Player Modeling: Clustering

Matthew Guzdial

guzdial@ualberta.ca

Announcements

- HW3 due tonight (11:55pm, regular grace period and late policy)
- HW3 and Quiz 3 returned early week
- HW4 released (and introduced) today

Review

- Model-based player modeling
  1. Performance rating
  2. Flow
  3. Drama Management
  4. Player types

Review

- Model-based player modeling
  1. Performance rating
  2. Flow
  3. Drama Management
  4. Player types

- Mentioned model-free approaches

Model free Approaches

Outside of the game, take some data about a playerbase and/or game world and use it to learn about the game…

- **Cluster** learn what types exist
- **Analyze** how does x relate to y?
  - X: Input about the playerbase/game world
  - Y: Something we want to predict/understand
- **Classification** which y is x a member of?

Machine Learning

Three Classes of Machine Learning Approaches:

1. **Supervised Learning**: Lots of inputs (X) and outputs (Y), try to learn a function to map X->Y, so that you can approximate Y' for new values of X.
2. **Unsupervised Learning:** Given inputs (X) attempt to learn a function to find categories/clusters/additional information (Y) using some metric.
3. **Reinforcement Learning**: Given an environment, attempt to approximate a function going from states (X) to the optimal actions (Y) based on observed reward from the environments.

