Data Mining: Successes and Failures

Christos Faloutsos

CMU

Failures

- non-technical: only ‘Public Relations’
- we are too modest to ‘brag’:
  - SAS, SPSS, +: (SAS on TV)
  - [Heckerman, KDD ‘04]: better cancer data analysis
- PLUS: companies often keep successes silent, to maintain edge
  - colleagues at search engines achieve $Ms in revenue increases

Successes

- merging of DB, ML, Stat
- excellent outreach:
  - bio-informatics
  - social networks
  - text / IR
  - game theory / economics
  - etc etc

Next steps, IMHO

- keep on the out-reach
- Large scale data mining (Tera and Peta bytes)
  - simple algorithms may give stunning results, when applied on massive data
  - scalability [in this KDD: Usama; Jon; ++]
  - parallelism

Scalability

- Google: > 450,000 processors in clusters of ~2000 processors each

Barroso, Dean, Hölzle, “Web Search for a Planet: The Google Cluster Architecture” IEEE Micro 2003

- target: hundreds of Tb, to several Peta-bytes
- (Netflix sample:2Gb uncompressed)
- Yahoo: ~5Pb [Usama’s keynote]

E.g.: self-* system @ CMU

- >200 nodes
- 40 racks of computing equipment
- 774kw of power.
- target: 1 PetaByte
- goal: self-correcting, self-securing, self-monitoring, self-...

DM for Tera- and Peta-bytes

Two-way street:

<- DM can use such infrastructures to find patterns

-> DM can help such infrastructures become self-healing, self-adjusting, ‘self-*’

Conclusion

- **Failures**: lack of ‘bragging’ ☺
- **Successes**: stunning out-reach + cross-disciplinarity
- **Next steps**: scalability: emphasis on Systems <–> DM collaboration