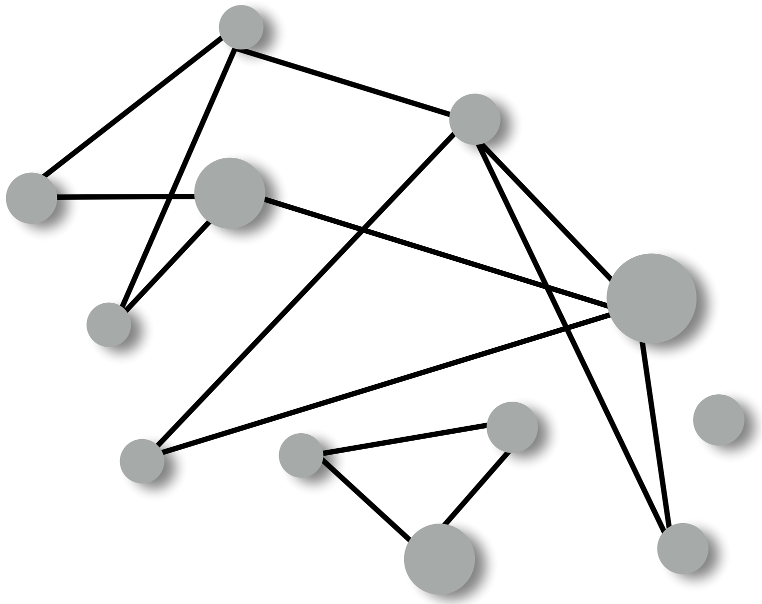


MOCHA: Federated Multi-Task Learning

NIPS ‘17

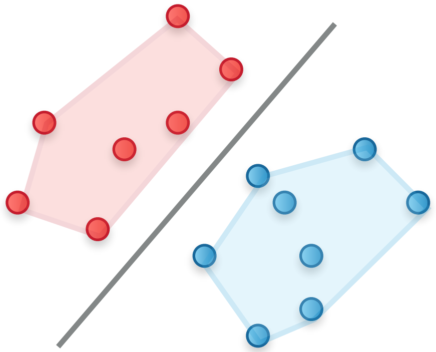
Virginia Smith
Stanford / CMU

Chao-Kai Chiang · *USC*
Maziar Sanjabi · *USC*
Ameet Talwalkar · *CMU*

MACHINE LEARNING WORKFLOW

data & problem

machine learning model



optimization algorithm

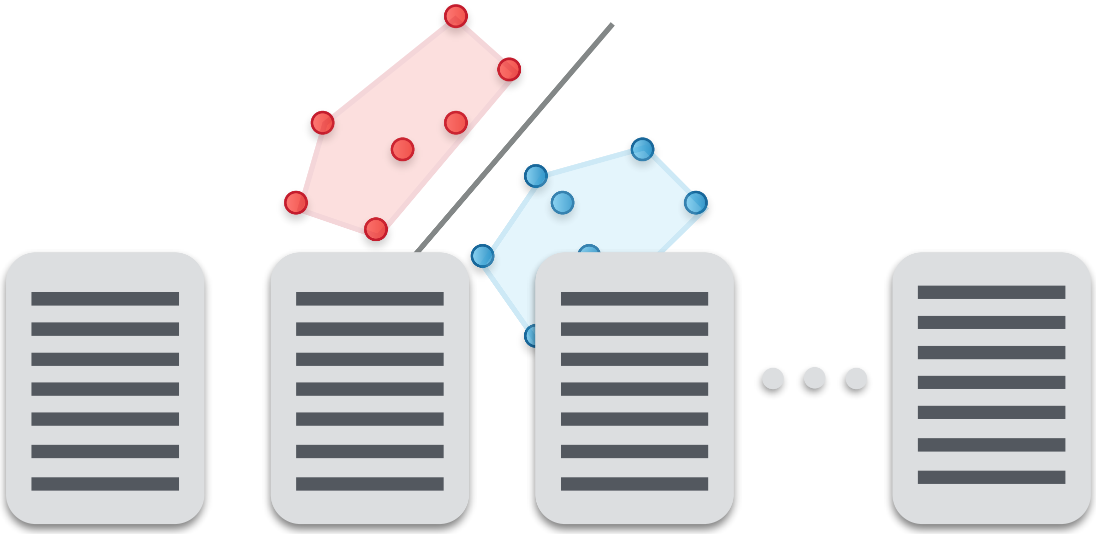
$$\min_{\mathbf{w}} \sum_{i=1}^{n} \ell(\mathbf{w}, x_i) + g(\mathbf{w})$$

MACHINE LEARNING WORKFLOW ^ IN PRACTICE

data & problem

machine learning model

systems setting



optimization algorithm

$$\min_{\mathbf{w}} \sum_{i=1}^{n} \ell(\mathbf{w}, x_i) + g(\mathbf{w})$$

how can we perform fast distributed optimization?

google.com/datacenters

BEYOND THE DATACENTER

- Massively Distributed
- Node Heterogeneity
- Unbalanced
- Non-IID
- Underlying Structure

BEYOND THE DATACENTER

- Massively Distributed
- Node Heterogeneity

Systems Challenges

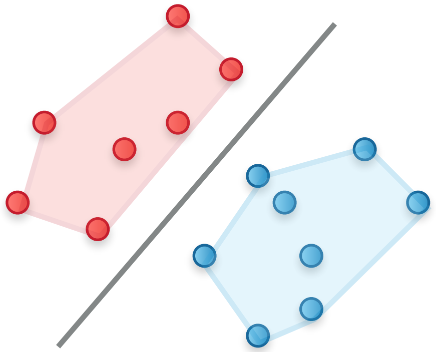
- Unbalanced
- Non-IID
- Underlying Structure

Statistical Challenges

MACHINE LEARNING WORKFLOW ^ IN PRACTICE

data & problem

machine learning model



systems setting

optimization algorithm

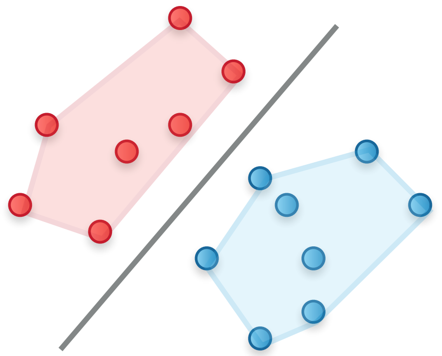
$$\min_{\mathbf{w}} \sum_{i=1}^{n} \ell(\mathbf{w}, x_i) + g(\mathbf{w})$$

