EDGAR Financial Analysis Computing Project

Robin Multon, Johanna Otto, Jose Zhu, Steven Kircher

The Problem We Saw

- Financial statement analysis

  Investment, competitive analysis, marketing …

- Security and Exchange Commission (SEC) : EDGAR database

- Decisions based on key financial ratios

Sample Products Co.
Income Statement
For the Five Months Ended May 31, 2015

|		
Sales		$100,000
Cost of goods sold		75,000
Gross profit		25,000
Operating expenses		
Selling expenses		
Advertising expense	2,000	
Commissions expense	5,000	7,000
Administrative expenses		
Office supplies expense	3,500	
Office equipment expense	2,500	6,000
Total operating expenses		13,000
Operating income		12,000
Non-Operating or other		
Interest revenues		5,000
Gain on sale of investments		3,000
Interest expense		(500)
Loss from lawsuit		(1,500)
Total non-operating		6,000
Net Income		$ 18,000

What is the Status Quo?

- Manually download data
- Time consuming, costs firms money and human capital
- Uncertainty, people make mistakes

Our Solution

Security and Exchange Commission (SEC) : EDGAR database
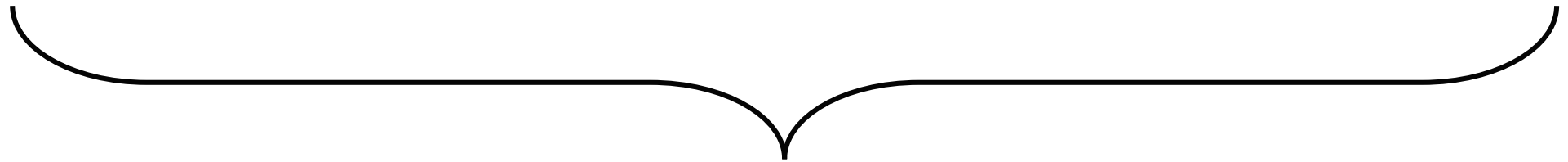
How about using it to automate financial and market analysis?

The Data Sets

- We need to control information in quantity and quality
- Analysis of the USA market

- What it is the data?
- What are our objectives?
- **What are the problems with the database?**
  - Some information is not mandatory
  - It is updated often

The options we have:

Compiled .txt database files ➔ it is not real time data's. However it is lighter ✔

XBRL files ➔ it is real time data however it is real heavy ✘
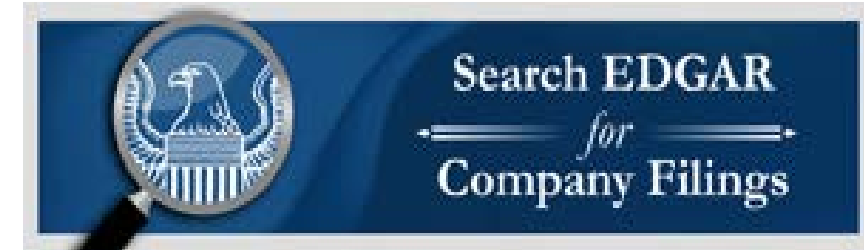
Our Desired Output – Where the Value Lies

- I'm lazy and my time costs money

- I want my computer to ease my job

- My job is to calculate the ratios... →
  Then to analyse the output... →
  To make choices →

1 . Computed ratios	
2 . Graph to ease analysis	
3 . Automate choices	

How Does Our Program Work?

