CS341: Project in Mining Massive Datasets

Info session

Jure Leskovec
Chris Re
Anand Rajaraman
Rok Sosic
Jeff Ullman
Andreas Paepcke



CS341: Project in Data Mining

- **Data mining research project on real data**
  - Teams of **3 students** **(Use Piazza to form teams)**
  - We have room for **10-15 teams**
- **We provide:**
  - **Data**
  - **Computers** (Amazon EC2, **3k$ per team**)
  - **Mentoring:** Each group will have an assigned mentor that they meet on a weekly basis
- **You provide:**
  - Project proposals
  - Work

CS341: Schedule

- Today (3/12): Info session
- Monday 3/23: Project proposals due
- Mon 3/30: Admission results
  - 10 to 15 projects will be admitted
- Mon 3/30: First class meeting in 380-380F
- Mon 5/4 and Weds 5/6: Midterm presentations
- Week of June 1: Presentations, poster session

More info: http://cs341.stanford.edu

Project Types

- **Data analysis/Modeling project:**
  - Discovers interesting relationships within a significant amount of data
- **Algorithmic project that extends/builds on what was covered in CS246**
  - Extend/Improve/Speed-up some existing algorithm
  - Define a new problem and solve it

Projects: Proposal has to address

- **(1) What is the problem/question your team is solving?**
  - **Give a brief but precise description or definition of the problem or question**
    - **Examples:**
      - **(a)** Analyze the data to understand why editors are leaving Wikipedia
      - **(b)** Build a social recommender engine for movies
      - **(c)** Design a MapReduce algorithm for finding clusters in graphs
- **(2) What data will you use?**
  - **Why is the data you plan to use appropriate? Does it have the right labels/information?**
  - **It is ok to use your own data** (give detailed description)!
    - **Examples:**
      - **(a)** Wikipedia edit history where every action of every user is recorded
      - **(b)** We **crawled** Yelp and obtained X million reviews from Y million users
      - **(c)** We will use the Altavista web graph on X million nodes.

Projects: Proposal has to address

- **(3) How will you solve the problem? What is your plan of action?**
  - **Describe and think about your approach!**
    - What method, algorithm, technique? How will you scale it up?
    - **Be as specific as you can!**
    - **Examples:**
      - **(a)** We will create edit histories of every article. We will then compare article edit histories and argue that users are leaving since all the "easy/obvious" articles have already been written
      - **(b)** Our hypothesis is that friends have similar tastes. We will include a regularization term to a Latent Factor Rec. Sys. which will encourage neighboring users to have similar parameters
      - **(c)** We will implement a scalable Frequent-itemset-based approach to identify cluster seeds (complete bipartite subgraphs). In the second pass we will then use a random walk based approach to expand around the seed and extract the clusters

Projects: Proposal has to address

- **(4) How will you evaluate your method?**
  - **How will you measure performance or success of your method? What baselines will you use?**
    - **Examples:**
      - **(a)** Using insights from our analysis we will build a model that will predict how complete is the article (much the article will change in the future). We will evaluate predictive accuracy of the model
      - **(b)** We will measure **RMSE** of our system. As a baseline for comparison will use traditional latent factor recommender
      - **(c)** We will measure resource usage and execution time of our algorithm and compare it to open source algs. Metis and Graclus
- **(5) What do you expect to submit/accomplish by the end of the quarter?**

Projects: Proposals

- **Submit** to cs341-spr1415-staff@lists.stanford.edu
  - **PDF should include**
    - **Project title**
    - **Project narrative addressing the 5 Qs**
    - **Information about team members:**
      - **For each team member: 5 line CV/Bio** about prior experience, and why are you suitable to take this course
  - **No page limit** (but we don't promise to read past page 3)
  - **Due Monday 3/23 11:59pm Pacific time**
- **We will let you know whether you got in by Monday March 31**
- **More info at: http://cs341.stanford.edu/info.html**

Generic Places to Look

- Kaggle (www.kaggle.com) runs competitions.
  - You can get both data + ideas + possibly win.
- Yahoo (http://webscope.sandbox.yahoo.com/)
  - Interesting datasets, no problem suggestions, but some ideas should be obvious.
- TREC (http://trec.nist.gov/).
  - Current and historical competitions.
  - May take a week or more to get authorization for data.

