

ILLINOIS INSTITUTE OF TECHNOLOGY

CS520



Data Integration, Warehousing, and Provenance

Course Info

IIT DBGroup

Boris Glavic

<http://www.cs.iit.edu/~glavic/>
<http://www.cs.iit.edu/~glavic/cs520/>
<http://www.cs.iit.edu/~dbgroup/>





0

ILLINOIS INSTITUTE OF TECHNOLOGY

Outline

- 0) Course Info
- 1) Introduction
- 2) Data Preparation and Cleaning
- 3) Data Translation: Schema mappings, Virtual Data Integration, and Data Exchange
- 4) Data Warehousing
- 5) Big Data Analytics
- 6) Data Provenance



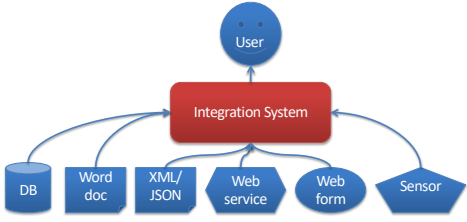

1

1

ILLINOIS INSTITUTE OF TECHNOLOGY

What is information integration?

- Combination of data and content from multiple sources into a common format
 - Completeness
 - Correctness
 - Efficiency


2

2

ILLINOIS INSTITUTE OF TECHNOLOGY

Why Information Integration?

- Data is already available, right?
- ..., but
- **Heterogeneity**
 - Structural
 - Data model (relational, XML, unstructured)
 - Schema (if exists)
 - Semantic
 - Naming and identity conflicts
 - Data conflicts
 - Syntactic
 - Interfaces (web form, query language, binary file)




3

3

ILLINOIS INSTITUTE OF TECHNOLOGY

Why Information Integration?

- **Autonomy**
 - Sources may not give you unlimited access
 - Web form only support a fixed format of queries
 - Does not allow access to unlimited amounts of data
 - Source may not be available all the time
 - Data, schema, and interfaces of sources may change
 - Potentially without notice




4

4

ILLINOIS INSTITUTE OF TECHNOLOGY

“Real World” Examples?

- **Portal websites**
 - Flight websites (e.g., Expedia) gather data from multiple airlines, hotels
- **Google News**
 - Integrates information from a large number of news sources
- **Science**
 - Biomedical data sources
- **Business**
 - Warehouses: integrate transactional data



5

5


Example Integration Problem [1]

ILLINOIS INSTITUTE OF TECHNOLOGY

- Integrate stock ticker data from two web services A and B
- Service A: Web form (Company name, year)
- Service B: Web form (year)

Steps

- Interfaces
- Schema integration
- Translate queries
- Optimization
- Send queries to sources
- Gather query results
- Entity resolution
- Fusion
- Curation / Cleaning
- Return final results



6

Example Integration Problem [2]

ILLINOIS INSTITUTE OF TECHNOLOGY


- Service A:


```
<Stock>
  <Company>IBM</Company>
  <DollarValue>155.8</DollarValue>
  <Month>12</Month>
</Stock>
```
- Service B:


```
<Stock>
  <Company>International Business Machines</Company>
  <Date>2014-08-01</Date>
  <Value>106.8</Value>
  <Currency>Euro</Currency>
</Stock>
```

Steps

- Interfaces
- Schema integration
- Translate queries
- Optimization
- Send queries to sources
- Gather query results
- Entity resolution
- Fusion
- Curation / Cleaning
- Return final results



7

Example Integration Problem [2]

ILLINOIS INSTITUTE OF TECHNOLOGY


- Service A:


```
<Stock>
  <Company>
  <DollarValue>
  <Month>
</Stock>
```
- Service B:


```
<Stock>
  <Company>
  <Date>
  <Value>
  <Currency>
</Stock>
```

Steps

- Interfaces
- Schema integration
- Translate queries
- Optimization
- Send queries to sources
- Gather query results
- Entity resolution
- Fusion
- Curation / Cleaning
- Return final results



8

Example Integration Problem [2]


ILLINOIS INSTITUTE OF TECHNOLOGY

Global Schema

```
<Stock>
  <Company>
  <DollarValue>
  <Month>
  <Value>
  <Year>
</Stock>
```

