Data Science with Visual Analytics

30 Min Teaser Talk

Klaus Mueller
Stony Brook University and SUNY Korea

SUNY Korea 2015 HOT-T-CS
Hot Topics in Computer Science

July 13 – 17, 2015
SUNY Korea, Songdo, Korea

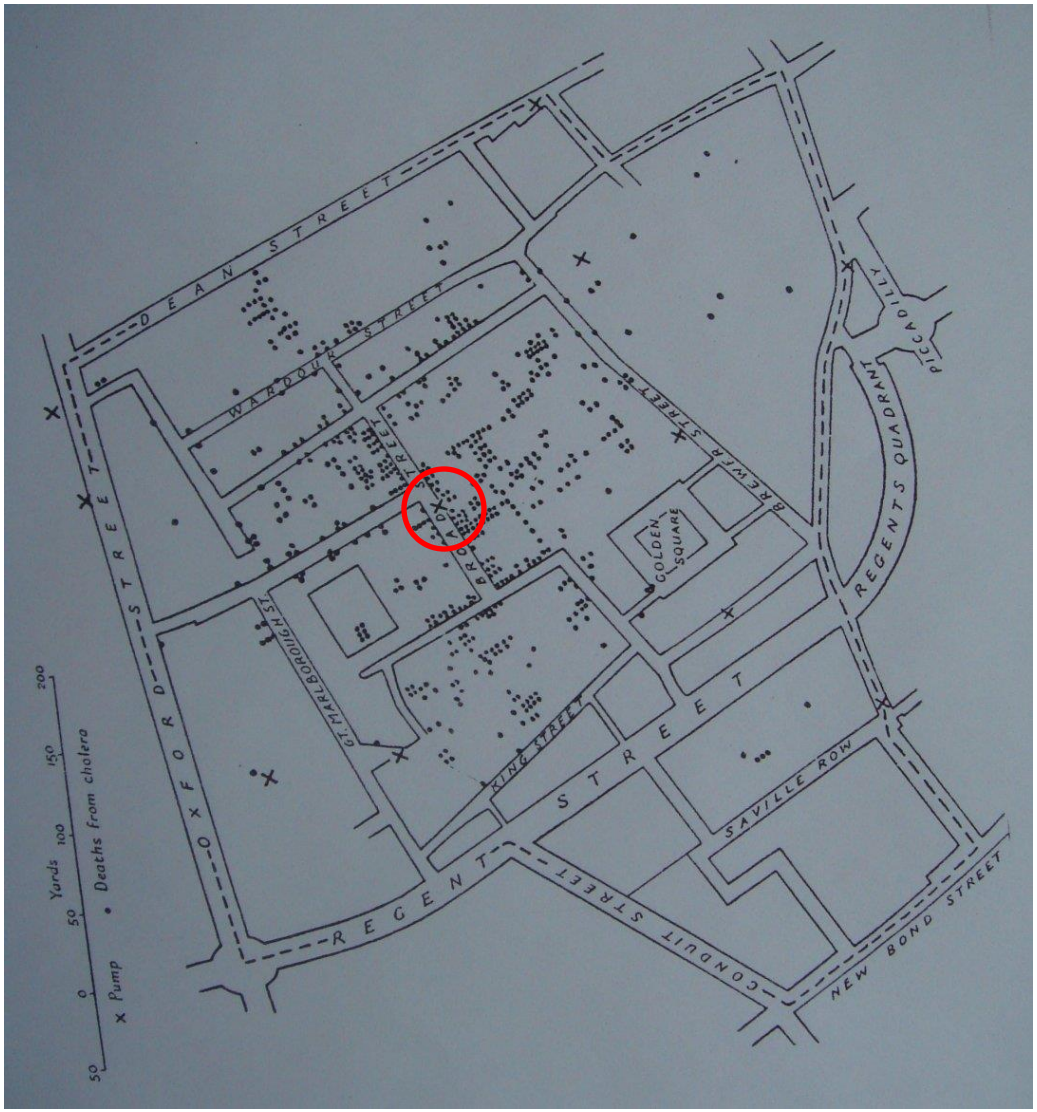
Data Science – Why All The Excitement?

Small Data Example

Dr. John Snow's London Cholera Map (1854)

- data collection
- data assimilation
- statistical testing
- visualization
- computational analysis (brain)
- domain knowledge

Very early example of data science



MODERN DATA SCIENTIST

21th century, requires a mixture of
, computer science, communicatio
who a data scientist is, is equally h
the modern data scientist really i

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
-
-

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages, e.g., R
- Databases: SQL and NoSQL
- Relational algebra
-

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

Google Flu Trends

Predict emerging flu from search terms in specific regions

Could predict regional outbreaks of flu up to 10 days before reported by the CDC

Nate Silver's Election Predictions

elections2012

Live results | President | Senate | House | Governor | Choose your

Numbers nerd Nate Silver's forecasts prove all right on election night

FiveThirtyEight blogger predicted the outcome in all 50 states, assuming Barack Obama's Florida victory is confirmed

Luke Harding
guardian.co.uk, Wednesday 7 November 2012 10.45 EST

Takes a big-picture approach

- use multiple sources of unique data
- combine with historical data
- apply principles of sound statistical analysis

Opportunities galore

IBM

Government achieves significant cost savings and ability to react to potential threats quickly

Government cuts acoustic analysis from hours to **70 Milliseconds**

Utility provider improves prediction of power outages

Utility avoids power failures by analyzing **10 PB** of data in minutes

Hospital detects and intervenes in potentially life-threatening conditions

Hospital analyzes streaming vitals to intervene **24 hours earlier**

Retailer optimizes inventory levels and product mix

Retailer reduces time to run queries by **80%**

Stock exchange reduces time to insights to achieve optimal buying / selling strategies