Send in your proposals

- Proposals due Monday 3/23 midnight

- For more detail on a dataset or problem, please contact the appropriate instructor
  - Andreas Paepcke (paepcke@cs.stanford.edu)
  - Anand Rajaraman (datawocky@gmail.com)
  - Chris Re (chrismre@cs.stanford.edu)
  - Rok Sosic (rok@cs.stanford.edu)
  - Jeff Ullman (ullman@gmail.com)

Project ideas / datasets

(1) News Media (Jure)

- **Online media**
  - **Collection of 6B news documents and 300M short textual phrases that appear in them**
    - **Think of this as a complete trace of U.S. news media space for the last 5 years!**
- **Goal:**
  - Detect trending topics and explores the dynamics of online news
    - Based on time, named entities, mutation of information
- **If interested send email and we can talk more, rok@cs.stanford.edu**

(2) Large Graph Processing (Jure)

- **Distributed processing of large graphs**
  - System and algorithms for processing graphs
  - We have some ideas here, **talk to us**
- **Graph anomaly detection**
  - Imagine a communication network
    - Email, phone calls (we have both)
  - Can you spot and identify anomalies in communication patterns
  - **If interested send us email and we can talk more, rok@cs.stanford.edu**

(3) Online Product Reviews (Jure)

- **Ratings and reviews of Beer, Wine, Movies**
  - Plus user data, temporal information
  - http://snap.stanford.edu/data/
- **Questions:**
  - Evolution of user tastes and opinions
  - User contribution to the community
  - Beer is reviewed along **5** distinct dimensions
    - Taste, smell, feel, color and overall

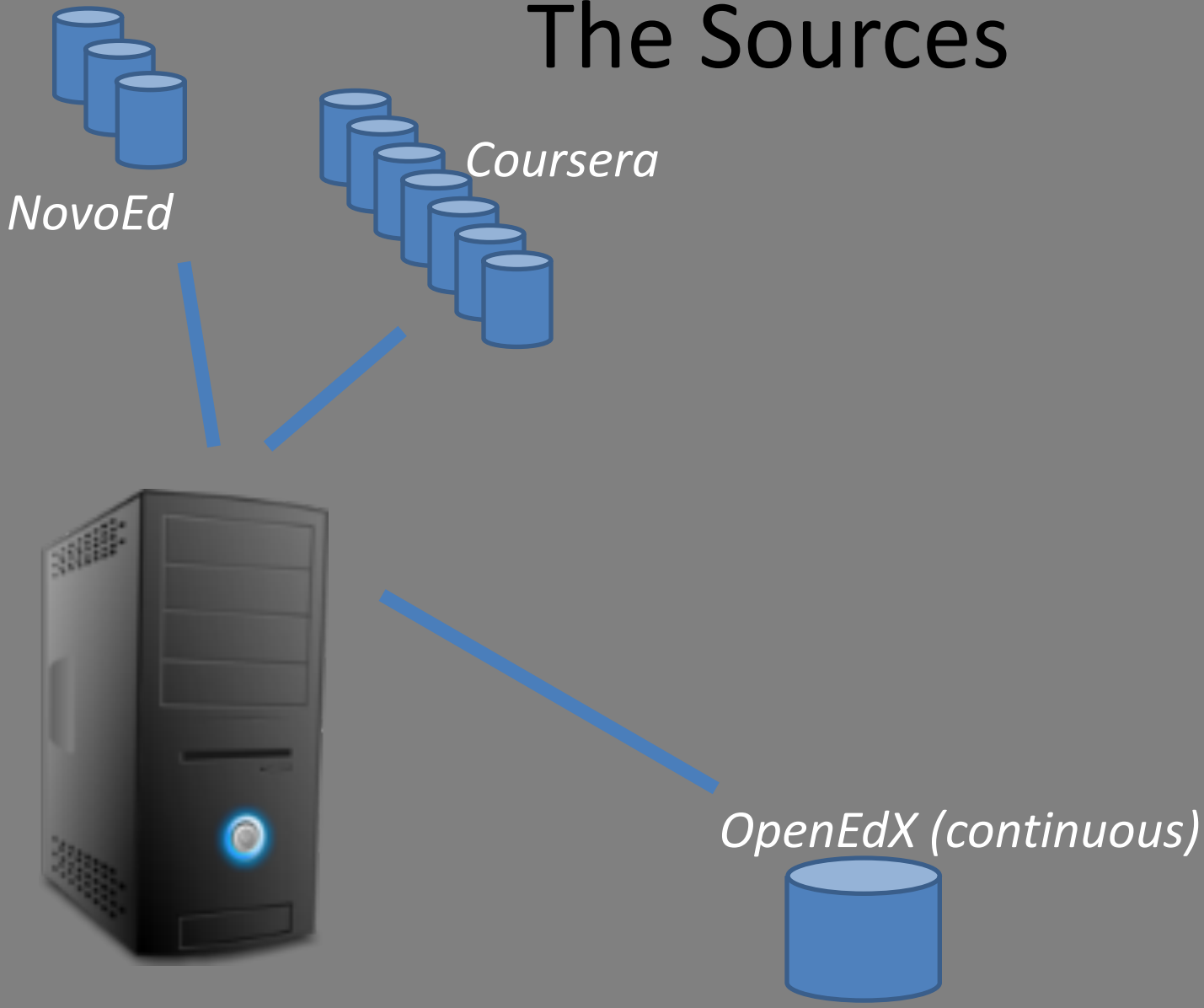
(4) Events and Locations (Jure)