Part 1 : Find & clean the data

Part 2 : Analyze the data

Part 1 : Finding & Cleaning the Data

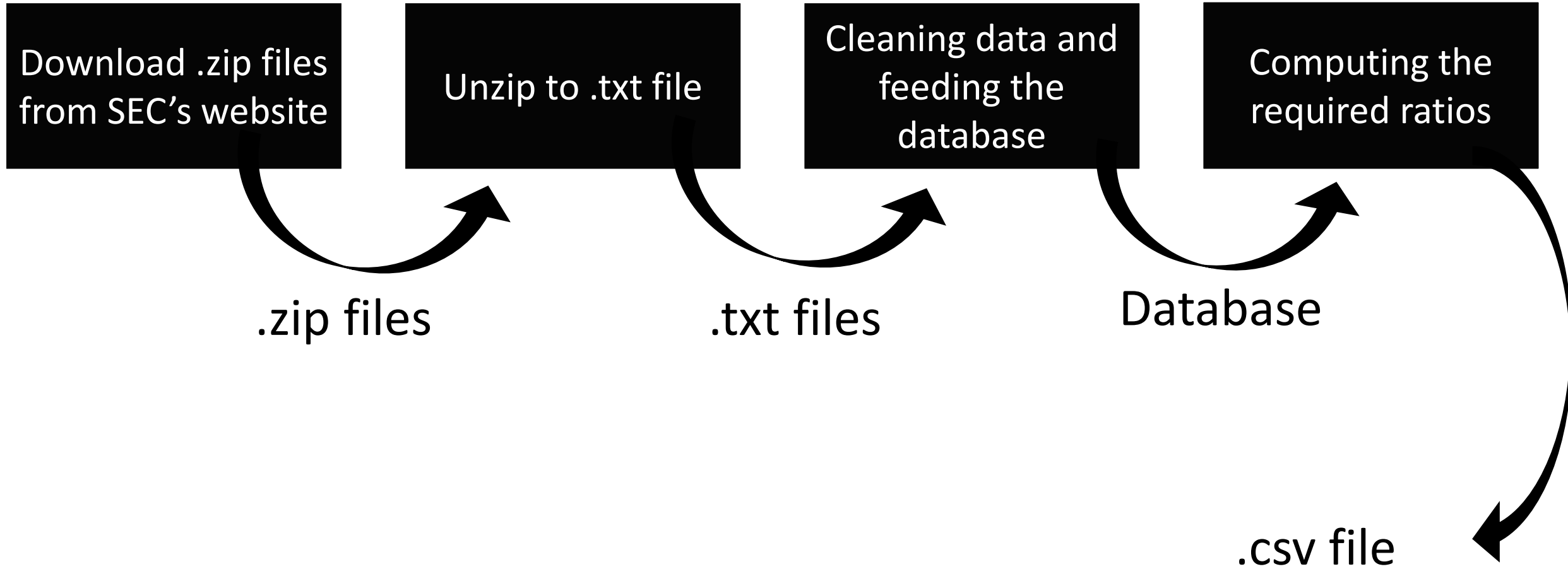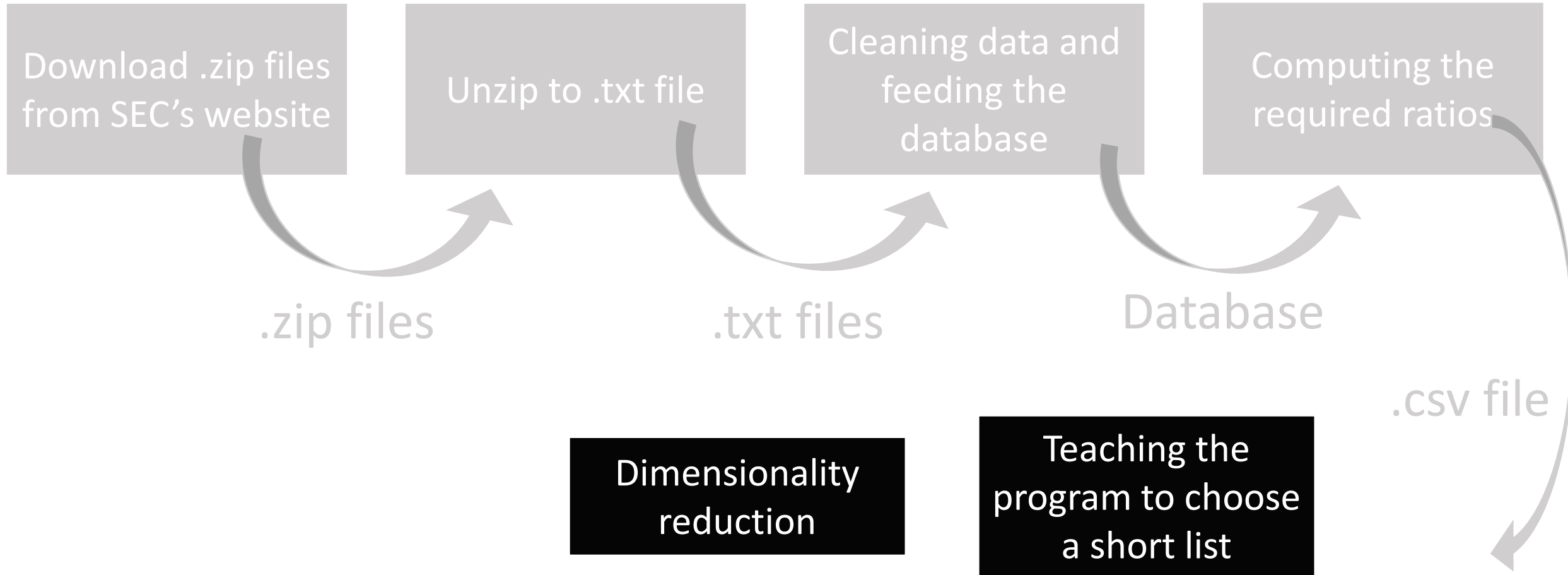| Download .zip files from SEC's website	Unzip to .txt file	Cleaning data and feeding the database	Computing the required ratios

