70-455
Modern Data Management

Wolfgang Gatterbauer

(gatt@cmu.edu)

Spring 2014

Why Data Management

In today's business world, you will have to use various forms of data to drive decisions.

www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html

HOME PAGE | TODAY'S PAPER | VIDEO | MOST POPULAR | TIMES TOPICS

The New York Times
SundayReview | The Opinion Pages

WORLD U.S. N.Y. / REGION BUSINESS TECHNOLOGY SCIENCE HEALTH SPORTS OPINION

NEWS ANALYSIS

The Age of Big Data

By STEVE LOHR
Published: February 11, 2012 | 82 Comments

GOOD with numbers? Fascinated by data? The sound you hear is opportunity knocking.

RECOMMEND
TWITTER
LINKEDIN
COMMENTS (82)
SIGN IN TO E-MAIL
PRINT
SINGLE PAGE
REPRINTS
SHARE

Enlarge This Image

Chad Hagen

Mo Zhou was snapped up by I.B.M. last summer, as a freshly minted Yale M.B.A., to join the technology company's fast-growing ranks of data consultants. They help businesses make sense of an explosion of data — Web traffic and social network comments, as well as software and sensors that monitor shipments, suppliers and customers — to guide decisions, trim costs and lift sales. "I've always had a love of numbers," says Ms. Zhou, whose job as a data analyst suits her skills.

Multimedia

Siera = 6.145−16.986×
−1.858×((GB-FB-PU)÷P
×(((GB-FB-PU)÷PA)^2)
÷PA)−5.195×(BB÷PA)×

Graphic

Play (Data-Driven) Ball!

To exploit the data flood, America will need many more like her. A report last year by the McKinsey Global Institute, the research arm of the consulting firm, projected that the United States needs 140,000 to 190,000 more workers with "deep analytical" expertise and 1.5 million more data-literate managers, whether retrained or hired.

The impact of data abundance extends well beyond

bits.blogs.nytimes.com/2012/10/24/big-data-in-more-hands/

The New York Times Technology | Personal Tech | Business Day

Bits

OCTOBER 24, 2012, 9:00 AM | 4 Comments

Big Data in More Hands

By QUENTIN HARDY

FACEBOOK
TWITTER
GOOGLE+
SAVE
E-MAIL
SHARE
PRINT

Business people, Big Data is coming for you.

Software that captures lots of data and uses it to make predictions has mostly been the province of engineers skilled in arcane databases and statisticians capable of developing complex algorithms. As the business gets bigger, however, software makers are domesticating their products in the hope they will prove attractive to a broader population.

Cloudera, which offers a popular version of the open source database called Hadoop, released software on Wednesday that makes it possible to run queries from a more mainstream SQL programming language interface. SQL, thanks to its adoption by Oracle, Microsoft and others, is known to millions of business analysts.

"This enables us to talk to a whole other class of customer," said Mike Olson, the chief executive of Cloudera. "The knock against Hadoop was that it is too complex."

There is a reason for that. Hadoop is one of several so-called unstructured databases that were created at Yahoo and Google, after those two companies found they had previously unimaginable amounts of data about activities like people's Web-surfing habits. Put into databases designed to handle this unstructured behavior, then analyzed, this information was valuable for figuring out things like what advertisement to put in front of each individual Web surfer.

Three Integrated Parts of "Data Management"

Organize Data → **Analyze Data** → **Synthesize Data**

- Data & Self-Org.
- E/R diagrams
- Database admin

- Excel
- SQL

- Structured & Visual Communication

Five Corresponding Class Modules

Organize Data → **Analyze Data** → **Synthesize Data**

- Data & Self-Org.
- E/R diagrams
- Database admin

- Excel
- SQL

- Structured & Visual Communication

Class Modules

1. Use of Excel
2. "Organization & Synth."
3. Use of SQL
4. Data modeling
5. Database admin

Module 1: Excel

Basic Knowledge of Excel is assumed! We cover advanced functions such as:

- Pivot tables
- Excel as database
- Lookup Tables
- Array Fomulas
- Advanced formulas
- Basic VBA

Example Array Formulas

	$/piece	quantity	
Item 1	10	1	10
Item 2	20	3	60
Item 3	30	2	60
Item 4	40	4	160
		Total $	**290**

How to create the weighted sum without the intermedate results?

	$/piece	quantity
Item 1	10	1
Item 2	20	3
Item 3	30	2
Item 4	40	4
Total $		**290**

Module 2: "Organization & Synth."

