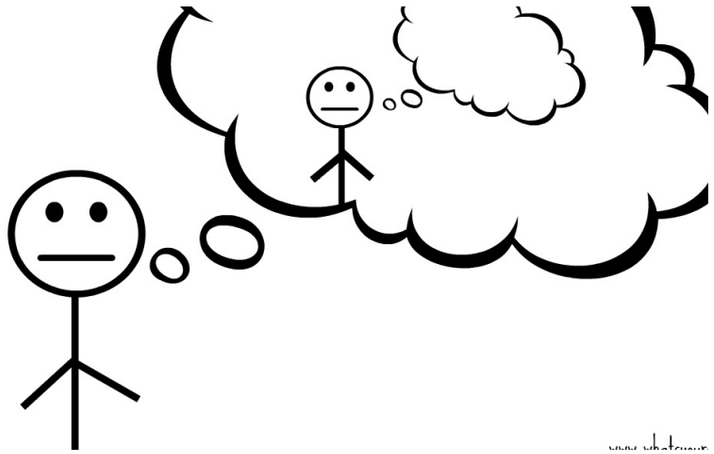
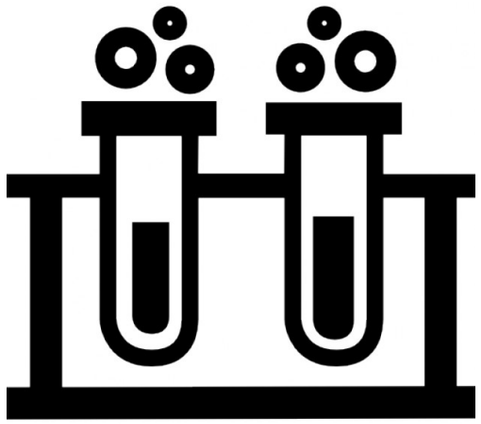
Predicting Slow Judgment

Owain Evans
FHI, University of Oxford

Collaborators: Andreas Stuhlmüller (ought.org)
Ryan Carey, Neal Jean, Andrew Schreiber, Girish Sastry, Chris Cundy



Introduction

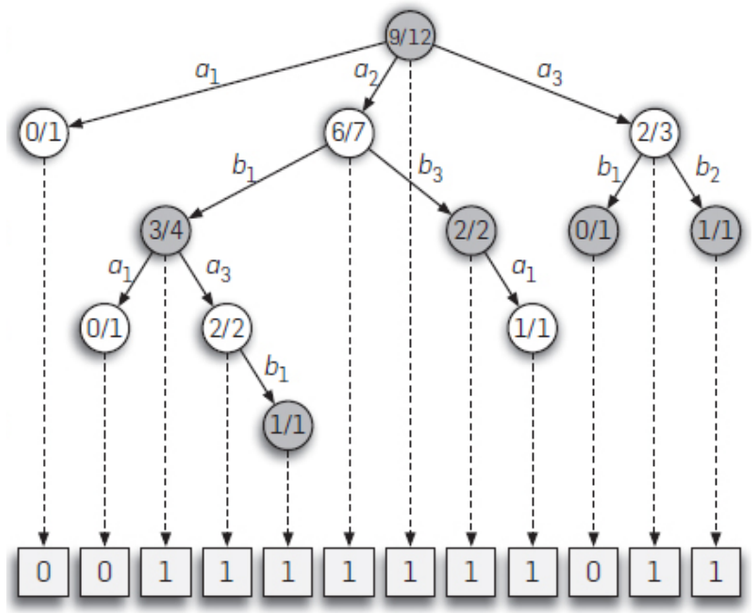


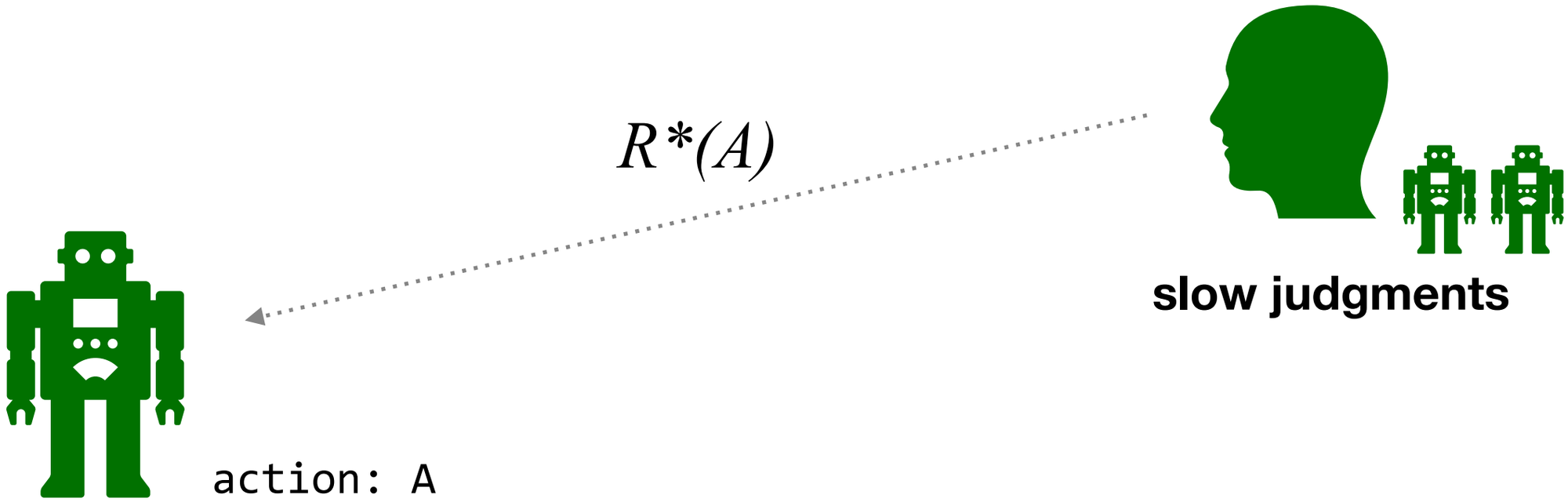
“slow judgment”

judgment that takes a long time due to serial reasoning, meta-reasoning, running experiments, discussing with experts, etc.

Related:

- Kahneman “Thinking, fast and slow”
- Christiano “Turning reflection up to 11”
- AlphaGo: imitation of human experts and imitation of MCTS by neural net



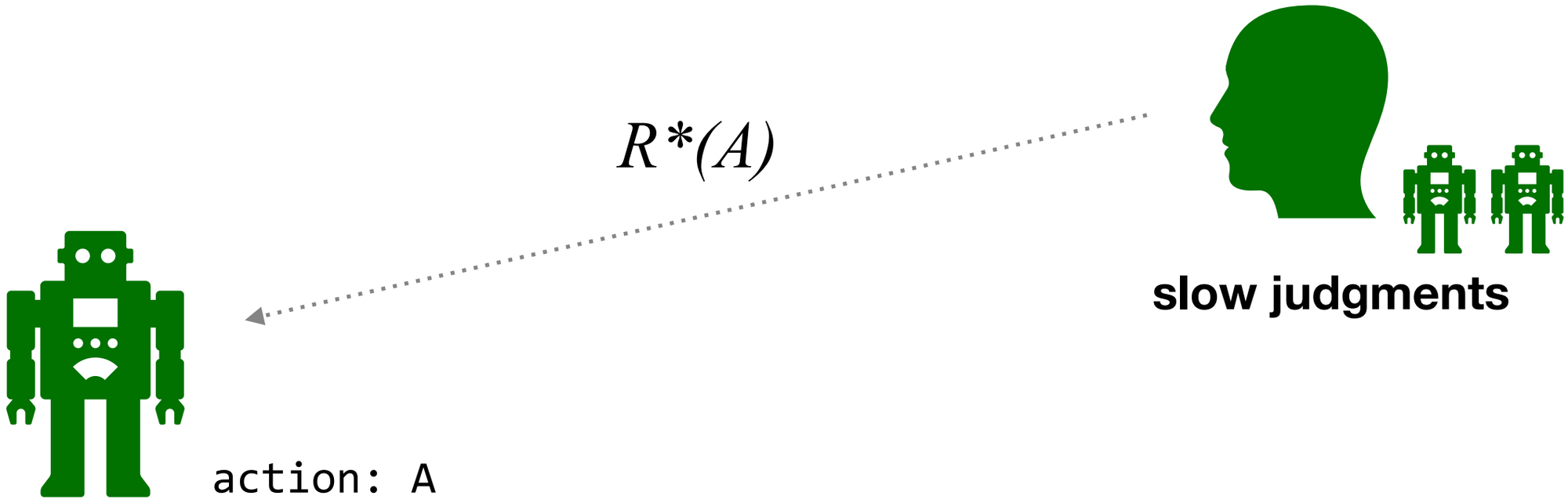


Optimizing for slow judgments

- Agent takes action with highest approval (reward) after slow judgment (e.g. thinking for 5 days).

