What is Data Science?

CMSC320 Spring 2017
Hector Corrada Bravo
University of Maryland

For today

- What is data science?
- One use case
- Check on R installation

Why Data Science?

- “I keep saying that the sexy job in the next 10 years will be statisticians”
  - Hal Varian, Chief Economist at Google
  - (http://www.nytimes.com/2009/08/06/technology/06stats.html?_r=0)

Why data science?

- “The ability to take data—to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it—that’s going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids.”

- Hal Varian

    - (http://www.mckinsey.com/insights/innovation/hal_varian_on_how_the_web_challenges_managers)

Why Data Science

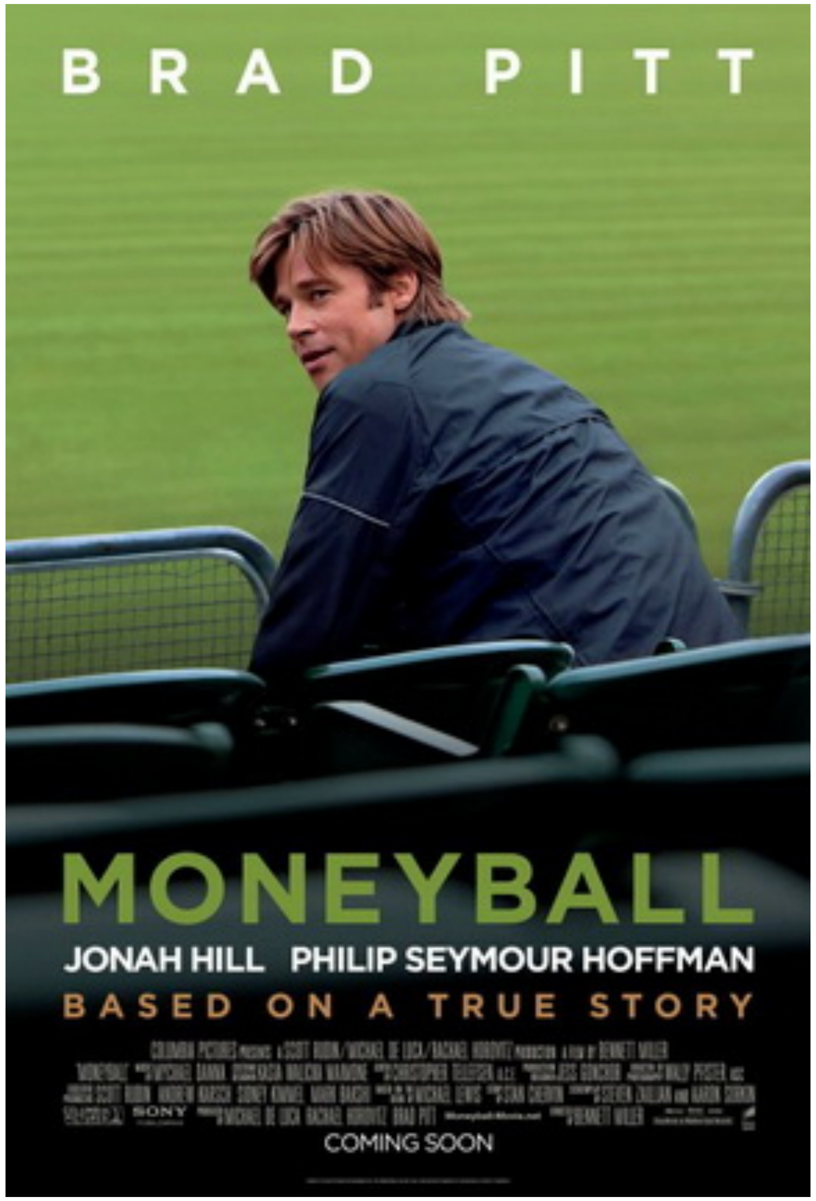
- “Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it.”

- Hal Varian

  - (http://www.mckinsey.com/insights/innovation/hal_varian_on_how_the_web_challenges_managers)

Data Science Success Stories

Rafael Irizarry, http://cs109.github.io/2014/

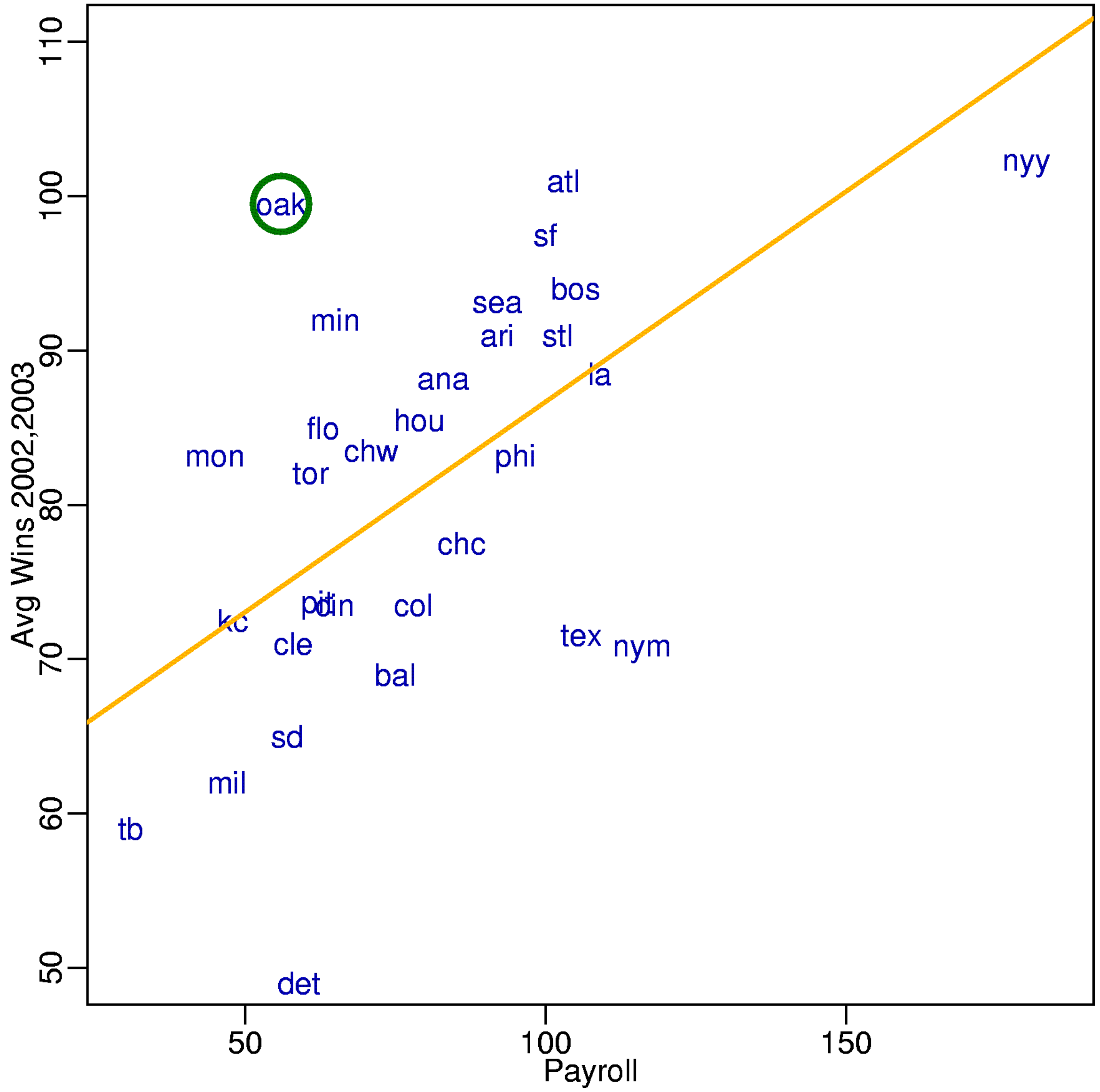

BRAD PITT
MONEYBALL
JONAH HILL PHILIP SEYMOUR HOFFMAN
BASED ON A TRUE STORY
COLUMBIA PICTURES presents a SCOTT RUDIN/MICHAEL DE LUCA/RACHAEL HOROVITZ production a film by BENNETT MILLER
SONY
Moneyball-Movie.net
COMING SOON

