Statistical Approaches for Entity Resolution under Uncertainty

Neil Marchant

School of Computing and Information Systems, University of Melbourne

PhD advisors: Ben Rubinstein and Rebecca Steorts

Savage Award (Applications), ISBA, 29 June 2022

Talk overview

Background on entity resolution (ER)

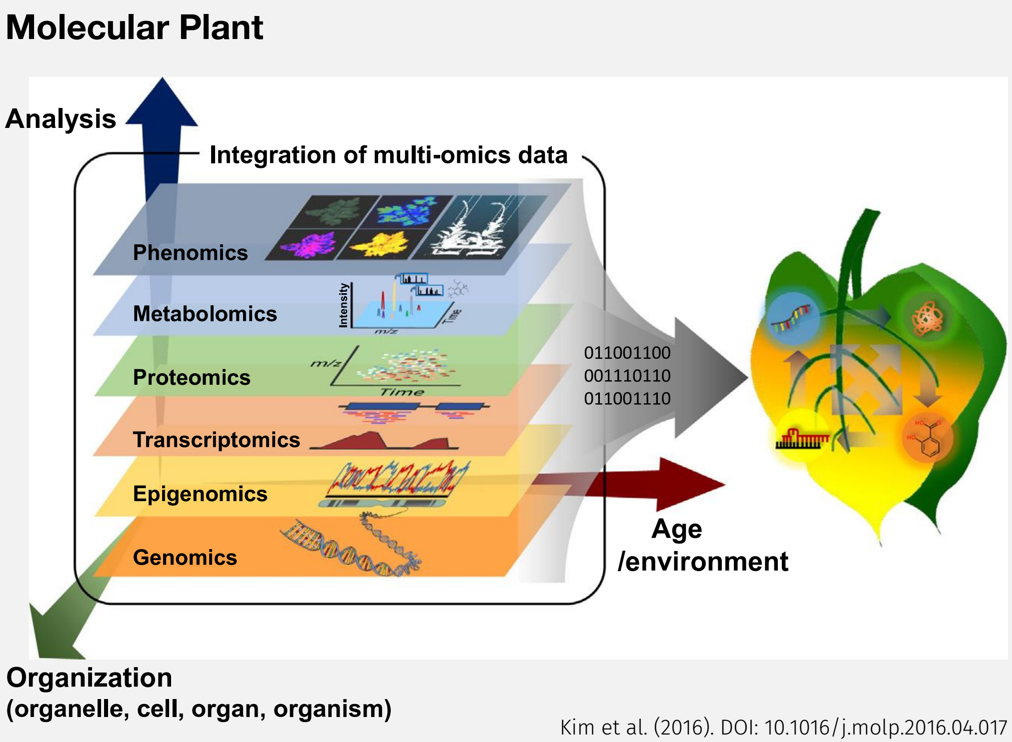
Summary of key contributions

1. Scalable unsupervised Bayesian ER
2. A refined model for unsupervised Bayesian ER
3. A theoretical framework for evaluation of ER

Conclusion

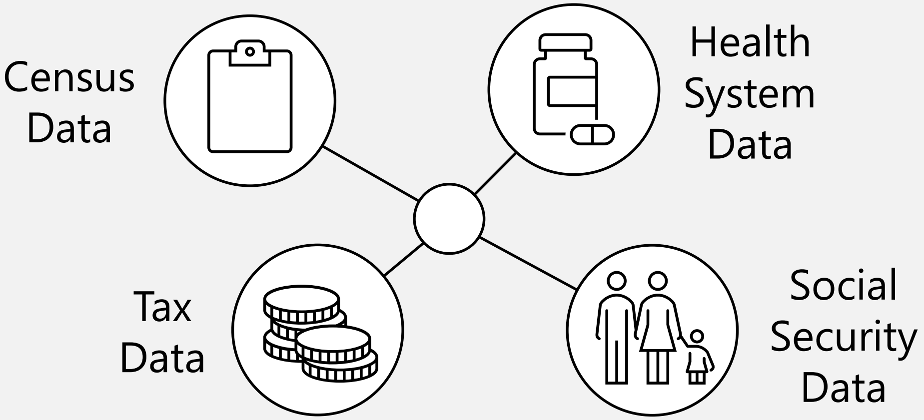
Data integration: a ubiquitous problem

Multi-omics data for biological research

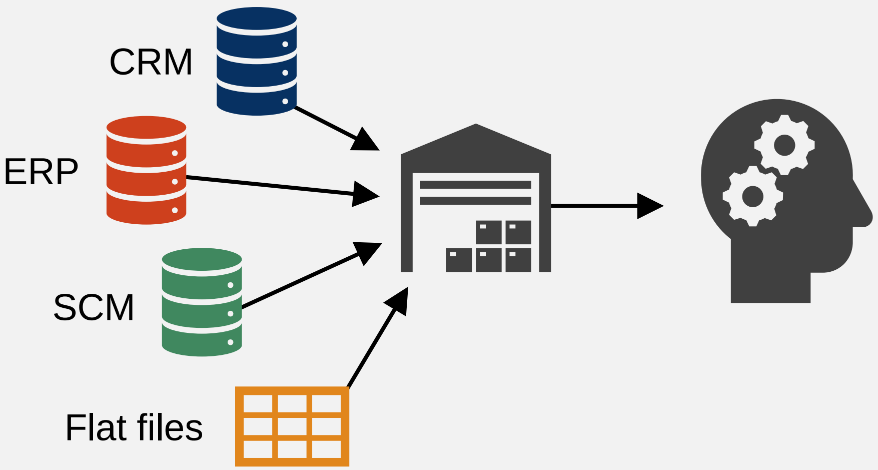


Kim et al. (2016). DOI: 10.1016/j.molp.2016.04.017

Data sharing across government

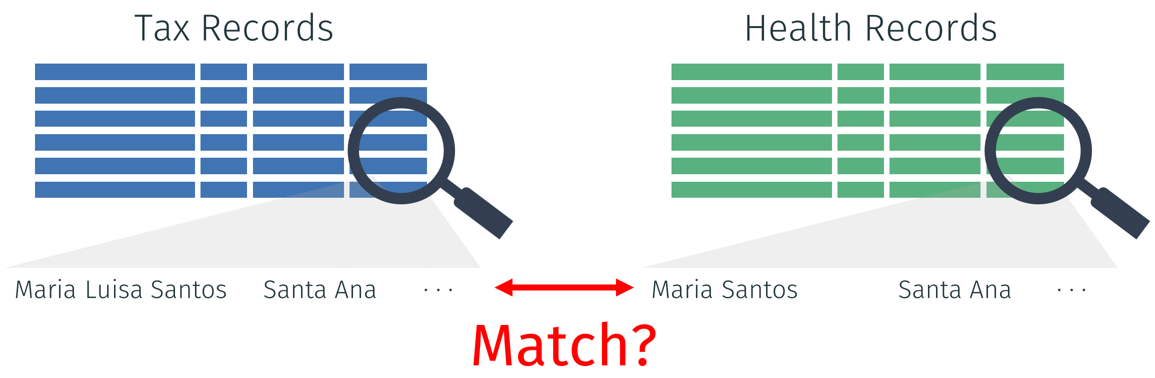


Enterprise information integration



Entity resolution: a key step in data integration

*Entity resolution (ER) links records that **relate to the same entity***



- Also known as: record linkage, data matching, merge/purge, deduplication
- Statistical approach due to Fellegi & Sunter (1969) still widely used today
- Other methods include: supervised machine learning, probabilistic graphical models, distance-based clustering, human-in-the-loop methods, rule-based methods etc.

Pain points for entity resolution

Costly manual labelling

Vast amounts of manually-labelled data are typically required for supervised learning and evaluation.

Scalability/computational efficiency

Approximations are required to avoid quadratic scaling. Need to ensure impact on accuracy is minimal.

Limited treatment of uncertainty

Given inherent uncertainties, it’s important to output predictions with confidence regions.

Unreliable evaluation

Standard evaluation methods return imprecise estimates of performance.

