PyData Talk

Replicating Google Correlate with Wikipedia Data

The Project

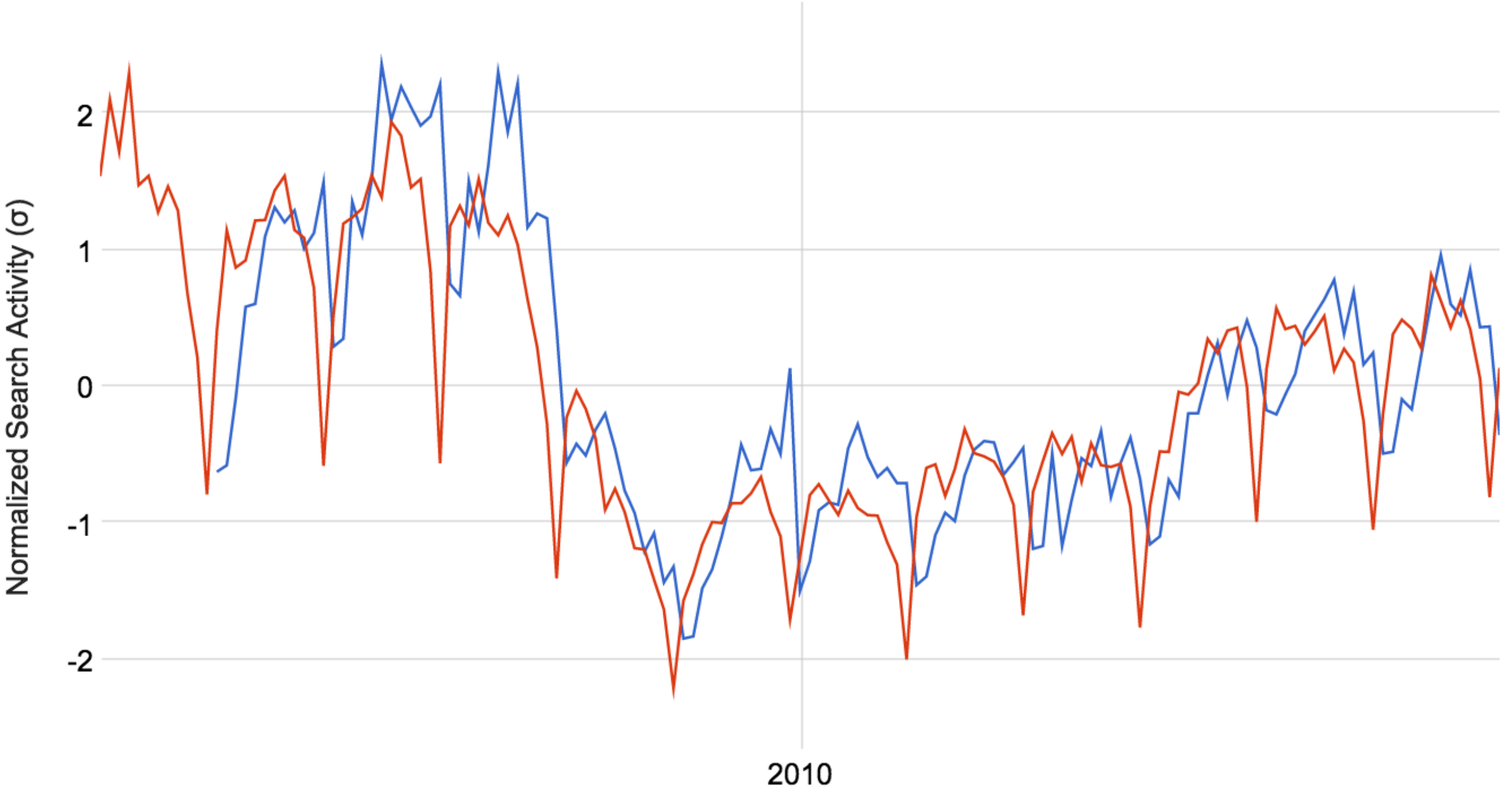
- Data Issues in Macroeconomics
- Many theories, little data – Overfitting models
- Fortunately everything we do is tracked