The Data Scientist

Actual

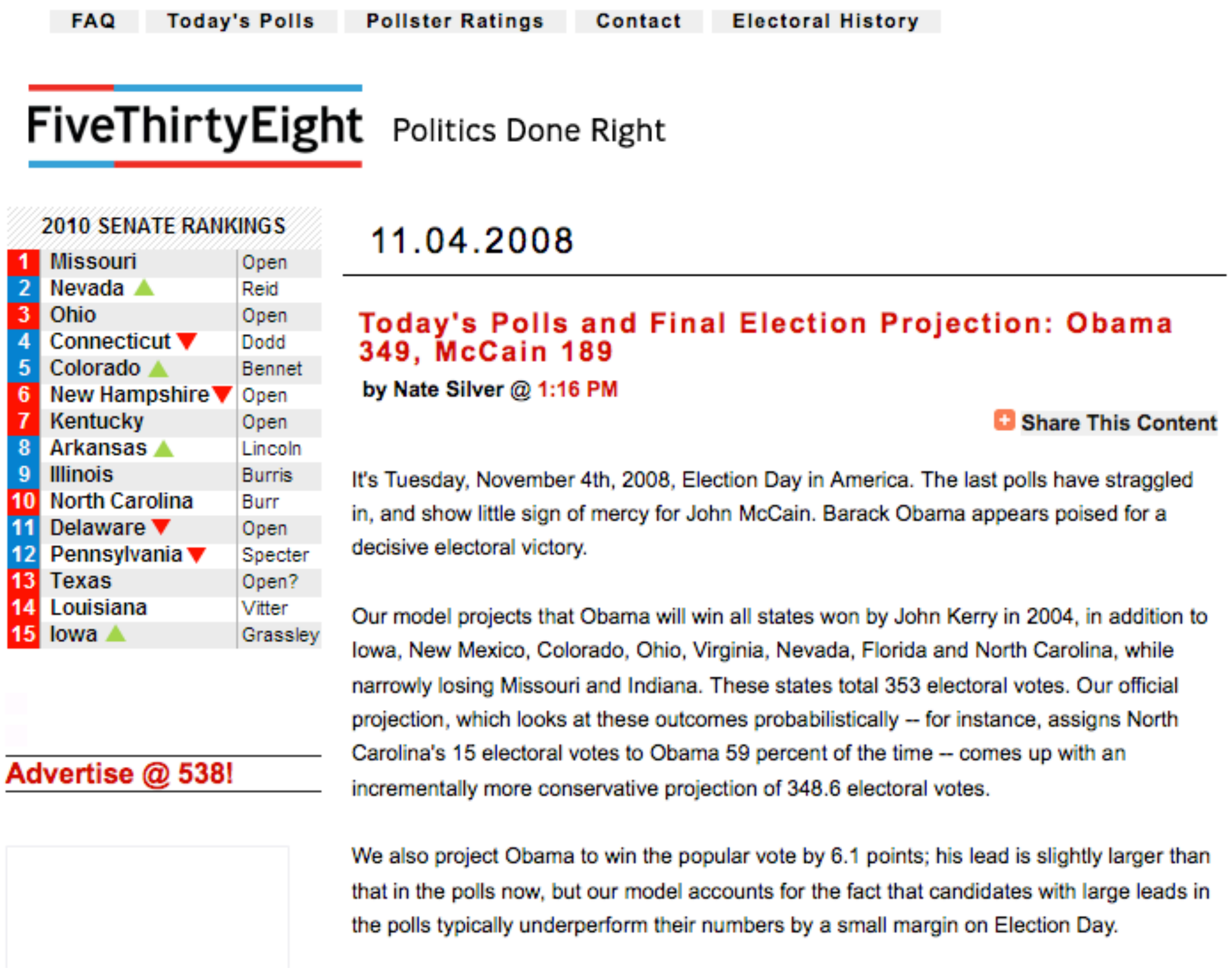
Hollywood

Money Ball



Starting around 2001, the Oakland A's picked players that scouts thought were no good but data said otherwise

“Nate Silver won the election” – Harvard Business Review



2010 SENATE RANKINGS

1	Missouri	Open
2	Nevada ▲	Reid
3	Ohio	Open
4	Connecticut ▼	Dodd
5	Colorado ▲	Bennet
6	New Hampshire ▼	Open
7	Kentucky	Open
8	Arkansas ▲	Lincoln
9	Illinois	Burris
10	North Carolina	Burr
11	Delaware ▼	Open
12	Pennsylvania ▼	Specter
13	Texas	Open?
14	Louisiana	Vitter
15	Iowa ▲	Grassley

11.04.2008

Today's Polls and Final Election Projection: Obama 349, McCain 189

by Nate Silver @ 1:16 PM

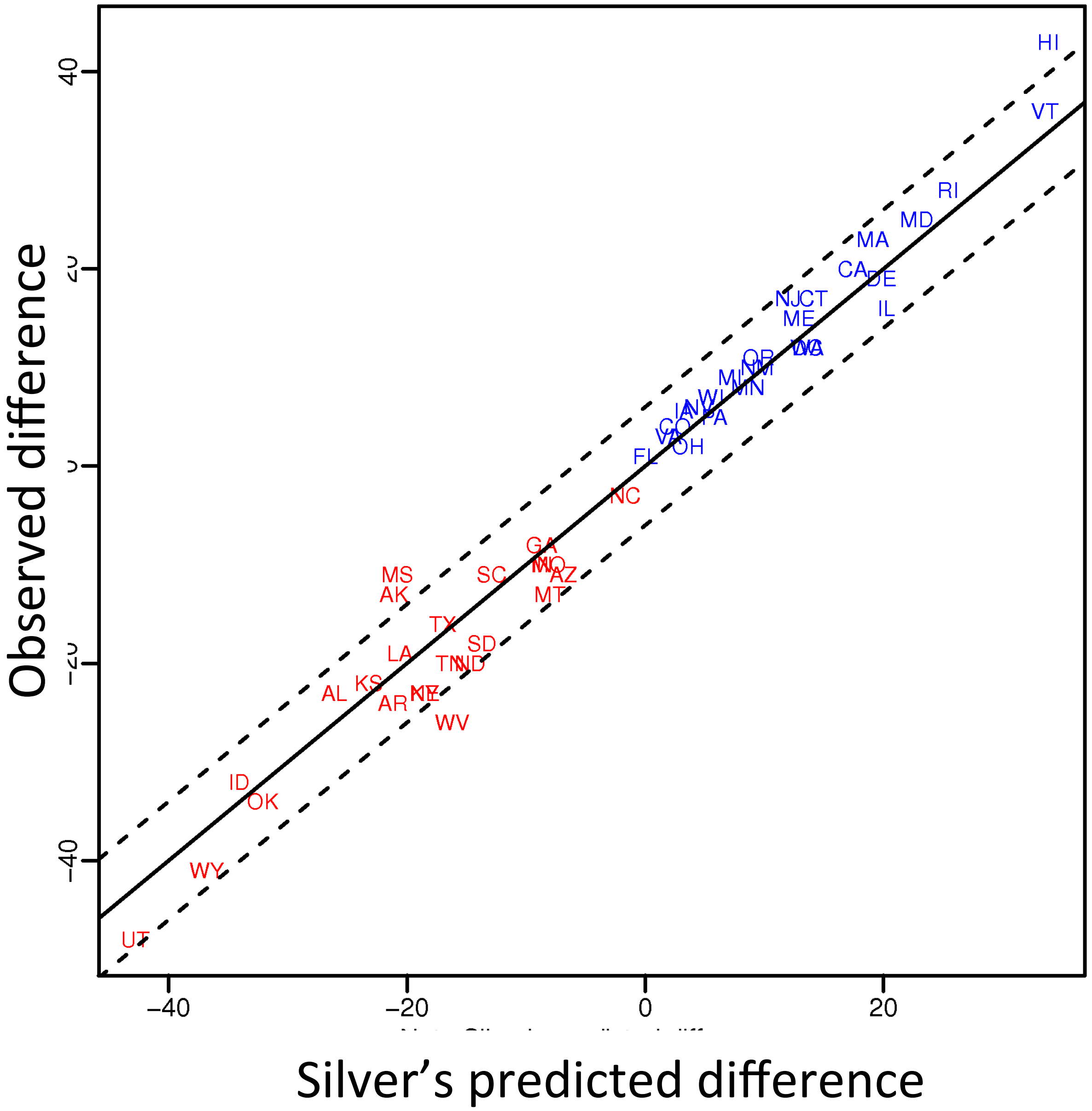
It's Tuesday, November 4th, 2008, Election Day in America. The last polls have straggled in, and show little sign of mercy for John McCain. Barack Obama appears poised for a decisive electoral victory.

Our model projects that Obama will win all states won by John Kerry in 2004, in addition to Iowa, New Mexico, Colorado, Ohio, Virginia, Nevada, Florida and North Carolina, while narrowly losing Missouri and Indiana. These states total 353 electoral votes. Our official projection, which looks at these outcomes probabilistically -- for instance, assigns North Carolina's 15 electoral votes to Obama 59 percent of the time -- comes up with an incrementally more conservative projection of 348.6 electoral votes.