- Slow judgment considers short-term progress on task (e.g. personal assistant), safety constraints, long-term considerations.

- (Standard RL vs. Christiano, “Approval Directed Agents”)

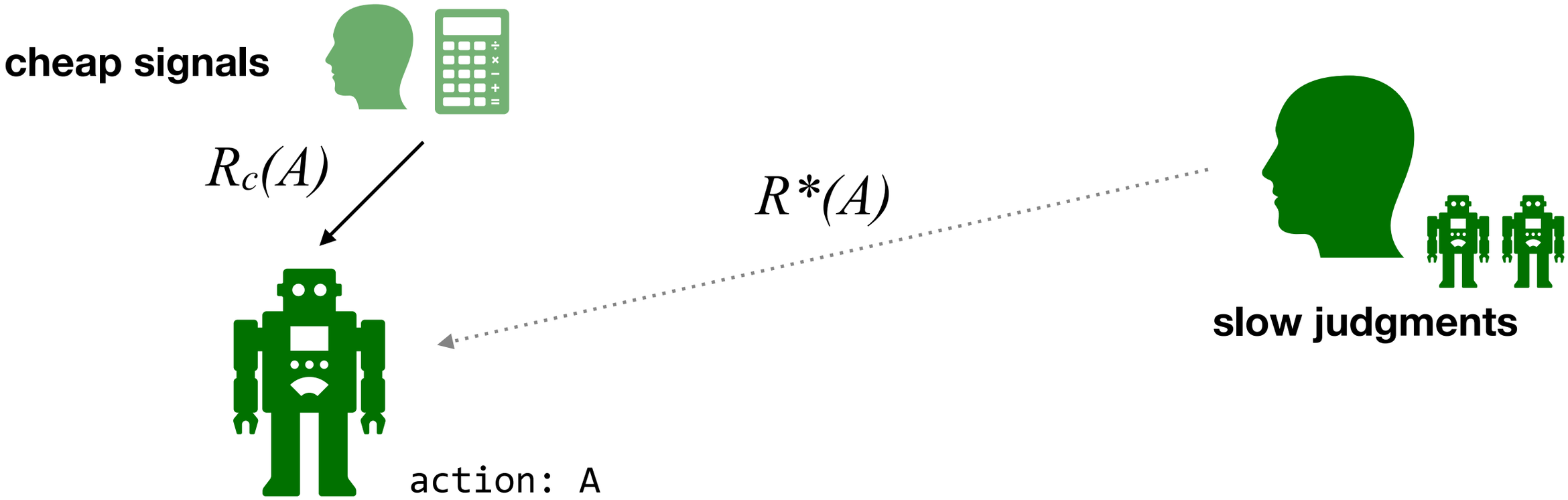


Optimizing for slow judgments

- Agent takes action with highest approval (reward) after slow judgment.

Problems

- **AI-complete:** slow judgments involve math reasoning, scientific experiments, moral deliberation.
- **missing data:** slow judgments are intrinsically expensive



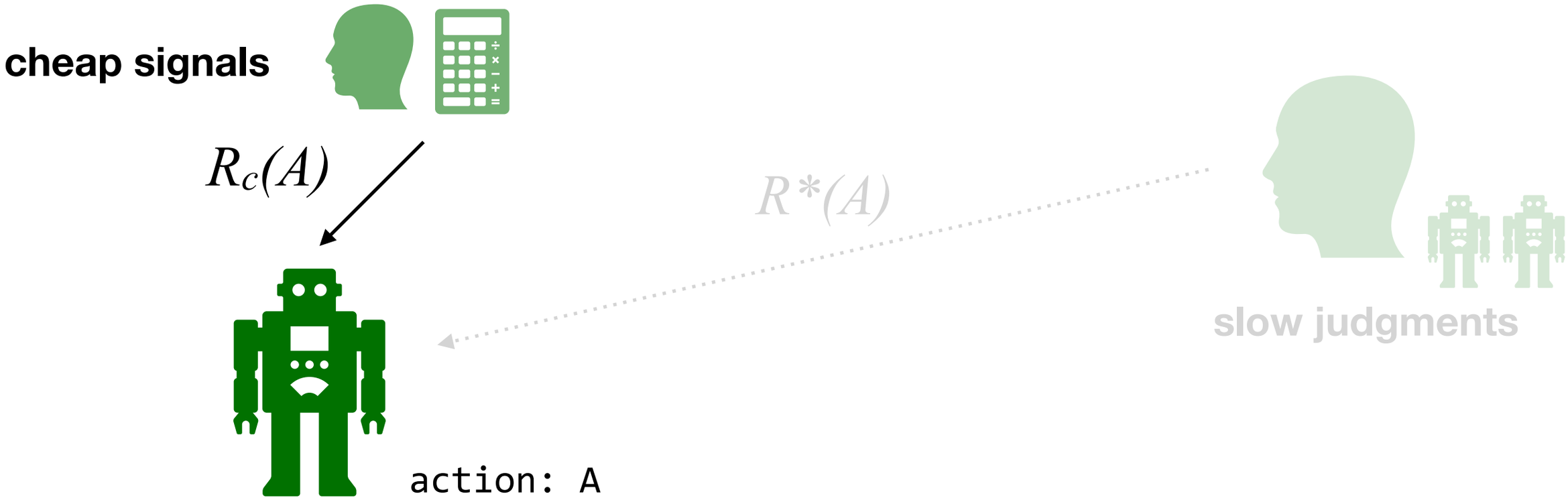
Optimizing for slow judgments

- Agent takes action with highest reward after slow judgment.

Approach

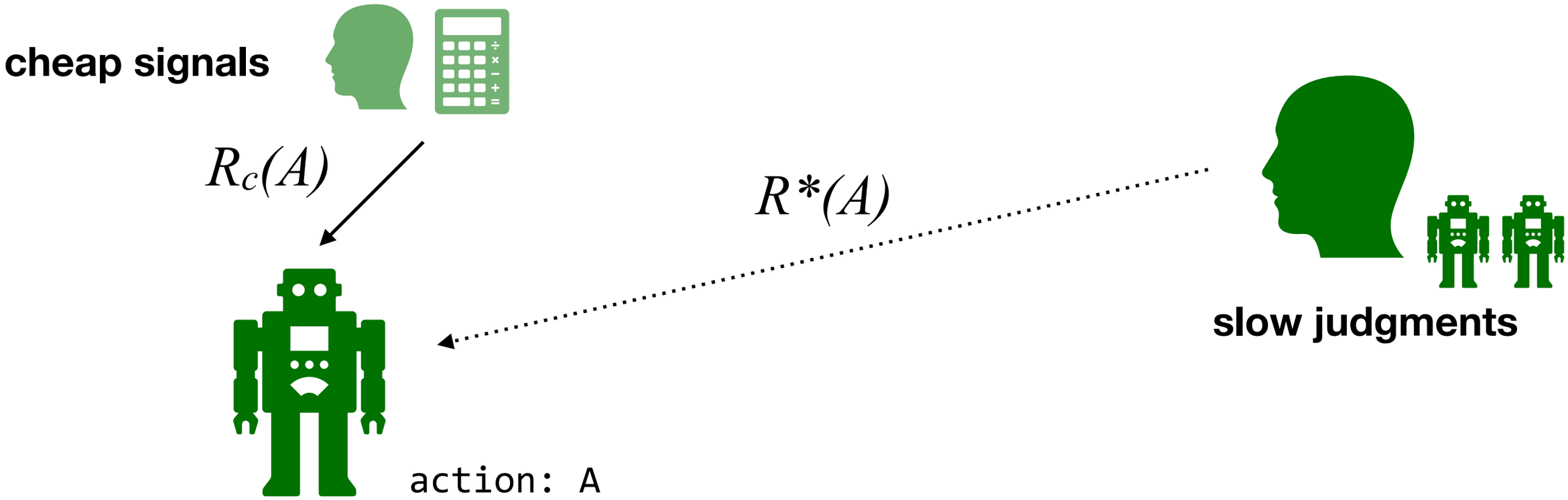
- **Cheap signals** = human quick judgment, ad revenue, stock price, hand-engineered shaped rewards.

