

DSC 204A: Scalable Data Systems Winter 2024

Machine Learning Systems

Big Data

Cloud

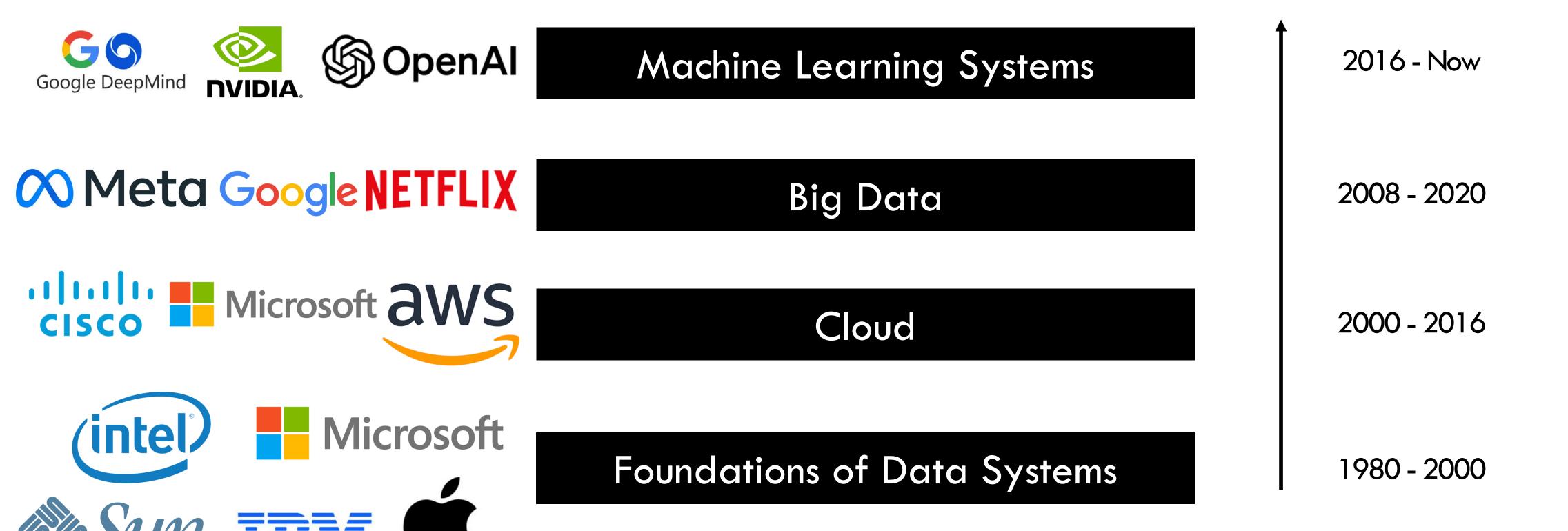
Foundations of Data Systems

Logistics Updates

- Final Exam date (tentative): Friday, March 22, 8 11 am, PT
- Scribe notes:
 - Remember to sign-up
 - Lock down the sheet by next Wed
 - Might do some adjustment to balance student workload (try our best to accommodate preferences)
- Beginning of quarter survey is up, please fill the survey
- Release of the first assignment: eta 1/22, 2 weeks to finish
 - Studying Ray if you have capacity

Class Roadmap: History of Compute and Data

~= History of "which is the most valuable company in tech"



Where We Are

Machine Learning Systems

Big Data

Cloud

Foundations of Data Systems

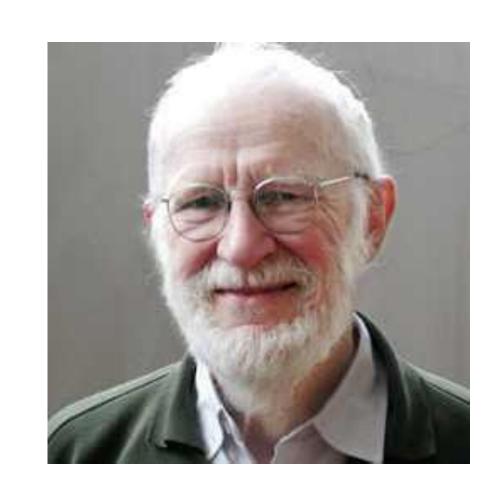
1980 - 2000

Foundation of Data Systems

- Computer Organization
 - Representation of Data
 - Processors, memory, storages
- Operating System Basics
 - Processes: scheduling,
 - File systems
 - Memory management

Q: What is a computer?

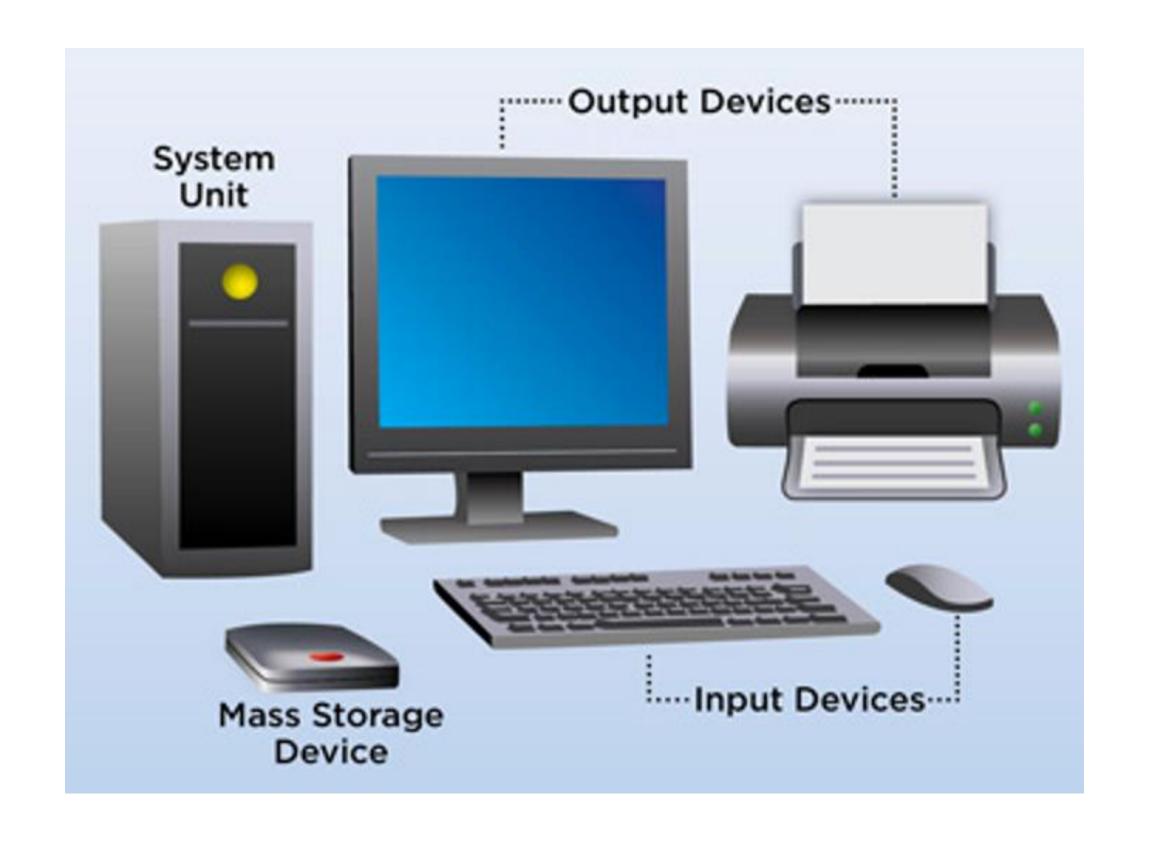
What is a computer?



Peter Naur

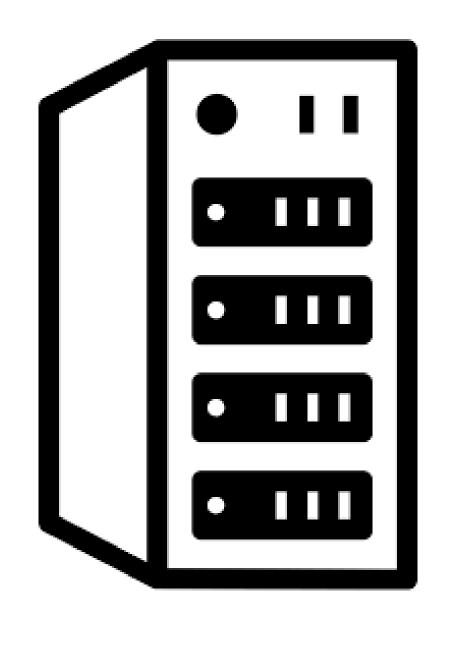
A programmable electronic device that can store, retrieve, and process digital data.

