Introduction
Breakout 1

• What is data analysis for game development?
• Where does this data come from?
• What can game analysis do for game development?

• Icebreaker, Groupwork, Questions
  https://web.cs.wpi.edu/~imgd2905/d20/breakout/breakout-1.html
What is data analysis for game development?
What is data analysis for game development?

- Using **game data** to inform the game development process

Where does this data come from?

- Players, actually playing game
  - Quantitative (instrumented)
  - Qualitative (subjective evaluation)
- (But often lots more of former!)
What is data analysis for game development?

• Using game data to inform the game development process
• Where does this data come from?
What is data analysis for game development?

- Using **game data** to inform the game development process
- Where does this data come from?
  - **Players**, actually playing game
    - Quantitative (instrumented)
    - Qualitative (subjective evaluation)
    -(But often lots more of former!)
What can game analysis do for game development?
What can game analysis do for game development?

- **Improve level design** – e.g., see where players are getting stuck
- **Focus development on critical content** – e.g., see what game modes or characters are not used
- **Balance gameplay** – e.g., tune parameters for more competitive and fun combat
- **Broaden appeal** – e.g., hear if content/story is engaging or repulsing
- **Note**: game data often informs players, too
  - Analytics not dissimilar
Why is data analysis for game development needed?
Why is data analysis for game development needed?

• **Challenge**
  – Games now larger & more complex
    + Number of reachable states, characters
    → Game balance harder to achieve
  – Need for metrics to make sense of player behavior has increased

• **Opportunity**
  – New technologies enable aggregation, access and analysis
IMGD 2905 – Doing Data Analysis for Game Development

- Data analysis pipeline – get data from games, through analysis, to stakeholders
  - Summary statistics – central tendencies of data
  - Visualization of data – how to display analysis, illustrate messages
  - Statistical tests – quantitatively determine relationships (e.g., correlation)
  - Probability needed as foundation (also used for game rules)
- Regression – model relationships
- More advanced topics (e.g., ML, Data management, etc.)

For this class:
- Described in lecture
- Discussed in class
- Read about in book
- Applied in projects & homework
Foundations for Data Analysis @ WPI

• Statistics classes
  – MA 2610 Applied Statistics for Life Sciences
  – MA 2611 Applied Statistics I
  – MA 2612 Applied Statistics II

• Probability classes
  – MA 2621 Probability for Applications

• Data Science (minor and major)
  – DS 1010 Introduction to Data Science
  – DS 2010 Modeling and Data Analysis
  – DS 3010 Computational Data Intelligence
  – DS 4433/CS4433 Big Data Management and Analytics

• Data Mining
  – CS 4445 Data Mining and Knowledge Discovery in Databases

• Other
  – CS 1004 Introduction to Programming for Non-Majors
  – CS 3431 Database Systems I

Note – other Stats and Probability classes geared for Math majors
Outline

• Overview (done)
• Game Analytics Pipeline (next)
• Examples
Sources of Game Data

Quantitative (Objective)
- Internal Testing
  - Developers
  - QA
- External Testing
  - Usability testing
  - Beta tests
  - Long-term play data

Qualitative (Subjective)
- Surveys
- Reviews
- Online communities
- Postmortems

From data to dissemination?

Game analytics pipeline
Game Analytics Pipeline

Game

Raw Data

Extracted Data

Analysis

Exploratory Graphs/Stats

Charts and Tables

Statistical Tests

Dissemination

Presentation

Report
Game Analytics Pipeline – Example

- Track-o-Bot
- Collect-o-Bot
- Python
- JSON
- Analysis
- Excel
- Dissemination
  - Word
  - PowerPoint
- Proj 3!
Game Analytics Components

• **Games** – breadth of experience with games, specific experience with game to be analyzed

• **Tools** – import, clean, filter, format data so can analyze

• **Statistics** – measures of central tendency, measures of spread, statistical tests

• **Probability** – rules, distributions

• **Data Visualization** – bar chart, scatter plot, histogram, error bars

• **Technical Writing and Presentation** – white paper, technical talk; audience is peer group, developers, boss
Outline

• Overview (done)
• Game Analytics Pipeline (done)
• Examples (next)
Example:
Project Gotham Racing 4

http://dl.acm.org/citation.cfm?id=1985952

- **Publisher** – Microsoft 2007
  - 134 vehicles, 9 locations, 10 game modes

- **Analyzed data**
  - (Authors worked at Microsoft)
  - 3.1 million log entries, 1000s of users
### Project Gotham Racing 4: Results

