SAND2016-11372PE

Exceptional service in the national interest

Sandia National Laboratories

Photos placed in horizontal position with even amount of white space between photos and header

Introduction to Social Analytics

Jacob Caswell

To be included in three part talk at LANL, July 28, 2016

With Richard Barrett and George Stelle

U.S. DEPARTMENT OF ENERGY

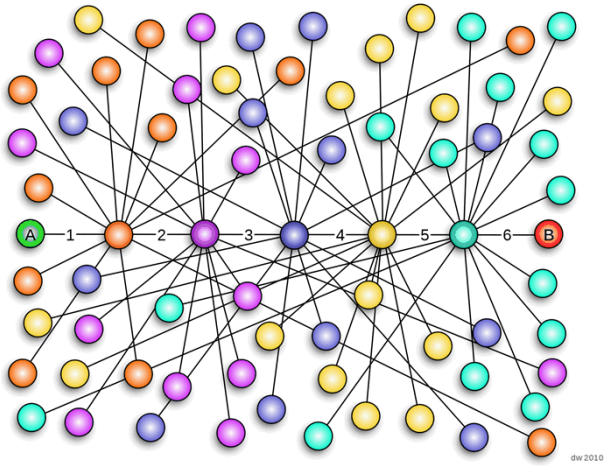
NNSA National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND NO. 2011-XXXXP

Introduction to Social Analytics

- What is Social Analytics?
  - Small World Experiment
  - Network Theory
  - Information Theory
  - Machine Learning / Textual Analysis

- We describe patterns in human relationships and interactions

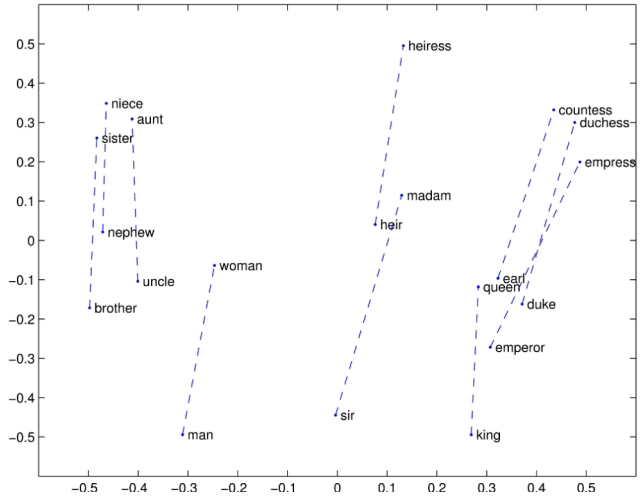


6 Degrees of Separation,
Image from Wikimedia Commons

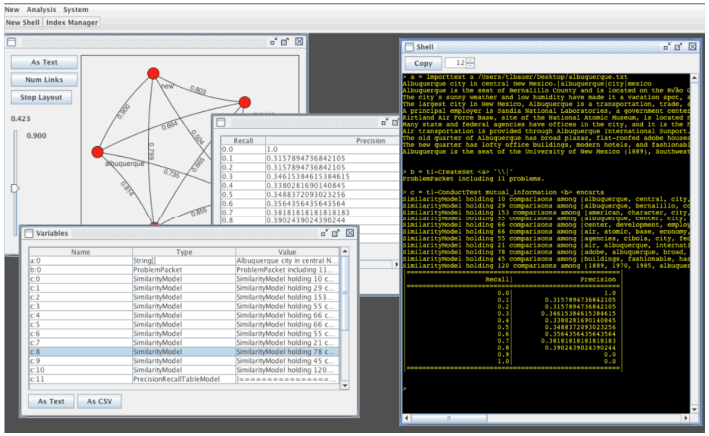
6 Degrees of Kevin Bacon,
Ian McKellen, Michael Fassbender Kevin Bacon,
Image from Wikimedia Commons

Citrus

- Robust text analysis java packages
  - Word vectorization
  - Document Similarity
  - Social Network Extraction
  - Web Crawling
- Aims to be a one-stop shop for all things text analytics



