida State, having shown little curiosity about the rape accusation, suddenly took a keen interest in the journalist seeking to report it, according to emails obtained by The Times.

"Can you share any details on the requesting source?" David Perry, the university's police chief, asked the Tallahassee police. Several hours later, Mr.



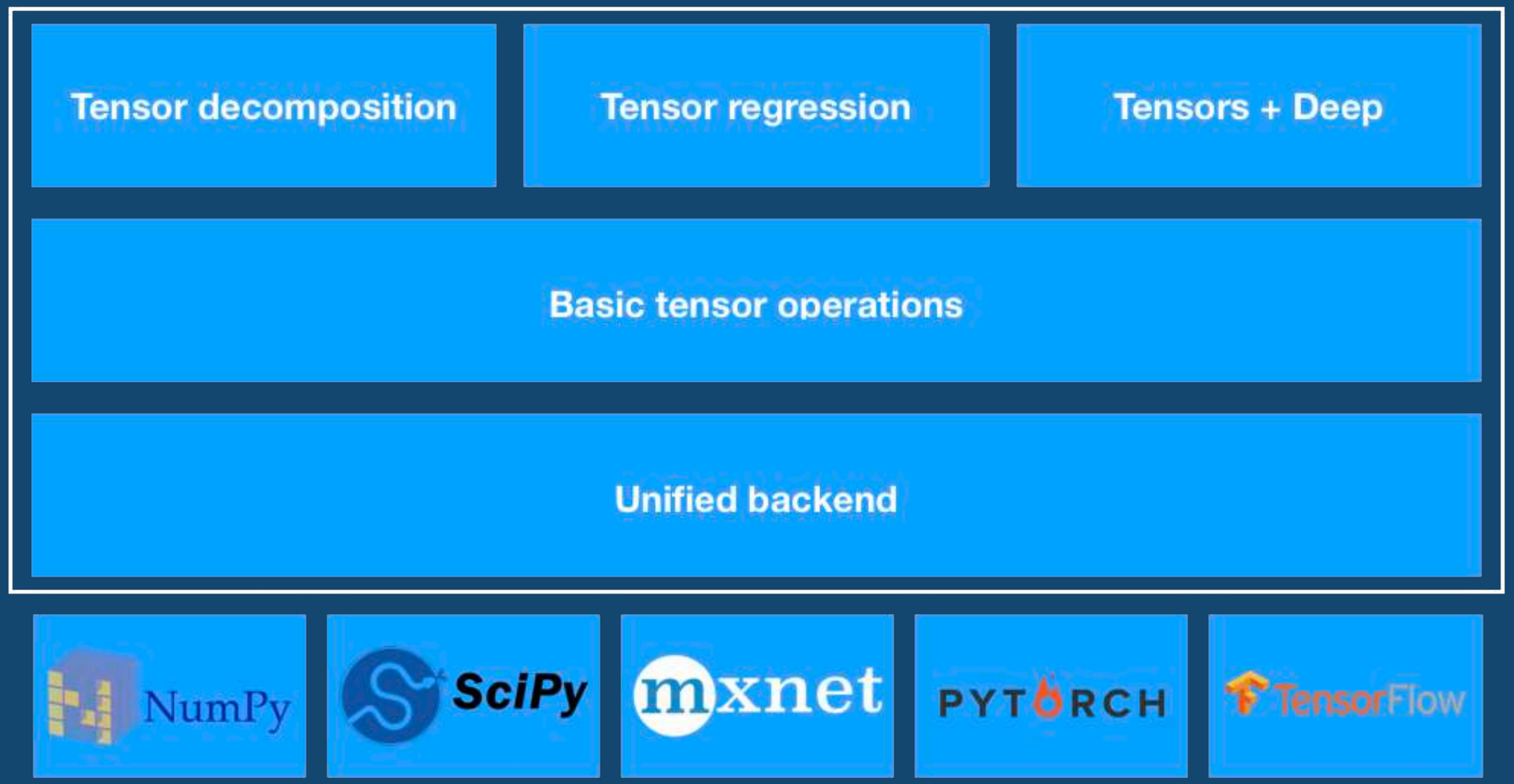
TENSOR-BASED LDA TRAINING IS FASTER



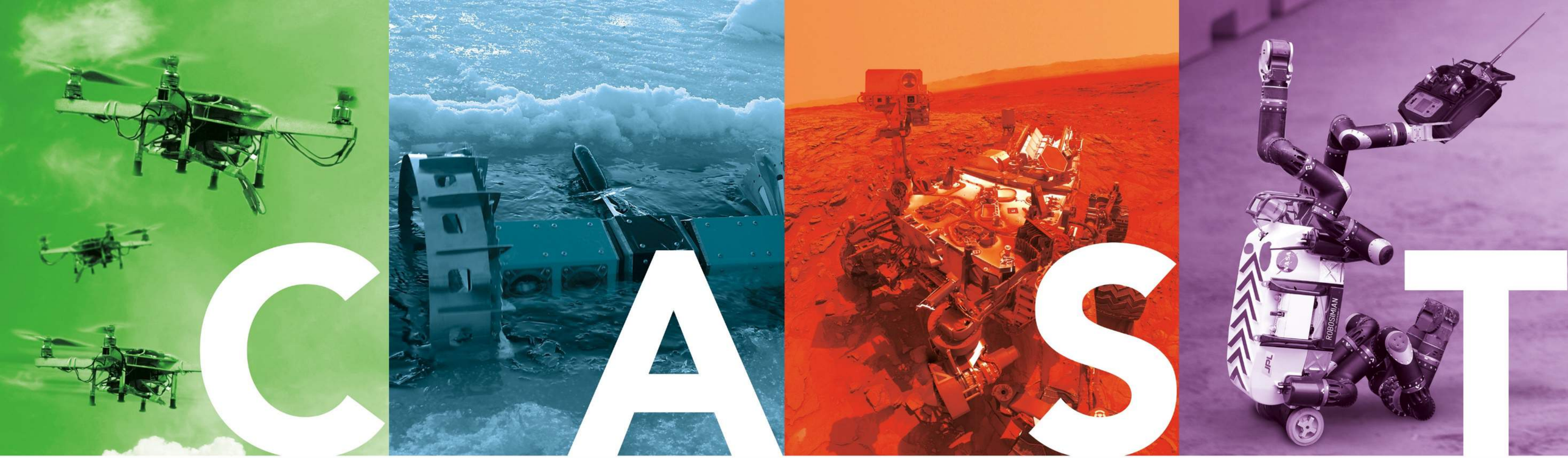


- Mallet is an open-source framework for topic modeling
- Benchmarks on AWS SageMaker Platform