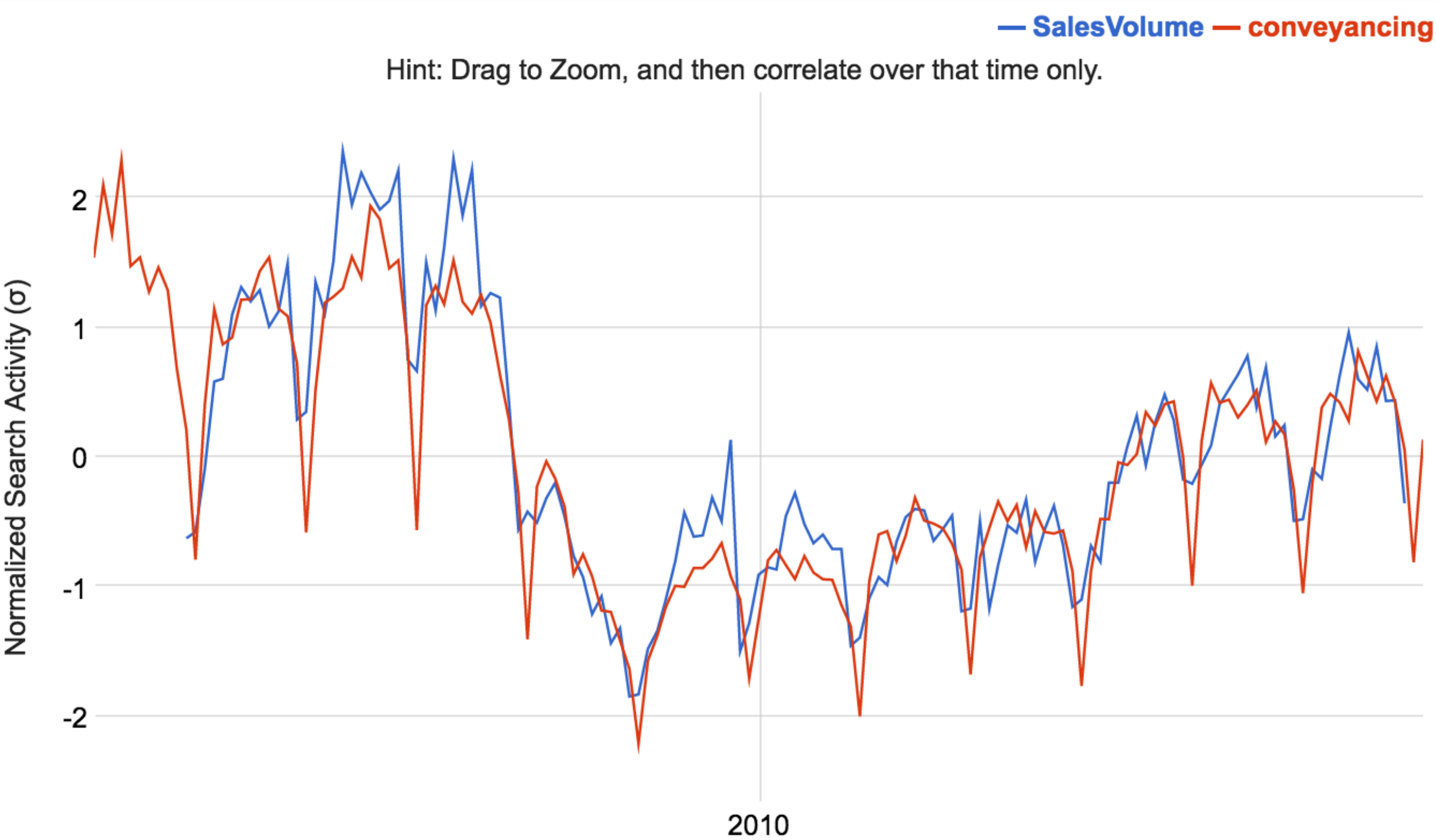
Google Correlate

- Google has lots of data
- Trends/correlate is one of the few places to get a view
- Pick an indicator and test it
- Monthly data, available, significant... Housing

— SalesVolume — conveyancing

Hint: Drag to Zoom, and then correlate over that time only.




