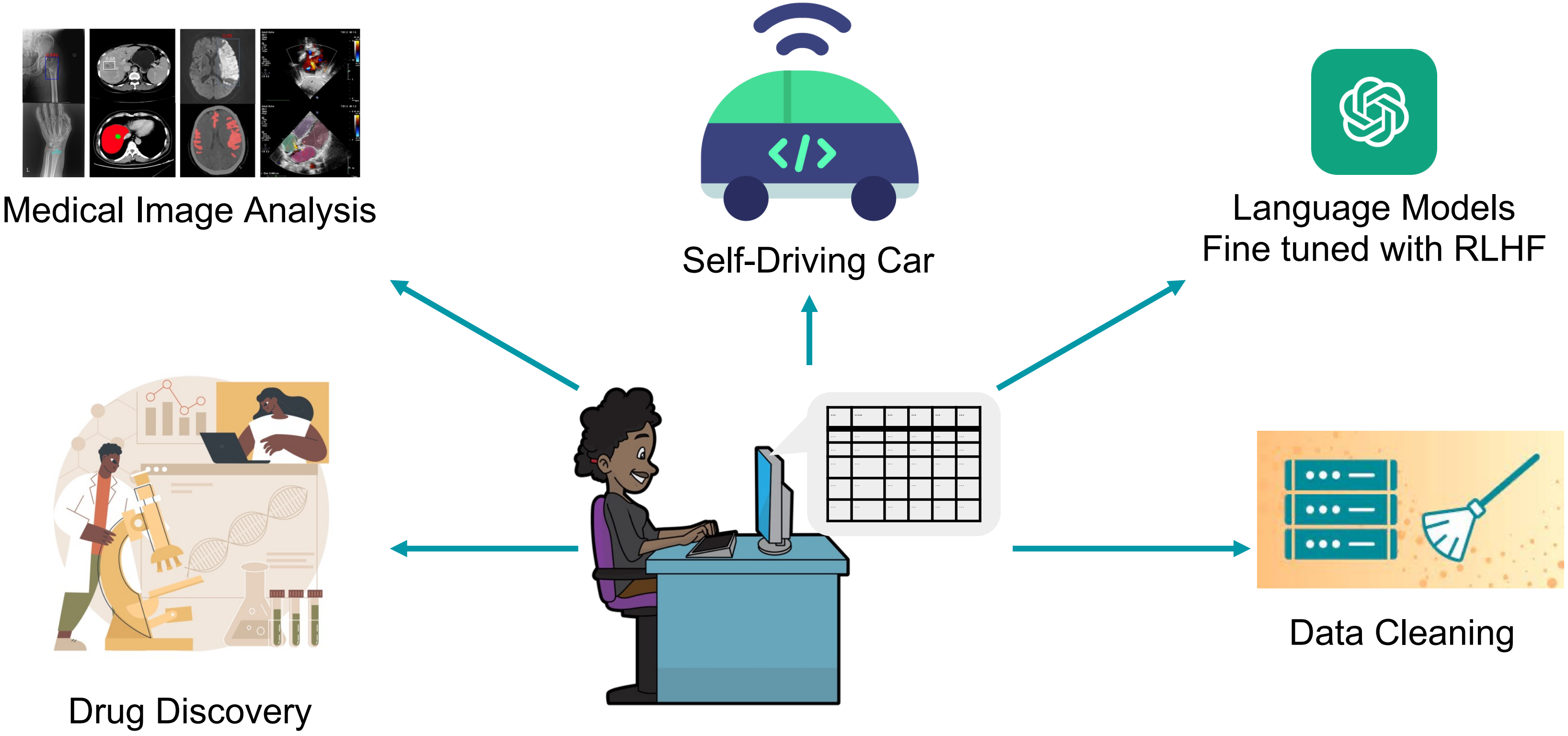
Exploratory Training: When Annotators Learn About Data

Rajesh Shrestha — Oregon State University
Omeed Habibelahian — Oregon State University
Arash Termehchy — Oregon State University
Paolo Papotti — EURECOM

Human labeling powers important ML applications

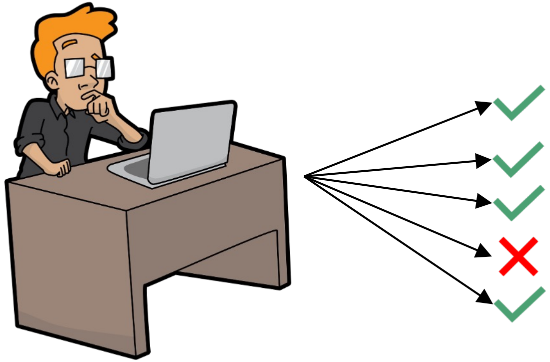


https://commons.wikimedia.org/wiki/Main_Page

Challenge: humans are not perfect!

