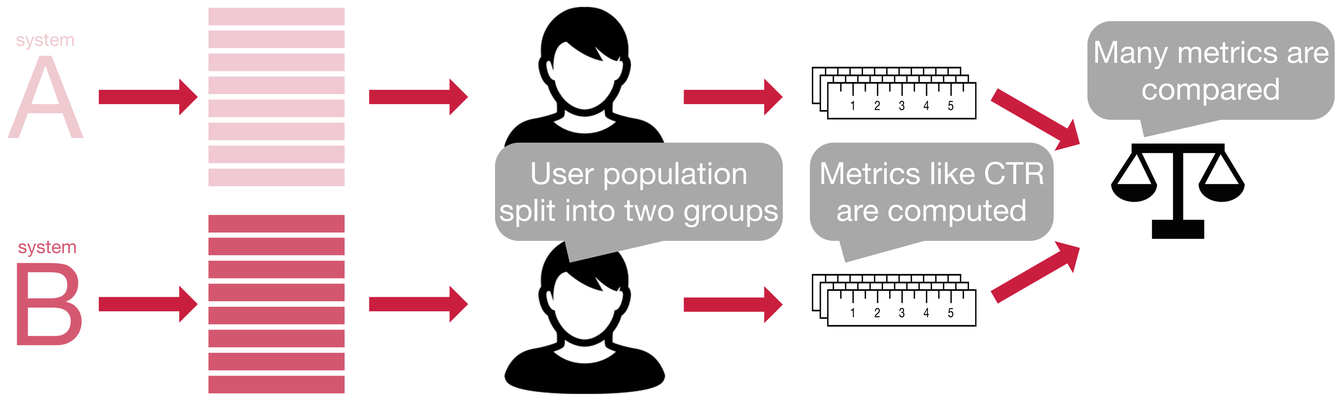
Probabilistic Multileave for Online Retrieval Evaluation

Anne Schuth, Robert-Jan Bruintjes, Fritjof Büttner, Joost van Doorn,
Carla Groenland, Harrie Oosterhuis, Cong-Nguyen Tran, Bas Veeling,
Jos van der Velde, Roger Wechsler, David Woudenberg, Maarten de Rijke

anne.schuth@uva.nl firstname.lastname@student.uva.nl derijke@uva.nl

University of Amsterdam

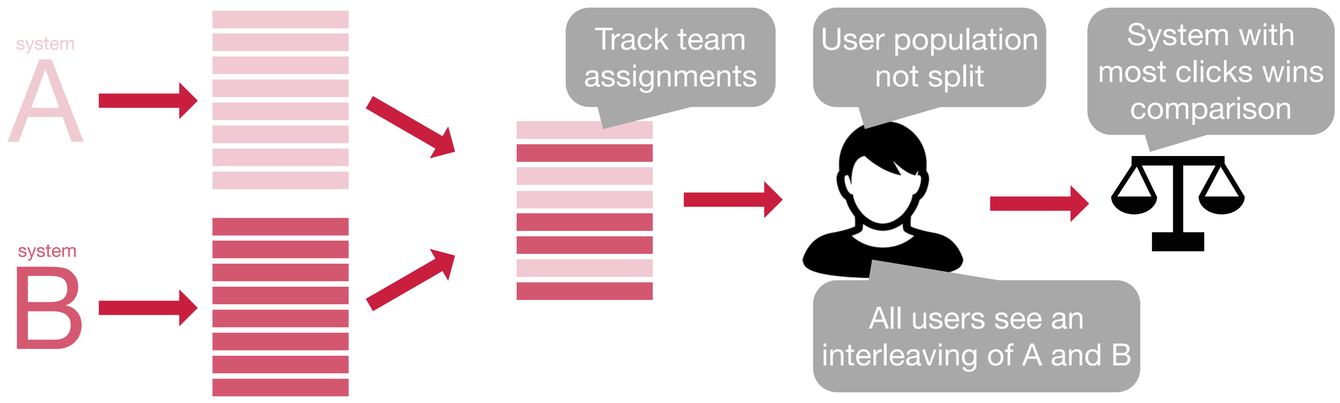
A/B Testing



+ **Any metric** can be measured using A/B testing
- **Not very sensitive,** *between subject design*.
  Noise coming from differences between users and their queries.

[1] R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne. Controlled experiments on the web: survey and practical guide. In Data Mining and Knowledge Discovery, 2009.

