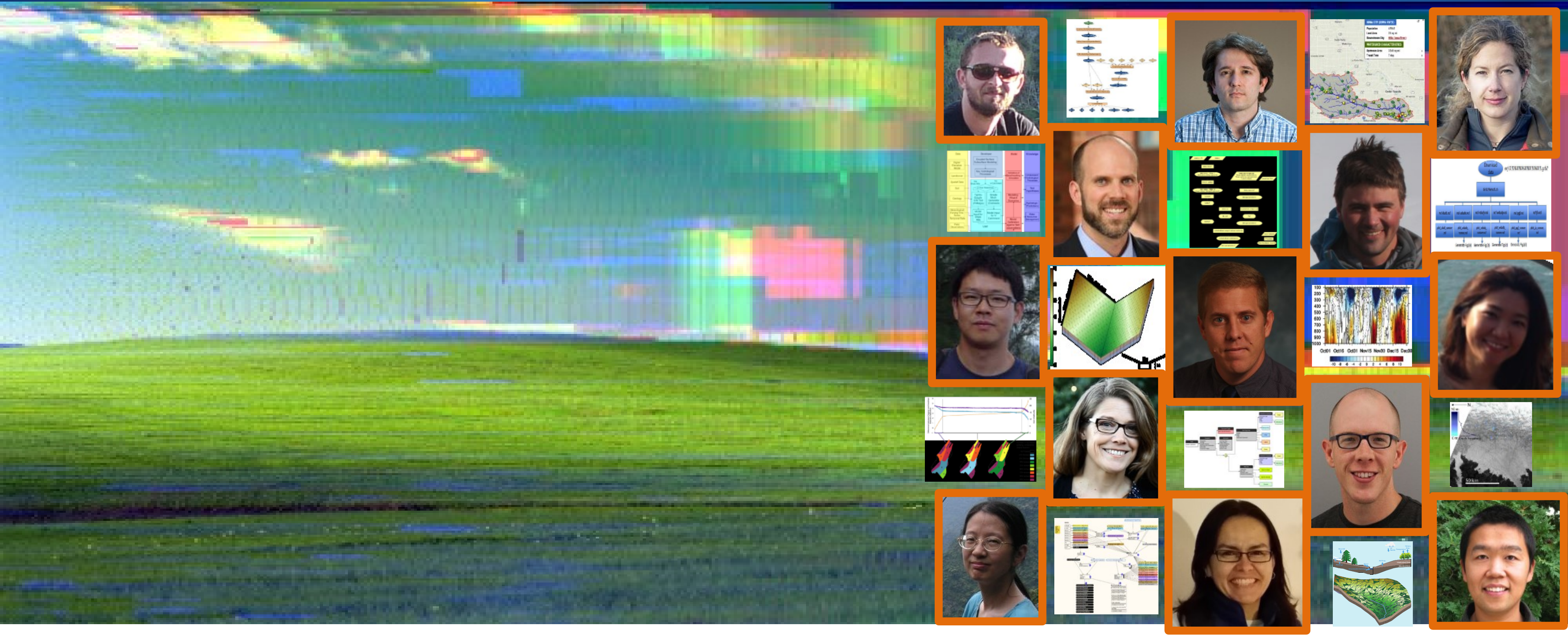
Course: "Introduction to Computational Thinking and Data Science"

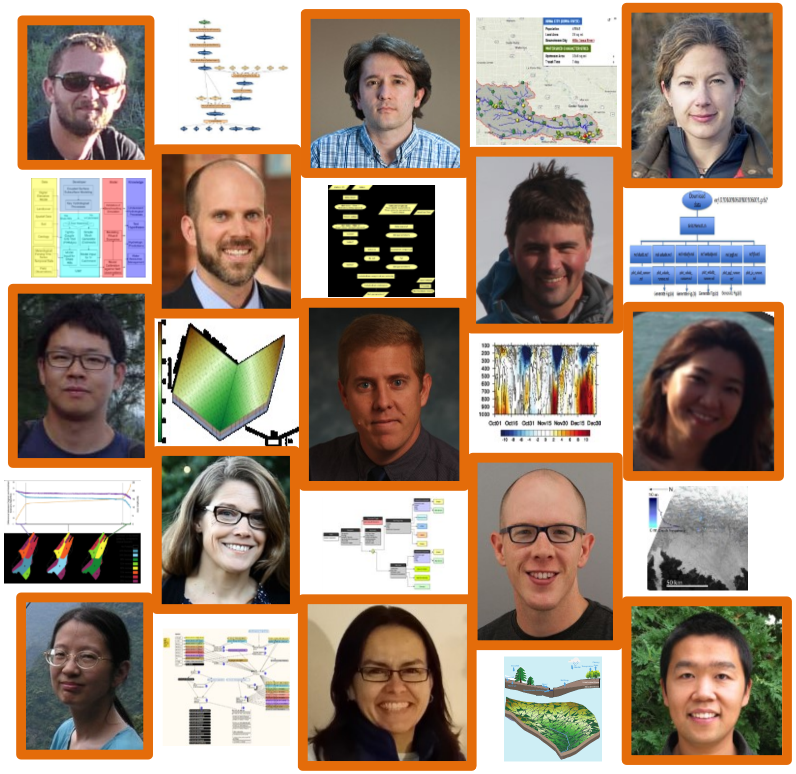
Yolanda Gil
gil@isi.edu



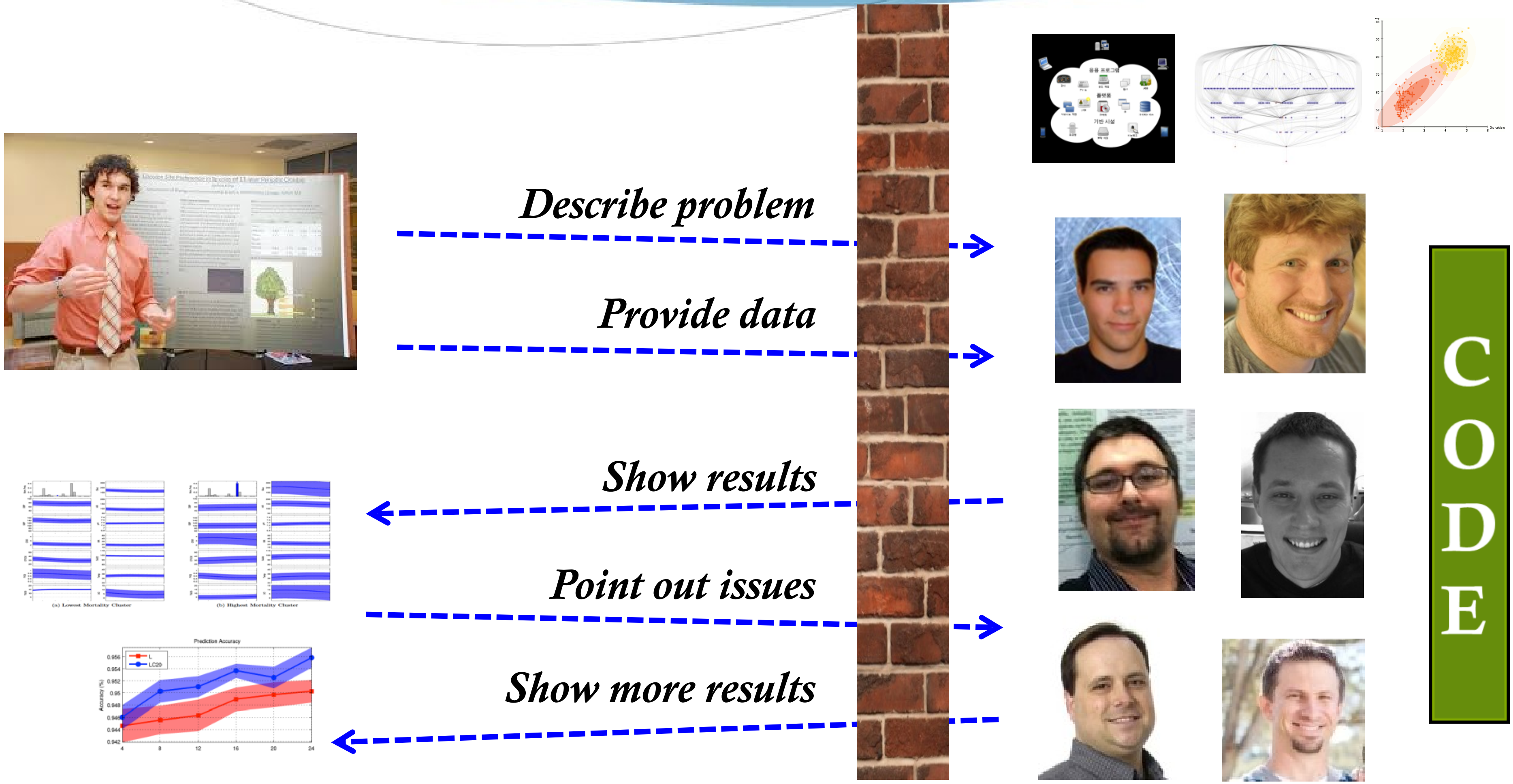
https://www.flickr.com/photos/hellocatfood/10545396704

