UC San Diego

DSC 102
Systems for Scalable Analytics

Rod Albuyeh

Topic 1: Basics of Machine Resources
Part 1: Computer Organization

Ch. 1, 2.1-2.3, 2.12, 4.1, and 5.1-5.5 of CompOrg Book

A few administrative items...

Looking ahead, let's take a peek at AWS EC2 instance types:

https://instances.vantage.sh

Q: *What is a computer?*

A programmable electronic device that can store, retrieve, and process digital data.

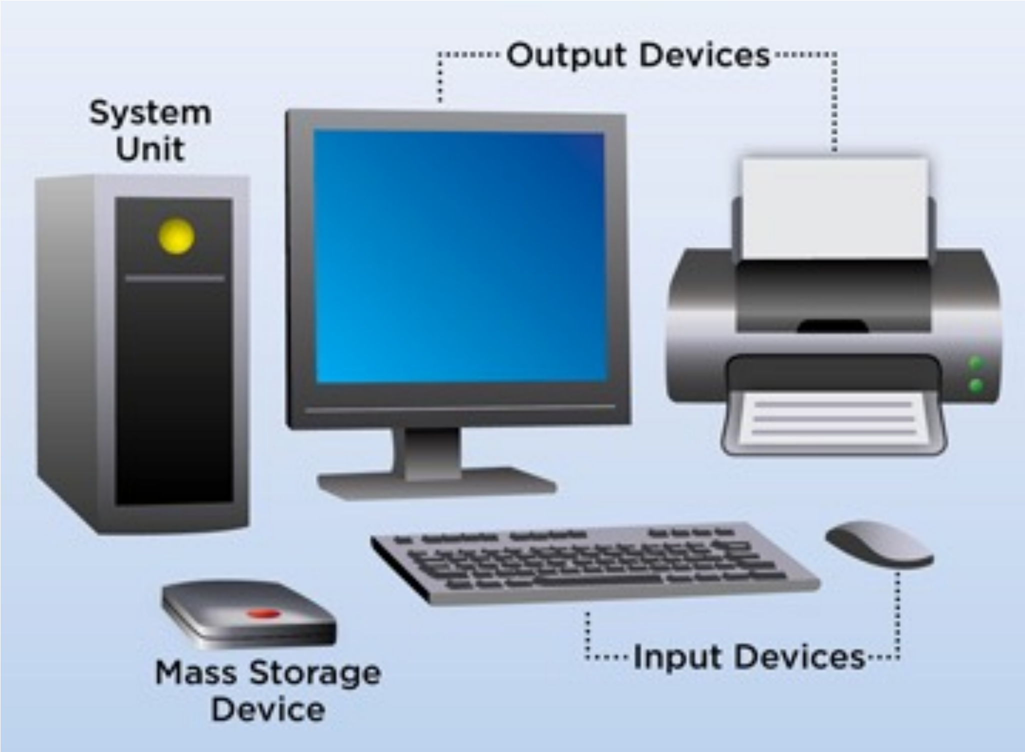
Peter Naur (1928-2016)

Computer science pioneer; proposed alternative term “Datalogy” (still used in his native country Denmark) or “Data Science”.

Outline

- Basics of Computer Organization
  - Digital Representation of Data
  - Processors and Memory Hierarchy
- Basics of Operating Systems
  - Process Management: Virtualization; Concurrency
  - Filesystem and Data Files
  - Main Memory Management
- Persistent Data Storage

Parts of a Computer



Hardware:
The electronic machinery (wires, circuits, transistors, capacitors, devices, etc.)

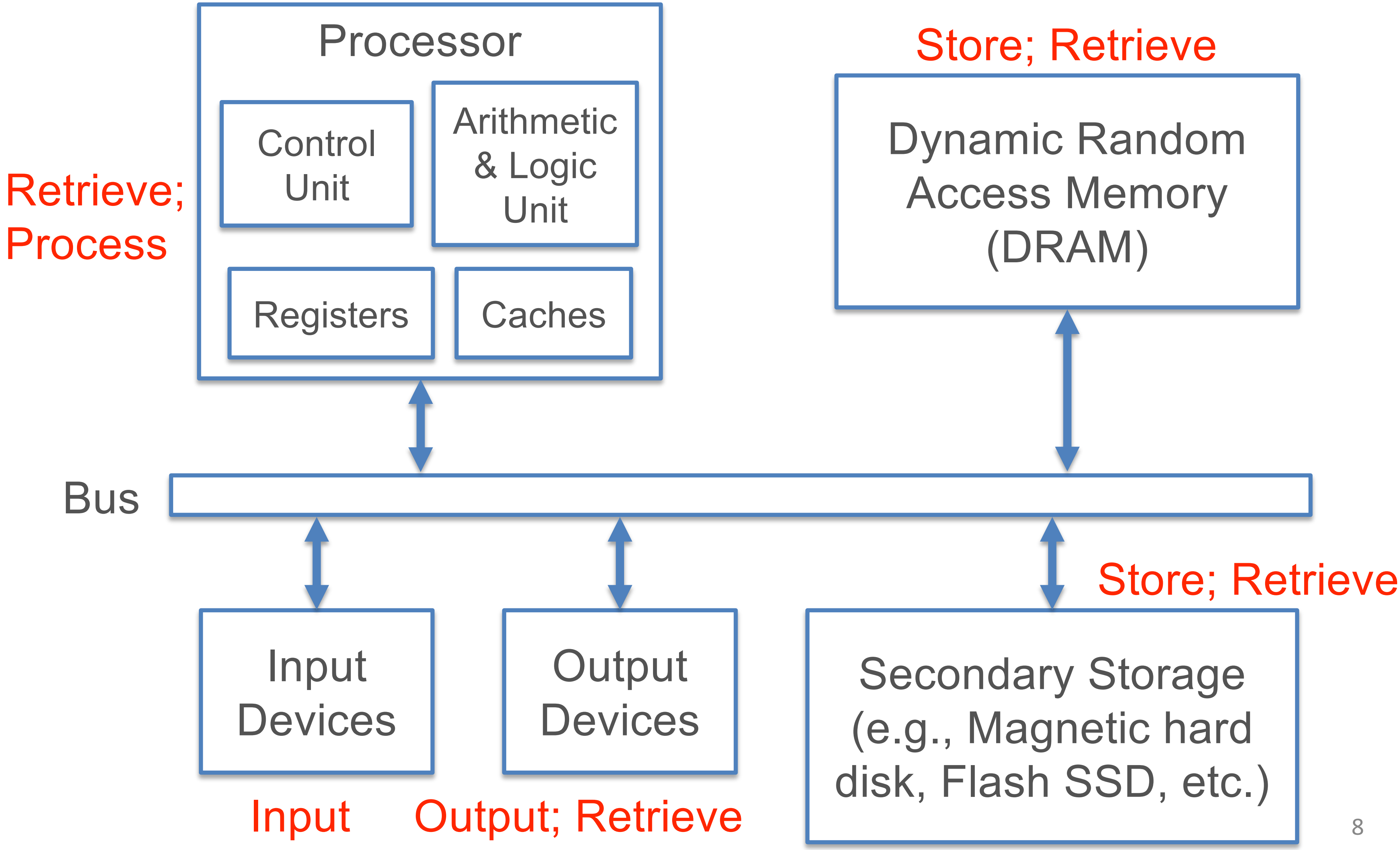
Software:
Programs (instructions) and data