MACHINE LEARNING WORKFLOW ^ IN PRACTICE

data & problem

systems setting

machine learning model



optimization algorithm

$$\min_{\mathbf{w}} \sum_{i=1}^{n} \ell(\mathbf{w}, x_i) + g(\mathbf{w})$$

OUTLINE

- Unbalanced
- Non-IID
- Underlying Structure

Statistical Challenges

- Massively Distributed
- Node Heterogeneity

Systems Challenges

OUTLINE

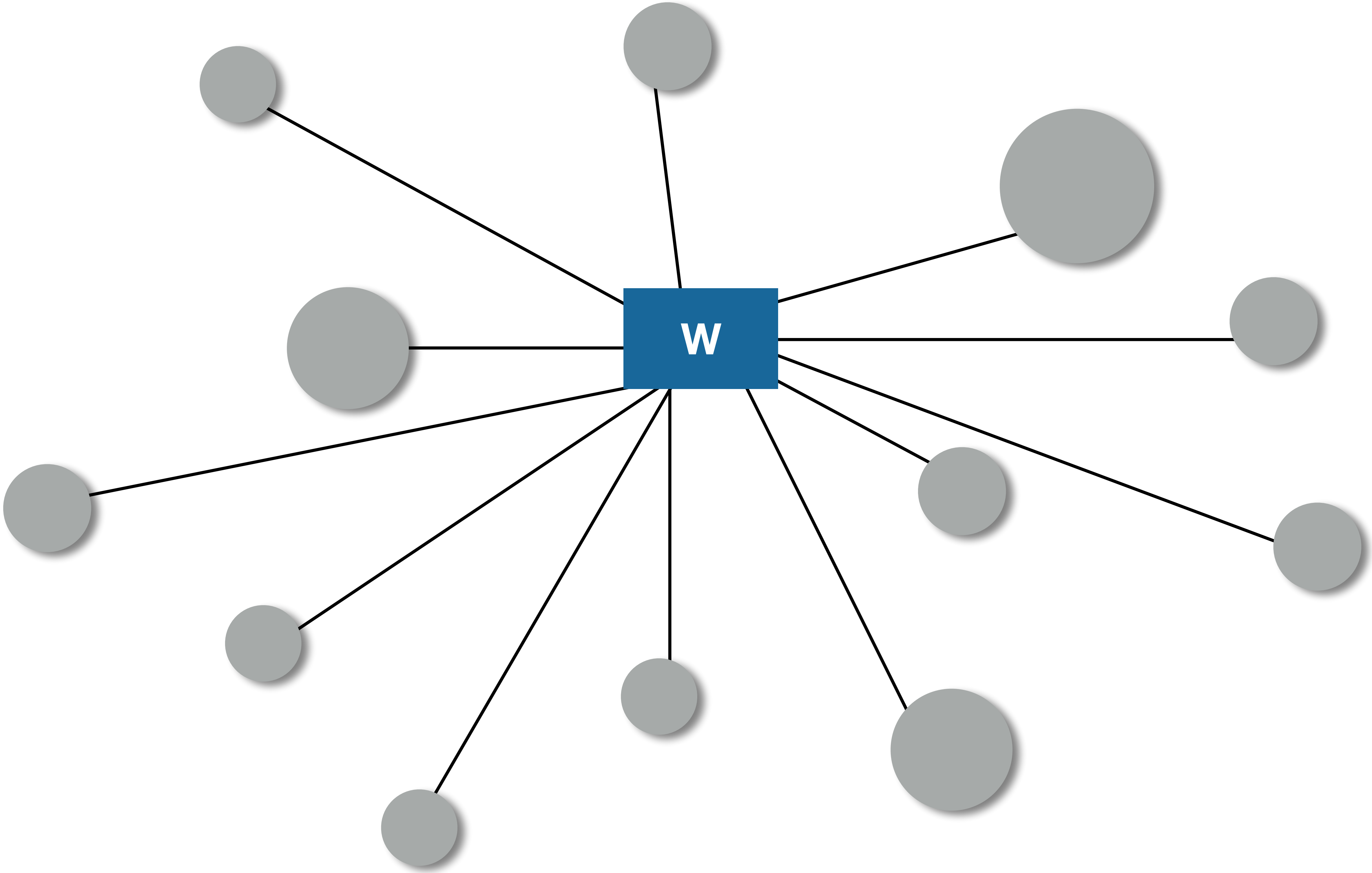
- Unbalanced
- Non-IID
- Underlying Structure