- **Location Data from Instagram, Twitter, NYC Taxi**
  - 3.3B+ tweets
  - 426M photos
  - 187M taxi trips
- **Challenge - find interesting patterns**
  - Interesting locations
  - Anomalous events
  - Anomalous movements
  - **If interested send us email and we can talk more, rok@cs.stanford.edu**

OpenEdx

Andreas Paepcke

The Sources



datastage.stanford.edu

Learner Course Interaction Data

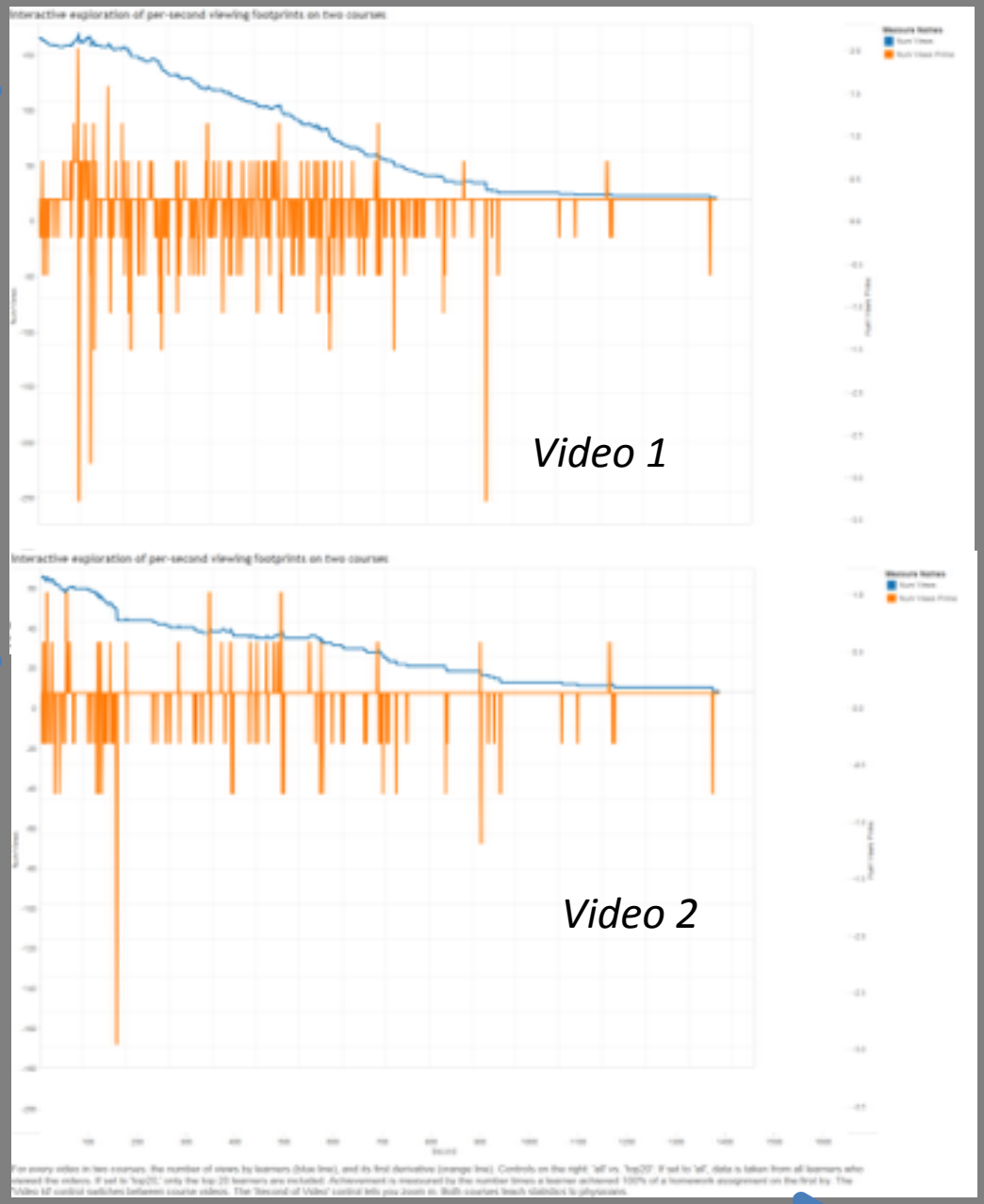
- Atomic events (~417M)
  - Video pause/stop/seek
  - Switch learning modules/pages
  - Problem checks
  - Time
- Submit assignment (~27M)
  - Correct/wrong/#Attempts
- Country of interaction origin (~580K)

More Data

- Demographics (~690K but self-reported)
  - Gender
  - Year of birth
  - Level of education
- Forum contents (~127K)
  - Time
  - Text
  - Up/down votes
  - Comments

Questions to Ask

- Forum participation ↔ learner success?
- Demographics ↔forum participation?
- Demographics across courses?
- Nature of forum posts *across* courses?
- Do forum topics change between course *runs*?
- Discover/define affinity groups?
- Video viewing profiles ↔learner success?



