



Learning to Rank with Apache Spark

A Case Study in Production Machine Learning
#SAISML12

Adam Davidson and Anna Bladzich, Elsevier





Empowering Knowledge

Elsevier is a global information analytics business that helps institutions and professionals advance healthcare, open science, and improve performance for the benefit of humanity

ScienceDirect®



Scopus®

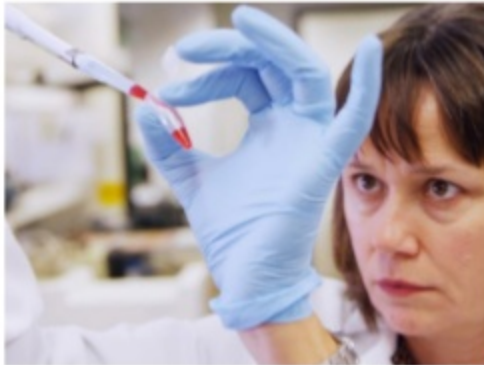
THE LANCET



#SAISML12

What do we do?

We combine content and data with analytics and technology to help:



RESEARCHERS

to make new discoveries and
have more impact on society



CLINICIANS

to treat patients better
and save more lives



NURSES

throughout their careers
and to help save lives



#SAISML12

Why do we need recommendations?



THE LD₅₀ OF TOXICITY DATA IS
2 KILOGRAMS PER KILOGRAM.

ScienceDirect

- Scientific publication database
- 15 million articles
- Millions of visitors every month

The screenshot shows the ScienceDirect interface for an article in the journal 'Big Data Research'. The article title is 'Machine Learning with Big Data An Efficient Electricity Generation Forecasting System'. The authors listed are Mohammad Naimur Rahman, Amir Esmailpour, and Junhui Zhao. The article is part of a special issue on 'Big data analytics and applications'. On the right side, there is a 'Recommended articles' section highlighted with a red box, listing three related articles with their titles, volumes, and page numbers, along with 'Download PDF' and 'View details' links. The main article's abstract is visible at the bottom, discussing the use of Machine Learning (ML) for electricity generation forecasting.

Download PDF Export

Search ScienceDirect Advanced

Big Data Research
Volume 5, September 2016, Pages 9-15

Machine Learning with Big Data An Efficient Electricity Generation Forecasting System ☆

Mohammad Naimur Rahman, Amir Esmailpour, Junhui Zhao

Show more

<https://doi.org/10.1016/j.bdr.2016.02.002> Get rights and content

Abstract

Machine Learning (ML) is a powerful tool that can be used to make predictions on the future nature of data based on the past history. ML algorithms operate by building a model from input examples to make data-driven predictions or decisions for the future. The growing concept "Big Data" has brought much success in the field of data science; it provides data scalability in a variety of ways that empower data science. ML can also be used in conjunction with Big Data to build effective predictive systems or to solve complex data analytic problems. In this work, we propose an electricity generation forecasting system that could predict the amount of power required at a rate close to the electricity consumption for

Part of special issue:
Big data analytics and applications
Edited by Jian Pei, Guoliang Li, Hanghang Tong

Download full issue

Other articles from this issue

Recommended articles

From Big Data to Data Science: A Multi-disciplinary ...
Big Data Research, Volume 1, 2014, p. 1
Download PDF View details

A Hybrid Data Center Architecture for Big Data
Big Data Research, Volume 3, 2016, pp. 29-40
Download PDF View details