- Humans may not know the data well enough
  - learn about data during labeling
  - provide incorrect labels

Example: Anomaly detection of account records in ABC bank



Name	Age	Education	Job	Salary	Current Balance
Bob	30	Bachelors	Intern	\$1k	-\$10k
Alex	21	Masters	Teacher	\$3k	\$30k
Diana	41	High School	Photographer	\$2k	\$10k
Lisa	36	High School	Sales Manager	\$7k	\$25k
Jane	29	No education	StoreKeeper	\$2k	\$9k

https://commons.wikimedia.org/wiki/Main_Page

Example: user observes first set of training data

*Negative balances (**-$10k**) seem to be anomalies.*

Name	Age	Education	Job	Salary	Current Balance
Bob	30	Bachelors	Intern	$1k	**-$10k**
Alex	21	Masters	Teacher	$3k	$30k
Diana	41	High School	Photographer	$2k	$10k
Lisa	36	High School	Sales Manager	$7k	$25k
Jane	29	No education	StoreKeeper	$2k	$9k

Name	Age	Education	Job	Salary	Current Balance
Bob	30	Bachelors	Intern	$1k	**-$10k**
Alex	21	Masters	Teacher	$3k	$30k
Diana	41	High School	Photographer	$2k	$10k
Lisa	36	High School	Sales Manager	$7k	$25k
Jane	29	No education	StoreKeeper	$2k	$9k
Adele	30	Bachelors	Librarian	$3k	**-$8k**
Haley	36	High School	Construction Laborer	$7k	$25k
James	21	Masters	Surgeon	$15k	$80k
Mark	19	No education	Electrician	$5k	$25k
Peter	41	High School	No Job	$0	**-$2k**
Adele	30	Bachelors	Librarian	$3k	**-$8k**
Haley	36	High School	Construction Laborer	$7k	$25k
James	21	Masters	Surgeon	$15k	$80k
Mark	19	No education	Electrician	$5k	$25k
Peter	41	High School	No Job	$0	**-$2k**
Adele	30	Bachelors	Librarian	$3k	**-$8k**
Haley	36	High School	Construction Laborer	$7k	$25k
James	21	Masters	Surgeon	$15k	$80k
Mark	19	No education	Electrician	$5k	$25k
Peter	41	High School	No Job	$0	**-$2k**
Adele	30	Bachelors	Librarian	$3k	**-$8k**
Haley	36	High School	Construction Laborer	$7k	$25k
James	21	Masters	Surgeon	$15k	$80k
Mark	19	No education	Electrician	$5k	$25k
Peter	41	High School	No Job	$0	**-$2k**

.
.
.

Dataset

Example: user labels based on initial belief

*Negative balances (**-$10k**) seem to be anomalies.*

Name	Age	Education	Job	Salary	Current Balance
Bob	30	Bachelors	Intern	$1k	**-$10k**
Alex	21	Masters	Teacher	$3k	$30k
Diana	41	High School	Photographer	$2k	$10k
Lisa	36	High School	Sales Manager	$7k	$25k
Jane	29	No education	StoreKeeper	$2k	$9k

Name	Age	Education	Job	Salary	Current Balance
Bob	30	Bachelors	Intern	$1k	**-$10k**
Alex	21	Masters	Teacher	$3k	$30k
Diana	41	High School	Photographer	$2k	$10k
Lisa	36	High School	Sales Manager	$7k	$25k
Jane	29	No education	StoreKeeper	$2k	$9k
Adele	30	Bachelors	Librarian	$3k	**-$8k**
Haley	36	High School	Construction Laborer	$7k	$25k
James	21	Masters	Surgeon	$15k	$80k
Mark	19	No education	Electrician	$5k	$25k
Peter	41	High School	No Job	$0	**-$2k**
Adele	30	Bachelors	Librarian	$3k	**-$8k**
Haley	36	High School	Construction Laborer	$7k	$25k
James	21	Masters	Surgeon	$15k	$80k
Mark	19	No education	Electrician	$5k	$25k
Peter	41	High School	No Job	$0	**-$2k**
Adele	30	Bachelors	Librarian	$3k	**-$8k**
Haley	36	High School	Construction Laborer	$7k	$25k
James	21	Masters	Surgeon	$15k	$80k
Mark	19	No education	Electrician	$5k	$25k
Peter	41	High School	No Job	$0	**-$2k**

.
.
.

Dataset

Example: user observes new training data

*What about the one with **-$8k** and **-$2k** balance in this sample?*

Name	Age	Education	Job	Salary	Current Balance
Adele	30	Bachelors	Librarian	$3k	**-$8k** ?
Haley	36	High School	Construction Laborer	$7k	$25k
James	21	Masters	Surgeon	$15k	$80k
Mark	19	No education	Electrician	$5k	$25k
Peter	41	High School	No Job	$0	**-$2k** ?