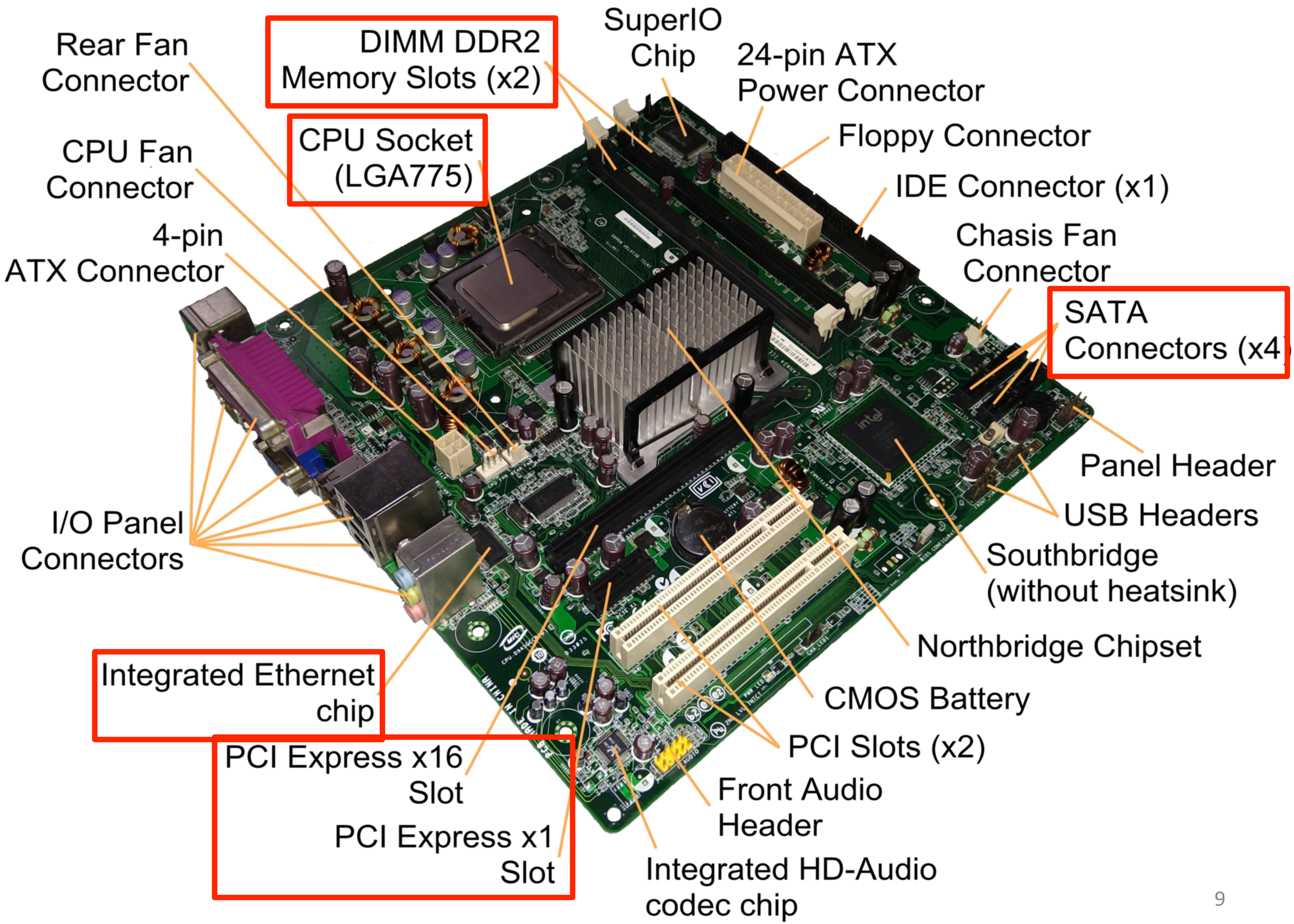
https://www.webopedia.com/TERM/C/computer.html

Key Parts of Computer Hardware

- **Processor** (CPU, GPU, etc.)
  - Hardware to orchestrate and execute *instructions* to manipulate *data* as specified by a *program*
- **Main Memory** (aka Dynamic Random Access Memory)
  - Hardware to store *data* and *programs* that allows very fast location/retrieval; byte-level *addressing* scheme
- **Disk** (aka secondary/persistent storage)
  - Similar to memory but *persistent*, *slower*, and higher capacity / cost ratio; various addressing schemes
- **Network** interface controller (NIC)
  - Hardware to send data to / retrieve data over network of interconnected computers/devices

