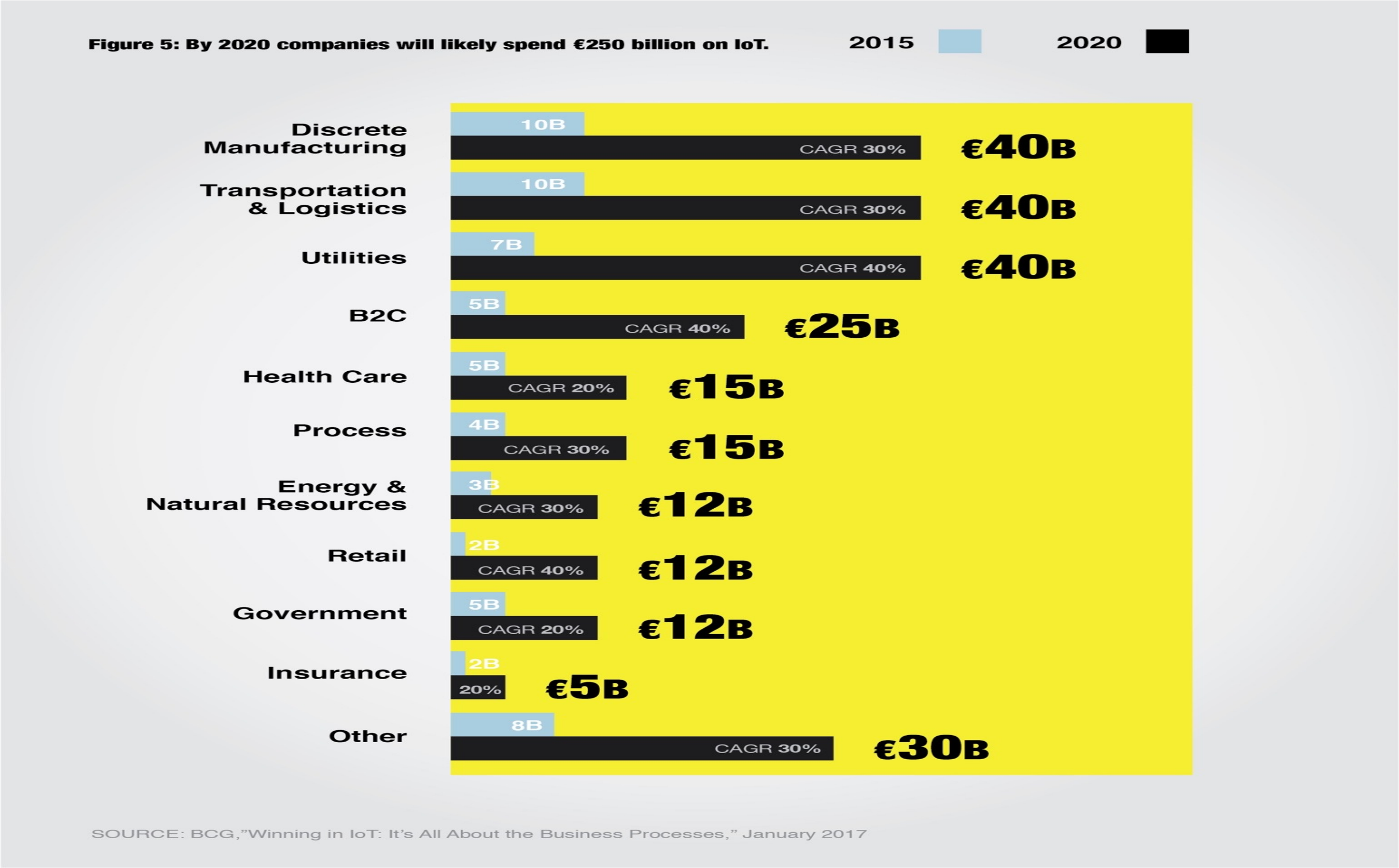
Human Automation Teams in Cyber Physical Societal Systems Transformation

Chinmay Maheshwari (Berkeley)
Eric Mazumdar (Caltech)
Manxi Wu (Cornell)
Lillian Ratliff (UW)
Kshitij Kulkarni (Berkeley)
Shankar Sastry (Berkeley)

Spending on IoT

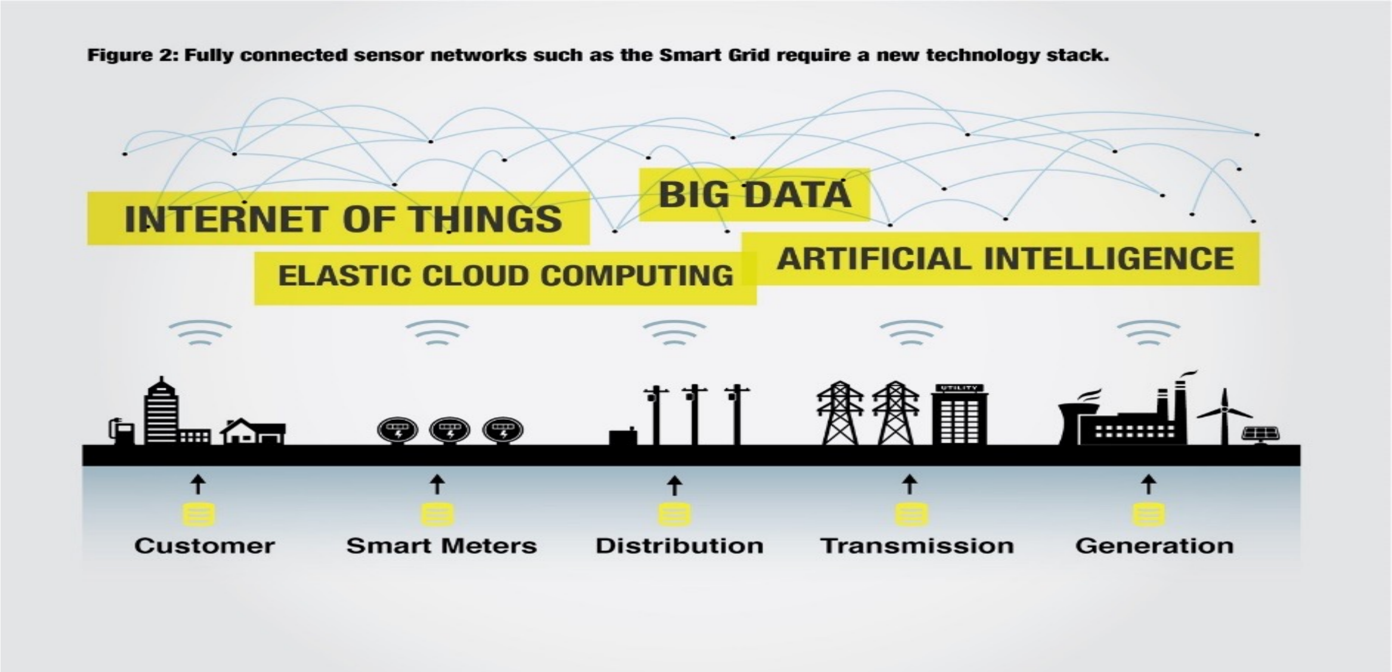


The Power Grid Example

- The electric power grid designed by Edison and Westinghouse 100 years ago was billed by NAE the most significant invention of the 20th Century. The 21st century development of the smart grid is the $ 2 Trillion IoT sensoring of the electric utility value chain.

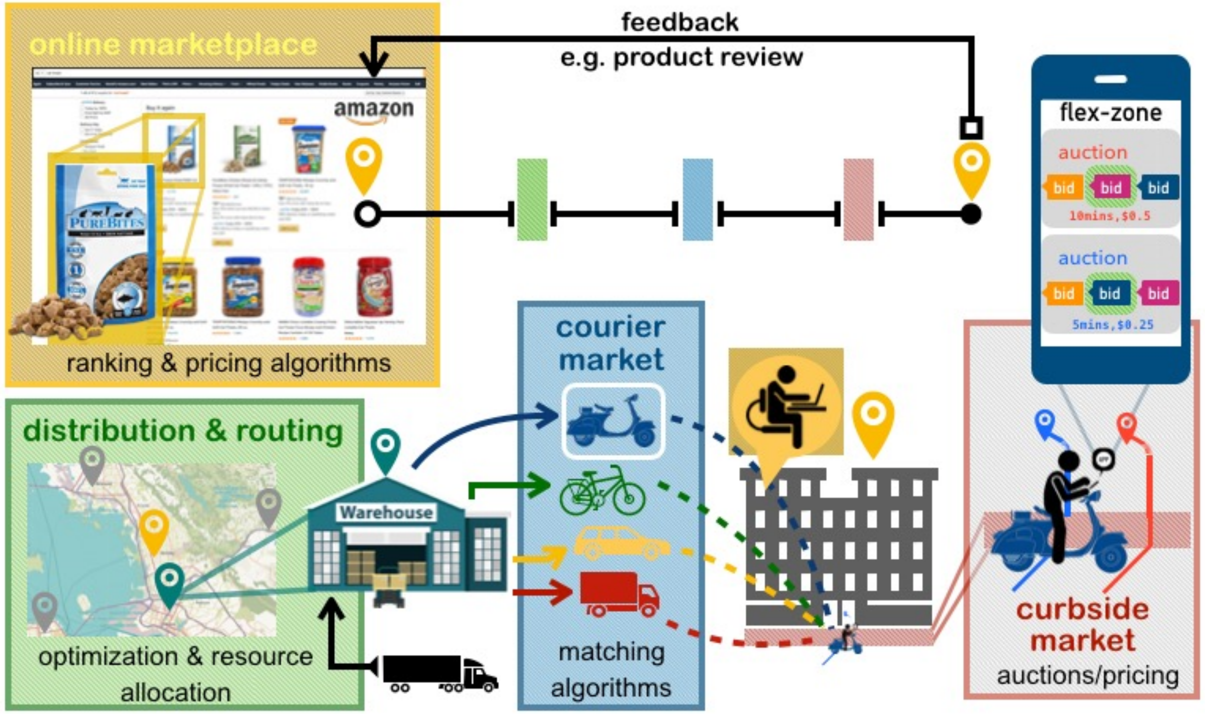
Century of Innovation: Twenty Engineering Achievements That Transformed Our Lives,"NAE 2003.

"*Estimating the Costs and Benefits of the Smart Grid*," Electric Power Research Institute (EPRI), March 2011.

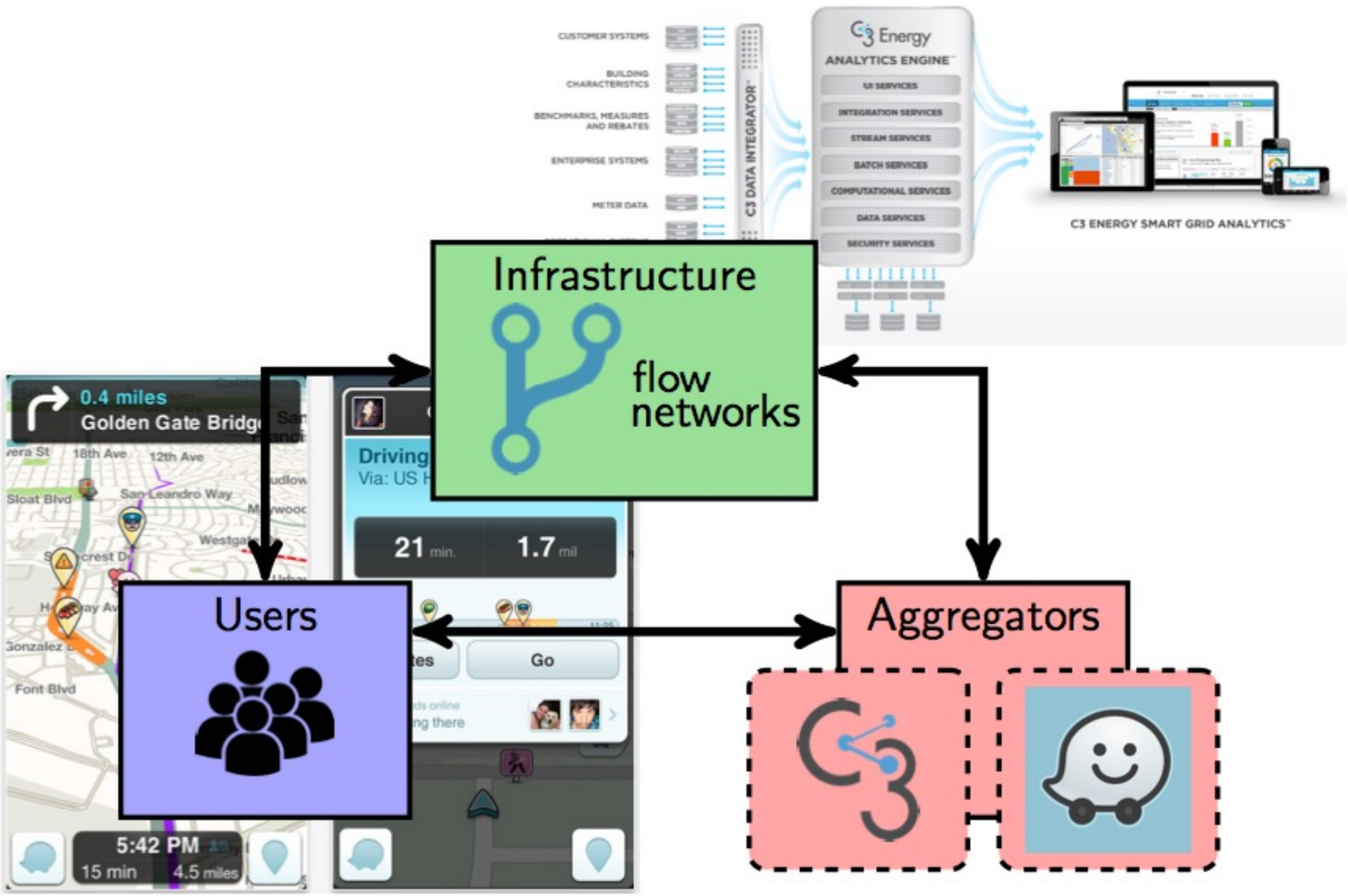


Intelligent Transportation and Logistics: From Suppliers to the Curb

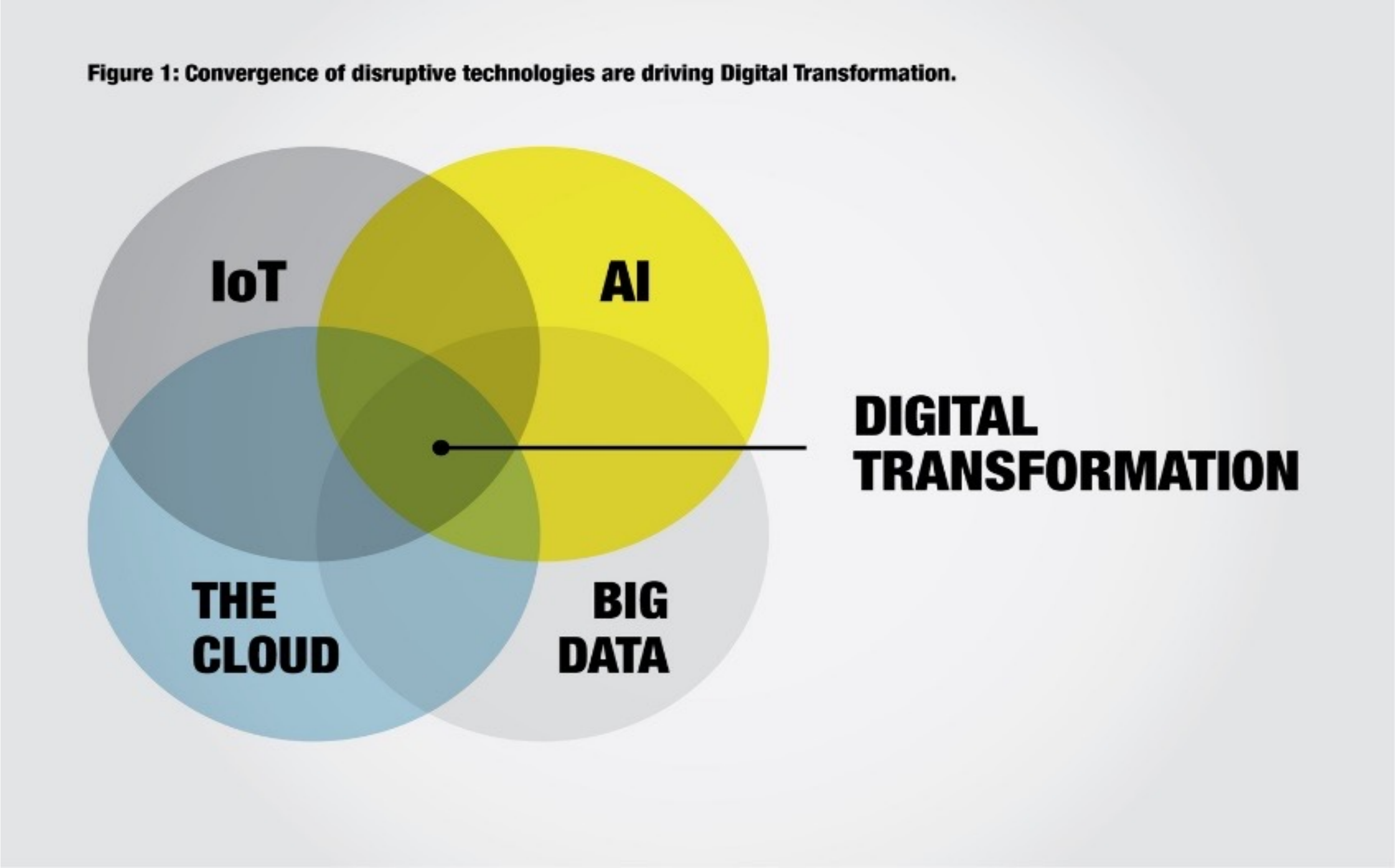
- How do AI and edge computing fit into Intelligent Transportation Systems?
  - 75% of enterprise-generated data will be created at the edge by 2025
  - 4TB data generated by one AV in one day
  - 1 in 10 vehicles will be AVs by 2030
  - “always-on” supply chains: IoT innovations are driving the future of logistics and supply chain management



Sharing Economy: Data as a Commodity



Digital Transformation of Societal Systems

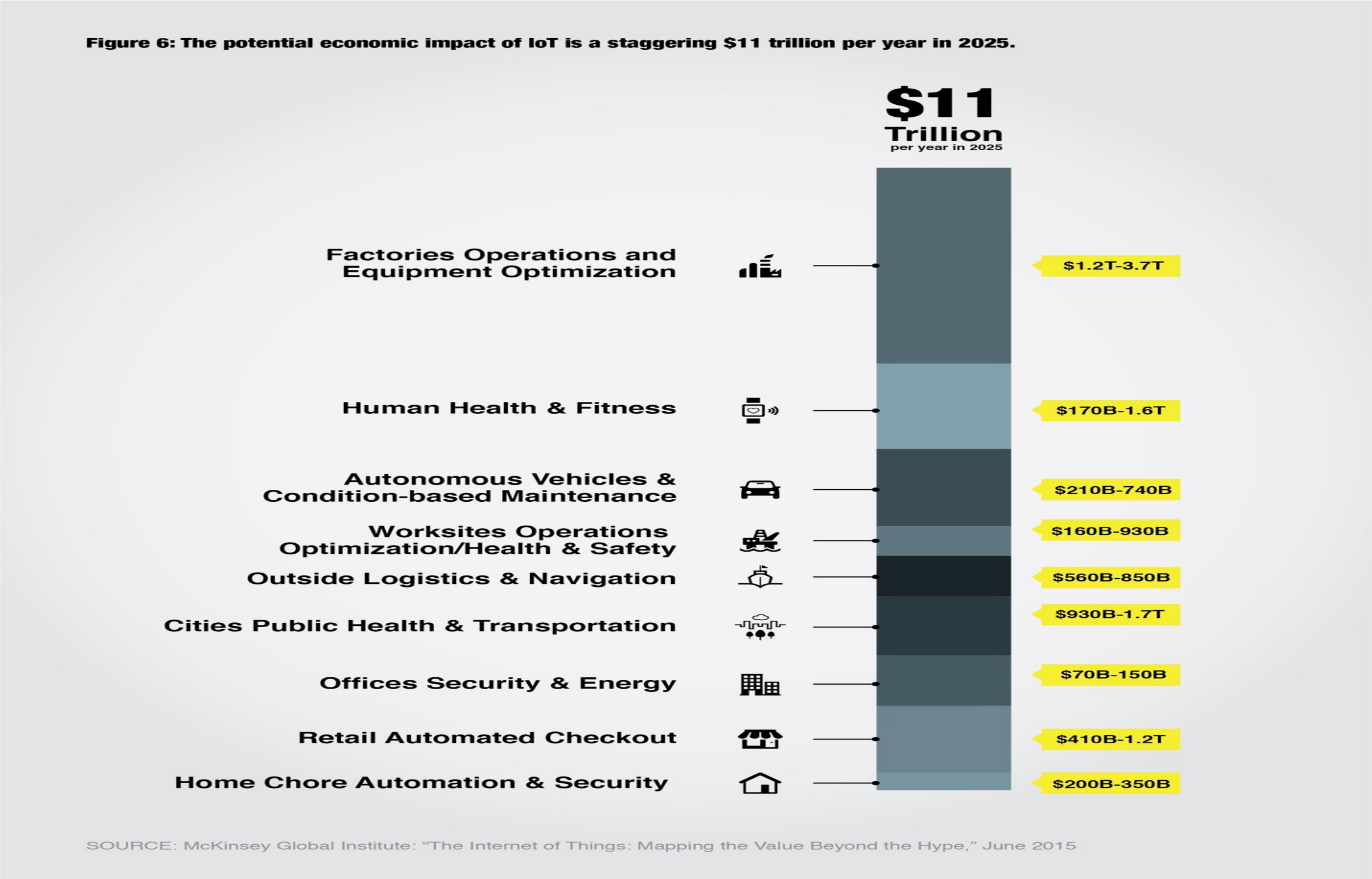


This is Much More than Big Data!!

- With "Big Data" we perform calculations on all the data. This brings "back again" a renaissance to the promise of AI to evolve a new kind of CPS machine learning to perform precise predictive analytics.
- At the convergence of IoT, Cloud Computing, Data Analytics, and AI is Digital Transformation.
- The value that industries and governments will create from IoT Digital Transformation will range from $3- $11 trillion per year in 2025.

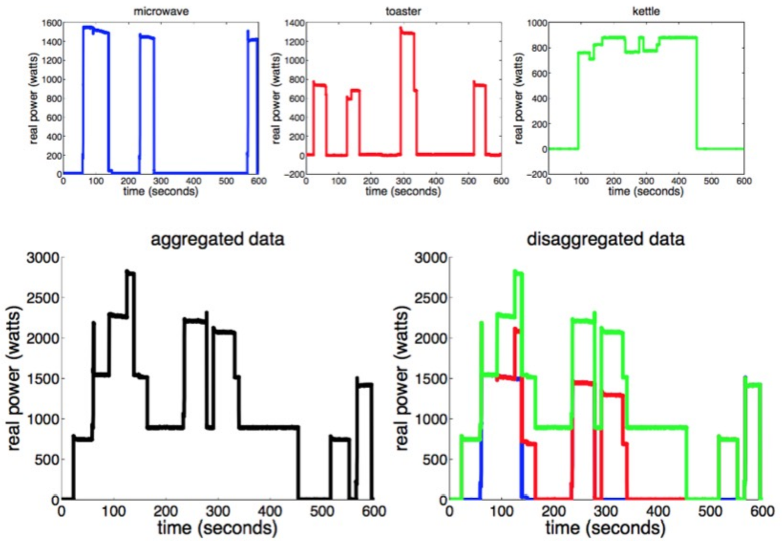
"*The Internet of Things: Mapping the Value Beyond the Hype*," McKinsey Global Institute, June 2015.

