Georgia Institute of Technology

Computational Investing, Part I

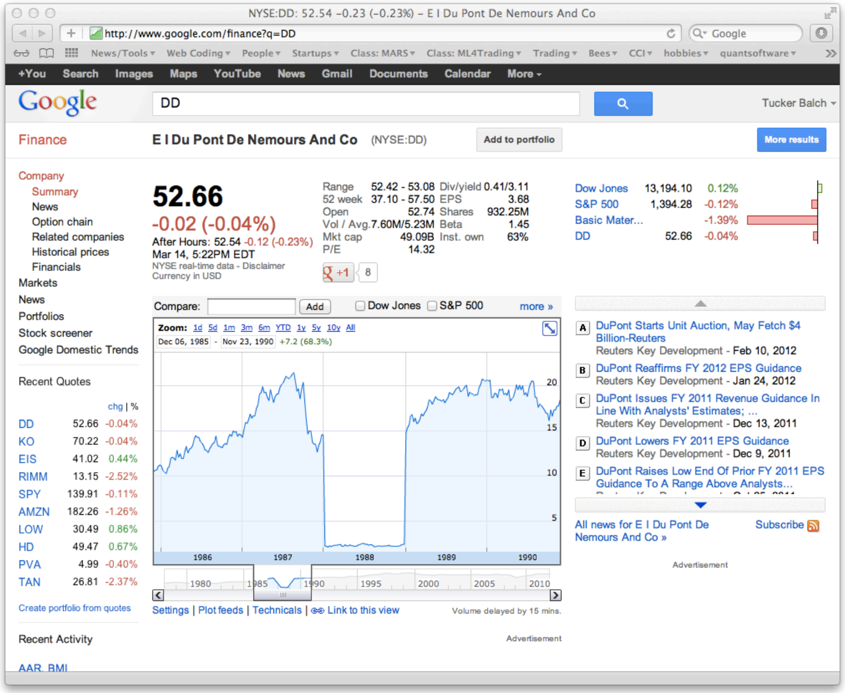
Dr. Tucker Balch
Associate Professor
School of Interactive Computing

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



E I Du Pont De Nemours And Co (NYSE:DD)

52.66
-0.02 (-0.04%)
After Hours: 52.54 -0.12 (-0.23%)
Mar 14, 5:22PM EDT
NYSE real-time data - Disclaimer
Currency in USD

Range	52.42 - 53.08	Div/yield	0.41/3.11
52 week	37.10 - 57.50	EPS	3.68
Open	52.74	Shares	932.25M
Vol / Avg.	7.60M/5.23M	Beta	1.45
Mkt cap	49.09B	Inst. own	63%
P/E	14.32		

Dow Jones	13,194.10	0.12%
S&P 500	1,394.28	-0.12%
Basic Mater...		-1.39%
DD	52.66	-0.04%

Recent Quotes

		chg | %
DD	52.66	-0.04%
KO	70.22	-0.04%
EIS	41.02	0.44%
RIMM	13.15	-2.52%
SPY	139.91	-0.11%
AMZN	182.26	-1.26%
LOW	30.49	0.86%
HD	49.47	0.67%
PVA	4.99	-0.40%
TAN	26.81	-2.37%

A DuPont Starts Unit Auction, May Fetch $4 Billion-Reuters
Reuters Key Development - Feb 10, 2012

B DuPont Reaffirms FY 2012 EPS Guidance
Reuters Key Development - Jan 24, 2012

C DuPont Issues FY 2011 Revenue Guidance In Line With Analysts' Estimates; ...
Reuters Key Development - Dec 13, 2011

D DuPont Lowers FY 2011 EPS Guidance
Reuters Key Development - Dec 9, 2011

E DuPont Raises Low End Of Prior FY 2011 EPS Guidance To A Range Above Analysts...

- 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.