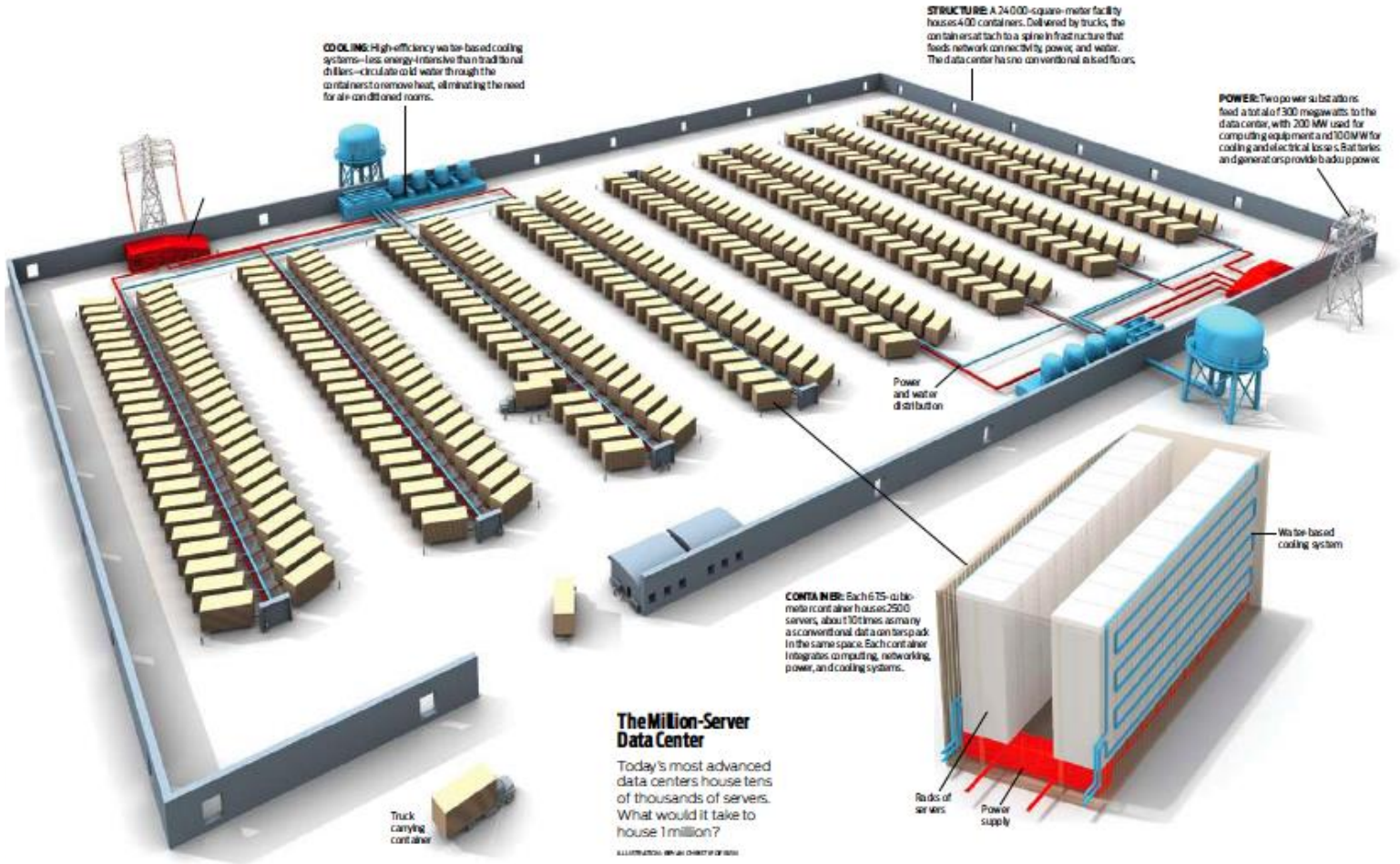
Stock Exchange cuts queries from 26 hours to **2 minutes** on **2 PB**

Telco provider improves ability to quickly address network issues / opportunities

Telco analyses streaming network data to reduce hardware costs by **90%**

Million Server Data Center

Not always needed



Data Scientists

The U.S. will need 140,000-190,000 predictive analysts and 1.5 million managers/analysts by 2018

- McKinsey Global Institute's June 2011

New Data Science institutes being created or repurposed – NYU, Columbia, Washington, UCB,...

New degree programs, courses, boot-camps:

- e.g., at Berkeley: Stats, I-School, CS, Astronomy…

Stony Brook as well as SUNY Korea will offer an MS Specialization in Data Science starting now (Fall 2015)

Characteristics of Big Data

Volume

Data at scale

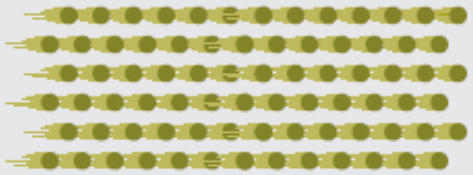
Terabytes to petabytes of data

Variety

Data in many forms

Structured, unstructured, text, multimedia

Velocity



Data in motion

Analysis of streaming data to enable decisions within fractions of a second

Veracity

Data uncertainty

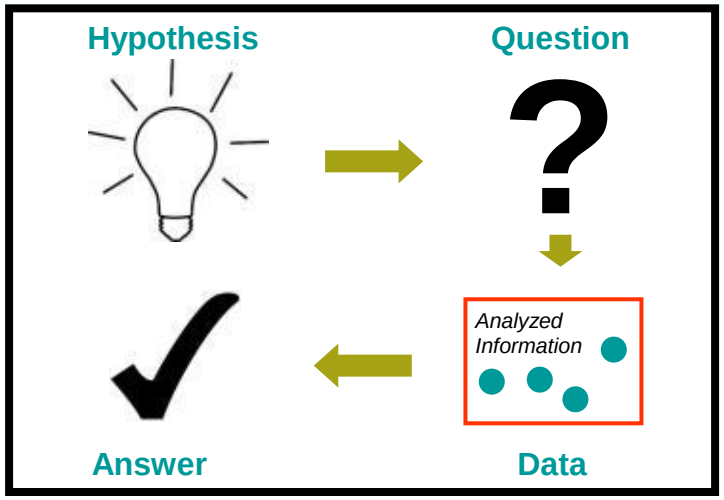
Managing the reliability and predictability of inherently imprecise data types

