Georgia Institute of Technology

Dr. Tucker Balch
Associate Professor
School of Interactive Computing

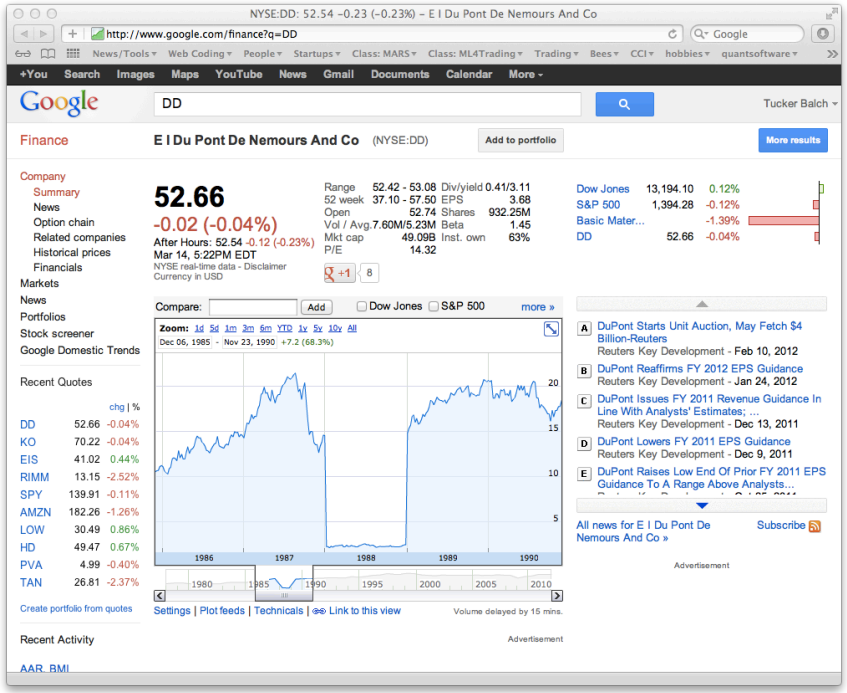
Computational Investing, Part I

123: Data Sanity and Scrubbing

Find out how modern electronic markets work, why stock prices change in the ways they do, and how computation can help our understanding of them. Learn to build algorithms and visualizations to inform investing practice.

School of Interactive Computing

Bad Data Example



- Symbol: DD
- 1988

Examples of Bad Data

- Failure to adjust for splits.
- Orders of magnitude drops, followed by offsetting orders of magnitude climbs.
- Database updates missing significant chunks of data/symbols.

Why Bad Data is Bad

- Automated strategies may exploit bad data, then fail with real data.
- You might think you've "discovered" something.

Sanity Checks

- Scan new data for ~50% drops or 200% gains (probably a split). Very rare for real data.
- NaNs in DOW stocks (probably data feed bad).
- Recent adjusted prices less than 0.01
- NaNs > 20 trading days?

Scrubbing

- Remove or repair?
  - Easier, more reliable to remove.
- Can only repair if you have multiple sources.

Summary

- Good data is important.
- You may discover false strategies otherwise.

Next

- Overview of homework 3.