A Spatial Agent-based Model and A Multi-dimensional Data Warehouse for Malaria Research

S. M. Niaz Arifin
Advisor: Dr. Gregory R. Madey
Department of Computer Science and Engineering
University of Notre Dame

Ph.D. Defense
June 18, 2013
Motivation: Malaria

• one of the oldest and deadliest infectious diseases in humans
• one million deaths per year\(^1\)
• a child dies every 30 seconds in Africa\(^2\)

\(^1\) World Health Organization (WHO)
\(^2\) UNICEF
Human Malaria

• by female mosquitoes of genus *Anopheles*

• *Anopheles gambiae* complex
  • most important malaria vector in sub-Saharan Africa

• of the most dangerous malaria parasite, *Plasmodium falciparum*
Overview

Part 1

Agent-Based Model (ABM)
Overview

Part 1

Agent-Based Model (ABM)

Part 2

Data Warehouse (DW)

Today’s focus
Part 1

Computational Biology

Agent-Based Model

Spatial Agent-Based Model

Bioinformatics

Malaria Epidemiology

Vector Control Interventions

Geographic Information System

Global Health
Part 2

Computational Biology

Agent-Based Model

Spatial Agent-Based Model

Vector Control Interventions

Geographic Information System

Bioinformatics

Malaria Epidemiology

Dimensional Modeling

Data Warehouse

Global Health
Part 1 (ABM) Summary

- ABM of *An. gambiae*
  - from the biological core model
- verification & validation
- a spatial extension
- a landscape generator tool
- vector control interventions
- integration with a geographic information system (GIS)
Part 2: A Data Warehouse for VECNet CI

- **VECNet**: Vector Ecology and Control Network
  - a consortium of institutions
  - analyze malaria transmission and its reduction

- **CI**: cyberinfrastructure
  - an online collaboration, storage, and compute engine
  - enables users to:
    - conduct research
    - share results from models
    - access shared data and documents
VECNet Partners

• James Cook University (JCU), Australia
• Swiss Tropical and Public Health Institute
• University of Pittsburgh
• Pittsburgh Supercomputing Center
• University of Oxford
• R. Farlow Consulting, LLC
• Intellectual Ventures Lab
• University of Notre Dame
• Bill & Melinda Gates Foundation
VECNet User Groups

• *researchers*
  • query the DW/DL or run simulations
• *modelers*
• *malaria control managers*
  • from ministries of health
• *product developers*
  • novel interventions products
• *funding agencies and policy makers*
  • decision-making on investment strategies
Motivation

• biological data storage systems
Motivation

- biological data storage systems
- **BioMart** ([www.biomart.org](http://www.biomart.org))
  - allows biologists to create complex datasets through a web interface
- **Chado** ([www.gmod.org](http://www.gmod.org))
  - an integrated database to manage biological knowledge
  - genome sequences, gene ontologies, etc.
- **BioWarehouse** ([www.biowarehouse.org](http://www.biowarehouse.org))
  - an open-source software environment
  - integrates biological databases for enzymes, gene ontologies, etc.
Research Questions

• we didn’t have good examples in the literature for solving the exact set of problems that VECNet poses
Research Questions

• we didn’t have good examples in the literature for solving the *exact* set of problems VECNet poses

• for example, how do we:

  1. integrate heterogeneous data sources from various domains (mosquito biology, epidemiology, weather, etc.)?
Research Questions

• we didn’t have good examples in the literature for solving the *exact* set of problems VECNet poses

• for example, how do we

  1. integrate heterogeneous data sources from various domains (mosquito biology, epidemiology, weather, etc.)?

  2. make the data understandable for a diverse group of users?
Research Questions

• we didn’t have good examples in the literature for solving the *exact* set of problems VECNet poses

• for example, how do we

  1. integrate heterogeneous data sources from various domains (mosquito biology, epidemiology, weather, etc.)?

  2. make the data understandable for a diverse group of users?

  3. make best use of the data by transforming it into meaningful aggregates?
Research Questions

• we didn’t have good examples in the literature for solving the *exact* set of problems VECNet poses

• for example, how do we

  1. integrate heterogeneous data sources from various domains (mosquito biology, epidemiology, weather, etc.)?

  2. make the data understandable for a diverse group of users?

  3. make best use of the data by transforming it into meaningful aggregates?

