# Literature Survey & Where to find ideas for a baseline

“Enterprise Data Science” (M-DS, WS19/20)

Mahan Ghashghaie, Rupali Sharma, Florian Becker

# Outline

# Intro

1. Baselines
    a. What is it? Why are they necessary?
    b. Baselines in Data Science
2. Papers
    a. What and where to find them?
    b. Example query
    c. How do we know they are any good?
3. Conferences
    a. How can i find them?
    b. Hirsch Index
    c. NIPS by Example

# Baselines

# Why do we need baselines?

- Each project is unique: setting, available data, problem to tackle, available features et cetera (e.g. your Master Thesis)
- Lots of models and algorithms
- Just apply some models and see what they'll return
  1) Time/resource consuming
  2) How to evaluate the results?
  3) How do we know, that they are actually adding value (e.g. improve our accuracy)?

> **an initial value for benchmarking could help**

# Baseline - a definition

- Techopedia defines a baseline as *“ […] the visible measure or progress and often marks milestones. In other words, a baseline serves as a crucial input for performing analysis to evaluate current performance against anticipated levels for the specific tasks in an established time-phase. […]”*

  (https://www.techopedia.com/definition/6148/baseline)

**How to translate this now to the DS domain?**

# Baselines in DS

- Main idea still holds: we need a reference for comparison, as simple as possible

- Known task
- Known data set
- Come up with better model

> compare performance model of on well-known data set to others

- New project
- Unknown data set
- Find a baseline algorithm/score or build one yourself

> baseline performance for improvement comparison

# Baseline data sets

- Some examples for widely established data sets:
  - Recommender systems using the Netflix Prize dataset
  - Speech recognition using the TIMIT corpus
  - Natural language understanding using the GLUE Benchmark
  - Image Recognition with MNIST

(https://towardsdatascience.com/beating-state-of-the-art-by-tuning-baselines-74ec6ad2cd59)

- Reproducibility
- https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research ( list overview)
- http://mlr.cs.umass.edu/ml/ (UCI ML Repository)

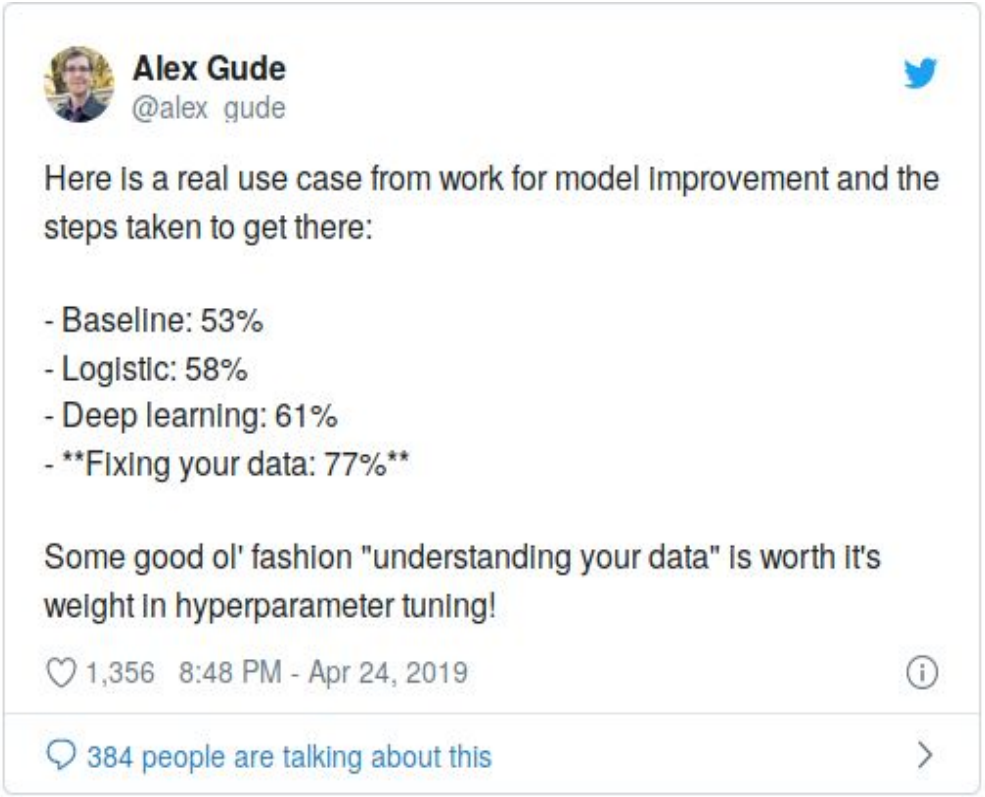
# Baseline models

- Some examples for baseline models:
  - Regression with Linear Regression/GLM
  - Classification with Logistic Regression
  - NLP using Naive Bayes
  - Neural Networks: CNN
- Reproducibility
- https://paperswithcode.com/sota (more on that later)
- https://en.wikipedia.org/wiki/Outline_of_machine_learning#Machine_learning_algorithms (list of algorithms for different purposes)