Name	Age	Education	Job	Salary	Current Balance
Bob	30	Bachelors	Intern	$1k	**-$10k**
Alex	21	Masters	Teacher	$3k	$30k
Diana	41	High School	Photographer	$2k	$10k
Lisa	36	High School	Sales Manager	$7k	$25k
Jane	29	No education	StoreKeeper	$2k	$9k
Adele	30	Bachelors	Librarian	$3k	**-$8k**
Haley	36	High School	Construction Laborer	$7k	$25k
James	21	Masters	Surgeon	$15k	$80k
Mark	19	No education	Electrician	$5k	$25k
Peter	41	High School	No Job	$0	**-$2k**
Adele	30	Bachelors	Librarian	$3k	**-$8k**
Haley	36	High School	Construction Laborer	$7k	$25k
James	21	Masters	Surgeon	$15k	$80k
Mark	19	No education	Electrician	$5k	$25k
Peter	41	High School	No Job	$0	**-$2k**
Adele	30	Bachelors	Librarian	$3k	**-$8k**
Haley	36	High School	Construction Laborer	$7k	$25k
James	21	Masters	Surgeon	$15k	$80k
Mark	19	No education	Electrician	$5k	$25k
Peter	41	High School	No Job	$0	**-$2k**
Adele	30	Bachelors	Librarian	$3k	**-$8k**
Haley	36	High School	Construction Laborer	$7k	$25k
James	21	Masters	Surgeon	$15k	$80k
Mark	19	No education	Electrician	$5k	$25k
Peter	41	High School	No Job	$0	**-$2k**

.
.
.

Dataset

Example: user recalls the previous data rows

Wait… I have seen people like this with negative balance.

Name	Age	Education	Job	Salary	Current Balance
Adele	30	Bachelors	Librarian	$3k	**-$8k** ?
Haley	36	High School	Construction Laborer	$7k	$25k
James	21	Masters	Surgeon	$15k	$80k
Mark	19	No education	Electrician	$5k	$25k
Peter	41	High School	No Job	$0	**-$2k** ?

Name	Age	Education	Job	Salary	Current Balance
Bob	30	Bachelors	Intern	$1k	**-$10k**
Alex	21	Masters	Teacher	$3k	$30k
Diana	41	High School	Photographer	$2k	$10k
Lisa	36	High School	Sales Manager	$7k	$25k
Jane	29	No education	StoreKeeper	$2k	$9k
Adele	30	Bachelors	Librarian	$3k	**-$8k**
Haley	36	High School	Construction Laborer	$7k	$25k
James	21	Masters	Surgeon	$15k	$80k
Mark	19	No education	Electrician	$5k	$25k
Peter	41	High School	No Job	$0	**-$2k**
Adele	30	Bachelors	Librarian	$3k	**-$8k**
Haley	36	High School	Construction Laborer	$7k	$25k
James	21	Masters	Surgeon	$15k	$80k
Mark	19	No education	Electrician	$5k	$25k
Peter	41	High School	No Job	$0	**-$2k**
Adele	30	Bachelors	Librarian	$3k	**-$8k**
Haley	36	High School	Construction Laborer	$7k	$25k
James	21	Masters	Surgeon	$15k	$80k
Mark	19	No education	Electrician	$5k	$25k
Peter	41	High School	No Job	$0	**-$2k**
Adele	30	Bachelors	Librarian	$3k	**-$8k**
Haley	36	High School	Construction Laborer	$7k	$25k
James	21	Masters	Surgeon	$15k	$80k
Mark	19	No education	Electrician	$5k	$25k
Peter	41	High School	No Job	$0	**-$2k**

.
.
.

Dataset

Example: user recalls the previous data rows

They also had very low salary.

Salary	Current Balance
$1k	**-$10k**

Name	Age	Education	Job	Salary	Current Balance
Adele	30	Bachelors	Librarian	$3k	**-$8k**
Haley	36	High School	Construction Laborer	$7k	$25k
James	21	Masters	Surgeon	$15k	$80k
Mark	19	No education	Electrician	$5k	$25k
Peter	41	High School	No Job	$0	**-$2k**

Name	Age	Education	Job	Salary	Current Balance
Bob	30	Bachelors	Intern	$1k	**-$10k**
Alex	21	Masters	Teacher	$3k	$30k
Diana	41	High School	Photographer	$2k	$10k
Lisa	36	High School	Sales Manager	$7k	$25k
Jane	29	No education	StoreKeeper	$2k	$9k
Adele	30	Bachelors	Librarian	$3k	**-$8k**
Haley	36	High School	Construction Laborer	$7k	$25k
James	21	Masters	Surgeon	$15k	$80k
Mark	19	No education	Electrician	$5k	$25k
Peter	41	High School	No Job	$0	**-$2k**
Adele	30	Bachelors	Librarian	$3k	**-$8k**
Haley	36	High School	Construction Laborer	$7k	$25k
James	21	Masters	Surgeon	$15k	$80k
Mark	19	No education	Electrician	$5k	$25k
Peter	41	High School	No Job	$0	**-$2k**
Adele	30	Bachelors	Librarian	$3k	**-$8k**
Haley	36	High School	Construction Laborer	$7k	$25k
James	21	Masters	Surgeon	$15k	$80k
Mark	19	No education	Electrician	$5k	$25k
Peter	41	High School	No Job	$0	**-$2k**
Adele	30	Bachelors	Librarian	$3k	**-$8k**
Haley	36	High School	Construction Laborer	$7k	$25k
James	21	Masters	Surgeon	$15k	$80k
Mark	19	No education	Electrician	$5k	$25k
Peter	41	High School	No Job	$0	**-$2k**

.
.
.

