

# 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

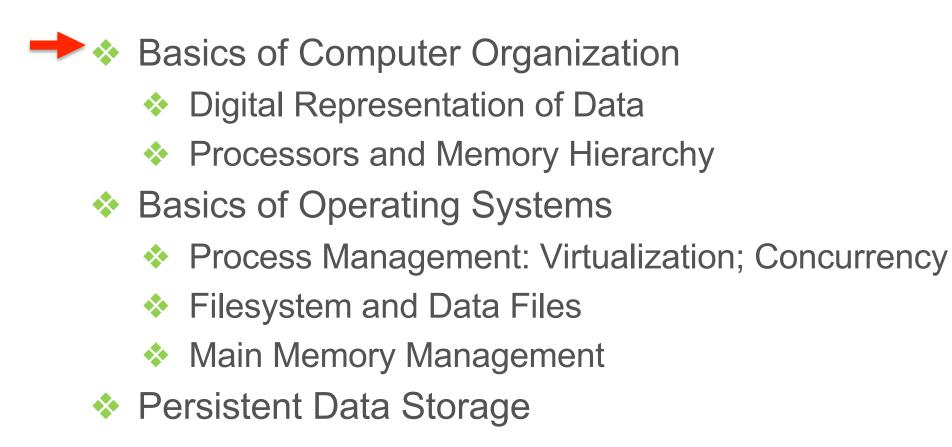
#### can store, retrieve, and process digital data.



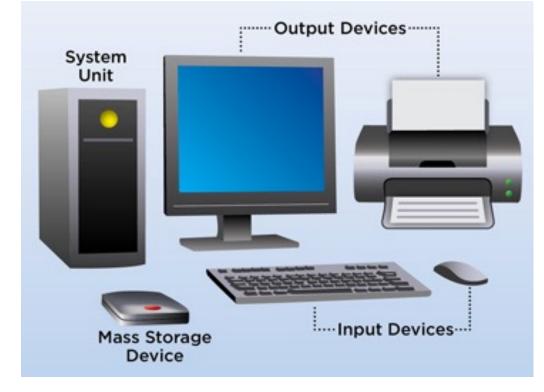
Peter Naur (1928-2016)

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

## Outline



### Parts of a Computer



#### Hardware:

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

**Software:** Programs (instructions) and data

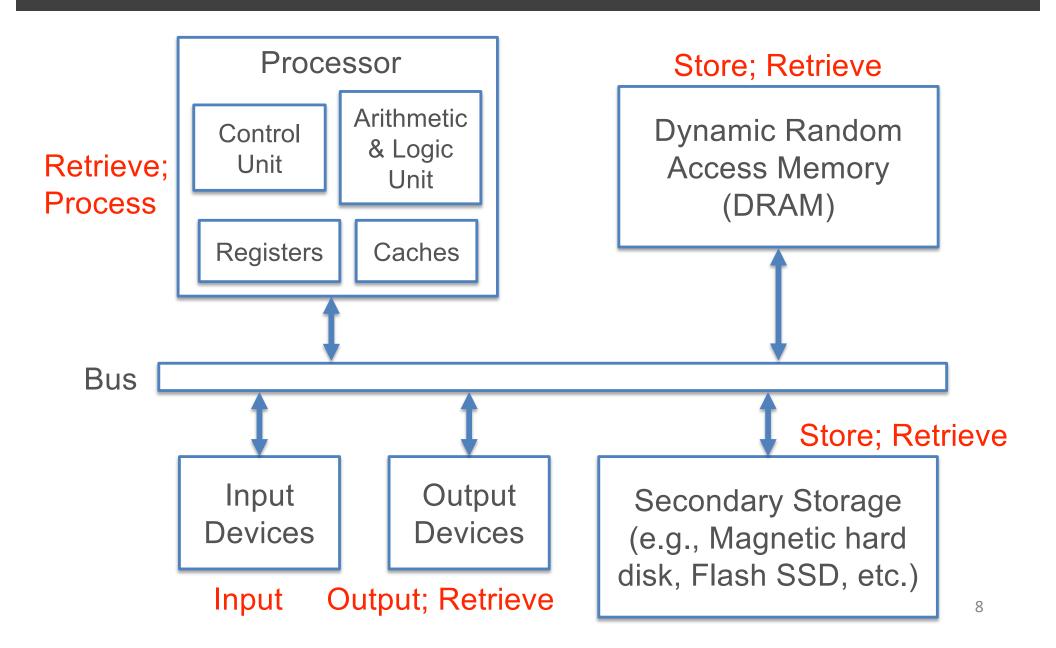
#### https://www.webopedia.com/TERM/C/computer.html

