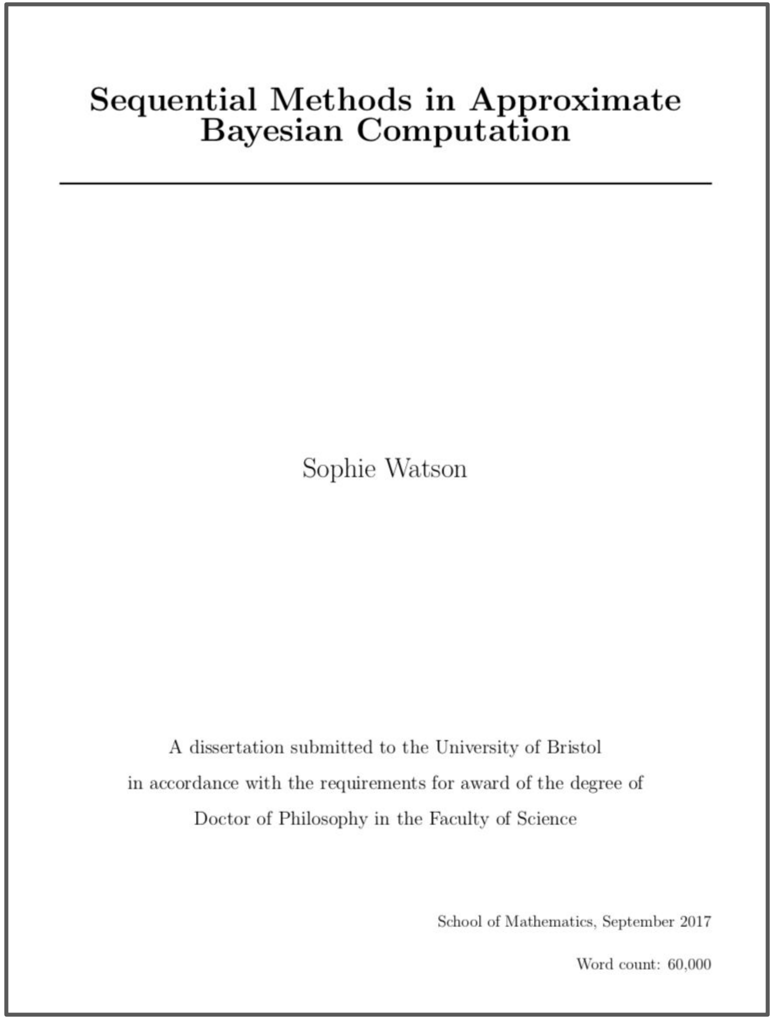
From Research to Production

What they didn’t teach you at grad school

Sophie Watson
sophie@redhat.com



Sequential Methods in Approximate Bayesian Computation

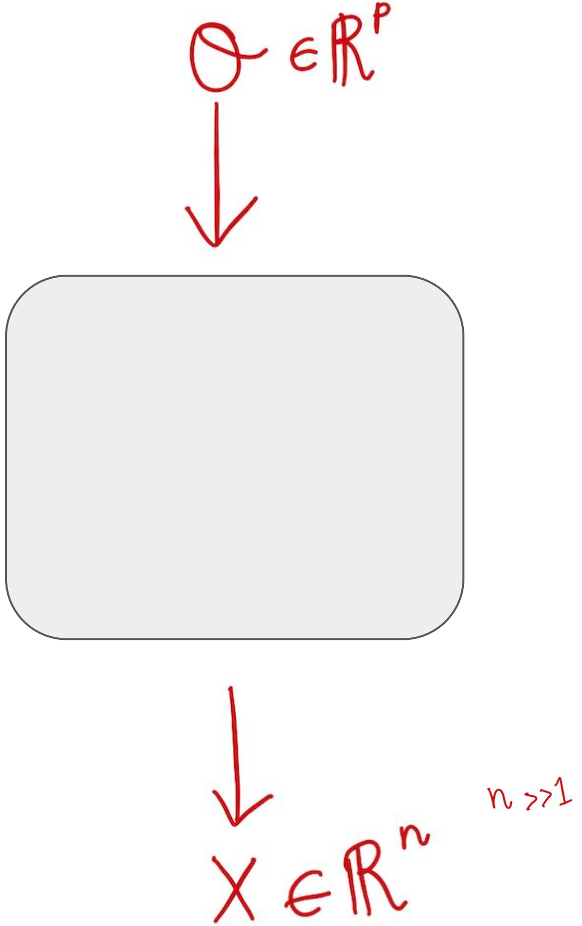
Sophie Watson

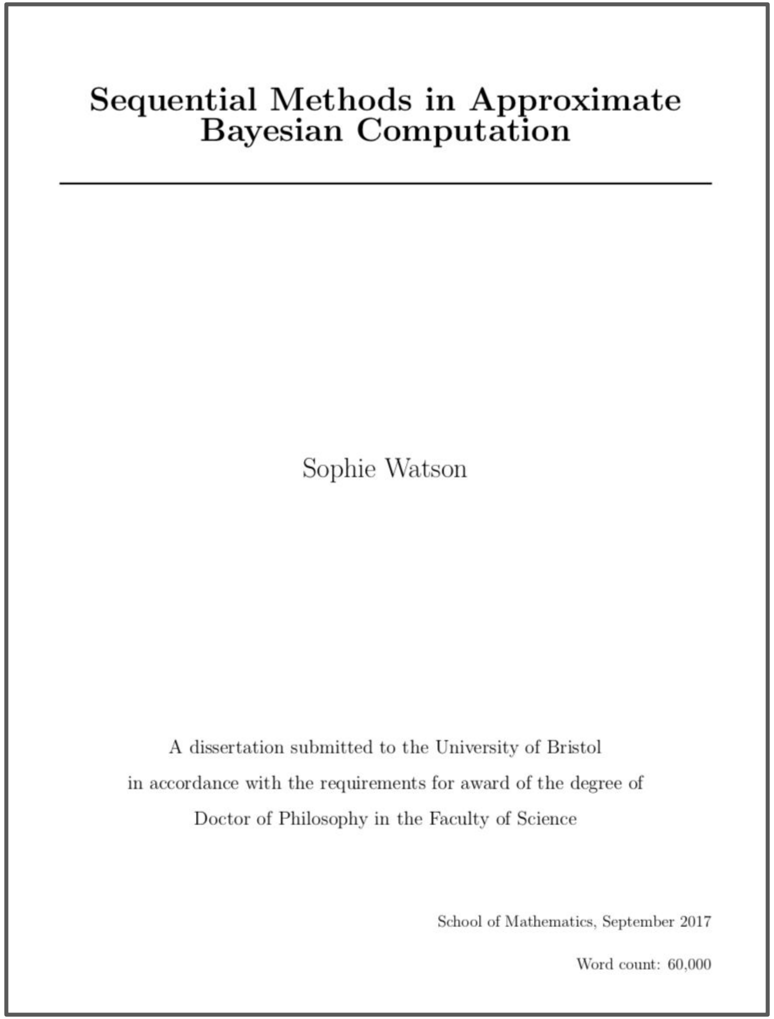
A dissertation submitted to the University of Bristol
in accordance with the requirements for award of the degree of
Doctor of Philosophy in the Faculty of Science