Goals of the Course

Course is **designed for students with no programming background who want to have literacy in computing and data science** to better approach data-rich problems



Barriers of Non-Programmers to Data Science



Describe problem

Provide data

Show results

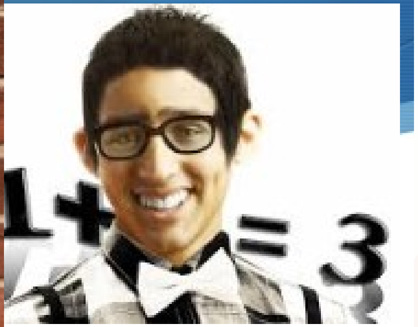
Point out issues

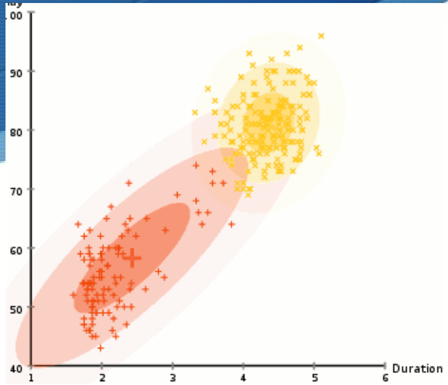
Show more results

CODE

Distinct Expertise in Data Science

Domain knowledge



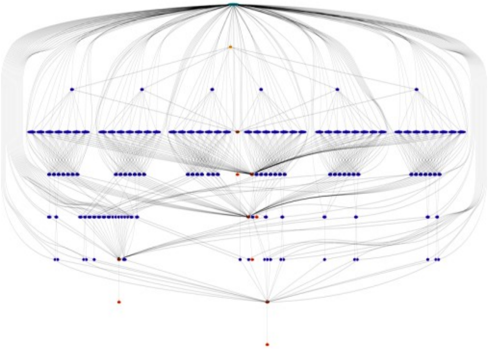


Statistics, data mining

Semantics and distributed data

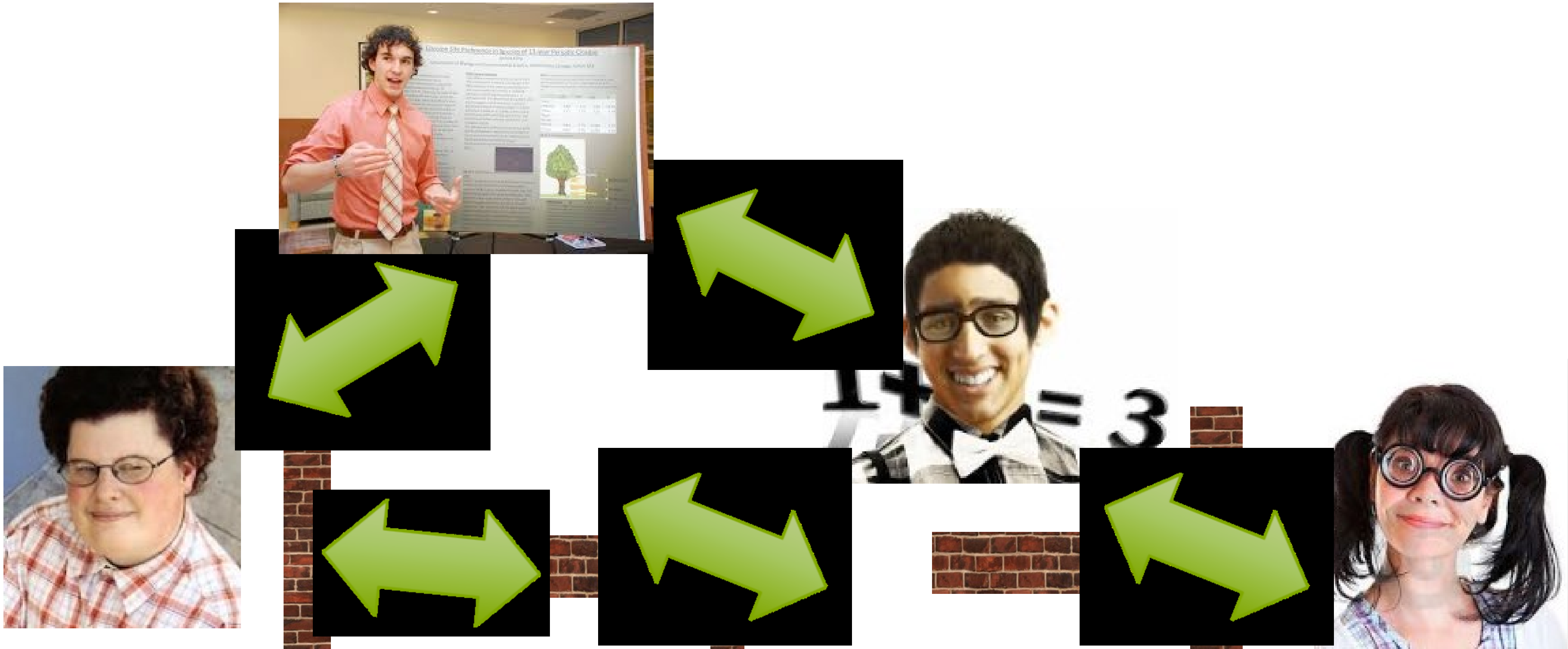
응용 프로그램
감시
콘텐츠
공동 작업
통신
금융
플랫폼
오브젝트 저장
신원
런타임
큐
데이터베이스
기반 시설
컴퓨팅
블록 저장
네트워크

Large-Scale Data Processing



Becoming Data Scientists: Overcoming the Barriers

The goal of the class is to empower non-programmers
to communicate with computer scientists
so they can collaborate in real-world data science projects



Students Learn to Channel Their Domain Expertise into Data Science

- USC course attended by graduate students in:
  - Political sciences, social sciences, education, biology, medicine, engineering
- Palpable trajectory:
  - Week 1: Sketch a data science project
    - Good goals, but impractical, nonsensical, unmanageable
  - Week 7: Revisit the data science project
    - Sensible approach, technical vocabulary
  - Week 12: Revisit again
    - Practical implementable approach

