



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®

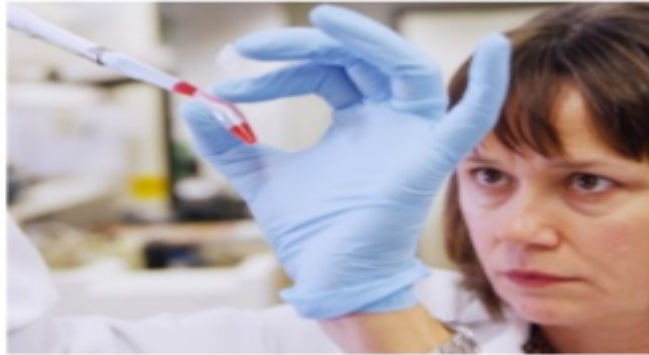
THE LANCET



#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. At the top, there are links for 'Download PDF' and 'Export', and a search bar with 'Search ScienceDirect' and an 'Advanced' option. The article is from 'Big Data Research', Volume 5, September 2016, Pages 9-15. The title is 'Machine Learning with Big Data An Efficient Electricity Generation Forecasting System'. The authors are Mohammad Naimur Rahman, Amir Esmailpour, and Junhui Zhao. The abstract discusses the use of Machine Learning (ML) for electricity generation forecasting. On the right, there is a 'Recommended articles' section with three articles listed, each with a 'Download PDF' and 'View details' link. The 'Recommended articles' section is highlighted with a red box.

Download PDF Export

Search ScienceDirect Advanced

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

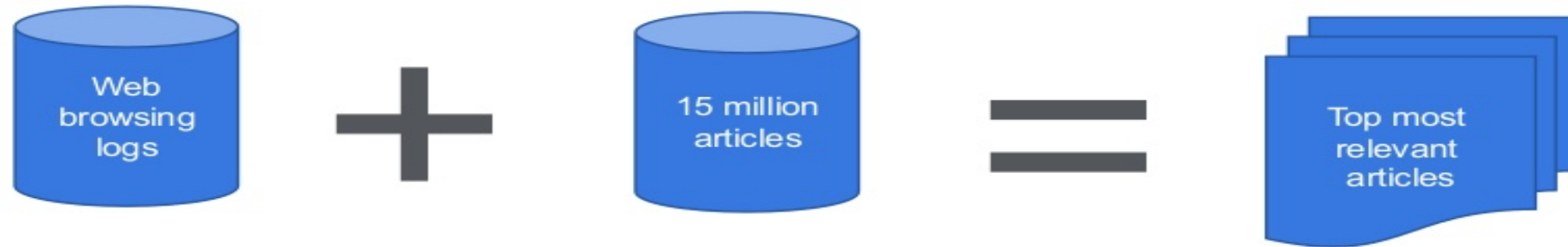
Closed-loop Big Data Analysis with Visualization a...
Big Data Research, Volume 8, 2017, pp. 12-26
Download PDF View details

1 2 Next



#SAISML12

How did we build recommendations for ScienceDirect?



#SAISML12



Collaborative Filtering



Learning to Rank



Model Evaluation



#SAISML12

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

Collaborative Filtering



Learning Spark: Lightning-Fast Big Data Analysis Paperback – 27 Feb 2015
by [Holden Karau](#) (Author), [Andy Konwinski](#) (Author), [Patrick Wendell](#) (Author), [Matei Zaharia](#) (Author)
Get a £3 promo code by listening to Prime Music by Oct 15th. [Learn more.](#)
★★★★☆ 12 customer reviews
See all 4 formats and editions
Kindle Edition £19.11 Paperback ~~£20.55~~ [prime](#)
Read with the [Free App](#) 18 Used from £18.29 30 New from £17.72
Promotional Message: Prime Students get 10% off on Books. 1 promotion.
Note: This item is eligible for click and collect. [Details](#)
Data in all domains is getting bigger. How can you work with it efficiently? Recently updated for Spark 1.3, this book introduces Apache Spark, the open source cluster computing system that makes data

Customers who bought this item also bought



Hadoop: The Definitive Guide
Tom White
★★★★☆ 7
Paperback
£25.99 [prime](#)



Programming in Scala, 3rd Edition
Martin Odersky
★★★★☆ 12
Paperback
£26.99 [prime](#)