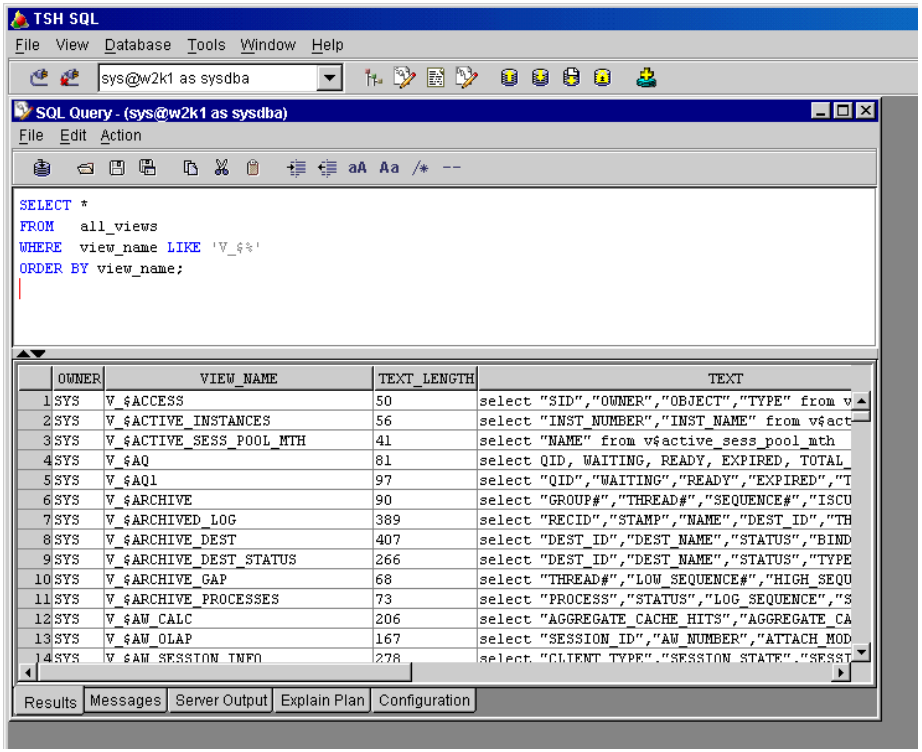
- How do you organize yourself and your data?
- How do you synthesize your results into a concise recommendation?
- What is an appropriate data-driven chart?

Profession	*Frequency of recent citations*	*1996 total employed (1,000)*	*Relative frequency*
Lawyers	8101	880	9.2
Economists	1201	148	8.1
Architects	1097	160	6.9
Physicians	3989	667	6.0
Statisticians	34	14	2.4
Psychologists	479	245	2.0
Dentists	165	137	1.2
Teachers (not university)	3938	4724	0.8
Engineers	934	1960	0.5
Accountants	628	1538	0.4
Computer programmers	91	561	0.2
Total	20,657	11,034	1.9

We have 3 numbers per item. How can we represent all of them in a 2-dimensional plane?

