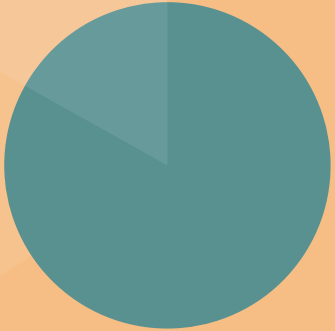
Datasets to use and avoid in quantitative portfolios



Ankit Awasthi

Quantitative Portfolio Manager, qplum

www.qplum.co/events

qplum

Investment Mandate

Goal is to build a high capacity data driven investment strategy that scale up to multi billion dollars.

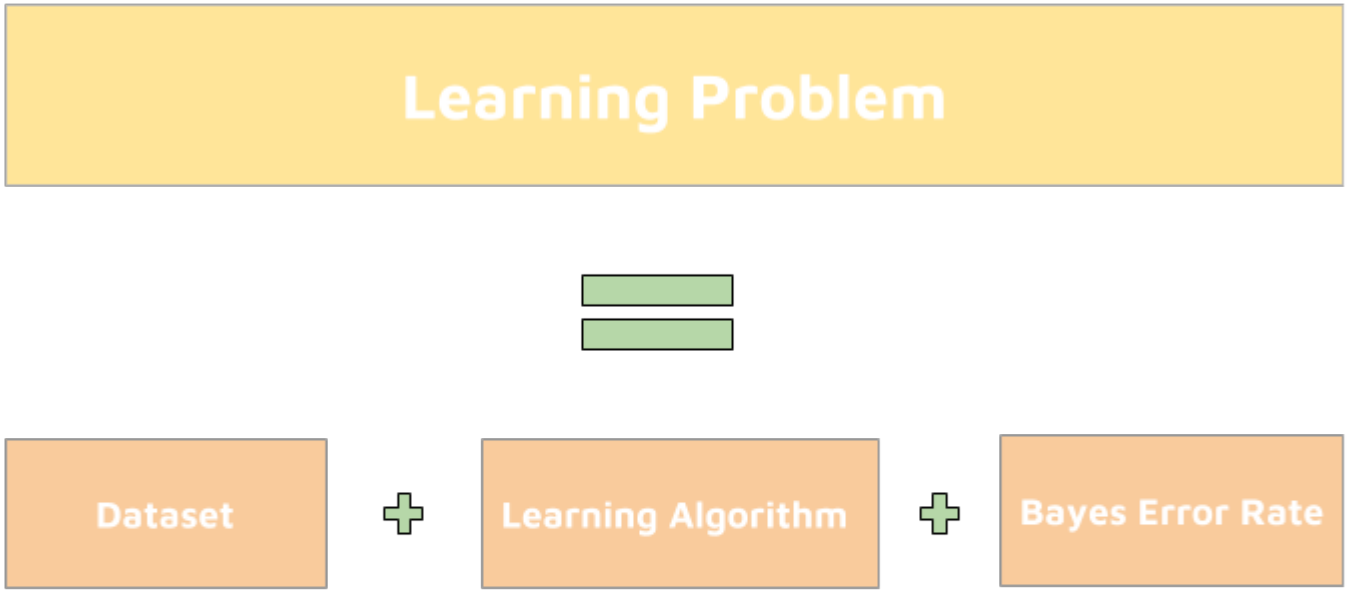
What data do you need ?

- > 30 years of EOD data
- > 5 years of intraday high frequency data
- > 30 years of Macroeconomic data
- Across Equities, Stock Indices, Futures, Mutual Funds, Fixed Income, Currencies

Outline of the talk

- Components of a generic learning problem
- How to select datasets for investment strategies
- Concerns on the use of alternative datasets
- Which datasets does qplum use

A first principles approach to selecting datasets



Bayes or Irreducible or Lowest Possible Error

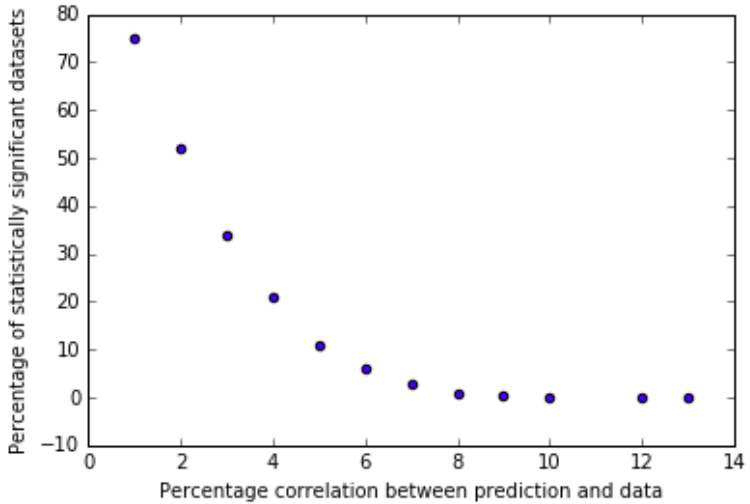
- Serves as a benchmark to improve performance of learning algorithms.
- Very hard to estimate for most problems.
- Human error rate is often used as proxy for tasks as which humans are good at.
- Predicting prices/returns is a hard problem - expected R-squared is quite low