Pain points for entity resolution

Costly manual labelling

Scalability/computational efficiency

Limited treatment of uncertainty

Unreliable evaluation

Thesis contributions

1. Scalable unsupervised Bayesian ER
2. Modelling improvements for unsupervised Bayesian ER
3. A theoretical framework for label-efficient evaluation

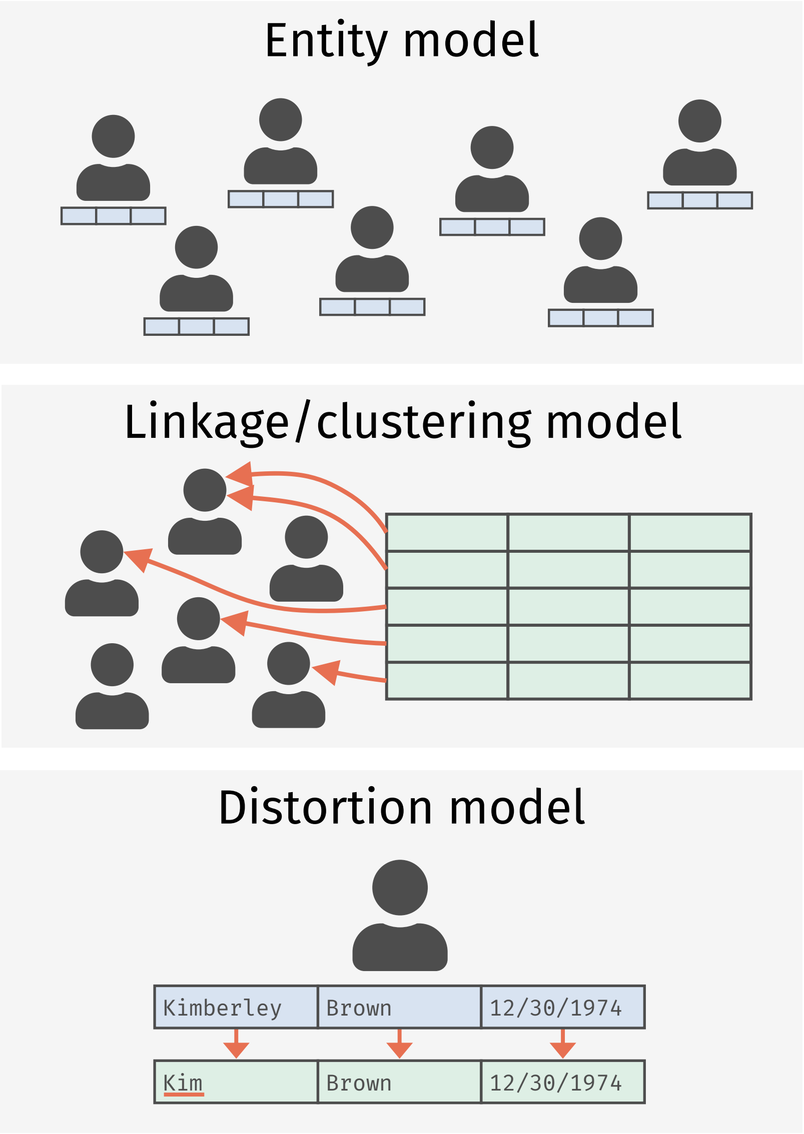
1. Scalable unsupervised Bayesian ER

N. G. Marchant, A. Kaplan, D. N. Elazar, B. I. P. Rubinstein and R. C. Steorts (2021) “d-blink: Distributed End-to-End Bayesian Entity Resolution,” J. Comp. Graph. Stat., 30:2, 406-421.

U.S. Census Bureau DRB No: CBDRB-FY20-309

`blink` ER model

- A Bayesian model proposed by Steorts (2015)
- Key features:
  - Assumes records are generated by sampling from a population of latent entities
  - Record attributes may be distorted (e.g. typos) when copied from the entity
  - Supports multiple structured data sources
  - Predicted coreference relation is transitive (no conflicts)
- Problem: difficulty scaling beyond ~1000 records



Can we scale `blink` to 1 million records?

Current state of affairs:

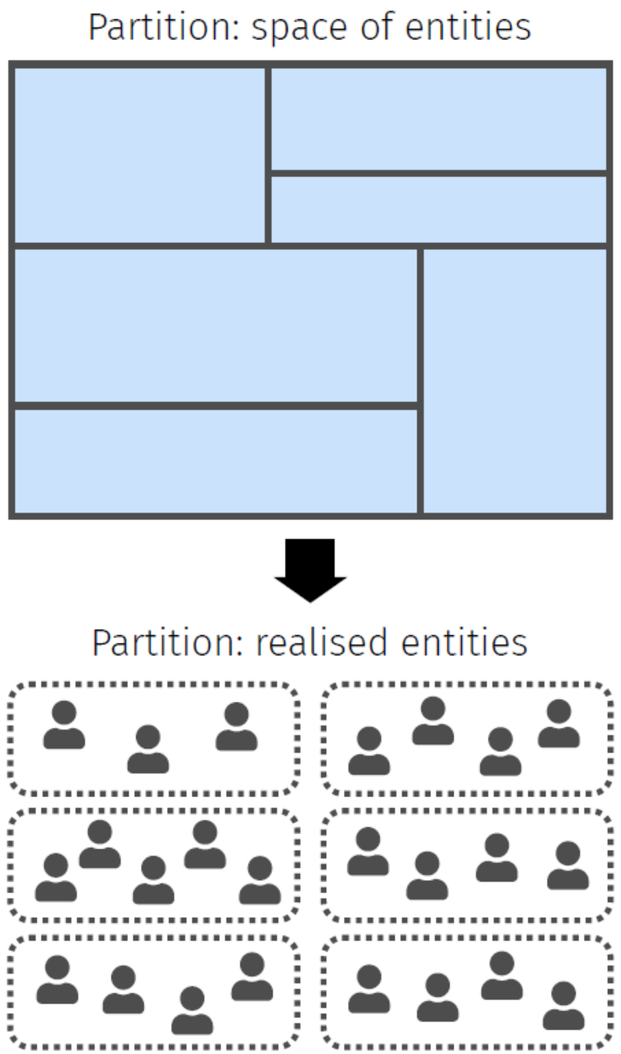
- Gibbs sampling is used for inference. Need to run for many iterations (e.g. 100,000).
- Gibbs update for the entity assignments scales roughly quadratically in the # records

We propose `d-blink`:

- Computational speed-ups:
  - Incorporate probabilistic blocking
  - Sub-quadratic entity assignment update via indexing
  - Perturbation sampling for entity attribute update
  - Distributed/parallel inference
- Partially-collapsed Gibbs sampling for improved statistical efficiency
- Also add support for:
  - missing values
  - arbitrary attribute similarity functions

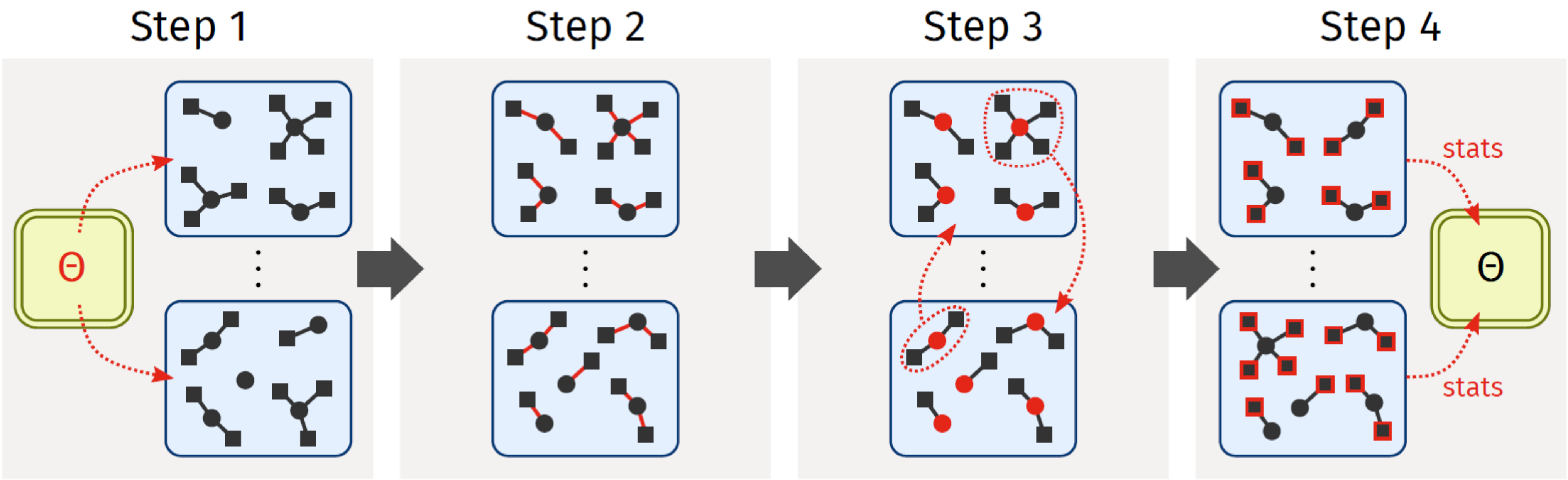
Probabilistic blocking

- Partition the space of entities into auxiliary blocks using a user-specified blocking function
- By careful design, can ensure the posterior is unchanged when the auxiliary blocks are marginalized out
- Asymptotically, inferred parameters are the same as for the original `blink` model
- Also, enables distributed/parallel inference at the block-level



Distributed inference

Records/entities are conditionally independent across blocks



Update distortion probs on the manager and broadcast to the workers

Update entity assignments on the workers. Records may only be assigned to entities within their block.