Abstract Computer Parts and Data




Rear Fan Connector
DIMM DDR2 Memory Slots (x2)
SuperIO Chip
24-pin ATX Power Connector
Floppy Connector
IDE Connector (x1)
CPU Fan Connector
CPU Socket (LGA775)
4-pin ATX Connector
Chasis Fan Connector
SATA Connectors (x4)
Panel Header
USB Headers
I/O Panel Connectors
Southbridge (without heatsink)
Northbridge Chipset
Integrated Ethernet chip
CMOS Battery
PCI Slots (x2)
PCI Express x16 Slot
Front Audio Header
PCI Express x1 Slot
Integrated HD-Audio codec chip

Key Aspects of Software

- **Instruction**
  - A command understood by hardware; finite vocabulary for a processor: Instruction Set Architecture (ISA); bridge between hardware and software
- **Program** (aka code)
  - A collection of instructions for hardware to execute
- **Programming Language (PL)**
  - A human-readable *formal* language to write programs; at a much higher level of *abstraction* than ISA
- **Application Programming Interface (API)**
  - A set of functions (“interface”) exposed by a program/set of programs for use by humans/other programs
- **Data**
  - Digital representation of *information* that is stored, processed, displayed, retrieved, or sent by a program

Main Kinds of Software

- **Firmware**
  - Read-only programs “baked into” a device to offer basic hardware control functionalities
- **Operating System (OS)**
  - Collection of interrelated programs that work as an intermediary platform/service to enable application software to use hardware more effectively/easily
  - Examples: Linux, Windows, MacOS, etc.
- **Application Software**
  - A program or a collection of interrelated programs to manipulate data, typically designed for human use
  - Examples: Excel, Chrome, PostgreSQL, etc.

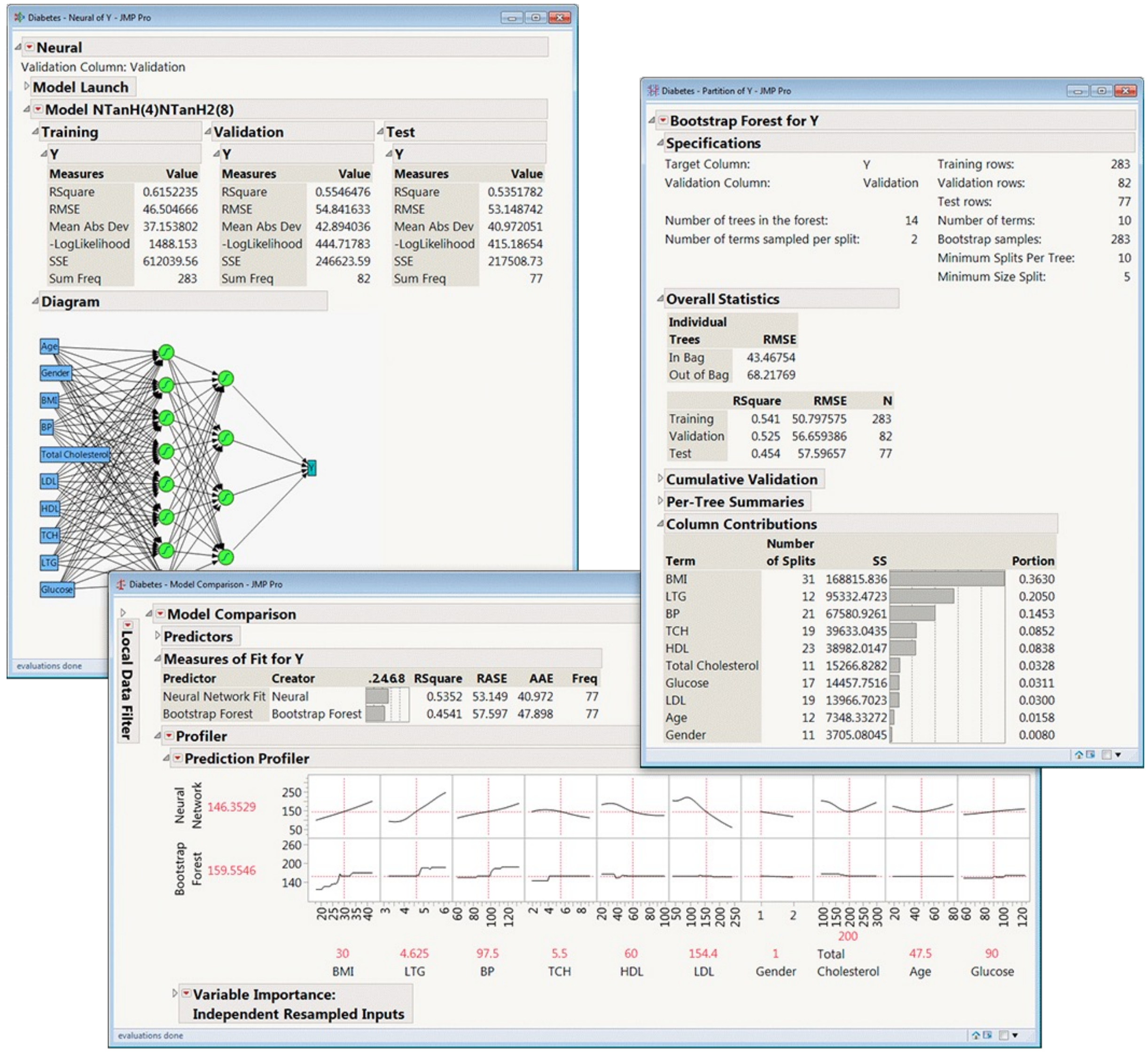
Outline

- Basics of Computer Organization
  - Digital Representation of Data
  - Processors and Memory Hierarchy
- Basics of Operating Systems
  - Process Management: Virtualization; Concurrency
  - Filesystem and Data Files
  - Main Memory Management
- Persistent Data Storage

