DATA SET TYPES

SYRACUSE UNIVERSITY
School of Information Studies

STUDY GUIDE: KEY CONCEPTS

Make sure you understand the following key concepts by the end of Week 2:

- Data set types
  - Records, transactions, images, sequences, audios
- Variable types
  - Nominal or categorical, ordinal, numeric (interval and ratio)
- Data quality issues
  - Outliers, missing values, duplicate data
- Data summary and visualization
- Data transformation

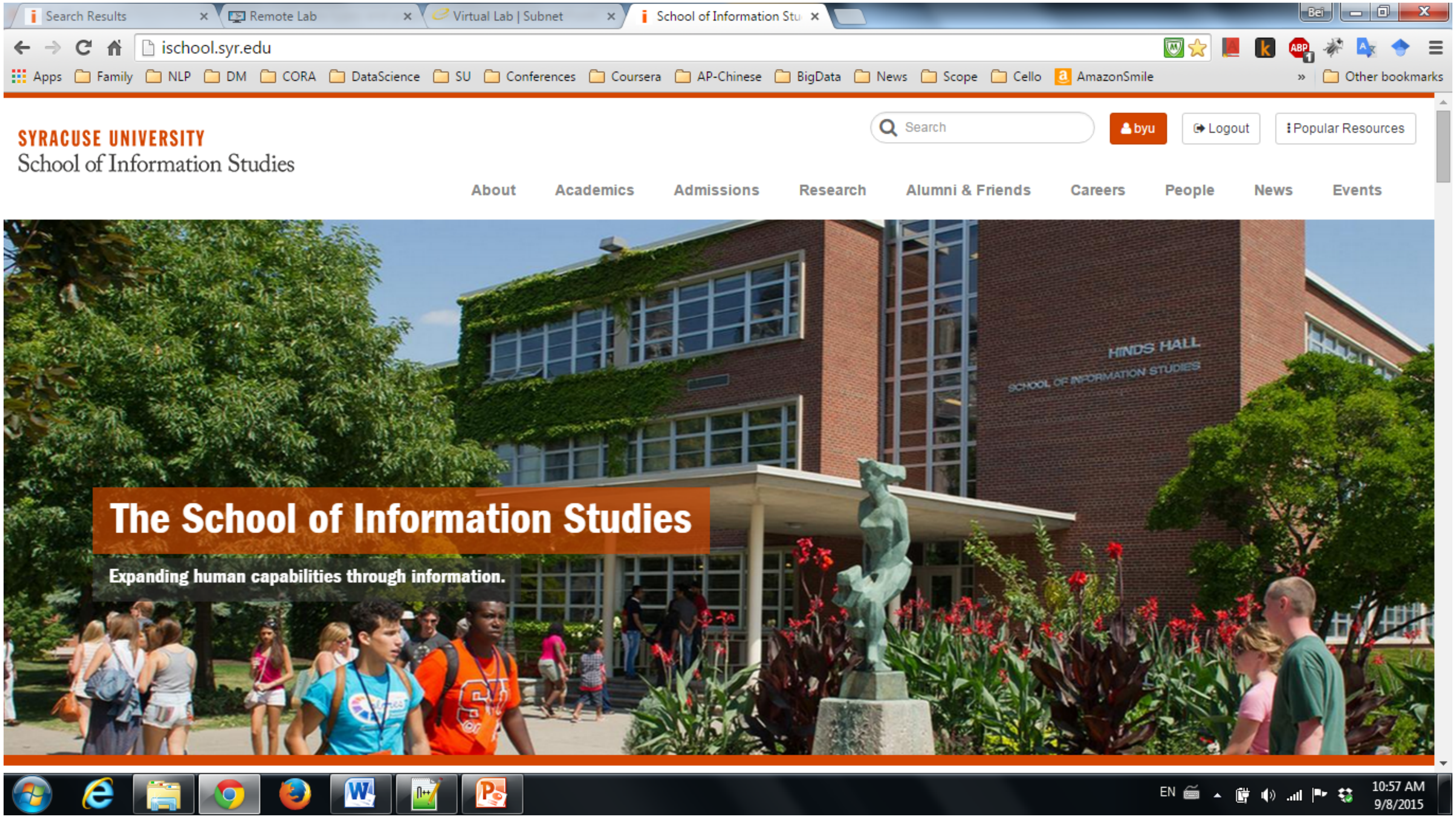
DATA SET TYPES

