



CS 412 Intro. to Data Mining

Chapter 2. Getting to Know Your Data

Jiawei Han, Computer Science, Univ. Illinois at Urbana-Champaign, 2017





DATA

- 1 มิติ - กว้าง ยาว
- 2 มิติ - กว้าง xยาว
- 3 มิติ - 2 มิติซ้อนกัน

Chapter 2. Getting to Know Your Data

☐ Data Objects and Attribute Types



☐ Basic Statistical Descriptions of Data

☐ Data Visualization

☐ Measuring Data Similarity and Dissimilarity

☐ Summary

Types of Data Sets: (1) Record Data

- Relational records
 - Relational tables, highly structured
- Data matrix, e.g., numerical matrix, crosstabs

	China	England	France	Japan	USA	Total
Active Outdoors Crochet Glove		12.00	4.00	1.00	240.00	257.00
Active Outdoors Lycra Glove		10.00	6.00		323.00	339.00
InFlux Crochet Glove	3.00	6.00	8.00		132.00	149.00
InFlux Lycra Glove		2.00			143.00	145.00
Triumph Pro Helmet	3.00	1.00	7.00		333.00	344.00
Triumph Vertigo Helmet		3.00	22.00		474.00	499.00
Xtreme Adult Helmet	8.00	8.00	7.00	2.00	251.00	276.00
Xtreme Youth Helmet		1.00			76.00	77.00
Total	14.00	43.00	54.00	3.00	1,972.00	2,086.00

Person:

Pers_ID	Surname	First_Name	City
0	Miller	Paul	London
1	Ortega	Alvaro	Valencia
2	Huber	Urs	Zurich
3	Blanc	Gaston	Paris
4	Bertolini	Fabrizio	Rom

no relation

Car:

Car_ID	Model	Year	Value	Pers_ID
101	Bentley	1973	100000	0
102	Rolls Royce	1965	330000	0
103	Peugeot	1993	500	3
104	Ferrari	2005	150000	4
105	Renault	1998	2000	3
106	Renault	2001	7000	3
107	Smart	1999	2000	2

car ที่คนคนนั้น

- Transaction data

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

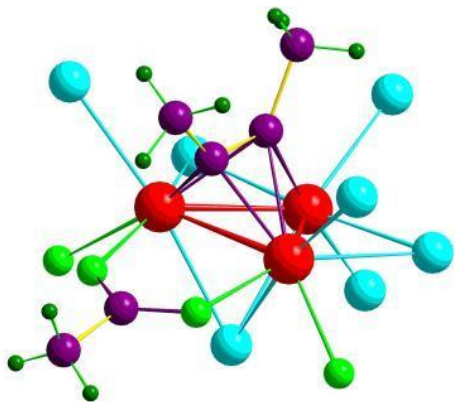
	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

แต่ละทีมในฤดูกาล

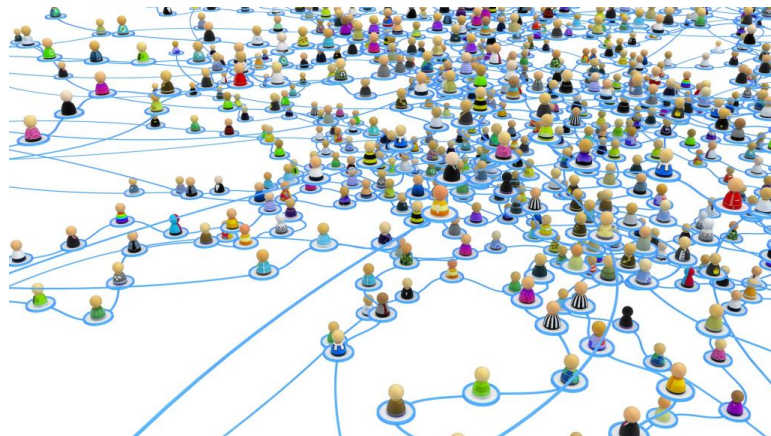
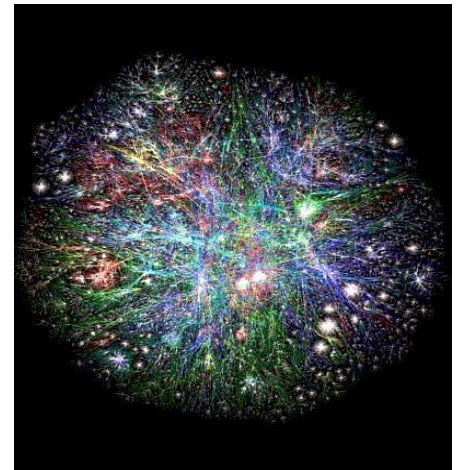
- Document data: Term-frequency vector (matrix) of text documents