Statistical Challenges

- Massively Distributed
- Node Heterogeneity

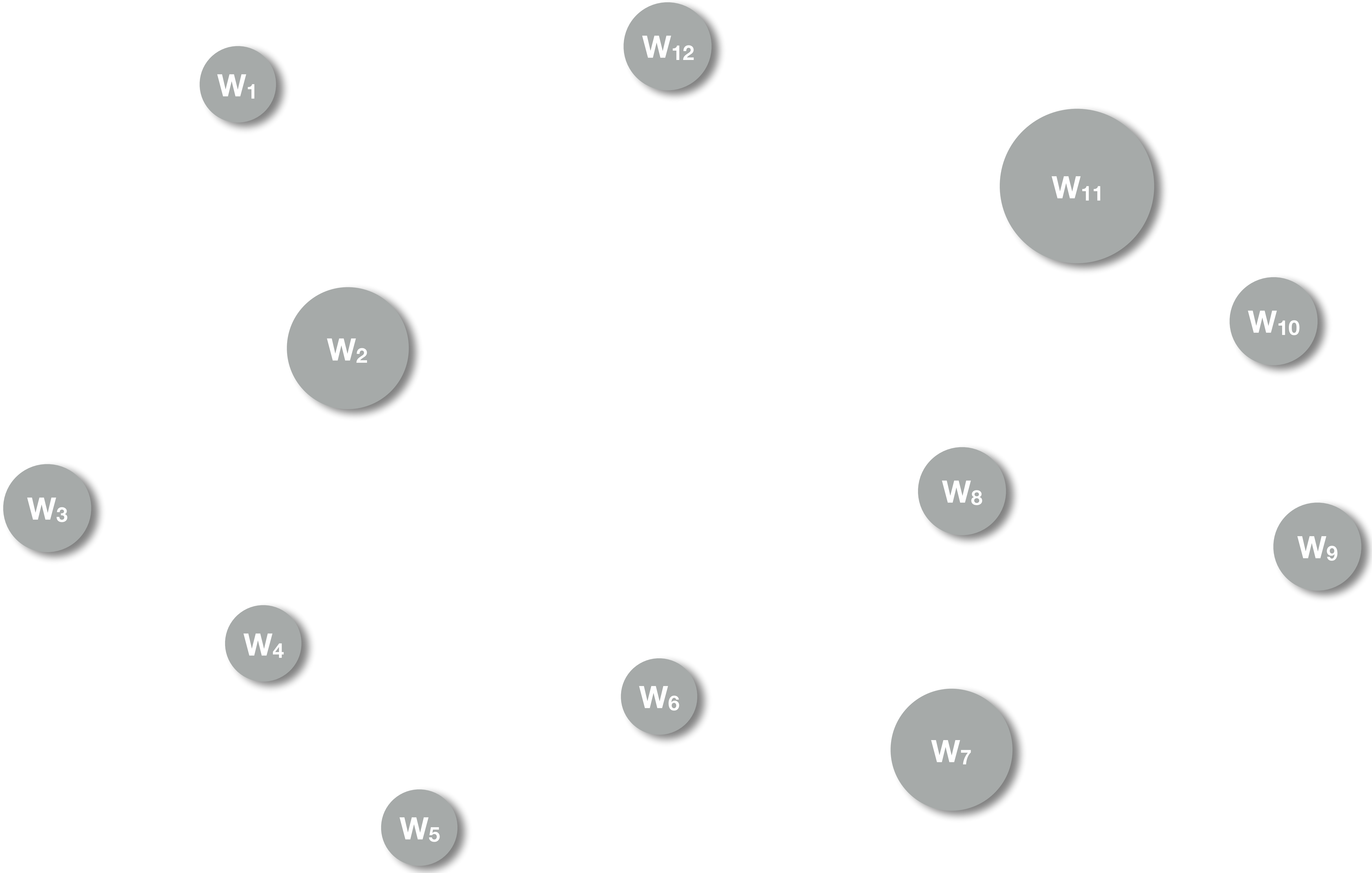
Systems Challenges

A GLOBAL APPROACH

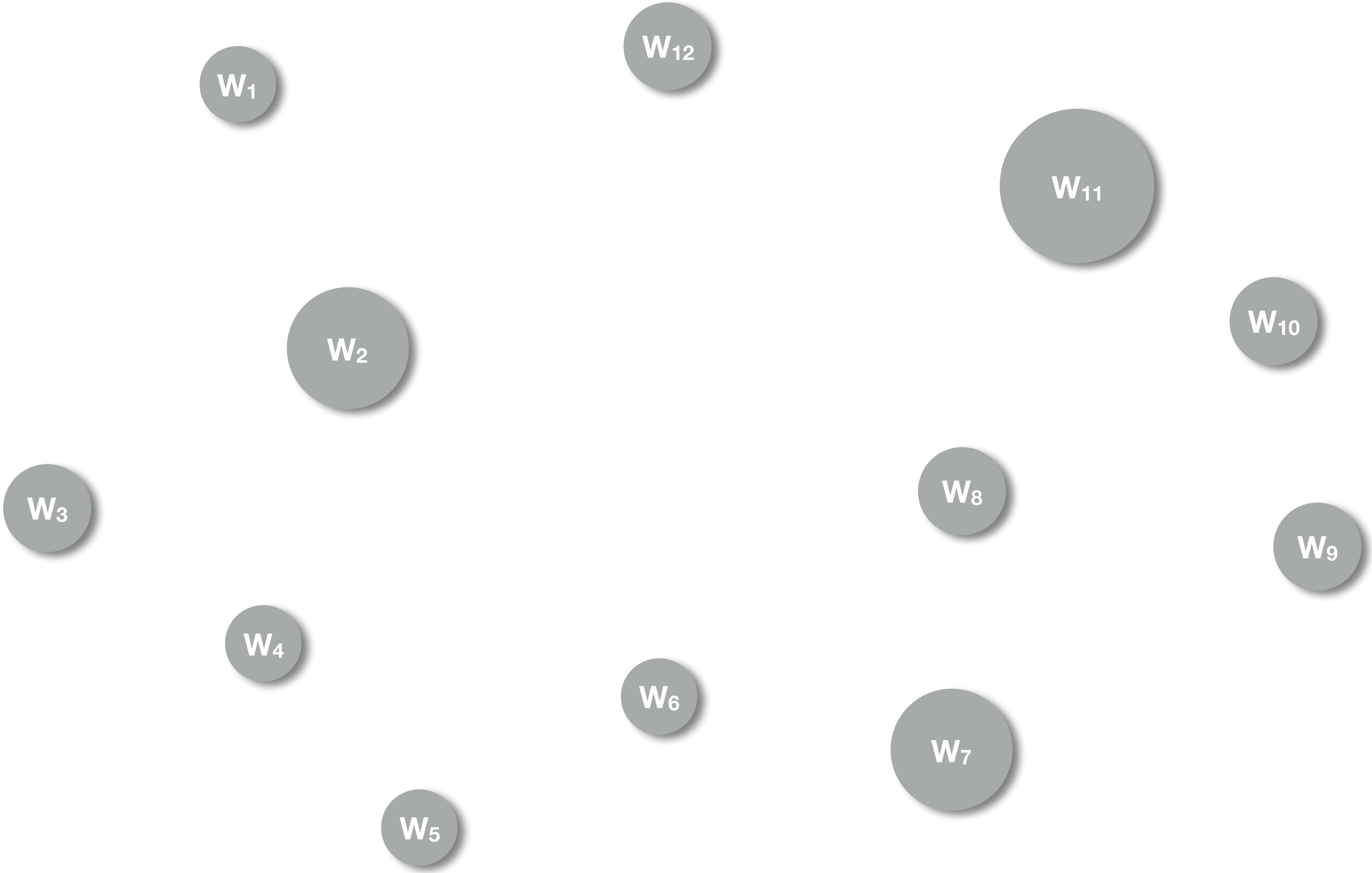


[MMRHA, AISTATS 16]