  4. link appropriate datasets with model-specific inputs-outputs?
Our Approach

• we borrowed ideas and knowledge from
  • business intelligence,
  • decision support systems, and
  • dimensional modeling
Our Approach

- we borrowed ideas and knowledge from
  - business intelligence,
  - decision support systems, and
  - dimensional modeling
- to build a data warehouse for malaria-related data:

  the VECNet Data Warehouse
Our Approach

• the pilot DW I designed and developed influenced production activities in a very large research grant on VECNet
Our Approach

• the pilot DW I designed and developed influenced production activities in a very large research grant on VECNet

• we designed a solution which:
  • is new and novel
Our Approach

- the pilot DW I designed and developed influenced production activities in a very large research grant on VECNet
- we designed a solution which:
  - is new and novel
  - can be generalized to other biological domains of knowledge
Our Approach

• the pilot DW I designed and developed influenced production activities in a very large research grant on VECNet

• we designed a solution which:
  • is new and novel
  • can be generalized to other biological domains of knowledge
  • can serve as a *proof-of-concept* for future biological data storage systems
DW within VECNet CI

**Back End Services**
- Servers
- Databases: relational, geospatial, GIS, etc.
- Data storage systems: digital asset management software, etc.
- Compute nodes: High-performance computing (HPC) resources, Cloud-computing services
- Data analysis

**Integrated Modeling Platform**
- Simulations
- Analysis

**Data Warehouse**
- Data marts
- Facts and dimensions
- Lookup tables

**Digital Library**
- Documents
- Published studies

**Front End Services**
- Graphical user interfaces (GUIs)
- Data visualization
- Data downloads
- Search (query): basic, slice and dice, aggregation
- Job submission and retrieval
**Historical data**
- ranging over several years
- to analyze trends of malaria data
- *example*: recent data on existing interventions to support introduction of new interventions in a location
**Predictive/Synthetic data**
- outputs of simulation models
- scenarios, campaigns
- can help to:
  - discover new insights
  - ask *what-if questions*
- example models: EMOD, OpenMalaria
Data Warehouse (DW)

- a central repository of data
- integrates data from multiple disparate sources
- supports management decisions
- also known as:
  - *on-line analytical processing (OLAP)* systems
- collection of data:
  - subject-oriented
  - integrated
  - nonvolatile
  - time-varying
Data Warehouse (DW)

• a central repository of data
• integrates data from multiple disparate sources
• collection of data: subject-oriented data:
  • subject-oriented
  • integrated
  • nonvolatile
  • time-varying
  • from multiple related subjects or domains
  • examples:
    • vector control interventions,
    • antimalarial drugs, etc.
Data Warehouse (DW)

• a central repository of data
• integrates data from multiple disparate sources

collection of data: **integrated data:**
• subject-oriented
• **integrated**
• nonvolatile
• time-varying

• from external operational source systems (OSSs)

• example OSSs:
  • World Malaria Report Data,
  • Pacific Rainfall Database, etc.
Data Warehouse (DW)

- a central repository of data
- integrates data from multiple disparate sources
- collection of data: **nonvolatile data:**
  - subject-oriented
  - integrated
  - **nonvolatile**
  - time-varying
- persistent
- retained for long-term analysis and aggregation purposes
Data Warehouse (DW)

- a central repository of data
- integrates data from multiple disparate sources

Collection of data:
- subject-oriented
- integrated
- nonvolatile

*time-varying (temporal) data:*
- records temporal evolution of data
- for a long period of time
- examples: data on
  - malaria incidence,
  - malaria prevalence, etc.
  ranging over several decades
Operational Source Systems (OSSs)

• traditional databases
• operational or transactional databases
• online transaction processing (OLTP) systems
• examples:
  • World Malaria Report
  • Pacific Rainfall Database
## OSS vs. DW

<table>
<thead>
<tr>
<th>Feature</th>
<th>OSS</th>
<th>DW</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Source of data</strong></td>
<td>from original sources</td>
<td>consolidated from OSSs</td>
</tr>
<tr>
<td><strong>Purpose</strong></td>
<td>day-to-day operations</td>
<td>decision support</td>
</tr>
<tr>
<td><strong>Type of data</strong></td>
<td>snapshot</td>
<td>historical, predictive, etc.</td>
</tr>
<tr>
<td><strong>Queries</strong></td>
<td>involves few records</td>
<td>involves aggregations of many records</td>
</tr>
<tr>
<td><strong>Design</strong></td>
<td>highly normalized</td>
<td>de-normalized</td>
</tr>
<tr>
<td><strong>Model</strong></td>
<td>entity-relationship (ER)</td>
<td>dimensional</td>
</tr>
<tr>
<td><strong>Users deal with</strong></td>
<td>one record at a time</td>
<td>almost never one record at a time</td>
</tr>
</tbody>
</table>
Dimensional Modeling (DM)

• a design technique
• used in DW design
• different from entity-relationship (ER) modeling
Dimensional Modeling (DM)

• a design technique
• used in DW design
• different from entity-relationship (ER) modeling

