MAS DSE 230
Scalable Analytics

Introduction

Mai H. Nguyen

SESSION 1 TOPICS

- Introductions
- Introduction to Big Data
- Course Overview
- Container Setup
- Computer Systems & Parallelism
- Guest Lecture
- Exercise
- Assignments

SESSION 1 TOPICS

- Introductions
- Introduction to Big Data
- Course Overview
- Container Setup
- Computer Systems & Parallelism
- Guest Lecture
- Exercise
- Assignments

INTRODUCTIONS

- Mai Nguyen, Ph.D.
  - San Diego Supercomputer Center
  - Lead for Data Analytics

- T.A.: Sagar Hathwar
  - Computer Science & Engineering
  - M.S. Student

MY BACKGROUND

- Education
    - B.S. in Computer Science from Colorado State University
    - M.S. & Ph.D. in CSE from UCSD
- Work Experience
    - Worked in industry for many years
    - Teaching since 2009
    - At SDSC since 2014
- Research
    - Machine learning, deep learning, data science
    - Application areas: Medical image analysis, satellite image analysis, hazards science, NLP, ...

SESSION 1 TOPICS

- Introductions
- Introduction to Big Data
- Course Overview
- Computer Systems & Parallelism
- Guest Lecture
- Exercise
- Assignments

WHAT IS THIS COURSE ABOUT?

- DSE 230: Scalable Analytics
- From course description:
  - This course is designed to provide students with skills and knowledge to perform analytics at scale… Students will get hands-on experience on distributed and cloud-based platforms to perform scalable analytics
- In a nutshell:
  - You will learn techniques and tools for analyzing big data

WHAT IS BIG DATA?

http://www.digitalzenway.com/2011/12/data-diet-a-resolution-you-can-stick-to/

- "Growing torrent" of data
- Data
  - Comes in large volumes
  - Continuous
  - Complex

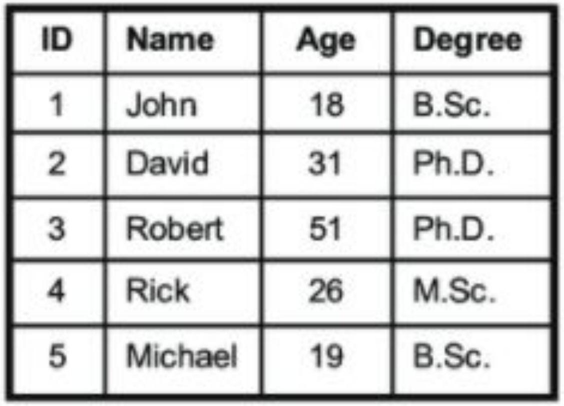
M. H. Nguyen

WHAT IS BIG DATA?

- Big data analytics is the use of advanced analytic techniques against very large, diverse data sets, including structured/unstructured and streaming/batch. (ibm.com)
- Data that is too large and complex to be dealt with by traditional data processing application software. (wikipedia.org)
- Big data is larger, more complex data sets… These massive volumes of data can be used to address business problems you wouldn’t have been able to tackle before. (oracle.com)

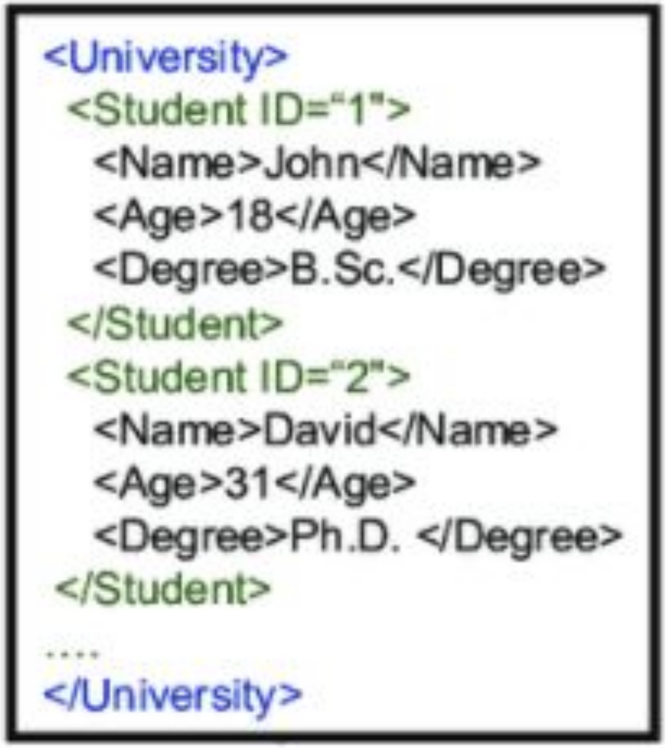
TYPES OF DATA

Structured Data



ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

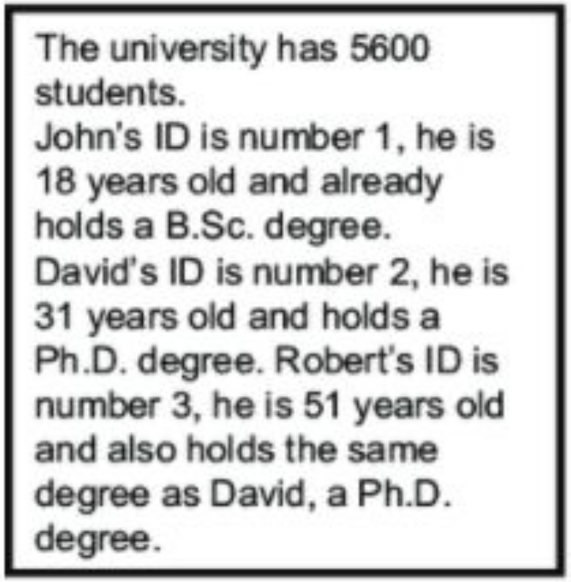
Semi-Structured Data



```
<University>
 <Student ID=“1">
   <Name>John</Name>
   <Age>18</Age>
   <Degree>B.Sc.</Degree>
 </Student>
 <Student ID=“2">
   <Name>David</Name>
   <Age>31</Age>
   <Degree>Ph.D. </Degree>
</Student>

....
</University>
```

