

# The Data Challenge

Panel at IEEE Hong Kong Poly U  
DM Forum

May 28, 2008

# Schedule

- [Keith C.C. Chan](#), The Hong Kong Polytechnic University
- [Nick Cercone](#), York University, Canada
- [Ramamohanarao Kotagiri](#), University of Melbourne, Australia
- [Qing Li](#), The City University of Hong Kong
- [Gregory Piatetsky-Shapiro](#), KDnuggets, USA
- [Michele Sebag](#), Universit Paris-Sud, France
- [Dacheng Tao](#), The Hong Kong Polytechnic University
- [Philip S. Yu](#), University of Illinois at Chicago, USA
- Panel Chair: Qiang Yang, HKUST
- 5 minutes each
- Q/A from the floor

# Central Question

- What are the main challenges and opportunities that come from the **data aspect** of data mining research?
- issues:
  - What kind of data
    - should we be looking for to help move data mining forward?
  - Simulated data or real data,
    - what are the pros and cons?
  - What can we learn from *other* sciences
    - in their use of data?

# Another way to ask

- Too much water, too little to drink!
  - In Clearwater Bay...
  
- Too much data, too little to be of use!

# Over 10 years of ACM KDDCUP

- ACM KDD Conferences
- IEEE ICDM Conferences
- SIAM DM and other Conferences

## A Gap Exists

- Mismatch between DM Conf. and KDD CUP topics?
  - Many of the DM topics do not show up on KDD-CUP
  - Typically, in KDDCUP
    - Simpler models win
    - Ensembles win
  - Typically in research
    - Complex algorithms more accepted
- KDD-CUP 2007, Netflix Data Challenge: Social Networks
  - [KDD-Cup 2006](#), Pulmonary embolisms detection from image data
  - [KDD-Cup 2005](#), Internet user search query categorization
  - [KDD-Cup 2004](#), Particle physics; plus Protein homology prediction
  - [KDD-Cup 2003](#), Network mining and usage log analysis
  - [KDD-Cup 2002](#), BioMed document; plus Gene role classification
  - [KDD-Cup 2001](#), Molecular bioactivity; plus Protein locale prediction.
  - [KDD-Cup 2000](#), Online retailer website clickstream analysis
  - [KDD-Cup 1999](#), Computer network intrusion detection
  - [KDD-Cup 1998](#), Direct marketing for profit optimization
  - [KDD-Cup 1997](#), Direct marketing for lift curve optimization