Steps

- Interfaces
- Schema integration
- Translate queries
- Optimization
- Send queries to sources
- Gather query results
- Entity resolution
- Fusion
- Curation / Cleaning
- Return final results



9

Example Integration Problem [3]

ILLINOIS INSTITUTE OF TECHNOLOGY


- SQL interface for integrated service

```
SELECT month, value
FROM ticker
WHERE year = 2014
      AND cmp = 'IBM'
```

- Service A: (IBM, 2014)
- Service B: (2014)

Steps

- Interfaces
- Schema integration
- Translate queries
- Optimization
- Send queries to sources
- Gather query results
- Entity resolution
- Fusion
- Curation / Cleaning
- Return final results



10


Example Integration Problem [4]

ILLINOIS INSTITUTE OF TECHNOLOGY

- For web service A we can either
 - Get stocks for IBM in all years
 - Get stocks for all companies in 2014
 - Get stocks for IBM in 2014
- Trade-off between amount of processing that we have to do locally, amount of data that is shipped, ...

Steps

- Interfaces
- Schema integration
- Translate queries
- Optimization
- Send queries to sources
- Gather query results
- Entity resolution
- Fusion
- Curation / Cleaning
- Return final results



11


Example Integration Problem [5]

ILLINOIS INSTITUTE OF TECHNOLOGY

- **Service A:** (IBM, 2014)
- **Service B:** (2014)

Steps

- 1) Interfaces
- 2) Schema integration
- 3) Translate queries
- 4) Optimization
- 5) Send queries to sources
- 6) Gather query results
- 7) Entity resolution
- 8) Fusion
- 9) Curation / Cleaning
- 10) Return final results



12 CSS20 - Course Info

12

Example Integration Problem [6]

ILLINOIS INSTITUTE OF TECHNOLOGY

- **Service A:**

```
<Stock>
  <Company>IBM</Company>
  <DollarValue>155.8</DollarValue>
  <Month>12</Month>
...


```
- **Service B:**

```
<Stock>
  <Company>International Business
  Machines</Company>
  <Date>2014-12-01</Date>
  <Value>106.8</Value>
  <Currency>Euro</Currency>
...

```

Steps

- 1) Interfaces
- 2) Schema integration
- 3) Translate queries
- 4) Optimization
- 5) Send queries to sources
- 6) Gather query results
- 7) Entity resolution
- 8) Fusion
- 9) Curation / Cleaning
- 10) Return final results



13 CSS20 - Course Info

13


Example Integration Problem [7]

ILLINOIS INSTITUTE OF TECHNOLOGY

- **IBM vs. Integrated Business Machines**

Steps

- 1) Interfaces
- 2) Schema integration
- 3) Translate queries
- 4) Optimization
- 5) Send queries to sources
- 6) Gather query results
- 7) Entity resolution
- 8) Fusion
- 9) Curation / Cleaning
- 10) Return final results



14 CSS20 - Course Info

14


Example Integration Problem [8]

ILLINOIS INSTITUTE OF TECHNOLOGY

- Granularity of time attribute
 - Month vs. day
- What if both services return different values (after adapting granularity)
 - Average?
 - Median?
 - Trust-based?

Steps

- 1) Interfaces
- 2) Schema integration
- 3) Translate queries
- 4) Optimization
- 5) Send queries to sources
- 6) Gather query results
- 7) Entity resolution
- 8) Fusion
- 9) Curation / Cleaning
- 10) Return final results



15 CSS20 - Course Info

15

Example Integration Problem [9]

ILLINOIS INSTITUTE OF TECHNOLOGY


- **“Dirty Data”**
 - **Outliers**
 - E.g., \$10M / unit not realistic
 - **Violations of constraints**
 - E.g., stock value has to be positive
 - **Format and type errors**
 - E.g., include \$ in value or not
 - Value has to be a number
- **Service A:**

```
<DollarValue>-15</DollarValue>
<DollarValue>1000000.8</DollarValue>
<DollarValue>$24</DollarValue>
<DollarValue>five dollar</DollarValue>
<DollarValue>fad23e19hasd</DollarValue>
...