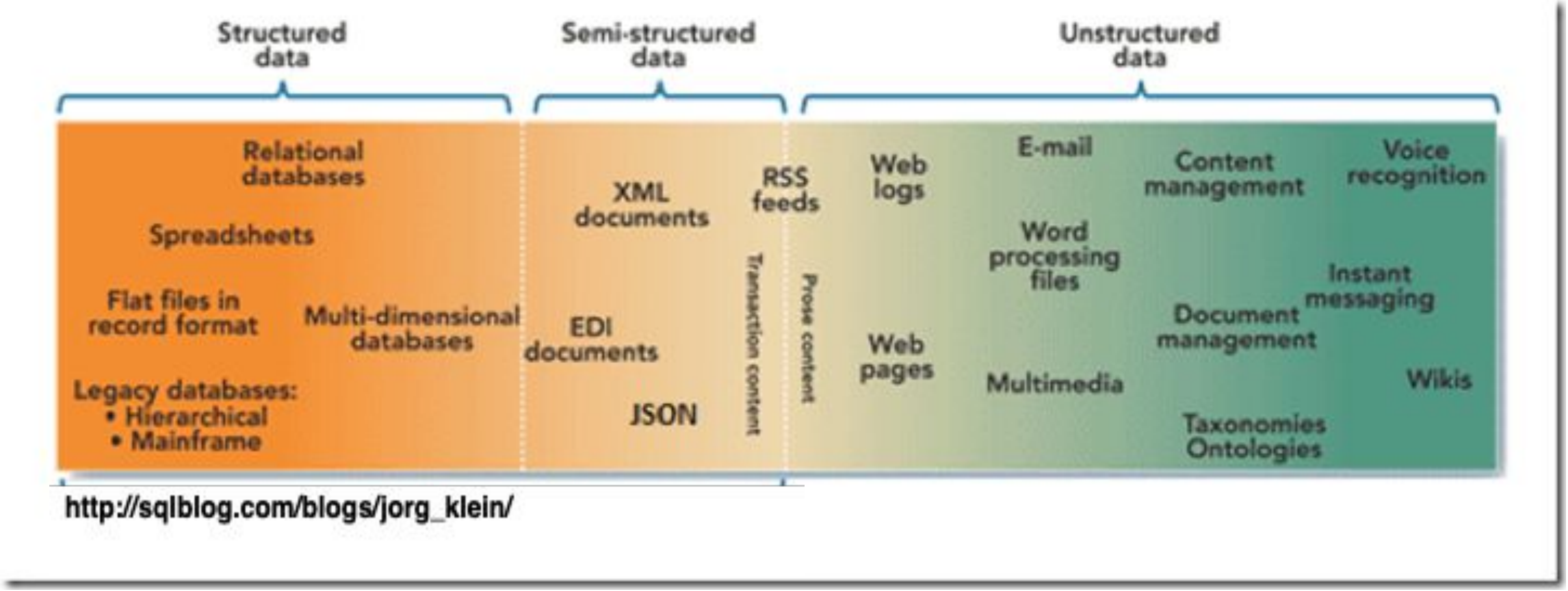
Unstructured Data



The university has 5600 students.
John's ID is number 1, he is 18 years old and already holds a B.Sc. degree.
David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

https://www.researchgate.net/figure/Unstructured-semi-structured-and-structured-data_fig4_236860222

TYPES OF DATA



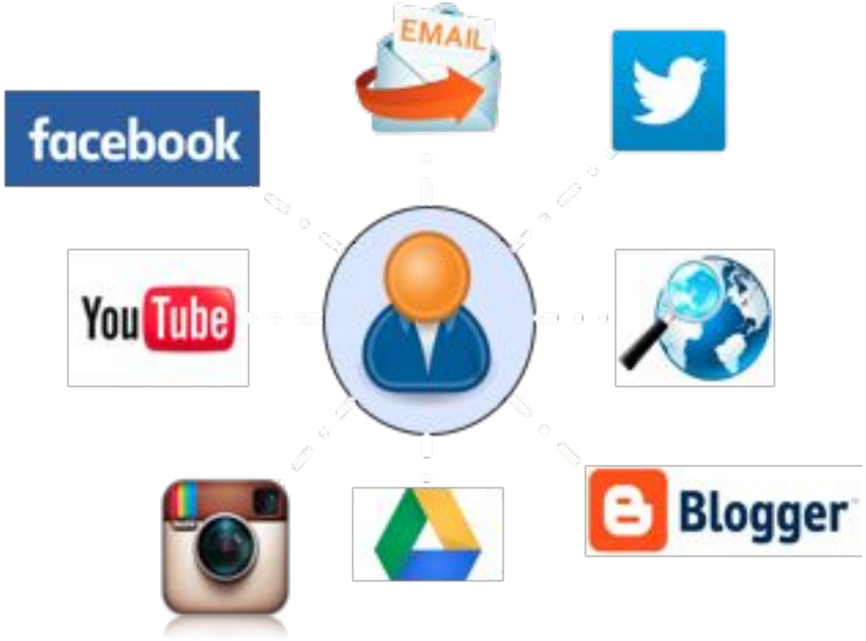
http://sqlblog.com/blogs/jorg_klein/

WHERE DOES BIG DATA COME FROM?

