Welcome to Big Data Analytics

https://tinyurl.com/cis545-lecture-01-12-22

Susan B. Davidson and Zachary G. Ives
University of Pennsylvania

CIS 545: Big Data Analytics

Portions of this lecture were developed for the OpenDS4All project, piloted by Penn, IBM, and the Linux Foundation

January 12, 2022

Welcome to CIS 545, Offered in Hybrid Mode

- Current plan: until 1/24 this and all courses will be remote; thereafter, Section 1 will be in-person and Section 2 will be remote
- We have support for remote asynchronous (prerecorded videos), remote synchronous (live stream), and in person participation!
  - Content at www.cis.upenn.edu/~cis545
  - Self-check quizzes are in Canvas, as will be recorded synchronous meetings
  - We'll count many different things (Piazza, in class questions, online quizzes) towards your participation score
- Zoom livestream: https://upenn.zoom.us/j/99014422529?pwd=R2ZYU0FXL3EvSDh1N3puc0R5bDErUT09

  (When we go hybrid, if you are in the classroom – *please* do not join audio!!!)

Data Is Driving Everything

1. Modern data acquisition is inexpensive!
   - Smartphones, embedded systems, inexpensive sensors,
   - Medical devices, simulators, …
2. Data storage is inexpensive!
3. Parallel (compute cluster) computation is inexpensive
   - The Cloud, clusters of computers, GPUs, tensor processors, …

Science only has explanatory and predictive models in a few (mostly physical sciences-related) domains

… So: can we use **algorithms** + **data** to understand phenomena? Build or augment **models**? Build **detectors**? Make **diagnoses**?

Data Is Driving Everything

"Big data"
"Data science"
"Data lakes"
"Visual analytics"

"Deep learning"
"Statistical analysis"
"Biomedical informatics"
"Business analytics"

Lots of trends in pursuit of the same goals!
Discovery, models, decision-making, …

Also, new issues -
"Ethical algorithms"
"Reproducibility"

The Key Question in Big Data Analytics: How Do We *Understand* and *Predict*?

Much of science and engineering derives from **physics**, which is "model-first"

- Newton's laws, the theory of relativity, optics, how materials react under stress, etc.
- Here, the basis of prediction – even with stochastic processes – tends to be **simulation**
  - Weather forecasting, simulating water in the movie Moana, etc.

We want predictions where we don't have good models

e.g., behavior, biology, the brain, whether a product will be a success, what to invest in

- We need to use **sampling, statistics, data-first** approaches
- The big data revolution is mostly about how to acquire and handle **enough** data, and ask the right questions, for these models to be useful!

- Of course, in the real world we often want to combine models and data!

Outline for Today

More on Big Data – why it's hard to:

- Capture and model data, and get into into an analyzable form
- Visualize, understand, and analyze data

Course logistics, goals, and focus

The Need to Model, Clean, and Integrate Data

Susan B. Davidson and Zachary G. Ives
University of Pennsylvania
CIS 545: Big Data Analytics

Portions of this lecture were developed for the OpenDS4All project, piloted by Penn, IBM, and the Linux Foundation

Our Focus in Big Data Analytics

Our goal: understand how to manage and create predictive models from big data

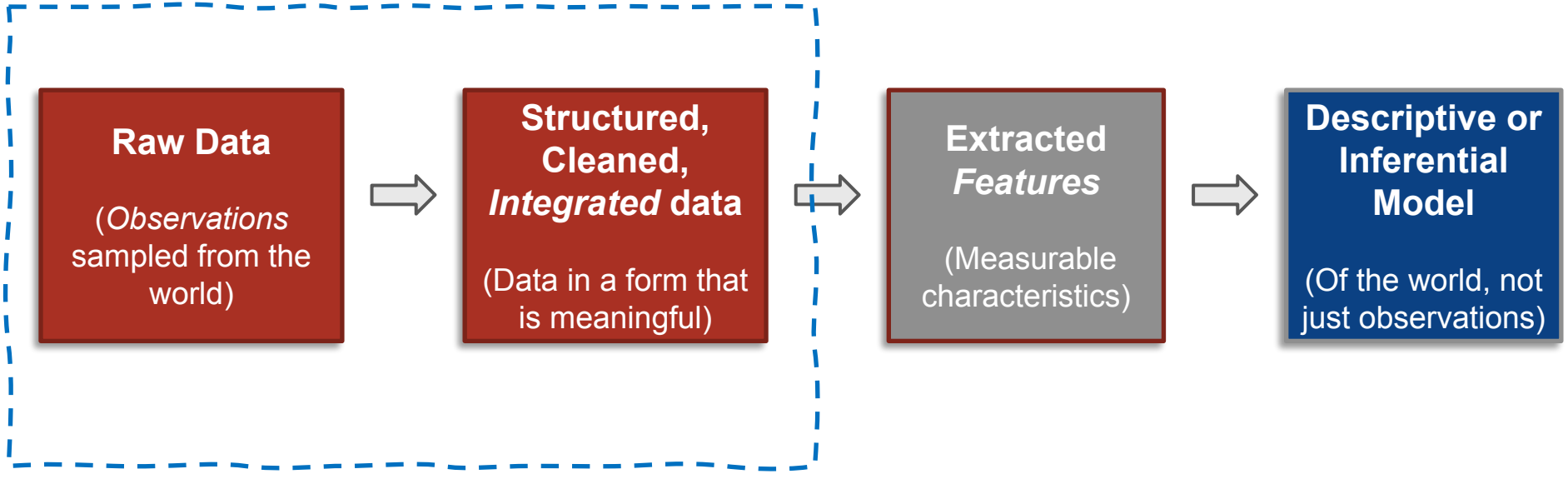
machine learning, distributed computing, distributed algorithms, data ethics, …

You'll understand foundations and principles of data management, machine learning, cloud systems, and more

- You'll be able to do a lot, and be equipped to rapidly learn even more!