Dataset

Example: user learns and changes their belief

Negative Balance is okay for them.
These aren't anomaly.

Name	**Age**	**Education**	**Job**	**Salary**	**Current Balance**
Adele	30	Bachelors	Librarian	$3k	**-$8k**
Haley	36	High School	Construction Laborer	$7k	$25k
James	21	Masters	Surgeon	$15k	$80k
Mark	19	No education	Electrician	$5k	$25k
Peter	41	High School	No Job	$0	**-$2k**

Name	**Age**	**Education**	**Job**	**Salary**	**Current Balance**
Bob	30	Bachelors	Intern	$1k	**-$10k**
Alex	21	Masters	Teacher	$3k	$30k
Diana	41	High School	Photographer	$2k	$10k
Lisa	36	High School	Sales Manager	$7k	$25k
Jane	29	No education	StoreKeeper	$2k	$9k
Adele	30	Bachelors	Librarian	$3k	**-$8k**
Haley	36	High School	Construction Laborer	$7k	$25k
James	21	Masters	Surgeon	$15k	$80k
Mark	19	No education	Electrician	$5k	$25k
Peter	41	High School	No Job	$0	**-$2k**
Adele	30	Bachelors	Librarian	$3k	**-$8k**
Haley	36	High School	Construction Laborer	$7k	$25k
James	21	Masters	Surgeon	$15k	$80k
Mark	19	No education	Electrician	$5k	$25k
Peter	41	High School	No Job	$0	**-$2k**
Adele	30	Bachelors	Librarian	$3k	**-$8k**
Haley	36	High School	Construction Laborer	$7k	$25k
James	21	Masters	Surgeon	$15k	$80k
Mark	19	No education	Electrician	$5k	$25k
Peter	41	High School	No Job	$0	**-$2k**
Adele	30	Bachelors	Librarian	$3k	**-$8k**
Haley	36	High School	Construction Laborer	$7k	$25k
James	21	Masters	Surgeon	$15k	$80k
Mark	19	No education	Electrician	$5k	$25k
Peter	41	High School	No Job	$0	**-$2k**

.
.
.

Dataset

Current methods assume **correct** and **fixed** labeler's belief

- Users know correct labels from the start
  - Potentially low fixed chance of making mistake

- Users don't change belief about data during labeling
  - Fixed belief

- **They may fail to learn accurate models**
- Evolving user belief, many errors, non-stationary errors, ...

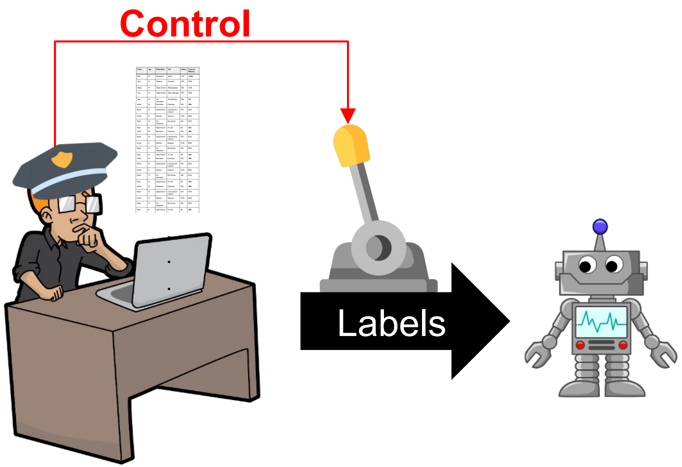
https://depositphotos.com/

Current solution: manual exploration then labeling

- User decides when to start labeling

- **Step 1:** user explores data till confident on labeling
- **Step 2:** provides labeling

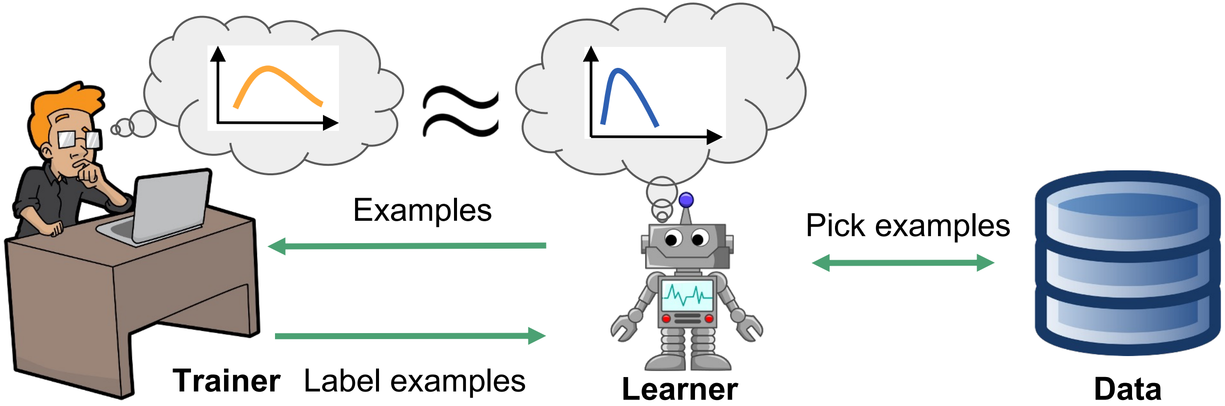
- **Shortcomings:**
  - take long time
    - users don't know when to stop exploring
  - might still change belief on labeling
  - forget information on long exploration