Q: *But why bother learning such low-level computer sciencey stuff in Data Science?*

Luxury of "Statisticians"/"Analysts" of Yore

- **Methods:** Sufficed to learn just math/stats, maybe some SQL
- **Types:** Mostly tabular (relational), maybe some time series
- **Scale:** Mostly small (KBs to few GBs)
- **Tools:** Simple GUIs for both analysis and deployment; maybe an R-like console



https://www.jmp.com/en_au/offers/jmp-pro-for-academic-research.html
https://www.technologymagazine.com/data-and-data-analytics/sas-tops-worldwide-advanced-and-predictive-analytics-market-share

Reality of Today’s “Data Scientists”

Data Scientists (DS):

Key team players for success in the Age of Big Data

Data Scientist: Coined as a person with a most versatile skillset to perform all-in-one tasks such as

- handling computationally any size of datasets
- possessing statistical prowess and modeling skills
- understanding and programming with databases
- solving problems; visualize, communicate well
- extracting business value from data

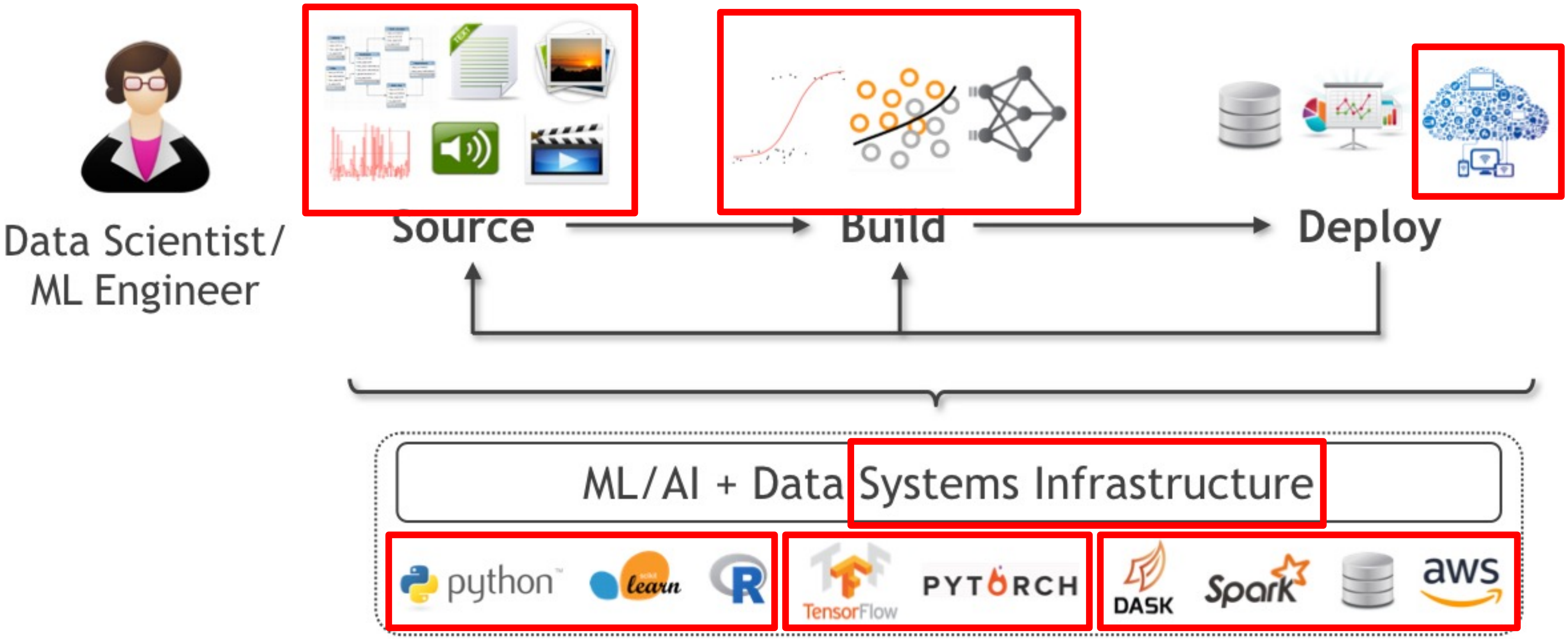
Reality of Today's "Data Scientists"

On basis of this incredibly broad skillset: "Unicorn" status:

Highly sought after, but scarce...

- Unrealistic to train / find practitioners with as broad a range and depth of knowledge
- Ideal teams combine efficiently the talents of multiple data scientists in desired focus areas

Reality of Today's "Data Scientists"



Data acquisition
Data preparation

Feature Engineering
Training & Inference
Model Selection

Serving
Monitoring

Why bother with these in Data Science?

- Basics of Computer Organization
  - Digital Representation of Data
  - Processors and Memory Hierarchy
- Basics of Operating Systems
  - Process Management: Virtualization; Concurrency
  - Filesystem and Data Files
  - Main Memory Management
- Persistent Data Storage

You will face myriad and new data types

Compute hardware is evolving fast

You will need to use new methods on evolving data file formats on clusters / cloud

Storage hardware is evolving fast

Let’s talk about money

Recall our discussion of different data scientist personas

Statistician

Analyst

Product DS

Half-stack DS

Full-stack DS

Machine Learning Engineer

Software Engineer, AI/ML (usually synonymous with MLE)

Typical Levels

Associate Data Scientist

Data Scientist

Senior Data Scientist

Staff Data Scientist <-> Manager DS

(Senior Staff Data Scientist) <-> Senior Manager DS

Principal Data Scientist <-> Director DS, Sr Director, Or VP

(Senior Principal Data Scientist)

(Distinguished / Architect / etc)

(Chief AI Officer

glassdoor

staff statistician | Location

Staff Statistician Salaries United States

For

Overview | **Salaries** | Interviews | Insights | Career Path

How much does a Staff Statistician make?