School of Mathematics, September 2017

Word count: 60,000

My Thesis





Sequential Methods in Approximate Bayesian Computation

Sophie Watson

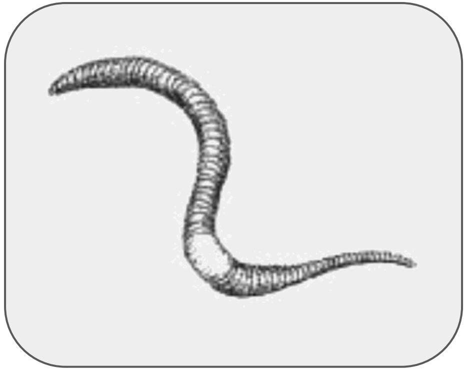
A dissertation submitted to the University of Bristol
in accordance with the requirements for award of the degree of
Doctor of Philosophy in the Faculty of Science

School of Mathematics, September 2017

Word count: 60,000

My Thesis

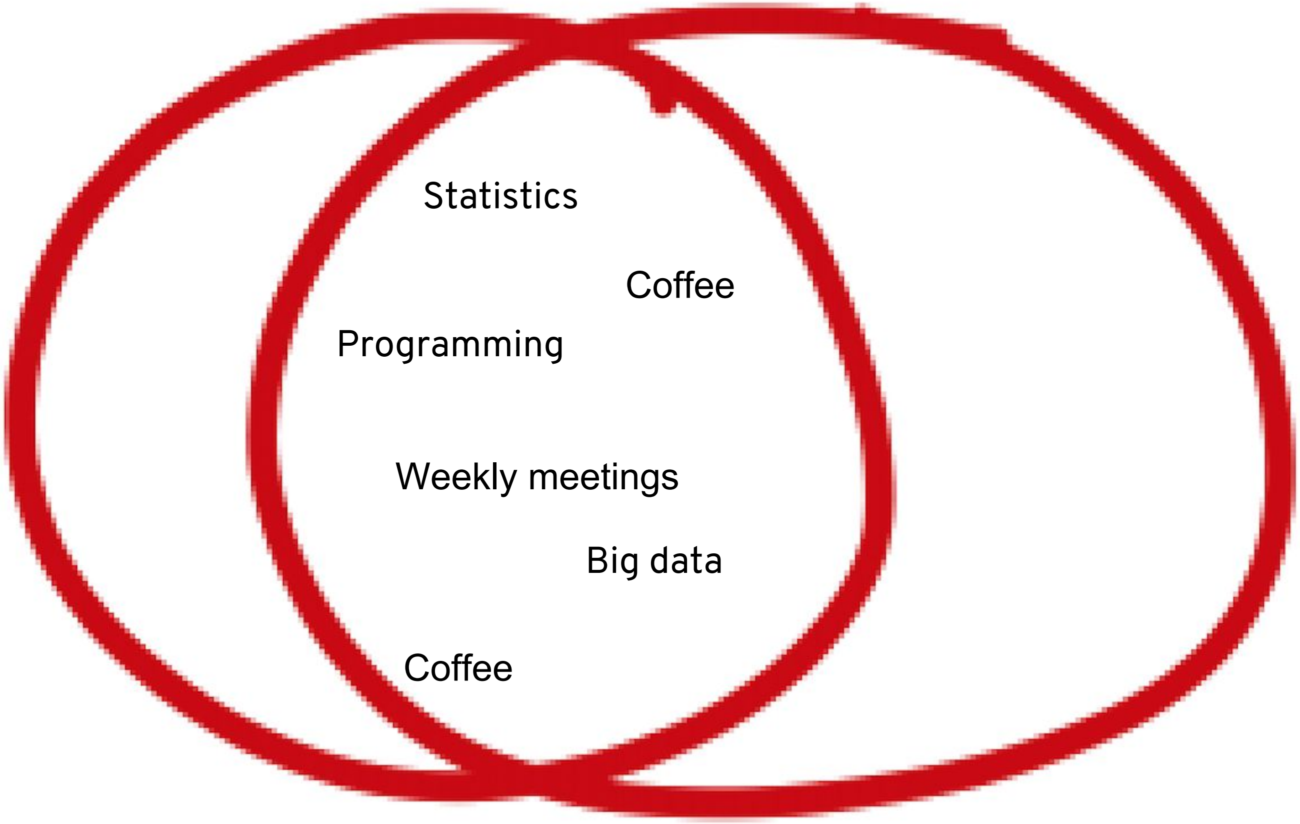
$\theta \in \mathbb{R}^p$



$X \in \mathbb{R}^n$

$n >> 1$

Research Student



Statistics

Coffee

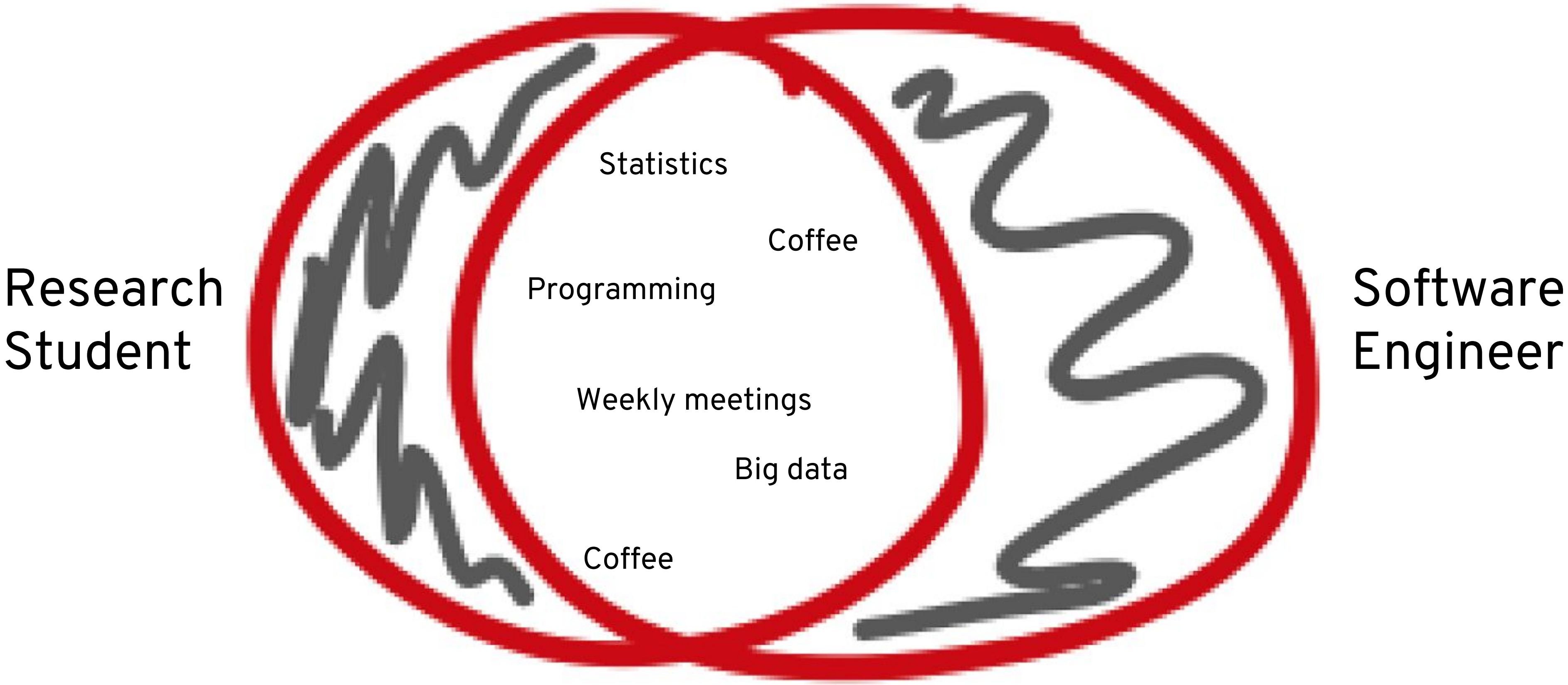
Programming

Weekly meetings

Big data

Coffee

Software Engineer