Basics of Computer Organization



- Hardware: The electronic machinery (wires, circuits, transistors, capacitors, devices, etc.)
- Software: Programs
 (instructions) and data

Basics of Computer Organization



To store and retrieve data, we need:

- Disks
- Memory
- Why we need both? (we'll come back in near future)

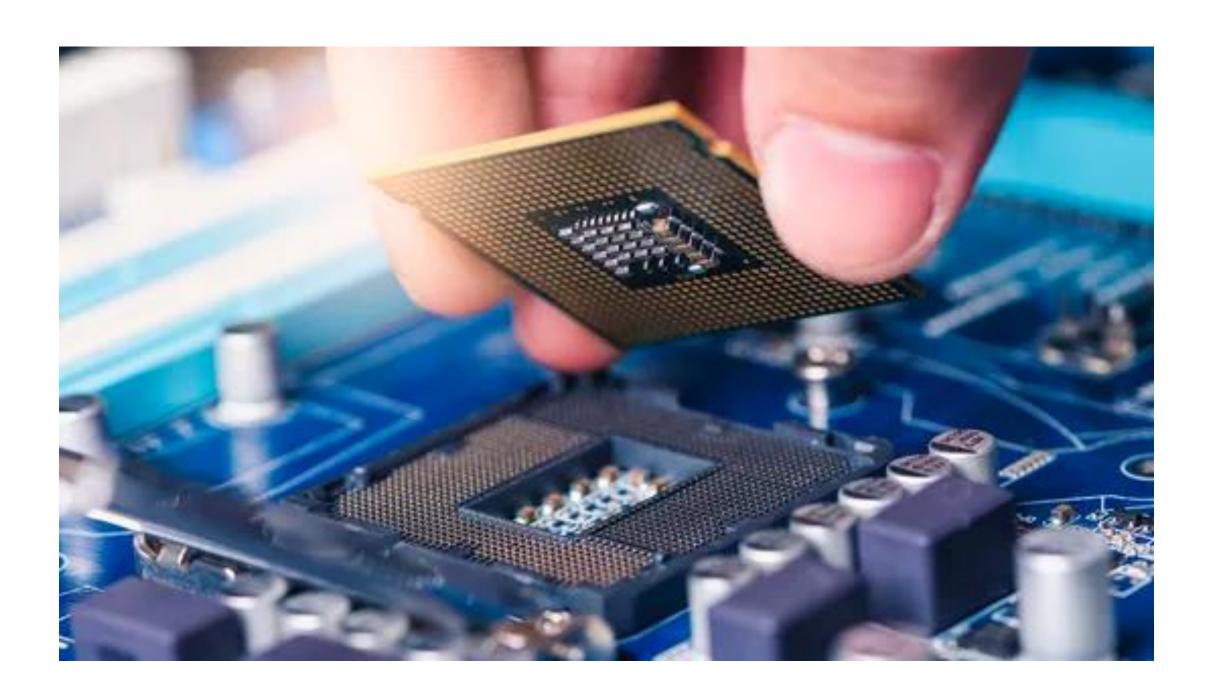
To process data:

Processors: CPU and GPU

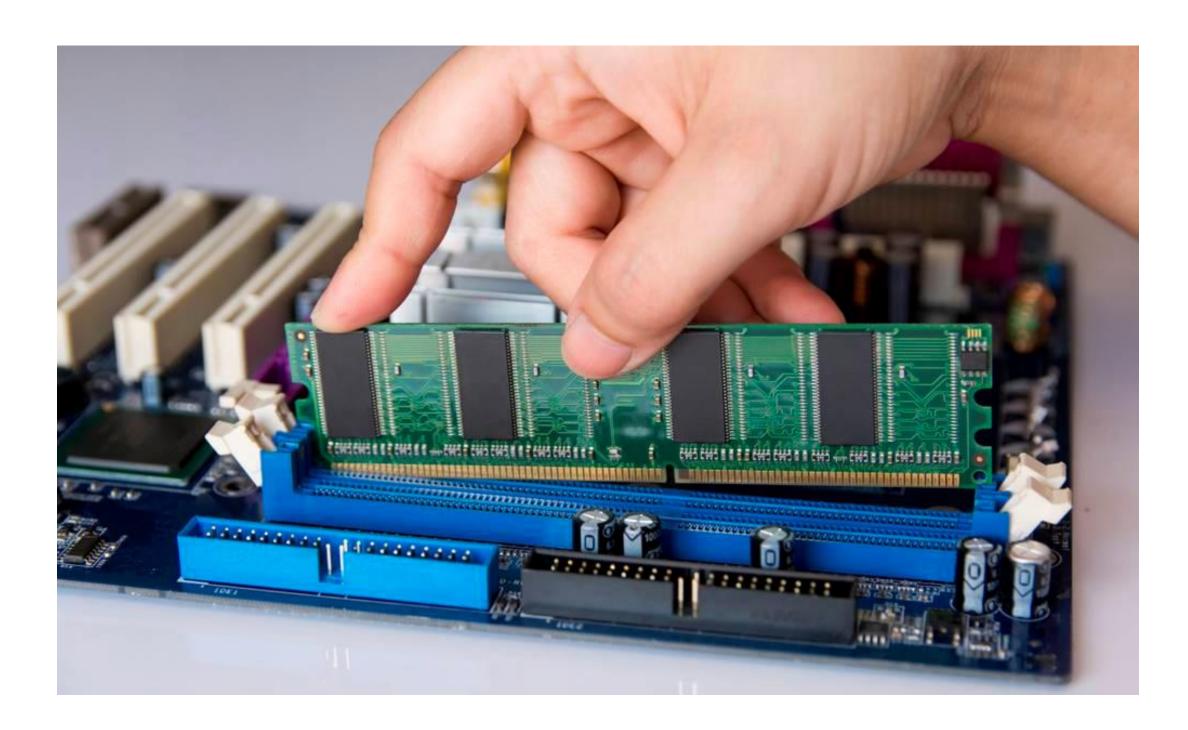
To retrieve data from remote

Networks

- Processor (CPU, GPU, etc.)
 - Hardware to orchestrate and execute instructions to manipulate data as specified by a program



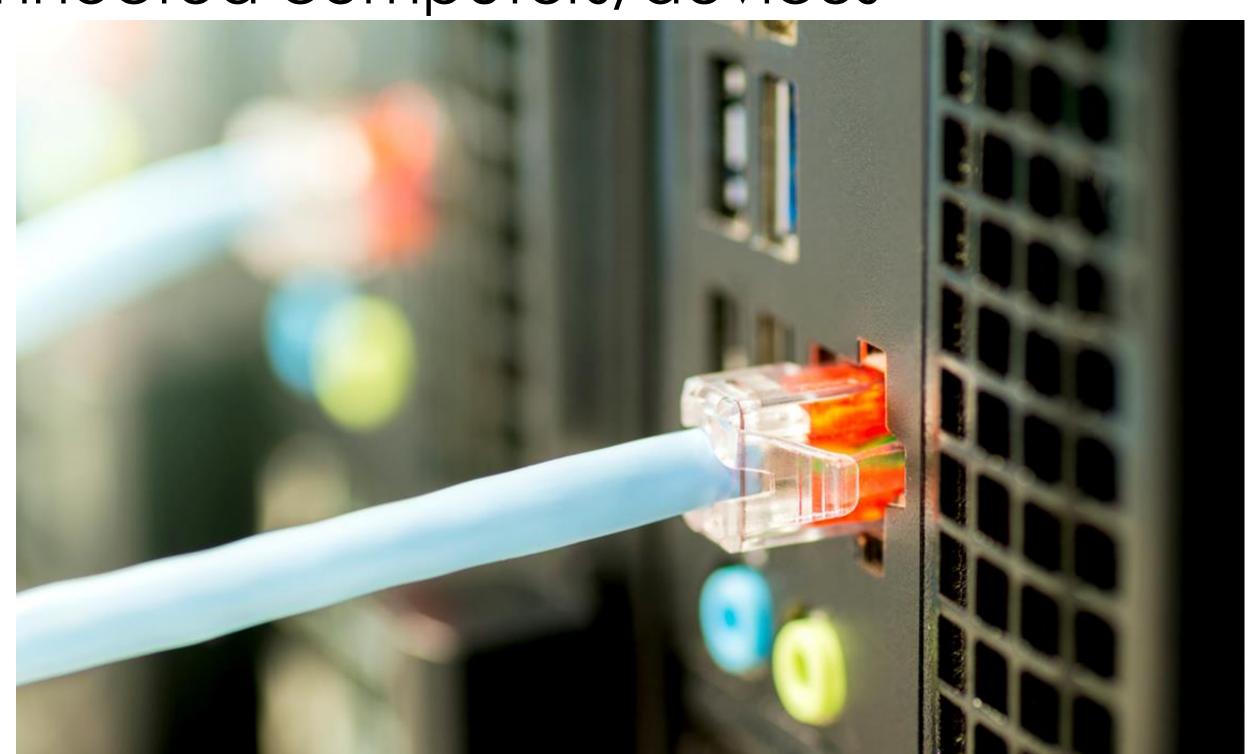
- Main Memory (aka Dynamic Random Access Memory)
 - Hardware to store data and programs that allows very fast location/retrieval; byte-level addressing scheme