# Baselines - how to start?

- Keep in mind to start with a stupid model first
- Some example baselines you could fit yourself:
  - Regression: Mean, Median
  - Classification: ZeroR (Mode)
  - “Human-based”
- Then check if you can beat your baseline result with a more complex, sophisticated model
- Iterate by optimizing your methods, data, parameters et cetera and see if you are able to improve
  - If not you may should overthink your problem/approach

# Baselines




*(via https://www.jeremyjordan.me/ml-projects-guide/)*

# Baselines - what comes next?

- Your baseline is set up, you have a score for comparison and you know the problem
  - You are now ready to use a more advanced model
  - Use a well-studied baseline as approximation

> But where to find them? And how do we know, that there are reliable, accepted and proven?

**Papers and Conferences as valuable sources!**

# Papers

# What and where to find them?

- Tangible source: Scientific research papers

- Where to find them?
  - Semantic Scholar: https://www.semanticscholar.org/
  - dblp : https://dblp.uni-trier.de/
  - Google Scholar: https://scholar.google.com/
  - Papers with code: https://paperswithcode.com/

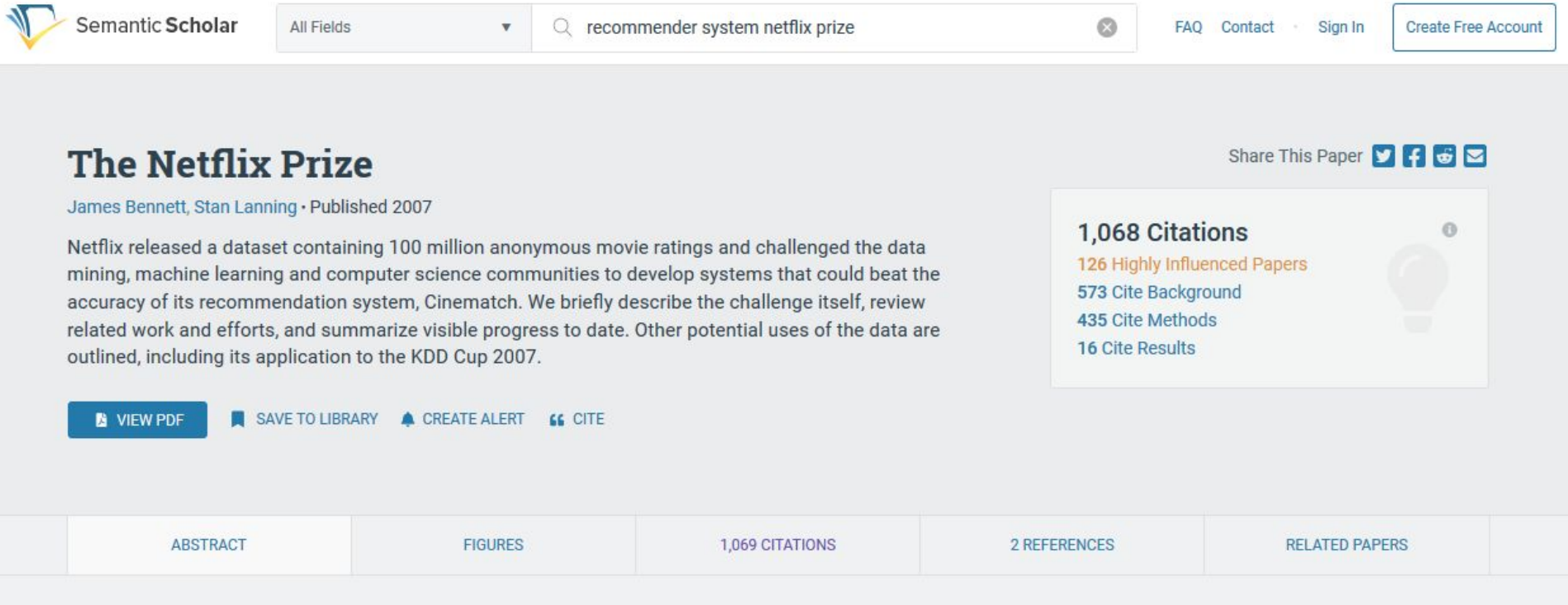
Example query:
recommender system netflix prize