Number of views

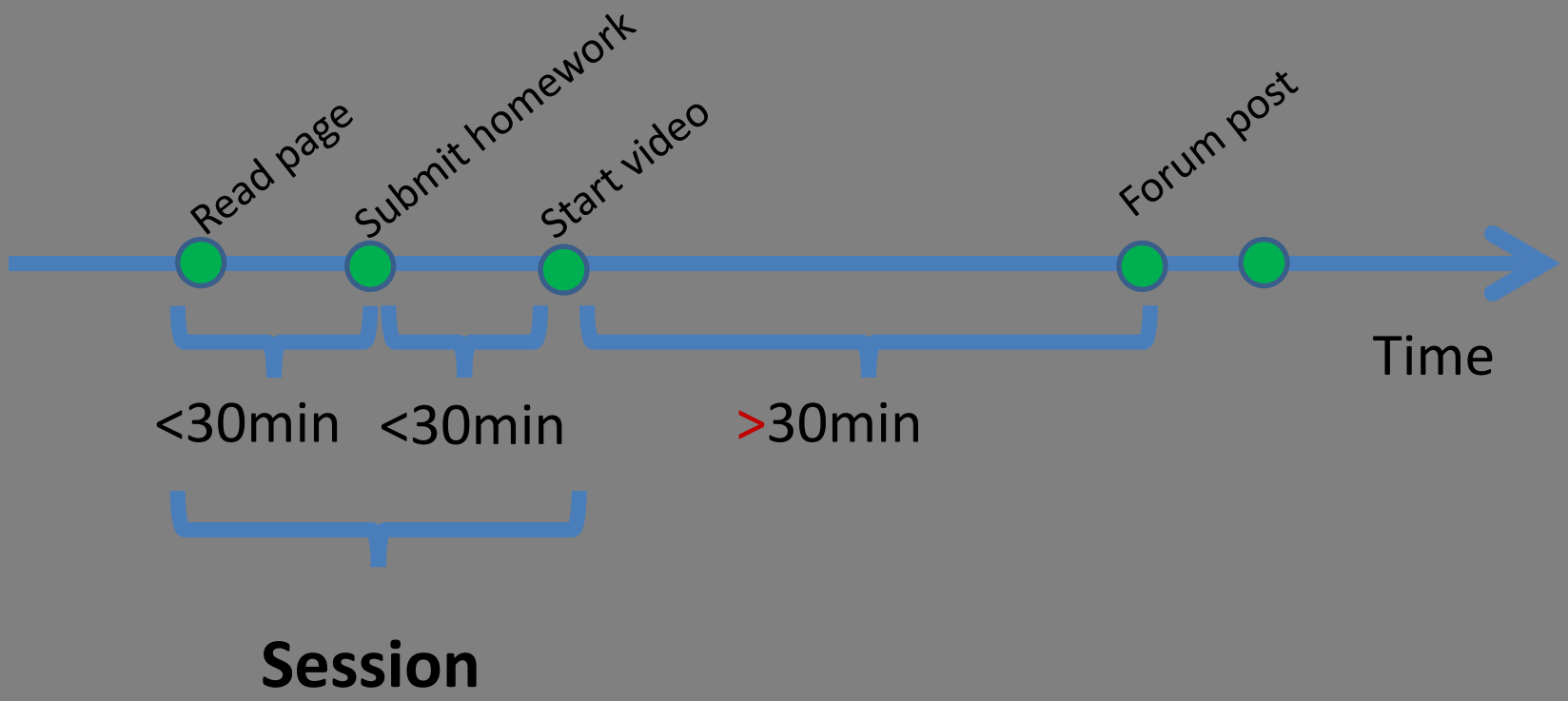
dV/dt

Time into video

More Questions

- Draw paths learners take through material
  - Web pages
  - Course modules
  - Videos
- Engagement…