Update entity attributes and block assignments on the workers. Move the entities and records to their newly-assigned blocks.

Update distortion indicators, then calculate summary stats on the worked. Broadcast to the manager.

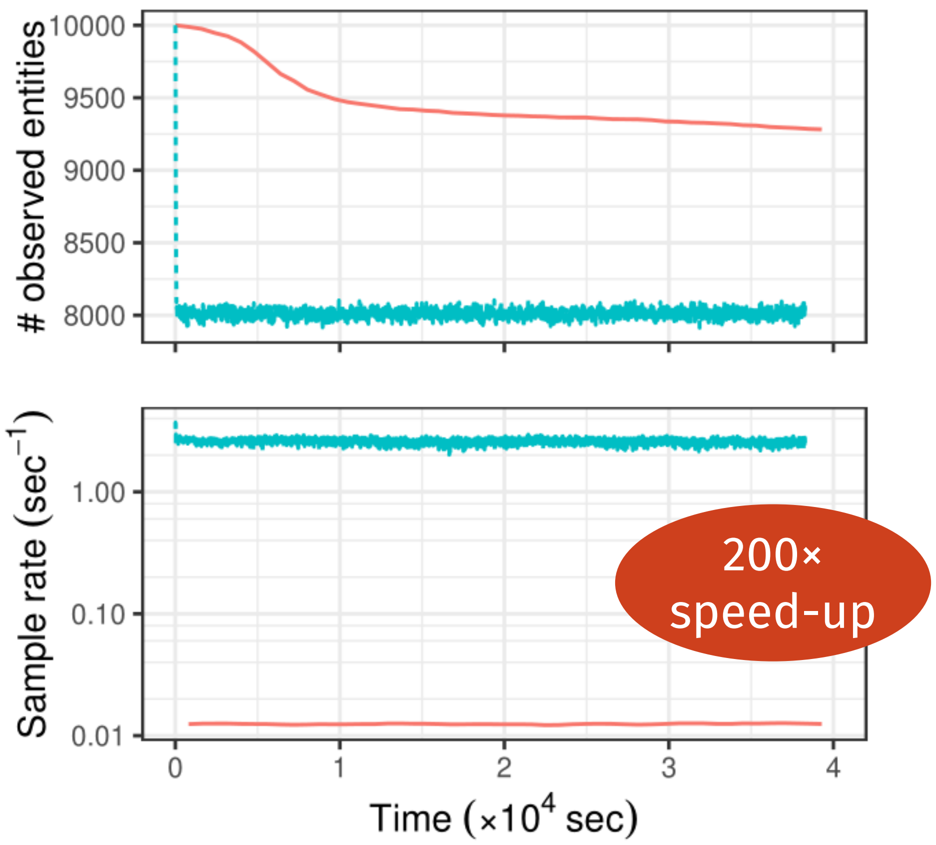
Empirical study

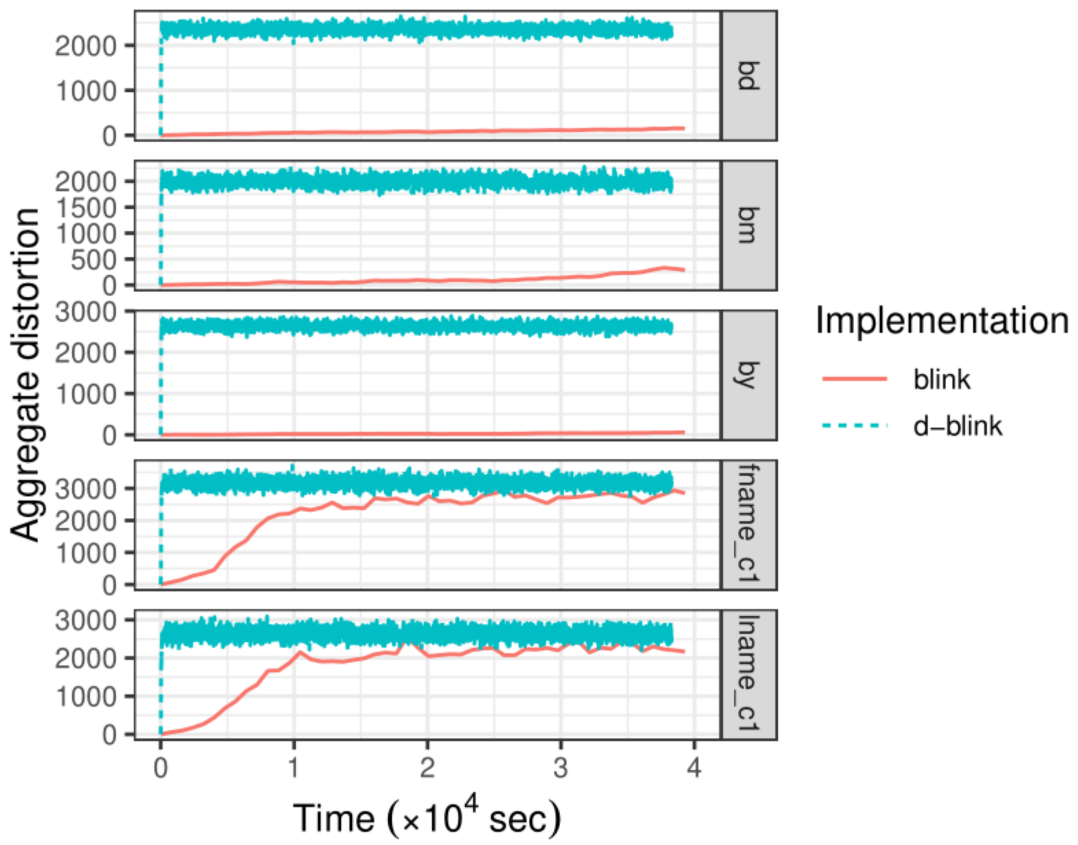
- Open-source implementation in Apache Spark
- Tested on local server + Amazon EMR
- Five synthetic/publicly-available data sets
- Comparison with 3 baseline methods
- Recent application to population enumeration using U.S. 2010 Decennial Census + admin records from the U.S. Social Security Administration

Data set	Description	Num. records	Num. sources	Num. entities
ABSEmployee	Synthetic employee data	600,000	3	400,000
NCVR	Voter records	448,134	2	296,433
NLTCS	Longitudinal health survey	57,077	3	34,945
SHIW0810	Longitudinal survey	39,743	2	28,584
RLdata10000	Synthetic personal data	10,000	1	9,000