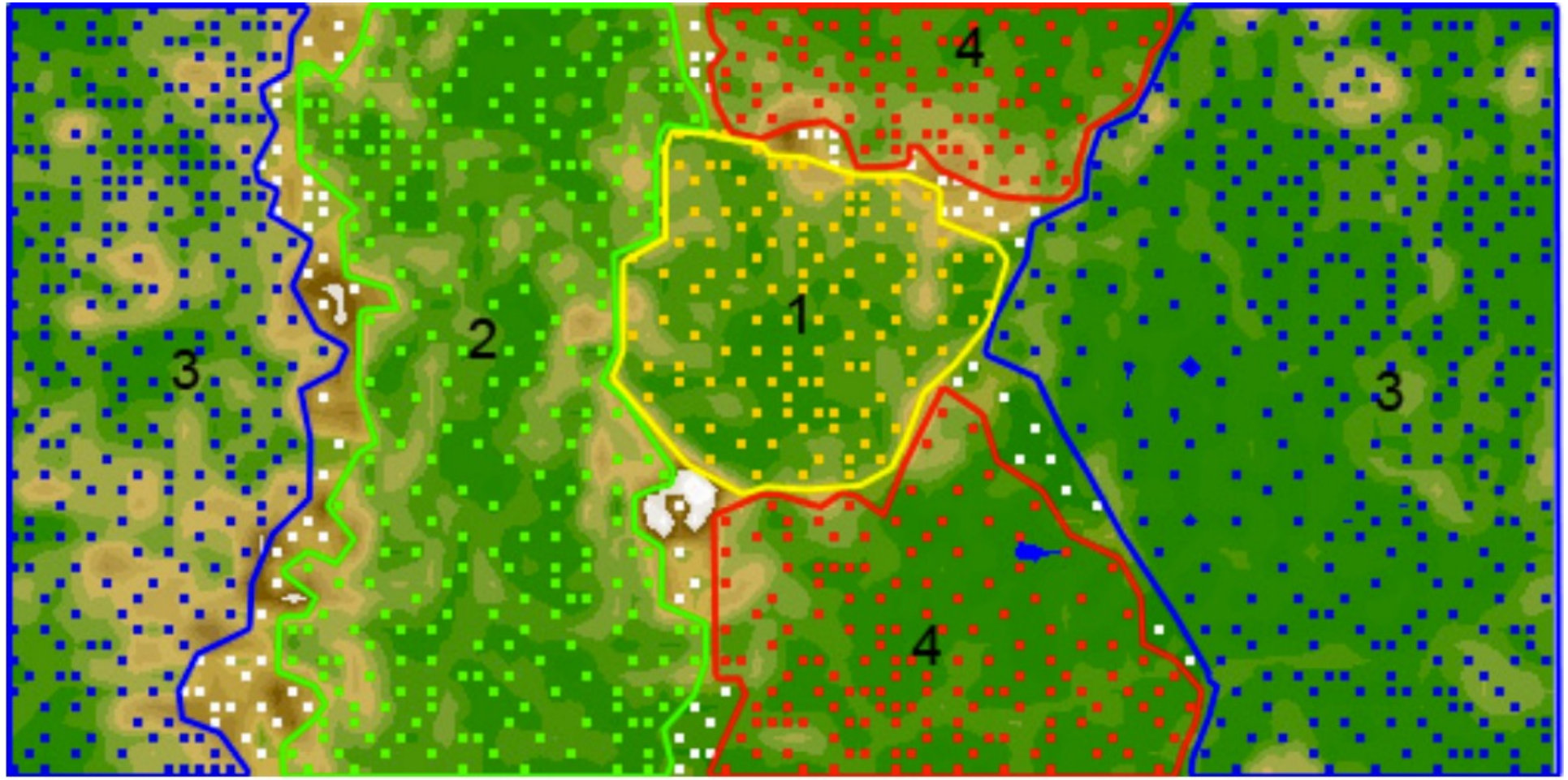
Machine Learning

Three Classes of Machine Learning Approaches:

1. **Supervised Learning**: Lots of inputs (X) and outputs (Y), try to learn a function to map X->Y, so that you can approximate Y' for new values of X.
2. **Unsupervised Learning: Given inputs (X) attempt to learn a function to find categories/clusters/additional information (Y) using some metric.**
3. **Reinforcement Learning**: Given an environment, attempt to approximate a function going from states (X) to the optimal actions (Y) based on observed reward from the environments.

Unsupervised Learning Example

https://youtu.be/HJS-SxgXAI4?t=6



Today: Clustering!

- Unsupervised Learning approach (From X, Approx. Y)
- For each input point (x), approximate what category (y) it belongs to.

- Uses (for the clusters):
  - Automatically discover categories
  - Group similar entities
  - Reduce complexity/variance

Clustering Approaches

- K means
- K medoids
- Single-linkage clustering
- Gaussian Mixture Models

K means

```
initialize centroids randomly
oldcentroids = []
while not centroids==oldcentroids:
    oldcentroids = centroids
    calculateClusters();//cluster each element to closest centroid
    centroids = average of each cluster

return centroids
```

Video Example

https://youtu.be/nXY6PxAaOk0

Demonstrate on white board

In some situation “mean” makes no sense

How would you calculate the average of a cow and a chicken?

K medoids

initialize centroids randomly

oldcentroids = []

while not centroids==oldcentroids:

    oldcentroids = centroids

    calculateClusters();//cluster each element to closest centroid

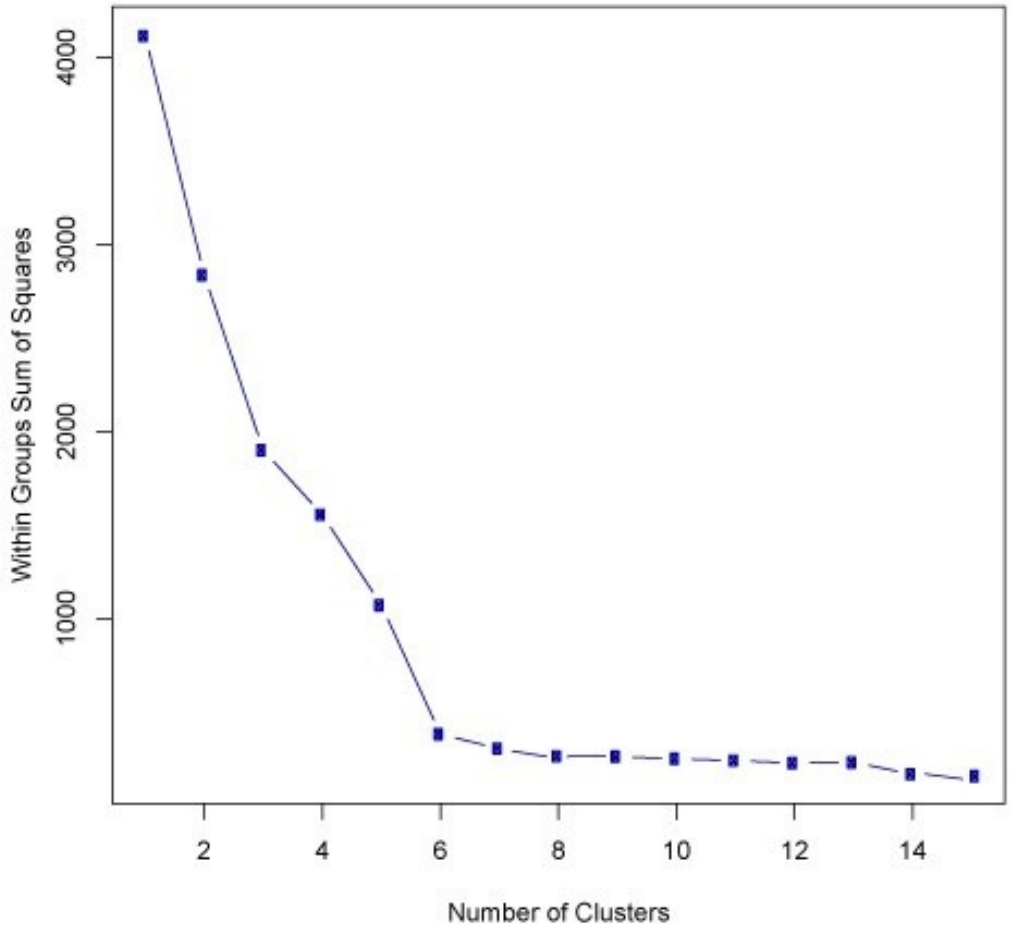
    centroids = median of each cluster

return centroids

Determine K: Elbow Method

Using some metric that approximates cluster quality, plot out the average value across values of K.

Elbow: K value that maximizes abs(metric(K-1)-metric(K)) /abs(metric(K)-metric(K+1))