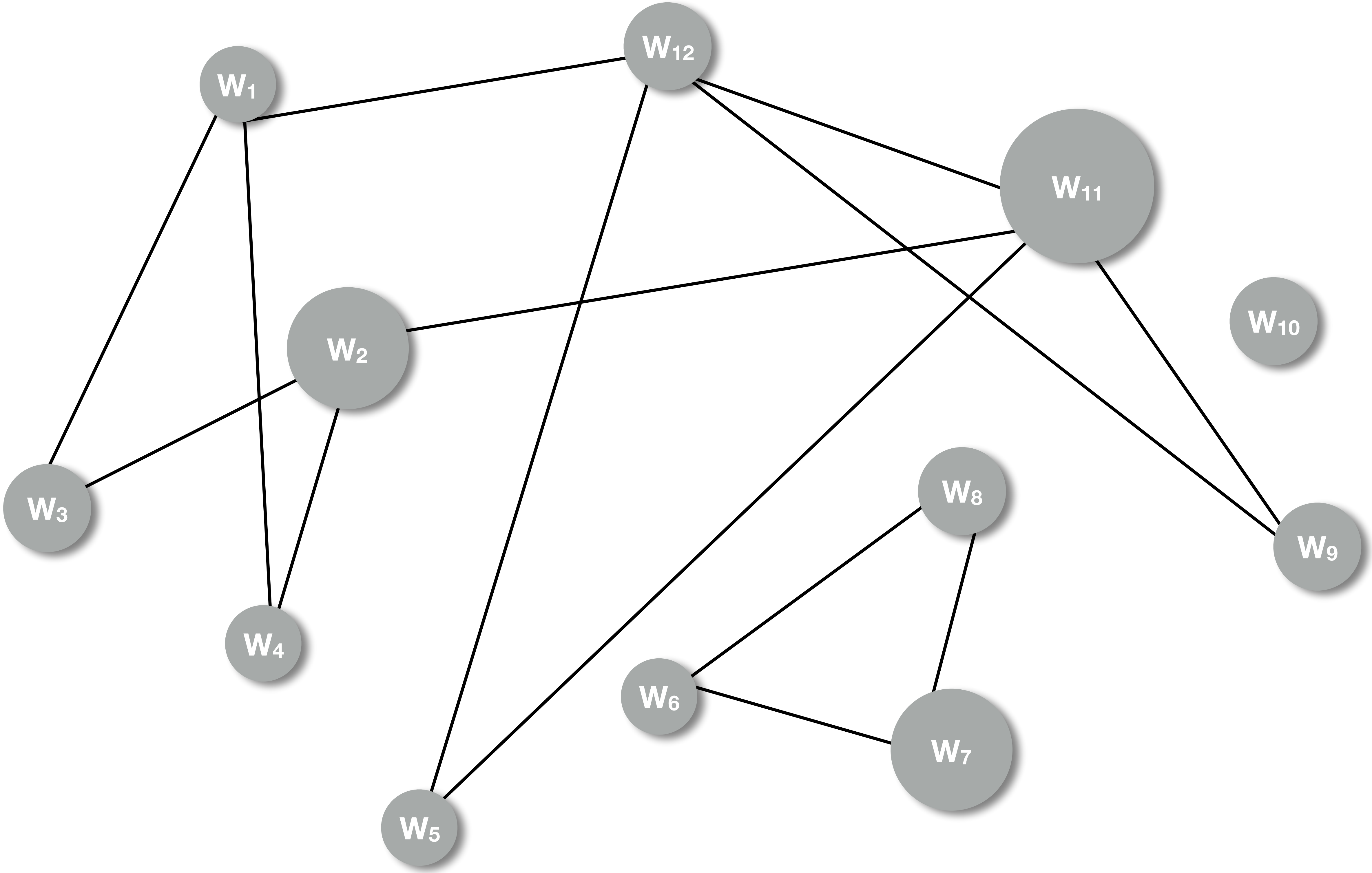
A LOCAL APPROACH



OUR APPROACH: PERSONALIZED MODELS



OUR APPROACH: PERSONALIZED MODELS



MULTI-TASK LEARNING

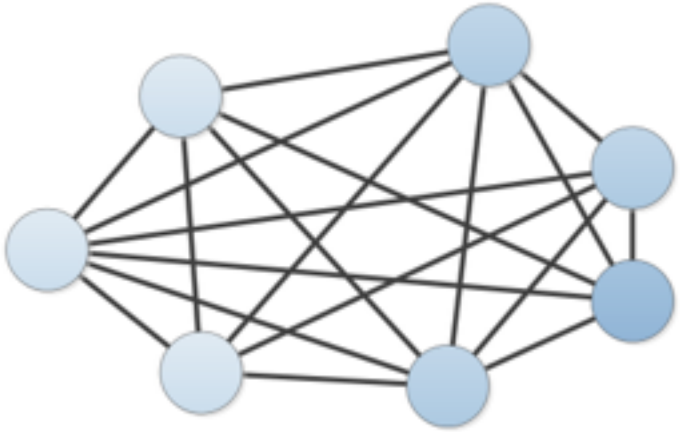
$$\min_{\mathbf{W},\boldsymbol{\Omega}} \sum_{t=1}^{m} \sum_{i=1}^{n_t} \ell_t(\mathbf{w}_t, \mathbf{x}_t^i) + \mathcal{R}(\mathbf{W}, \boldsymbol{\Omega})$$

models

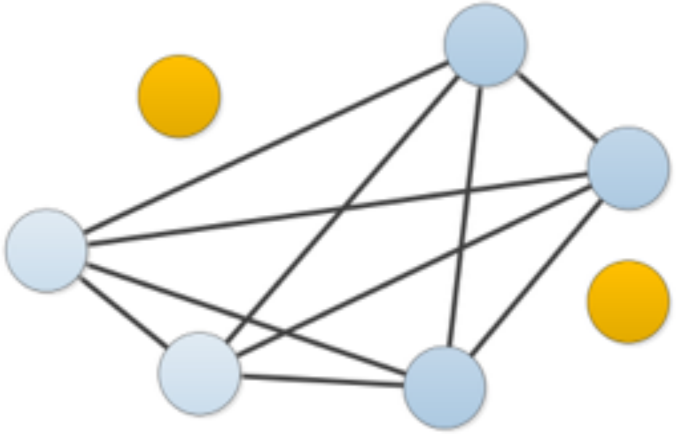
task relationship

losses

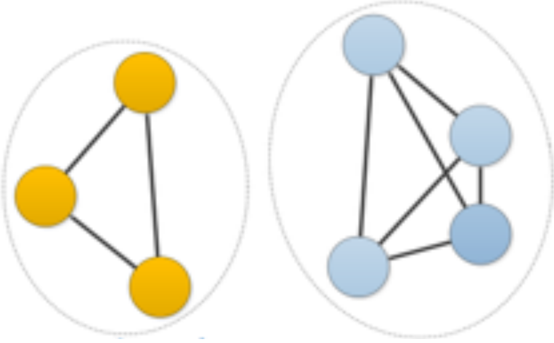
regularizer



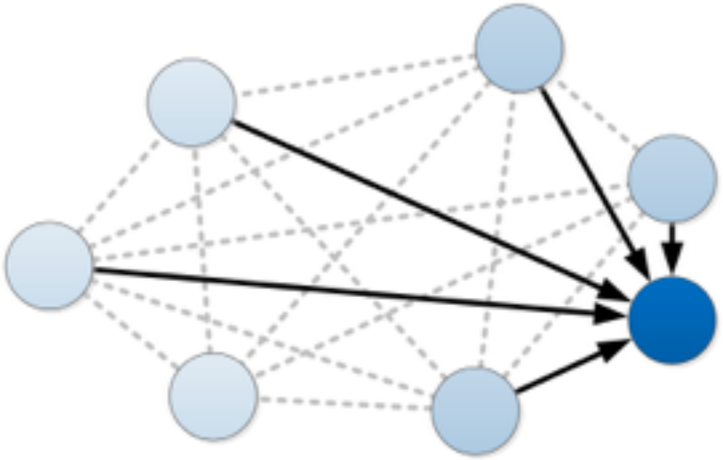
All tasks related



Outlier tasks



Clusters / groups



Asymmetric relationships

[ZCY, SDM 2012]

FEDERATED DATASETS

Human Activity

Google Glass

Land Mine

Vehicle Sensor

PREDICTION ERROR

	Global	Local	MTL
Human Activity	2.23 (0.30)	1.34 (0.21)	**0.46** **(0.11)**
Google Glass	5.34 (0.26)	4.92 (0.26)	**2.02** **(0.15)**
Land Mine	27.72 (1.08)	23.43 (0.77)	**20.09** **(1.04)**
Vehicle Sensor	13.4 (0.26)	7.81 (0.13)	**6.59** **(0.21)**

OUTLINE

- Unbalanced
- Non-IID
- Underlying Structure

Statistical Challenges

- Massively Distributed
- Node Heterogeneity

Systems Challenges

OUTLINE

- Unbalanced
- Non-IID
- Underlying Structure

Statistical Challenges

- Massively Distributed
- Node Heterogeneity

Systems Challenges

GOAL: FEDERATED OPTIMIZATION FOR MULTI-TASK LEARNING

$$\min_{\mathbf{W},\mathbf{\Omega}} \sum_{t=1}^{m} \sum_{i=1}^{n_t} \ell_t(\mathbf{w}_t^T \mathbf{x}_t^i) + \mathcal{R}(\mathbf{W}, \mathbf{\Omega})$$

- Solve for $\mathbf{W}, \mathbf{\Omega}$ in an alternating fashion
    - $\mathbf{\Omega}$ can be updated centrally
    - $\mathbf{W}$ *needs to be solved in federated setting*

Challenges:

- Communication is expensive
- Statistical & systems heterogeneity
    - Stragglers
    - Fault tolerance

GOAL: FEDERATED OPTIMIZATION FOR MULTI-TASK LEARNING

Idea:

Modify a *communication-efficient* method for the data center setting to handle:

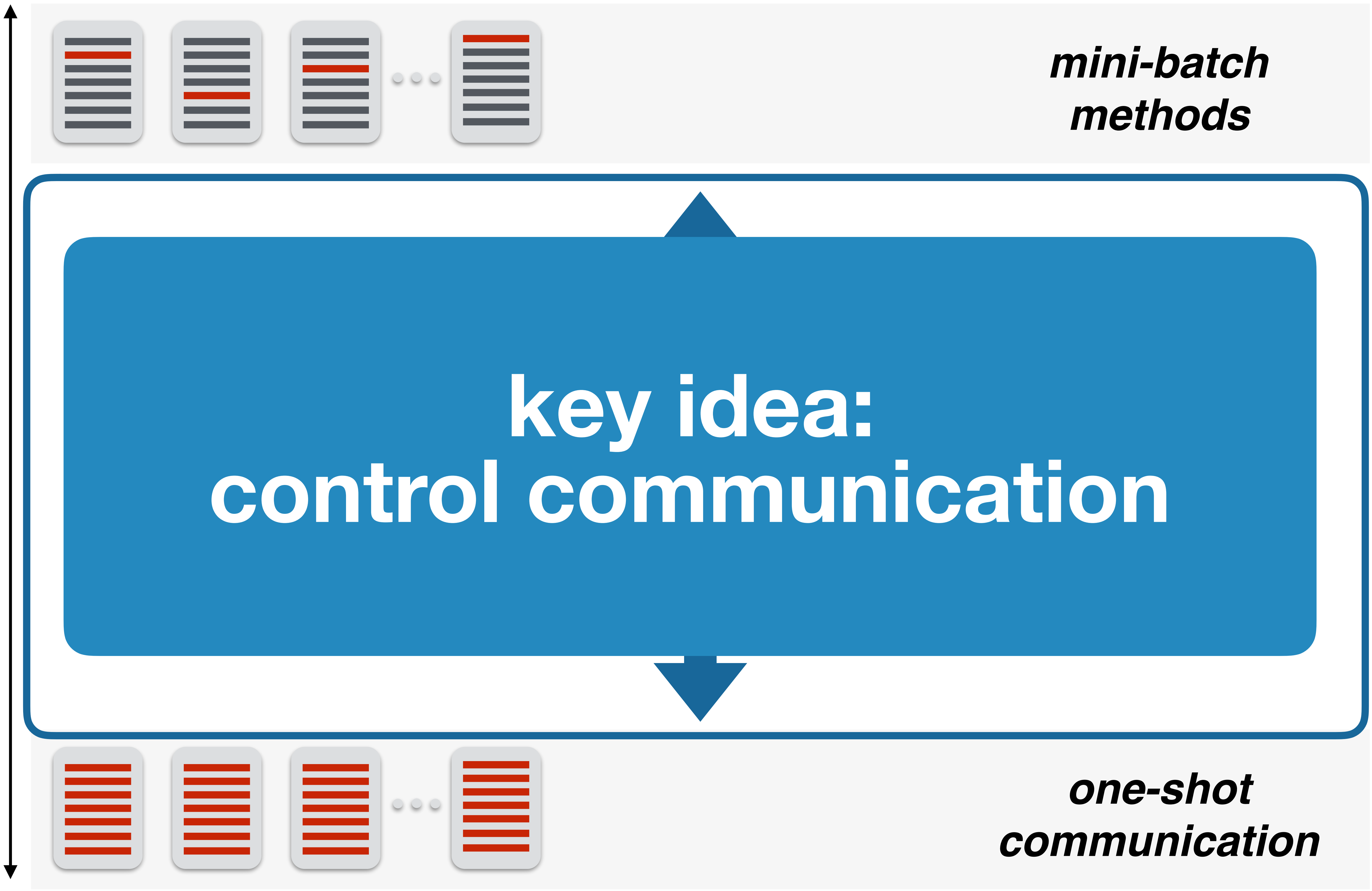
✔ Multi-task learning

✔ Stragglers

✔ Fault tolerance

- Fault tolerance

COCOA: COMMUNICATION-EFFICIENT DISTRIBUTED OPTIMIZATION



mini-batch methods

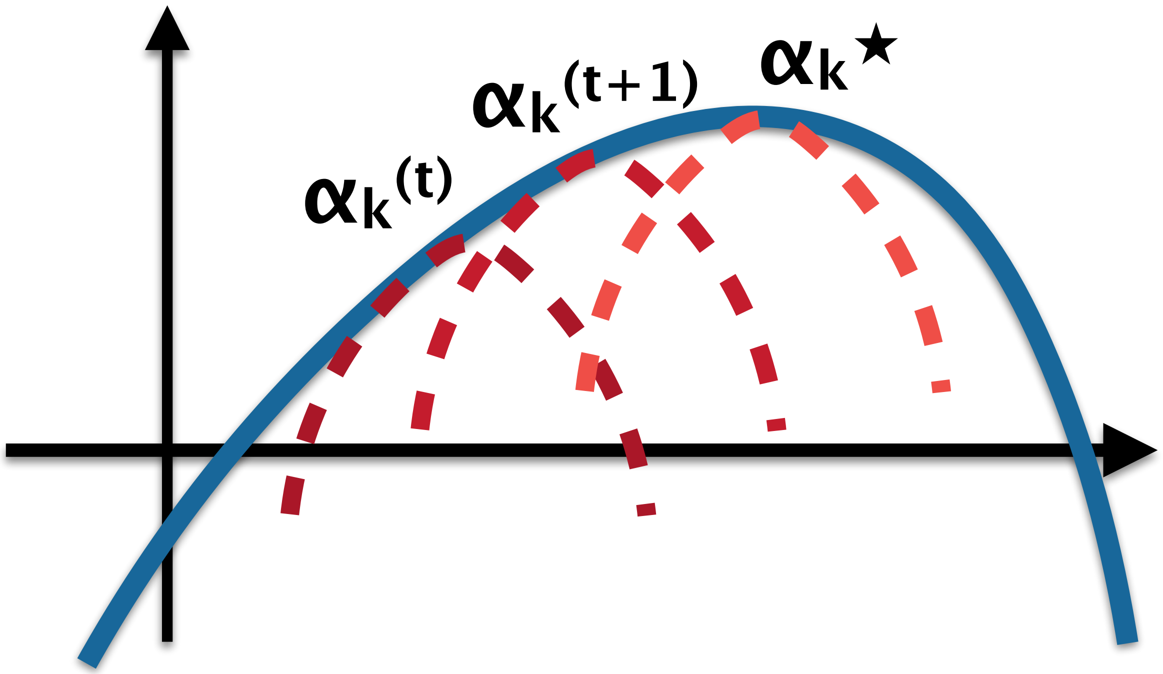
key idea: control communication

one-shot communication

COCOA: PRIMAL-DUAL FRAMEWORK

PRIMAL $\geq$ DUAL

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{w}^T x_i) + \lambda g(\mathbf{w}) \qquad \max_{\boldsymbol{\alpha}\in\mathbb{R}^n} -\frac{1}{n}\sum_{i=1}^{n} \ell^*(-\alpha_i) - \overset{\sum_{k=1}^{K} \tilde{g}^*(X_{[k]}, \boldsymbol{\alpha}_{[k]})}{\cancel{\lambda g^*(X\boldsymbol{\alpha})}}$$