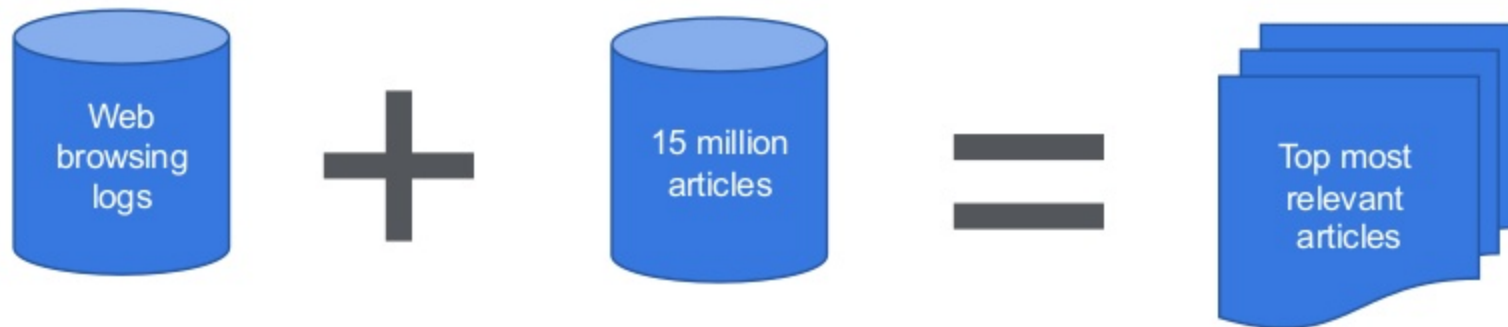
Closed-loop Big Data Analysis with Visualization a ...
Big Data Research, Volume 8, 2017, pp. 12-26
Download PDF View details

1 2 Next



#SAISML12

How did we build recommendations for ScienceDirect?





Collaborative Filtering



Learning to Rank



Model Evaluation



#SAISML12

Images from: [josuthea](#), [kittyfiction](#), [dailypakistan](#)

Collaborative Filtering



- Widely used in the industry
- No knowledge about items or users
- Using the wisdom of crowds

Customers who bought this item also bought



#SAISML12

Collaborative Filtering

- Usage matrix
- Browsing history
- User's who bought X also bought Y



#SAISML12

Item Based Collaborative Filtering

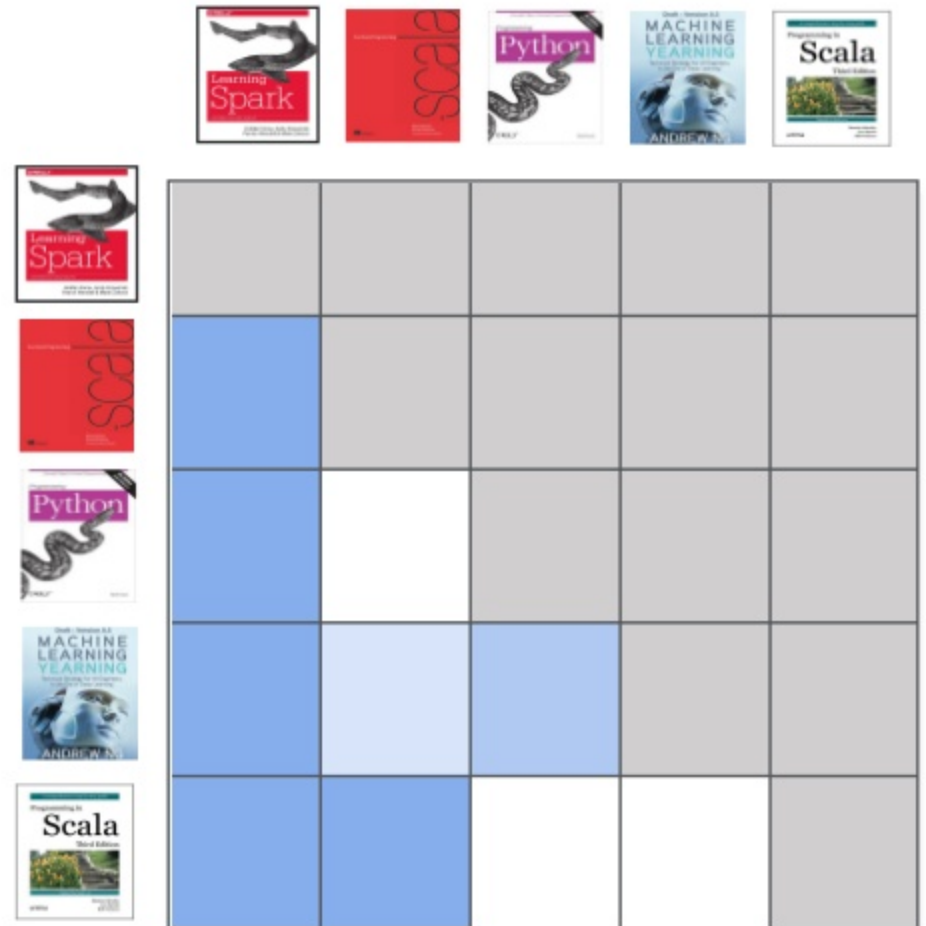
- Pairwise cosine similarity
- Similarity matrix
- K nearest-neighbors