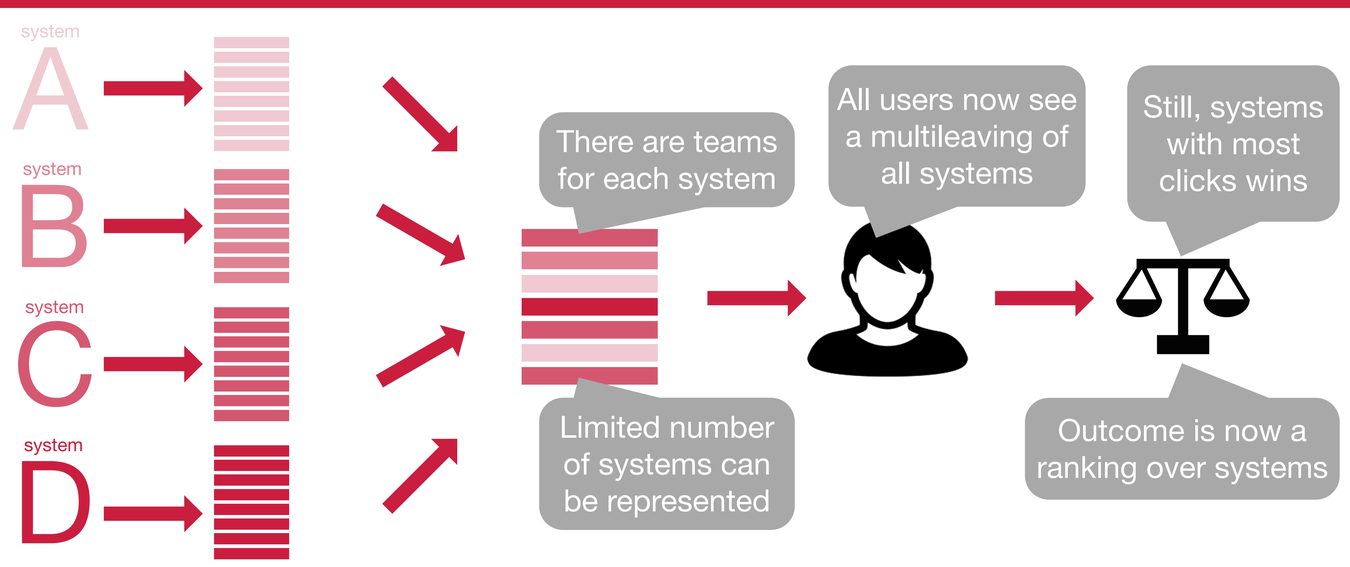
Interleaved Comparisons



+ **Sensitive**, w*ithin subject design*.
  About 100 times less interactions needed compared to A/B testing.
- **Only pairwise.** Given a set of systems, quadratic comparisons are required**.** Often prohibitive.

[2] T. Joachims. Optimizing search engines using clickthrough data. In KDD, 2002.
[3] T. Joachims, L. A. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search. In ACM TOIS, 2007.
[4] F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In CIKM, 2008.

