Learning to Rank with Apache Spark

A Case Study in Production Machine Learning

#SAISML12

Adam Davidson and Anna Bladzich, Elsevier

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®

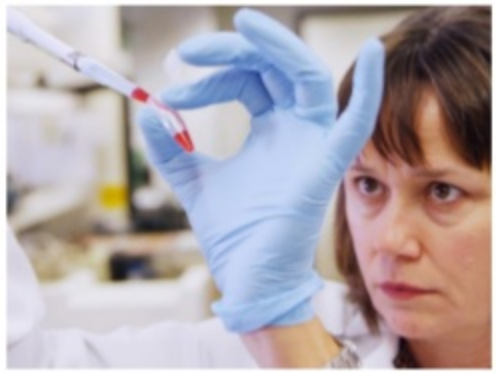
MENDELEY

Scopus®

THE LANCET

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

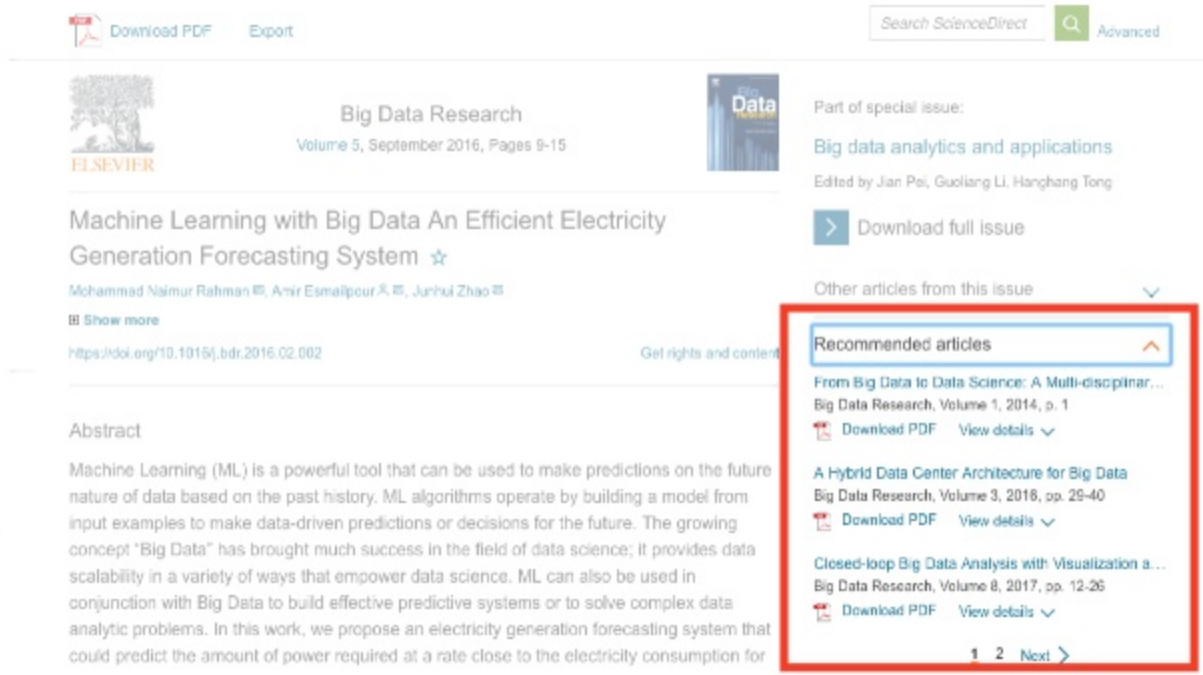
Why do we need recommendations?

THE LD_{50} OF TOXICITY DATA IS
2 KILOGRAMS PER KILOGRAM.