Learner Engagement



Data sets from Rocketship.vc

smart investing through data science

Sailesh Ramakrishnan
sailesh@rocketship.vc

Context

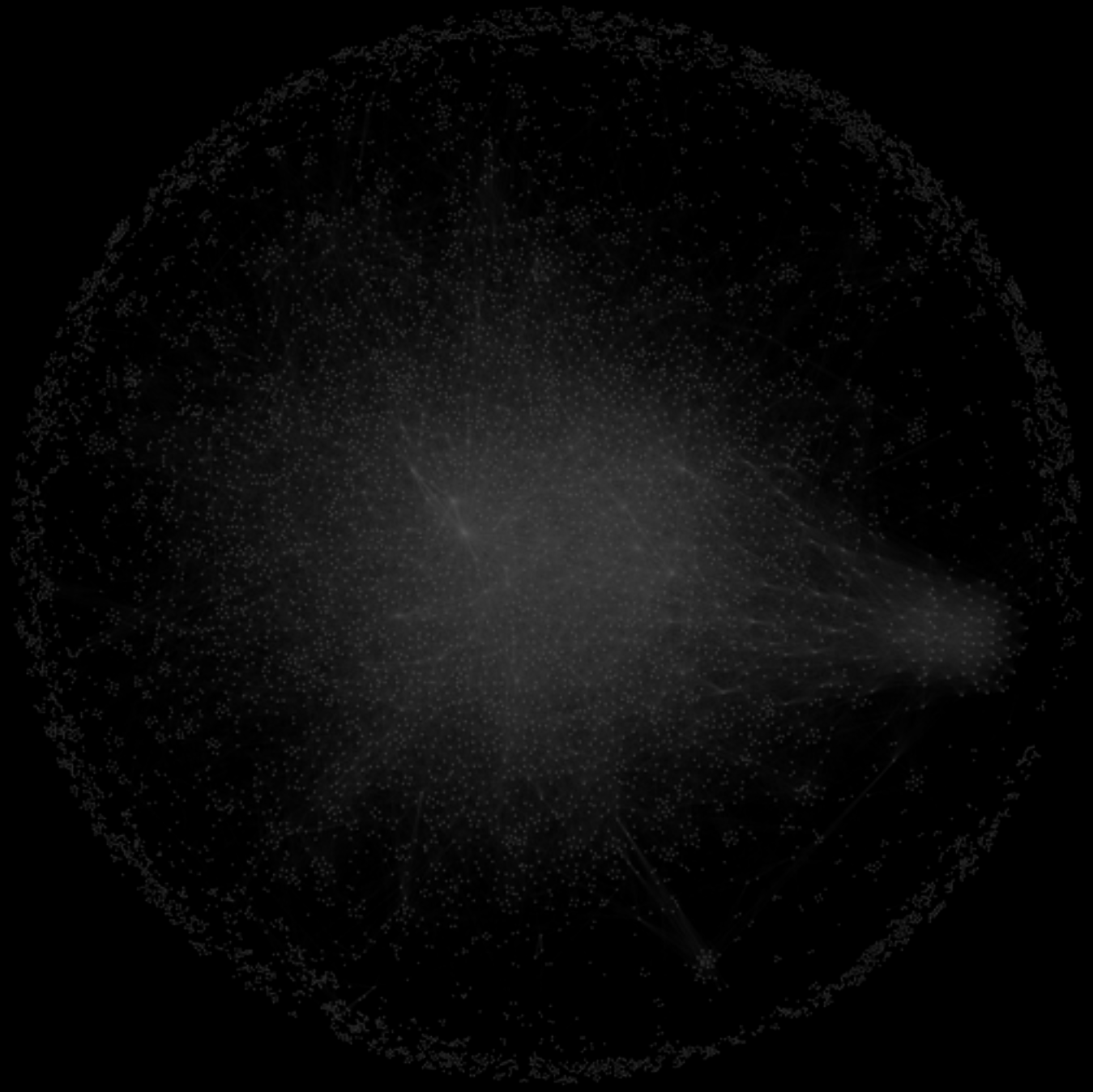
- Identify companies/products with sustainable growth
- Convince companies to accept our investment
- Help companies succeed post investment
- Do all of the above though data science

Companies

- Startups, Users
- Users follow other users, startups
- Users are followed by other users
- Users have many different relationships with startups (Investor, advisor, board member)

Problem: Can we predict the likelihood of a startups success based on the about graph?

Success = Next round of funding, Size of Funding, Exit etc



Twitter

- 10% of the twitter firehose
- Data since 2014-02-01 (~ 10 TB)
- Tweets = stream of consciousness