Wings workflow system

Defin questions

Collect/find data

Protégé ontology editor

collaborative projects

provenance

Store data

standards

ontologies

databases

Extract data

text, network, geospatial, timeseries

Pre-process data

cleaning, imputation, reformatting

Analyze data

classification, clustering, causality, simulation

parallel processing

workflows

Present results

visualization

Publish data

repositories, licenses, DOIs

Design Principles for the Course

Conceptual Learning

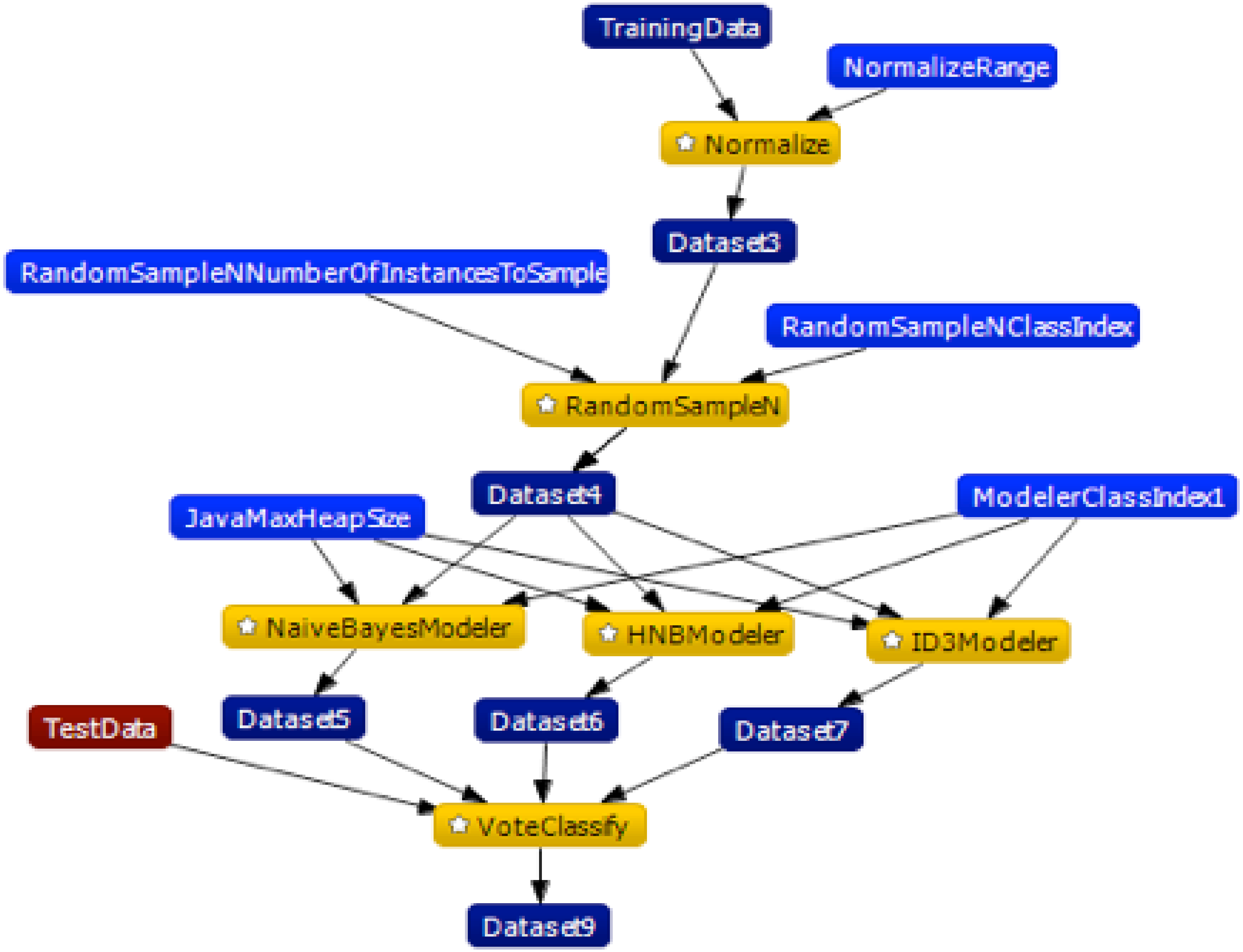
- **Computational thinking**: a new way to approach problems through computing
  - Simulation, data analysis, data mining
- **Data science**: a cross-disciplinary approach to solving data-rich problems
  - Machine learning, large-scale computing, semantic metadata

Practical Learning

- **Workflows**: a graphical programming environment that enables non-programmers to experiment with complex multi-step data analysis environments
- **Application domains**: exposure to past and ongoing projects where data science exposes multi-disciplinary challenges
  - Social networks, hydrology, proteomics, genomics, medicine, etc.

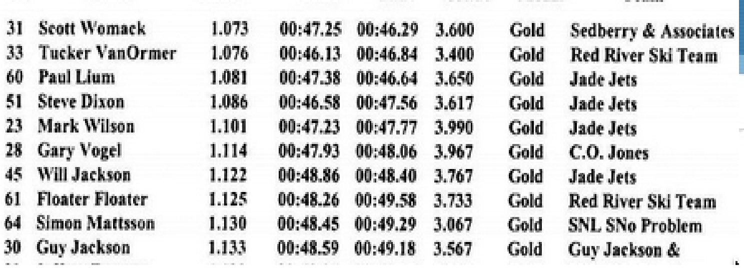
Practical Learning:

Workflows Make Data Science Accessible to Non-Programmers

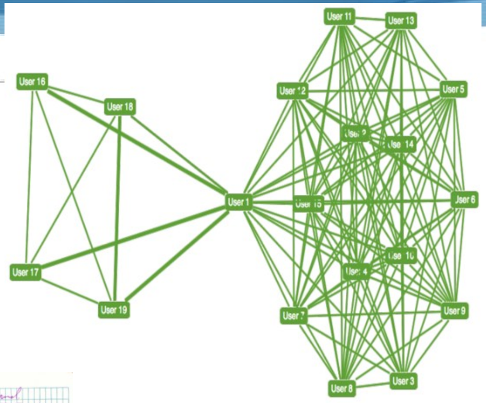


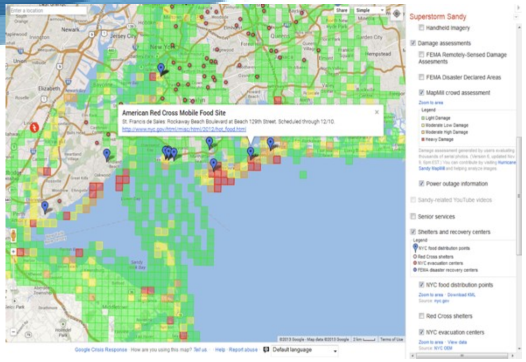
Practical Learning:

Workflows Enable Non-Programmers to Process Real-World Data



31	Scott Womack	1.073	00:47.25	00:46.29	3.600	Gold	Sedberry & Associates
33	Tucker VanOrmer	1.076	00:46.13	00:46.84	3.400	Gold	Red River Ski Team
60	Paul Lium	1.081	00:47.38	00:46.64	3.650	Gold	Jade Jets
51	Steve Dixon	1.086	00:46.58	00:47.56	3.617	Gold	Jade Jets
23	Mark Wilson	1.101	00:47.23	00:47.77	3.990	Gold	Jade Jets
28	Gary Vogel	1.114	00:47.93	00:48.06	3.967	Gold	C.O. Jones
45	Will Jackson	1.122	00:48.86	00:48.40	3.767	Gold	Jade Jets
61	Floater Floater	1.125	00:48.26	00:49.58	3.733	Gold	Red River Ski Team
64	Simon Mattsson	1.130	00:48.45	00:49.29	3.067	Gold	SNL SNo Problem
30	Guy Jackson	1.133	00:48.59	00:49.18	3.567	Gold	Guy Jackson &





American Red Cross Mobile Food Site
St. Francis de Sales: Rockaway Beach Boulevard at Beach 129th Street. Scheduled through 12/10.

Superstorm Sandy

☐ Handheld imagery

☑ Damage assessments

☐ FEMA Remotely-Sensed Damage Assesments

☐ FEMA Disaster Declared Areas

☑ MapMill crowd assessment

Zoom to area

Legend
- 🟩 Light Damage
- 🟨 Moderate Low Damage
- 🟧 Moderate High Damage
- 🟥 Heavy Damage

Damage assessment generated by users evaluating thousands of aerial photos. (Version 6, updated Nov 9, 6pm EST.) You can contribute by visiting Hurricane Sandy MapMill and helping analyze images.