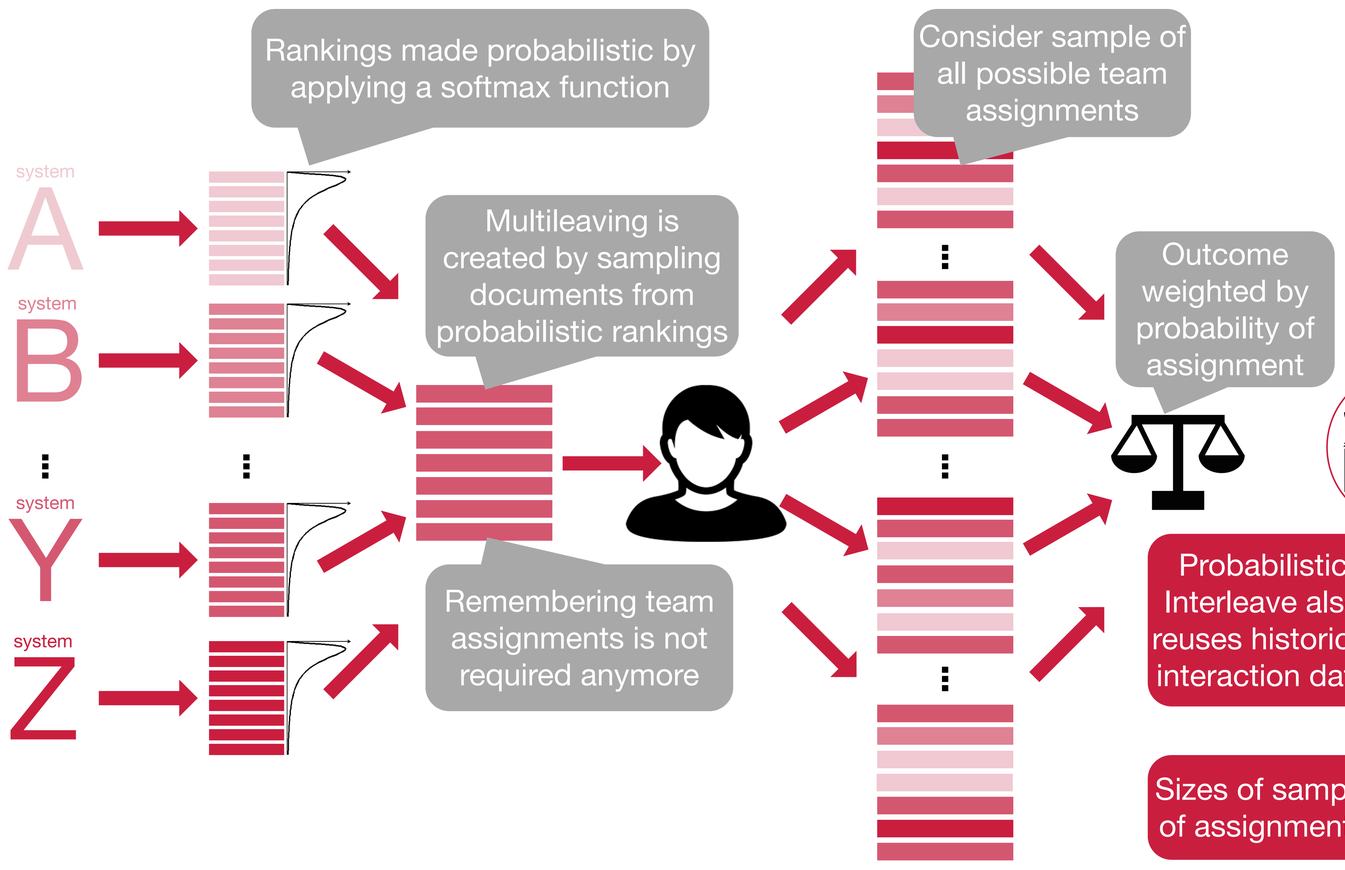
Multileaved Comparisons (TDM)



+ **Highly sensitive**, *within subject design*.
  Even more sensitive than interleaving, depending on the number of systems and result list length.
+/- **Many rankings at a time.** But not many more than can be represented in the result list.
- **No reuse** of historical interaction data.
  Comparisons always involve a user.

[5] A. Schuth, F. Sietsma, S. Whiteson, D. Lefortier, and M. de Rijke. Multileaved comparisons for fast online evaluation. In CIKM, 2014.

Probabilistic Multileaved Comparisons (PM)



+ **Highly sensitive**, *within subject design*.
  As sensitive as TDM Multileaved comparisons.
+ **Unlimited number of systems at a time.**
+ **Reuse** of historical interaction data is possible**.**
  Sets of new systems can be compared using historical clicks.

Clicks from these users are very noisy

Preference Error after 500 impressions

	perfect	navigational	informational
PI	0.085 (0.08)	0.137 (0.11)	0.363 (0.15)
TDM	0.037 (0.06)	0.038 (0.05)	0.099 (0.09)
PM($n = 10^2$)	0.062 (0.07) ▼▲	0.073 (0.07) ▼▲	0.162 (0.10) ▼▲
PM($n = 10^3$)	0.054 (0.05) ▼▲	0.060 (0.06) ▼▲	0.117 (0.09) ▼△
PM($n = 10^4$)	0.046 (0.05) ▼	0.054 (0.05) ▼▲	0.090 (0.08) ▼
PM($n = 10^5$)	0.046 (0.05) ▼	0.039 (0.05) ▼	0.087 (0.08) ▼

This research was partially supported by Amsterdam Data Science, the Dutch national program COMMIT, Elsevier, the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement nr 312827 (VOX-Pol), the ESF Research Network Program ELIAS, the HPC Fund, the Royal Dutch Academy of Sciences (KNAW) under the Elite Network Shifts project, the Microsoft Research Ph.D. program, the Netherlands eScience Center under project number 027.012.105, the Netherlands Institute for Sound and Vision, the Netherlands Organisation for Scientific Research (NWO) under project nrs 727.011.005, 612.001.116, HOR-11-10, 640.006.013, 612.066.930, CI-14-25, SH-322-15, the Yahoo! Faculty Research and Engagement Program, Yandex, and an ACM SIGIR Student Travel Grant.