Research Student

Software Engineer

Statistics

Coffee

Programming

Weekly meetings

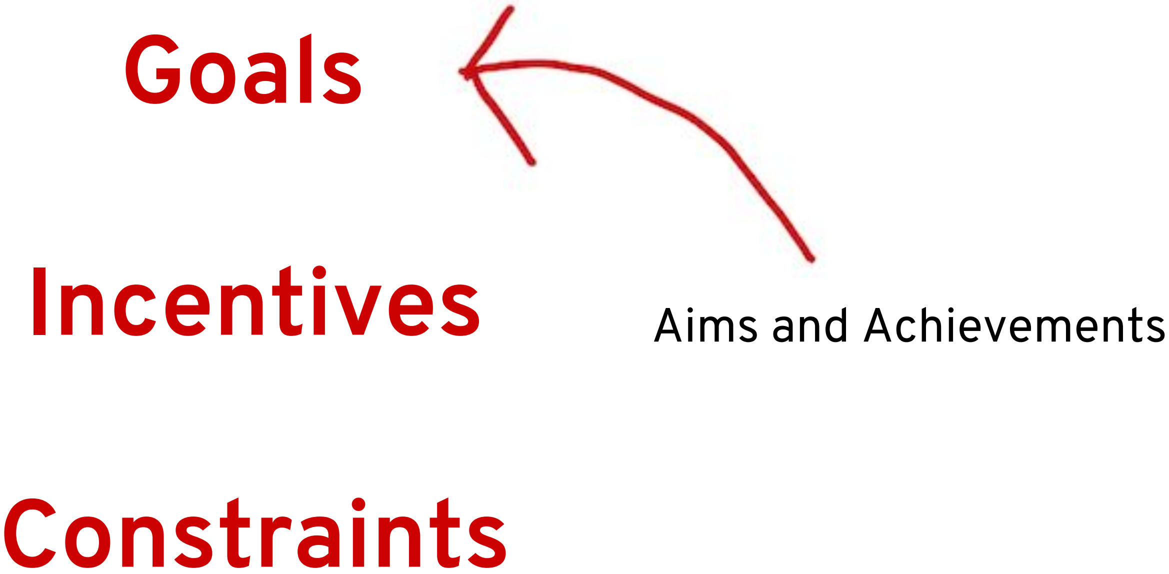
Big data

Coffee

Goals

Incentives

Constraints



Goals

Incentives

Constraints

Aims and Achievements

Goals

Incentives

Constraints

Drive

Goals

Incentives

Constraints

↑

Barriers and Borders

Goals

Aims and Achievements

Incentives

Constraints

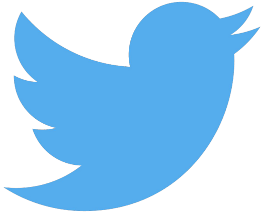
Recommendation Engines

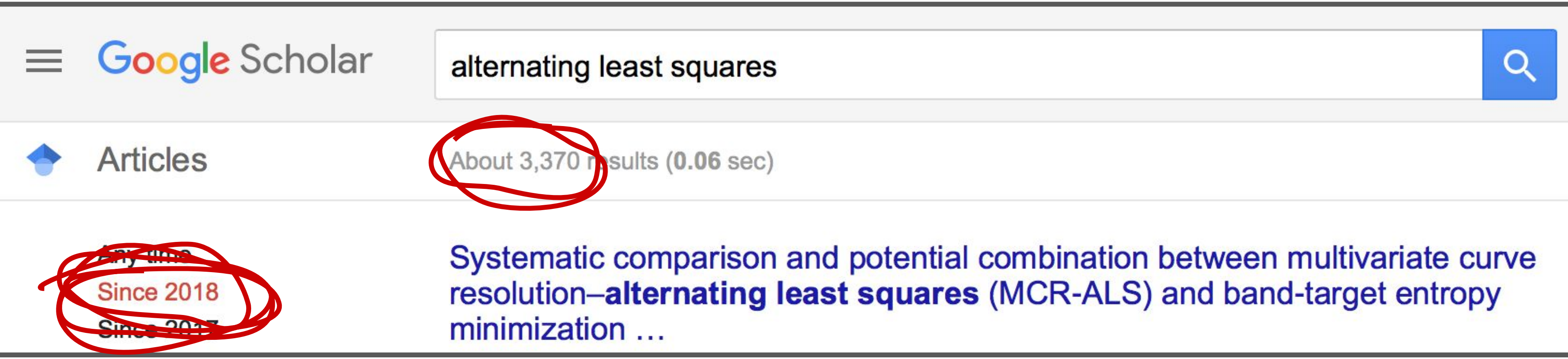
NETFLIX

Expedia®

goodreads

TESCO





Google Scholar

alternating least squares

Articles

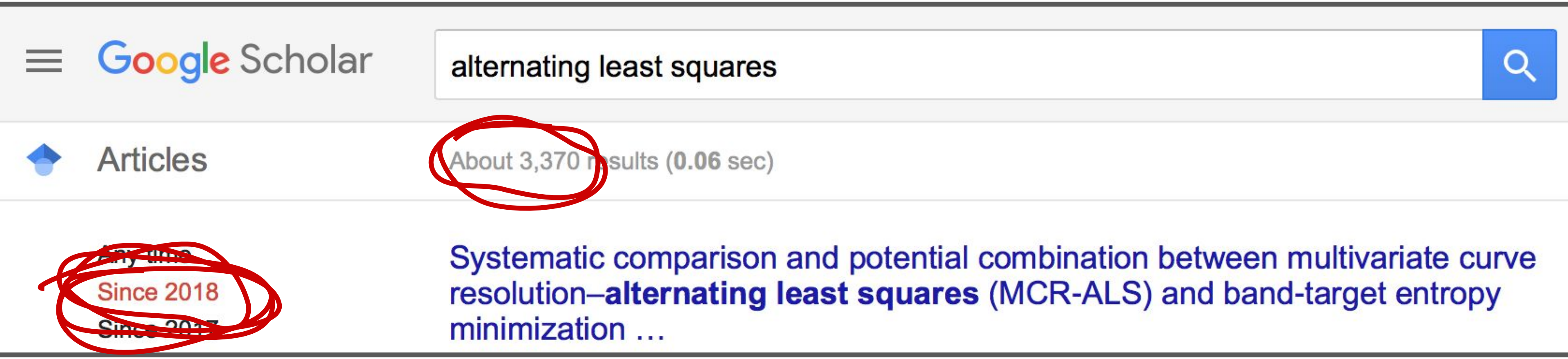
About 3,370 results (0.06 sec)