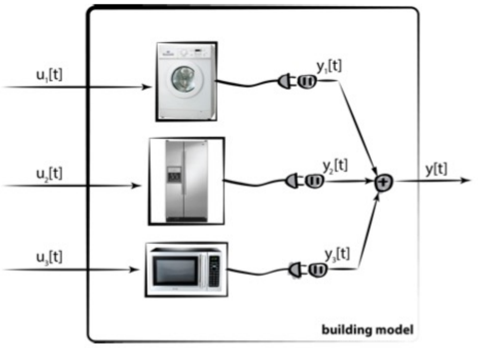
Economic Impact: Off the Charts!



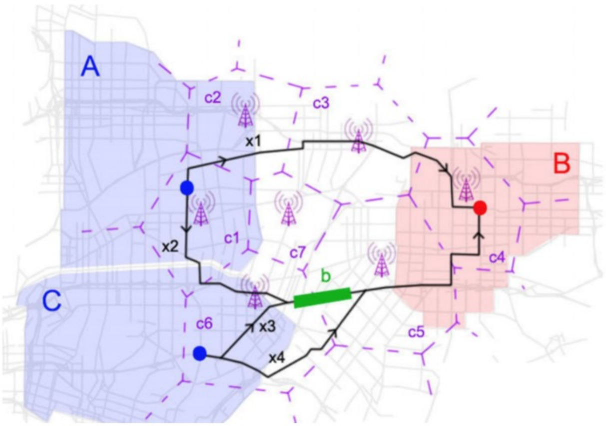
Issues: Usage Modeling—Disaggregation

Energy Disaggregation



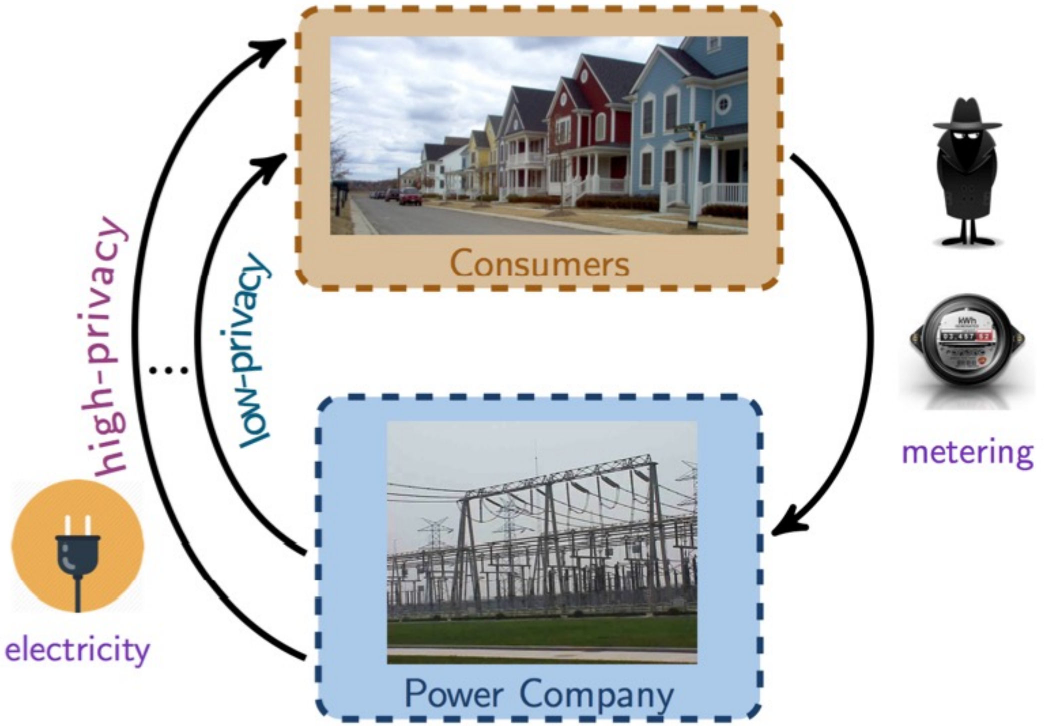


Route Disaggregation



Privacy Contracts

Design service contracts differentiated according to the fidelity of the data collected



- We find that those that value privacy very highly free ride on society.
- Privacy risk leads to tradeoff between investment in security and insurance.
- User valuations of data need to be factored into the design of service models in order to increase social welfare!

Humans and Digital Transformation: Traffic apps (courtesy Prof Alex Bayen)

Fundamental premise of routing services

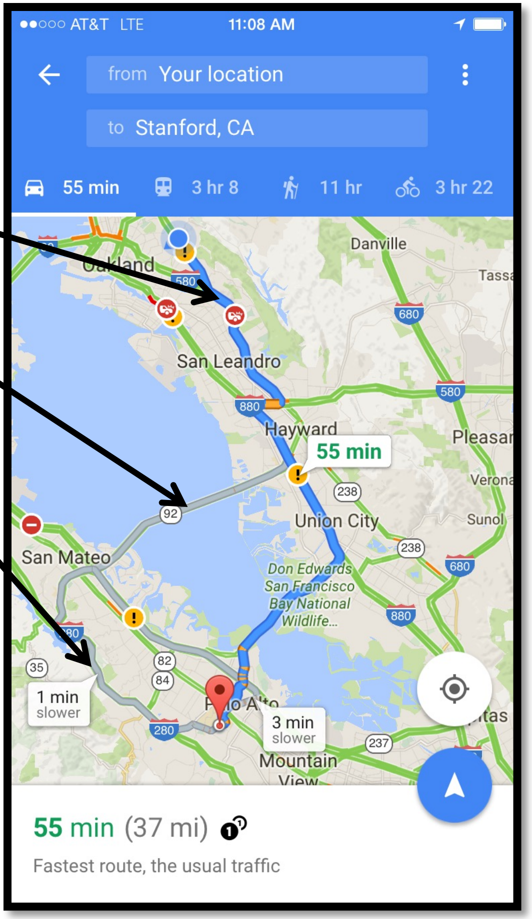
- Each app enabled user receives a [SOTA] shortest path
- Some follow the recommendations

All paths proposed are nearly equal:

- Shortest path (55mins)
- Third shortest path (58 mins)
- Second shortest path (56 mins)

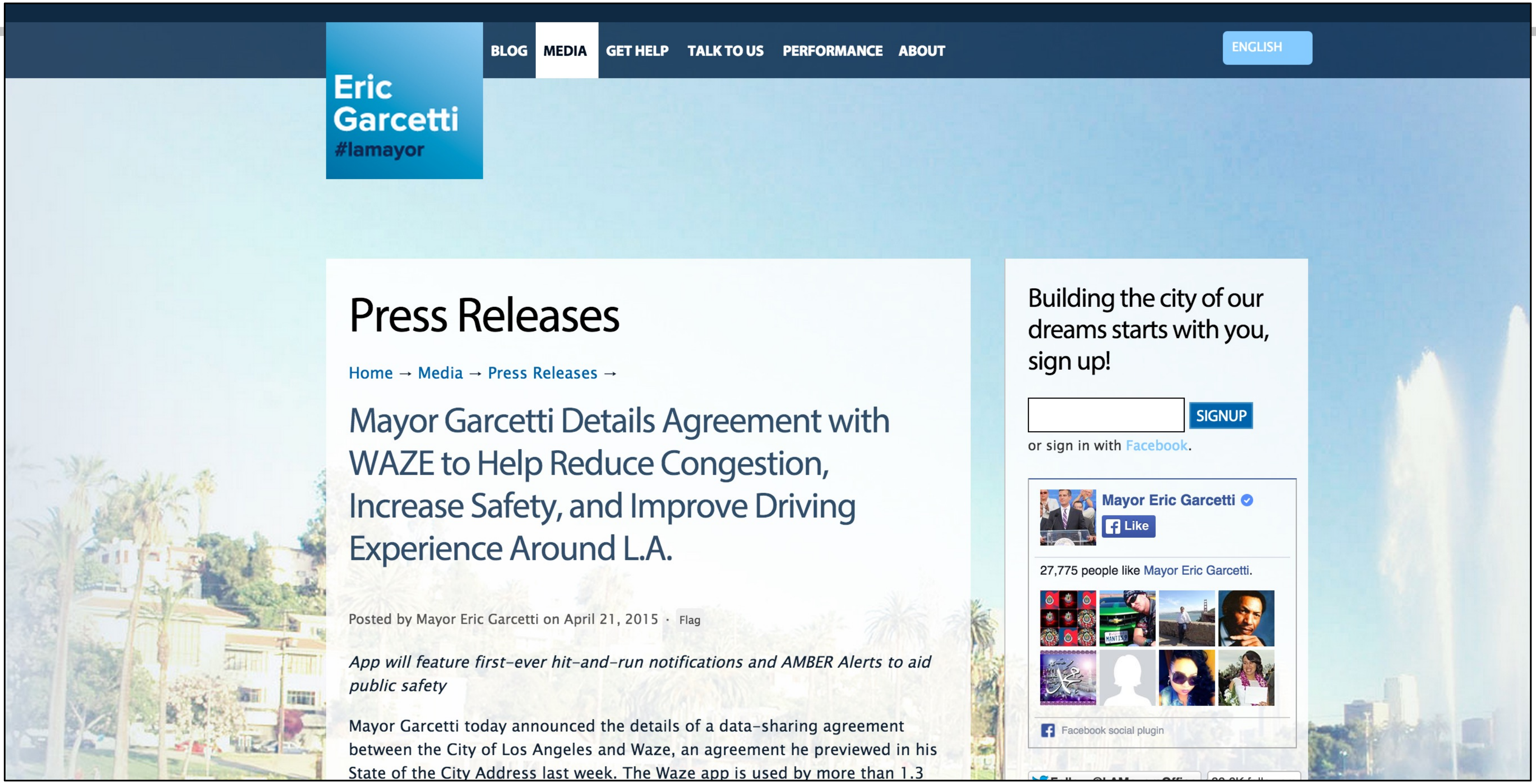
Routing does in general not depend on

- Forecast of the network loading using demand data (incomplete today)
- Forecast of the network using potential impact of routing (i.e. routed users) on the network
- Knowledge of what competitors of the app are doing (in the present case, Apple, INRIX, 511, etc.)

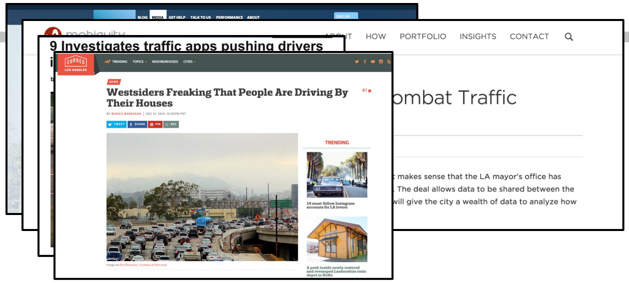


[Samaranayake et al., TR-C, 2012, ALANEX 2014, SIAM MAEE 2014]

Initially people “thought” app helped



Until more and more people started using it



SECTIONS SEARCH **Los Angeles Times** SUBSCRIBE 4 weeks for 99¢

SDAY FEB. 14, 2017 MOST POPULAR LOCAL SPORTS ENTERTAINMENT POLITICS OPINION PLACE AN AD

California Commute New traffic apps may be pushing cars into residential areas

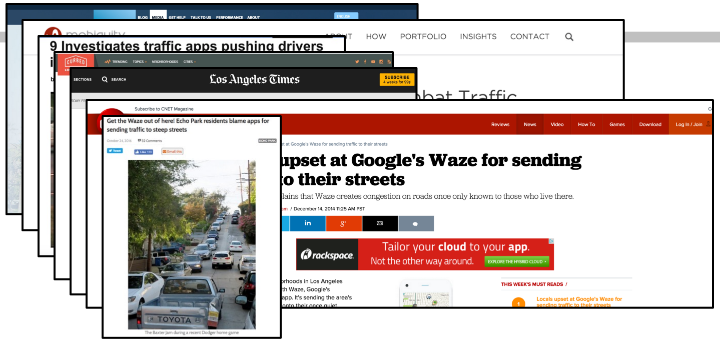
Vehicles crowd the intersection of Cody and Woodcliff roads in Sherman Oaks. Residents say GPS apps are to blame for the new

Related Coverage

Stuck in bad traffic? Good chance it's Thursday evening

NOV. 11, 2014

Specific apps are identified as responsible




Readers React How an app destroyed their streets: Readers count the Waze

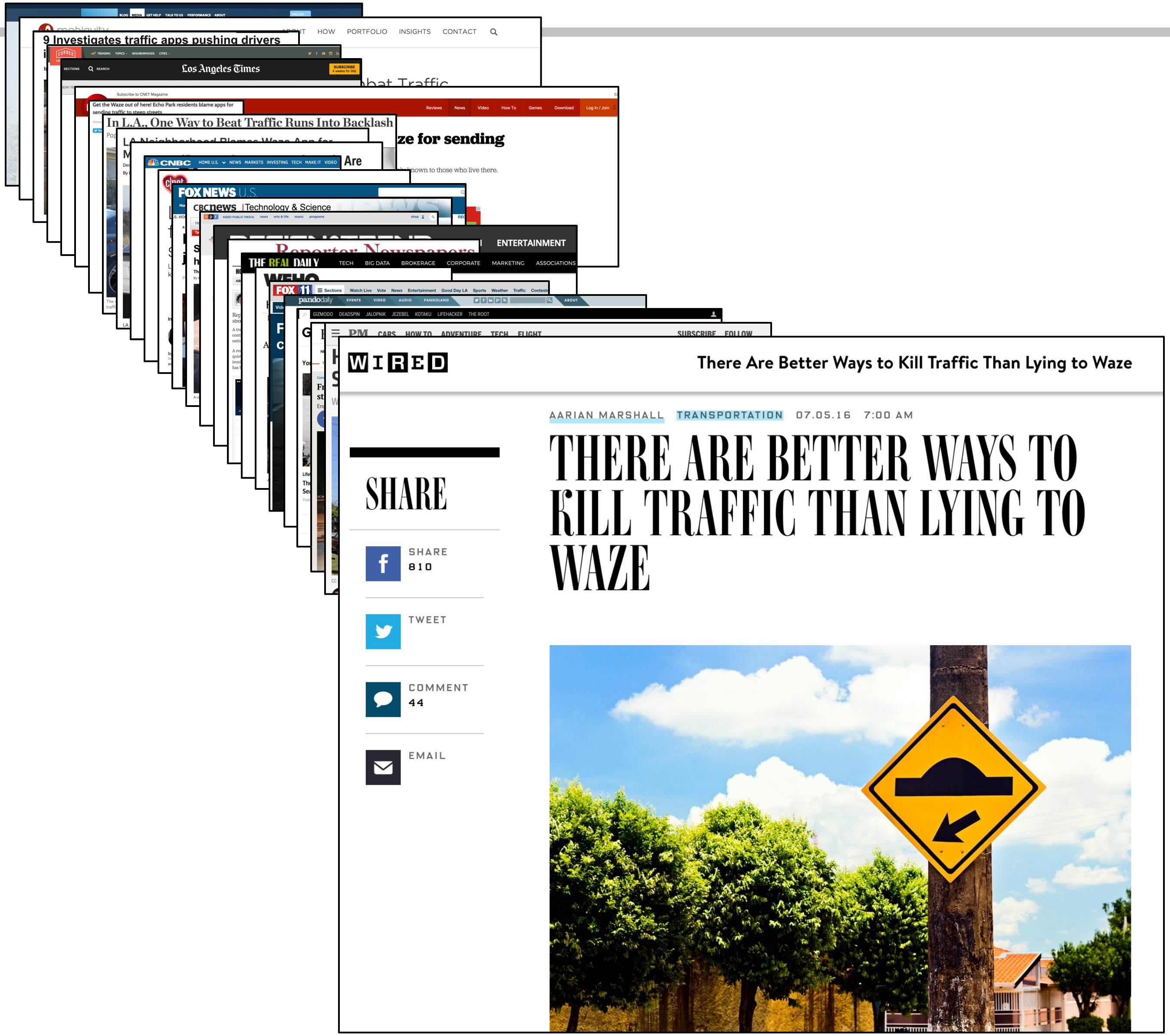
Vehicles crowd the intersection of Cody Road and Woodcliff Road in Sherman Oaks on Jan. 5. Residents say the worsening traffic on side streets is partially to blame on Waze. (Los Angeles Times)

Related Coverage

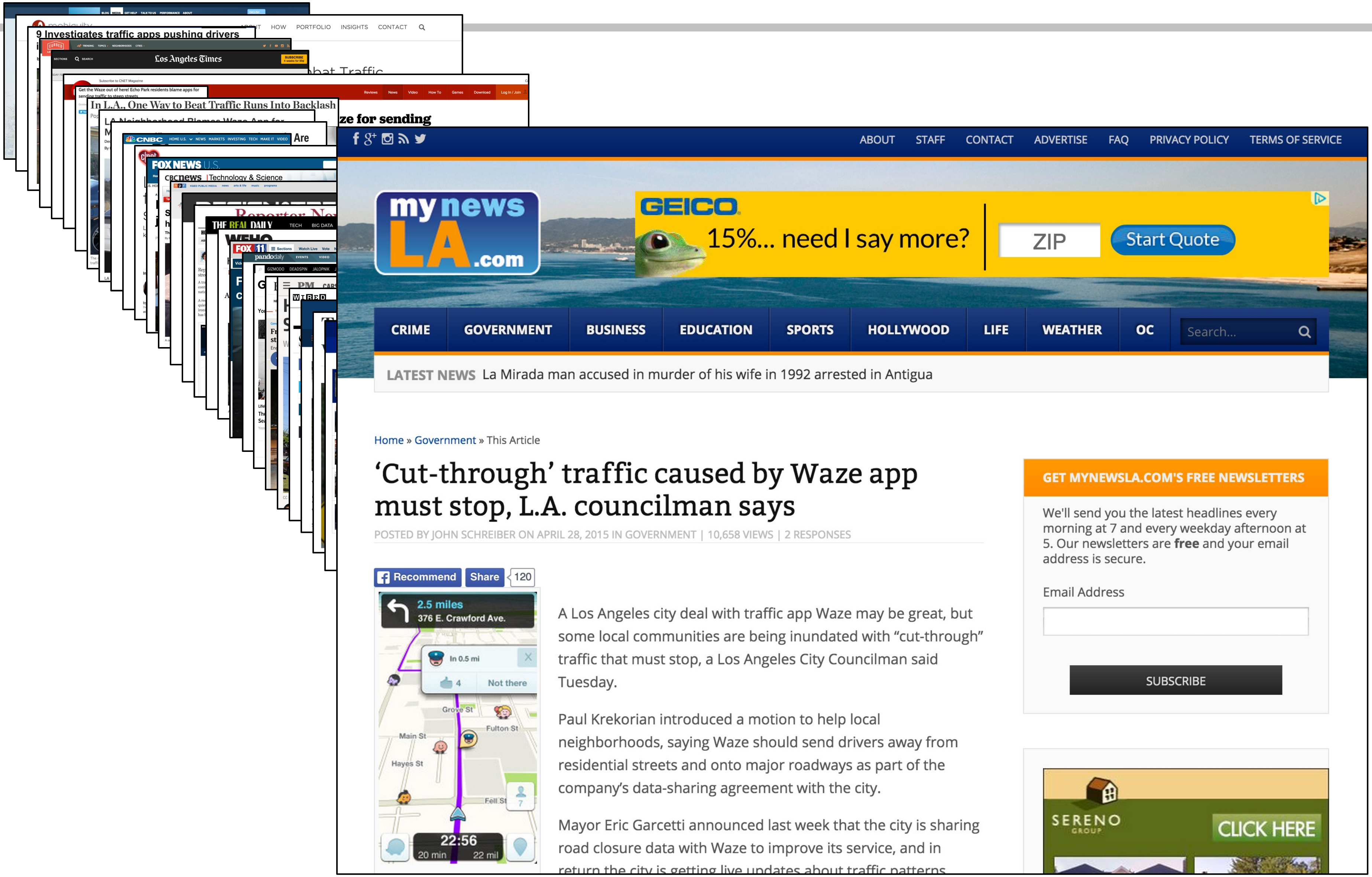
Time to rein in California's traffic ticket surcharges

MAY 1, 2015

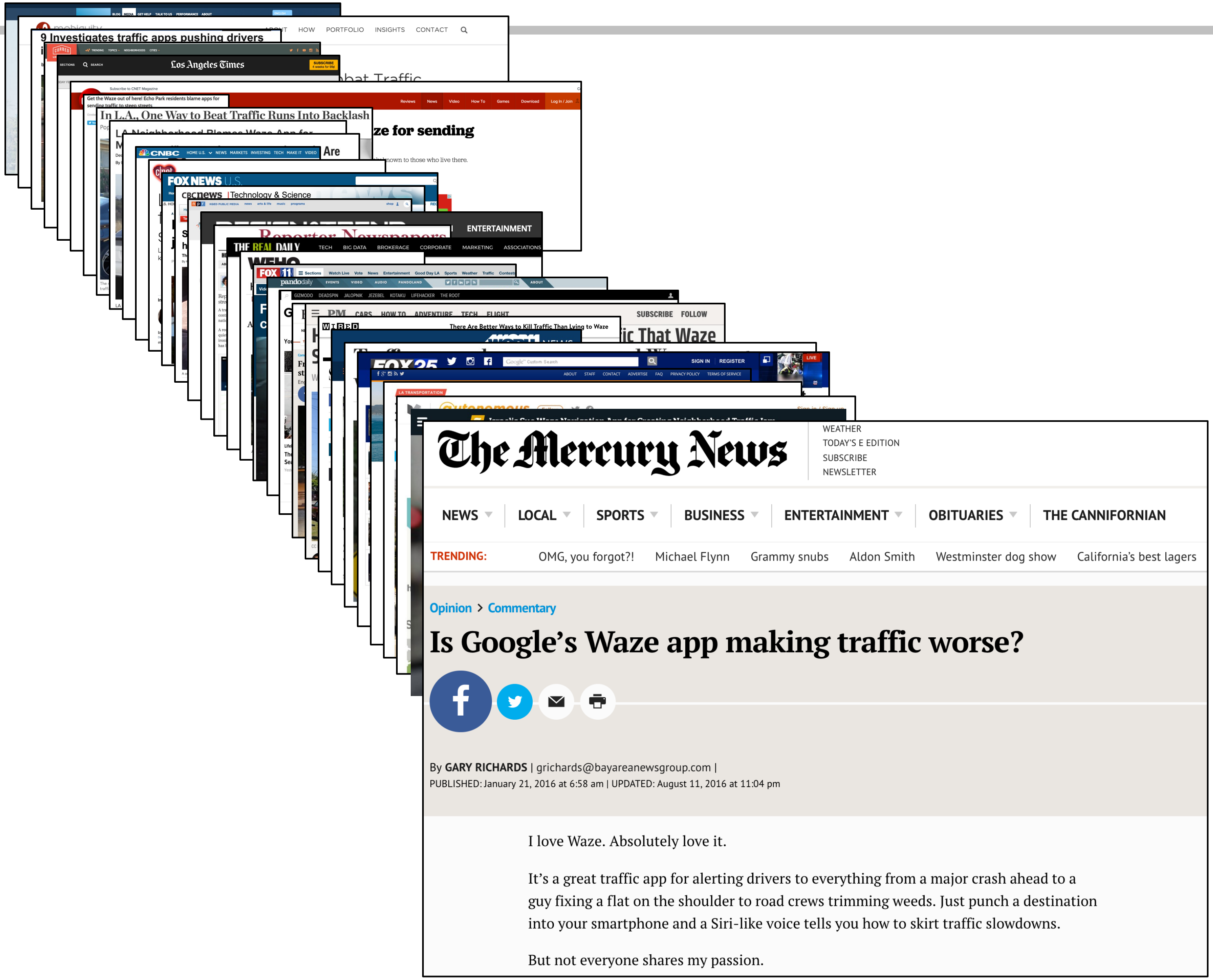
Neighborhoods and cities start to resist