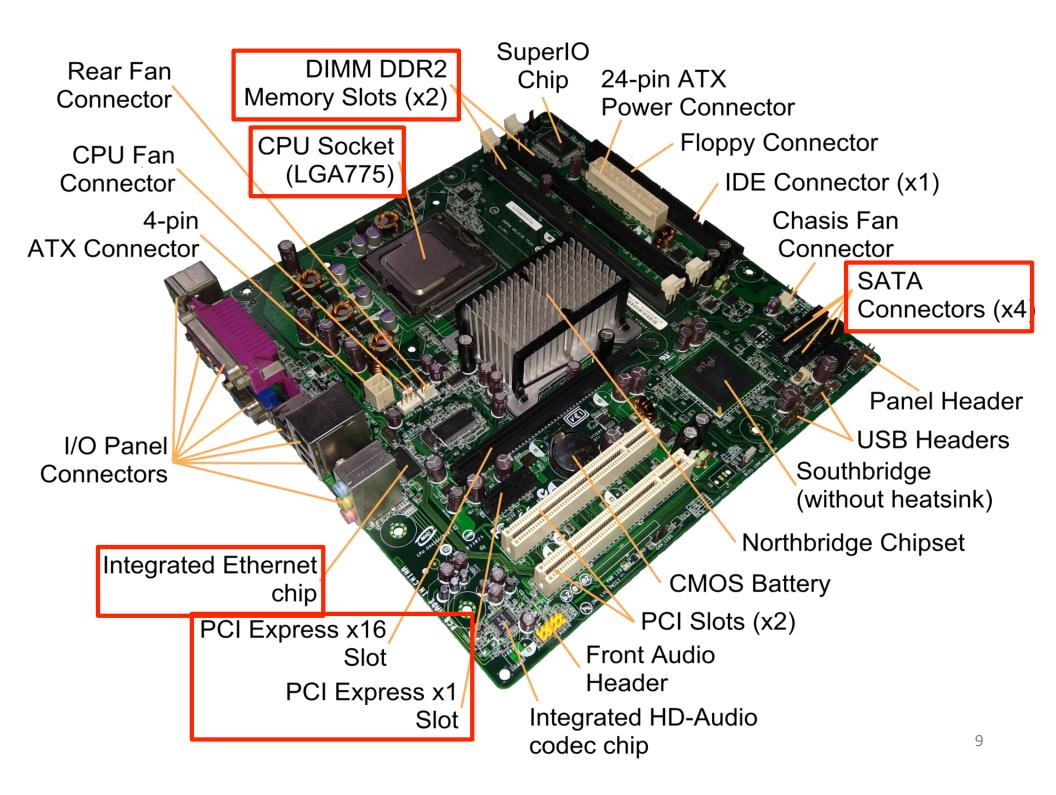
# Key Parts of Computer Hardware

#### 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





# 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

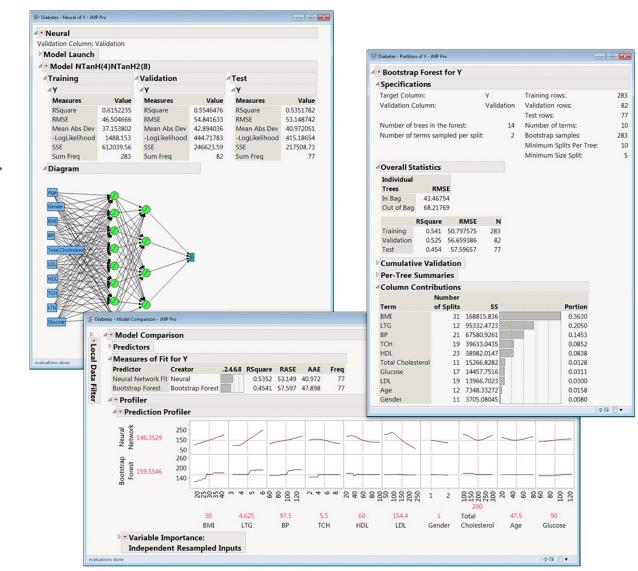


- Digital Representation of Data
- Processors and Memory Hierarchy
- Basics of Operating Systems
  - Process Management: Virtualization; Concurrency
  - Filesystem and Data Files
  - Main Memory Management
- Persistent Data Storage