Any time

Since 2018

Since 2017

Systematic comparison and potential combination between multivariate curve resolution–**alternating least squares** (MCR-ALS) and band-target entropy minimization …



alternating least squares

About 3,370 results (0.06 sec)

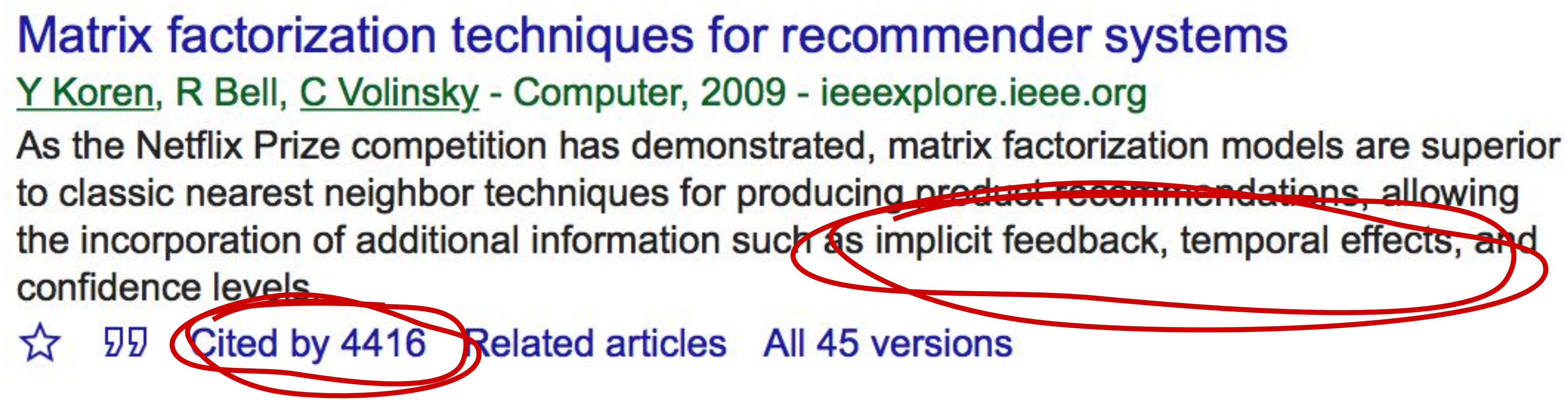
Any time

Since 2018

Since 2017

Systematic comparison and potential combination between multivariate curve resolution–**alternating least squares** (MCR-ALS) and band-target entropy minimization …

code it up → tune params → fiddle around → do better



Matrix factorization techniques for recommender systems

Y Koren, R Bell, C Volinsky - Computer, 2009 - ieeexplore.ieee.org

As the Netflix Prize competition has demonstrated, matrix factorization models are superior to classic nearest neighbor techniques for producing product recommendations, allowing the incorporation of additional information such as implicit feedback, temporal effects, and confidence levels.

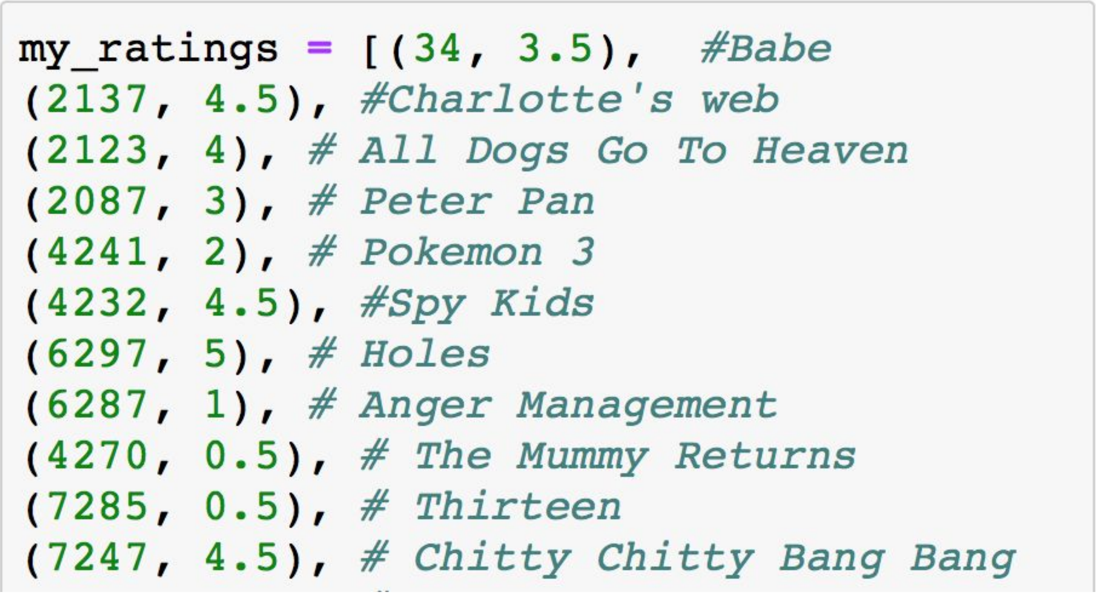
☆ ❞ Cited by 4416 Related articles All 45 versions

Data

- **MovieLens** [1]
- Widely used in recommendation engine research
- 26 million ratings / 45,000 movies / 270,000 users
- Ratings.csv
    - `(userId, movieId, rating, timestamp)`
    - `(100, 200, 3.5, 2010-12-10 12:00:00)`

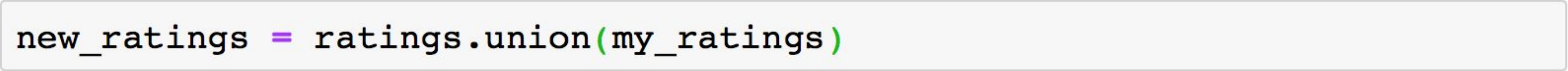
[1] - https://grouplens.org/datasets/movielens/

Modelling

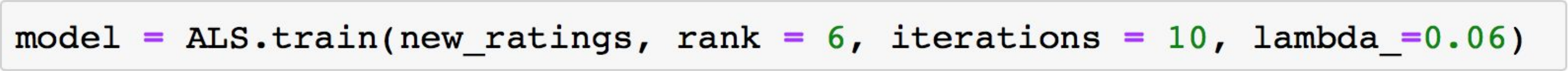


```
my_ratings = [(34, 3.5),  #Babe
(2137, 4.5), #Charlotte's web
(2123, 4), # All Dogs Go To Heaven
(2087, 3), # Peter Pan
(4241, 2), # Pokemon 3
(4232, 4.5), #Spy Kids
(6297, 5), # Holes
(6287, 1), # Anger Management
(4270, 0.5), # The Mummy Returns
(7285, 0.5), # Thirteen
(7247, 4.5), # Chitty Chitty Bang Bang
```

(film id, rating)

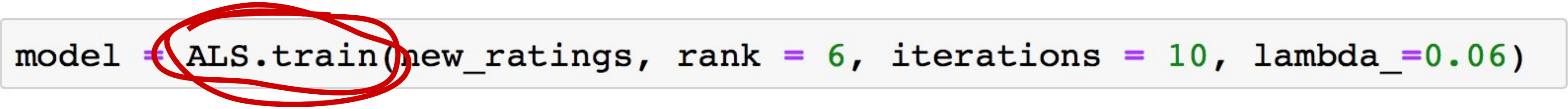


```
new_ratings = ratings.union(my_ratings)
```