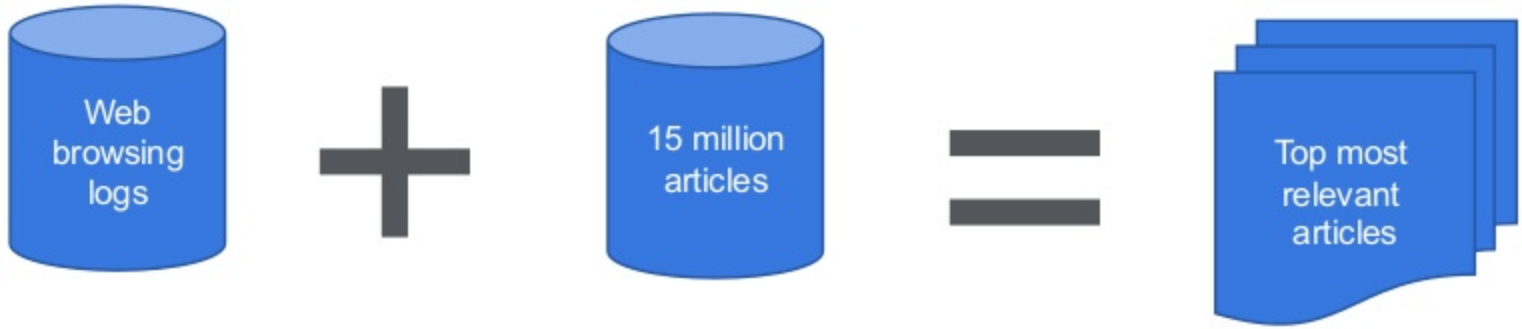
Image: xkcd

ScienceDirect

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



How did we build recommendations for ScienceDirect?



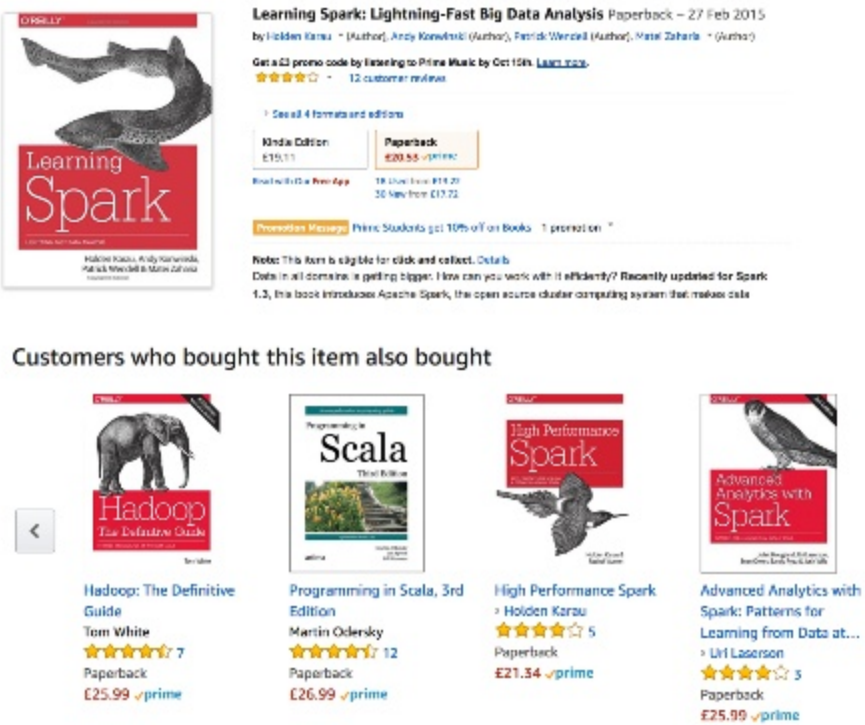
Collaborative Filtering

Learning to Rank

Model Evaluation

Images from: josuthea, kittyfiction, dailypakistan

Collaborative Filtering



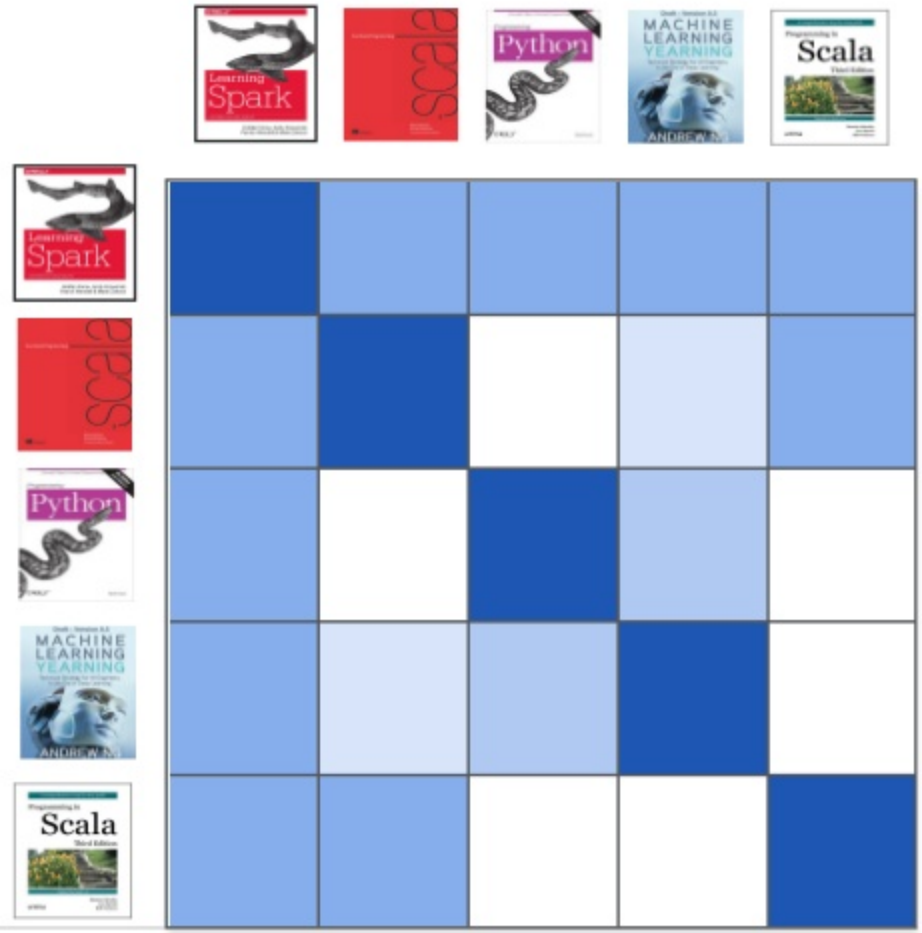
- Widely used in the industry
- No knowledge about items or users
- Using the wisdom of crowds