Experience: All years of Experience

Industry: All industries

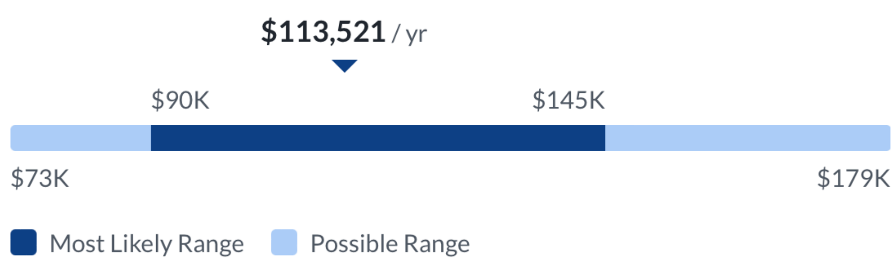
$113,521 / yr

Total Pay

Confident

$93,349 / yr
Base Pay

$20,172 / yr
Additional Pay



The estimated total pay for a Staff Statistician is $113,521 per year in the United States area, with an average salary of $93,349 per year. These numbers represent the median, which is the midpoint of the ranges from our proprietary Total Pay Estimate model and based on salaries collected from our users. The estimated additional pay is $20,172 per year. Additional pay could include cash bonus, commission, tips, and profit sharing. The "Most Likely Range" represents values that exist within the 25th and 75th percentile of all pay data available for this role.

How accurate does $113,521 look to you?

Right | High | Low

glassdoor

staff data scientist | Location

Staff Data Scientist Salaries United States

For Employers

Overview | **Salaries** | Interviews | Insights | Career Path

Upd

How much does a Staff Data Scientist make?

Experience: All years of Experience

Industry: All industries

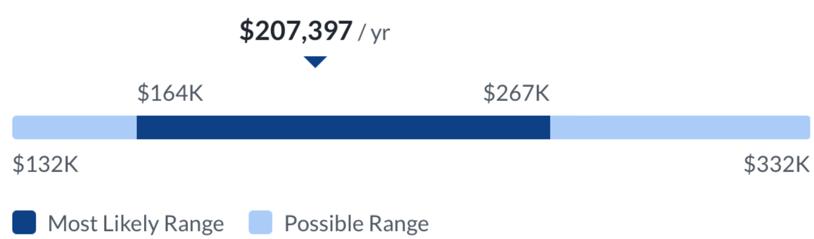
Confident

$207,397 / yr

Total Pay

$148,850 / yr
Base Pay

$58,547 / yr
Additional Pay



The estimated total pay for a Staff Data Scientist is $207,397 per year in the United States area, with an average salary of $148,850 per year. These numbers represent the median, which is the midpoint of the ranges from our proprietary Total Pay Estimate model and based on salaries collected from our users. The estimated additional pay is $58,547 per year. Additional pay could include cash bonus, commission, tips, and profit sharing. The "Most Likely Range" represents values that exist within the 25th and 75th percentile of all pay data available for this role.

How accurate does $207,397 look to you?

