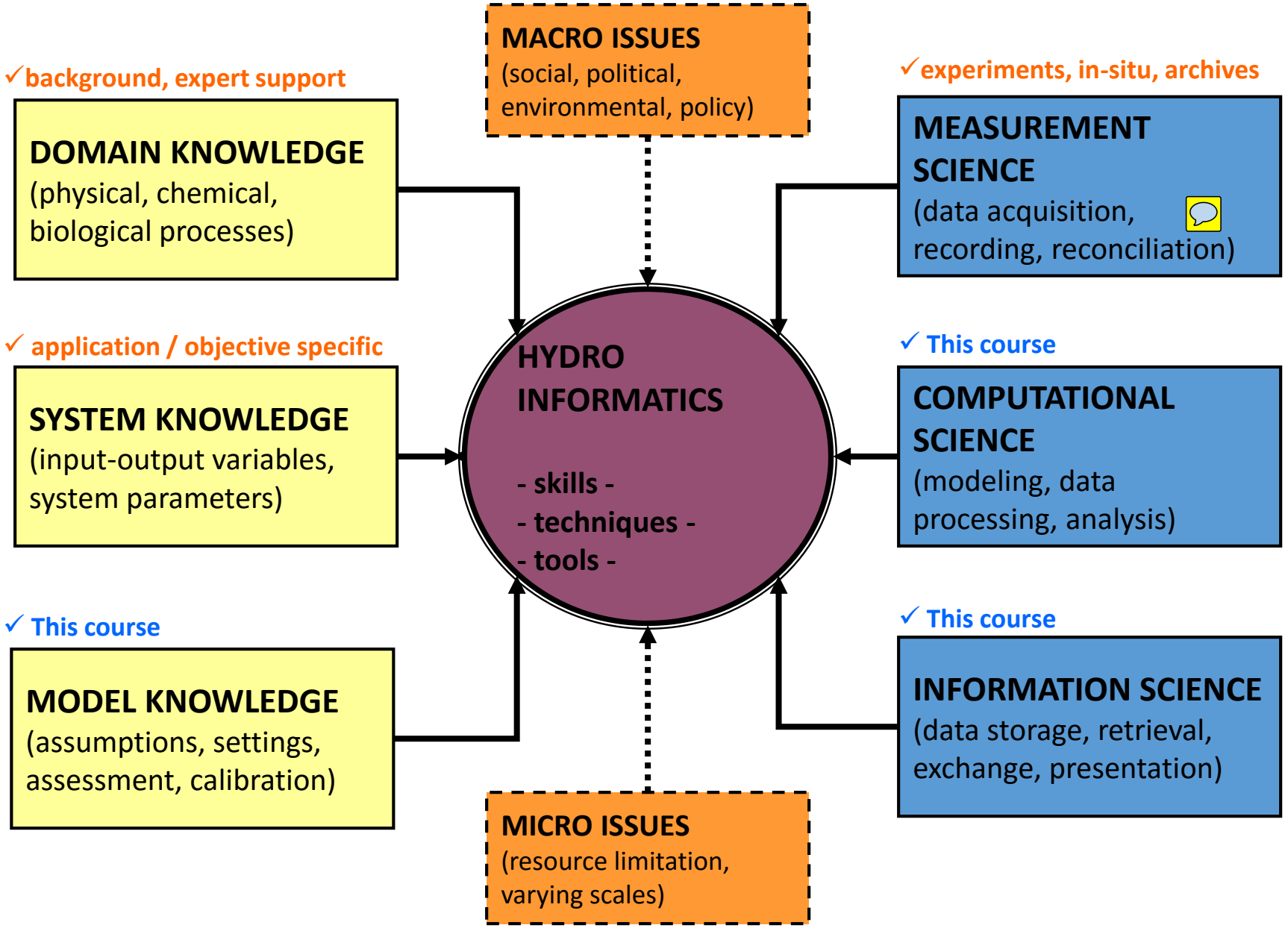
CE5310 - Hydroinformatics

Data Handling and Analysis

Refreshing Hydroinformatics – *what does it involve?*



Refreshing Hydroinformatics – *why is it important?*

• Hydrodynamic: ocean currents prediction for navigation, analysis of flow patterns in a canal for irrigation, water level dynamics in a reservoir, eddy simulations for sluice gate operation

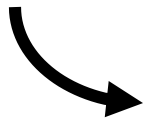
• Hydrological: modeling of transport processes for rainfall forecasting, run-off calculations using catchment modeling, operation multiple reservoirs and their interactions

• Multi-physics: air-water interaction for storm surge modeling, diffusion-hydrodynamic interaction for water quality modeling, energy-mass interaction for implementing mixing in lake

•Eco-hydraulic: effect of vegetation on flow patterns, wave energy attenuation studies for tsunami damage control, algae growth prediction for water quality monitoring

The common theme in all these applications

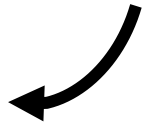
- Understand existing systems (OBSERVE/ANALYZE)
- Represent existing systems (MODEL)
- Improve existing systems (RETROFIT/OPTIMIZE)
- Regulate existing system (CONTROL)
- Develop new/better systems (DESIGN)



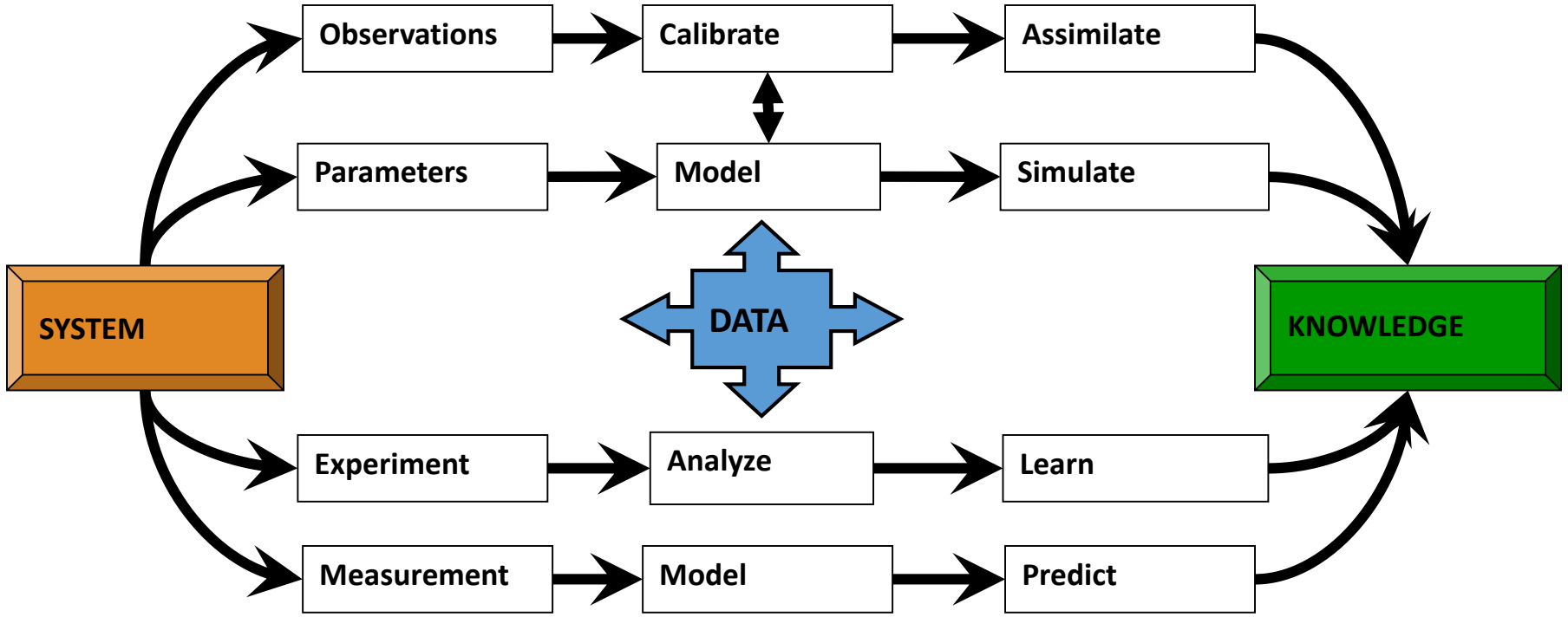
These issues depend and benefit largely from

Hydroinformatics

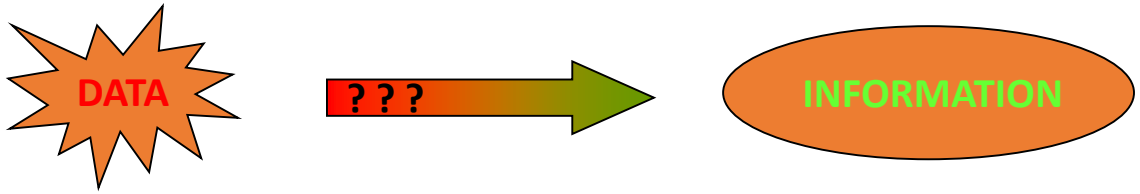
skills, techniques and tools



Refreshing Hydro*informatics* – *data, data, data everywhere!*



Big question: How to make sense out of all these various, huge, complex sets of DATA?



Data handling, processing, visualization, analysis

What is Data?

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, or feature
- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity, or instance

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Attribute Values

- Attribute values are numbers or symbols assigned to an attribute
- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
    - But properties of attribute values can be different
      - ID has no limit but age has a maximum and minimum value

Types of Attributes

- There are different types of attributes
  - Nominal
    - Examples: ID numbers, eye color, zip codes
  - Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
  - Interval
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - Ratio
    - Examples: temperature in Kelvin, length, time, counts

Discrete and Continuous Attributes

- Discrete Attribute
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes
- Continuous Attribute
  - Has real numbers as attribute values
  - Examples: 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.

Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male*, *female*}	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, {*good*, *better*, *best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

Data are Big

- Europe's Very Long Baseline Interferometry (VLBI) has 16 telescopes, each of which produces **1 Gigabit/second** of astronomical data over a 25-day observation session
  - storage and analysis a big problem
- AT&T handles billions of calls per day
  - so much data, it cannot be all stored -- analysis has to be done “on the fly”, on streaming data

Largest databases in 2003

- Commercial databases:
  - Winter Corp. 2003 Survey: France Telecom has largest decision-support DB, ~30TB; AT&T ~ 26 TB
- Web
  - Alexa internet archive: 7 years of data, 500 TB
  - Google searches 4+ Billion pages, many hundreds TB
  - IBM WebFountain, 160 TB (2003)
  - Internet Archive (www.archive.org),~ 300 TB

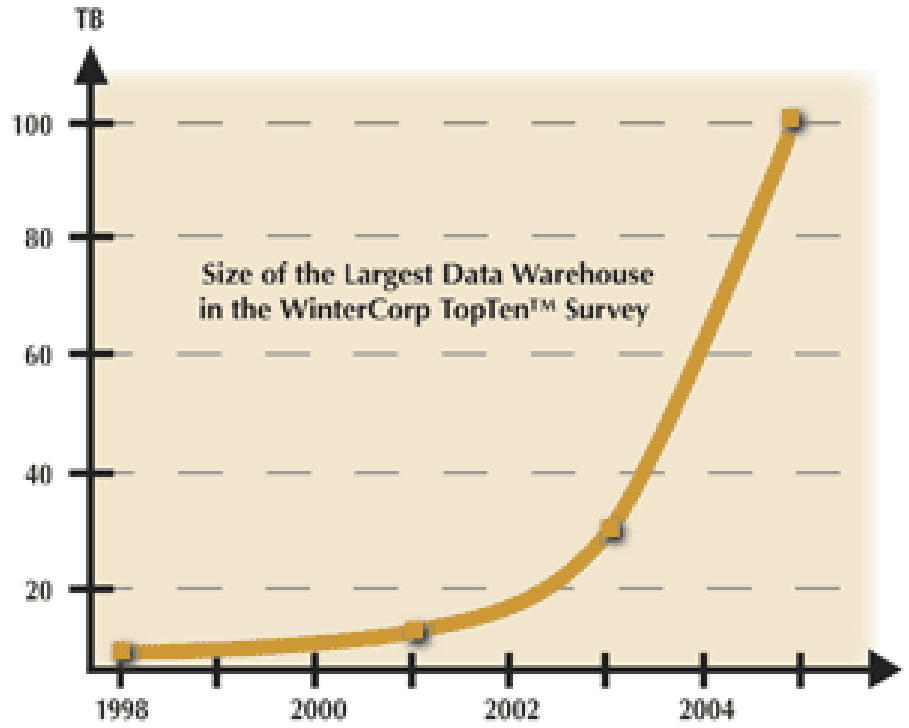
From terabytes to exabytes to …

- UC Berkeley 2003 estimate: 5 exabytes (5 million terabytes) of new data was created in 2002.

www.sims.berkeley.edu/research/projects/how-much-info-2003/

- US produces ~40% of new stored data worldwide
- 2006 estimate: 161 exabytes (IDC study)
  - www.usatoday.com/tech/news/2007-03-05-data_N.htm
- 2010 projection: 988 exabytes

Data Growth



In 2 years, the size of the largest database TRIPLED!

Data in Hydroinformatics – *basic classification*

- **based on processes :**

Hydrological – (discharge flows, ground water levels, runoff time, evaporation rate,...)

Meteorological – (wind, temperature, pressure, precipitation,...)

Oceanographic – (sea level, currents, tides, residual currents, bathymetry, salinity, SST,...)

Environmental – (water quality parameters, concentrations of solutes, pH, TOC,…)

Hydrodynamic – (fluid properties, velocity profiles, drag measurements, wave properties,...)

Biological – (microbial characteristics, ecosystem diversity, plant properties, 'omics' data,...)

Geographical – (latitude/longitude, soil/sediment properties, terrain usage, seismic data,...)

- **based on nature of data:**

Time series – (observations are recorded in the order of time Eg. - tide, wind, discharge,...)

Spatial series – (grids, altimeter track, ship mounted measurements)

Continuous – (can take any real value, Eg. - concentration, pH, temperature)

Discrete – (represents units or counts, Eg. - gate opening, valve positions, station index)

Parameters – (represent system characteristics Eg. - bathymetry, drag coeff., roughness)

Constants – (does not change universally Eg. - g, R, astronomical settings, …)

Data in Hydroinformatics – *basic classification (contd)*

- **based on data types:**

Numeric – all the instrument readings (most of the data in hydroinformatics)
Text - (tags, station names, gene sequence)
Date/Time - (reference dates, different formats)
Image- (microscopy, satellite image, scanners, RADAR imagery, maps, meteo/oceanographic data) ;
Audio/video (ecosystem recordings, sonar, PIV, tracer experiment recordings)

- **based on source:**

In-situ measurements – (ADV, tide gauges, UVM, ADCP, weather stations)
Altimetry - (radar, satellites)
Standards - (handbooks, admiralty charts)
Simulation results - (numerical model outputs, model forecasts)
Residuals - (model system mismatches, non-tidal component in sea level changes)
Processed data - (data obtained after processing the raw data, filtered/normalized/scaled,…)
Synthetic data - (approximations, missing value replacements)

- **based on data structure:**

Relational – (mySQL, O-ap, MS access, netCDF)
Hierarchical - (mat files, workbook)
Unstructured/Manual – (tables, parameters, settings, data in flat text files)

Data resources for Hydroinformatics applications – *data at fingertips*

- **In-situ Sea Level measurements –** UHSLC (http://uhslc.soest.hawaii.edu/); GLOSS (http://www.gloss-sealevel.org/) ;
- **Satellite Altimetry** (SSH/SST/Ocean Color data from TOPEX/Poseidon/Jason) **–** Univ CCAR altimetry center (http://argo.colorado.edu/~realtime/welcome/) ; NASA (http://podaac.jpl.nasa.gov/) ; CNES (http://www.aviso.oceanobs.com/) ; RADS (http://rads.tudelft.nl/rads/rads.shtml)
- **Standardized records** (admiralty charts, tide tables) – UHO - UK (http://www.ukho.gov.uk/Pages/Home.aspx) , Singapore – MPA (http://www.mpa.gov.sg/sites/global_navigation/publications/singapore_tide_tables.page)
- **Model (assimilated) meteorological data** (wind/pressure/temperature/precipitation) **:** NCEP (http://www.ncep.noaa.gov/) ; ECMWF (http://www.ecmwf.int/) ;
- **In-situ weather measurements -** WMO(http://www.wmo.int/pages/index_en.html);NOMAD (http://www.nomad3.ncep.noaa.gov/) METAR (http://weather.noaa.gov/weather/metar.shtml); NUS (http://courses.nus.edu.sg/course/geomr/front/fresearch/metstation/data01.htm)
- **Satellite imagery** (google earth – www.earth.google.com, wikimapia - www.wikimapia.org, ISRO Bhuvan - http://isrobhuvan.in/)
- **Coast Line boundaries –** NGDC Shoreline (http://www.ngdc.noaa.gov/mgg/shorelines/shorelines.html)
- **Bathymetry –** ETOPO (http://www.ngdc.noaa.gov/mgg/global/global.html) ; IHO (http://www.ngdc.noaa.gov/mgg/bathymetry/iho.html)
- **Geographical data –** GLOBE (http://www.ngdc.noaa.gov/mgg/topo/globe.html) ; LIDAR (http://www.ngdc.noaa.gov/mgg/bathymetry/lidar.html) ; GIS data / maps (http://data.geocomm.com/ ; http://www.maproom.psu.edu/dcw/ ; google map)
- **Hydrological data** – Links (http://www.nerc-wallingford.ac.uk/ih/devel/wmo/hhcdbs.html) ; GHRC (http://ghrc.msfc.nasa.gov/) ; Univ. Texas atlas (http://www.crwr.utexas.edu/gis/gishyd98/atlas/Atlas.htm)
- **ALSO can request local ministries/authorities/centers/research groups for relevant data – generally provided free for research use. Some universities also have repositories of data on specific hydroinformatics applications.**

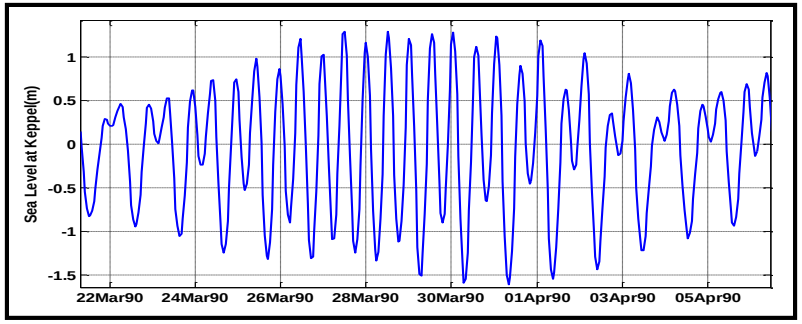
Sample data in Hydroinformatics applications

Sea Level data (UHSLC) - Malaysian peninsula

- **Oceanographic, Relational database**
- **Numerical + Text + date, continuous + discrete**

CODE	Name	Country	Lat	Long	Start	End
699	Keppel	Singapore	1.26	103.85	01.01.1988	15.05.2008
141	Ko Lak	Thailand	11.8	99.81	01.05.1985	12.04.2006
.	…	…	…	…	…	…

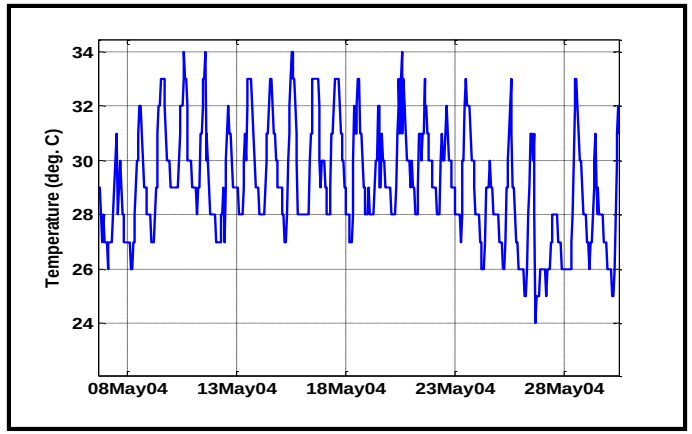
Date	Time	WL (m)
01.01.1988	01:00	1.225
01.01.1988	02:00	1.225
01.01.1988	03:00	9999
01.01.1988	04:00	1.225
…	…	..
15.05.2008	23:00	1.225



Sample data in Hydroinformatics applications

Temperature at Changi Airport

- meteorological, in-situ measurements
- Numerical, Time series, Hierarchical (mat file)

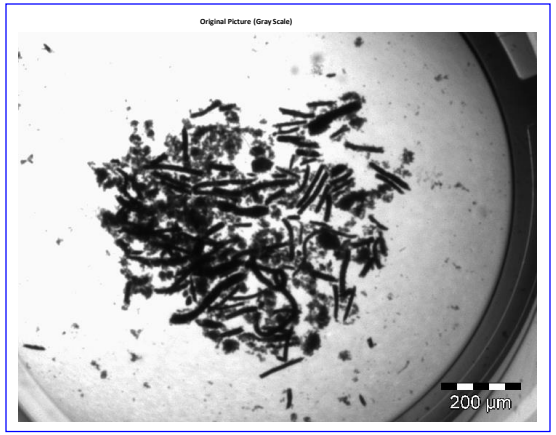


Singapore satellite image

- Geographical, image data

Microbial growth in canal water

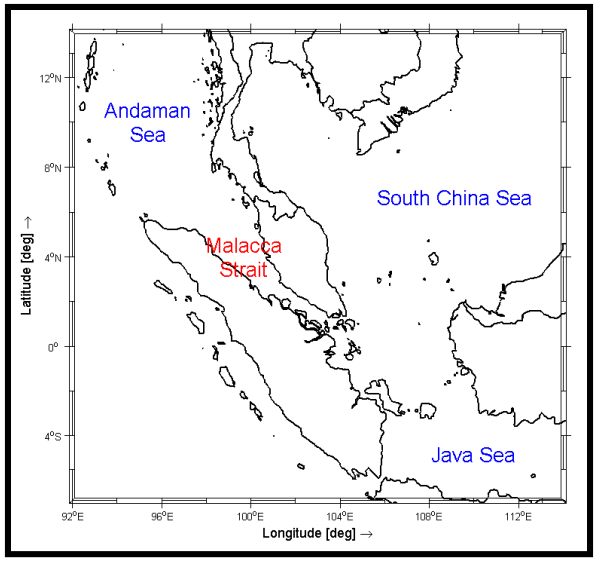
- Environmental/biological, Image data



Sample data in Hydroinformatics applications – *(continued)*

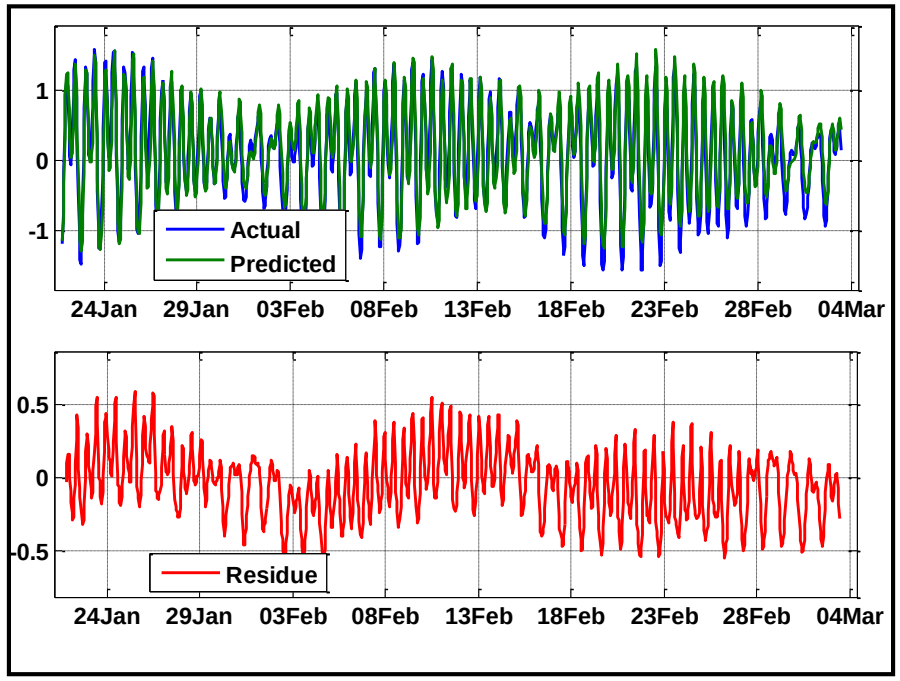
South East Asia regional map

- **Coastal boundaries, observations**
- **Numerical, spatial series**



Sea Level Anomaly at Keppel Harbor

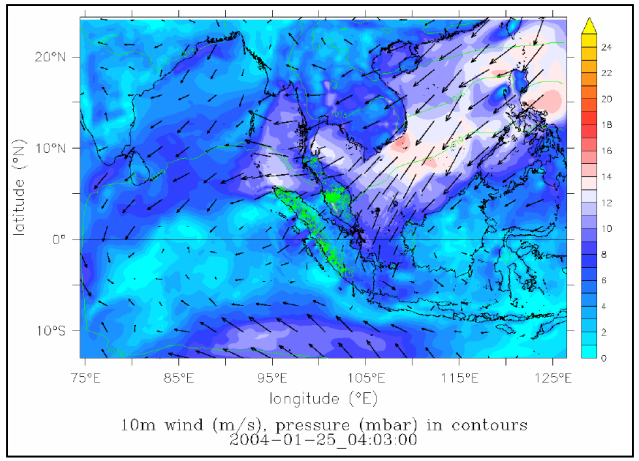
- **Oceanographic, model output/residuals**
- **Numerical + date/time, continuous**



Sample data in Hydroinformatics applications – *(continued)*

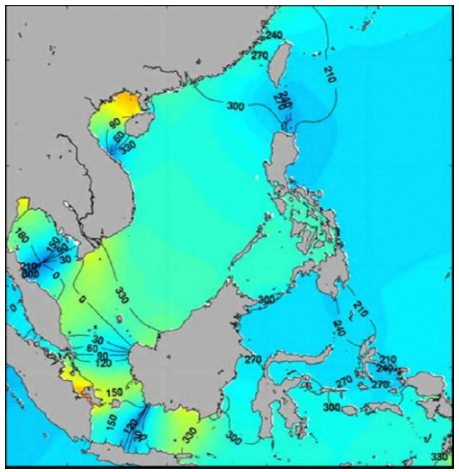
Regional Wind Data (METAR)

- Meteorological, Image data
- Spatially continuous data, directions



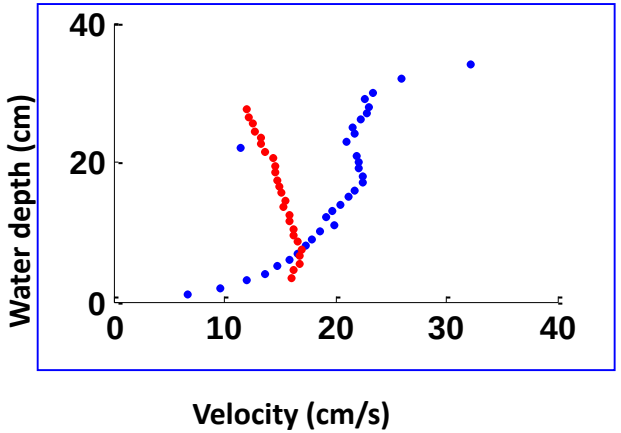
Diurnal tide in SEA

- oceanographic, Image data



Flume experiment for vegetated flow

- eco-hydraulics, numerical data
- multiple measurements



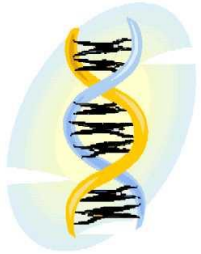
Which Level ? How much ? What not ?

molecule

A
G
C
T

X

gene



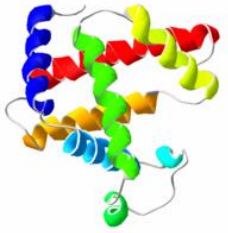
X

Amino Acid

A
C
D
.
Z

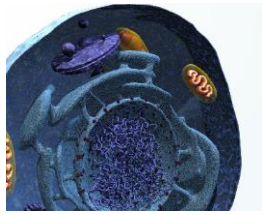
X

Protein



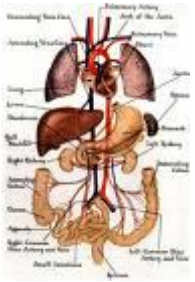
X

Cell



X

Organ



X

Universe

X

Nature

X

Eco System

X

Species

X

Organism

Huge, High Dimensional, Complex Data at each Level → ALL FOR TAKING

Data analysis issues

- Data management : storage, distribution, authenticity
- Data complexity : voluminous, noise, inconsistency, form
- Data Integration : varying time scales, spatial locations, randomness
- Data analysis and visualization
- Knowledge inference : system identification, modeling, network design
- Basic mathematical issues related to Data processing : tools
- Verification of analysis outcome – collaborative research
- Publishing

data analysis techniques/tools are inevitable

Data *activities* in Hydroinformatics – *basic tasks and important issues*

- **Data acquisition (collection):**
  - instrument selection (accuracy, precision and resolution, redundancy)
  - direct/remote/track measurements (calibration, noise, mounting)
  - temporal (sampling rate) and spatial (locations) resolution