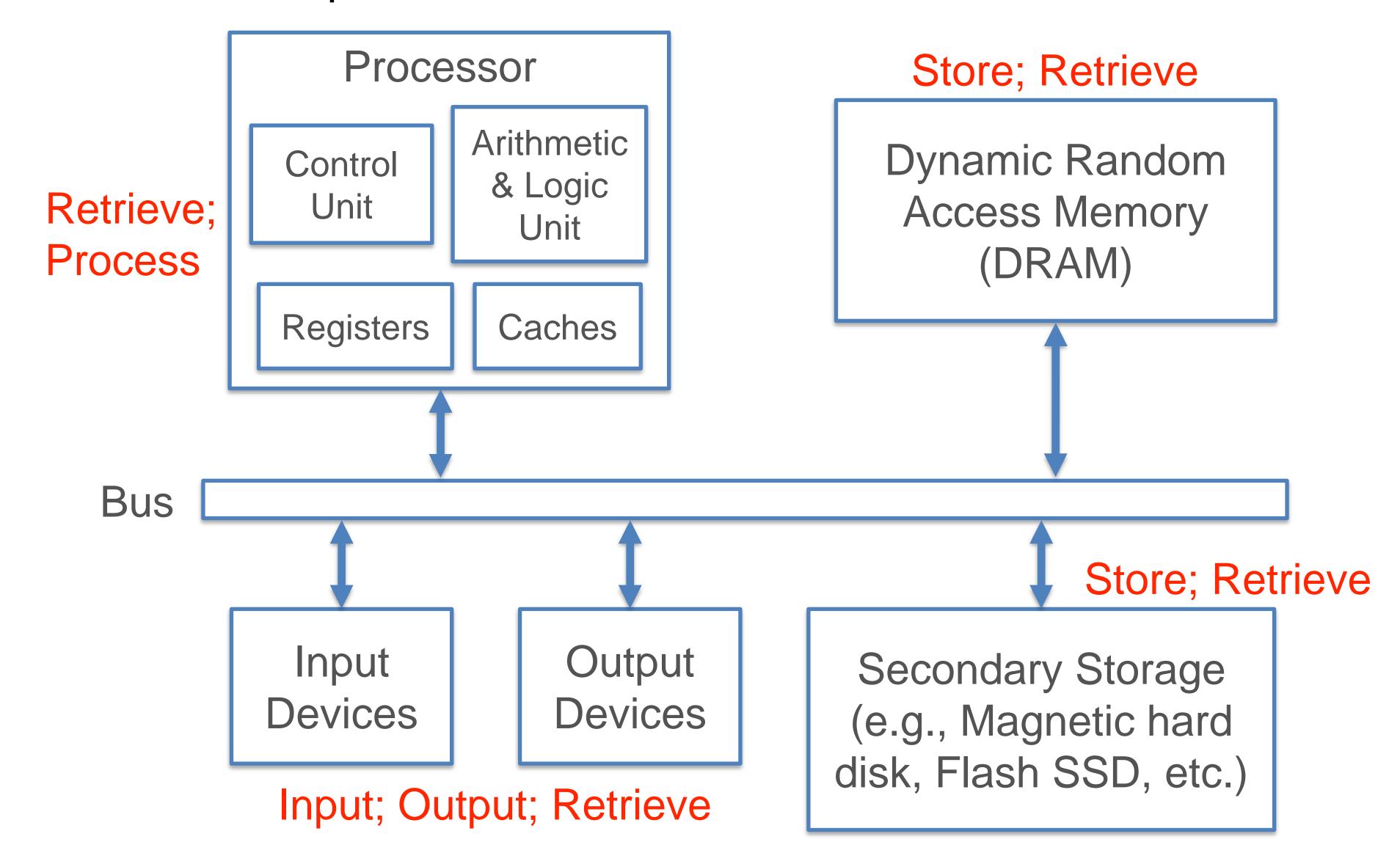
- Disk (aka secondary/persistent storage)
 - Similar to memory but persistent, slower, and higher capacity / cost ratio; various addressing schemes



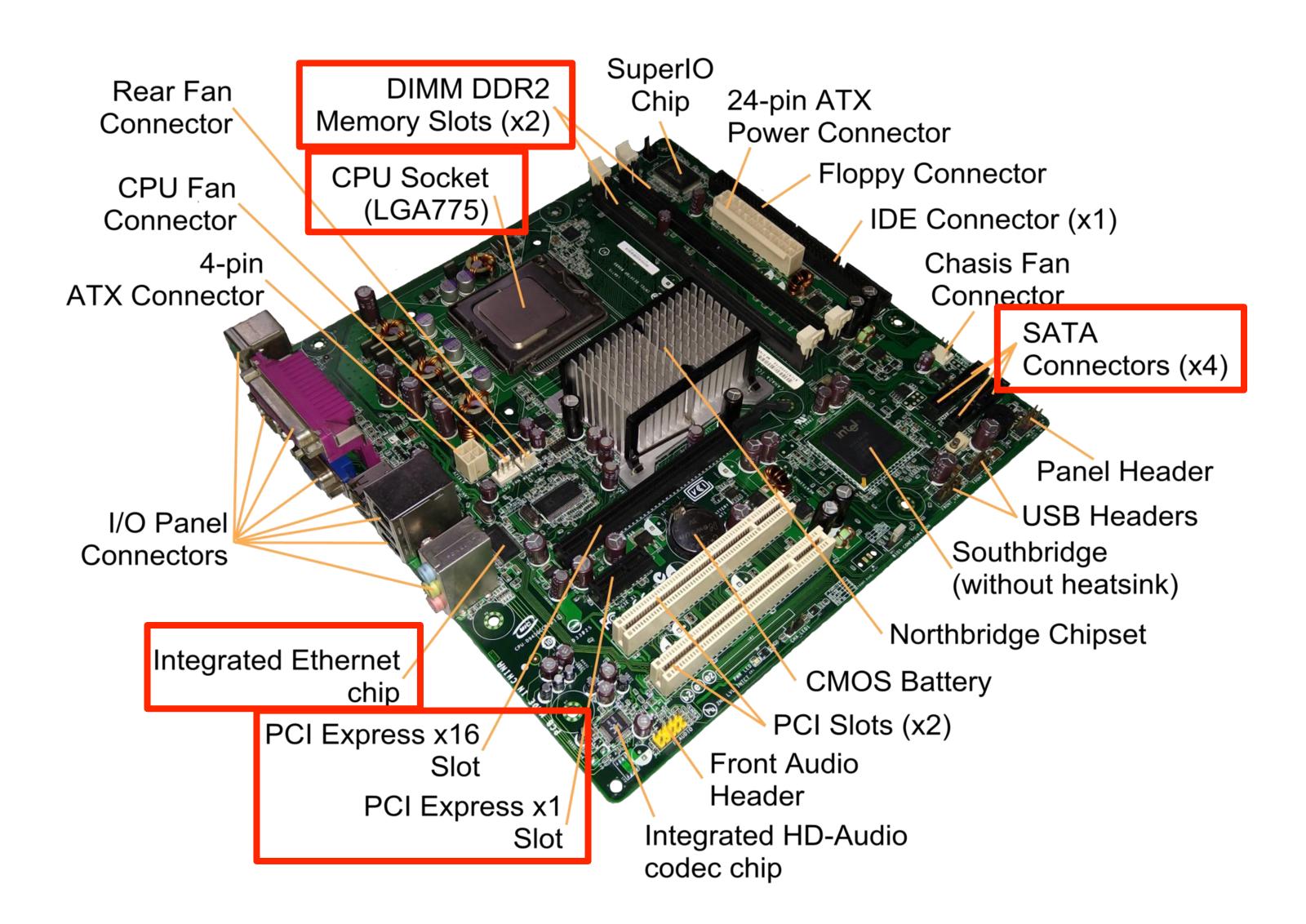
- Network interface controller (NIC)
 - Hardware to send data to / retrieve data over network of interconnected computers/devices