Two main aspects: **getting data into the right form** and **building models…**

The Journey from Data to Models



Part I

Part II

What Makes Data "Big"?

A thousand rows? A GB? A PB? A YB?

A Million records? Billion records? Trillion records?

No consensus definition, but from our perspective:

- Too **complex for a human** to understand directly
- Doesn't fit into a **single computer's memory**
- Needs more than **brute force** algorithms to analyze

- May require **multiple computers** to work together to process
- May have **many dimensions**
- May be changing rapidly (high **velocity**)

Even at Scale, Raw Data Is Hard to Interpret

- If you're given: 66 105 102 – what does it mean?

Interpretability Is Helped by *Metadata*

- If you're given: 66 105 103 – what does it mean?
- What if I tell you these are character codes?

B i g

"Raw" data is less useful than data with information about how to interpret it!

65	A	97	a
66	B	98	b
67	C	99	c
68	D	100	d
69	E	101	e
70	F	102	f
71	G	103	g
72	H	104	h
73	I	105	i
74	J	106	j
75	K	107	k
76	L	108	l
77	M	109	m
78	N	110	n
79	O	111	o
80	P	112	p
81	Q	113	q
82	R	114	r
83	S	115	s
84	T	116	t
85	U	117	u
86	V	118	v
87	W	119	w
88	X	120	x
89	Y	121	y
90	Z	122	z

Interpretability Is Helped by *Metadata*

Smithson	1829
Adams	1848
Morgan	1928

What is this a list of?

Interpretability Is Helped by *Metadata*

Notable figures in Washington DC:

Name	Died
Smithson	1829
Adams	1848
Morgan	1928

What is this a list of?

The names and years of death for people whose names appeared on Washington DC buildings…

Smithsonian Castle, John Quincy Adams Elementary School, Thomas P Morgan Elementary School (now closed)

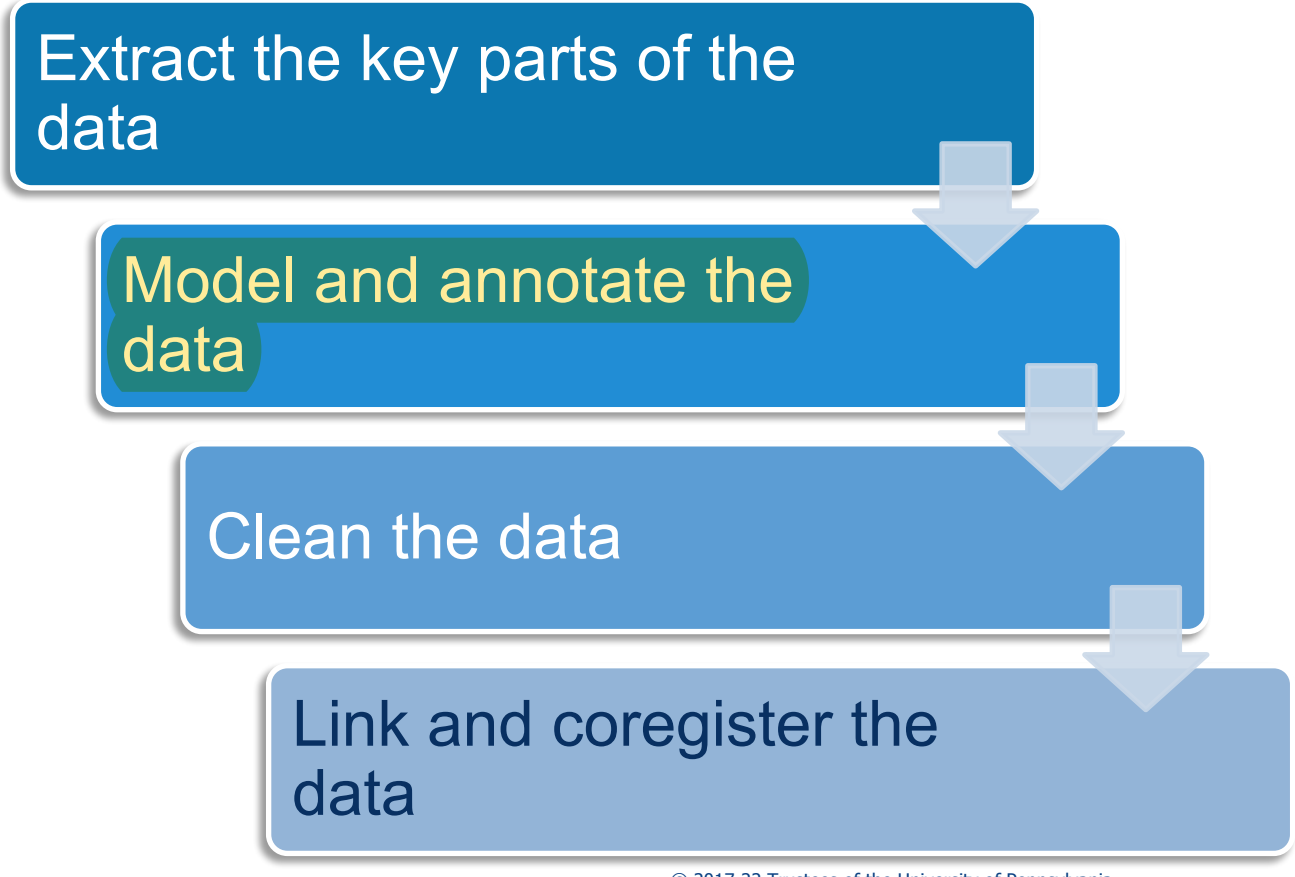
We Want to Know Structure, Meaning, and *Provenance*

Knowing these are names + dates helps interpret the data

Also important *why these names and dates* were collected!

And useful to know *who* collected and provided the data (and how)!