#### Thoughts?

#### What are some main messages?

<table>
<thead>
<tr>
<th>Game Mode</th>
<th>Races</th>
<th>% Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>OFFLINE_CAREER</td>
<td>1,479,586</td>
<td>47.63%</td>
</tr>
<tr>
<td>PGR_ARCADE</td>
<td>566,705</td>
<td>18.24%</td>
</tr>
<tr>
<td>NETWORK_PLAY</td>
<td>584,201</td>
<td>18.81%</td>
</tr>
<tr>
<td>SINGLEPLAYER_PLAY</td>
<td>185,415</td>
<td>5.97%</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>NET_TOURNY_ELIM</td>
<td>2,713</td>
<td>0.09%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group</th>
<th>Races</th>
<th>% Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>STREET_RACE</td>
<td>795,334</td>
<td>25.60%</td>
</tr>
<tr>
<td>NET_STREET_RACE</td>
<td>543,491</td>
<td>17.50%</td>
</tr>
<tr>
<td>ELIMINATION</td>
<td>216,042</td>
<td>6.95%</td>
</tr>
<tr>
<td>HOTLAP</td>
<td>195,949</td>
<td>6.31%</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>TESTTRACK_TIME</td>
<td>7,484</td>
<td>0.24%</td>
</tr>
<tr>
<td>CAT_N_MOUSE_FREE</td>
<td>3989</td>
<td>0.13%</td>
</tr>
<tr>
<td>CAT_N_MOUSE</td>
<td>53</td>
<td>0.00%</td>
</tr>
</tbody>
</table>
Project Gotham Racing 4: Results

<table>
<thead>
<tr>
<th>Game Mode</th>
<th>Races</th>
<th>% Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>OFFLINE_CAREER</td>
<td>1479586</td>
<td>47.63%</td>
</tr>
<tr>
<td>PGR_ARCADE</td>
<td>566705</td>
<td>18.24%</td>
</tr>
<tr>
<td>NETWORK_PLAY</td>
<td>584201</td>
<td>18.81%</td>
</tr>
<tr>
<td>SINGLE_PLAYER_PLAY</td>
<td>185415</td>
<td>5.97%</td>
</tr>
<tr>
<td>NET_TOURNY_ELIM</td>
<td>2713</td>
<td>0.09%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group</th>
<th>Races</th>
<th>% Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>STREET_RACE</td>
<td>795334</td>
<td>25.60%</td>
</tr>
<tr>
<td>NET_STREET_RACE</td>
<td>543491</td>
<td>17.50%</td>
</tr>
<tr>
<td>ELIMINATION</td>
<td>216042</td>
<td>6.95%</td>
</tr>
<tr>
<td>HOTLAP</td>
<td>195949</td>
<td>6.31%</td>
</tr>
<tr>
<td>TESTTRACK_TIME</td>
<td>7484</td>
<td>0.24%</td>
</tr>
<tr>
<td>CAT_N_MOUSE_FREE</td>
<td>3989</td>
<td>0.13%</td>
</tr>
<tr>
<td>CAT_N_MOUSE</td>
<td>53</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

- **Mode**
  - *Offline career* dominates
  - *Network tournament* hardly used

- **Events**
  - *Street race* and *network street race* dominate
  - *Cat and mouse* never used

- **Vehicles (not shown)**
  - 1/3 used in less than 0.1% of races
Project Gotham Racing 4: Conclusion

• Content underused - 30-40% of content in less than 1% of races

• Use to shift emphases for DLC, next version
  – Asset creation costs significant, so even 25% reduction noticeable

• Other (not shown)
  – Encouraging new players to play career mode
    † Increasing likelihood of continuing play
  – Encouraging new players to stay with F Class longer
    † Rather than move to more difficult to control A Class
Example:
Halo 3


• Publisher – Microsoft 2007
  – Achievements: single player missions, challenges such as finding skulls, multiplayer accomplishments...

• Analyzed data
  – (Author worked at Microsoft)
  – 18,0000 players
Halo 3: Results

Thoughts?
Main messages?
Halo 3: Results

- 73% of players completed campaign
  - Can compare to other Xbox games
- Took 26 days to accomplish
- Double that time for all original content
- DLC provides users up to 2 years of content

Good Descriptive Example
• Publisher – Riot Games 2009
  – Rank: ~5 Tiers, 5 divisions each → 25

• User study (52 players)
  – Play LoL in controlled environment
  – Record objective data
    + (e.g., player rank and game stats)
  – Survey for subjective data
    + (e.g., match balance and enjoyment)

League of Legends: Results

Objective

Cumulative Distribution

Team Imbalance
(difference in average player rank)

Game Imbalance
(difference in average team rank)

Main messages?
League of Legends: Results

Objective

Most teams are balanced
But about 10% more than 3 from mean

Most games evenly matched
But about 5% difference of 2 from mean
League of Legends: Results

Subjective

Cumulative Distribution

Balance

- My team was much better
- The teams were even
- The other team was much better

Cumulative Distribution

Enjoyment

- Least
- Most

Main messages?
League of Legends: Results

Subjective

Win? Game is balanced
Lose? Game is imbalanced

Main messages?

Win? Game is fun (70%), never not fun
Lose? Game is almost never fun (90%)
League of Legends: Results

Imbalance in player’s favor the most fun!

Matchmaking systems may want to consider - e.g., balance not so important, so long as player not always on imbalanced side
Summary

• Data analysis for games increasingly important
  – Has potential to improve game development

• Knowledge and skills required
  – Scripting
  – Statistics
  – Data analysis
  – Writing and presentation

“Let’s get to it, already!”
-- Tracer (Overwatch)