Right | High | Low

Total Pay Trajectory

For Staff Data Scientist

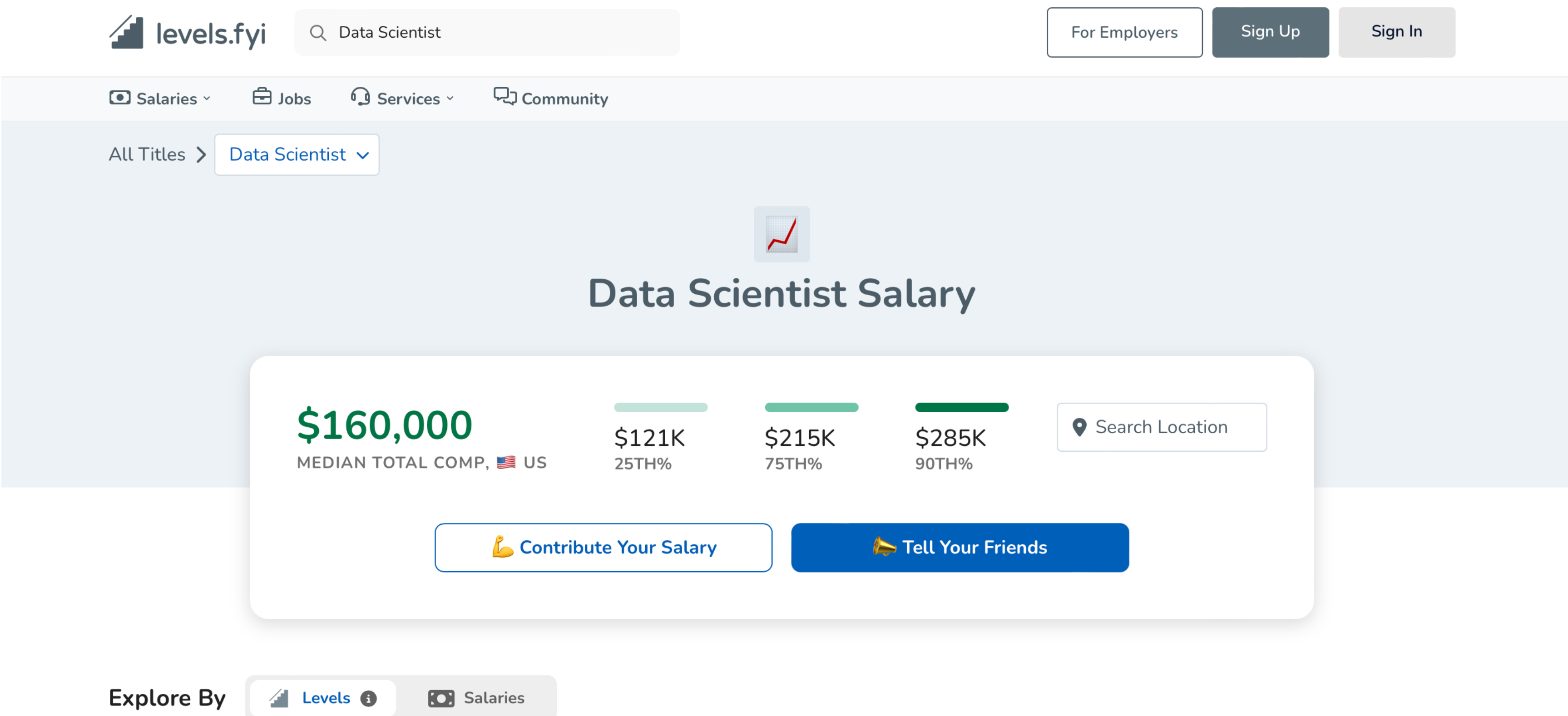
$207,397 /yr
Staff Data Scientist

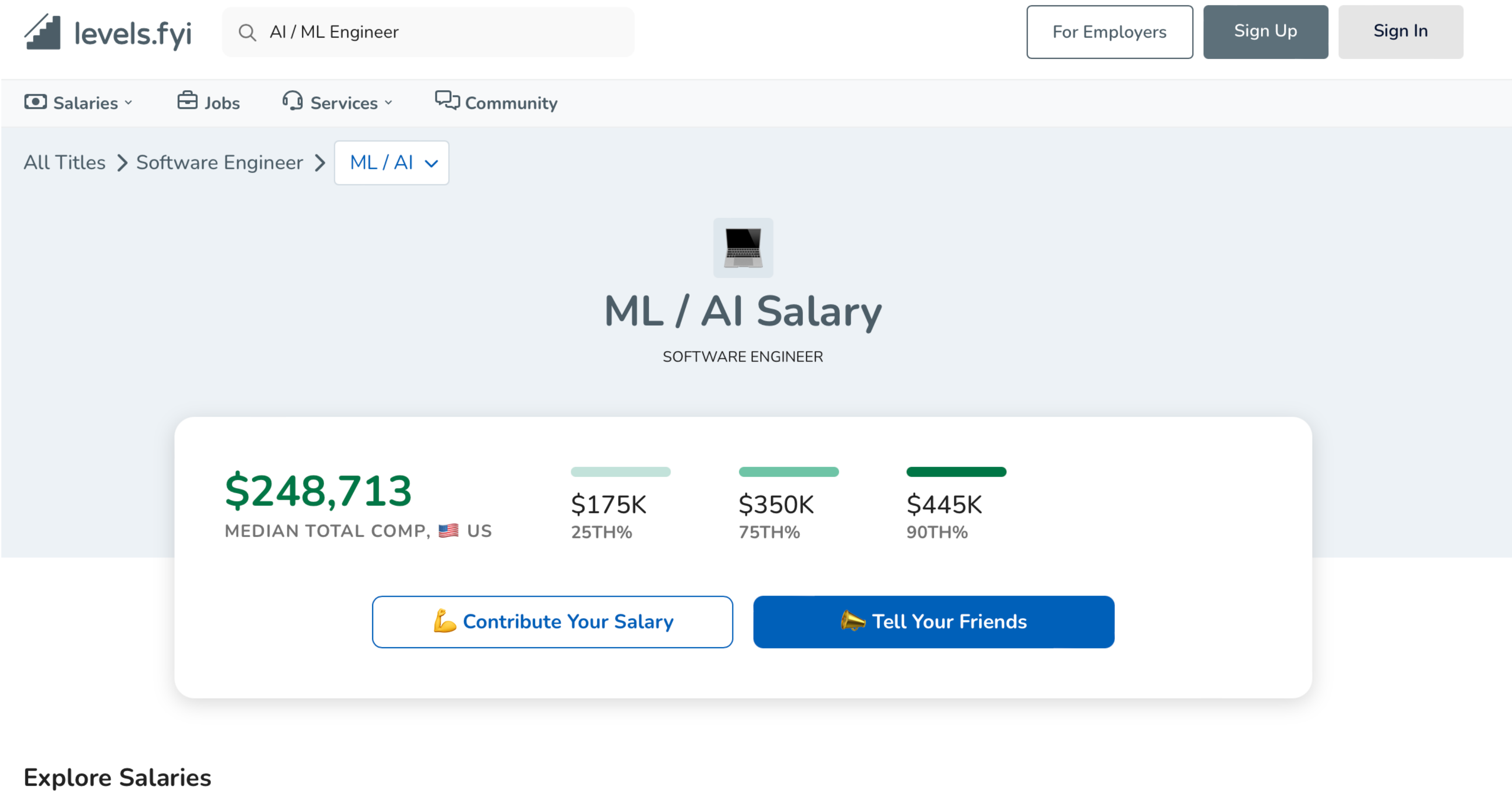
$222,935 /yr
Principal Data Scientist

$240,140 /yr
Senior Principal Data Scientist

See Full Career Path

Download as data table

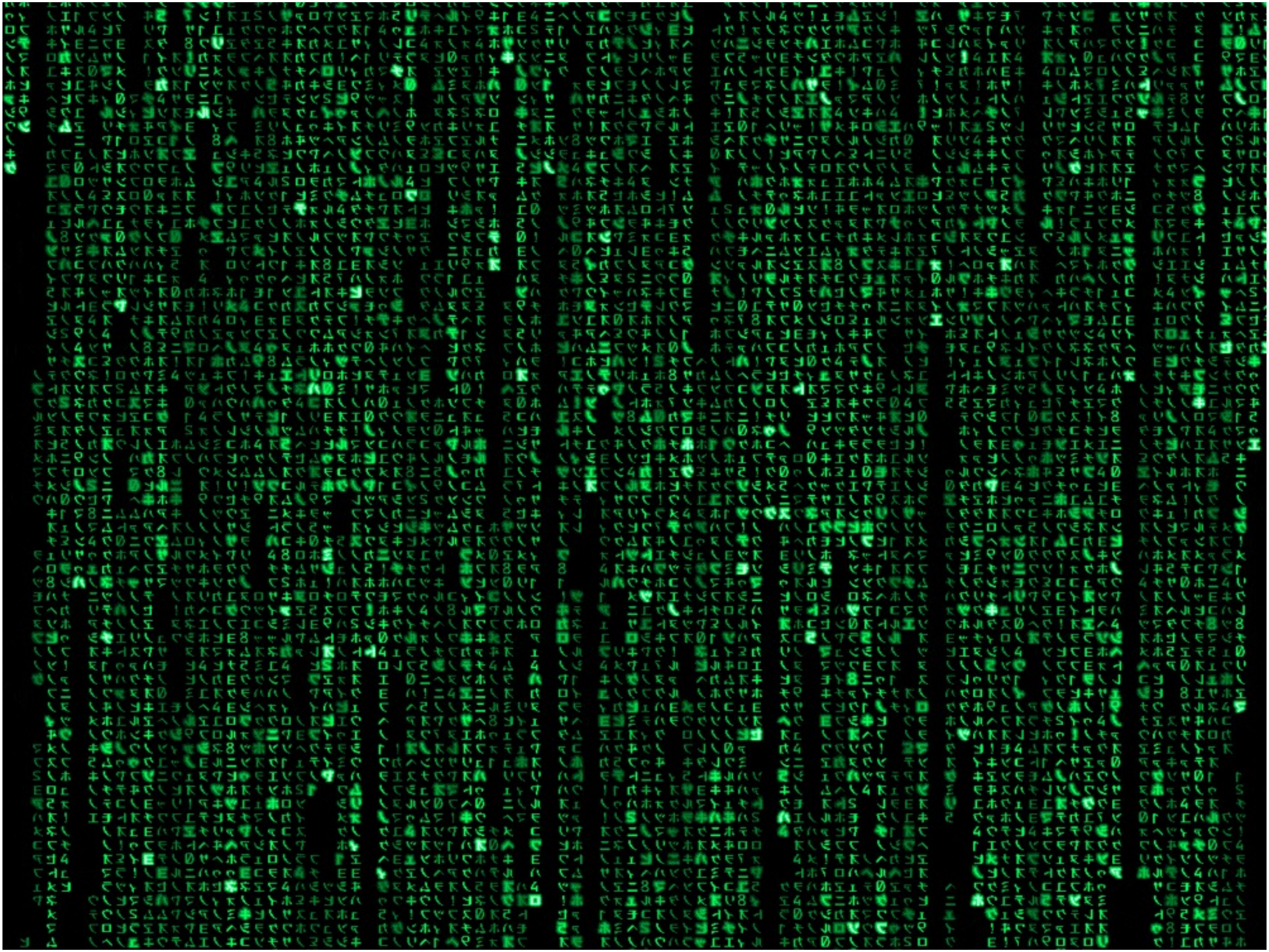

levels.fyi
Data Scientist
For Employers
Sign Up
Sign In
Salaries
Jobs
Services
Community
All Titles
Data Scientist
Data Scientist Salary
$160,000
MEDIAN TOTAL COMP, US
$121K
25TH%
$215K
75TH%
$285K
90TH%
Search Location
Contribute Your Salary
Tell Your Friends
Explore By
Levels
Salaries


levels.fyi
AI / ML Engineer
For Employers
Sign Up
Sign In
Salaries
Jobs
Services
Community
All Titles
Software Engineer
ML / AI
ML / AI Salary
SOFTWARE ENGINEER
$248,713
MEDIAN TOTAL COMP, US
$175K
25TH%
$350K
75TH%
$445K
90TH%
Contribute Your Salary
Tell Your Friends
Explore Salaries

Outline

- Basics of Computer Organization
  - Digital Representation of Data
  - Processors and Memory Hierarchy
- Basics of Operating Systems
  - Process Management: Virtualization; Concurrency
  - Filesystem and Data Files
  - Main Memory Management
- Persistent Data Storage