Let’s do two experiments

- We sample a large number (N) of points (x - independent, y - dependent) randomly from a multivariate normal distribution where ‘x’ and ‘y’ are uncorrelated
- Next, we divide these points into datasets of ‘n’ points
- Let ‘r’ be an estimate of maximum correlation (in other words, irreducible error rate) for the concerned learning problem
- A dataset is called statistically significant if correlation for the dataset > r
- In the first experiment, we check how the percentage of statistically significant datasets changes as ‘r’ changes
- In the second experiment, we check how percentage of statistically significant datasets changes as ‘n’ changes

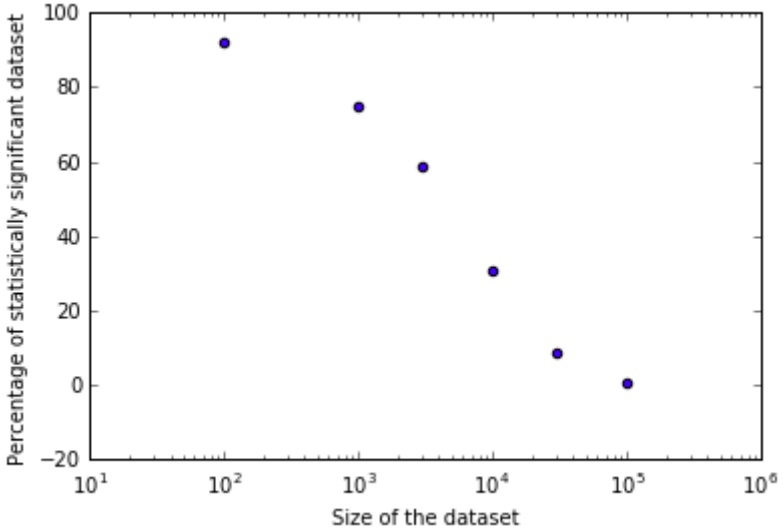
Hard problems are easy (to overfit)!

As the expected accuracy increases, the chances of a random dataset qualifying as statistically significant reduces drastically.



Here, for a dataset of 1000 data points, given an expected correlation of 1%, around 75% of datasets could appear to be statistically significant, whereas for expected correlation values of 13%, it is close to 0.

Hard problems are easy (to overfit)!



Given expected correlation of 1%, as the size of the dataset increases from 100 data points to 100,000 data points, the chances of a random dataset appearing statistically significant goes from around 92% to close 0%

Select a dataset that...

- Is clean and reliable
- Has high coverage
- Has long history
- Is predictive over long horizons

Data Cleaning

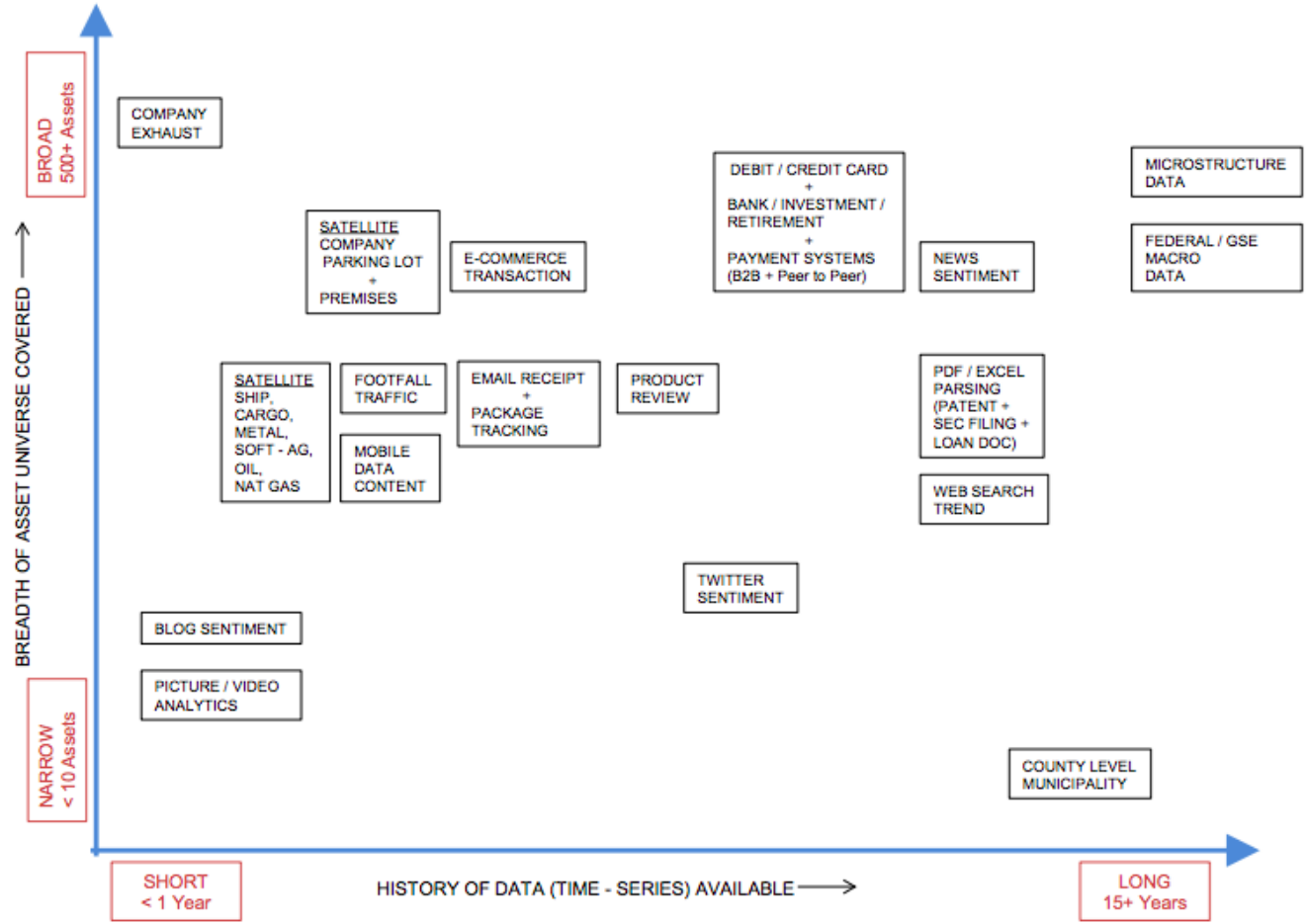
- Addressing lookahead bias
  - Revisions
  - Normalization
- Filling missing data carefully - ‘0’, NA, previous value, moving average
- Detecting outliers
- Real Time Data Feed Vs Historical Data Feed

