CSCI 585, Final Exam, 6/30/20

Please read the following carefully, before starting the test.

1. The exam is open books/notes/devices - feel free to look up **whatever** you want!

2. There are 12 questions, 5 points each - you can choose any 7 (for a total of 35). **You can answer up to 10, we will add up your scores from all of them (with a cap of 35)!** If you turn in 11, we will leave out the highest score and count the other 10 (again for 35 max); if you submit all 12, we'll omit the top two highest scores and count the rest (again, 35 is the max). In other words, **don't turn in more than 10.**

3. There are no 'trick' questions, or ones with long calculations or formulae, nothing that requires you to needlessly write a lot.

4. **Please do NOT cheat** - this means NOT communicating with anyone via any device/medium/channel - you will get a 0, and be reported to SJACS, if you are found to have cheated; ANY attempt to get help from others in any form is a VIOLATION, as per https://policy.usc.edu/scampus-part-b/ (https://policy.usc.edu/scampus-part-b/), sections 11.11 through 11.14 [read it, if you are not familiar with it].

5. When the time is up (75 minutes), stop your work, then spend the rest of time (30 minutes) on submission.

6. Good luck! Hope you do well, and enjoy coming up with the answers.

Q0 [0 points]. You NEED TO turn this in - DO NOT omit doing so - there is a penalty of 2 points if you omit this.

Please write the following line, and sign it - it is your acknowledgment of having read USC's policies on academic misconduct (https://policy.usc.edu/scampus-part-b/ (https://policy.usc.edu/scampus-part-b/), 11.11-11.14) and agreement to honor them: **I have read USC's standards on academic integrity, and agree to abide by them.**

Q1. Of the various forms of AI, you know that supervised ML is data-driven. What are 5 different problems/issues/dangers/... that stem from supervised ML, and ways to fix/address them?

A.

1. Low recognition accuracy; better algorithms/architectures, eg using CapsNet, tweaking the loss function, etc.

2. Biased classification, stemming from the use of a biased dataset; audit the dataset and populate it to reduce or eliminate the bias.

3. 'Stale ML' - the trained network is no longer relevant to a changed input set; retrain, using newer data.

4. Disinformation production, in the form of deepfakes etc; laws can be passed to contain or eliminate their production and use.

5. Autonomous weapons ('killer bots'); a total ban.

6. Ubiquitous surveillance; enact laws to curb/eliminate excessive use [depends on the country!]

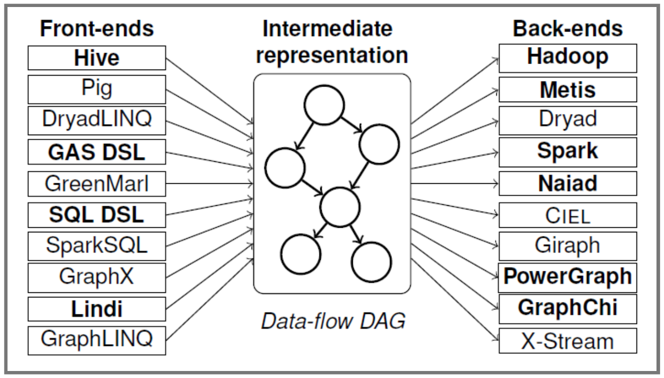
Q2. When we build a data warehouse, we ETL lots of operational data to build a multi-dimensional table - we augment each ’fact’ with a lot of additional dimensions (add multiple columns). What is the problem with using this (resulting table) for BI, and what is the fix?

A. The problem is that the drilling down and rolling up requires multiple passes through redundant data (which are in the columns we added, eg. to put in a location hierarchy); the redundant data is itself a problem (uses up storage, increases load time, uses up main memory, etc). The fix is to normalize the dimensions, creating a snowflake schema, ie. create multiple fact tables (at various aggregated levels).