```

Steps

- 1) Interfaces
- 2) Schema integration
- 3) Translate queries
- 4) Optimization
- 5) Send queries to sources
- 6) Gather query results
- 7) Entity resolution
- 8) Fusion
- 9) Curation / Cleaning
- 10) Return final results



16 CSS20 - Course Info

16

Example Integration Problem [10]

ILLINOIS INSTITUTE OF TECHNOLOGY


- **Return final results:**

```
<Stock>
  <Month>01</Month>
  <Value>105</Value>
</Stock>
...
<Stock>
  <Month>12</Month>
  <Value>107</Value>
</Stock>

```

Steps

- 1) Interfaces
- 2) Schema integration
- 3) Translate queries
- 4) Optimization
- 5) Send queries to sources
- 6) Gather query results
- 7) Entity resolution
- 8) Fusion
- 9) Curation / Cleaning
- 10) Return final results



17 CSS20 - Course Info

17

Why hard?

ILLINOIS INSTITUTE
OF TECHNOLOGY

- System challenges
 - Different platforms (OS/Software)
 - Efficient query processing over multiple heterogeneous systems
- Social challenges
 - Find relevant data
 - Convince people to share their data
- Heterogeneity of data and schemas
 - A problem that even exists if we use same system



18

CSS20 - Course Info

18

Why hard? Cont.

ILLINOIS INSTITUTE
OF TECHNOLOGY

- Often called **AI-complete**
 - Meaning: “It requires human intelligence to solve the problem”
 - Unlikely that general completely automated solutions will exist
- So why do you still sit here
 - There exist automated solutions for relevant less general problems
 - Semi-automated solutions can reduce user effort (and may be less error prone)



19

CSS20 - Course Info

19

AI completeness

ILLINOIS INSTITUTE
OF TECHNOLOGY

- Yes, but still why is this problem really so hard?
 - **Lack of information:** e.g., the attributes of a database schema have only names and data types, but no machine interpretable information on what type of information is stored in the attribute
 - **Undecidable computational problems:** e.g., to decide whether a user query can be answered from a set of sources that provide different views on the data requires **query containment** checks which are undecidable for certain query types



20

CSS20 - Course Info

20

Relevant less general problems

ILLINOIS INSTITUTE
OF TECHNOLOGY

- **Data Extraction**
 - Extract data from unstructured sources / text
- **Data cleaning:**
 - Clean dirty data before integration
 - Conformance with a set of constraints
 - Deal with missing and outlier values
- **Entity resolution**
 - Determine which objects from multiple dataset represent the same real world entity
- **Data fusion**
 - Merge (potentially conflicting) data for the same entity



21

CSS20 - Course Info

21

Relevant less general problems

ILLINOIS INSTITUTE
OF TECHNOLOGY

- **Schema matching**
 - Given two schemas determine which elements store the same type of information
- **Schema mapping**
 - Describe the relationships between schemas
 - Allows us to rewrite queries written against one schema into queries of another schema
 - Allows us to translate data from one schema into



22

CSS20 - Course Info

22

Relevant less general problems

ILLINOIS INSTITUTE
OF TECHNOLOGY

- **Virtual data integration**
 - Answer queries written against a **global mediated schema** by running queries over **local sources**
- **Data exchange**
 - Map data from one schema into another
- **Warehousing: Extract, Transform, Load**
 - Clean, transform, fuse data and load it into a data warehouse to make it available for analysis



23


CSS20 - Course Info

23

Relevant less general problems

ILLINOIS INSTITUTE OF TECHNOLOGY

- **Integration in Big Data Analytics**
 - Often “pay-as-you-go”:
 - No or limited schema
 - Engines support wide variety of data formats
- **Provenance**
 - Information about the origin and creation process of data
 - Very important for integrated data
 - E.g., “from which data source is this part of my query result”




24 CSS20 - Course Info

24

Webpage and Faculty

ILLINOIS INSTITUTE OF TECHNOLOGY

- **Course Info**
 - **Course Webpage:** <http://cs.iit.edu/~glavic/cs520>
 - **Discord:**
 - Used for announcements
 - Use it to discuss with me, TA, and fellow students
 - **Syllabus:** <http://www.cs.iit.edu/~glavic/cs520/2023-fall/syllabus/>
- **Faculty**
 - **Boris Glavic** (<http://cs.iit.edu/~glavic>)
 - **Email:** bglavic@iit.edu
 - **Phone:** 312.567.5205
 - **Office:** SB 206B