COCOA: PRIMAL-DUAL FRAMEWORK

$, \boldsymbol{\alpha}_{[k]})$

$\underset{\mathbf{w}\in}{\mathrm{m}}$

$, \boldsymbol{\alpha})$

**challenge #1:
extend to MTL setup**

COCOA: COMMUNICATION PARAMETER

Main assumption:

each subproblem is solved to accuracy $\boldsymbol{\theta}$

$$\boldsymbol{\theta} \in [0, 1) \approx$$

amount of local
computation
vs.
communication

exactly solve

inexactly solve

COCOA: COMMUNICATION PARAMETER

Main assumption:

challenge #2:
make communication
more flexible

exactly solve

inexactly solve

MOCHA: COMMUNICATION-EFFICIENT FEDERATED OPTIMIZATION

$$\min_{\mathbf{W},\boldsymbol{\Omega}} \sum_{t=1}^{m}\sum_{i=1}^{n_t} \ell_t(\mathbf{w}_t^T \mathbf{x}_t^i) + \mathcal{R}(\mathbf{W},\boldsymbol{\Omega})$$

- Solve for **W, Ω** in an alternating fashion
- Modify CoCoA to solve **W** in federated setting

$$\min_{\boldsymbol{\alpha}} \sum_{t=1}^{m}\sum_{i=1}^{n_t} \ell_t^*(-\boldsymbol{\alpha}_t^i) + \mathcal{R}^*(\mathbf{X}\boldsymbol{\alpha})$$

$$\min_{\Delta\boldsymbol{\alpha}_t} \sum_{i=1}^{n_t} \ell_t^*(-\boldsymbol{\alpha}_t^i - \Delta\boldsymbol{\alpha}_t^i) + \langle \mathbf{w}_t(\boldsymbol{\alpha}), \mathbf{X}_t\Delta\boldsymbol{\alpha}_t\rangle + \frac{\sigma'}{2}\|\mathbf{X}_t\Delta\boldsymbol{\alpha}_t\|_{\mathbf{M}_t}^2$$

MOCHA: PER-DEVICE, PER-ITERATION APPROXIMATIONS

New assumption: $\theta_t^h \in [0, 1]$

each subproblem is solved to accuracy ~~$\theta \in [0, 1)$~~

Stragglers (Statistical heterogeneity)

- Difficulty of solving subproblem
- Size of local dataset

Stragglers (Systems heterogeneity)

- Hardware (CPU, memory)
- Network connection (3G, LTE, …)
- Power (battery level)

Fault tolerance

- Devices going offline

CONVERGENCE

New assumption:
each subproblem is solved to accuracy θ_t^h

and assume: $\mathbb{P}[\theta_t^h := 1] < 1$

Theorem 1. Let ℓ_t be L-Lipschitz, then

$$T \geq \frac{1}{(1-\bar{\Theta})}\left(\frac{8L^2n^2}{\epsilon} + \tilde{c}\right)$$