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?

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

Temporal

Subject

Popularity

Text

Images: wsj, alamy, bookedelic

Learning to Rank (LtR)

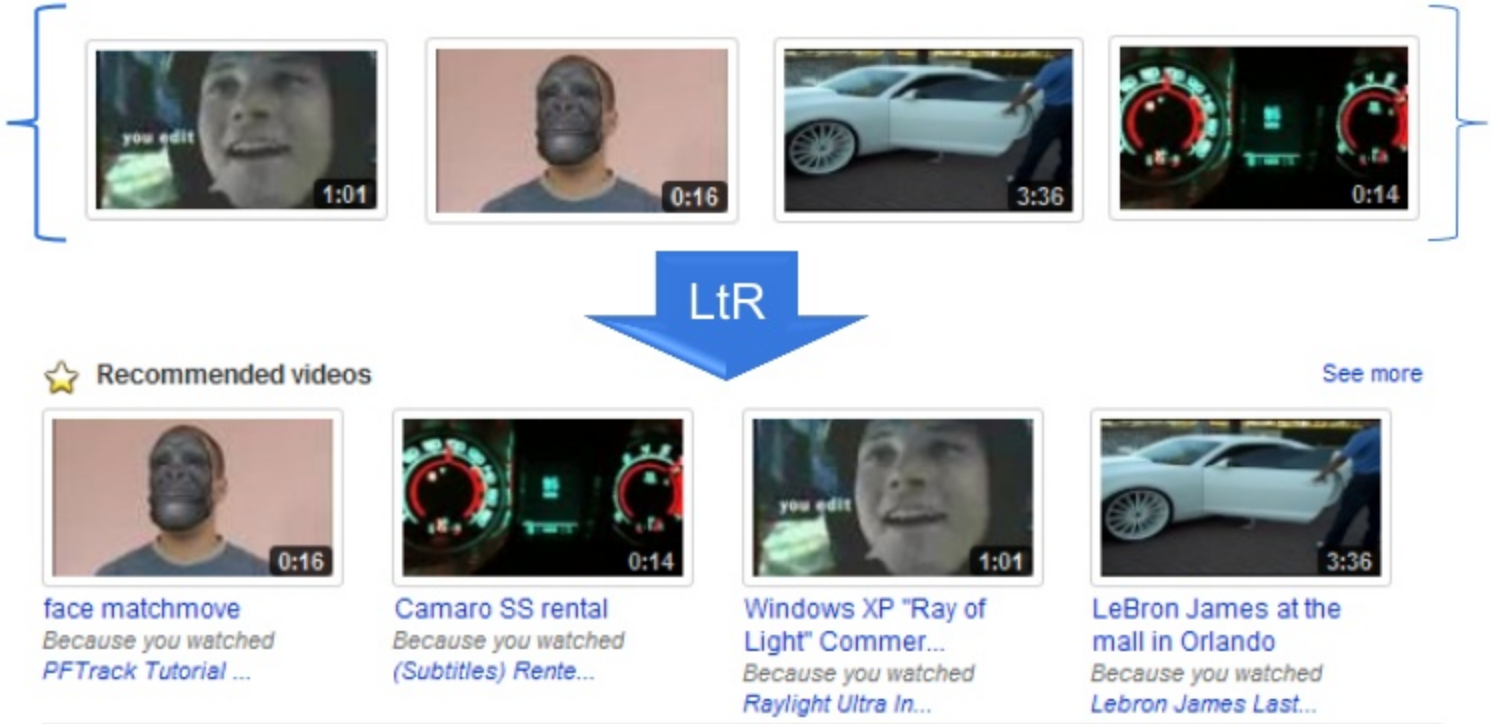


Image: hunterwalk.com

LtR Model – Decision Tree

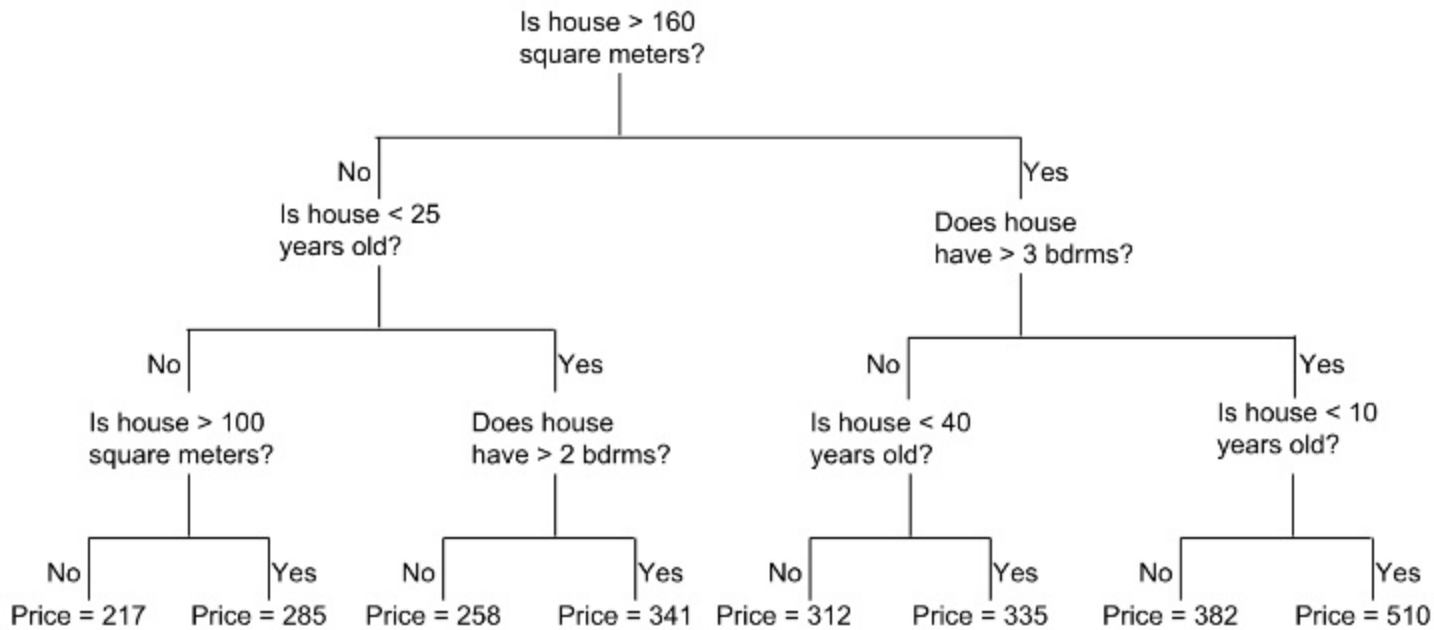
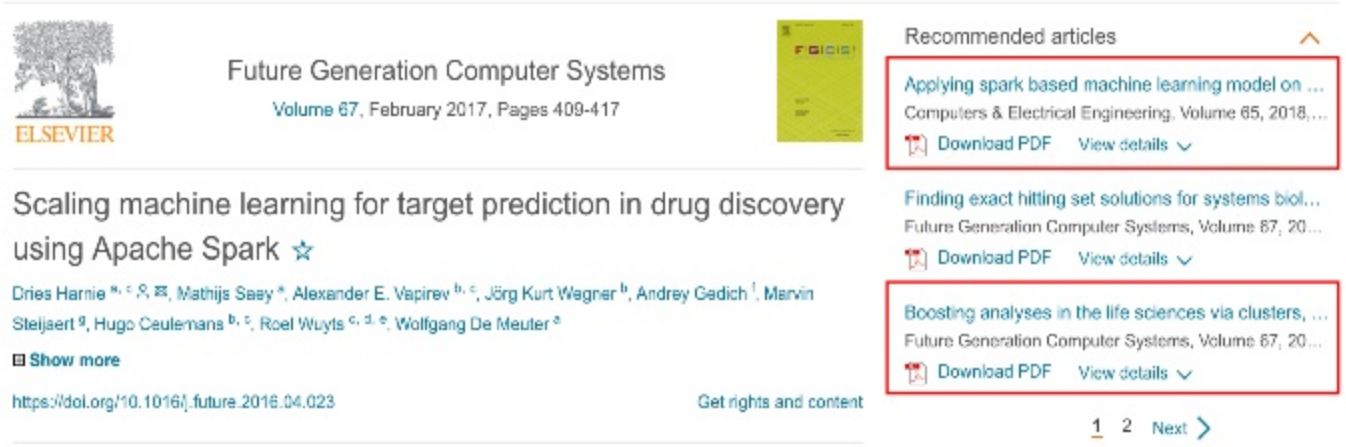