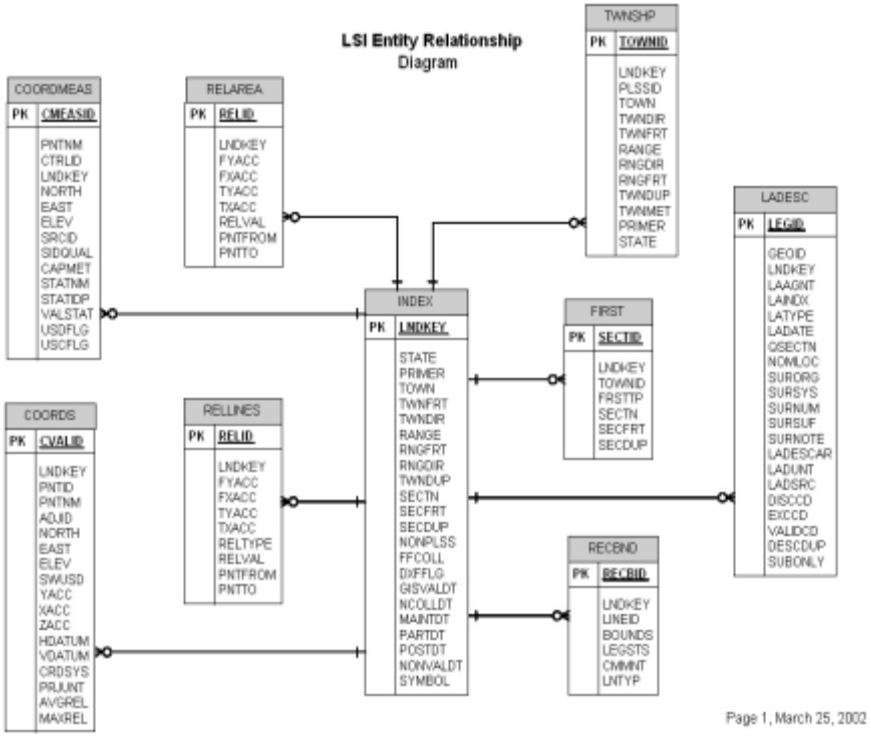
Module 3: SQL

- What is SQL? What can it do that Excel cannot?
- How do I put data into my database? How do I update what is already there?
- How do I query my database to get the information that I am seeking?



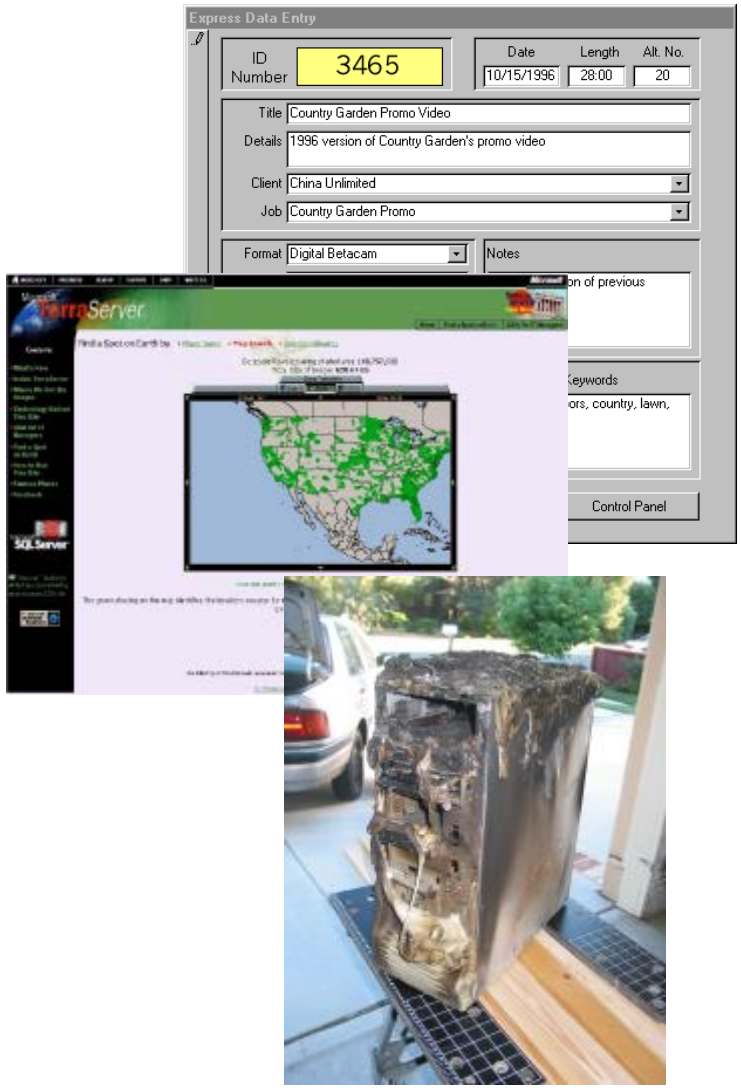
Module 4: Data Modeling / Designing Databases

- What is *Data Modeling* and why is it useful?
- How do I create a data model? How can I tell if my data model is "good"?
- What is the *relational data model* and why is it useful?
- What is *normalization*? Why should I care?
- How do I create a database based on my data?

