Educating a New Breed of Data Scientists for Research Data Management

Jian Qin

School of Information Studies
Syracuse University, NY, USA

Escola de Ciência da Informação da Universidade Federal de Minas Gerais
October 28, 2014

DS

Talk points

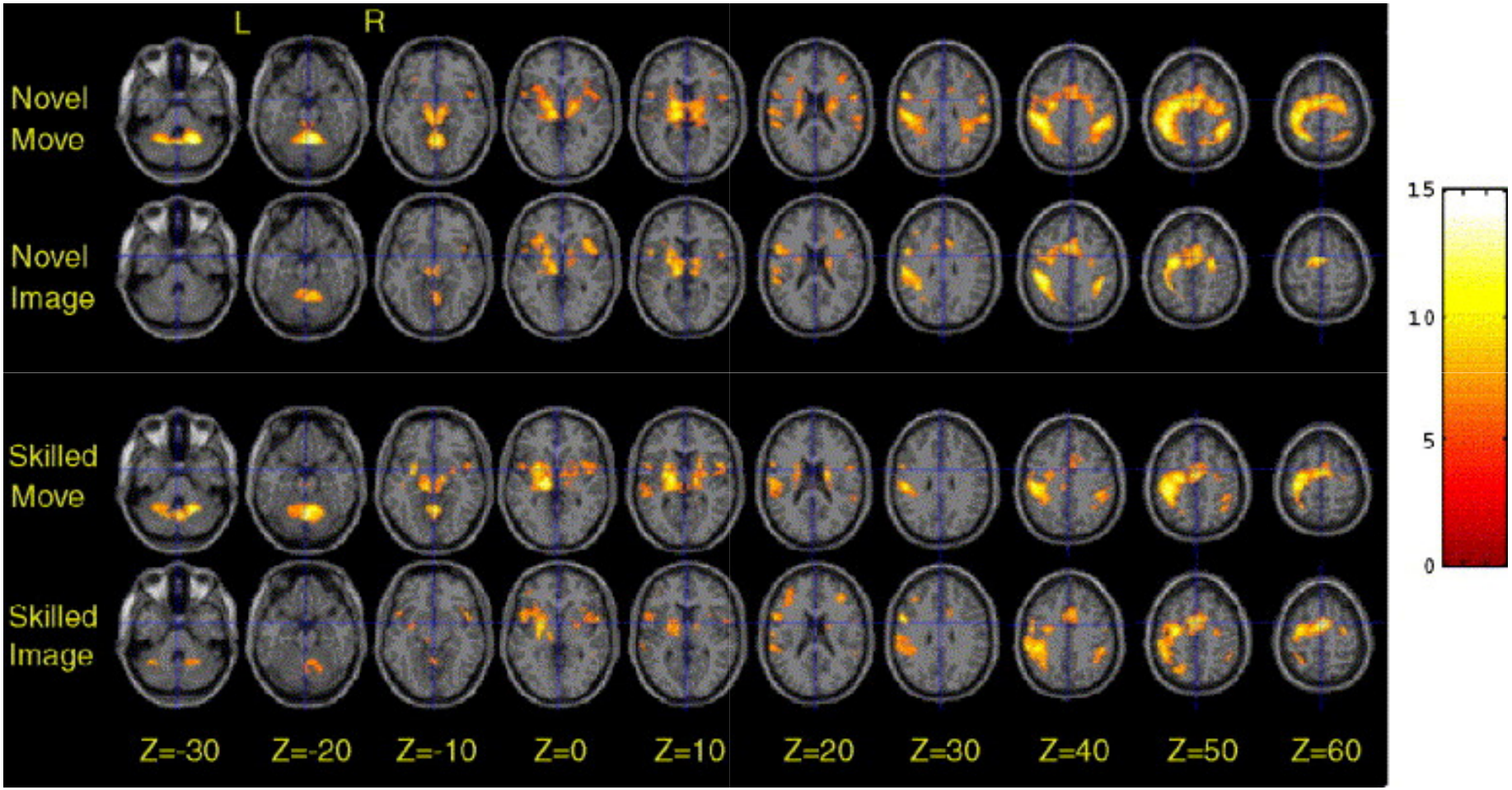
- Data science (DS) and data scientists in the context of research data
- An iSchool version of the DS curriculum
- Findings and lessons from implementing the DS curriculum
- A new breed of data scientists: the iSchool approach