# What is the Netflix Prize?

- Open competition (2006-2009)
  - Training (100,000,000 ratings)
  - Qualifying set (2,800,000 ratings)
    - Test (determine winners)
    - Quiz (calculate leaderboard scores)
- Best collaborative filtering algorithm
- Predict user ratings for films based on previous ratings without any other information about the users or films
- Improvement over Netflix's own recommendation system, Cinematch

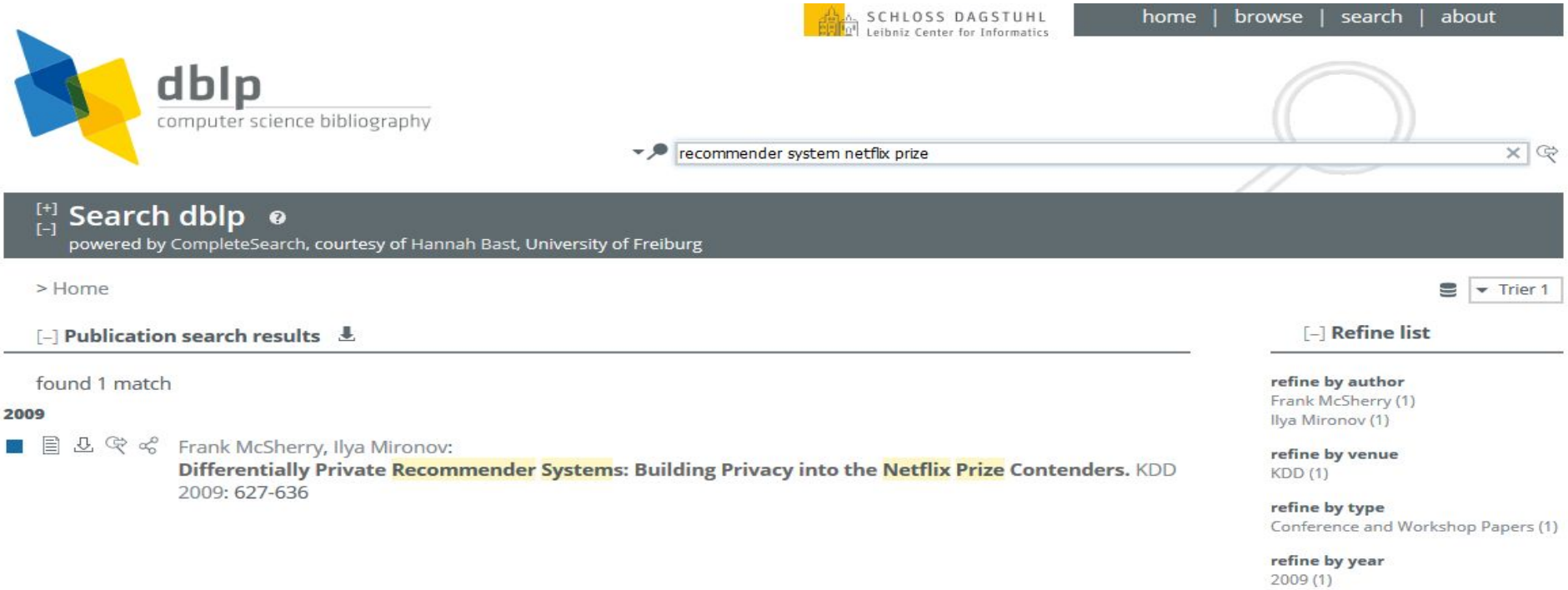
- Prizes awarded for improvements in the RMSE
  - 1st (2006): 1.06%
  - Last (2009): 10.09%