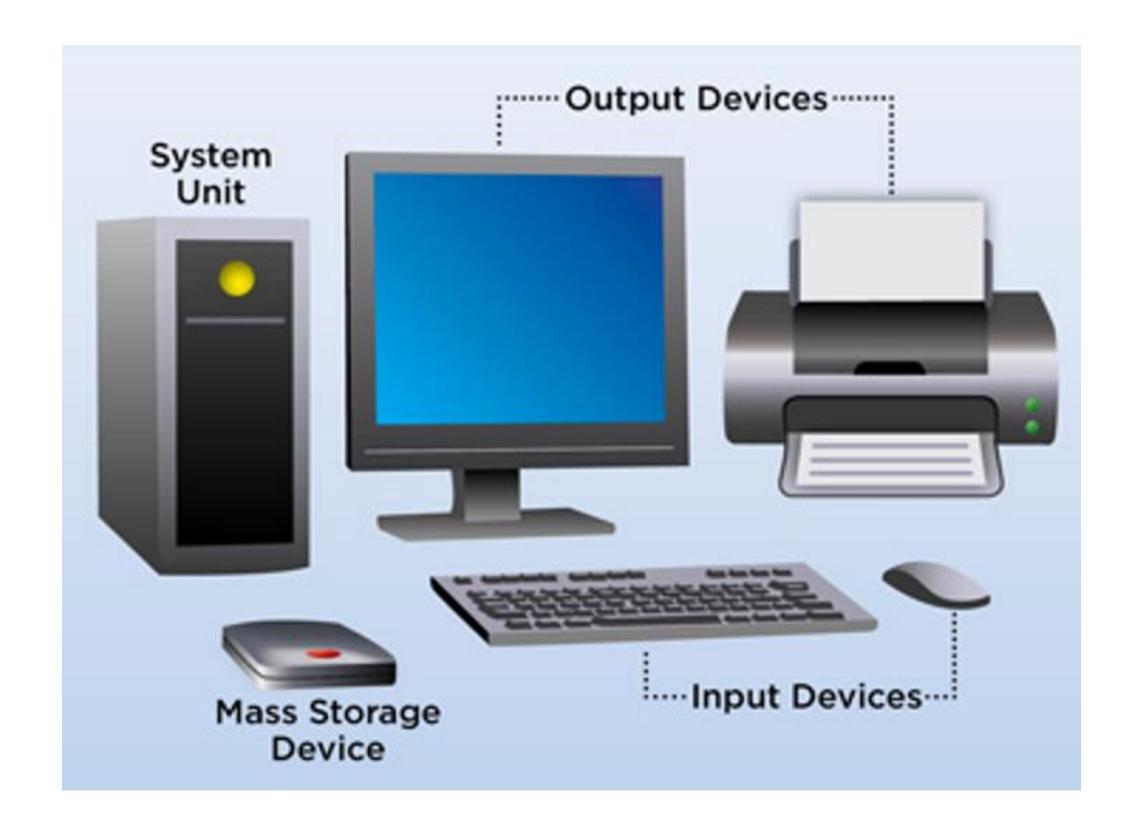
Abstract Computer Parts and Data



In Reality



Parts of a Computer



- Hardware: The electronic machinery (wires, circuits, transistors, capacitors, devices, etc.)
- Software: Programs
 (instructions) 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

Key Aspects of Software

- 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.

Main kinds of Software

- Application Software
 - A program or a collection of interrelated programs to manipulate data, typically designed for human use
 - Examples: Excel, Chrome, PostgreSQL, etc.

Foundation of Data Systems

- Computer Organization
 - Representation of Data
 - Processors, memory, storages
- Operating System Basics
 - Processes: scheduling,
 - File systems
 - Memory management