Data (Pre)processing Tasks



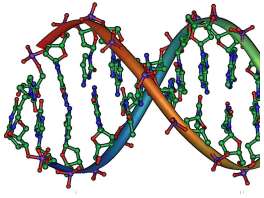
Many Data *Modalities* Contain Structured Data

Do not be like the cat who wanted a fish but was afraid to get his paws wet.

William Shakespeare

Text

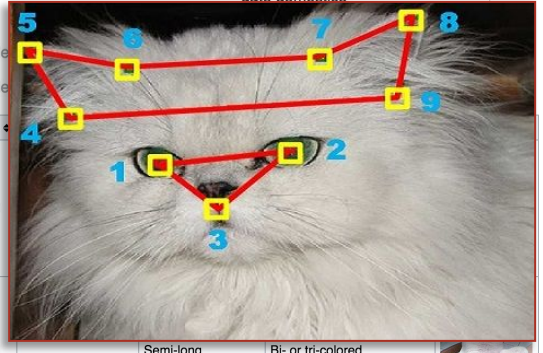
Images



Genes

Extract structured data from raw

Language	Label	Description	Also known as
English	cat	domesticated species of feline	housecat Felis silvestris catus Felis catus house cat domestic cat
Spanish	gato	especie doméstica de felino	Felis catus minino gato doméstico
Traditional Chinese	貓	No description de	
Chinese	猫	No description de	



Breed	Country	Origin				
Abyssinian	Ethiopia	Natural/Standard				
Aegean	Greece	Natural/Standard		Semi-long	Bi- or tri-colored	
American Curl	United States	Mutation		Short/Long	All	
American Bobtail	United States	Mutation		Short/Long	All	

Data often Needs to be Linked

New York Taxi Data

passenger_count	trip_distance	pickup_longitude	pickup_latitude	RateCodeID	store_and_fwd_flag	dropoff_longitude	dropoff_latitude	fare_am
1	15.1	-73.98356628417969	40.749881744384766	2	N	-73.80455780029297	40.649757385253906	
1	2.9	-74.00045776367188	40.727294921875	1	N	-73.97980499267578	40.761722564697266	
1	1	-73.99409484863281	40.74161148071289	1	N	-74.00382232666016	40.73116683959961	
2	1.2	-73.9753646850586	40.78733444213867	1	N	-73.9654541015625	40.80290603637695	
1	3.6	-74.00634002685547	40.73313522338867	1	N	-73.9741439819336	40.759849548339844	
1	4.8	-73.97996520996094	40.73479461669922	1	N	-74.01618957519531	40.711151123046875	

Reverse Geocode Data

Geocoder.ca Services | Products Terms Login Create Account

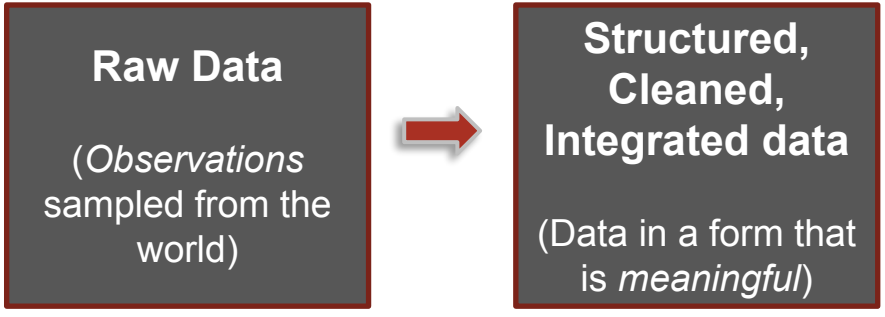
NEW YORK, NY » 393 5TH AVE , NEW YORK, NY » 10016-3325 » 40.74988,-73.98357

This is your location (40.74988, -73.98357) Directions
This is the nearest street address. [393 5TH AVE NEW YORK NY, 40.749848, -73.983532].
Neighborhoods:

1. Tenderloin,New York New York US
2. Tudor City,NY
3. New York,NY

Street View

The Story So Far: Big Data Must be *Cleaned & Integrated*



We seldom have all of our "big data" directly available

- Pieces in different systems or organizations, across documents, databases, APIs, etc.

Requires *extraction, modeling, cleaning, and linking*

May also do *exploratory data analysis* and build descriptive models over the data

Part I of the course will focus on this, at scale!

Quiz 01B on Canvas, linked to the Course schedule.

https://canvas.upenn.edu/courses/1636888/quizzes/2771598

Brief self-check

Which of the following is least likely to be a good example of big data?

A. Points on the trajectory of a falling object

B. Purchase behavior for 1000 users

C. 1000 YouTube videos

D. 100s of gene sequences

Which of the following tasks is not part of *preprocessing* data?