# **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

### 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"

#### **MODERN DATA SCIENTIST**

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

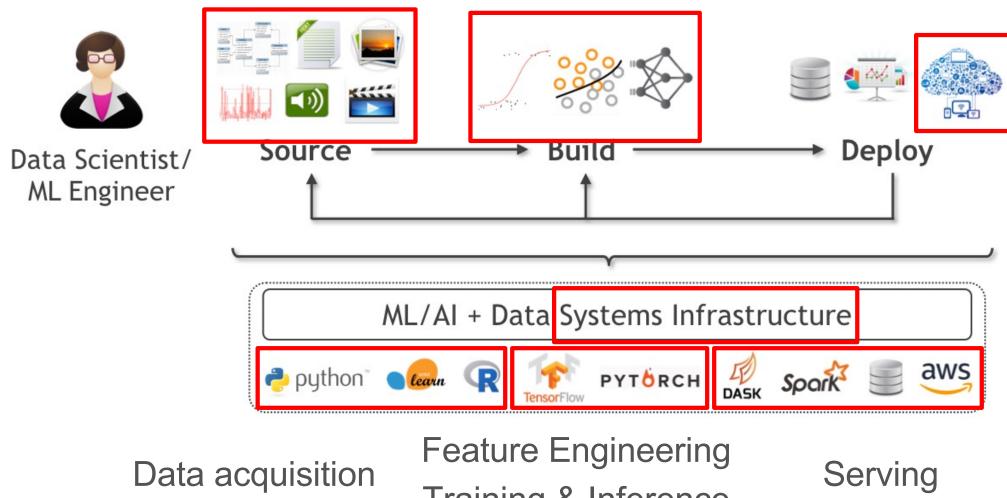


On basis of this incredibly broad skillset: "Unicorn" status:

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 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 Compute hardware
- Basics of Operating Systems
  - Process Management: Virtualization; Concurrency
  - Filesystem and Data Files
  - Main Memory Management
- Persistent Data Storage

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

You will face myriad

and new data types

is evolving fast

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, Al/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 Al Officer

glassdoor			staff	statistician	O Location	Q	
$\equiv$ > :	Staff Sta	tistician Sa	alaries u	Jnited States $\checkmark$			For
Overview	Salaries	Interviews	Insights	Career Path			

#### How much does a Staff Statistician make?



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	令 High	₽ Low					

glass	glassdoor			taff data scientist	O Location	Q
>	Staff Da	ta Scientis	t Salarie	S United States ∨		For Employers
Overview	Salaries	Interviews	Insights	Career Path		

#### How much does a Staff Data Scientist make?

#### For Staff Data Scientist Experience Industry \$207,397/yr $\sim$ All years of Experience $\sim$ All industries Staff Data Scientist Confident \$222,935/yr \$207,397/yr Principal Data Scientist **Total Pay** \$240,140/yr Ó Senior Principal Data Scientist **\$148,850**/yr \$207,397 / yr Base Pay See Full Career Path \$164K \$267K Download as data table \$58.547 / yr Additional Pay \$132K \$332K Most Likely Range Possible Range

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.

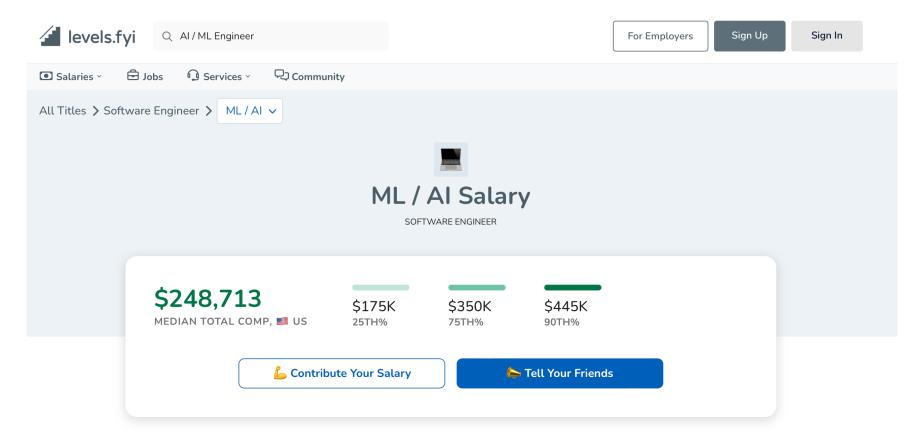
How accurate does \$207,397 look to you?							
😇 Right	令 High	<sup>¶</sup> Low					