• views data as **facts** linked to several **dimensions**
Connections

Business Intelligence

Dimensional Modeling

Data Warehouse

Decision Support Systems

uses

roots

roots
Connections

- **Business Intelligence**
  - Allows an organization to make better business decisions

- **Decision Support Systems**
  - Computer-based information systems to support decision-making activities

- **Dimensional Modeling**
  - Uses

- **Data Warehouse**
  - Roots

---

**Business Intelligence**: A computer-based system that supports decision making activities by providing an organization with the ability to make better business decisions.
Connections

- **Business Intelligence**
  - ability of an organization to leverage its information assets
  - for making better business decisions

- **Dimensional Modeling**
  - uses

- **Data Warehouse**
  - roots

- **Decision Support Systems**
  - roots
Connections

- **Business Intelligence**
  - roots
  - uses
  - roots

- **Dimensional Modeling**
  - uses

- **Data Warehouse**
  - uses

- **Decision Support Systems**
  - uses

  - computer-based information systems
  - support organizational decision-making activities
DW Vocabulary
DW Vocabulary

Fact
• a quantitative measure
• is stored in a **fact table**
• usually numeric and additive

Dimension
• a descriptive attribute
• provides context to the facts
• is stored in a **dimension table**
• primary source of query constraints
• can be **hierarchical** or **non-hierarchical**
DW Vocabulary

Fact
- a quantitative measure
- is stored in a *fact table*
- usually

Dimension
- a descriptive attribute
- provides context to the facts
- is stored in a *dimension table*
- primary source of query constraints
- can be *hierarchical* or *non-hierarchical*
DW Vocabulary

**Location Dimension**
- Continent
- Country
- Province/State
- Region

**Date Dimension**
- Year
- Month
- Day

**Vector Control Intervention Facts**
- Coverage
  - Number sold or delivered
  - Number of people protected
  - Duration
  - ...

**Intervention Dimension**
- Name
- Abbreviation
- Type
- Method
- Description
- Synonyms

**Species Dimension**
- Species name
- Form
- Vector status
DW Vocabulary

Dimension

• a descriptive attribute that provides context to the facts
• describes who, what, when, where, why, and how about facts
• primary source of query constraints, groupings, and report labels
• can be hierarchical and non-hierarchical
• examples:
  • hierarchical: date with hierarchy year-month-day
  • non-hierarchical: land use, chemical class
Fact table
• a primary table to store the facts
• each fact has a composite key
• foreign keys from the dimension tables
• examples: Household Surveys

Dimension table
• a table with descriptive (textual) attribute columns
• has a single-part primary key
• integral companion to a fact table
• entry point into the fact table
• examples: location, date
Data mart

- a logical and physical subset of the DW
- data from a single process of interest
- may contain one or more fact tables
- example: Operational Coverages
Data mart

- a logical and physical subset of the DW
- data from a single process of interest
- may contain one or more fact tables

example:
Operational Coverages
**DW Vocabulary**

**Lookup table**
- reference tables of static data
- related to dictionaries
- used to translate obscure terms
- saves space, improves flexibility
- example: Species

<table>
<thead>
<tr>
<th>Species Lookup</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Human blood index</td>
<td>12 - 69</td>
</tr>
<tr>
<td>Flight range</td>
<td>0.9 - 6.0</td>
</tr>
<tr>
<td>Sugar sources</td>
<td>Plants, fruits and flowers</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
DW Vocabulary

Lookup table

- reference tables of static data
- related to dictionaries
- used to translate obscure terms
- saves space, improves flexibility
- example: Species

Dictionary

- contains definitions of terms
- explanatory descriptions
- example: Species Bionomics

<table>
<thead>
<tr>
<th>Species Lookup</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Human blood index</td>
<td>12 - 69</td>
</tr>
<tr>
<td>Flight range</td>
<td>0.9 - 6.0</td>
</tr>
<tr>
<td>Sugar sources</td>
<td>Plants, fruits and flowers</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Species Bionomics Dictionary</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Term</td>
<td>...</td>
</tr>
<tr>
<td>Description</td>
<td>...</td>
</tr>
<tr>
<td>Reference</td>
<td>...</td>
</tr>
<tr>
<td>Units</td>
<td>...</td>
</tr>
<tr>
<td>Methods</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
VECNet DW Ontology

Data Warehouse

- has Data Mart
- has Data Cube
- has Fact Table
- has Dimension Table
- has Record

Data Mart

- has Data Cube
- has Fact Table
- has Dimension Table

Data Cube

- has Fact Table
- has Definition

Fact Table

- has Definition
- has Record

Definition

- is a Table (Relational)
- is a Dictionary

Table (Relational)

- is a Dictionary
- has Definition

Dictionary

- has Lookup Table
- has Table (Relational)

Lookup Table

- is a Dictionary
- has Table (Relational)