A. Clean the data

B. Model and annotate the data

C. Extract the key parts of the data

D. Run machine learning algorithms on the data

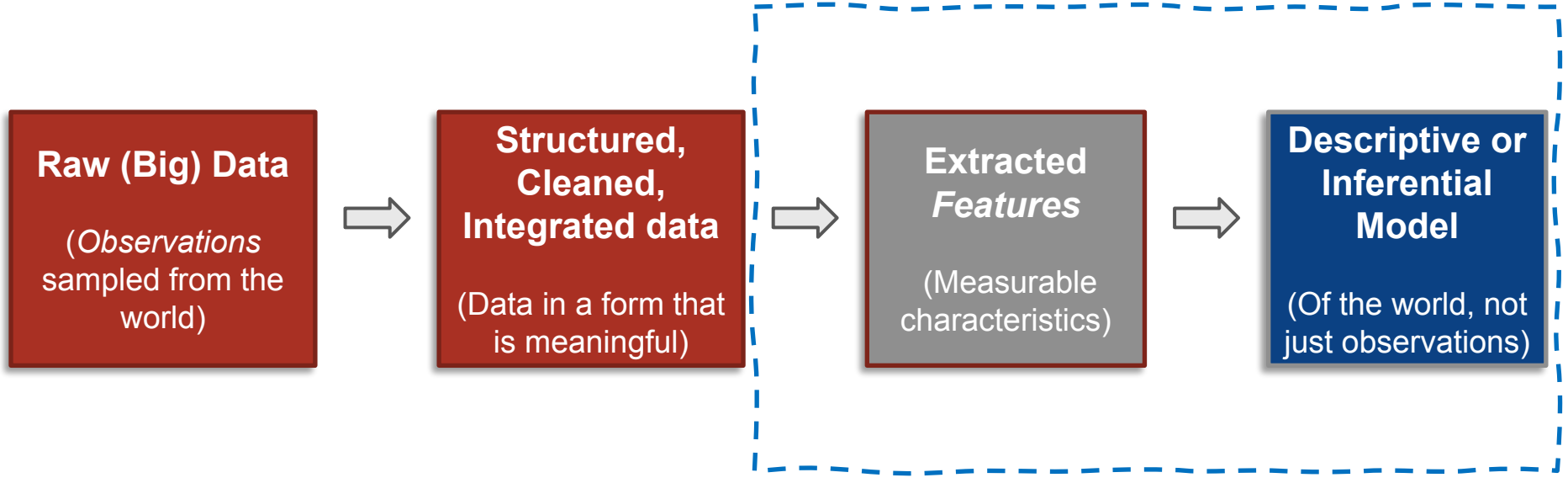
(transition)

Features and Learning

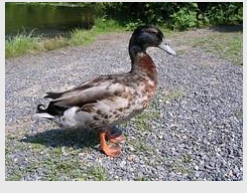
Susan B. Davidson and Zachary G. Ives
University of Pennsylvania
CIS 545: Big Data Analytics

Portions of this lecture were developed for the OpenDS4All project, piloted by Penn, IBM, and the Linux Foundation

From Data to *Features* for Modeling



Models Predict Relationships between *Features* and *Classes* or Groups

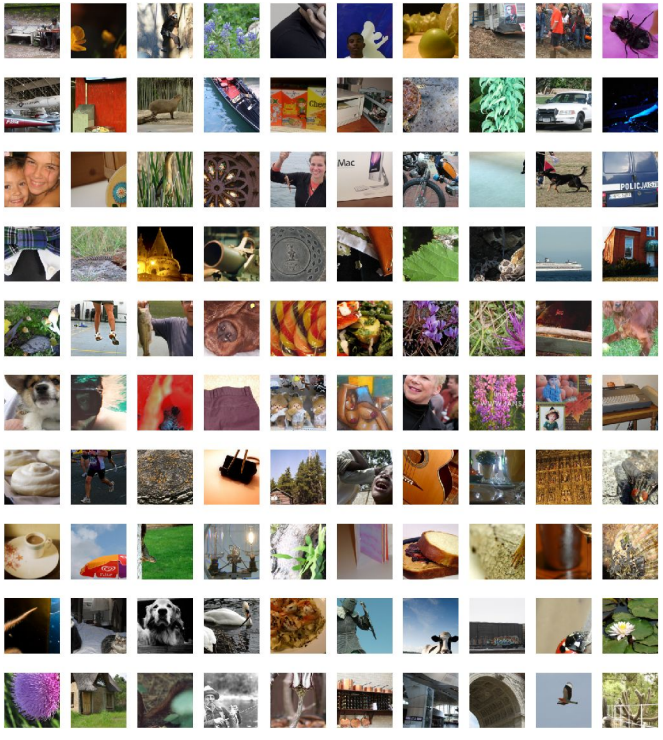
Instance	Features	Class
	✔ Beak ? Webbed feet ✔ Quacks	✔ Is a duck
	✔ Beak ✔ Webbed feet ✔ Quacks	✔ Is a duck
	☐ Beak ☐ Webbed feet ☐ Quacks	☐ Is NOT a duck

