Big data analytics
People, infrastructure, provenance

Electronics and Computer Science
Web and Internet Science (WAIS)

Prof Luc Moreau
l.moreau@ecs.soton.ac.uk
Big data analytics

- Multimedia Analytics
- People and machines
  - Social media
  - Crowd sourcing
  - Human agent collectives
- Infrastructure work
- Provenance
Unstructured Data Analytics

• Focus on Multimedia data (images, video, audio, text, metadata, etc)
  – Necessarily “big” by their nature

• Looking at:
  – Event mining (esp. from streams of media)
  – Representation learning
    • Particularly in the context of aerial photoreconnaissance (with Ordnance Survey)
  – Scalable approaches
    • Distributed & GPGPU


Unsupervised segmentation of an aerial photo using a learned feature representation
Information Cascades

• Turning flat data streams into networks preserving the temporal order and showing patterns of information co-occurrence
• Varying configurations of the matching functions allow to derive different structures from the same source data stream
• The Transcendental Information Cascades method has been applied to:
  – social media data: capture collective action
  – urban traffic data: mobility management and resilience
  – EEG data: longitudinal sampling to find spatial-temporal relationships in brain signals

TREC Crowdsourcing Challenge

- Task: Label documents relevant to a complex query (15K documents)
- Combining Bayesian classifiers with crowdsourcing → minimize crowdsourcing costs

User engagement with datasets and analytics

- trends across social media, Wikipedia and other web resources (Southampton WO)
- identify and respond to natural disasters combining social media and IoT data sources (Korean WO)
- improving government by measuring how the elderly feel about the government services available to them (Adelaide WO)

http://online.liebertpub.com/doi/pdfplus/10.1089/big.2014.0035
Provenance

World Wide Web Consortium:

Provenance is a record that describes the people, institutions, entities, and activities, involved in producing, influencing, or delivering a piece of data or a thing in the world.

Understanding Provenance at Scale

• Provenance Network Metrics:
  – summary of topological structure of provenance graphs
  – network metrics that are specific to provenance graphs

• Predictive Models:
  – network metrics inputs to construct predictive models
  – to gain useful knowledge about the data described by provenance

• Summarisation:
  – extracts outliers
  – finds common pattern

• Radical approach:
  – not relying on any knowledge about the application,
  – except ground truth, labeling of data on a training set
Public ledgers & Provenance

• Block chain technology offers unforgeable public ledger
• Combine private/public provenance with public ledgers to make provenance trustable
• Doesn’t have to be on Bitcoin’s blockchain, but could be hosted on trusted host.