- Bulit into AWS Comprehend NLP service.

TENSORLY: HIGH-LEVEL API FOR TENSOR ALGEBRA



- Python programming
- User-friendly API
- Multiple backends: flexible + scalable
- Example notebooks in repository



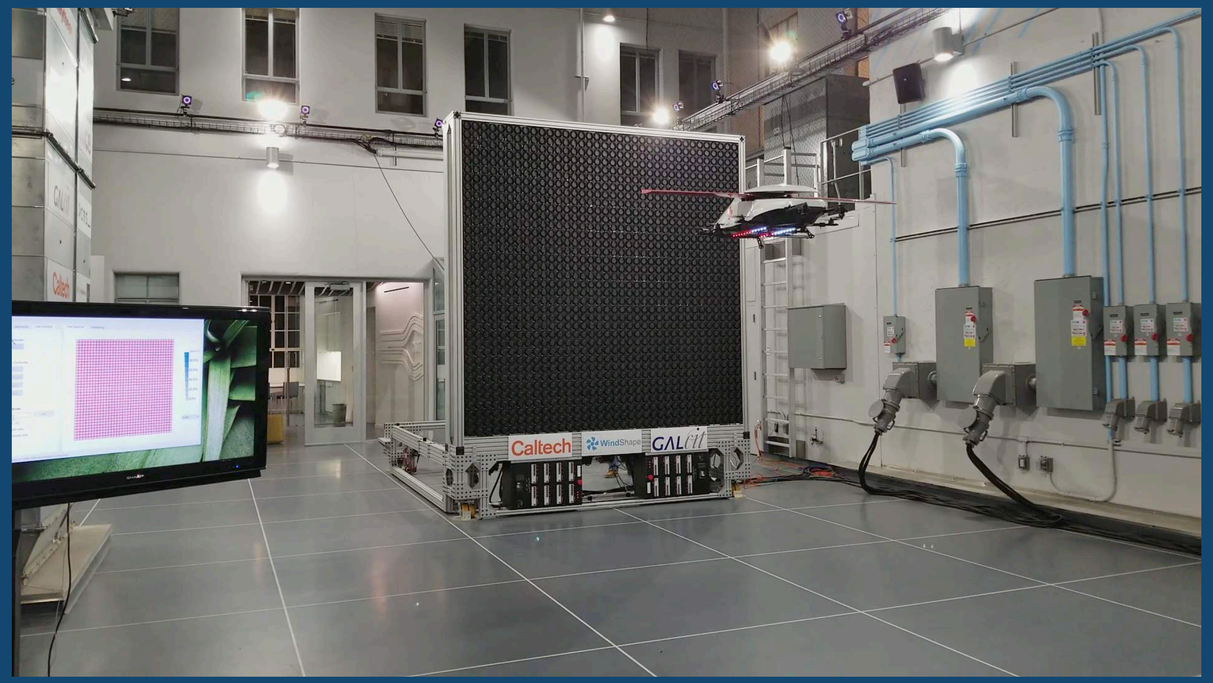
Center for Autonomous Systems and Technologies

A New Vision for Autonomy

Caltech

CHALLENGES IN LANDING A QUADROTOR DRONE

- Unknown aerodynamic forces & moments.
- *Example 1*: when drone is close to ground.
- *Example 2*: as velocity goes up, air drag.
- *Example 3*: external wind conditions.