# Semantic Scholar



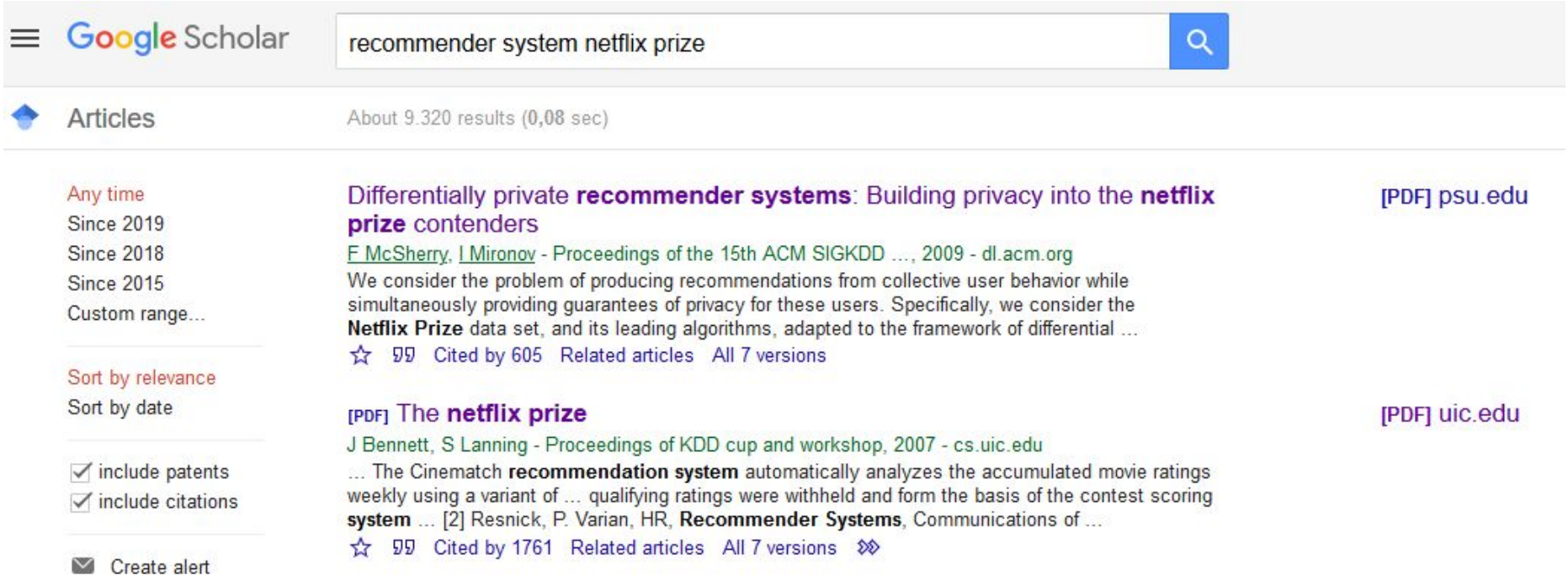

Source: https://www.semanticscholar.org/paper/The-Netflix-Prize-Bennett-Lanning/31af4b8793e93fd35e89569ccd663ae8777f0072