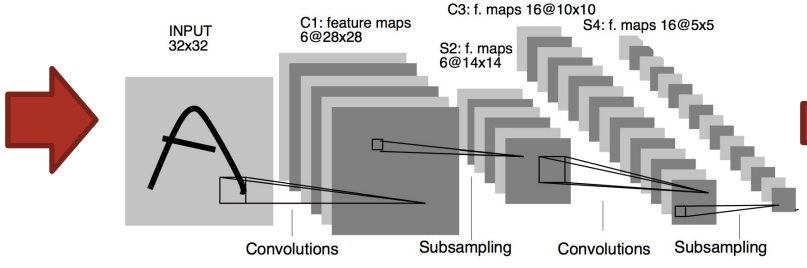
Text Data

Features

Conference	Year	Deadline 1	Held in
SIGMOD	2021	2020-09-22	Xi'an

Image Data

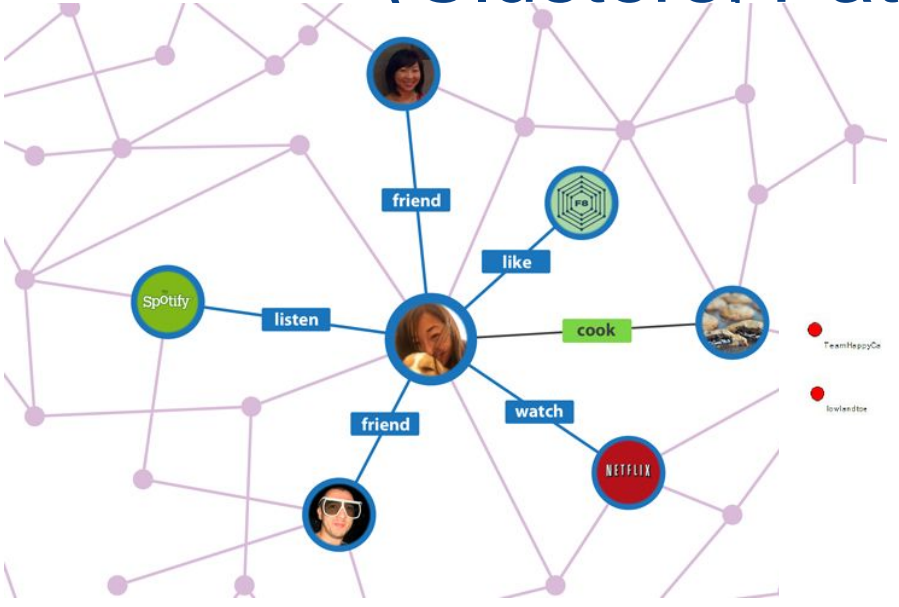


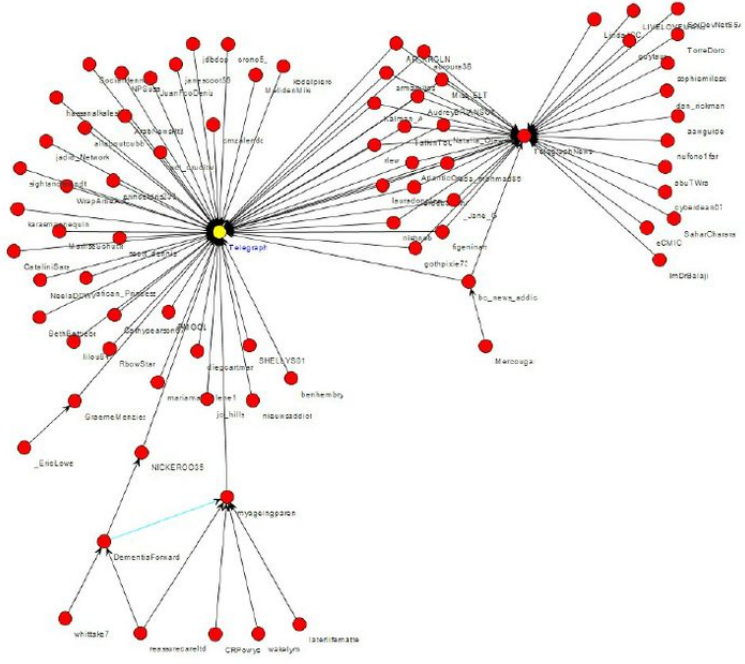


Features
representing
shapes,
colors,
boundaries

Convolutional neural network

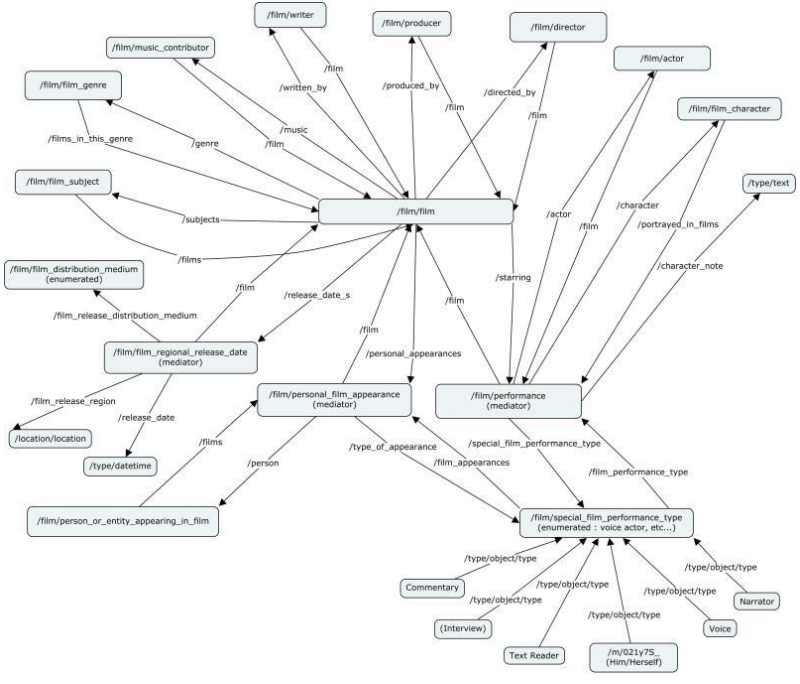
Network Data: Find Patterns in Connectivity (Clusters. Paths, …)





Data with Complex Semantics: Knowledge Graphs

Classes, subclasses, instances, and properties



EducationalOrganization

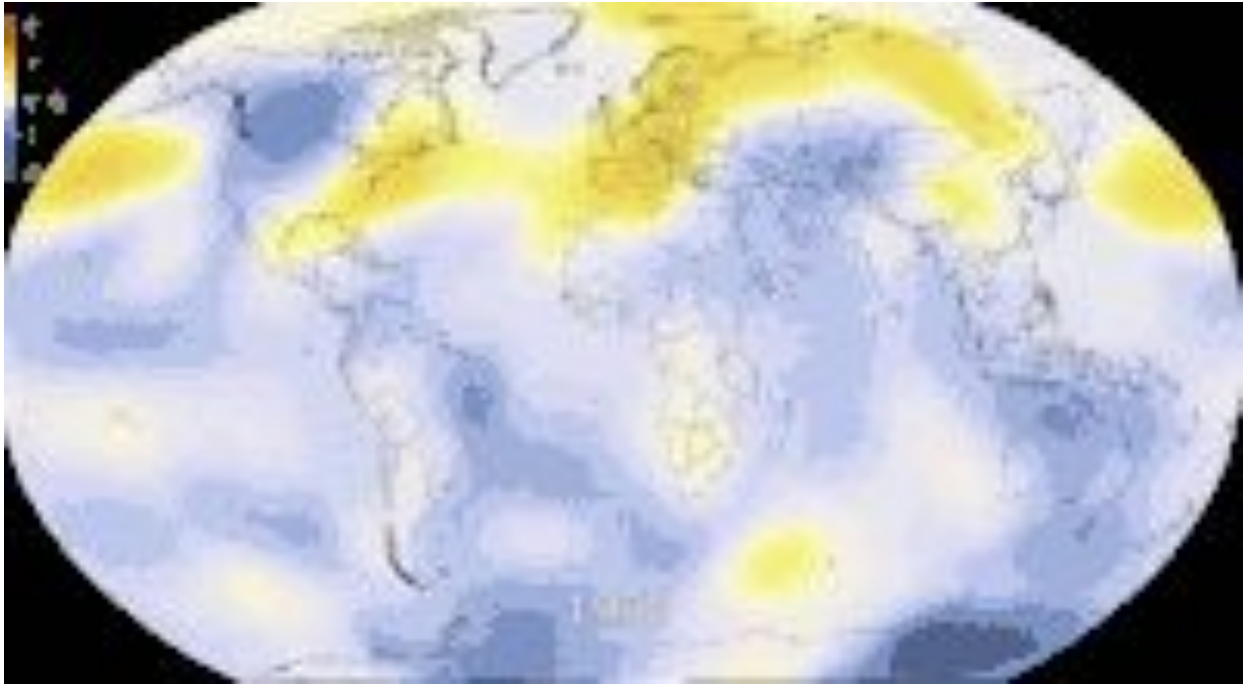
Thing > Organization > EducationalOrganization

