The Challenge

Elm City Stories is an educational video game designed for middle and high school students. The game was developed for researchers at the Play2Prevent Lab within the Yale School of Medicine. The overall goal for the designers of the game is to prevent negative health outcomes such as HIV, sexually transmitted infections, and substance misuse in young at-risk teens (ages 11-14 years) by increasing their perception of risk and acquiring healthy skills to prevent engaging in risky behavior. A key component to this change is a set of skills that allow teens to predict and understand future consequences of actions

From a research point of view, one interesting aspect about this game, and games like it, is that researchers usually have to rely on questionnaires to try to predict future behaviors. But people are not always the most reliable reporters of their own beliefs and actions, particularly young people. One promise of using games to understand real-life behavior is that patterns of behavior within a game might predict real-life behavior. For example, perhaps risk-taking in gameplay, or caution, or persistence, reflect behavior in real-life situations. Some examples of low-level cognitive processes that play a role in playing this game are attention, impulsivity, working memory, and others.)

The challenge in this DataFest is to help the researchers who created Elm City Stories to see if their game might be useful in understanding real-life behavior. To do this, they're asking you to try to characterize, measure, observe, and display patterns of play within the game. One goal for future games is to design them so that the games produce real-time data that is useful to psychology researchers. Your answers will help the Play2Prevent Lab researchers better understand what types of data this might be.

The primary data set consists of player logs for a 166 players. The players were recruited from middle schools in districts throughout the state of Connecticut roughly ten years ago. We provide the data in two modes. In the first mode, one single file contains all of the logs for all of the players. In the second mode, we provide separate files for each player. The log records each "event" that a player performs in the game, as well as the time the action was performed. It also records supplementary information, such as the players' schools, dates of play, etc. We advise working with the log for a single player when just starting out.

You'll probably be struck by the paucity of numerical variables. The key to success for this challenge will be to come up with meaningful metrics that can be used to evaluate play. For example, you might want to measure how long it takes before certain actions happen, or how much time between actions or events, or even just count certain events. Think of how you might compare two players to see if they are similar or different. Think of different ways you might rank players to see if any are "better"

than others. Or think in more qualitative terms: do some players favor different aspects of the game over others?

This is one of the more open-ended and complex challenges DataFest has provided. The researchers who created the game have had very few opportunities to examine the log data, and you will in effect be the first to take a serious look at it. In that sense, almost everything you discover will be new and potentially notable.

Game Play Overview

Here's a brief overview of gameplay, excerpted from the IVY Design Document (which is provided for you in its entirety in the documentation.) We added references to provided variables where possible.

This is the typical gameplay experience of the player in a play session:

1. Begin Session
    a. If first session: create Aspirational Avatar (event_id in range 600-605) and play tutorial Challenge Scene (stack_id = 0)
    b. If not: briefly review game so far
2. Enter an unlocked, uncompleted Challenge Stack (event_id in range of 200-221).
    a. Explore the Challenge Scenes (animatic_id) within the stack
    b. Find new Key Points and new Minigame Cards (these are not directly viewable in the logs)
    c. Change some Decision Points within the Challenge Stack (not directly viewable in the logs.)
3. Return to Lifeline (event_id in the range 100-105).
4. Enter a Minigame (event_id in ranges 300-517, 800-912; results from these are provided in event_id 1000-1005).
    a. Pick a minigame level by selecting a minigame Card [variable: Minigame_id]
    b. Play minigame level & earn additional Stars to increase overall Skill Level represented by that minigame [variable: minigame_level_id]
    c. Earn additional star increases that Skill one level.
    d. Find new Key Points and new Minigame Cards (these are not directly viewable in the logs)
    e. Change some Decision Points within the Challenge Stack (not directly viewable in the logs.)
5. Return to Lifeline (event_id in the range 100-105).
6. Re-enter uncompleted Challenge Stack
    a. Use new Skill Levels to unlock additional Key Points and complete Stack
7. Repeat from Step 2 until near end of session
8. Enter the Epilogue (event_id in range 703-708) and see the changes from this session
    a. Customize anything customizable
9. End Session (event_id = 1)

There are 12 stacks, including the "welcome" stack, which serves as an orientation. Play continues until all 12 stacks are completed. Stacks contain scenes. One scene introduces the situation, and another always shows a negative outcome. Later, after leveling up, players will be able to view the positive outcome for that scene.

Stacks

1. Welcome to PlayForward
2. Make the grade, make a friend
3. My friend's birthday
4. Grandma's pills
5. Summer sneaking
6. New Year's Eve
7. When Things Got Serious
8. Car Crash Curiosity
9. When HIV Hits Close to Home
10. The Prom
11. Honest Conversations
12. My First Paycheck

Log Structure

Each event has an id number (event_id), and whenever a defined event occurs, it is logged, along with a time stamp and any data values that result. These event id's are provided within the data set and are listed in the **Event ID**s spreadsheet. The core action takes place in the Challenge Stack events, which all have event id's in the 200-221 range. We provide fairly complete information about these events.

Notable events:

Most of the gameplay can be followed through the 200-221 events. A typical play will see players engaging with the Challenge Stack by entering particular scenes (stack_id), "panning" the scene (scrolling from left to right to make the entire scene visible), "scanning" the scene (clicking on different parts of the scene), clicking on objects and/or people with the result that those objects are "activated" or they remain "locked", and then moving on to "mini games". After completing mini games their skill levels are (ideally) increased and they return to the stack to continue clicking, this time activating objects for which their newly increased skill levels have unlocked. When all objects are unlocked, they

"complete" the stack. They then will usually go to the Lifeline (100-105 events) to choose the next stack.