— SalesVolume — conveyancing
Hint: Drag to Zoom, and then correlate over that time only.
Normalized Search Activity (σ)
2
1
0
-1
-2
2010

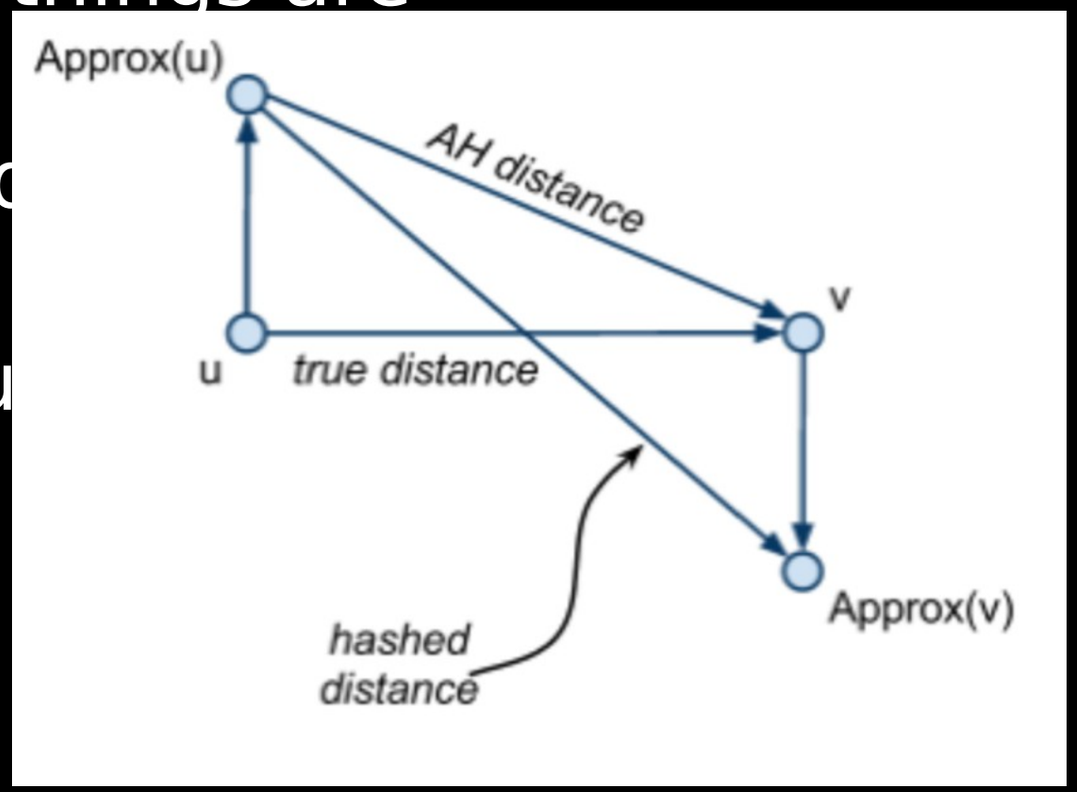
Wikidata

- Google data is private – Wikimedia Foundation publish theirs
- Page Views, by the hour;
- RDF context, SPARQL search;
- Supporting tools, APIs and documentation

“Nearest Neighbor Search in Google Correlate”

- Google published paper, by D Vanderkam et al, 2013

- Correlation – How similar two things are
- Pearsons (∈ [-1, 1]) and Euclic
- Quantise everything, precalcu



Practically

- Single language site – Loading Swedish data into pandas
- `df.pivot`, into a single time-series table
- `df.transform` to normalise, μ=0 σ=1

- Create a `pd.Series` for approximation values
- Build an index table for nearest approximation for each value
- Distance mapping table for series
- Hash incoming series, lookup per date with index value and sum distances

Fast and Accurate

- Quantise based on distribution (KDE `scipy.stats` or Harrell-Davis Quantiles `scipy.stats.mstats.hdquantiles`)
- Minimise loss with different distribution per period or collection of periods
- Early exit
- 2nd pass with Pearson’s with original values

Results

- Poor implementation - slower for than just using `pandas.corr` with
- Plausible results
  - Cars, Accounting, Computing and Biosciences
- Not possible to prove – Autocorrelation, small dataset
- $R^2 = 0.985$ for Konfirmeringsbias, so maybe not.