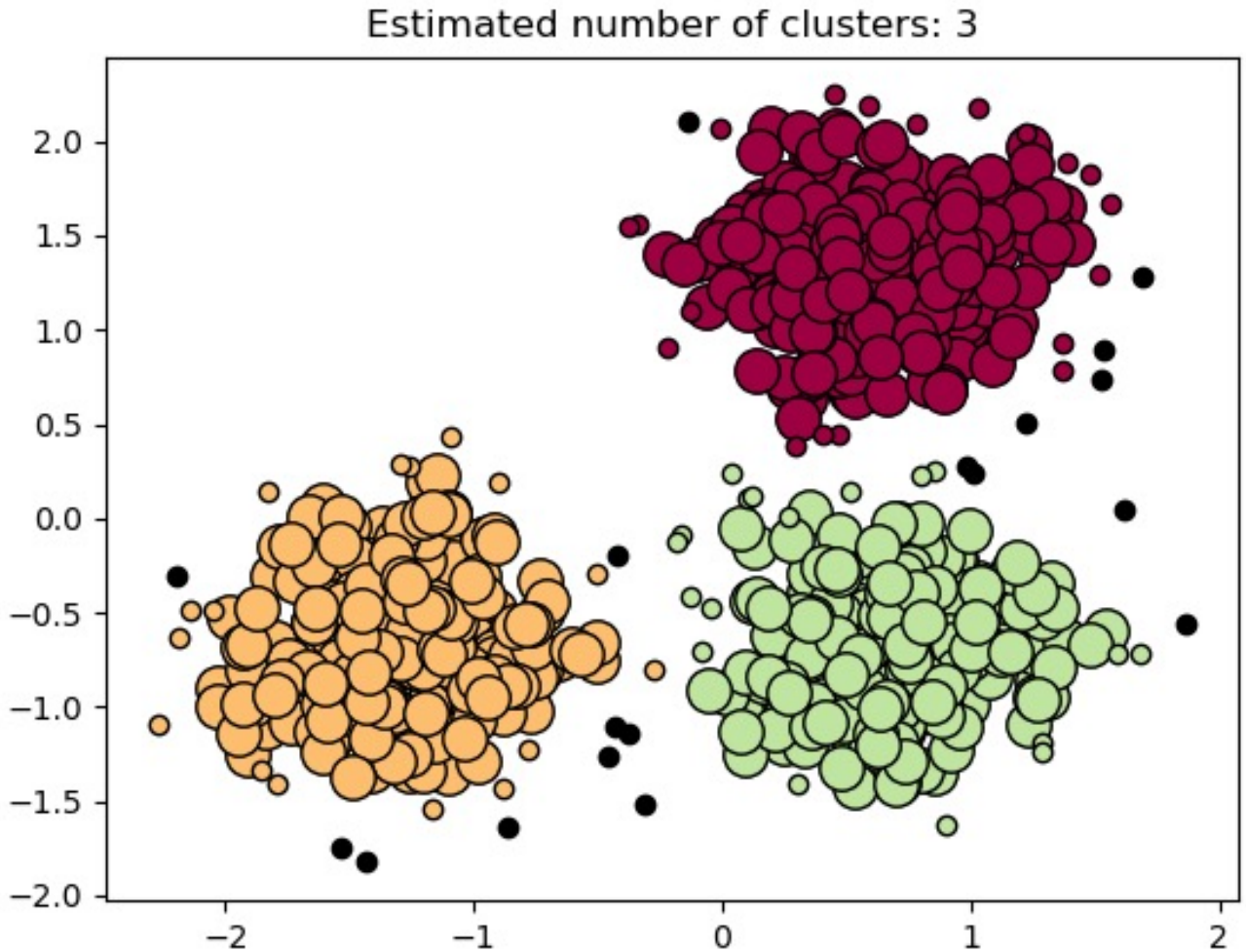
Elbow Method: Metrics

- **Average Distortion**: Average distance between each member of a cluster and its center.
- **Silhouette**: The degree to which elements of a cluster are more similar to their own cluster (cohesion) than to other clusters (separation)
- **Calinski-Harabasz Index:** Similar to Silhouette, but takes into account size of dataset and K.

Elbow Method: Metrics

- **Average Distortion: Average distance between each member of a cluster and its center.**
- **Silhouette**: The degree to which elements of a cluster are more similar to their own cluster (cohesion) than to other clusters (separation)
- **Calinski-Harabasz Index:** Similar to Silhouette, but takes into account size of dataset and K.

Visual Inspection



Other clustering techniques?

- There are a lot!
- But this isn't a machine learning course
- Plus…

K-Means Algorithms | Unsupervised Learning | Cluster analysis | Data Mining | Machine Learning

Why is k-means clustering so popular, given that more sophisticated alternatives exist?

1 Answer

Muktabh Mayank, Data Scientist,CoFounder @ ParallelDots, BITSian for life, love new technology trends

Answered Jan 23, 2015

The answer to any "why <a simpler Machine Learning technique> is more popular than <an advanced Machine Learning technique> is generally that the gains in performance for a complex algorithm are not worth the complexity of implementation for most people"