IBM

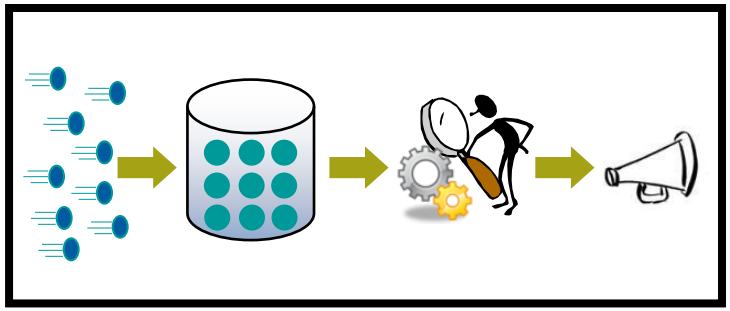
Big Data Approach to Science

Traditional Analytics

Structured & Repeatable
Structure built to store data



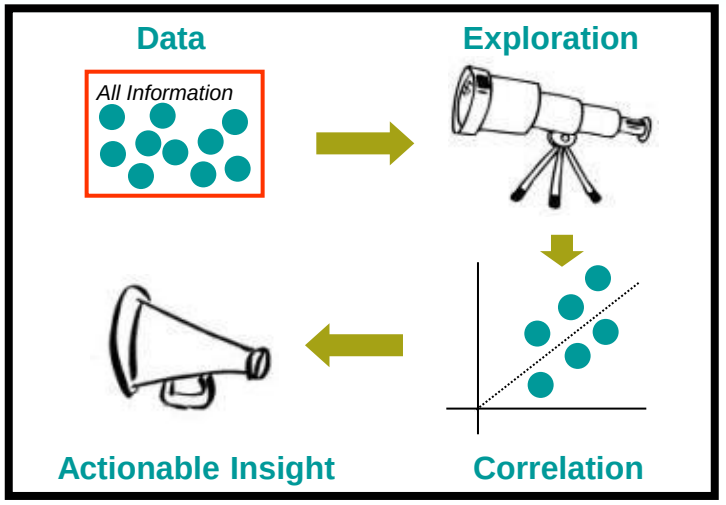
Start with hypothesis
Test against selected data



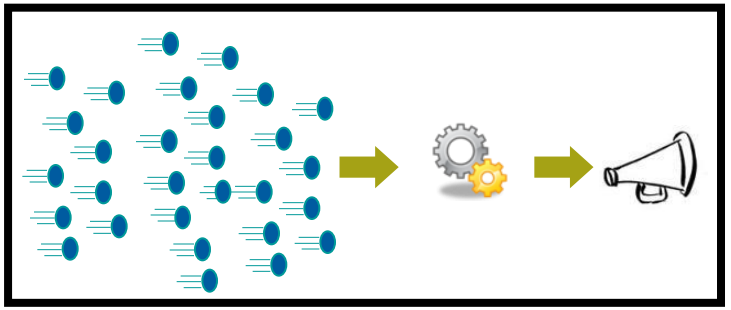
Analyze after landing…

Big Data Analytics

Iterative & Exploratory
Data is the structure



Data leads the way
Explore *all* data, identify correlations



Analyze in motion…

Databases vs. data Science

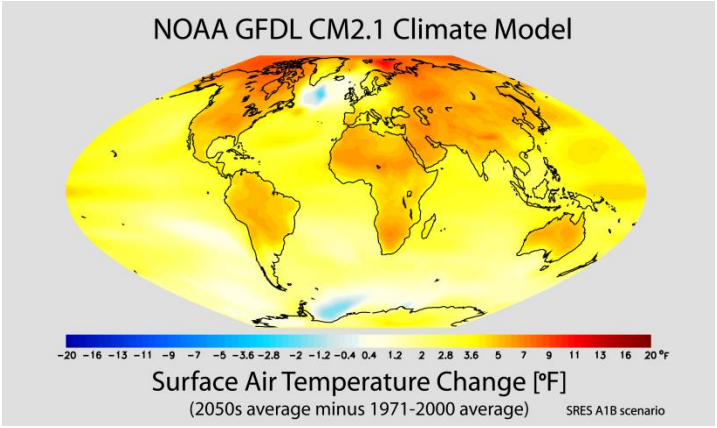
	Databases	Data Science
Data Value	"Precious"	"Cheap"
Data Volume	Modest	Massive
Examples	Bank records, Personnel records, Census, Medical records	Online clicks, GPS logs, Tweets, Building sensor readings
Priorities	Consistency, Error recovery, Auditability	Speed, Availability, Query richness
Structured	Strongly (Schema)	Weakly or none (Text)
Properties	Transactions, ACID*	CAP* theorem (2/3), eventual consistency
Realizations	SQL	NoSQL: Riak,Memcached, Apache River, CouchDB. etc.
Approach	Query the past	Query the future

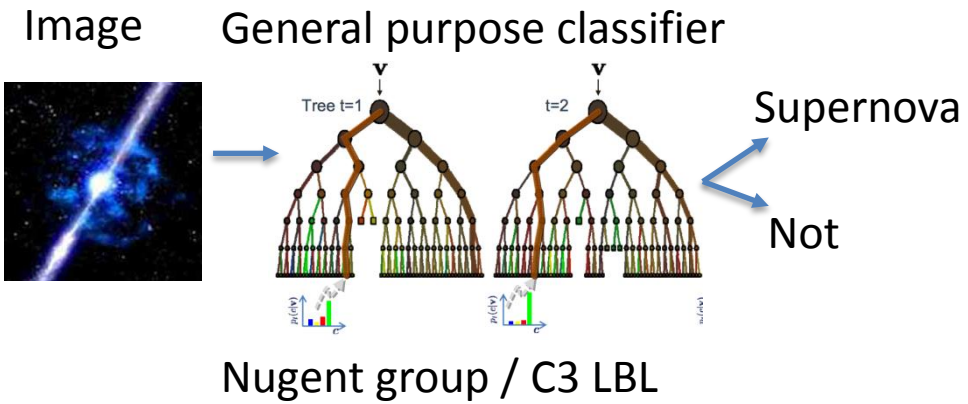
CAP = Consistency, Availability, Partition Tolerance

ACID = Atomicity, Consistency, Isolation and Durability

John Canny, Berkeley

Scientific vs. data-Driven Modeling





Scientific Modeling	Data-Driven Approach
Physics-based models	General inference engine replaces model
Problem-Structured	Structure not related to problem
Mostly deterministic, precise	Statistical models handle true randomness, and unmodeled complexity
Run on Supercomputer or High-end Computing Cluster	Run on cheaper computer Clusters