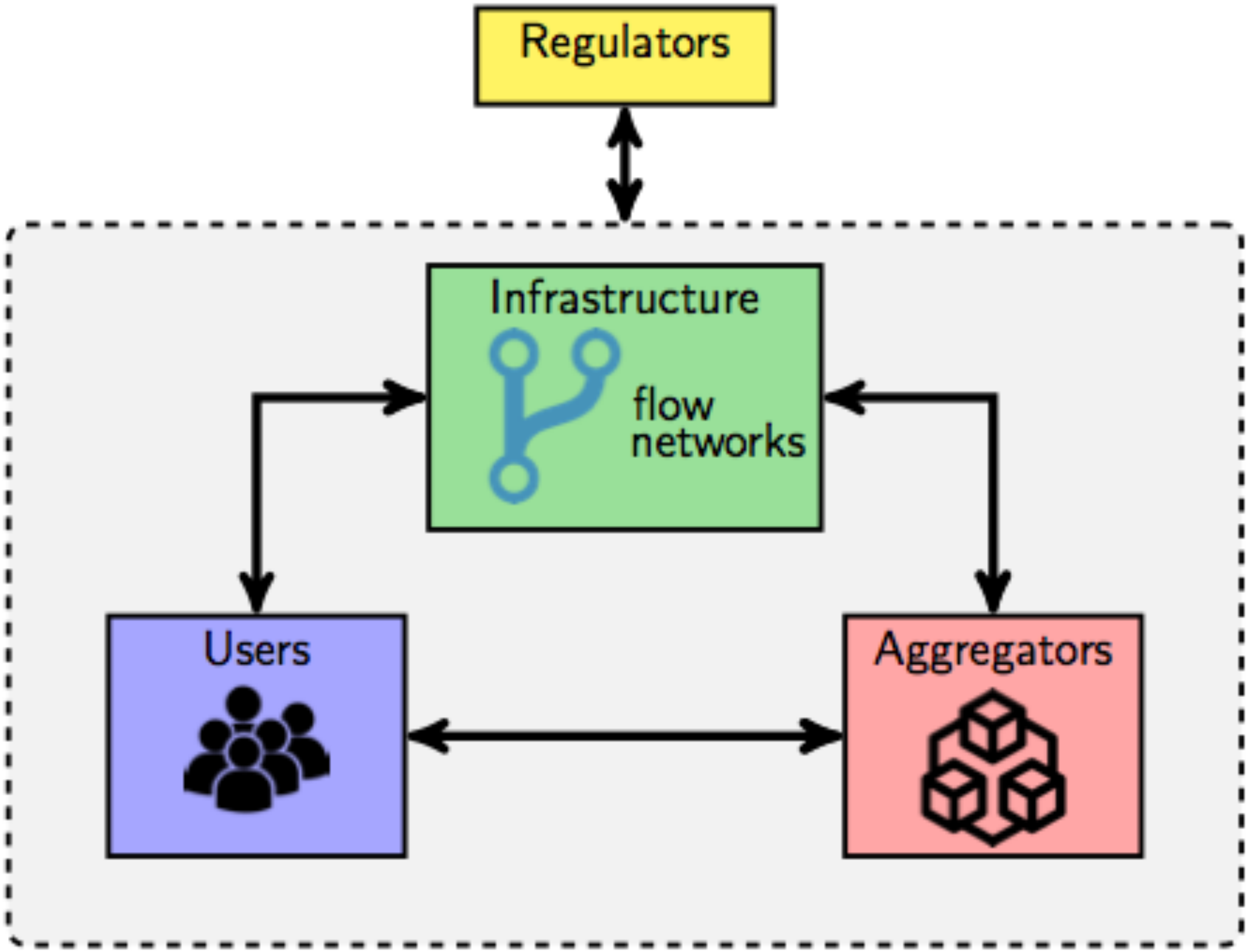
No real policy to help elected officials



But few people are asking the right question



Emerging Data Market—Regulation & Policy



Challenge:

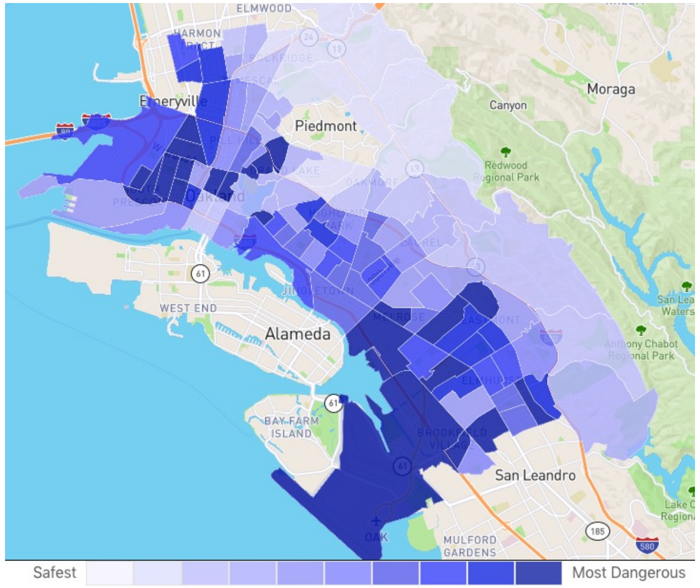
Humans adapt their behavior to AI/ML systems

Addressing this requires closing the loop in Machine Learning

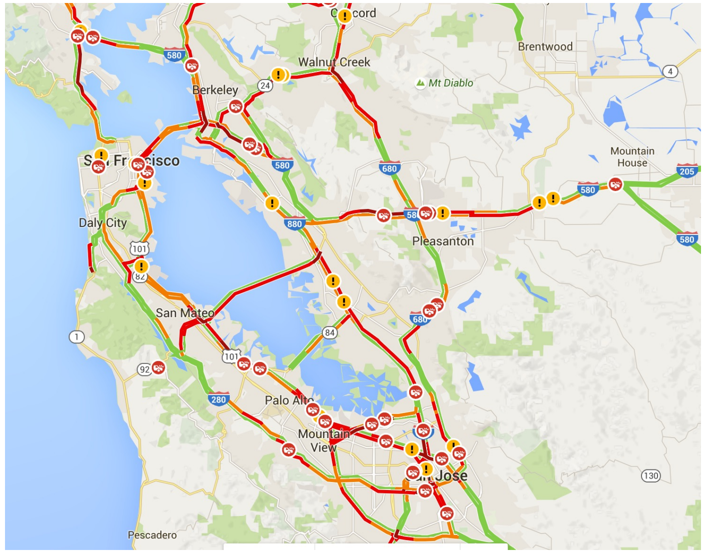
Intelligent Systems Require Rethinking ML

A Central Tenet of Classical ML

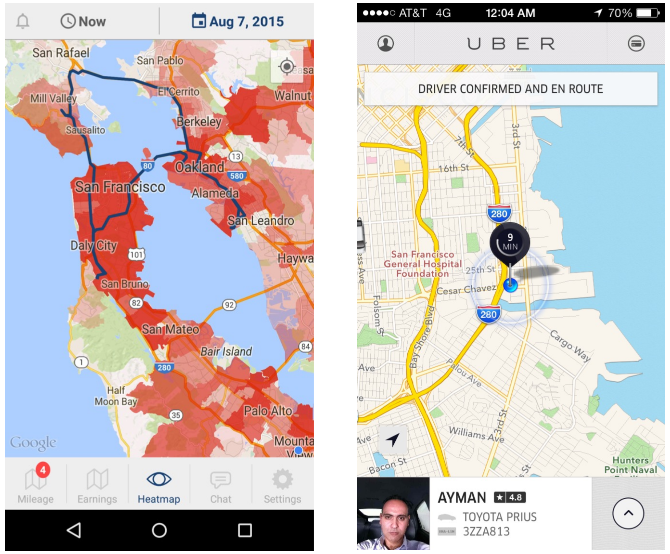
Classical ML assumes the past is representative of the future: When it is arduous to model a real phenomena, observations thereof are representative samples from static distribution



Oakland Safety Index



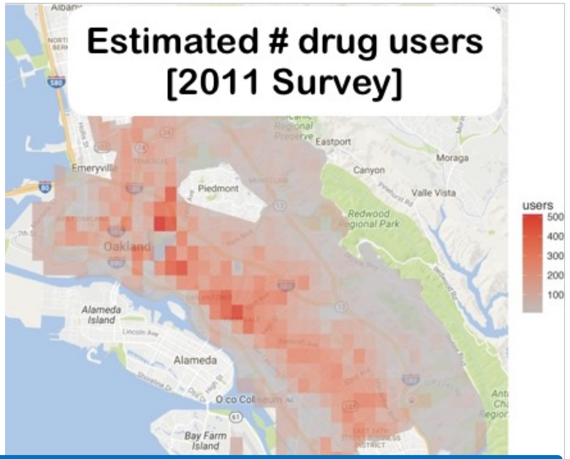
Bay Area Traffic



Ride-Share Supply/Demand

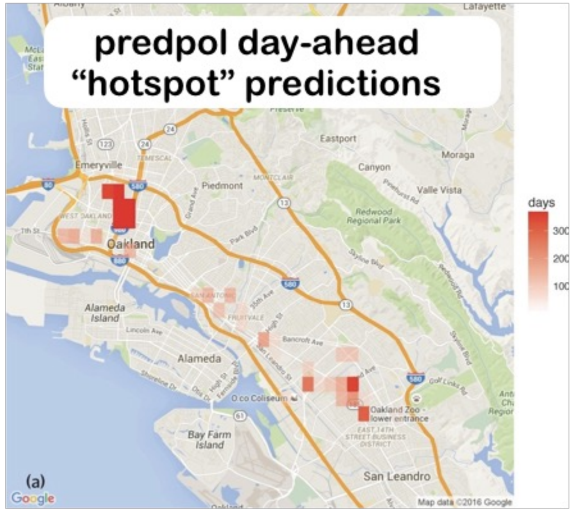
Unintended Consequence: Feedback Reinforced Bias

Actual drug arrests: concentrated in “hotspot”

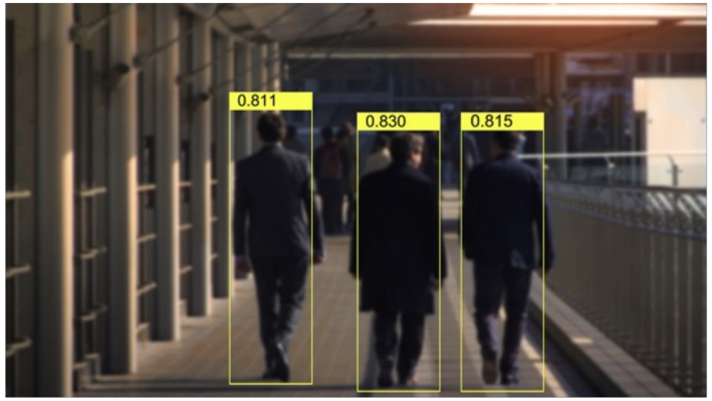


Estimated drug use: **not** concentrated in “hotspot”

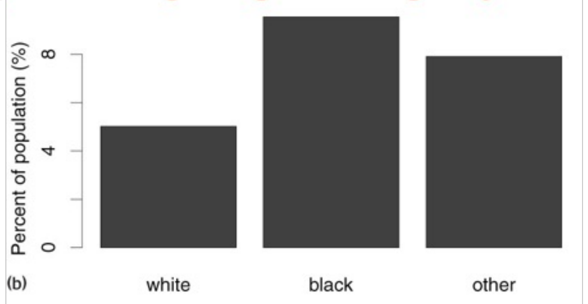
PredPol
Predict Crime in **Real Time**

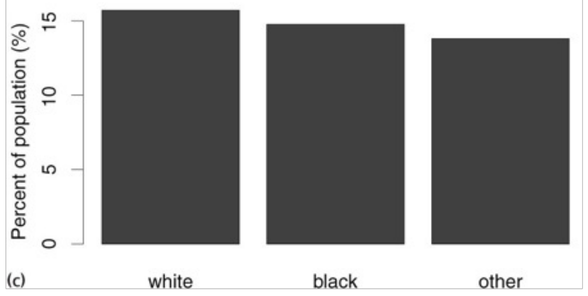


Take-away: Ill-designed predictive policing algorithms can reinforce institutional bias



police targeting for drugs by race





estimated drug use by race

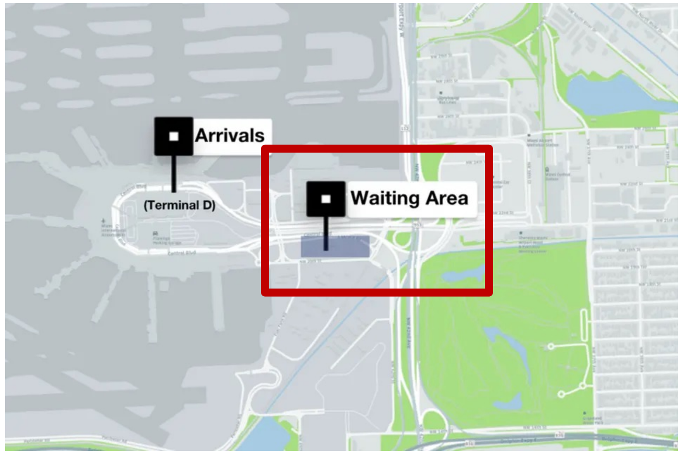
”To Predict and Serve?” Kristian Lum and William Isaac, Royal Statistical Society, 2016

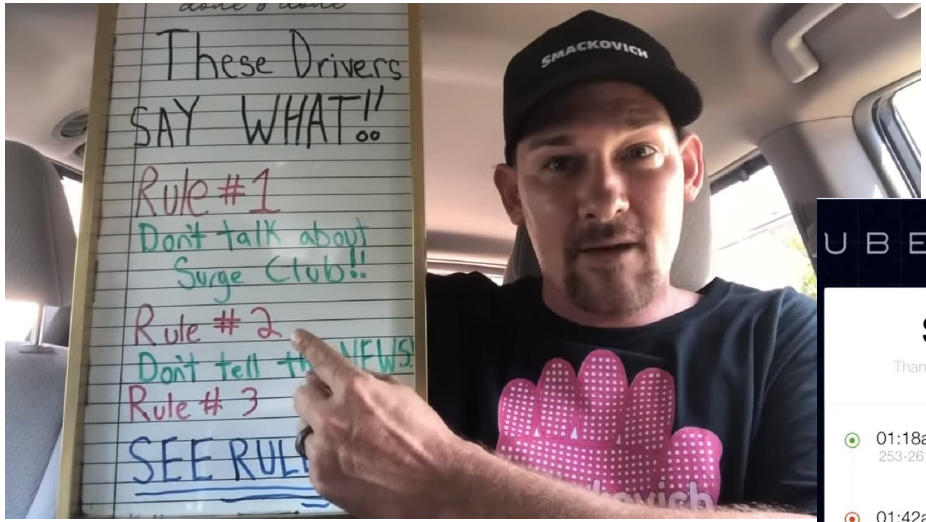
Unmodeled Strategic Behavior: Collusion Triggered Inequities

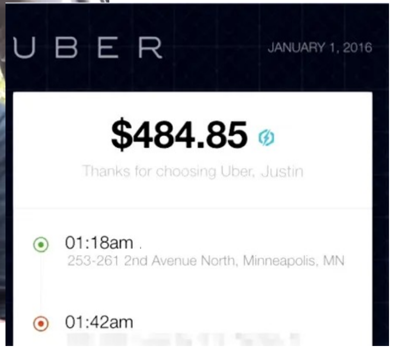
HOME > TECH

Uber drivers are reportedly colluding to trigger 'surge' prices because they say the company is not paying them enough

Isobel Asher Hamilton Jun 14, 2019, 5:19 AM





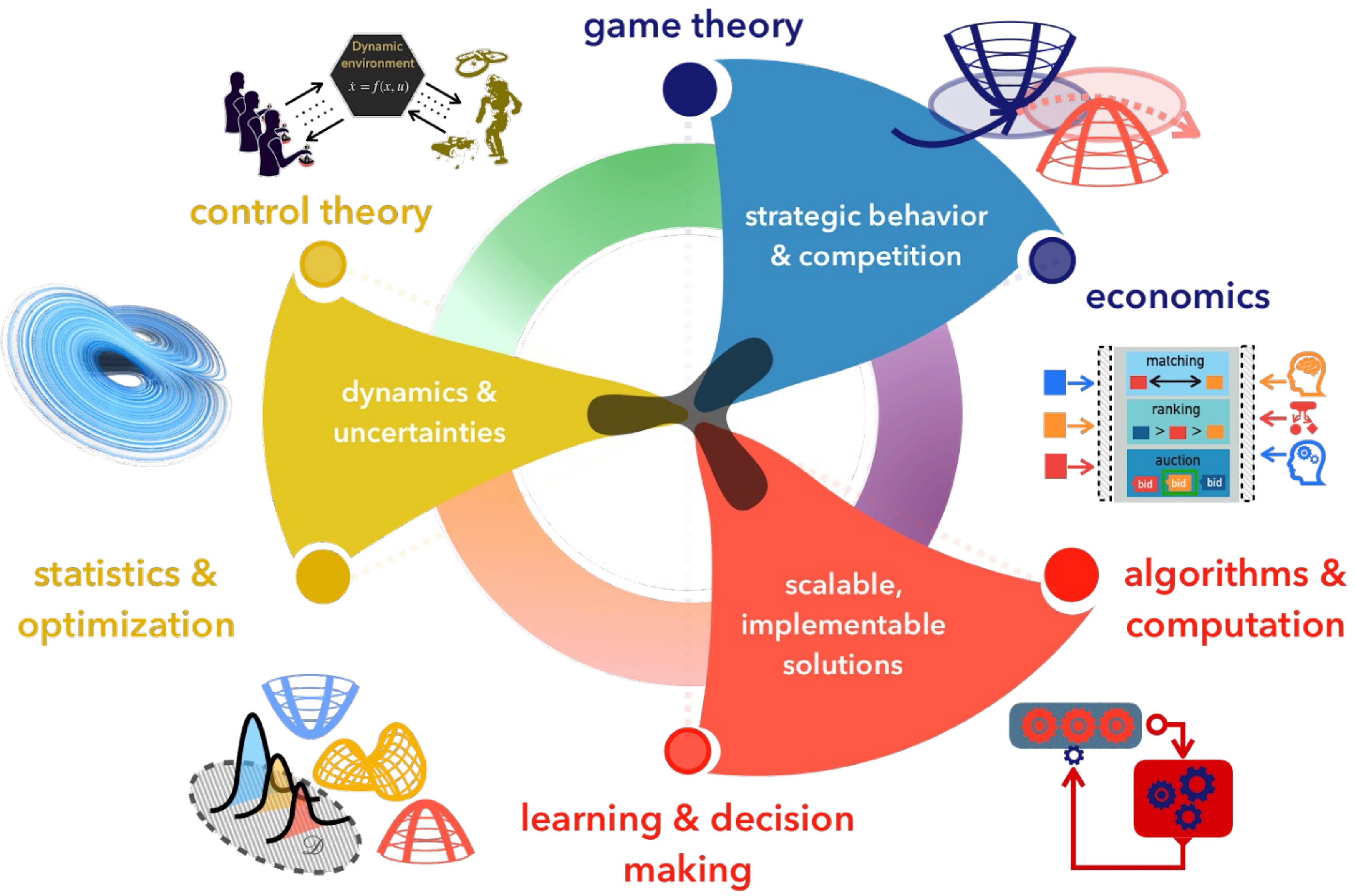


"Surge" Club

- Drivers "caught" colluding to trigger surge prices in high demand locations
- Unintended consequence: increased prices get offloaded on passenger side of market

Take-away: Ill-designed pricing algorithms can exacerbate inequities

Emerging New Domain: Learning-Enabled Intelligent Systems



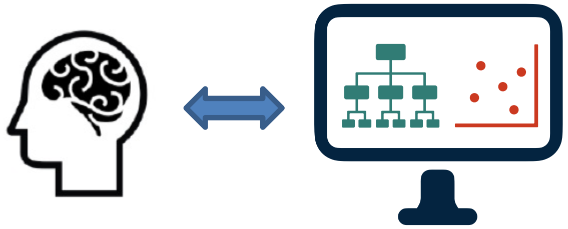
Designing AI/ML-Enabled Systems requires tools from several core domains

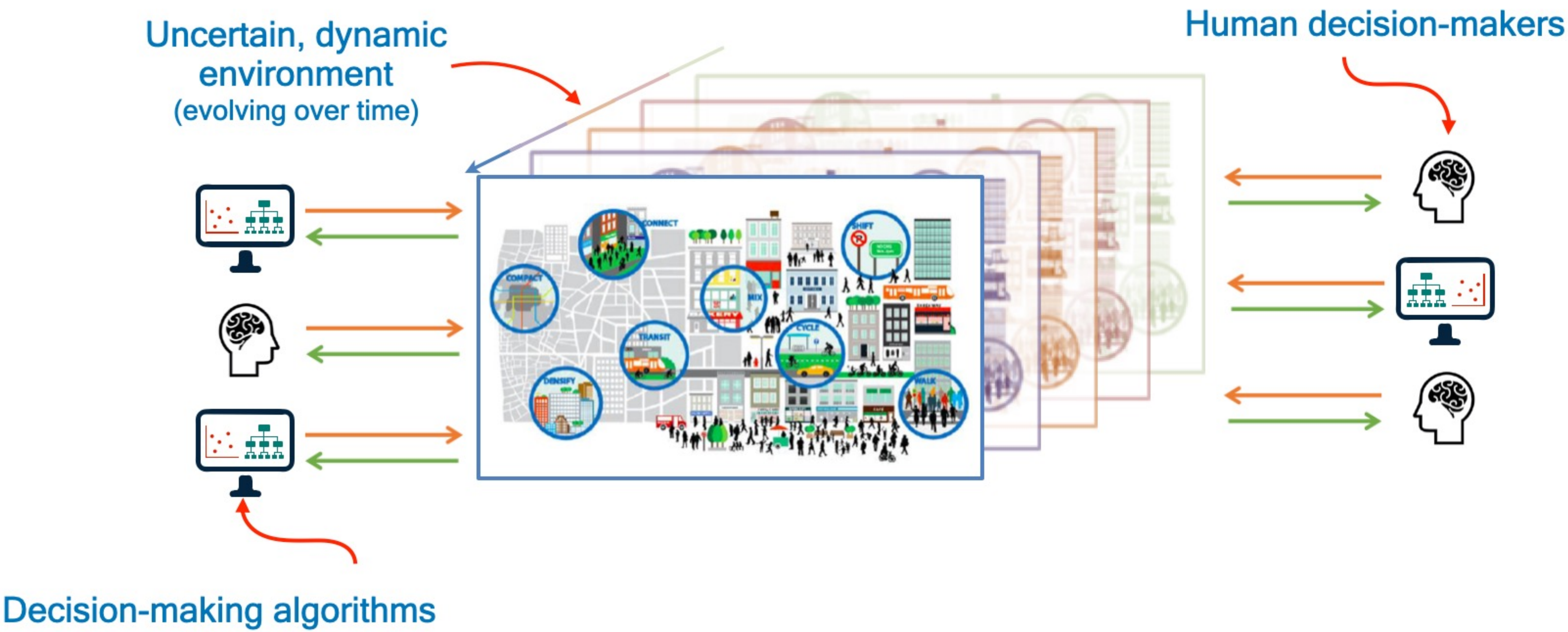
Today's Talk: Results on How to Deal with Humans in the Loop

- Key issues we are addressing
  1. Multiple decision-makers (algorithms) interacting, and potentially competing
  2. Considerations when learning in the presence of dynamically adaptive agents.
  3. Robustness to model misspecification

Insight	Game theoretic abstractions and dynamic models of interaction are crucial in addressing many of these challenges

Age of algorithmic automation is here

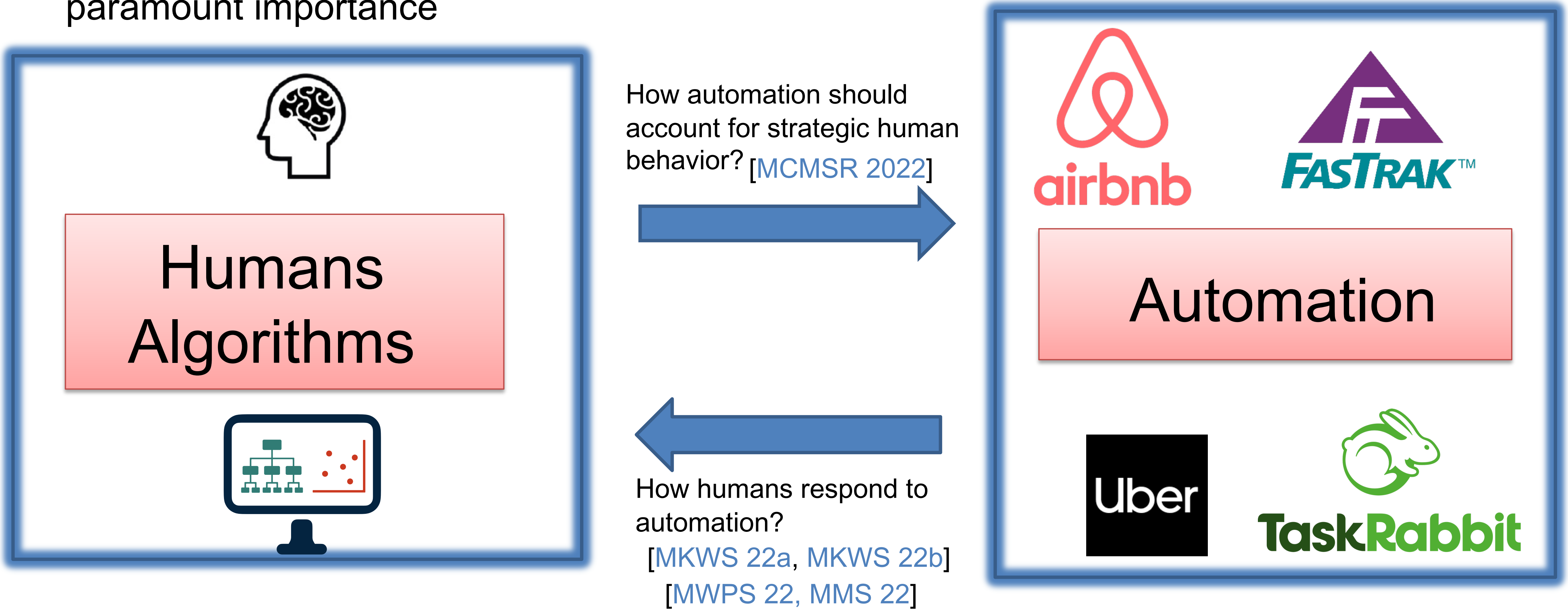




- Humans, algorithms (acting on their behalf) and automation interact with one another in today's societal scale systems. Eg: transporation network, online marketplaces, electric grid, stock exchanges etc.