Table (Relational)
DW Schemas

Star schema

• simplest type of DM schema
• resemblance to a star:
  • *fact table*: at center
  • *dimension tables*: the star’s points
DW Schemas

Constellation schema (galaxy schema)

- collection of star schemas
- ties together:
  - fact tables,
  - dimension tables, and
  - lookup tables
- uses *shared* dimensions
Constellation Schema

Star Schema 1

Star Schema 2

Star Schema 3

Dimension 1

Fact Table 1

Dimension 2

Fact Table 2

Dimension 3

Lookup Table 1

Dimension 4

Dimension 5

Dimension 6

Dimension 7

Dimension 8
DW Operations
DW Operations

- Roll Up and Roll Down
- Slice and Dice
Roll Up and Roll Down

- along a specific dimension
- from the most summarized (up)
- to the most detailed (down)
Roll Up and Roll Down

<table>
<thead>
<tr>
<th>Year</th>
<th>Continent</th>
<th>Country</th>
<th>An. arabiensis</th>
<th>An. gambiae</th>
<th>An. dirus</th>
<th>An. darlingi</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>Africa</td>
<td>Angola</td>
<td>75</td>
<td>89</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2010</td>
<td>Africa</td>
<td>Benin</td>
<td>41</td>
<td>55</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2010</td>
<td>Africa</td>
<td>Kenya</td>
<td>139</td>
<td>257</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2010</td>
<td>Africa</td>
<td>Nigeria</td>
<td>121</td>
<td>189</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2010</td>
<td>Asia</td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>1055</td>
<td>0</td>
</tr>
<tr>
<td>2010</td>
<td>South America</td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1613</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2011</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2012</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

hypothesised data of mosquito abundances
### Roll Up and Roll Down

<table>
<thead>
<tr>
<th>Year</th>
<th>Continent</th>
<th>Country</th>
<th>An. arabiensis</th>
<th>An. gambiae</th>
<th>An. dirus</th>
<th>An. darlingi</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>Africa</td>
<td>Angola</td>
<td>75</td>
<td>89</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2010</td>
<td>Africa</td>
<td>Benin</td>
<td>41</td>
<td>55</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2010</td>
<td>Africa</td>
<td>Kenya</td>
<td>139</td>
<td>257</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2010</td>
<td>Africa</td>
<td>Nigeria</td>
<td>121</td>
<td>189</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2010</td>
<td>Asia</td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>1055</td>
<td>0</td>
</tr>
<tr>
<td>2010</td>
<td>South America</td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1613</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2011</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2012</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- **data cube**, derived from the data, shows mosquito abundance **facts**
- **three dimensions**: 
  - Year
  - Continent
  - Species
three dimensions:
• Year
• Continent
• Species
Roll Up and Roll Down

three dimensions:
- Year
- Continent
- Species
Roll Up and Roll Down

along the hierarchical dimension *Location*, abundances are rolled down from **Continent** (Africa, topmost row) level to **Country** level

three dimensions:
- Year
- Continent
- Species
Slice and Dice

• **Slice:**
  - selects a *single* value for a *single* dimension
  - creates a new cube with one fewer dimension

• **Dice:**
  - produces a sub-cube
  - selects values of *multiple* dimensions
### Slice and Dice

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>Site</th>
<th>Abundance</th>
</tr>
</thead>
<tbody>
<tr>
<td>19 May 1970</td>
<td>7PM</td>
<td>Ifakara</td>
<td>20</td>
</tr>
<tr>
<td>20 May 1970</td>
<td>7PM</td>
<td>Ifakara</td>
<td>13</td>
</tr>
<tr>
<td>21 May 1970</td>
<td>7PM</td>
<td>Ifakara</td>
<td>17</td>
</tr>
<tr>
<td>17 June 1970</td>
<td>7PM</td>
<td>Ifakara</td>
<td>19</td>
</tr>
<tr>
<td>19 June 1970</td>
<td>7PM</td>
<td>Ifakara</td>
<td>22</td>
</tr>
<tr>
<td>19 May 1970</td>
<td>7PM</td>
<td>Garki</td>
<td>11</td>
</tr>
<tr>
<td>20 May 1970</td>
<td>7PM</td>
<td>Garki</td>
<td>15</td>
</tr>
<tr>
<td>21 May 1970</td>
<td>7PM</td>
<td>Garki</td>
<td>25</td>
</tr>
<tr>
<td>17 June 1970</td>
<td>7PM</td>
<td>Garki</td>
<td>13</td>
</tr>
<tr>
<td>19 June 1970</td>
<td>7PM</td>
<td>Garki</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19 May 1970</td>
<td>10PM</td>
<td>Asembo</td>
<td>10</td>
</tr>
<tr>
<td>20 May 1970</td>
<td>10PM</td>
<td>Asembo</td>
<td>17</td>
</tr>
<tr>
<td>21 May 1970</td>
<td>10PM</td>
<td>Asembo</td>
<td>13</td>
</tr>
<tr>
<td>17 June 1970</td>
<td>10PM</td>
<td>Asembo</td>
<td>24</td>
</tr>
<tr>
<td>19 June 1970</td>
<td>10PM</td>
<td>Asembo</td>
<td>19</td>
</tr>
</tbody>
</table>