1/ε rate

Theorem 2. Let ℓ_t be $(1/\mu)$-smooth, then

$$T \geq \frac{1}{(1-\bar{\Theta})} \frac{\mu+n}{\mu} \log \frac{n}{\epsilon}$$

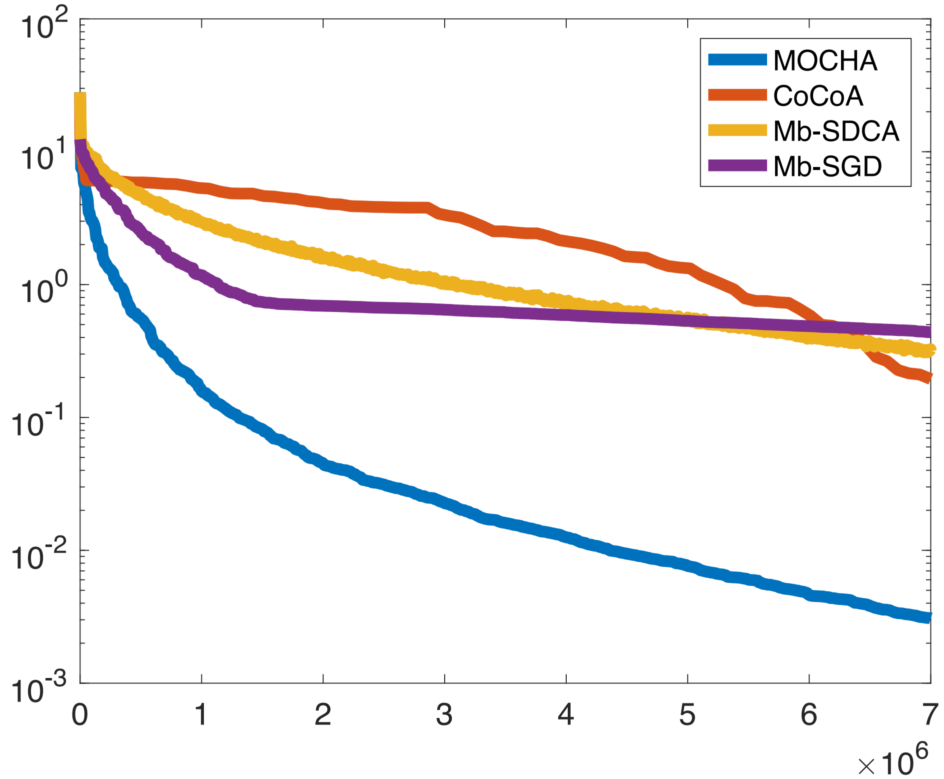
linear rate

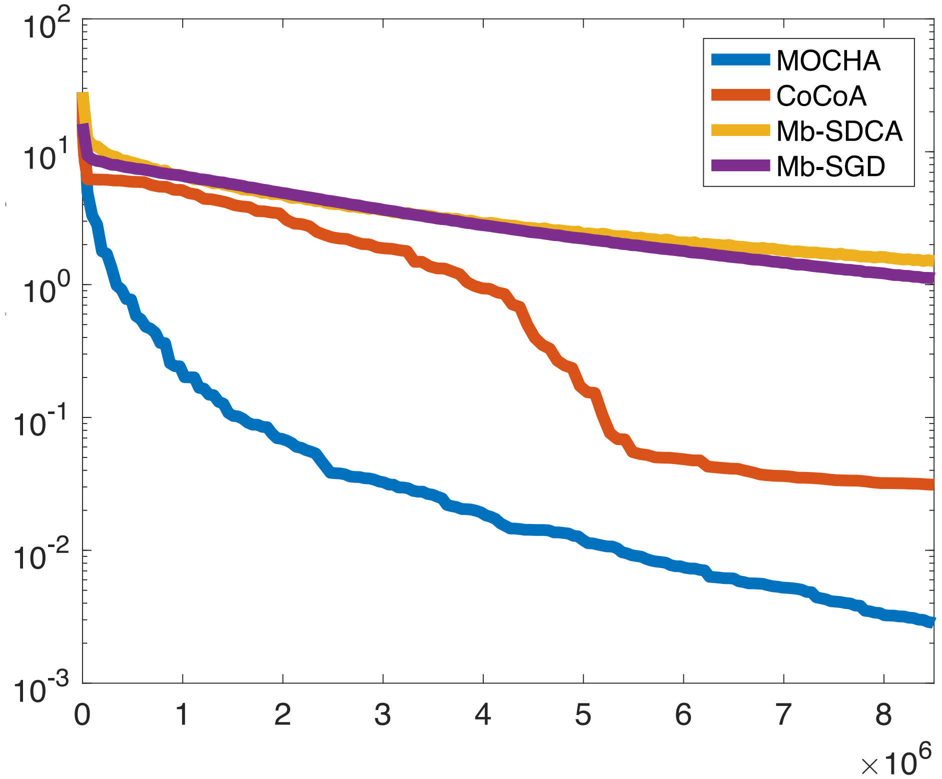
MOCHA: COMMUNICATION-EFFICIENT FEDERATED OPTIMIZATION

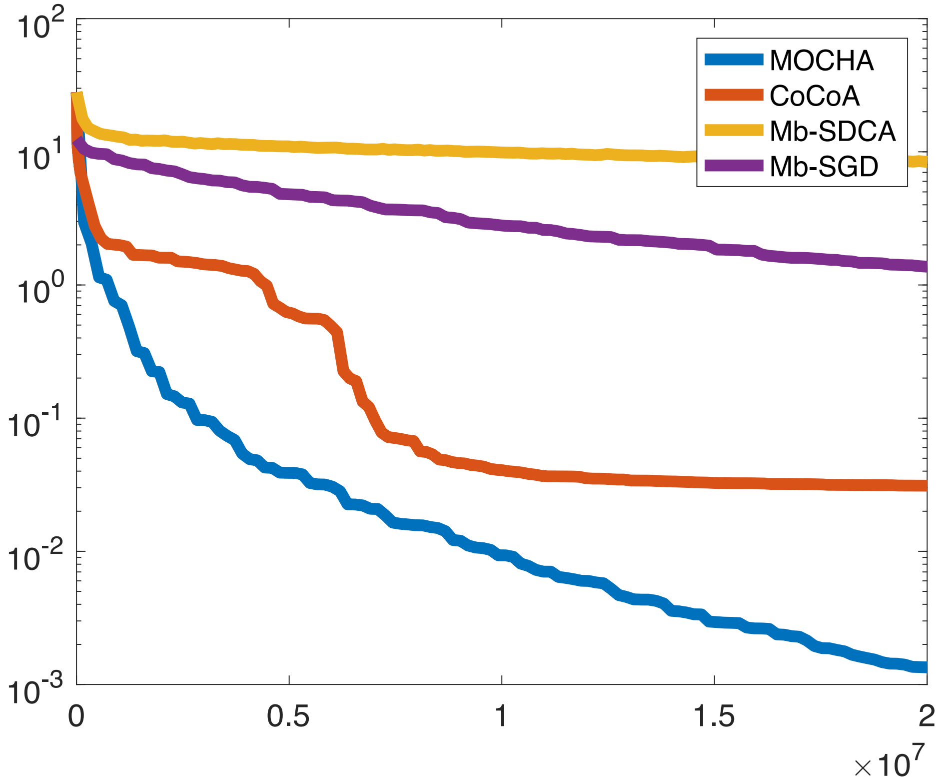
Algorithm 1 MOCHA: Federated Multi-Task Learning Framework

```
**Input:** Data $\mathbf{X}_t$ stored on $t = 1, \dots, m$ devices
Initialize $\boldsymbol{\alpha}^{(0)} := \mathbf{0}$, $\mathbf{v}^{(0)} := \mathbf{0}$
**for iterations** $i = 0, 1, \dots$ **do**
    **for iterations** $h = 0, 1, \cdots, H_i$ **do**
        **for devices** $t \in \{1, 2, \dots, m\}$ **in parallel do**
            call local solver, returning $\theta_t^h$-approximate solution $\Delta\boldsymbol{\alpha}_t$
            update local variables $\boldsymbol{\alpha}_t \leftarrow \boldsymbol{\alpha}_t + \Delta\boldsymbol{\alpha}_t$
        **reduce:** $\mathbf{v} \leftarrow \mathbf{v} + \sum_t \mathbf{X}_t \Delta\boldsymbol{\alpha}_t$
    Update $\boldsymbol{\Omega}$ centrally using $\mathbf{w}(\mathbf{v}) := \nabla\mathcal{R}^*(\mathbf{v})$
Compute $\mathbf{w}(\mathbf{v}) := \nabla\mathcal{R}^*(\mathbf{v})$
**return:** $\mathbf{W} := [\mathbf{w}_1, \dots, \mathbf{w}_m]$
```