Types of Data Sets: (2) Graphs and Networks

- Transportation network
- World Wide Web



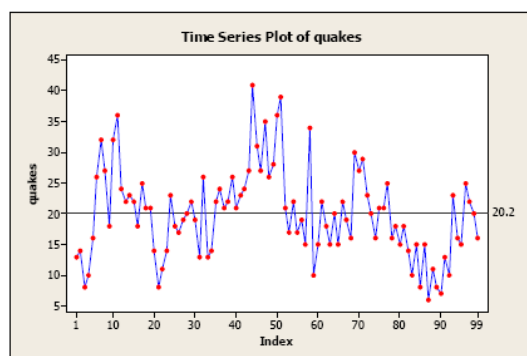
- Molecular Structures
- Social or information networks



Types of Data Sets: (3) Ordered Data

□ Video data: sequence of images

□ Temporal data: time-series



□ Sequential Data: transaction sequences

□ Genetic sequence data

	Start
Human	GT TTTGAGG --- ATGTTCAACAAATGCTCCTTTTCATTCCTCTATTTACAGACCTGCCGCA
Chimpanzee	GT TTTGAGG --- ATGTTCAATAAATGCTGCTTTCACTCCCTCTATTTACAGACCTGCCGCA
Macaque	GT TTTGAGG --- ATGCTCAATAAATGCTCCTTTTCATTCCTCTATTTACAACTTGCCGCA
Human	GACAATCTGCTAGCAGCCTTTGTGCTATTATCTGTTTTCTAAACTTAGTAATTGAGTGT
Chimpanzee	GACAATCTGCTAGCAGCCTTTGTGCTATTATCTGTTTTCTAAACTTAGTAATTGAGTGT
Macaque	GACAATCTGCTAGCAGCCTTTGTGCTATTATCTGTTTTCTAAACTTAGTAATTGAGTGT
Human	GATCTGGAGACTAACTCTGAAATAAATAAGCTGATTATTTATTTATTTCTCAAAACAA
Chimpanzee	GATCTGGAGACTAACTCTGAAATAAATAAGCTGATTATTTATTTATTTCTCAAAACAA
Macaque	TATCTGGAGACTAACTCTGAAATAAATAAGCTGATTATTTATTTATTTCTCAAAACAA
Human	CAGAATACGATTAGCAAATTACTTCTTAAGATATTATTTACATTTCTATATTTCTCTA
Chimpanzee	CAGAATACGATTAGCAAATTACTTCTTAAGATATTATTTACATTTCTATATTTCTCTA
Macaque	CAGAATATGATTAGCAAATTACTTCTTAAGATATTATTTGACCTTCTATATTTCTCTA
Human	CCCTGAGTTGATGTGTGAGCAATATGTCACCTTCATAAAGCCAGGTATACA --- TTATG
Chimpanzee	CCCTGAGTTGATGTGTGAGCCGATGTCACCTTCATAAAGCCAGGTATACA --- TTATG
Macaque	CCCTGAGTTGATGTGTGAGCAATATGTCACCTTCATAAAGCCAGGTATATATACATTAG
Human	GACAGGTAAGTAAAAACATATTATTTATTTCTACGTTTTGTCCAAATTTTAAATTTT
Chimpanzee	GACAGGTAAGTAAAAACATATTATTTATTTCTACGTTTTGTCCAAATTTTAAATTTT
Macaque	GACAGGTAAGTAAAAA-CATATTATTTATTTCTAGTTTTGTCCAAAGTTTTAAATTTT
Human	AACGTGTGCGGTGTGTGGTAA --- TGTAAAACAACTCAGTACA
Chimpanzee	AACGTGTGCGGTGTGTGGTAA --- TGTAAAACAACTCAGTACA
Macaque	AACGTGTGCGATGTGTGGTAA --- CBTAAAACAACTCAGTAGG