Results

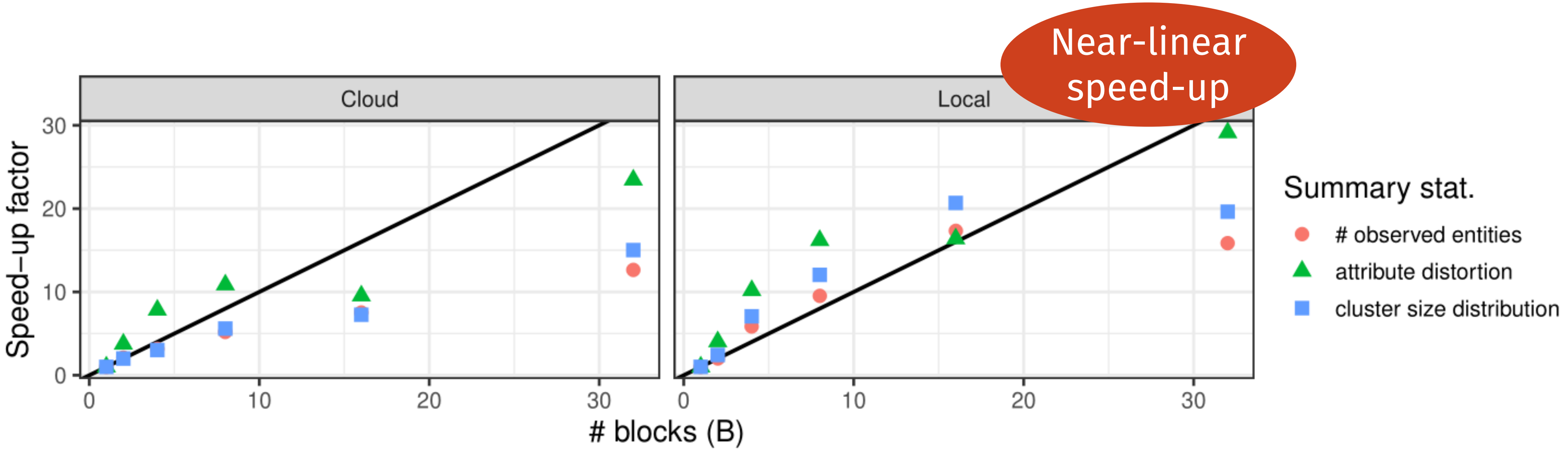
Convergence and efficiency of d-blink (no blocking) versus blink





Results

Efficiency gains due to blocking



Summary

- Achieved a significant speed-up, e.g. by a factor of 300×
- All of our ideas contributed to the speed-up: blocking, partially-collapsed Gibbs sampling, fast algorithms for Gibbs updates, parallelisation
- `d-blink` is promising for ER of moderately-sized data (~1 million records)
- Future work:
  - Variational Bayes as an alternative to MCMC
  - Applying to other models

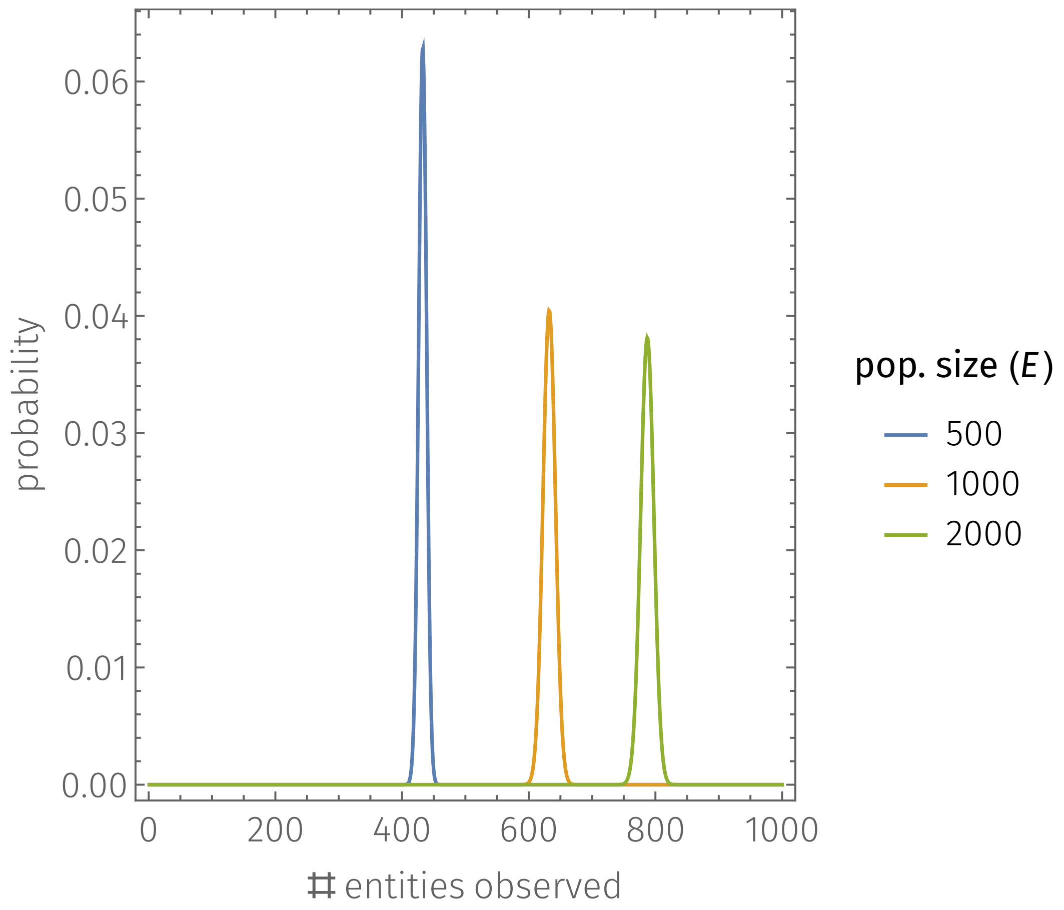
2. A refined model for unsupervised Bayesian ER

N. G. Marchant, B. I. P. Rubinstein and R. C. Steorts (2021) “Bayesian Graphical Entity Resolution using Exchangeable Random Partition Priors,” Under review

Can we improve the `blink` ER model?

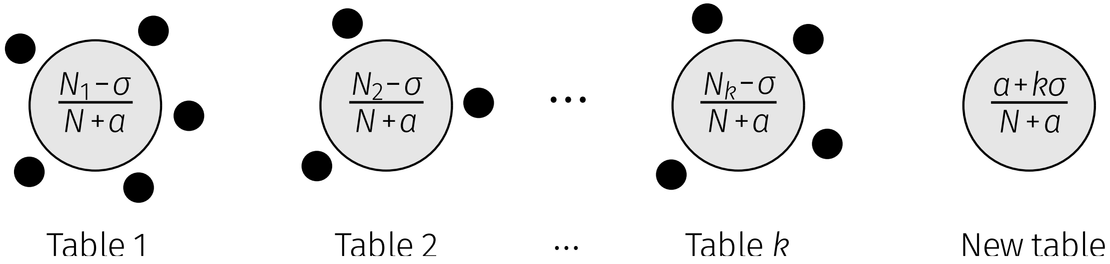
Criticisms:

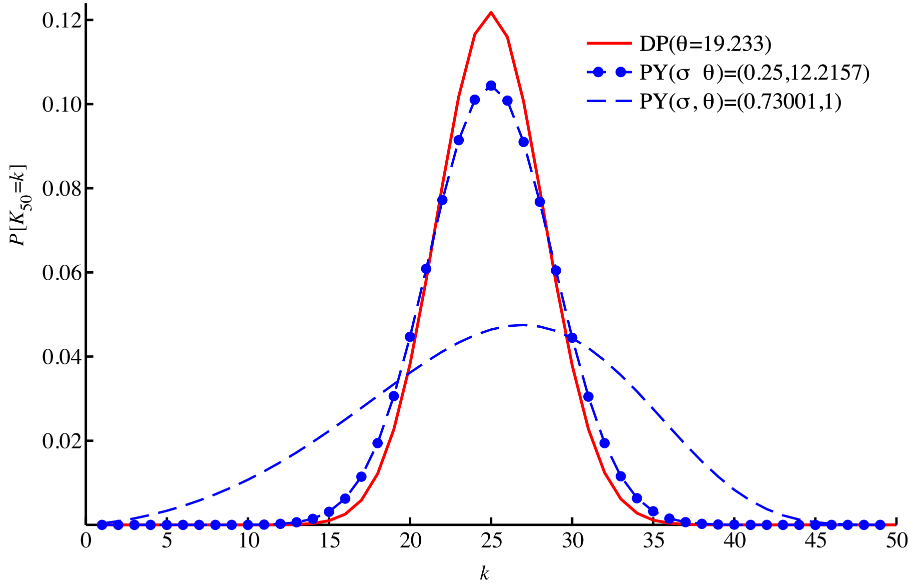
- Several parameters are set empirically
- Informative priors
- Sensitivity to hyperparameters



Flexible priors on the linkage structure

- Assuming (1) *exchangeability* and (2) *Kolmogorov consistency*, the family of *Ewens-Pitman random partitions* is the most general class of priors
- Parametrised by σ, α. Differing asymptotic regimes:
    - GenCoupon ($\sigma < 0$): num. entities is finite $-{}^{\alpha}/_{\sigma}$ a.s.
    - Ewens ($\sigma = 0$): num. entities is $\alpha \log N$ a.s.
    - Pitman-Yor ($0 < \sigma < 1$): num. entities is $S_\sigma N^\sigma$ a.s.
- Hyperpriors improve flexibility