Machines

Sensors

People



HOW IS BIG DATA USED?

- What are some applications that use big data?

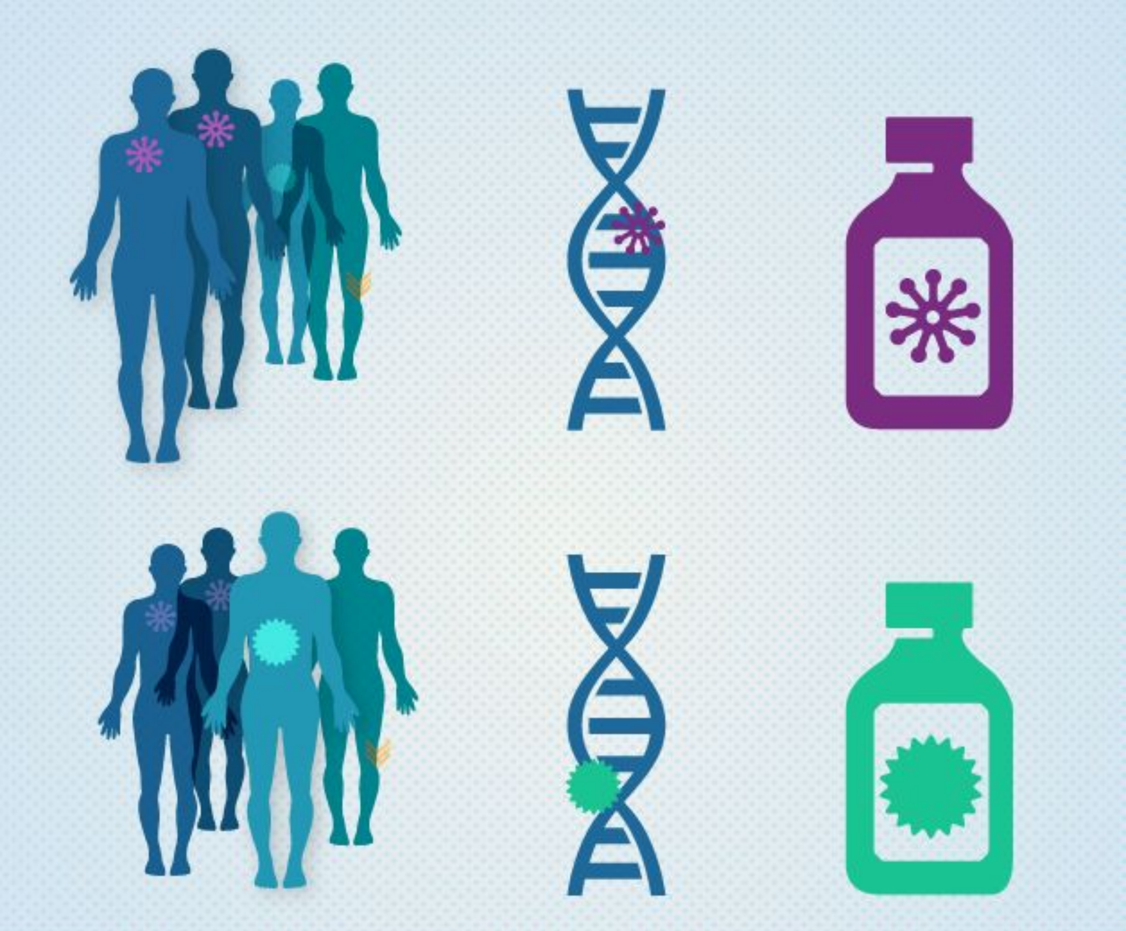
ASTRO-PHYSICS

LIGO: Laser Interferometer Gravitational-Wave Observatory

Artist's rendition of two colliding neutron stars. Credit: National Science Foundation/LIGO/Sonoma State University/A. Simonnet