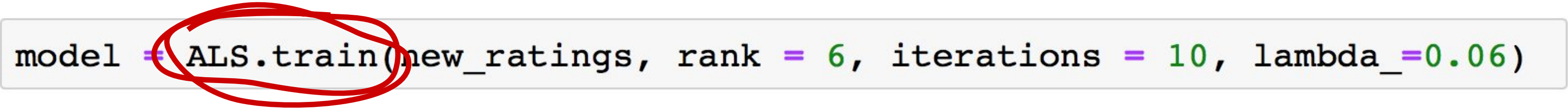
```
model = ALS.train(new_ratings, rank = 6, iterations = 10, lambda_=0.06)
```

Scaling Out

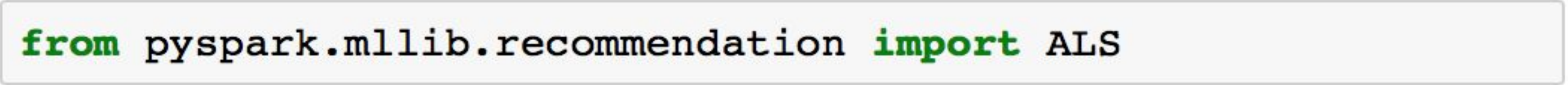


```
model = ALS.train(new_ratings, rank = 6, iterations = 10, lambda_=0.06)
```

Scaling Out

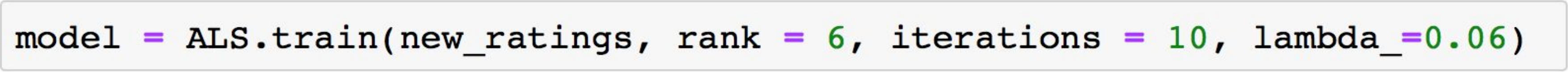


```
model = ALS.train(new_ratings, rank = 6, iterations = 10, lambda_=0.06)
```

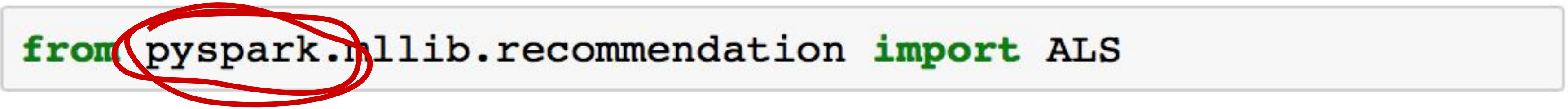


```
from pyspark.mllib.recommendation import ALS
```

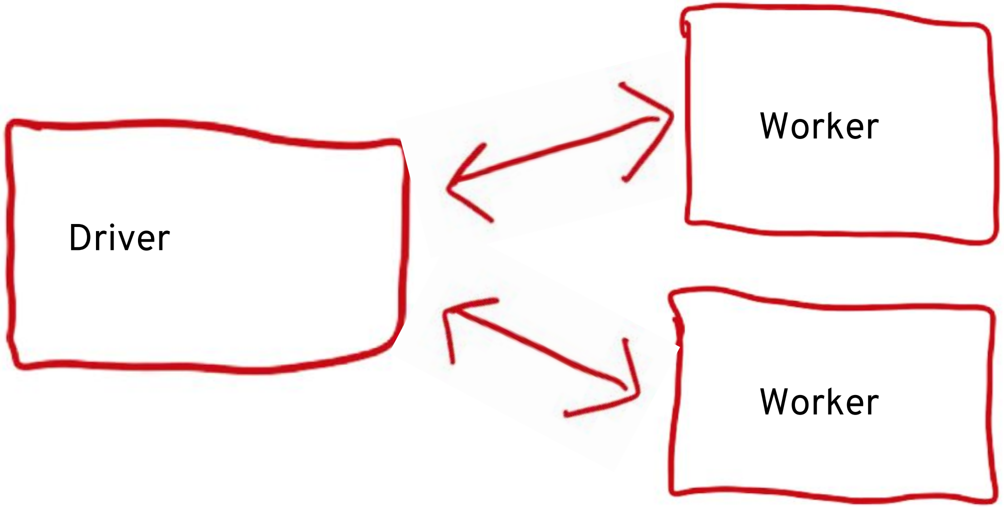
Scaling Out



```
model = ALS.train(new_ratings, rank = 6, iterations = 10, lambda_=0.06)
```



```
from pyspark.mllib.recommendation import ALS
```