Hypothetical data of mosquito abundances
### Slice and Dice

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>Site</th>
<th>Abundance</th>
</tr>
</thead>
<tbody>
<tr>
<td>19 May 1970</td>
<td>7PM</td>
<td>Ifakara</td>
<td>20</td>
</tr>
<tr>
<td>20 May 1970</td>
<td>7PM</td>
<td>Ifakara</td>
<td>13</td>
</tr>
<tr>
<td>21 May 1970</td>
<td>7PM</td>
<td>Ifakara</td>
<td>17</td>
</tr>
<tr>
<td>17 June 1970</td>
<td>7PM</td>
<td>Ifakara</td>
<td>19</td>
</tr>
<tr>
<td>19 June 1970</td>
<td>7PM</td>
<td>Ifakara</td>
<td>22</td>
</tr>
<tr>
<td>19 May 1970</td>
<td>7PM</td>
<td>Garki</td>
<td>11</td>
</tr>
<tr>
<td>20 May 1970</td>
<td>7PM</td>
<td>Garki</td>
<td>15</td>
</tr>
<tr>
<td>21 May 1970</td>
<td>7PM</td>
<td>Garki</td>
<td>25</td>
</tr>
<tr>
<td>17 June 1970</td>
<td>7PM</td>
<td>Garki</td>
<td>13</td>
</tr>
<tr>
<td>19 June 1970</td>
<td>7PM</td>
<td>Garki</td>
<td>16</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>19 May 1970</td>
<td>10PM</td>
<td>Asembo</td>
<td>10</td>
</tr>
<tr>
<td>20 May 1970</td>
<td>10PM</td>
<td>Asembo</td>
<td>17</td>
</tr>
<tr>
<td>21 May 1970</td>
<td>10PM</td>
<td>Asembo</td>
<td>13</td>
</tr>
<tr>
<td>17 June 1970</td>
<td>10PM</td>
<td>Asembo</td>
<td>24</td>
</tr>
<tr>
<td>19 June 1970</td>
<td>10PM</td>
<td>Asembo</td>
<td>19</td>
</tr>
</tbody>
</table>

#### Data Cube

Data cube, derived from the data

- **Time**
- **Site**
- **Date**

The data cube is three-dimensional, representing the data from different dates, sites, and times.
abundances of all sites, for all dates, **at 7PM**
are sliced out
abundances for sites Ifakara and Garki, for all dates, at 8PM and 10PM are diced out
VECNet
Operational Source Systems (OSSs)
Household Surveys

- **source**: World Malaria Report (WMR)
  - World Health Organization (WHO)

- mosquito nets ownership and usage data
  - single data mart, single cube

### Household Surveys of Mosquito Nets Ownership and Usage

<table>
<thead>
<tr>
<th>Location</th>
<th>Date</th>
<th>Source</th>
<th>Subgroup</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Household with ≥ 1 any net</td>
<td>% Household with ≥ 1 ever-treated net</td>
<td>% Household with ≥ 1 ITN</td>
<td>% Total population who slept under an ITN</td>
</tr>
<tr>
<td>% Children &lt; 5 years who slept under any net</td>
<td>% Children &lt; 5 years who slept under ever-treated net</td>
<td>% Children &lt; 5 years who slept under an ITN</td>
<td>% Pregnant women who slept under any net</td>
</tr>
<tr>
<td>% Pregnant women who slept under ever-treated net</td>
<td>% Pregnant women who slept under an ITN</td>
<td>annex 6a</td>
<td></td>
</tr>
</tbody>
</table>
Operational Coverages

- **source**: World Malaria Report (WMR)
- single data mart, multiple cubes

**ITNs Operational Coverage**
- Number of LLINs sold or delivered
- Number of ITNs sold or delivered
- Number of ITNs & LLINs sold or delivered
- % ITN coverage

**IRS Operational Coverage**
- Number of people protected by IRS
- % IRS coverage

**ACT Operational Coverage**
- Any 1st line treatment courses delivered (including ACT)
- ACT treatment courses delivered
- % Any antimalarial coverage total
- % ACT coverage total

**ACT**: Artemisinin-Combination Therapy; possess the most rapid action of most current drugs against *Plasmodium falciparum* malaria
Rainfall Data from PacRain