DS

Feeling the pressure of data deluge in the digital information world …

http://readwrite.com/2011/11/17/infographic-data-deluge---8-ze

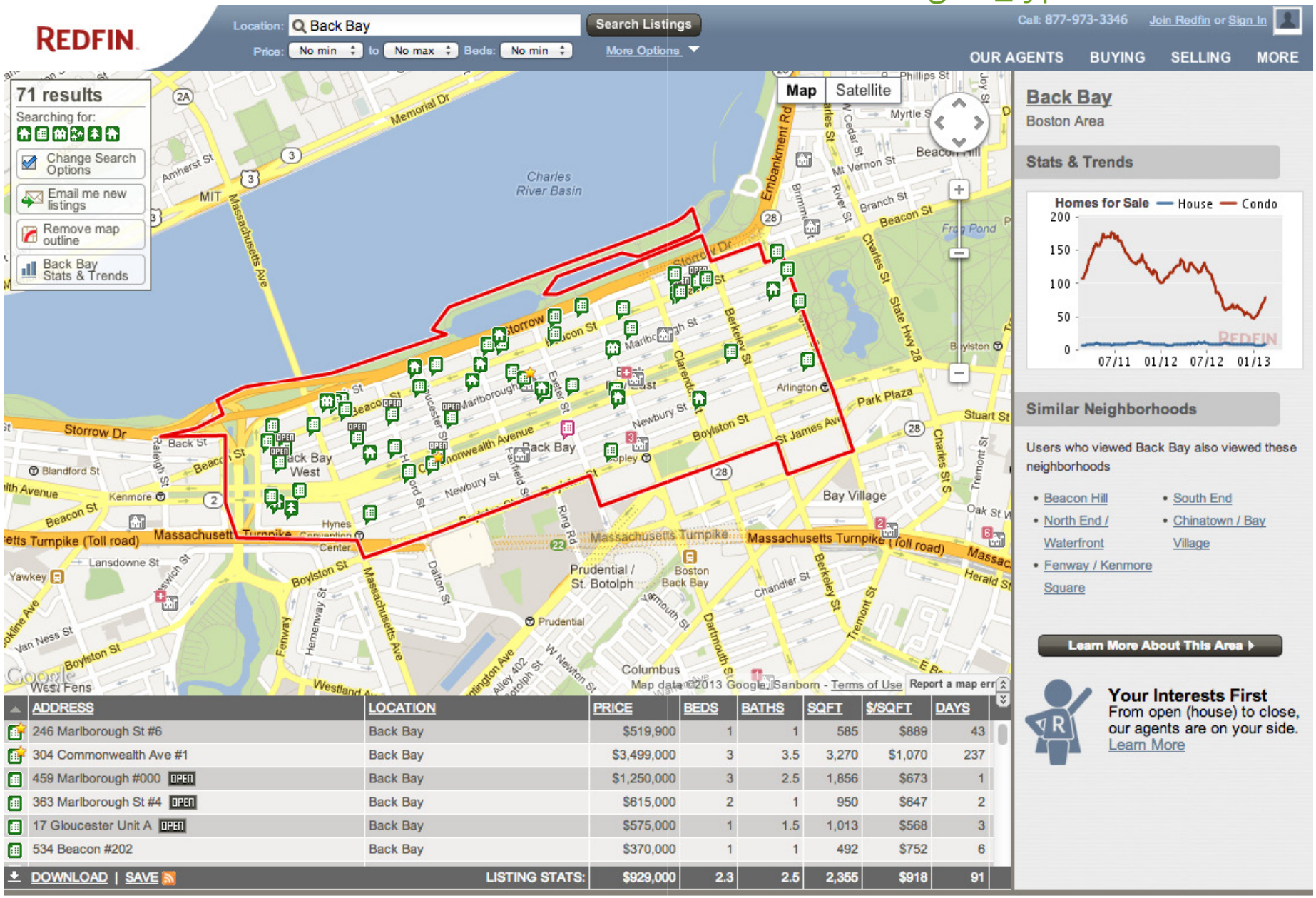
...in our health care



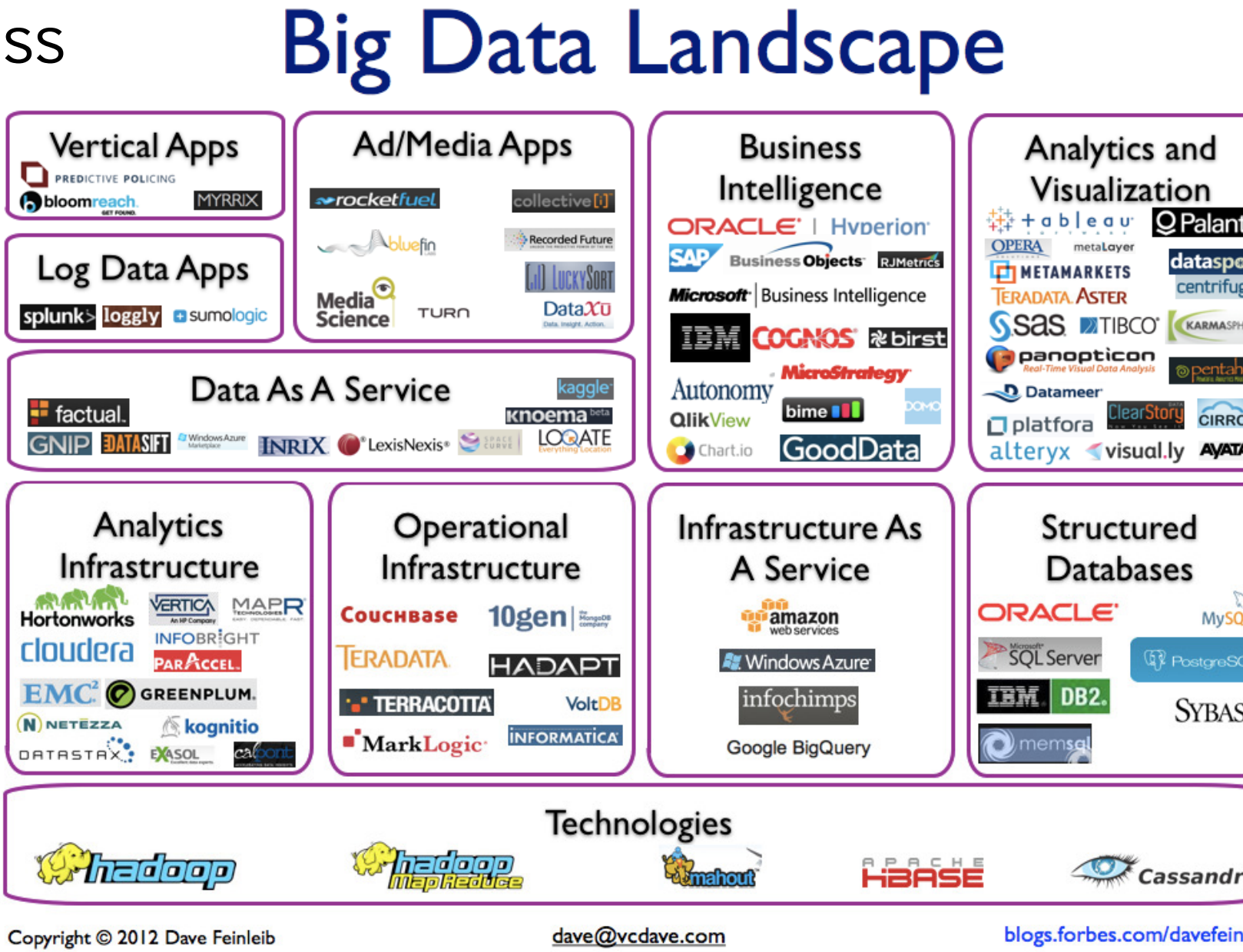
http://ars.els-cdn.com/content/image/1-s2.0-S1053811905002508-gr4.jpg

…in our neighborhood

http://www.redfin.com/homes-for-sale#!market=boston&region_id=112 region_type=1&v=8