High Coverage leads to..

- Better diversification
- Robustness against overfitting
- Higher Capacity

Longer History

- Allows use more sophisticated learning algorithms - deep learning
- Less chances of overfitting
- Persistence of risk premium
- Coverage over different economic regimes



Datasets mapped on the scale of coverage and history (*Source: JPMorgan)*

Concerns

- Alpha Decay
- Overfitting
- Spurious Relationships
- Lack of transparency
- Escalating costs
- Compliance

Alpha Decay

- Datasets which deliver alpha over high coverage are extremely rare
- It is inevitable
- Lifecycle of datasets getting increasingly small
- Need to constantly look for new edge - worthy data sources

Overfitting

- Shorter histories makes use methods like walk-forward optimization infeasible
- Datasets evolve drastically over time
- Higher frequency predictions reduce overfitting but also reduce capacity
- State of the art learning algorithms like deep neural networks can’t be used

Spurious Relationships

Oct. 3, 2008 - *Rachel Getting Married* opens: BRK.A up .44%

Jan. 5, 2009 - *Bride Wars* opens: BRK.A up 2.61%

Feb. 8, 2010 - *Valentine's Day* opens: BRK.A up 1.01%

March 5, 2010 - *Alice in Wonderland* opens: BRK.A up .74%

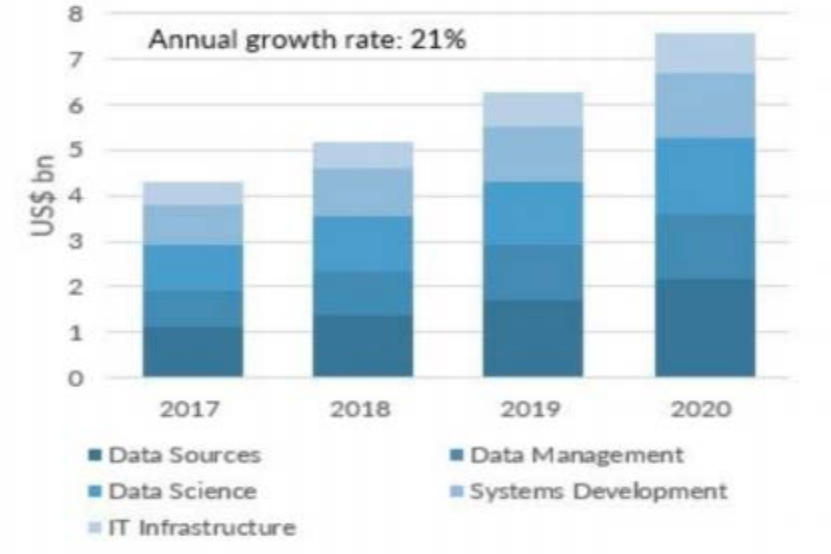
Nov. 24, 2010 - *Love and Other Drugs* opens: BRK.A up 1.62%

Nov. 29, 2010 - Anne announced as co-host of the Oscars: BRK.A up .25%

Source : Huffington Post

Source : The Atlantic

Escalating Costs



Source: Opimas

Opimas forecast spend on alternative datasets to reach $2bn by 2020

Which datasets does qplum use ?

- 25 years of EOD price - volume data
- 7 years of high frequency tick data
- > 50 years of macroeconomic data
- > 1000 years of hypothetical price - volume data*

** Our use of hypothetical data to be discussed in a future webinar!*

Please email your questions to: contact@qplum.co

qplum