#SAISML12

Item Based Collaborative Filtering

- Pairwise cosine similarity
- Similarity matrix
- K nearest-neighbors



#SAISML12

Collaborative Filtering in production





Can we do any better?



#SAISML12

Image: [shutterstock](#)

A wealth of features

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

CF score



Popularity

Temporal



Text



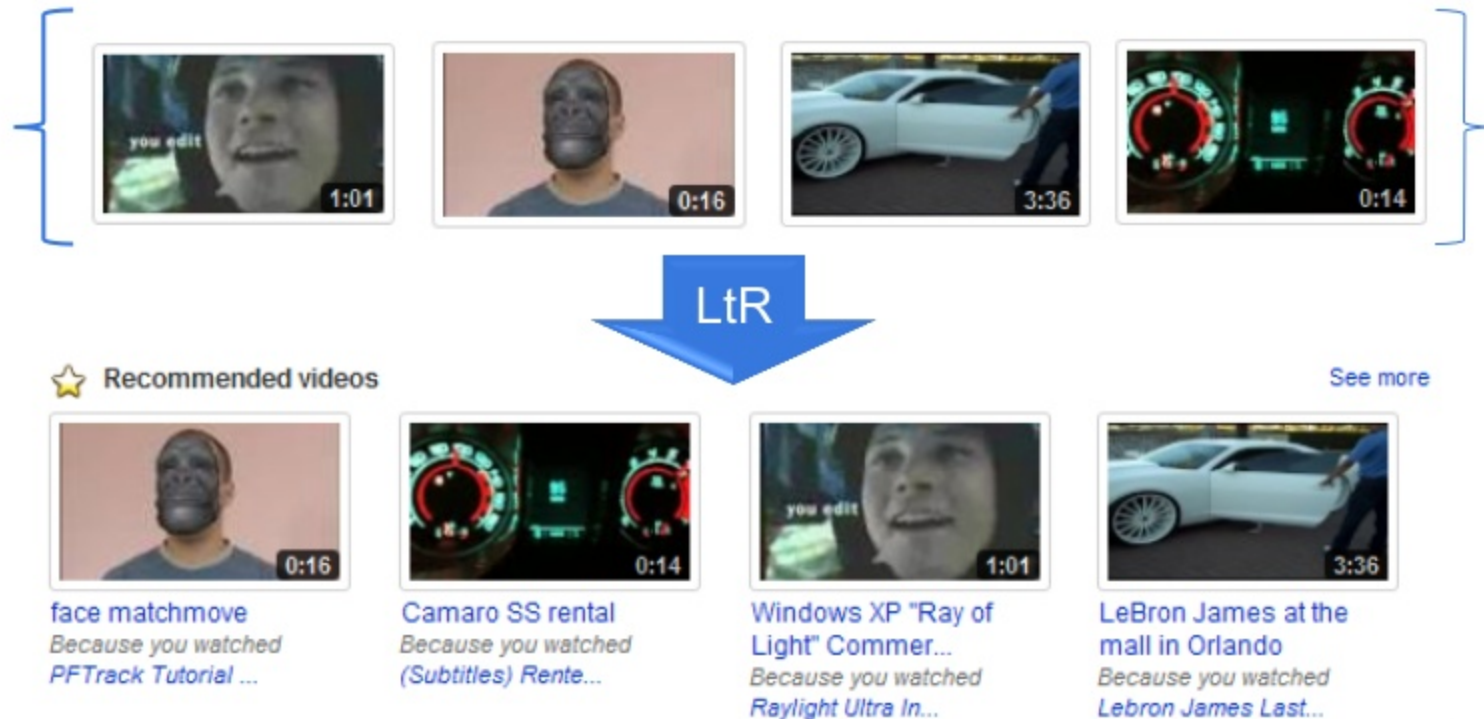
Subject



#SAISML12

Images: [wsj](#), [alamy](#), [bookedelic](#)