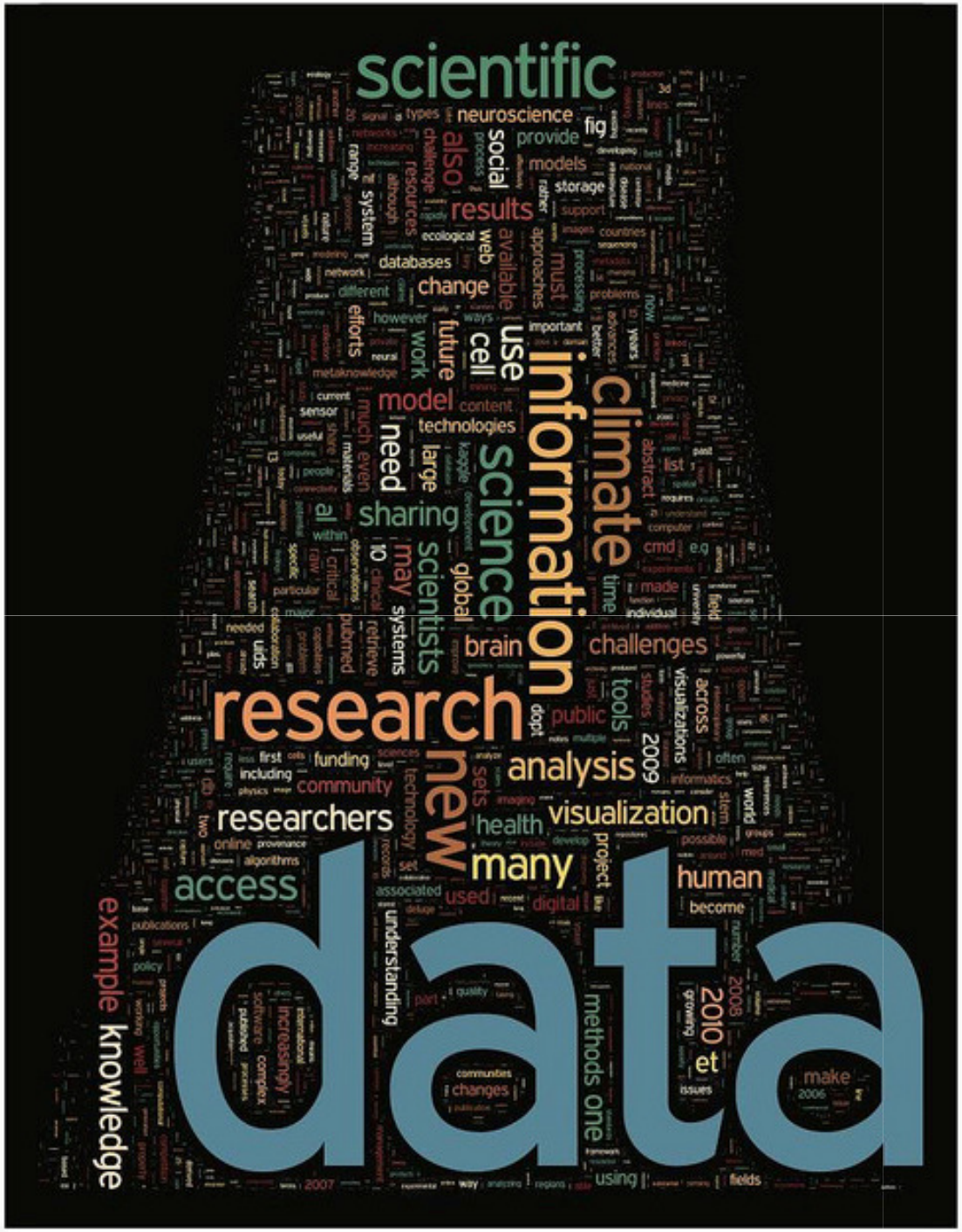


In the business world...



(Feinleib, 2012)

DS



...in science research

http://www.sciencemag.org/content/331/6018.cover-expansion

Shift in Science Paradigms

Thousand years ago	A few hundred years ago	A few decades ago	Today

Science was empirical
describing natural phenomena

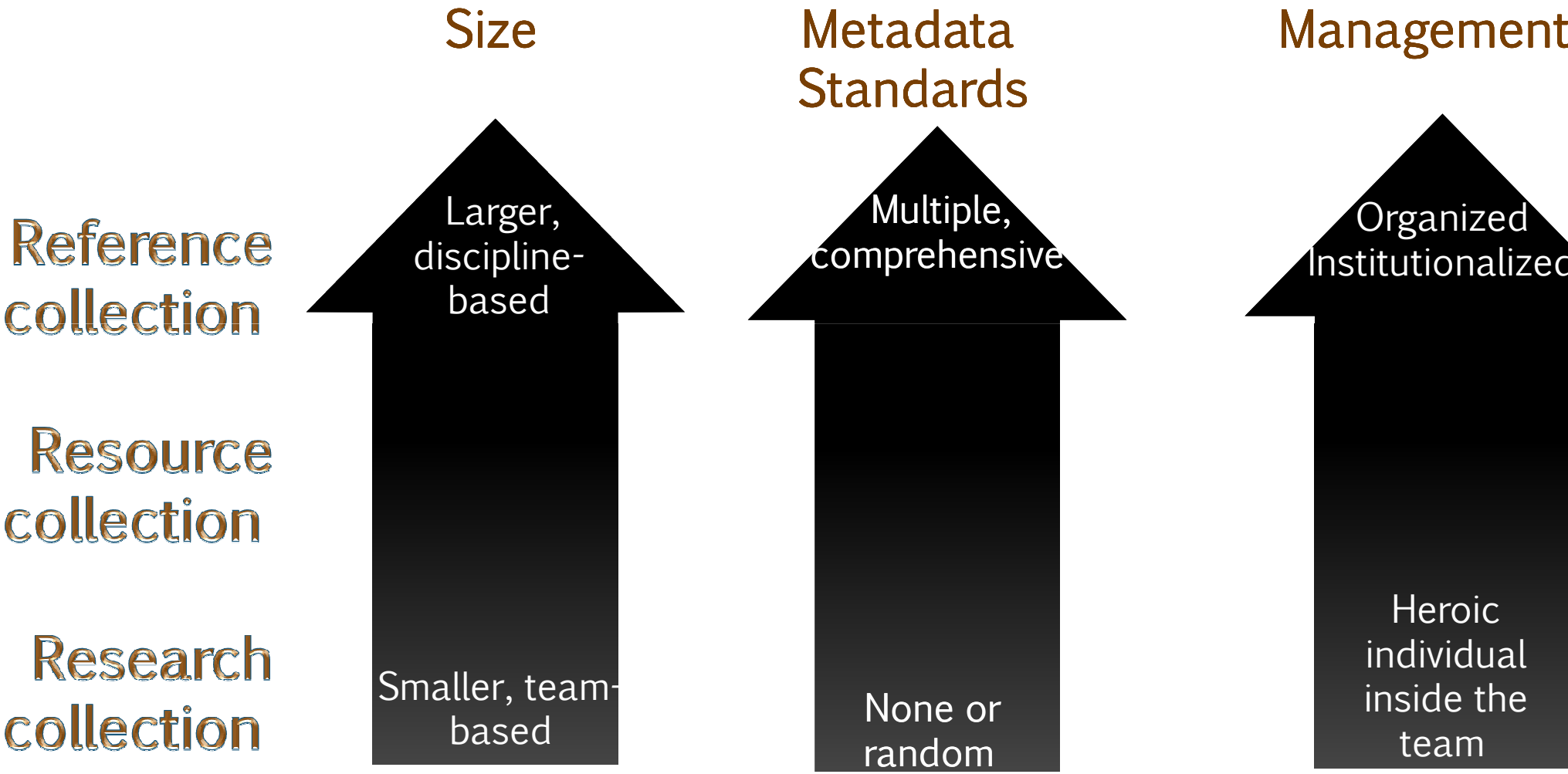
Theoretical branch
using models, generalizations

A computational approach
simulating complex phenomena

Data exploration (eScience)
unify theory, experiment, and simulation
-- Data captured by instruments or generated by simulator
-- Processed by software
-- Information/Knowledge stored in computer
-- Scientist analyzes database/files using data management and statistics

