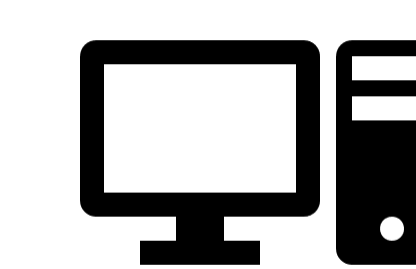# RepoMining – Repository and Mining of Temporal Data

Siomara Nieves, Jessica Nguy
Faculty Advisor/s: Dr. Philip Chan, Dept of Computer Science, Florida Institute of Technology

Objective

Create a web application that allows and simplifies the visualization and analysis of data by answering the following questions:

- ✓ Q1: Is there a significant change on my data?
- ✓ Q2: What are the top-k variables affecting it?
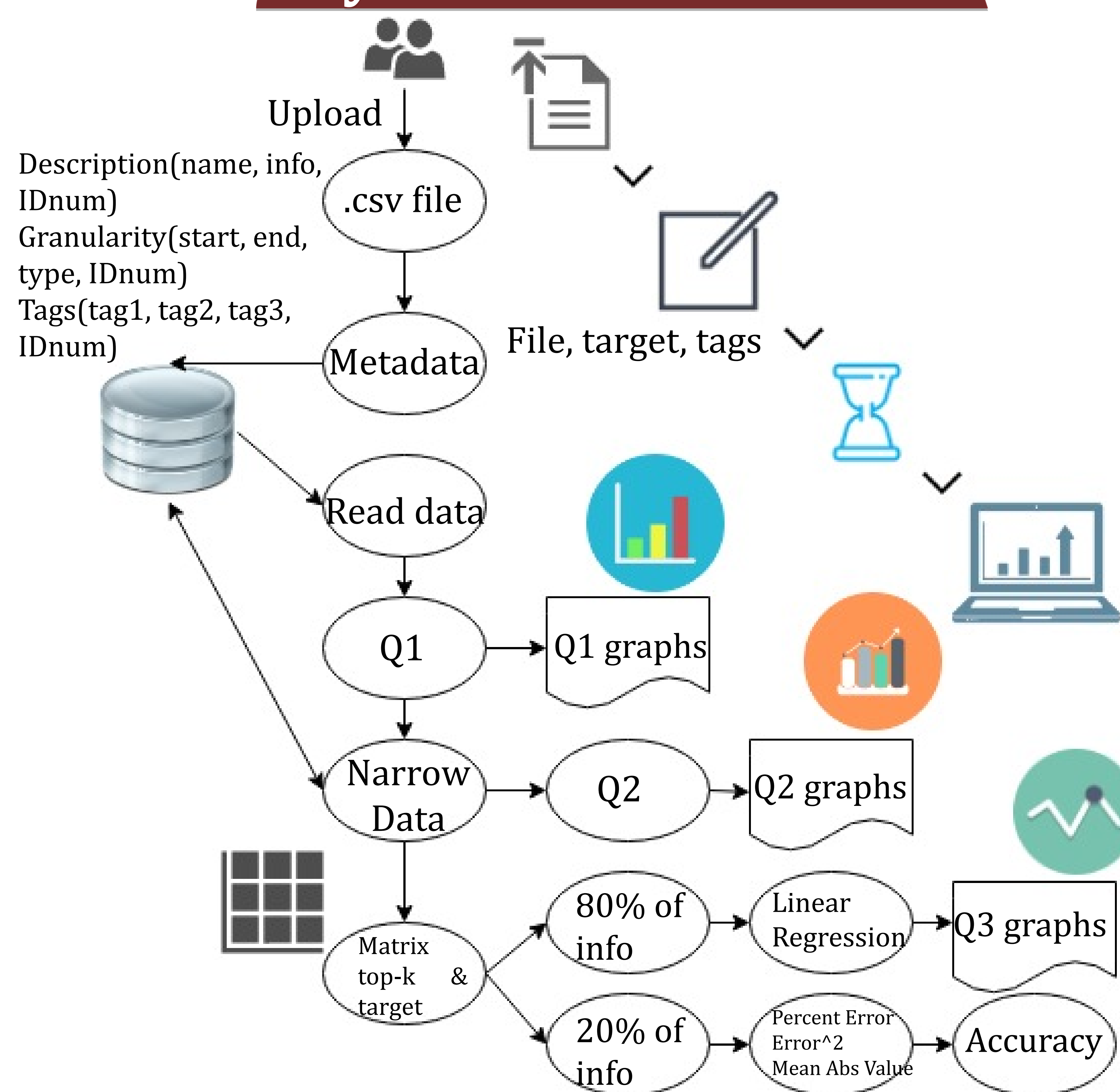- ✓ Q3: What is the next possible value?

