CSE 40647/60647 Data Science (Spring 2018)
Lecture 1: Introduction to Data Science

Goals:

- Describe what is data science;
- Describe components of data science research;
- Describe data science functionalities.

Goals specific to the first lecture:

- Know general/concrete learning goals of the course;
- Know syllabus and class schedule;
- Know course project and project schedule;
- Know grading policy;
- Know time, location, and textbook.

Part I: What is Data Science?

> **"...the process of automatically discovering *useful information* in *large* repositories of data."**
> — *Introduction to Data Mining* (Tan, Steinbach, & Kumar)
> **"...the process of discovering *patterns* in *big* data."**
> — *Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition* (Witten, Frank, & Hall)
> **"...the process of discovering *interesting patterns and knowledge* from *large* amounts of data."**
> — *Data Mining: Concepts and Techniques, 3rd Edition* (Han, Kambler, & Pei)

Our definition of the Data Science course:

> "...the art and craft of extracting ***knowledge*** from ***large*** bodies of ***structured and unstructured*** data using methods from many disciplines, including (but not limited to) machine learning, databases, probability and statistics, information theory, and data visualization."

Question I: What is/isn't Data Science?

- [ ] Looking up a record in a database.
- [ ] Noting that some last names occur in certain geographical areas.
- [ ] Searching for a term on Google.
- [ ] Taking all query results from Google and discovering that they can be grouped or categorized.
- [ ] Testing a two-sample hypothesis in a clinical trial.
- [ ] Identifying strongly significant genes when doing multiple tests across many genes.
- [ ] Finding the most popular hobby among us.
- [ ] Inferring a student's hobby.

Part II: What are the components of data science research?

A good example: Netflix prize

- Wikipedia: https://en.wikipedia.org/wiki/Netflix_Prize
  - *The Netflix Prize was an open competition for the best collaborative filtering algorithm to predict user ratings for films, based on previous ratings without any other information about the users or films, i.e. without the users or the films being identified except by numbers assigned for the contest. The competition was held by Netflix, an online DVD-rental and video streaming service, and was open to anyone who is neither connected with Netflix (current and former employees, agents, close relatives of Netflix employees, etc.) nor a resident of certain blocked countries (such as Cuba or North Korea). On September 21, 2009, the grand prize of US$1,000,000 was given to the BellKor's Pragmatic Chaos team which bested Netflix's own algorithm for predicting ratings by 10.06%.*
- Data sets on Kaggle: https://www.kaggle.com/netflix-inc/netflix-prize-data

Question II-1:

Component 1: "… *to* ***predict user ratings for films****, based on previous ratings without any other information about the users or films …"*

[ ]

Component 2: *"… Netflix provided a training data set of 100,480,507 ratings that 480,189 users gave to 17,770 movies. Each training rating is a quadruplet of the form* ***<user, movie, date of grade, grade>****. The user and movie fields are integer IDs, while grades are from 1 to 5 (integral) stars …"*

[ ]

Component 3:

What if your dataset has a non-Netflix user (say, MovieLens user)?
What if your dataset has an invalid date (say, 01/01/2019)?
What if your dataset has an invalid grade (say, 0 or 6)?

[ ]

Component 4:

What if you have datasets from different geographical areas, such as United States, Japan, Hong Kong…?
What if you have datasets from Netflix Streaming, Netflix TV, Netflix movies, Netflix App…?
What if you have external datasets from AT&T, T-Mobile, and Verizon?

[ ]

Component 5: *"****Movie rating prediction task:*** *Given a training set of <user, movie, grade> triplets, predict grades in a testing set of triplets."*

[ ]

Component 6: *Select the triplet data from your database for movie rating prediction.*

[ ]

Component 7:

(1) ***Classification:*** *Given two features, user and movie,* ***classify*** *a triplet into five categories (1 to 5).*
(2) ***Matrix completion:*** *Given a user-movie rating* ***matrix****,* ***complete*** *missing entries in the matrix.*

[ ]

Component 8:

(1) ***Feature transformation + Naïve Bayes or Support Vector Machines*** *for classification.*
(2) ***Non-negative matrix factorization*** *for matrix completion.*

[ ]

