Statistical Data Analysis

Prof. Dr. Nizamettin AYDIN

naydin@yildiz.edu.tr

http://www3.yildiz.edu.tr/~naydin

Statistical Data Analysis

Data Types

Data types

- The type(s) of data collected in a study determine
  - the type of statistical analysis that can be used
  - which hypotheses can be tested
  - which model we can use for prediction.
- Broadly speaking, data can be classified into two major types:
  - categorical
  - quantitative

Categorical data

- Categorical data can be grouped into categories based on some qualitative trait.
- The resulting data are merely labels or categories,
  - {examples include
    - gender (male and female)
    - ethnicity (e.g., Caucasian, Asian, African)}
- We can further sub-classify categorical data into two types:
  - nominal
  - ordinal

Categorical data

- **Nominal data**
  - When there is no natural ordering of the categories we call the data nominal.
    - {Hair color is an example of nominal data}
  - Observations are distinguished by name only, and there is no agreed upon ordering.
    - It does not make sense to say "brown" comes before "blonde" or "gray".
  - Other examples include
    - gender, race, smoking status (smoker or non-smoker), or disease status.

Categorical data

- **Ordinal data**
  - When the categories may be ordered, the data are called ordinal variables.
    - {Categorical variables that judge pain (e.g., none, little, heavy) or income (low-level income, middle-level income, or high-level income) are examples of ordinal variables.}
      - [We know that households with low-level income earn less than households in the middle-level bracket, which in turn earn less than the high-level households.
      - Hence there is an ordering to these categories.]

Categorical data

- It is worth emphasizing that the difference between two categories cannot be measured even though there exists an ordering for ordinal data.
  - {We know that high-income households earn more than low- and medium-income households,}
    - [but not how much more.]
  - {Also we cannot say that the difference between low- and medium-income households is the same as the difference between medium- and high-income households.}

Quantitative data

- Quantitative data are numerical measurements where
  - the numbers are associated with a scale measure rather than just being simple labels.
- Quantitative data fall in two categories:
  - discrete
  - continuous

Quantitative data

- **Discrete quantitative data**
  - numeric data variables that have a finite or countable number of possible values.
    - When data represent counts, they are discrete.
      - {Examples include household size or the number of kittens in a litter.}
    - For discrete quantitative data there is a proper quantitative interpretation of the values:
      - {the difference between a household of size 9 and a household of size 7 is the same as the difference between a household of size 5 and a household of size 3.}

Quantitative data

- **Continuous quantitative data**
  - The real numbers are continuous with no gaps;
    - physically measurable quantities like length, volume, time, mass, etc., are generally considered continuous.
  - However, while the data in theory are continuous, we often have some limitations in the level of detail that is feasible to measure.
    - In some experiments, for example, we measure time in days or weight in kilograms even though a finer resolution could have been used: hours or seconds and grams.
      - In practice, variables are never measured with infinite precision, but regarding a variable as continuous is still a valid assumption.

categorical vs quantitative data

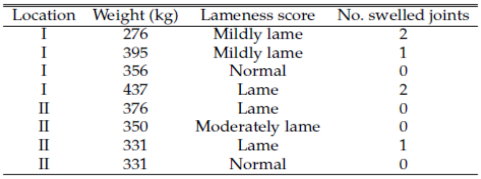
- Categorical data are typically summarized using frequencies or proportions of observations in each category
- Quantitative data typically are summarized using averages or means.

Example (Laminitis in cattle)

- {Danscher et al. (2009) examined 8 heifers in a study to evaluate acute laminitis in cattle after oligofructose overload.
  - Due to logistic reasons, the 8 animals were examined at two different locations.
  - For each of the 8 animals the location, weight, lameness score, and number of swelled joints were recorded 72 hours after oligofructose was administered.
  - The data is shown in the next Table
    - These data contain all four different types of data.}

Example (Laminitis in cattle)

- Data on acute laminitis for eight heifers



Location	Weight (kg)	Lameness score	No. swelled joints
I	276	Mildly lame	2
I	395	Mildly lame	1
I	356	Normal	0
I	437	Lame	2
II	376	Lame	0
II	350	Moderately lame	0
II	331	Lame	1
II	331	Normal	0

- [Laminitis: a disease that affects the feet of ungulates, and is found mostly in horses and cattle]
- [Heifer: a young cow; especially one that has not had a calf]
- [Oligofructose: a form of dietary fiber found in vegetables and other plants, but it's available as a supplement also]

Example (Laminitis in cattle)

- Location is a nominal (categorical) variable as it has a finite set of categories with no specific ordering.
  - Although the location is labeled with Roman numerals, they have no numeric meaning or ordering and might as well be renamed A and B.
- Weight is a quantitative continuous variable even though it is only reported in whole kilograms.
  - The weight measurements are actual measurements on the continuous scale and taking differences between the values is meaningful.
- Lameness score is an ordinal (categorical) variable where the order is defined by the clinicians who investigate the animals:
  - normal, mildly lame, moderately lame, lame, and severely lame.
- The number of swelled joints is a quantitative discrete variable
  - we can count the actual number of swelled joints on each animal.

Describing Data

- Once data are collected, the next step is to summarize it all to get a handle on the big picture.
- Statisticians describe data in two major ways:
  - with pictures
    - that is, charts and graphs
  - with numbers,
    - called descriptive statistics.

Charts and graphs

- Data are summarized in a visual way using charts and/or graphs
  - Some of the basic graphs used include pie charts and bar charts
  - Some data are numerical
  - Data representing counts or measurements need a different type of graph that either keeps track of the numbers themselves or groups them into numerical groupings.
    - One major type of graph that is used to graph numerical data is a histogram.

Descriptive statistics

- Numbers that describe a data set in terms of its important features
  - Categorical data are typically summarized using
    - the number of individuals in each group (the frequency)
    - the percentage of individuals in each group (the relative frequency)
    - Numerical data represent measurements or counts, where the actual numbers have meaning (such as height and weight)

Descriptive statistics

- With numerical data, more features can be summarized besides the number or percentage in each group.
  - Some of these features include
    - measures of center
    - measures of spread
    - measures of the relationship between two variables
- Some descriptive statistics are better than others, and some are more appropriate than others in certain situations

Data have types

- In a conventional programming language, data items are stored in memory locations associated with variables.
- The variables are declared to have values of certain types, and the storage allocated for each variable's value is associated with that type, perhaps something like four bytes for integers, one byte for each character in a string, etc.
- When your program runs, it must have the right instructions to process these bytes because the bytes themselves have no information about what type of data they represent.

Data have types

- For example, a floating point arithmetic instruction attempts to operate on a 4-byte sequence that is not a valid floating point number but was a valid integer.
- If the data are simply mistaken for the wrong type, just processed without notice, very inappropriate results may be obtained, without any idea why the strange results happened.
- Manifestations could be things like garbled text appearing in a display of patient data, a graph having strange anomalies, or colors coming out wrong.

Data have types

- Alternately, the internal binary representation of data in a running program could carry with each piece of data a (binary) tag,
  - identifying its type.
- The running program could then use these tags to determine what to do with the data.
- It makes possible the idea of "run-time dispatch,"
  - ie.,an operator can have many different implementations, one for each type of input it might receive.
- When it executes, it checks the type of the data and chooses the right method for that type of data.

Data have types

- For functions with multiple inputs, there could be methods for all the combinations of different types for each of the inputs.
- Such functions are called generic functions because their source code is organized to perform a generic operation but with different details depending on what kind of data are input.

Data have types

- A system where the data themselves carry type is the basis for key ideas of object-oriented programming.
- The association of types with data is consistent with the idea of a type hierarchy, with general types, subtypes, sub-subtypes, etc.
- The specialized types inherit the properties associated with their parent types.
- A generic function method applicable to a type will also be applicable to all its subtypes.
- The ideas of object-oriented programming depend on having type hierarchies and user-definable types that extend the type hierarchy.

The Nature and Representation of Biomedical Data

- The first things to consider:
  - the various forms of biomedical data
  - how such data can be
    - represented in a computer
    - manipulated by a computer program
- In a typical biology/medical textbook
  - photographs, diagrams, drawings, chemical formulas, and lots of description
    - about the attributes of biological entities such as cells, organs, tissues, fluids, chemical compounds found in all those and about the relations between these entities and their properties.

The Nature and Representation of Biomedical Data

- Some of the properties of biological objects
  - numerical (quantities),
    - the concentration of certain chemicals in blood,
    - the size of a tumor,
    - the pH (degree of acidity) in a cell, tissue, organ, or body fluid.
  - qualities that can only be named but not quantified
    - the protein(s) produced by gene transcription,
    - the presence or absence of an organ in an organism,
    - the parts of an organ.
      - [These all have names, not numerical values]

The Nature and Representation of Biomedical Data

- Cell types
  - squamous cells,
  - epithelial cells,
  - muscle cells,
  - blood cells.
    - red cells, white cells
- These categorical attributes are also related back to numerical values.
  - Since we can count cells, it is possible to report for an individual the concentrations of different types of blood cells in the blood.

The Nature and Representation of Biomedical Data

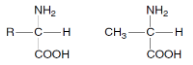
- Protein
  - One of the most important classes of constituents of living organisms
- Many hundreds of thousands of proteins found in living organisms have been named, catalogued, and their properties recorded.
  - These properties include
    - the function(s) of the protein,
    - the gene that codes for it,
    - diseases that may relate to its absence or mutation.

The Nature and Representation of Biomedical Data

- Proteins have many different roles in cells and tissues.
  - They can serve as enzymes to facilitate biochemical reactions,
    - such as metabolism of glucose.
  - They can regulate other processes by serving as signals.
- Proteins also are components of cell and tissue structure.

The Nature and Representation of Biomedical Data

- Proteins are large molecules made up of long sequences of small units (amino acids).
  - the general structure of an amino acid and a specific example, Alanine.



  - These molecules are called amino acids because of the presence of the NH_2 group, an amine group, and the presence of a COOH group, which typically gives an organic molecule properties associated with acids.

The Nature and Representation of Biomedical Data

- Each amino acid can be symbolized by a name
  - full name, an abbreviated name, or a single letter,
  - so the amino acid sequence of the protein can be represented by a sequence of names or letters.
    - This sequence is called its primary structure.
    - It is not a sequence of numbers.
- One possible way to encode this information would be to assign a number to each of the 20 different amino acids,
  - then the protein would be a number sequence.

The Nature and Representation of Biomedical Data

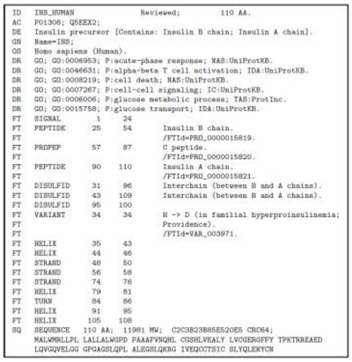
- However, it is meaningless to do arithmetic operations with these numbers.
  - They are just names.
- It is meaningful to compare two sequences to see if they are similar, that is,
  - they have the same or almost the same amino acids in the same order.
- This is a comparision operation between lists of named things or objects,
  - not a comparison between numbers.

The Nature and Representation of Biomedical Data

- Large digital data repositories are available containing information about proteins,
  - the UniProt / Swiss-Prot Knowledge Base,
    - a project of the Swiss Institute for Bioinformatics.
  - The UniProt data are downloadable in a variety of formats from the ExPASy web site

    http://www.expasy.ch,
    - maintained by the Swiss Institute for Bioinformatics.
- Next slide shows an abbreviated excerpt from a Swiss-Prot entry for the precursor protein from which human insulin is formed.

The Nature and Representation of Biomedical Data



```
ID   INS_HUMAN               Reviewed;         110 AA.
AC   P01308; Q5EEX2;
DE   Insulin precursor [Contains: Insulin B chain; Insulin A chain].
GN   Name=INS;
OS   Homo sapiens (Human).
DR   GO; GO:0006953; P:acute-phase response; NAS:UniProtKB.
DR   GO; GO:0046631; P:alpha-beta T cell activation; IDA:UniProtKB.
DR   GO; GO:0008219; P:cell death; NAS:UniProtKB.
DR   GO; GO:0007267; P:cell-cell signaling; IC:UniProtKB.
DR   GO; GO:0006006; P:glucose metabolic process; TAS:ProtInc.
DR   GO; GO:0015758; P:glucose transport; IDA:UniProtKB.
FT   SIGNAL        1     24
FT   PEPTIDE      25     54       Insulin B chain.
FT                                /FTId=PRO_0000015819.
FT   PROPEP       57     87       C peptide.
FT                                /FTId=PRO_0000015820.
FT   PEPTIDE      90    110       Insulin A chain.
FT                                /FTId=PRO_0000015821.
FT   DISULFID     31     96       Interchain (between B and A chains).
FT   DISULFID     43    109       Interchain (between B and A chains).
FT   DISULFID     95    100
FT   VARIANT      34     34       H -> D (in familial hyperproinsulinemia;
FT                                Providence).
FT                                /FTId=VAR_003971.
FT   HELIX        35     43
FT   HELIX        44     46
FT   STRAND       48     50
FT   STRAND       56     58
FT   STRAND       74     76
FT   HELIX        79     81
FT   TURN         84     86
FT   HELIX        91     95
FT   HELIX       105    108
SQ   SEQUENCE   110 AA;  11981 MW;  C2C3B23B85E520E5 CRC64;
     MALWMRLLPL LALLALWGPD PAAAFVNQHL CGSHLVEALY LVCGERGFFY TPKTRREAED
     LQVGQVELGG GPGAGSLQPL ALEGSLQKRG IVEQCCTSIC SLYQLENYCN
```

- For a complete explanation of the individual field items, consult the documentation available at the ExPASy web site

The Nature and Representation of Biomedical Data

- The DR records are cross-references,
  - in this case to the Gene Ontology (GO).
    - It is useful to be able to have a computer program look up these cross-references so that information can then be used in combination with the data shown here.
- The FT records are feature descriptions.
  - Some of the features shown are the places in the amino acid sequence where different types of structures occur, such as α-helix structures, β strands, and turns.
    - In this record, the locations of disulfide linkages are also reported.

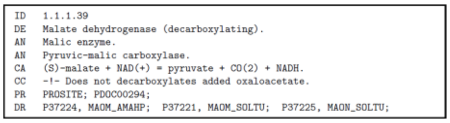
The Nature and Representation of Biomedical Data

- The names following the FT tags are the feature types.
- The numbers are the sequence start and end points for each feature.
  - This particular entry describes a polypeptide that is a precursor for the insulin protein.
  - The molecule folds up, forming the disulfide bonds indicated, and the section marked PROPEP in the FT records is spliced out, leaving the two PEPTIDE sections, linked by the disulfide bridges.
- Finally, the SQ section contains the actual sequence of amino acids, one letter for each.

The Nature and Representation of Biomedical Data

- Many proteins function as enzymes,
  - chemical compounds that facilitate chemical reactions
    - Many biologically important reactions do not proceed without the presence of the corresponding enzymes.
  - So, an important piece of information about a protein is its function.
    - Is it an enzyme, and what type of enzyme is it?
    - If it is an enzyme, what reaction(s) does it facilitate?
- Next slide shows an example, a small excerpt from the Enzyme database, at the ExPASy web site.

The Nature and Representation of Biomedical Data



```
ID   1.1.1.39
DE   Malate dehydrogenase (decarboxylating).
AN   Malic enzyme.
AN   Pyruvic-malic carboxylase.
CA   (S)-malate + NAD(+) = pyruvate + CO(2) + NADH.
CC   -!- Does not decarboxylates added oxaloacetate.
PR   PROSITE; PDOC00294;
DR   P37224, MAOM_AMAHP;  P37221, MAOM_SOLTU;  P37225, MAON_SOLTU;
```

- The line beginning with ID is the Enzyme Commission number, unique to each entry.
- The DE line is the official name,
- The AN lines are alternate names or synonyms.
  - This enzyme catalyzes the reaction that removes a carboxyl group from the malate molecule, leaving a pyruvate molecule, and in the process also converting an NAD+ molecule to NADH.

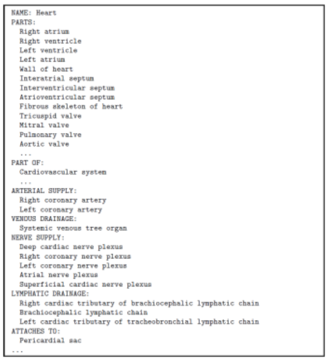
The Nature and Representation of Biomedical Data

- The NAD+ and NADH molecules are coenzymes,
  - molecules that participate in the reaction
- The reaction catalyzed by malate dehydrogenase is described in a kind of stylized symbolic form on the line beginning with CA.
- The CC line is a comment,
  - not meant for use by a computer program.
- The PR line is a cross-reference to the Prosite database,
  - which has other information about proteins
- The DR line is a set of cross-references to entries in the Swiss-Prot database
  - where the sequences corresponding to various versions of this protein may be found.

The Nature and Representation of Biomedical Data

- Bodies, organs, and tissues also have a lot of data and knowledge associated with them
  - Organs, for example, also have names, lists of their constituent parts, location within the body of the organism.
  - This information is also symbolic and consists of items that need to be grouped together in lists or more complex structures.
- Next slide shows some information about the human heart, taken from the University of Washington Foundational Model of Anatomy (FMA).

The Nature and Representation of Biomedical Data



```
NAME: Heart
PARTS:
  Right atrium
  Right ventricle
  Left ventricle
  Left atrium
  Wall of heart
  Interatrial septum
  Interventricular septum
  Atrioventricular septum
  Fibrous skeleton of heart
  Tricuspid valve
  Mitral valve
  Pulmonary valve
  Aortic valve
  ...
PART OF:
  Cardiovascular system
  ...
ARTERIAL SUPPLY:
  Right coronary artery
  Left coronary artery
VENOUS DRAINAGE:
  Systemic venous tree organ
NERVE SUPPLY:
  Deep cardiac nerve plexus
  Right coronary nerve plexus
  Left coronary nerve plexus
  Atrial nerve plexus
  Superficial cardiac nerve plexus
LYMPHATIC DRAINAGE:
  Right cardiac tributary of brachiocephalic lymphatic chain
  Brachiocephalic lymphatic chain
  Left cardiac tributary of tracheobronchial lymphatic chain
ATTACHES TO:
  Pericardial sac
...
```

- Some of this information can be used for spatial reasoning about anatomy, as applicable to surgical procedures, consequences of bodily injury, etc.
- Information about the connectivity of the lymphatic system and lymphatic drainage of various organs and anatomic locations makes it possible to construct computational models for spread of tumor cells.
- This in turn helps to determine the target for radiation therapy for cancer, resulting in improved accuracy of treatment.

The Nature and Representation of Biomedical Data

- Electronic medical records are rapidly becoming the standard of practice for managing clinical data, in medical offices and hospitals.
- Some of the information stored is numeric,
  - such as the results of many laboratory tests.
    - counts of the number of various types of blood cells, concentration of drugs, sugars, important proteins, etc.
- Some nonnumerical information is also important in these laboratory tests,
  - the units used for the tests as well as tests for the presence of bacteria and the identities of the bacteria
- Electronic medical record systems also include enormous amounts of textual information,
  - physician’s dictation of findings about the patient

The Nature and Representation of Biomedical Data

- Gene and protein sequence data use a small “alphabet” of symbols,
  - four for elements of gene sequences and 20 for the amino acids that make up proteins.
    - This encoding problem is simple by comparison with the problem of representing clinical laboratory test data.
- Laboratory data are the results of tests that the clinical laboratory performs when a sample of a patient’s blood is drawn from a vein and put in a tube or tubes to be sent to the laboratory.

The Nature and Representation of Biomedical Data

- Some examples of blood tests are
  - red and white blood cell counts,
  - hemoglobin,
  - creatinine level,
  - glucose level,
  - etc.
- These tests are usually grouped into panels that can be ordered as a unit.
- Next slide shows some values from a complete blood count for a patient, along with the standard (normal) range for each test.
  - Note that the units in which the values are reported are different for the different tests.

The Nature and Representation of Biomedical Data

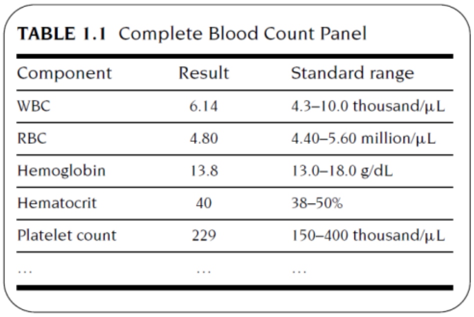


TABLE 1.1 Complete Blood Count Panel

Component	Result	Standard range
WBC	6.14	4.3–10.0 thousand/µL
RBC	4.80	4.40–5.60 million/µL
Hemoglobin	13.8	13.0–18.0 g/dL
Hematocrit	40	38–50%
Platelet count	229	150–400 thousand/µL
…	…	…

The Nature and Representation of Biomedical Data

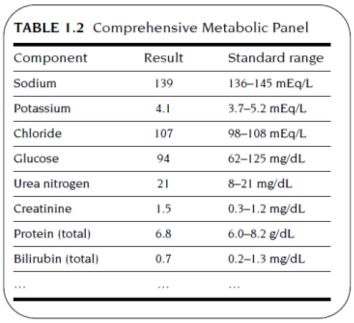


TABLE 1.2 Comprehensive Metabolic Panel

Component	Result	Standard range
Sodium	139	136–145 mEq/L
Potassium	4.1	3.7–5.2 mEq/L
Chloride	107	98–108 mEq/L
Glucose	94	62–125 mg/dL
Urea nitrogen	21	8–21 mg/dL
Creatinine	1.5	0.3–1.2 mg/dL
Protein (total)	6.8	6.0–8.2 g/dL
Bilirubin (total)	0.7	0.2–1.3 mg/dL
…	…	…

- Example of a panel that lists the names, values, and standard ranges for some of the tests in a comprehensive metabolic panel ordered on a patient.

The Nature and Representation of Biomedical Data

- Electronic medical record (EMR) systems
  - complex database systems that
    - attempt to integrate and provide ease of access to all the data needed for care providers and patients in the medical setting.
- They include facilities for
  - acquiring data from instruments,
  - data entry by practitioners in the clinic and on the wards
  - generation of reports for use by managers and others.

The Nature and Representation of Biomedical Data

- The core challenges in designing EMR systems are to come up with logical and consistent ways
  - to organize the data on all the patients, procedures, facilities, providers, etc.,
  - to easily retrieve the particular subsets of data needed at different places and times in the health care setting,
  - to organize and present the information
    - so providers and patients can easily use it.

The Nature and Representation of Biomedical Data

- the electronic health record should be a lifelong record,
    - including some things that are not now part of existing systems.
  - genome analysis,
  - family relationships (pedigree),
  - ethical conventions,
  - the patient’s own observations.
- It is also important to have EMR systems that can easily support anonymization of data for research use.

The Nature and Representation of Biomedical Data

- The basic science underlying much of public health practice is epidemiology,
  - the study of the spread of diseases and environmental toxic substances throughout populations and geographic areas.
- In addition, public health studies the organizational aspects of health care practice and how it affects the health of populations.
- Many kinds of data are used in public health research and practice.

The Nature and Representation of Biomedical Data

- Examples include
  - time series of various sorts,
    - used for things like syndromic surveillance, outbreak detection, longterm trends in morbidity and mortality, etc.,
  - vital statistics,
    - that is, birth and death records and other related data,
  - immunization records,
  - reportable disease records
    - such as tumor registries and STD registries,
  - risk factor data.

What Can Be Represented in a Computer?

- In order for computer programs to deal with biomedical data,
  - the data must be encoded according to some plan so that
    - the binary numbers in the computer’s memory represent numeric data or represent text or possibly more abstract kinds of data.
- Many encoding schemes have been used for biomedical data.

What Can Be Represented in a Computer?

- Examples include biomolecular sequence data (DNA and proteins), laboratory data from tests done on blood samples and other substances obtained from patients in a clinic or hospital, and medical image data.
- In addition, everyone needs methods for searching and sorting through the vast collections of bibliographic reference data now available such as the MEDLINE database of journal articles and other informational items.
- The entries in such bibliographic databases are definitely not numeric but are indexed by keywords, and the search methods used depend on being able to manipulate lists and complex structures.

DNA and the Genetic Code

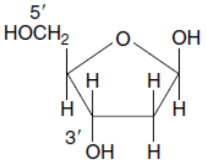
- Much of the key information that controls the operation of cells is in DNA (Deoxyribo-Nucleic Acid)
  - One of the functions of DNA is to encode (and transmit to the next generation) information from which the cell can produce proteins.
  - The information sections in the DNA that correspond to and encode for proteins are called “genes.”
    - Some of these correspond to the genes of classical genetics though we now know that the operation of inheritance and expression of genetic characteristics are very much complicated than the initial ideas discovered by Mendel much earlier.

DNA and the Genetic Code

- DNA is
  - a sequence of small molecules connected together in a kind of polymer.
  - a long double-stranded helix, where each strand is a sequence of units called nucleotides.
- Only four different kinds of nucleotides are found in DNA.
  - These nucleotides are composites of a sugar molecule (β-D-2-deoxyribose), a phosphate (PO_3) group, and one of four compounds called "bases," from the purine or pyrimidine family.

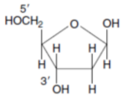
DNA and the Genetic Code

- β-D-2-deoxyribose, the sugar component of the nucleotide units of DNA, with labels to show the locations of the 3′ and 5′carbon atoms, where the phosphate groups attach.



DNA and the Genetic Code

- The "backbone" of the DNA molecule is a sequence consisting of alternating phosphate groups and sugar (deoxyribose) molecules.



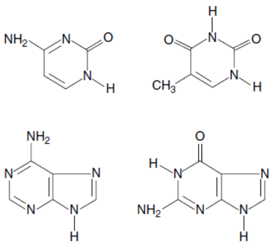
- The phosphate group connects the sugar molecules together by bonding at the carbon atoms labeled 3′ and 5′.
- The OH at the 3′ carbon connects to the phosphate group (PO_3), and the OH at the 5′ carbon connects to the other end of the phosphate group, splitting out a water (H_2O) molecule in the process.
- The bases connect to the sugar molecules at the 1′ position.

DNA and the Genetic Code

- The bases are guanine, adenine, cytosine, and thymine
- The DNA sequences are typically described by letter sequences naming the bases at each position in one of the two strands making up the DNA molecule.
- The nucleotides themselves are sometimes also called bases because they are uniquely identified by which base each includes.

DNA and the Genetic Code

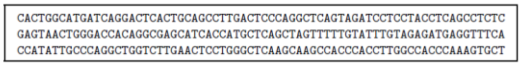
The chemical structures of the bases



- The pyrimidine bases,
  - cytosine and thymine,
    - two of the four bases that are constituents of the nucleotide units of DNA.
- The purine bases,
  - adenine and guanine,
    - the other two bases that are constituents of the nucleotide units of DNA.

DNA and the Genetic Code

- In most data sources, a typical representation of a DNA molecule or sequence would consist of a sequence of letters, G, C, A, and T, to represent each of the possible four nucleotide (also called "base") pairs that could appear in a double helix DNA strand.
  - Although the DNA molecule is double-stranded, the bases are paired uniquely,
    - A with T and G with C, so that only the bases on one strand need to be represented.



```
CACTGGCATGATCAGGACTCACTGCAGCCTTGACTCCCAGGCTCAGTAGATCCTCCTACCTCAGCCTCTC
GAGTAACTGGGACCACAGGCGAGCATCACCATGCTCAGCTAGTTTTTGTATTTGTAGAGATGAGGTTTCA
CCATATTGCCCAGGCTGGTCTTGAACTCCTGGGCTCAAGCAAGCCACCCACCTTGGCCACCCAAAGTGCT
```

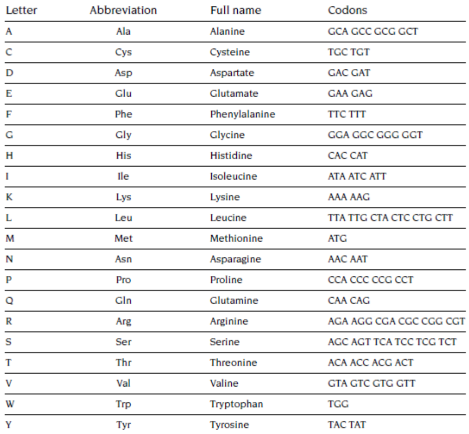
- a small fragment of the region around a well-known gene associated with breast cancer called BRCA

The Fundamental Dogma of Molecular Biology

- The relation between DNA and proteins is called the "genetic code."
- Each amino acid corresponds to one or more patterns of three nucleotides.
  - For example, the nucleotide sequence GGA corresponds to the amino acid glycine, and TTC corresponds to the amino acid phenylalanine.
- Each combination of three nucleotides is called a codon.

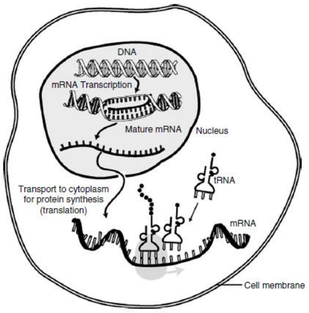
The Fundamental Dogma of Molecular Biology

- With four possible nucleotides in three places, there are 64 (4×4×4) codons.
- Not all codons correspond to amino acids;
  - there are three special codons that signal the end of a sequence, TAA, TAG, and TGA.
- For most of the amino acids, there are several codons that represent the same amino acid.
  - For example, the amino acid lysine is represented by two codons, AAA and AAG, and leucine is represented by any one of six.



Letter	Abbreviation	Full name	Codons
A	Ala	Alanine	GCA GCC GCG GCT
C	Cys	Cysteine	TGC TGT
D	Asp	Aspartate	GAC GAT
E	Glu	Glutamate	GAA GAG
F	Phe	Phenylalanine	TTC TTT
G	Gly	Glycine	GGA GGC GGG GGT
H	His	Histidine	CAC CAT
I	Ile	Isoleucine	ATA ATC ATT
K	Lys	Lysine	AAA AAG
L	Leu	Leucine	TTA TTG CTA CTC CTG CTT
M	Met	Methionine	ATG
N	Asn	Asparagine	AAC AAT
P	Pro	Proline	CCA CCC CCG CCT
Q	Gln	Glutamine	CAA CAG
R	Arg	Arginine	AGA AGG CGA CGC CGG CGT
S	Ser	Serine	AGC AGT TCA TCC TCG TCT
T	Thr	Threonine	ACA ACC ACG ACT
V	Val	Valine	GTA GTC GTG GTT
W	Trp	Tryptophan	TGG
Y	Tyr	Tyrosine	TAC TAT

The Fundamental Dogma of Molecular Biology



- Illustration of the process of transcription and translation from DNA to mRNA to proteins

Representing DNA in Computer Programs

- In a typical computerized encoding, each letter is represented by its ASCII code,
  - each occupies 8 bits in a text file
    - although ASCII is a 7-bit code, it is usual to use 8 bit "bytes".
- We can represent base sequences as letter sequences or long strings.
- Another way to represent such sequences is to represent each nucleotide (or base) as a symbol in a list.

Representing DNA in Computer Programs

- The entire sequence then becomes a list of symbols and looks like this:
  - (C A C T G G C A T G A T C A G G A C T C A C T G C A G C C T T G A C T C C C A G G C T C A G T A G A T C C T C C T A C C T C A G C C T C T C G A G T A A C T G G G A C C A C A G G C G A G C A T C A C C A T G C T C A G C T A G T T T T T G T A T T T G T A G A G A T G A G G T T T C A C C A T A T T G C C C A G G C T G G T C T T G A A C T C C T G G G C T C A A G C A A G C C A C C C A C C T T G G C C A C C C A A A G T G C T)

Anatomy

- Anatomy is the science that studies the structure of the body
- The process of mapping out the structure of the human body has been going on for several thousand years.
  - As early as 1600 B.C.E., the ancient Egyptians apparently knew how to deal with a considerable range of injuries by surgical procedures as well as other basic medical knowledge, such as the role of the heart in circulating body fluids.

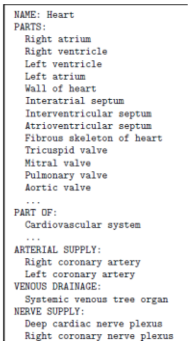
Anatomy

- Development of X-ray imaging
- Development of cross-sectional imaging using CT scanner, and later the MRI scanner
- The ability to produce images of living animals (including humans), without invasive surgical procedures,
- has revolutionized diagnostic medicine as well as opened new possibilities for visualization of anatomy for teaching purposes.

Anatomy

- Two separate but related kinds of information about anatomy are useful to represent in computerized form.
- First,
  - the logical relationships discovered in ancient times, and revised many times, express what kinds of things are contained in a human body how these things are related.
    - This is the kind of information found in the FMA

An excerpt from the University of Washington Foundational Model of Anatomy (FMA)



```
NAME: Heart
PARTS:
  Right atrium
  Right ventricle
  Left ventricle
  Left atrium
  Wall of heart
  Interatrial septum
  Interventricular septum
  Atrioventricular septum
  Fibrous skeleton of heart
  Tricuspid valve
  Mitral valve
  Pulmonary valve
  Aortic valve
  ...
PART OF:
  Cardiovascular system
  ...
ARTERIAL SUPPLY:
  Right coronary artery
  Left coronary artery
VENOUS DRAINAGE:
  Systemic venous tree organ
NERVE SUPPLY:
  Deep cardiac nerve plexus
  Right coronary nerve plexus
```

- The essential idea here is to be able to represent symbols and use lists to group them into entities, attributes, and relations.
- Thus, the heart information could become a list structure with the attributes labeled by symbolic tags and the values implemented as lists of symbols

Anatomy

- In addition to the logic of anatomy, however, two additional very important kinds of data need representation as well.
  - the image data,
    - X-ray projected images and cross-sectional images
    - photographic images,
      - for example, of skin lesions or open views into a body during surgery.
      - Microscope images provide views of anatomy at the tissue and cellular level.

Anatomy

- So, we need a representation and methods for computing with (and display of) image data.
- Images are usually represented as arrays of numbers, with each number or set of numbers corresponding to a color or gray level for a particular spot or pixel in the image.
- The image is then considered to consist of a rectangular array of such spots.
- If the resolution is high enough, a considerable amount of detail can be discerned.

Anatomy

- Finally, the anatomic objects that appear in image data sets often need to be identified and delineated with their exact shapes represented, for purposes of illustration, and more importantly to perform measurements on these organs and other components (e.g., tumors).
- The exact location of such objects in a particular cancer patient's body is crucial information for designing and delivering accurate and effective radiation treatment.

Medical Laboratory Data

- An example to represent laboratory results,
  - the red and white cell counts in the blood of a patient.

```
(cbc
(wbc 6.14 thou-per-microltr)
(rbc 4.80 mill-per-microltr)
(hemoglobin 13.8 grams-per-dl)
(hematocrit 40 percent)
(platelets 229 thou-per-microltr))
```

cbc: complete blood count
wbc: white cell count
rbc: red cell count
thou-per-microltr: thousands per microliter

- This expression encodes part of the results of a complete blood count

Medical Laboratory Data

- For the laboratory tests, one would want to look up a particular value by name,
  - for example, the white blood cell count, which will give some indication of the presence of infection.
    - These laboratory tests are all numerical values and have a certain regularity,
    - There are thousands of such tests, many with nonnumeric result types and with highly variable structure.
    - A lot of important information about the test are not incuded, for example,
      - how it was performed, under what conditions, and what constitutes a normal range of results for the particular laboratory.

Medical Laboratory Data

- In a typical clinical laboratory, the instruments produce results in digital form, and these results are stored in computers running clinical laboratory system software.
- These systems have in the past been made accessible through local area network access but that required clinicians and other staffs on the wards to learn how to use the specific laboratory system software and, of course, required the clinical laboratory to manage access by issuing user IDs and passwords to everyone.

Medical Laboratory Data

- The next step in evolution was to provide some way to transmit the laboratory data to another system, the electronic medical record,
  - which would provide an integrated view of all the sources of medical data as well as provide a central system for registration of patients, scheduling clinic visits, tracking inpatients, accepting orders for medications, and generating the bills to the patients or their insurance plans.
    - At present, the only generally accepted standard for constructing, sending, and interpreting messages with this data in them is the Health Level 7 (HL7) standard, developed and managed by a consortium known as HL7.

Medical Laboratory Data

- The HL7 standard does not specify the names, encoding of values, grouping of tests into panels, or other low-level details.
- Many systems have been put into use, and the definitions of these panels may vary from one institution to the next or even from one device manufacturer to the next.
- So, to integrate many such systems even in a single institution, it is necessary to translate one set of definitions into the others.

Medical Laboratory Data

- Electronic medical record systems cannot, therefore, be standardized and simply installed.
- Each hospital needs to define its own tables of tests, conventions for displaying results, and display screens that organize the data the way it is needed for that institution.
- To exchange data between institutions is of course very much more difficult.

Medical Laboratory Data

- To achieve meaningful data interchange, one of the first steps needed is
  - to create a standard system of nomenclature for the tests and for the encoding of the test result values.
- A well-known scheme for labeling and encoding laboratory test results is
  - the Logical Observation Identifier Names and Codes (LOINC) system.

Medical Laboratory Data

- The LOINC system proposes a set of names and codes for labeling laboratory results and clinical observations.
- This involves achieving agreement about
  - the type of a datum,
  - its external representation,
  - the units in which it is measured,
  - other factors that affect the interpretation of the result being reported.

Medical Laboratory Data

- The goal of LOINC is
  - to assign unique codes and names to each of the thousands of observations and laboratory tests in use in clinical practice
  - so that the associated data for a particular patient can be labeled in a computerized message transmitting the data between various electronic medical record (EMR) systems and other application programs using medical data.

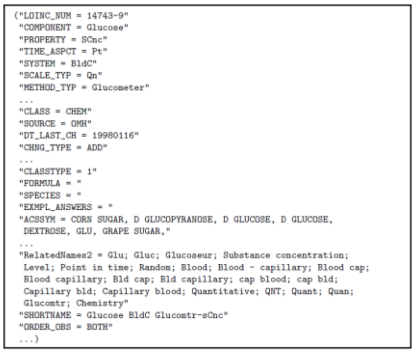
Medical Laboratory Data

- These same codes can be used by an electronic medical test order entry system
  - to identify what a care provider has ordered for a patient.
- Although LOINC started out with a focus only on the clinical laboratory, today it encompasses a broad range of tests and observations throughout medical practice.
- The current set of identifiers has over 50,000 entries.

Medical Laboratory Data

- The LOINC table is available from the Regenstrief Institute as a text file.
  - https://loinc.org/downloads/loinc
- The entries are organized into one record per line of text
  - lines are delimited by a <CR><LF> two-character sequence.
- Within a record or single line, individual fields are delimited by <tab> characters, used as separators,
  - two successive <tab> characters indicate a blank or omitted field.
- Strings are represented as text surrounded by double-quote characters.
- The few fields that are integers.
- This makes it easy to read in the LOINC codes into a list for processing in a program.

An excrept from a record in LOINC



```
("LOINC_NUM = 14743-9"
 "COMPONENT = Glucose"
 "PROPERTY = SCnc"
 "TIME_ASPCT = Pt"
 "SYSTEM = BldC"
 "SCALE_TYP = Qn"
 "METHOD_TYP = Glucometer"
 ...
 "CLASS = CHEM"
 "SOURCE = OMH"
 "DT_LAST_CH = 19980116"
 "CHNG_TYPE = ADD"
 ...
 "CLASSTYPE = 1"
 "FORMULA = "
 "SPECIES = "
 "EXMPL_ANSWERS = "
 "ACSSYM = CORN SUGAR, D GLUCOPYRANOSE, D GLUCOSE, D GLUCOSE,
  DEXTROSE, GLU, GRAPE SUGAR,"
 ...
 "RelatedNames2 = Glu; Gluc; Glucoseur; Substance concentration;
  Level; Point in time; Random; Blood; Blood - capillary; Blood cap;
  Blood capillary; Bld cap; Bld capillary; cap blood; cap bld;
  Capillary bld; Capillary blood; Quantitative; QNT; Quant; Quan;
  Glucomtr; Chemistry"
 "SHORTNAME = Glucose BldC Glucomtr-sCnc"
 "ORDER_OBS = BOTH"
 ...)
```

Medical Images

- Objective is to obtain images of the internal structure of the human body by using one of the imaging modalities
  - X-Ray
  - MRI (Magnetic Resonance Imaging)
  - Ultrasound imaging
  - PET (Positron Emission Tomography)
  - Electrical Empedans Tomoghraphy
  - Nuclear imaging

Medical Images

- Digital images consist of two kinds of information:
- First, the image itself, consisting of an array of numbers, which can be integer or decimal.
  - Each number represents a picture element (pixel).
  - At that spot in the image,
    - the display may show a monochrome brightness corresponding to the value of the number or
    - a color with brightness, hue, and saturation qualities
      - a mixture of red, green, and blue intensities

Medical Images

- The size of the array may vary for a given modality.
  - For CT images, an array of 512 rows by 512 columns is common.
- The array may not be square;
  - computed radiographs that correspond to chest films are typically portrait style,
    - with more pixels in the vertical dimension than in the width, which corresponds to the typical viewport shape of an X-ray film.

Medical Images

- Second, the descriptive data that specifies
  - what kind of image it is
  - how it was obtained
  - who the subject (patient) is
  - where in some patient or machine-centric coordinate system the image is located
  - its orientation
  - many other possibly useful items
- These are encoded in many different ways

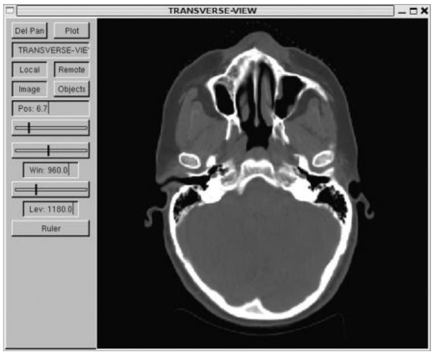
Medical Images

- Images may be aggregated into sets,
  - the set of images may have its own attributes, which apply in common to all the images.
- One very important concept is that of a position-related set,
  - where all the images share the same coordinate system, and are oriented and located in that coordinate system, relative to each other.
    - This allows the possibility that the set of 2D images can be reformatted to provide 2D images in other planes through the volume that the set represents.

Medical Images

- The images can be coalesced or interpolated to form a regular 3D array of image data.
- Solid models or surfaces that represent the 3D structures seen in cross-section in the images can then be computed from it.
- The advent of X-ray transaxial CT scanners in the 1970s marked the beginning of the digital era of medical radiology.
  - The first scanners produced transverse cross-sectional images of heads, an example of which is in the next slide

A CT image of a person’s head, showing the cerebellum, sinus cavities, and nose



Medical Images

- These images are produced by projecting a very thin “fan” X-ray beam through a patient’s head, from each of many angles.
- The transmitted X-ray intensities are measured by photon (X-ray) detectors on the opposite side of the patient from the X-ray tube.
- Each intensity is related approximately to the integral of the X-ray linear attenuation coefficients of the tissues along the line from the X-ray source to the detector.

Medical Images

- The attenuation coefficients are in turn related to the tissue densities since the attenuation of the X-rays is mostly due to Compton scattering, which depends only on the electron density, not on the composition of the tissue.
- The image data, or, equivalently, the tissue densities can be reconstructed from the projected data, using a method called filtered back-projection.

Medical Images

- This method relies on a property of the Fourier transform called the projection-slice theorem, which relates the one-dimensional Fourier transform of the line integral data mentioned above to the 2D Fourier transform of the original function.
- So, if you can measure the projections in all directions, you can recover the original function.

Medical Images

- The first person to demonstrate the effectiveness of this idea as a practical medical device was Geoffrey Hounsfield
  - for which he was awarded the Nobel Prize.
- Hounsfield’s research and development was supported by EMI,
  - the same company that produced and marketed the first commercial CT scanner.
    - This company was also the publisher of the music of The Beatles. It has been noted that the CT scanner is the Beatles’ greatest legacy since Hounsfield’s work was funded by EMI’s profits from the Beatles’ phenomenal record sales.

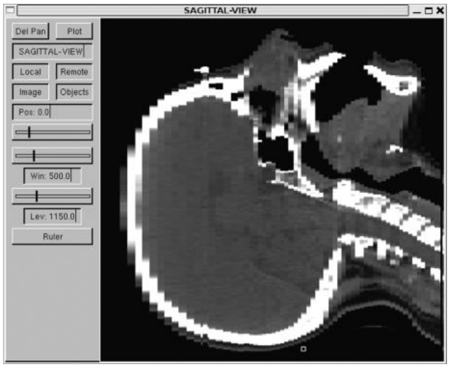
Medical Images

- The images are 2D arrays of numbers representing the image intensity
  - the spots in the image that the numbers represent are called pixels.
  - In a typical CT image, each pixel may represent a 1 to 2 mmsquare cross-sectional region.
- The image pixel numbers represent average tissue densities over some thickness,
  - roughly the width of the fan-shaped X-ray beam used to produce the data from which the image is computed.
  - This width is called the slice thickness.

Medical Images

- When the slice thickness is large, it will be difficult to detect small spots,
  - as they will be averaged with the surrounding tissue and not stand out so much in contrast.
- As the ability to produce images with thinner and thinner X-ray beams developed, it became possible to get high resolution in the direction perpendicular to the image plane,
  - this made it possible to construct images in orthogonal planes from the transverse images by selecting pixels from a series of transverse images as if they were stacked into a 3D array and reorganizing them for display.
- Following slide shows a sagittal cross-sectional image constructed in this way.

Asagittal cross-section through the midline of the same person, constructed from a series of 80 transverse images



Pos: 0.0

Win: 500.0

Lev: 1150.0

Medical Images

- In order to properly display the image data, a computer program needs substantial information about the image pixel array and the image it represents.
- This includes such things as
  - the dimensions of the array
    - how many pixels are in a row, how many rows are in the image
  - what the image pixel numbers represent
    - what numbers should correspond to maximum brightness,
    - where should the midrange be, etc.

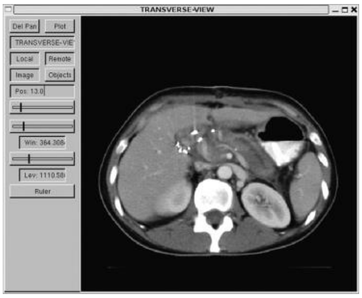
Medical Images

- To reorganize the transverse image data into sagittal images, it is also necessary to know the table index of each image so that
  - the pixels from that image can be placed properly along the perpendicular axis.
- Other information that is important are
  - the name of the patient
  - other identifying information
  - the reason for getting the images
  - techniques used to get image

Medical Images

- One use of cross-sectional images is to construct organ models and identify tumor locations for radiation therapy planning (RTP)
- Much work has been done on image processing algorithms to automate the process of finding organ boundaries so that 3D models can be constructed.
- Next slide shows a cross-section of an abdomen, where the patient’s left kidney can easily be identified
  - (the view is from feet to head, and the patient is on his/her back, so the patient’s left is to your right).

Across-section of a patient’s abdomen



Medical Images

- The image pixel values are represented in files and in computer programs as binary numbers, not as text.
  - This saves space and is also important because image processing algorithms do arithmetic on these data.
- Radiology departments today are almost completely filmless.
  - The development of high-resolution X-ray detector arrays has made digital radiography (DR) a reality.
- In addition, many other digital imaging systems are part of the standard of practice, including ultrasound, MRI, and various systems using radioisotopes, such as PET.

Medical Images

- As an example, in CT imaging;
  - Pixel values are typically small integers in the range from 0 to 4095 or sometimes larger up to 65,535
    - i.e., they can be represented in 12-bit or 16-bit integer
  - A typical image consists of a 512 row by 512 column 2D array of such numbers.
- A simple way to encode these is as nested lists of lists,
  - a list of numbers for each row of pixels and a list of such lists for all the rows of an image.
  - Columns correspond to the individual elements of the row lists.

Metadata

- A set of data that describes and gives information about other data.
  - Descriptive metadata
    - For finding or understanding a resource
  - Administrative metadata
    - Technical metadata
      - For decoding and rendering files
    - Preservation metadata
      - Long-term management of files
    - Rights metadata
      - Intellectual property rights attached to content
  - Structural metadata
    - Relationships of parts of resources to one another
  - Markup languages
    - Integrates metadata and flags for other structural or semantic features within content

Metadata

- Metadata is defined as the data providing information about one or more aspects of the data
- It is used to summarize basic information about data which can make tracking and working with specific data easier
- Some examples include:
  - Means of creation of the data
  - Purpose of the data
  - Time and date of creation
  - Creator or author of the data
  - Location on a computer network where the data was created
  - Standards used
  - File size

- http://www.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf

Metadata

- There are several ways to organize biomedical data, including the use of **tags** to label items and the use of structure to organize data.
- Although the tag representation seems appealing and is easy for a person to read and understand, the tags have no particular meaning to a computer program.
- Data are meaningful only when accompanied by some interpretation that defines what the symbols mean.
- While this may seem obvious, it is easy to overlook in the rush to create repositories of “self-describing data.”

Metadata

- For example,
  - what does the following list of numbers mean?

    5984107
    8278719
    2214646
    8220013
    5433362
    4274566
    ...

Metadata

- What really are they?
  - telephone numbers?
  - patient identification numbers assigned by a hospital?
  - identification numbers assigned to terms in a terminology system?
  - a series of pixel values in a medical image,
  - a microscope image?
  - a packed, encoded representation of expression array data?
  - a very compact encoding of laboratory test data?
- Some conventions are needed as well as labeling to clearly identify the meaning.

Tags as Metadata

- The description, interpretation, or identification of the meaning of a particular set of data can take many forms.
- The data set in previous slide may be accompanied by a written document specifying the meaning and format of the data.
- Another approach is to include with the data some identifying tags that label the elements,
  - so-called self-describing data.
  - Next slide shows how the numbers above might appear using such a scheme:

Tags as Metadata

```
<phone numbers>
        <work>5984107</work>
        <home>8278719</home>
</phone numbers>
<phone numbers>
        <work>2214646</work>
        <home>8220013</home>
</phone numbers>
<phone numbers>
        <work>5433362</work>
</phone numbers>
<phone numbers>
        <cell>4274566</cell>
</phone numbers>
...
```

Tags as Metadata

- Now, we can see that indeed they are telephone numbers,
  - They are organized so that numbers belonging to a single person or entity are grouped together.
- Of course, there must be some agreement as to the meaning of the tags, which is still outside the context of the data and tags.
  - The words “phone numbers” and the other tags have no intrinsic meaning to a computer program.
- This idea of grouping items and providing tags is such a powerful way to organize information that an internationally popular syntax called XML was invented to support the idea.

Tags as Metadata

- The XML standard provides two methods for specifying the structure of the data labeled by a tag,
  - XML data type definitions (DTD)
  - XML Schemas
- One might then say, “We just include schemas or DTD that define the meaning of the tags.”
  - Such schemas do not define the meaning of the tags.
  - These only define the allowed or expected syntax or structure of the data.
    - A computer program can use a schema or other such “metadata” to check if a document or data structure is correctly constructed, but it cannot discern what to do with the data.

Tags as Metadata

- In any case, there must be some external document describing the agreement on the meaning of the elements of the data type description language.
- There is always some boundary beyond which a separate documented agreement about meaning must exist.
- The ongoing debate about self-describing data is not about whether data should be self-describing (which is impossible), but about where this boundary will be, and what description languages will be created for use inside the boundary (and documented outside the boundary).

Tags as Metadata

- In order for a computer program to process the data, the program itself must recognize the tags,
  – so the meaning of each tag is encoded in the program.
- Using human readable tags can mislead the person unfamiliar with how computer programs work.
  – A computer program does not understand the meaning of the phrase “phone numbers.”
- It can be written to take some particular action when it encounters that text string as a piece of data, but the way it recognizes the tag is by comparing with some other information containing the tag.
  – So, a computer program is also in a sense a data description.
- The designer(s) of such systems and data repositories will have to choose how this matter is addressed.

Tags as Metadata

- Computer programs are notoriously intolerant of misspellings or other variations.
  – A slight change in the tag creates an entirely different distinct tag.
- Humans have the ability to handle variant words and phrases, but, in general, computer programs do not.
  – If the tag read “phone number” (with the “s” left out), a program written to process the tags used above would likely break.
- A program can be written to handle a range of such variations resulting more sophisticated program, but that too must be completely and precisely specified

Source Code as Metadata

- If a record structure is defined without tags, in which the data follow some fixed conventions for format and encoding, as in DNA and protein sequence data, there may still be metadata.
- In a relational database, the kind of data stored in each field of a record is specified by a statement in the standard relational database language, SQL.
- This SQL statement also is a piece of metadata.
- The distinction is important in designing systems.

Source Code as Metadata

- One can design a system that has a high level of generality by making the system use the metadata as a guide to how to handle the data.
- This idea makes a database system possible
  – you write a description of your data in SQL, and the database system interprets the SQL to facilitate
    - entry,
    - storage,
    - retrieval

  of your data, using the tags you specified.

Source Code as Metadata

- The use of structured data types in ordinary programming languages is an even more generalized version of the same thing.
  – However, in this case, we are referring to parts of the program itself as metadata although normally we think of structure definitions as code.
- The distinction between code and data in a program is not always as one conventionally learns in introductory C or java programming.
- The code that defines structures in a program becomes metadata when the program can actually use it explicitly during execution.

121