We also provide detailed information about the Avatar events (which happen only at the start of the very first session, events 600-605) and Minigame events (events 100-1005). The minigames themselves are extremely complex. Each Minigame a has its own structure and its own set of data generated, and these are not documented in a consistent fashion. We therefore advise tackling these only in very particular cases. The primary data will provide the values generated by these minigames in a single column; if you're interested in these, you'll have to write your own code to put these in an analyzable format.

DataFiles Provided

logs.csv	The primary data set. Contains all logs for all players. Roughly 1.5 Gb. 132 columns, 2,106,597 rows.
player-6607011.csv player-6486029.csv player-6427031.csv	Player logs extracted from logs.csv. Players selected randomly.
S5_scores.csv	Supplementary data. See "additional data" below for details.
raw-data.zip	compressed file containing multiple folders of raw data
logdata.zip	compressed version of logs.csv, plus the three individual player logs
log-headers.csv	list of variable names. These are defined in the Data Dictionary

Documentation Provided: Spreadsheets of key values

Data Dictionary.csv	a list of variables in the order provided with brief description and relation to event ID codes.
Event_IDs.xlsx	a spreadsheet listing all event id's, event categories, and data generated by each event and examples.

IVY Game Data Logged Events.xlsx	One of the more central spreadsheets. Various definitions.
Keys spreadsheet.xlsx	a collection of translations for coding various flags and keys in the data set. It also includes an index for which spreadsheets to use to find descriptions of some Minigame variables
AspirationalAvatarData.xlsx	spreadsheet to document the Aspirational Avatar minigame
ChallengeStacksData.xlsx	provides context for the scenes in the Challenge Stack
KnowledgeMinigameData.xlsx	documentation for Knowledge Minigame. Includes 10 sheets: an overview sheet plus an additional sheet for each level of play
PeopleMiniGameData.xlsx	10 sheets
PriorityMiniGameData.xlsx	14 sheets
RefusalMiniGameData.xlsx	11 sheets

Documentation Provided: Descriptive documents

"Section 5 Self-Efficacy for Drug Survey.pdf" paper	supplemental; see "additional data" below for details.
Clin Trials-2016-Fiellin.pdf	journal paper that describes the clinical trial that produced the data in S5_scores_cleaned.csv
IVY Challenge Stack Design Document.pdf	a written description of how players interact with the Challenge stack
IVY Design Document.pdf	a written description of how the entire game is designed
Playforward Quick Manual V11.pdf	How-to-play guide for students and teachers. Section 4 is particularly helpful.

Advice

Where to begin? First, become acquainted with the data dictionary and the Playforward Quick Manual. Of the two "Design Documents", the one for the Challenge Stack is probably the most useful. Of the

spreadsheets, the data dictionary is most important followed by the event id spreadsheet. The Ivy Game Data Logged Events spreadsheet serves as a sort of index to the other spreadsheets, and it and the rest should just be referred to as needed. (In fact, most of the useful information in the Ivy Game Data Logged Events spreadsheet is organized more nicely in the Event ID spreadsheet and the data dictionary.)

Because of the complexity of the data, we strongly urge you to "think small" when you begin. You might, for example, focus on a very high level of behavior concerning entering and exiting scenes. Or maybe you'll focus on a single player and provide a very detailed examination of their play. It might be that the piece you choose is too small (but we doubt it), and if so you'll know because you'll quickly finish it up and so can then add another small piece. The small pieces, when put together, will give you insight and ideas for further investigation or for a conclusion.

Computing Elapsed Time

A technical challenge that many will wrestle with is computing elapsed time between events and until events. Here are some things you need to know in order to do this:

1) The file logs.csv, which contains all of the data, is highly structured. The first entries are from one player, and each of that player's events are provided in the order that they occurred. After that player's entries, the next player's entries are provided, also in the order that they occurred.

2) Gameplay takes place across several sessions, and the sessions themselves may span weeks. Thus, you will see several different dates of play for a single player.

3) The elapsed time variable ignores dates. For example:

Event	Date	Time	Elapsed Time (seconds)
starts session	April 1	8:00:00	0
ends session	April 1	8:01:00	60
starts session	April 2	8:00:00	60
ends session	April 2	8:02:00	120

So although 24 hours passed since the start of session 1 and session 2, the elapsed time includes only those seconds for which the player was logged onto the game.

Reading in the Data

The full data file, with all 166 players included, should have 2,106,597 rows and 131 columns. We strongly advise that you compare your uploaded data structure with the description provided in the data dictionary.

For Rstudio users, please note that tidyverse::read_csv may read in some variables incorrectly if guess_max is set too small. (The default is 1000 and that is too small). This command should work, although it will be slow (maybe 10-20 minutes):

```
library(tidyverse)
mydata <- read_csv("filename",guess_max=2106600)
```

Additional Data Set

The players in this data set were part of a randomized clinical trial. 166*2 middle school students were recruited for the study. The 166 in this data set were those randomly assigned to the treatment group. The treatment in this case was to play the game. As a response measure, both groups of students took a survey on efficacy i resisting drugs. The mean scores of only the treatment group students are provided in the S5_scores_cleaned.csv. A **higher** mean scores indicates that the student has **lower** efficacy in resisting drugs.