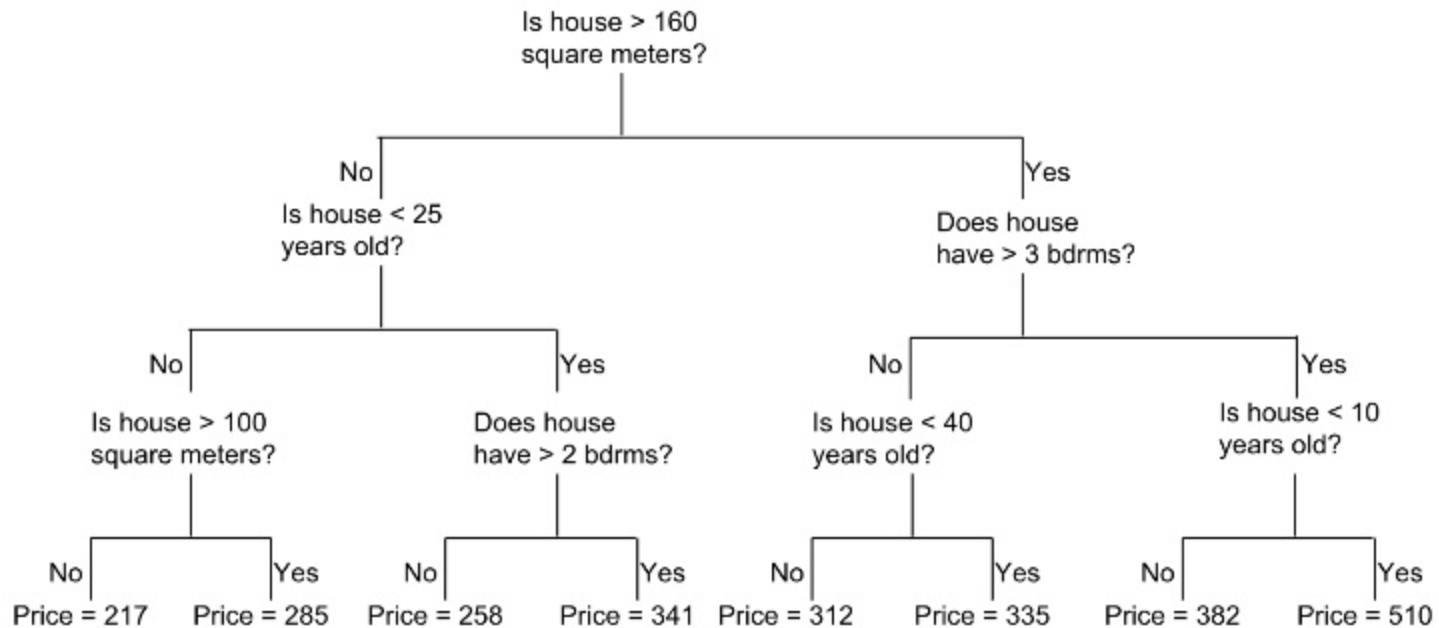
Learning to Rank (LtR)



#SAISML12

Image: hunterwalk.com

LtR Model – Decision Tree





Scaling machine learning for target prediction in drug discovery using Apache Spark ☆

Dries Harnie^{a,*,}, Mathijs Saey^{a,}, Alexander E. Vapirev^{b, c,}, Jörg Kurt Wagner^{b,}, Andrey Gadich^{1,}, Marvin Steijaert^{a,}, Hugo Ceulemans^{b, c,}, Roel Wuyts^{c, d, e,}, Wolfgang De Meuter^a

[Show more](#)

<https://doi.org/10.1016/j.future.2016.04.023>

[Get rights and content](#)

Recommended articles

Applying spark based machine learning model on ...
Computers & Electrical Engineering, Volume 65, 2018, ...