High Performance Spark
Holden Karau
★★★★☆ 5
Paperback
£21.54 [prime](#)



Advanced Analytics with Spark: Patterns for Learning from Data at Scale
Uri Laserson
★★★★☆ 5
Paperback
£25.99 [prime](#)



#SAISML12

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

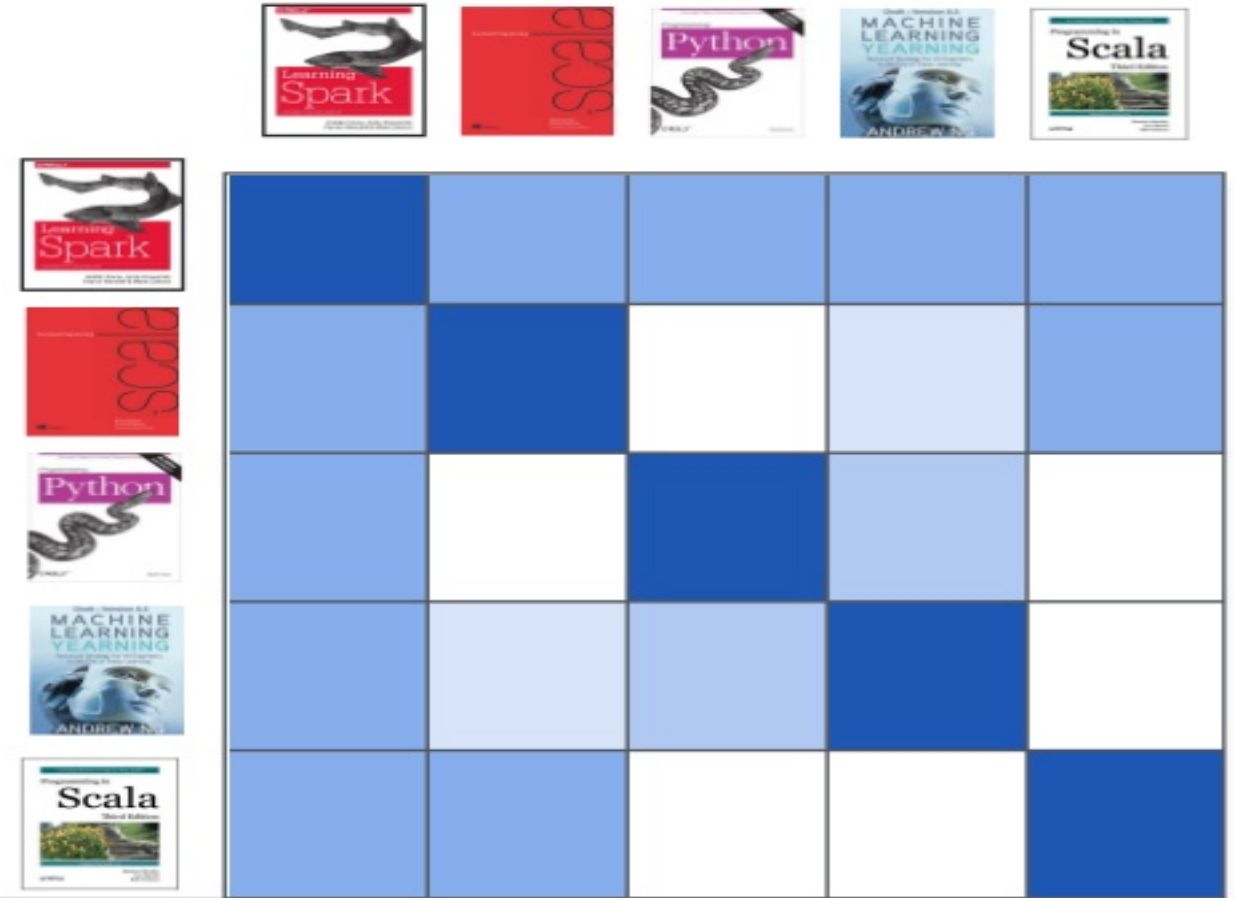