- **source**: Pacific Rainfall Database

PacRain Area Map
Rainfall Data from PacRain

- *source*: Pacific Rainfall Database

<table>
<thead>
<tr>
<th>Site Data Source Dimension</th>
<th>PacRain Rainfall Facts</th>
<th>Beginning of observation date Dimension</th>
<th>Last modification date Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site Dimension</td>
<td>Length of Observation (in hours)</td>
<td>Rainfall amount (in mm)</td>
<td></td>
</tr>
</tbody>
</table>
Georeferenced EIR Data

- **source**: Mapping Malaria Risk in Africa (MARA) collaboration
- Entomological Inoculation Rate (EIR)

<table>
<thead>
<tr>
<th>Location</th>
<th>Seasonality Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Interval</td>
<td>Sporozoite Index Method</td>
</tr>
<tr>
<td>Georeference</td>
<td>Biting Rate Method</td>
</tr>
<tr>
<td>Land Use</td>
<td>Georeference Source</td>
</tr>
</tbody>
</table>

**Georeferenced EIR Facts**

- Sporozoite index
- Biting Rate
- EIR
- Seasonality
- Relative EIR transmission by *Anopheles gambiae s.l.* (percentage)
- Relative EIR transmission by *Anopheles funestus* (percentage)
- Relative EIR transmission by other important local vectors (percentage)
Georeferenced EIR Data

- **Date Dimension**
  - Start Date
  - End Date

- **Time Interval Dimension**
  - Time Interval Key (PK)
    - Start Month
    - Start Year
    - End Month
    - End Year

- **Georeferenced EIR Facts**
  - Sporozoite index
  - Biting Rate
  - EIR
  - Seasonality
  - Relative EIR transmission by *Anopheles gambiae* s.l. (percentage)
  - Relative EIR transmission by *Anopheles funestus* (percentage)
  - Relative EIR transmission by other important local vectors (percentage)

- **Location**
- **Time Interval**
- **Georeference**
- **Land Use**
- **Citation**
- **Seasonality Meaning**
- **Sporozoite Index Method**
- **Biting Rate Method**
- **Georeference Source**
Weather Data from GSOD

- **source**: Global Surface Summary of Day (GSOD)
- over 9000 weather stations

<table>
<thead>
<tr>
<th>GSOD Weather</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean temperature</td>
</tr>
<tr>
<td>Mean temperature observation count</td>
</tr>
<tr>
<td>Mean dew point</td>
</tr>
<tr>
<td>Mean dew point observation count</td>
</tr>
<tr>
<td>Mean sea level pressure</td>
</tr>
<tr>
<td>Mean sea level pressure observation count</td>
</tr>
<tr>
<td>Mean station pressure</td>
</tr>
<tr>
<td>Mean station pressure observation count</td>
</tr>
<tr>
<td>Mean visibility</td>
</tr>
<tr>
<td>Mean visibility observation count</td>
</tr>
<tr>
<td>Mean wind speed</td>
</tr>
<tr>
<td>Mean wind speed observation count</td>
</tr>
<tr>
<td>Maximum sustained wind speed</td>
</tr>
<tr>
<td>Maximum wind gust</td>
</tr>
<tr>
<td>Maximum temperature</td>
</tr>
<tr>
<td>Minimum temperature</td>
</tr>
<tr>
<td>Total precipitation</td>
</tr>
<tr>
<td>Snow depth</td>
</tr>
<tr>
<td>Indicators</td>
</tr>
</tbody>
</table>

**Date**

**Dimension**

**GSOD Station**

**Dimension**
Insecticides

- **source**: R. Farlow Consulting, LLC
- common insecticides used in vector control
DW Bus Matrix

• tabular tool
• all data marts at a glance
• rows: data marts
• columns: dimensions
• shared, uniform architecture
• conformed dimensions and facts
# DW Bus Matrix

<table>
<thead>
<tr>
<th>Data Mart</th>
<th>Location</th>
<th>Date</th>
<th>Source</th>
<th>Subgroup</th>
<th>Time Interval</th>
<th>Georeference</th>
<th>Land Use</th>
<th>Manufacturer</th>
<th>Intervention</th>
<th>GSOD Station</th>
<th>Site</th>
<th>Site Data Source</th>
<th>Seasonality Meaning</th>
<th>Sporozoite Index Method</th>
<th>Biting Rate Method</th>
<th>Georeference Source</th>
<th>Citation</th>
<th>Active Ingredient</th>
<th>Chemical Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Household Surveys</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ITNs Operational Coverage</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IRS Operational Coverage</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACT Operational Coverage</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PacRain Rainfall</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Georeferenced EIR</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSOD Weather</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insecticides</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Rows are denoted as:

- -----
- -----:
- -----:
- -----:

Columns are denoted as:

- Species Lookup
- Location Dimension
- Intervention Dimension
- Entomological Endpoints Lookup
- Entomological Endpoints Phases
- Species Parameters Dictionary
- Entomological Endpoints Dictionary

Species:

Location:

Entomological Endpoints:

Intervention:

Species Parameters:

Entomological Endpoints Phases:

Dictionary:

Constellation Schema for Lookup Tables
VECNet DW
Demonstration
DW Demonstration

- Pilot site
- VECNet Development Site
VECNet DW
Implementation
DW Implementation

Operational Source systems

- Household Surveys
- Operational Coverage
- PACRAIN Rainfall
- Georeferenced EIR
- GSOD Weather
- Insecticides
- Species
- Entomological Endpoints

Logical Modeling Schemas

Relational Database
PostgreSQL

Extract
### DW Implementation

**Operational Source systems**
- Household Surveys
- Operational Coverage
- PACRAIN Rainfall
- Georeferenced EIR
- GSOD Weather
- Insecticides
- Species
- Entomological Endpoints

**Logical Modeling Schemas**

**Data Staging Area**
- **Transform**: Clean, combine, and standardize
- Conform dimensions

**Data Store**: Flat files
- Relational tables

**Processing**: Sorting
- Sequential processing

**Extract-Transformation-Load (ETL)**
- **Python**

**Relational Database**
- PostgreSQL
DW Implementation

Operational Source systems
- Household Surveys
- Operational Coverage
- PACRAIN Rainfall
- Georeferenced EIR
- GSOD Weather
- Insecticides
- Species
- Entomological Endpoints

Logical Modeling Schemas

Data Staging Area
- Transform: Clean, combine, and standardize Conform dimensions
- Data Store: Flat files Relational tables
- Processing: Sorting Sequential processing

Extract-Transformation-Load (ETL) Python

Data Presentation Area
- Household Surveys
- ITNs Operational Coverage
- IRS Operational Coverage
- ACT Operational Coverage
- PACRAIN Rainfall
- Georef. EIR
- GSOD Weather
- Insecticides
- Lookup Tables

Cubes Framework

Relational Database PostgreSQL

Extract

Load
Cubes Framework

- **backend**: logical model
- open-source
- OLAP Python framework
- independent of the physical implementation
**Django Web Framework**

- **frontend**: web-based user interfaces
- open-source
- Python web application framework
- **model-view-controller** software pattern
- **Don’t Repeat Yourself (DRY)** principle
  - avoids repeating HTML code

---

**Logical Modeling Schemas**

**Extract-Transformation-Load (ETL) Python**

**Insecticides**

**Lookup Tables**

**Cubes Framework**

**Data Access Tools**

- Web UI
- Aggregated results
- Reports
- Visualization
  - Graphs
  - Charts

**Django Web Framework**

Apache

Web UI

Aggregated results

Reports

Visualization

Graphs

Charts
**DW Data Flow**

**Data Sources (OSSs)**

- WMR
- MARA
- PACRAIN

**Backend:**

- Cubes Framework
- Extract-Transformation-Load (ETL)
- Python

- Relational Database
- PostgreSQL

- Aggregation Browser

**Logical Model**

- JSON

- Cell

- Aggregates

**Frontend:**

- Django Web Framework

- Graphical User Interfaces (GUIs)

- HTTP Request

- Results

**Web Browsers**

**DW Users**

**Backend:** Cubes Framework
Summary
Contributions: Part 1

- ABM of *An. gambiae*
- verification & validation of ABMs
- a spatial extension
- a landscape generator tool
- vector control interventions
- integration of the ABM with a GIS
Contributions: Part 2

• a multi-dimensional DW
• for malaria-related data
• logical, conceptual design
  • dimensional modeling
Contributions: Part 2

• a multi-dimensional DW
• for malaria-related data
• logical, conceptual design
  • dimensional modeling
• implementation: Cubes, Django
Contributions: Part 2

• the DW:
  • is new and novel
Contributions: Part 2

• the DW:
  • is new and novel
  • can be generalized to other biological domains
Contributions: Part 2

• the DW:
  
  • is new and novel
  
  • can be generalized to other biological domains
  
  • can serve as a proof-of-concept for future biological data storage systems
Contributions: Part 2

• the DW:
  • is new and novel
  • can be generalized to other biological domains
  • can serve as a proof-of-concept for future biological data storage systems
  • the logical design is currently being used by the VECNet CI team
Publications

A Spatial Agent-Based Model of Malaria: Model Verification and Effects of Spatial Heterogeneity

S. M. Niaz Arifin, Gregory J. Davis, and Ying Zhou

Modeling Space in an Agent-Based Model of Malaria: Comparison between Non-Spatial and Spatial Models

S. M. Niaz Arifin, Gregory J. Davis, Ying Zhou, Gregory R. Madey
2011 Spring Simulation Multiconference (SpringSim '11), April 4-7, 2011, Boston, MA, USA.