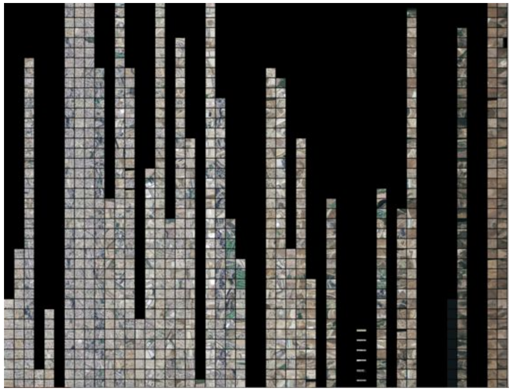
PRECISION MEDICINE

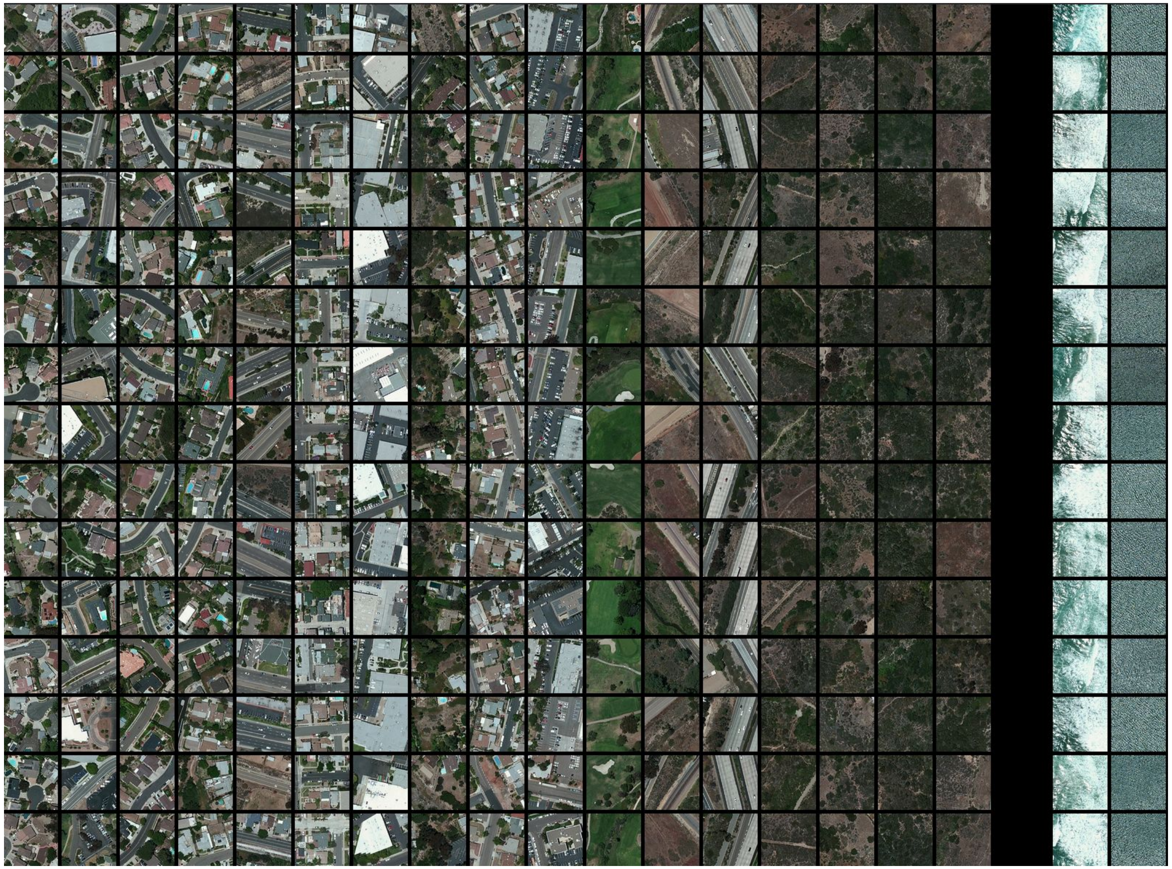
- Patients with tumors that share the same genetic change receive the drug that targets that change, no matter the type of cancer



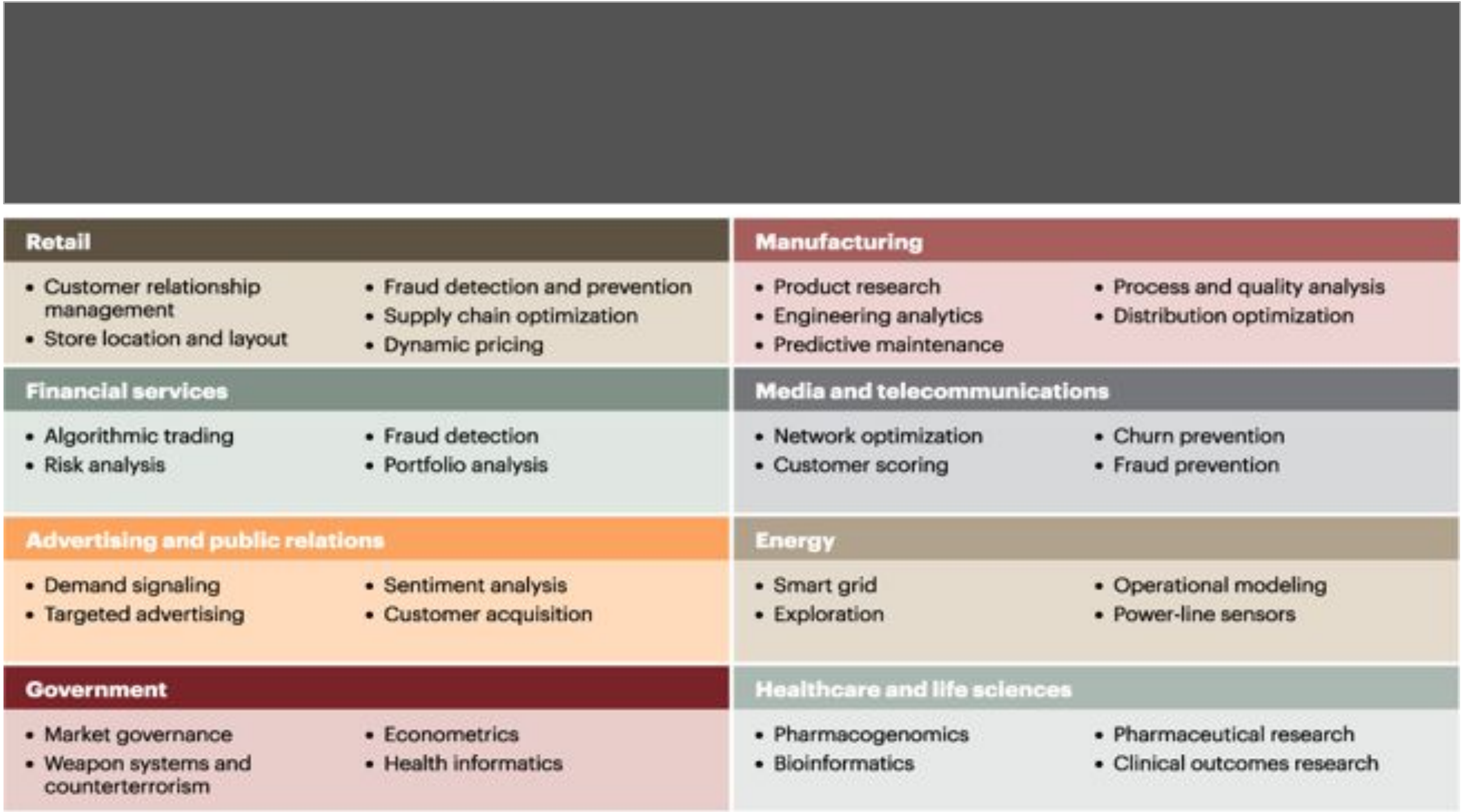
https://www.cancer.gov/news-events/cancer-currents-blog/2015/precision-medicine-initiative-2016