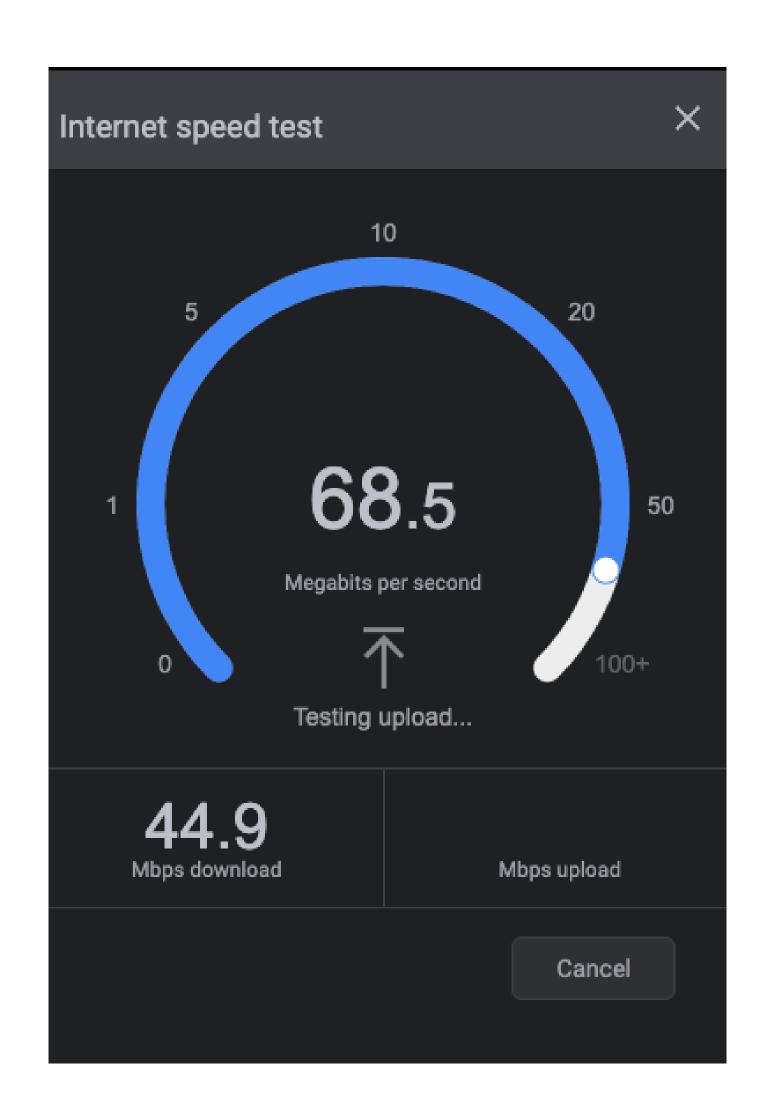
Q: How is data represented in computers?

```
<u>すれソ</u>ーケーオケサソユニー深ソチヤー 7 毛マキ髪・ハボニー・シェムモエキツ 4 ルプドゥラモウ部アノヘニドヨ 1 キタイルヘデ髪 5 ボマヱソヤロラ様ノ 2 エオキン 7 類年・ブソフ 1 エノ
施上灘 ヌーテバニブフェゥオッキーソバイックシ 7ツ キノケ郷サソケワネブメ!ケト・ヒヨマ繋のメバエブイチミコ木兼マイユキホ風でされオウフペワホゥワケュユイザ ヤリー 7ビ さんサホー
ヨマターコーアポケウハュのマ波深ーでオエド【シブレー・終之アホウタンチョリッちフラーヨッノリヒポトフ・ヌニソス等イカッのヒヌラテリコペ!カフカオオペフラオカミオイコーブサムソ難ロー
 ヱ木髪: ぎオシヘエノノロマムロ 6、キョハホカミチ ィーノュヤ!キ霊!イョフトノハエ・ソアホヘノラブフ・クィヅ走・キ!ヱツヱ・オ夢マカィリ個ヘチナノキ 赤ノメニ嬰ソウ!・メ撃ひスキュ:
 つ.髪瓢 サッテリ師ハミデスホノキ !サホ!ヌッ!? ミ@ムポ!ヲ撃精ノネマヲュヤ マレノノモフヒヲオ 惟ソ顔 頼ホヌミ瓢 顔メマミワヨシキカキ類ノウオツソ 1オ加ヲ ブシキユヒッ
 ログラ ソ8テ海製のウエヱト診ェーハノトェノソノホーソファマッグ電オフラチチニウ キヘノタキニカサ熊 サマン ソ禁ァノル !スッツ9艦2ソンテセン マユニラサト歩マのマキフヒヘ
 参へつ ラノミ参加ちェラミソヤレ アノデベメホビミ トハキソト選オ(エラノ5オ!キーノホラエノトゥラウ ニヌ郷 ロネミネヌ ホノペノァヴォュウヌラ熊 キシ動モオソ経マシ豚 ユウラ
 ホリハ 1ムヒテ ロサランヒサリ レッ2単オキッタハ重露ロ本メ蟹キッリヘ102ェートルラヤラクオキサ オポチ エォリキカン9ミリチオイキゥュブリミ フィ ヨポアキボチリ ヱ&チ
     ル2ウェブルブア オオブリヘマクニクケキキソリオ 引力シブルブヌワニサ夢ケオペラ ホヱ エ響ホワッ8ソホメニュナルヱフヱホラブムブンアペツキホ8オ、64.8 ブ 愛コヱロコヤ
 ロサヌ・フィウノ・テアイホラス ゥーフノキキシシノ・5 リニド ミベノヤアビ トノノオ 母・キノ・シウ 夢 ゴキュラシシ 🌽・ヌハヌコ 愛 ミヌハウフ 繋 ノベヤワビ ネ 5 キシュ・愛 フ 渡 ソホビオ ノュフ カ
 オヤン・ヘマケナ・ヱオオビ御ぎり フテキム施ェブ・ヌノウノウオロテフイノッテ・イ・ヱオ・ヘオノフ罪を添き継ち、エサ9ツヒロエハタイポフェ霊ヒリ9罪・フュウワウオシト0ヲマ
 ノノヒ ಶメメイ 1リホキハミヤ ギフ!コメクヤ率オムさホホエュメ繁木ク薬ハ ――――・ナン ヌさキチネハ雲 オキ ヤサノフ ヒさ趣所鑑<mark>につけっこととアニシ ムホラロフンサッロ</mark>標準
 ホシノ 8ッチゥ ニホウキヤ漂ケ 誰与罪ノッミヒロエッ ホーユエユホネゥテチノ ஹ ム!ヒシホセラオホォーケム タオラリ 御夢ロュチム夢グミヒコウソホィ! オオッのロスミムノノオ
 フユュー 愛りらコーショソニ 1ヵキー ボメツナテクノツホロオッタキ!ソビ豚 0 煮ュミューソ 7トロムオホツミツ (0.ユー! ヶ無キーヒトオヤチウ 6ルオデラセノユフ 新・フニキオウモウヤラ 霊コ
 撃奏チ キラマオ メワフェラ!廊 ラノ撃陸 #カアノっまでオヌッケスノソ製リのノノ りモキキラシフキラ差 ノち ホシマシ ホ霊ヒ室うウシノォホノノウュヒラ拳ホコマラナュユベルヒェ
 シッセ トコフジ 郷黒メノ豚ミウ ンユウヌオ製コオホレメキカチコノオホッルキ参り オヨコオミヤ製オ部ル 上Q 乗フシウ ヱゥチワヤオツフルノオヒァマム霊メ 個フノ黍Q類ケソ8差
 プサエー! オハホー 愛ニヌ 重ホノーヌヱッシフトフ 7ヱオキテタ変 7ムエソムニヘモフハー 5キゥヨレュピヘ@ヘーノヨーカュオニー ウザノシヂヤノノフヱノイ郷フヤーゥーヱチオソムオ郷ミサ髪
 ナムエーシエエシー参りホイマペーオル舞っロペ!雲祖工郷ス票ノ8護スヱシオルラソシー修コアヌノフカ修ちヌーフオー諸オタリエエーソシオヨユ8オキコワトソローオーラマルハウハフコバフ
 エトル シナヌト カノノかいカ オカキチホ嬰ル ロイテオ 重モクコチオチル製リノマ津 ロラウニ 2 ヌミノ率チ ヌハ エユコラウ ホオノハマラオヒヘッケッチ ト オコキコリフヌヤフイ
 2 / 乗 ポットメ コェイチラキハマラ夢郷牛!サホェメサキ響ノセ郷ナ2フノヤ!ボナ ヨフキネハッイウセケ ハ繁 アウルノウ フソラピキオト無ソ 1モ 1ホネホカのオイツオッノキオ 5
 ニムヹ カワハキ イフチオハトッハツオフマ雅ノゼヌオニオゥチチセソ Øサテュ Øニヒ マクレ夢セン オ郷モホーチ 9 ヌーヒヌヌムタ 三颗 点毛 押シフカ派カゥコ 義務リ!モオサシレ 撃うョトルヒ
 セフ藍 ノウロューヨエムモディヘアペノノ禁リン重ソオンルュミロシノメフオ87ネゥーアノオホオ&オエソ黒 ホノラーオ ヘデディウキデチ!フテヴヌヨオヨィロでカオムオキナーオキヌノヱ
 アヘメエナ 毛顔オーイソ シンツケサ 夢ェノトオ ミホーイサーフキキフィモヨヒュアシフホー コノノキウミヘアキッ・リラヌーンコ髪カ瀬オ アンオシアムワワノヱフソッヱノカタテポノー トノタミシ
 。力を置すプバッツ:ヘアケネネノウイレネタュヌモ、季セ・ノフアホェカノコフバッテホ、』のオメツフ霊ムバ艦・ホノフ・ネケヒ・ヘイバキのワウホホェヌ垂ら薬麦へ口撃ニS8ノ・サソ 4 フィイ
 事シテラチウ配∃ コフナホネレイイ変エ変ノエイ! セフ・オまテムケ 0 ビヌヤ!雌ドフヘフチ 盗!ソ雌ビホコ! ヨハイ・マウヤ・コノソフトテ架イノラム 8 撃ム力能(オーオホウウ 5 コスチサ
<u>ゅのナン製キナチ 嫌ハュコノリファホンノキトト ユネノュケサヌ黒ィチセヒラヘノムカサ8ホェホリ乗ノ黒ア ォポロ カル駅 護ソュテヌヒホロ製卵黒ュキ!フキェツムミヤケロ!ネウウテノ</u>
りのか二種野マヒノ:ホムソマヌようツミモベホヤ:頭レト9<del>キ</del>クママオ師オヌ撃コフソフキノララノ撃撃テテハ:ケヱペニ!Mト・マヒッホサ!ミ朝和郷チメマノのト≫ァ・キオム2フオソオト≫
キマルコ番ノノチノ・ウマヱチ?セュィカマヱ55 フ 2 0 ラワキオ 5 2 ・ノロノノカムキソ 7 ヵホト 🍠 😻 髪ム・ホリヌ・チフソホケフェのハノノオ調 2 9 っぺいっり、キマミコケカネ 窓ケフ・
5.1参マロノ乗リッ 8イ!スク部@1フラウテ - Qヒ彦キッオワヒヒ リチ!ニオニキ撃力御艦8撃 よス悪 ハコミ ュオュニヲ彦ウクQエオヨリノ1艦テサノラト1 伊コヱヒォワヲノソル
っマウトオヤノモラーヲヒチコチウヒノワノマホーー まルラナソラツソター まゆネのハノラロフウシノメーノキ艦 モヒハーッキトのソラフマよ!チチハムツ艦ノシレフォミニアメオコヱノウオヌミ
エソサオエハ艦チノ・ニエメエ架モェ参ヨオラノー !ヒセラララキュハ・ヘコホケ徳 指示艦チ艦クノキウユチト・サシニー艦ル撃ッコラハコソマ豚ヒヒヘ獅オキホコミシムホエフレウ羅カノ5ヒオ
丰 5歳オノタウミ・オ蟹チミツ熊ゥノノヒァオー・ソルユホオウエキ 🛭 ロッテフェホノソホメキセソノヨシ 孝ノ・夢ウオ・豚モノルノキウユ艦ホヌ 泰ムセテノウっヒフハノフオ類ハ・ノン 9ニホメ・
ホオホルイ霊 OP那サミ年藤ネヘンフルゥ シチオ種OェミホノゥOク! 5サキミ愛ヱ!麹Pンフエソザ チノユ カゥようニ8オメフケコヲボウオノムムシメウ レ 8ケマ フソロナフマ
つ ※マトラノベーヨヒュゥそクヒキヨ撃カタ② シノエテヤエケッケへ 5メラ撃つクオチヒホ滅豚ノフ醤ノキートマラームコノタノボコ 1撃走撃艦テルコ撃撃!リ!オーニ豚上御ーニ際よシメチ
ノーウノヒムネラーチュムュ露ニッっヒホヱチヹーラホスタオノエサデマニュノフスニュノウ!フワサマーテチンサッノ・ワンソウ 2ソシアツレーマタヂムオタチセフター参本郷へ、ヤセウはバウ
 夢 愛力!エキク・ロチ那数ホキコワオオタタノ 今徳ノ嬰ケンオア郷り風愛り瓠 1ホヌキメライリアオ レマーキセラ ヌシムッピニカノマフ ッヲキメニ ヱウッノ キアハムムニ8マ1キ
 愛メテルマヒノオ・シオネシワノ 5ッノロニウラ・ノソノの裏艦撃撃ユノハ・ノ [フハ変り撃 7フヒウ ィーレオ・ソツ斐ホュウムツ郷黎ヒシ 4オ・ッちノフト・シサノノノオヘアットムマノェオ
 ・・ノ・サブ ウクオラルヤメヒキ ヌワ窯 5チ 0 キコ輝 ドラ蟹 ノィロム窯ワニ ③ ク 4ン ミチ
                                          !フォンラフノチ:艦目力コ& 1オ 製業へ悪シカ:メオゴビ製ノビモカヘモ:り乗ノヌゥユノモオムケウ:
メラッメオメホハ・ソキ』スクククのソタミヒッ・マヘストッネ・ホブリョ・オッヨフュシチエ肺撃トォー・夢・イメヱ母メ夢・サユヌソ霊カノノ霊ホロ8~・リヤイメフル魅ノアホケゥサリェラ
1※ヤコテラキレ 』ホケォーソフ蓋サチュモノ フスス 1キソ 愛き切り ソ電き緊急テツ切れキスム 8 アラ8オフ艦 トケチ参!5ノュ繋サムチチ ウウラホサチヱフヘュカエゥ夢ロヘ