Record data: Data in the tabular format

- Each row is a data example.
- Each column is an attribute.
- Most common type of data set.

NAME	HIGHEST DEGREE	AGE	BLOOD TYPE
Jane	Middle School	25	A
John	High School	30	B
Amy	College	34	O
Larry	Grad School	31	AB

NONRECORD DATA



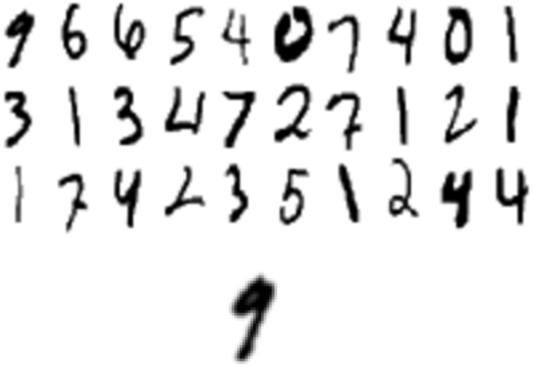
NONRECORD DATA: TEXT DOCUMENTS

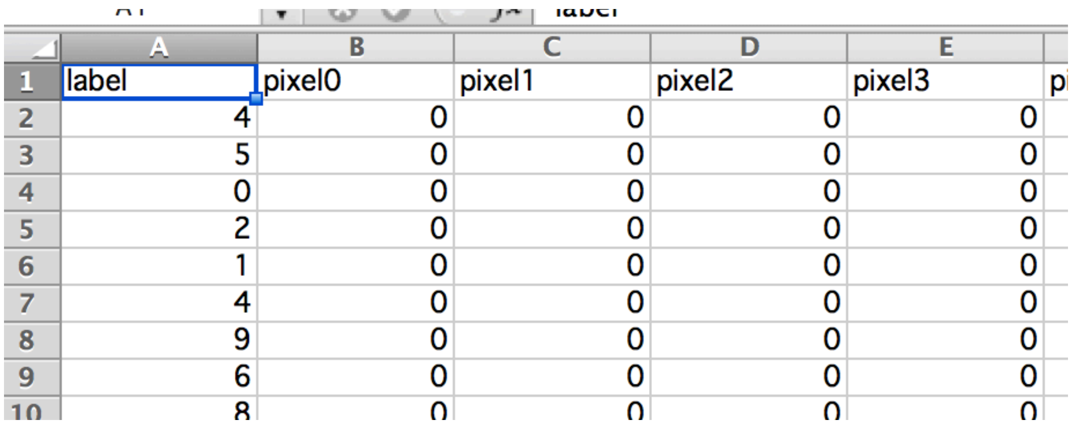
Some data sets are not born as record data but can be converted to record format.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