STATISTICAL HETEROGENEITY


Wifi


LTE


3G

MOCHA & COCOA PERFORM PARTICULARLY WELL IN HIGH-COMMUNICATION SETTINGS

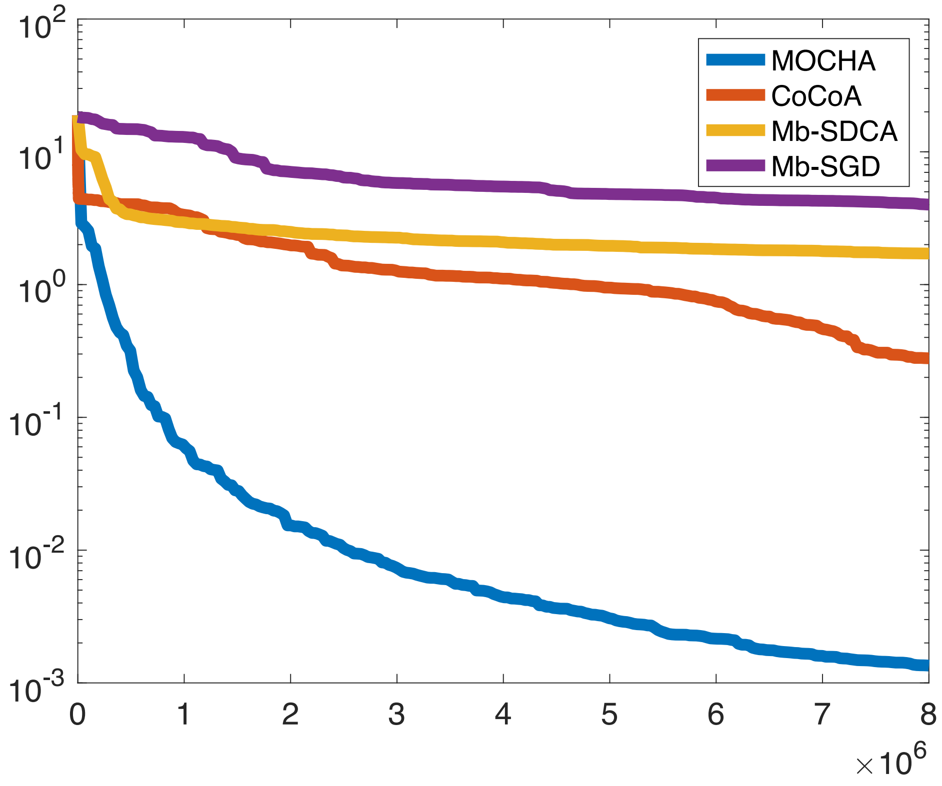
MOCHA IS ROBUST TO STATISTICAL HETEROGENEITY

SYSTEMS HETEROGENEITY

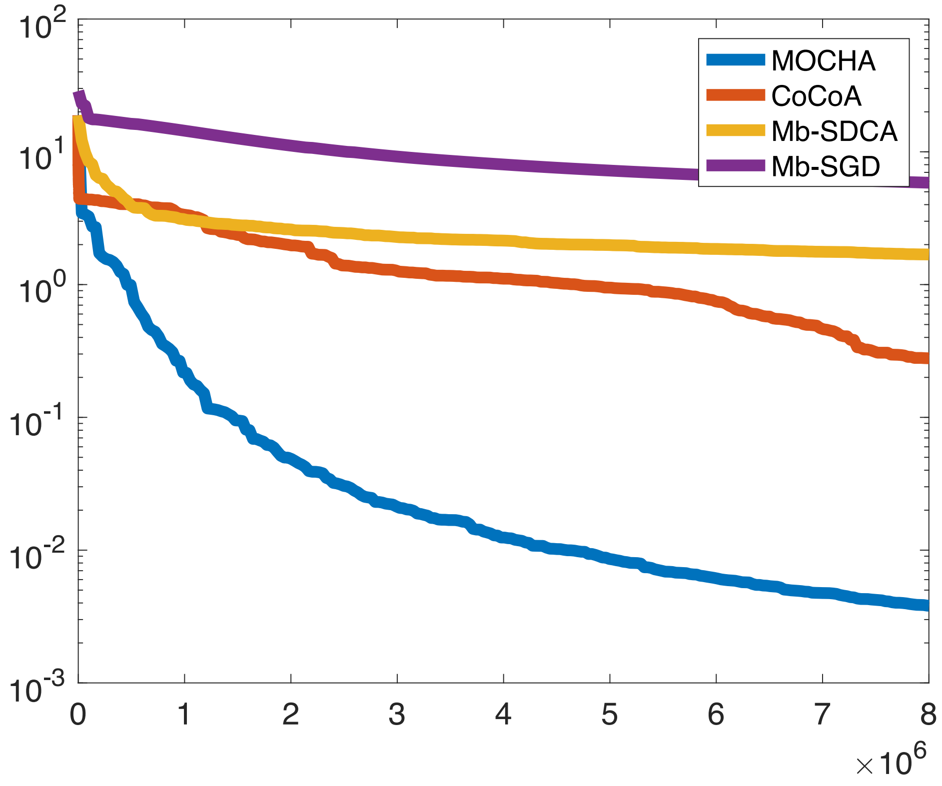
MOCHA SIGNIFICANTLY OUTPERFORMS ALL COMPETITORS

[BY 2 ORDERS OF MAGNITUDE]

Low

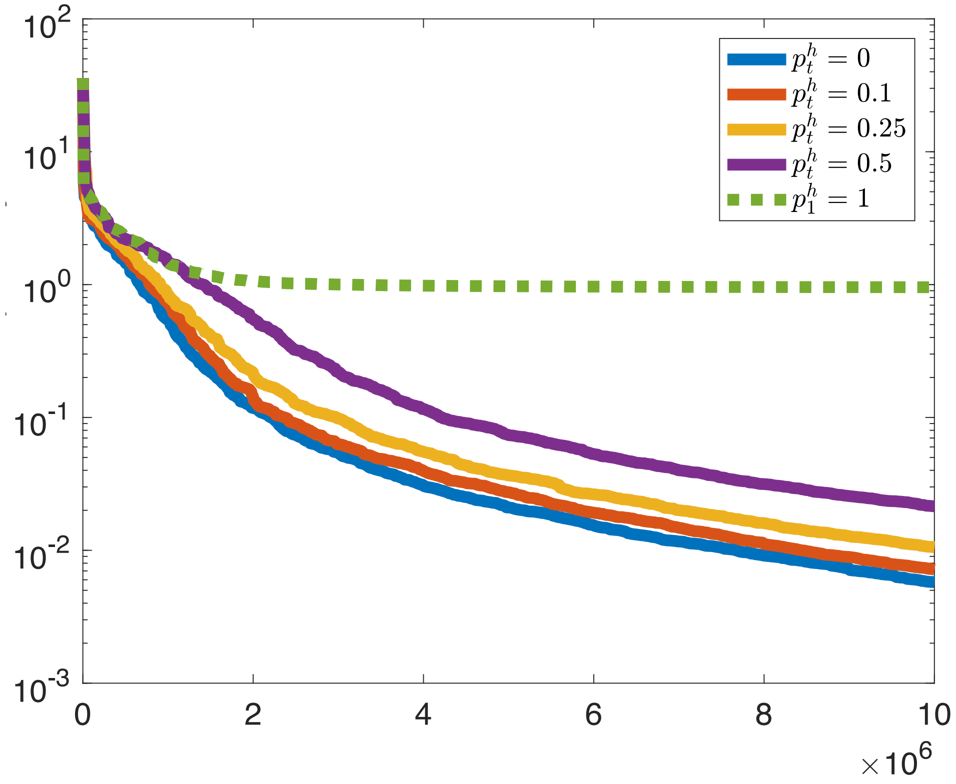


High

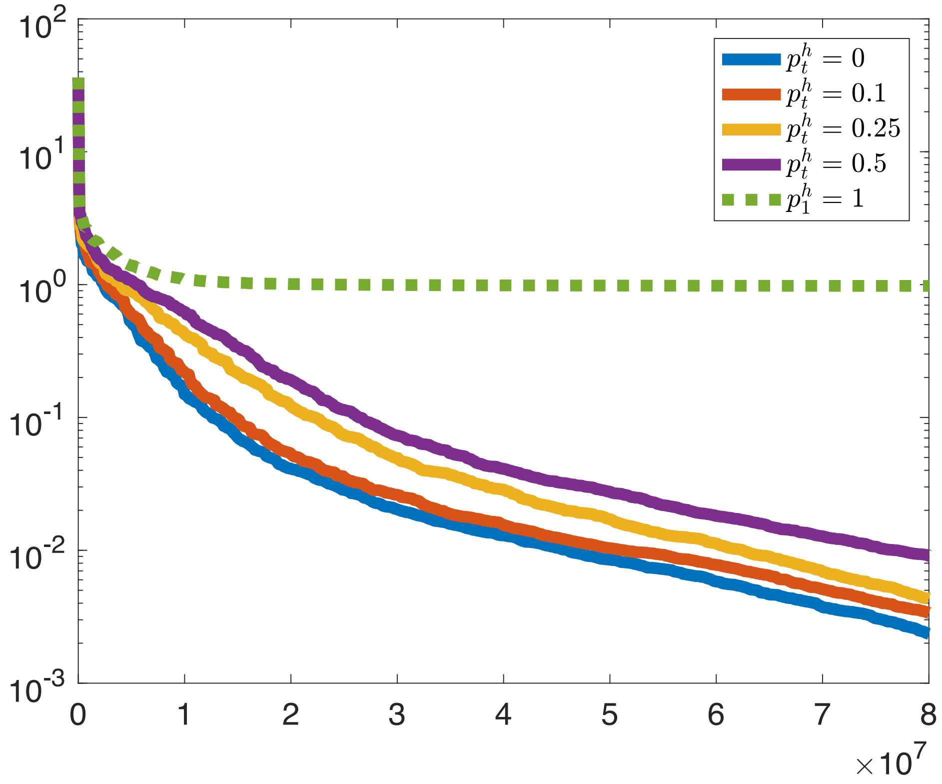


FAULT TOLERANCE

W-Step



Full Method



MOCHA IS ROBUST TO DROPPED NODES

OUTLINE

- Unbalanced
- Non-IID
- Underlying Structure

Statistical Challenges

- Massively Distributed
- Node Heterogeneity

Systems Challenges

WWW.SYSML.CC

Virginia Smith

Stanford / CMU

CODE & PAPERS

cs.berkeley.edu/~vsmith