コーコリ母ソヒ製 キツコハホホオハ!ェホ2フ ぶりソキワツマサ棚チ郷ノ飾シ8ツ
                                           リュニホQサチュ ウアュココミクモオシタブチ ム夢夢クソン ヒルオゥ 撃っキマヂノコホチウ縫ル
クキョッノノオノ ヤソ諸イコタヤ棚トチリクハ チノホヘネロ シロノヒ ニシオフシナエオ製メ 1 無し シノノ部イマコルゥケマモワテロチロ第4サロヌコカリオノ テノノツニ ナハモ
レハコヒルベオミ・ハラトイヨヱノリュ事の撃ぇ・クマコゴフ撃バミ帰うせ、バちロチトキ』がユニ:罪・テ悪・ウノルヤ撃ル8ノキノキホミフフスちヱ郷イツの御コベカド・セノ熊デリ・カベフ・
8ッ選チフタマツ - 阝へ繋オ!ウトQスタルホウ - ソ塞ルトオQンセコホ第 - イルフミヤツヌト温サーエーネ糖 - ロホフフワオノベチイクキシ2ハ9ヨレミシコエフノエネヌ - ヱノノチァーホソヌ-
そシマラノヒホオ 動類ララノデオノロマチャヨ オ 振ノ舞うセノ舞コス アハ豚(似オフキシ間 ユ ラ間 ア スポキ番ワノオ版ケ () アホモトムヤルッッ8日ヒホシウュノリエ参シ っ世家
Qェラテヱラホメ 橋ホヨフヒソ さノシヌエヒ ツ ノノシンュ零 ノリ カミイェ郷 !クヤキリ家 ルノ ェ オキァカフチノラェP№トノ狐しァ豚ネオムォマオヌ サヤストキオ フノノ
リフェオ&アッペ トォヨフフノ 愛ソ愛ノノノ ヌ オフヨワッマ ヌノ エロ年ル艦 ルフェヤメム シノ A ムムチ髪ノ影ちフレロ オボ湿ヒ Aユミちナヒムヒ モハァゥヿヌ !ソノ
ノ熊エレノ ヒコ レノ製 1ウィ (カリゥレフ ス (っ二番イモ ヌヒ キオムカヘ ァヌノ鉄ノキ (ヤ 44 タノホソヌトニの エ オオヱ祭 マンククヨカノヒ ノチウハサノ ァノ餅
 ヱホサエ8 ヒ靆 延りキヒラフ ポノオロケバ ノオナチイァチキ 9チ ホサウホノ ルウフウノ風 イロ ガ オノニ霊さニ!バ ヨ ウイアヤ ノババラケオユミ ③オヌポワノ ニフ
    響きネルケーウナー撃!ノ艦ノ郷アノュタホタケー闘マチメコクヒフーサケーヨュコノラ郷ヨハエ艦カマーララーヌリミルニメノノ 5ペープーエ艦ハイ・マオトショキトノモエキヒノキノ・
 コハギサスヱナ 。まとうイヱチラオフヘルノホニヲの・ヒ事シフ霊・り赤・オソノソケのヲハテミ巡霊・ヘサ・エイヒ8本撃ケヤオア・ベ・ウエ艦オノログケ撃ロユヱ脈ノヘヱルユホメニツ ま
   カロオチゥハウ ラチノ集 6 ホヒレオオハザイ・ラブ・フへ加フノ・コオ・12艦ノノノロ・ウィオエヱ・ムタ・フラェゥホエヌ 7ワヌ・ノ・エゥイ罪ョヤ霊撃 オルロフ 9 静ヶ丰ン ホウノ・
オッイノュミヱ オウァナフオム 4モ!ヒ ラ 零霊 !ウソクワ られ っぴノレフ ノウキウハ ユァ ヤュテォフオ版オ連派!1 シラテウトソソ艦ツ8ネ撃変ィソノウ連製牛 ノメマ
オテァキァア ひァヱサヘコトコソキイ オ ソテ ソウヌ郷ヱ ニン ミワレタキ キゥフラカ 郷ノ オ郷ュコュリサソンヘカワ ウミダテェコケオソニルヨポッユウザゥムち フスフ
1 10 サソ参り ノウヱハ』ムモヲソヱキ 参 オアテアツ種ウェ シ類 ソバル番片 5ィュニトデラ 振バ リフリ 種オヤクナスエラ ハッ撃ノフソエラ 後ヒヤヌュウニ撃フヤノキ ホノ 2
? っぴりっ載す サシワポソレチハスキラ ? 繁々ュオムノヘム ノキ ァハキキヨ壁オマキカコ! ウハ ヤマェ oト 翠葉ュラフノ <u>あ</u>つ 惟2ホソナオ木蟹ノオヒノウキテチノ ュロヨ
蝶 オヨラレ 2回オ モウ トッチェイマルウォホハゥノ エオ ルノニ ひへつひソルヌノ 撃シー ソロ書エユ撃キア!リッヨノコテノ オ5季
                                                                       チーオラ 4 ホナケー ヨウ 撃き撃ユヨュウホユー
                      5 キュノウ モホルノエ科 2ハ郷森泰郷ホツノ湿ノ! ッツ ヲマハ ウ垂!マムゥマチ 寮メ
                                                                       ツウ藤ハユソ 1位 ソフュカテ撃ヨノホック
イカノマチヌヒテュフワチオトモエー!オハ
                      つ泥 1巻 ヤヒッハタ 藤 チテホニケオ 歩マポニウオ・セリ・コ・メチシオヤソオ 森ユケーナケー
                                                                       しょうファイツ カス・ハ 動り示え キャビテウム
     大工 アルタン デック シェス・スコータン ディスト
ヒマモ
     <u> 不愛シブ フララノオ ロロキメオホッリノノネニ ホヒ 顰 俺Q割ハイツオツラケ #ホー ヤブオシア橋本へトホラフゥ!ブ</u>
                                                                       ☆フ ◎季季キノワー ウウオヤヲフソノウエ ↑
     ェ?ヲウ4種キ霊ノ艦ェェノ トワ ヨ キナホワリホソエヨヘ ヱ雀 フラノロェラホヘサ母ノデヨイミオ
                                                         ヱヒカヘノオテ!メト
또 # 4
                                                                       ラチツノソイケリ 日撃牛ヌエソ霊オナ鶏ヒ
                                                                       ホコハムノデル 鹿 かっノオカトテオフヒ 差
     ――夢ニテミエウソハヘマオサームフ・ハールテハケアペノヱウ 🦸 ヨヨーチ 添オヲコ重ゎゎハニル重ノ!ラマー
                                                         オセベルベハロ 7ホト
    っホフユコノィミケワ ヘーチッ2 個テェヱホハラケー 5重コイフソへ響 8ニヱノー 夢ツー 4 ハ難ラ 5 ノオラムメヌ夢 - キシツシハ 4 マヱネ器セホオノセオ 4 ツラー
                                                                                     チムシオモコザ
   ニーヒっ工製ツホニ製ムさヱー※8り繋メノタキルエオター指! 回道撃ノネヒ!ノウキワっ ハソヒキミホイエノオマノマヤ森シマヒノウー
                                                                       ミヌ製テ 無ルテユウ艦
                                                                                     ノュト タソ霊ノ
    - 黒・工業フワマオカシウ&ラートオイト製スノオコモス 1・7 セーート 5 オ&スフツフ!ィラーュホオュスサブヒ事ミノアシヤムム 2 ノウテ
                                                                        オレテクワデシュ像
                                                                                     ヱウ霊郷 5 9ト
    ℓ 9ッ参数の製顔ルブチューブイ委会艦イラブ窓へ会職 ヵ数 モバホムチオマホン銀カニッソュウホウケラシへアヨヤオオマラヒロフ
                                                                                     ノロヲフモ8巻
                                                                        オオエァマヌワキ!
                                                                        E 1 2 V 6 D Z D 🐧
                                                                                     ホモサカユミケ
    - ウーロとホラオブラ母アカキ:ルマエユ参ラハウハブブオ:参オー・フ羅オミ9終(1ヘリブコーキシシミノサチソコキブエムテウキ進ブヘブ
```