IMAGE DATA

https://www.kaggle.com/c/digit-recognizer

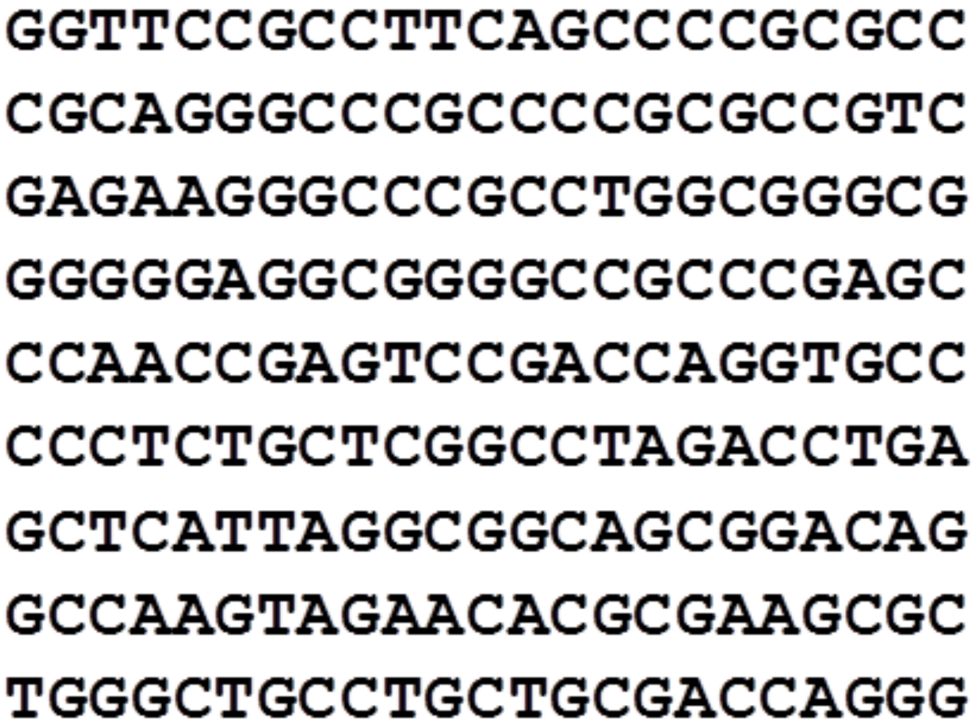




	A	B	C	D	E	
1	label	pixel0	pixel1	pixel2	pixel3	pi
2	4	0	0	0	0	
3	5	0	0	0	0	
4	0	0	0	0	0	
5	2	0	0	0	0	
6	1	0	0	0	0	
7	4	0	0	0	0	
8	9	0	0	0	0	
9	6	0	0	0	0	
10	8	0	0	0	0	

Each image is 28*28 pixels = 784 total.
Each pixel has a single pixel value [0, 255] associated with it, indicating the lightness or darkness of that pixel, with higher numbers meaning darker.