Problems

- Missing data for the ratios
- GAAP accounting standards are loose
- Save or keep in RAM computed data
- Data structure

Part 1: Finding & Cleaning the Data

Part 2: Analysing the Data

Download .zip files from SEC's website

Unzip to .txt file

Cleaning data and feeding the database

Computing the required ratios

.zip files

.txt files

Database

.csv file

Dimensionality reduction

Teaching the program to choose a short list

Part 2: Analysing the Data

Download .zip files from SEC's website

Unzip to .txt file

Cleaning data and feeding the database

Computing ratios

1 . Computed ratios

.zip files

.txt files

Database

.csv file

Teaching the computer to choose a short list

3 . Automate choices

Dimensionality reduction

2 . Graph to ease analysis

Part 2: Analyzing the data

Teaching the program to choose a short list	Dimensionality reduction

Why? Automate a part of analysts' job

Ex : Creating a short list of companies to invest in

How? K Nearest Neighbors method

Two classes supervised learning : Good / Bad
K optimised in [1,20]

For whom?

Investors, competitors analysts, entrepreneurs …

What is produced? **A decision model**

Why? Help analysts understand a market

Ex : What are the strategies of a sector's companies ? Is this sector a monoply ? Which company should I invest in ? What are the key figures that make a company different from the other ?

How? Principal Component Analysis

For whom?