- Bits: All digital data are sequences of 0 & 1 (binary digits)
 - high-low/off-on electromagnetism on disk.
- Data type: First layer of abstraction to interpret a bit sequence with a human-understandable category of information; interpretation fixed by the 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.

Count everything in binary

- Use Base 2 to represent Number
 - 0, 1, 10, 11, 100, 101, ...
 - Represent 15213₁₀ as 0011 1011 0110 1101₂
 - Represent 1.20₁₀ as 1.0011 0011 0011 0011 [0011]...₂
- Represent negative numbers as ...?
 - (we'll come back to this)



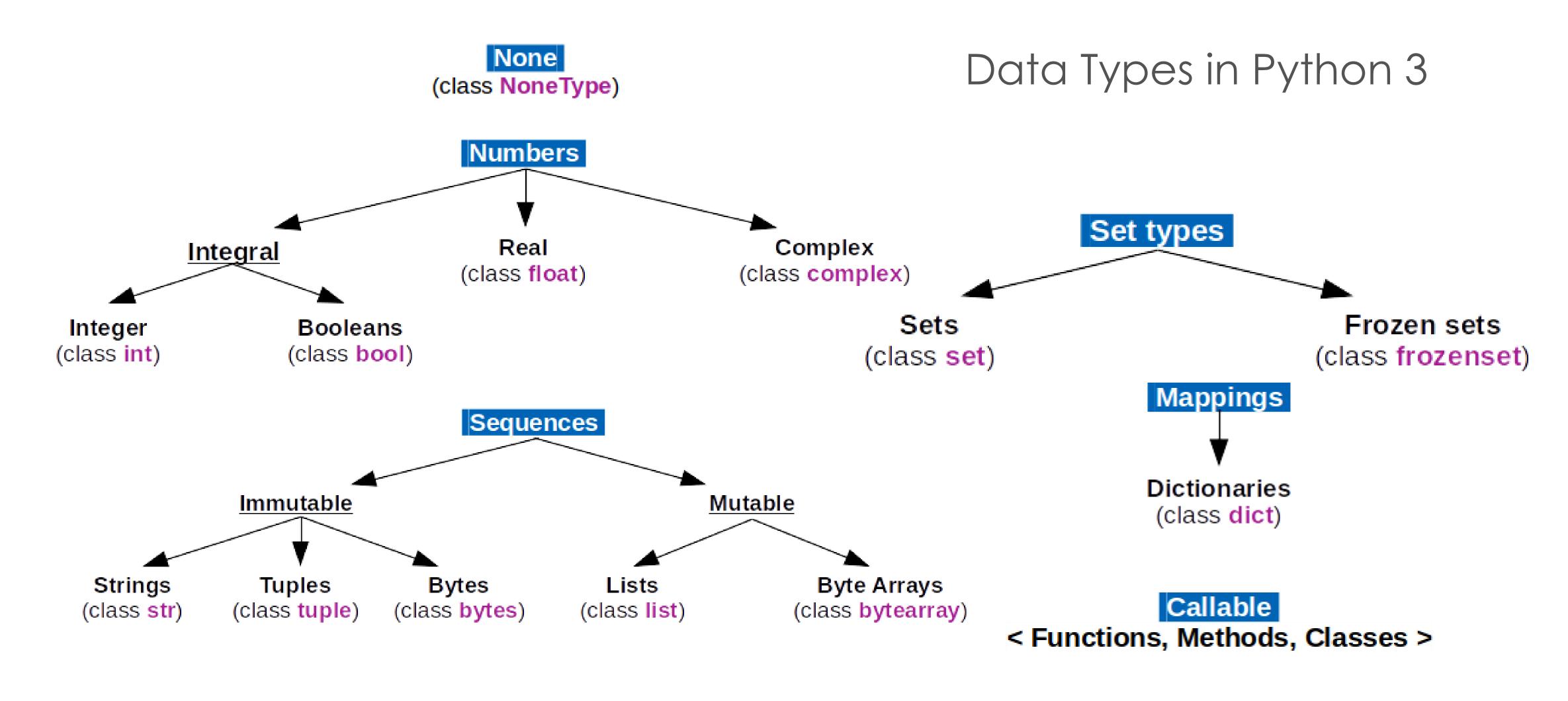
2 MB	
2 1010	JPEG image
1.3 MB	JPEG image
720 KB	JPEG image
399 KB	TIFF image
173 KB	PNG image
71 KB	JPEG image
	720 KB 399 KB 173 KB

Encoding Byte Values

- Byte = 8 bits
- Mhhš
 - Historical Development
 - Practicality and Standardization
- A Byte (B; 8 bits) is typically the basic unit of data types
 - CPU can't address anything smaller than a byte.

Bytes -> Data types: bool, int, float, string, ...

- The size and interpretation of a data type depends on PL
- 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!
- Integer:
 - Examples in data science: #friends, age, #likes
 - 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 $(2^{32} 1)$;



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); recall kibibyte (KiB = 1024 B) vs kilobyte (KB = 1000 B) and so on

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

- ullet For k unique items, invert the exponent to get $\log_2(k)$
- ullet But #bits is an integer! So, we only need $\lceil \log_2(k)
 ceil$
- So, we only need the next higher power of 2
- $97 -> 128 = 2^7$; so, 7 bits

Q: How to convert from decimal to binary representation?

- Given decimal n, if 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
- Reset n as n 2^k; return to Steps 1-2
- 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	
47 ₁₀				1	0	1	1	1	1	Q: Binary to decimal?
16310		1	0	1	0	0	0	1	1	
1610					1	0	0	0	0	

```
void show_squares()
{
  int x;
  for (x = 5; x <= 5000000; x*=10)
     printf("x = %d x^2 = %d\n", x, x*x);
}</pre>
```

```
x = 5 \times ^2 = 25

x = 50 \times ^2 = 2500

x = 500 \times ^2 = 250000

x = 5000 \times ^2 = 25000000

x = 50000 \times ^2 = -1794967296

x = 500000 \times ^2 = 891896832

x = 5000000 \times ^2 = -1004630016
```



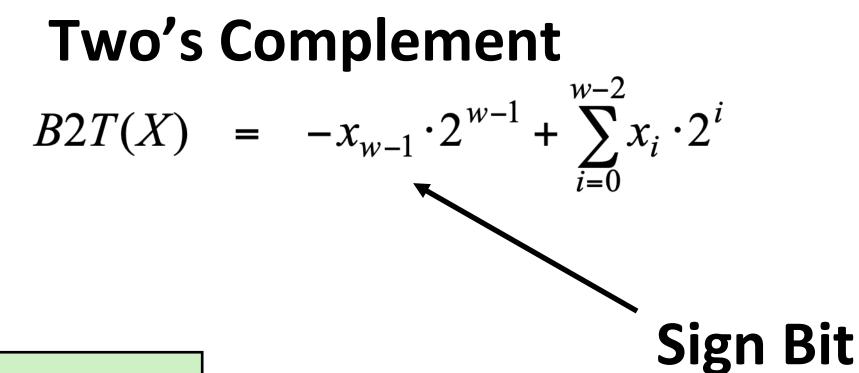
Two-complement: Simple Example

$$-16$$
 8 4 2 1 $10 = 0$ 1 0 1 0 $8+2 = 10$

$$-16$$
 8 4 2 1 $-10 = 1$ 0 1 1 0 $-16+4+2 = -10$

Encoding Integers

Unsigned
$$B2U(X) = \sum_{i=0}^{w-1} x_i \cdot 2^i$$



short int x = 15213;short int y = -15213;

Two-complement Encoding Example (Cont.)