Collaborative Filtering

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



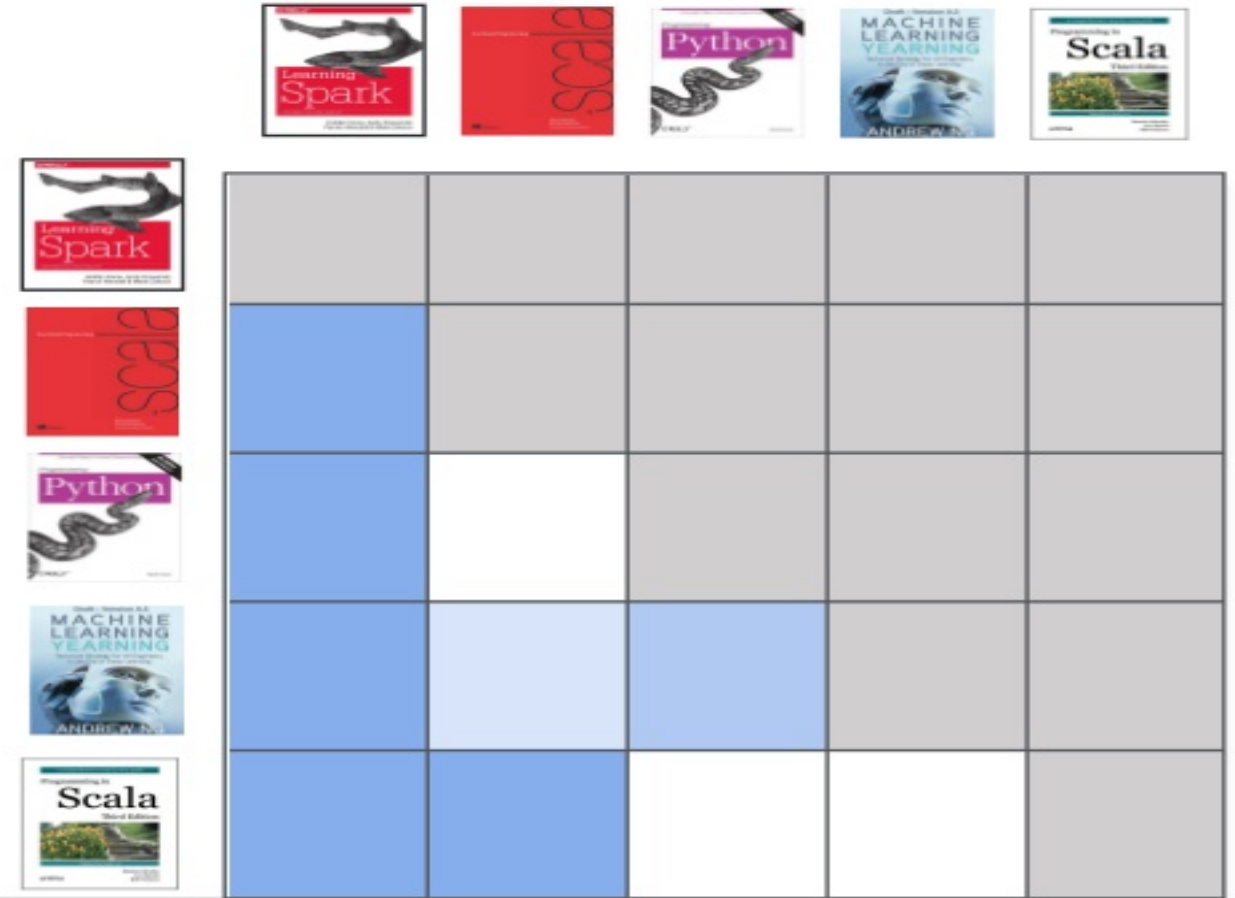
Item Based Collaborative Filtering

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

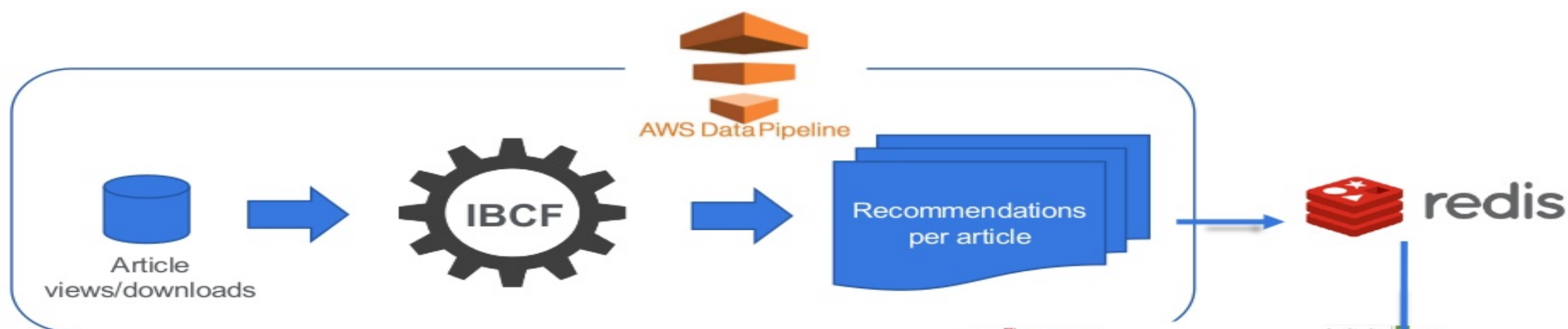


Item Based Collaborative Filtering