Image: Google machine learning glossary



Gather data

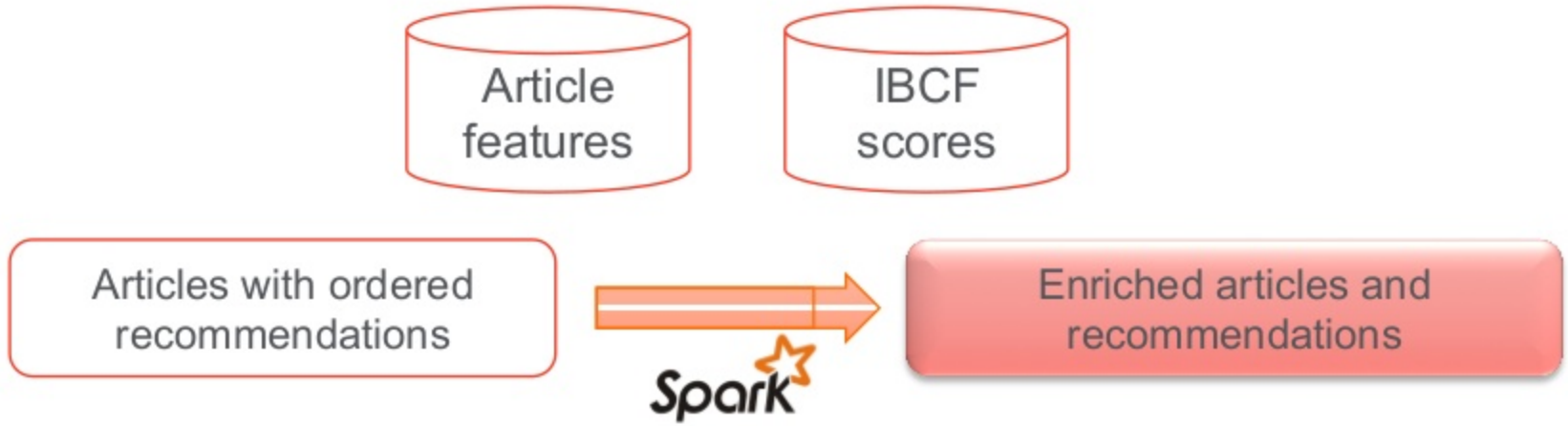
Calculate CTR for recommendations by article