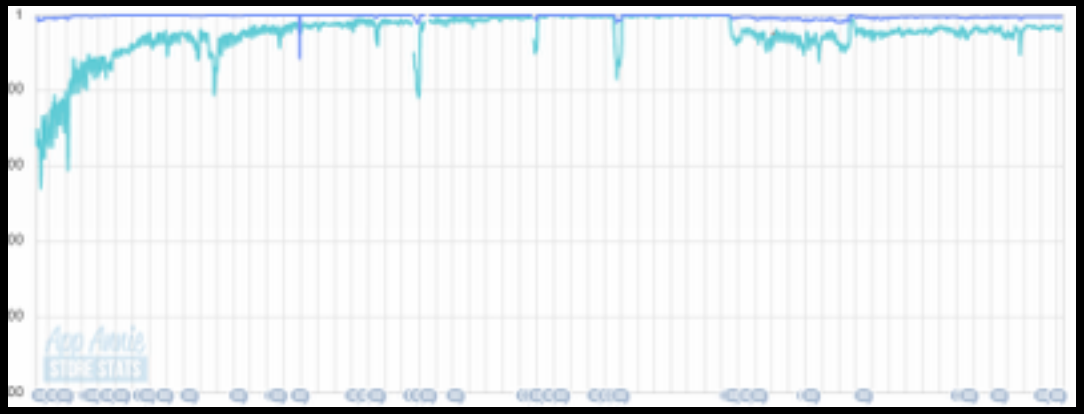
Problem: Can we predict the next break out company or product from the twitter stream?

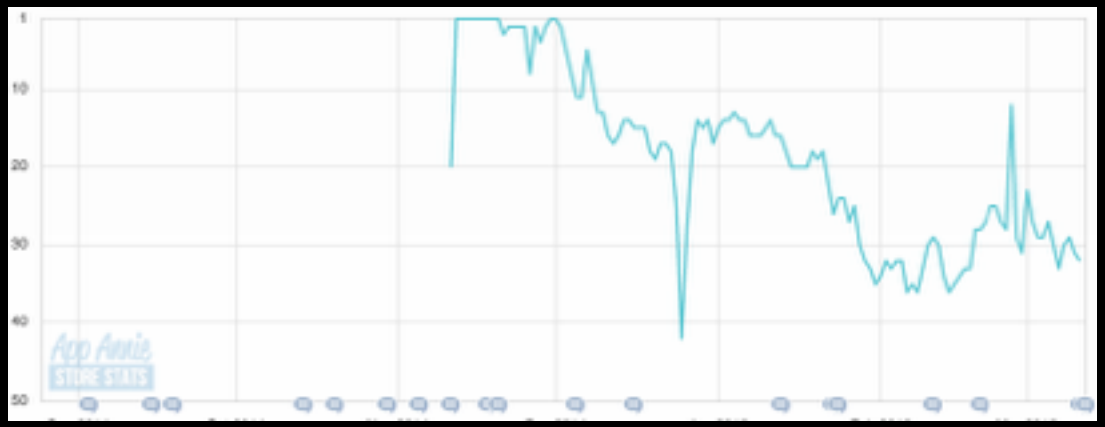
```
{
  "body": "De la mente :))",
  "retweetCount": 0,
  "generator": {
    "link": "http:\/\/twitter.com\/download\/android",
    "displayName": "Twitter for Android"
  },
  "twitter_filter_level": "low",
  "gnip": {
    "language": {
      "value": "es"
    }
  },
 ....,
  "actor": {
    "preferredUsername": "LuqiitasJohan",
    "displayName": "Vaguiitto #1 ",
    "links": [
      {
        "href": null,
        "rel": "me"
      }
    ],
    "twitterTimeZone": "Brasilia",
    "image": "https:\/\/pbs.twimg.com\/profile_images\/570965583912693761\/tdMoinLr_normal.jpeg",
    "verified": false,
    "statusesCount": 3227,
    "summary": "Un wacho piiola JAJA ;)",
    "languages": [
      "es"
    ],
    "utcOffset": "-10800",
    "link": "http:\/\/www.twitter.com\/LuqiitasJohan",
    "followersCount": 176,
```

App Data

- iOS and Google Play
- Over 1.3 M apps
- Downloads, Paid Downloads, Revenue estimates
- Data from 2013-01-01

Problem: Can we predict the next break out app from this data?


App Annie
STORE STATS


App Annie
STORE STATS

Company Datasets

- OpsClarity
  - Contact: "Emre Velipasaoglu" <emre@opsclarity.com>
- TwinPrime
  - Contact: "Tejaswini Ganapathi" <tejaswini@twinprime.com>