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



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

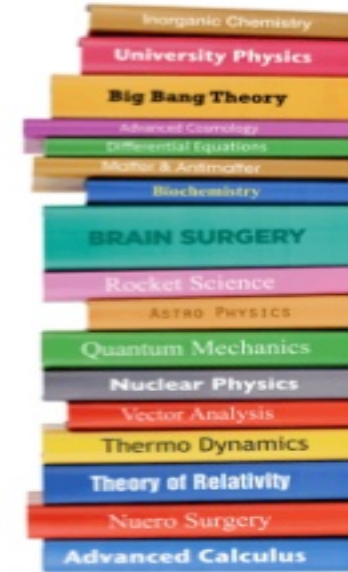


Popularity

Temporal



Text



Subject



#SAISML12

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

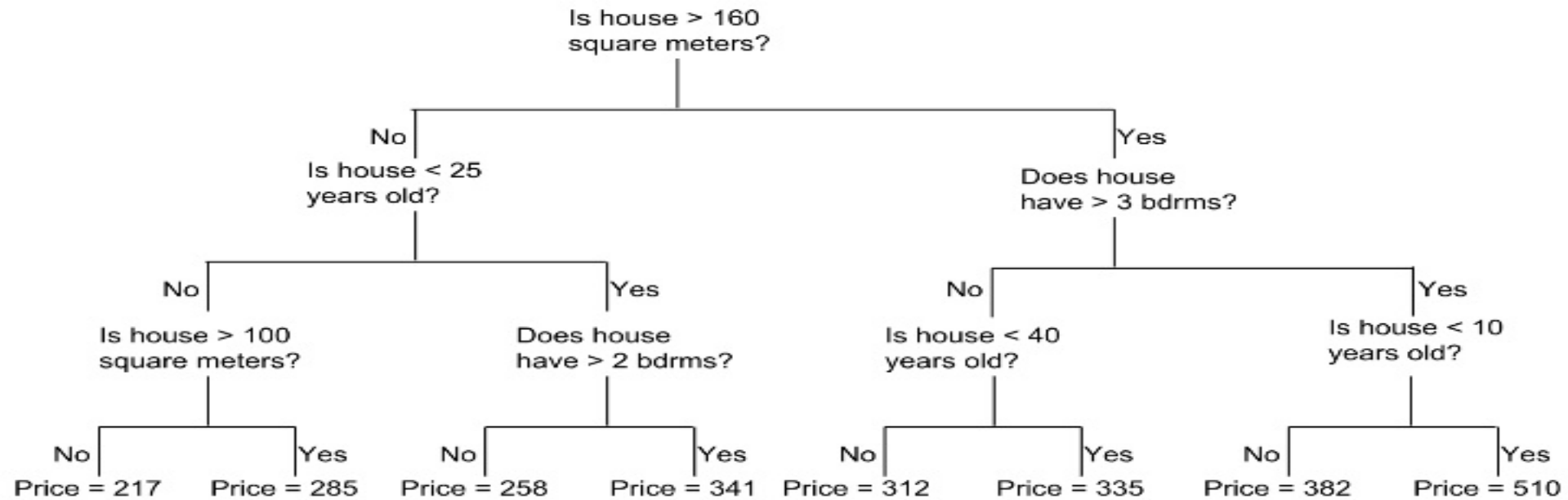
Learning to Rank (LtR)



