The Rensselaer IDEA
The data “forest”
Unfortunately

This leads to “brittle” data applications

• Complicated to develop
• Expensive to maintain
• Hard to validate
• Difficult to archive
• Impossible to share

As well as duplication of effort between apps and a lack of common infrastructure between research fields
Data to the rescue…

This leads to “brittle” data...

- Collection
- Exchange
- Handling
- Distribution
- Implementation

As well as a high level of effort and infrastructure between...
The Rensselaer IDEA

HPC: Computational Science and Engineering
+ Data Science and Predictive Analytics
+ Cognitive Computing
+ Perceptualization

DATA
Why Rensselaer? Why now?

Healthcare Analytics  Business Systems  Built and Natural Environments  Virtual and Augmented Reality  Cyber-Security  Network Science  4th Paradigm Physical/Life Sciences  Policy/Ethics/Open Govt

TWC  CSE  DSRC  CCNI  EMPAC  CDS  CBIS  Watson  CCC  NeST  CFES  ERC/LRC  TWC  CSE  DSRC  CCNI  EMPAC  CDS  CBIS  Watson  CCC  NeST  CFES  ERC/LRC  TWC  CSE  DSRC  CCNI  EMPAC  CDS  CBIS  Watson  CCC  NeST  CFES  ERC/LRC
Key Applications Highlight RPI Strengths

- Data-driven Medical and Healthcare Applications
- Business Analytics
- Built and Natural Environments/Smart cities
- Agents in Virtual and Augmented Reality
- Cybersecurity
- Network Science
- Data-Driven (4th Paradigm) Basic Science
- Policy, Ethics and Open Data
- …
Applications: Healthcare Analytics

- Extensible Mashups via Linked Data
  - Diverse datasets from NIH
  - Potentially linking to “unemployment rate”
- Accountable Mashups via Provenance
  - Annotate datasets used in demos
  - Feedback users’ comment to gov contact (e.g. %)

Deborah McGuinness, John Erickson & students
Applications: Built and Natural Environments

Why Park and Ride (P&R)?
- P&R systems enable potential users to develop to a P&R facility where they could use a transit service to their destinations.
- P&R systems provide a good alternative to minimize high levels of congestion in cities such as New York, Stockholm, and London.

P&R is a way to provide suburban commuters an attractive transit alternative.

Abstract
The main objective of this paper is to develop a methodology to assess where P&R facilities should be located to maximize their accessibility to New York city. To do so, it is necessary to reach a set of analytical formulations that provide insight into the factors that would contribute to the success of P&R and the implications of location decisions.

The novel approach developed by the team is based on the explicit considerations of the constraints under which a mixed use would find it beneficial to use P&R and translates these findings into a set of policy recommendations. The project builds on these results by considering the role that P&R location has on the determination of the potential demand for the system.

Project tasks
- Develop a comprehensive literature review in P&R design, location, operation, and plan.
- Suggest best practices for location analyses.
- Provide guidelines on how to design, and plan P&R facilities.
- Scan of potential sites in New York Metropolitan Transportation Council area.
- Propose a methodology to evaluate P&R facilities.
- Apply the evaluation methodology and give a set of recommendations for specific P&R.

Mathematical analysis of market potential in a corridor
The objective of this simplified analysis is to obtain insight in the location of P&R location. The P&R can be located before, at, or after the beginning of the congested area.


determination

Title:

This is a schematic corridor.

Analytical results

Insight from corridor analysis
- Minimum wait and walk times are needed.
- The lower the scheduled delays and walk time, the more flexible the PR facility would be as they would lead to a lower value of the right hand side. This implies that P&R facilities located near transit hubs with high frequency service would work better than those located near poorly served areas.

- Fast transit system
- A P&R facility would work best when it replaces a slow auto trip with a faster and shorter auto trip, plus a faster transit trip. These conditions happen when the Park facility is located at the middle of a severely congested, where it connects to an efficient transit service.

- As a consequence of the previous item, the result implies that it is hard to make an economic case for use of the P&R facility if the transit service is not different than auto service.

- Incentives
- The lower the out of pocket expenses of PR and, in comparison with the auto-only alternative, the more attractive it becomes. As a result, an offer must be made to ensure that using PR facilities are as inexpensive as possible.

- Long distance in fast transit
- The longer the auto trip replaced by the transit trip, the better as this would increase the savings in distance traveled costs.

- Work in progress
- The team is reviewing the NYMC data in order to evaluate the proposed methodology.
- Next step is to test computer programs to identify PR pairs that benefit from P&R, in terms of both, generated cost and travel time.
- Sensitivity analysis of model clinics to compare cost-benefit for scenario.

Conclusions and further research
This paper developed analytical formulations to gain insight into the optimal location, i.e., the one that maximizes the potential market, P&R facilities. The analysis is based on a fundamental principle of rationality, i.e., that a traveler would only consider using a P&R facility if the corresponding generalized cost is lower than the auto only alternative, which it takes an acquires necessary economic condition.

This implies that the P&R system must provide a service that is fast and frequent enough to overcome the travel costs (waiting in the P&R, and waiting for the transit vehicle). As a result, it is very difficult to make an economic case for P&R when the transit system does not travel faster than the general traffic. Under the conditions, the maximum condition is not met.