Gray, J. & Szalay, A. (2007). eScience – A transformed scientific method. http://research.microsoft.com/en-us/um/people/gray/talks/NRC-CSTB_eScience.ppt

Research data collections



DS

Emerging concepts

that are going to stay and matter to your career

eScience

Data management is essential

DS

entific Data Management ecialist

esign, develop, implement, and **anage** high-throughput automatic ata processing infrastructure for rge databases in a mature system

evelop and improve the frastructure supporting this stem

terface with multiple data oviders to design, build, and aintain their customized atabases

arify requirements, feature quests and bug reports for ftware developers and assist in sting code.

http://www.bioinformatics.org/forums/forum.php?forum_id=9670

Laboratory Data Management Specialist

- **Administer** operational database
- **Assure** the quality of data database content
- **Interact** closely with researchers, lab managers, and platform coordinators
- **Track** deliverables against budget and prepare data reports
- **Collaborate** closely with IT and bioinformatics colleagues
- **Assist** IT in gathering workflow requirements
- **Test** changes and updates in IT systems
- **Create and maintain** app documentation

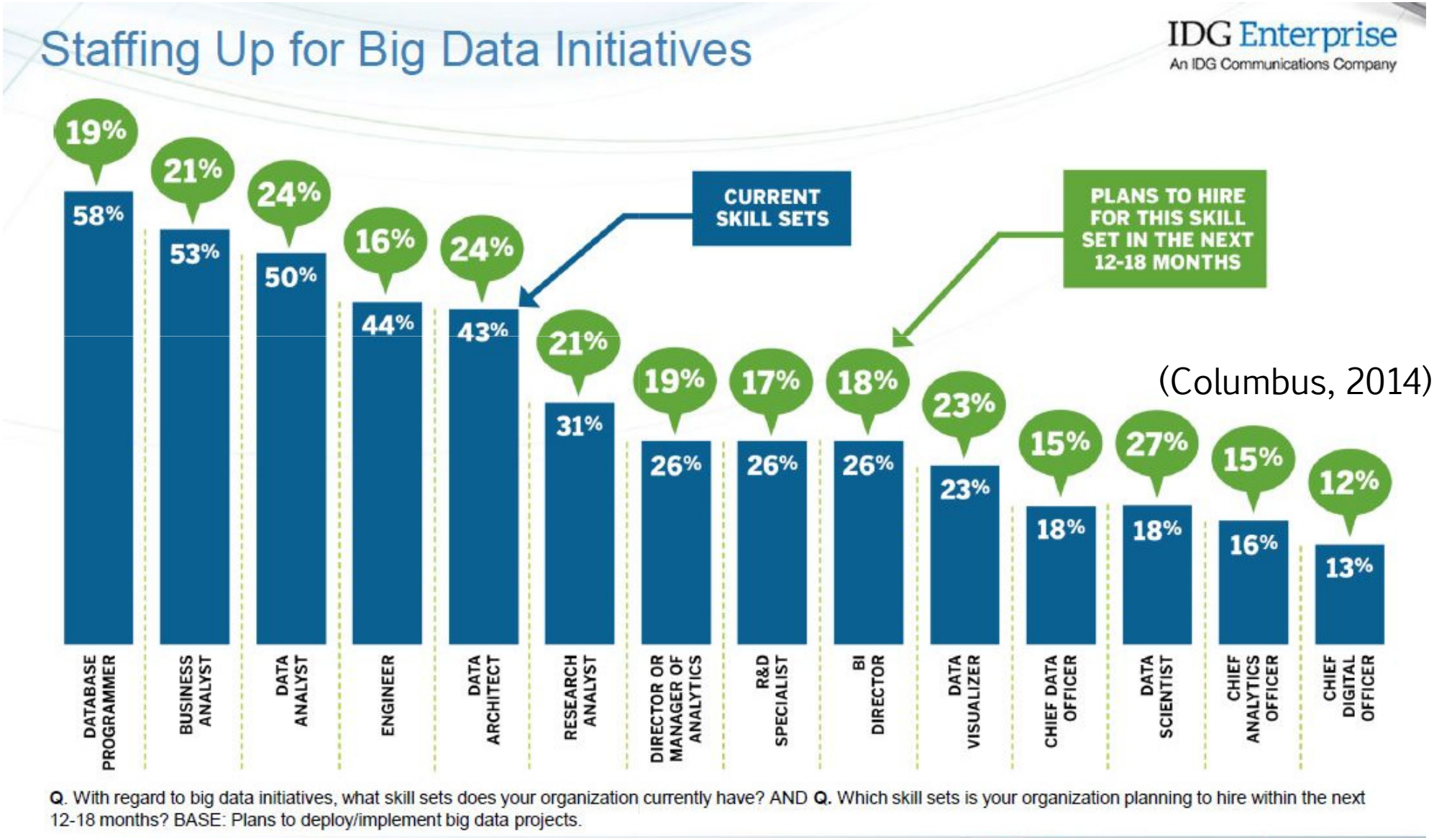
Data Modeling/ Management Specialis

- **Work** closely with the high performance computing a the IT manager
- **Develop** a data model for complex multi-scale rocks
- **Design and organize** a database and complex queries
- **Integrate and mange** mult scale rocks subjected to large-scale scientific computing applications

http://www.ingrainrocks.com/data-management-specialist/

DS

Market demand for DS skills



(Columbus, 2014)

Data science

What is it ?

What does it mean for future information science workforce education and employment?

DS

What is data science?

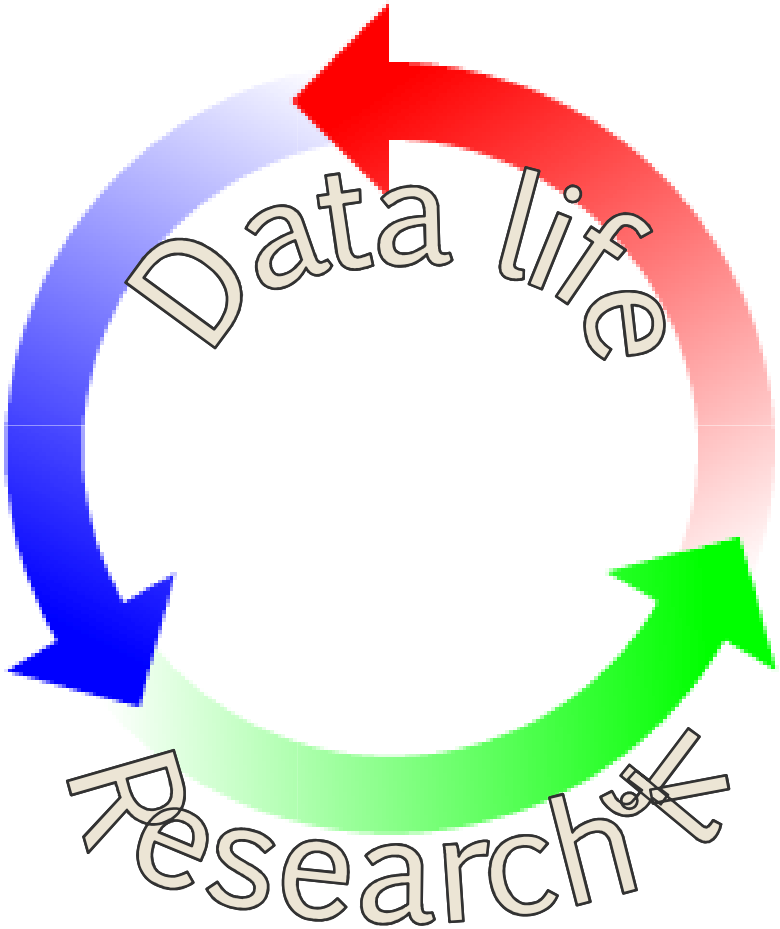
"An emerging area of work concerned with the collection, presentation, analysis, visualization, management, and preservation of large collections of information."