PQ1 (time permitting)

https://forms.gle/ZdBQtrpws6za5UMF8

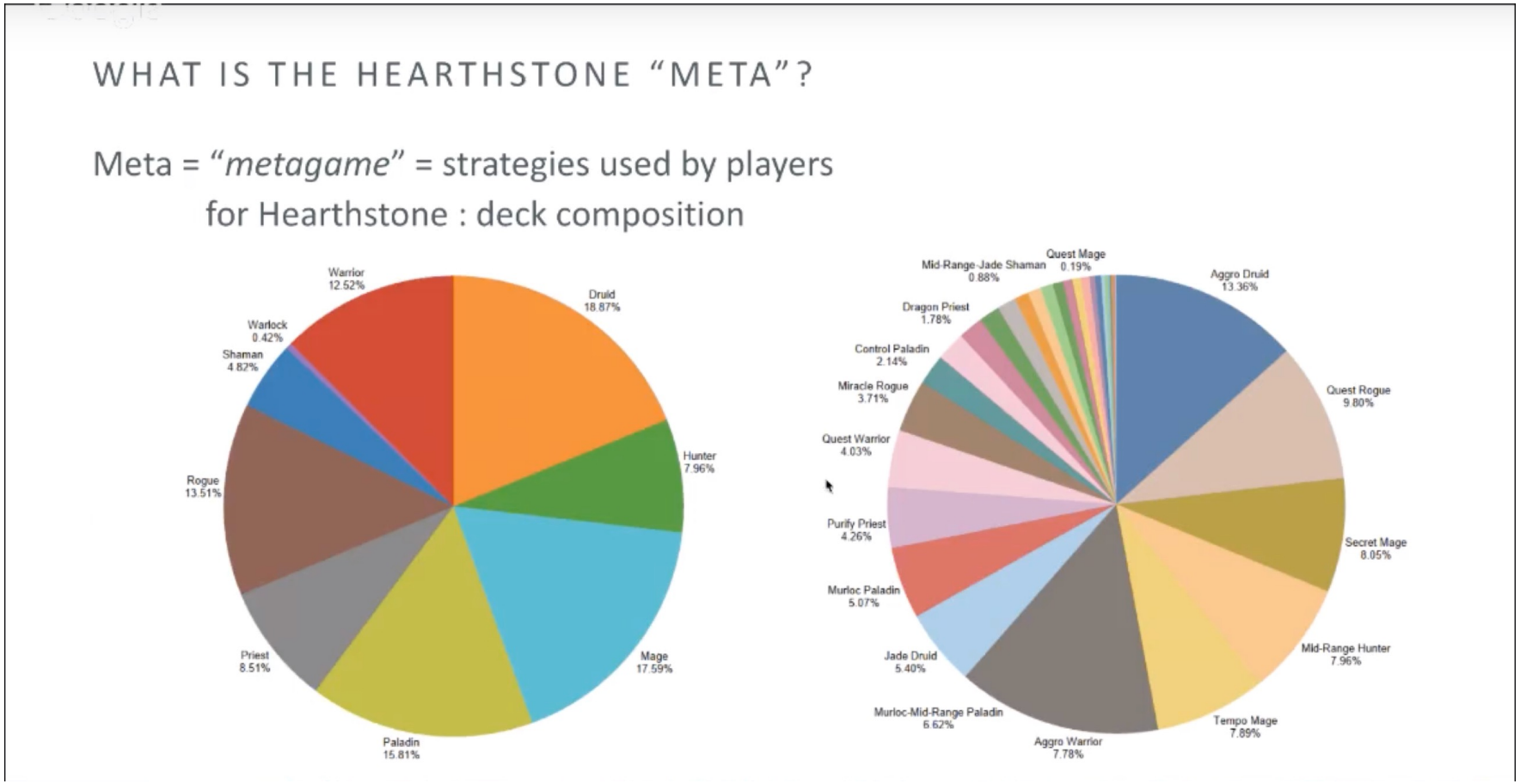
https://tinyurl.com/guz-pq22a

Think of a game and a particular element of that game (player strategy/player deaths) you'd apply clustering to.

What distance function would you use? Why?

How could these clusters benefit the game devs?