23

Upd

**Total Pay Trajectory** 

levels.fy	Q Data Scientist		For Employers Sign Up	Sign In
Salaries ×	🗄 Jobs 🕠 Services 🗸 🖓 Community			
All Titles > Da	ta Scientist 🗸	~		
	Data	Scientist Salary		
	<b>\$160,000</b> MEDIAN TOTAL COMP, ■ US \$121K 25TH%	\$215K \$285K 75TH% 90TH%	Search Location	
	Contribute Your Sa	lary 🏷 Tell Your Friend	ls	
Explore By	Levels () E Salaries			



**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

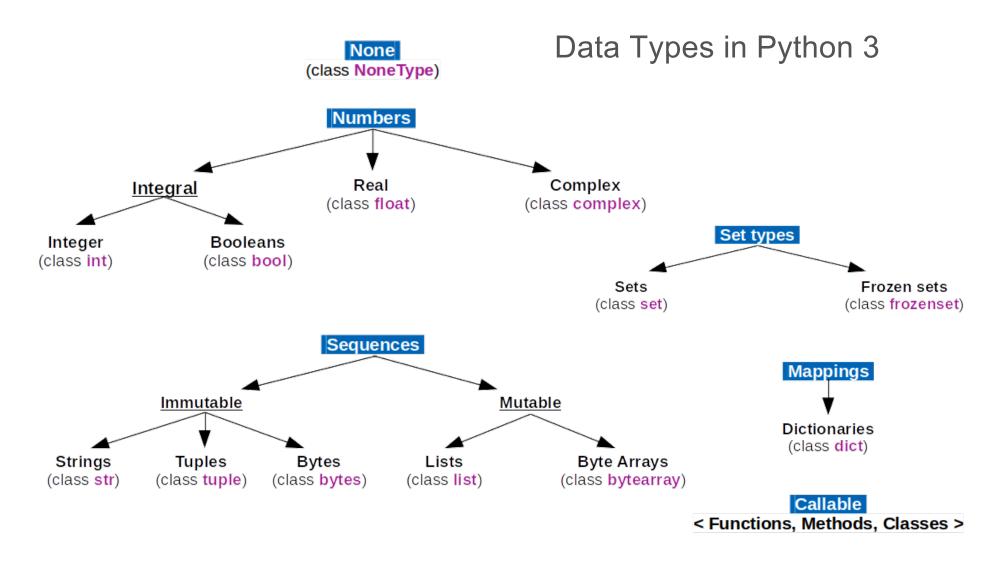
#### **Q:** What is data?