# dblp



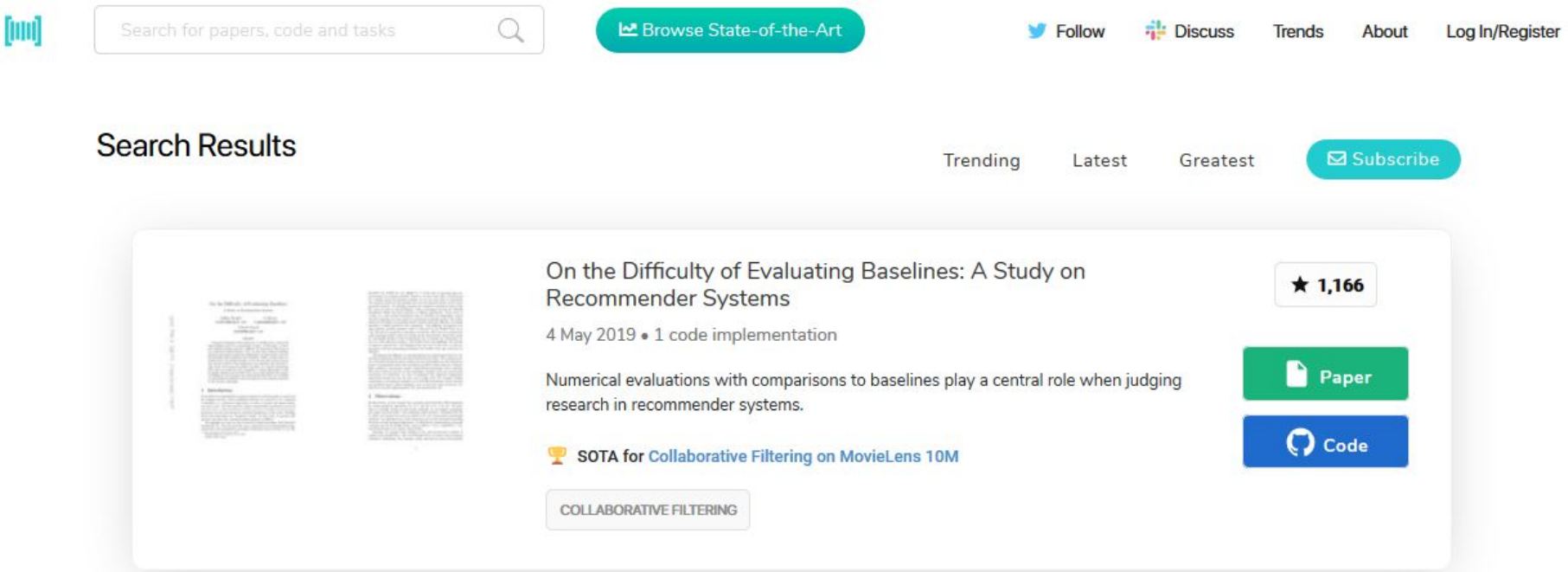

Source: https://dblp.uni-trier.de/search?q=recommender%20system%20netflix%20prize

# Google Scholar




Source: https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=recommender+system+netflix+prize&btnG=&oq=recommender+system+netflix+pri

# Papers with code




Source: https://paperswithcode.com/search?q_meta=&q=recommender+system+netflix+prize

# How do we know they are any good?

*Just because a modelling technique was proposed more recently doesn't mean it's necessarily going to outperform an older method (even if the results in the paper suggest that it can)*

*~ Rachael Tatman*

- # citations
- Reference list
- Authors and papers with multiple references
- Overlaps between authors, journals, conferences, etc (good thing)
- Journals and conferences the papers were published in

# Conferences

# Conferences

- Have a focus on subareas in computer science
- How to find a suitable one
  - Online conference aggregators
    - https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng
    - https://www.scimagojr.com
    - http://www.guide2research.com
    - https://aminer.org/ranks/conf
  - University curated commendation list
    - http://www.core.edu.au/conference-portal
  - Senior colleagues in the same field

# Hirsch-Index

- Attempts to measure productivity and citation impact of the publications  of an author
- Calculated by:
  - Sorting Publications from highest cited to lowest
  - Choose the number of citations such that the most publications fullfill:
    - n_paper_cited <= citation_threshold
- Example:
  - https://scholar.google.de/citations?user=am2ohp0AAAAJ&hl=de

# Example conference: Neural Information Processing Systems ( NIPS )

# NIPS

## Conference Information

| | |
|---|---|
| Submission Deadline | Thursday 23 May 2019 |
| Conference Dates | Dec 10, 2019 - Dec 12, 2019 |
| Conference Address | Vancouver, Canada |
| Conference & Submission Link | https://nips.cc |

Proceedings indexed by : Neural Information Processing Systems Foundation

Conference Organizers : ( *Deadline extended ? Click here to edit* )