SATELLITE IMAGE ANALYSIS





MANY INDUSTRIES USE BIG DATA



Retail

- Customer relationship management
- Store location and layout
- Fraud detection and prevention
- Supply chain optimization
- Dynamic pricing

Manufacturing

- Product research
- Engineering analytics
- Predictive maintenance
- Process and quality analysis
- Distribution optimization

Financial services

- Algorithmic trading
- Risk analysis
- Fraud detection
- Portfolio analysis

Media and telecommunications

- Network optimization
- Customer scoring
- Churn prevention
- Fraud prevention

Advertising and public relations

- Demand signaling
- Targeted advertising
- Sentiment analysis
- Customer acquisition

Energy

- Smart grid
- Exploration
- Operational modeling
- Power-line sensors

Government

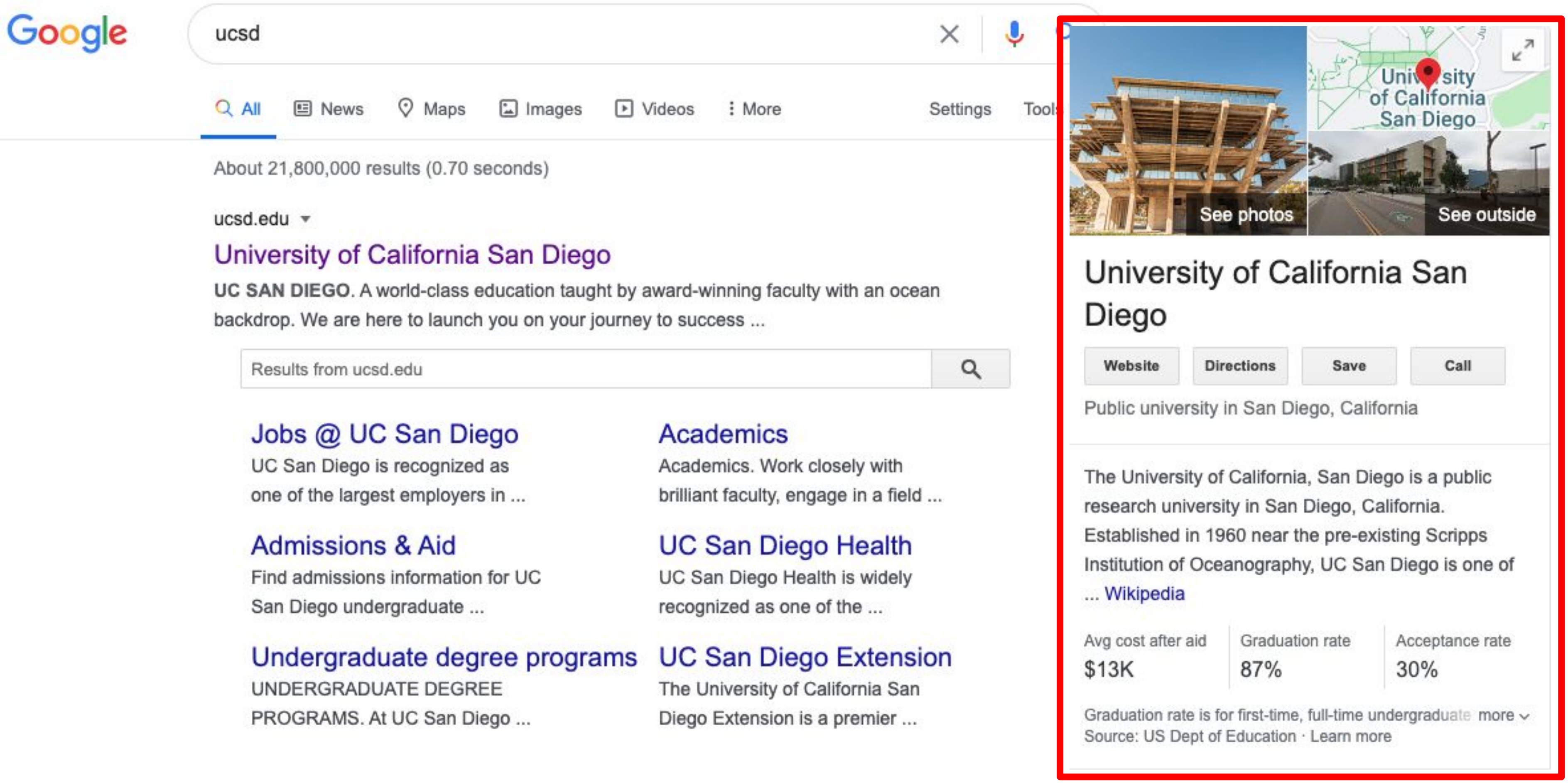
- Market governance
- Weapon systems and counterterrorism
- Econometrics
- Health informatics

Healthcare and life sciences

- Pharmacogenomics
- Bioinformatics
- Pharmaceutical research
- Clinical outcomes research

Source: A.T. Kearney analysis

INTERNET SEARCH



About 21,800,000 results (0.70 seconds)

ucsd.edu ▾

University of California San Diego

UC SAN DIEGO. A world-class education taught by award-winning faculty with an ocean backdrop. We are here to launch you on your journey to success ...