John Canny, Berkeley

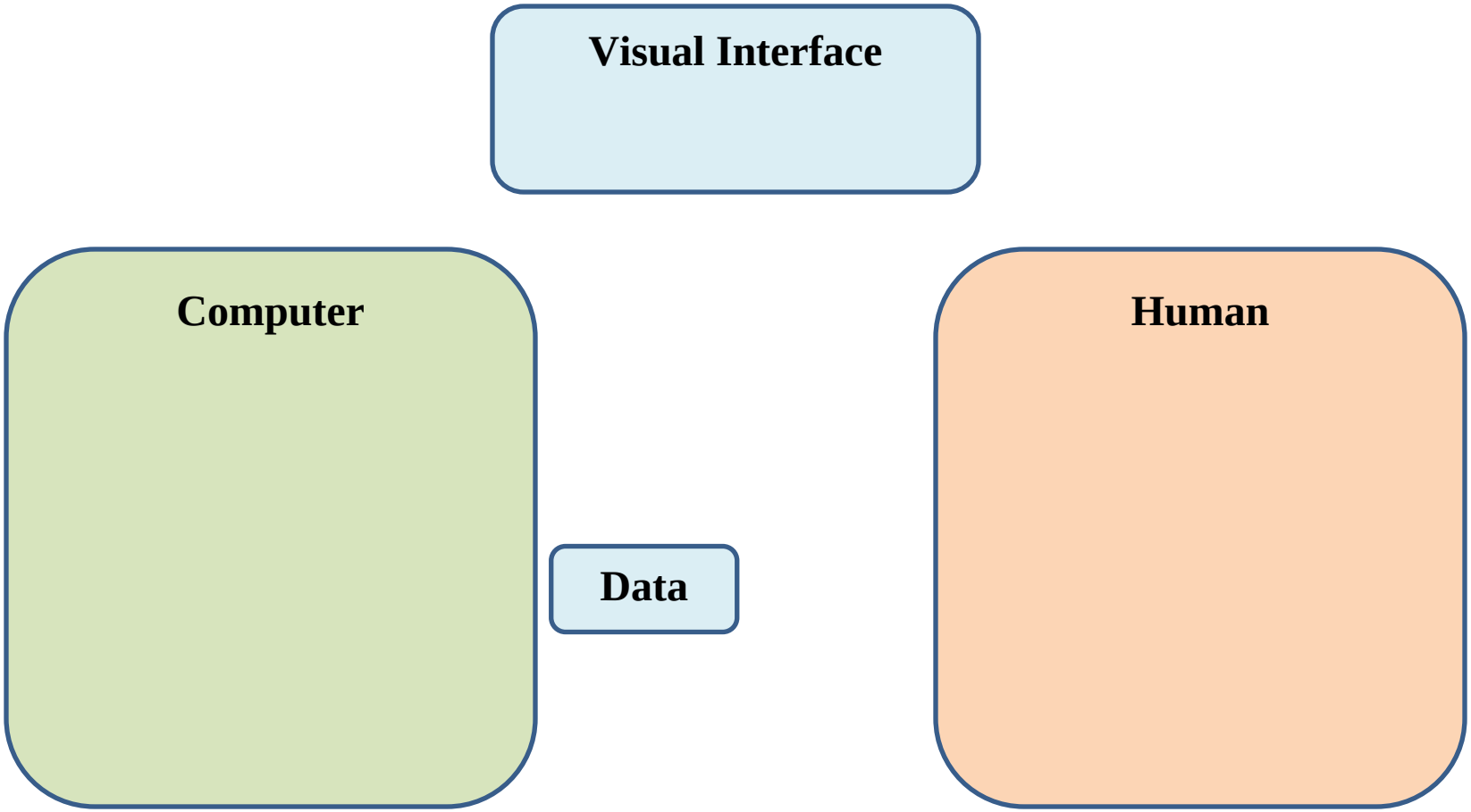
Visual Data Science
or
Visual Analytics

An Introductory Example