Stanton, J. (2012). Introduction to Data Science. http://ischool.syr.edu/media/documents/2012/3/DataScienceBook1_1.pdf

DS

Data science and scientific research

Plan, design, consult for, implement, and evaluate data management projects and services



Ingest, store, organize, merge, filter, and transform data and create analysis-ready dat

DS

"Data scientists are the people who understand how to fish out answers to important business questions from today's tsunami of unstructured information."

Davenport, T.H. & Patil, D.J. (October 2012). Data scientist: the sexiest job of the 21st century. Harvard Business Review, 70-76.

How should educational programs address the challenge?

A case of the CAS in Data Science program at Syracuse iSchool


DATA SCIENCE AT THE iSCHOOL

DS

Development of the data science program

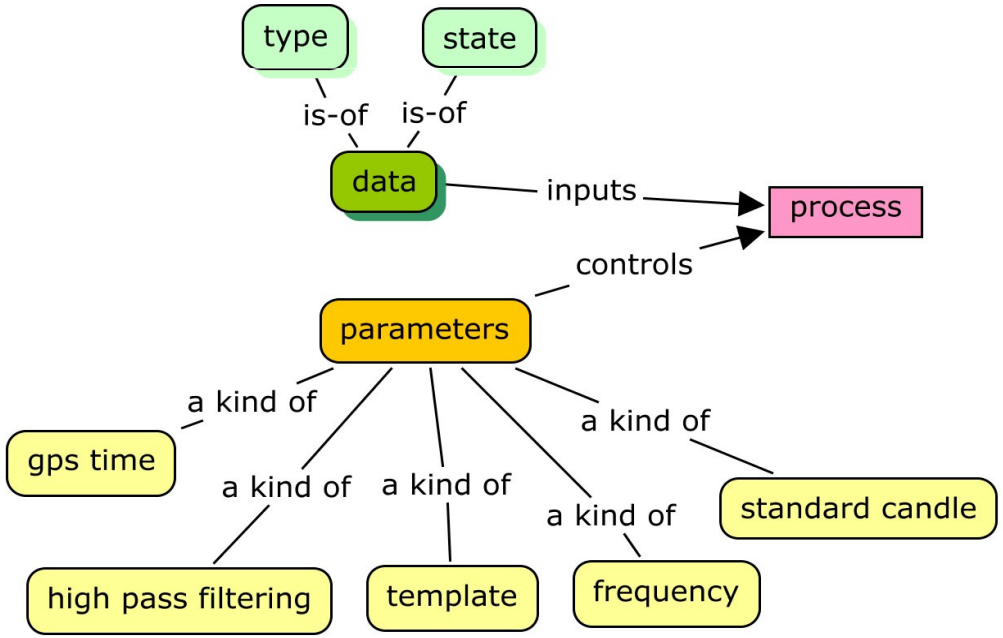
- CI-Facilitators: Information Architects across the STEM Disciplines
- Scientific Data Literacy curriculum development project: http://sdl.syr.edu/
- eScience Librarianship: Education and Training http://eslib.ischool.syr.edu/wp/
- CAS in Data Science http://ischool.syr.edu/future/cas/datascience.aspx

Story 1: cognitive-demanding workflows and data management

- *Domain:* Thermochronology and tectonics
- *What's involved*: rock samples from drilling and field observation, sliced and grained rock samples
- *Data types*: Excel data files (lots of them), spectrum and microscopic images annotations
- *Analysis*: modeling and sensemaking by combining data from multiple data files with specialized software
- *Bottleneck problem*: manually matching/merging/filtering data is extremely cumbersome and the problem is compounded by the difficulty finding the right data files