Wind generation in
CALTECH CAST wind tunnel

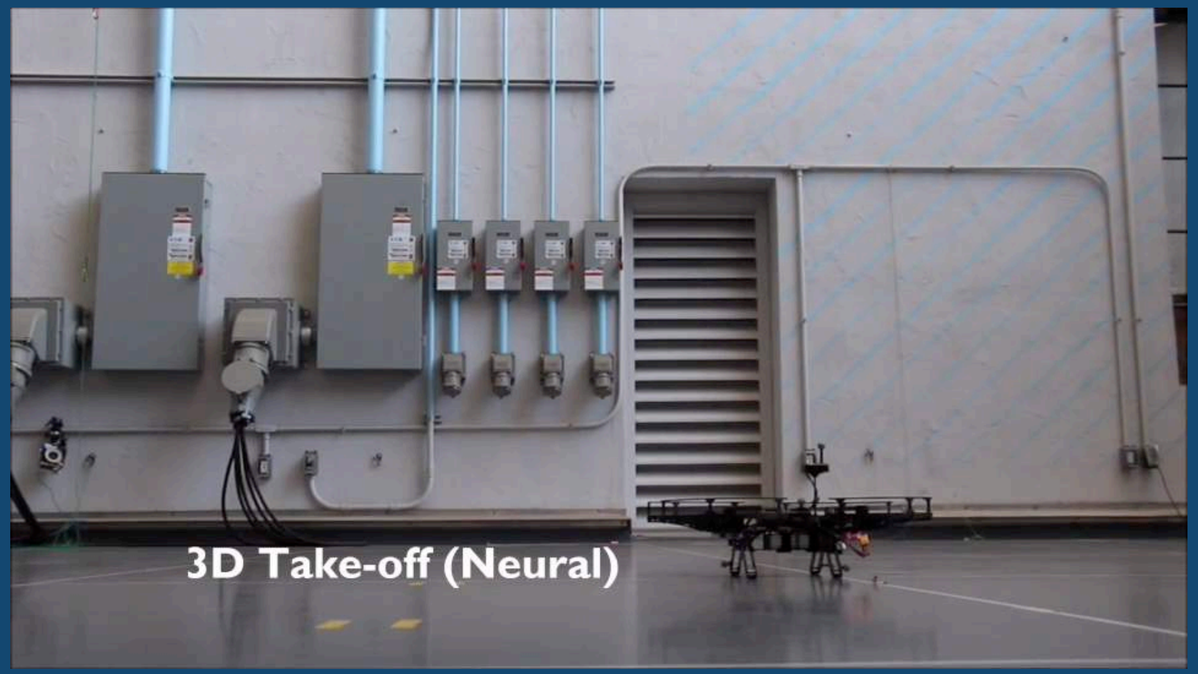
CHALLENGES IN USING DNNS TO LEARN UNKNOWN DYNAMICS

- Our idea: use DNNs to learn unknown aerodynamic forces and then design nonlinear controller to cancel it (unknown moments are very limited in landing)

- *Challenge 1*: DNNs are data-hungry
- *Challenge 2*: DNNs can be unstable and generate unpredictable output
- *Challenge 3*: DNNs are difficult to analyze and design provably stable controller based on them

- **Our approach:** using **Spectral Normalization** to control Lipschitz property of DNNs and then design stable nonlinear controller (**Neural-Lander**)

FIRST SET OF RESULTS: LEARNING TO LAND




NVIDIA

SOME RESEARCH LEADERS AT NVIDIA

Chief Scientist

Bill Dally

Graphics

Dave Luebke

Alex Keller

Aaron Lefohn

Learning & Perception

Jan Kautz

Robotics

Dieter Fox

Computer vision

Sanja Fidler

Core ML

Me !

Applied research

Bryan Catanzaro

Programming

Michael Garland

Networks

Larry Dennison

Architecture

Steve Keckler

Dave Nellans

Mike O'Connor

VLSI

Brucek Khailany

Circuits

Tom Gray

TRINITY FUELING ARTIFICIAL INTELLIGENCE

ALGORITHMS

- OPTIMIZATION
- SCALABILITY
- MULTI-DIMENSIONALITY

INFRASTRUCTURE

FULL STACK FOR ML

- APPLICATION SERVICES
- ML PLATFORM
- GPUS

DATA

- COLLECTION
- AGGREGATION
- AUGMENTATION

COLLABORATORS (LIMITED LIST)

Thank you