Divide and Conquer: A Four-Fold Docking Experience of Agent-Based Models

S. M. Niaz Arifin, Gregory J. Davis, Steve Kurtz, James E. Gentile, Ying Zhou, Gregory R. Madey
Winter Simulation Conference (WSC), December 5-8, 2010, Baltimore, Maryland.

Verification & Validation by Docking: A Case Study of Agent-Based Models of Anopheles gambiae

S. M. Niaz Arifin, Gregory J. Davis, Ying Zhou and Gregory R. Madey
Summer Computer Simulation Conference (SCSC 2010), Ottawa, Canada.
Publications

P-SAM: A Post-Simulation Analysis Module for Agent-Based Models

S. M. Niaz Arifin, Ryan C. Kennedy, Kelly E. Lane, Gregory R. Madey, Agustin Fuentes and Hope Hollocher
Summer Computer Simulation Conference (SCSC 2010), Ottawa, Canada.

An Agent-based Model of the Anopheles gambiae Mosquito Life Cycle

Ying Zhou, S. M. Niaz Arifin, James Gentile, Steven Kurtz, Gregory Davis, Barbara Wendelberger and Greg Madey
Summer Computer Simulation Conference (SCSC 2010), Ottawa, Canada.

A GIS Aware Agent-Based Model of Pathogen Transmission

Ryan C. Kennedy, Kelly E. Lane, S. M. Niaz Arifin, Agustin Fuentes, Hope Hollocher and Gregory R. Madey
Publications

Examining the impact of larval source management and insecticide-treated nets using a spatial agent-based model of *Anopheles gambiae* and a landscape generator tool

*S. M. Niaz Arifin*, Gregory R. Madey and Frank H. Collins
Malaria Journal. (Under review)

Integrating an Agent-based Model of Malaria Mosquitoes with a Geographic Information System

*S. M. Niaz Arifin*, Rumana Reaz Arifin, Dilkushi De Alwis Pitts and Gregory R. Madey
25th European Modeling and Simulation Symposium (EMSS), September 2013, Athens, Greece. (Under review)

An Online Analytical Processing (OLAP) System for a Multi-dimensional Data Warehouse for Malaria Data

*S. M. Niaz Arifin*, Benoit Raybaud, Gregory R. Madey
PLOS ONE. (In preparation)
Future Work

• predictive data
  • priorities: OpenMalaria and EMOD
Future Work

• predictive data
  • priorities: OpenMalaria and EMOD

• progressively advance the DW
  • incorporate new data sources
Future Work

• predictive data
  • priorities: OpenMalaria and EMOD

• progressively advance the DW
  • incorporate new data sources

• content integration with other tools in VECNet CI
Future Work

• features under development
  • aggregation, charting (graphs), exporting

• advanced data browsing techniques
  • faceted navigation of attributes

• geo-spatial aspects of navigation
  • maps for data browsing
## Acknowledgements

<table>
<thead>
<tr>
<th>Committee Members</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Dr. Gregory R. Madey</td>
<td>Computer Science and Engineering</td>
</tr>
<tr>
<td>Dr. Frank H. Collins</td>
<td>Biological Sciences</td>
</tr>
<tr>
<td>Dr. Patrick J. Flynn</td>
<td>Computer Science and Engineering</td>
</tr>
<tr>
<td>Dr. Jessica J. Hellmann</td>
<td>Biological Sciences</td>
</tr>
</tbody>
</table>
# Acknowledgements

<table>
<thead>
<tr>
<th>Colleagues</th>
<th>VECNet CI Team</th>
<th>College of Science</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ryan C. Kennedy</td>
<td>Benoit Raybaud</td>
<td>Neil F. Lobo</td>
</tr>
<tr>
<td>Rumana Reaz Arifin</td>
<td>Natalie Meyers</td>
<td>Kelly Lane</td>
</tr>
<tr>
<td>Ying Zhou</td>
<td>Timothy Wright</td>
<td>Hope Hollocher</td>
</tr>
<tr>
<td>Gregory J. Davis</td>
<td>Lawrence Selvy</td>
<td>Agustin Fuentes</td>
</tr>
<tr>
<td>Steve Kurtz</td>
<td>Zachary Torstrick</td>
<td></td>
</tr>
<tr>
<td>James E. Gentile</td>
<td>Alex Vyushkov</td>
<td></td>
</tr>
</tbody>
</table>

**Bill & Melinda Gates Foundation**

*and others...*
Thank you!