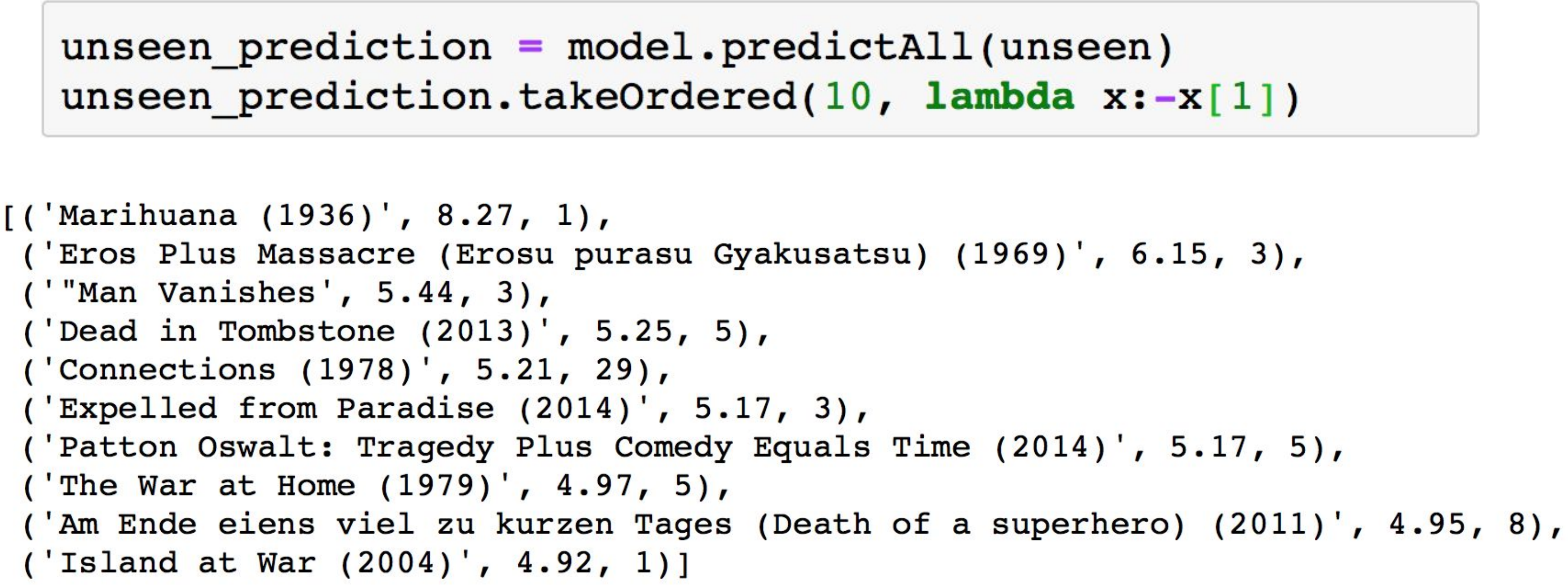


Driver

Worker

Worker

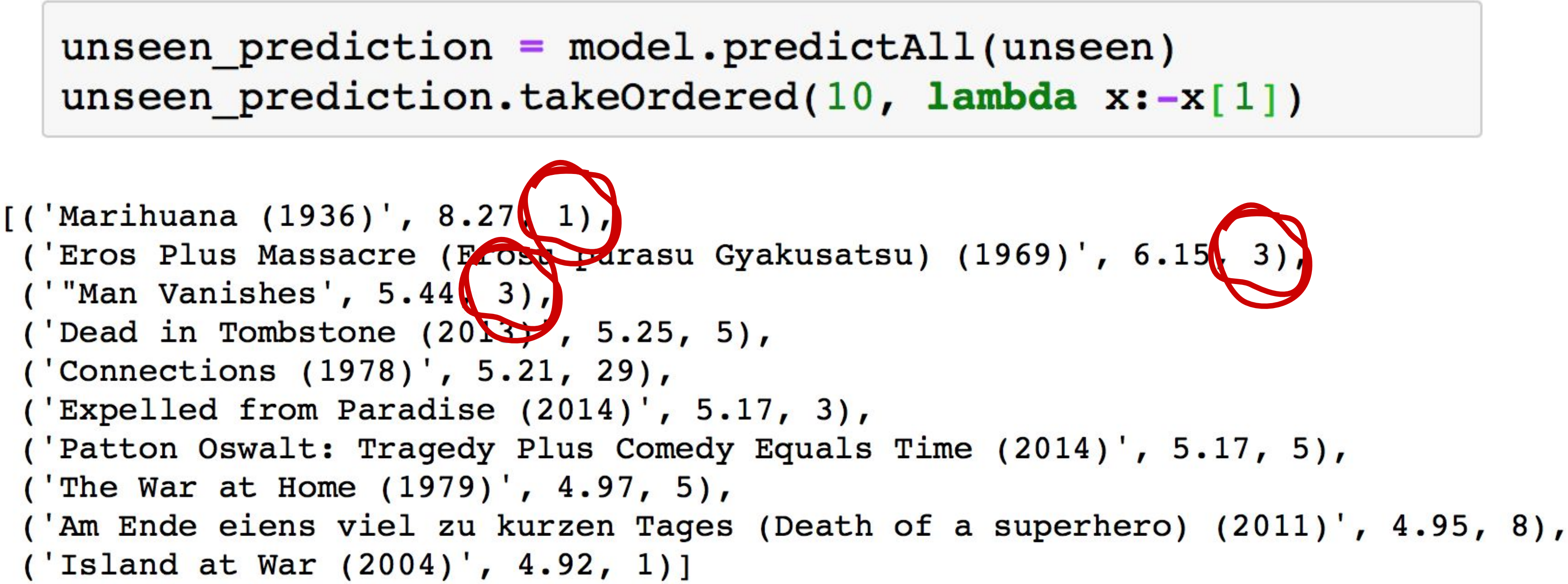
Prediction



```
unseen_prediction = model.predictAll(unseen)
unseen_prediction.takeOrdered(10, lambda x:-x[1])
```

```
[('Marihuana (1936)', 8.27, 1),
 ('Eros Plus Massacre (Erosu purasu Gyakusatsu) (1969)', 6.15, 3),
 ('"Man Vanishes', 5.44, 3),
 ('Dead in Tombstone (2013)', 5.25, 5),
 ('Connections (1978)', 5.21, 29),
 ('Expelled from Paradise (2014)', 5.17, 3),
 ('Patton Oswalt: Tragedy Plus Comedy Equals Time (2014)', 5.17, 5),
 ('The War at Home (1979)', 4.97, 5),
 ('Am Ende eiens viel zu kurzen Tages (Death of a superhero) (2011)', 4.95, 8),
 ('Island at War (2004)', 4.92, 1)]
```

Prediction



```
unseen_prediction = model.predictAll(unseen)
unseen_prediction.takeOrdered(10, lambda x:-x[1])
```

```
[('Marihuana (1936)', 8.27, 1),
 ('Eros Plus Massacre (Erosu purasu Gyakusatsu) (1969)', 6.15, 3),
 ('"Man Vanishes', 5.44, 3),
 ('Dead in Tombstone (2013)', 5.25, 5),
 ('Connections (1978)', 5.21, 29),
 ('Expelled from Paradise (2014)', 5.17, 3),
 ('Patton Oswalt: Tragedy Plus Comedy Equals Time (2014)', 5.17, 5),
 ('The War at Home (1979)', 4.97, 5),
 ('Am Ende eiens viel zu kurzen Tages (Death of a superhero) (2011)', 4.95, 8),
 ('Island at War (2004)', 4.92, 1)]
```