Our vision: understand and adapt to human learning

Training: reaching an agreement between labeler & learner about the model

- ➢ Both start with some prior belief
- ➢ Learner picks some examples and ask for labels
- ➢ Labeler observes examples, updates its belief, and labels examples accordingly
- ➢ Learner updates its model and policy of picking examples
- ➢ Until both reach the same belief about the model



Components of our proposal

Part 1: Model human learning

- Model learning and belief change for humans

Part 2: Collaborative learning

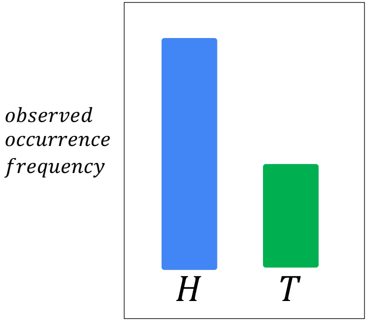
- Pick & show examples adaptively based on human learning
- Update model belief and policy of choosing examples
- **Goal**: human and system converge to an accurate model

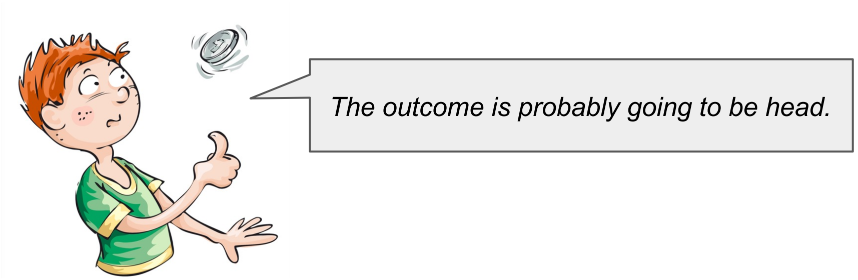
Human learning in interactive setting

- Cognitive psychology/ economics
- **Two components**
  - **Prediction model:** updates belief using interaction data
  - **Response model:** chooses policy based on updated belief

Human prediction model: Fictitious Play(FP)

- ➢ Simplest method
- ➢ Updates belief based on observed empirical frequencies





Human prediction model: Bayesian

- Use Bayesian rule to modify prior belief

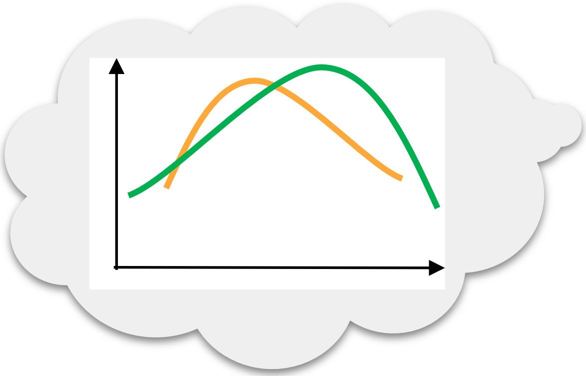
Prior Belief **Observe Data Sample** **Posterior Belief**

- Under mild assumptions in discrete space, it is equivalent to FP.

