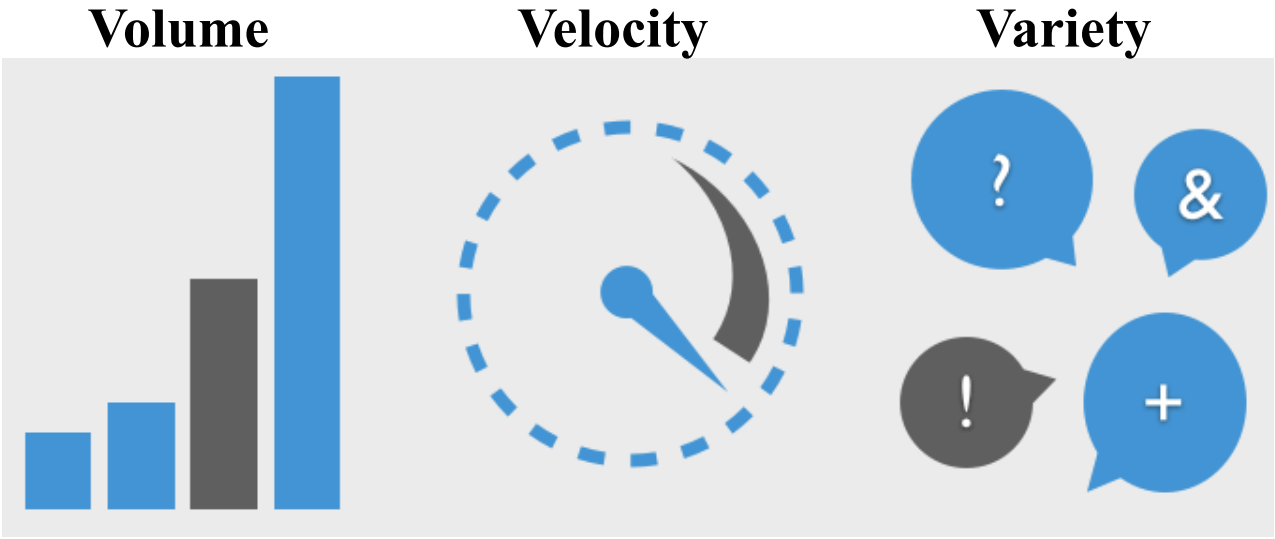
The Vs of Big Data

School of Information Studies
Syracuse University

The Three Vs of Big Data



"Too Much" "Too Fast" "Too Complicated"

Three Vs

Volume: Sheer size of data is too large to be processed by a single system.

Velocity: The rate at which new data arrives is too frequent to be processed by a single system.

Variety: Data are unstructured or semistructured, so compute resources must add structure before it can be used.

- E.g., a company's website generates more logging data in 24 hours than its ETL system can process in a 24-hour period.

Implications of the Three Vs

1. Require a scale-out and distribute data over several systems to meet scalability demands.
2. Must have all three Vs. Having one or two of the three can still be met through other methods.
3. Do not scale out because you can; do it because you must.

Other Vs of Big Data

Veracity: Uncertainty of your data. How can we be confident in the trustworthiness of our data sources?

- E.g., matching a tweet to a customer, without knowing his or her Twitter handle.

Viability: Can we predict results from the data? Can we determine which features serve as predictors?

- E.g., discovering patterns among customer purchase habits and unfavorable weather conditions.

Value: What meaning can we derive from our data? Can we use them to make good business decisions?

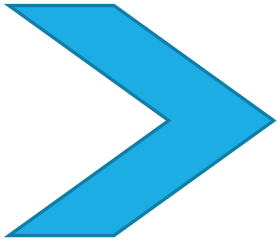
- E.g., increase inventory levels of potato chips two weeks before the Super Bowl.

Birthplace of Big Data

Three companies: Google, Facebook, Yahoo!

These companies had so much data that **enterprise DBMSs** could not meet their reporting requirements.

Time to process one day of data



Number of hours in a day

But, I'm Not Google. Do I Need This?

Struggling with data volume, velocity, or variety

Insufficient resources to store process, and analyze your data