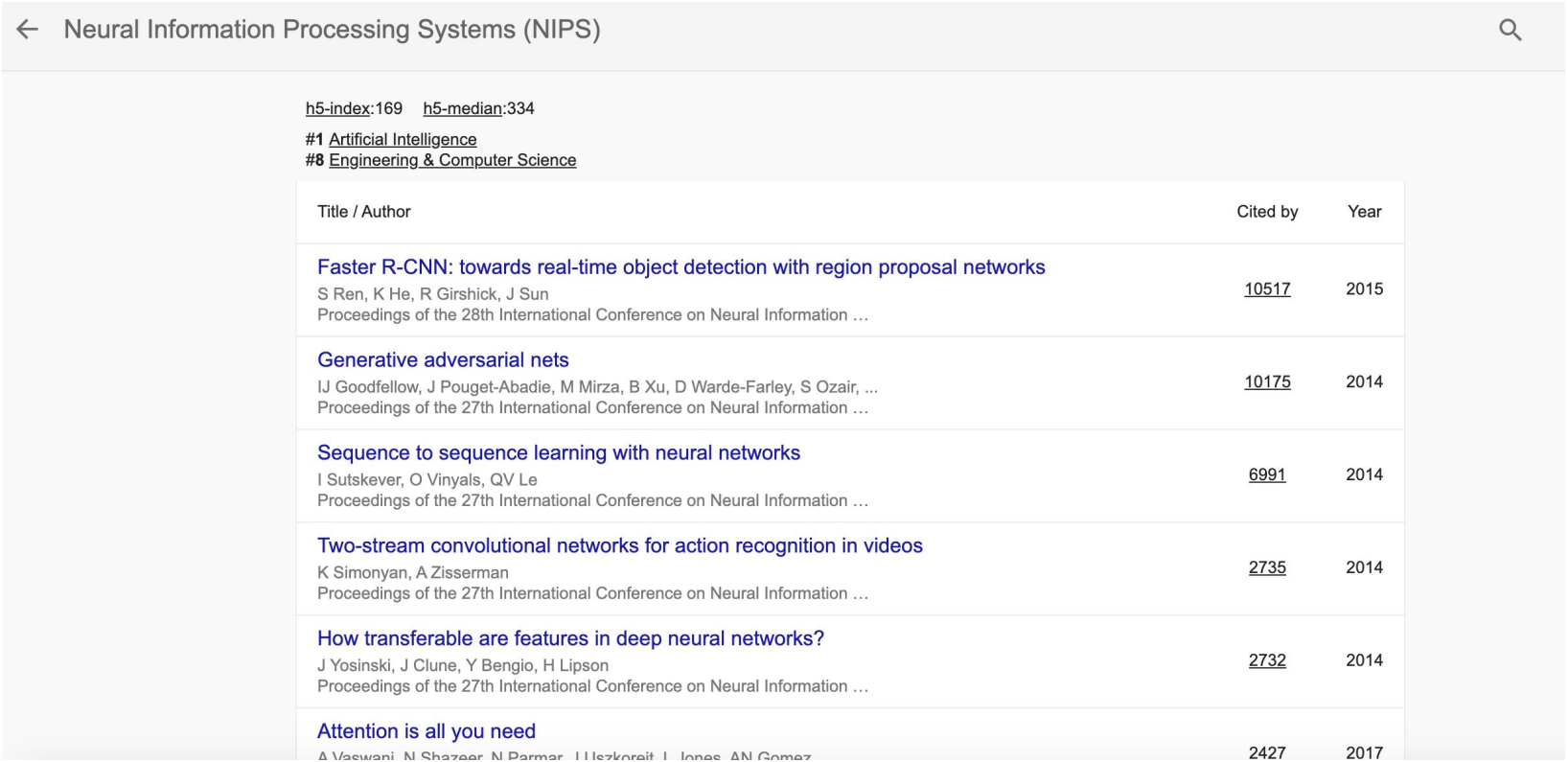
## Conference Ranking & Metrics (This is a TOP Conference)

| | |
|---|---|
| Guide2Research Overall Ranking: | 2 |
| *Category Rankings* | |
| *Machine Learning & Arti. Intelligence* | 2 |
| *Signal Processing* | 2 |
| *Computational Theory and Mathematics* | 1 |
| Number of Editions: | 33 ( since 1987 ) |
| CORE 2018 Rating: | A* |
| DBPL Proceedings | |

Google Scholar H5-index: 169

Source: http://www.guide2research.com/conference/nips-2019

# NIPS



Neural Information Processing Systems (NIPS)