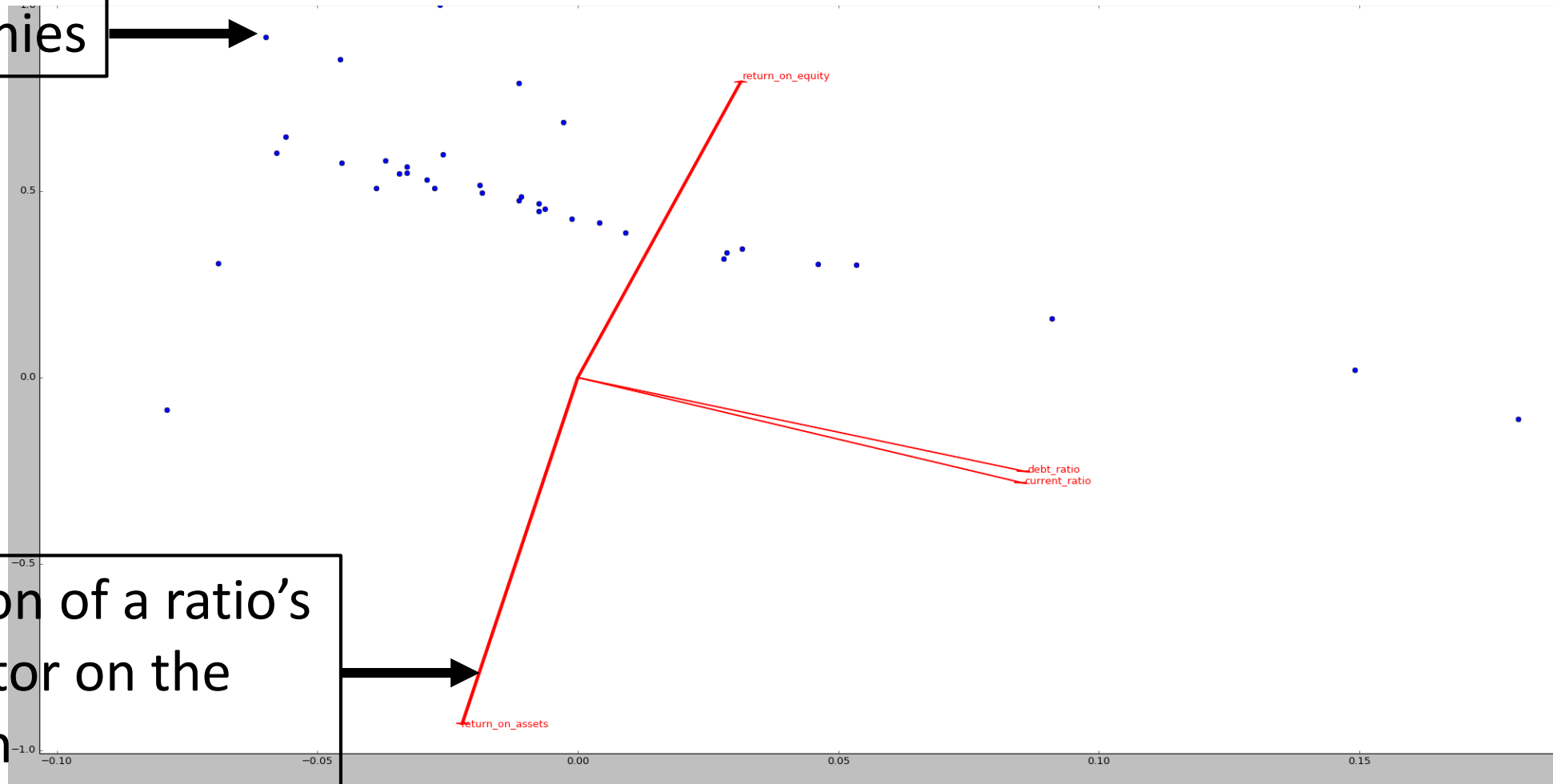Investors, competitors analysts, entrepreneurs …

What is produced? **A graph**

PCA Results for Computer Companies 2015 Q4

Companies

Projection of a ratio's unit vector on the PCA plan

How the Data will be Used

- Key financial ratios:
  - Analysis of successful industry investment compositions
  - Position in the industry
- Published reports
- Automation and more accurate

Problems with Our Program and Potential for Version 2

- GAAP loose standards

- Other methods than K Nearest Neighbors are possible
Random tree, …

- Accuracy is not great (0.66 in our example)

- Poor graph designs
Sorry for your eyes

- Needs improvements on the way data are handled
More pandas, less SQL

- But it is a **proof of concept**!

Acknowledgments for Code Inspiration

SciKit Learn website

Stack Overflow website

Our teacher's instruction