[Download PDF](#) [View details](#) ▾

Finding exact hitting set solutions for systems biol...
Future Generation Computer Systems, Volume 67, 20...

[Download PDF](#) [View details](#) ▾

Boosting analyses in the life sciences via clusters, ...
Future Generation Computer Systems, Volume 67, 20...

[Download PDF](#) [View details](#) ▾

1 2 [Next](#) >

Gather data

Calculate CTR for
recommendations
by article

Enrich

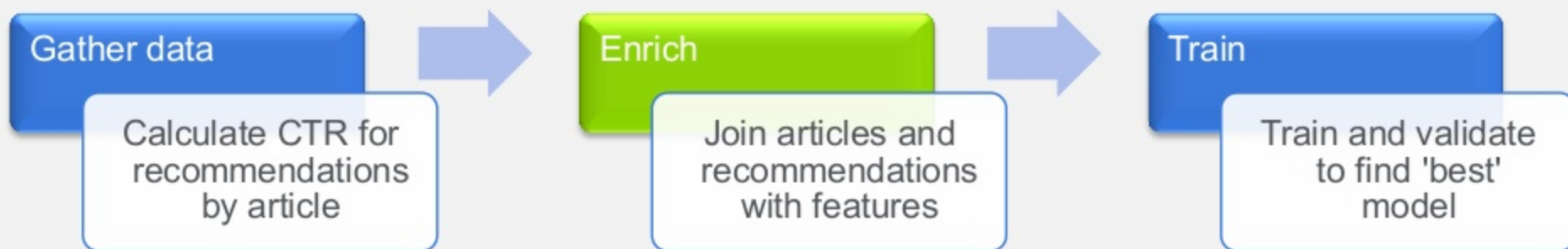
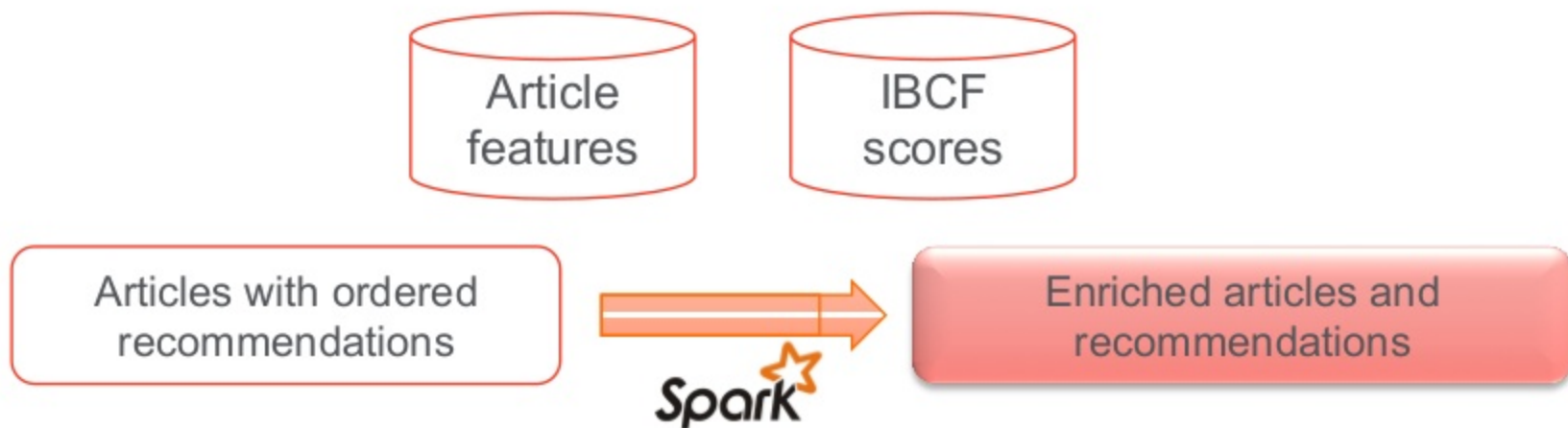
Join articles and
recommendations
with features