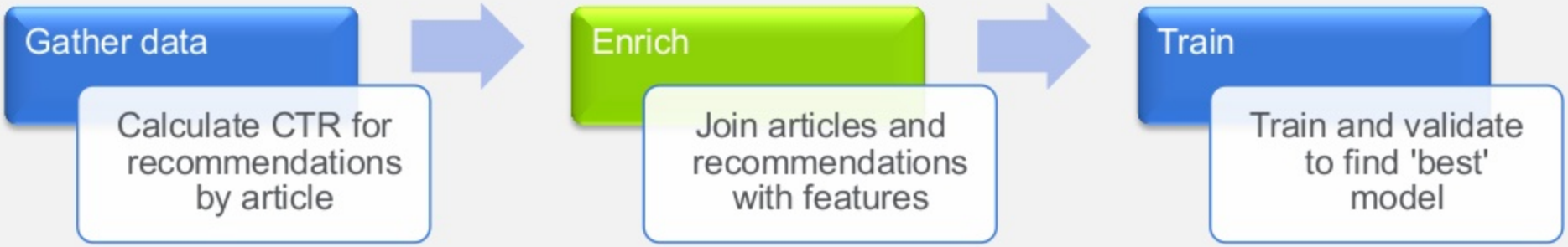
Enrich

Join articles and recommendations with features

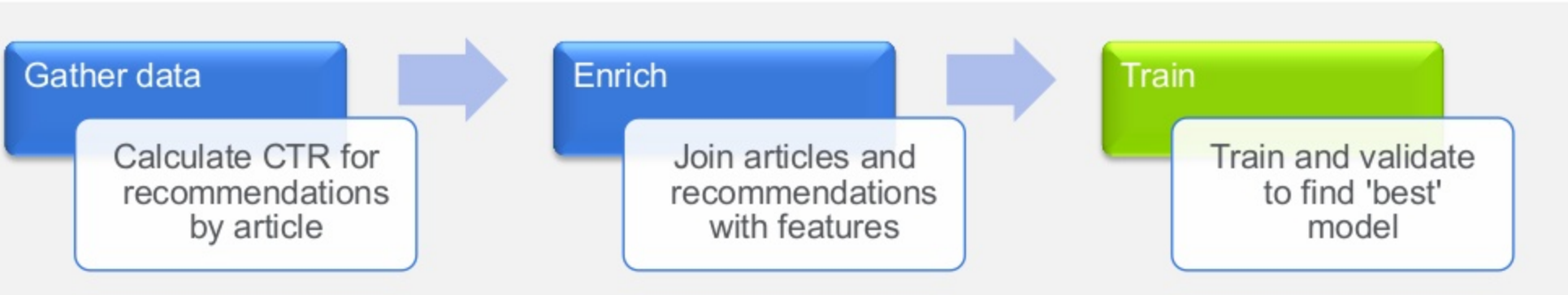
Train

Train and validate to find 'best' model

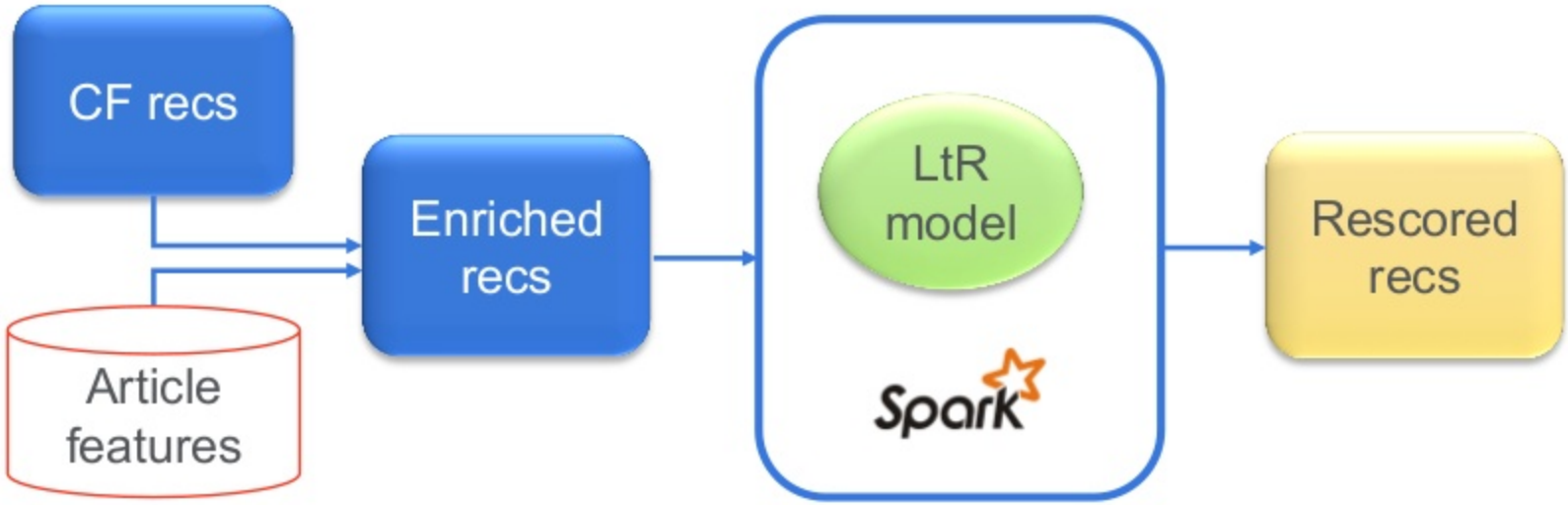

Article features
IBCF scores
Articles with ordered recommendations
Spark
Enriched articles and recommendations


Gather data
Calculate CTR for recommendations by article
Enrich
Join articles and recommendations with features
Train
Train and validate to find 'best' model