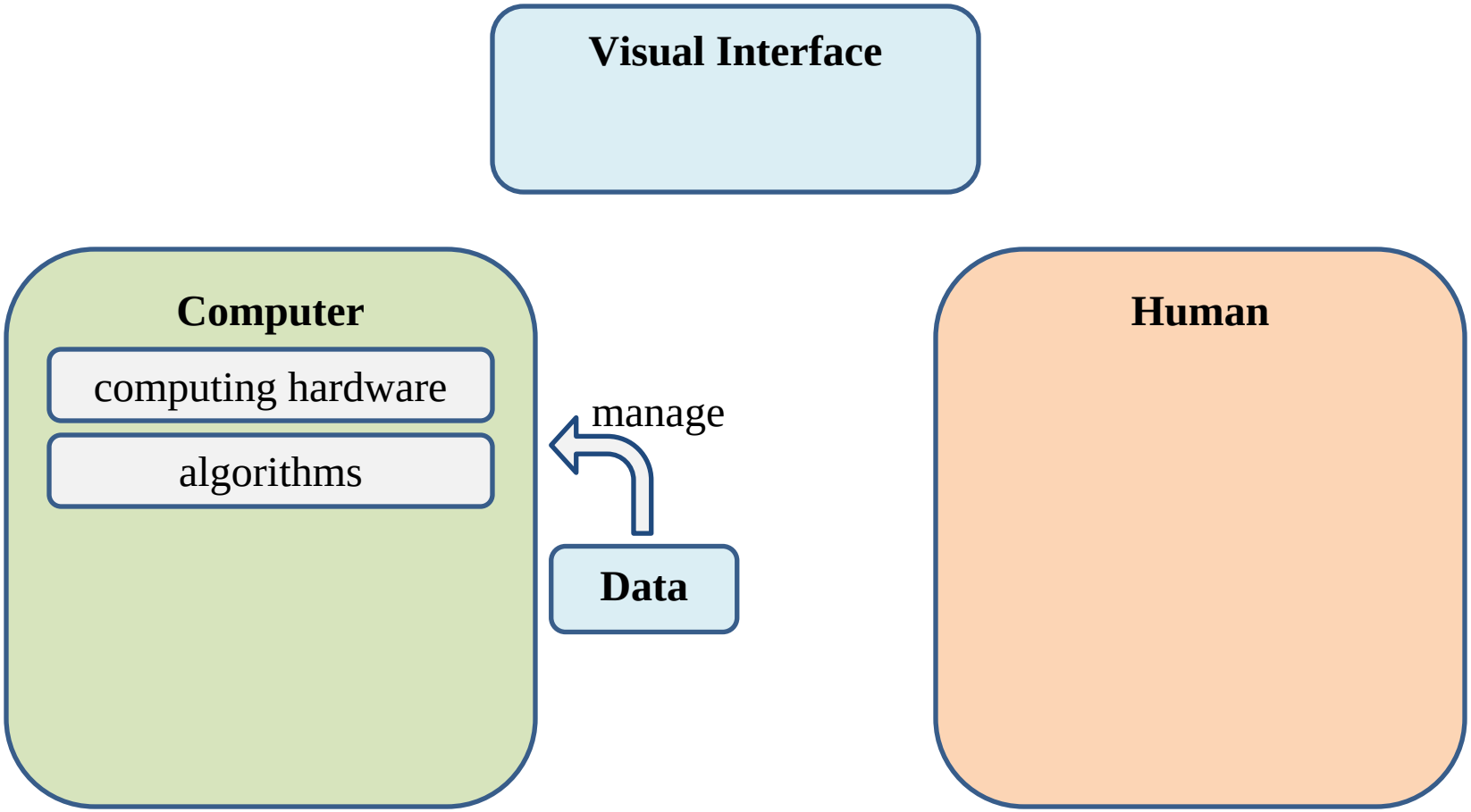
The Georgia Tech Jigsaw System

Next a more Conceptual View