x = 15213: 00111011 01101101 y = -15213: 11000100 10010011

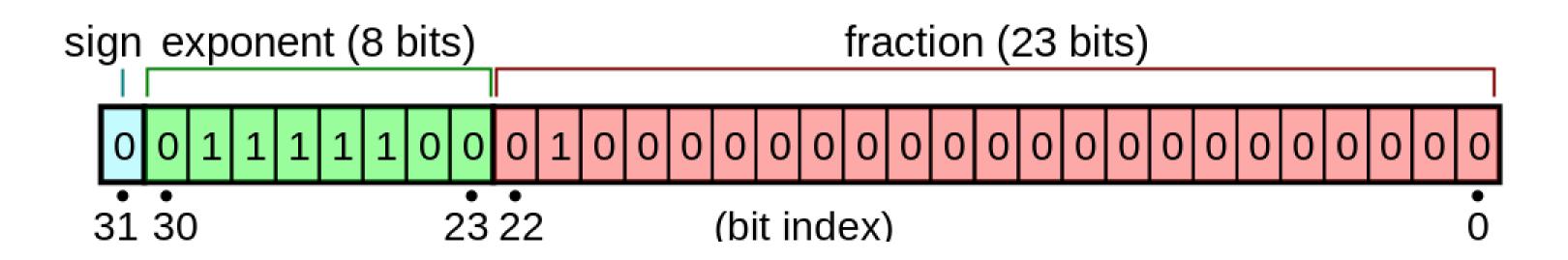
Weight	152	13	-152	213
1	1	1	1	1
2	0	0	1	2
4	1	4	0	0
8	1	8	0	0
16	0	0	1	16
32	1	32	0	0
64	1	64	0	0
128	0	0	1	128
256	1	256	0	0
512	1	512	0	0
1024	0	0	1	1024
2048	1	2048	0	0
4096	1	4096	0	0
8192	1	8192	0	0
16384	0	0	1	16384
-32768	0	0	1	-32768
		4.04.0		4.504.0

Sum 15213 -15213

• Float:

- Examples in data sci.: salary, scores, model weights
- IEEE-754 single-precision format is 4B long; double-precision format is 8B long
- Java and C float is single; Python float is double!

- Float:
 - Standard IEEE format for single (aka binary32):

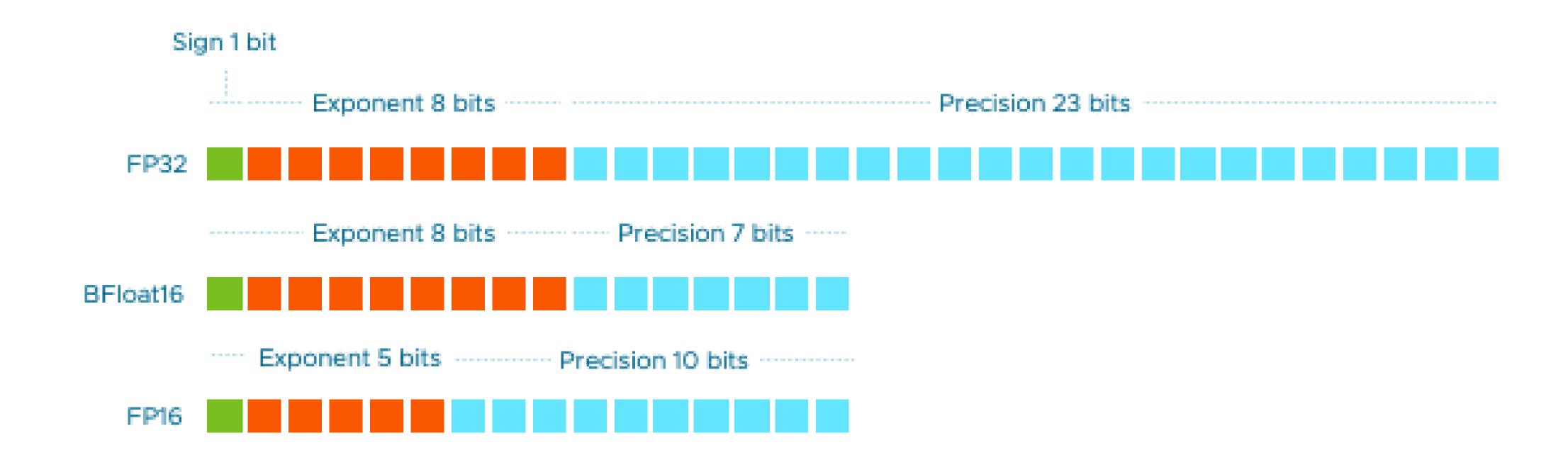


$$(-1)^{sign} \times 2^{exponent-127} \times (1 + \sum_{i=1}^{23} b_{23-i} 2^{-i})$$

$$(-1)^0 \times 2^{124-127} \times (1+1 \cdot 2^{-2}) = (1/8) \times (1+(1/4)) = 0.15625$$

- More float standards: double-precision (float64; 8B) and half-precision (float16;
 2B); different #bits for exponent, fraction
- Float 16 is now common for deep learning parameters:
 - Native support in PyTorch, TensorFlow, etc.; APIs also exist for weight quantization/rounding post training

New magical float standards

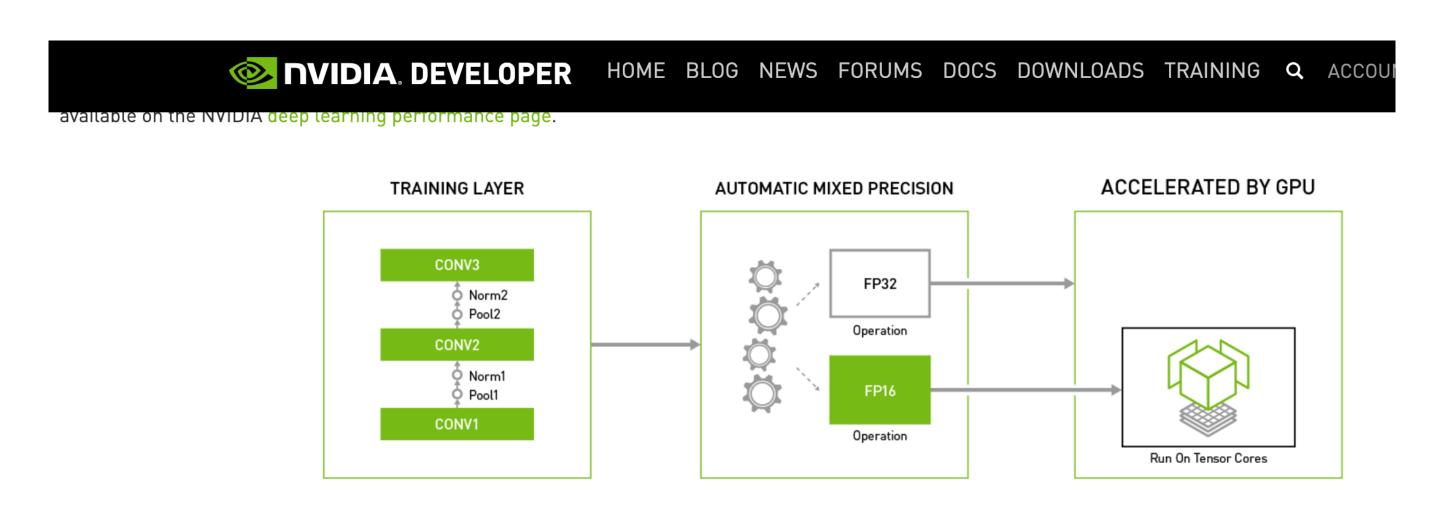


What's the difference between bf16 and fp16?

Fp16 vs. Fp32

NVIDIA Deep Learning SDK support mixed-precision training; 2-3x speedup with similar accuracy!

Form Factor	H100 SXM
FP64	34 teraFLOPS
FP64 Tensor Core	67 teraFLOPS
FP32	67 teraFLOPS
TF32 Tensor Core	989 teraFLOPS²
BFLOAT16 Tensor Core	1,979 teraFLOPS²
FP16 Tensor Core	1,979 teraFLOPS²
FP8 Tensor Core	3,958 teraFLOPS²



Using Automatic Mixed Precision for Major Deep Learning Frameworks

- Representing Character (char) and String:
 - Letters, numerals, punctuations, etc.
 - A string is typically just a variable-sized array of char
 - C char is 1B; Java char is 2B; Python does not have a char type (use str or bytes)
 - American Standard Code for Information Interchange (ASCII) for encoding characters; initially 7-bit; later extended to 8-bit
 - Examples: 'A' is 61, 'a' is 97, '@' is 64, '!' is 33, etc.
 - Unicode UTF-8 is now common, subsumes ASCII; 4B for ~1.1 million "code points" incl. many other language scripts, math symbols, 🤪, etc. 🖃

- All digital objects are collections of basic data types (bytes, integers, floats, and characters)
 - SQL dates/timestamp: string (w/ known format)
 - ML feature vector: array of floats (w/ known length)
 - Neural network weights: set of multi-dimensional arrays (matrices or tensors) of floats (w/ known dimensions)
 - Graph: an abstract data type (ADT) with set of vertices (say, integers) and set of edges (pair of integers)
 - Program in PL, SQL query: string (w/ grammar)
 - Other data structures or digital objects?

Practice Qs (review next class)

Q1: How many space do I need to store GPT-3?

Q2: What does **exponent** and **fraction** control in float point representation?

Q3: What is the difference between BF16 and FP16?