- Use cheap signals (not sparse) to help predict slow judgment



Alternative: Optimize for cheap signals

- No need to collect expensive data
- Easier to apply standard ML
- Cheap signals are correlated with slow judgments. Optimizing for cheap might initially do well for slow judgments (but will ultimately diverge— **mis-alignment).**



Research Goal

- Optimizing for slow judgments should be **competitive** with optimizing for cheap signals (to achieve practical tasks, e.g. Siri).

- **Competitive** = similar speed when deployed, training set has similar size, training time is not 20x slower, etc.

Plan for rest of talk

1. Describe concrete tasks where objective is predicting slow judgment. For tractability, these are classification problems rather than RL.

2. Explain how we constructed our datasets

3. Show pilot results for predicting slow judgments

Problem Statement

Frame “Predicting Slow Judgment” as classification problem:

x_i ⟶ $h^*(x_i)$

input (feature vector) target

$h^*(x_i)$: Alice’s judgment about x_i after long deliberation and research (e.g. 5 days)

Guiding example:

x_i = “When I was governor of Massachusetts, we didn’t just slow the rate of growth of our government, we actually cut it.” Mitt Romney (CBN Interview, 2012).

Problem Statement

Frame “Predicting Slow Judgment” as classification problem:

$x_i \longrightarrow h^*(x_i)$

input (feature vector) target

$h^*(x_i)$: Alice’s judgment about x_i after long deliberation and research (e.g. 5 days)

Problem

ML would do poorly due to few h^* labels and AI-completeness:

- Cheap signals at train time (mitigate few h^* labels)
- Cheap signals at test time (mitigate AI-completeness).

Problem Statement

Predicting slow judgment:

$$x_i\,,\quad \boldsymbol{h(x_i\,;\,20s)} \longrightarrow h^*(x_i)$$

input (feature vector) target

$h^*(x_i)$: Alice's judgment about x_i after long deliberation and research

$\boldsymbol{h(x_i\,;\,t)}$: Alice's guess about x_i after time t.

ML would do poorly due to few h^* labels and AI-completeness:

- Cheap signals at train time (mitigate few h^* labels)
- Cheap signals at test time (mitigate AI-completeness).

Problem Statement

Predicting slow judgment:

$$x_i\,,\quad \boldsymbol{h(x_i\,;\,20s)} \longrightarrow h^*(x_i)$$

input (feature vector) → target

$h^*(x_i)$: Alice's judgment about x_i after long deliberation and research

$\boldsymbol{h(x_i\,;\,t)}$: Alice's guess about x_i after time t.

- Shouldn't target h^* be whether statement is true?
- We are ultimately interested in slow judgments that depend on preferences, which aren't objectively true/false. Also: for this task, our actual labels are just expert judgments —some are wrong.

Problem Statement

x_i = "When I was governor of Massachusetts, we didn't just slow the rate of growth of our government, we actually cut it." Mitt Romney (CBN Interview, 2012).

x_i, $h(x_i; 20\text{s})$, $h(x_i; 60\text{s})$, $h(x_i; 320\text{s})$, $h(x_i; 540\text{s})$ $\longrightarrow$ $h^*(x_i)$

	$h(x_i; 20\text{s})$	$h(x_i; 60\text{s})$	$h(x_i; 320\text{s})$	$h(x_i; 540\text{s})$	$h^*(x_i)$
x_i	50%	40%	20%	25%	0%

Problem Statement

Add more questions for Alice:

		$h(x_i; 20s)$	$h(x_i; 60s)$	$h(x_i; 320s)$	$h(x_i; 540s)$	$h^*(x_i)$
"Romney cut MA government"	x_1	50%	40%	20%	25%	0%
"Rubio skipped 18 defense votes"	x_2	50%	55%	55%	90%	100%
"The Keynesian response is right"	x_3	50%	50%	50%	50%	20%
"Obama will enforce Sharia Law"	x_4	0%	0%	0%	0%	0%

Problem Statement

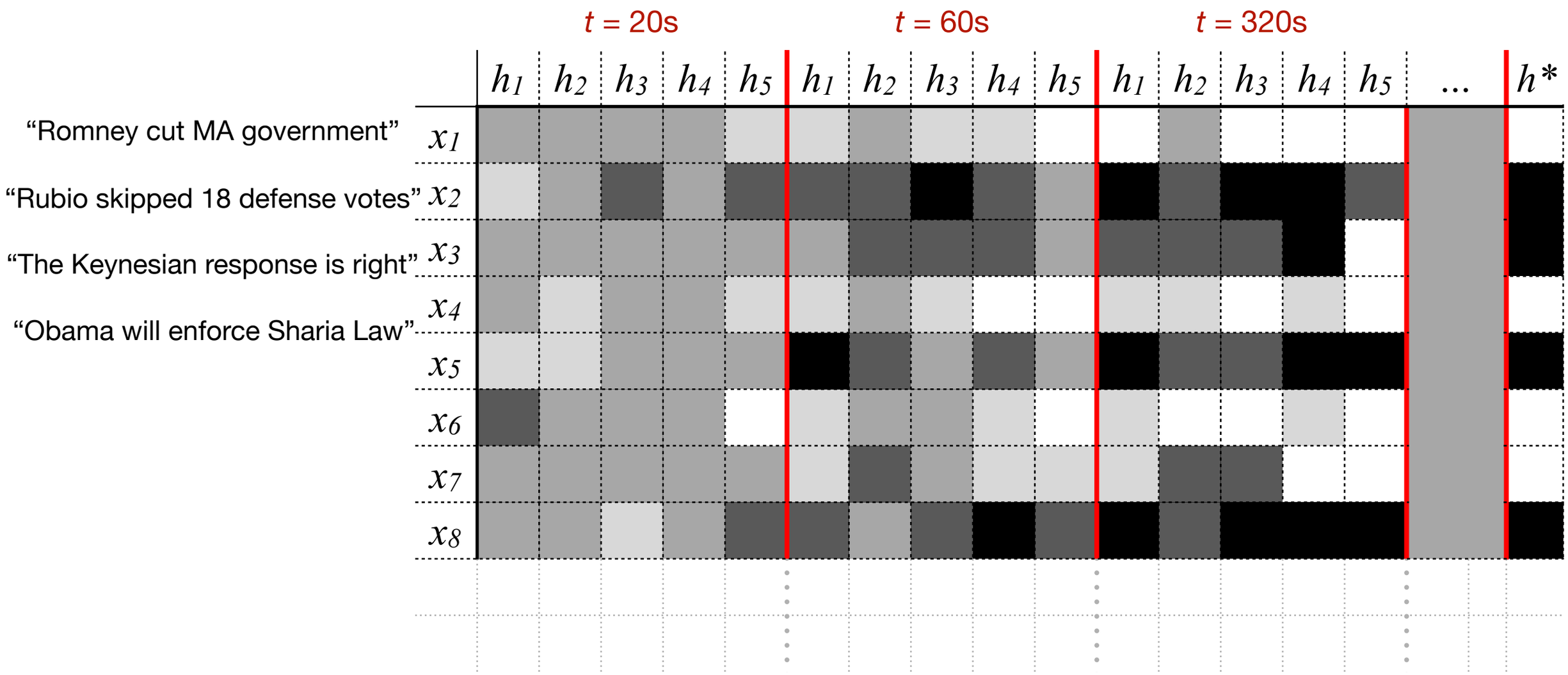
Get judgments from Bob as well as Alice:

		$h(x_i; 20)$		$h(x_i; 60)$		$h(x_i; 320)$		$h(x_i; 540)$		$h^*(x_i)$
		A	B	A	B	A	B	A	B	A
“Romney cut MA government”	x_i	50%	55%	40%	70%	20%	70%	25%	70%	0%

objective is still Alice’s slow judgment

Problem Statement

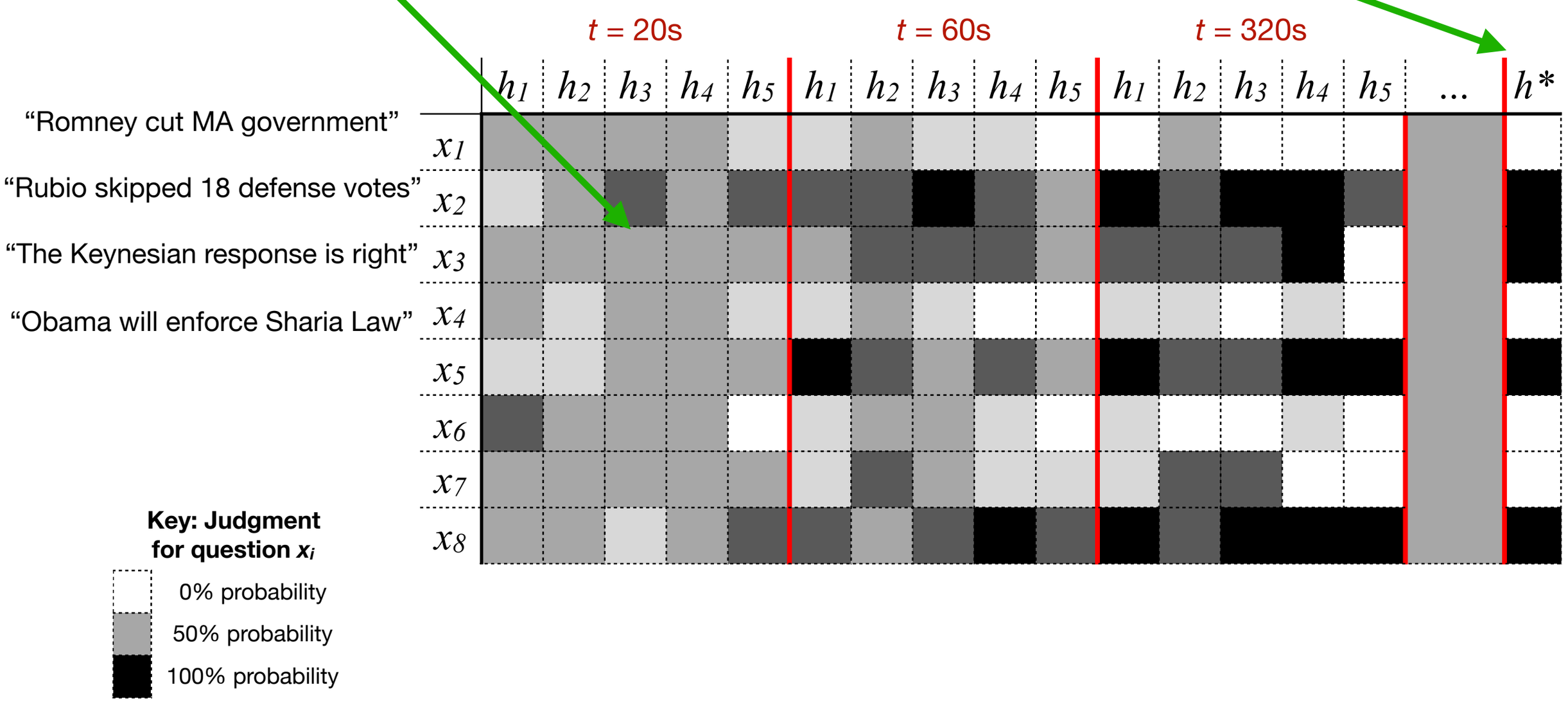
Multiple questions and multiple people:



Problem Statement

Collect most of cheapest signal

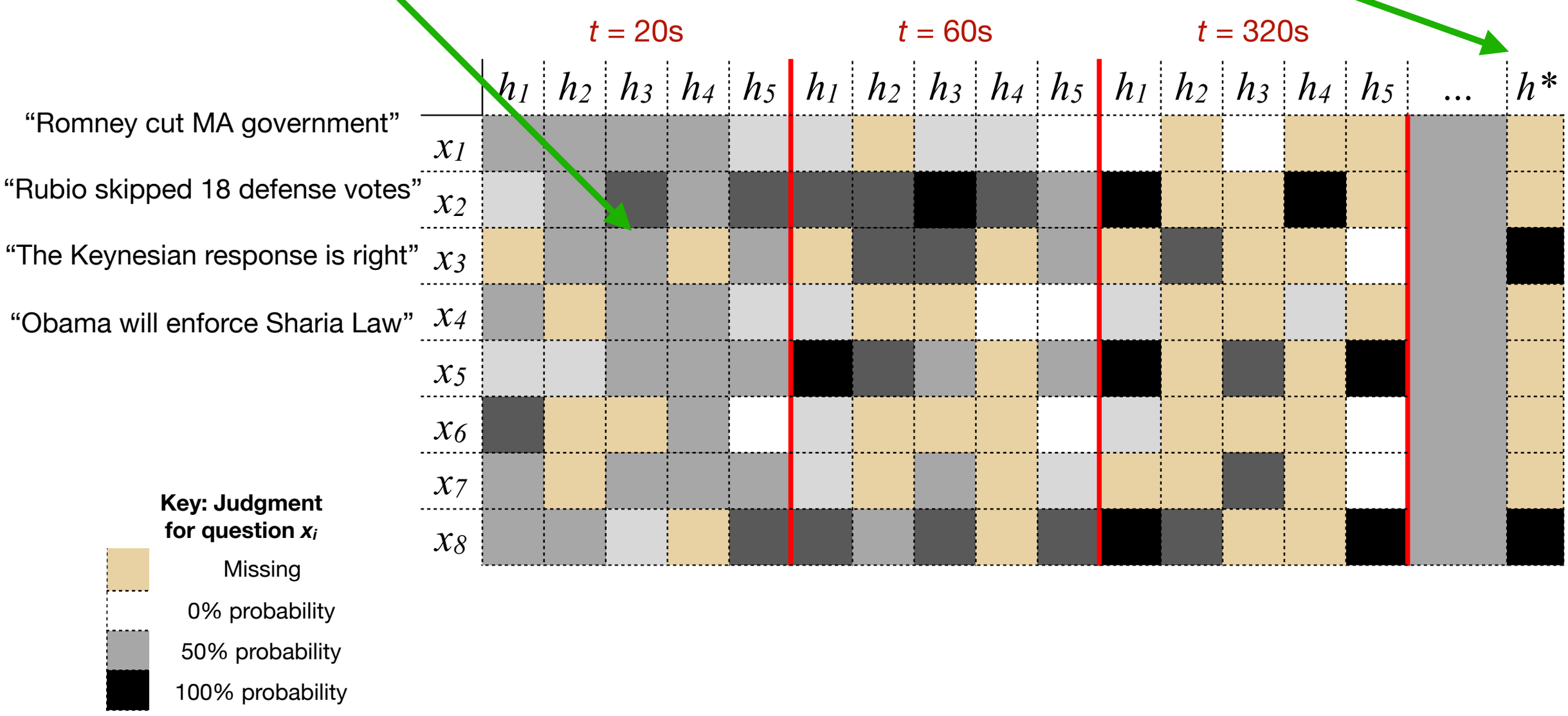
Collect least of slowest judgment



Problem Statement

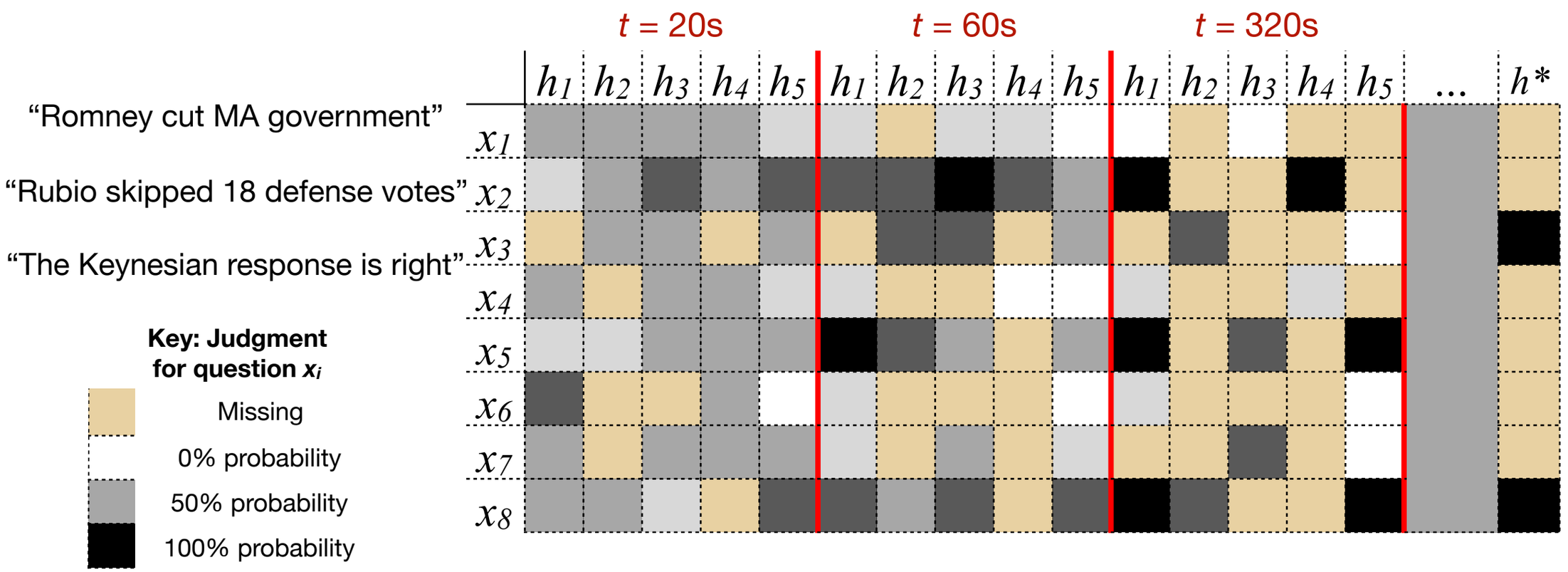
Collect most of cheapest signal

Collect least of slowest judgment



Problem statement

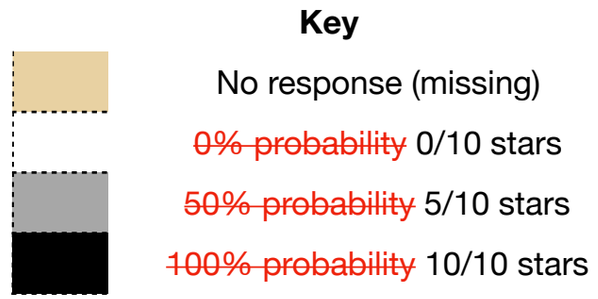
- Relation to standard ML problems:
  - Most training data is unlabeled: **Semi-supervised learning**
  - For unlabeled data, we have noisy/biased labels: **Weak supervision**
  - Item-user (sparse) matrix: **Collaborative Filtering**