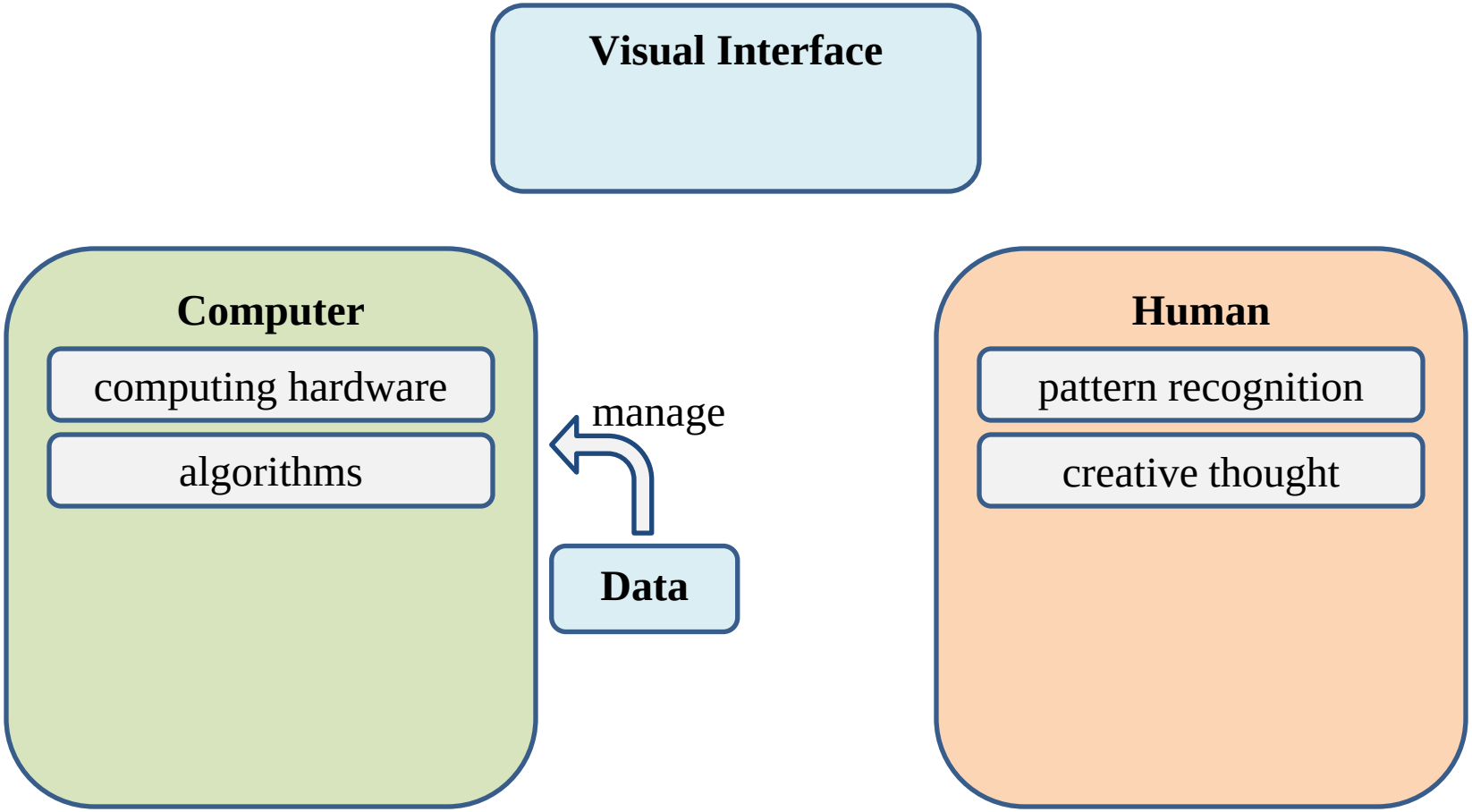
Visual Analytics



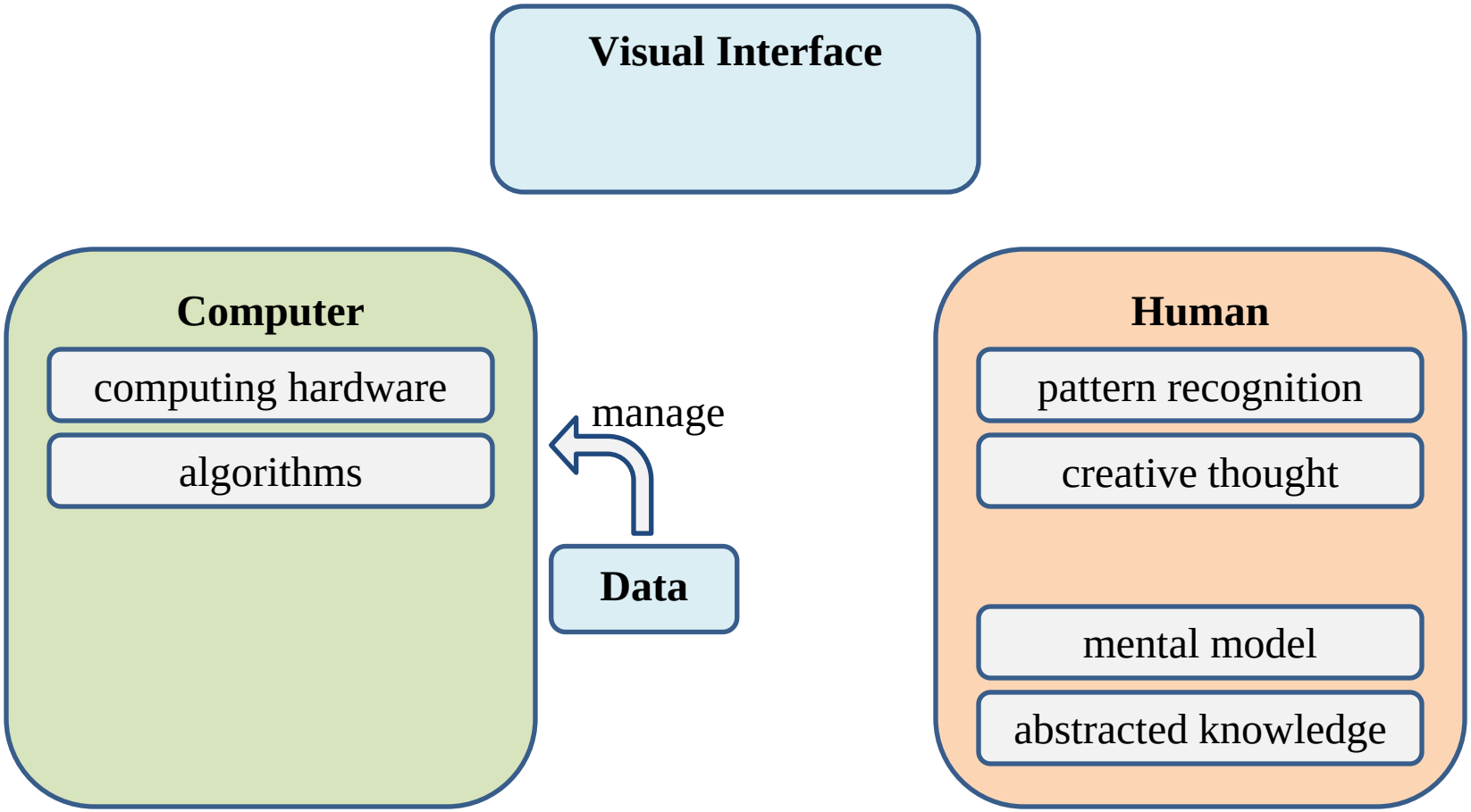
Visual Analytics



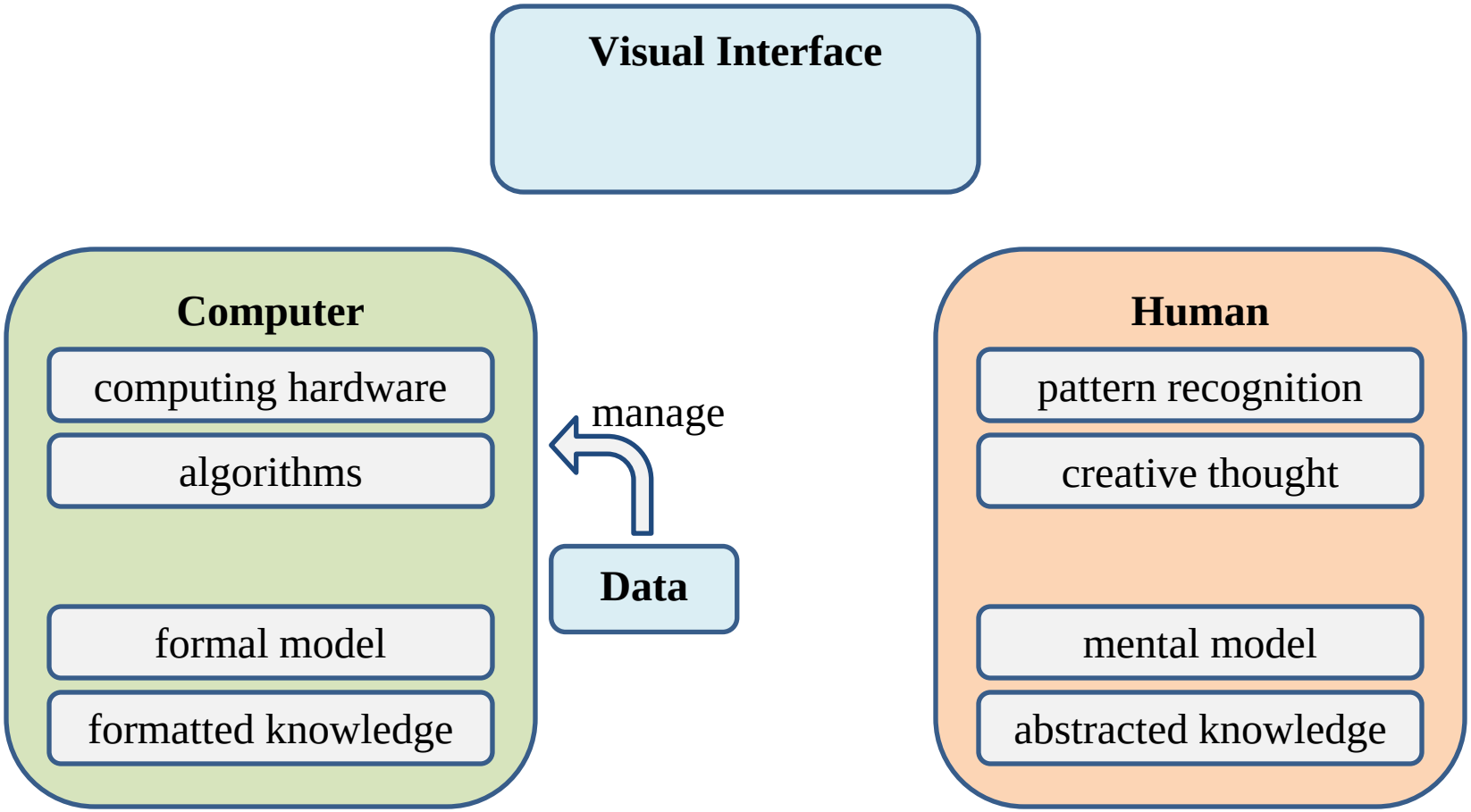
Visual Analytics



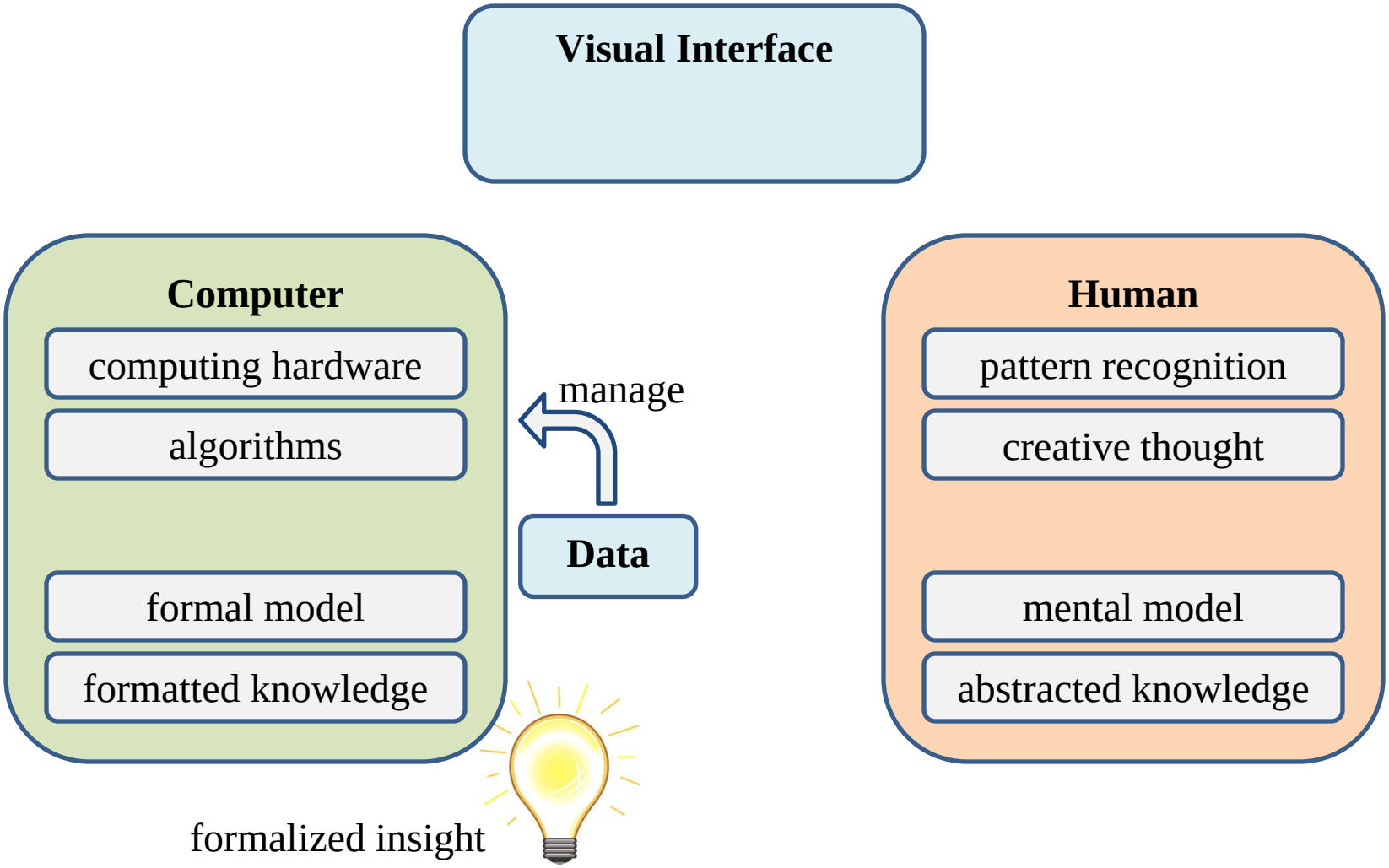
Visual Analytics