Q: *What is data?*



Digital Representation of Data

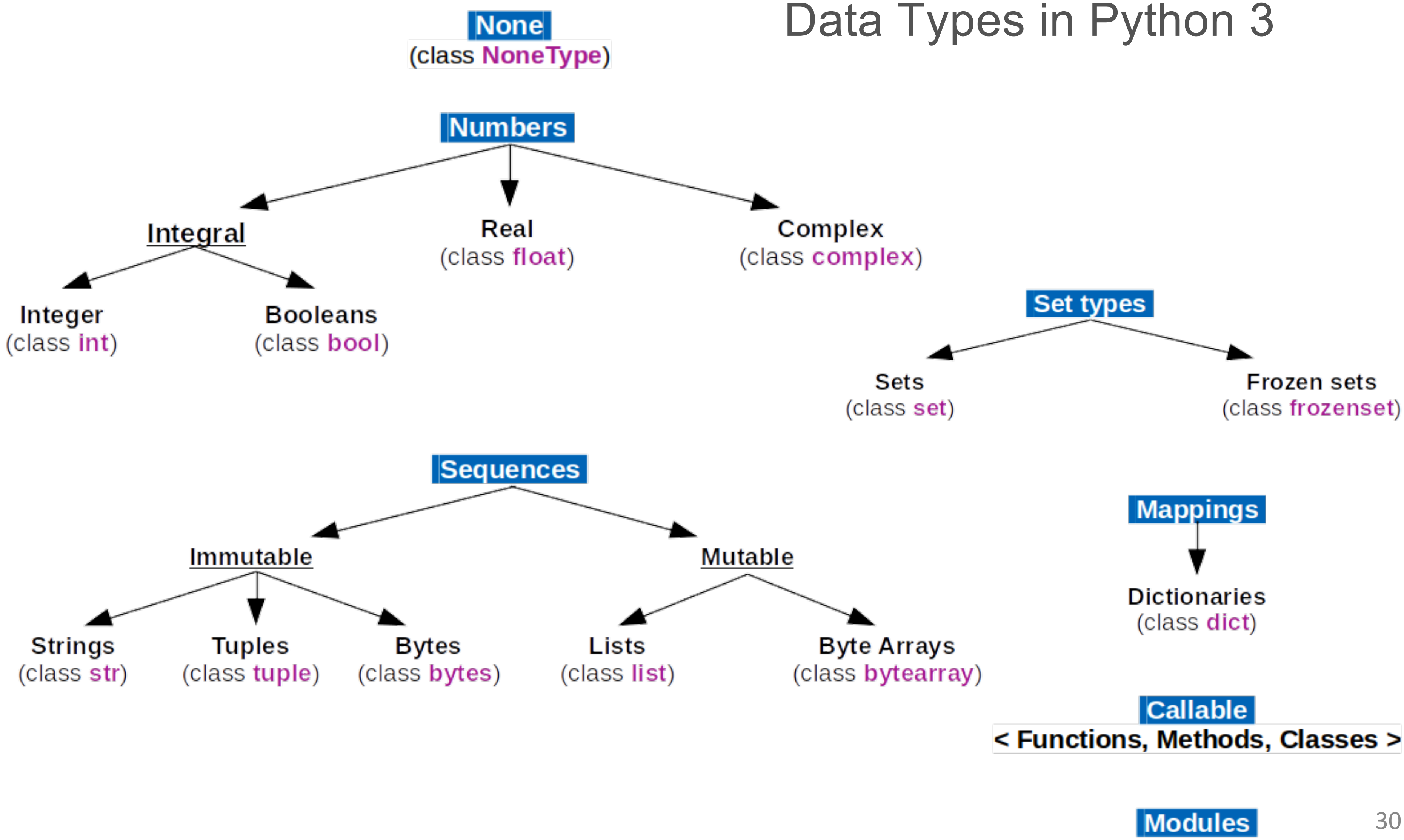
- **Bits:** All digital data are sequences of 0 & 1 (binary digits)
  - Amenable to high-low/off-on electromagnetism

Layers of *abstraction* to interpret bit sequences

- **Data type:** First layer of abstraction to interpret a bit sequence with a human-understandable category of information; interpretation fixed by the programming language (PL).
  - Example common datatypes: Boolean, Byte, Integer, “floating point” number (Float), Character, and String
- **Data structure:** A second layer of abstraction to *organize* multiple instances of same or varied data types as a more complex object with specified properties
  - Examples: Array, Linked list, Tuple, Graph, etc.

Digital Representation of Data

Data Types in Python 3



Digital Representation of Data

- The *size* and *interpretation* of a data type depends on PL
- A **Byte** (B; 8 bits) is typically the basic unit of data types
- **Boolean**:
  - Examples in data sci.: Y/N or T/F responses
  - Just 1 bit needed but actual size is almost always 1B, i.e., 7 bits are wasted! (***Q:*** *Why?*)
- **Integer**:
  - Examples in data science: # of friends, age, # oflikes
  - Typically 4 bytes; many variants (short, unsigned, etc.)
  - Java *int* can represent -2^{31} to $(2^{31} - 1)$;
  - C *unsigned int* can represent 0 to $(2^{32} - 1)$;
  - Python3 *int* is effectively unlimited length (PL magic!)

Digital Representation of Data

***Q:** How many unique data items can be represented by 3 bytes?*

- Given k bits, we can represent 2^k unique data items
- 3 bytes = 24 bits => 2^{24} items, i.e., 16,777,216 items
- Common approximation: 2^{10} (i.e., 1024) ~ 10^3 (i.e., 1000); kibibyte (KiB) = 1024 bytes, vs kilobyte (KB) = 1000 bytes

***Q:** How many bits are needed to distinguish 97 unique items?*

- For k unique items, invert the exponent to get $\log_2(k)$
- But #bits is an integer! So, we only need $\lceil \log_2(k) \rceil$
- So, we only need the next higher power of 2
- So... 7 bits

Digital Representation of Data

Q: *How to convert from decimal to binary representation?*

1. Given decimal n
   - if n is power of 2 (say, 2^k), put 1 at bit position k; if k=0, stop; else pad with trailing 0s till position 0
   - if n is not power of 2, identify the power of 2 just below n (say, 2^k); #bits is then k; put 1 at position k
2. Reset n as n - 2^k; return to Steps 1-2
3. Fill remaining positions in between with 0s

	7	6	5	4	3	2	1	0	Position/Exponent of 2
Decimal	128	64	32	16	8	4	2	1	Power of 2
5_{10}						1	0	1	
47_{10}			1	0	1	1	1	1	
163_{10}	1	0	1	0	0	0	1	1	
16_{10}				1	0	0	0	0	

Q: *Binary to decimal?*

Digital Representation of Data

What about if we have fractional component?

Same idea. But we append from the right. First, use the previous algorithm to calculate the left side of the decimal place. Let's think of it as a python function if that is more intuitive to some:

```python
def decimal_to_binary_fraction(n, precision):
    k = -1
    binary_string = "."

    while n != 0 and len(binary_string) < precision:
        if n >= 2**k:
            binary_string += "1"
            n -= 2**k
        else:
            binary_string += "0"
        k -= 1

    return binary_string
```

Why the heck are we talking about this?

In machine learning, we work with large amounts of numerical data represented in binary format. This has implications for:

- Memory efficiency
- Numerical accuracy
- Data compression
- Hardware compatibility: some GPUs may support different levels of precision