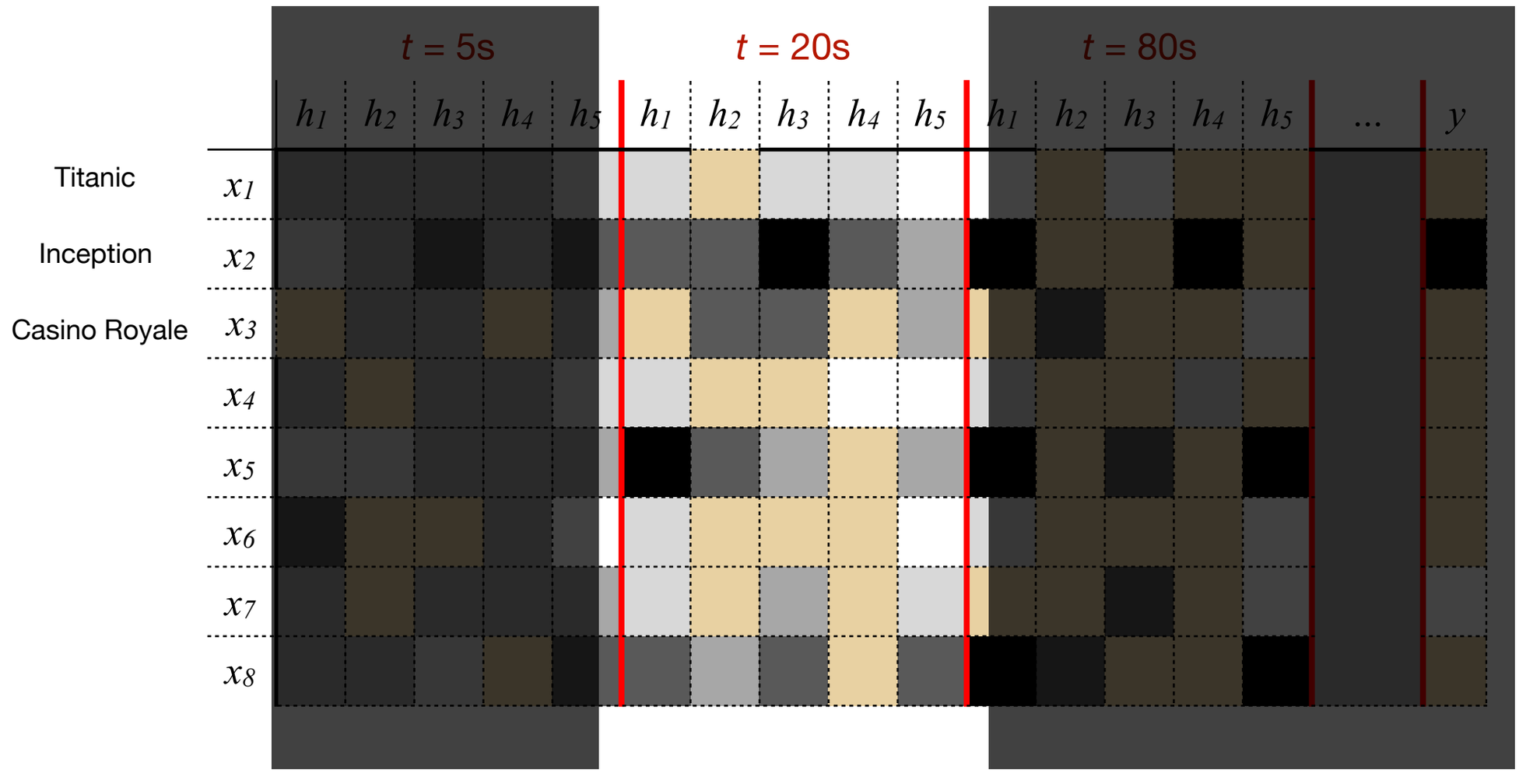


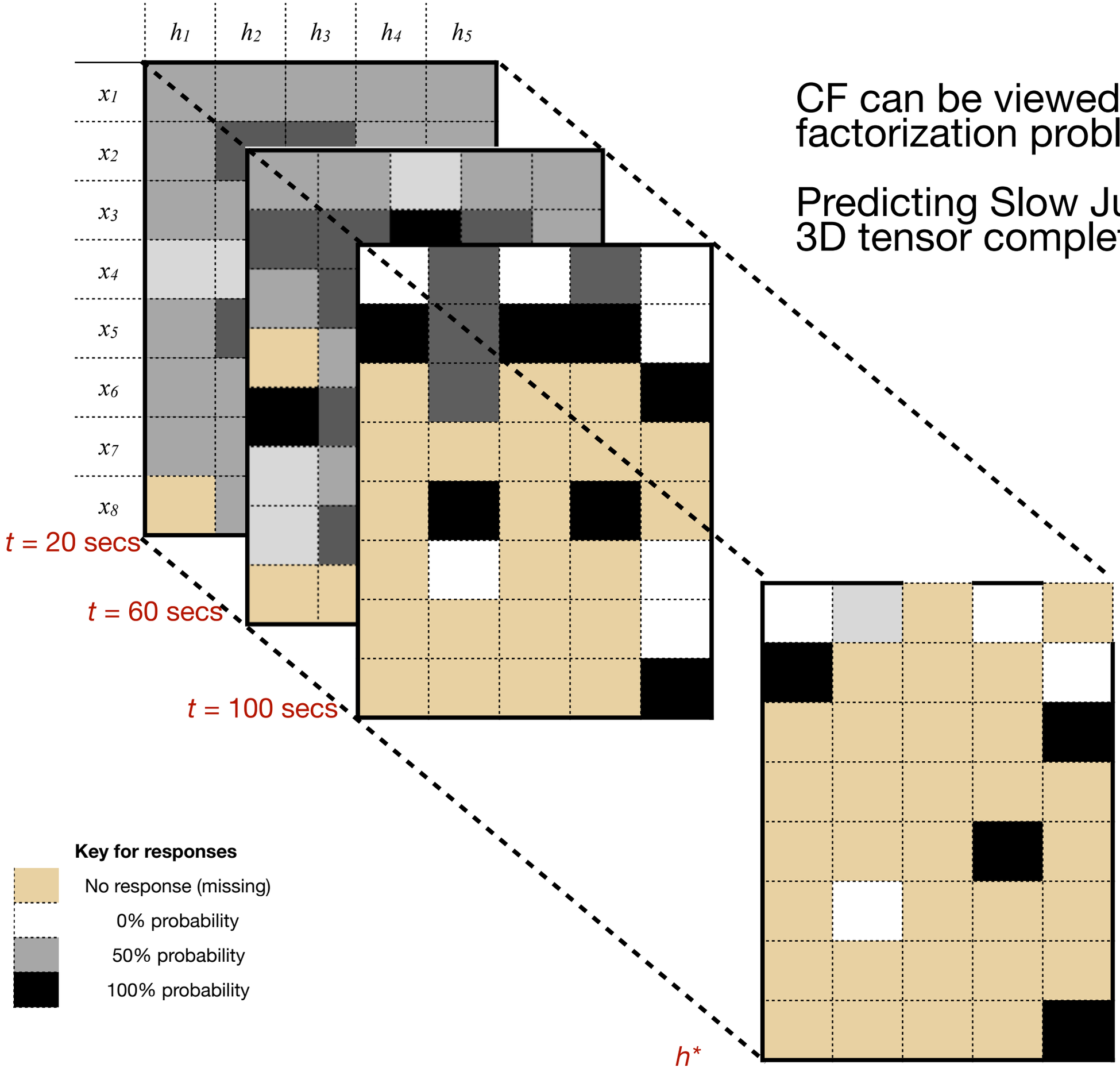
Modeling the data

Collaborative filtering (content-based)

- Human users rate movies
- ML objective (Netflix) is to predict all missing ratings







CF can be viewed as a matrix factorization problem (SVD)

Predicting Slow Judgments is a 3D tensor completion problem.

Datasets for PSJ

We are creating two new datasets:

1. **Fermi:** Fermi estimation comparisons (no research)
2. **Politifact**: Judge truth of political statement (use Google)

Fermi Examples

- weight of a bushpig in kg $<$ 99
- production budget for The Force Awakens in millions - value of 3 dimes and 2 pennies in cents - maximum speed of a horsefly $<$ 1343
- driving distance in miles between London and Amsterdam $<$ 371

Datasets for PSJ

	Fermi	Politifact
Depends on:	pure thinking / calculating	researching, interpreting language
Related to:	Related: math, science, games	Related: law, academic research

Datasets for PSJ: Fermi

Done: 0 / 10 | Score: -20.00 | Bonus: $0.00 | Total pay: $1.00

12 seconds remaining

number of letters in the English alphabet +

(173 - number of copies sold worldwide in millions of "Daydream Believer" by The Monkees)

197

What is the probability that the quantity on the right is bigger (in %)?

0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

Left is definitely bigger

No idea

Right is definitely bigger

This is your scratchpad. You may jot down any notes you have while thinking here.

Datasets for PSJ: Politifact

13 seconds remaining

Stephen Hayes (Senior writer for the weekly standard) on December 27th, 2015:

> Guantanamo has never been a key component of ISIS or al-Qaida propaganda.

Context
On "fox news sunday".

Topics
Terrorism

How likely is it that this statement is true (in %)?

0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

Definitely false — No idea — Definitely true

Google | Browse

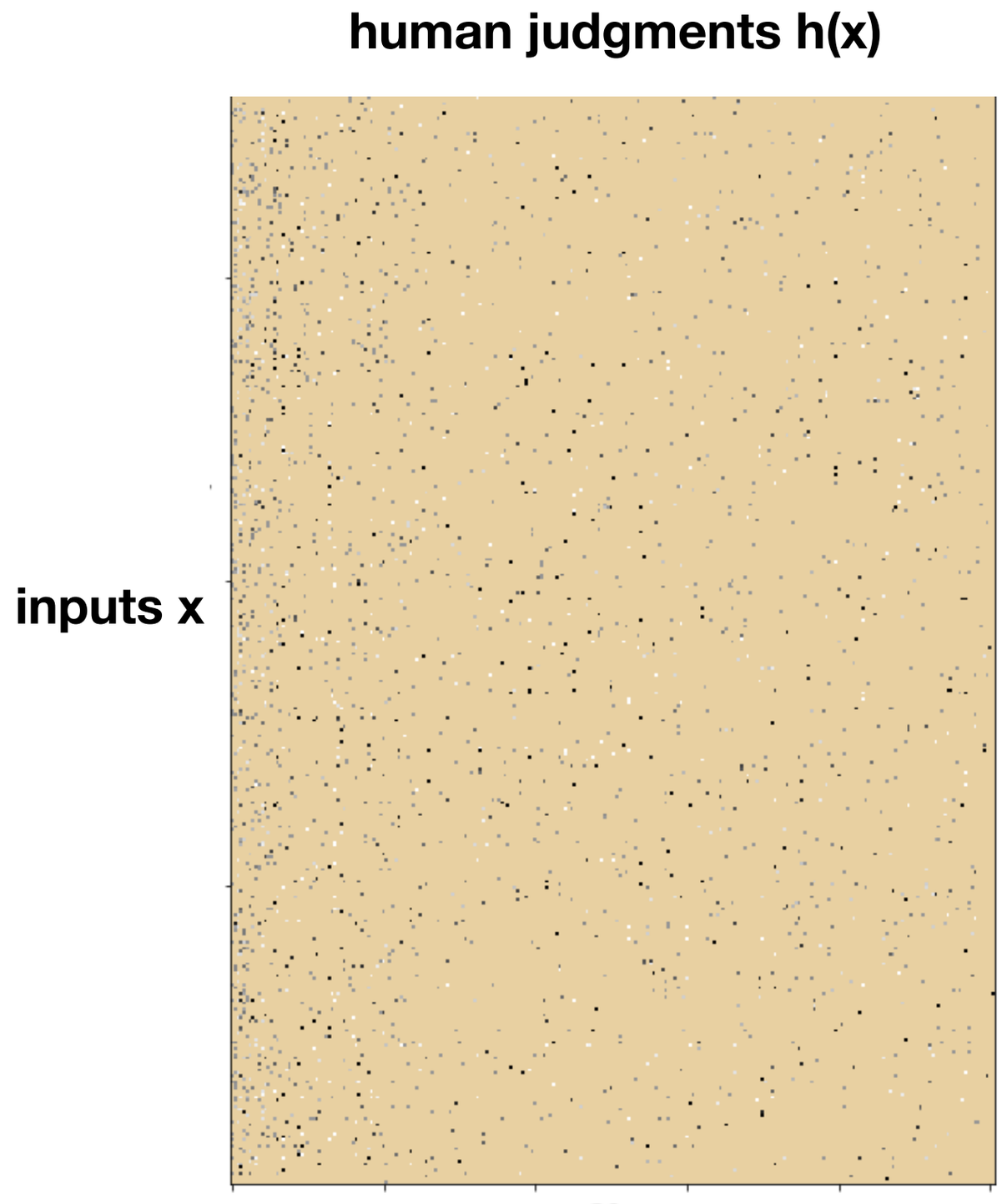
Your Google Query

Search Google

Results (Pilot)

Politifact Data

- ≈ 20,000 judgments
- ≈ 150 human judges
- ≈ 11 judgments per question (most data missing)
- Quick judgments less confident than slow



Results (Pilot)

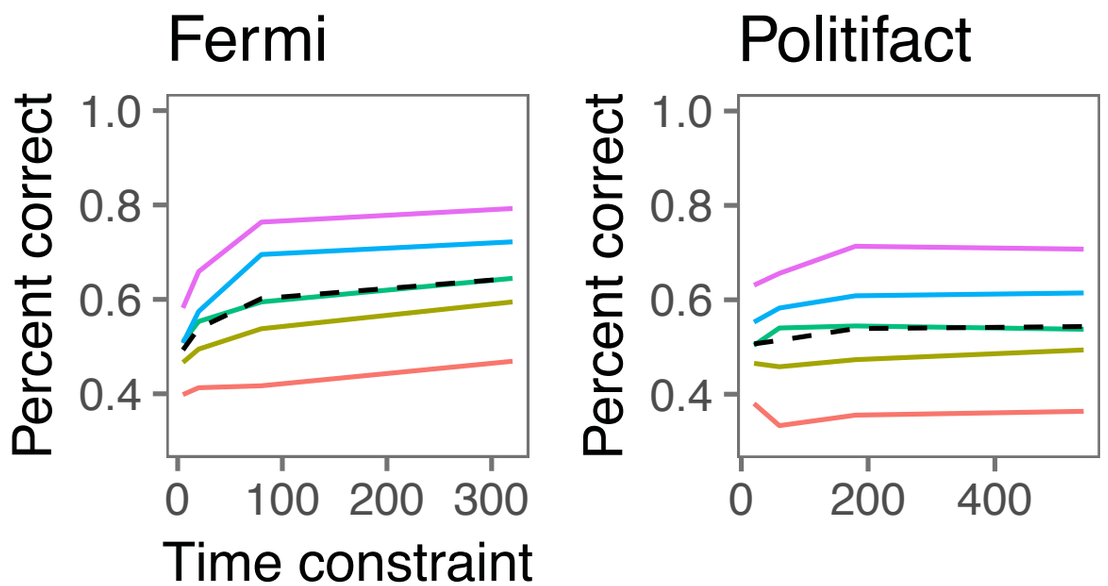
- Are the quickest judgments informative?
  YES
- Did individuals get more accurate/calibrated when given more time?