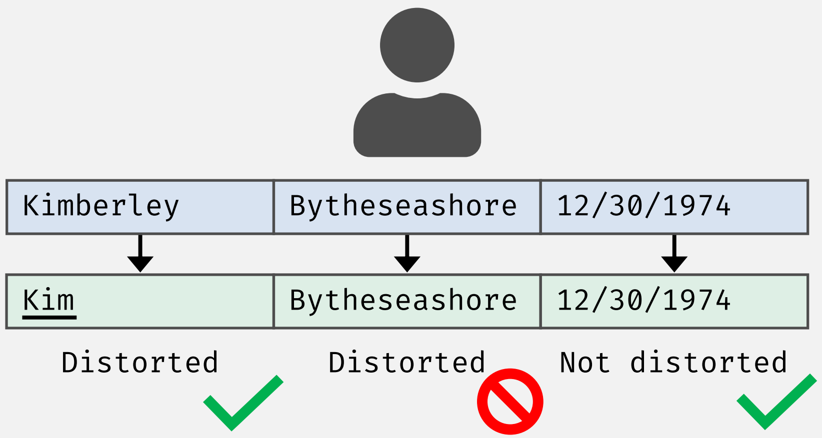




Other improvements

Corrected distortion model

- Make the probability of distortion depend on the entity attribute
- If a record attribute is "distorted" it *must differ* from the entity attribute



Deepen the model

- Place Dirichlet process priors on:
  - the entity attribute distribution (generates an entity attribute)
  - the distortion distribution (generates a distorted record value conditional on the entity attribute value)
- These were set empirically in `blink`

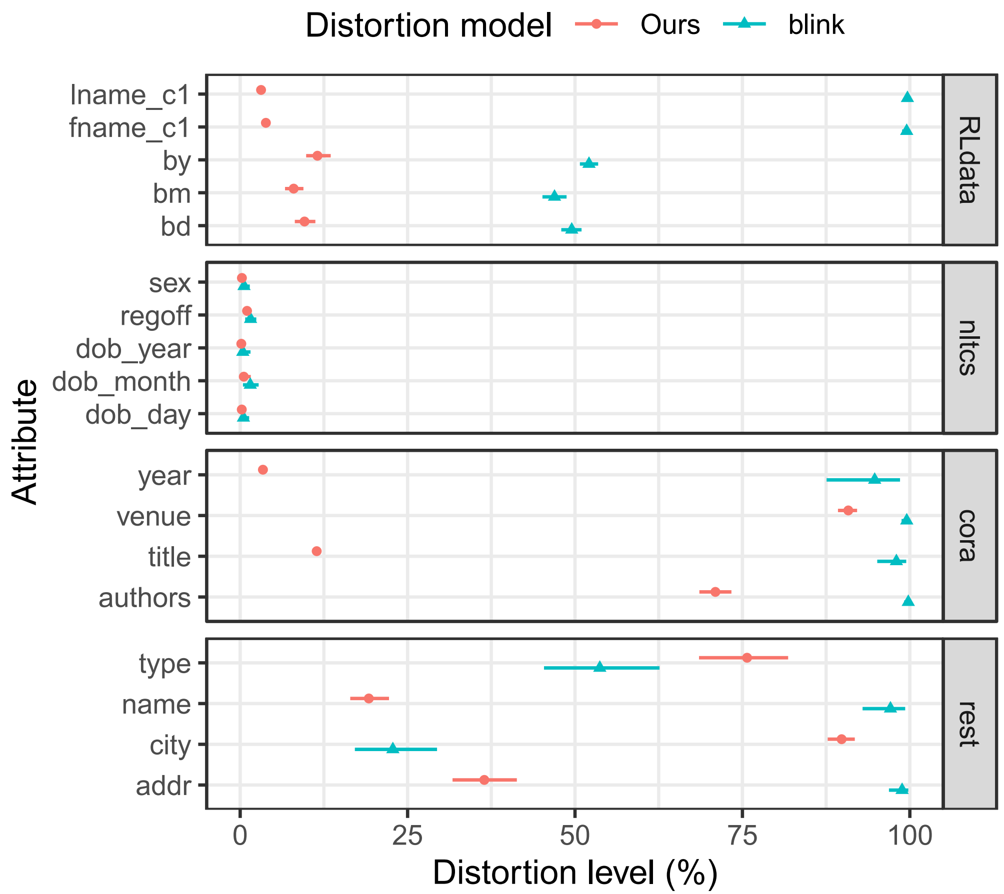
Empirical study
Effect of flexible Ewens-Pitman priors

- Compared the Ewens-Pitman priors in three regimes (PY, Ewens, GenCoupon) against `blink`'s Coupon prior
- Find that `blink`'s Coupon prior performs worse, especially when misspecified
- PY, Ewens, GenCoupon perform similarly, but only if vague hyperpriors are used

		Performance measure		
Data set	EP regime	Precision	Recall	F1 score
RLdata	PY	0.896 (0.879,0.917)	0.961 (0.952,0.972)	0.928 (0.918,0.939)
	Ewens	0.870 (0.853,0.893)	0.970 (0.961,0.978)	0.917 (0.908,0.931)
	GenCoupon	0.903 (0.886,0.920)	0.966 (0.955,0.975)	0.933 (0.923,0.941)
	Coupon	0.402 (0.396,0.410)	0.987 (0.982,0.993)	0.572 (0.565,0.580)
nltcs	PY	0.921 (0.908,0.933)	0.924 (0.915,0.934)	0.923 (0.915,0.930)
	Ewens	0.921 (0.910,0.932)	0.925 (0.915,0.934)	0.923 (0.916,0.930)
	GenCoupon	0.902 (0.879,0.918)	0.935 (0.926,0.944)	0.918 (0.906,0.927)
	Coupon	0.919 (0.908,0.930)	0.926 (0.916,0.935)	0.923 (0.915,0.930)
cora	PY	0.971 (0.963,0.979)	0.671 (0.647,0.696)	0.794 (0.776,0.813)
	Ewens	0.974 (0.965,0.981)	0.673 (0.645,0.697)	0.796 (0.775,0.813)
	GenCoupon	0.973 (0.965,0.981)	0.657 (0.632,0.683)	0.784 (0.766,0.804)
	Coupon	0.978 (0.971,0.986)	0.173 (0.164,0.181)	0.294 (0.281,0.306)
rest	PY	0.770 (0.735,0.824)	0.812 (0.759,0.884)	0.795 (0.755,0.828)
	Ewens	0.770 (0.711,0.823)	0.830 (0.781,0.875)	0.798 (0.760,0.838)
	GenCoupon	0.794 (0.742,0.850)	0.821 (0.777,0.875)	0.807 (0.773,0.849)
	Coupon	0.637 (0.602,0.674)	0.911 (0.893,0.938)	0.750 (0.722,0.781)

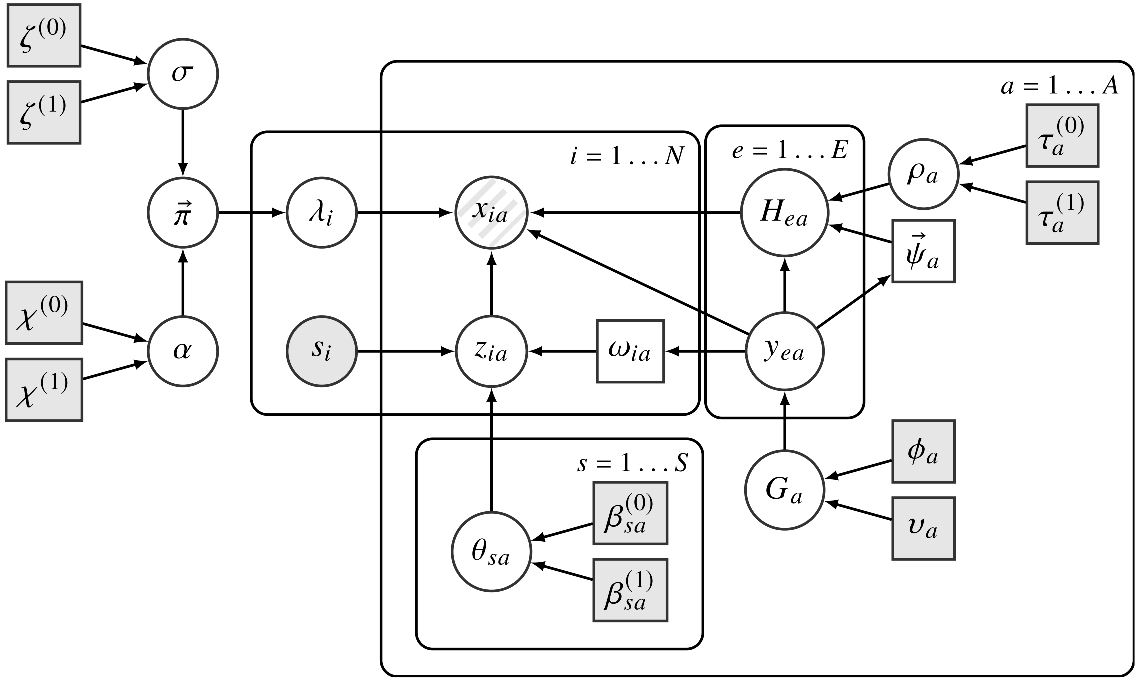
Empirical study
Effect of the distortion model