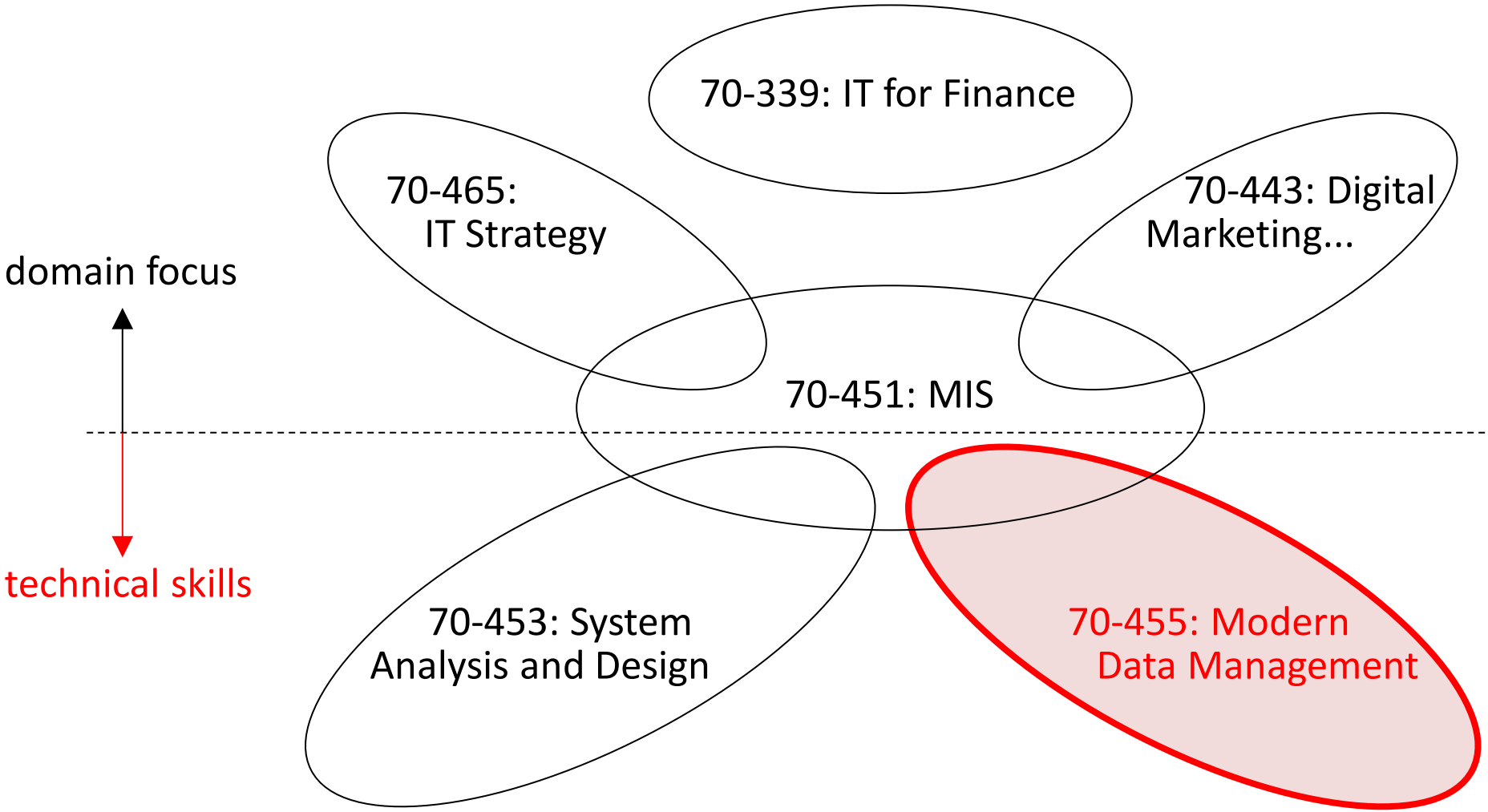


Module 5: Managing Database Systems

- How can I tune and scale my databases?
- What are *transactions*?
- When is it better to *denormalize*?
- How can I distribute a database across multiple machines or locations?
- What is the fuss about *NOsql databases*?



How does this class fit with other Tech classes



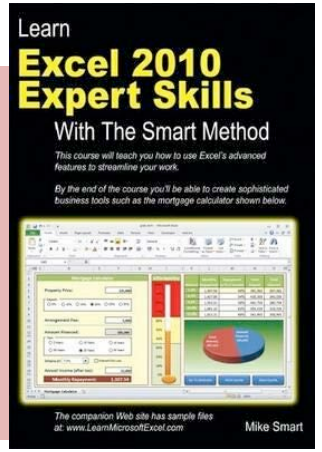
Administrative

- Most learning will happen in-class with hands-on exercises and student presentations throughout the course. Bring your laptops with a Windows partition!
- Prerequisites: Basic Excel and programming
- Workload outside of class: ~10 hrs/week
  - Flipped classroom: small preparations before class
  - 5-6 smaller assignments
  - one project in groups of 2-3
  - In-class Midterm and Final exams
- Syllabus: to be posted on Blackboard
- Class: Tue & Thu 1:30-2:50 pm (location TBD)