An educational organization.

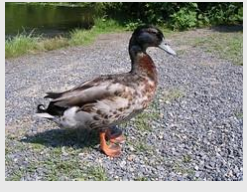
[more...]

Property	Expected Type	Description
Properties from EducationalOrganization		
alumni	Person	Alumni of an organization. Inverse property: alumniOf.
Properties from Organization		
actionableFeedbackPolicy	CreativeWork or URL	For a NewsMediaOrganization or other news-related Organization, a statement about public engagement activities (for news media, the newsroom's), including involving the public - digitally or otherwise -- in coverage decisions, reporting and activities after publication.
address	PostalAddress or Text	Physical address of the item.
aggregateRating	AggregateRating	The overall rating, based on a collection of reviews or ratings, of the item.
alumni	Person	Alumni of an organization. Inverse property: alumniOf.

Spatiotemporal data: Track over Time, Potentially Forecast the Future



Machine Learning

Instance	Features		Class
	✔ Beak ? Webbed feet ✔ Quacks	f: features □ class	✔ Is a duck
  	✔ Beak ✔ Webbed feet ✔ Quacks		✔ Is a duck
	□ Beak □ Webbed feet □ Quacks		□ Is NOT a duck

Summary

- Part II involves taking data, extracting relevant *features*
- Then building models via machine learning algorithms

Quiz 01C on Canvas, linked to the Course schedule.

Brief Self-Check

https://canvas.upenn.edu/courses/1636888/quizzes/2771527

Features are:

A. Unique properties of the data

B. Any kind of property of the data

C. Always learned

D. Values used to predict classes or groups

Which of the following is not a common type of feature:

A. Specific words in text

B. Paths associated with a node in a network

C. Particular shapes or boundaries in an image

D. A document

Data Analytics Tasks and Goals

Susan B. Davidson and Zachary G. Ives
University of Pennsylvania
CIS 545: Big Data Analytics

Portions of this lecture were developed for the OpenDS4All project, piloted by Penn, IBM, and the Linux Foundation

The Goal of Data Analytics: From Data to "Knowledge" or Action

Pattern detection: Raw data □ patterns □ partial understanding

- "Show me sales by region by product category"
- "Show me clusters of documents by concept"
- "Data cubes" (sales by region by quarter by type of product)
- Typically, *descriptive* statistics

- Sometimes we can find patterns automatically and *cluster* data

Hypothesis □ experiment over sample □ significance

- "Behavioral factor F leads to higher risk of outcome O"
- Do statistical test, measure significance vs. *null hypothesis*
- Typically, *inferential* statistics

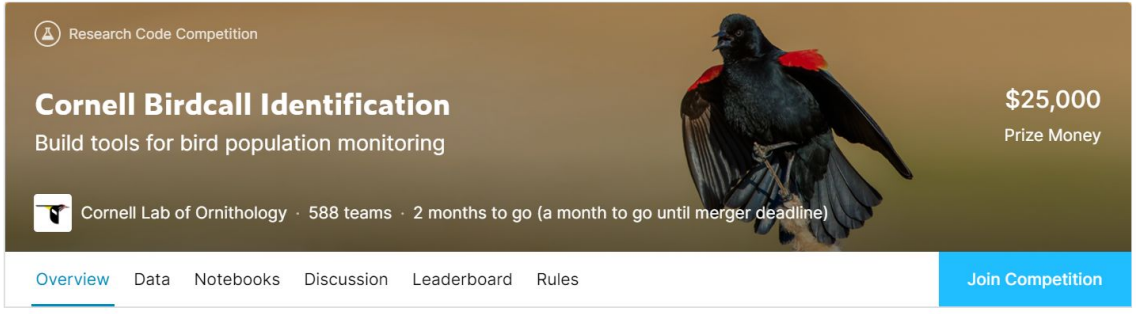
- Sometimes we use this to build a *predictive model* or *classifier*

What Does Big Data Analytics Involve?

- Acquisition, access – data may exist without being accessible
- Wrangling – data may be in the wrong form
- Integration, representation – data relationships may not be captured
- Cleaning, filtering – data may have variable quality
- Hypothesizing, querying, analyzing, modeling – from data to info
- Understanding, iterating, exploring – helping build knowledge

- And: ethical obligations – need to protect data, follow good statistical practices, present results in a non-misleading way

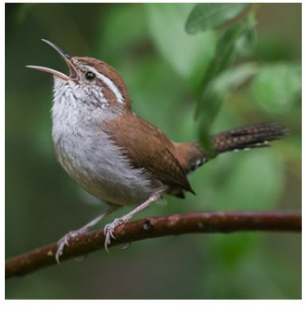
An Example from Kaggle.com



Overview

- Description
- Evaluation
- Timeline
- Prizes
- Code Requirements
- Acknowledgements

Do you hear the birds chirping outside your window? Over 10,000 bird species occur in the world, and they can be found in nearly every environment, from untouched rainforests to suburbs and even cities. Birds play an essential role in nature. They are high up in the food chain and integrate changes occurring at lower levels. As such, birds are excellent indicators of deteriorating habitat quality and environmental pollution. However, it is often easier to hear birds than see them. With proper sound detection and classification, researchers could automatically intuit factors about an area's quality of life based on a changing bird population.