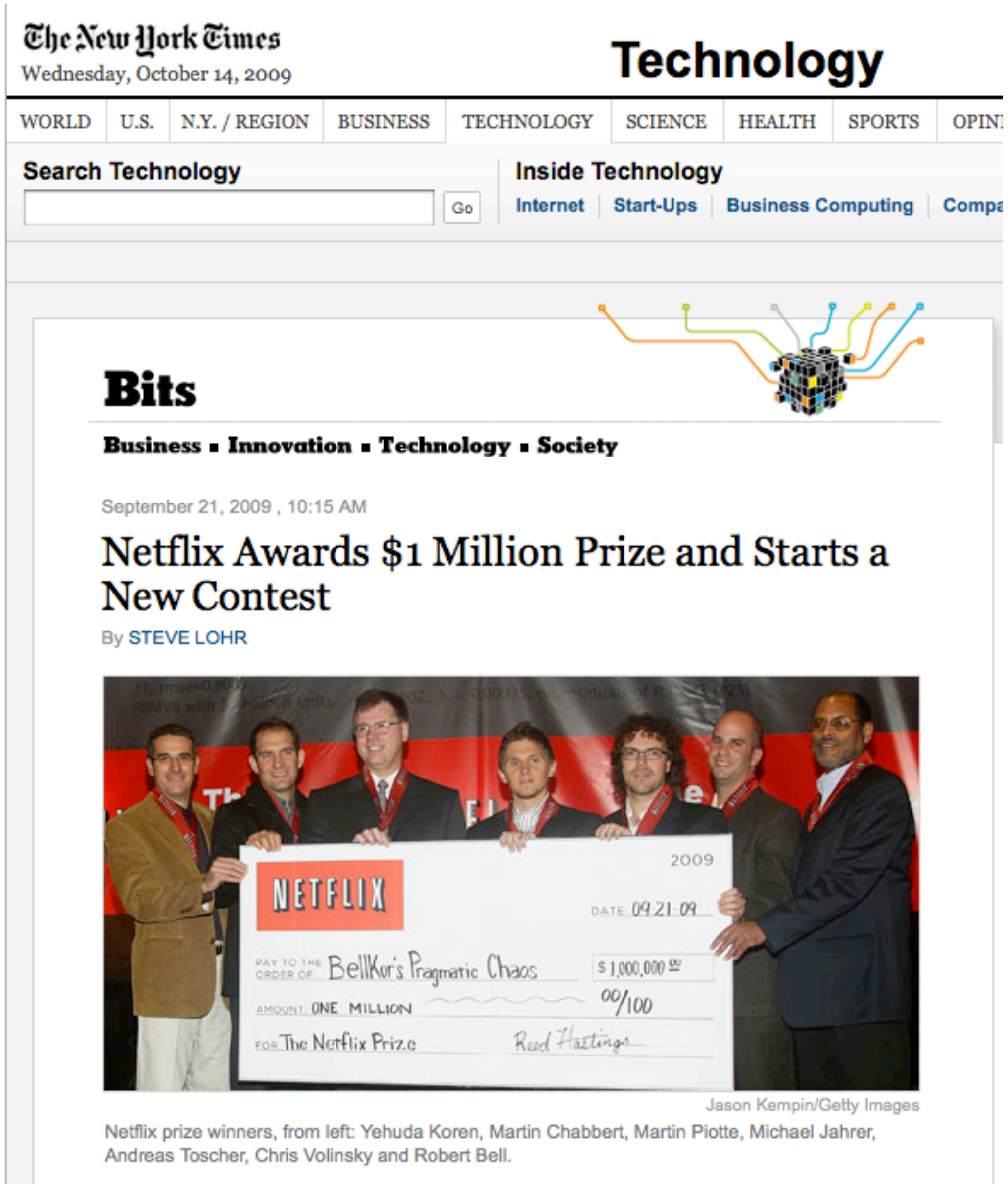
We also project Obama to win the popular vote by 6.1 points; his lead is slightly larger than that in the polls now, but our model accounts for the fact that candidates with large leads in the polls typically underperform their numbers by a small margin on Election Day.

Prediction: 349 to 189, 6.1% difference.
Actual: 365 to 173, 7.2% difference

2012 results



Netflix Challenge



September 21, 2009 , 10:15 AM

Netflix Awards $1 Million Prize and Starts a New Contest

By STEVE LOHR

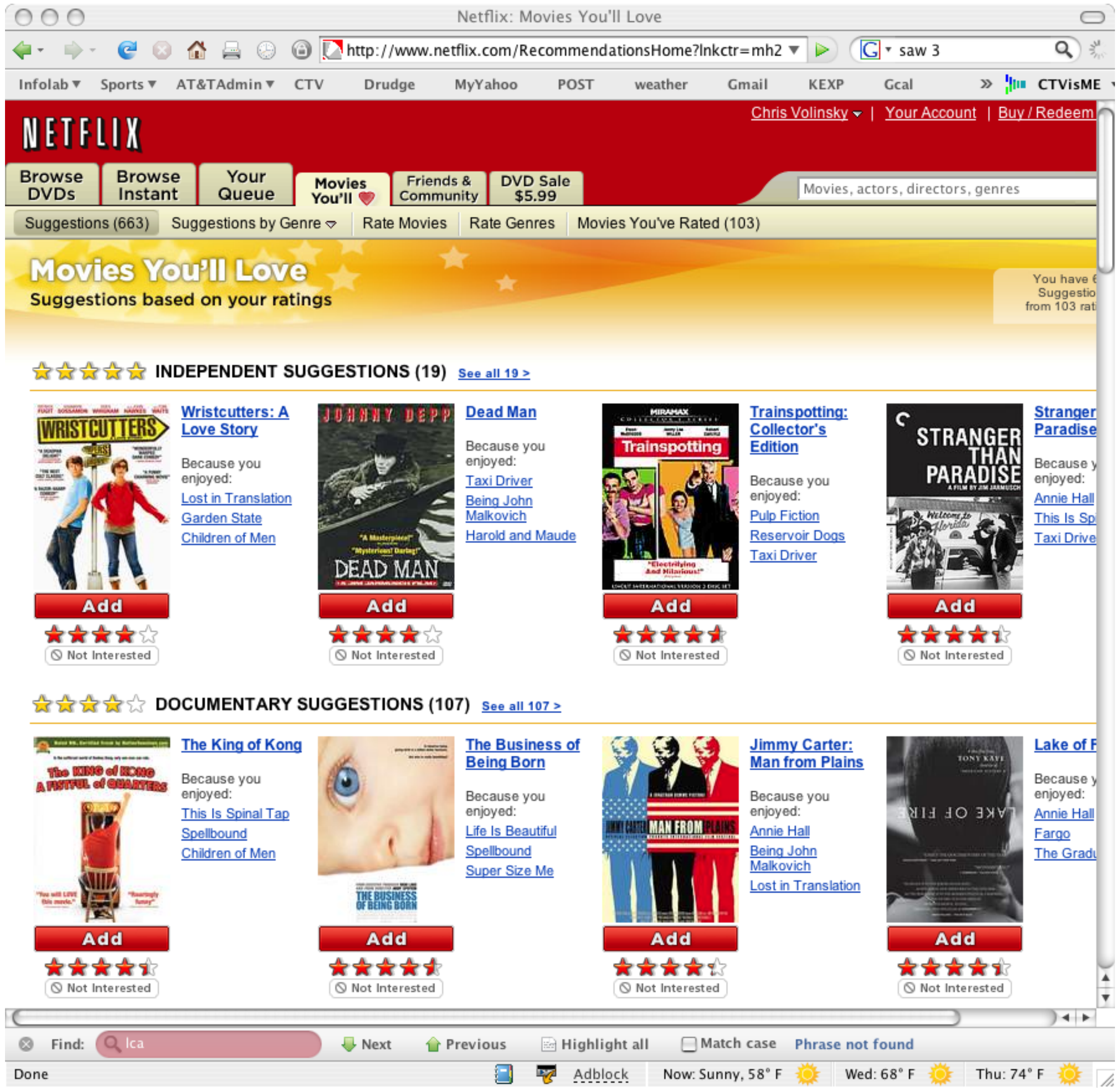
Jason Kempin/Getty Images

Netflix prize winners, from left: Yehuda Koren, Martin Chabbert, Martin Piotte, Michael Jahrer, Andreas Toscher, Chris Volinsky and Robert Bell.

In Sept 2009 a team lead by Chris Volinsky from Statistics Research AT&T Research was announced as winner!

Netflix

- A US-based DVD rental-by mail company
- >10M customers, 100K titles, ships 1.9M DVDs per day



Movies You'll Love

Suggestions based on your ratings

You have 6… Suggestio… from 103 rati…

★★★★★ INDEPENDENT SUGGESTIONS (19) See all 19 >

Wristcutters: A Love Story — Because you enjoyed: Lost in Translation, Garden State, Children of Men

Dead Man — Because you enjoyed: Taxi Driver, Being John Malkovich, Harold and Maude

Trainspotting: Collector's Edition — Because you enjoyed: Pulp Fiction, Reservoir Dogs, Taxi Driver

Stranger Paradise — Because y… enjoyed: Annie Hall, This Is Sp…, Taxi Drive…

★★★★☆ DOCUMENTARY SUGGESTIONS (107) See all 107 >

The King of Kong — Because you enjoyed: This Is Spinal Tap, Spellbound, Children of Men

The Business of Being Born — Because you enjoyed: Life Is Beautiful, Spellbound, Super Size Me

Jimmy Carter: Man from Plains — Because you enjoyed: Annie Hall, Being John Malkovich, Lost in Translation

Lake of F… — Because y… enjoyed: Annie Hall, Fargo, The Gradu…

Good recommendations = happy customers

Courtesy of Chris Volinsky

Netflix Prize

- October, 2006:
  - Offers **$1,000,000** for an improved recommender algorithm

- Training data
  - 100 million ratings
  - 480,000 users
  - 17,770 movies
  - 6 years of data: 2000-2005

- Test data
  - Last few ratings of each user (2.8 million)
  - Evaluation via RMSE: root mean squared error
  - Netflix Cinematch RMSE: 0.9514

- Competition
  - **$1 million** grand prize for **10% improvement**
  - If 10% not met, $50,000 annual “Progress Prize” for best improvement

user	movie	score	date
1	21	1	2002-01-03
1	213	5	2002-04-04
2	345	4	2002-05-05
2	123	4	2002-05-05
2	768	3	2003-05-03
3	76	5	2003-10-10
4	45	4	2004-10-11
5	568	1	2004-10-11
5	342	2	2004-10-11
5	234	2	2004-12-12
6	76	5	2005-01-02
6	56	4	2005-01-31

Courtesy of Chris Volinsky

Netflix Prize

- October, 2006:
  - Offers **$1,000,000** for an improved recommender algorithm

- Training data
  - 100 million ratings
  - 480,000 users
  - 17,770 movies
  - 6 years of data: 2000-2005

- Test data
  - Last few ratings of each user (2.8 million)
  - Evaluation via RMSE: root mean squared error
  - Netflix Cinematch RMSE: 0.9514