  Fermi: yes for top 20% of people.

  Politifact: Even top 20% don't improve after 200s.

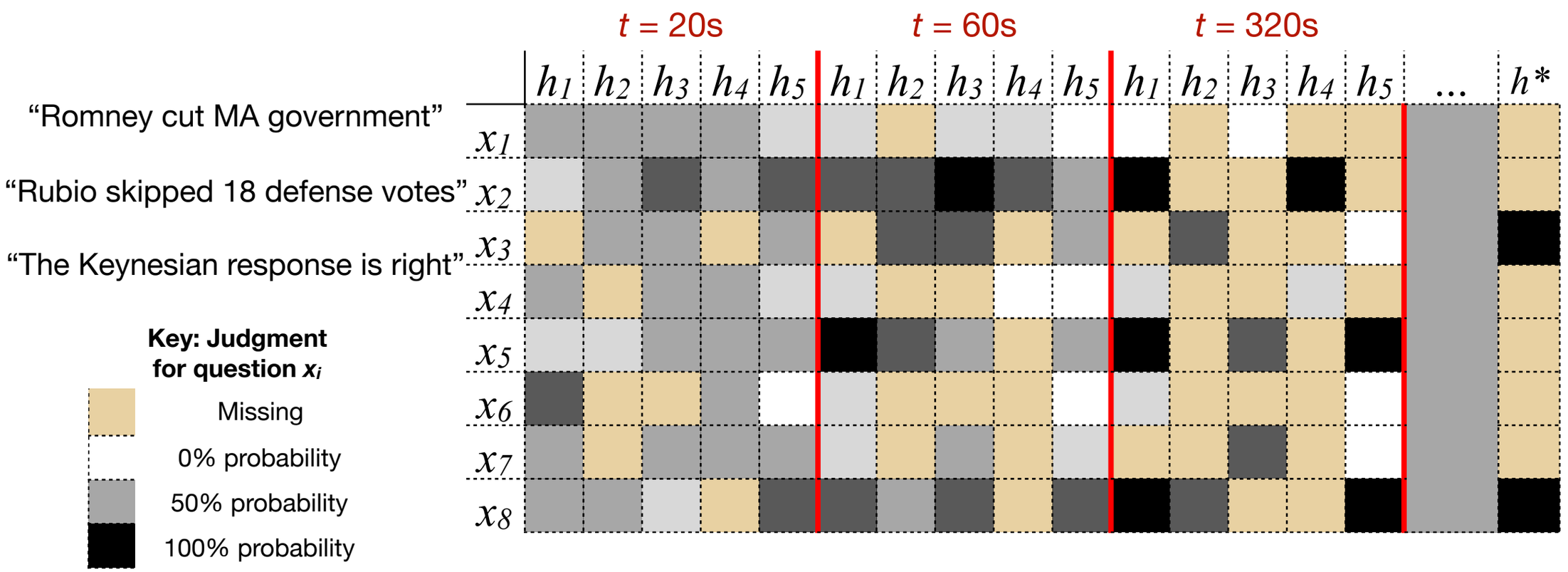
0-1 Accuracy scores

	Base-rate	Mean of quickest judgment	Mean of slowest judgment
Fermi	51%	59%	79%
Politifact	52%	56%	65%



Problem statement

- Relation to standard ML problems:
  - Most training data is unlabeled: **Semi-supervised learning**
  - For unlabeled data, we have noisy/biased labels: **Weak supervision**
  - Item-user (sparse) matrix: **Collaborative Filtering**



Modeling the data

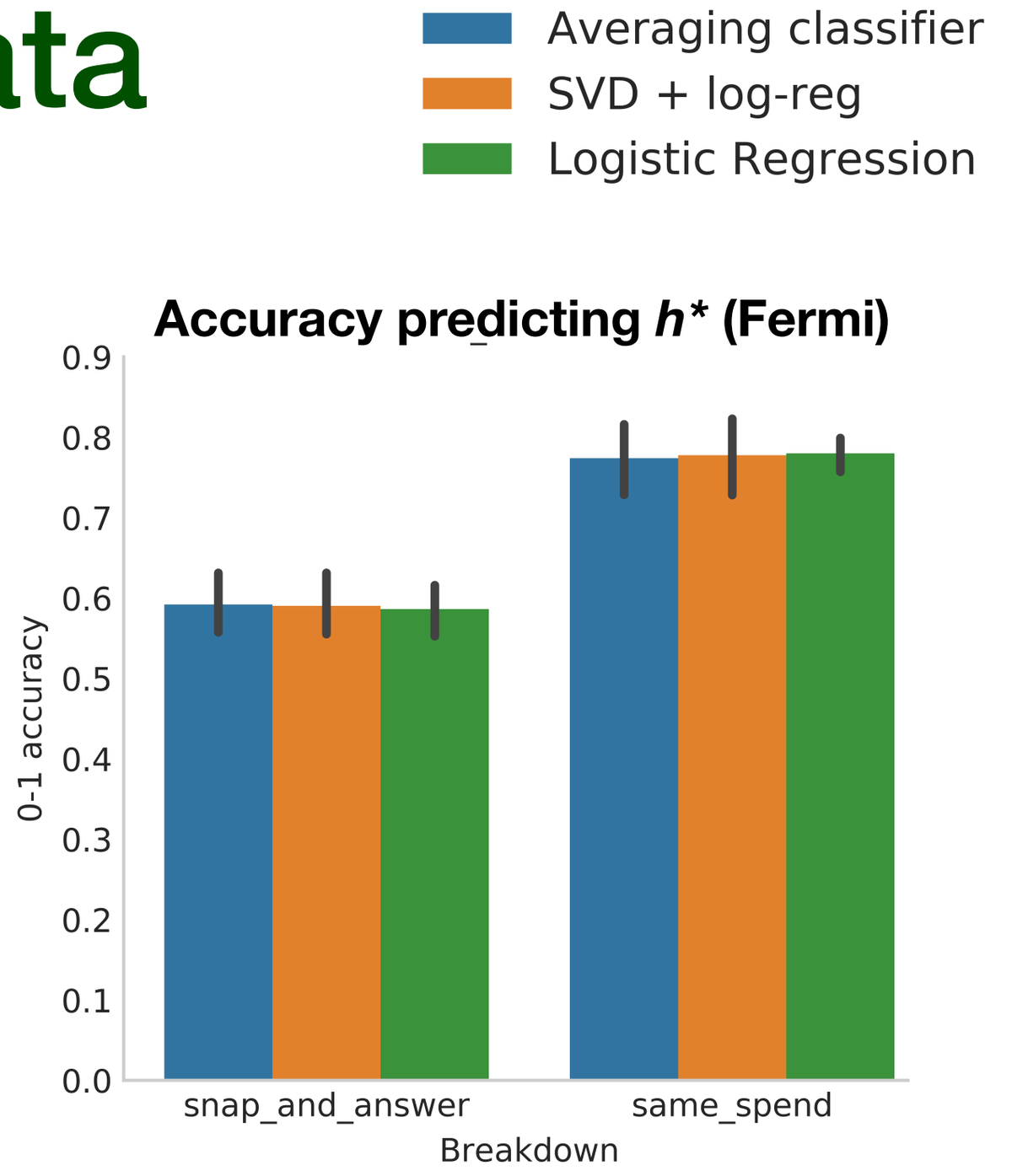
Pilot study (only 500 questions)

Fix total cost of human data, vary "breakdown":

- **snap_and_answer**: mostly quickest judgments

- **same_spend**: mix of quick and slower judgments (number of judgments proportional to cost)

