Data Science at U of U

Jeff M. Phillips
Assistant Professor, School of Computing
Center for Extreme Data
Management, Analysis, and Visualization
Director, Data Management and Analysis Track
University of Utah
Classic Scientific Paradigm

make hypothesis
Classic Scientific Paradigm

- make hypothesis
- gather data
- data
Classic Scientific Paradigm

- Make hypothesis
- Gather data
- Data
- Compute statistics
Classic Scientific Paradigm

make hypothesis

gather data

data
compute
statistics

draw conclusion
Classic Scientific Paradigm

- Make hypothesis
- Gather data
- Data
- Compute statistics
- Draw conclusion
Modern Data Science

enormous
noisy data
Modern Data Science

enormous noisy data → data squashing → working data
Modern Data Science

enormous noisy data

data squashing

working data

data mining
Modern Data Science

enormous noisy data

data squashing

working data

data mining

draw conclusion
Modern Data Science

- enormous noisy data
- data squashing
- working data
- data mining
- draw conclusion

model uncertainty and bound error
Modern Data Science

enormous noisy data → data squashing → working data → data mining → draw conclusion

model uncertainty and bound error

Tools

Hadoop! | Data Munging | Tool Boxes (R, SciKit) | Illustrations and Plots
Modern Data Science

enormous noisy data → data squashing → working data → data mining → draw conclusion

model uncertainty and bound error

Tools
- Hadoop!
- Data Munging
- Tool Boxes (R, SciKit)
- Illustrations and Plots

Fundamentals
- Databases (no SQL)
- Modern Algorithms
- Machine Learning & Data Mining
- Visualization & Statistics
Data Science
Algorithms
Statistics
Data Mining
Machine Learning
Databases
Data Science
Visualization
Data Scientist

Data Science
Algorithms
Fundamentals
Fundamentals

Databases

Redis
HBase
Cassandra
mongoDB
CouchDB
Scalaris
Neo4j the graph database
membase
Tokyo Cabinet
Project Voldemort
Computing at U.Utah View
www.cs.utah.edu/bigdata

Fundamentals

Databases

Algorithms

Data Mining

Isotonic regression

Data
Isotonic Fit
Linear Fit
Computing at U.Utah View
www.cs.utah.edu/bigdata

Fundamentals

Databases

Algorithms

Data Mining

Machine Learning

Visualization
Modern Data Science

enormous noisy data

data squashing

working data

data mining

draw conclusion

model uncertainty and bound error

Big Data, Big Questions

1. If we get good enough at finding pattern, will we find patterns that are not there? (False Discovery)
Modern Data Science

enormous noisy data

data squashing

working data

data mining

draw conclusion

model uncertainty and bound error

Big Data, Big Questions

1. If we get good enough at finding pattern, will we find patterns that are not there? (False Discovery)

2. Can you analyze big data without violating personal privacy?
Modern Data Science

- enormous noisy data
- data squashing
- working data
- data mining
- draw conclusion

model uncertainty and bound error

Big Data, Big Questions

1. If we get good enough at finding pattern, will we find patterns that are not there? (False Discovery)

2. Can you analyze big data without violating personal privacy?

3. As we get more sources of data, how will we know which to trust?
Teaching Data Science
3 Bold Proclamations

1. Data Science should be taught in **Computer Science** departments.
   *(others can teach domain specific tools)*
Teaching Data Science
3 Bold Proclomations

1. Data Science should be taught in **Computer Science** departments.
   *(others can teach domain specific tools)*

2. Every domain of Science & Engineering will **center** on Data Science in 5 years.
Teaching Data Science
3 Bold Proclamations

1. Data Science should be taught in Computer Science departments. (others can teach domain specific tools)

2. Every domain of Science & Engineering will center on Data Science in 5 years.

3. All of Data Science will be applying Machine Learning in 10 years.