Towards Privacy-Preserving Explanations for Recommender Systems

Anupam Datta | Sophia Kovaleva | Jenna MacCarley | Shayak Sen | **Daniel Calderón Villela**

Due to the ubiquity of automated decision-making today, a demand for algorithmic transparency arises

- Datta et. al's algorithmic transparency tool, Quantitative Input Influence (QII), explains how classification decisions arise in a black-box manner[1]
- In our work, we extend QII to generate explanations for recommender system decisions
- Naive extension infeasible, so we build off scalable MLlib library[2]
- Given privacy risks in data, we seek privacy-preserving explanations, and thus build sensitivity measurements needed for Differentially Private QII

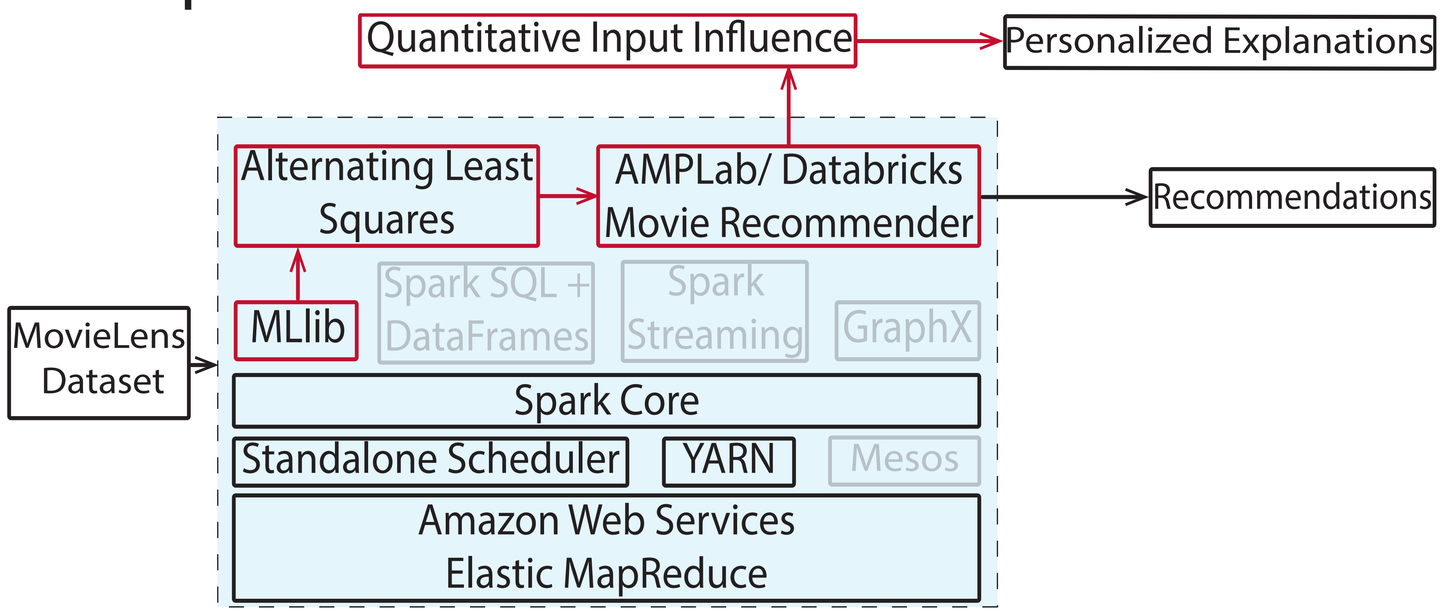
What is Algorithmic Transparency?

Harms arise from Inappropriate Information Use

Integrity	Was incorrect data influential in a recommendation?
Privacy	Is a sensitive attribute influential in recommendation?
Fairness	Do sensitive attributes result in disparate impact?

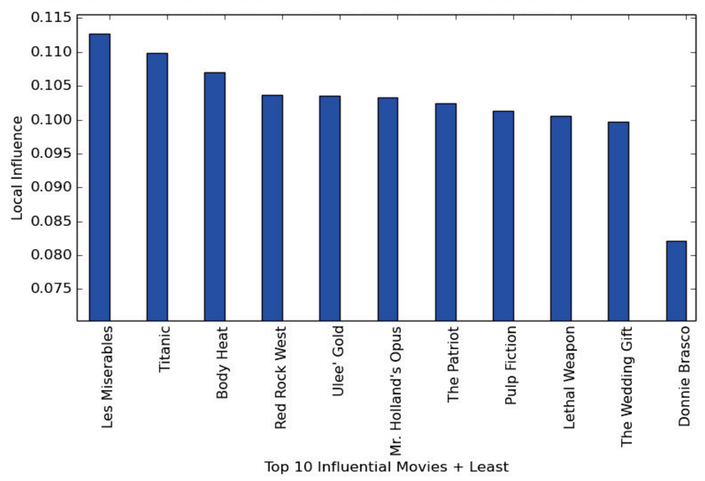
The goal of algorithmic transparency is to minimize harms by providing explanations for system decisions

Leveraged MLlib to Speed Up QII Computation to Feasible Timeframe



Personalized Explanation | User X

Local Influence of Rated Movies on User X



Recommendations for User X

Rank	Movie Name
1	Bewegte Mann, Der (1994)
2	Sanjuro (1962)
3	For All Mankind (1989)
4	Chushingura (1962)
5	Sixth Sense, The (1999)
6	Man of the Century (1999)
7	Godfather, The (1972)
8	Leather Jacket Love Story (1997)
9	Raiders of the Lost Ark (1981)
10	Sting, The (1973)

[1] Anupam Datta et al. "Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems," IEEE SS&P, 2016

[2] Xiangrui Meng et al. "MLlib: Machine Learning in Apache Spark." JMLR 17.34: 1-7, 2016

[3] Kobbi Nissim et al. "Smooth Sensitivity and Sampling in Private Data Analysis." STOC, 2007

Sensitivity Analysis

- Differential Privacy (DP) has been established as powerful definition for ensuring rigorous data privacy
- DP depends on sensitivity definition used for utility and privacy guarantees
- Nissim et al expand on classical global sensitivity definition by introducing a definition of local sensitivity[3]:

$$LS_f(x) = \max_{y:d(x,y)=1} ||f(x) - f(y)||_1$$

Local Influence is More Sensitive Than Recommendations

	Max	Mean	STD
QII LS	4.84e-2	2.08e-2	1.26e-2
Recs LS	1.65e-1	6.5e-2	5.41e-2
QII LS norm	1.1e-3	3.73e-4	3.27e-4
Recs LS norm	1.13e-5	4.49e-6	3.73e-6

QII Can Provide Explanations for System Decisions

Extensions can be pursued by:

- Scaling up to larger datasets
- Using alternative sensitivity metrics
- Implementing differentially private explanations