SEQUENCE DATA



```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

PLAGIARISM DETECTION

Edit distance: The minimum number of steps needed to transform one sequence to the other

- E.g., to transform “ABCD” to “ABCE,” one step is needed to transform “D” to “E.”

The algorithms used for comparing genomic sequences were used to detect plagiarism (e.g., turnitin.com) by replacing the nucleotides A, T, C, and G with words in text documents.

TRANSACTION DATA

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

TRANSACTION DATA

TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

Converted to record data

TID	Bread?	Coke?	Milk?	Diaper?	Beer?
1	1	1	1	0	0
2	1	0	0	0	1
3	0	1	1	1	1
4	1	0	1	1	1
5	0	1	1	1	0

SPARSE MATRIX

Most values in the matrix are “0”

Too many columns

Too few with nonzero values

TID	Bread?	Coke?	Milk?	Diaper?	Beer?
1	1	1	1	0	0
2	1	0	0	0	1
3	0	1	1	1	1
4	1	0	1	1	1
5	0	1	1	1	0

STORAGE OF SPARSE MATRIX

TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

Requires less space

TID	Bread?	Coke?	Milk?	Diaper?	Beer?
1	1	1	1	0	0
2	1	0	0	0	1
3	0	1	1	1	1
4	1	0	1	1	1
5	0	1	1	1	0

Requires more space

NETWORK DATA

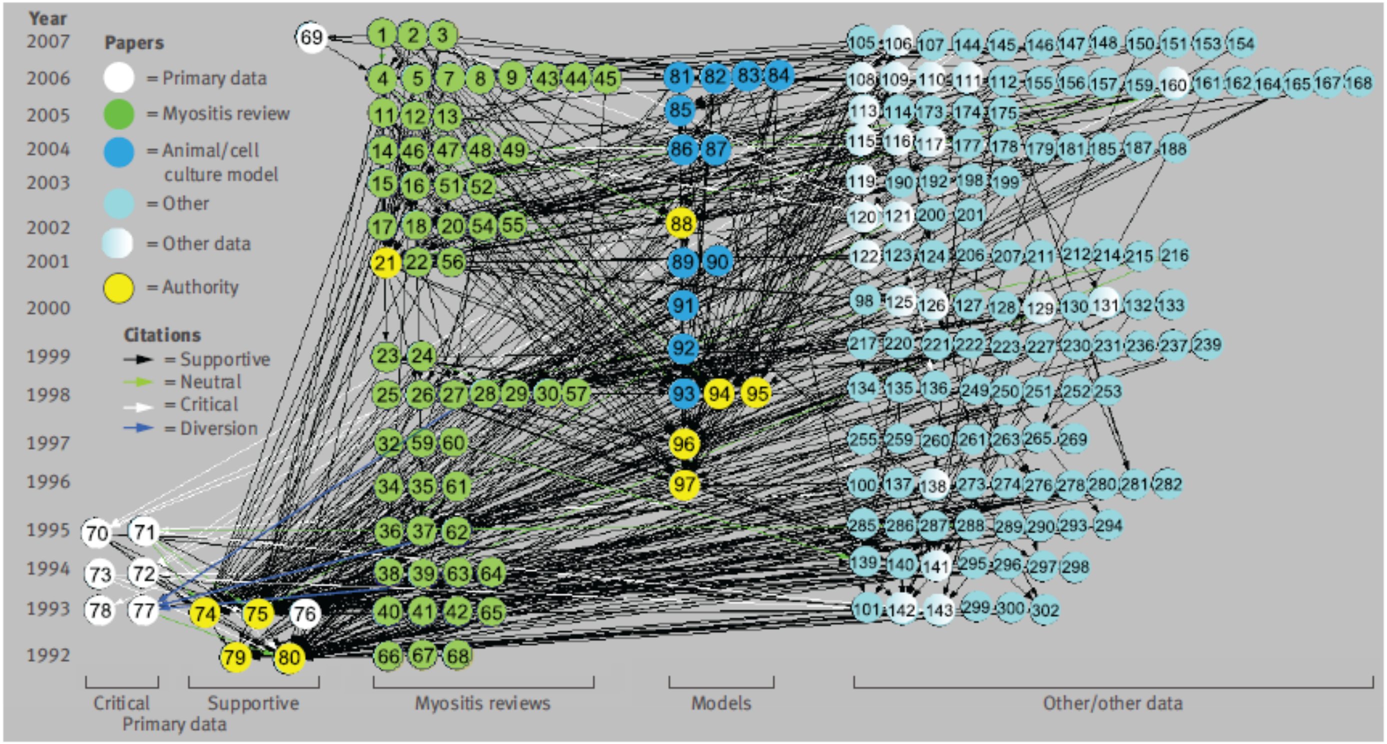


Fig 1 | Claim specific citation network. Citations regarding claim that β amyloid precursor protein mRNA or protein, or β amyloid protein, is abnormally present in inclusion body myositis muscle. The network is organised according to paper category and year of publication. Authority status (yellow) was defined computationally by network theory. Many citations flow to supportive primary data but not critical data. Papers are represented as nodes (n=218) and citations as directed edges (supportive n=636, neutral n=18, critical n=21, diversion n=3). Twenty four papers contain statements pertaining to claim but do not make or receive citations about it (not shown).

Steven G. Greenberg (2009). How citation distortions create unfound authority: analysis of citation network. BMJ 2009; 339: b2680 http://dx.doi.org/10.1136/bmj.b2680

REVIEW OF DATA SET TYPES

Record data

Nonrecord data

- Text data
- Image data
- Sequence data
- Transaction data
- Network data