There are already many projects underway to extensively monitor birds by continuously recording natural soundscapes over long periods. However, as many living and nonliving things make noise, the analysis of these datasets is often done manually by domain experts. These analyses are painstakingly slow, and results are often incomplete. Data science may be able to assist, so researchers have turned to large crowdsourced databases of focal recordings of birds to train AI models. Unfortunately, there is a domain mismatch between the training data (short recording of individual birds) and the soundscape recordings (long recordings with often multiple species calling at the same time) used in monitoring applications. This is one of the reasons why the performance of the currently used AI models has been subpar.

To unlock the full potential of these extensive and information-rich sound archives, researchers need good machine listeners to reliably extract as much information as possible to aid data-driven conservation.

Data Science / Data Analytics: Beware Over-Hyped Expectations!

Data science myth:

- We'll learn everything "bottom up" using fancy statistics and machine learning
- Basically we "turn the crank" and out pop insights!

Data + algorithms □ knowledge

Data science reality:

- We'll typically rely on human expertise to impose **models** over the data, the features, etc.
- Deep learning can do feature selection – but why throw away what we know!

Data + human insight + algorithms + iteration □ information □ knowledge

Data Science Process

- What **question** are you answering?
- What is the right **scope** of the project?
- **What data** will you use?
- What **techniques** are you going to try?
- How will you **evaluate** your results?

- What **maintenance** will be required?

The Word from Our Students and Collaborators at Data Science Companies

80-90% of their work involves:

- Working with experts to understand the domain, assumptions, questions, etc.
- Trying to catalog and make sense of the data sources
- Wrangling, extracting, and integrating the data
- Cleaning the wrangled data

… Before we get to feature extraction and machine learning!

Recapping Our Motivations for Big Data Analytics (and this Course)

Big data:

- High-dimensional
- Hard to understand
- Requires understanding computation and I/O costs

Need to add semantic structure, integrate data, extract features

Discovery, hypothesis testing, clustering, classification, recommendation!

Quiz 01D on Canvas, linked to the Course schedule.

Brief Self-Check

https://canvas.upenn.edu/courses/1636888/quizzes/2771510

In data science, human expertise in the *problem domain* helps with all of the following *except*:

A. Filling in missing data

B. Determining the right features

C. Supplying computational resources

D. Formulating the appropriate hypothesis

Understanding, cleaning, and wrangling data account for roughly how much of the data science process?

A. 40-60%

B. 80-90%

C. 10-20%

D. 30-40%

(transition)

Course Logistics

Susan B. Davidson and Zachary G. Ives
University of Pennsylvania
CIS 545: Big Data Analytics

Portions of this lecture were developed for the OpenDS4All project, piloted by Penn, IBM, and the Linux Foundation

Course Website: cis.upenn.edu/~cis545

Logistics

Instructors: Profs. Susan Davidson and Zachary Ives

Your fantastic TAs: Vian Djianto (Head TA), Carol Li (Head TA), Phillip Chau, Kunaal Chaudhari, Calvin Hu, Michael Lau, Parmita Mishra, Ernest Ng, Prakruthi Raghav, Aditya Rathi, Shreyans Tiwari, Warren Wang, Kailin Zheng

This class will be offered in **hybrid** format, with both on-campus and remote learning opportunities. **Given Penn's new COVID-19 guidelines, the first 2 lectures will be delivered remotely, 12:00 - 1:30pm, via Zoom.** Thereafter:

- **Lectures** will be Mondays and Wednesdays, 12:00pm - 1:30pm, in **Towne 100 (Heilmeier Hall)**
  Lectures will also be live-streamed via Zoom.
  Prerecorded video lectures will also be made available, linked through the syllabus.
- **Recitations** will be Fridays, **1:45pm - 3:15pm** in Towne 100. **Note the different time slot from the lecture!**
  The presentation portions of recitations will be recorded. Recitations are optional, but highly recommended.

Course Description

In the era of big data, we are increasingly faced with the challenges of converting massive amounts of data to actionable *knowledge*. Given the limits of individual machines (compute power, memory, bandwidth), increasingly the solution is to clean, integrate, and process the data using statistical machine learning techniques, in parallel on many machines. This course focuses on the fundamentals of scaling computation to handle common data analytics tasks. You will learn about basic tasks in collecting, wrangling, and structuring data; programming models for performing certain kinds of computation in a scalable way across many compute nodes; common approaches to converting algorithms to such programming models; standard toolkits for data analysis consisting of a wide variety of primitives; and popular distributed frameworks for analytics tasks such as filtering, graph analysis, clustering, and classification.