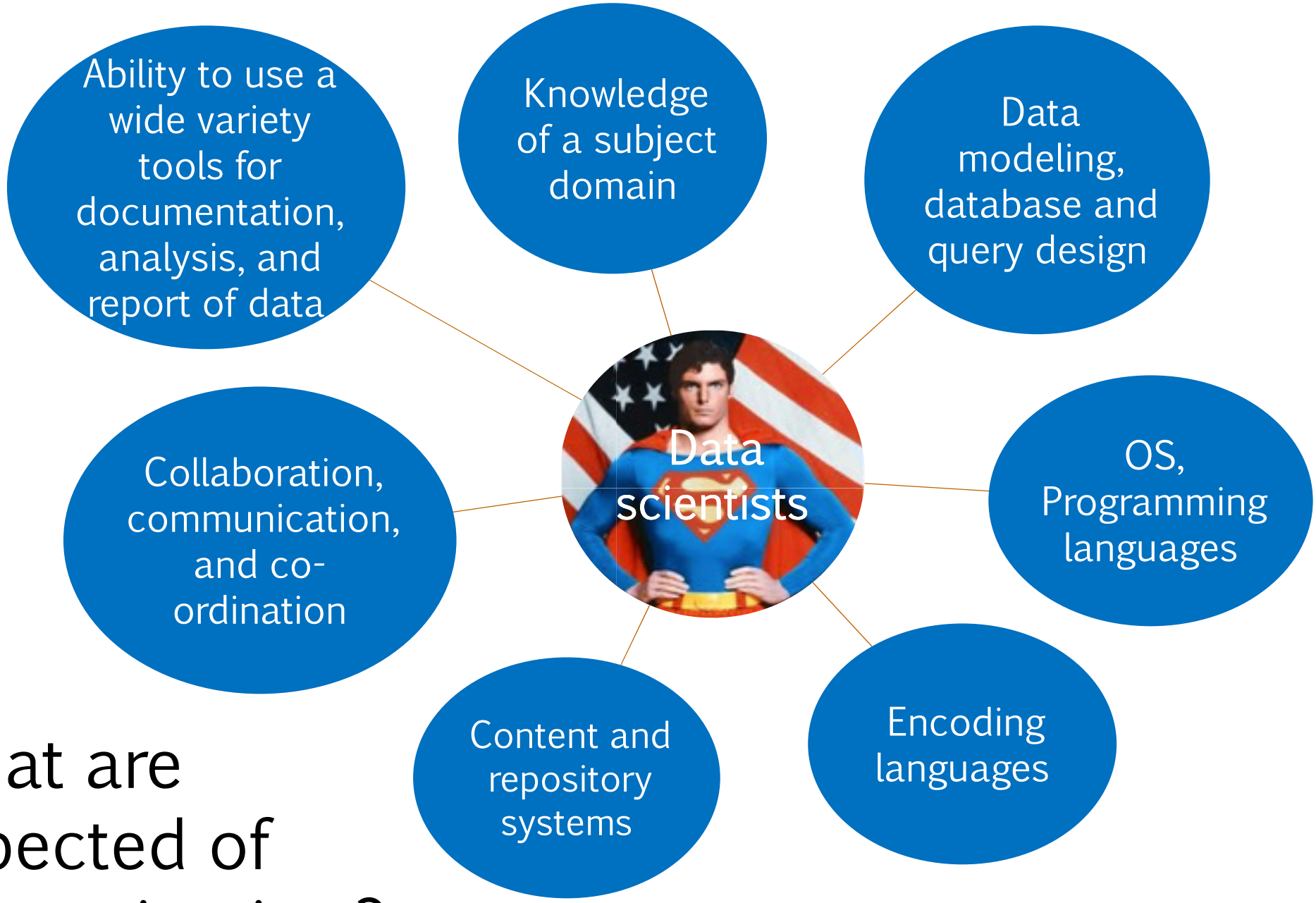
Story 2: highly automated workflows

- *Domain:* Astrophysics: gravitational wave detection
- *What's involved*: data ingestion from laser interferometers, raw data calibration and segmentation, workflow management, provenance
- *Data types*: streaming data from the laser interferometers, images
- *Analysis*: detection of "events"
- *Bottleneck problem*: tracking of data and processes and the relationships between them



DS

What are expected of data scientists?

