# Data Science in Action

**Linda Zhao**

**Jeff Cai**

Wharton
University of Pennsylvania | Lifelong Learning

CELEBRATING 50 WHARTON GLOBAL FORUMS

GLOBAL FORUM
HONG KONG 2017
JUNE 22-24

# Date Science in Action

## Granville Room

# ABOUT US

## Linda Zhao | Jeff Junhui Cai

### ME

- Have a **DREAM** job:
  a **Wharton Professor**
- Have a **wonderful** life
- A competitive **dancer**

### Jeff CAI

- A **Ph.D.** student
- **Computer Science** + **Statistics**
- **Brain** + **Effort**
  = **Star Modern Statistician**

# DATA SCIENCE
## CS + STAT + DOMAIN KNOWLEDGE

- **Goal of the Study**
- **Data**
  - Acquisition, collection
  - Cleaning (wrangling)
  - Understanding
- **Analyses**
  - EDA (**E**xploratory **D**ata **A**nalysis)
  - Analyses
- **Presentation**
- **Business decisions**

# DATA SCIENCE

## CASE STUDIES

1. Small Token but Big Effect: Call Center Quick Hang
2. Quick Study: Business Radio Powered by the Wharton School
3. Complete Case Study: Lending Club P2P
4. Big Data: Lung Cancer Micro array, MRI
5. Statistics is NOT Magic: Google Success/Failure

# DATA SCIENCE

## CASE STUDIES

1. **Small Token but Big Effect: Call Center Quick Hang**
2. Quick Study: Business Radio Powered by the Wharton School
3. Complete Case Study: Lending Club P2P
4. Big Data:  Lung Cancer Micro array, MRI
5. Statistics is NOT magic: Google Success/Failure

# CALL CENTER QUICK HANG

## DATA DOES NOT LIE

### Goal

- An interesting B-side story from a research of efficiency of a call center

### Background

- A small call center
- Data: detailed information of **each call**
  - Call time, wait time, conversation time
  - Business type, operators…

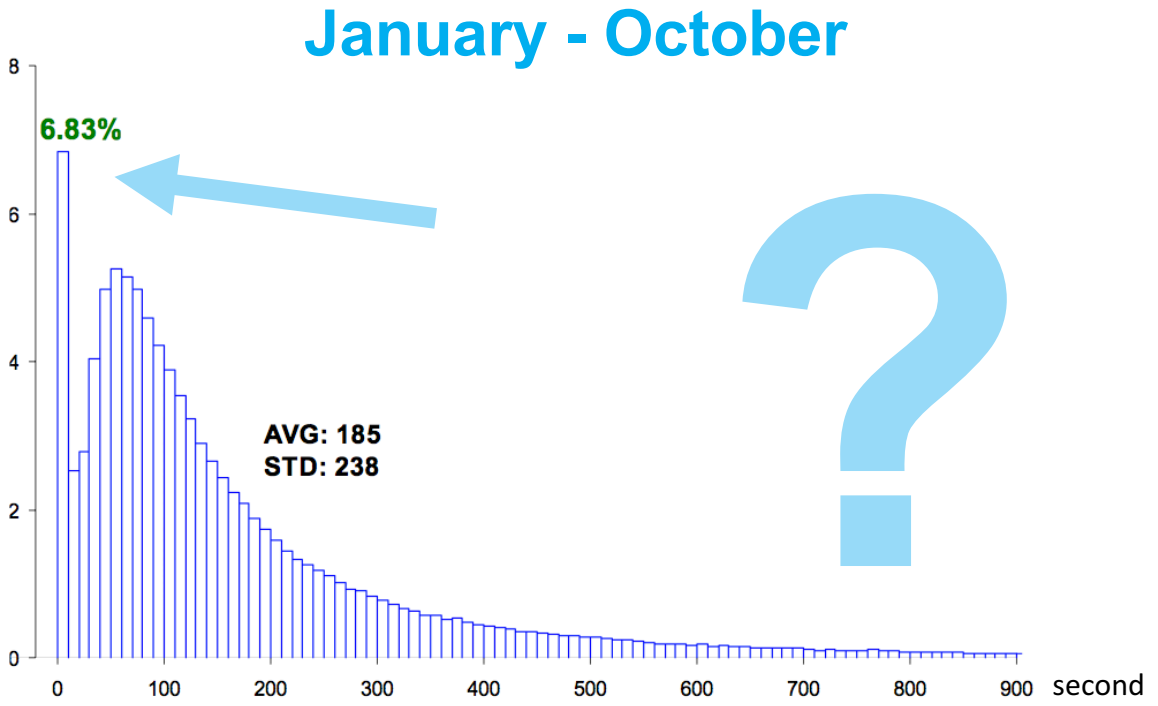
Brown, Mandelbaum, Shen, Zhao, et al. (2005)

# CALL CENTER QUICK HANG

## DATA DOES NOT LIE

**Service time**: conversation time between a customer and a service rep

- Look at the data first: while the average call time is 185 seconds, 7% calls only last for 10 seconds
- We discovered quick hang phenomena: operators hang up on customers!!!
- Most of quick hangs were made among 3 operators

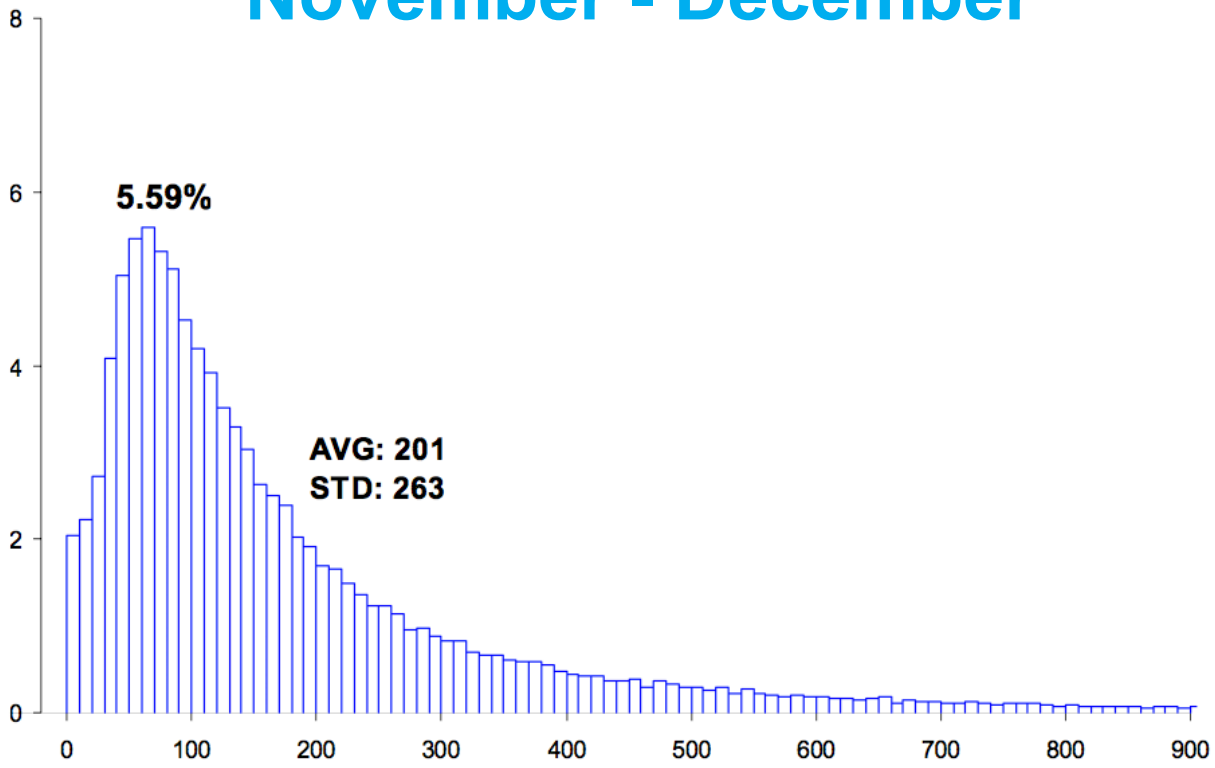


# CALL CENTER QUICK HANG

## DATA DOES NOT LIE

**Service time**: conversation time between a customer and a service rep

### November - December



# DATA SCIENCE

## CASE STUDIES

1. Small Token but Big Effect: Call Center Quick Hang
2. **Quick Study: Business Radio Powered by the Wharton School**
3. Complete Case Study: Lending Club P2P
4. Big data: Lung Cancer Micro array, MRI
5. Statistics is not magic: Google Success/Failure

# CASE STUDY: AUDIENCE ESTIMATION

## Sirius XM: Business Radio Powered by the Wharton School

### About Sirius XM

- Satellite radio
- Over 100 channels
- Subscription
- Commercial free
- Team up with automobile manufacturers

### Growing Rapidly (by 2014)

- 28.5 million subscribers
- 51.6 million listeners (conservative)

# CASE STUDY: AUDIENCE ESTIMATION

## CHANNEL 111: Business Radio Powered by the Wharton School

- For the purpose of national presence, social impact and innovation
  - The first business talk show in the world was born in January 2014
  - Hosted by Wharton professors
- 17 programs by Wharton professors
  - Management
  - Marketing
  - Entrepreneurship
  - Personal & Corporate Finance
  - Economic & Business Policy
- Question by Vice Dean Karl Ulrich:
  - **Do we have audience?**

Business RADIO
POWERED BY THE WHARTON SCHOOL

https://businessradio.wharton.upenn.edu

# CASE STUDY: AUDIENCE ESTIMATION
## DATA IS NEEDED!

### Ideally, a random sample is needed

- n=2000 should be enough
- Use a sample proportion of the Wharton listeners

### Restrictions

- A quick study
- Cost efficiency

### Solution: Go to MTURK

# CASE STUDY: AUDIENCE ESTIMATION
# AMAZON MECHANICAL TURK (MTURK)

## Brief Introduction

- An online marketplace for work
- Approximately 150,000 HITs (jobs/questions)
- Quick feedback, cheap but good

## MTURK Population

- At least 18 years old
- Higher education level than the US population

# CASE STUDY: AUDIENCE ESTIMATION

## AMAZON MECHANICAL TURK (MTURK)


MTurk

# CASE STUDY: AUDIENCE ESTIMATION QUESTIONNAIRE

## Key Questions

- Have you ever listened to Sirius Radio?
- Have you ever listened to the Wharton Business Radio?

## Important Controlling Factors

- Age
- Gender
- Income
- Education

# CASE STUDY: AUDIENCE ESTIMATION

## Survey Launched 05/24/2014 MTURK

### Validity of the Data

- A random sample from the US? + Probably not
- A random sample form the MTURK? + Might be

### Our Approach

- Utilize the fact that SIRIUS has 56 mm listeners
- Try to estimate the ratio of Wharton/Sirius
    - Assume the ratio remains same

# CASE STUDY: AUDIENCE ESTIMATION

## SURVEY RESULT

**1764** SURVEYS WITHIN 2 DAYS

**1362** SiriusXM SATELLITE RADIO LISTENERS

**70** Business RADIO POWERED BY THE WHARTON SCHOOL LISTENERS

# CASE STUDY: AUDIENCE ESTIMATION

## ESTIMATION

$$\frac{70}{1362} \times 51.6\text{MM} =$$

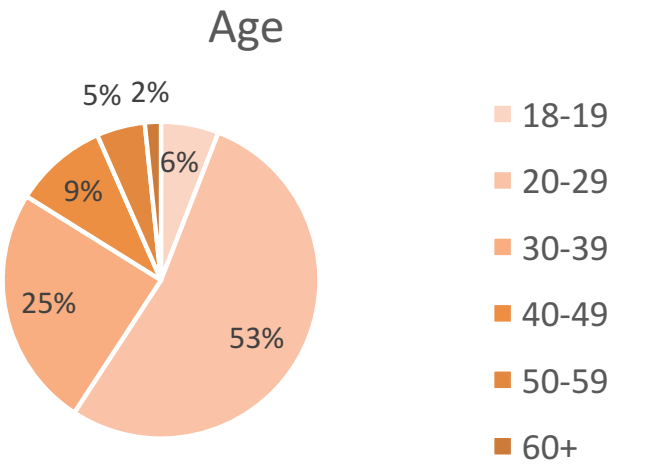
**2.65M**

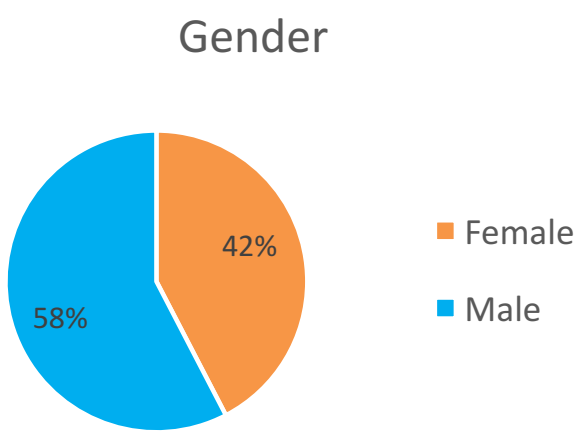
**± 0.0311**

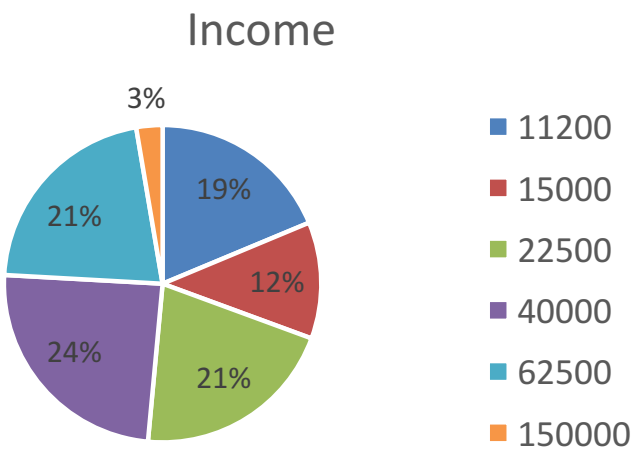
**WHARTON LISTENERS**

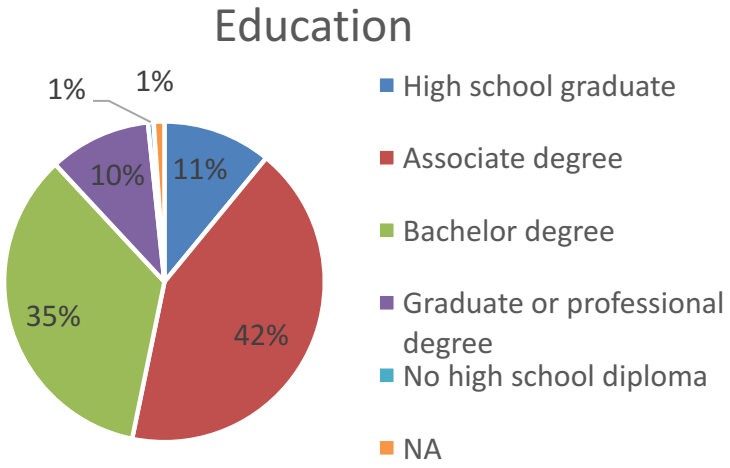
# CASE STUDY: AUDIENCE ESTIMATION

## REPRESENTATIVE SAMPLE AMONG MTURK?



# CASE STUDY: AUDIENCE ESTIMATION
## THE PROGRAM TODAY

### Leadership

- Created by Professor Karl Ulrich
- Currently led by Professor Peter Cappelli

Business RADIO
POWERED BY THE WHARTON SCHOOL

### Current 23 Programs

- Launch Pad (Karl Ulrich, …)
- Behind the Market (Jeremy Siegel, …)
- Leadership in Action (Michael Useem, …)
- In the Workplace (Peter Cappelli, …)
- Wharton Moneyball (Eric Bradlow, Shane Jensen, Cade Massey, Adi Wyner)
- Measured Thoughts (David Reibstein)
- …

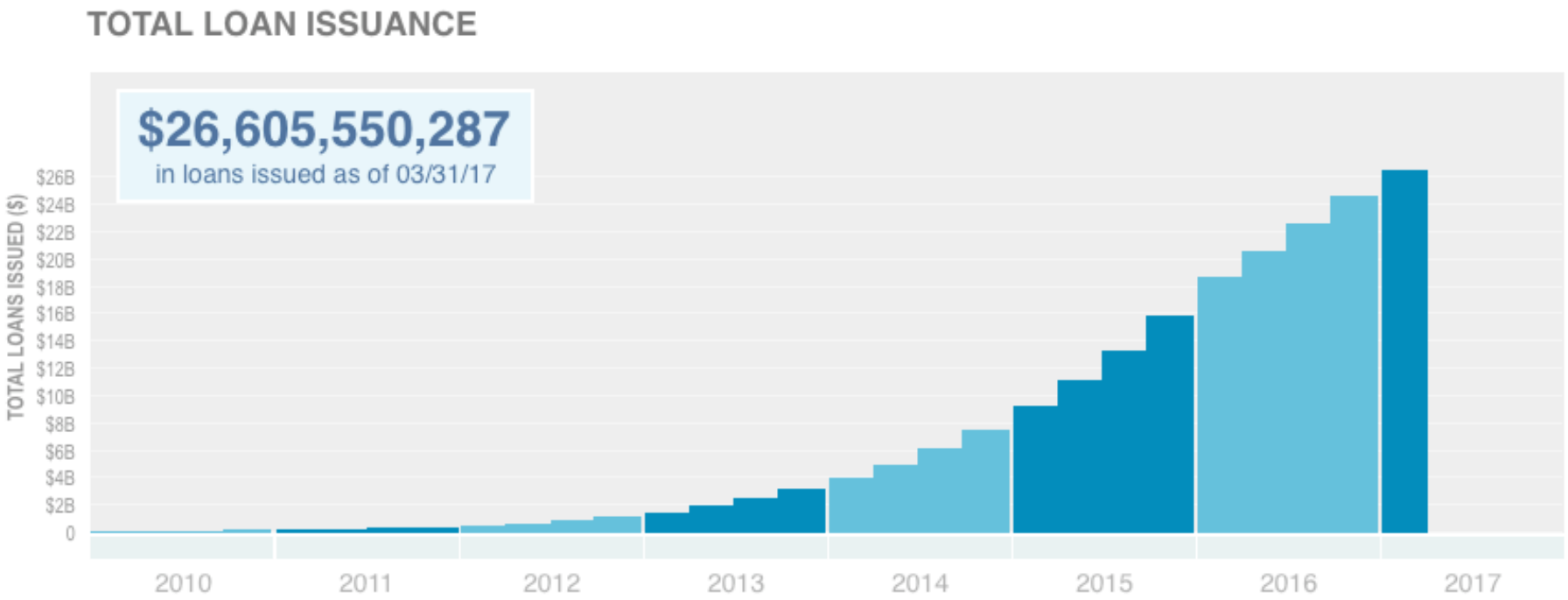
# DATA SCIENCE

## CASE STUDIES

1. Small Token but Big Effect: Call Center Quick Hang
2. Quick Study: Business Radio Powered by the Wharton School
3. **Complete Case Study: Lending Club P2P**
4. Big Data: Lung Cancer Micro array, MRI
5. Statistics is NOT magic: Google Success/Failure

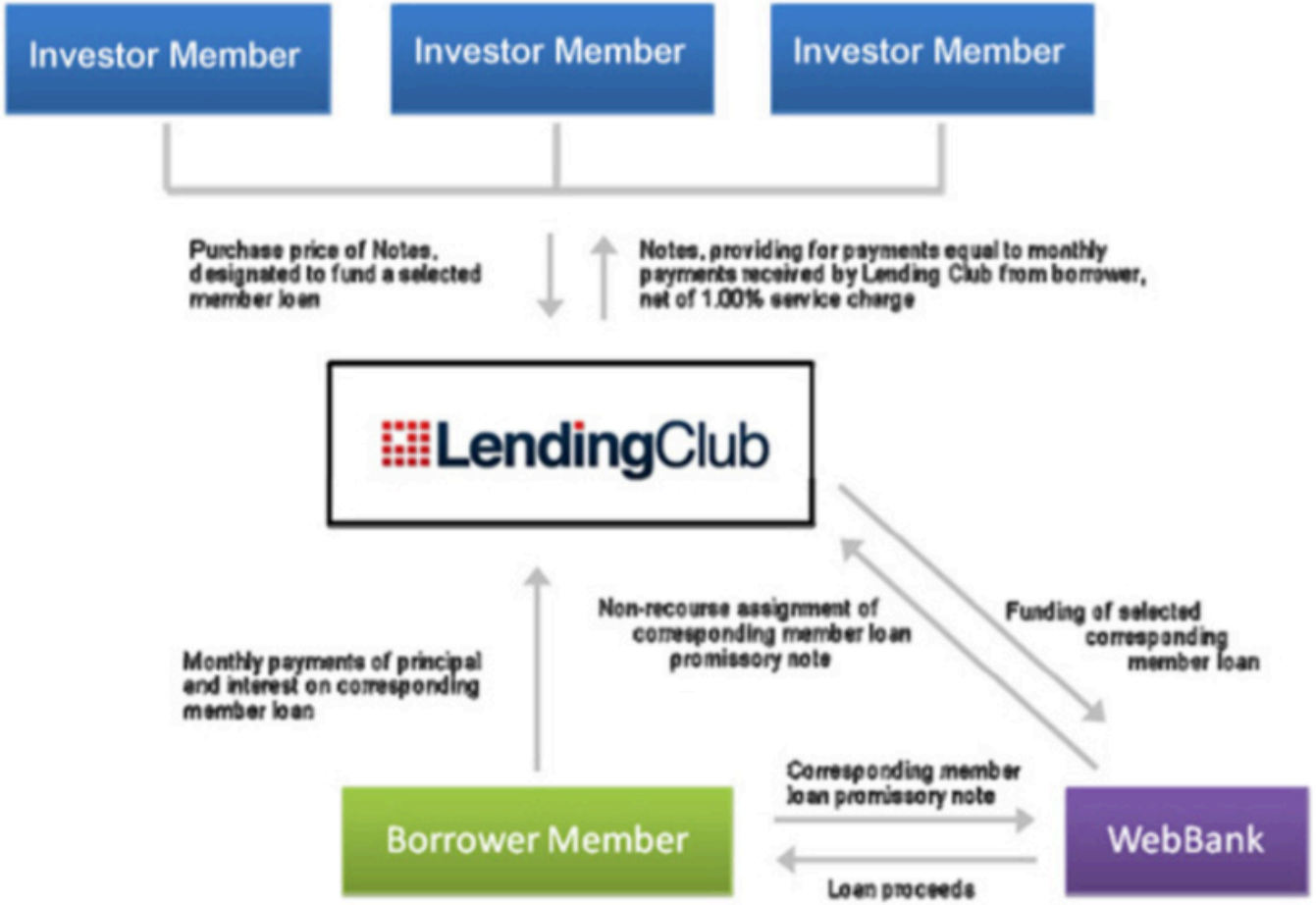
# CASE STUDY: LENDING CLUB

## A FAST GROWING P2P LOAN PLATFORM




Source: https://www.lendingclub.com

# CASE STUDY: LENDING CLUB
## A P2P LOAN PLATFORM

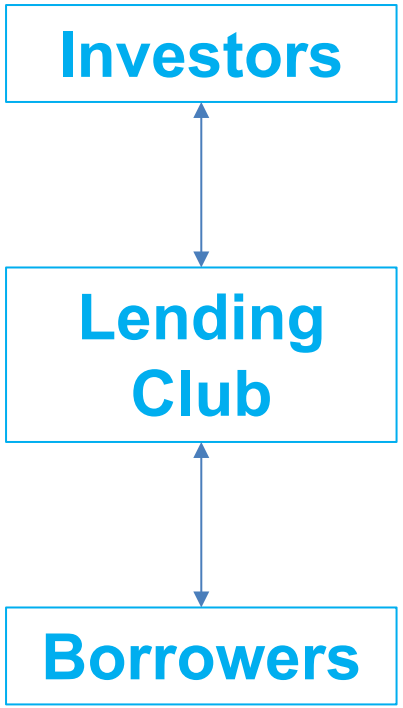



Source: https://www.lendingclub.com

# CASE STUDY: LENDING CLUB
## A P2P LOAN PLATFORM

### Lending Club

- **3-year** and **5-year loans**
- **Transparent** for **borrowers** and **investors**
- Classify loans into **grades** with various **interest rates**
- **1.1- 5%** (mostly 5% with an average of 4%) origination fee for each successfully closed loan from **borrowers**
- **1%** service charge for each payment from **investors**



# CASE STUDY: LENDING CLUB
## A P2P LOAN PLATFORM

### Borrowers

- Loans with a **fixed rate**
- Simplify with a **single payment**
- **Quick** and **easy** to apply
- No prepayment penalties or hidden fees
- Up to **$40K** for personal loans
- **$5K** to **$300K** for business loans

Source: https://www.lendingclub.com

# CASE STUDY: LENDING CLUB
## A P2P LOAN PLATFORM

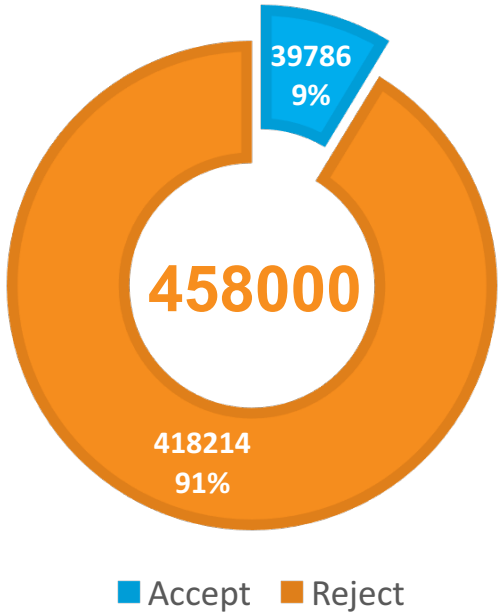
### Investors

- **5-7%** historical returns
- **Diversified** loans to reduce risk
- **2-5%** monthly cash flow
- API to access the platform programmatically and to execute investments automatically

# CASE STUDY: LENDING CLUB
## A CLOSER LOOK

### Complete Data 2008 - 2011

- Only **9%** got accepted
  - Minimum risk score
  - Debt-to-Income ratio
  - Credit report




### Goal of the Study

- Based on the **accepted** loan:
  - Performance of investors' portfolio
  - Room for improvement?

# CASE STUDY: LENDING CLUB
## DETAILED ANALYSIS

- Description of the Data
- Explore Data Analysis (EDA)
- Identify risk factors of a good customer vs a bad one
- Suggest an optimal classification criterion to minimize the loss (for investors)
  - **53.6% increase of the current return rate on average**

# CASE STUDY: LENDING CLUB
## EDA: DATA DESCRIPTION

**OVERVIEW**

- 38958 closed loans issued in 2008-2011
- 122 variables

**PRE-LOAN**

- Nature of the loan: amount, term, grade, interest rate…
- Borrower's information: annual income, home ownership, credit record…

**POST-LOAN**

- Fully paid/Charged off
- Performance: total interest, total principle, overdue charges

# CASE STUDY: LENDING CLUB
## DATA CLEANING

**MISSING**

- 70 out of 122 variables are all missing
- Lots of missing values in 3 variables

**FORMAT**

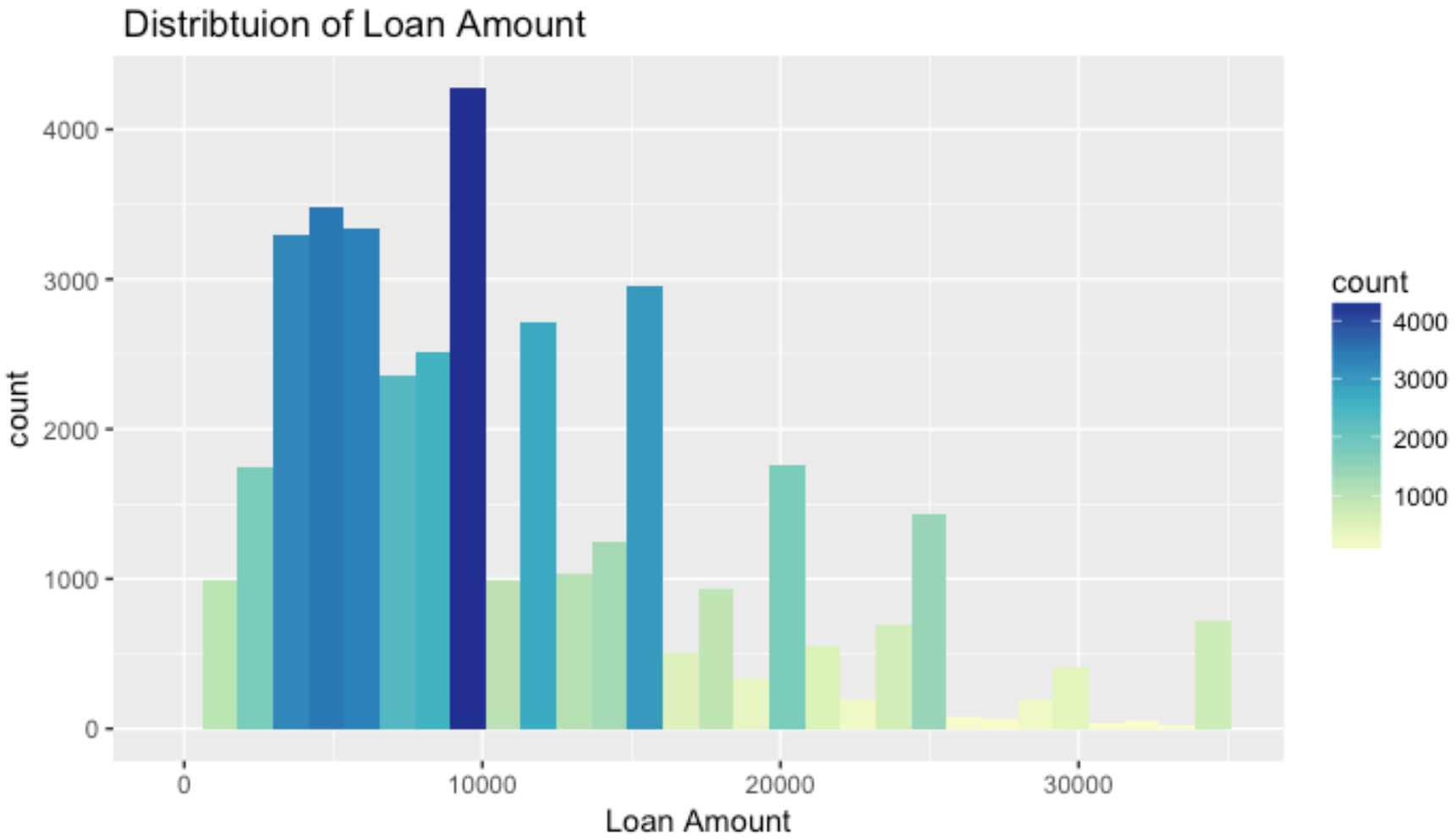
- Data type: numerical, categorical, time, text

**VARIABILITY**

- 11 variables are of low variability
- 1 variable of duplicate purpose

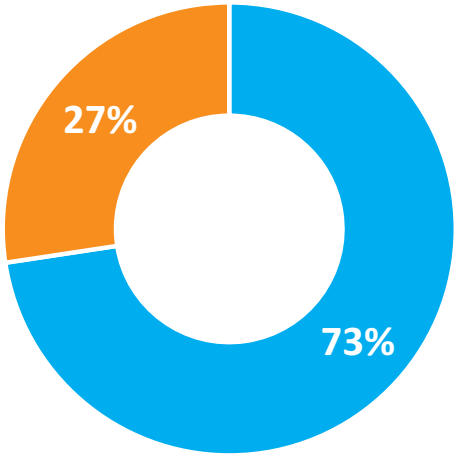
122 → 40 VARIABLES

# CASE STUDY: LENDING CLUB
## EDA: BRIEF DATA SUMMARY



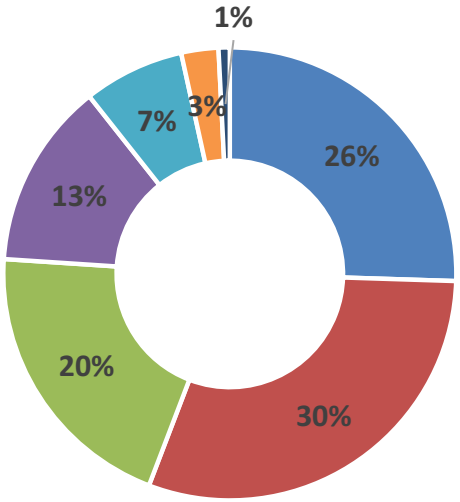
# CASE STUDY: LENDING CLUB
## EDA: BRIEF DATA SUMMARY

Term



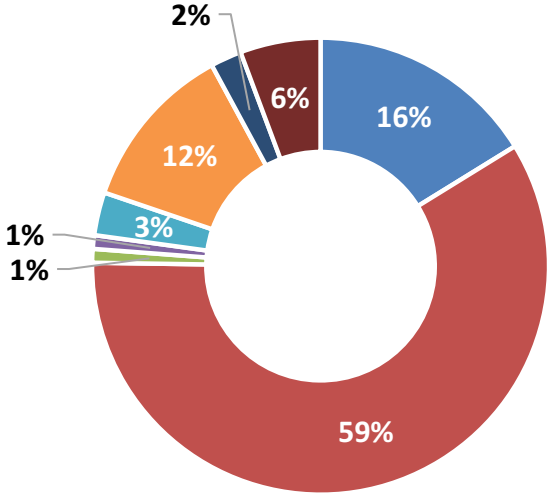

36 months 60 months

Grade




A B C D E F G

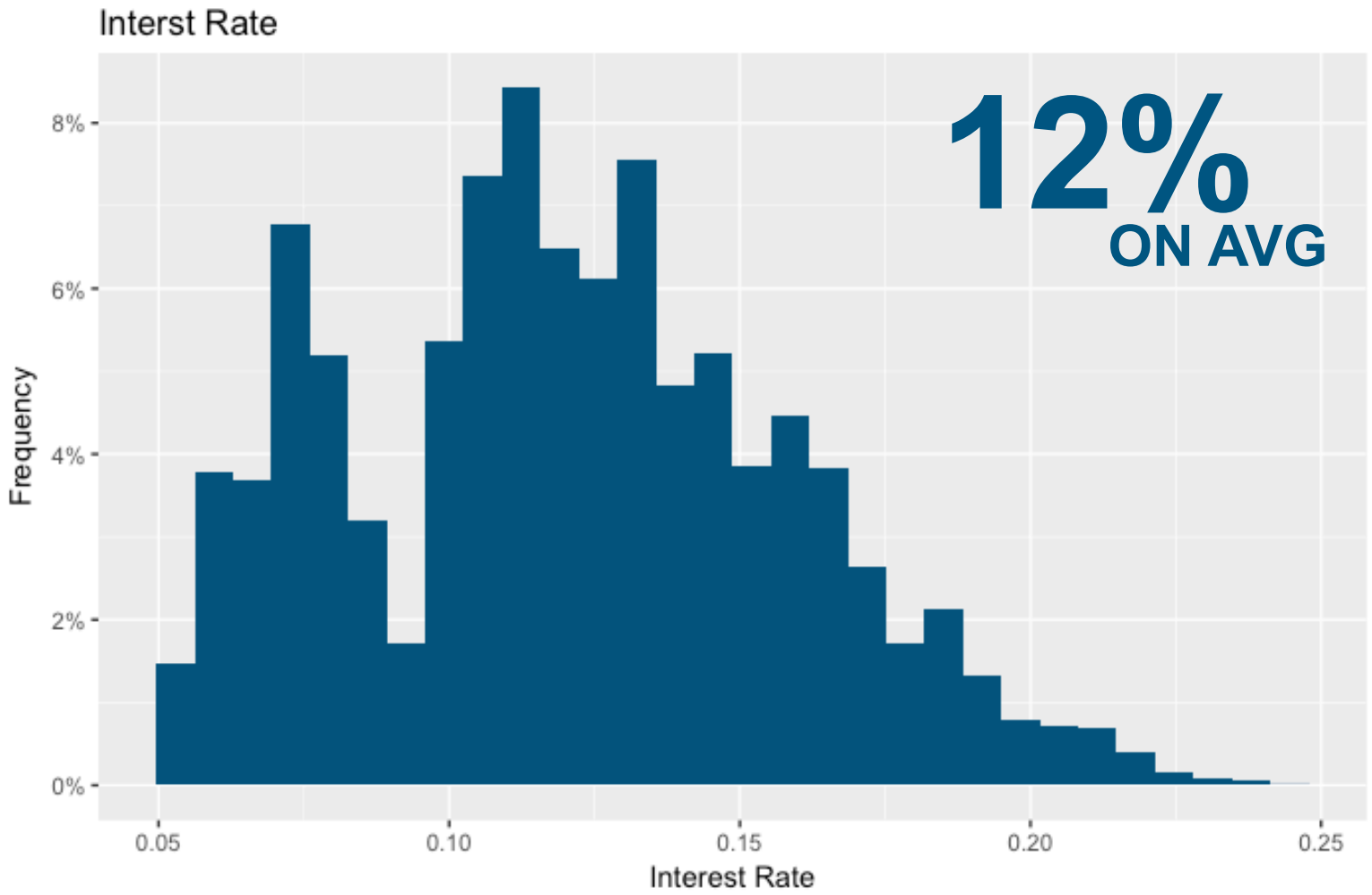
Purpose



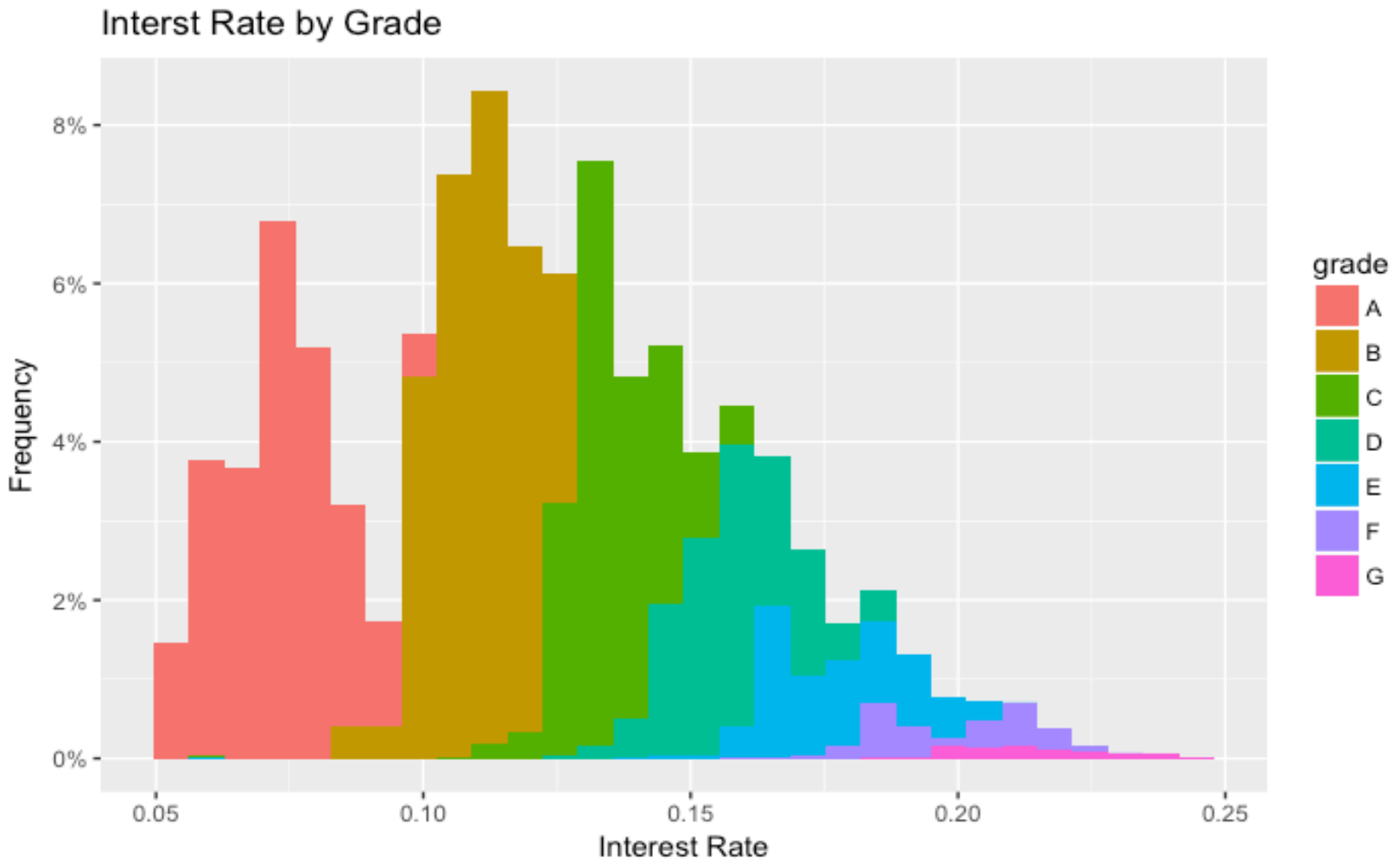

Credit Card
Debt Consolidate
Edu
Home Improve
House/Moving
Major Purchase
Medical
Small biz
Vacation
Wedding
Other

# CASE STUDY: LENDING CLUB

## EDA: BRIEF DATA SUMMARY

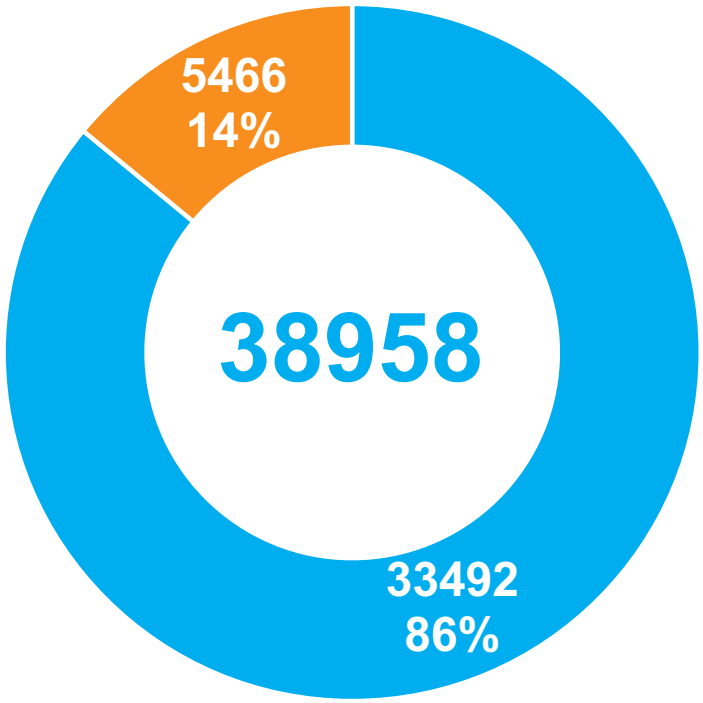


# CASE STUDY: LENDING CLUB
## EDA: BRIEF DATA SUMMARY



# CASE STUDY: LENDING CLUB
## EDA: BRIEF DATA SUMMARY

### Loan status

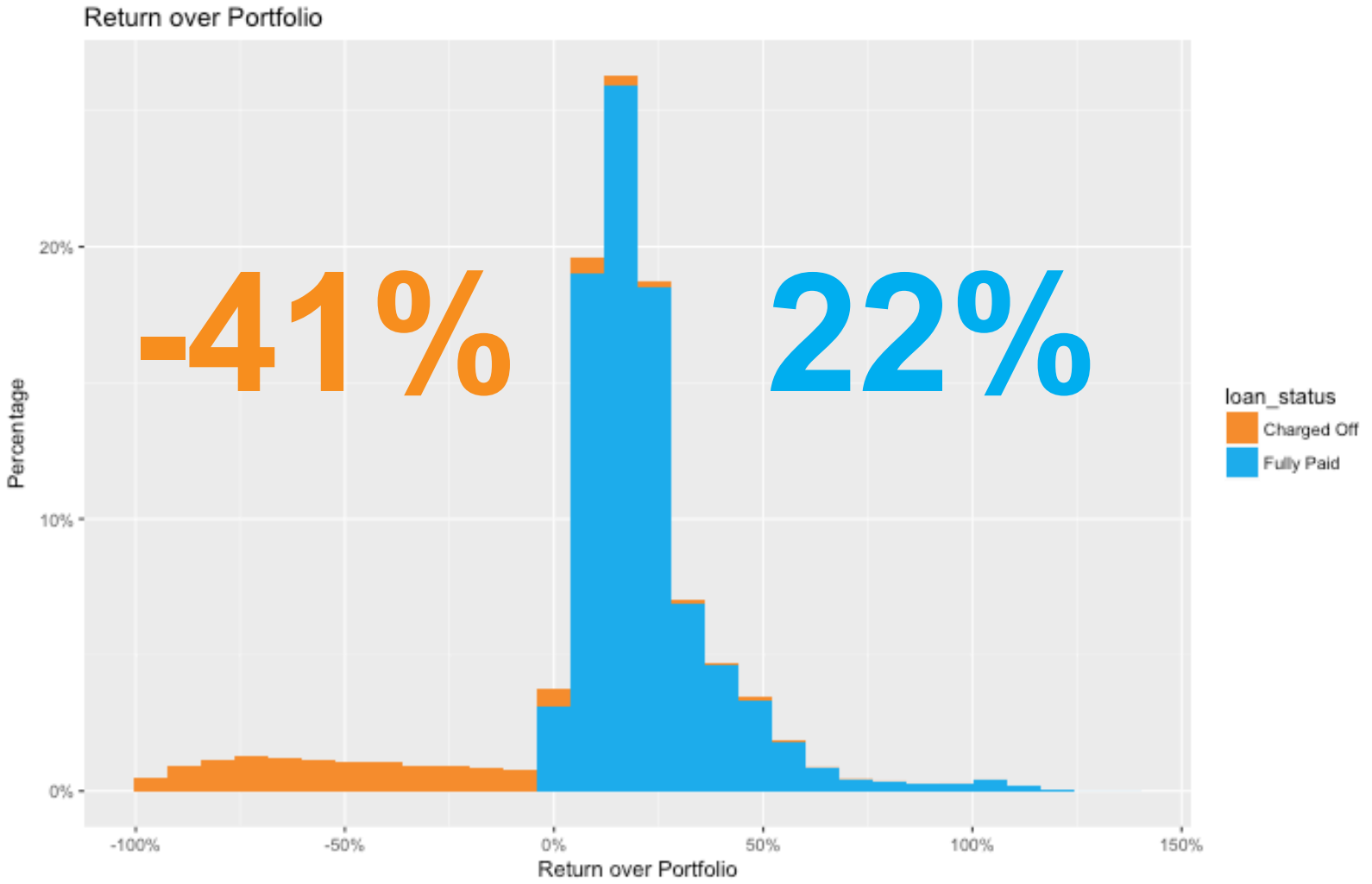



**22%**
AVG GAIN

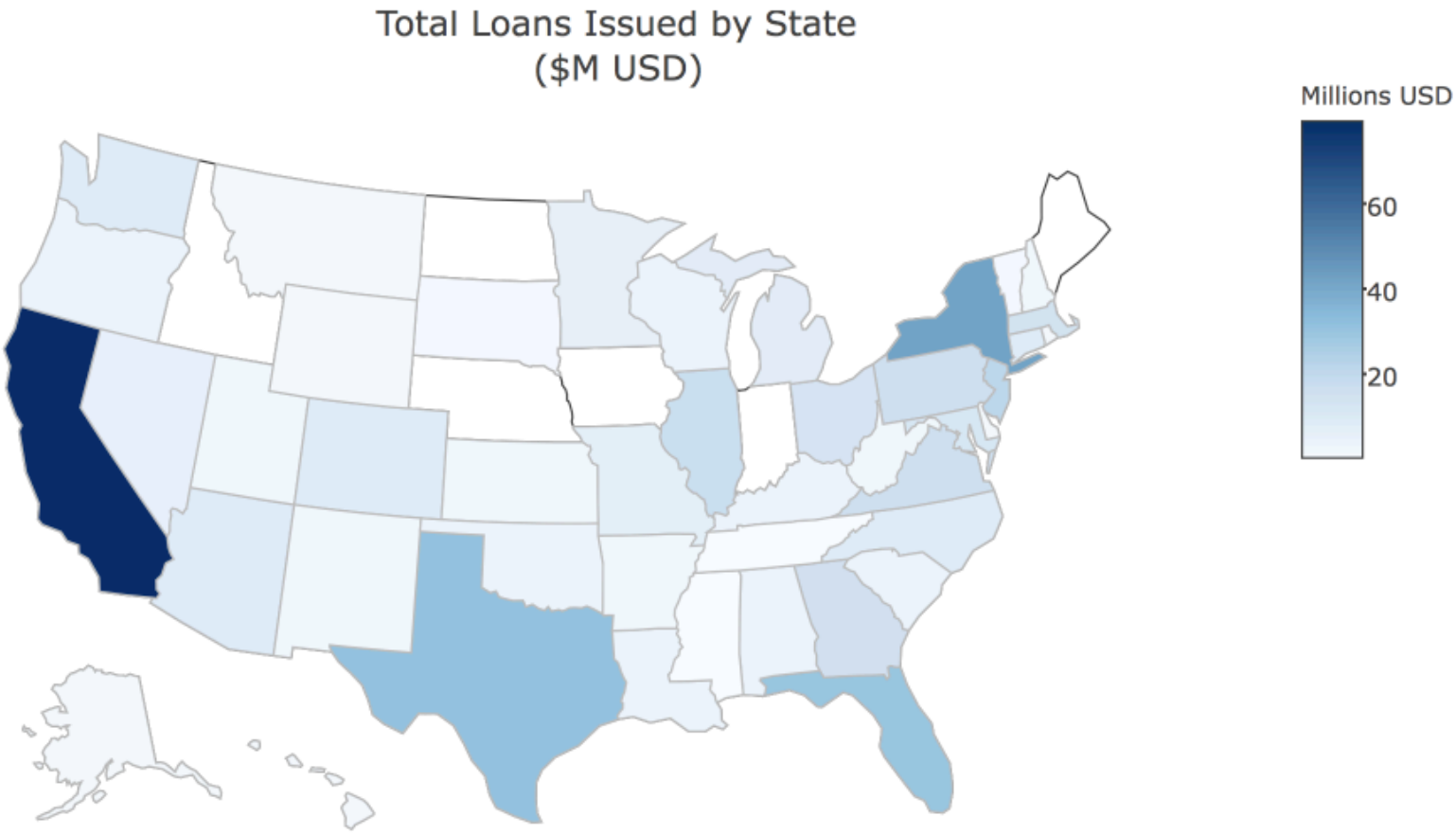
**-41%**
AVG LOSS

# CASE STUDY: LENDING CLUB

## EDA: BRIEF DATA SUMMARY

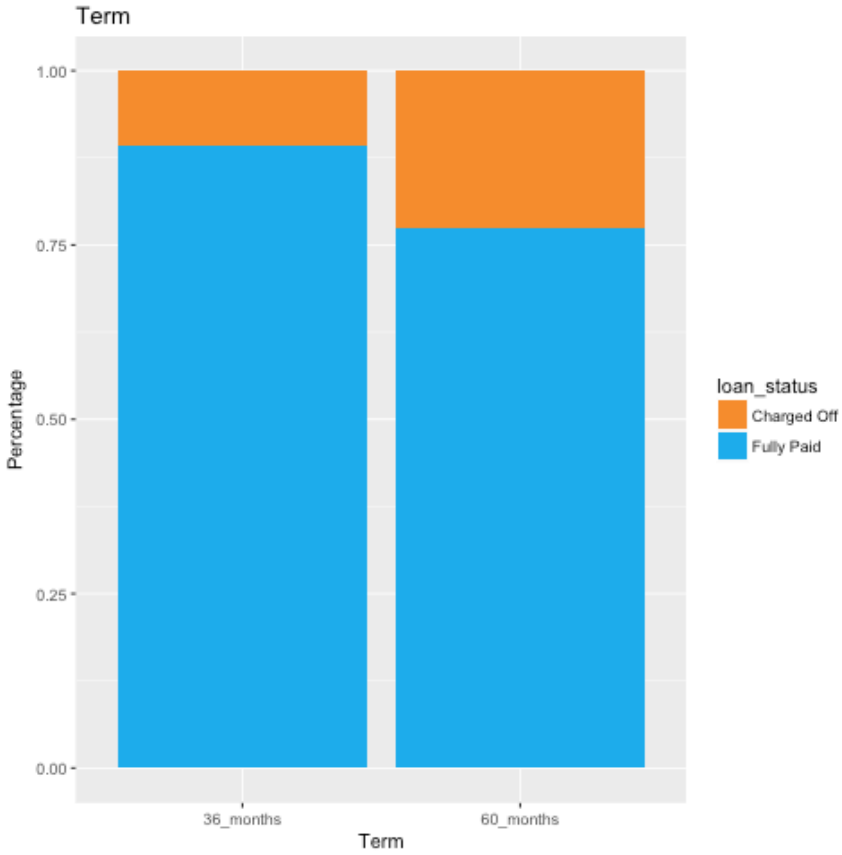


# CASE STUDY: LENDING CLUB
## EDA

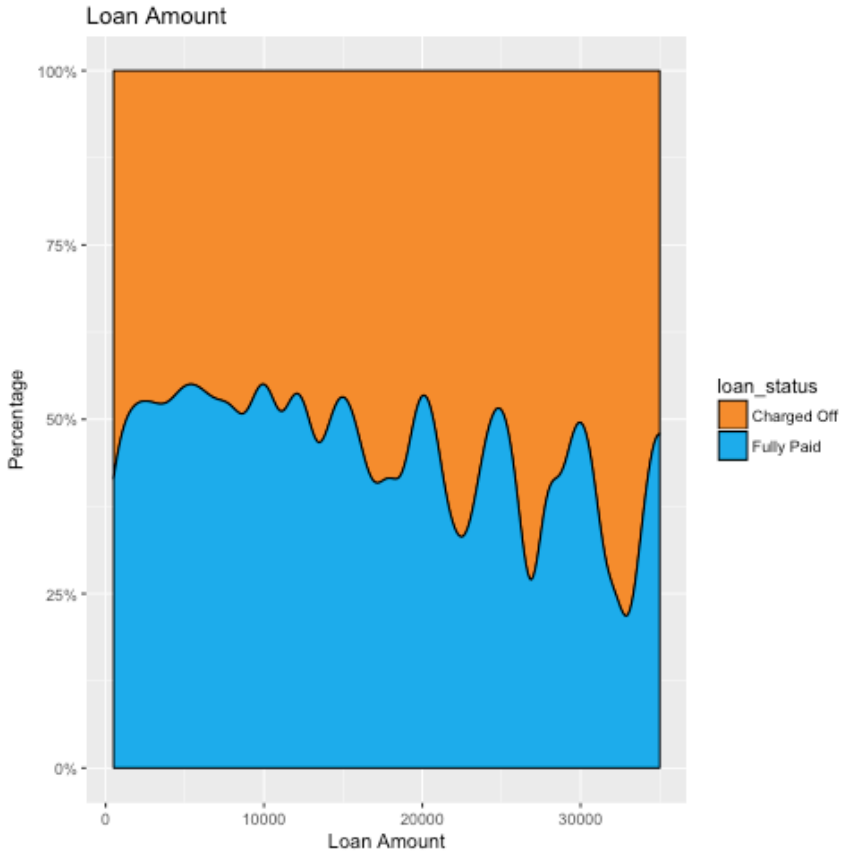


# CASE STUDY: LENDING CLUB
## EDA

### Term by Status



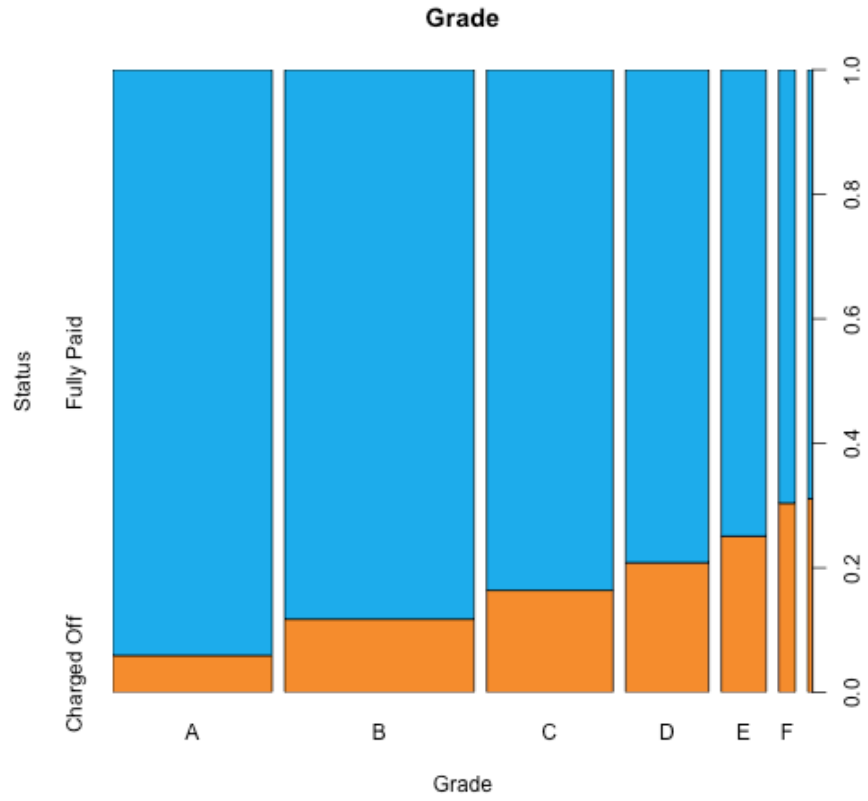

### Loan Amount by Status



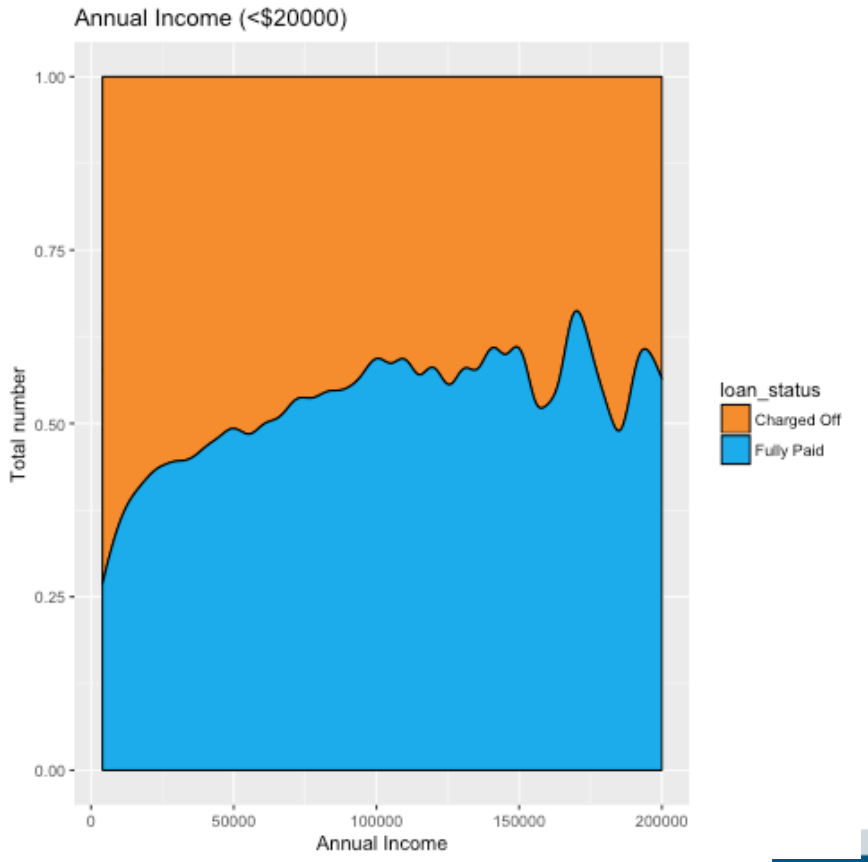
# CASE STUDY: LENDING CLUB

## EDA

### Grade by Status




### Annual Income by Status



# CASE STUDY: LENDING CLUB

## OUR MODEL: GOAL

$$\mathcal{P}(BAD\ LOAN \mid RISK\ FACTORS)$$

### Risk Identification

- 122 variables provided
- 40 variables after elimination
- 3764 dimensions involved

# CASE STUDY: LENDING CLUB
## OUR MODEL: METHOD

$$\mathcal{P}(BAD\ LOAN \mid RISK\ FACTORS)$$

### Risk Identification / Dimension Reduction

- Elastic Net
- AIC/BIC
- Cross Validation

# CASE STUDY: LENDING CLUB
## OUR MODEL: RESULT

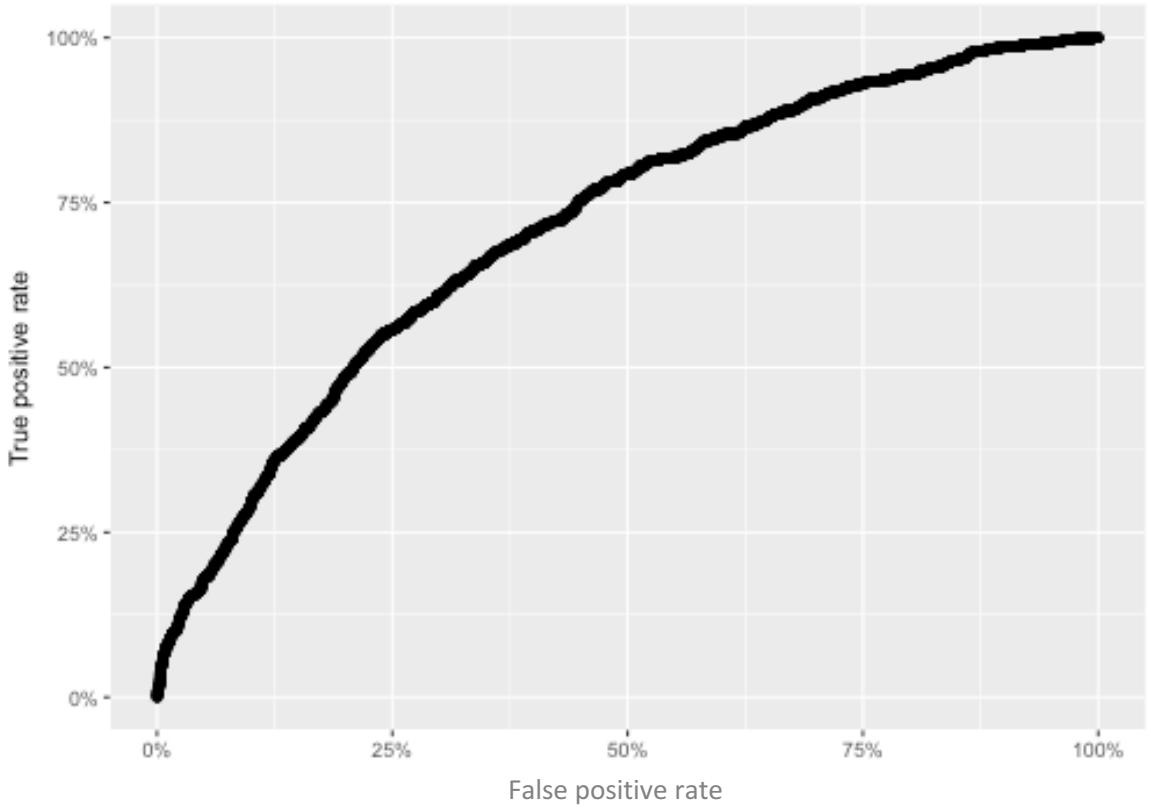
### Risk Identification

Controlling important factors (e.g. Location) based on our estimated probability equation:

- **Annual Income**
- **Grade**
- **Term**
- **Purpose**
- **Interaction of Loan Amount and Grade**
- **Degree of Desperation of the Borrower**
- **…**

# CASE STUDY: LENDING CLUB
## OUR MODEL: RESULT

### Risk Identification / Estimation Evaluation (ROC)




Purpose of **ROC**

- Controlling **false positive**
- Evaluate **accuracy** of the probability estimate

# CASE STUDY: LENDING CLUB
## OUR MODEL

### Our Optimal Customer Segment Rule

- **Define loss function**

| Loss | | Prediction | |
|---|---|---|---|
| | | Good | Bad |
| True | Good | -22% | 0 |
| | Bad | 41% | 0 |

- **Bayes rule**
  - Minimize the **average loss**
  - **Deny** a loan if the **probability** is **larger** than a threshold
  - Based on the loss above, the threshold is **0.35**
  - Deny a loan if our estimated probability is larger than 0.35

# CASE STUDY: LENDING CLUB
## OUR MODEL

### Performance Comparison

If our rule were applied, the relative increase of the average return from the original return is

**53.6%**
**INCREASE**

# CASE STUDY: LENDING CLUB
## OUR MODEL

### Conclusion and Remark (1/2)

- Use machine learning tools **to identify risk factors** and **to build a model** on top or LC's screening of the customers
- Would have increased the investor performance by **53.6%** on average
- **Caveats and Limitations:**
  - Built on the accepted loans (lack of information of general customers)
  - The time period is right during the financial crisis
  - A sample from general applicants is much more desirable
    - Affirm by Max Levchin (co-founder of PayPal)

# CASE STUDY: LENDING CLUB
## OUR MODEL

### Conclusion and Remark (2/2)

- Can be readily extended
  - A dynamic system (based on ongoing loans)
  - More elaborate tools involved (e.g. survival data analysis)
  - Factor-dependent loss
    - Loss evaluation by grades/terms/income…
  - Other possible models (e.g. directly targetting the mean gains)

# DATA SCIENCE

## CASE STUDIES

1. Small Token but Big Effect: Call Center Quick Hang
2. Quick Study: Business Radio Powered by the Wharton School
3. Complete Case Study: Lending Club P2P
4. **Big Data: Lung Cancer Micro array, MRI**
5. Statistics is NOT magic: Google Success/Failure

# CASE STUDY: LUNG CANCER
## WHICH GENES TO BLAME

**12,625** GENES

**56** SAMPLES

**4** SUBGROUPS

# CASE STUDY: LUNG CANCER
## WHICH GENES TO BLAME

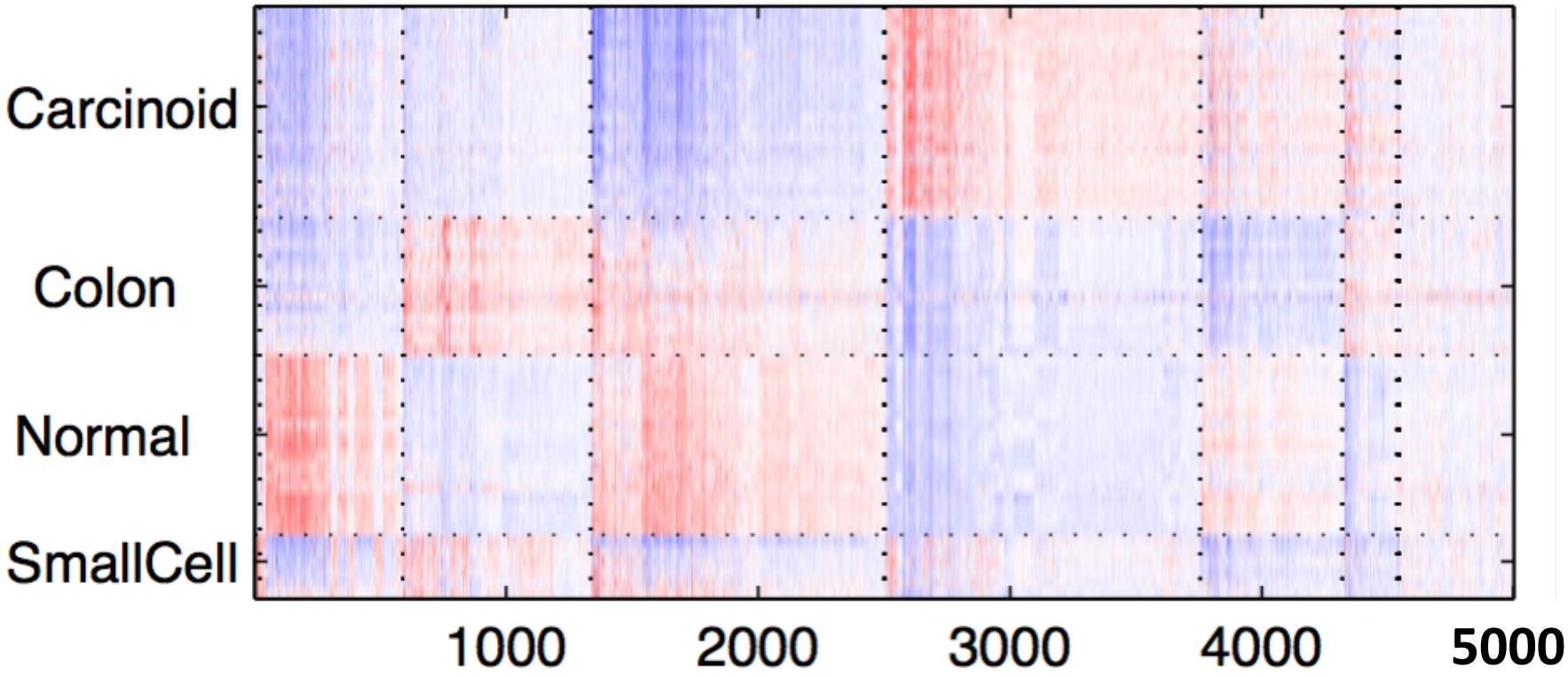
4 SUBGROUPS

- Normal
- Pulmonary Carcinoid Cancer
- Colon Cancer
- Small Cell Cancer

# CASE STUDY: LUNG CANCER
## RAW DATA

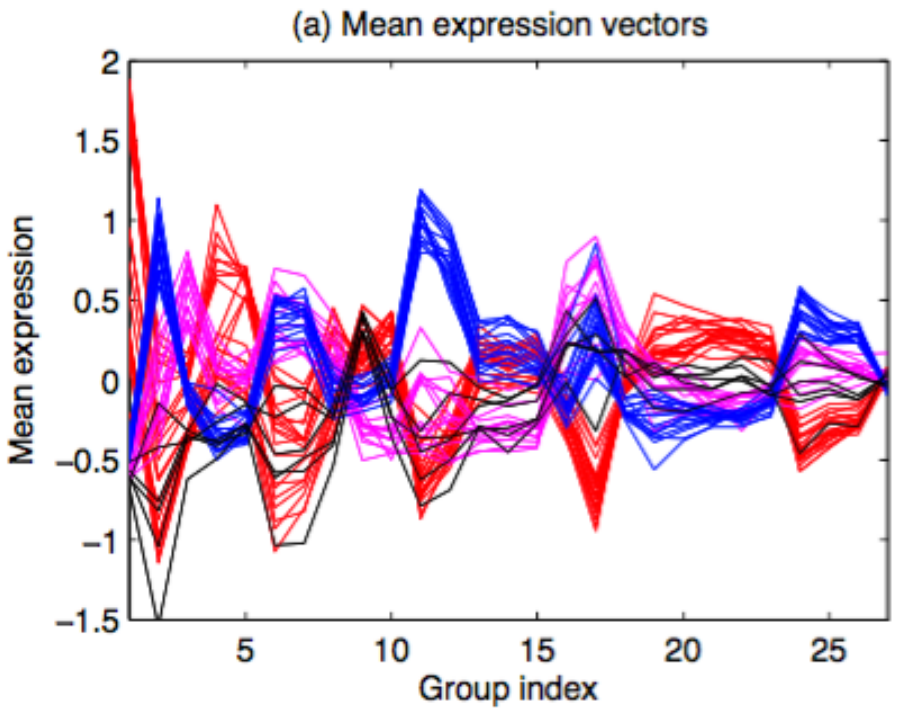
# CASE STUDY: LUNG CANCER
## SPARSE PCA

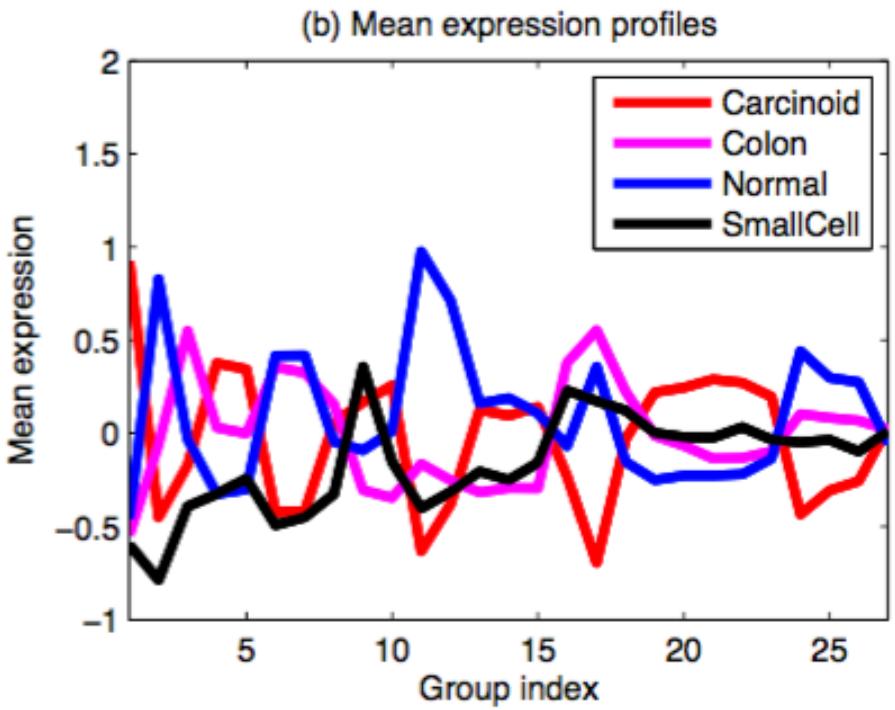
### SELECT AND REORDER



# CASE STUDY: LUNG CANCER
## SPARSE PCA

### GROUP THE GENES



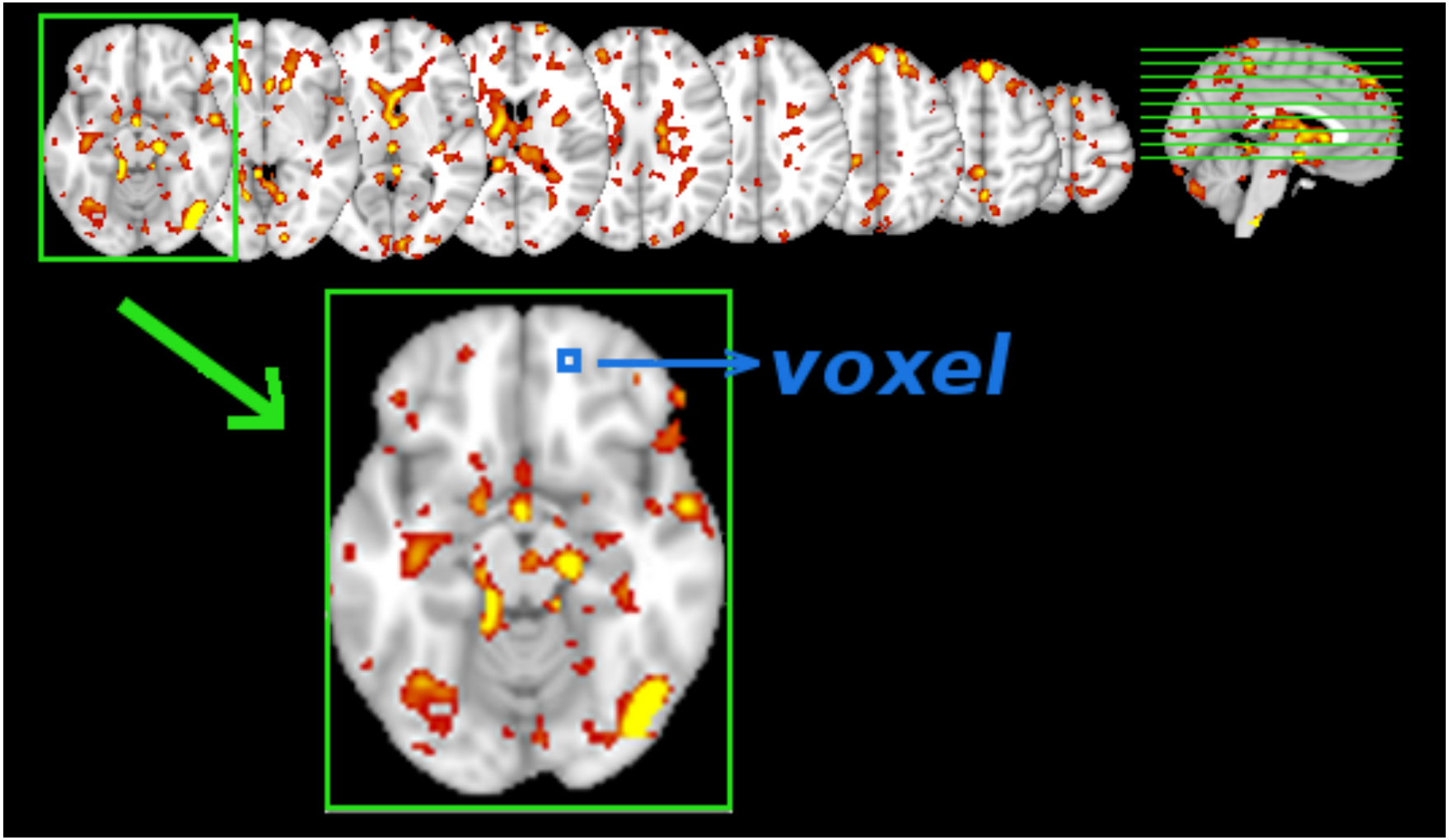




Shen et al. (2010)

# CASE STUDY: IMAGING DATA

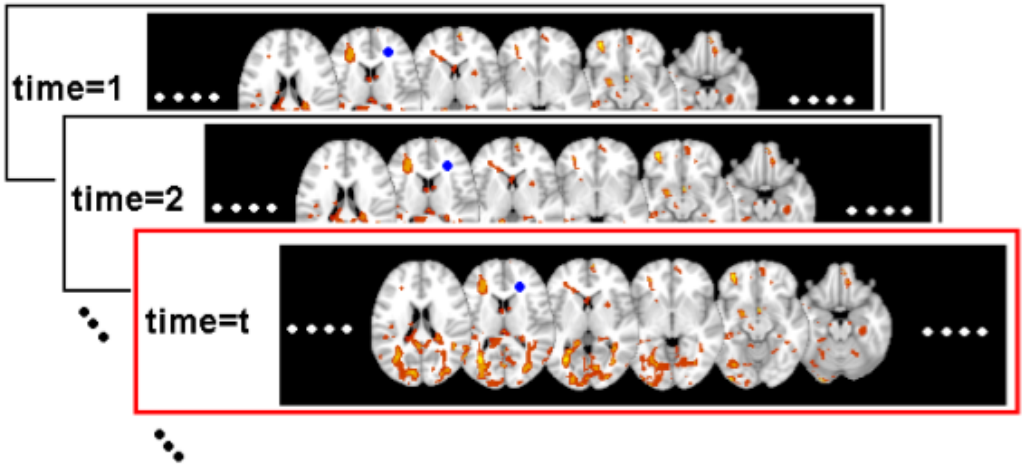
- Multiple modality
  - fMRI, EEG/MEG, DTI, PET, CT, …
- **Large size, high dimensionality**
  - Take fMRI for example
    - 100 individuals,  200, 000 voxel/individual,  200 time points
    - 30GB storage
- Dimension reduction, parallel computing

# CASE STUDY: IMAGING DATA
## FUNCTIONAL MRI



# CASE STUDY: IMAGING DATA

## SPATIAL-TEMPORAL MATRIX, TENSOR



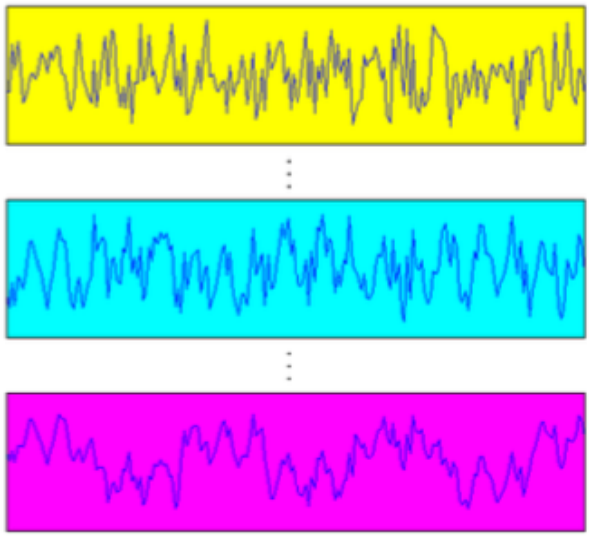

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1t} & \cdots & X_{1T} \\ \vdots & \vdots & & \vdots & & \vdots \\ X_{v1} & X_{v2} & \cdots & X_{vt} & \cdots & X_{vT} \\ \vdots & \vdots & & \vdots & & \vdots \\ X_{V1} & X_{V2} & \cdots & X_{Vt} & \cdots & X_{VT} \end{bmatrix}$$

Voxel Time Series

T=200

Spatial Pattern

V=64·64·49



# DATA SCIENCE

## CASE STUDIES

1. Small Token but Big Effect: Call Center Quick Hang
2. Quick Study: Business Radio Powered by the Wharton School
3. Complete Case Study: Lending Club P2P
4. Big data: Lung Cancer Micro array, MRI
5. **Statistics is not magic: Google Success/Failure**

# CASE STUDY: GOOGLE FLU TREND SUCCESS

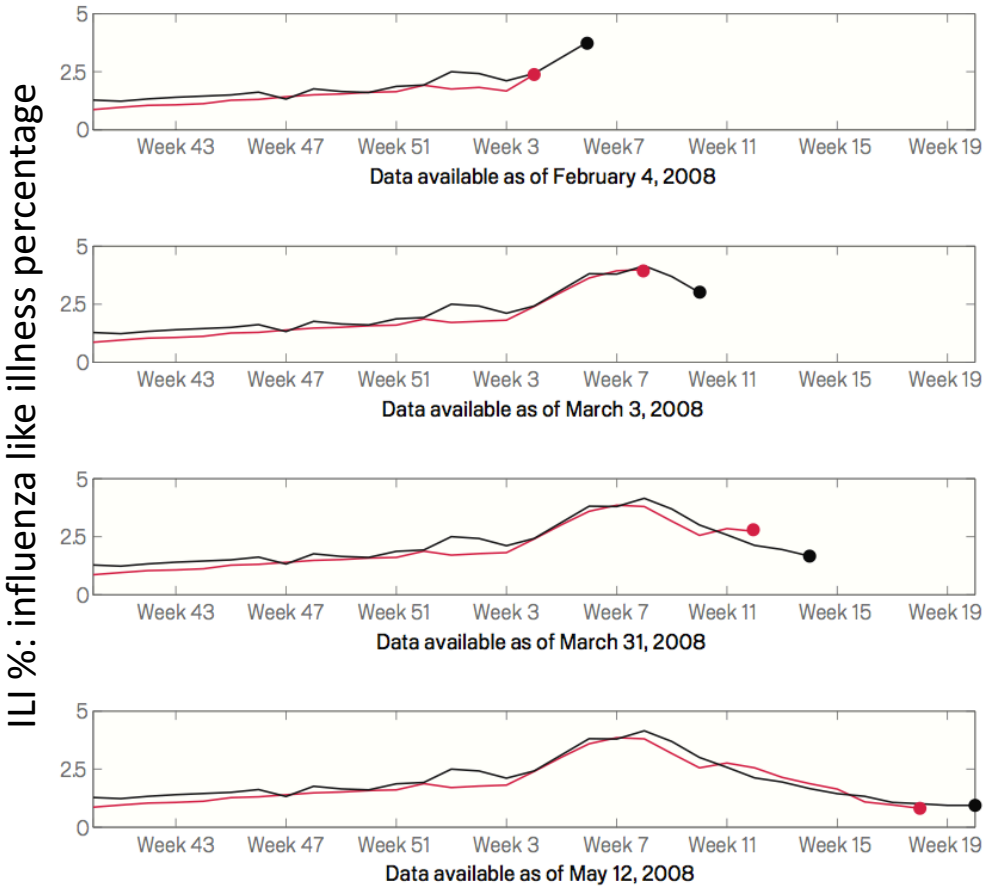
## US Center for Disease Control (CDC)

- Statistics of current number of flu cases
- Lagged data

## Google Flu Trends (GFT)

- Started from 2007
- Users' search records (**50 million most common queries**)
- The prediction worked for some time

# CASE STUDY: GOOGLE FLU TREND SUCCESS




**45 search queries selected that provided strongest prediction power for ILI from 50 million**

| Search Query Topic | Top 45 Queries N | Weighted |
|---|---|---|
| Influenza Complication | 11 | 18.15 |
| Cold/Flu Remedy | 8 | 5.05 |
| General Influenza Symptoms | 5 | 2.60 |
| Term for Influenza | 4 | 3.74 |
| Specific Influenza Symptom | 4 | 2.54 |
| Symptoms of an Influenza Complication | 4 | 2.21 |
| Antibiotic Medication | 3 | 6.23 |
| General Influenza Remedies | 2 | 0.18 |
| Symptoms of a Related Disease | 2 | 1.66 |
| Antiviral Medication | 1 | 0.39 |
| Related Disease | 1 | 6.66 |
| Unrelated to Influenza | 0 | 0.00 |
| | **45** | **49.40** |

Ginsberg et al. (Nature, 2009) : GFT, CDC

# CASE STUDY: GOOGLE FLU TREND FLOP

## Google Flu Trends (GFT)

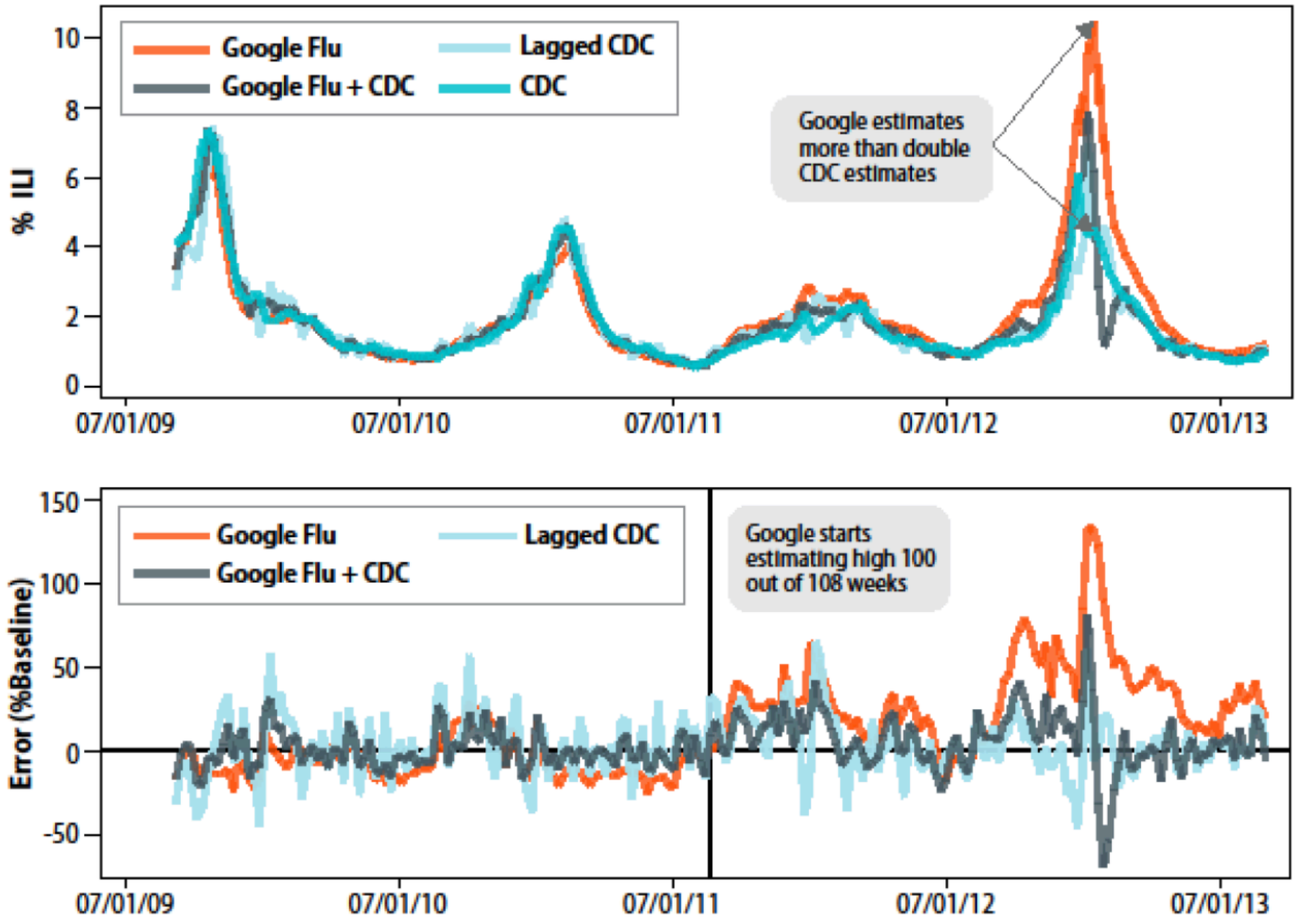
- Worked well for some time
- Overestimated by more than **50%** by 2011-2012

## GFT Fails: terminated in 2013

- No semantic analysis
- Not based on Science
- Correlation is NOT same as causality!

# CASE STUDY: GOOGLE FLU TREND

## FLOP

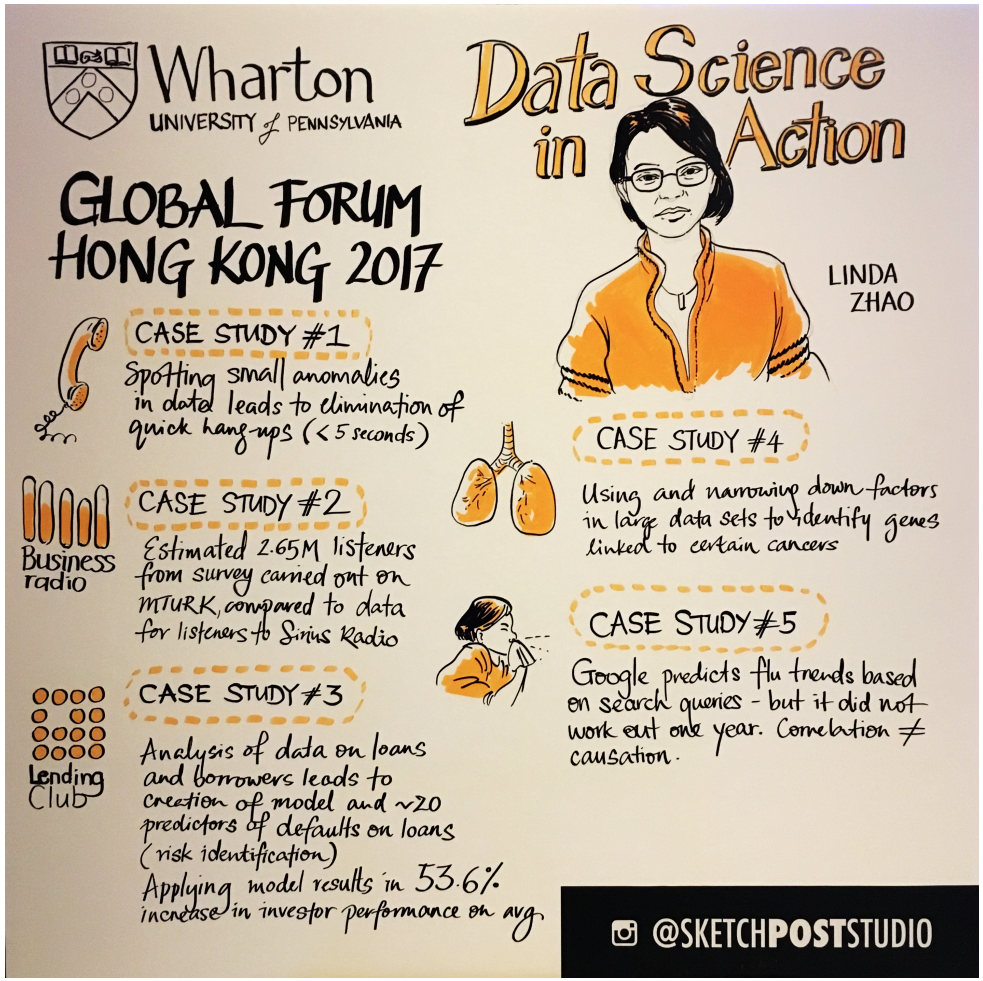



Lazer et al. (Science, 2014) : overfitting, change in search behavior

# TOOLBOX

- Reproducible report
  - Knitr
  - R markdown
  - Jupyter
- Database
  - RDBMS (MySQL, PostgreSQL, …)
  - NoSQL(MongoDB, …)
  - Hbase, …

# Recap


Wharton
UNIVERSITY of PENNSYLVANIA
GLOBAL FORUM HONG KONG 2017
Data Science in Action
LINDA ZHAO
CASE STUDY #1
Spotting small anomalies in data leads to elimination of quick hang-ups (< 5 seconds)
CASE STUDY #2
Business radio
Estimated 2.65M listeners from survey carried out on MTURK, compared to data for listeners to Sirius Radio
CASE STUDY #3
Lending Club
Analysis of data on loans and borrowers leads to creation of model and ~20 predictors of defaults on loans (risk identification)
Applying model results in 53.6% increase in investor performance on avg.
CASE STUDY #4
Using and narrowing down factors in large data sets to identify genes linked to certain cancers
CASE STUDY #5
Google predicts flu trends based on search queries - but it did not work out one year. Correlation ≠ causation.
@SKETCHPOSTSTUDIO