Who This Class is For

Students from engineering / science, applied math, social science, …

- With solid but basic background in statistics, plus experience in Python programming
- STAT 430, EAS 301 or equivalent, CIS 105 or MCIT 590 or equivalent

Students from computer science who want to understand how to analyze data

Many of you have had CIS 419/519/520, some have had CIS 450/550

… Please bear with us as we try to accommodate both audiences! We'll try to make sure we **add a different perspective in each area**

Relationship to Other Courses at Penn

CIS 450/550 focuses on data modeling, querying

- Data representation and management
- Relational querying with SQL; XML querying with XQuery
- DBMS-backed web sites
- We will cover some data modeling and querying, but our focus is on **Apache Spark** and understanding how Python Pandas and SQL relate

CIS 419/519 and 520 focus on machine learning techniques, some tools

- This course uses machine learning tools, and gives "just enough detail" to have a sense of the basic algorithms
- Our focus is on dealing with **compute clusters, heterogeneous data, scale, and visualization**

CIS 455/555 shows how to build the platforms

- Under the covers: how big data analytics systems are built
- Build your own Web server, Google MapReduce

What You'll Learn Here

- Data manipulation in Python, its libraries, and the Spark platform
- Data representation and wrangling
- Data as graphs and matrices
- Data ethics and privacy
- Data cleaning, hypothesis testing, …
- Use of common algorithmic patterns and learning primitives
  - Clustering and unsupervised learning
  - Training and classification
  - Managing time-varying data
  - Deep learning
- Some data visualization

This is not a machine learning course
This is not a database course

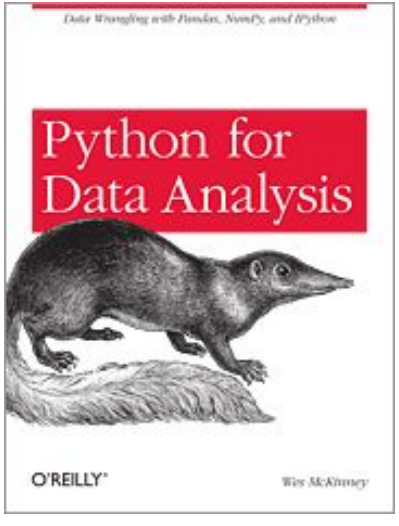
But you'll learn how to apply both!

Texts / Reference Books

- We will use the following books, plus supplemental readings

Data Science from Scratch, by Joel Grus. O'Reilly

(if you have a non-CS



Python for Data Analysis, 2nd ed by Wes McKinney. O'Reilly

(for everyone)

Prerequisites, Workload, etc.

Necessary skills:

- Comfort with coding in Python (we won't be writing huge programs)
- Attention to detail
- A willingness to "push the envelope"

Workload:

- Multiple notebook-based homework assignments
- A term project with an experimental report (roughly analogous to Kaggle competitions)
- Two midterms, one taken as a final exam

Payoff is based on how much you put in:

- Lots of hands-on experience with data analysis and data tools / techniques

A Disclaimer…

The field continues to evolve, and so does this course

We are trying to balance breadth and depth, and a diverse set of students and goals!

We will be using some immature technology

- Not everything will have been validated ahead of time
- We'll do the best we can to smooth over the bugs!

We hope it will be a fun course, though…

… And an interesting one!

Technologies

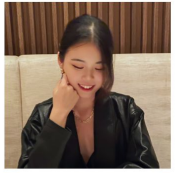
- We'll make heavy use of the Python data science ecosystem
  - Generally hosted on **Google Colab**
  - Using **Jupyter** as an integrated environment, with a cloud based **PennGrader**
  - We'll do some tasks on **Amazon & Google**'s cloud services
- And **Apache Spark** for parallel programming, D3 for custom visualization
- **Scikit-learn, Pandas, PySpark, …**
- Also, some use of Apache mxNet, Tensorflow, and Keras

Get to Know Your Fantastic TAs, and Stop by Our Office Hours

Head Teaching Assistant

Vian Djianto

djianto@seas



Head Teaching Assistant

Carol Li

caroljli@seas

Teaching Assistant

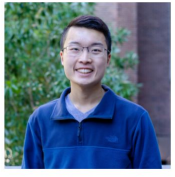
Parmita Mishra

pmish@sas

Teaching Assistant

Ernest Ng

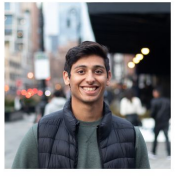
ngernest@seas



Teaching Assistant

Phillip Chau

pchau@seas



Teaching Assistant

Kunaal Chaudhari

kuchaud@seas

Teaching Assistant

Prakruthi Raghavendra

praghav@wharton

Teaching Assistant

Aditya Rathi

adityara@seas

Teaching Assistant

Adiwid (Boom) Devahastin Na Ayudhya

adiwid@seas

Teaching Assistant

Calvin Hu

calvinhu@seas

Teaching Assistant

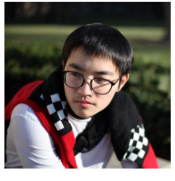
HyungSeok (Paul) Roh

hyroh@seas

Teaching Assistant

Kavish Shah

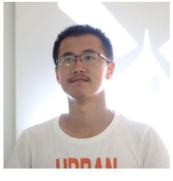
kavish@seas



Teaching Assistant

Michael Lau

lyming@sas



Teaching Assistant

Ang Li

liang810@seas

Teaching Assistant

Shreyans Tiwari

shreytiw@seas

Teaching Assistant

Warren Wang

gmwwang@wharton

For All Meetings with Us (In-Person + Remote)

- Please go to https://ohq.io and sign up for CIS 545
  (OHQ doesn't allow signups to direct course links)

- Then add yourself to the queue, so we can balance between in-person and remote visitors!

Grading (Tentative)

- 5 homework assignments + Homework 0 (40%)
- Midterm 1, week of 3/2, 80 minutes (15%)
- Midterm 2, finals week, 80 minutes (15%)
- Kaggle-competition-style term project (20%)

- Participation: a combination of
  Canvas self-checks,
  participating in class or synchronous activities,
  viewing asynchronous videos and doing quizzes,
  engagement with tools,
  activity on Piazza (10%)

Before Next Week:
Get Started with Google Colab and HW0

As you've probably discovered: course web is at www.cis.upenn.edu/~cis545
Sign up for Piazza, log into Canvas

Please read "What Data Scientists Really Do, What Does a Data Engineer Do?"

Get started on Homework 0 to familiarize yourself with Colab

A word about the waitlist…

The course is full.

To get a permit for the course, you must be on the waitlist

The waitlist is very long. We are processing it as space allows

There are two sections of the course 001 (in-person), 002 (virtual, for grad students only; please do **not** come to in-person lectures until we tell you there is room).