With the advent of new automation technologies, understanding how humans respond to automation and how automation should respond to humans is of paramount importance

With the advent of new automation technologies, understanding how humans respond to automation and how automation should respond to humans is of paramount importance



[MCMSR 22] Maheshwari C.*, Chiu C-Y*, Mazumdar E, Sastry S, Ratliff L. Zeroth Order Methods for Convex Concave Minmax problems: Applications to Decision Dependent Risk Minimization. Published in proceedings of AISTATS 2022
[MKWS 22a] Maheshwari C*., Kulkarni K*., Wu M., Sastry S. Dynamic Tolling for inducing socially optimal traffic loads. Published in proceedings of ACC 2022
[MKWS 22b] Maheshwari C., Kulkarni K., Wu M., Sastry S. Inducing Social Optimality in Games via Adaptive Incentive Design. To appear in CDC 2022
[MWPS 22] Maheshwari C.*, Wu M.*, Pai D., Sastry S. Independent and Decentralized Learning in Markov Potential Games. Arxiv 2205.14590
[MMS 22] Maheshwari C., Mazumdar E.., Sastry S. Decentralized, Communication and Coordination free learning in Markov Potential Games. Arxiv 2206.02344

Key Vignettes

- Vignette 1: How to align societal objectives with selfish objectives by suitably modifying the incentives of humans / algorithms (acting on their behalf) participating in a societal scale system? [MKWS 22a, MKWS 22b]

- Vignette 2: How does humans /algorithms (acting on their behalf), who act independently and in a decentralized manner, make decisions so as to ensure “stability” in the system? [MWPS 22, MMS 22]

- Vignette 3: How to ensure societal scale systems be robust to strategic behavior of humans/ algorithms (acting on their behalf)? [MCMSR 2022]

[MCMSR 22] Maheshwari C.*, Chiu C-Y*, Mazumdar E, Sastry S, Ratliff L. Zeroth Order Methods for Convex Concave Minmax problems: Applications to Decision Dependent Risk Minimization. Published in proceedings of AISTATS 2022

[MKWS 22a] Maheshwari C*., Kulkarni K*., Wu M., Sastry S. Dynamic Tolling for inducing socially optimal traffic loads. Published in proceedings of ACC 2022

[MKWS 22b] Maheshwari C., Kulkarni K., Wu M., Sastry S. Inducing Social Optimality in Games via Adaptive Incentive Design. To appear in CDC 2022

[MWPS 22] Maheshwari C.*, Wu M.*, Pai D., Sastry S. Independent and Decentralized Learning in Markov Potential Games. Arxiv 2205.14590

[MMS 22] Maheshwari C., Mazumdar E.., Sastry S. Decentralized, Communication and Coordination free learning in Markov Potential Games. Arxiv 2206.02344

Two Vignettes Today

- Vignette 1: How to align societal objectives with selfish objectives by suitably modifying the incentives of humans / algorithms (acting on their behalf) participating in a societal scale system? [MKWS 22a, MKWS 22b]

- Vignette 2: How does humans /algorithms (acting on their behalf), who act independently and in a decentralized manner, make decisions so as to ensure "stability" in the system? [MWPS 22, MMS 22]

- Vignette 3: How to ensure societal scale systems be robust to strategic behavior of humans/ algorithms (acting on their behalf)? [MCMSR 2022]

[MCMSR 22] Maheshwari C.*, Chiu C-Y*, Mazumdar E, Sastry S, Ratliff L. Zeroth Order Methods for Convex Concave Minmax problems: Applications to Decision Dependent Risk Minimization. Published in proceedings of AISTATS 2022

[MKWS 22a] Maheshwari C*., Kulkarni K*., Wu M., Sastry S. Dynamic Tolling for inducing socially optimal traffic loads. Published in proceedings of ACC 2022

[MKWS 22b] Maheshwari C., Kulkarni K., Wu M., Sastry S. Inducing Social Optimality in Games via Adaptive Incentive Design. To appear in CDC 2022

[MWPS 22] Maheshwari C.*, Wu M.*, Pai D., Sastry S. Independent and Decentralized Learning in Markov Potential Games. Arxiv 2205.14590

[MMS 22] Maheshwari C., Mazumdar E.., Sastry S. Decentralized, Communication and Coordination free learning in Markov Potential Games. Arxiv 2206.02344

Vignette 1: Dynamic Tolling for Inducing Socially Optimal Traffic Loads

Published in the American Control Conference 2022

Transportation

Virginia begins last piece of Beltway toll lanes expansion, reaching the American Legion Bridge

The $600 million project will widen the Beltway in one of the Washington region's busiest corridors

VIEWPOINTS

COLUMBIA CLIMATE SCHOOL
Climate, Earth, and Society

SUSTAINABILITY, URBANIZATION

Congestion Pricing is Slowly Coming to New York City

BY STEVE COHEN | OCTOBER 4, 2021

4 Comments

Los Angeles Times

BUSINESS

Traffic is terrible again. Here's how to get it closer to spring 2020 levels

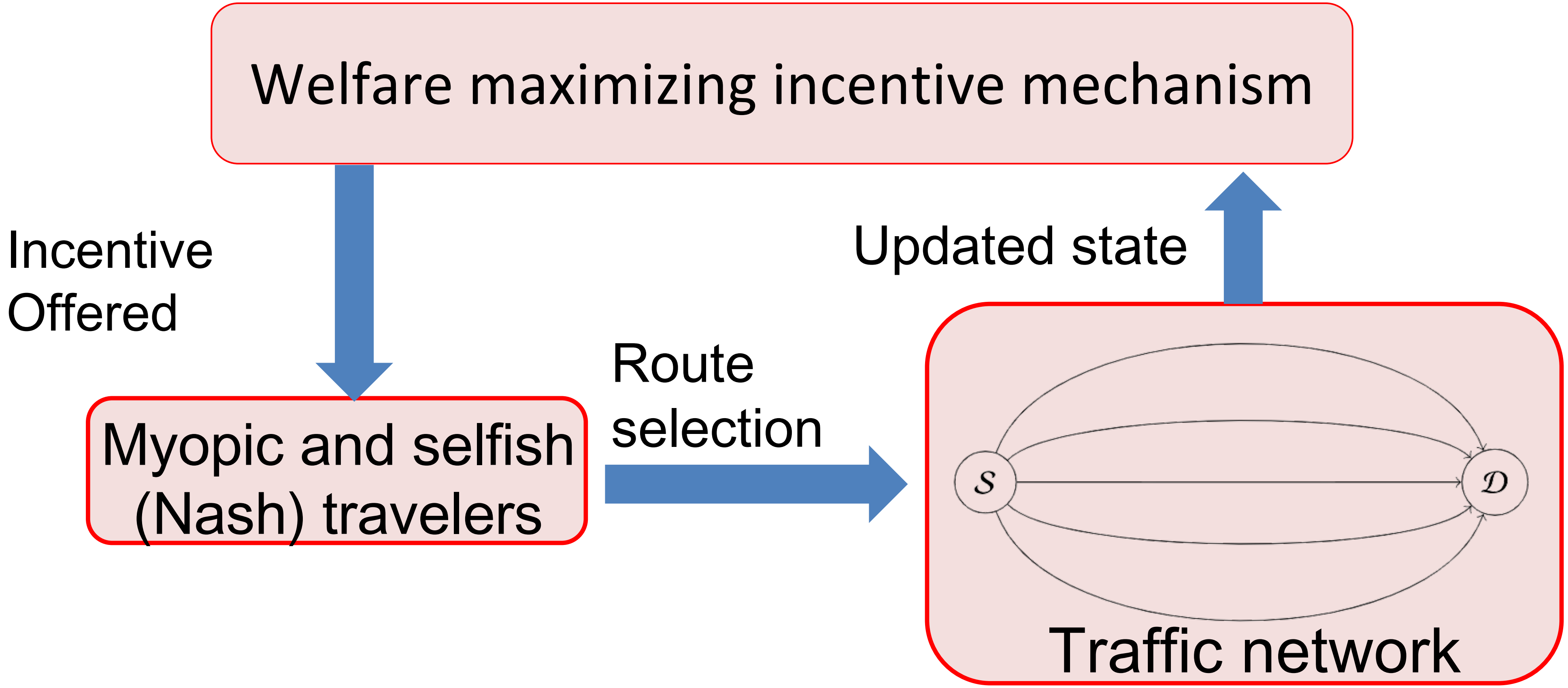
The Washington Post
Democracy Dies in Darkness

D.C. is working on a futuristic plan: Less parking, taller buildings and a transformed city

By Julie Zauzmer Weil

May 2, 2021 at 9:00 a.m. EDT

Societal Problem → Modeling



- ❖ Stochastic arrival and departure of travelers updates the state of congestion on the traffic network
- ❖ A central planner who wants to minimize the overall congestion on the network levies tolls on the travelers

Overview

Key Question

How to design toll prices on a traffic network which

1. account for dynamically changing congestion levels due to incoming and outgoing traffic comprised of myopic and selfish travelers?
2. ensure that eventually the congestion levels are socially optimal
3. are economically motivated

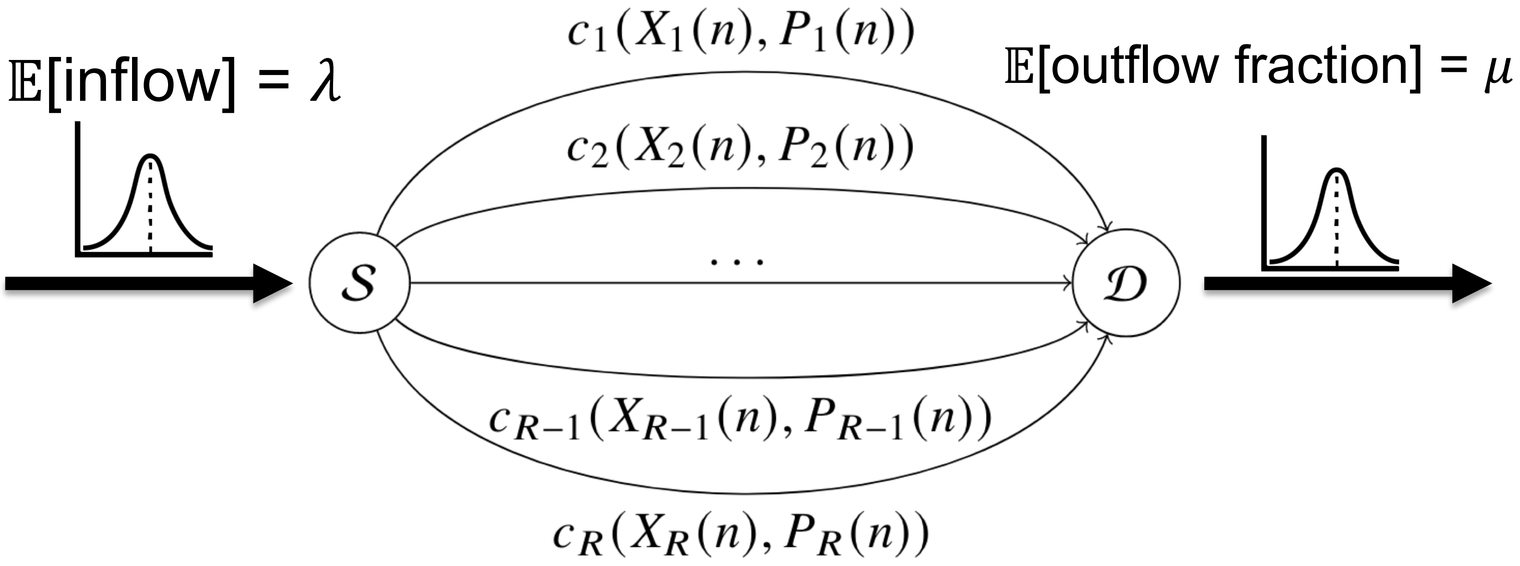
Key features of the proposed approach

1. The toll prices are updated at a slower timescale than the dynamically changing congestion levels
2. The toll prices are updated based on marginal increment in travel time at the current congestion levels

Discrete time stochastic dynamics

Notation:

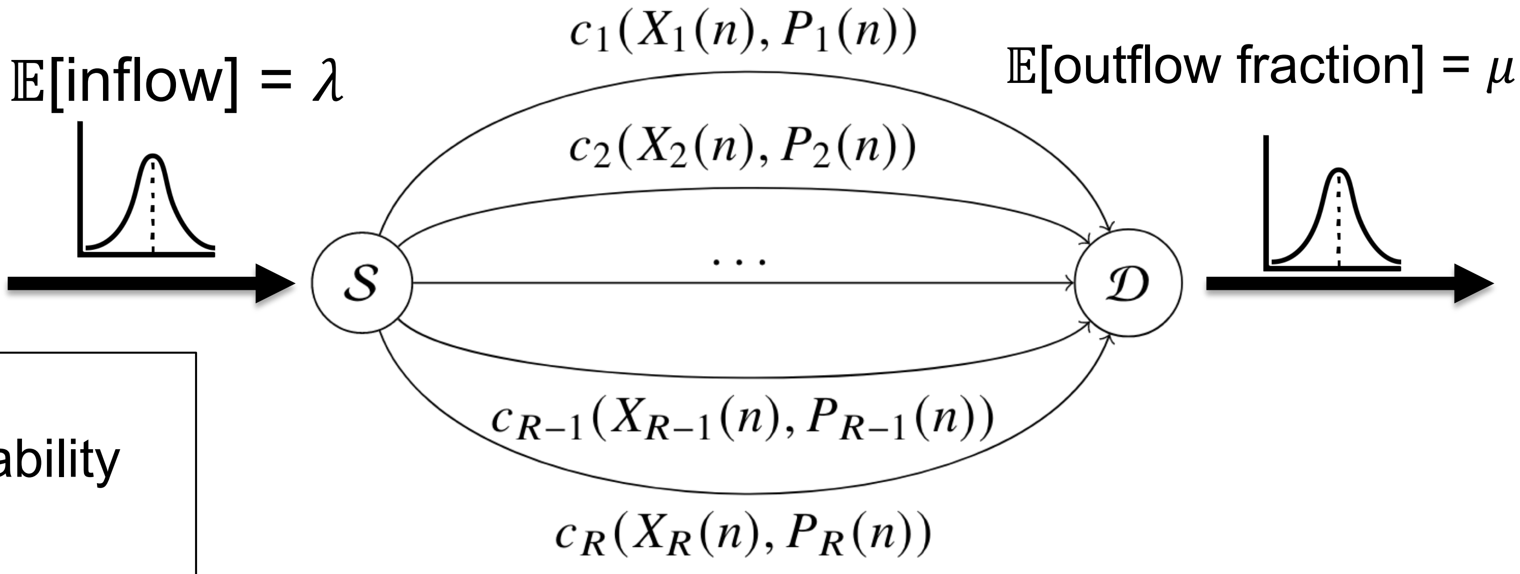
$X_i(n)$: Congestion on link i at time n
$P_i(n)$: Toll price on link i at time n
c_i : cost of using link i



Discrete time stochastic dynamics

Notation:

$X_i(n)$: Congestion on link i at time n
$P_i(n)$: Toll price on link i at time n
c_i : cost of using link i



Perturbed best response

Incoming traffic chooses link i with probability

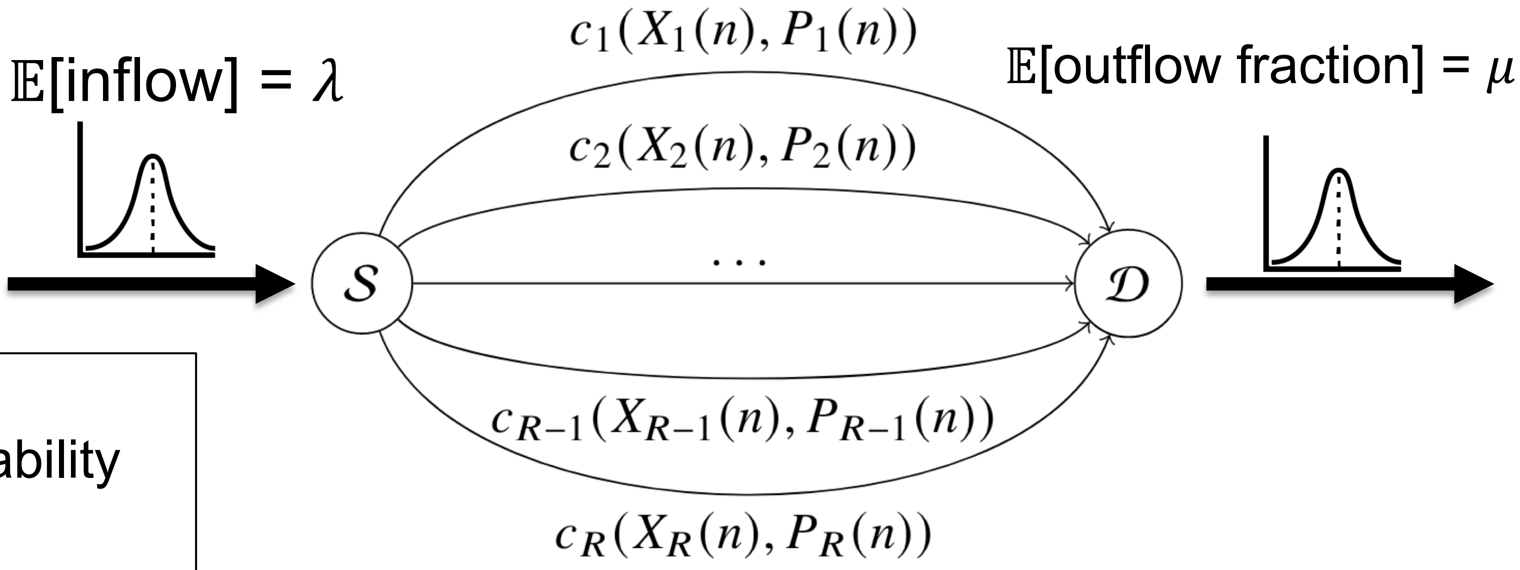
$$\frac{\exp(-\beta c_i(x_i, p_i))}{\sum_{j=1}^{R} \exp(-\beta c_j(x_j, p_j))}$$

for some $\beta > 0$

Notation:

$X_i(n)$: Congestion on link i at time n
$P_i(n)$: Toll price on link i at time n
c_i : cost of using link i

Discrete time stochastic dynamics



Perturbed best response

Incoming traffic chooses link i with probability

$$\frac{\exp(-\beta c_i(x_i, p_i))}{\sum_{j=1}^{R} \exp(-\beta c_j(x_j, p_j))}$$

for some $\beta > 0$

System state update: $X_i(n+1) = X_i(n) + \mu\left(\frac{\lambda}{\mu}\frac{\exp(-\beta c_i(x_i,p))}{\sum_j \exp(-\beta c_j(x_j,p))} - X_i(n)\right) + \mu M_i(n+1)$

Price update: $P_i(n+1) = (1-a)P_i(n) + \boxed{\boldsymbol{a X_i(n) d\ell(X_i(n))/dx}}$ Optimal prices

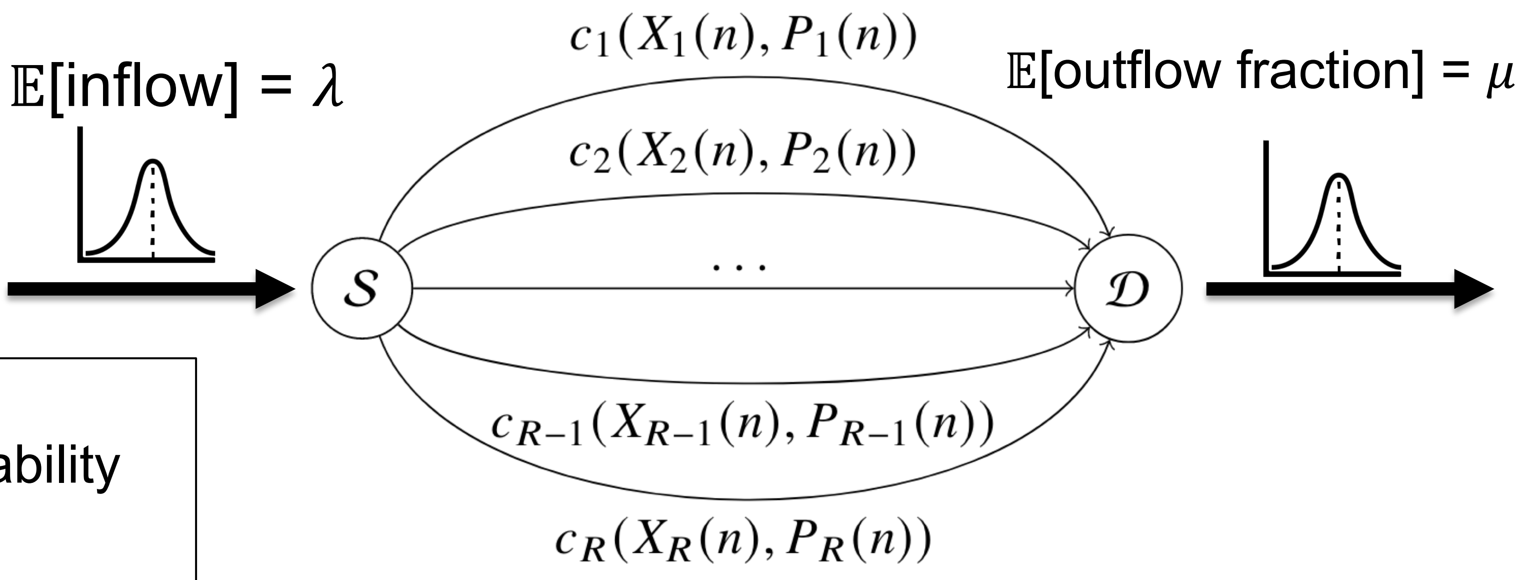
Notation:

$X_i(n)$: Congestion on link i at time n

$P_i(n)$: Toll price on link i at time n

c_i : cost of using link i

Discrete time stochastic dynamics



Perturbed best response

Incoming traffic chooses link i with probability

$$\frac{\exp(-\beta c_i(x_i, p_i))}{\sum_{j=1}^{R} \exp(-\beta c_j(x_j, p_j))}$$

for some $\beta > 0$

$c_i(X_i(n), P_i(n)) = \ell_i(X_i(n)) + P_i(n)$

Two timescale

$$\epsilon = \frac{a}{\mu} \ll 1$$

Martingale noise

System state update: $X_i(n+1) = X_i(n) + \mu\left(\frac{\lambda}{\mu}\frac{\exp(-\beta c_i(x_i,p))}{\sum_j \exp(-\beta c_j(x_j,p))} - X_i(n)\right) + \mu M_i(n+1)$

Price update: $P_i(n+1) = (1-a)P_i(n) + \boxed{\boldsymbol{a X_i(n) d\ell(X_i(n))/dx}}$ Optimal prices

Convergence Theorem

System state update: $X_i(n+1) = X_i(n) + \mu\left(\frac{\lambda}{\mu}\frac{\exp(-\beta c_i(x_i,p))}{\sum_j \exp(-\beta c_j(x_j,p))} - X_i(n)\right) + \mu M_i(n+1)$

Price update: $P_i(n+1) = (1-a)P_i(n) + a\boldsymbol{X_i(n)}\boldsymbol{d}\ell(\boldsymbol{X_i(n)})/\boldsymbol{dx}$

Theorem: As $\beta \to \infty$, the congestion levels and the toll prices converge in a neighborhood of socially optimal levels $\bar{x}, \bar{p}$

$$\limsup_{n\to\infty} E\left[\left|\left|X(n) - \bar{x}\right|\right|^2 + \left|\left|P(n) - \bar{p}\right|\right|^2\right] \leq O(\mu) + O(a/\mu)$$

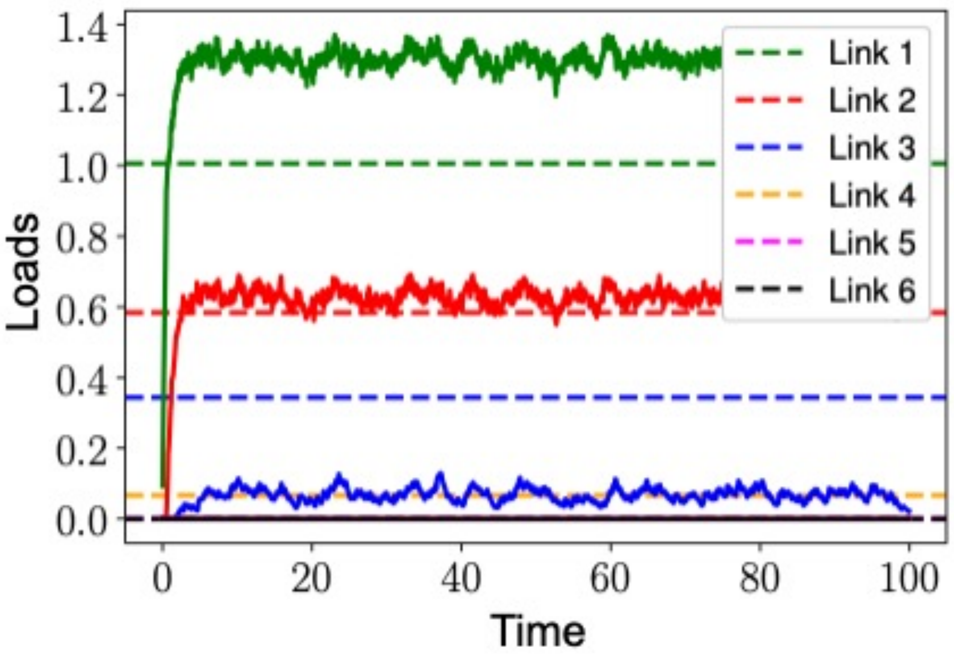
Proof Technique

- The proof is based on using two-timescale constant step-size stochastic approximation
- It is sufficient to analyze the asymptotic behavior of continuous time dynamical system
- Convergence of trajectories is established by using cooperative dynamical systems theory and variational inequality based analysis of perturbed equilibrium.

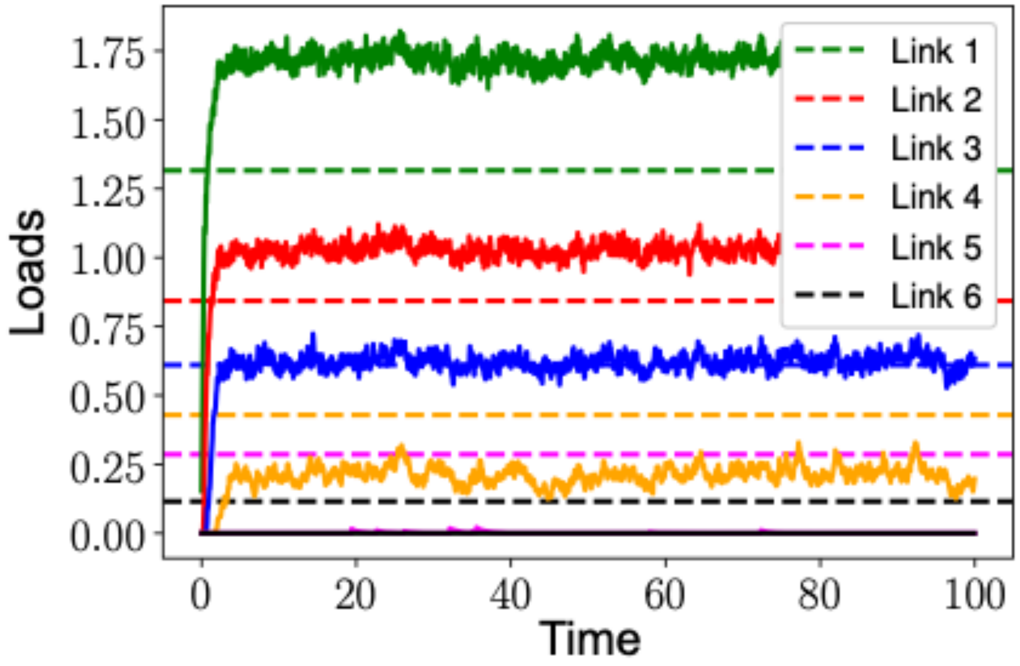
Results extend to general network with routing decisions made at source nodes.

Experiments: No tolls

- Consider quadratic costs $\ell_i(x) = ix^2 + i$
- We first plot the discrete time update and the socially optimal load levels with no tolls ($a = 0$) for $R = 6$ and $\beta = 100$.



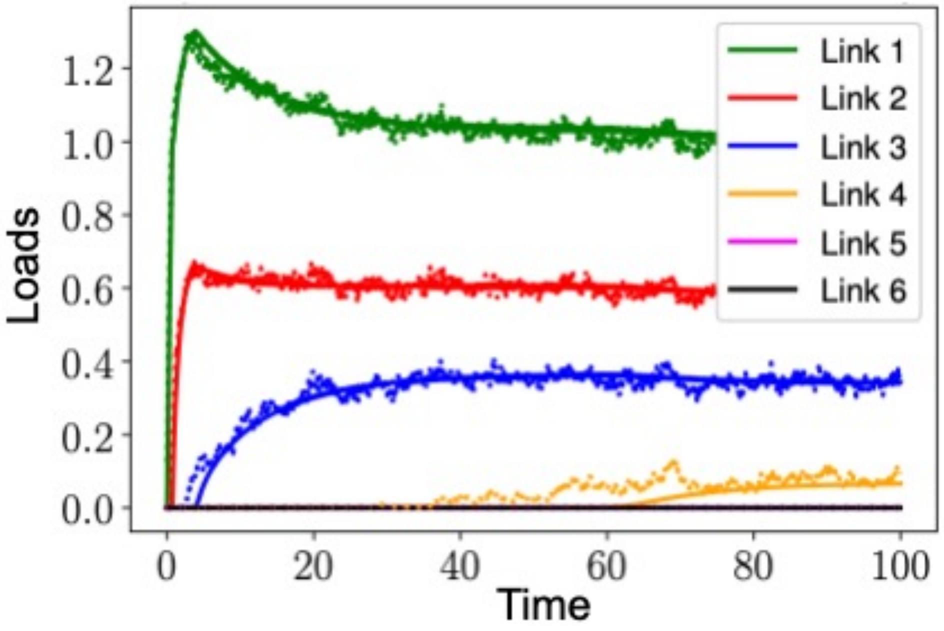
$\lambda = 0.1, \mu = 0.05$



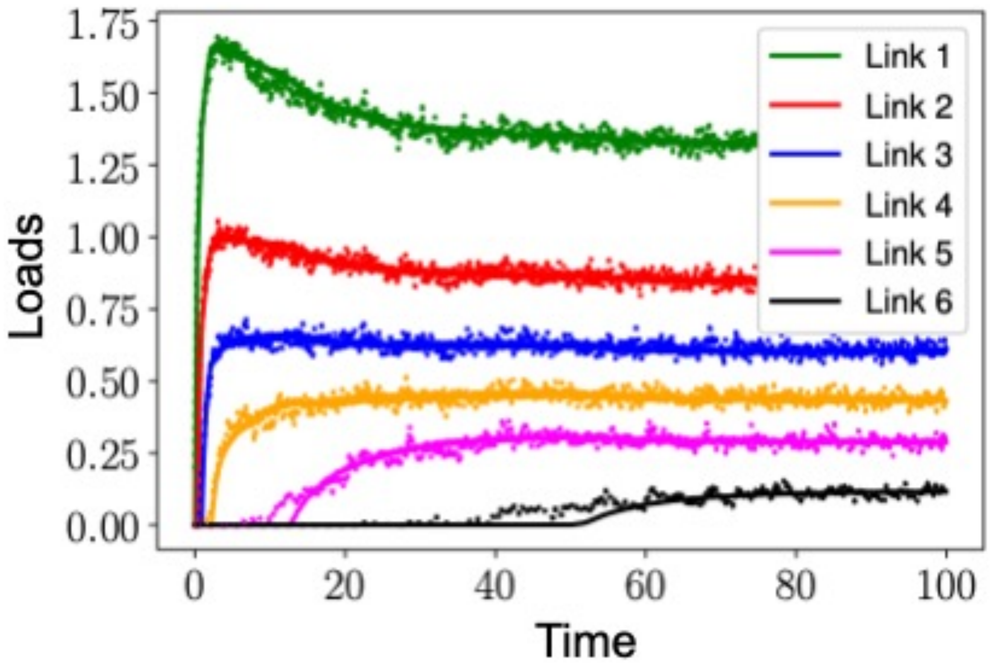
$\lambda = 0.2, \mu = 0.05$

Experiments: With tolls

- For the same quadratic costs $\ell_i(x) = ix^2 + i$
- We first plot the discrete time update and the continuous time dynamical system with no tolls ($a = 0.0015$) for $R = 6$ and $\beta = 100$.



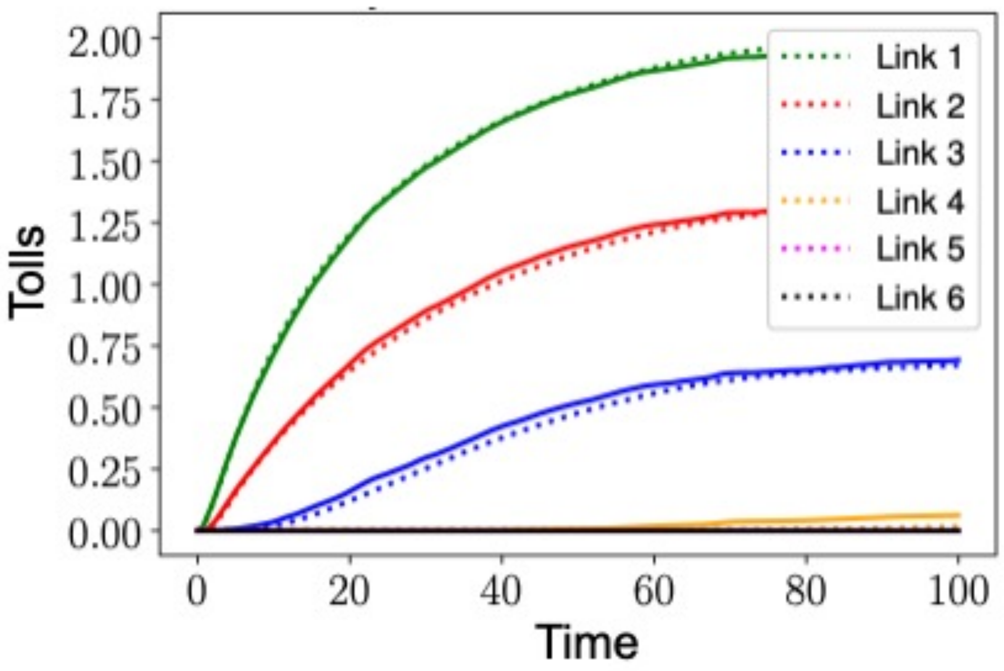
$\lambda = 0.1, \mu = 0.05$



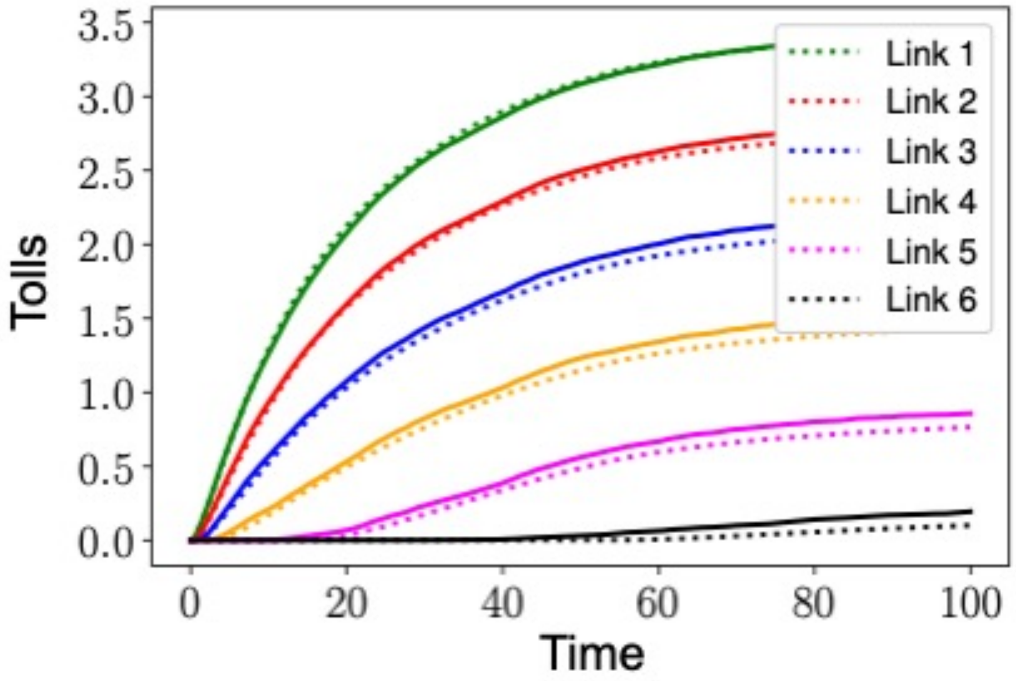
$\lambda = 0.2, \mu = 0.05$

Experiments: Toll update

- We also plot the toll update for $a = 0.0015$. We see that the tolls updates <u>slowly</u> as compared to the loads, and reach their equilibria.



$\lambda = 0.1, \mu = 0.05$



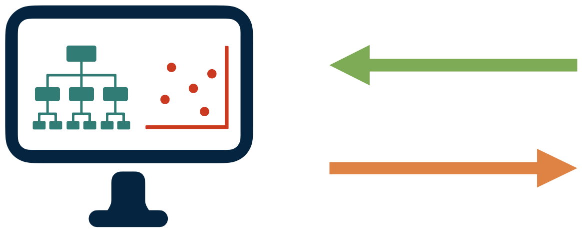
$\lambda = 0.2, \mu = 0.05$

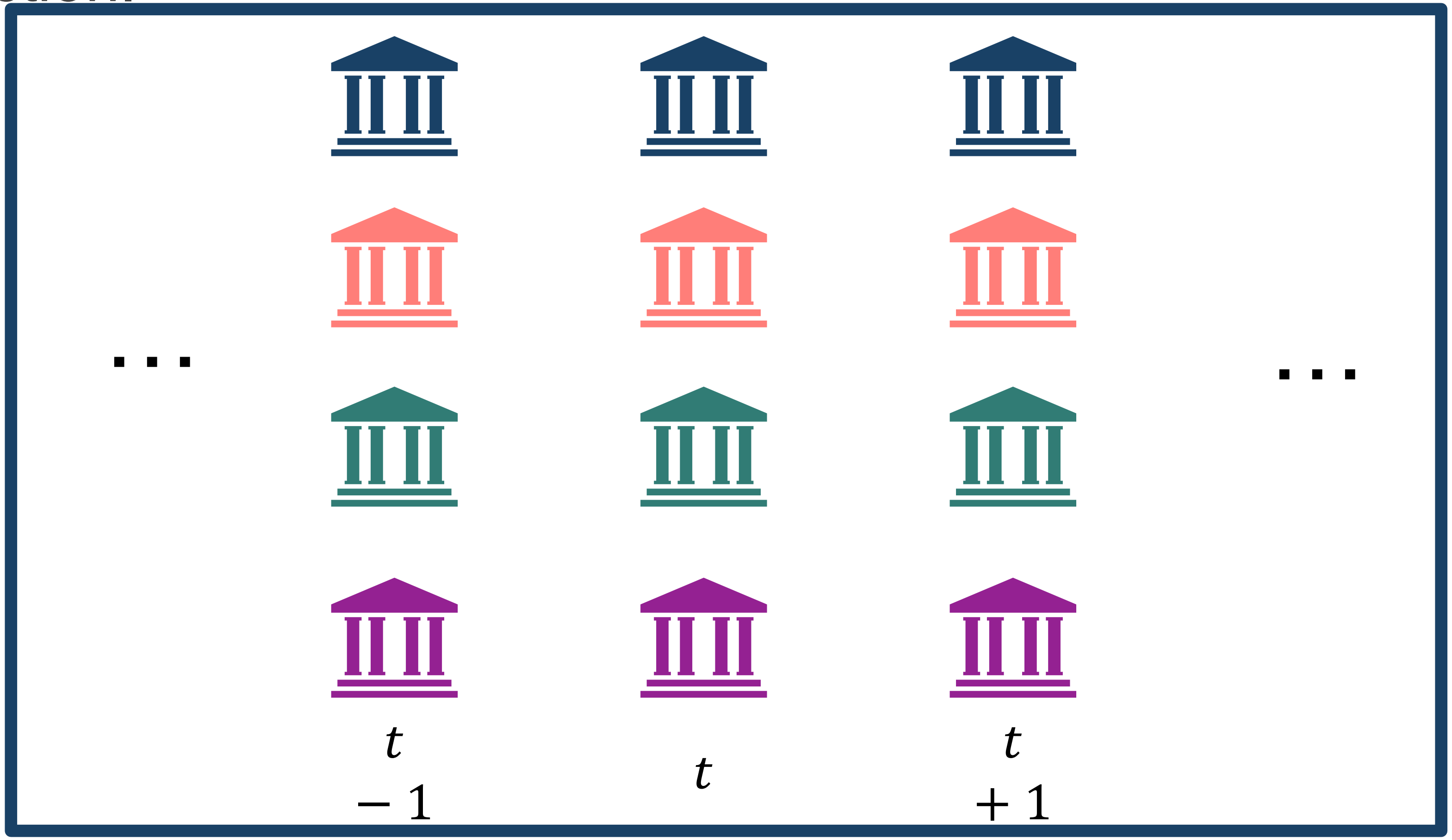
Vignette 2.1:
Decentralized Communication and Coordination free learning in matching markets

To appear in Neurips 22. Available at https://arxiv.org/abs/2206.02344

Learning through interaction

Classically, machine learning has looked at how single agents can learn about their preferences through interaction.