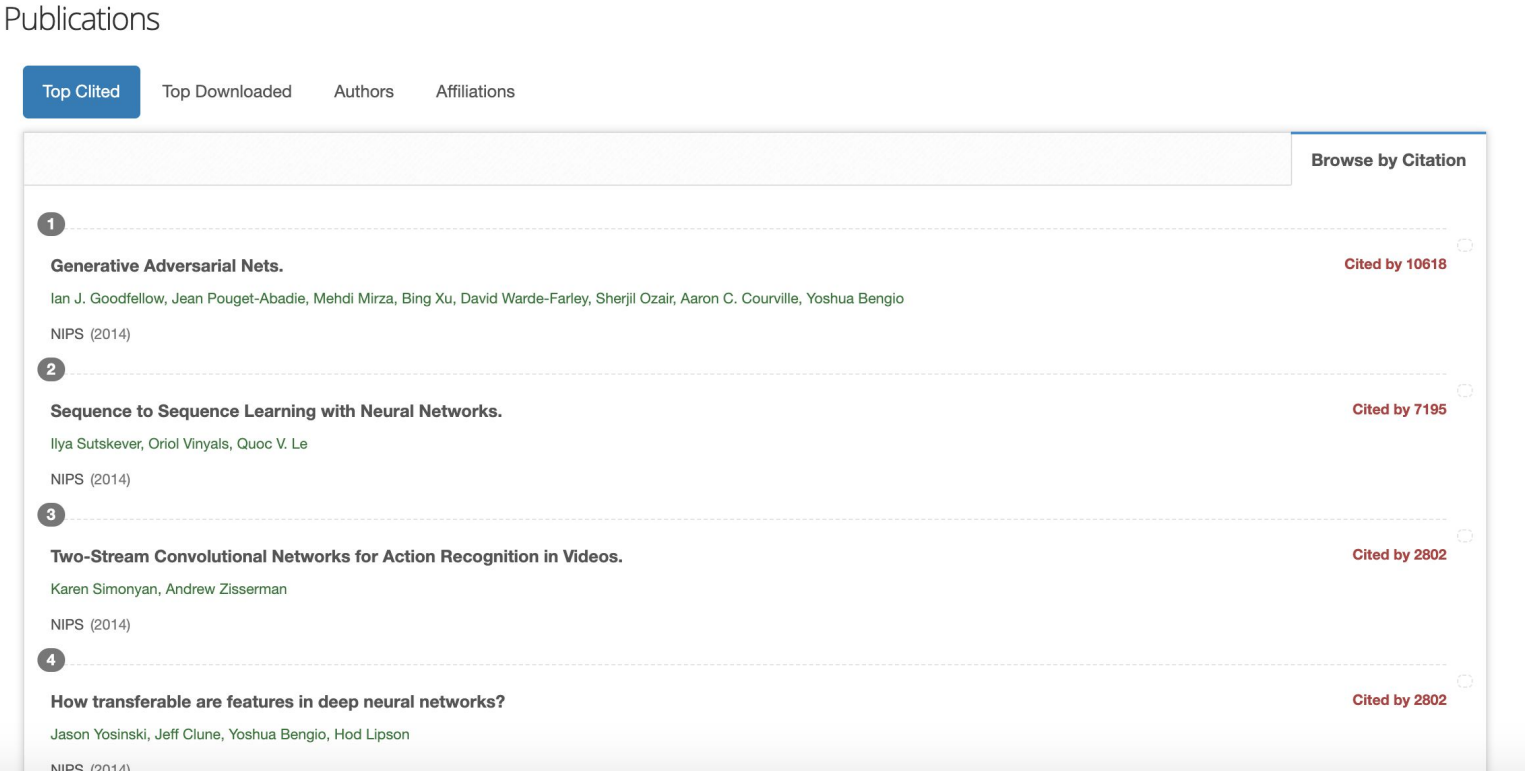
h5-index:169 h5-median:334

#1 Artificial Intelligence
#8 Engineering & Computer Science

| Title / Author | Cited by | Year |
|---|---|---|
| Faster R-CNN: towards real-time object detection with region proposal networks<br>S Ren, K He, R Girshick, J Sun<br>Proceedings of the 28th International Conference on Neural Information … | 10517 | 2015 |
| Generative adversarial nets<br>IJ Goodfellow, J Pouget-Abadie, M Mirza, B Xu, D Warde-Farley, S Ozair, ...<br>Proceedings of the 27th International Conference on Neural Information … | 10175 | 2014 |
| Sequence to sequence learning with neural networks<br>I Sutskever, O Vinyals, QV Le<br>Proceedings of the 27th International Conference on Neural Information … | 6991 | 2014 |
| Two-stream convolutional networks for action recognition in videos<br>K Simonyan, A Zisserman<br>Proceedings of the 27th International Conference on Neural Information … | 2735 | 2014 |
| How transferable are features in deep neural networks?<br>J Yosinski, J Clune, Y Bengio, H Lipson<br>Proceedings of the 27th International Conference on Neural Information … | 2732 | 2014 |
| Attention is all you need<br>A Vaswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, AN Gomez | 2427 | 2017 |