https://depositphotos.com/8009188/stock-illustration-coin-toss.html

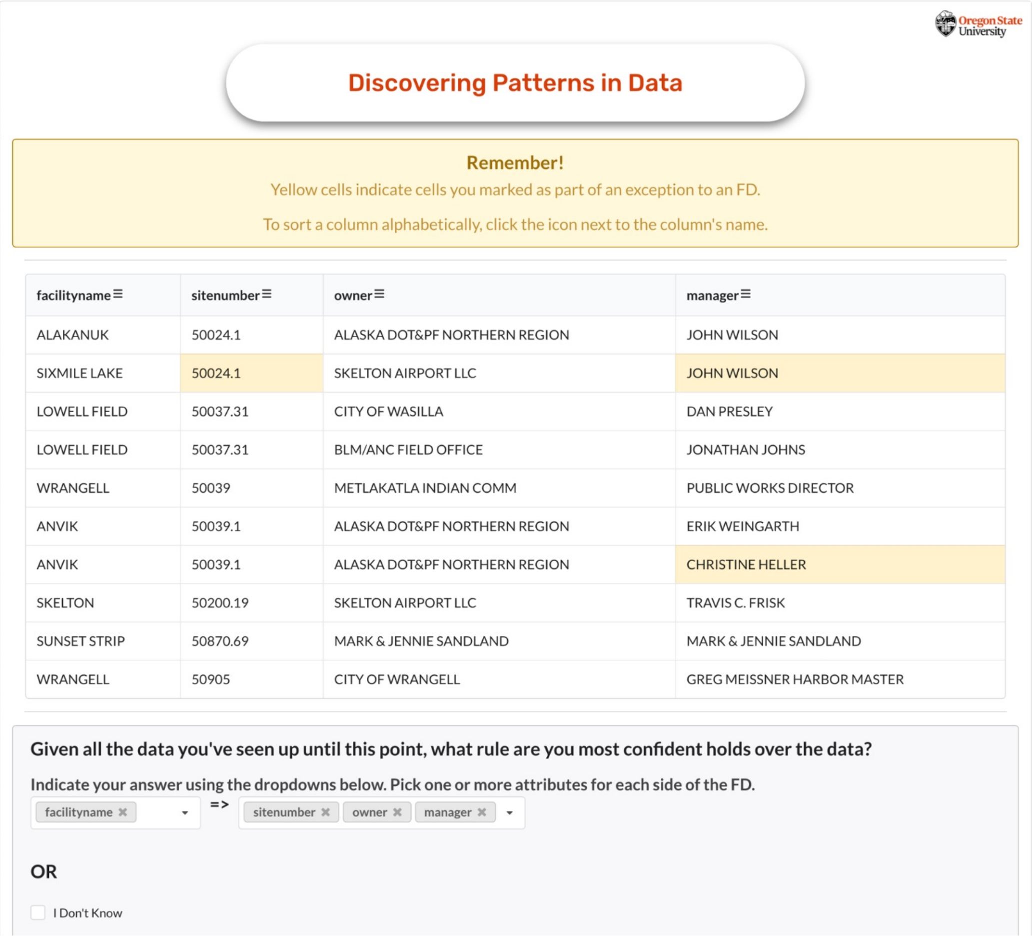
Human prediction model: hypothesis testing

- Keep plausible model based on the observed information
  - Select initial model
  - Validate current model once every **K** interactions
  - Keep model till it explains the observed data **(error < threshold)**
  - Choose new model if threshold exceeded



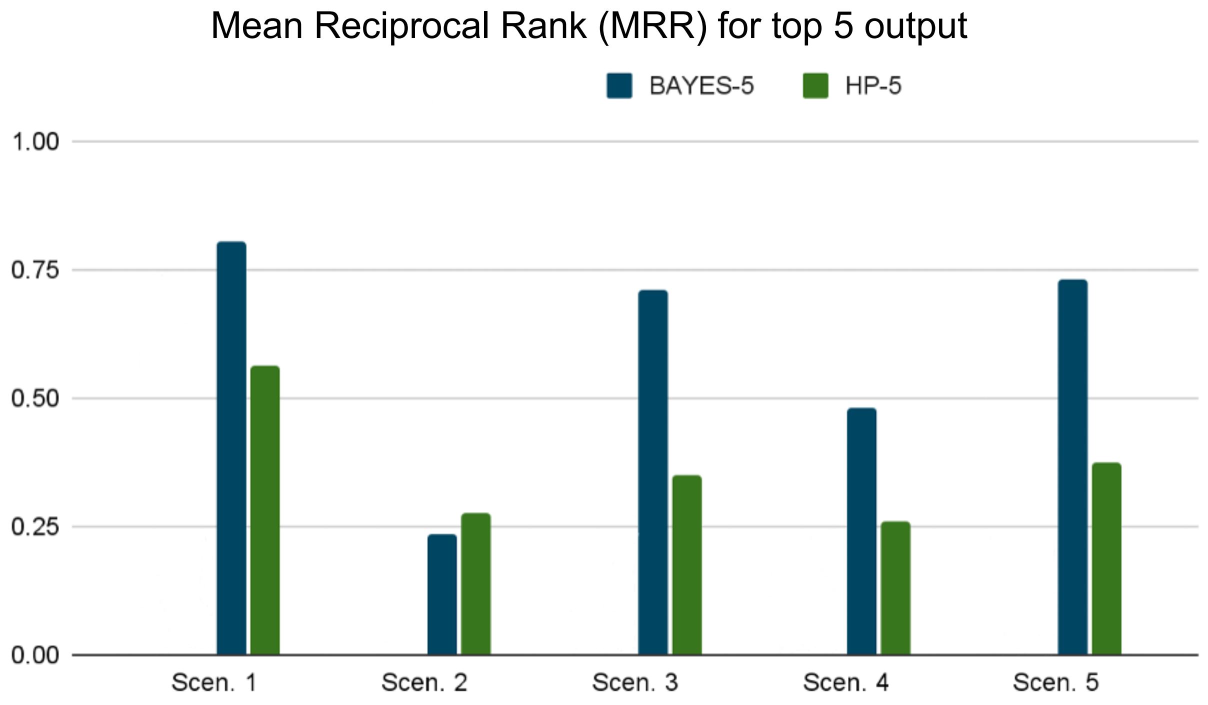
User study: Setting

- 17 participants
- Learning FDs over noisy data
- 5 scenarios:
  - Participants familiarity
  - Degree of difficulty
- Users asked for initial belief
- Users mark violations and give updated belief in each interaction