Results from ucsd.edu

Jobs @ UC San Diego

UC San Diego is recognized as one of the largest employers in ...

Academics

Academics. Work closely with brilliant faculty, engage in a field ...

Admissions & Aid

Find admissions information for UC San Diego undergraduate ...

UC San Diego Health

UC San Diego Health is widely recognized as one of the ...

Undergraduate degree programs

UNDERGRADUATE DEGREE PROGRAMS. At UC San Diego ...

UC San Diego Extension

The University of California San Diego Extension is a premier ...

University of California San Diego

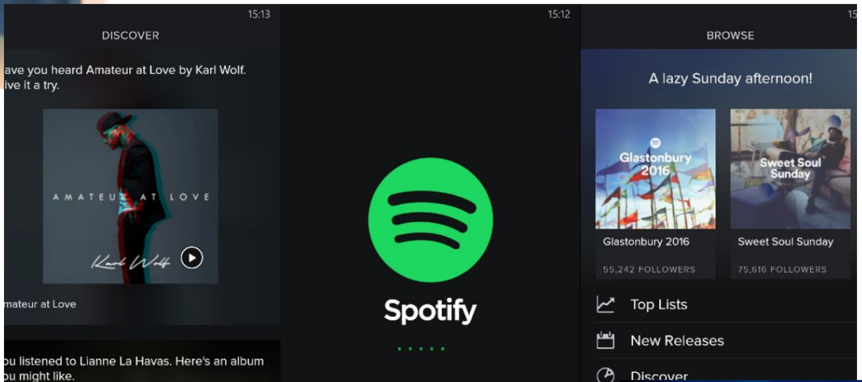
Public university in San Diego, California

The University of California, San Diego is a public research university in San Diego, California. Established in 1960 near the pre-existing Scripps Institution of Oceanography, UC San Diego is one of ... Wikipedia

Avg cost after aid	Graduation rate	Acceptance rate
$13K	87%	30%

Graduation rate is for first-time, full-time undergraduate more ⌄
Source: US Dept of Education · Learn more

EVERYDAY APPLICATIONS



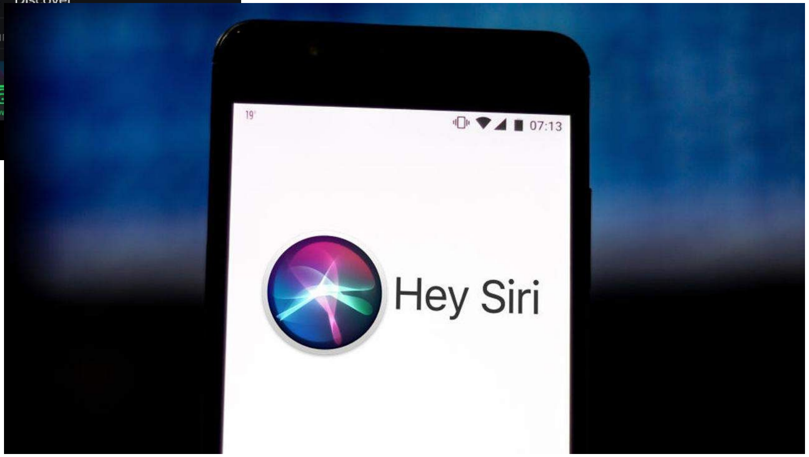
ave you heard Amateur at Love by Karl Wolf.
ive it a try.

mateur at Love

u listened to Lianne La Havas. Here's an album
u might like.

A lazy Sunday afternoon!

Glastonbury 2016	Sweet Soul Sunday
55,242 FOLLOWERS	75,616 FOLLOWERS



WHY BIG DATA NOW?

- Advances in processing power, storage capacity, mobile computing, interconnectivity
  - Create unprecedented data
  - Can store and process more data
- Data-driven applications in all areas
  - Science: bioinformatics, image analysis
  - Medicine: drug design, healthcare
  - Retail: targeted advertisement, dynamic pricing
  - Finance: fraud detection, risk analysis
  - Manufacturing: preventive maintenance, supply chain management
  - Law enforcement: crime pattern detection
  - Others ...

ANALYZING BIG DATA

- Requires scalable techniques and tools
- That’s what we’ll cover in this course!

SESSION 1 TOPICS

- Introductions
- Introduction to Big Data
- Course Overview
- Container Setup
- Computer Systems & Parallelism
- Guest Lecture
- Exercise
- Assignments

COURSE OVERVIEW

- MAS DSE 230 - Scalable Analytics
  - This course is designed to provide students with the skills and knowledge to perform analytics at scale. Topics cover both systems and analytics, and include basic principles of computer systems and parallelism; analytics process; analytics algorithms; scalable computing; and cloud-based analytics. Tools and techniques to perform analytics on large-scale data will be introduced. Students will get hands-on experience on distributed and cloud-based platforms to perform scalable analytics.