Word vector visualization
Image by http://nlp.stanford.edu/projects/glove/

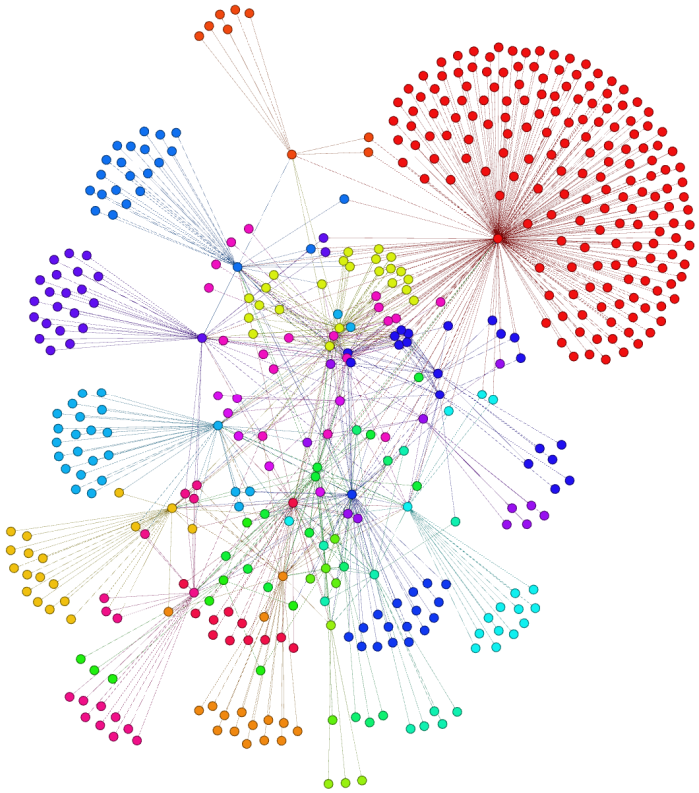


Citrus screenshot

Social Analytics & Chapel

Sandia National Laboratories

- Why HPC and Social Analytics?
  - Huge Data
  - Clustering
  - Multi-task problems
  - Verification, Validation, and Uncertainty Quantification (VVUQ)
- How can these benefit from using Chapel?

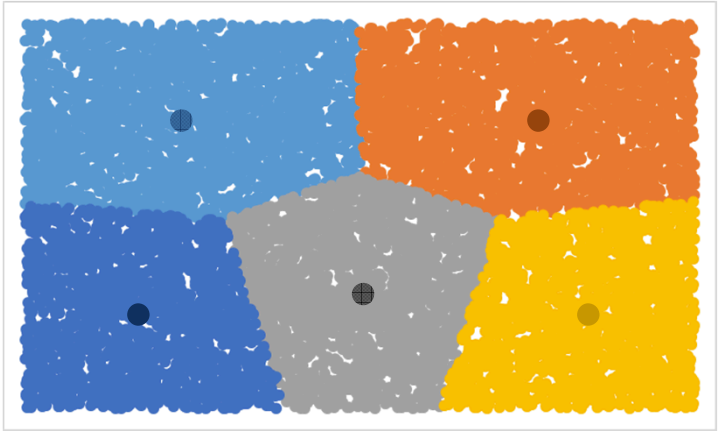


Network visualization

Example: K-Means Clustering

Sandia National Laboratories

- How do you simplify data?
  - How to make Wikipedia scale corpuses understandable?
  - Initialize k representative points
  - Assign points to closest centroids
  - Find “center of mass”
- Issues: time consuming, requires specified k, only finds local optimum

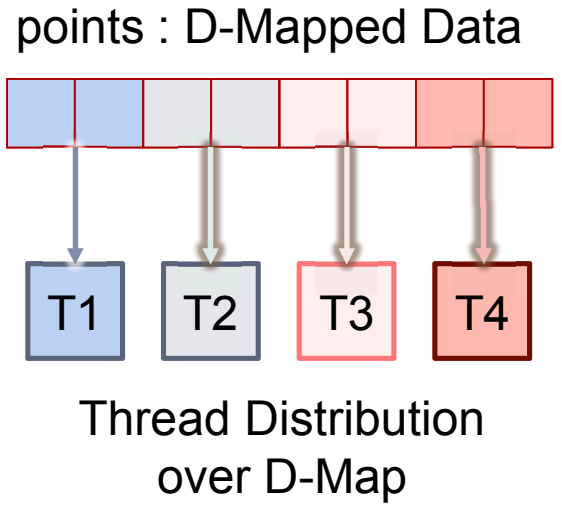


Sample Clustering from Chapel Implementation
10,000 random 2D data points, k=5

Chapel K-Means Data Parallelism

Sandia National Laboratories

- Share the load
  - Specify or modify desired d-map
  - Can distribute data over threads or locales
  - Cluster data on Locale
  - Repeat
- Code:



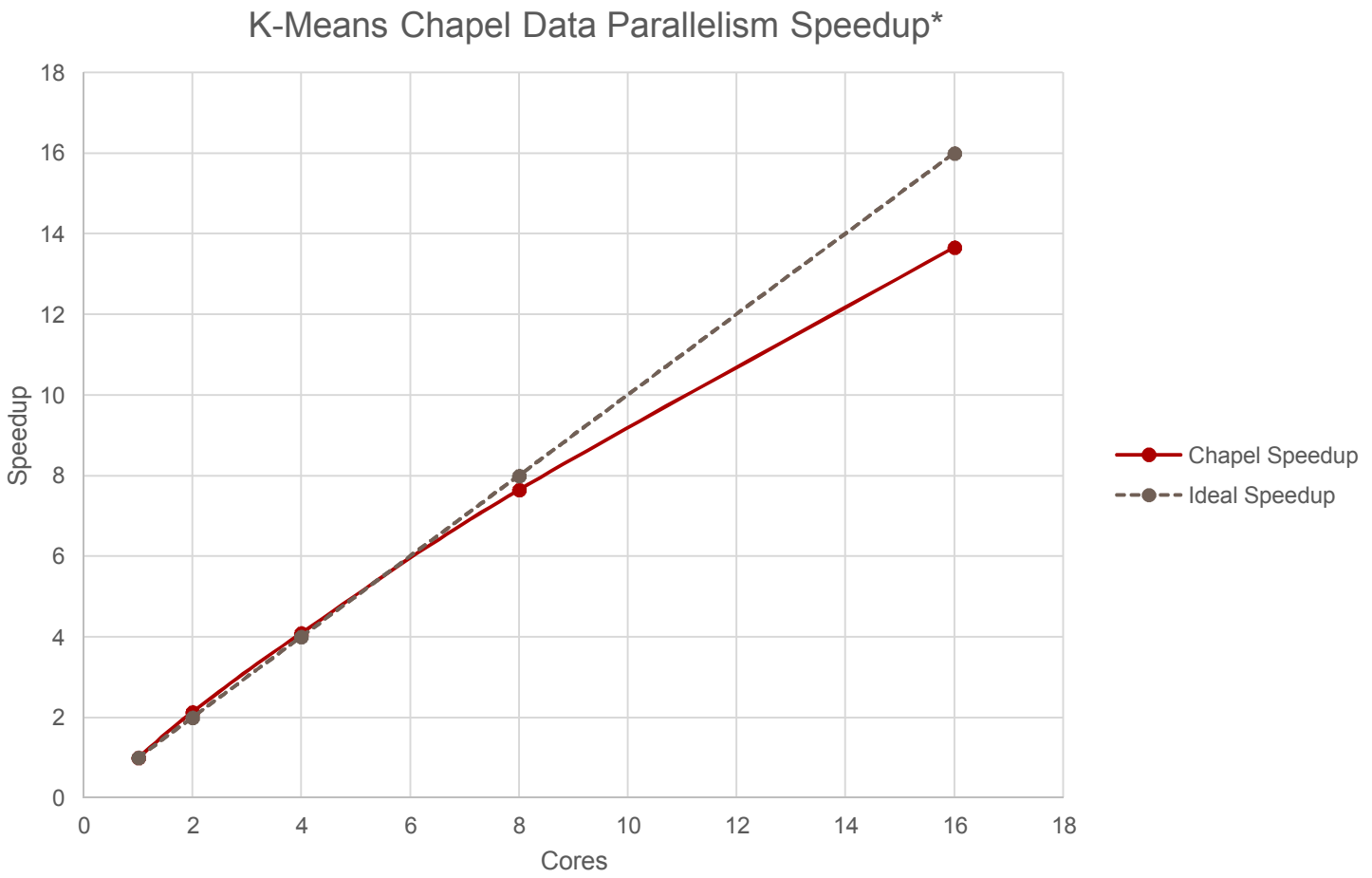

```
use BlockDist;
config var numPoints = 1000000, numClusters=10;

var Space  = {1..numPoints};
const
    P : domain(1) dmapped Block(boundingBox=Space) = Space;

var points : [P] dataPoint;

forall p in points {
        // Find the closest centroid
        // Update assignment
}
```

Chapel K-Means Speedup

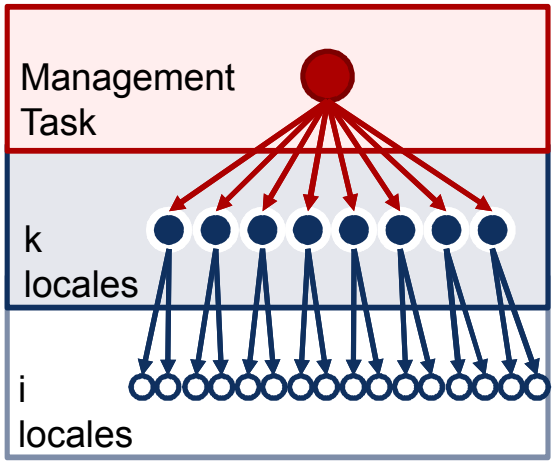


**Speedup compared to time per iteration for one core. Averaged over 10 runs per core.*
Executed on an Intel Xeon E5-2670 processor
Tested by grouping 2000 random points in 300D space into 40 clusters

Chapel K-Means Task Propagation

Sandia National Laboratories

- Task-Data Cascade Parallelism
  - Accept desired parameters from Citrus
  - Spawn off different k-locales as desired
  - Spawn off different i-locales as desired
  - Cluster data using on locale



- Code:

```
config const var ni=4, lowK=2, highK=82;
const taskGrid = reshape(Locales, {lowK..highK, 1..ni});
const kLocales = taskGrid[..,1];

coforall kl in kLocales {
        initializations(k, taskGrid[k,..]);
}

proc initializations(k, initArray) {
        coforall i in initArray {
                // Organize data into k clusters
        }
}
```

Benefits

- Options, options, options
  - Minimal disruption of main code body
  - Large datasets
  - Parameter Sweep
  - VVUQ