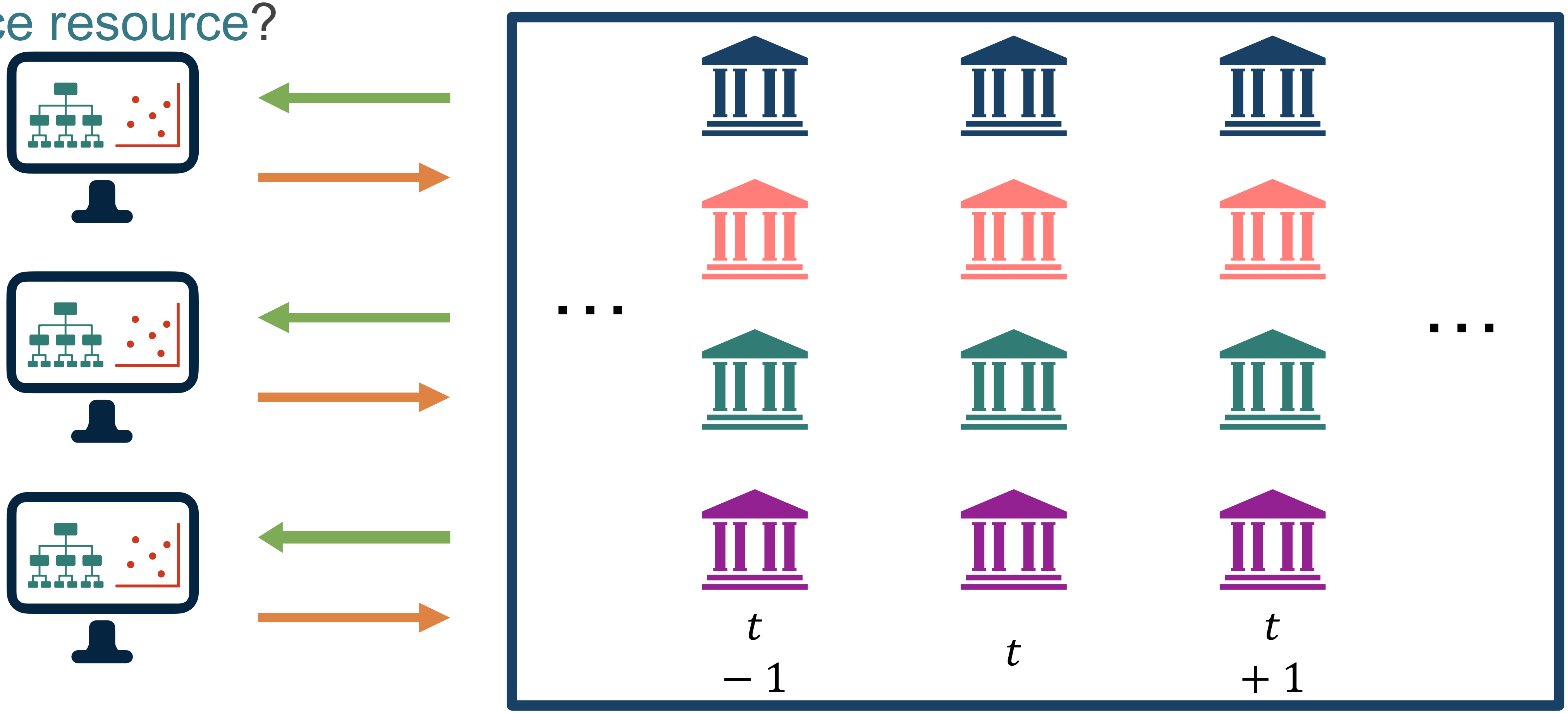




- Exploration/Exploitation Trade-offs.
- Optimal Algorithms

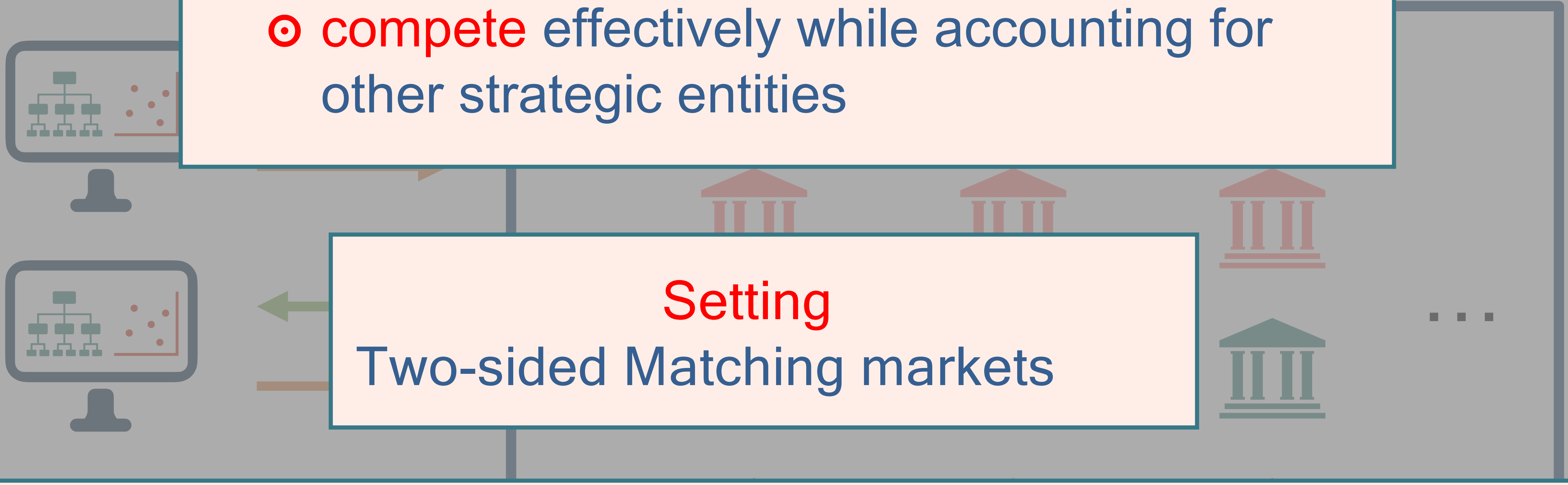
Learning through interaction

- Competition usually leads to externality in the learning process
- What happens if multiple agents compete to learn their preferences for some scarce resource?



•Learning through interaction

Need for algorithms that
- efficiently explore for new information
- exploit the available information
- compete effectively while accounting for other strategic entities



Setting

Two-sided Matching markets

When one side of the market needs to learn to match to a desirable option while other agents are also competing for it

Emerging two sided matching markets

Features

- Resource constraint: Firms cannot get matched with arbitrary agents
- Two-sided preference: Both sides of market have preference over one another

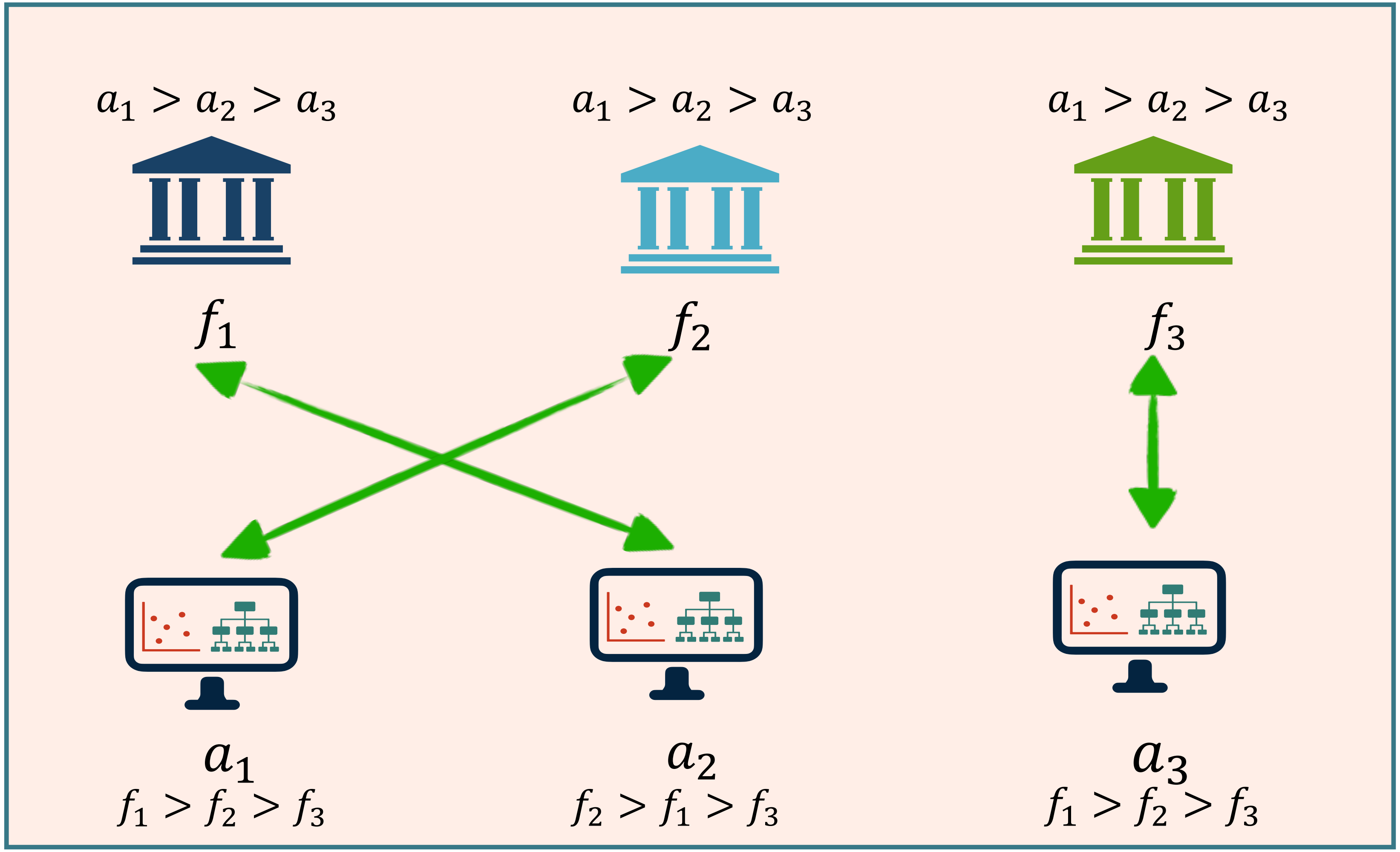
Question

How should agents interact in decentralized manner with no coordination or communication in such markets learn their preferences while accounting competition from other agents for limited resource on other side of market?

Agent $a \in A$

Two-sided matching market

Firm $f \in F$



$$M: A \to F$$

- Injective mapping
- Example

$$M(a_1) = f_1$$

$$M(a_2) = f_2$$

$$M(a_3) = f_3$$

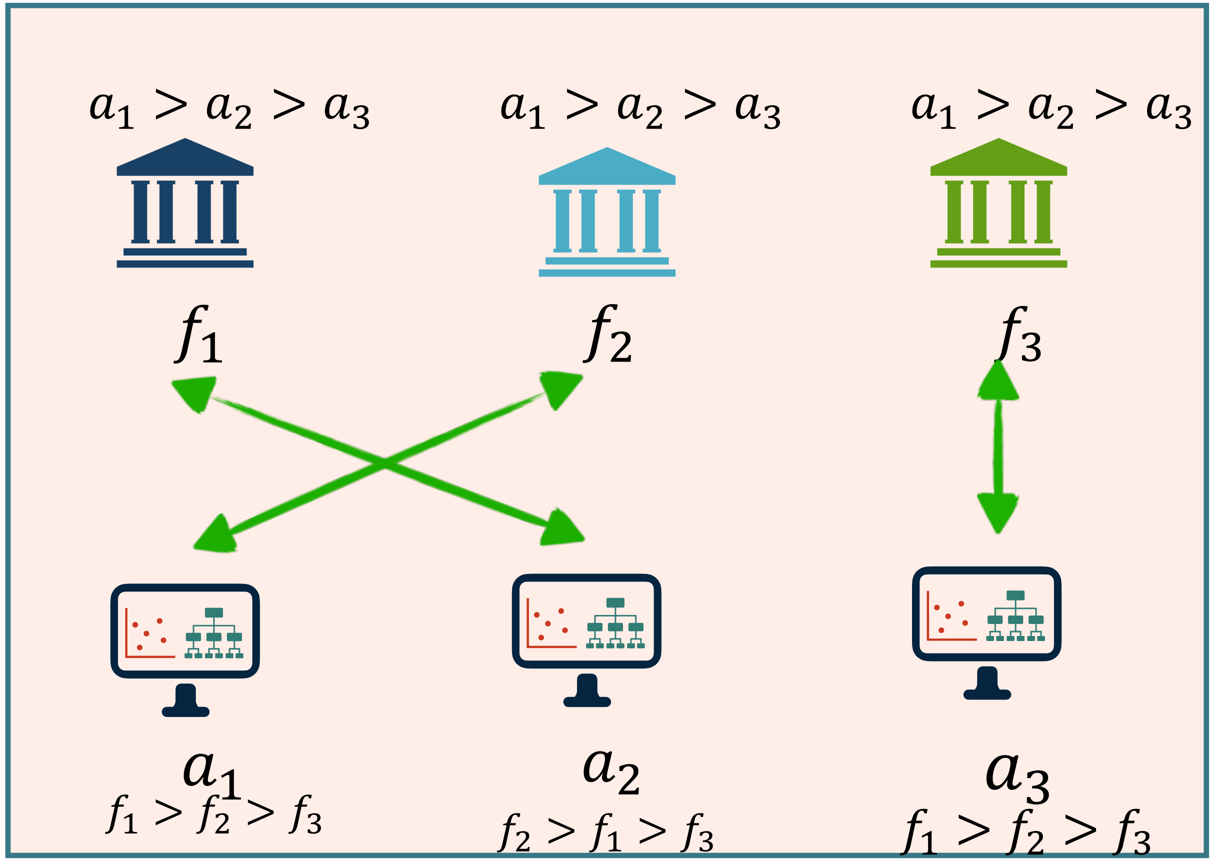
Blocking pair

A tuple (a, f) is a blocking pair with respect to a matching M if both of them prefer each other over their current match

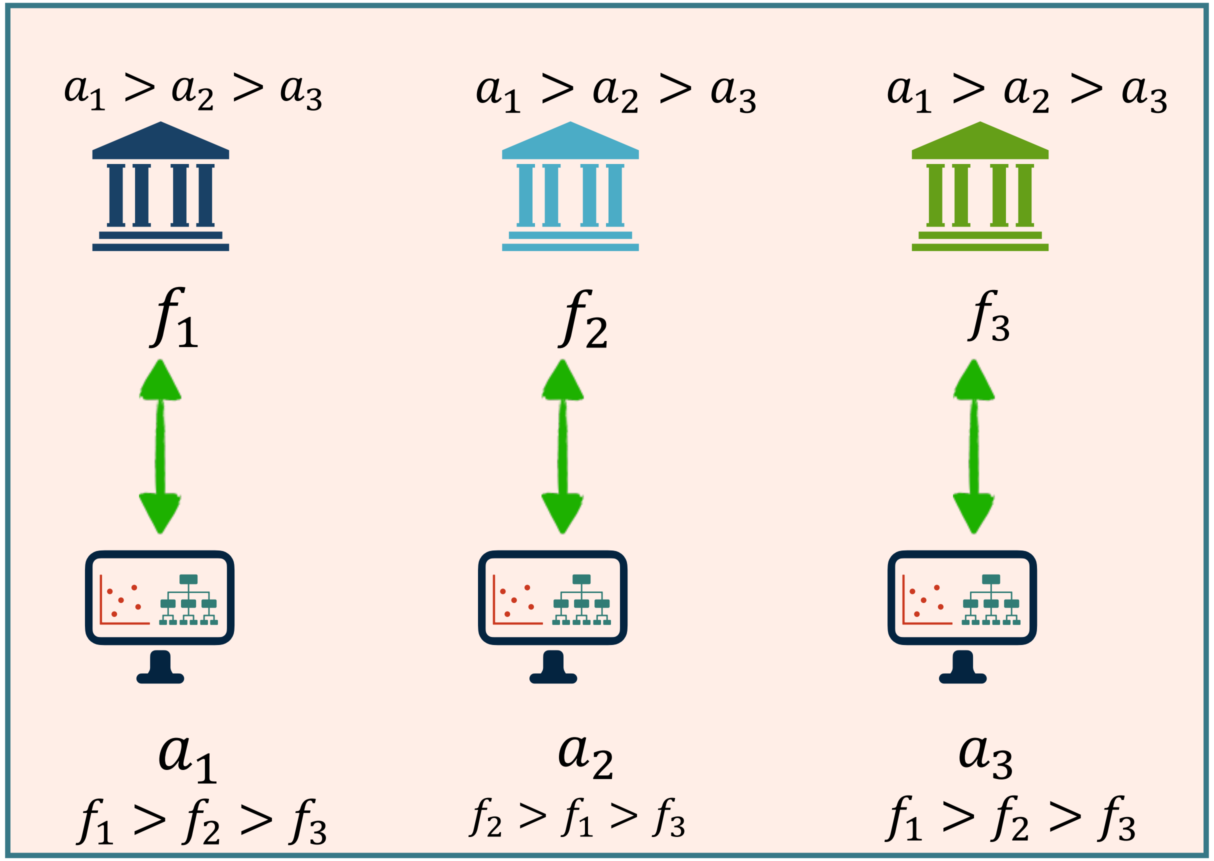
Stable Matching

A matching is called stable if there exists no blocking pairs

NOT Stable matching



Stable matching



[Gale and Shapley 1962]

Stable matching exists and can be non unique

Deferred Acceptance Algorithm: Known preference

[Gale and Shapley 1962]

1. Everyone starts unmatched
2. Each agent queries the most preferred firm that has not rejected it
3. Firm reviews list of queries and gets tentatively matched with best agent who queried and rejects other agents
4. Repeat from step 2

It is polynomial time algorithm and achieves a stable match

Deferred Acceptance Algorithm: Known preference

[Gale and Shapley 1962]

- Decentralized
  - Agents make their own decisions
- No Coordination
  - Agents do not need to coordinate actions across rounds
- No Communication
  - Based only on local past information and does not need to communicate with others
- Converges to a stable match
  - No (agent, firm) pair would abandon their current match for each other and be better off.

• Full Information Solution: Deferred Acceptance

- Decentralized

Develop an algorithm that learns agents preferences and quickly identifies stable match in a

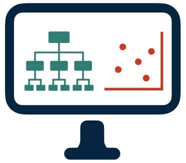
- decentralized,
- communication-free and
- coordination free manner

  - Agents do not need to see who they collide with or know the firms' preferences.

- Convergent to a stable match

  - No (agent, firm) pair would abandon their current match for each other and be better off.

Agent $a \in A$



Firm $f \in F$

Setup

- Set of agents A and a set of firms F form a market $\mathcal{M} = A \cup F$
- Firms have a fixed known preference on agents
- Agents have fixed but unknown preferences over firm
- Agents repeatedly query firms in order to learn preferences
- Agent a receives a noisy utility on successfully interacting with firm f

$$U_{a,f} = u_{a,f} + \epsilon_{a,f}$$

Unknown

- Firm queried by agent a at time t be $f_a(t)$
- Set of agents who query firm f at time t is given by $A_f(t) = \{a \in A : f_a(t) = f\}$

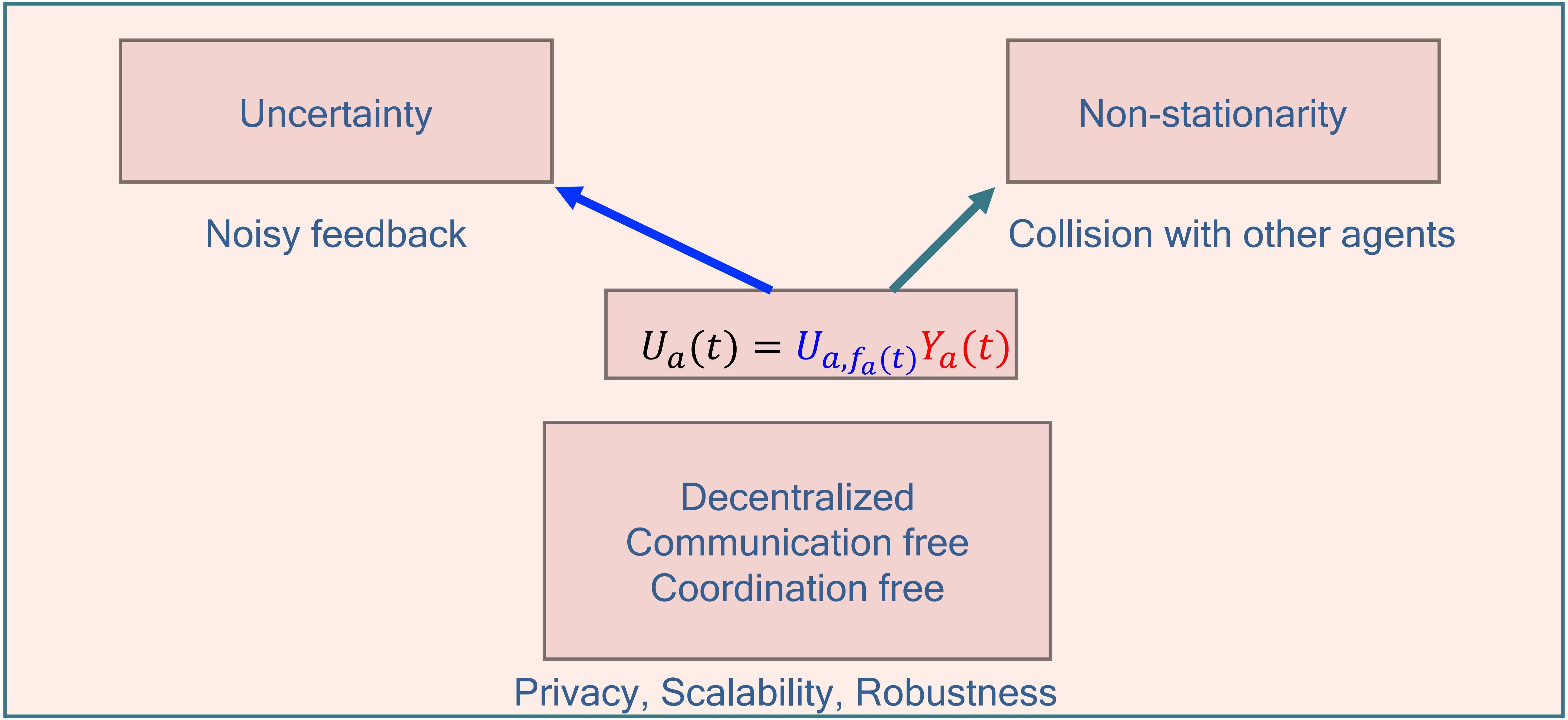
Setup continued...

- At time t if agent a queries firm f it gets a utility $U_a(t) = Y_a(t) U_{a,f_a(t)}$
  - $Y_a(t) = 1$ if agent a is most preferred amongst $A_f(t)$ by firm $f_a(t)$ [Matching]
  - $Y_a(t) = 0$ otherwise [Collision]
- Assume that there is a unique stable matching

- Let the stable matching firm for agent a be denoted by $f_a^\star$

Performance measure (Regret) $$R_a(T) = \mathbb{E}\left[\sum_{t=1}^{T} u_{a,f_a^\star} - U_a(t)\right]$$

Challenges


Uncertainty
Non-stationarity
Noisy feedback
Collision with other agents
$U_a(t) = U_{a,f_a(t)} Y_a(t)$
Decentralized
Communication free
Coordination free
Privacy, Scalability, Robustness

Challenges

Uncertainty

Stochastic Bandit Module

Non-stationarity

Adversarial Bandit Module

$$U_a(t) = U_{a,f_a(t)} Y_a(t)$$

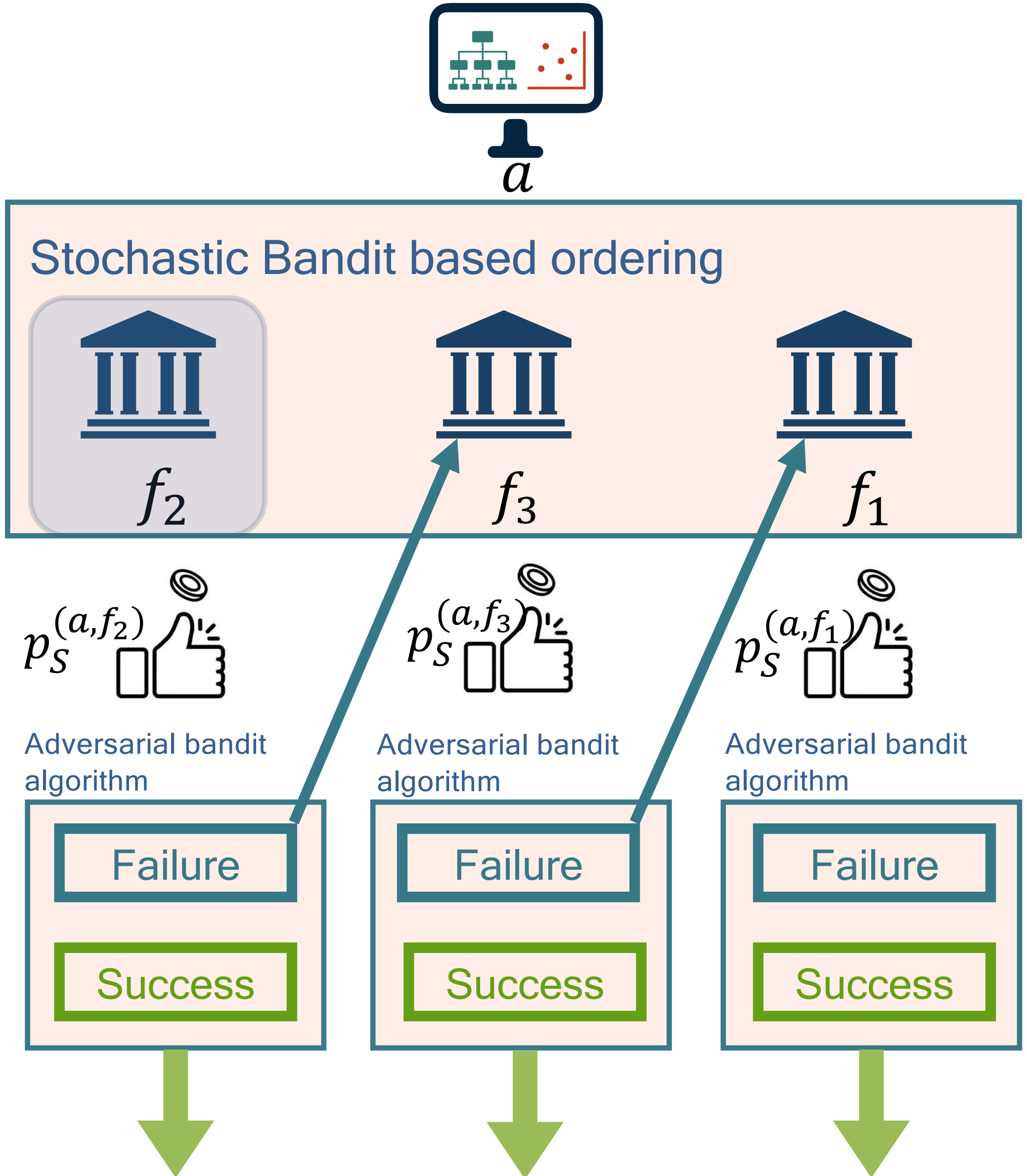
Decentralized
Communication free
Coordination free