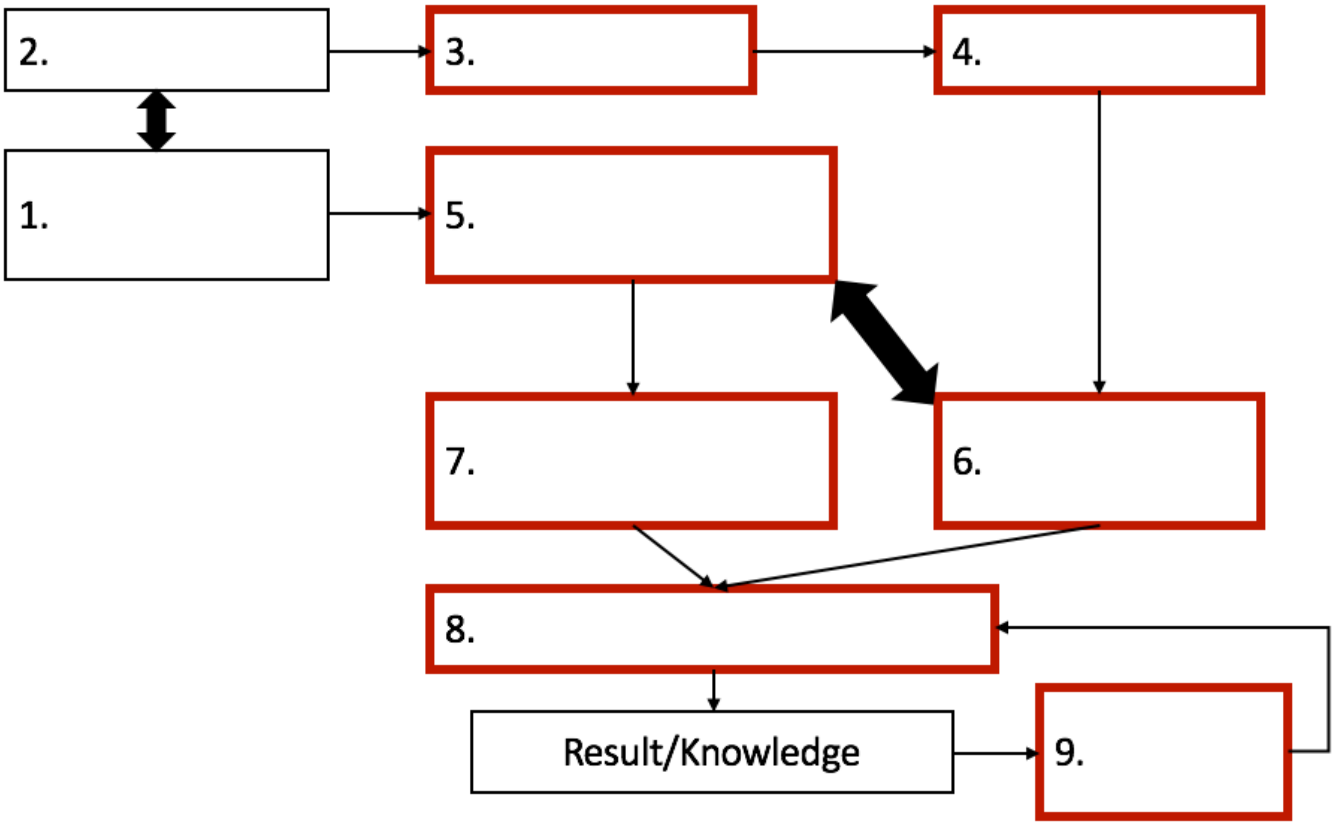
Component 9:

"A trivial algorithm that predicts for each movie in the quiz set its average grade from the training data produces an ***RMSE of 1.0540****."*
"Using only the training data, Cinematch scores an ***RMSE of 0.9514*** *on the quiz data, roughly a* ***10%*** *improvement over the trivial algorithm."*
"In order to win the grand prize of $1,000,000, a participating team had to improve this by another ***10%****, to achieve* ***0.8572*** *on the test set."*
"On September 2, 2007.... At the beginning of this period the leading team was BellKor, with an ***RMSE of 0.8728*** *(8.26% improvement)."*
"The 2008 Progress Prize was awarded to the team BellKor. Their submission combined with a different team, BigChaos achieved an ***RMSE of 0.8616*** *with 207 predictor sets."*
"On June 26, 2009 the team "BellKor's Pragmatic Chaos", a merger of teams "Bellkor in BigChaos" and "Pragmatic Theory", achieved a 10.05% improvement over Cinematch (a Quiz ***RMSE of 0.8558****)."*
"The Netflix Prize competition then entered the "last call" period for the Grand Prize."

[ ]

Component 3 to 9 are important steps of Data Science Research.

Question II-2:
Fill in with component names:



Part III: What are data science functionalities? In other words, what are data mining tasks?

- _____
- _____
- _____
- _____
- _____
- _____
- …

Name: **NetID:**

Please write down whatever question you have about this course: