SHHS Data Dissemination to the Research/Academic Community: Needs and Opportunities

Kenneth A. Loparo
Nord Professor of Engineering
EECS Department
Case Western Reserve University

SHHS Data Dissemination Project:

- The objective of the web-based SHHS Reading Center data dissemination project is to provide researchers with PSG and covariate data collected from participating SHHS centers over many years.
- The current implementation supports sleep research directed at (1) investigating clinical or epidemiological associations, (2) developing or testing scoring algorithms, or (3) cross-validating findings.
- Due to the large number of subjects, the web site provides a very convenient way to query and select studies that meet the current needs of sleep researchers.
- The electronic distribution of the data makes the data accessible to researchers worldwide.

Challenges

- Data Format: PSG data from various studies are collected with a variety of proprietary data formats depending upon the collection and scoring software utilized. A common non-proprietary format for signal data is required.

(2) Matching PSG with Covariate Data: Covariate data is essential to ensure that the population of PSG subjects represents the desired population of interest. The selection of studies from a large dataset for any research project involves applying often complex exclusion and inclusion criteria.

Implementation

- Data is downloadable in EDF format with and without annotations
- The SHHS database contains raw PSG signals, summary PSG variables, covariate data, and additional data including which includes other physiological measures and questionnaire data. The dynamic query tool allows the investigator to specify research criteria online using key covariate data to determine how many available studies meet their criteria.

Web Site Operation:

Step 1: Create query based on demographic and covariate variables (e.g. age, gender, race, BMI, …). For example, a researcher can select all the subjects who are in the age range of 60 to 70 years old, BMI between 25 to 30, Male with certain Arousal Index.

Step 2:Choose the data format for download. Currently the data can be obtained in European Data Format (EDF) with and without annotations in the following formats:

a) EDF (European Data Format): Raw PSG data in binary format

b) XML: Study annotations

c) Individual Level Dataset (TXT): Covariate Variables in text format

d) Individual Level Dataset (XML): Covariate Variables in XML format

e) Summary Dataset (TXT): Statistical Summary of the covariate

Web Site Operation (cont.)

Step 3: Variable selection for an individual level dataset or summary dataset (access to individual level data is restricted to SHHS collaborators, non collaborators only receive a statistical summary of the dataset).

Note: Selection of covariate variables is accomplished in a hierarchical manner- variable classes, then variable sub- classes and then individual variables.

Step 4: After variable selection a summary of the query is presented for verification before submission for processing. After the query is processed and their data is compiled users will receive an email with the link to the FTP site to download their studies.

Step 5: With a username and password, a user can log into the FTP server and download their studies.

EDF Data Format for SHHS Data Dissemination:

- European Data Format (EDF) has been in use for PSG data since the 1987 international Sleep Congress in Copenhagen
- EDF is a simple and straightforward format for storing and exchanging multi-channel biological signals over a variety of platforms
- EDF does not easily handle annotations/events directly

A Typical PSG Study and EDF Data Specification:

CHANNEL	SAMPLING RATE	RESOLUTION
SaO2 (Systemic arterial Oxygen saturation)	1 Hz	16 bit
H.R. (Heart Rate)	1 Hz	16 bit
EEG (Electroencephalogram)	125 Hz	8 bit
EEG (second)	125 Hz	8 bit
ECG (Electrocardiogram)	250 Hz	8 bit
EMG (Electromyogram) (Chin or Leg Movement)	125 Hz	8 bit
EOG (L) (Electrooculography) (Eye Movement)	50 Hz	8 bit
EOG (R)	50 Hz	8 bit
SOUND	10 Hz	8 bit
AIRFLOW	10 Hz	8 bit
THOR RES (Thoracic Respiratory)	10 Hz	8 bit
ABDO RES (Abdominal Respiratory)	10 Hz	8 bit
POSITION	1 Hz	16 bit
LIGHT	1 Hz	16 bit
NEWAIR	10 Hz	16 bit
OX stat	1 Hz	16 bit

General Format of an EDF file:

HEADER
Version of the data format
Local patient identification
Local recording identification
start date of recording
Start time of recording
Number of bytes in header record
Number of data records
Duration of a data record, in seconds
Number of signals in data record
HEADER (Specification of each individual signal)
Label
Transducer type
Physical dimension
Physical minimum
Physical maximum
Digital minimum
Digital maximum
Pre-filtering
Number of samples in each data record
DATA RECORD
First signal in the data record
Second signal in the data record
.
.
.
First signal in the data record
Second signal in the data record
.
.

Drawbacks of the EDF Format:

- All signal data are saved as 16 bit in two's complement. However, not all the requires 16 bit resolution. For example, light status is a binary variable.
- Annotations must be saved in a different file (This problem is addressed in EDF+ format)
- Data recording for EDF must be continuous (This problem is addressed in EDF+ format)

Challenges for Data Dissemination to the Research/Academic Community

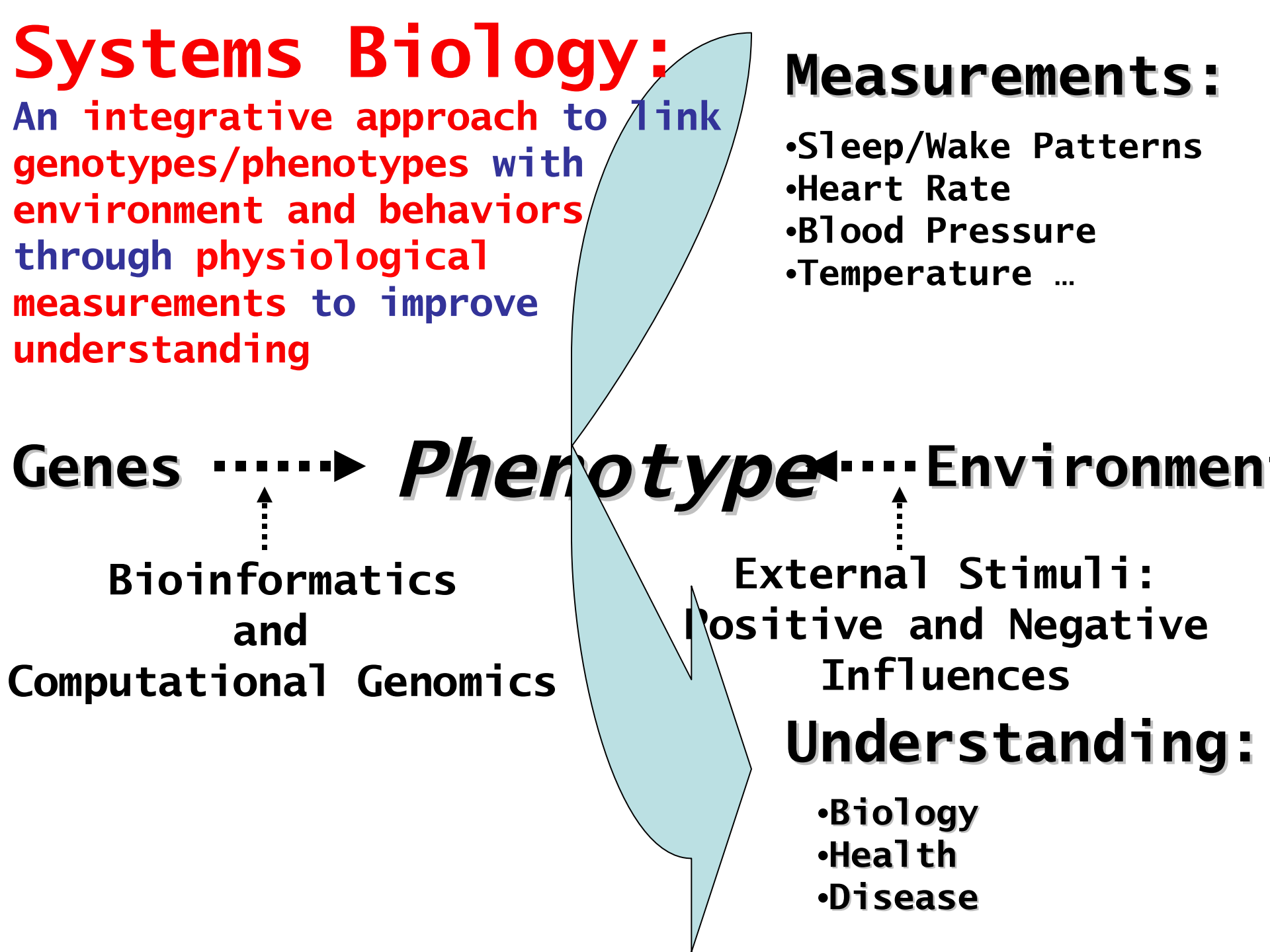
- Enabling the new Systems Biology research agenda in sleep
- Translating this agenda into the data needs of the Research/Academic Community
- Translating these data needs into a data format/structure that can support the questions being addressed

Systems Biology:

An integrative approach to link genotypes/phenotypes with environment and behaviors through physiological measurements to improve understanding

Measurements:

- Sleep/Wake Patterns
- Heart Rate
- Blood Pressure
- Temperature …



Understanding:

- Biology
- Health
- Disease

Example Research Problem

Investigate the relationship between time series features of EEG sleep and RDI across a population

Needs:

(1) Search data according to demographic and covariate data

(2) Identify epochs across studies where RDI criteria is met, label these epochs according to sleep state

(3) Collect all labeled epochs into a file(s) for download

Select and Query the PSG studies based on Events

Current Status

- Users can search and query the database based on Demographic and Covariate variables.
- Demographic variables such as: Participant ID, Age, Education Level, Gender, Marital Status, Race
- Covariate variables such as: Co-morbidities, any Additional Measurements, PSG Study Outcome Data, PSG Study Quality

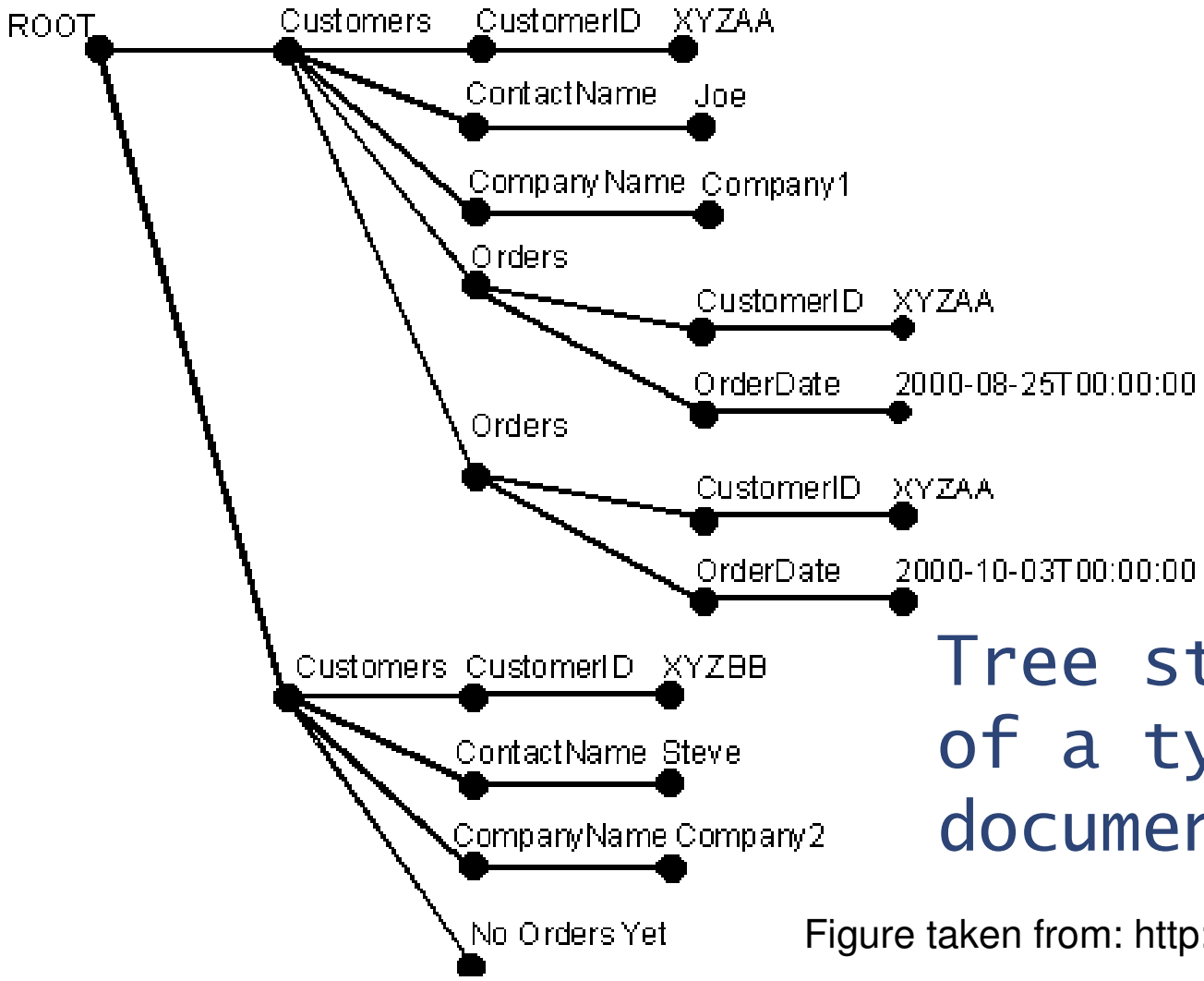
Needs

- Search according to Events such as:
  - different sleep stages
  - mark the epochs in the studies for specific events such as Apnea, Hypopnea, etc.
- Provide a tool to select and query the database based on the events in conjunction with demographic and covariate variables.

Methodology

- First Step: Develop an application for event extraction
- Second Step: Design a database structure for events
- Third Step: Develop an application to query the events, find the corresponding studies, extract and compile data in open exchange format for users

Current Situation: Annotations in XML file



Tree structure of a typical XML document

Figure taken from: http://msdn.microsoft.com/

The process of structuring XML documents to database tables

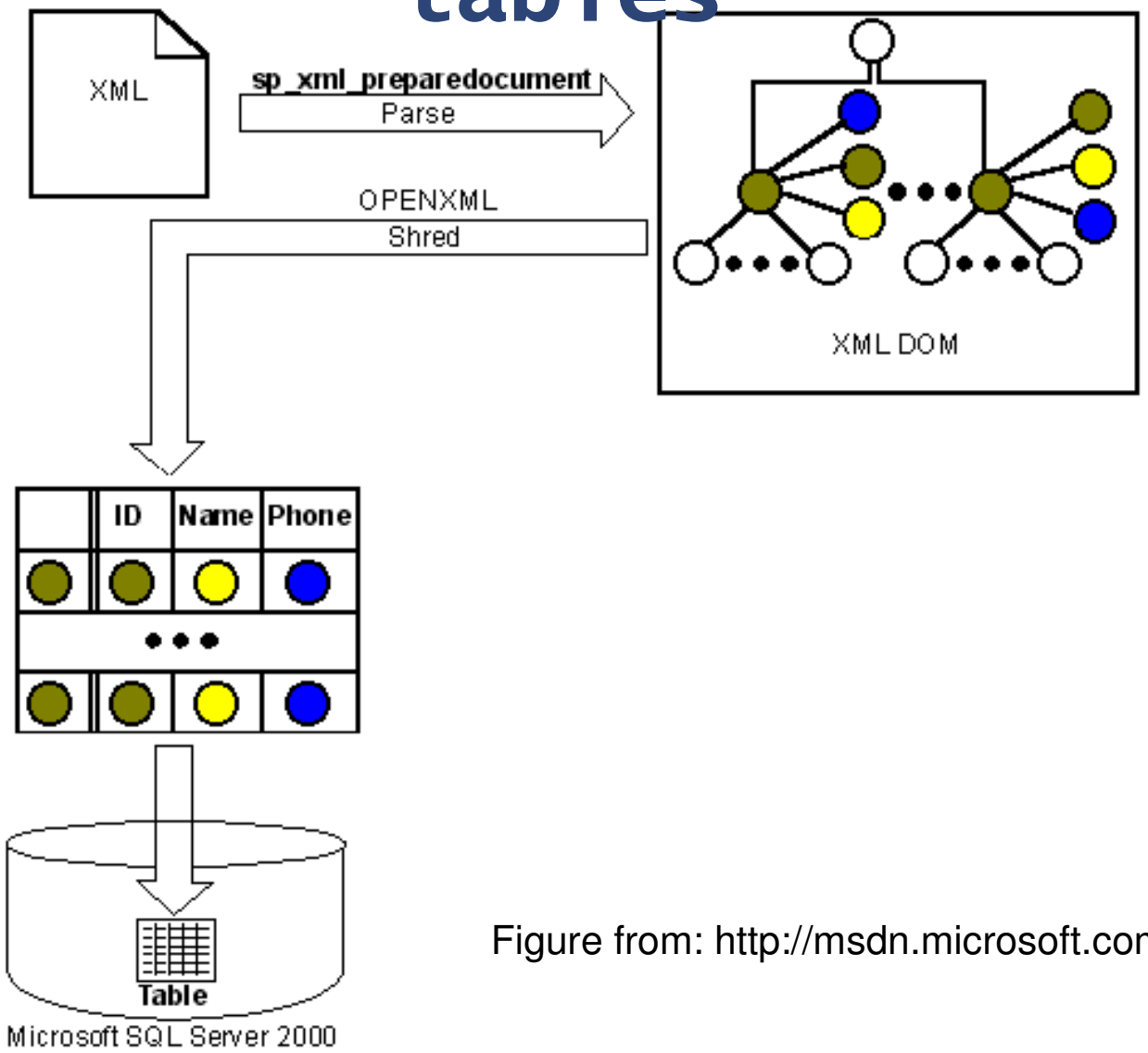


Figure from: http://msdn.microsoft.com/