- **Data storage:**
  - hard records (archives as tables/visuals) - soft records (digitizing, database design, multimedia data, file formats, nomenclature, data entry) - hardware (memory, backup)

- **Data handling:**
  - transfer (access, sharing, speed) - file format conversion (compatibility) - hosting (web)

- **Data reconciliation:**
  - alignment (time zone, consistency, nested grid data, different sources, averaging repeats)
  - missing value imputation (averaging, spatial/temporal interpolation, deletion)
  - noise removal (noise to signal ratio, cutoff limits, noise source, uncertainty analysis)
  - outlier detection (validation, distribution, range, statistical significance, outlier correction)
  - trend/drift correction (floating instruments, moving tracks, instrument bias, baseline shift)
  - type conversion (image to numeric data, non-dimensionalizing, continuous to discrete)

- **Data visualization:**
  - plotting/charting - multimedia (animation/movies) - web based (DSS, hosting)

Data *activities* in Hydroinformatics – *basic tasks and important issues (continued)*

- **Data Mining:**
  - querying (database query, summary, row/column filtering, sorting, event selection)
  - pattern recognition (supervised and unsupervised, scales (time/space), data availability)
  - trend analysis (qualitative, quantitative, seasonal, event specific, periodic/non-stationery)
  - feature extraction (influential variables, sensitive parameter, dominant location/period)

- **Data processing:**
  - normalization (bringing different variables to similar scale, variable/global)
  - scaling (mean centering, offsetting,, log transformation)
  - filtering (frequency based, moving averages, Godin tide filter, wave-let transforms)
  - projection (transforms, variance based, eigen/singular values, scores and weights)

- **Data analysis:**
  - statistical (descriptive, comparative, hypothesis testing)
  - correlation (nature of interaction, scale free measures, dependency (auto/cross correlation)
  - regression (linear/non-linear, least squares, residual analysis)
  - uncertainty analysis (error estimation, model sensitivity, confidence intervals)
  - performance assessment (model assessment, indices, system performance)
  - calibration (model parameter tuning, estimation methods, optimization schemes)
  - assimilation (model order reduction, error correction, state/parameter updating)
  - modeling (data driven, black box/statistical/evolutionary methods, validation, over-fitting)
  - frequency response analysis (tidal analysis, periodicity, disturbance response, stability)

Tools to carry out Hydroinformatics tasks – *major support*

Software – (search google / wikipedia for more details on these and download links for few)

- **Editors** – note pad, word pad, image editors, audio/video editors, …
- **Spreadsheets** – MS Excel, MiniTab, Lotus 123, Star Office, …
- **Databases** - mySQL, MS Access, O-ap, …
- **Statistical** - R, SPSS, SAS, STAT, MS Excel, MiniTab, …
- **Visualization** - gnuPLOT, Pajek, google APIs, ArcGIS, D3D quickplot, datPAV (SDWA), FEWS, ..
- **Specialized** - classification (SIMCA / WEKA / SVMlight), modeling (Neurosolutions for ANN, GP tools), optimization (GAMS), tidal analysis (D3D TIDE, TRIANA)

Desired Features

- **facility to handle/process/analyze/visualize data**
- **interactive and integrated environment**
- **fast and accurate**
- **easy implementation /compatibility / generalizibility / repeatability**

- **MATLAB – very generic to high end specialized applications (expensive/available in NUS)**
- **R/Octave – free, can do many of MATLAB tasks but less powerful toolboxes/visualization.**

R package: http://www.r-project.org/ ; **Octave:** http://www.gnu.org/software/octave/index.html

**We shall see sample data, study a few techniques,
use some tools and explore interesting HI case studies**