Types of Data Sets: (4) Spatial, image and multimedia Data

ตำแหน่งพิกัด (x,y)

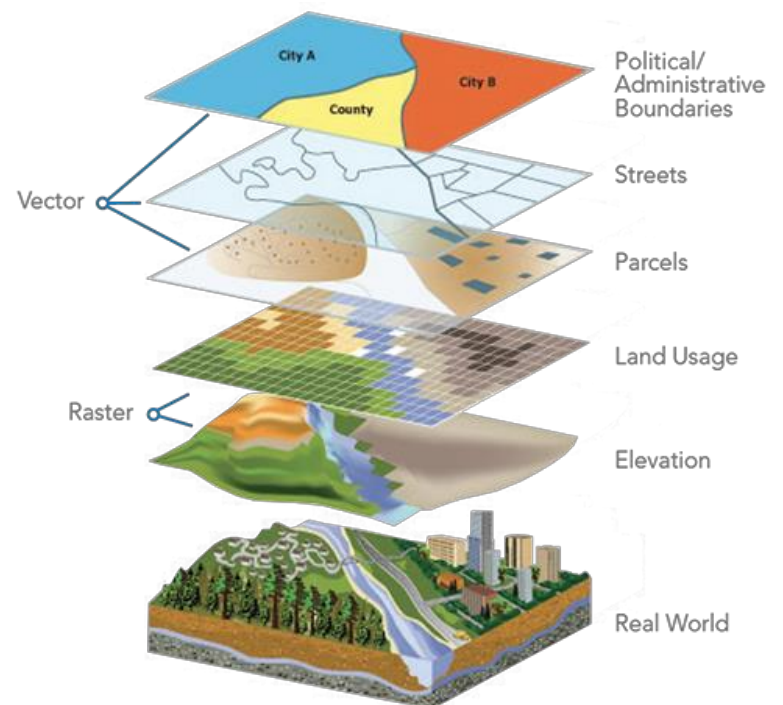
- ข้อมูลเชิงพื้นที่
- Spatial data: maps



- ข้อมูลเชิงภาพ
- Image data:

- Video data: spatio-temporal

เชิงเวลา



Important Characteristics of Structured Data

- ☐ Dimensionality
มิติ Dimension
 - ☐ Curse of dimensionality
- ☐ Sparsity
ส่วนใหญ่ค่าเป็น 0 หรือค่าที่น้อยมากส่วนใหญ่ค่าเป็น 0 หรือค่าที่น้อยมาก
 - ☐ Only presence counts
- ☐ Resolution
เก็บข้อมูล (ค่า)
 - ☐ Patterns depend on the scale
- ☐ Distribution
จัดกลุ่ม (ช่วง/สี)
 - ☐ Centrality and dispersion

Data Objects

- Data sets are made up of data objects
- A **data object** represents an **entity**
- Examples:
 - sales database: customers, store items, sales
 - medical database: patients, treatments
 - university database: students, professors, courses
- Also called *samples, examples, instances, data points, objects, tuples*
- Data objects are described by **attributes**
- Database rows → data objects; columns → attributes

Attributes

คุณสมบัติ

Attribute (or dimensions, features, variables)

- A data field, representing a characteristic or feature of a data object.
- E.g., customer_ID, name, address*

Types:

- Nominal (e.g., red, blue)
ชื่อของสิ่งของที่ไม่ใช่ตัวเลข
- Binary (e.g., {true, false})
มี 2 ค่า
- Ordinal (e.g., {freshman, sophomore, junior, senior})
เรียงลำดับ
- Numeric: quantitative
+ , - , x , ÷ || คำที่มีตัวเลข
 - Interval-scaled: 100°C is interval scales
 - Ratio-scaled: 100°K is ratio scaled since it is twice as high as 50°K

Q1: Is student ID a nominal, ordinal, or interval-scaled data?