#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^d, Marvin Steljaert^a, Hugo Caulemans^{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

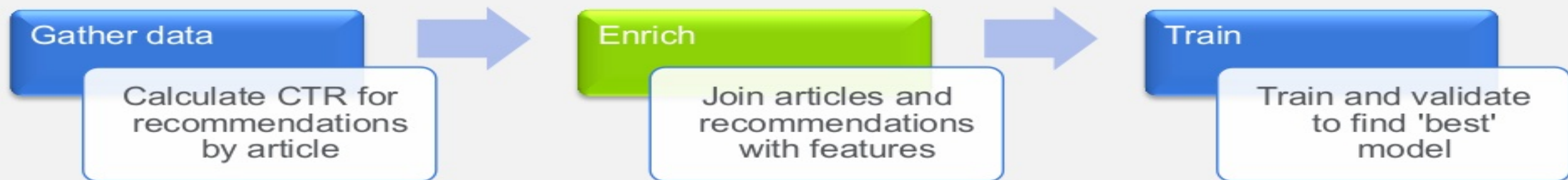
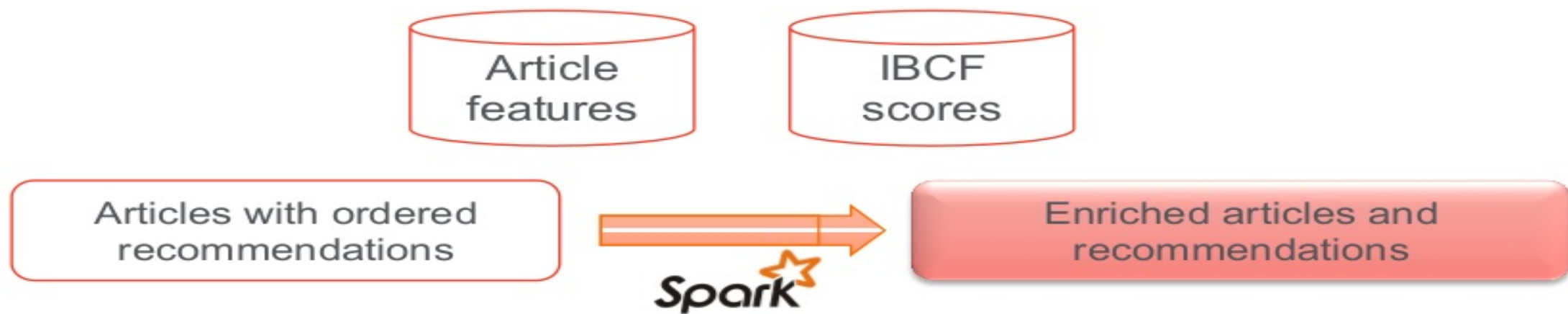
Join articles and
recommendations
with features

