CSE 344

MAY 30TH – ANALYSIS

ADMINISTRIVIA

- **HW8 Due Friday, 11:30**
- **OQ7 Due Tonight, 11:00**
- **Course evaluations!**
- **Exam review**
  - Section tomorrow
    - Problems from previous midterms
  - Friday June 1, in-class
    - Topics list

EXAM

- **June 6th, 8:30 – 10:20 am**
  - Please don't show up at 9:30!
- **Cumulative final**
  - Focus on second half of the course
    - Database Design
    - Transactions
  - Covers through the end of isolation last Friday
  - Practice exam + solutions out Friday

EXAM

- **One sheet of notes**
  - Front and back
  - Written or typed
- **Length**
  - Roughly 50% longer
  - Twice the time
  - Watch out for tricky questions

EXAM

- **Free to meet Friday**
  - Now is the time to discuss grades, not after the final
  - All posted grades will be final at end of day on Friday
- **HW7 Graded before final**
- **HW8 Graded by end of quarter**

DATABASES

- **Two types of data base usage**
  - Transactional
  - Analytical
- **How does the usage differ?**

DATABASES

- **Transactional**
  - Maximizing throughput
  - Ensuring consistency
  - No “right way” to deploy a DB

DATABASES

- **Transactional**
  - Maximizing throughput
  - Ensuring consistency
  - No “right way” to deploy a DB
- **Things to consider**
  - Relational v. Semi-structured
  - Parallelism and Multi-national distribution
  - Durability and analytical accumulation

ANALYTICAL DB USAGE

- **Web DB are primarily transactional**
  - Need to be processed to be analyzed
- **Facebook**

ANALYTICAL DB USAGE

- **Web DB are primarily transactional**
  - Need to be processed to be analyzed
- **Facebook**
  - Transactional element: free-user service
  - Analytical element: paid-user service

ANALYTICAL DB USAGE

- **Web DB are primarily transactional**
  - Need to be processed to be analyzed
- **Facebook**
  - Transactional element: free-user service
  - Analytical element: paid-user service
  - *Which do you think is most important?*

ANALYTICAL DB USAGE

- **Data analysis is big business**
  - Machine learning
  - Data mining
  - “Captcha” tests

ANALYTICAL DB USAGE

- **Data analysis is big business**
  - Machine learning
  - Data mining
  - “Captcha” tests
- **Analysis in SQL?**
  - Very limited capacity
  - Combine with other data processing

DATA ANALYSIS

- **Data analysis is big business**
  - Machine learning
  - Data mining
  - “Captcha” tests
- **Analysis in SQL?**
  - Very limited capacity
  - Combine with other data processing

ANALYTICAL TOOLS

- Common data analysis tools

ANALYTICAL TOOLS

- **Common data analysis tools**
  - Python -- great for those with CS background
  - R – great for those with social science background

ANALYTICAL TOOLS

- **Common data analysis tools**
  - Python -- great for those with CS background
  - R – great for those with social science background
- **“Common” data analysis tools**
  - Excel – see “Growth in a Time of Debt”
  - Tableau – graphs are not analysis

ANALYTICAL TOOLS

- **Python**
  - Good scripting language to learn
  - Interfaces well with other languages
- **R**
  - Designed for data science
  - Well supported public packages
  - Very pretty graphs

DATA ANALYTICS VS DATA SCIENCE

- **Analysis is often heuristic**
  - If you cannot explain “why” your model predicts behavior, then you don’t understand it
- **Data science**
  - Combination of statistical analysis and cooperation with relevant experience

A NOTE ABOUT ETHICS

- **Computer aided decision making affects lives**
  - ”Growth in a time of debt”
  - NYC CompStat
- **Essential to be certain in the quality of your work and the impacts that your decisions will have**
  - Present data ethically

A NOTE ABOUT ETHICS

- **Computer aided decision making affects lives**
  - ”Growth in a time of debt”
  - NYC CompStat
- **“95% accuracy” can be a very misleading statement**
  - Cancer screening
  - Precision

ANALYTICAL JOBS

- Statistics
- Data cleaning
- Reconciling multiple data sources
- Verification
- Visualization

CONCLUSION

- **While most DB usages need to be optimized for transactional usage, they must also be ready for analytical tools**
  - This is separate from most of the course as demonstrated
  - Important and common application of data
  - Big money for analytics, but also important to understand error and verification