Motivation

To simplify the appearance of data mining, collection, and analysis as well as to be able to predict future trends based on different types of information.

Tools

- ➤ Django – Web Application Framework
- ➤ Python – Programming Language
- ➤ JSON – Handling data objects
- ➤ HTML/Bootstrap/CSS/JS – Front-end

System Architecture

Upload

Description(name, info, IDnum)
Granularity(start, end, type, IDnum)
Tags(tag1, tag2, tag3, IDnum)

.csv file → Metadata

File, target, tags

Read data

Q1 → Q1 graphs

Narrow Data → Q2 → Q2 graphs

Matrix top-k & target → 80% of info → Linear Regression → Q3 graphs

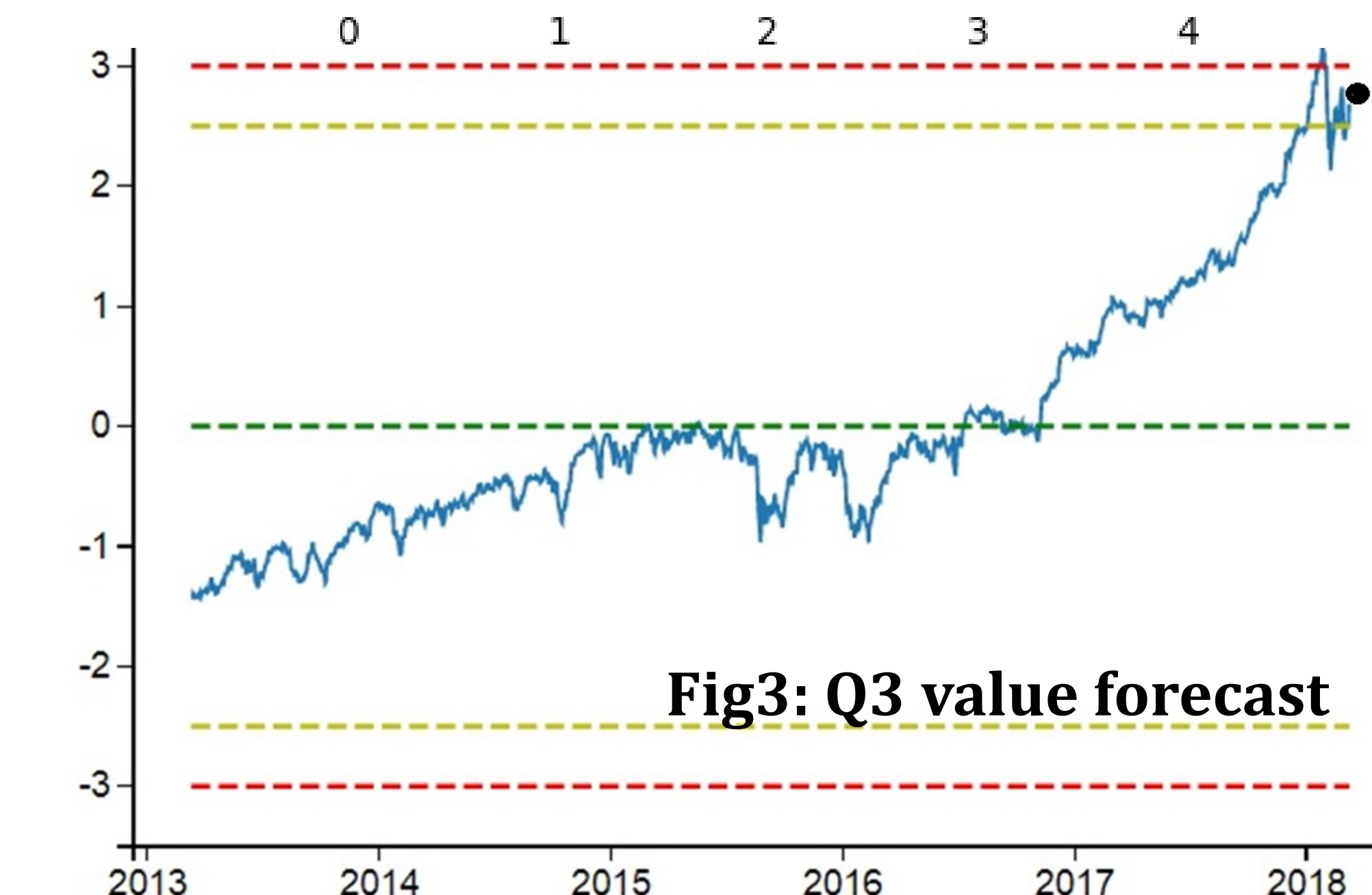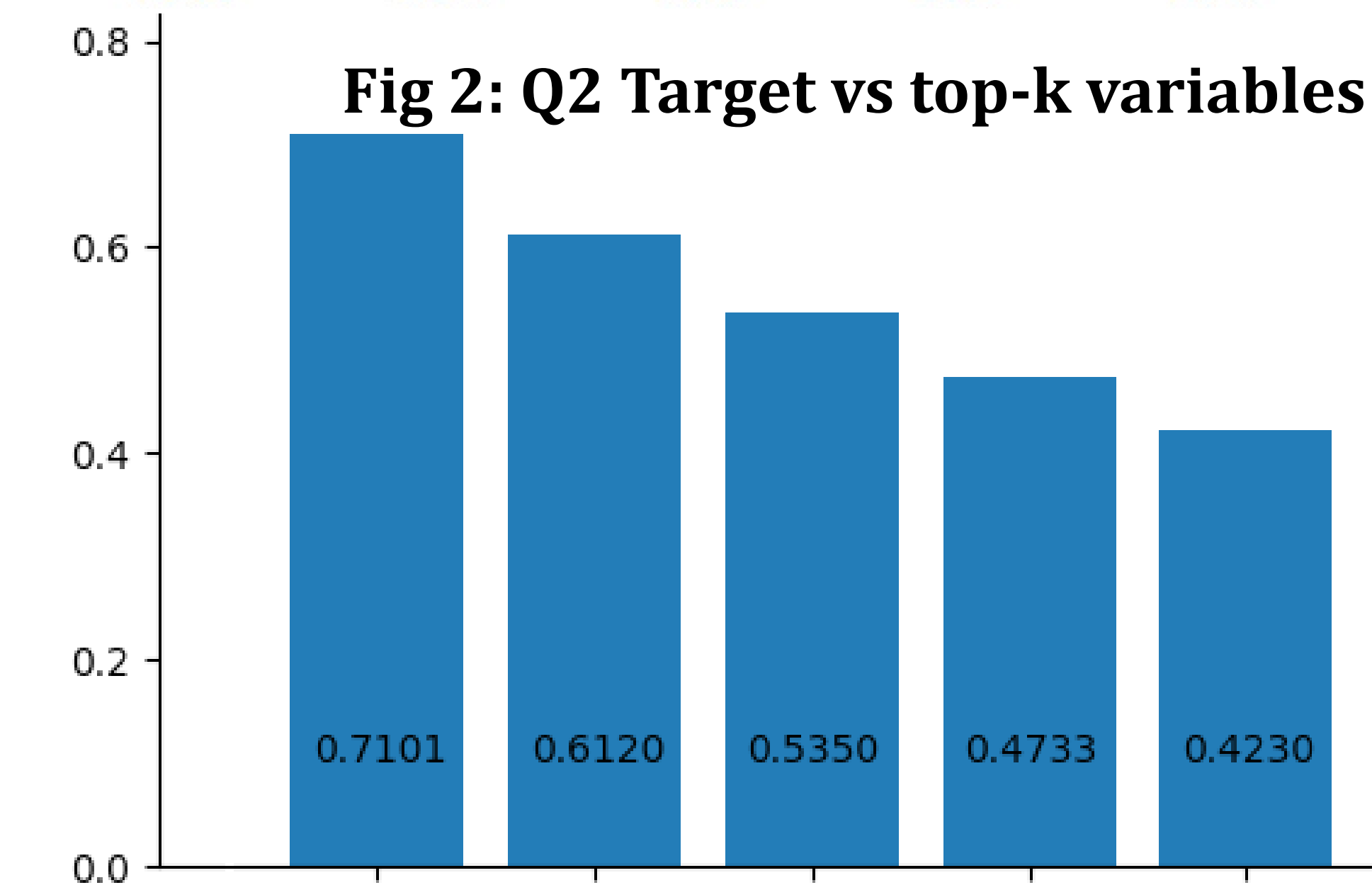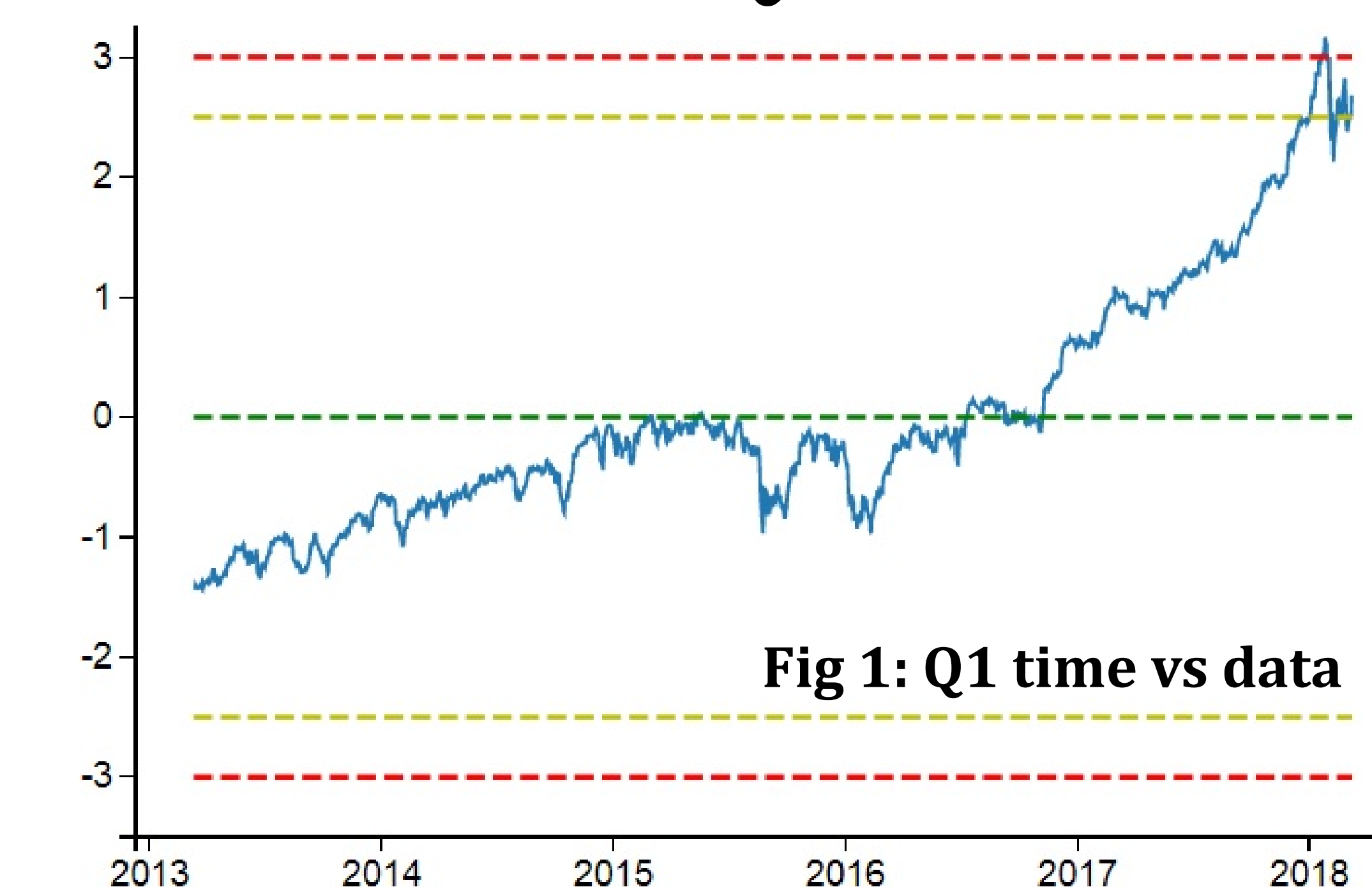Matrix top-k & target → 20% of info → Percent Error Error^2 Mean Abs Value → Accuracy

Algorithms

- **Q1**: time vs data, z-score, Δx, Δz-score
- **Q2**: NarrowData, Pearson Correlation
- **Q3**: Linear Regression
- **NarrowData**: Compares the timestamps of all data in storage and returns the ones whose starting timestamp falls before the target's and whose end timestamp falls after the target's. Uses lag of n-times to determine if past data affected present data. Calculates the correlation value of the matching data and returns the top-k variables affecting the target.
- **Linear Regression**: Used along with top-k values from to predict the target's future timestamp.

Results

Q1: Time vs data
Time vs z-score
Time vs Δdata
Time vs Δ z-score

Q2: Data vs top-k var.
Data vs Corr. top-k
Top-1 corr. top-k

Q3: Forecast

Fig 1: Q1 time vs data

Fig 2: Q2 Target vs top-k variables

0	1	2	3	4
0.7101	0.6120	0.5350	0.4733	0.4230

Fig3: Q3 value forecast