Novel Algorithmic Design

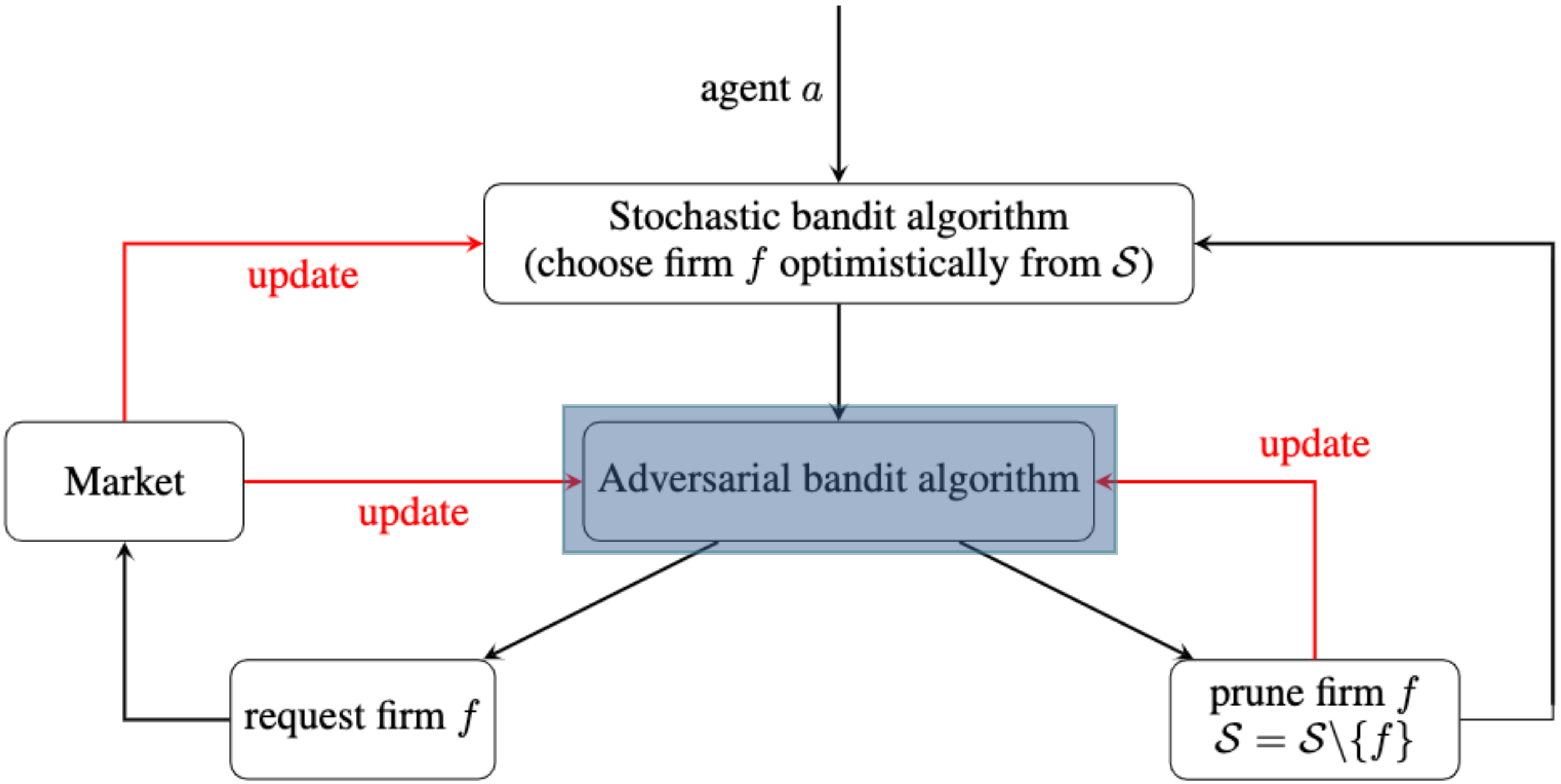
Algorithm

At every time t = 1,2,....

- Each agent maintains an ordering of firms based on past rewards and confidence
- Agent considers firms as per ordering one by one
- Flips a coin with a success probability $p_S^{(a,f)}$ associated with firm f
  - Failure: Move to next best firm (Prune)
  - Success: Query the current firm and obtain reward to update the ordering and the success probability
- If all the firms fail at time t then pick best firm from ordering



Algorithmic Paradigm



Exp3 based adversarial bandit module

Two actions i_t

Q Query the firm

P Prune the firm

$$p_S(t+1) = \frac{1}{1+\exp(\eta(\hat{L}_{query}(t) - \hat{L}_{prune}(t)))}$$

Loss estimators

$$\ell_Q^{adv} = \begin{cases} 0 & : successful\ match \\ 1 & : collision \end{cases}$$

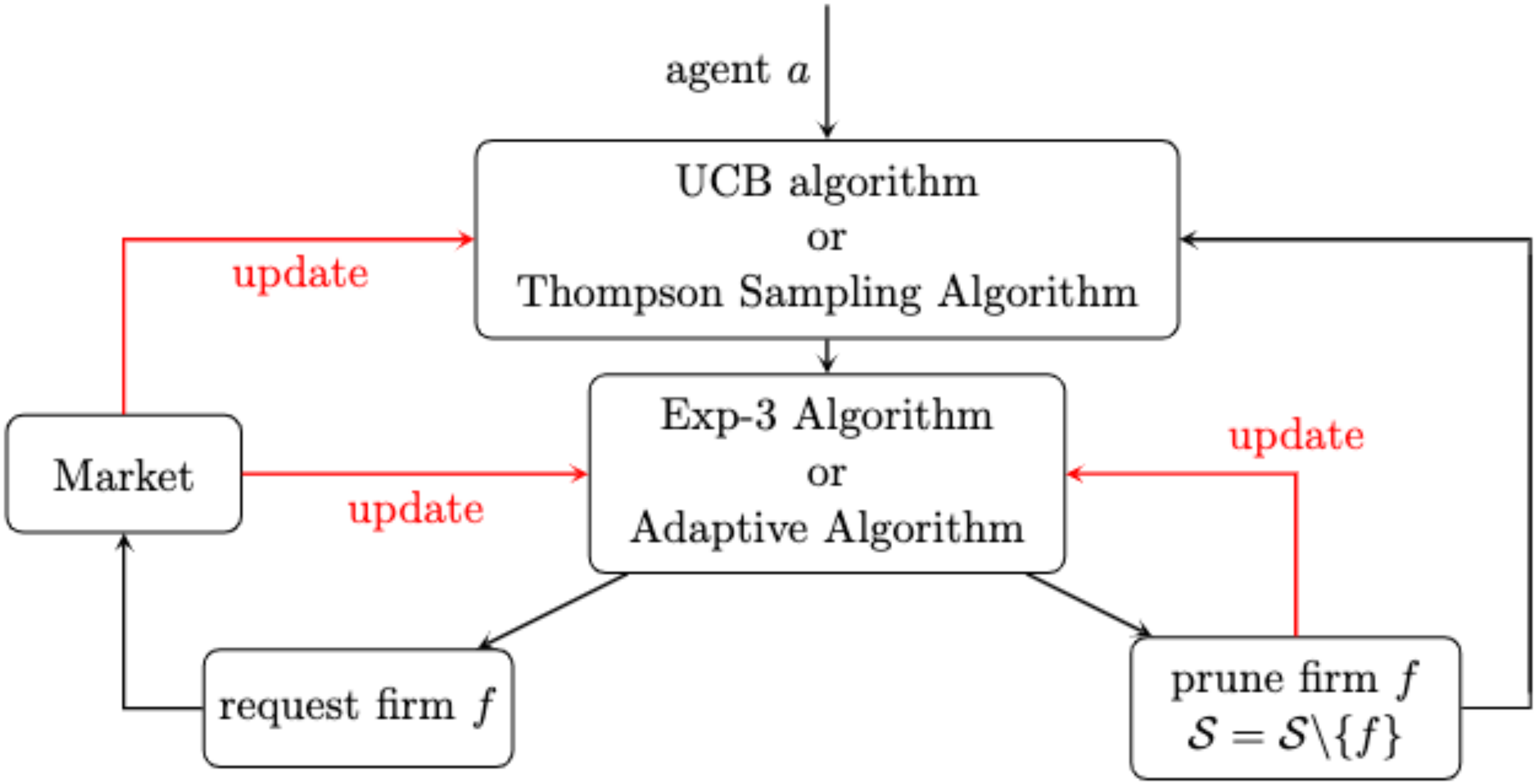
$$\ell_P^{adv} = 0.5$$

Loss structure

$$\hat{L}_{query}(t+1) = \hat{L}_{query}(t) + 1 - \frac{\mathbb{I}(i_{t+1} = Q)(1 - \ell_Q^{adv}(t+1))}{p_S(t+1)}$$

$$\hat{L}_{prune}(t+1) = \hat{L}_{prune}(t) + 1 - \frac{\mathbb{I}(i_{t+1} = P)(1 - \ell_P^{adv}(t+1))}{1 - p_S(t+1)}$$

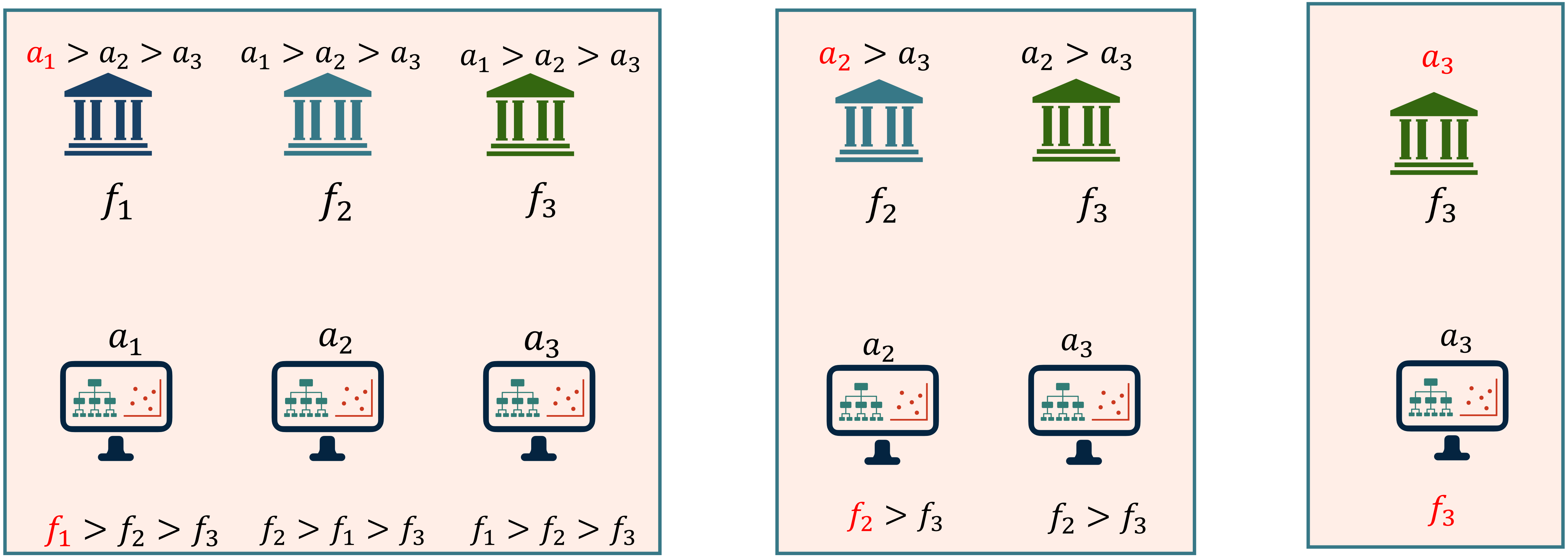
Modular Algorithmic Structure



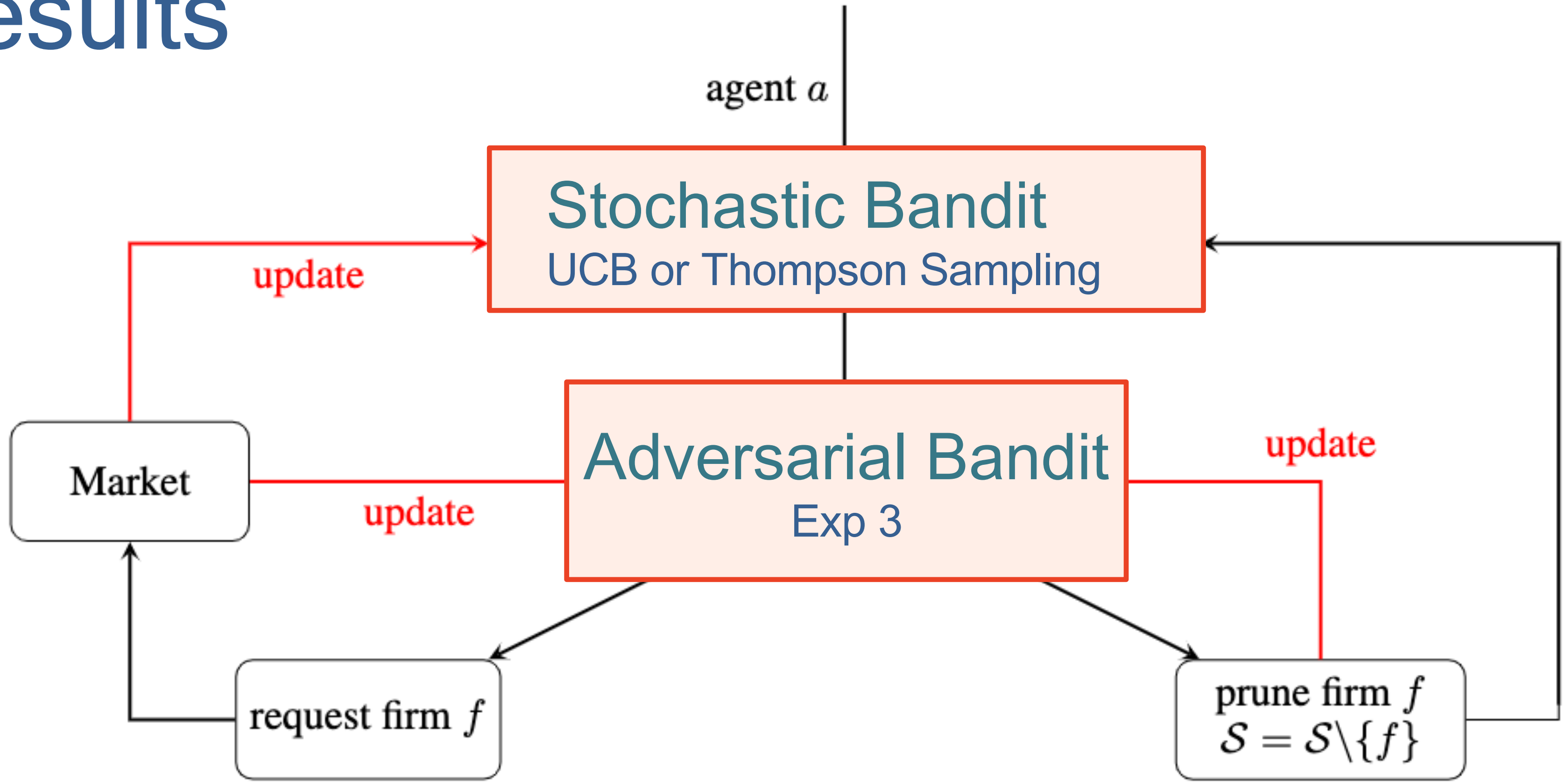
α-reducible markets

Definition: A tuple (a, f) is called fixed pair if f is most preferred by a and vice versa

Definition: A market is α −reducible if every submarket has a fixed pair.



Main Results



Theorem (informal) Under α −reducibility assumption the regret

$$R_a(T) \leq \mathcal{O}\left(\frac{1}{(\Delta^*)^2}\left(\sqrt{T}C\,(|F|, |A|)\right)\right)$$

where $\Delta * = \min_{a,f:\Delta_{a,f}>0} \Delta_{a,f} = u_a(f_a^\star) - u_a(f)$

Maheshwari, Mazumdar, Sastry. Decentralized, Coordination-, and Communication-Free Algorithms for Learning Structured Matching Markets (under submission 2022)

Main Results

agent a

Theorem (informal) Under α −reducibility assumption the regret

$$R_a(T) \leq \mathcal{O}\left(\frac{1}{(\Delta^*)^2}\left(\sqrt{T}C\ (|F|, |A|)\right)\right)$$

where $\Delta * = \min_{a,f:\Delta_{a,f}>0} \Delta_{a,f} = u_a(f_a^\star) - u_a(f)$

Curse of multi-agents

request firm f

prune firm f
$\mathcal{S} = \mathcal{S}\backslash\{f\}$

Can we at least improve dependence on T?

Theorem (informal) Under α −reducibility assumption the regret

$R_a(T)$

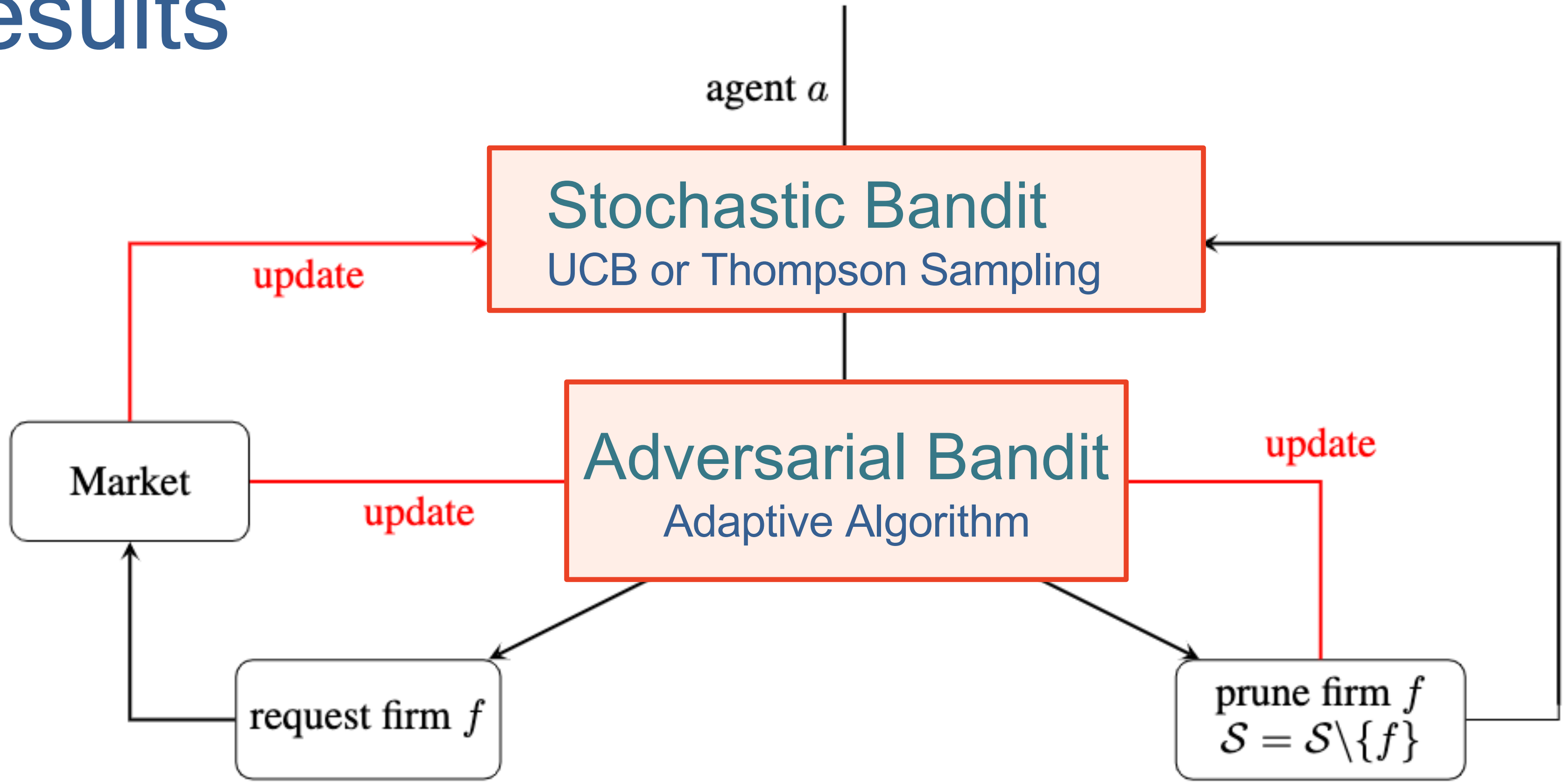
where $\Delta * = \min_{a,f:\Delta_{a,f}>0} \Delta_{a,f} = u_a($

Key idea

Collisions are not adversarial

Maheshwari, Mazumdar, Sastry. Decentralized, Coordination-, and Communication-Free Algorithms for Learning Structured Matching Markets (under submission 2022)

Main Results



Theorem (informal) Under α −reducibility assumption the regret

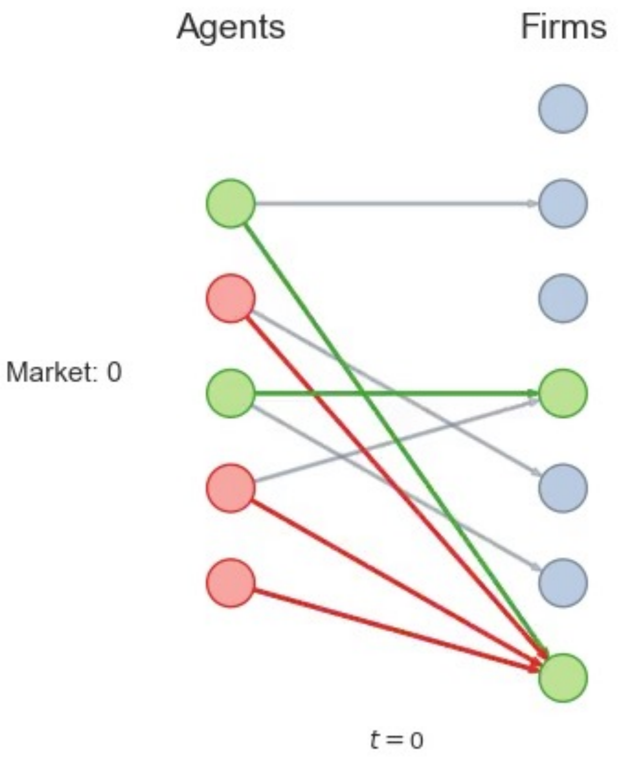
$$R_a(T) \leq \mathcal{O}\left(\frac{1}{(\Delta^*)^2}\left(\sqrt{\log T}\, C\,(|F|, |A|)\right)\right)$$

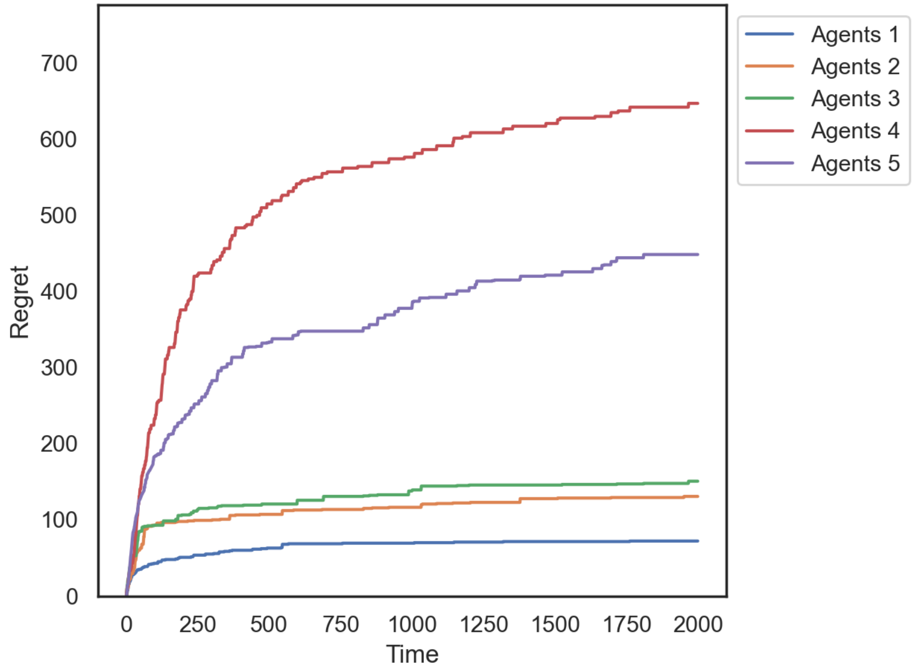
where $\Delta * = \min_{a,f:\Delta_{a,f}>0} \Delta_{a,f} = u_a(f_a^\star) - u_a(f)$

Maheshwari, Mazumdar, Sastry. Decentralized, Coordination-, and Communication-Free Algorithms for Learning Structured Matching Markets (To qppear in neurips2022)

• Numerical Experiments

- 5 agents with randomly generated preferences. 7 firms with fixed preferences.
- Each agent is randomly chosen to use either Thompson sampling or UCB.
- All agents use mirror-descent with the log-barrier regularizer.