☑ Power outage information

☐ Sandy-related YouTube videos

☐ Senior services

☑ Shelters and recovery centers

Legend
- NYC food distribution points
- ○ Red Cross shelters
- ● NYC evacuation centers
- ● FEMA disaster recovery centers

☑ NYC food distribution points

Zoom to area · Download KML

Source: nyc.gov

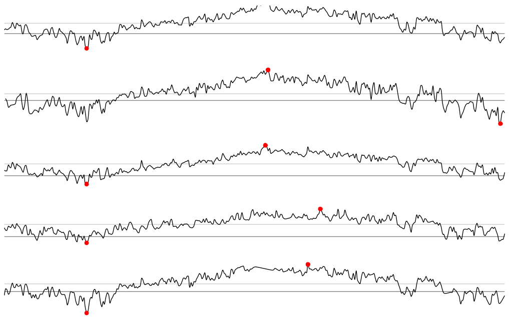
☐ Red Cross shelters

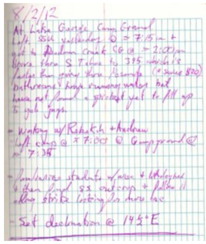
☑ NYC evacuation centers

Zoom to area · View data

Source: NYC OEM

Google Crisis Response · How are you using this map? Tell us · Help · Report abuse · Default language
Terms of Use





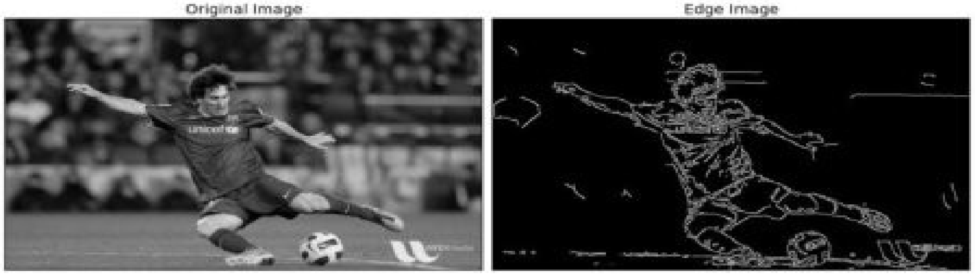
8/2/12

At Lake Garda Camp Ground
Left SSU yesterday @ ≈ 7:15 am +
got to Deadman Creek CG @ ≈ 2:00 pm
Drove thru S Tahoe to 395 which is
faster than going thru Placerville (+ saves $20)
bathrooms have running water but
have not found a spicket yet to fill up
to gal jugs.

- Working w/ Rebekah + Andrew
- left camp @ ≈ 7:00 @ Campground @
≈ 7:35

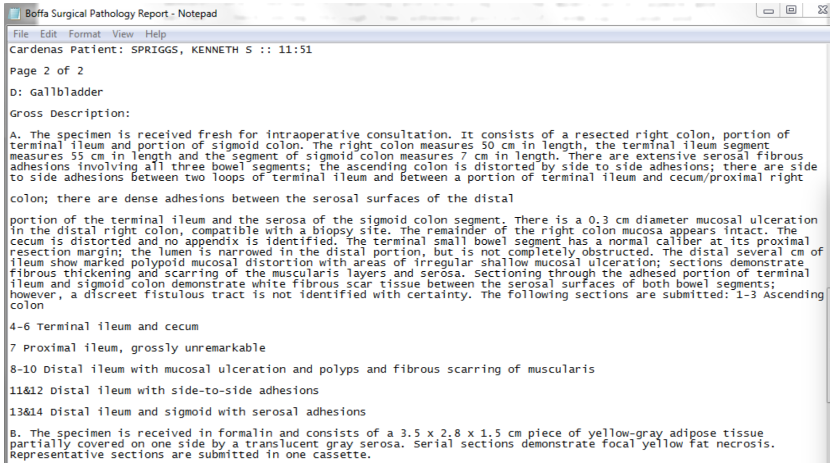
- Familiarize students w/ area + lithologies
+ then find SS outcrop + follow it
along strike looking for hinge line

- Set declination @ 14½°E



Original Image

Edge Image

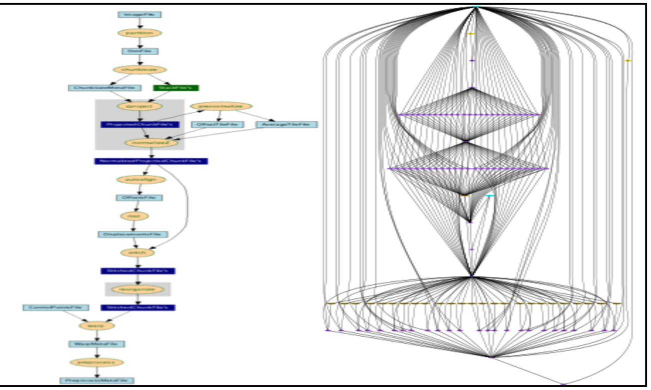


Boffa Surgical Pathology Report - Notepad
File Edit Format View Help

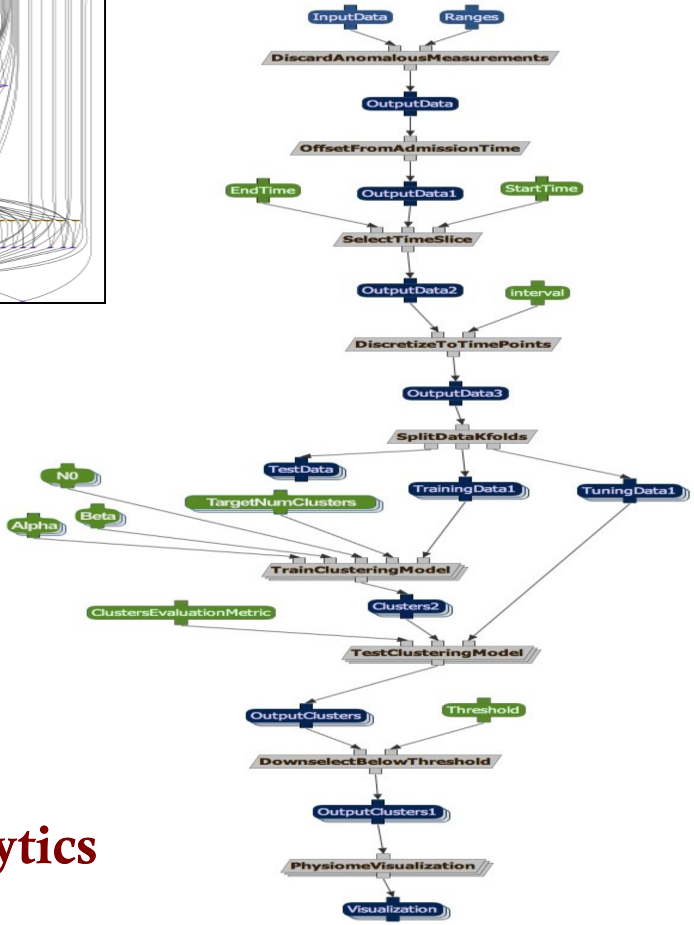
```
Cardenas Patient: SPRIGGS, KENNETH S :: 11:51

Page 2 of 2

D: Gallbladder

Gross Description:

A. The specimen is received fresh for intraoperative consultation. It consists of a resected right colon, portion of
terminal ileum and portion of sigmoid colon. The right colon measures 50 cm in length, the terminal ileum segment
measures 55 cm in length and the segment of sigmoid colon measures 7 cm in length. There are extensive serosal fibrous
adhesions involving all three bowel segments; the ascending colon is distorted by side to side adhesions; there are side
to side adhesions between two loops of terminal ileum and between a portion of terminal ileum and cecum/proximal right

colon; there are dense adhesions between the serosal surfaces of the distal

portion of the terminal ileum and the serosa of the sigmoid colon segment. There is a 0.3 cm diameter mucosal ulceration
in the distal right colon, compatible with a biopsy site. The remainder of the right colon mucosa appears intact. The
cecum is distorted and no appendix is identified. The terminal small bowel segment has a normal caliber at its proximal
resection margin; the lumen is narrowed in the distal portion, but is not completely obstructed. The distal several cm of
ileum show marked polypoid mucosal distortion with areas of irregular shallow mucosal ulceration; sections demonstrate
fibrous thickening and scarring of the muscularis layers and serosa. Sectioning through the adhesed portion of terminal
ileum and sigmoid colon demonstrate white fibrous scar tissue between the serosal surfaces of both bowel segments;
however, a discreet fistulous tract is not identified with certainty. The following sections are submitted: 1-3 Ascending
colon

4-6 Terminal ileum and cecum

7 Proximal ileum, grossly unremarkable

8-10 Distal ileum with mucosal ulceration and polyps and fibrous scarring of muscularis

11&12 Distal ileum with side-to-side adhesions

13&14 Distal ileum and sigmoid with serosal adhesions

B. The specimen is received in formalin and consists of a 3.5 x 2.8 x 1.5 cm piece of yellow-gray adipose tissue
partially covered on one side by a translucent gray serosa. Serial sections demonstrate focal yellow fat necrosis.
Representative sections are submitted in one cassette.
```

