Case study: Crimson Hexagon

* **Tech** 
  + Java, with R for algorithms
  + Massive Lucene infrastructure with custom shard management
  + Distributed computing framework for analysis
  + Hadoop!
* **Algorithms from co-founder Gary King**
  + Bootstrapping big (feeding classified data back into classifier as part of training data)
  + Their conclusion:
    - Bigger data is important
    - Better data is important
    - Better algorithms are important
    - The sweet spot is when one leverages the other.
    - (bigger data can lead to better algorithms)

**Gary King: A Method for Computer Assisted Conceptualization (University of Kentucky, 4/20/2012)**

* Conceptualization through Classiﬁcation: “one of the most central and generic of all our conceptual exercises. . . . the foundation not only for conceptualization, language, and speech, but also for mathematics, statistics, and data analysis. . . . Without classiﬁcation, there could be no advanced conceptualization, reasoning, language, data analysis or, for that matter, social science research.” (Bailey, 1994).
  + Cluster Analysis: simultaneously (1) invents categories and (2) assigns documents to categories
* The problem with fully automated clustering
  + **The (Impossible) Goal:** optimal, fully automated, application-independent cluster analysis
  + **No free lunch theorem**: every possible clustering method performs equally well on average over all possible substantive applications
  + **Existing methods**:
    - **Many choices**: model-based, subspace, spectral, grid-based, graph-based, fuzzy k-modes, affinity propagation, self-organizing maps…
    - **Well defined** statistical, data analytic, or machine learning foundations
    - How to add substantive knowledge: With few exceptions, **unclear**
    - The literature: **little guidance on when methods apply**
    - Deriving such guidance: **difficult or impossible**
  + **Deep problem:** full automation requires more information
  + **No surprise:** everyone’s tried cluster analysis; very few are satisfied.
* **Switch from fully automated to computer assisted**
  + **Fully automated Clustering** may succeed sometimes, but fails in general: too hard to understand when each model applies
  + An alternative: **Computer assisted clustering** 
    - **Easy in theory:** List all clusterings; choose the best
    - **Impossible in practice:** Too hard for us mere humans!
    - **An organized list** will make the search possible
    - **Insight:** Many clusterings are perceptually identical
    - **E.**g.,: consider two clusterings that differ only because one document (of 10,000) moves from category 5 to 6
  + **Question:** How to organize clusterings so humans can understand?

**ReadMe: Software for Automated Content Analysis**

* Developed Daniel Gopkins, Gary King, Matthew Knowles, Steven Melendez