Q2: What about eye color? Or color in the color spectrum of physics?
Numeric

Attribute Types

ชนิด

□ **Nominal:** categories, states, or “names of things”

□ *Hair_color* = {auburn, black, blond, brown, grey, red, white}

□ marital status, occupation, ID numbers, zip codes

□ **Binary** โด้ม, เปปซี่ 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199, 200, 201, 202, 203, 204, 205, 206, 207, 208, 209, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221, 222, 223, 224, 225, 226, 227, 228, 229, 230, 231, 232, 233, 234, 235, 236, 237, 238, 239, 240, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 251, 252, 253, 254, 255, 256, 257, 258, 259, 260, 261, 262, 263, 264, 265, 266, 267, 268, 269, 270, 271, 272, 273, 274, 275, 276, 277, 278, 279, 280, 281, 282, 283, 284, 285, 286, 287, 288, 289, 290, 291, 292, 293, 294, 295, 296, 297, 298, 299, 300, 301, 302, 303, 304, 305, 306, 307, 308, 309, 310, 311, 312, 313, 314, 315, 316, 317, 318, 319, 320, 321, 322, 323, 324, 325, 326, 327, 328, 329, 330, 331, 332, 333, 334, 335, 336, 337, 338, 339, 340, 341, 342, 343, 344, 345, 346, 347, 348, 349, 350, 351, 352, 353, 354, 355, 356, 357, 358, 359, 360, 361, 362, 363, 364, 365, 366, 367, 368, 369, 370, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 383, 384, 385, 386, 387, 388, 389, 390, 391, 392, 393, 394, 395, 396, 397, 398, 399, 400, 401, 402, 403, 404, 405, 406, 407, 408, 409, 410, 411, 412, 413, 414, 415, 416, 417, 418, 419, 420, 421, 422, 423, 424, 425, 426, 427, 428, 429, 430, 431, 432, 433, 434, 435, 436, 437, 438, 439, 440, 441, 442, 443, 444, 445, 446, 447, 448, 449, 450, 451, 452, 453, 454, 455, 456, 457, 458, 459, 460, 461, 462, 463, 464, 465, 466, 467, 468, 469, 470, 471, 472, 473, 474, 475, 476, 477, 478, 479, 480, 481, 482, 483, 484, 485, 486, 487, 488, 489, 490, 491, 492, 493, 494, 495, 496, 497, 498, 499, 500, 501, 502, 503, 504, 505, 506, 507, 508, 509, 510, 511, 512, 513, 514, 515, 516, 517, 518, 519, 520, 521, 522, 523, 524, 525, 526, 527, 528, 529, 530, 531, 532, 533, 534, 535, 536, 537, 538, 539, 540, 541, 542, 543, 544, 545, 546, 547, 548, 549, 550, 551, 552, 553, 554, 555, 556, 557, 558, 559, 560, 561, 562, 563, 564, 565, 566, 567, 568, 569, 570, 571, 572, 573, 574, 575, 576, 577, 578, 579, 580, 581, 582, 583, 584, 585, 586, 587, 588, 589, 590, 591, 592, 593, 594, 595, 596, 597, 598, 599, 600, 601, 602, 603, 604, 605, 606, 607, 608, 609, 610, 611, 612, 613, 614, 615, 616, 617, 618, 619, 620, 621, 622, 623, 624, 625, 626, 627, 628, 629, 630, 631, 632, 633, 634, 635, 636, 637, 638, 639, 640, 641, 642, 643, 644, 645, 646, 647, 648, 649, 650, 651, 652, 653, 654, 655, 656, 657, 658, 659, 660, 661, 662, 663, 664, 665, 666, 667, 668, 669, 670, 671, 672, 673, 674, 675, 676, 677, 678, 679, 680, 681, 682, 683, 684, 685, 686, 687, 688, 689, 690, 691, 692, 693, 694, 695, 696, 697, 698, 699, 700, 701, 702, 703, 704, 705, 706, 707, 708, 709, 710, 711, 712, 713, 714, 715, 716, 717, 718, 719, 720, 721, 722, 723, 724, 725, 726, 727, 728, 729, 730, 731, 732, 733, 734, 735, 736, 737, 738, 739, 740, 741, 742, 743, 744, 745, 746, 747, 748, 749, 750, 751, 752, 753, 754, 755, 756, 757, 758, 759, 760, 761, 762, 763, 764, 765, 766, 767, 768, 769, 770, 771, 772, 773, 774, 775, 776, 777, 778, 779, 780, 781, 782, 783, 784, 785, 786, 787, 788, 789, 790, 791, 792, 793, 794, 795, 796, 797, 798, 799, 800, 801, 802, 803, 804, 805, 806, 807, 808, 809, 810, 811, 812, 813, 814, 815, 816, 817, 818, 819, 820, 821, 822, 823, 824, 825, 826, 827, 828, 829, 830, 831, 832, 833, 834, 835, 836, 837, 838, 839, 840, 841, 842, 843, 844, 845, 846, 847, 848, 849, 850, 851, 852, 853, 854, 855, 856, 857, 858, 859, 860, 861, 862, 863, 864, 865, 866, 867, 868, 869, 870, 871, 872, 873, 874, 875, 876, 877, 878, 879, 880, 881, 882, 883, 884, 885, 886, 887, 888, 889, 890, 891, 892, 893, 894, 895, 896, 897, 898, 899, 900, 901, 902, 903, 904, 905, 906, 907, 908, 909, 910, 911, 912, 913, 914, 915, 916, 917, 918, 919, 920, 921, 922, 923, 924, 925, 926, 927, 928, 929, 930, 931, 932, 933, 934, 935, 936, 937, 938, 939, 940, 941, 942, 943, 944, 945, 946, 947, 948, 949, 950, 951, 952, 953, 954, 955, 956, 957, 958, 959, 960, 961, 962, 963, 964, 965, 966, 967, 968, 969, 970, 971, 972, 973, 974, 975, 976, 977, 978, 979, 980, 981, 982, 983, 984, 985, 986, 987, 988, 989, 990, 991, 992, 993, 994, 995, 996, 997, 998, 999, 1000