Industry Goals

1. Build a recommendation engine that works.

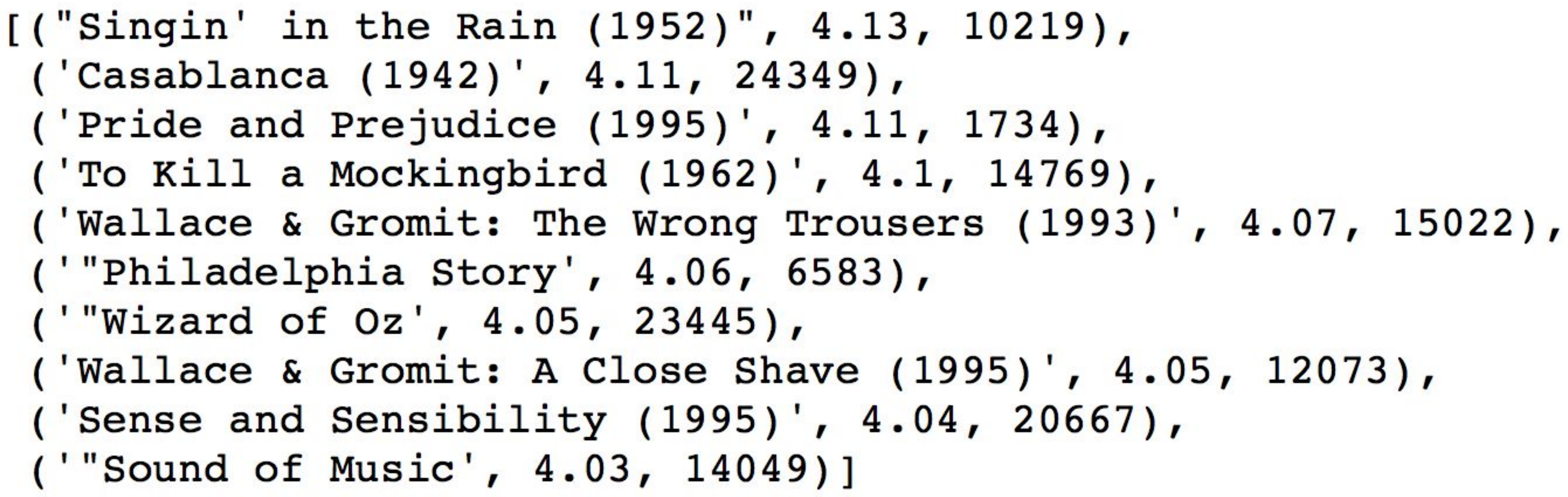
Scales out

fast

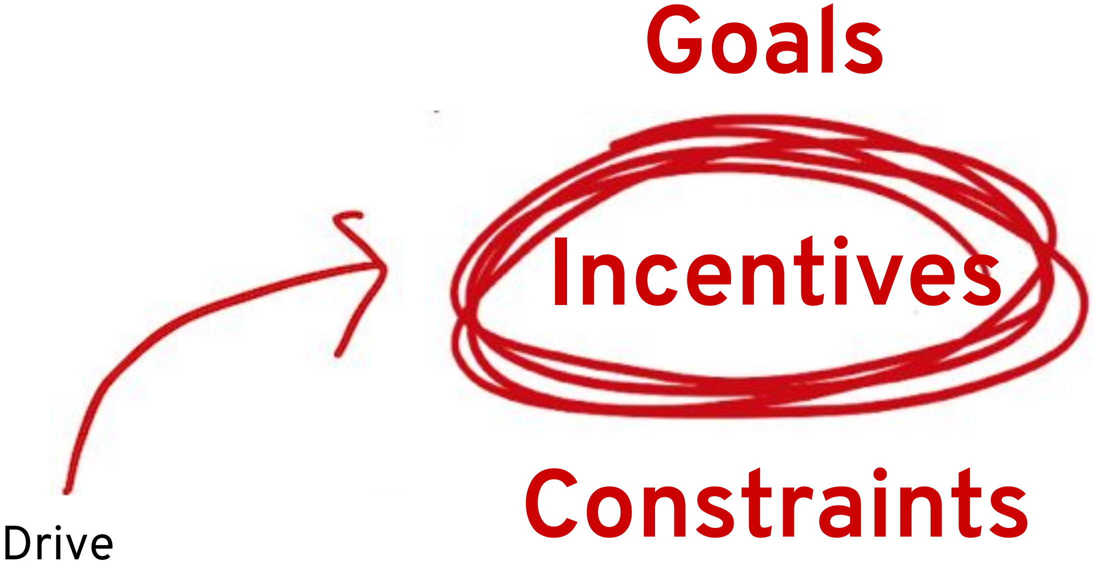
gives sensible recommendations

Industry Goals

1. Build a recommendation engine that works.



```
[("Singin' in the Rain (1952)", 4.13, 10219),
 ('Casablanca (1942)', 4.11, 24349),
 ('Pride and Prejudice (1995)', 4.11, 1734),
 ('To Kill a Mockingbird (1962)', 4.1, 14769),
 ('Wallace & Gromit: The Wrong Trousers (1993)', 4.07, 15022),
 ('"Philadelphia Story', 4.06, 6583),
 ('"Wizard of Oz', 4.05, 23445),
 ('Wallace & Gromit: A Close Shave (1995)', 4.05, 12073),
 ('Sense and Sensibility (1995)', 4.04, 20667),
 ('"Sound of Music', 4.03, 14049)]
```