Excel books

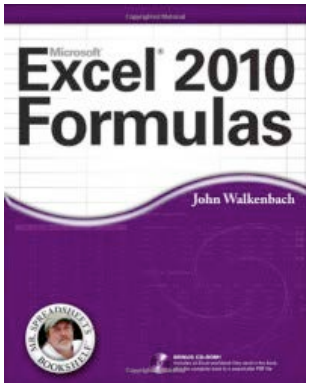
Required

- The Excel 2013 version (will appear ~Nov 2013) of: "Learn Excel 2010 Expert Skills with The Smart Method" by Mike Smart, 2011. (~$20) www.amazon.com/Learn-Excel-Expert-Skills-Method/dp/0955459982



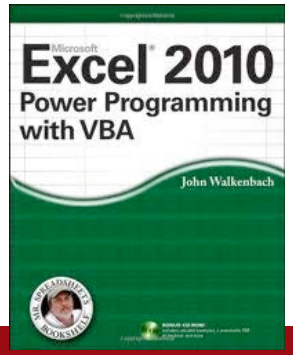
Other relevant books (not required)

- "Excel 2010 Formulas" by John Walkenbach, 2010 (~$30)



- "Excel 2010 Advanced" by Stephen Moffat, free at bookboon: http://bookboon.com/en/textbooks/it-programming/excel-2010-advanced

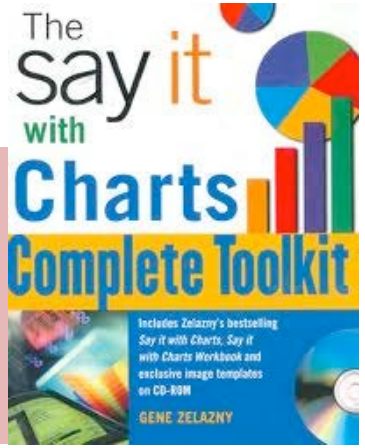
- "Excel 2010 Power Programming with VBA" by John Walkenbach (~$30)



"Synthesis" books

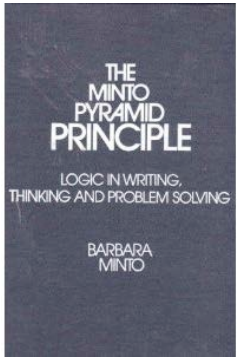
Not Required but highly recommended:

- "The Say it with Charts Complete Toolkit" by Gene Zelazny, 2006 (~$35)
  www.amazon.com/Say-Charts-Complete-Toolkit/dp/0071474706/

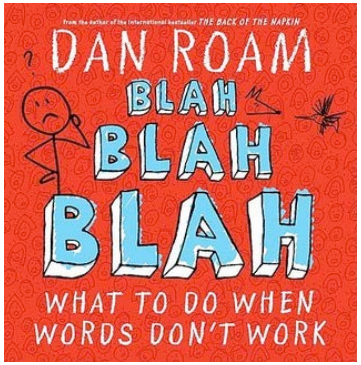


Other recommended books (not required):

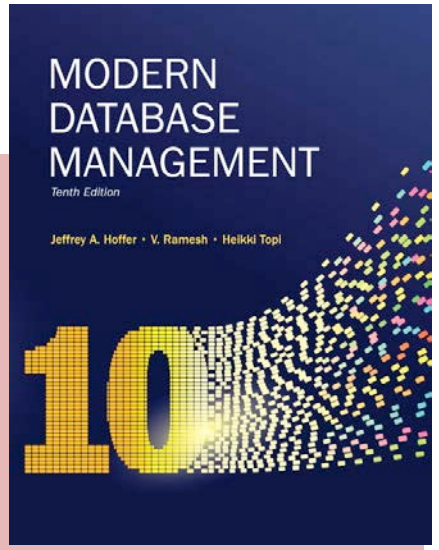
- "The Minto Pyramid Principle: Logic in Writing, Thinking, & Problem Solving" by Barbara Minto, 1996 (Used ~$75)



- "Blah Blah Blah: What To Do When Words Don't Work" by Dan Roam, 2011 (~$20)



Database books



Required:

- "Modern database management (10th ed)" by Hoffer, Ramesh, Topi, 2010 (used ~$70)
  www.amazon.com/Modern-Database-Management-10th-Edition/dp/0136088392/
  www.amazon.com/Modern-Database-Management-Jeffrey-Hoffer/dp/1408264315/

Other relevant book (not required):

- "Fundamentals of Database Management Systems by Mark Gillenson, 2011 (~$70)