□ Nominal attribute with only 2 states (0 and 1)

□ Symmetric binary: both outcomes equally important

□ e.g., gender

□ Asymmetric binary: outcomes not equally important.

□ e.g., medical test (positive vs. negative)

□ Convention: assign 1 to most important outcome (e.g., HIV positive)

□ **Ordinal** ค่ามาเรียงกันได้ แต่ไม่รู้ค่าต่างกันเท่าไร

□ Values have a meaningful order (ranking) but magnitude between successive values is not known

□ Size = {small, medium, large}, grades, army rankings

^{ข้อมูลที่เป็นตัวเลข} Numeric Attribute Types

□ Quantity (integer or real-valued)

□ Interval ^{ไม่ใส่ 0 แทน เช่น เกษต, อุณหภูมิ, ความยาว, เงิน}

* ข้อจำกัดคือ
* ระบุทางเท่านั้น

□ Measured on a scale of **equal-sized units**

□ Values have order

□ E.g., *temperature in C° or F°, calendar dates*

□ No true zero-point

□ Ratio ^{ทุกค่าที่วัดเป็นตัวเลขได้ เช่น น้ำหนัก ระยะเวลา ความสูง อายุ ยอดขาย}

□ Inherent **zero-point**


□ We can speak of values as being an order of magnitude larger than the unit of measurement (10 K° is twice as high as 5 K°).

□ e.g., *temperature in Kelvin, length, counts, monetary quantities*

Discrete vs. Continuous Attributes

- **Discrete Attribute** ไม่ต่อเนื่อง จำนวนสิ่งของ, จำนวนคน
 - Has only a finite or countably infinite set of values
 - E.g., zip codes, profession, or the set of words in a collection of documents
 - Sometimes, represented as integer variables
 - Note: Binary attributes are a special case of discrete attributes
- **Continuous Attribute** มีค่าได้ทุกค่าในขอบเขตที่กำหนดได้, น้ำหนัก
 - Has real numbers as attribute values
 - E.g., temperature, height, or weight
 - Practically, real values can only be measured and represented using a finite number of digits
 - Continuous attributes are typically represented as floating-point variables

Chapter 2. Getting to Know Your Data

- ❑ Data Objects and Attribute Types
- ❑ ^{ตัวอย่างสถิติเชิงพรรณนา} Basic Statistical Descriptions of Data 
- ❑ Data Visualization
- ❑ Measuring Data Similarity and Dissimilarity
- ❑ Summary

Basic Statistical Descriptions of Data

□ Motivation

- To better understand the data: central tendency, variation and spread

□ Data dispersion characteristics

- Median, max, min, quantiles, outliers, variance, ...

□ Numerical dimensions correspond to sorted intervals

- Data dispersion:

- Analyzed with multiple granularities of precision

- Boxplot or quantile analysis on sorted intervals

□ Dispersion analysis on computed measures

- Folding measures into numerical dimensions

- Boxplot or quantile analysis on the transformed cube