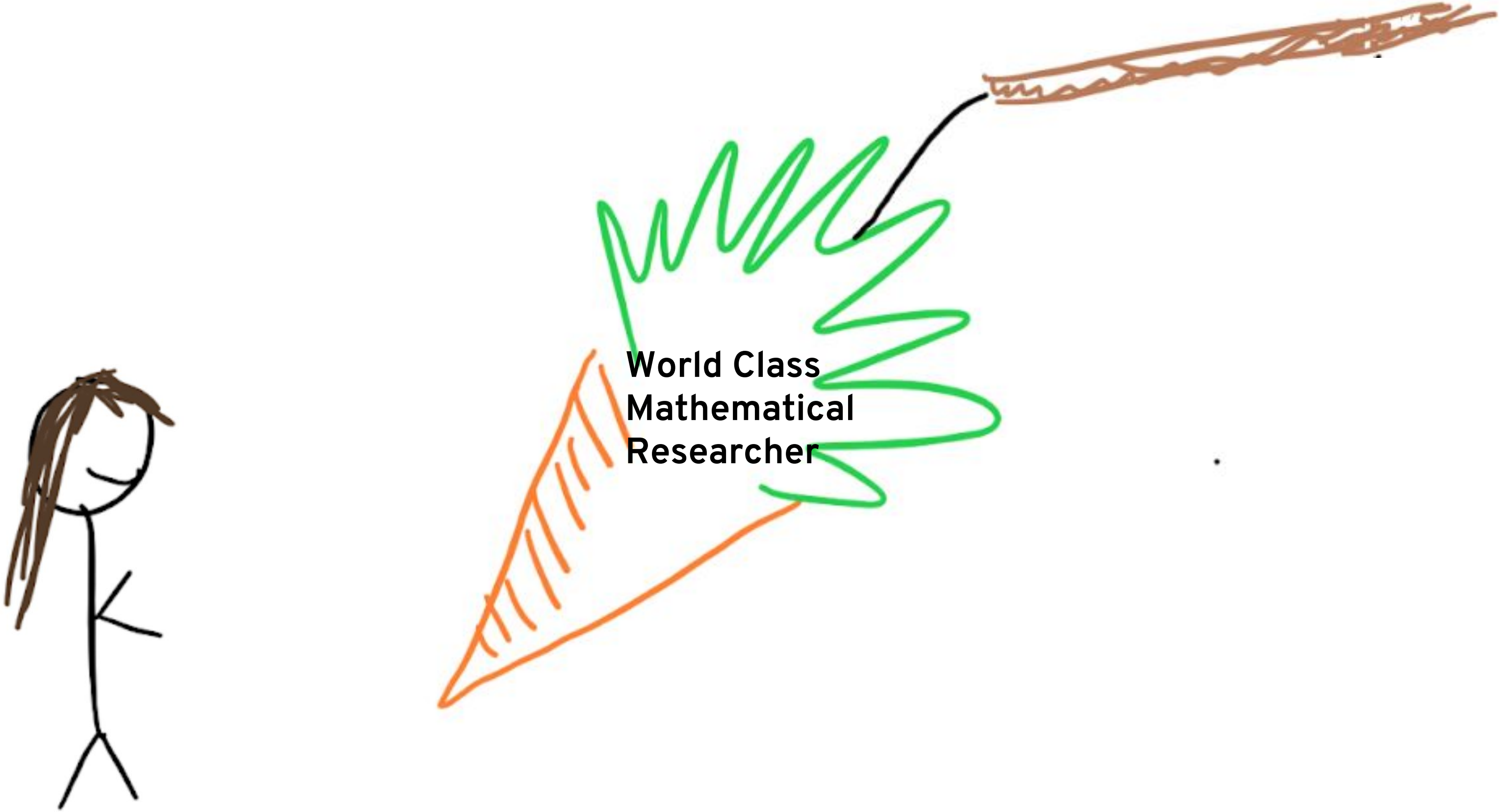
Goals

Incentives

Constraints

Drive

Research Incentives

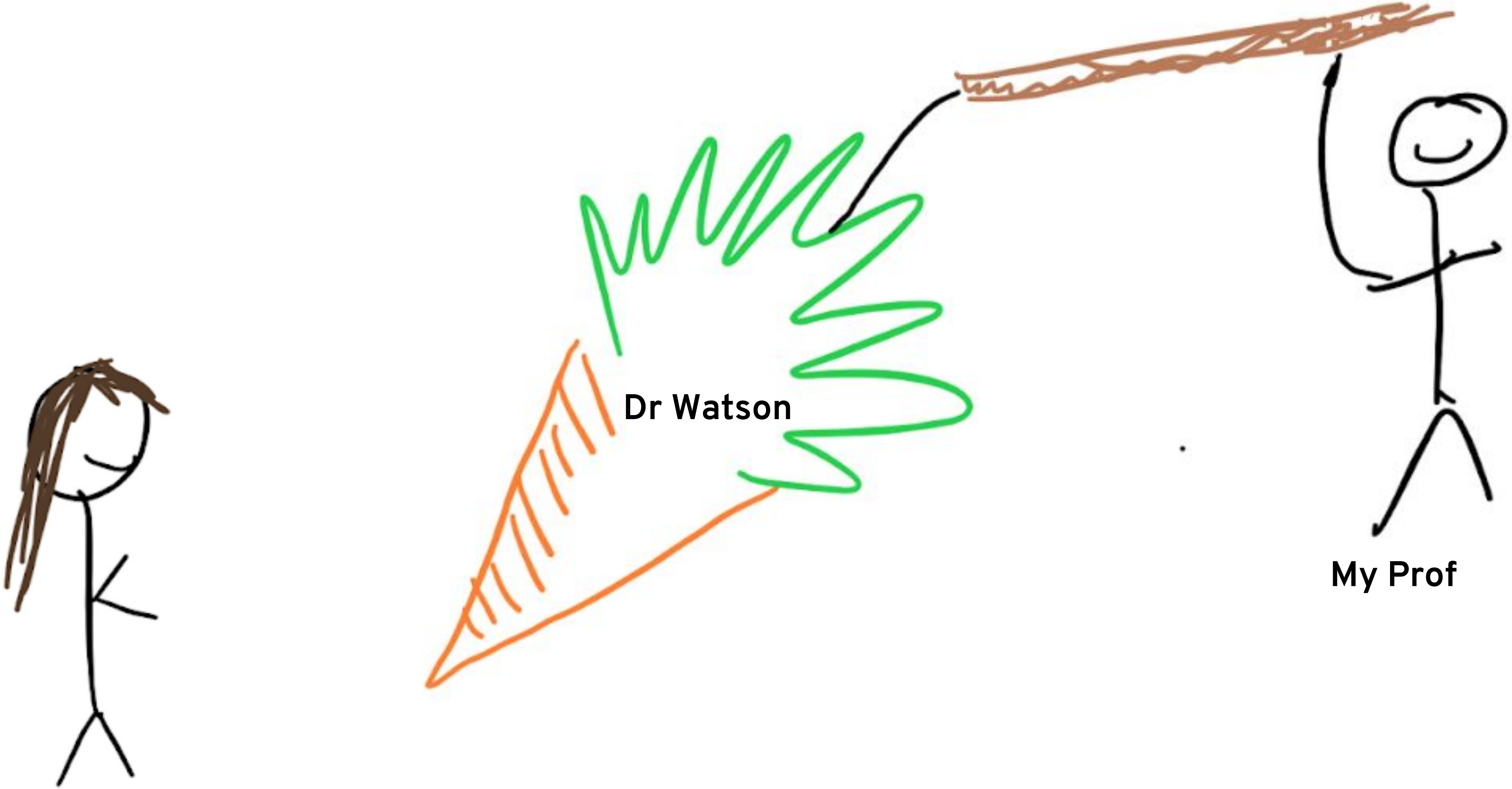


Research Incentives

‘Good’ PhD Student
A couple
of publications

Research Incentives

Research Incentives



Industry Incentives

$$$

Industry Incentives

Team

$$$

Company

Goals

Incentives

Constraints

Barriers and Boundaries

Constraints

Constraints

Research = Strict

Constraints

not so

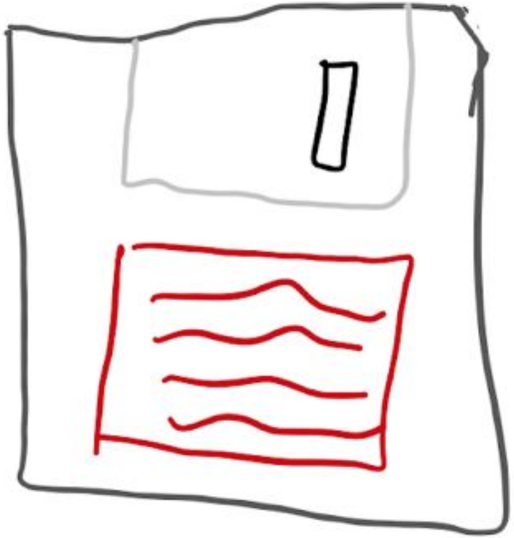
Research = Strict

Constraints

Research = not so Strict

Production = Strict

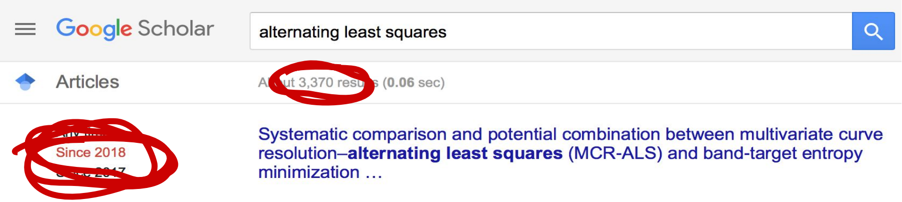
Constraints



Goals

Incentives

Constraints

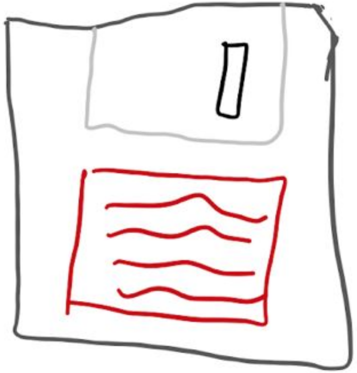


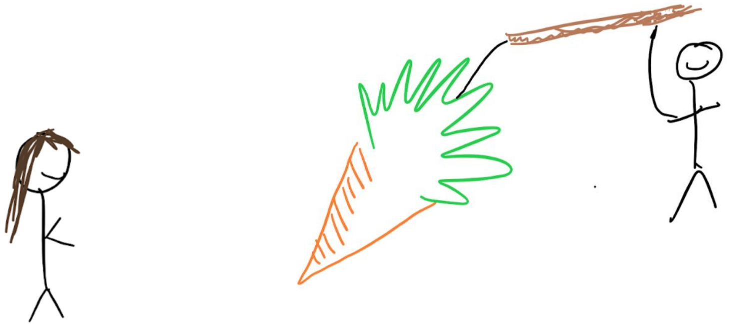
alternating least squares

About 3,370 results (0.06 sec)

Since 2018

Systematic comparison and potential combination between multivariate curve resolution–**alternating least squares** (MCR-ALS) and band-target entropy minimization …





sophie@redhat.com
@sophwats

radanalytics.io/applications/project-jiminy