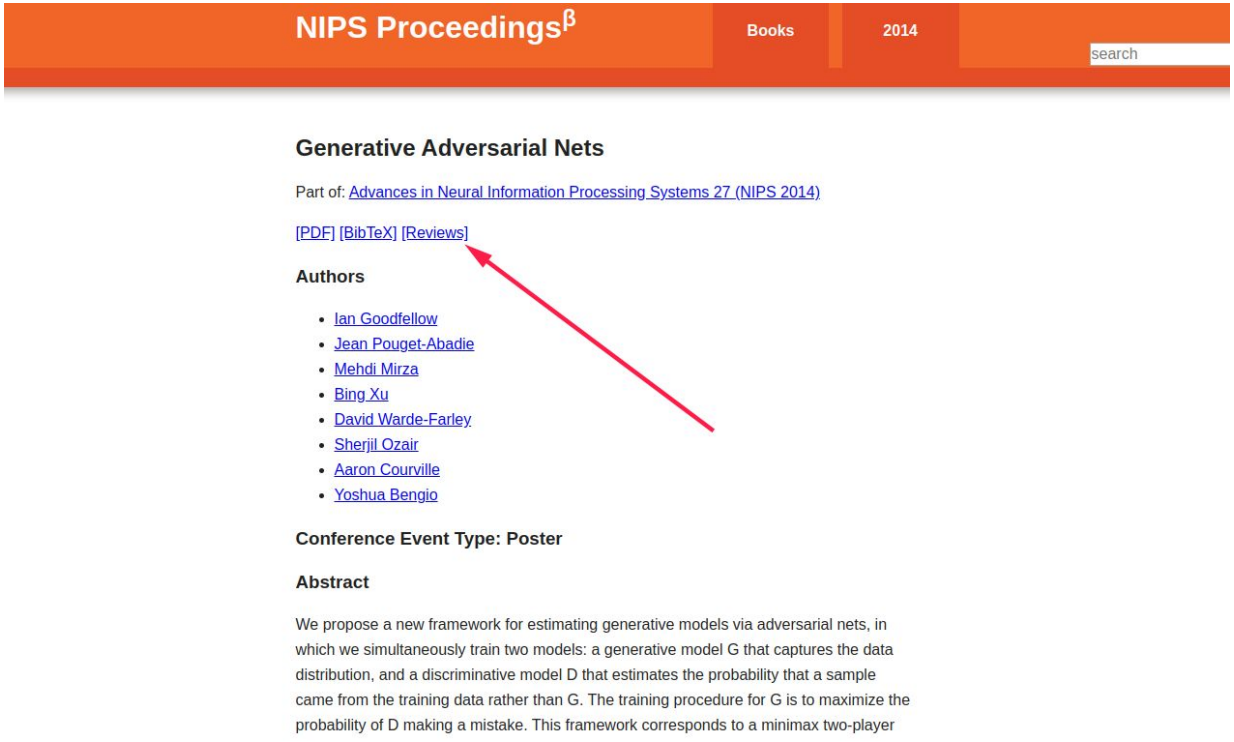
Source: https://scholar.google.com/citations?hl=en&vq=eng&view_op=list_hcore&venue=eqYFflc_uhEJ.2019

# NIPS




Source: https://aminer.org/conference/53a728e520f7420be8bbc4bb

# NIPS




Source: https://papers.nips.cc/paper/5423-generative-adversarial-nets

# How to choose a conference

- use h-index to get an impression
- search for top conferences in your field of interest
- check citations and authors of paper
    - check out comments from reviewers of the paper

# Summary

- baselines help in understanding the problem and the data
- compare your developed model (e.g. to established ones)
- a first (self-developed) baseline should be as simple as possible, reproducible
- Start with the well defined platforms (e.g. Semantic scholar) for finding relevant papers
- Look at the ‘Related Papers’ section in the *aforementioned* platforms

# Thank you!

*“All models are wrong but some are useful.”* (George Box)

# literature/sources


- *Steffen Rendle, Li Zhang, and Yehuda Koren. On the difficulty of evaluating baselines: A study on recommender systems.* (https://arxiv.org/abs/1905.01395v1)
- *Sida Wang, Christopher D. Manning. Baselines and Bigrams: Simple, Good Sentiment and Topic Classification* (https://nlp.stanford.edu/pubs/sidaw12_simple_sentiment.pdf)
- https://machinelearningmastery.com/how-to-get-baseline-results-and-why-they-matter/
- https://machinelearningmastery.com/why-you-should-be-spot-checking-algorithms-on-your-machine-learning-problems/
- https://machinelearningmastery.com/how-to-know-if-your-machine-learning-model-has-good-performance/
- https://medium.com/upwork-datascience/managing-data-science-with-v2mom-scrum-and-cross-functional-squads-655b08cf828b
- https://towardsdatascience.com/first-create-a-common-sense-baseline-e66dbf8a8a47
- https://blog.insightdatascience.com/always-start-with-a-stupid-model-no-exceptions-3a22314b9aaa
- https://towardsdatascience.com/beating-state-of-the-art-by-tuning-baselines-74ec6ad2cd59
- << **https://developers.google.com/machine-learning/guides/rules-of-ml** >>
- **<< https://www.jeremyjordan.me/ml-projects-guide/ >>**
- https://medium.com/@i.m.vivek/machine-learning-why-baseline-is-important-cc63c857a56d
- https://towardsdatascience.com/machine-learning-general-process-8f1b510bd8af
- https://developers.google.com/machine-learning/glossary
- *J.E. Hirsch. An index to quantify an individual's scientific research output* (https://arxiv.org/PS_cache/physics/pdf/0508/0508025v5.pdf)
- https://en.wikipedia.org/wiki/Netflix_Prize