facilityname	sitenumber	owner	manager
ALAKANUK	50024.1	ALASKA DOT&PF NORTHERN REGION	JOHN WILSON
SIXMILE LAKE	50024.1	SKELTON AIRPORT LLC	JOHN WILSON
LOWELL FIELD	50037.31	CITY OF WASILLA	DAN PRESLEY
LOWELL FIELD	50037.31	BLM/ANC FIELD OFFICE	JONATHAN JOHNS
WRANGELL	50039	METLAKATLA INDIAN COMM	PUBLIC WORKS DIRECTOR
ANVIK	50039.1	ALASKA DOT&PF NORTHERN REGION	ERIK WEINGARTH
ANVIK	50039.1	ALASKA DOT&PF NORTHERN REGION	CHRISTINE HELLER
SKELTON	50200.19	SKELTON AIRPORT LLC	TRAVIS C. FRISK
SUNSET STRIP	50870.69	MARK & JENNIE SANDLAND	MARK & JENNIE SANDLAND
WRANGELL	50905	CITY OF WRANGELL	GREG MEISSNER HARBOR MASTER

How well do models replicate human learning?



Bayesian model perform better in majority.

Human learning (FD discovery)

- **Prediction model:** FP (Bayesian)
- **Response model:** **Best response**
  - **Label examples according to current belief**
    - **Label a tuple as noisy if noisy with probability > 0.5**

Learner: prediction and response models

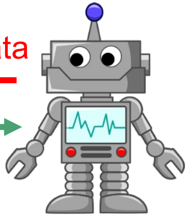
- Prediction model: updates belief about user's belief
- Response model: selects new examples for labeling based on belief

New Examples

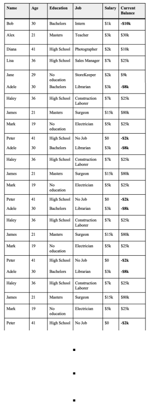
Name	Age	Education	Job	Salary	Current Balance
Bob	30	Bachelors	Intern	$1k	-$10k
Alex	21	Masters	Teacher	$3k	$30k
Diana	41	High School	Photographer	$2k	$10k
Lisa	36	High School	Sales Manager	$7k	$25k
Jane	29	No education	StoreKeeper	$2k	$9k