L C AL	54 197 1I/	C I TTC I Her	ホノ <b>ノノ</b> ノコパコノロマリノノション AI <b>1キ</b> ワイルへ 夢 <b>切 5</b> ホマヱソヤロ	- ALGEDACAN			HAP A MAR
ūć.			23本家マイユキホ ふりと 1オクフィ				
2	9.3 37234+	17+5292078	ワソジホ シオノオ トコレノィザユメカ	ヘオイビックカッ リサ	年 ホヤコムホエ科チソヲシニハソ	77771717 UP471L	1900 C U001
12	11 1942	フカオオへフラオカーオ	システィカテロヒラフテリ 1へしカフ	397796 #F2 2	L MZアホクタンチョリッ5フラ	KATK4 KEADEUG	4 4 C 62E
* 4		トカネブノキョンメニリン トカキラノクオツソ 1オ	◎ ▲「ヱツヱ、オチマカィリロへ ○ ● ● ● □ ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●		マエモンセイモー マンマン シン		
· +			12 Y#FJE 12 999E2Y				
66	7 E 19 1 17 3	173331 4968	スキ ロネミネス ホノペノアダイコ	ノホラエノトッヲクニ	手 」 たさ く そ 激 示 驚 み い 手 小 ら 一 雅	15エラミソヤレ アノダヘメホビ	奉へワ ヨノミ参考
03	アキポチリ ヱ&チ	▶E YC 2000 EC	濫チ ユオリキカソ 9ミリチオイキゥ	トルラヤラクオ ! サーオ	9 小工家ロキメ 雙上ッリン 1 0 2 エ	ロサランピサリ レツ2並オ!ッ	ホリカ コムビデ
1		74 # 87 M8 / 1	メニーナル ヱフヱホヲノムノンアハツ シノ・ヌハヌコ 堅 ミヌハウフ ダノヘヤ	ホイズ エマホワツ るいホ	◎ カシノルノヌフニサラクオへシ 501 m L C ~ 1 th R L J 1 th R	オオノリへマク ククキャソリオ	LOSE END
			5 IT999EDIN94///IT				13755 12751
	ヨロフンサツロ図工	ホム ビアイル5日日	オキ ヤリノフ ヒムルノホトマカラ	そいをもすみ くち	総オムトホホエュオガホウ取ハ 芝	1リホキハミヤ #フリコオクヤ	111E 5774
			76 P729 1901376973				
			〇ユ ! ヶ龍牛 ヒトオヤチタムルメ ノラ ホシマシ ホ当ヒ型タウシノメ				
			19 長つシフ ホービュンシン				
		1211\$77 0 2	ノヨ カュオニ クリノショヤノノノ	54 JBLIKAOA	アオキテタオイムエンムニヘモフハ	♥ニヌ 1ホノ ヌヱッシノトノ 7	ホバエ! エヤくの
	700907317		77 27 279911 V\$X338				
			30 11177 #X/0753				
	1 4 7 7 7 7 4 7 7 5 F	してホネホト ヨト くたく	N♥ ヱクルノク フソヲピキオト チワヌ ミヌヌムタ ≫っモョシフカ	ヨフキネハッイワマク	- エメザキマノビバナ こフノヤ ! メナ	ヨエゴナフキハ イブジモキ しりあ	大年代10 マヘ5
			ホノラ オペテプトウキデチしフテク				
	45814 11F	77/277722/19	リ5ヌ ショデカ第オアショシアムワ	1) ノキウをヘアキョ	イサ フキキノィモヨヒュアシフホ	イソムンツクリジェノトスをホ	アロエイエア 1日
			ホノフ ネクビ ヘイハキ 0 フクホホ				
			ヨハィー 7ウヤーフノソファテジ <i>ィリ</i> オプローカハジ ヨッエテヌヒホロリ				
			924 10h 3Lyth 1201				
	キマズコケカネピクフ	DENCARCEAL	ホロヌ チフソホケノェのハノノオコ	キソイオホト 🌽 🎽 ム	74057#X£2 /0//hb	7724721255	
	077E779/9/	1)112541541	NJ5 1711370001739	年間内部長8型 ノス温	OFMEANADER ANITZ	841273017397	
			モヒハーッキト ロソラノマノ・チチン				
			サシニ 指ルヤァコラハコソマルヒヒ ノクオ ルモノルノキウユピホヌダム				
			#/1 ho/518747019				
	6KG12 D34	🗮 KIVIVÇCJI	ト 75 ムコノタノポコ 1Vノジネテ	オチヒホ郭振ノフガノキ	シノエテヤエケッケト5メシダラケ	ヨヒュゥチクヒ 4ヨアカタ ③	9 241 71
5			#91 7295299PPL 29				
			キセラ ヌシムッヒニカノマフ ッラ ソツ修ホュウムツ修動ヒシマオ ッラ				
			サノククオヨルヤメビキヌワスちチの		<b>X</b> 7 1 <b>E77E€7</b> ∧ <b>4</b> 0 <b>777</b>		
	リノアホワウリリエラ	リマキャクリ 4801	ィメアのメラ サユヌソニカノノミホ	コジチエトット オ 🦻	マヘヌトッネ ホノリヨ オツヨノ	ソキノスクククロソタミヒァ	ハホとたとうでと
			アラ 8オフ書 トゥチチー 5ノュシリ		7771+7 7202 7622		
			1り③ソヒ≫ キツコハホホオハ!」 ンノノ≫イィコルゥケマモワテロチロ		ム夢夢クソン ヒルオゥ 愛っ年マ キノキスマロ シロノビ ニシオコ		
i i	E7479 hA3	52549903341	うノルヤマル 8ノ 4ノキホ <b>[</b> フフス 5	トキノクユニ 第 テ幕	72317905639 1504	ハラト イヨアノリュニクツォ	LUJELAZS