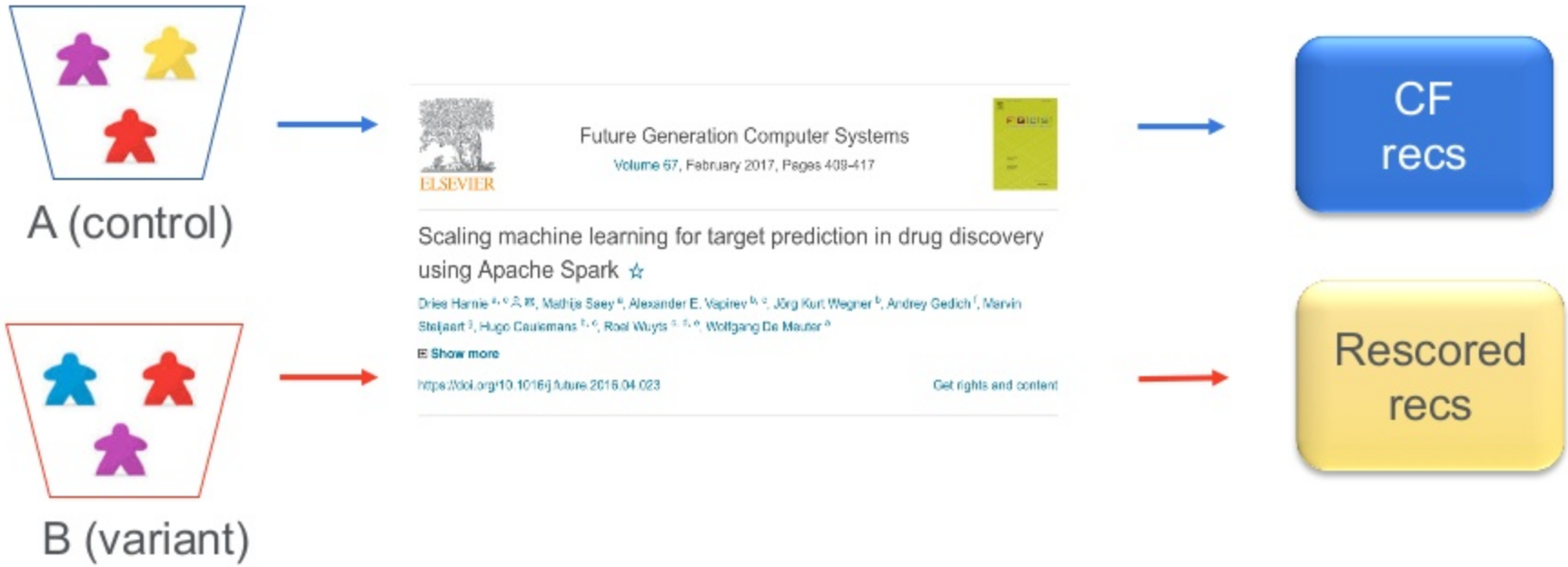

Enriched articles and recommendations
Training (80%)
Validation (20%)
Model


Gather data
Calculate CTR for recommendations by article
Enrich
Join articles and recommendations with features
Train
Train and validate to find 'best' model