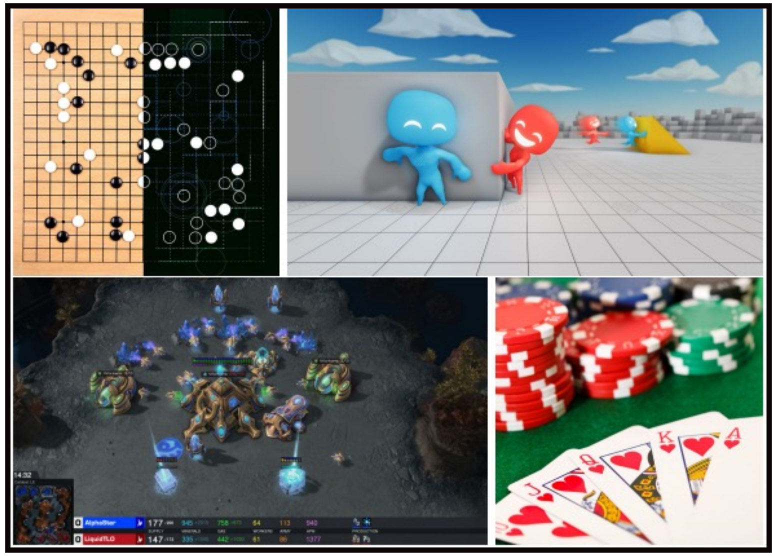


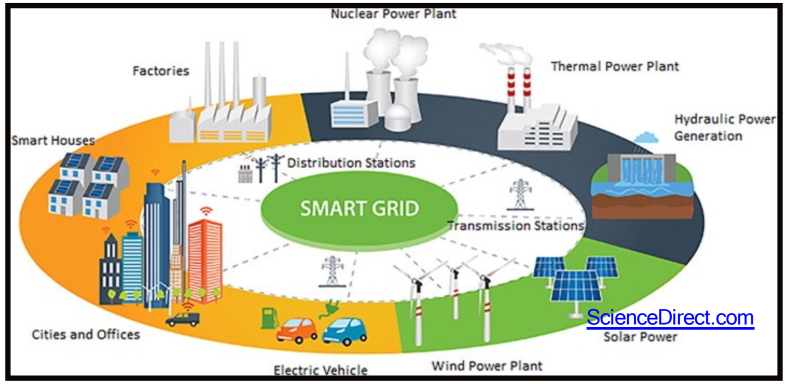


Maheshwari, Mazumdar, Sastry. Decentralized, Coordination-, and Communication-Free Algorithms for Learning Structured Matching Markets (under submission 2022)

Vignette 2.2:
Independent and decentralized learning in Markov Games

Under review. Available at https://arxiv.org/abs/2206.02344





Key Characteristics

- Non-myopic strategic agents
- Uncertain and dynamic environment
- Limited communication or coordination between agents
- Limited knowledge about other agents

Question

How can agents make decisions in such environment by effectively exploring and exploiting in presence of other agents?

Setup

[Markov Game] The game $G = \langle I, S, (A_i)_{i\in I}, (u_i)_{i\in I}, P, \delta\rangle$ where

- I : finite set of players
- S : finite set of states
- A_i is the set of available actions to player i
- $u_i: S\times A \to \mathbb{R}$ is the one-stage payoff of player i encodes preferences
- $P(s'|s,a)$ denote the transition probability to s' from state s under action a
- $\delta \in (0,1)$ is the discount factor

Setup

[Policy class] We restrict the players' policy to be stationary Markovian.

- $\pi_i(s, a_i)$ be a stationary Markov policy for player i which states the probability that player i chooses action a_i in state s
- Joint policy profile of players is $\pi = (\pi_i)_{i \in I}$

[Players' objective] Given the initial state distribution $\mu \in \Delta(S)$ the long-run expected payoff of any player $i \in I$ is given as

$$V_i(s, \pi) = \mathbb{E}\left[\sum_{k=0}^{\infty} \delta^k u_i(s^k, a^k)\right]$$

where $s^0 = s, a^k \sim \pi(s^k), and s^k \sim P(\cdot\,|s^{k-1}, a^{k-1})$

Solution Concepts

Nash equilibrium A policy π^* is stationary Nash equilibrium if for any player i, π_i and initial state distribution μ

$$V_i(\mu, \pi_i^*, \pi_{-i}^*) \geq V_i(\mu, \pi_i, \pi_{-i}^*)$$

Theorem

A stationary Nash equilibrium always exists for a Markov game with finite state and finite actions

ϵ –Nash equilibrium A policy π^* is stationary Nash equilibrium if for any player i, π_i and initial state distribution μ

$$V_i(\mu, \pi_i^*, \pi_{-i}^*) \geq V_i(\mu, \pi_i, \pi_{-i}^*) - \epsilon$$

Solution Concepts

Nash equilibrium A policy π^* is stationary Nash equilibrium if for any player i, π_i and initial state distribution μ

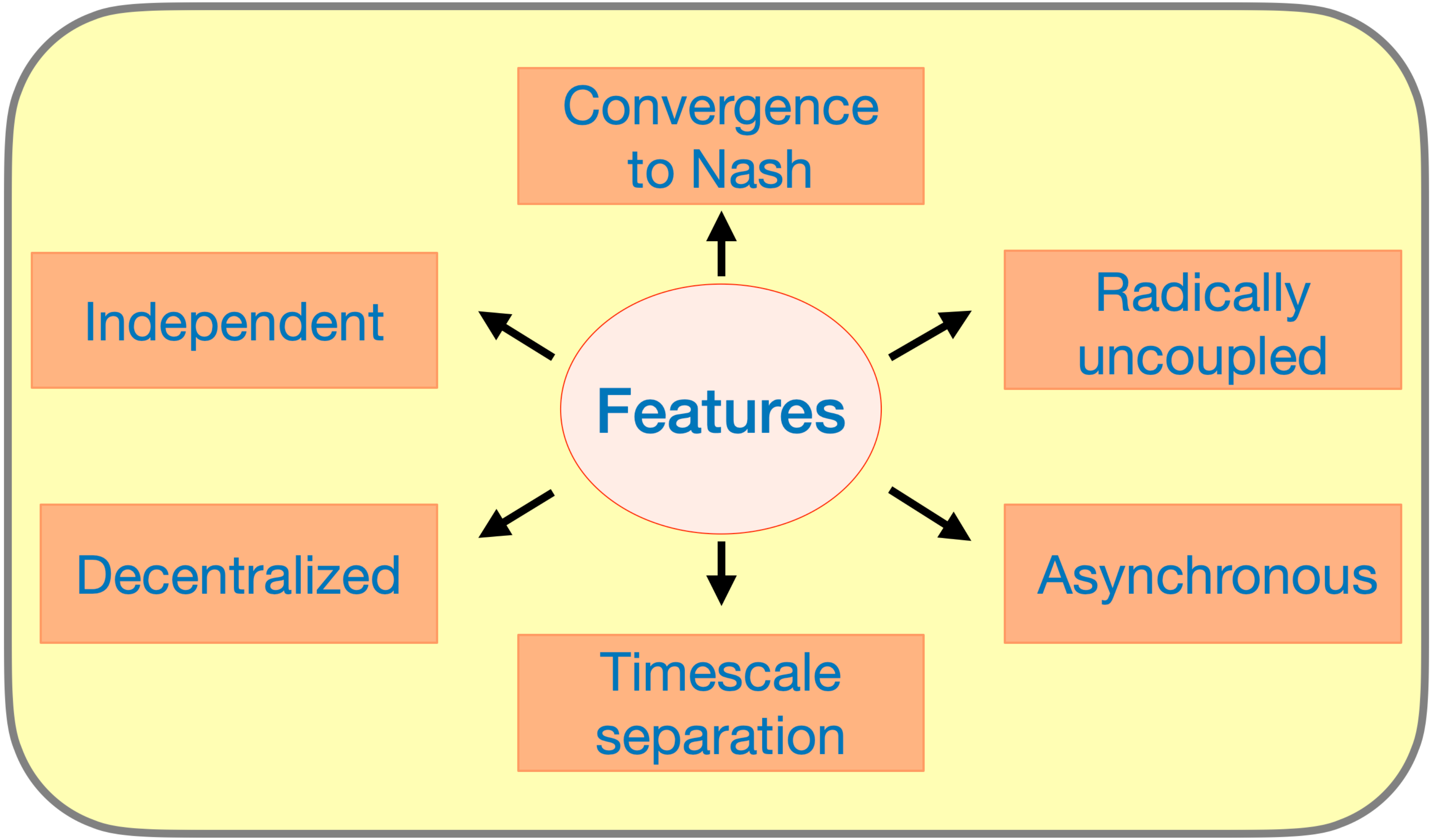
$$V_i(\mu, \pi_i^*, \pi_{-i}^*) \geq V_i(\mu, \pi_i, \pi_{-i}^*)$$

Develop learning dynamics which helps players learn about underlying environment and has following properties

- Decentralized and independent implementation
- Requires no information about the underlying structure of the game
- Converges to Nash equilibrium

ϵ —Nash equilibrium A policy π^* is stationary Nash equilibrium if for any player i, π_i and initial state distribution μ

$$V_i(\mu, \pi_i^*, \pi_{-i}^*) \geq V_i(\mu, \pi_i, \pi_{-i}^*) - \epsilon$$


Convergence to Nash
Independent
Radically uncoupled
Features
Decentralized
Asynchronous
Timescale separation

Prior Work

- (Borkar 02) proposed an actor-critic based algorithm with similar timescale separation and showed weighted empirical distribution of actions of players converge to generalized Nash equilibrium
- (Arslan and Yüksel 16) proposed decentralized algorithm in the context of acyclic Markov games which required coordination between players
- (Perolat et al 18) proposed decentralized actor-critic algorithm for finite length cooperative multistage games
- (Daskalakis et al 20) proposed decentralized learning dynamics for zero-sum games but requires one player to update slower than another
- (Sayin et al 21) proposed a decentralized and independent learning dynamics in the context of zero-sum games but with reversed timescale separation

Nash Equilibrium characterization

Q-function

$$Q(s, a_i; \pi) = \mathbb{E}_{a_{-i} \sim \pi_{-i}(s)} \left[u_i(s, a_i, a_{-i}) + \delta \mathbb{E}_{s' \sim P(\cdot | s, a)} [V_i(s', \pi)] \right]$$

Optimal one-stage deviation

$$\mathbf{br}_i(s; \pi) = \arg\max_{\hat{\pi}_i(s)} \sum_{a_i} \hat{\pi}_i(s, a_i) Q_i(s, a_i; \pi)$$

Let $\text{br}_i(\pi) = (\text{br}_i(s, \pi))_{s \in S}$

Proposition

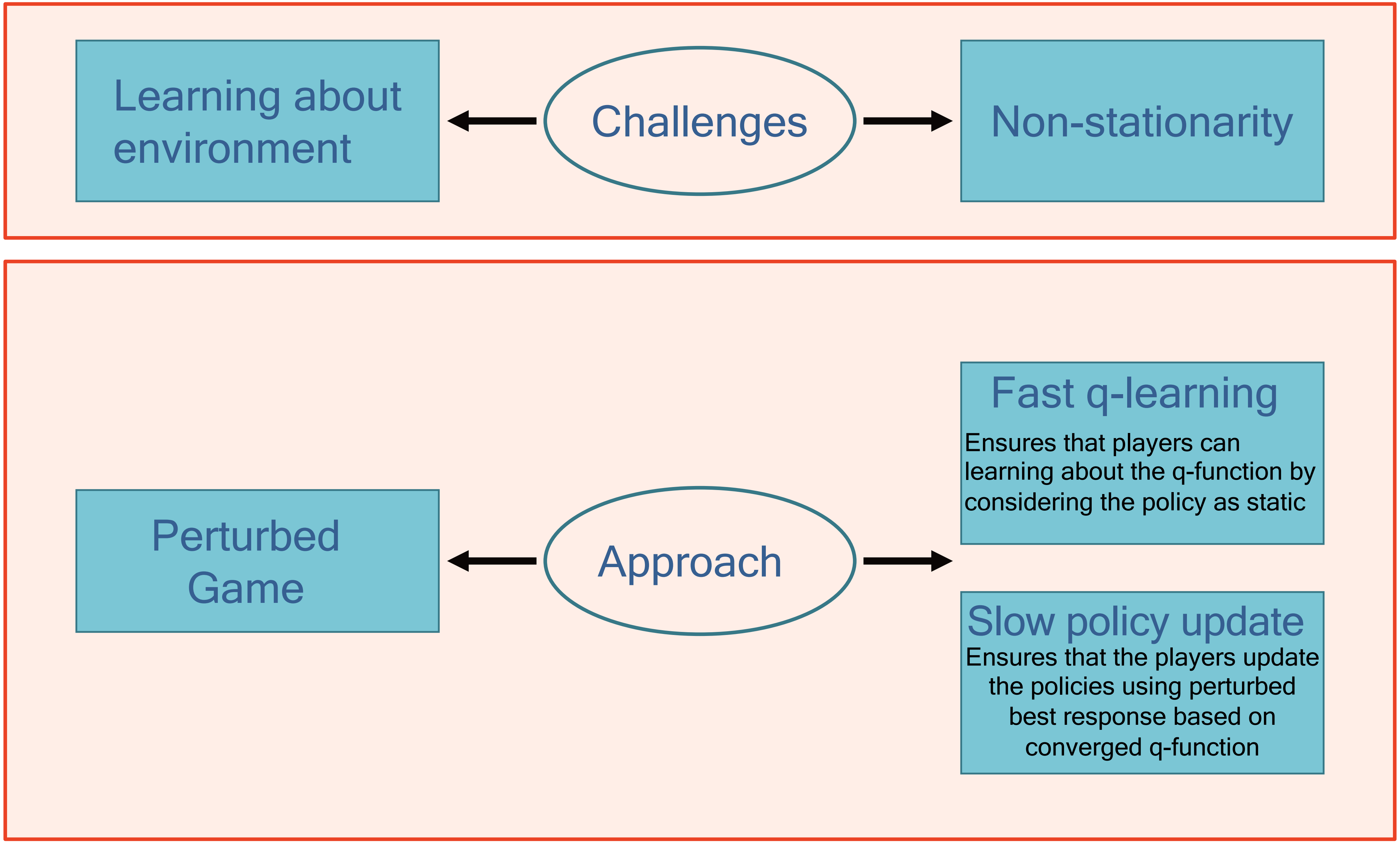
Any fixed point of the mapping $\text{br}_i(\cdot)$ is a Nash equilibrium of game G

Markov Potential Game

A game is Markov Potential Game if there exists a (potential) function $\Phi : S \times \Pi \to \mathbb{R}$ such that for every $s \in S, \pi_i, \pi_i', \pi_{-i}$

$$\Phi(s, \pi_i', \pi_{-i}) - \Phi(s, \pi_i, \pi_{-i}) = V_i(s, \pi_i', \pi_{-i}) - V_i(s, \pi_i, \pi_{-i})$$

- Example: Markov team games where $u_i = u$ for all $i \in I$
- The maximizer of potential function is a Nash equilibrium of the game
- The potential function is typically a non-concave function


Learning about environment
Challenges
Non-stationarity
Perturbed Game
Approach
Fast q-learning
Ensures that players can learning about the q-function by considering the policy as static
Slow policy update
Ensures that the players update the policies using perturbed best response based on converged q-function

Perturbed Game $\tilde{G}(\tau)$

Define $\nu_i(s, \pi_i) = \sum_{a_i} \pi_i(s, a_i) \log(\pi_i(s, a_i))$

Perturbed one stage payoff $\tilde{u}_i(s, \pi) = \mathbb{E}_{a \sim \pi(s)}[u_i(s, a)] - \tau \nu_i(s, \pi_i)$

Perturbed long-run payoff $\tilde{V}_i(s, \pi) = \mathbb{E}\left[\sum_{k=0}^{\infty} \delta^k \left(u_i(s^k, a^k) - \tau \nu_i(s^k, \pi_i)\right)\right]$

Perturbed Q-function $\tilde{Q}_i(s, a_i; \pi) = \mathbb{E}_{a_{-i} \sim \pi_{-i}(s)}\left[u_i(s, a) - \tau \nu_i(s, \pi_i) + \delta \mathbb{E}_{s' \sim P(\cdot|s,a)}[\tilde{V}_i(s', \pi)]\right]$

Proposition If G is a Markov potential game with potential function Φ then $\tilde{G}$ is also a Markov potential game with potential function $\tilde{\Phi}$

$$\tilde{\Phi}(s, \pi) = \Phi(s, \pi) - \tau \mathbb{E}\left[\sum_{i \in I} \sum_{k=0}^{\infty} \delta^k \nu_i(s^k, \pi_i)\right]$$

Perturbed game

- Perturbed optimal one-stage deviation

$$\tilde{\mathbf{br}}_i(s,\pi) = \arg\max_{\hat{\pi}_i(s)} \sum_{a_i} \hat{\pi}_i(s,a_i)\tilde{Q}_i(s,a_i,\pi) - \tau\nu_i(s,\hat{\pi}_i)$$

$$\Downarrow$$

$$\tilde{\mathbf{br}}_i(s,\pi) = \left(\frac{\exp(\tilde{Q}_i(s,a_i;\pi)/\tau)}{\sum_{a_i}\exp(\tilde{Q}_i(s,a_i;\pi)/\tau)}\right)$$

- Perturbed best response chooses every action with positive probability
- As $\tau \to \infty$ every action is chosen with equal probability in every state
- As $\tau \to 0$ the action with highest Q-value is chosen in every state

Learning Dynamics

Multi-agent Bellman operator

$$\mathcal{T}_i^{\pi}\tilde{q}_i(s,a_i) = u_i(s,a_i,\pi_{-i}) - \tau\nu_i(s,\pi_i) + \delta\mathbb{E}_{s'\sim P(\cdot|s,a_i,\pi_{-i})}\left[\sum_{a_i'}\pi_i(s',a_i')\tilde{q}_i(s',a_i')\right]$$

Multi-agent (sampled) Bellman operator

$$\hat{\mathcal{T}}_i^{\pi}\tilde{q}_i(s,a_i) = u_i(s,a_i,a_{-i}) - \tau\nu_i(s,\pi_i) + \delta\sum_{a_i'}\pi_i(s',a_i')\tilde{q}_i(s',a_i') \text{ where } a_{-i}\sim\pi_{-i}, s'\sim P(\cdot\,|\,s,a_i,a_{-i})$$

Fast timescale (q-updates)

$$\tilde{q}_i^k(s^{k-1},a_i^{k-1}) = \tilde{q}_i^{k-1}(s^{k-1},a_i^{k-1}) + \underbrace{\alpha(\textbf{Counter})\mathbb{I}((s,a_i)=(s^{k-1},a_i^{k-1}))}_{\text{Asynchronous update}}\left(\hat{\mathcal{T}}_i^{\pi^{k-1}}\tilde{q}_i^{k-1}(s,a_i) - \tilde{q}_i^{k-1}(s,a_i)\right)$$

Slow timescale (policy updates)

$$\pi_i^k(s^{k-1},a_i) = \pi_i^{k-1}(s,a_i) + \underbrace{\beta(\textbf{Counter})\mathbb{I}(s=s^{k-1})}_{\text{Asynchronous update}}\left(\frac{\exp(\tilde{q}_i^{k-1}(s,a_i)/\tau)}{\sum_{a_i}\exp(\tilde{q}_i^{k-1}(s,a_i)/\tau)} - \pi_i^{k-1}(s,a_i)\right)$$

Assumptions

- (A1) [Initial state distribution] Initial state distribution μ has full support

- (A2) [Transition kernel] There exists a joint action profile a such that the markov chain induced by $(P(s'|s,a))_{s,s'}$ is irreducible and aperiodic

- (A3) [Learning rates] The step size sequence $(\alpha(n), \beta(n))$ satisfy the following
  - [Infinite travel and decaying] $\sum_n \alpha(n) = +\infty, \sum_n \beta(n) = +\infty, \lim_{n\to\infty} \alpha(n) = \lim_{n\to\infty} \beta(n) = 0$
  - [Time scale separation] $\lim_{n\to\infty} \beta(n)/\alpha(n) = 0$
  - [Taming the asynchronicity] For any $x \in (0,1)$, $\sup_n \alpha([xn])/\alpha(n) + \beta([xn])/\beta(n) < \infty$

Main Result

Define $\tau^{\dagger} = \dfrac{e\epsilon(1-\delta)}{2max_i|A_i|}$

Theorem: Under (A1)-(A3), given any $\epsilon > 0$ and any $\tau \in (0, \tau^{\dagger})$ the sequence of policy profiles $(\pi^k)_{k=0}^{\infty}$ converges to ϵ −Nash equilibrium with probability 1.

Application Area: Mobility Systems

Deployment of autonomous vehicles into mobility infrastructure by effectively incorporating

- Continuous state and action spaces
- Partial information about state of the system
- Communication with neighbors
- Bounded rationality of human decision making

Inventing the Future

- ***Deep Technology Design Innovation***
  - **Robotics and Intelligent Machines --** *Advancing the state of art in robots working with humans, unmanned vehicles, air and ground, deep learning, new transportation methodologies,*
  - **Augmented Reality/Virtual Reality** – *Augmenting Cognition, redefining the future of brain machine interfaces, the future of performance, educational delivery.*
  - **IoT and Next Generation Infrastructure** – *Swarms of Sensors in Cyber Physical Systems creating the sharing economy.*
- ***Driving Societal Change***
  - **Better Health** – *Saving and extending lives with new tools for better diagnosis and care 24/7 in the hospital, clinic, and home*
  - **Improving the Human Experience** – *Enhancing the quality of life in work, family, and society with human-centered technology solutions*
- ***Societal Concerns***
  - **Collective Good and Individual Utility :** – *Individual utlility and societal good. How do you incentivize players to do the right thing.*
  - **Should the Future be like the Past** – *All supervised learning will incorporate the biases of "past" training data into predictions of the future*
  - **Humans Adapt to Automation --** *Should machine learning algorithms be robust to human adaptation*

Thank you!