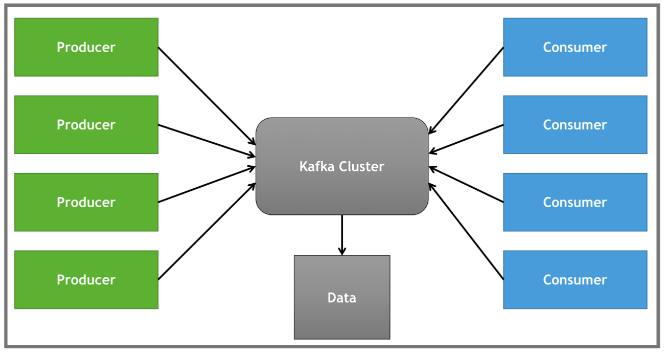
Q3. 'Many problems in CS can be solved using an extra layer of indirection', goes the adage. How does this apply to processing of data? Provide two examples - briefly discuss each, using diagrams.

A.

1. Use of an intermediate graph format, in Musketeer, to enable dataflow graphs specified in different formats to run on multiple backends.



2. Use of Kafka pub-sub, to forward messages from producers, to client consumers.



Q4. What are two different ways in which augmented reality (AR) can help deal with data handling?

A.

1. Table-top analysis, having data animate on a wall, etc. [use of AR for doing data analysis and visualization, using natural gestures, over real-world surfaces].

2. Digital Twins - specifically in manufacturing/assembly lines, to be able to map IoT sensor data back and forth between a real machine and its digital 3D equivalent ('twin').

Q5. As data science matures as a discipline, there is going to be less of a need for dedicated teams (the equivalent of an ’IT Department’ in the past) for maintaining and enhancing tools related to data processing. For the ’citizen data scientist’ (end users who are non-technical), what are various do-it-yourself (DIY) options? The end users can be individuals, small companies, startups... Name, and discuss, five different options (instead of having to write/run raw Python code customized by hand, to do ML, for ex). Think broadly!

A.

1. Use of pre-built apps on a smartphone (eg to do detection of mushrooms, clouds, birds; to do language translation...).

2. Cloud ML - upload data in a variety of formats, that is ingested using connectors.

3. Cloud BI - again, store and process data on the cloud, download results.

4. Voice assistants, eg. Alexa to help with queries.

5. Use of dataflow tools, eg. KNIME/RapidMiner/Baseet/...

6. Use of ’turnkey’ systems - local machines with pre-installed pipelines.

...

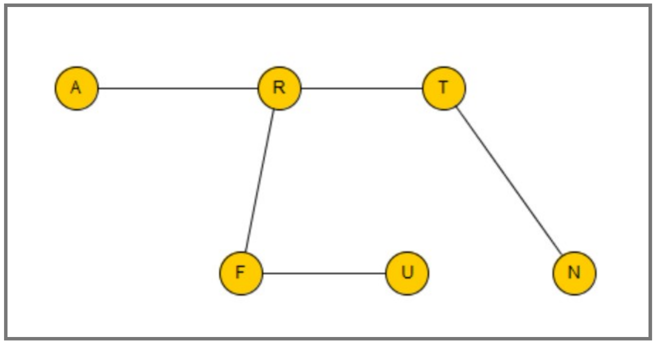
Q6. What is a unique feature of a choropleth map, for data viz? If you could animate a choropleth, discuss, using an example or two, what extra capability you would provide.

A.

a. The unique feature is this - every location on the map would have a value (eg for temperature, income, education level, or any other nominal/ordinal data).

b. We could ’zoom in’, eg. from a state-wide total (eg of COVID-19 cases) to county-wide totals; or, we could display values that change over time (again, for ex, COVID-19 cases in each state, daily, for the months of March ’20 - June ’20).

Q7. As you know, a graph offers a flexible format for storing and processing 'connected' data. How would you depict the following graph (labeled nodes, unlabeled, undirected edges), in three ways: using XML, using JSON, using a different plaintext format (of your own design)? Make sure to use valid syntax for XML and JSON.



A. There are several alternatives, for each. Here is one set of possibilities.

XML:

```
<graph>
<node id="A"> </node>
<node id="R"> </node>
<node id="T"> </node>
<node id="F"> </node>
<node id="U"> </node>
<node id="N"> </node>
<edge source="A" target="R" />
<edge source="R" target="T" />
<edge source="R" target="F" />
<edge source="F" target="U" />
<edge source="T" target="N" />
</graph>
```

JSON:

```
{
 'graph': {
            'nodes': ["A","R","T',"F","U","N"],
            'edges': [["A","R"], ["R","T"], ["R","F"], ["F","U"], ["T","N"]]
          }
}
```

Plaintext [assumes that nodes are listed first as a block, each line contains a nodename; edges are listed next as a block, using commas between nodenames:

```
A
R
T
F
U
N
A,R
R,T
R,F
F,U
T,N
```

Q8. Agricultural production is full of data processing possibilities, from sowing seeds/seedlings all the way to harvest and product sales. What are 5 different kinds of data you would capture, store, analyze+act on? Specify the type of data and acquisition, type of storage format, type of processing. Think broadly!!