COURSE LOGISTICS

- Lecture: Saturday 9am - 4pm
  - Dates: 4/03, 4/17, 5/01, 5/15, 5/29, 6/05
- Canvas: Zoom links, assignments
- Piazza: Announcements, Q&A
  - http://piazza.com/ucsd/spring2021/dse230/home
  - Access code: 203-2021
- Office Hours:
  - Prof. Nguyen: Monday 5 - 6 pm
  - TA Sagar: Tuesday & Thursday 7:30 - 8:30 pm
- All times are in Pacific Time

COURSE TOPICS

- S1 – Big Data Intro, Computer Systems & Parallelism
  - Introduction to Big Data
  - Computer systems
  - Parallelism principles
  - Speedup
- S2 – Big Data & Distributed Processing
  - Big Data challenges
  - Distributed processing
  - Hadoop
  - Spark
  - Analytics process
- S3 – Big Data Analytics
  - Spark core & libraries
  - Analytics with Spark MLlib
  - Model building, selection & evaluation
- S4 – Big Data Analytics & Cloud Computing
  - Dask
  - Cloud computing
  - AWS basics
- S5 – AWS Analytics, DL, Others
  - AWS analytics
  - Deep learning overview
  - Other topics

GRADING

- Grading components
  - Programming Assignments 65%
  - Project 35%
  - PAs are individual work. Can work in pairs for project.
  - PySpark or Dask
- Late Penalty - PAs
  - 10% per day, up to 3 days after due date
- Can ask conceptual or high-level questions on Piazza.

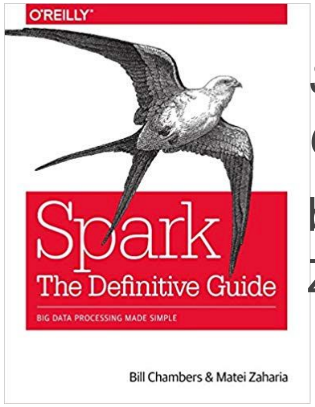
  Do NOT post any code on Piazza.
- Academic Integrity
  - Do your own work!
  - AI violations will be reported to university's AI Office

COURSE SCHEDULE

Week	Topic	Assignment	Points
S1	Big Data Intro Computer Systems & Parallelism	Spark	5
S2	Big Data & Distributed Processing	Spark	10
S3	Big Data Analytics	Spark Project Proposal	15 10
S4	Cloud Analytics	Dask AWS	15 10
S5	AWS Analytics	AWS	10
Finals	Project	Project Presentation	25

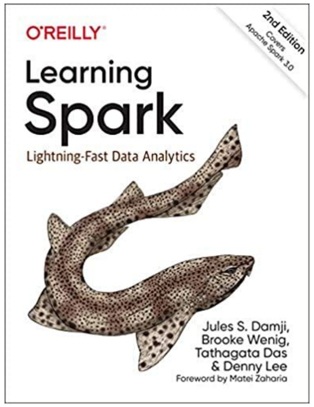
MATERIALS

- Required
  - The Data Scientist’s Guide to Apache Spark (PDF on Canvas)
  - Apache Spark: https://spark.apache.org/docs/latest/
  - Dask: https://dask.org/
  - AWS
    - EMR:
      - https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-overview.html
      - https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-spark.html
    - SageMaker
      - https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works.html
      - https://docs.aws.amazon.com/sagemaker/latest/dg/gs.html
- Recommended


O'REILLY
Spark
The Definitive Guide
BIG DATA PROCESSING MADE SIMPLE
Bill Chambers & Matei Zaharia

Spark: The Definitive Guide (1st edition)

by Chambers and Zaharia

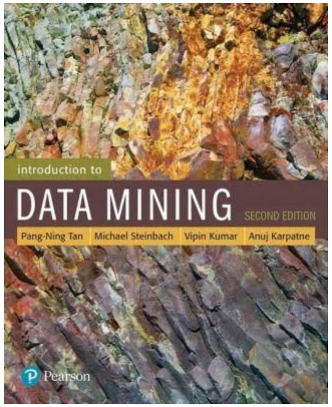

O'REILLY
2nd Edition
Covers Apache Spark 3.0
Learning
Spark
Lightning-Fast Data Analytics
Jules S. Damji, Brooke Wenig, Tathagata Das & Denny Lee
Foreword by Matei Zaharia

Learning Spark (2nd edition)

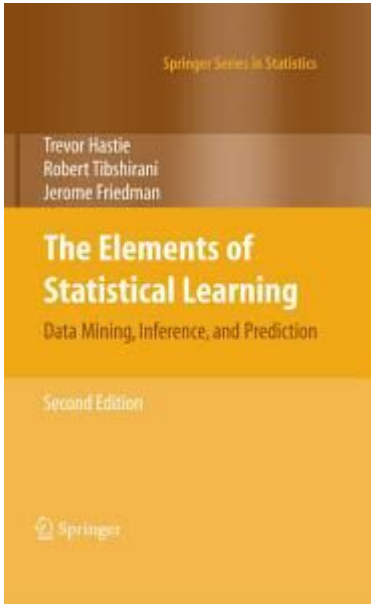
by Damji, Wenig, Das, & Lee

MATERIALS

- Reference

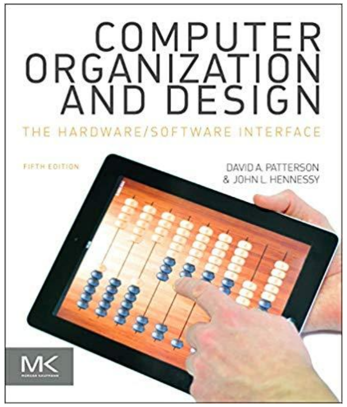


The Elements of Statistical Learning by Hastie, Tibshirani, & Friedman


Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer

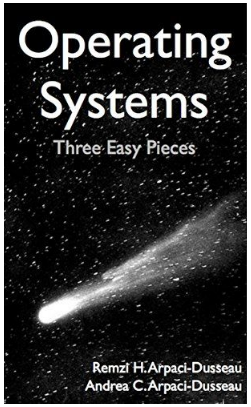
Introduction to Data Mining (2nd edition)