Jose Holguin-Vera et. al., '13
Applications: Network Science/ Cyber Security

Boleslaw Szymanski, Qingpeng Zhang, Jim Hendler, Deborah McGuinness and Students
Data Infrastructure for Science

Integrated Applications

- Discovery visualizations
- Semantic interoperability
- Analytics and mining
- Global Census, Virtual Mineral Laboratory, ...

Software, Tools & Apps

- Deep Energy/Life Applications
- Semantic interoperability
- Physics/Chemistry Models
- Res/Flux Applications
- Semantic query, hypothesis and inference
- Query, access and use of data

Data Repositories

- GVP
- MINDAT
- EOS
- EarthChem
- Emission/Compositions

Application-level mediation: vocabulary, mapping to science and data terms

Semantic mediation: physics, chemistry, mineral, emission data - ChemML,

Metadata, schema, data ...

Peter Fox, John Erickson & Students
The trunk: Data Technologies

High Performance Modeling and Simulation
  • Computation Center for Nanotechnology Innovation

Data Science
  • Data Science Research Center

Cognitive Computing
  • RPI and IBM Watson Partnership

Perceptualization
  • Center for Cognition, Communication and Culture
HPC: Computational Science and Engineering

SoS and SoE Data Visualizations, 2009-2013
Data Science at RPI: Analytics, Agents and HPC (DSRC)

**Tensor-Matrix Algorithms for Big Data**

- Regression on PetaByte (PB)-scale inputs by sampling and randomization.
- Implementations of these methods in cloud-like environments and/or massively parallel cyberinstruments to analyze PB-scale data in the near future.

**Data Fusion Algorithms**

- Joint analysis of data from multiple sources may capture hidden groups that would not be captured by the individual analysis of each data set.
- If a matrix or a higher-order tensor has a low-rank structure, it is possible to recover the missing entries.
- However, if there is a large amount of missing data, then analysis of a single data set is no longer enough for accurate data recovery. Data fusion can help to recover missing entries.

**Networks of Selfish Agents**

- Networks designed by or consisting of many self-interested agents.
- Understand the behavior and structural properties of such networks.
- Design Algorithms and Protocols for influencing these networks.
- Increasing the spread of diffusive processes (e.g., warnings, advertising).
- Inhibiting the spread of diffusive processes (gossip, viruses, etc).
- Improving global network properties through targeted incentives.

**Graph Mining Algorithms**

- Sampling Interesting Subgraph Patterns from Complex Graphs.
- Approximate and Exact Pattern Mining.
- Labeled and attributed graphs.
- Integrated mining of diverse and heterogeneous data via graph modeling (see figure in the context of integrated “omics” mining in Bioinformatics).
- Graph-based classification and clustering.
- Large-scale and parallel methods.
- Applications in social and biological networks.

**Networks of Selfish Agents**

- Networks designed by or consisting of many self-interested agents.
- Understand the behavior and structural properties of such networks.
- Design Algorithms and Protocols for influencing these networks.
- Increasing the spread of diffusive processes (e.g. warnings, advertising).
- Inhibiting the spread of diffusive processes (gossip, viruses, etc).
- Improving global network properties through targeted incentives.

**Statistical Analysis of Massive Networked Data**

- Algorithms to discover hidden groups in massive, networked, dynamic and noisy data.
- Identify statistical patterns indicative of hidden groups and evolution of such patterns, including growth, merging and dissolution of hidden groups.
- Identify internal command hierarchies of hidden groups.
- Algorithms scalable to multi-billion node networks and based on interactions and not semantics.
- Robust to missing nodes and links.
- Intuitive user interface for interaction with human analyst.
- Theorems governing accuracy of algorithms.

Bulent Yener, Elliot Anshelevich, and Petros Drineas, Malik Magdon-Ismail, Mohammed Zaki, & Students
Cognitive Computing: Watson and Beyond

Memory

Reasoning

Decision Making

Watson, Cogito, and Clarion
Beyond Visualization: Perceptualization
Geekopedia: Data exploration helps a data consumer focus an information search on the pertinent aspect of relevant data before true analysis can be achieved. In large data sets, data is not gathered or controlled in a focused manner. Even in smaller data sets, it is also true that data gathered are not in a very rigid and specific technique can result in a disorganized manner and a myriad of subsets each...
Make Data Available to a Wider Audience

**Rensselaer Data Services**

Rensselaer Data Services provides:

- a pilot data repository for management and dissemination of research data generated at Rensselaer and/or in conjunction with Rensselaer collaborators, along with a dataset identifier service
- information for Rensselaer researchers on management of research data, including citation guidelines

Want to join the conversation about research data at Rensselaer? Interested in making your data available through data.rpi.edu? Have questions? Contact us.

Rensselaer Data Services is currently supported by Rensselaer’s Tetherless World Constellation.
In short, the Rensselaer IDEA is

Data Technologies that allow researchers to

• Discover, Develop or Derive Data from multiple sources in multiple ways
• Process the Data through multiple pathways allowing for new analytic insights
• Provide new cognitively-informed means for interaction with Unstructured Data
• Validate Data against models at large scale
• Tools to Archive, Explain, Visualize, Repurpose, Store, Cite, Share, Catalog, Stream, Scrape and Supply Data for application developers who want to change the world
Any questions?