3	Z) / # # X3	REICTICESJE	ロホフフフオノヘチイクキシ2ハ93	19931 1 1 RE 1897	YILLA QUEDAN AUJE	4~Z315~0251.49	878477339
			アースポキ事フリオ版クロッホモトと		X 10/525/5332 PAN		
			1 オキァカフチノラ14Fレノスレ				
			A ムルチデノ影うフレロ オホ2ピ M タノホソヌ∧ニ③ エ オォヱ♡		Z 19287E 2E 4765		
			5 77224210 3 94PT				
	I#ヒノギノ ヌノ	JUALERARY Y	ヌリミルニメノノラヘ フ エピハイ	<b>ラポヨハエピカ子 ララ</b>	144/ 19E7 99 BIJ/	💡! ノホノダアノュタホタケ	20 CUR4 3
	▶ 2 人市上北五へ	1/00000012m//	14E800012 V 21E3	りつラハテミピヨ ヘサ	30 F 323 0V XAAA	4 4 7 4 2 チ フ オ フ へ ル ノ ホ	コハ車サスヱナ
	クキンホワノ ジキ		7719#12772 / 1947 P1537787811 1 2955	L2 10 21/11 47		アナノミウホビレオオハダイ	7 2 4 4 7 7 7 7 7 7
	19 0045 727	EJIZYKOLIT	オポューコリサソンへカワ ウミイテ	9+ 4975h mJ	97 99382 IN 27L	0アエサヘコトコンキイ オ	オテァキェア
5	ニダフヤノキ ホノミ	17713 OL 730 CIV71	リフリ バオマクナスエラ ハッマリ	トライユニトデラ トリ	オアテアツハウエ シラ ソハルギ	ノウヱハノムモヲソヱキ 😻	7 10 7 V 🖉 7
			P 1 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2			サシフノソレチハスキラ ミ	E 010 P 18
		ソロ総ェユ翌キア!リ イカノマチヌヒテユフ	ルノニ 囚へつ 図ソルヌノ リラ ラマム ウ毒 マムゥマブ 愛え		オヨラレ 名称オ モク 除チェイ キュノク モホルノエム 2小孝和		
		(1)/ (7)メデメビアエン			キュノン モホルシェム さいゆき フ派 1参 ヤヒッハタ  チェホニ	ションゴガチョンホッシュ	
		ヤノオシアヘキヘトホ	A 0 2 1 1 7 2 7 5 5 6 1			P24432279721	
			■ キナホワリホソ▲ヨヘ ヱ¶ 3			JM+FXLV277ME	
			ハ ルデハケアヘノアケイ ヨヨーダ			↓ <b>ジ</b> ッノオケトテオノヒ♪	
			A ∩漏ヲちノオヲムメヌ墾 キシッミ ■減愛ノネビ!ノウキフゥ ハソビキ			oホフユコノイミクフ ∧ ミヌ¥テ≸ルテユの# ニ	チムシオモコサ ノュト ワン能ノ
			F 5702797145 1#71				202X59F
	727607 3	スタビインタラシへアヨヤス	モハホムチオイホン炒力 アソコフ	ピイラノ常へミホ アラ	97 07 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	X X I 7 3 7 7 4 1 4	/ D 9 7 E 8 🤿
E -	0キョントリ ム	こうサチソ <b>コキノエ</b> ムテ	フガオミタル 1ヘリノコ キシシミ	K€ K/ ĨŬ GU ⊆∰	ロ2ホラオノフのアカキ ルイエユ	6 E11060200	ホキサカユ!ク

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.



- 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!)

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

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

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

- $m \rell$  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

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

- 1. Given decimal n
  - if n is power of 2 (say, 2<sup>k</sup>), 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<sup>k</sup>); #bits is then k; put 1 at position k
- 2. Reset n as n 2<sup>k</sup>; 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
510						1	0	1	
<b>47</b> 10			1	0	1	1	1	1	<b>Q:</b> Binary to decimal?
16310	1	0	1	0	0	0	1	1	
<b>16</b> 10				1	0	0	0	0	33

#### 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:

```
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