Experiment with Data Science in a Variety of Domains

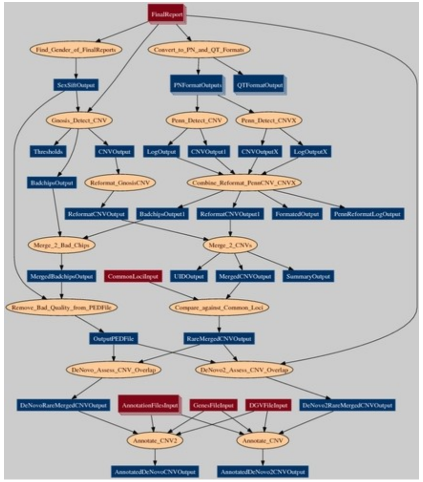
image processing

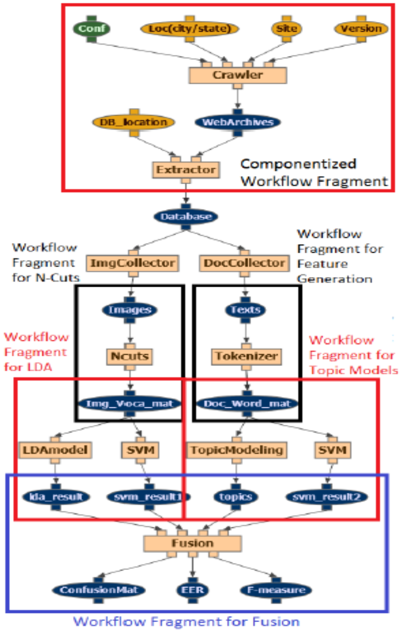


hydrology



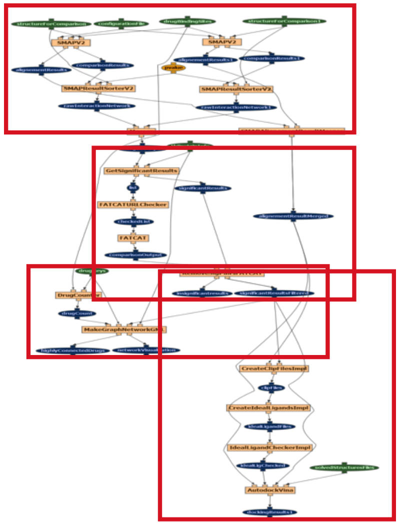
social network analysis





text analytics

biology

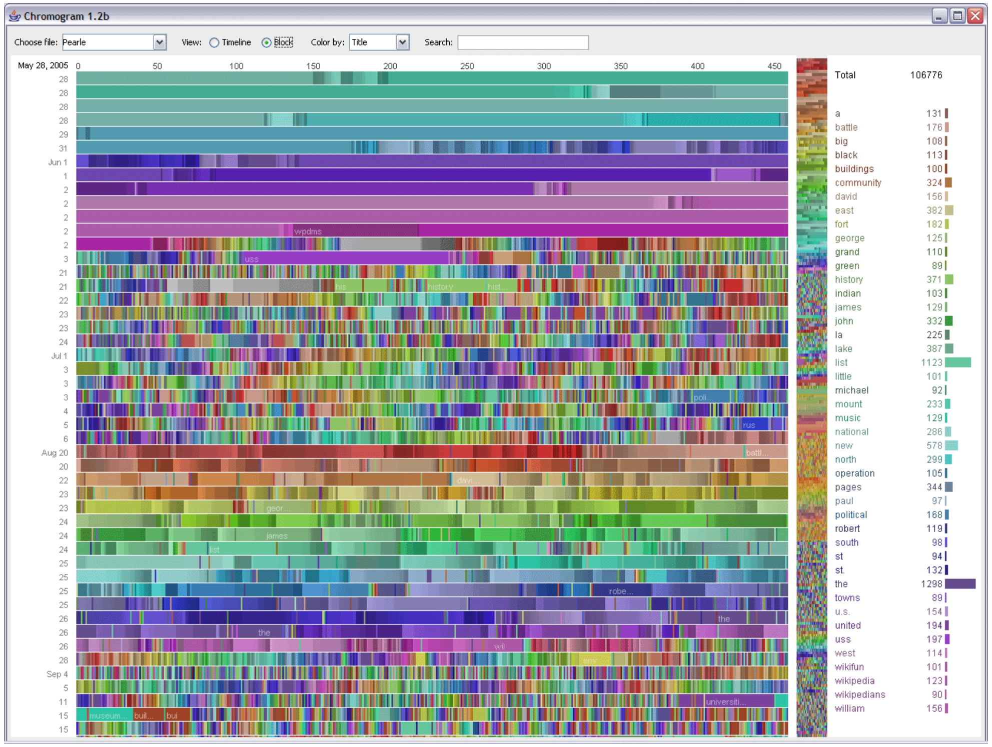


Section I: Introduction

1. Computational thinking and data science
2. Data
   - What is accessible data
   - Major types of data
   - Basic terminology
3. Data analysis software
   1. Algorithms vs code
   2. Programming languages
   3. Turing machines
4. Multi-step data analysis as workflows

http://blog.ginger.io/data-science-101/

II: Data Analysis



Choose file: Pearle

View: ○ Timeline ● Block

Color by: Title

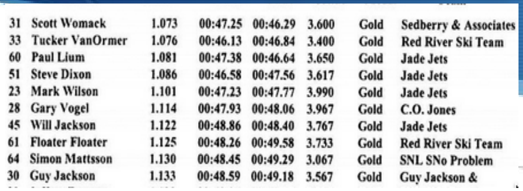
Search:

Total	106776
a	131
battle	176
big	108
black	113
buildings	100
community	324
david	156
east	382
fort	182
george	125
grand	110
green	89
history	371
indian	103
james	129
john	332
la	225
lake	387
list	1123
little	101
michael	92
mount	233
music	129
national	286
new	578
north	299
operation	105
pages	344
paul	97
political	168
robert	119
south	98
st	94
st.	132
the	1298
towns	89
u.s.	154
united	194
uss	197
west	114
wikifun	101
wikipedia	123
wikipedians	90
william	156

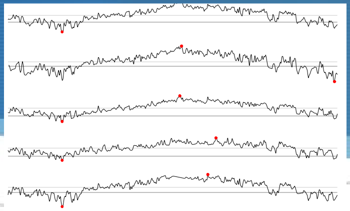
1. Data analysis tasks
   - Classification, clustering, pattern detection, causal discovery, simulation
2. Data pre-processing
3. Data visualization
4. Data lifecycle

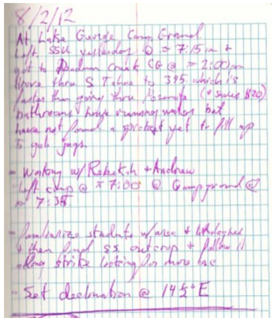
https://es.wikipedia.org/wiki/Big_data

III: Data Analysis in Practice



31	Scott Womack	1.073	00:47.25	00:46.29	3.600	Gold	Sedberry & Associates
33	Tucker VanOrmer	1.076	00:46.13	00:46.84	3.400	Gold	Red River Ski Team
60	Paul Lium	1.081	00:47.38	00:46.64	3.650	Gold	Jade Jets
51	Steve Dixon	1.086	00:46.58	00:47.56	3.617	Gold	Jade Jets
23	Mark Wilson	1.101	00:47.23	00:47.77	3.990	Gold	Jade Jets
28	Gary Vogel	1.114	00:47.93	00:48.06	3.967	Gold	C.O. Jones
45	Will Jackson	1.122	00:48.86	00:48.40	3.767	Gold	Jade Jets
61	Floater Floater	1.125	00:48.26	00:49.58	3.733	Gold	Red River Ski Team
64	Simon Mattsson	1.130	00:48.45	00:49.29	3.067	Gold	SNL SNo Problem
30	Guy Jackson	1.133	00:48.59	00:49.18	3.567	Gold	Guy Jackson &



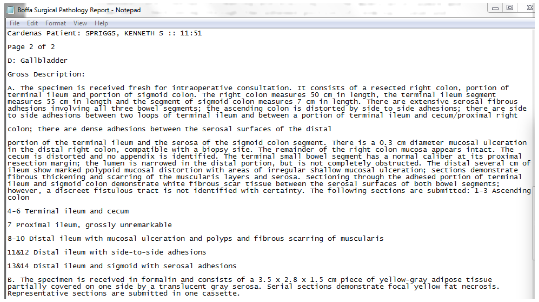


8/2/12
At Lake Garda Camp Ground
Left SSU yesterday @ ≈ 7:15 am &
got to Deadman Creek CG @ ≈ 2:00 pm
Drove thru S Tahoe to 395 which is
faster than going thru Yosemite (& saves $20)
bathrooms have running water but
have not found a spigot yet to fill up
to gal jugs.