My answer: “Meta” Identification in Hearthstone (We will watch this video together later)



https://youtu.be/_YSYVRdzUkE?t=3558

Clustering Pros and Cons

- Pros:
  - Can learn “big picture” information
  - Can cut down on complexity
  - Can still give good answer with errors in training data
- Cons:
  - Individuals can get lost in the noise of groups
  - Those decisions to the open questions can hugely impact results
  - Slow to build
    - Even slower if we want to objectively decide on K
  - Slow to use (but this is offline, not in game)

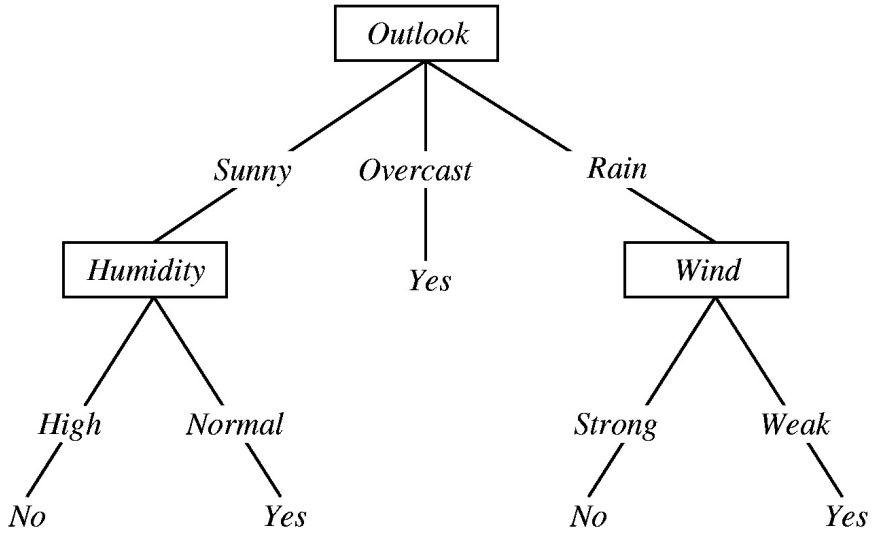
Open Questions

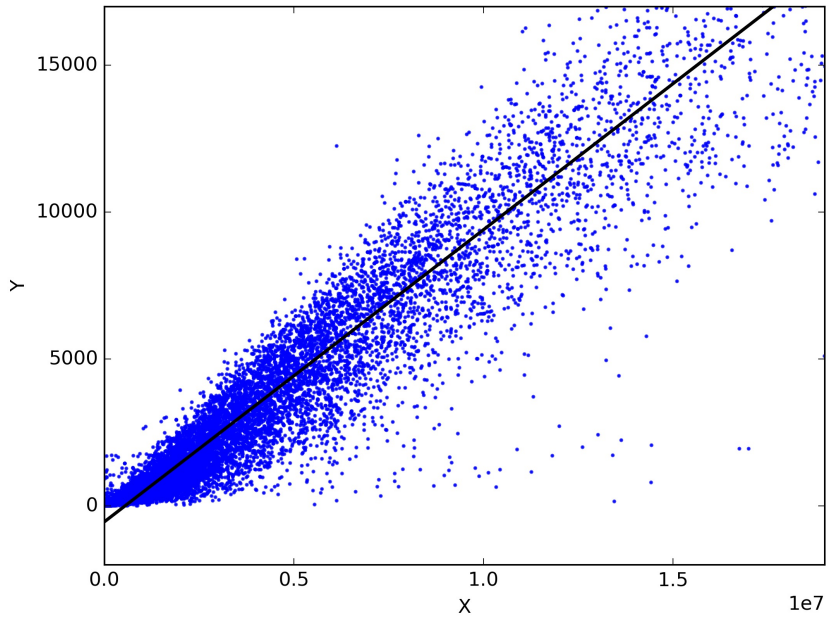
- How do we pick what variables to cluster on?
  - To normalize, or not to normalize?
- How do we pick a distance function?
- How do we use the clusters once we have them?

Most Commonly Desired Prediction? Churn

- Churn rate: The rate at which customers cut ties with a company
- In games this is how quickly a game loses players. After how long do they stop playing?
- Churn is one of the most common problems player analytics teams are tasked with

Wednesday: How can we predict churn?





Go over Assignment 4