25 CSS20 - Course Info

25

Workload and Grading

ILLINOIS INSTITUTE OF TECHNOLOGY

- **Exams (60%)**
 - Final (30%), Midterm (30%)
- **Homework Assignments** (preparation for exams!)
 - **Theory part:** Practice theory for final exam
 - **Lab part:** Practice the tools we discuss in class
- **Literature Review (20%)**
 - In groups of 3 students
 - Topics will be announced soon
 - You have to read a research paper
 - Papers will be assigned in the first few weeks of the course
 - You will give a short presentation (15min) on the topic in class
 - You will write a report summarizing and criticizing the paper (up to 4 pages)




26 CSS20 - Course Info

26

Workload and Grading

ILLINOIS INSTITUTE OF TECHNOLOGY

- **Data Curation Project(20%)**
 - In groups of 3 students (same groups as for literature review)
 - You will acquire and curate (clean, integrate, ...) a real world dataset
 - This is open-ended, you can choose whatever tools you need, whatever domain you think is interesting, ...
 - Only limitation is that you need to document your cleaning workflow using a **Vizier notebook** (so at least some python is required)
 - <https://vizierdb.info/>
 - Steps:
 - Acquire or extract one or more real world datasets for a domain of choice
 - Gain an understanding of the data and identify data quality issues
 - Research tools that are suited for the data cleaning, integration, extraction tasks that you need to apply to create a correct and clean output dataset
 - Apply the tools and produce an output
 - Work will be submitted through git repositories on bitbucket.org that we will create for each group




27 CSS20 - Course Info

27

Workload and Grading

ILLINOIS INSTITUTE OF TECHNOLOGY

- **Timeline:**
 - See course webpage for detailed dates
 - You are required to meet with the TA/Prof. several times for discussing the progress for the literature review and data curation project
 - Literature reviews and project presentations will be done in a block seminar towards the end of the semester (1-2 days)




28 CSS20 - Course Info

28

Course Objectives

ILLINOIS INSTITUTE OF TECHNOLOGY

- Understand the problems that arise with querying heterogeneous and autonomous data sources
- Understand the differences and similarities between the data integration/exchange, data warehouse, and Big Data analytics approaches
- Be able to build parts of a small data integration pipeline by “glueing” existing systems with new code



29 CSS20 - Course Info

29

Course Objectives cont.

ILLINOIS INSTITUTE
OF TECHNOLOGY

- Have learned formal languages for expressing schema mappings
- Understand the difference between virtual and materialized integration (data integration vs. data exchange)
- Understand notions of data provenance and know how to compute provenance

30

CSS20 - Course Info



30

Fraud Policies

ILLINOIS INSTITUTE
OF TECHNOLOGY

- All work has to be original!
 - Cheating = 0 points for review/exam
 - Possibly E in course and further administrative sanctions
 - Every dishonesty will be reported to office of academic honesty
- Late policy:
 - -20% per day
 - You have to give your presentation to pass the course!
 - No exceptions!

31

CSS20 - Course Info



31

Fraud Policies cont.

ILLINOIS INSTITUTE
OF TECHNOLOGY

- Literature Review:
 - Every student has to contribute in the presentation, report, and data curation project!
 - **Don't let others freeload on you hard work!**
 - Inform me or TA immediately

32

CSS20 - Course Info



32

Reading and Prerequisites

ILLINOIS INSTITUTE
OF TECHNOLOGY

- **Textbook:** Doan, Halevy, and Ives.
 - **Principles of Data Integration**, 1st Edition
 - Morgan Kaufmann
 - Publication date: 2012
 - ISBN-13: 978-0124160446
 - Prerequisites:
 - CS 425



33

CSS20 - Course Info



33

Additional Reading

ILLINOIS INSTITUTE
OF TECHNOLOGY

- Papers assigned for literature review
- Optional: Standard database textbook

34

CSS20 - Course Info



34

Outline

ILLINOIS INSTITUTE
OF TECHNOLOGY

- 0) Course Info
- 1) Introduction
- 2) Data Preparation and Cleaning
- 3) Schema mappings and Virtual Data Integration
- 4) Data Exchange
- 5) Data Warehousing
- 6) Big Data Analytics
- 7) Data Provenance

35

CSS20 - Course Info



35