Train

Train and validate
to find 'best'
model

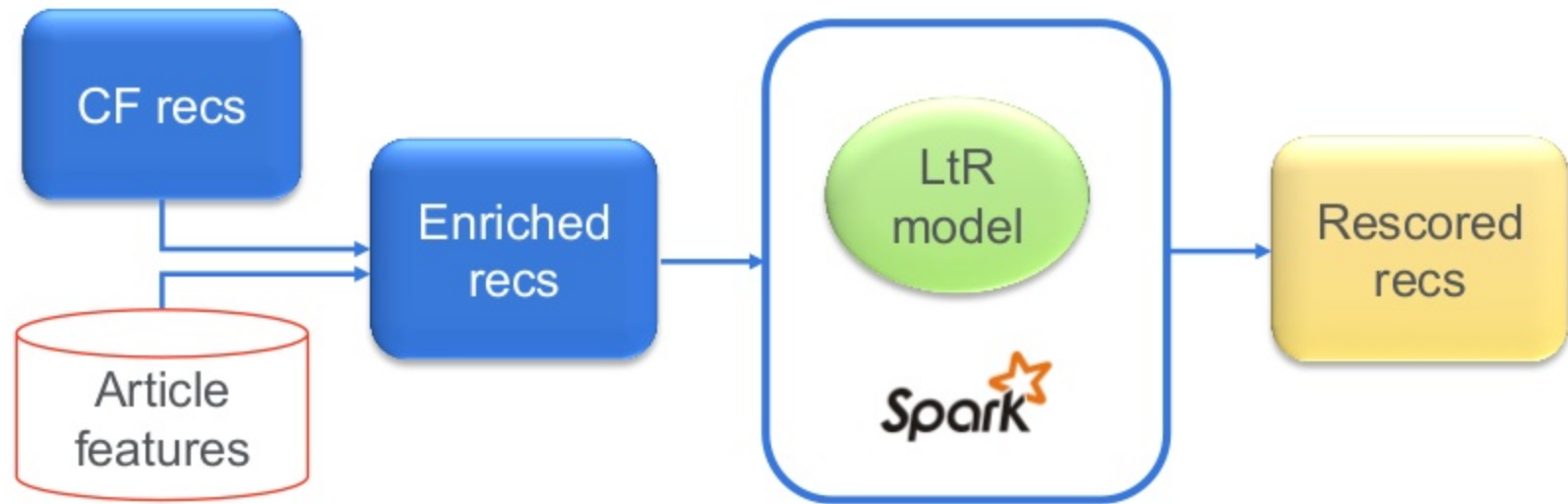


#SAISML12

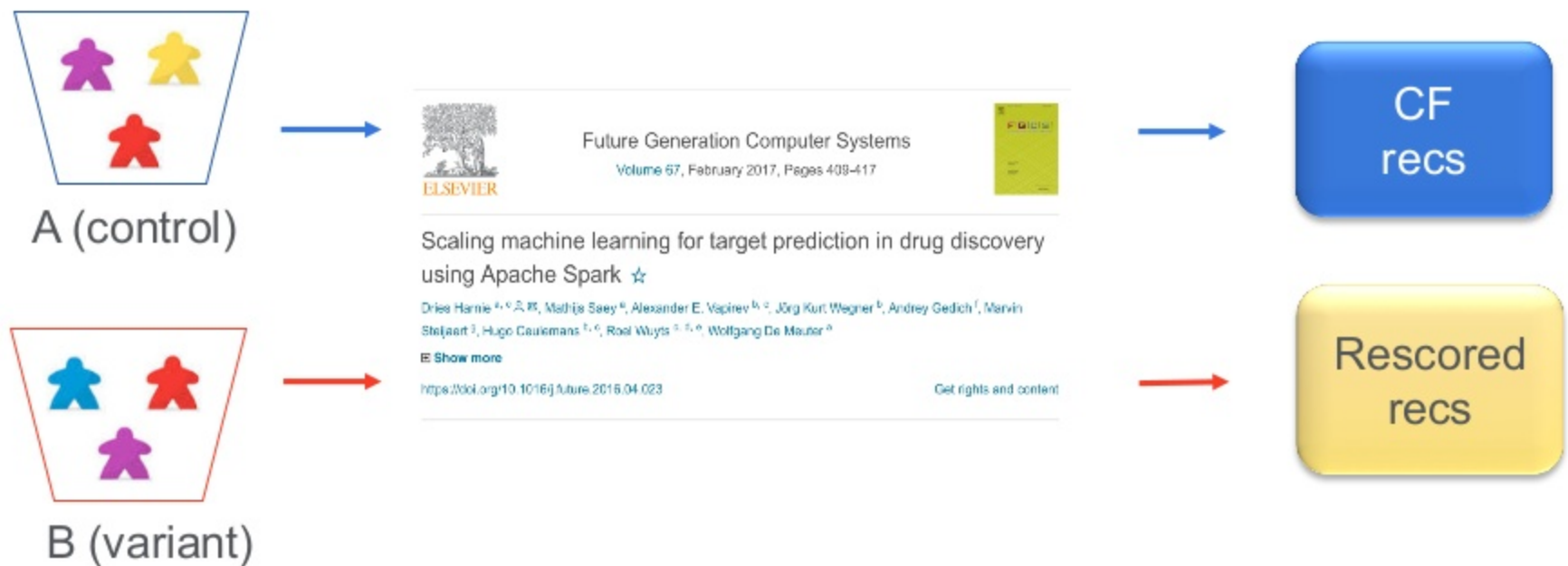


#SAISML12

Recommendation Rescoring



Online model evaluation - A/B testing



#SAISML12

Result: **7-10%** improvement
in user engagement



#SAISML12

GIF: [imgur](https://imgur.com)