A. Lots of possibilities...

1. Soil dryness, captured using sensors/drones; stored in a k:v DB; analyzed using regression or a neural net.

2. Crop disease, captured with cameras; stored in a document DB; analyzed using ML (CNNs).

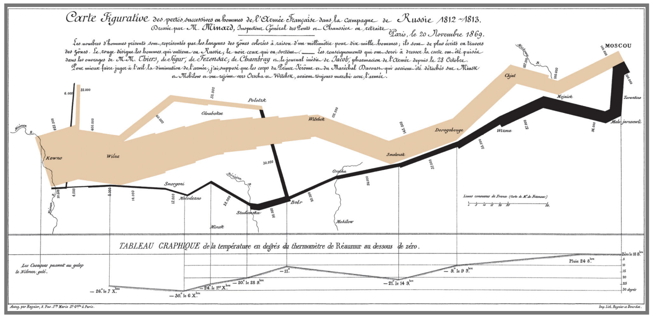
3. Crop yield, measured traditionally (eg by weighing), stored in a relational DB, analyzed using SQL.

4. Weed management - captured using cameras, stored as images in folders, used to train a binary classifier (eg SVM)

5. Harvesting - data captured using cams, analyzed on the fly using an NN trained on multi-label classification (eg not-ripe, ripe-for-harvest, damaged...)

...

Q9. Charles Minard’s 1869 visualization of multi-variate data, is a classic. What are two features of modern visualizations that are absent in a printed graphic such as Minard’s? Explain briefly, using examples.


Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.
Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. ___ Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.
Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mobilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.
MOSCOU
Niemen R.
Kowno
Wilna
Gloubokoe
Polotzk
Witebsk
Smolensk
Dorogobouge
Wizma
Chjat
Mojaisk
Tarantino
Malo-jarosewli
6.000
22.000
422.000
400.000
60.000
33.000
30.000
175.000
145.000
37.000
55.000
87.000
127.100
100.000
96.000
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

A.

1. Animation - eg. on a map, showing the passage of a storm (or the spread of disease…).

2. Interactivity - eg. zooming in for more detail, sliding along a timeline, etc. [LOTs of examples].

Q10. Our bodies are rich sources of incredibly valuable data, given the fundamental necessity of good health and hygiene, and disease prevention/management/cure [currently, data from medical procedures, tests etc are owned by the hospitals and clinics where the procedures are performed, and not by the patients; effort is underway to try to change this]. What kinds of data generated from our bodies, might be marketable, and for what purposes? Discuss 5 - type of data, benefit derived (use for it). You can answer from the perspective of a female, or male, or both; from the perspective of a healthy individual, or one with disease(s). Think broadly!

A.

1. Blood data

2. Urine data

3. Gut biomes

4. Heart

5. Lungs

6. Sleep

7. Brain

8. Cancer

9. Psychological data - eg. stress, trauma...

...

Q11. How does data (ease of collection, storage, processing) aid journalism? How does it hinder it? Be sure to discuss each ’side’, with examples.

A.

Aiding - it is possible to tell ’data stories’ - in-depth articles backed by interactive graphics, eg. on war, disease, pollution...

Hurting - disinformation, eg. fakevideos of election candidates, celebrities; altered photos that appear to counter climate change facts, etc.

Q12. With all the data in the world we have at our disposal for training ML algorithms, and with even more data we can collect/generate for this purpose, a variety of jobs in various industries are under threat - of being replaced by 'intelligent automation', ie. "IA". What cannot be fully automated, and why? Describe, in a paragraph or two.

A. Jobs/roles that involve the 'human touch' (eg. teaching, nursing), perspective (eg. in journalism - editorials, opinion pieces), critical expertise (eg. surgeons, lawyers), business acumen (eg upper level managers, CEOs)...