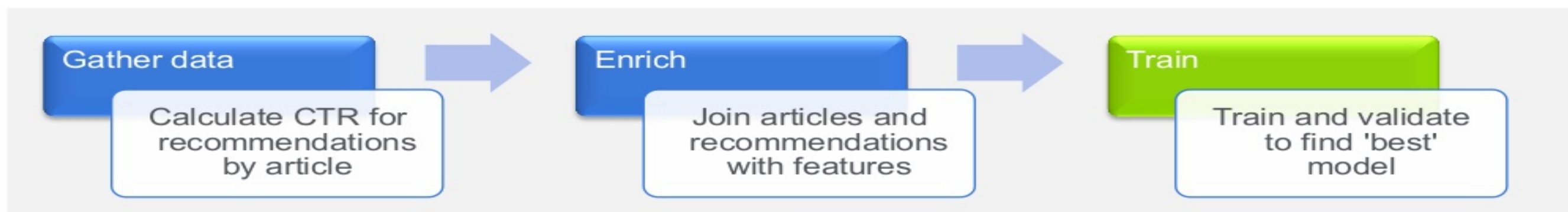
Train

Train and validate
to find 'best'
model



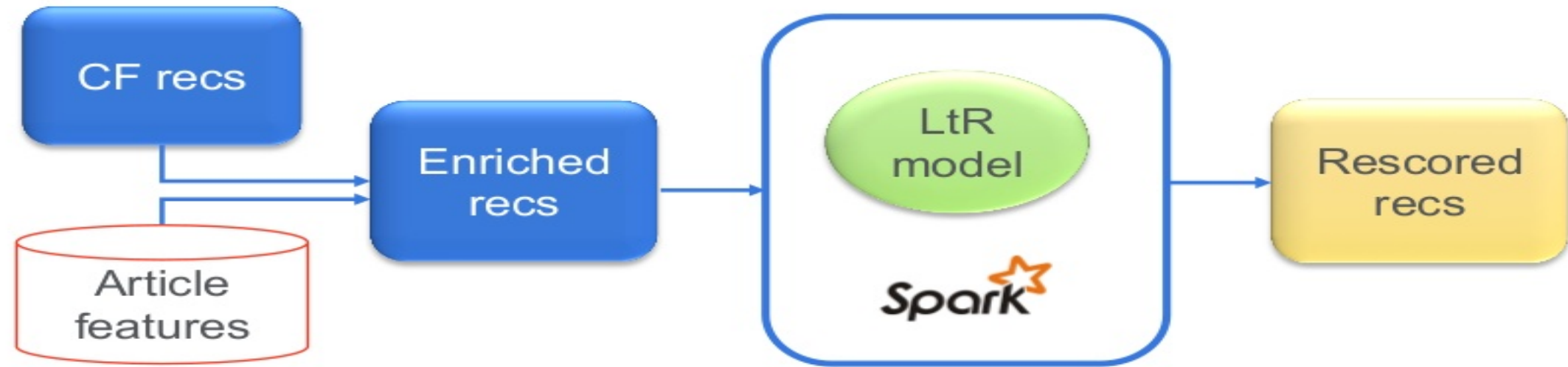
#SAISML12



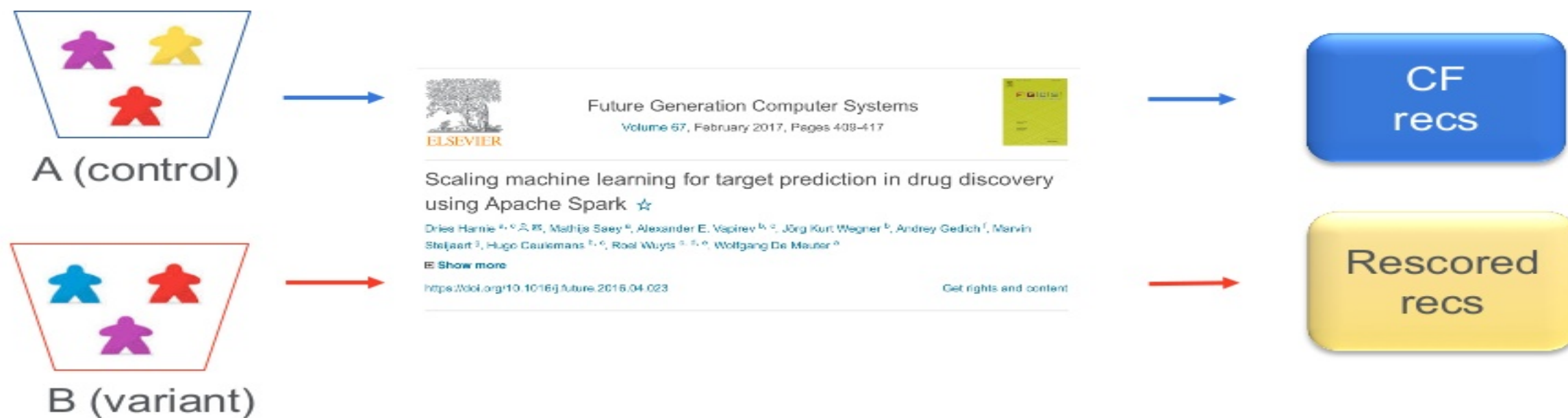


#SAISML12