- Working w/ Rakesh + Andrew
- Left camp @ ≈ 7:00 @ Campground @
≈ 7:35

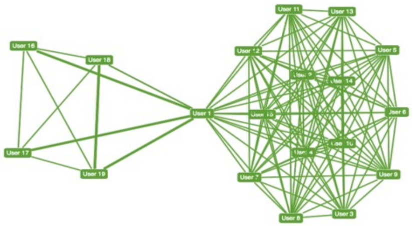
- familiarize students w/ area & lithologies
& then find SS outcrop & follow it
along strike looking for more [illegible]

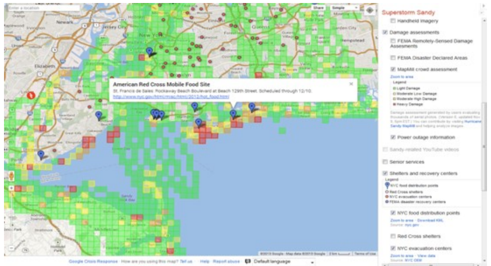
- Set declination @ 14½°E



Boffa Surgical Pathology Report - Notepad
File Edit Format View Help

```
Cardenas Patient: SPRIGGS, KENNETH S :: 11:51

Page 2 of 2

D: Gallbladder

Gross Description:

A. The specimen is received fresh for intraoperative consultation. It consists of a resected right colon, portion of
terminal ileum and portion of sigmoid colon. The right colon measures 50 cm in length, the terminal ileum segment
measures 55 cm in length and the segment of sigmoid colon measures 7 cm in length. There are extensive serosal fibrous
adhesions involving all three bowel segments; the ascending colon is distorted by side to side adhesions; there are side
to side adhesions between two loops of terminal ileum and between a portion of terminal ileum and cecum/proximal right

colon; there are dense adhesions between the serosal surfaces of the distal

portion of the terminal ileum and the serosa of the sigmoid colon segment. There is a 0.3 cm diameter mucosal ulceration
in the distal right colon, compatible with a biopsy site. The remainder of the right colon mucosa appears intact. The
cecum is distorted and no appendix is identified. The terminal small bowel segment has a normal caliber at its proximal
resection margin; the lumen is narrowed in the distal portion, but is not completely obstructed. The distal several cm of
ileum show marked polypoid mucosal distortion with areas of irregular shallow mucosal ulceration; sections demonstrate
fibrous thickening and scarring of the muscularis layers and serosa. Sectioning through the adhesed portion of terminal
ileum and sigmoid colon demonstrate white fibrous scar tissue between the serosal surfaces of both bowel segments;
however, a discreet fistulous tract is not identified with certainty. The following sections are submitted: 1-3 Ascending
colon

4-6 Terminal ileum and cecum

7 Proximal ileum, grossly unremarkable

8-10 Distal ileum with mucosal ulceration and polyps and fibrous scarring of muscularis

11&12 Distal ileum with side-to-side adhesions

13&14 Distal ileum and sigmoid with serosal adhesions

B. The specimen is received in formalin and consists of a 3.5 x 2.8 x 1.5 cm piece of yellow-gray adipose tissue
partially covered on one side by a translucent gray serosa. Serial sections demonstrate focal yellow fat necrosis.
Representative sections are submitted in one cassette.
```

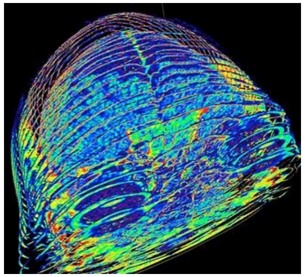




American Red Cross Mobile Food Site
St. Francis de Sales: Rockaway Beach Boulevard at Beach 129th Street. Scheduled through 12/10.
http://www.nyc.gov/html/misc/html/2012/hot_food.html

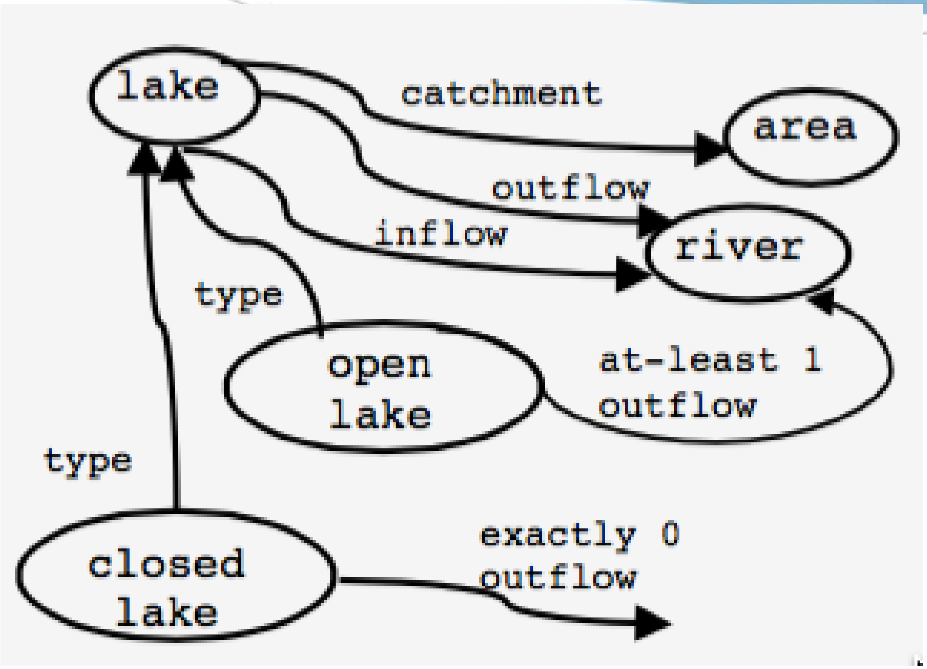
Superstorm Sandy
☐ Handheld imagery
☑ Damage assessments
☐ FEMA Remotely-Sensed Damage Assesments
☐ FEMA Disaster Declared Areas
☑ MapMill crowd assessment
Zoom to area
Legend
🟩 Light Damage
🟨 Moderate Low Damage
🟧 Moderate High Damage
🟥 Heavy Damage
Damage assessment generated by users evaluating thousands of aerial photos. [illegible] Hurricane Sandy MapMill and helping analyze images.
☑ Power outage information
☐ Sandy-related YouTube videos
☐ Senior services
☑ Shelters and recovery centers
Legend
NYC food distribution points
○ Red Cross shelters
● NYC evacuation centers
● FEMA disaster recovery centers
☑ NYC food distribution points
Zoom to area · Download KML
Source: nyc.gov
☐ Red Cross shelters
☑ NYC evacuation centers
Zoom to area · View data
Source: NYC OEM

©2013 Google · Map data ©2013 Google
Terms of Use
Google Crisis Response · How are you using this map? Tell us · Help · Report abuse
Default language



1. Analyzing different kinds of data
2. Parallel and distributed computing
   - Multi-core, clusters, grids, web services, …
   - Speedup, dependencies, critical paths, Amdahl’s law, MapReduce