- Inferred level of distortion is now consistent with expectations
- ER accuracy also improved: less susceptible to over-linkage



Summary

- Proposed modeling improvements to `blink`
- New model is less sensitive, achieves more accurate ER results
- Future work:
    - Scaling this model like we did for `blink`
    - Semi-supervised settings



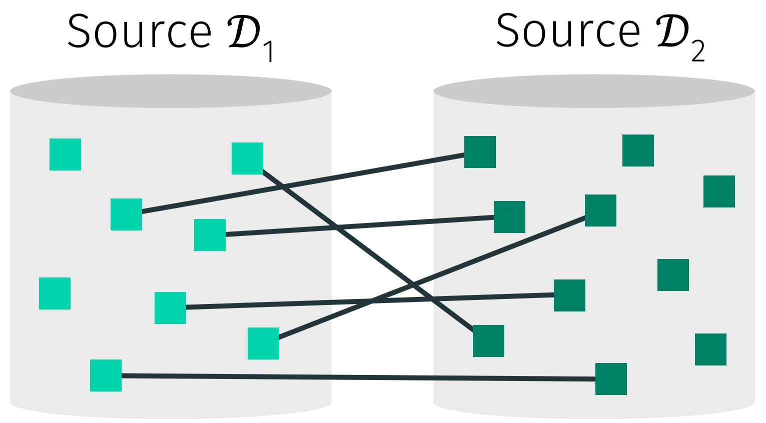
3. A theoretical framework for label-efficient evaluation

N. G. Marchant and B. I. P. Rubinstein (2020) *"Needle in a Haystack: Label-Efficient Evaluation under Extreme Class Imbalance,"* Proceedings of SIGKDD

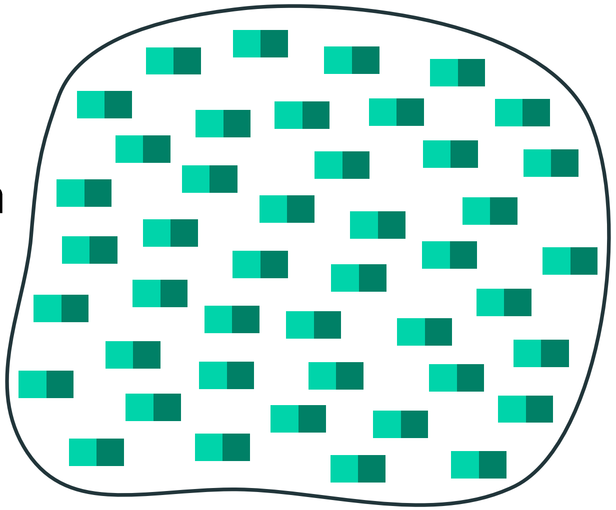
N. G. Marchant and B. I. P. Rubinstein (2017) *"In Search of an Entity Resolution OASIS: Optimal Asymptotic Sequential Importance Sampling,"* Proceedings of the VLDB Endowment

Why is ER evaluation challenging?

- Given an ER system to evaluate that predicts whether pairs of records are matches or non-matches (refer to the same entity or not)
- Standard evaluation approach:
  - Sample pairs of records uniformly at random
  - Ask humans to label as match/non-match
  - Compute performance measures on the sample



Pool of record pairs from $\mathcal{D}_1 \times \mathcal{D}_2$



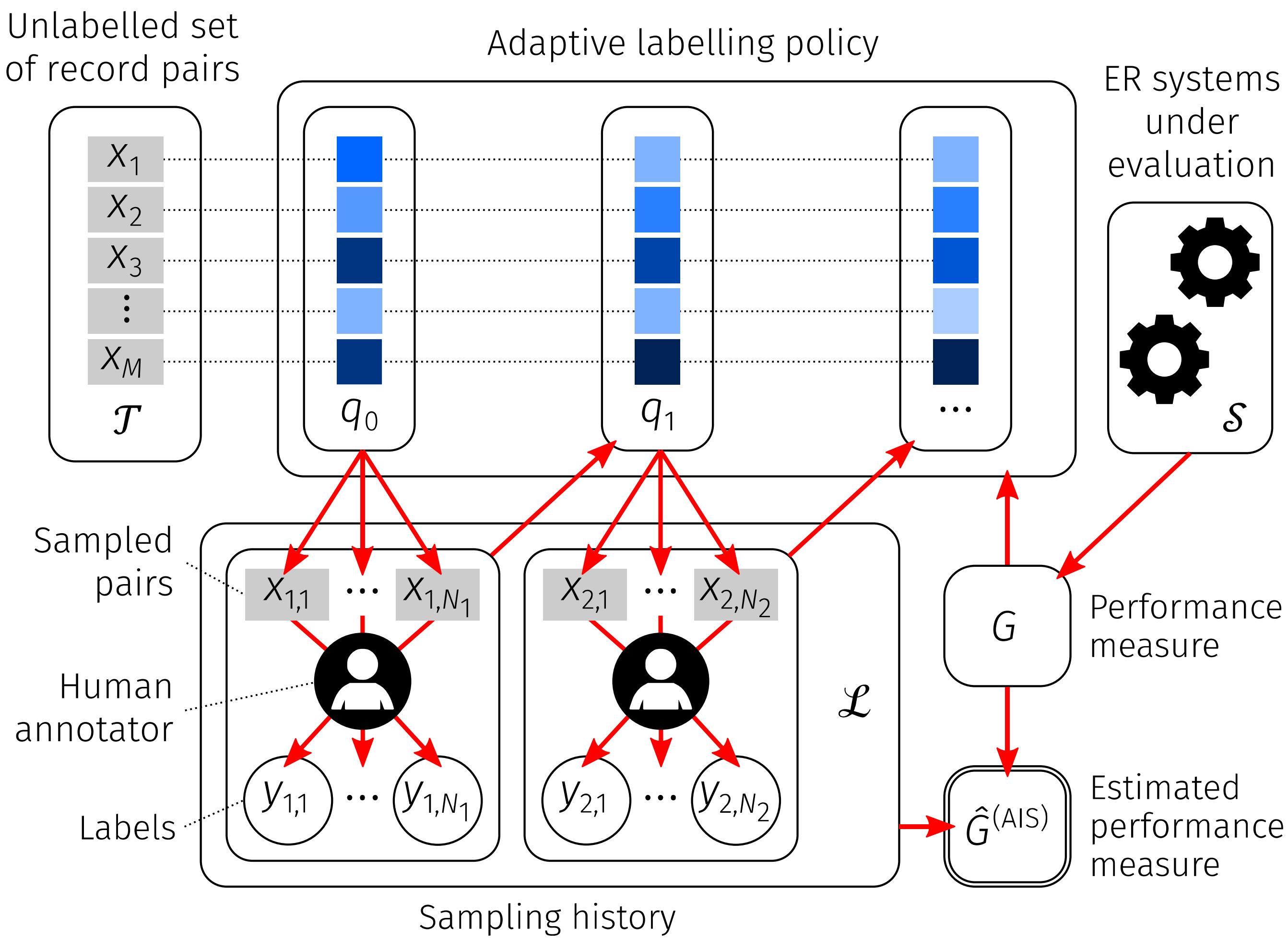
Imbalance problem:
For every match, there are roughly $N = \max(|\mathcal{D}_1|, |\mathcal{D}_2|)$ non-matches ⇒ need a huge labelled sample to get a precise performance estimate.