Recommendation Rescoring



Online model evaluation - A/B testing



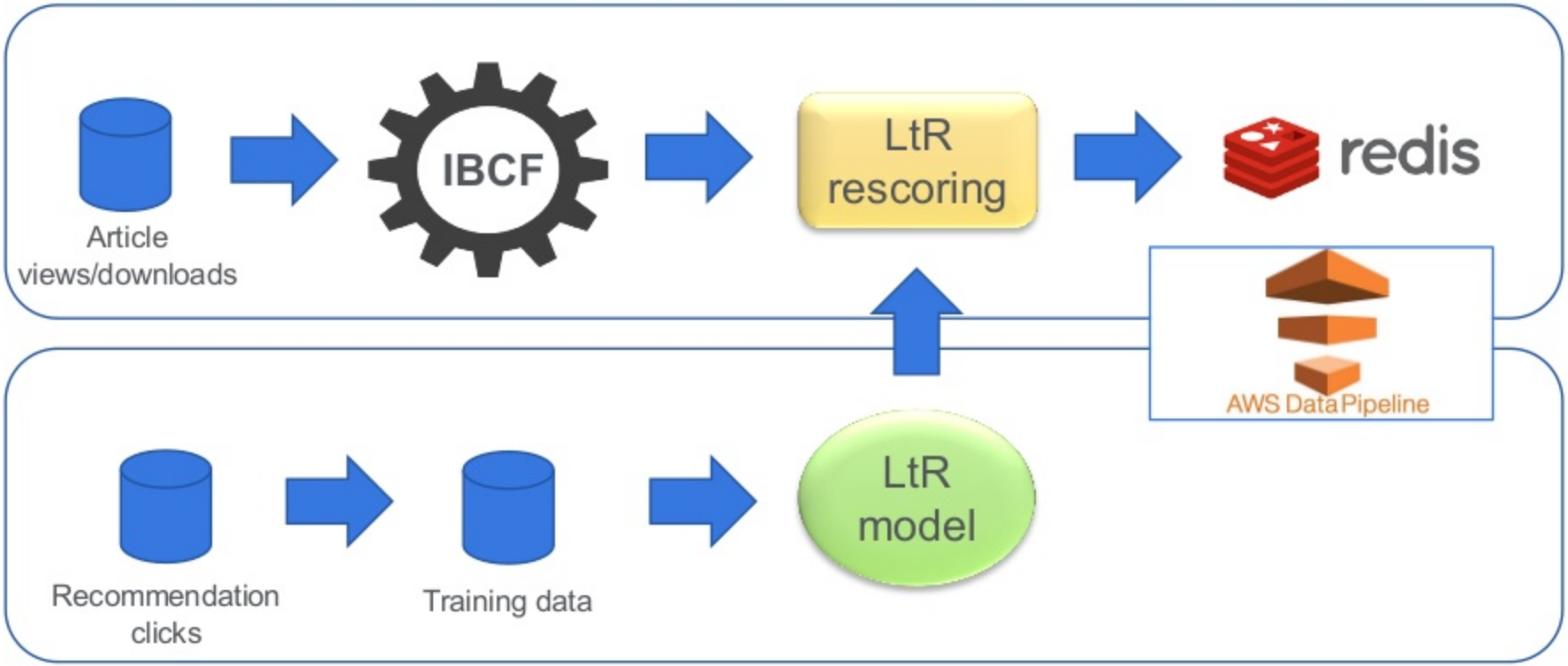
Result: **7-10%** improvement in user engagement

GIF: imgur

Adaptive LtR Model – keep training

Image: pixabay

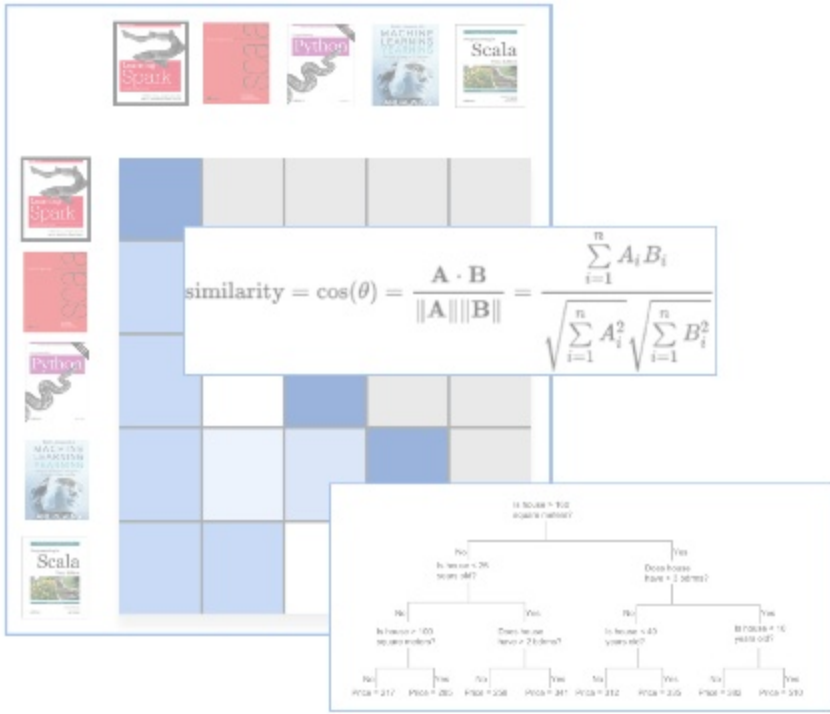
Collaborative Filtering & Learning to Rank



Conclusions

Good recommendations can make a difference

ELSEVIER



Collaborative filtering and Learning to Rank work great!




ELSEVIER

Apache Spark is the foundation for scalable machine learning

image: ocado

We're hiring, come speak to us!

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

ELSEVIER

Thank you

Adam Davidson - a.davidson.1@elsevier.com
Anna Bladzich - a.bladzich@elsevier.com

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