Unlabeled data

Labels



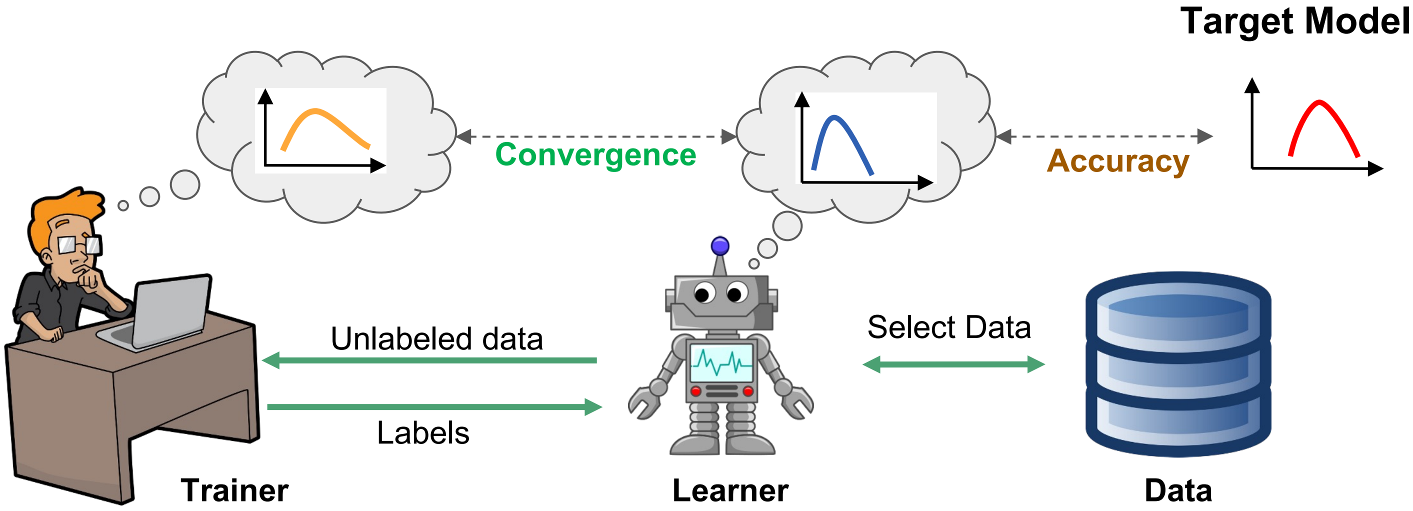
Select Data



Large Data

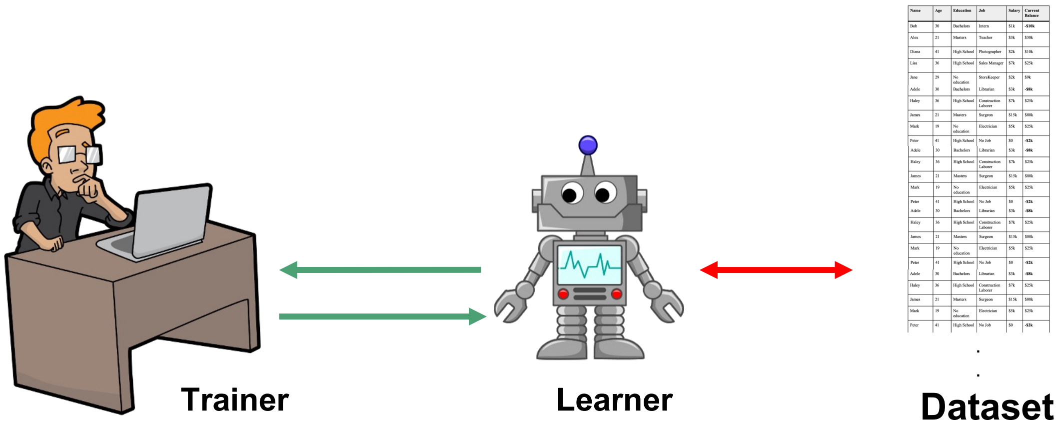
Learner: objective function

- **Agreement with user**
  - Converge to same belief
  - Show examples on which it agrees with user (**close to current user's labeling**)
- **Accurate belief**
  - Provide **informative examples**
  - Show diverse sets of examples (**entropy**)



Learner: proposed response models

- **Stochastic Best Response**
  - Pick examples stochastically using *softmax* of its predicted objective
  - Balances closeness and informativeness
- **Stochastic Uncertainty Sampling**
  - Stochastic form of *Uncertainty Sampling*
  - Pick a diverse set of uncertain examples



Convergence guarantee

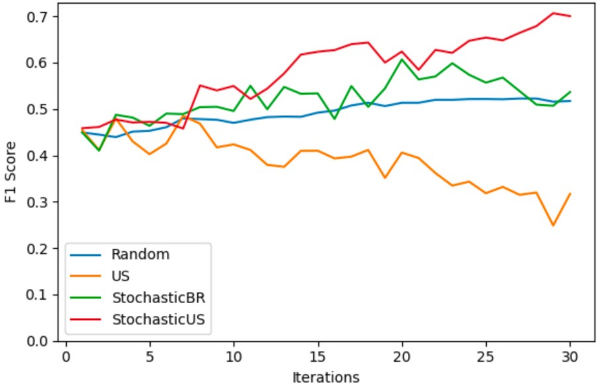
If the trainer and learner follow (FP, Best Response) and (FP, Stochastic Best Response), respectively, the empirical behavior of the game converges to an equilibrium.

Empirical study: setup

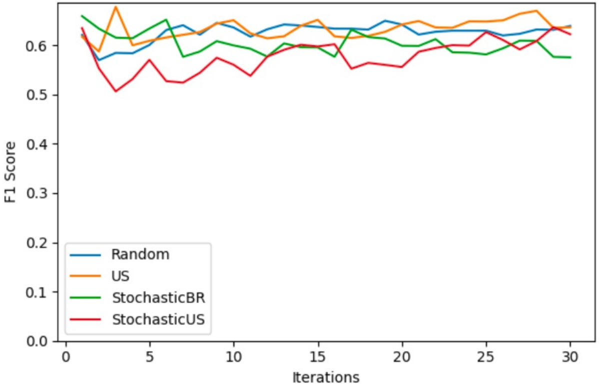
- **Task**
  - Learn approximate functional dependencies
  - Labeling the dirty tuples
- **Datasets**
  - Same as the user study
  - ***Hospital***, ***Tax*** datasets
- **Trainer based on results from user study**
  - Simulated using stochastic model
  - Uses Bayesian model as a backbone
- Study conducted with introduction of different degree of violations

Result: Accuracy

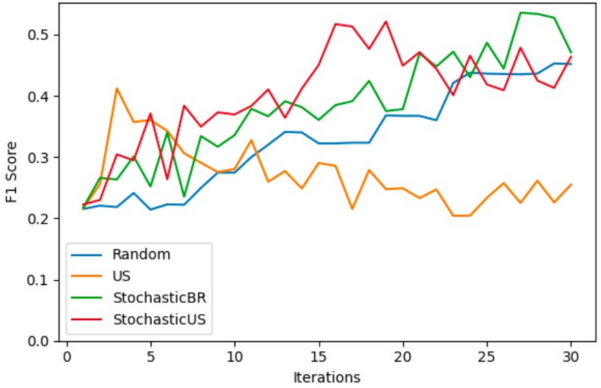
Average F1 Score of the labeling of Learner model



(a) OMDB



(b) Hospital



(c) Tax

The accuracy improvement over iteration is better in proposed methods.

Conclusion and Future Work

- Current methods learn inaccurate models when annotators learn
- Developed methods that adapts to human learning
- Show that our method converges quickly to accurate models

Future Works:

- other modes of interactions/ types of data
- more complex learning schemes: recursive reasoning
- data systems that adapt to non-stationary/ learning workload