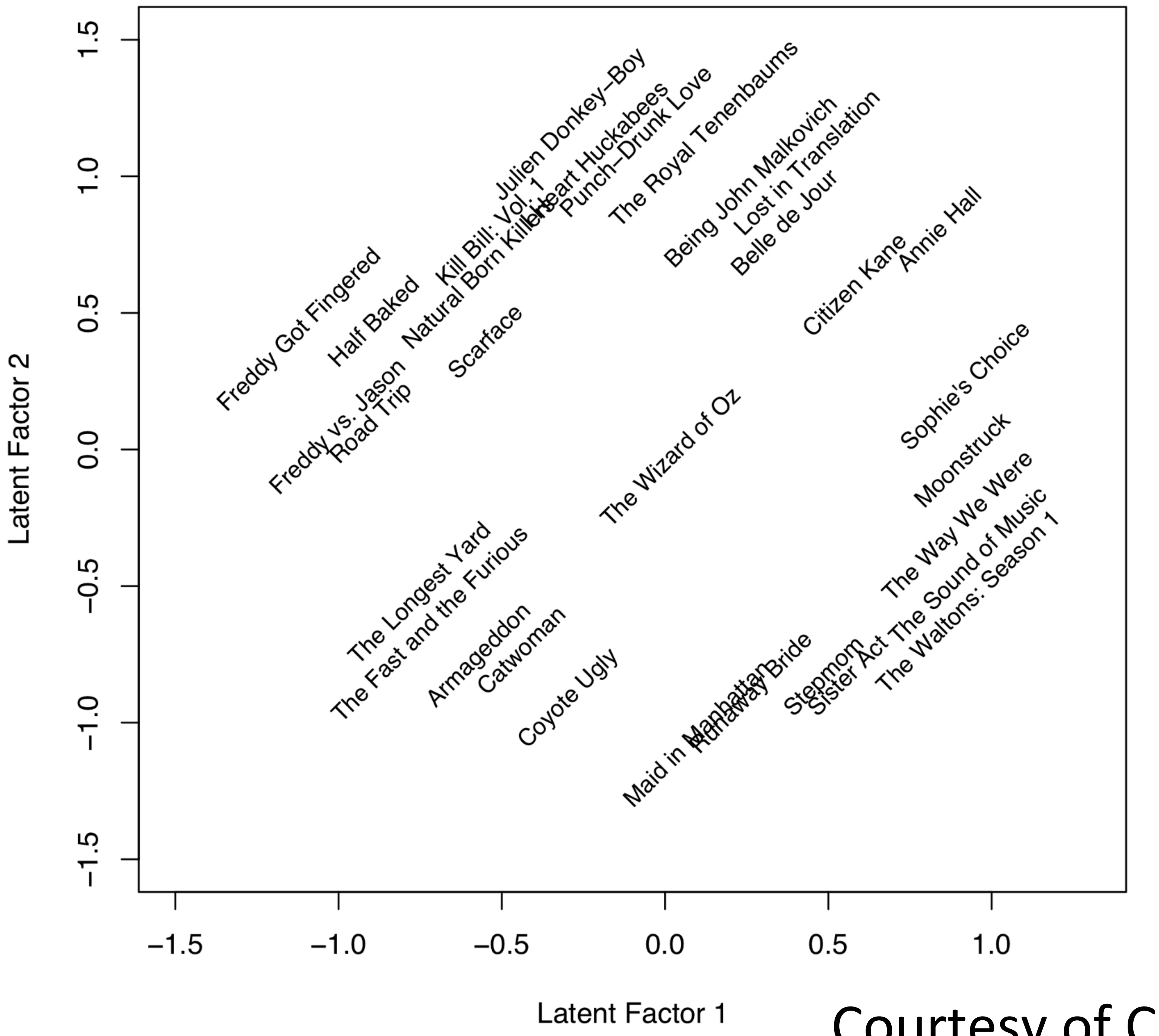
user	movie	score	date
1	21	1	2002-01-03

user	movie	score	date
1	212	**?**	2003-01-03
1	1123	**?**	2002-05-04
2	25	**?**	2002-07-05
2	8773	**?**	2002-09-05
2	98	**?**	2004-05-03
3	16	**?**	2003-10-10
4	2450	**?**	2004-10-11
5	2032	**?**	2004-10-11
5	9098	**?**	2004-10-11
5	11012	**?**	2004-12-12
6	664	**?**	2005-01-02
6	1526	**?**	2005-01-31

- Competition
  - **$1 million** grand prize for **10% improvement**
  - If 10% not met, $50,000 annual "Progress Prize" for best improvement

Courtesy of Chris Volinsky

Latent Factors Model

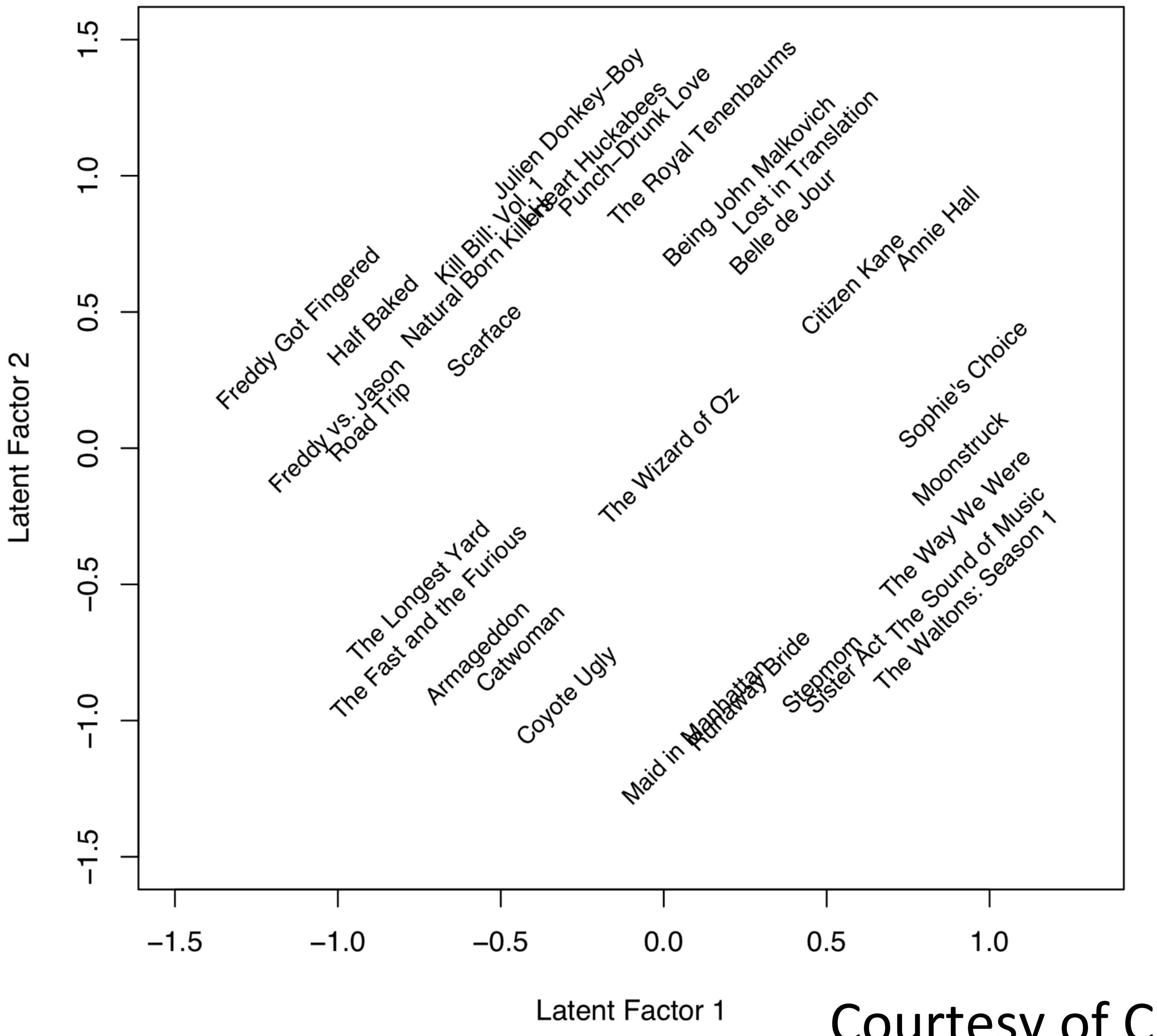


A **latent factors** model identifies factors with maximum discrimination between movies

Courtesy of Chris Volinsky

Latent Factors Model

Frat-house, gross-out comedies ⟷ Critically – acclaimed/ Strong female leads



A **latent factors** model identifies factors with maximum discrimination between movies