A snapshot of related work

- Variance reduction methods for evaluation:
  - Static importance sampling (Sawade et al., 2010; Schnabel et al., 2016)
  - Stratified sampling (Druck & McCallum, 2011)
  - Online stratified sampling (Bennett & Carvalho, 2010)
- These haven't been applied to ER
- Several limitations:
  - Lack of support for a broad range of performance measures
  - Lack of support for evaluating multiple systems/measures in parallel
  - Lack of support for interactive (adaptive) evaluation
  - Limited efficiency (stratified sampling)

An AIS-based evaluation framework

- We propose a framework based on *adaptive importance sampling* (AIS)
- Labels are collected in rounds by querying a human annotator
- The labelling policy (which selects items to label) is adapted based on labels collected in previous rounds
- Performance estimates are bias-corrected (can prove consistency + CLT)



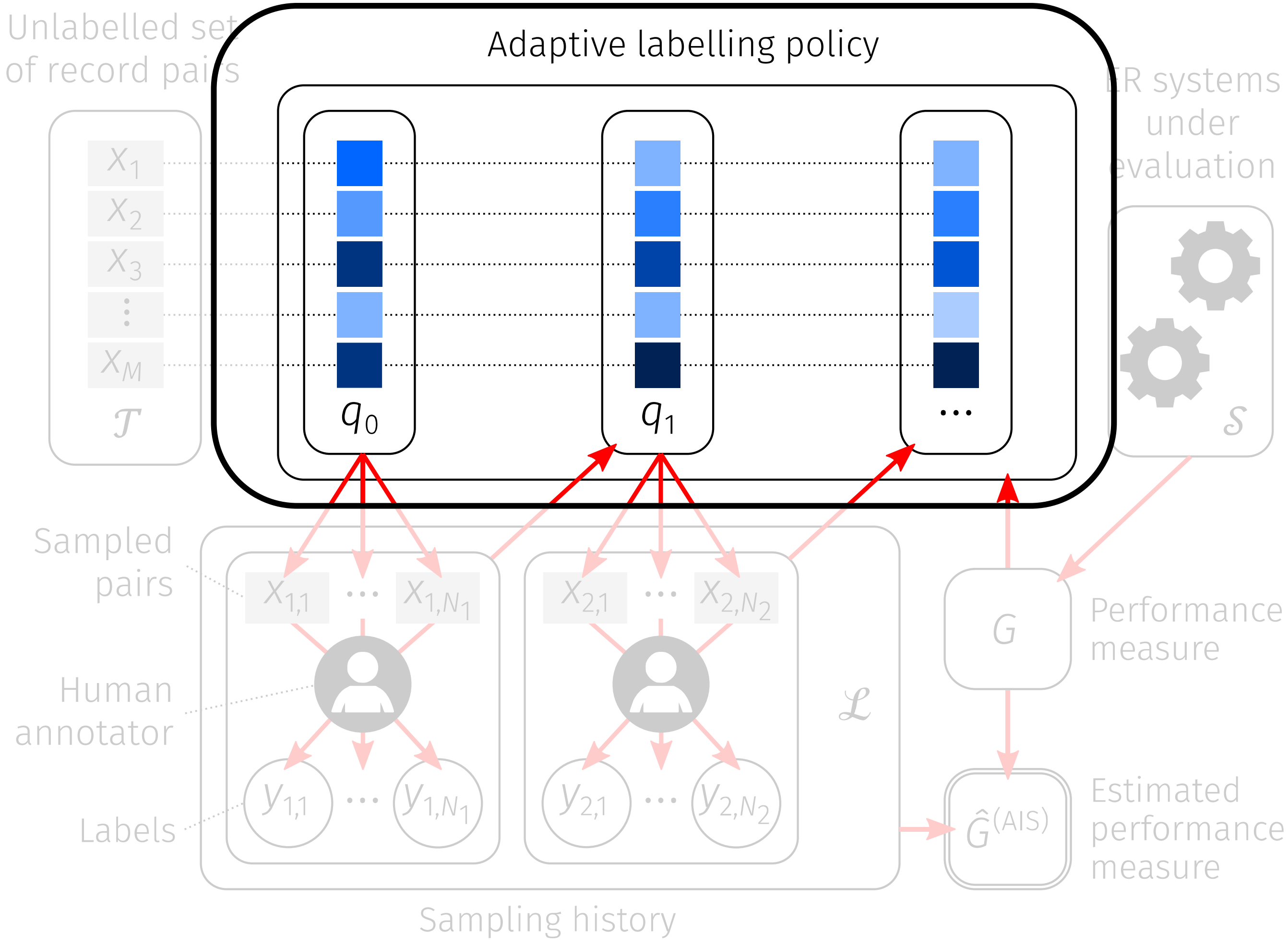
Which performance measures are covered?

We consider a family of *generalised measures* which corresponds to transformations of vector-valued risk functionals.

Measure	$\ell(x, y)^\intercal$	$g(R)$
Accuracy	$\mathbb{I}[y \neq f(x)]$	$1 - R$
Balanced accuracy	$[yf(x), y, f(x)]$	$\frac{R_1 + R_2(1-R_2-R_3)}{2R_2(1-R_2)}$
Precision	$[yf(x), f(x)]$	$\frac{R_1}{R_2}$
Recall	$[yf(x), y]$	$\frac{R_1}{R_2}$
F_β score	$\left[yf(x), \frac{\beta^2 y + f(x)}{1+\beta^2}\right]$	$\frac{R_1}{R_2}$
Matthews correlation coefficient	$[yf(x), y, f(x)]$	$\frac{R_1 - R_2R_3}{\sqrt{R_2R_3(1-R_2)(1-R_3)}}$
Fowlkes-Mallows index	$[yf(x), y, f(x)]$	$\frac{R_1}{\sqrt{R_2R_3}}$
Brier score	$2(\hat{p}_1(x) - y)^2$	R

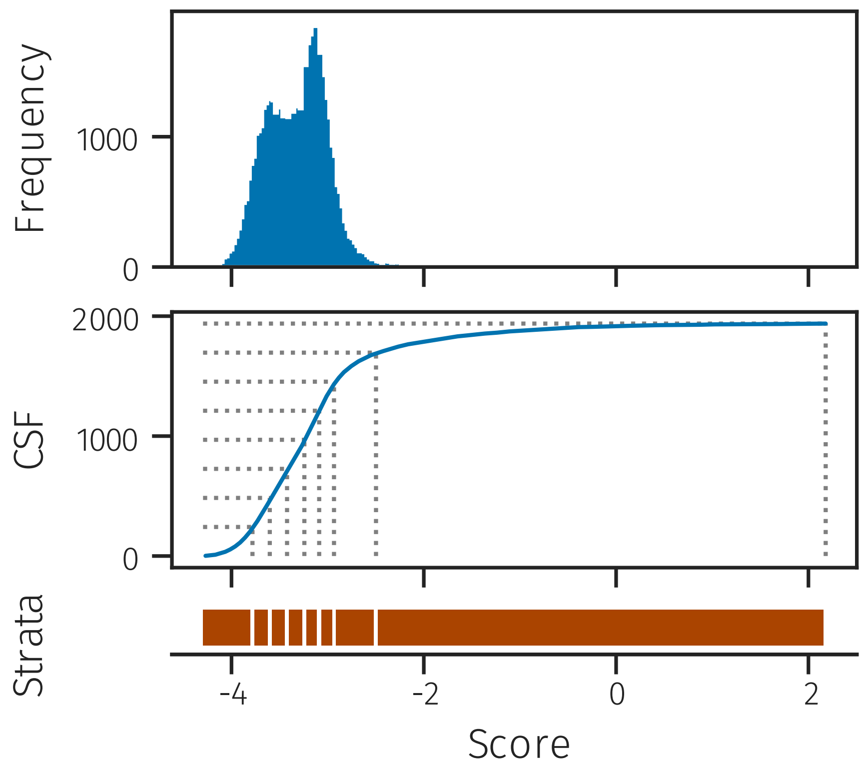
How to adapt the labelling policy?

- We'd like to target the asymptotically-optimal policy $q^\star(x)$, but it depends on the unknown human response $p(y|x)$
- Solution: plug-in online estimates of $p(y|x)$ using a Bayesian model.
- Technical point: need to ensure estimate of $q^\star(x)$ has the same support as $q^\star(x)$.