by Tan, Steinbach, Karpatne, & Kumar


COMPUTER ORGANIZATION AND DESIGN
THE HARDWARE/SOFTWARE INTERFACE
FIFTH EDITION
DAVID A. PATTERSON
& JOHN L. HENNESSY
MK

Computer Organization and Design (5th edition)

by Patterson & Hennessy



Operating Systems: Three Easy Pieces

by Remzi & Arpaci-Dusseau

SYLLABUS

- Be sure to review
- Available on Canvas
- Contents
  - Course logistics
  - Course description
  - Schedule
  - Materials
  - Grading
  - Academic Integrity

SESSION 1 TOPICS

- Introductions
- Introduction to Big Data
- Course Overview
- Container Setup
- Computer Systems & Parallelism
- Guest Lecture
- Exercise
- Assignments

SESSION 1 TOPICS

- Introductions
- Introduction to Big Data
- Course Overview
- Container Setup
- Computer Systems & Parallelism
- Guest Lecture
- Exercise
- Assignments

SESSION 1 TOPICS

- Introductions
- Introduction to Big Data
- Course Overview
- Container Setup
- Computer Systems & Parallelism
- Guest Lecture
- Exercise
- Assignments

GUEST LECTURE

- Ilkay Altintas, Ph.D.
    - Chief Data Science Officer, SDSC
    - Fellow, HDSI
    - Division Director, Cyberinfrastructure Research, Education, and Development
    - Founder and Director, Workflows for Data Science (WorDS) Center of Excellence
    - Founder and Director, WIFIRE Lab
    - Faculty Co-Director, Master of Advanced Study in Data Science and Engineering
- "Toward a Scalable Computing Ecosystem: Advancing Data-Integrated Applications for Science and Society"

SESSION 1 TOPICS

- Introductions
- Introduction to Big Data
- Course Overview
- Container Setup
- Computer Systems & Parallelism
- Guest Lecture
- Exercise
- Assignments

SESSION 1 TOPICS

- Introductions
- Introduction to Big Data
- Course Overview
- Container Setup
- Computer Systems & Parallelism
- Guest Lecture
- Exercise
- Assignments

PROJECT DESCRIPTION

- Team of 2 people

- Project
  - Proposal presentation
    - Presented in Session 3
  - Final presentation
    - Presented in Finals Week
  - Peer review
    - Each team evaluates 2 other teams - questions and feedback on presentation
  - Team evaluation
    - Evaluate your team partner

- PySpark or Dask code

PROJECT DESCRIPTION

- Problem description
- Analysis task
- Data
- Data preparation
- Analysis approach
- Challenges and solutions
- Analysis results and insights gained
- Future work

PROGRAMMING ASSIGNMENT 1

- Spark setup - Docker container
  - Read/Write to HDFS
  - Write simple PySpark code
- On Canvas
  - PA1 Instructions.pdf
  - BookReviews_1M.txt.zip
- Submit
  - Jupyter notebook
  - Python script (.py): This will be run to check your code
- Due Friday 2021-04-16 at 11:59pm Pacific Time

BRIEF SPARK INTRODUCTION

- Starting Spark session

```
import pyspark
from pyspark.sql import SparkSession

conf = pyspark.SparkConf().setAll([
         ('spark.master’,   'local[1]'),
         ('spark.app.name', 'App Name')])
spark = SparkSession.builder.config(conf=conf).getOrCreate()
```

BRIEF SPARK INTRODUCTION

- Loading data from local file system

```
df = spark.read.text(“file:///<path>/<file>.txt”)

df = spark.read.csv(“file:///<path>/<file>.csv”,
                    header=True).cache()
```

- Loading data from HDFS

```
df = spark.read.text(“hdfs:///<path>/<file>.txt”) \
                     .cache()

df = spark.read.csv(“hdfs:///<path>/<file>.csv”,
                    header=True)
```

SESSION 1 ASSIGNMENTS

- Programming Assignment 1
- Project
  - Read project description
  - Form team (2 people)
  - Find dataset
  - Formulate problem description
  - Start analyzing data

SPARK RESOURCES

- Spark Main Page
  - https://spark.apache.org/
- Spark Overview
  - https://spark.apache.org/docs/latest/index.html
- Spark Examples
  - https://spark.apache.org/examples.html
- Spark SQL, DataFrames and DataSets Programming Guide
  - https://spark.apache.org/docs/latest/sql-programming-guide.html
- Spark MLlib Programming Guide
  - https://spark.apache.org/docs/latest/ml-guide.html
- PySpark API Documentation
  - https://spark.apache.org/docs/latest/api/python/index.html
- PySpark SQL Basics Cheat Sheet
  - PDF on Canvas
- Note: Spark version 3.1.1, Python, DataFrame API

SESSION 1 TOPICS

- Introductions
- Introduction to Big Data
- Course Overview
- Container Setup
- Computer Systems & Parallelism
- Guest Lecture
- Exercise
- Assignments