Recommendation Rescoring



Online model evaluation - A/B testing



#SAISML12

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



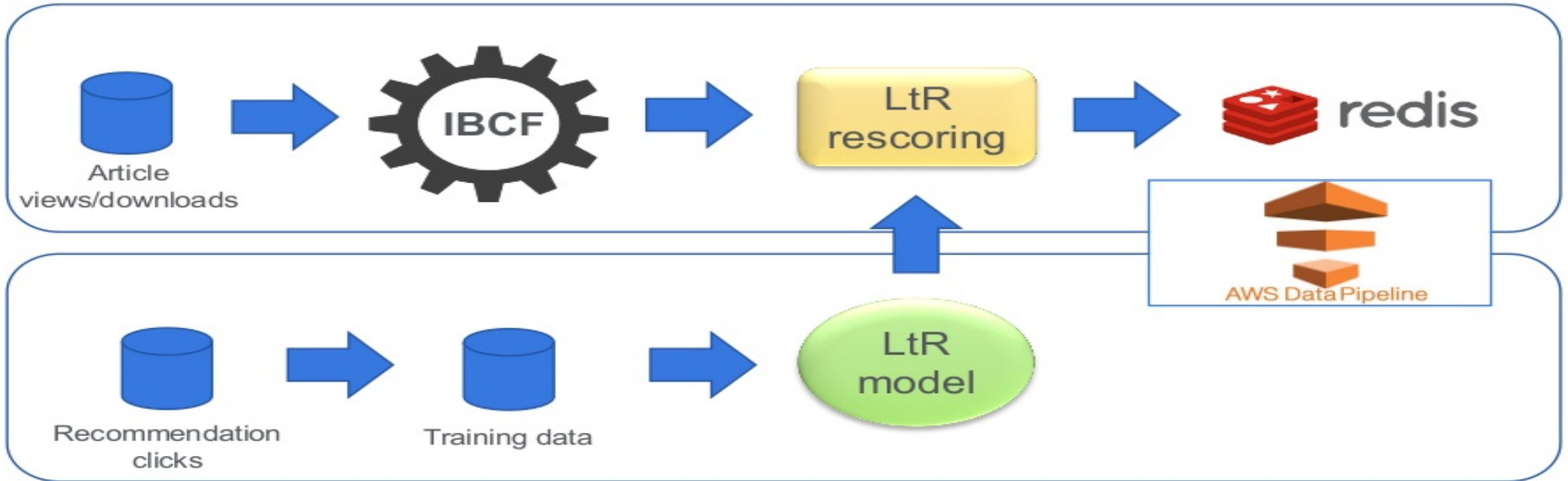
#SAISML12

GIF: [imgur](#)

