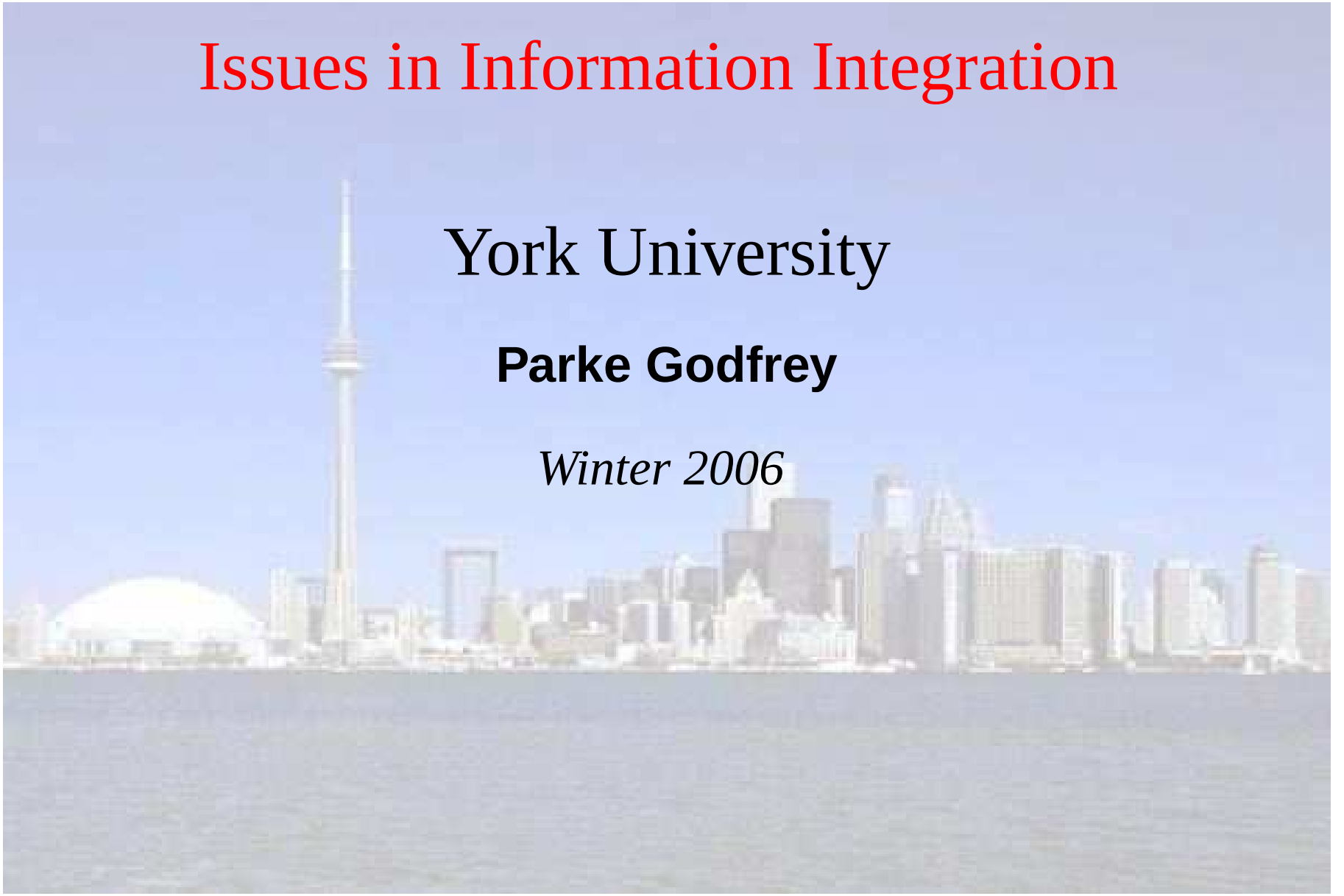
COSC-6490-B

Issues in Information Integration



York University

Parke Godfrey

Winter 2006

What is Information Integration?

And why do we need it?

"The center cannot hold." Yeats

Information is distributed over many sources / databases.

- Cannot have *everything* in one database under one schema.
- Sometimes the information we need must be composed from several sources.
- Sometimes we do not know *where* the information that we need resides.

The Katrina Disaster

- How to coördinate the different agencies? E.g., supplies.
- How to locate the missing people?

Schema Exercise:
For each located person, record his or her name, date-of-birth, some form of identification, and (previously) permanent address. Record contact information (e.g., phone numbers) for the located people. Record for each closest relatives. Record which shelter the person is residing in. (He or she may have moved from one shelter to another. Remember previous shelters and durations.)

Many Different Databases

all about the same thing

"You are in a maze of twisty little passages, all alike."

- **incompleteness**. The data state of each DB is incomplete. Does not contain information about *each* missing person.
- **schema heterogeneity**. The schemas of the DBs may vary, *even though* each is about the same "topic" (located people).
- **system heterogeneity**. The data models all may be the same ("object-relational"), but the systems and platforms differ (e.g., IBM DB2, Microsoft SQL, & Oracle).
- **model heterogeneity**. The data models may differ. E.g., one is a relational database, one an XML database, one a spreadsheet, and another a textfile.

Many Different Databases

all about the same thing

"You are in a maze of twisty little passages, all alike."

- **global inconsistency.** Each database alone is *locally* consistent, but taken together they are *globally* inconsistent.
  - E.g.,
    - One DB says that *Parke Godfrey* is in a temporary shelter in Baton Rouge. Another says he is in Atlanta.
    - The identity problem: Are the two *Parke Godfrey*'s the same person?
  - Is there any way to make the global view consistent?
  - Can we modify the query answering procedure to produce *consistent* answers even though the global view is *inconsistent*?

What is Information?

versus, say, data?

It is all about *schemas* and *queries*.

We need good tools for working with, and reasoning about, *schemas* and *queries*.

- When are two schemas the same?
- When are two queries logically the same?
- How can schemas be unified?
- How can a query be parsed into sub-queries for multiple sources?

Many Different Databases

each about something different (but overlapping)

"You are in a maze of little twisting passages, each different."

This is the common predicament of most companies / organizations.

Big issues:

- schema integration
- schema mapping
- optimization & caching