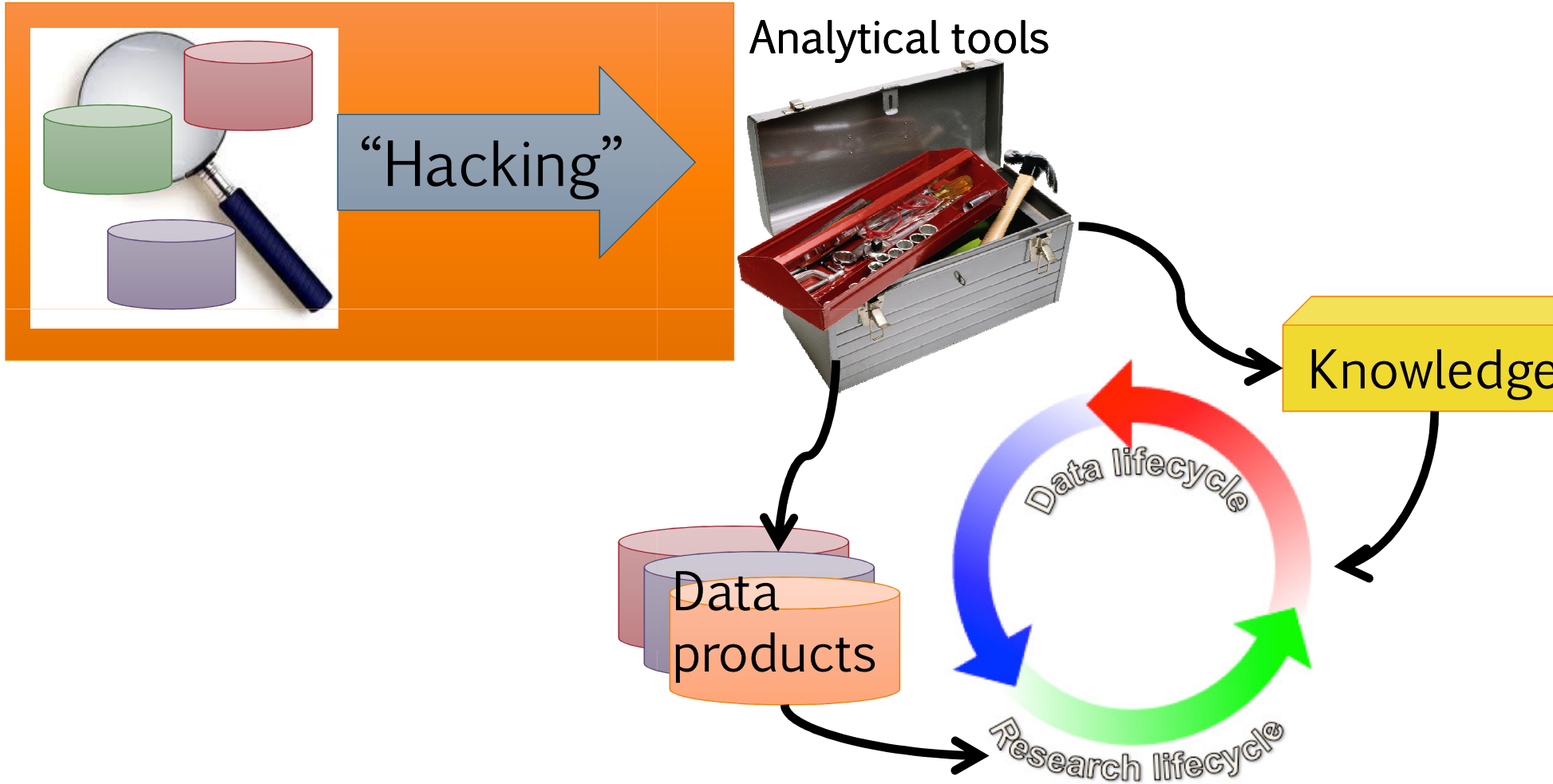


Analytical skills: domain modeling

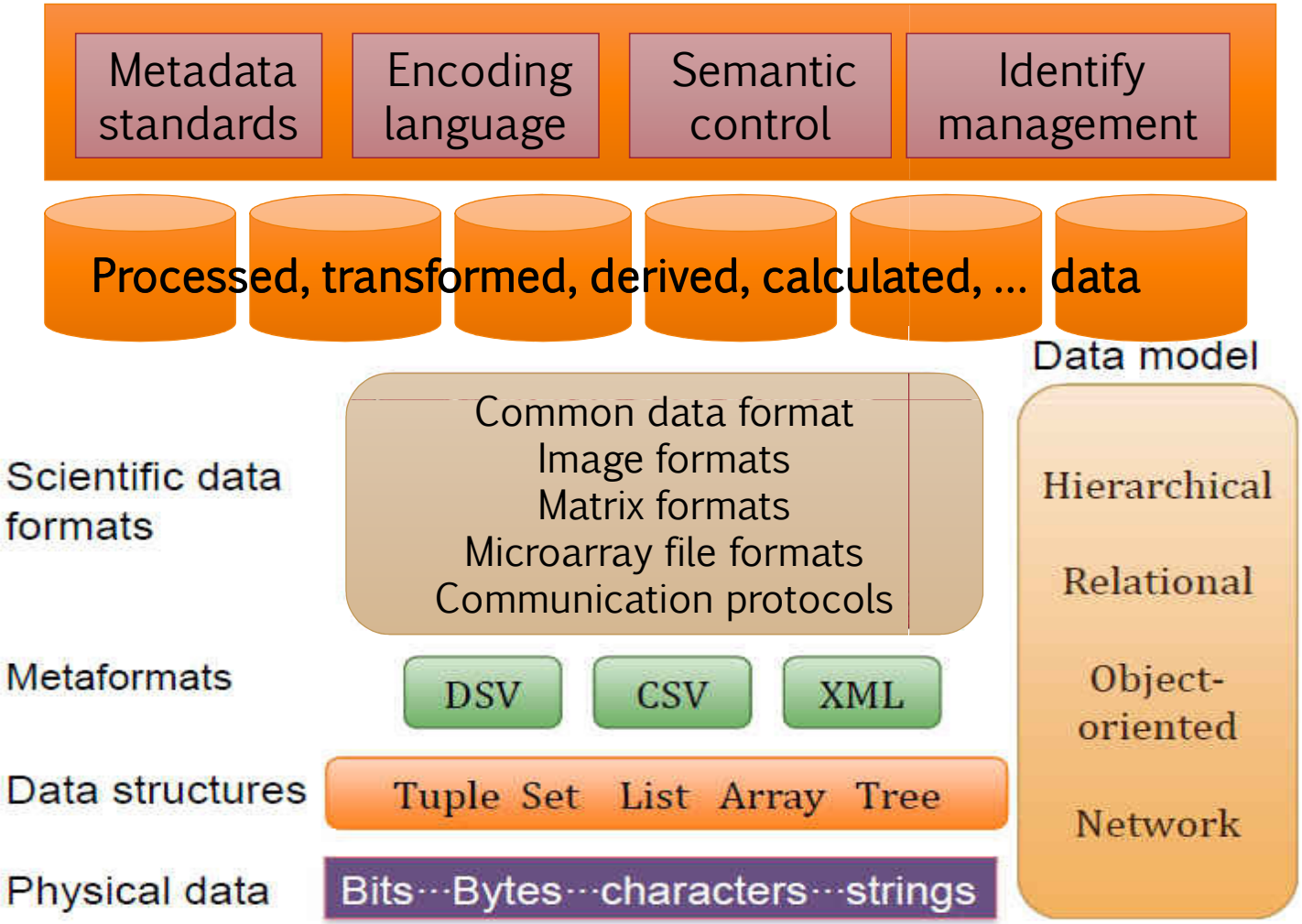
- Requirement analysis
- Workflow analysis
- Data modeling
- Data transformation needs analysis
- Data provenance needs analysis

- Interview skills, analysis and generalization skills
- Ability to capture components and sequences in workflows
- Ability to translate domain analysis into data models
- Ability to envision the data model within the larger system architecture

Analytical skills: from data sources to patterns, relationships, and trends



Data management skills: data lifecycle and infrastructural services



Infrastructural services

- Data source discovery
- Data curation
- Data preservation
- Data integration ar mashup
- Data citation, publication, and distribution
- Data linking and interoperability
- ...

DS

Technology skills with excellent communication skills

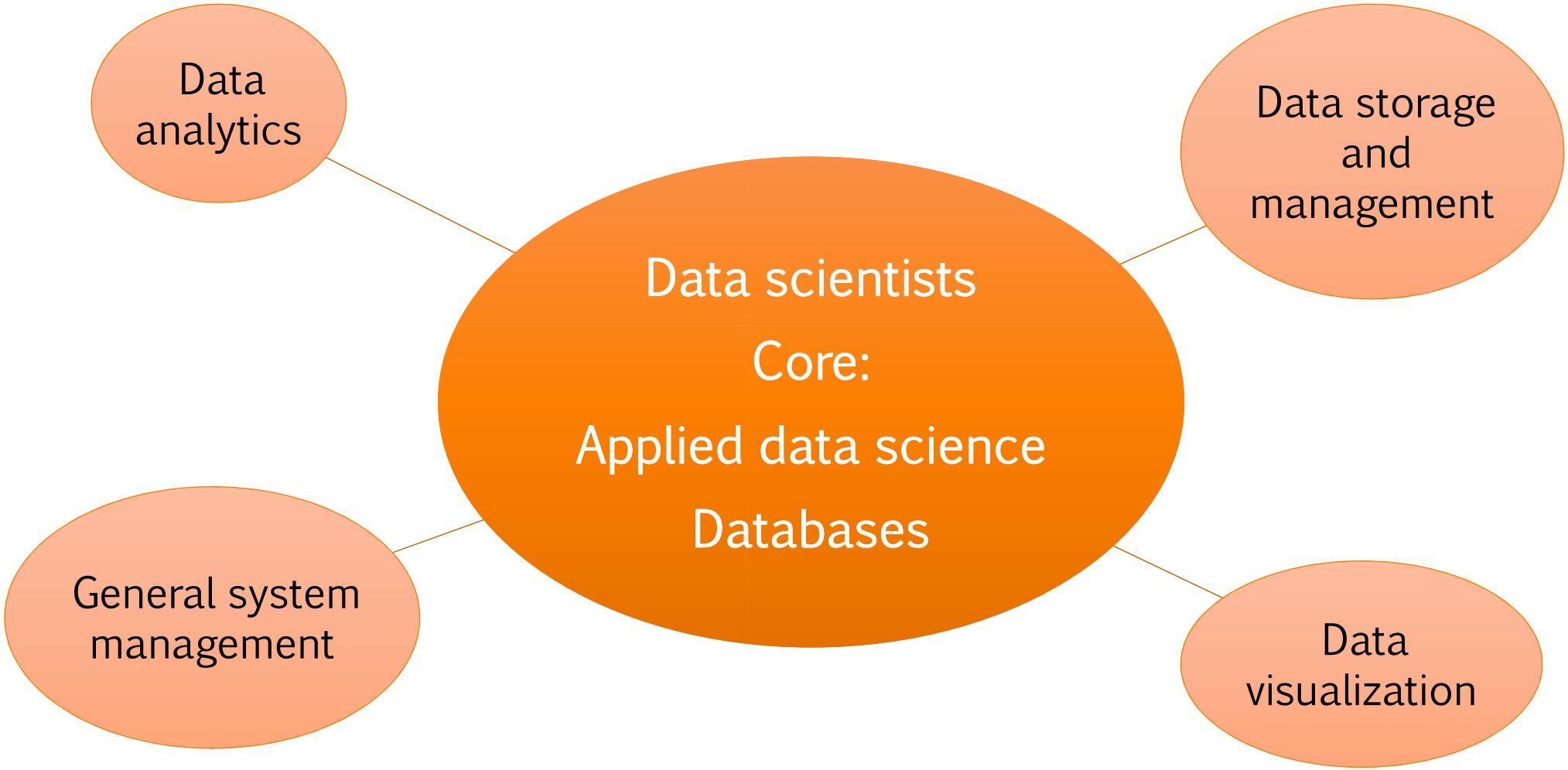
TECHNOLOGY SKILLS

- Operation systems
- Repository systems
- Database systems
- Programming languages
- Encoding languages
- Specialized programming