Adaptive LtR Model – keep training



Collaborative Filtering & Learning to Rank





Conclusions

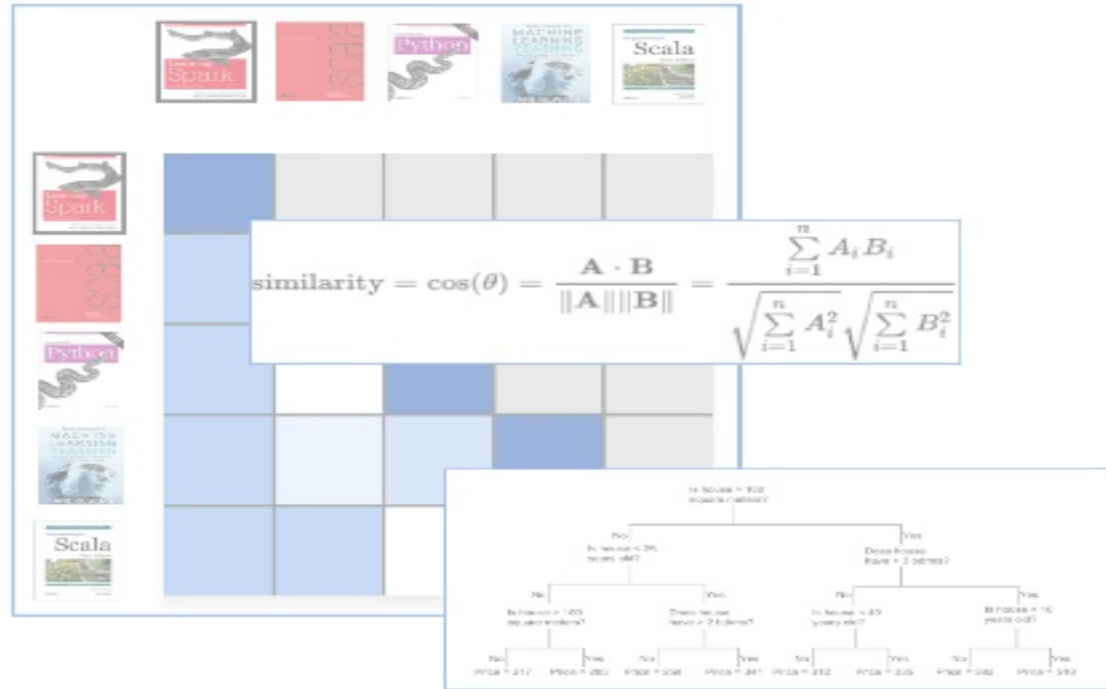


#SAISML12

Good
recommendations
can make a
difference



Collaborative filtering and Learning to Rank work great!



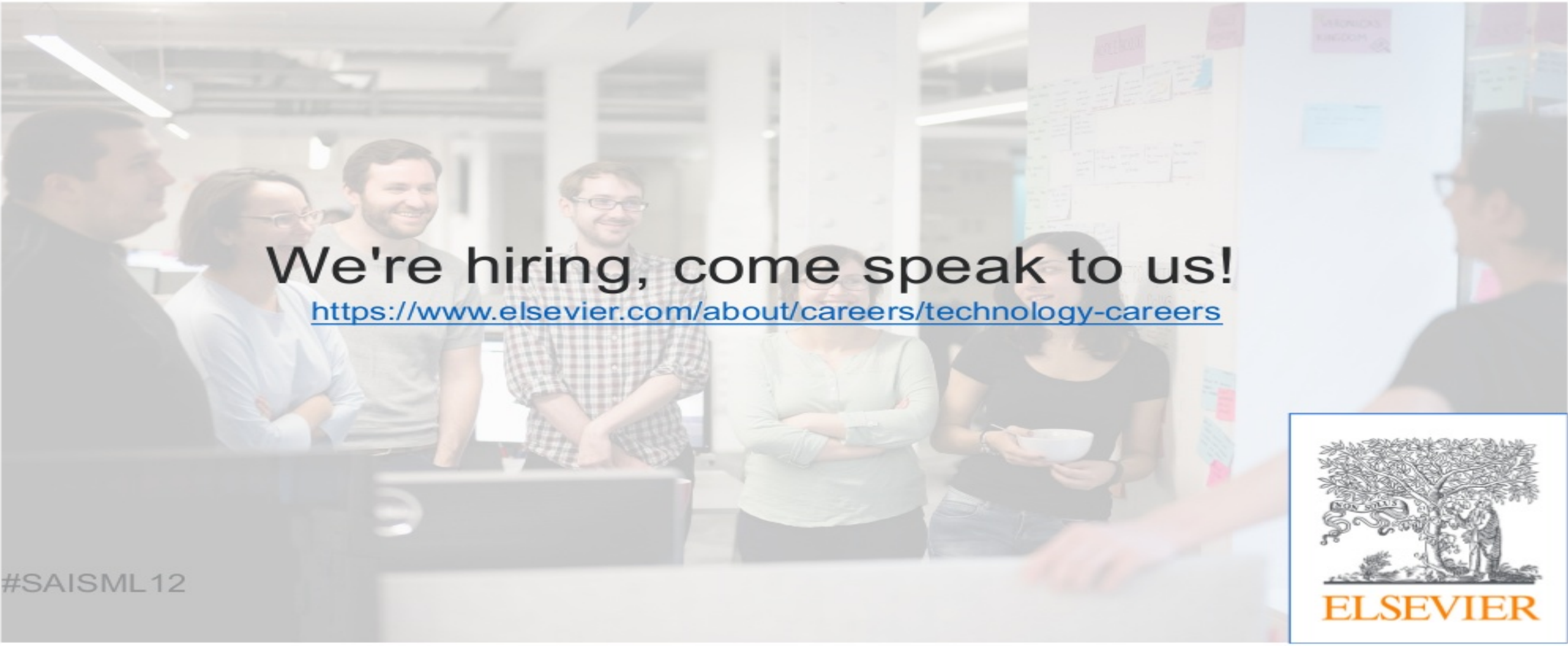
Apache Spark is
the foundation for
scalable machine
learning



#SAISML12



image: [ocado](#)



We're hiring, come speak to us!

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

#SAISML12



ELSEVIER



Thank you

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

Anna Bladzich - a.bladzich@elsevier.com

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