Courtesy of Chris Volinsky

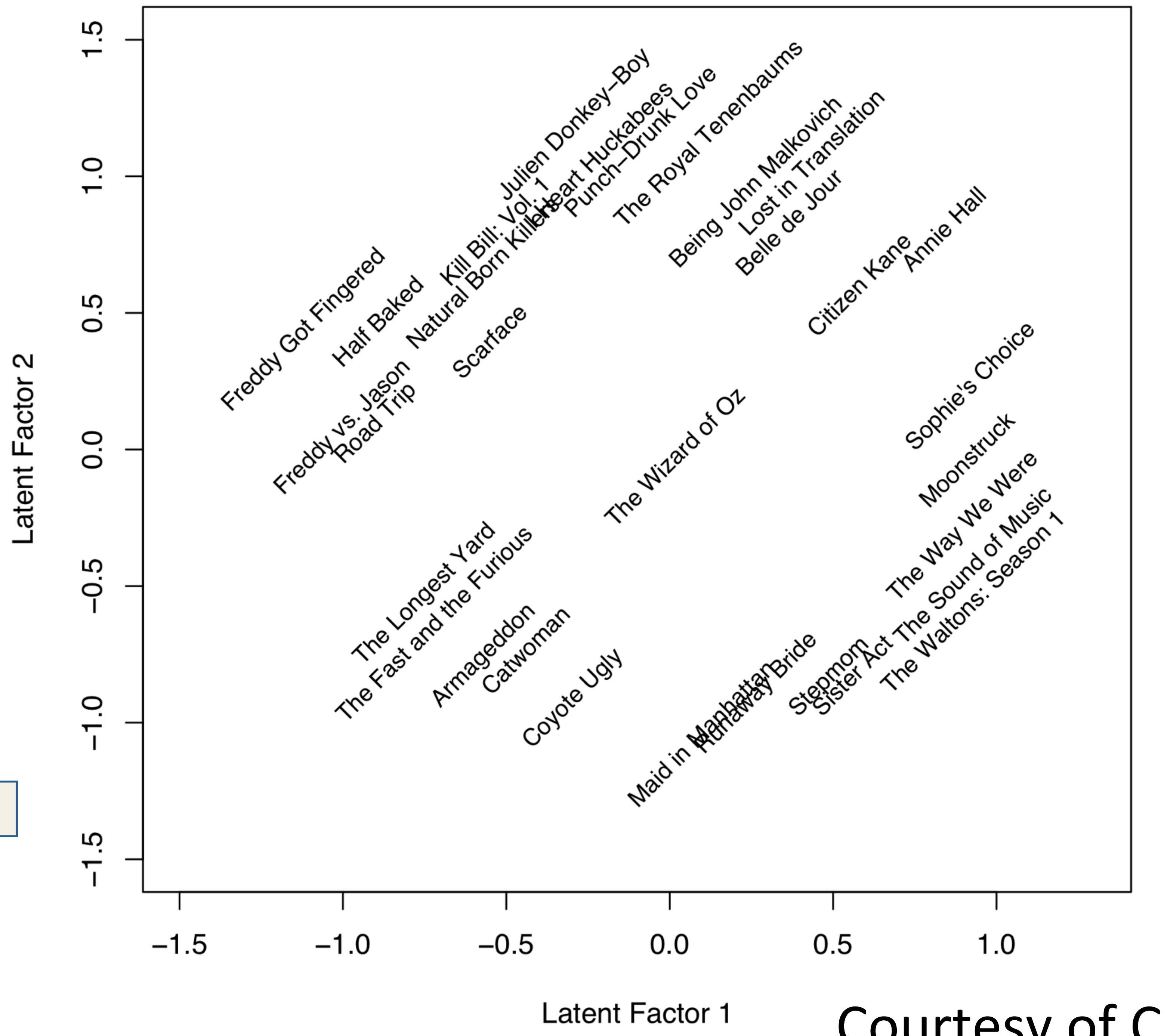
Latent Factors Model

Frat-house, gross-out comedies ⟷ Critically – acclaimed/ Strong female leads

Artsy, edgy, movies

↓

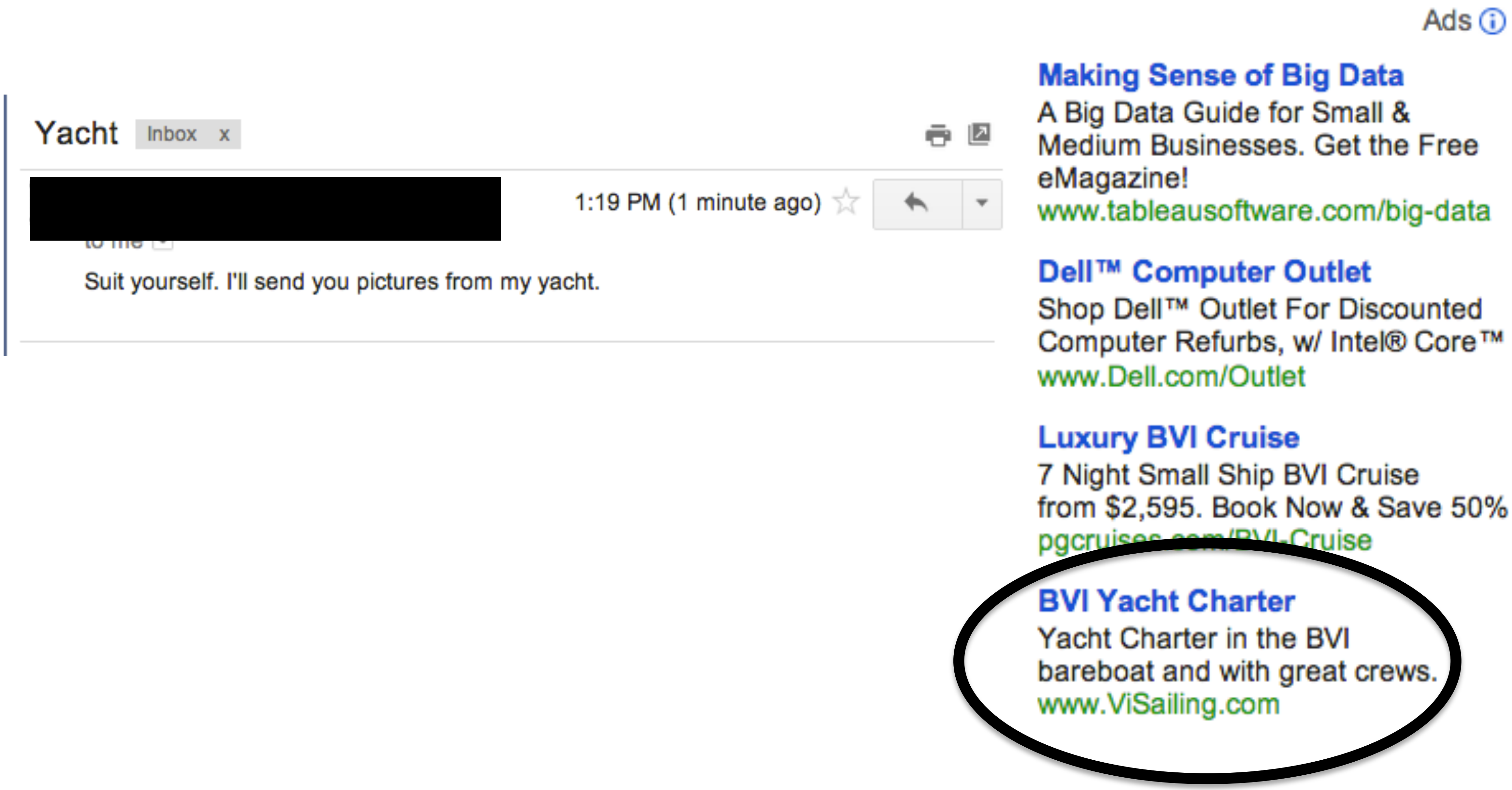
Hollywood Big budget



A **latent factors** model identifies factors with maximum discrimination between movies

Courtesy of Chris Volinsky

Ad-targeting



Yacht

[redacted] 1:19 PM (1 minute ago)

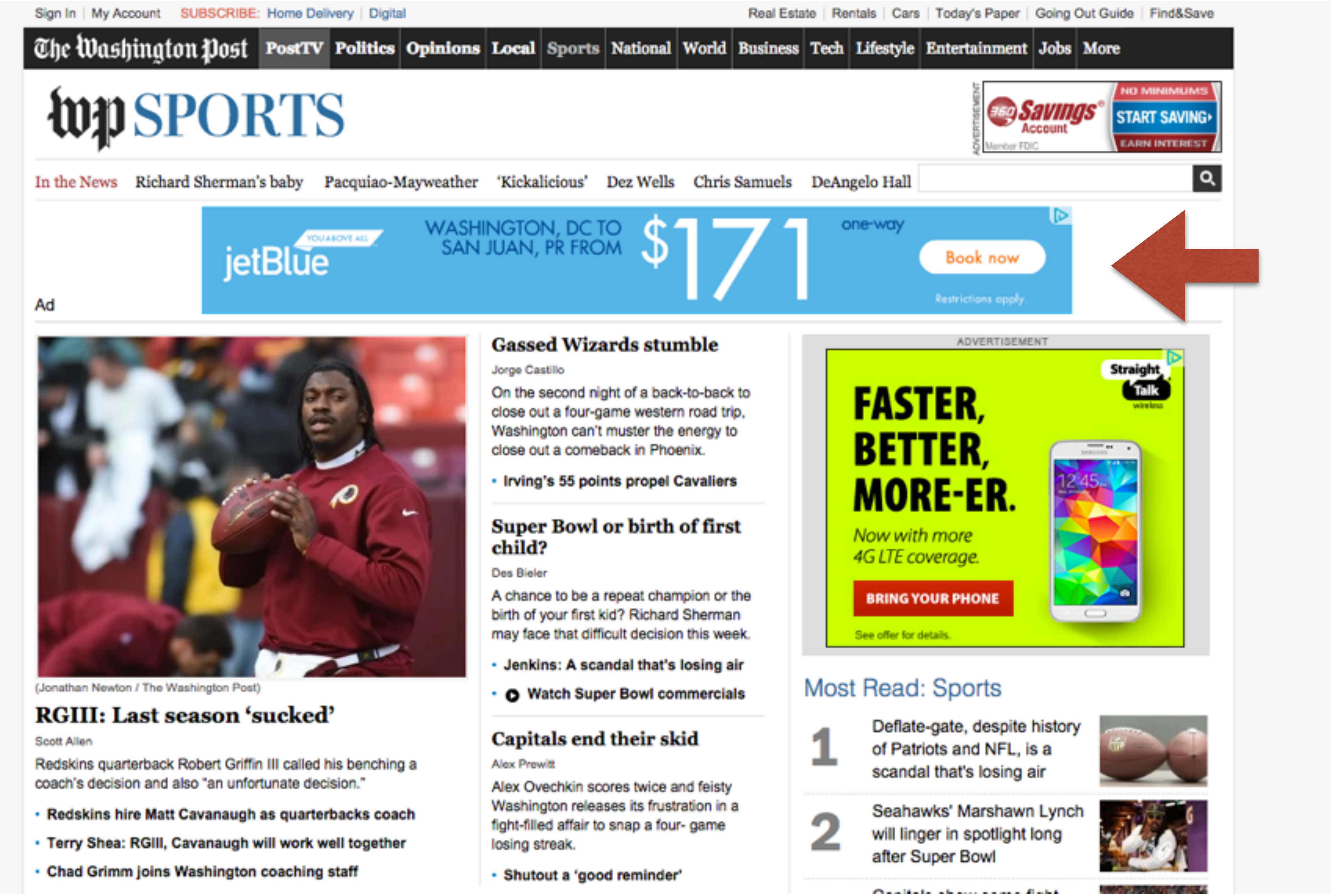
Suit yourself. I'll send you pictures from my yacht.

Making Sense of Big Data
A Big Data Guide for Small & Medium Businesses. Get the Free eMagazine!
www.tableausoftware.com/big-data

Dell™ Computer Outlet
Shop Dell™ Outlet For Discounted Computer Refurbs, w/ Intel® Core™
www.Dell.com/Outlet

Luxury BVI Cruise
7 Night Small Ship BVI Cruise from $2,595. Book Now & Save 50%
pgcruises.com/BVI-Cruise

BVI Yacht Charter
Yacht Charter in the BVI bareboat and with great crews.
www.ViSailing.com



Sign In | My Account SUBSCRIBE: Home Delivery | Digital
Real Estate | Rentals | Cars | Today's Paper | Going Out Guide | Find&Save
The Washington Post PostTV Politics Opinions Local Sports National World Business Tech Lifestyle Entertainment Jobs More
wp SPORTS

In the News Richard Sherman's baby Pacquiao-Mayweather 'Kickalicious' Dez Wells Chris Samuels DeAngelo Hall

(Jonathan Newton / The Washington Post)

RGIII: Last season 'sucked'

Scott Allen

Redskins quarterback Robert Griffin III called his benching a coach's decision and also "an unfortunate decision."

- **Redskins hire Matt Cavanaugh as quarterbacks coach**
- **Terry Shea: RGIII, Cavanaugh will work well together**
- **Chad Grimm joins Washington coaching staff**

Gassed Wizards stumble

Jorge Castillo

On the second night of a back-to-back to close out a four-game western road trip, Washington can't muster the energy to close out a comeback in Phoenix.

- **Irving's 55 points propel Cavaliers**

Super Bowl or birth of first child?

Des Bieler

A chance to be a repeat champion or the birth of your first kid? Richard Sherman may face that difficult decision this week.

- **Jenkins: A scandal that's losing air**
- **Watch Super Bowl commercials**

Capitals end their skid

Alex Prewitt

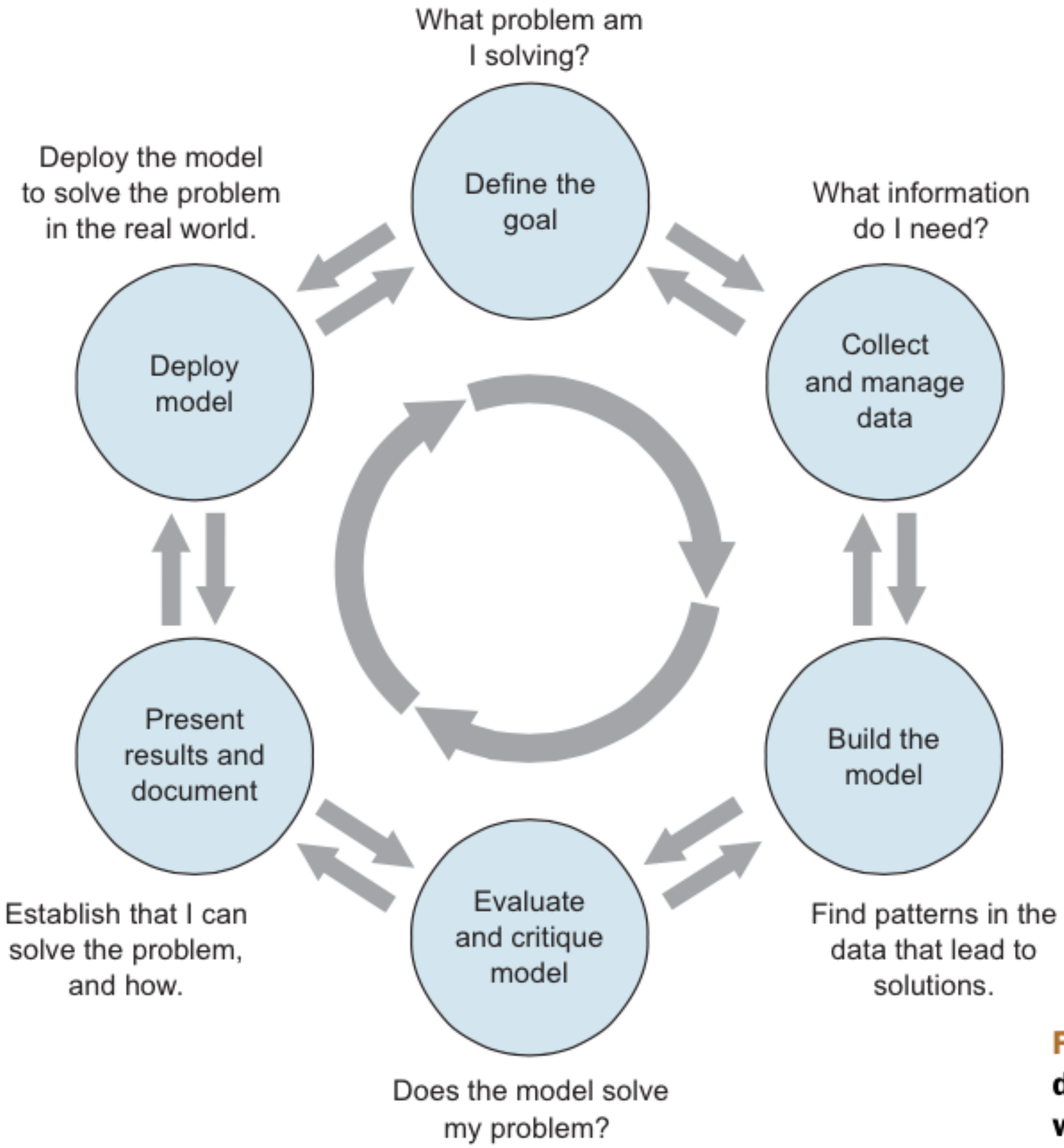
Alex Ovechkin scores twice and feisty Washington releases its frustration in a fight-filled affair to snap a four- game losing streak.

- **Shutout a 'good reminder'**

Most Read: Sports

1 Deflate-gate, despite history of Patriots and NFL, is a scandal that's losing air

2 Seahawks' Marshawn Lynch will linger in spotlight long after Super Bowl



What problem am I solving?

Define the goal

What information do I need?

Collect and manage data

Find patterns in the data that lead to solutions.

Build the model

Evaluate and critique model

Does the model solve my problem?

Present results and document

Establish that I can solve the problem, and how.

Deploy model

Deploy the model to solve the problem in the real world.

Figure 1.1 The lifecycle of a data science project: loops within loops

[Zumel and Mount]

Defining the goal

- What is the question/problem?
  - Who wants to answer/solve it?
  - What do they know/do now?
- How well can we *expect* to answer/solve it?
  - How well do they *want* us to answer/solve it?

Data collection and Management

- What data is available?
  - Is it good enough?
  - Is it enough?
- What are sensible *measurements* to derive from this data?
  - Units, transformations, rates, ratios, etc.

Modeling

- What kind of problem is it?
  - E.g., *classification, clustering, regression, etc.*
- What kind of model should I use?
  - Do I have enough data for it?
  - Does it really answer the question?

Model evaluation

- Did it work? How well?
- Can I interpret the model?
- What have I learned?