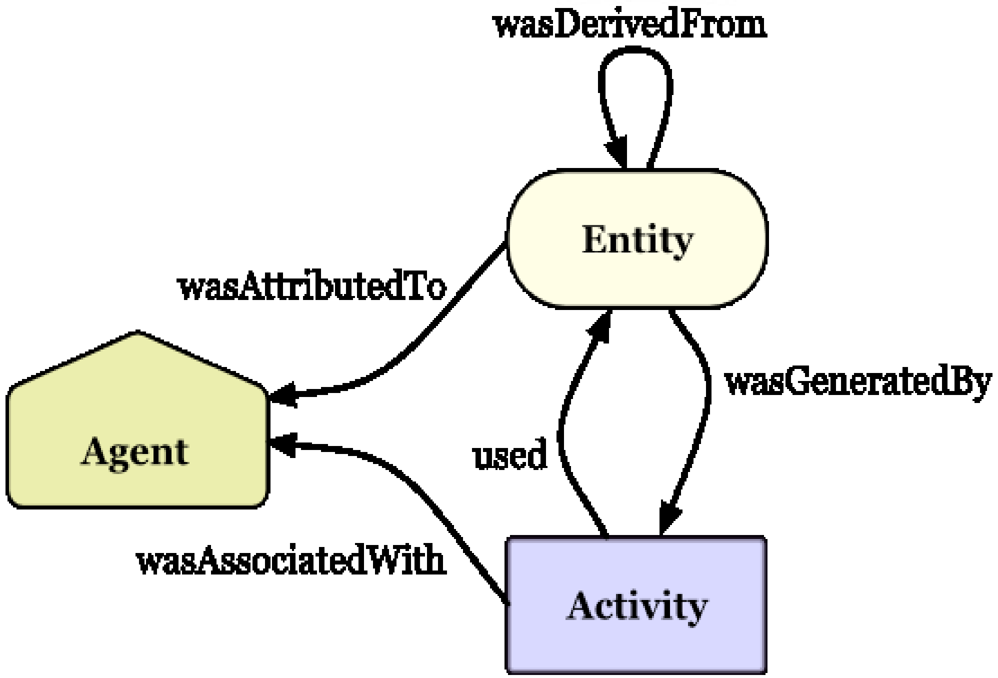
IV: Metadata



protégé

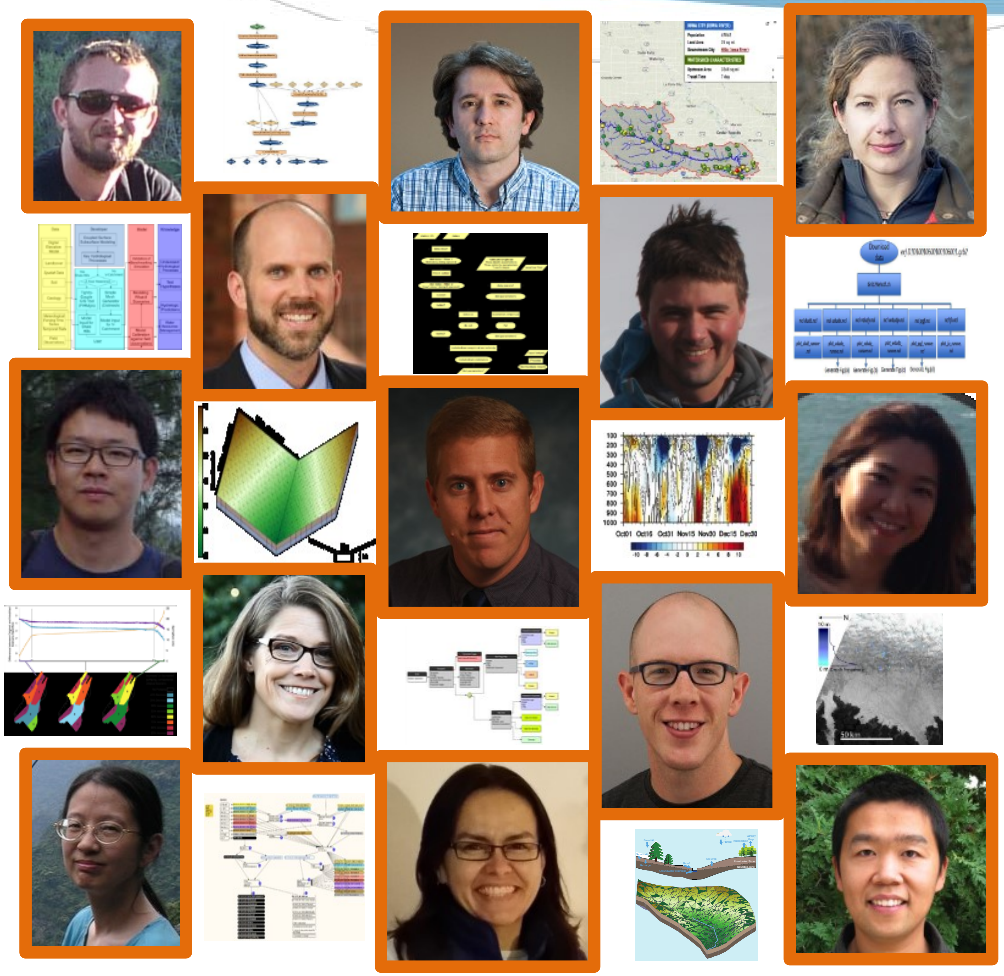
1. Semantic metadata
   - The DC standard
2. Ontologies
3. Provenance

V: Data Dissemination



1. Data formats and standards
2. Provenance
3. Data stewardship
   - Data identifiers, data citation

VI: Advanced Topics



- Privacy and sensitive data
- Introduction to databases
- Crowdsourcing data collection
- Multi-disciplinary collaboration
- Project management

Course Design

Focus on AI Topics

- No statistics
- No databases
- No programming skills taught
- Yes: scalable algorithms

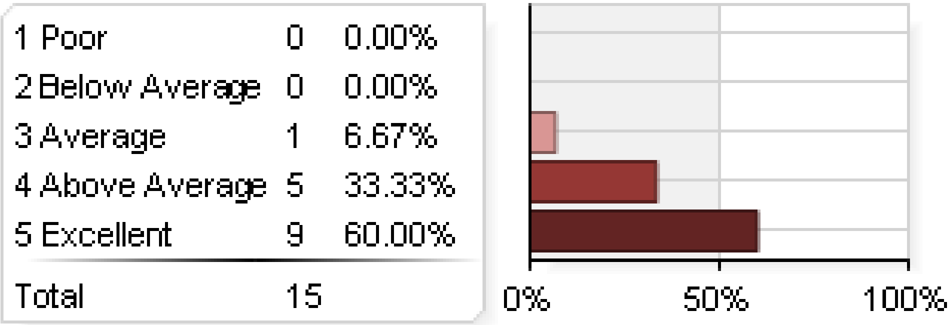
Constant Practice

- Class group activities followed by group reports
  - Learn to communicate and use technical terms
- Homeworks emphasize hypothesis formulation and testing

INF549 for USC Informatics Student Comments

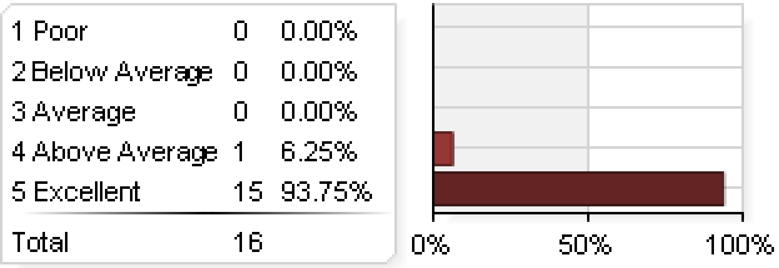
- *"I went to a big data and big analytics keynote by a SAS VP. The presenter used a lot of jargon vocabulary that we have seen in class, and it was very interesting to see the application of many of the class topics. I understood the talk!"*
- *"I attended a Big Data student poster session in Engineering, and I could understand the presentations!"*

The prerequisites for the course were adequate.



1 Poor	0	0.00%
2 Below Average	0	0.00%
3 Average	1	6.67%
4 Above Average	5	33.33%
5 Excellent	9	60.00%
Total	15	

Encouraged students to participate in their learning (e.g., through discussion, projects, study groups and other appropriate activities).