Conclusion: time for research/ calculation helps models



quick/slower = 9.8 **quick/slower = 2.3**

Modeling the data

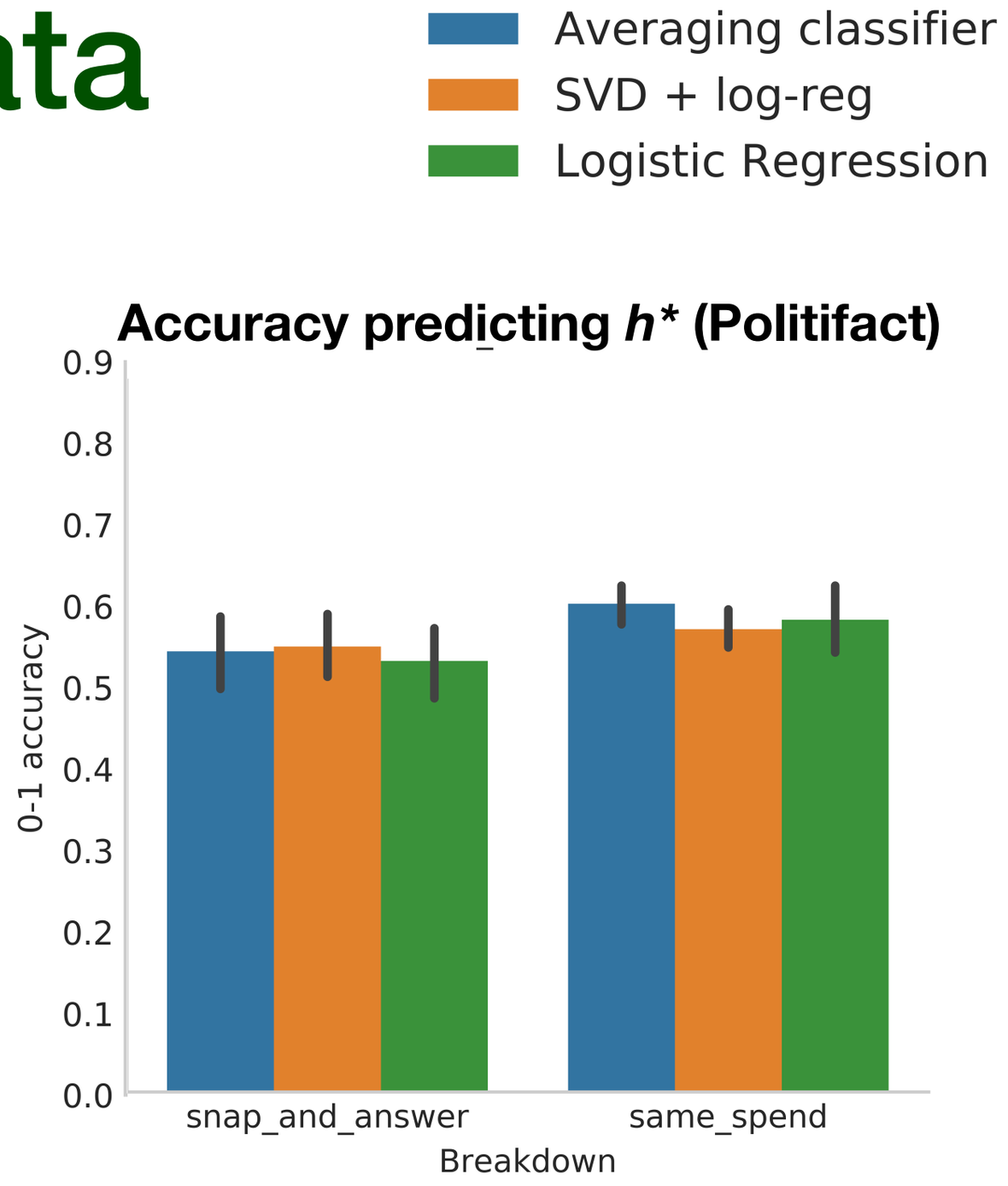
Pilot study (only 500 questions)

Fix total cost of human data, vary "breakdown":

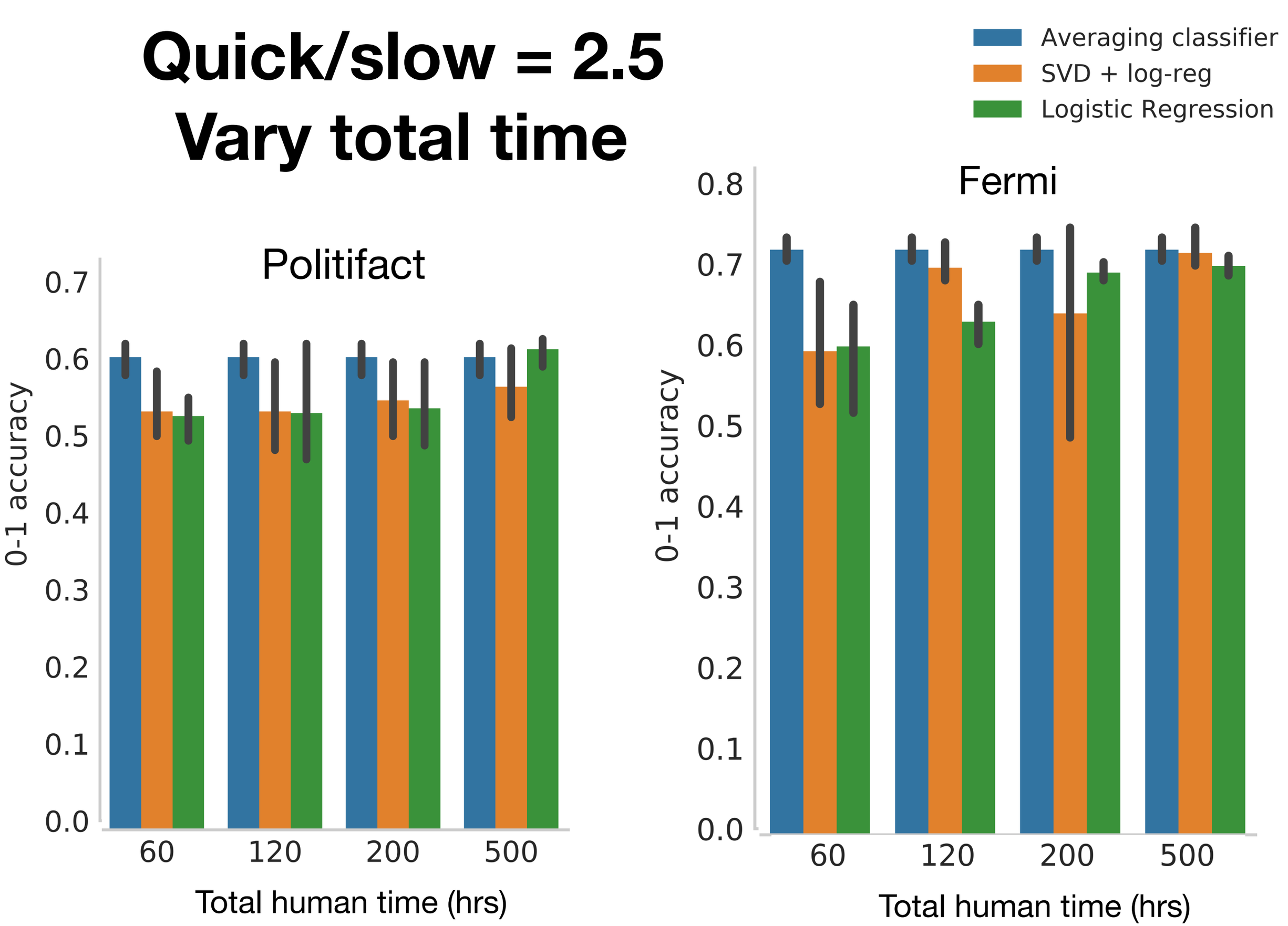
- **snap_and_answer**: mostly quickest judgments

- **same_spend**: mix of quick and slower judgments (number of judgments proportional to cost)

Conclusion: time for research/ calculation helps models



Quick/slow = 2.5
Vary total time



Future directions…

- Collect more data: see quiz.ought.org
- Try to predict slow judgments without having cheap signals at test time
- Thus need data on **how** people deliberate (e.g. scratchpad, Google history)
- Task where human judges whether A or B is better (involves preferences not just true/false)