Presentation

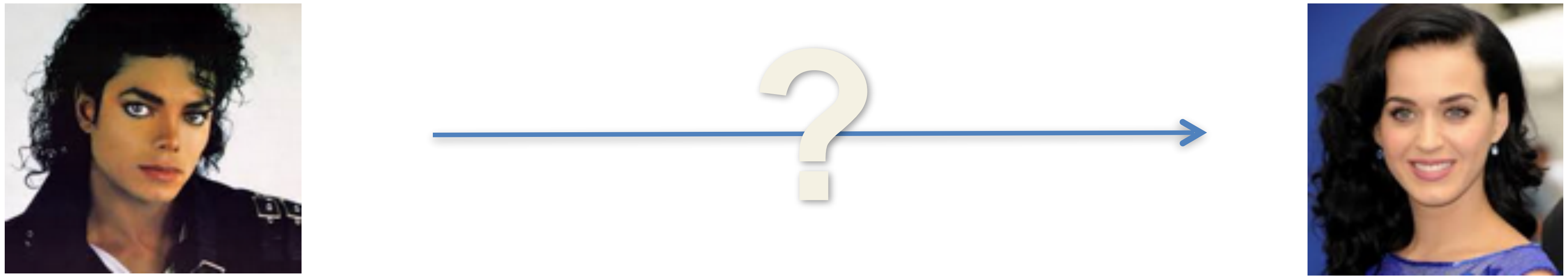
- Again, what are the *measurements* that tell the real story?
- How can I describe and visualize them effectively?

Deployment

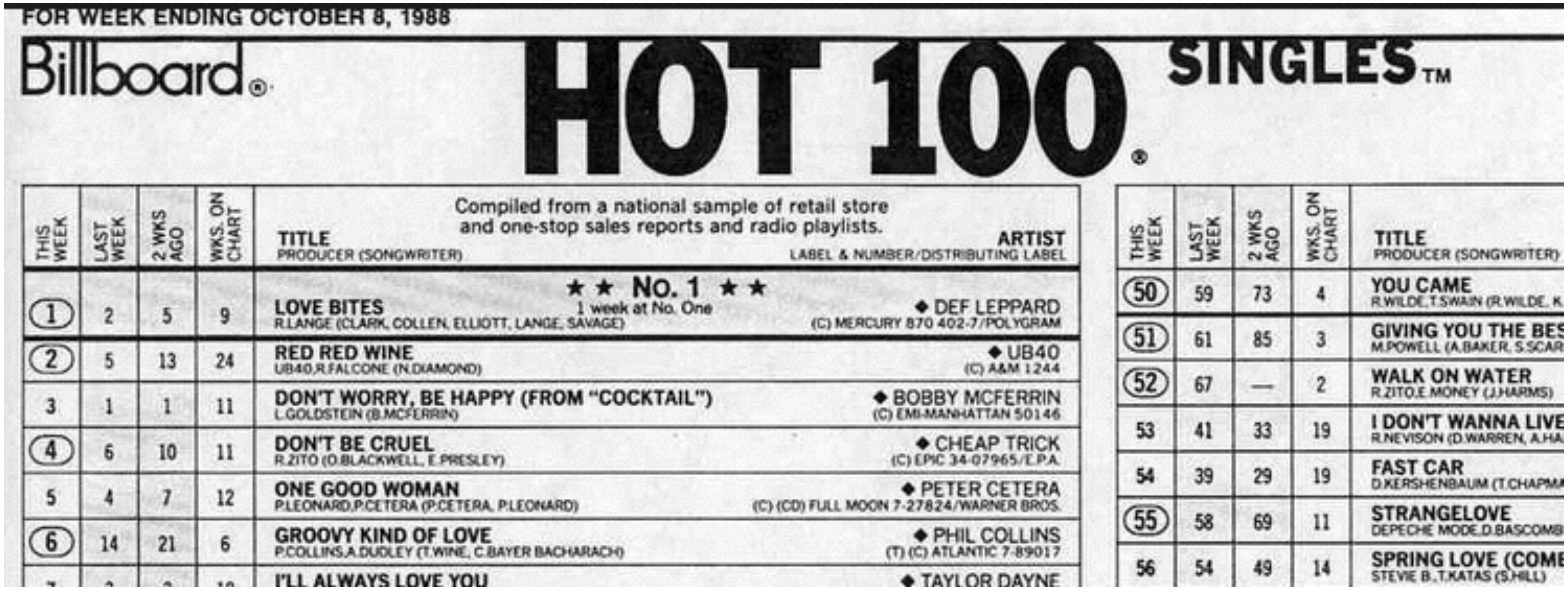
- Where will it be hosted?
- Who will use it?
- Who will maintain it?

Longitudinal network analysis shows the decline of pop music in the 21st century.

Talukder H., Corrada Bravo H.



Who are the writers of our favorite songs?



FOR WEEK ENDING OCTOBER 8, 1988

Billboard® HOT 100® SINGLES™

Compiled from a national sample of retail store and one-stop sales reports and radio playlists.

THIS WEEK	LAST WEEK	2 WKS AGO	WKS. ON CHART	TITLE PRODUCER (SONGWRITER)	ARTIST LABEL & NUMBER/DISTRIBUTING LABEL
1	2	5	9	★ ★ No. 1 ★ ★ 1 week at No. One LOVE BITES R.LANGE (CLARK, COLLEN, ELLIOTT, LANGE, SAVAGE)	◆ DEF LEPPARD (C) MERCURY 870 402-7/POLYGRAM
2	5	13	24	RED RED WINE UB40,R.FALCONE (N.DIAMOND)	◆ UB40 (C) A&M 1244
3	1	1	11	DON'T WORRY, BE HAPPY (FROM "COCKTAIL") L.GOLDSTEIN (B.MCFERRIN)	◆ BOBBY MCFERRIN (C) EMI-MANHATTAN 50146
4	6	10	11	DON'T BE CRUEL R.ZITO (O.BLACKWELL, E.PRESLEY)	◆ CHEAP TRICK (C) EPIC 34-07965/E.P.A.
5	4	7	12	ONE GOOD WOMAN P.LEONARD,P.CETERA (P.CETERA, P.LEONARD)	◆ PETER CETERA (C) (CD) FULL MOON 7-27824/WARNER BROS.
6	14	21	6	GROOVY KIND OF LOVE P.COLLINS,A.DUDLEY (T.WINE, C.BAYER BACHARACH)	◆ PHIL COLLINS (T) (C) ATLANTIC 7-89017
[illegible]	[illegible]	[illegible]	[illegible]	I'LL ALWAYS LOVE YOU	◆ TAYLOR DAYNE

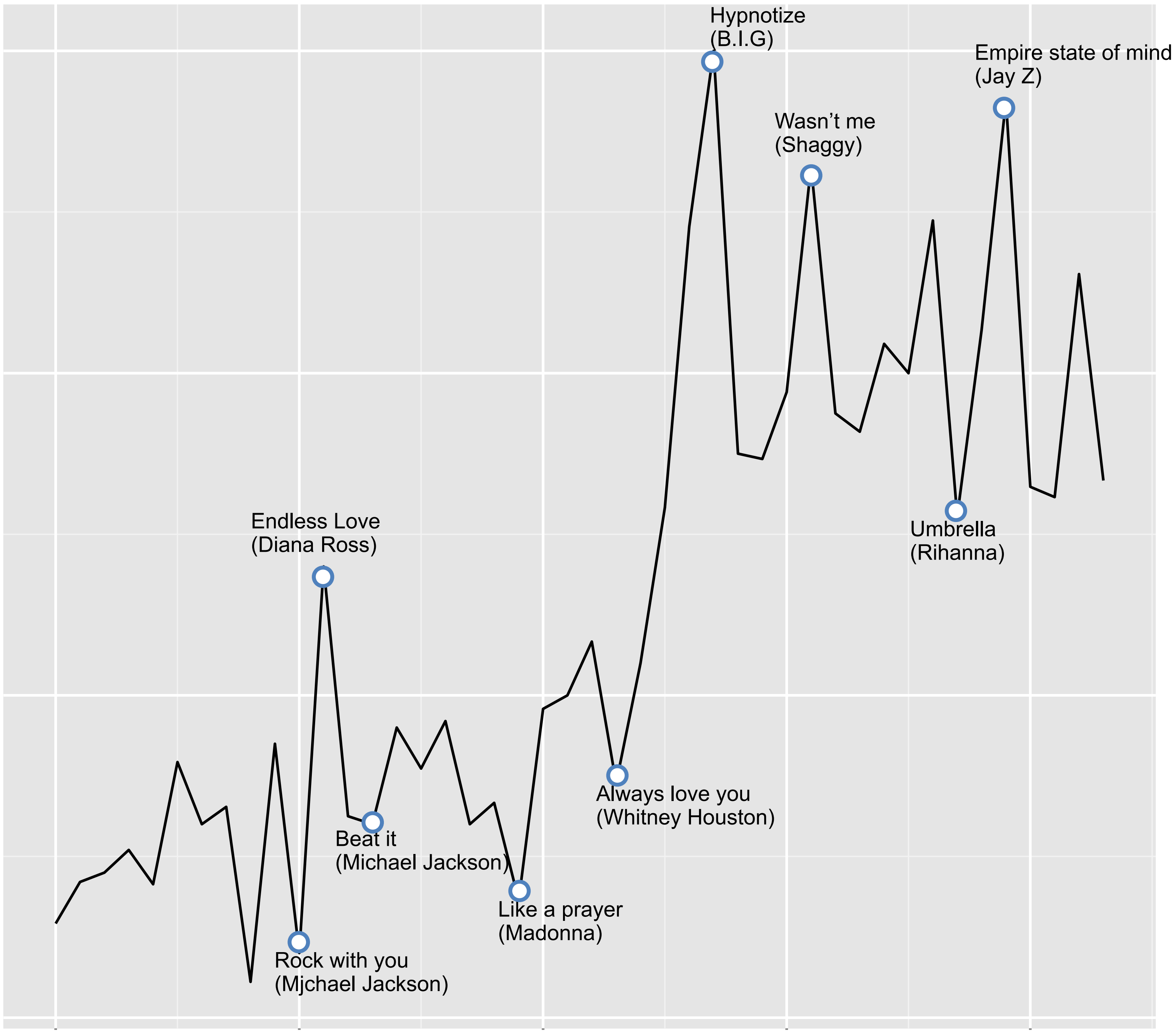
THIS WEEK	LAST WEEK	2 WKS AGO	WKS. ON CHART	TITLE PRODUCER (SONGWRITER)
50	59	73	4	YOU CAME R.WILDE,T.SWAIN (R.WILDE, K.
51	61	85	3	GIVING YOU THE BES M.POWELL (A.BAKER, S.SCAR
52	67	—	2	WALK ON WATER R.ZITO,E.MONEY (J.HARMS)
53	41	33	19	I DON'T WANNA LIVE R.NEVISON (D.WARREN, A.HA
54	39	29	19	FAST CAR D.KERSHENBAUM (T.CHAPMA
55	58	69	11	STRANGELOVE DEPECHE MODE,D.BASCOMB
56	54	49	14	SPRING LOVE (COME STEVIE B.,T.KATAS (S.HILL)