Bayesian model for the human response

- Stratify the set of pairs $\mathcal{T} = \cup_{k=1}^{K} \mathcal{T}_k$ using scores from the system(s) and assume $p(y|x) \approx p(y|x \in \mathcal{T}_k)$
- **Model 1:** assume each stratum is an independent source of labels (independent Dirichlet-Categorical models)
- **Model 2:** assume strata are hierarchically dependent (Dirichlet-tree model; two variants for stochastic/deterministic oracles)



Empirical study

- Implemented as open-source Python package called `activeeval`
- 4 ER data sets (highly imbalanced) + 3 non-ER data sets
- 5 evaluation methods

Evaluation methods

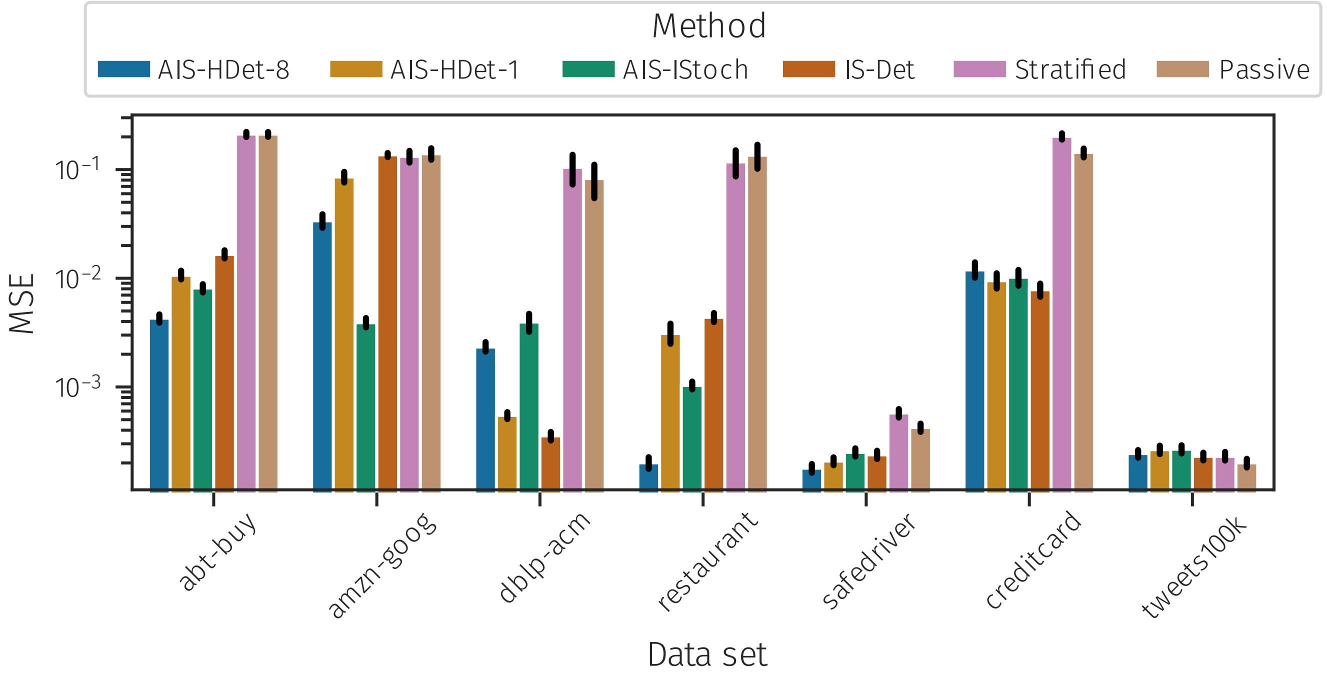
Name	Adaptive	Estimator for $q^\star(x)$	Estimator for $p(y\|x)$
AIS-HDet	Yes	Threshold deterministic	Hierarchical deterministic
AIS-IStoch	Yes	Stratified	Independent stratified
IS-Det	No	Threshold deterministic	Scores from system
Stratified	No	-	
Passive	No	-	

Data sets

Data set	Size	Imb. ratio	Classifier	True F1
abt-buy	53,753	1075	SVM	0.595
amzn-goog	676,267	3381	SVM	0.282
dblp-acm	53,946	2697	SVM	0.947
restaurant	149,747	3328	SVM	0.899
safedriver	178,564	26.56	XGB	0.100
creditcard	85,443	580.2	LR	0.728
tweets100k	20,000	0.990	SVM	0.770

Selected results

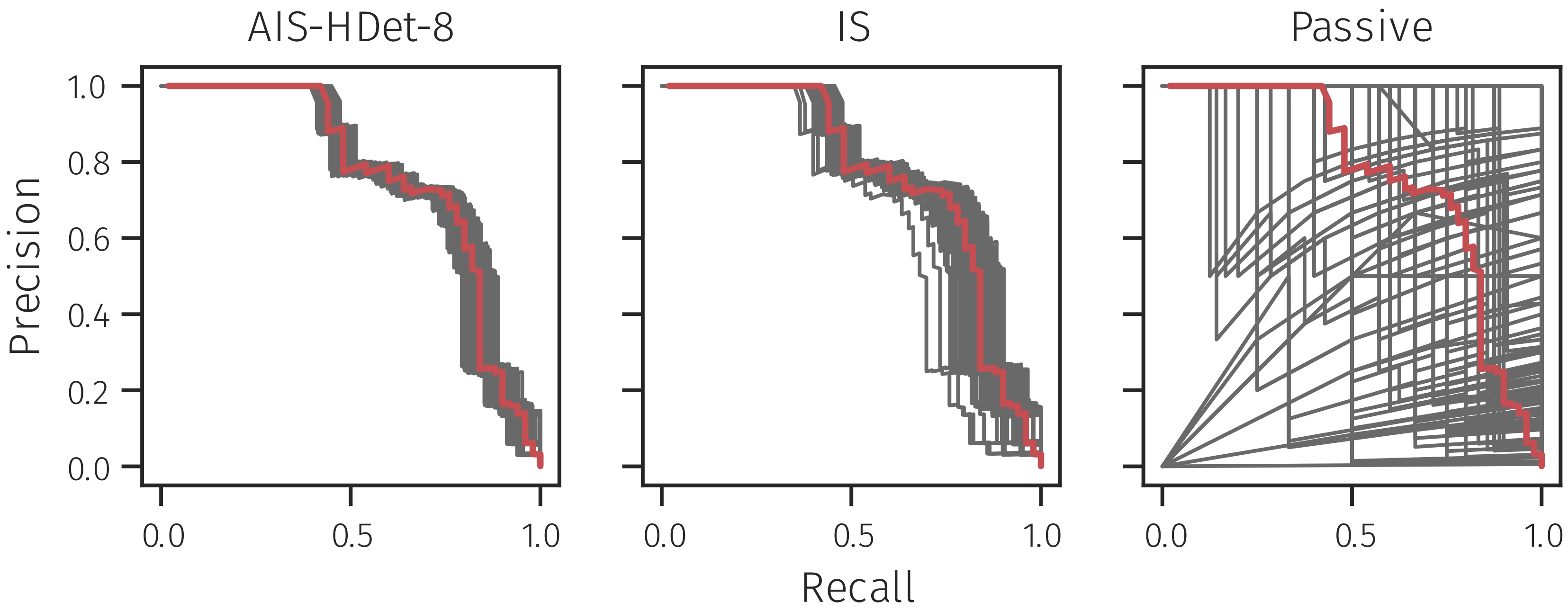
- Passive/stratified essentially unusable under extreme imbalance
- Adaptivity generally helps when estimates of $p(y|x)$ from the system are poor



MSE of estimated F1-score (over 1000 repeats) assuming a label budget of 1000

Selected results

- We can also estimate vector-valued measures using our framework
- Again, passive sampling is essentially unusable



A sample of 100 estimated precision-recall curves for abt-buy assuming a label budget of 5000. The red curve is the unknown true curve.

Summary

- Developed a statistically-grounded framework for evaluation with asymptotic guarantees
- Adaptive policy leverages a Bayesian model for the human response
- Increased statistical precision means
  - practitioners can be more confident in evaluation results
  - fewer labels are required

Conclusion

Summary of key contributions

Statistical methods for performing and evaluating entity resolution

1. Scalable and efficient inference for Bayesian ER
2. Modelling improvements for Bayesian ER: reduced sensitivity and improved accuracy
3. A statistical framework for evaluation with asymptotic guarantees

unsupervised, proper handling of uncertainty

reduced cost of manual labelling, improved reliability of evaluation

Open-source software published at github.com/ngmarchant and github.com/cleanzr

Questions?

Please contact me (Neil Marchant)

Email	nmarchant@unimelb.edu.au
Web	www.ngmarchant.net
GitHub	@ngmarchant