1 Poor	0	0.00%
2 Below Average	0	0.00%
3 Average	0	0.00%
4 Above Average	1	6.25%
5 Excellent	15	93.75%
Total	16	

Ongoing

- Making materials available at datascience4all.org
  - Include videolectures
- Topical tutorials at science meetings (EarthCube, NOAA)
  - Eg, ontologies, machine learning
- North American Summer School in Data Science (with Caltech)
  - Already used for the 2016 RDA Summer School in Data Science
  - Already used in 2016 IS-GEO Summer School

ADDITIONAL SLIDES

Section I: Introduction to Basic Concepts

1. Computational thinking and data science

2. Data
   - Accessible data
     - APIs, license,…
   - Major types of data
     - Networks, text, time series, geospatial,…
   - Data terms
     - Metadata, silos, sensitive data, big data, …

3. Data analysis software
   - Algorithms vs code
     - Algorithm design
   - Programming languages
   - Encapsulation
   - Turing machines, Turing-complete languages
4. Multi-step data analysis as workflows
   - Components, dataflow, intermediate data
   - **Practicum: WINGS**

Section II: Data Analysis

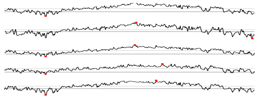
1. Data analysis tasks
   - Classification
     - Decision trees
     - Alternative methods
     - Accuracy and other metrics
   - Pattern detection
     - Clustering
     - Temporal patterns
     - Network patterns
   - Causal discovery
     - Graphical models
     - Bayesian networks
   - Simulation

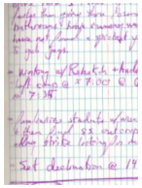
2. Data lifecycle
   - Data pre-processing
     - Data cleaning
   - Data wrangling
   - ETL
   - Collect, clean, analyze, visualize, deposit

3. Data visualization
   - Time series
   - Statistical
   - Maps, cartograms
   - Treemaps, heatmaps
   - Networks
   - Visual analytics

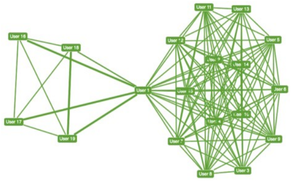
Section III: Data Analysis in Practice

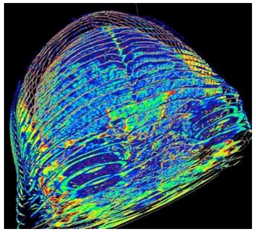
1. Analyzing different kinds of data
   - Time series data
     - Ecology
     - Medicine
   - Text data
     - Web
     - Archives
   - Network data
     - Social media
     - Web
   - Multimedia data
     - Images
     - Videos
   - Geospatial data

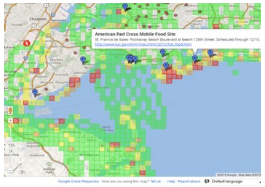




- Walking w/ Rebekah + And
- Left camp @ ≈ 7:00 @
- ~ 7:35
- Set declination @ 14

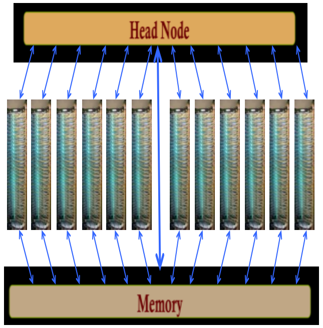






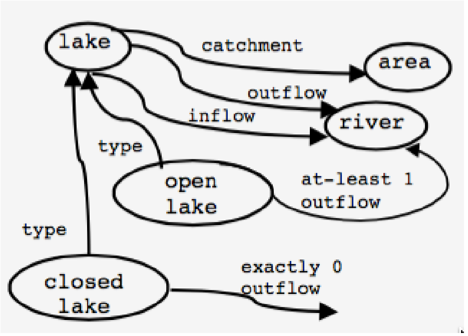
2. Parallel computing
   - Speedup estimates
   - Dependencies
   - Critical paths
   - Amdahl's law
   - MapReduce

3. Distributed computing
   - Multi-core computing, chips
   - Clusters
   - Grids
   - Web services
   - Cloud computing



Section IV: Metadata

1. Semantic metadata
   - Attribution metadata
     - The Dublin Core standard
   - Summary metadata
   - Provenance metadata
   - Metadata in workflows
     - Metadata capture
     - Metadata propagation



2. Ontologies
   - Taxonomies
   - Classes
   - Properties
   - Assertions
   - Definitions
   - Constraints and rules
   - Reasoning
   - **Practicum: PROTÉGÉ**

3. Provenance
   - Process provenance
   - Attribution provenance
   - Resource provenance
   - Provenance standards

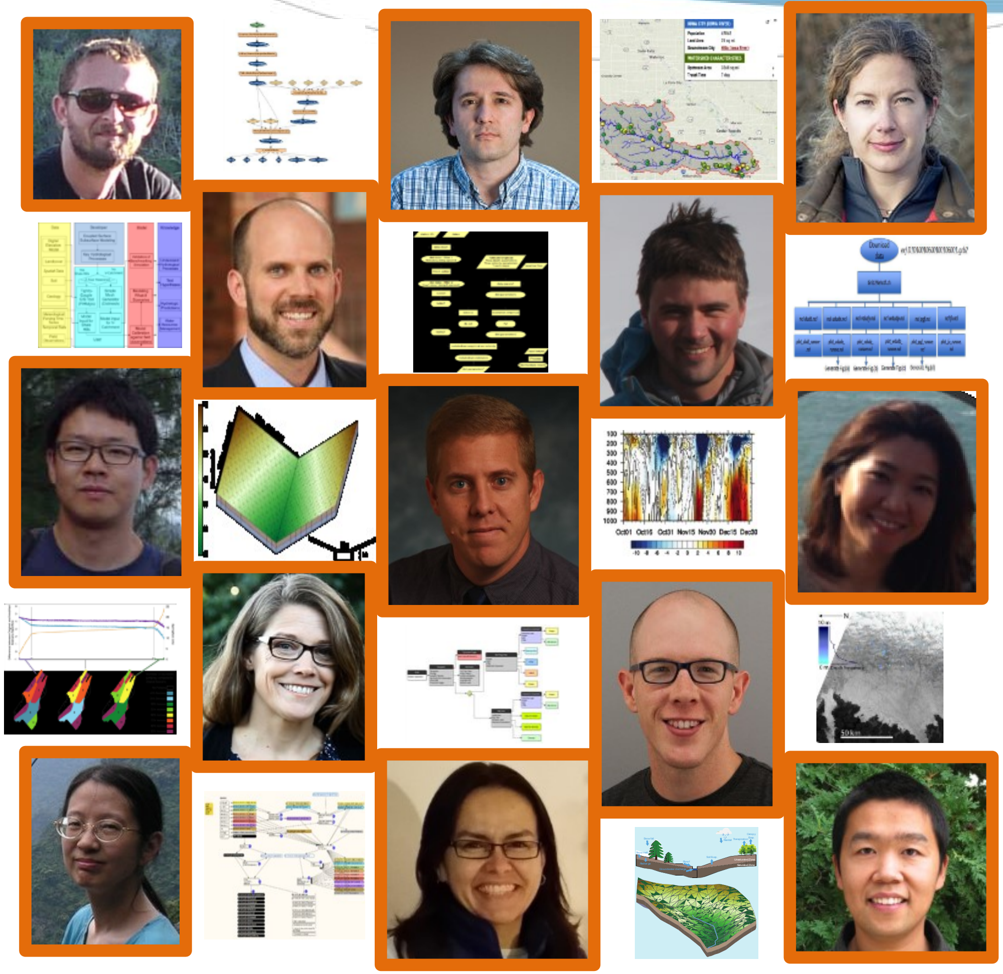
Section V: Data Dissemination

1. Data formats
   - Data standards
   - Data repositories
   - Data services
   - Web data

2. Combining metadata and provenance
   - Metadata propagation
   - Automatic method validation
   - Automatic generation of metadata
   - Automatic provenance tracking

3. Data stewardship
   - Data sharing
   - Data identifiers
   - Licenses for data
   - Data citation and attribution
   - Software publication

VI: Advanced Topics



- Privacy and sensitive data
- Introduction to databases
- Crowdsourcing data collection
- Multi-disciplinary collaboration
- Project management