COMMUNICATION SKILLS

- Interviews
- “Ice breaking”
- Community building
- Institutionalization
- Stakeholder buy-in

No superman model for beginning data scientists



DS

The CAS in Data Science program at SU

› Required:
  - Data Administration Concepts and Database Management
  - Applied Data Science

› Elective:

Data Analytics	*Data Storage and Management*	*Data Visualization*
• Data Mining • Basics of Information Retrieval Systems • Natural Language Processing • Advanced Information Analytics • Research Methods • Statistical Methods	• Technologies for Web Content Management • Foundations of Digital Data • Creating, Managing, and Preserving Digital Assets • Data Warehousing • Advanced Database Management	• Information Architecture for Internet Services • Information Visualization *General Systems Management* • Enterprise Technologies • Managing Information Systems Projects • Information Systems Analysis

DS

What we learned from the program development

- Data science is a moving target with multiple focal points
  - Versions from statistics, computer science, and library and information science
- Skills vs. theories
  - Students are anxious to learn skills but not so interested in theories
  - Theories help build visions
- Sufficient hands-on time for technologies and tools
- Authentic learning through real-world data management projects

DS

Reconciliation of the two views of data science

"An emerging area of work concerned with the collection, presentation, analysis, visualization, management, and preservation of large collections of information."

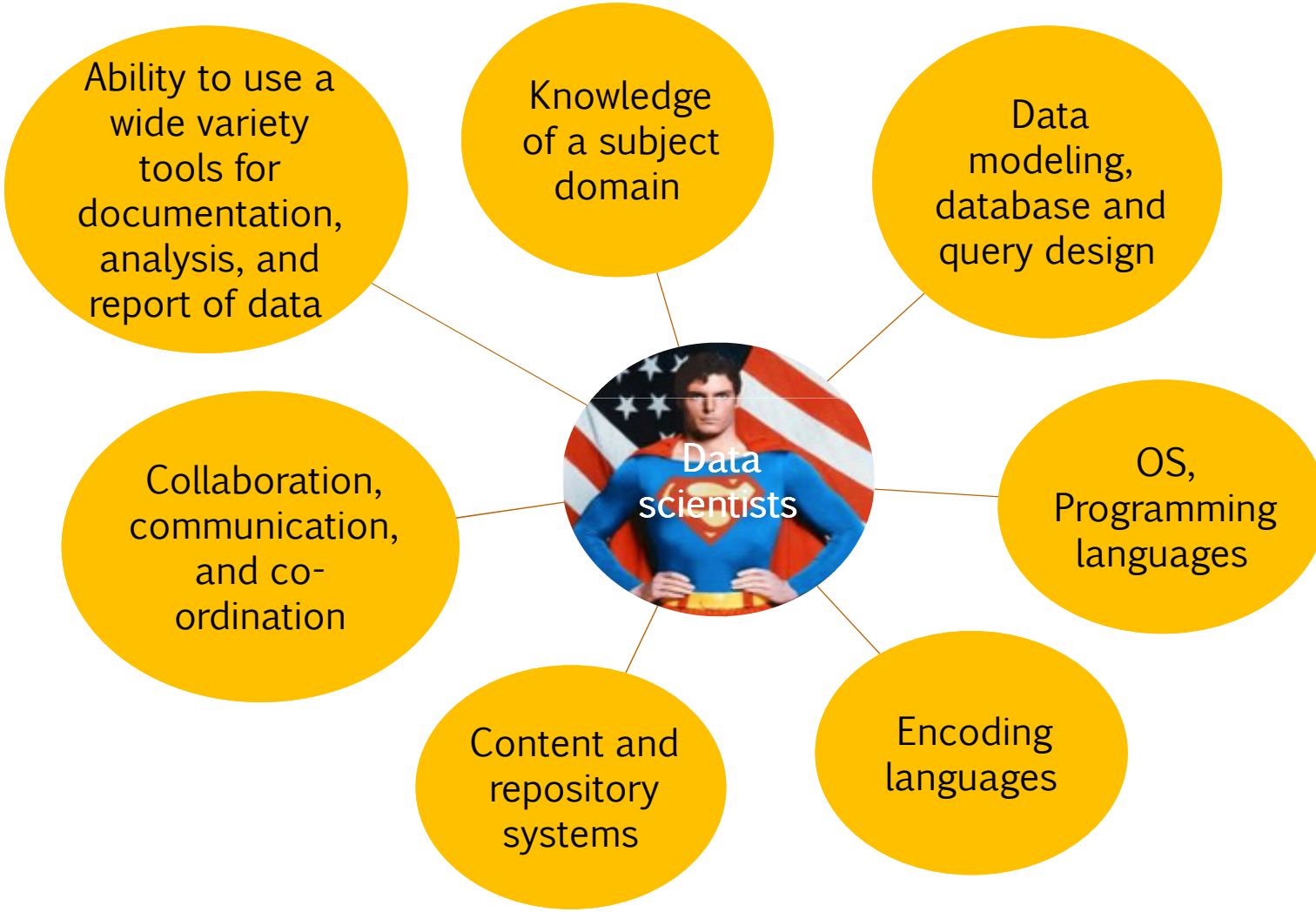
Stanton, J. (2012). Introduction to Data Science. http://ischool.syr.edu/media/documents/2012/3/DataScienceBook1_1.pdf

"We're increasingly finding data in the wild, and data scientists are involved with gathering data, massaging it into a tractable form, making it tell its story, and presenting that story to others."

Loukides, M. (2011). What is data science? Sebastopol, CA: O'Reilly.

The iSchool's version of data science education

Eventually the iSchool data science program will build the foundation for super data scientists…



DS

eScience Librarianship Curriculum Project:
http://eslib.ischool.syr.edu/

Science Data Literacy Project:
http://sdl.syr.edu/

CAS in Data Science:
http://ischool.syr.edu/future/cas/data-science.aspx

DS

References

- Columbus, L. (2014). 2014: The year Big Data adoption goes mainstream in the enterprise. Forbes, http://www.forbes.com/sites/louiscolumbus/2014/01/12/2014-the-year-big-data-adoption-goes-mainstream-in-the-enterprise/
- Feinleib, D. (2012). The Big Data landscape. Forbes, http://www.forbes.com/sites/davefeinleib/2012/06/19/the-big-data-landscape/