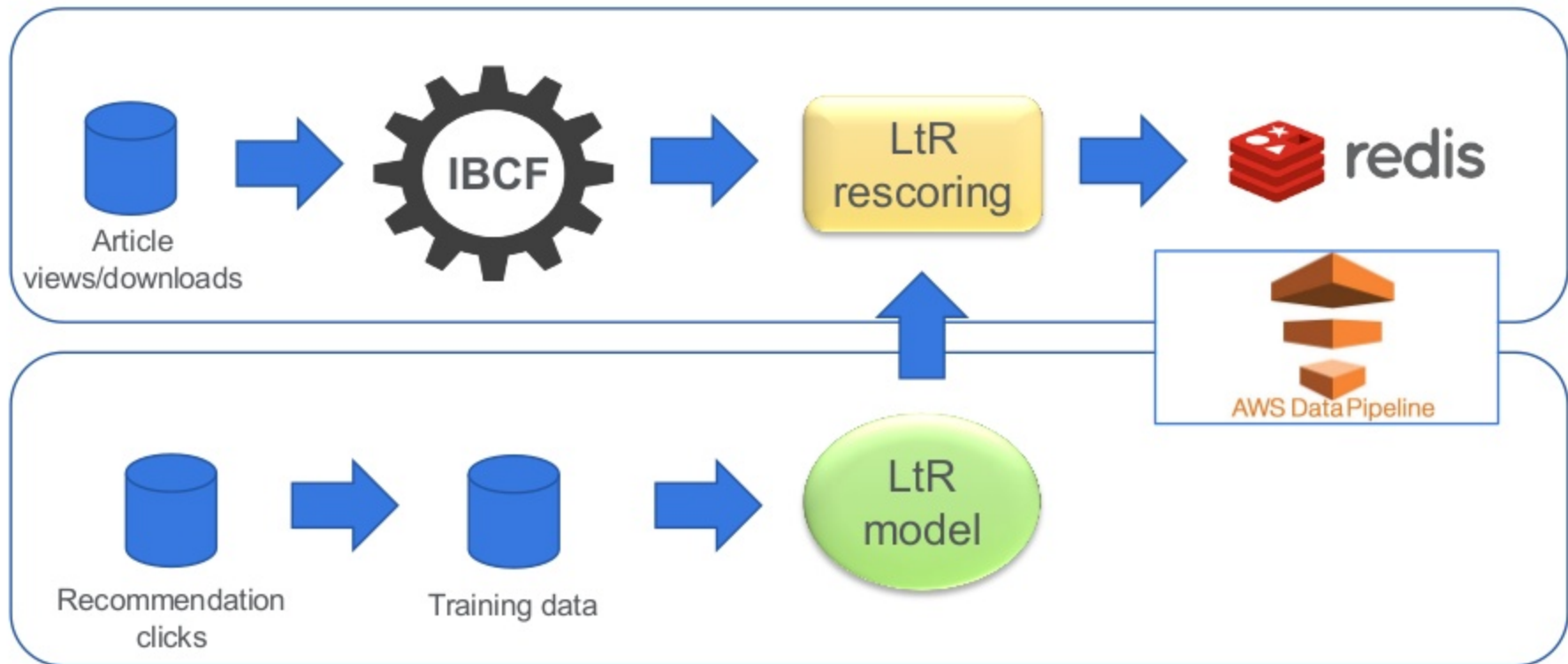
Adaptive LtR Model – keep training



#SAISML12

Image: [pixabay](https://pixabay.com/)

Collaborative Filtering & Learning to Rank



#SAISML12



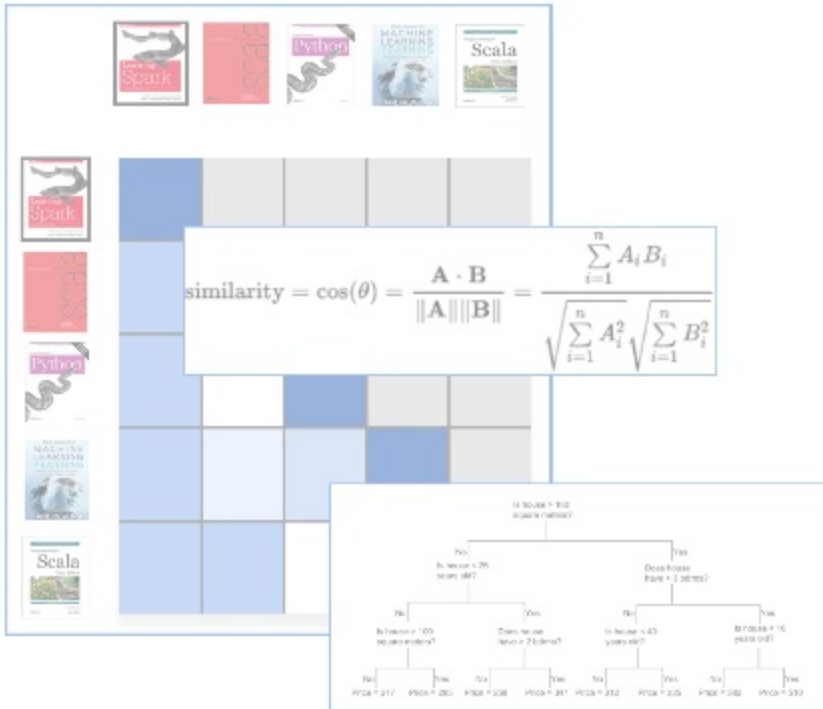
Conclusions



#SAISML12

Good
recommendations
can make a
difference





Collaborative
filtering and
Learning to Rank
work great!




Apache Spark is
the foundation for
scalable machine
learning



#SAISML12





We're hiring, come speak to us!

<https://www.elsevier.com/about/careers/technology-careers>



ELSEVIER

#SAISML12



Thank you

Adam Davidson - a.davidson.1@elsevier.com

Anna Bladzich - a.bladzich@elsevier.com

<https://www.elsevier.com/about/careers/technology-careers>