Billboard Hot 100 list

- Released weekly.
- Song is ranked by number of records sold, number of downloads, number of radio play and some other measures.
- Look at songs that hit number 1 in this list
  - At most 52 songs per year.

http://www.billboard.com/charts/hot-100

Average writer of songs per year



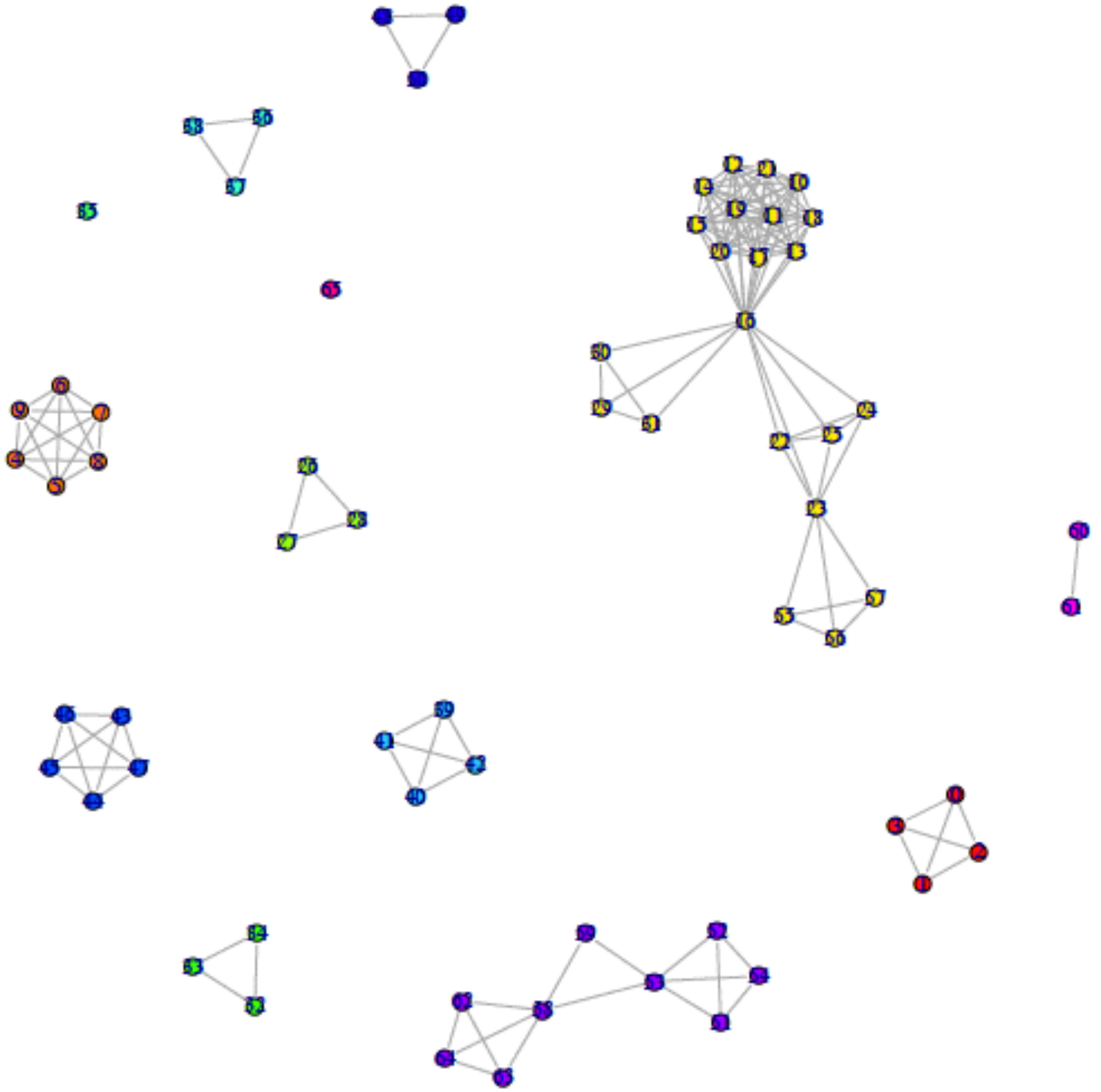
Number of writers per songs people are listening to over time is increasing.

Building Networks

- Network of music writers for top hits from 1970 to 2013.
  - – Nodes: writers
  - – Edges: collaboration in a top hit song

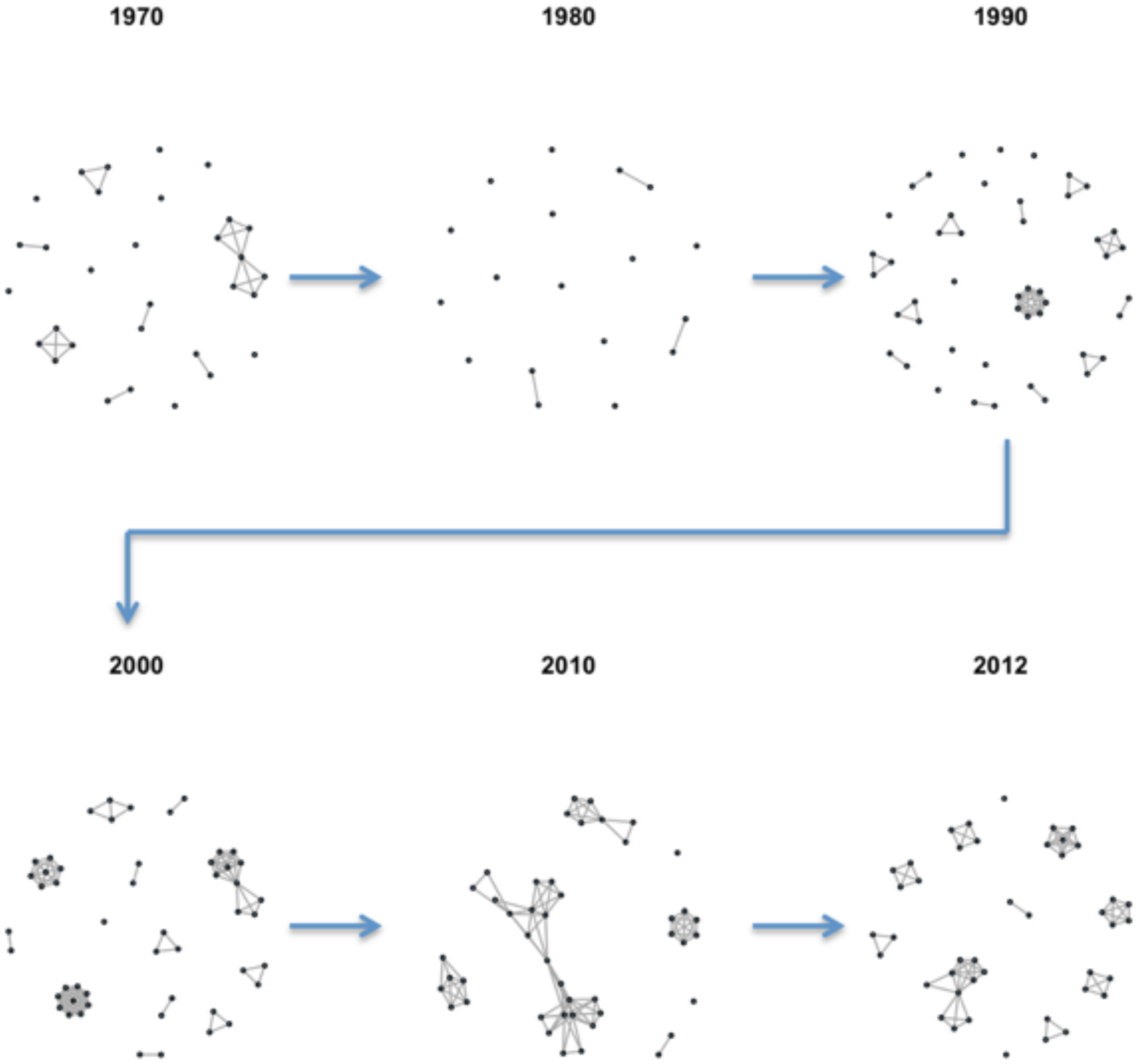
- Goals:
  - – How are network characteristics changing over time?
    - Node Degree: Number of collaborators for each writer.
    - Network density: Measure of how many writers are working on a given song on average.
  - – Can we predict these changes with other covariates?

Example of a music writer network



2006

Network of Writers



R-Shiny

https://github.com/htalukder/musicwriters