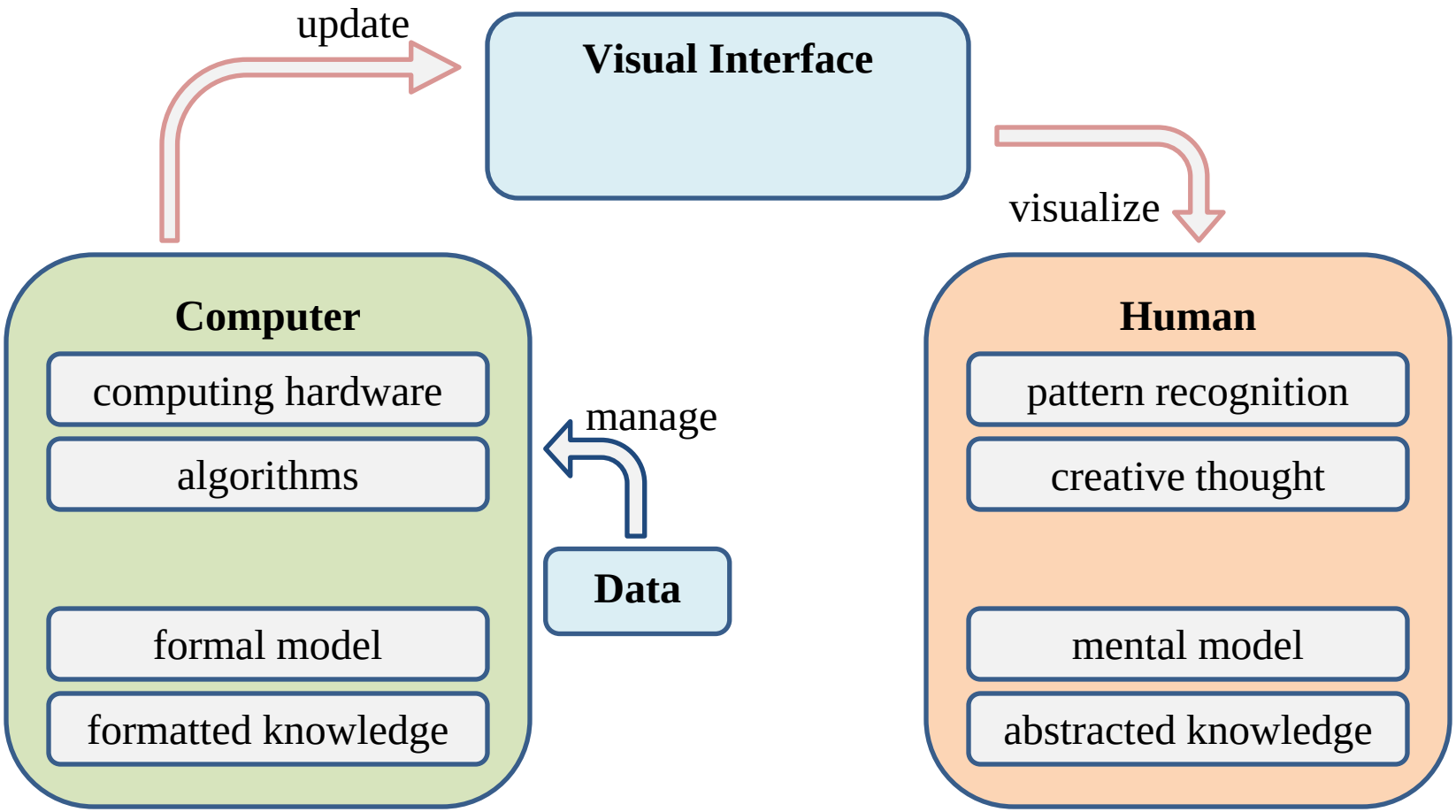
Visual Analytics



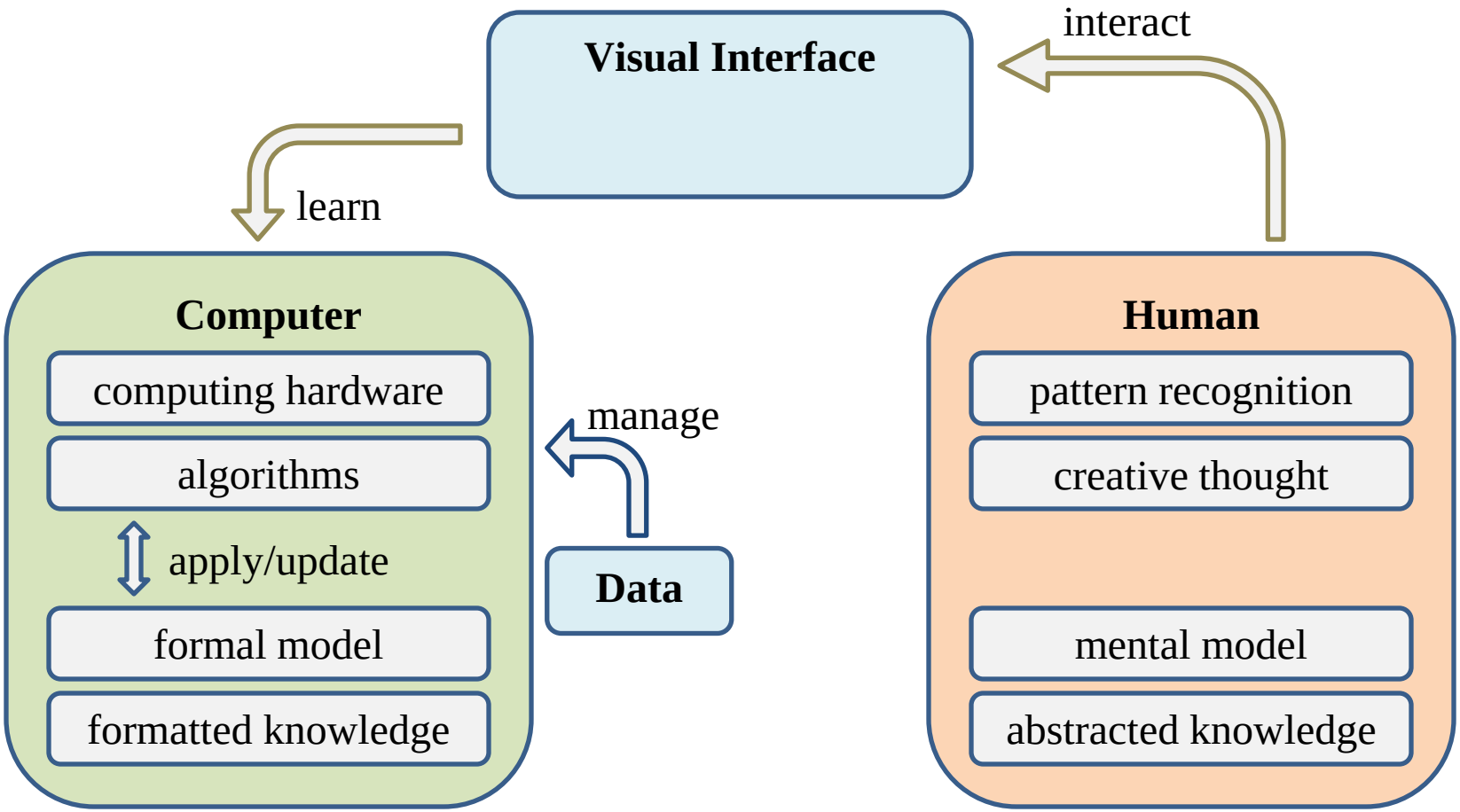
Visual Analytics



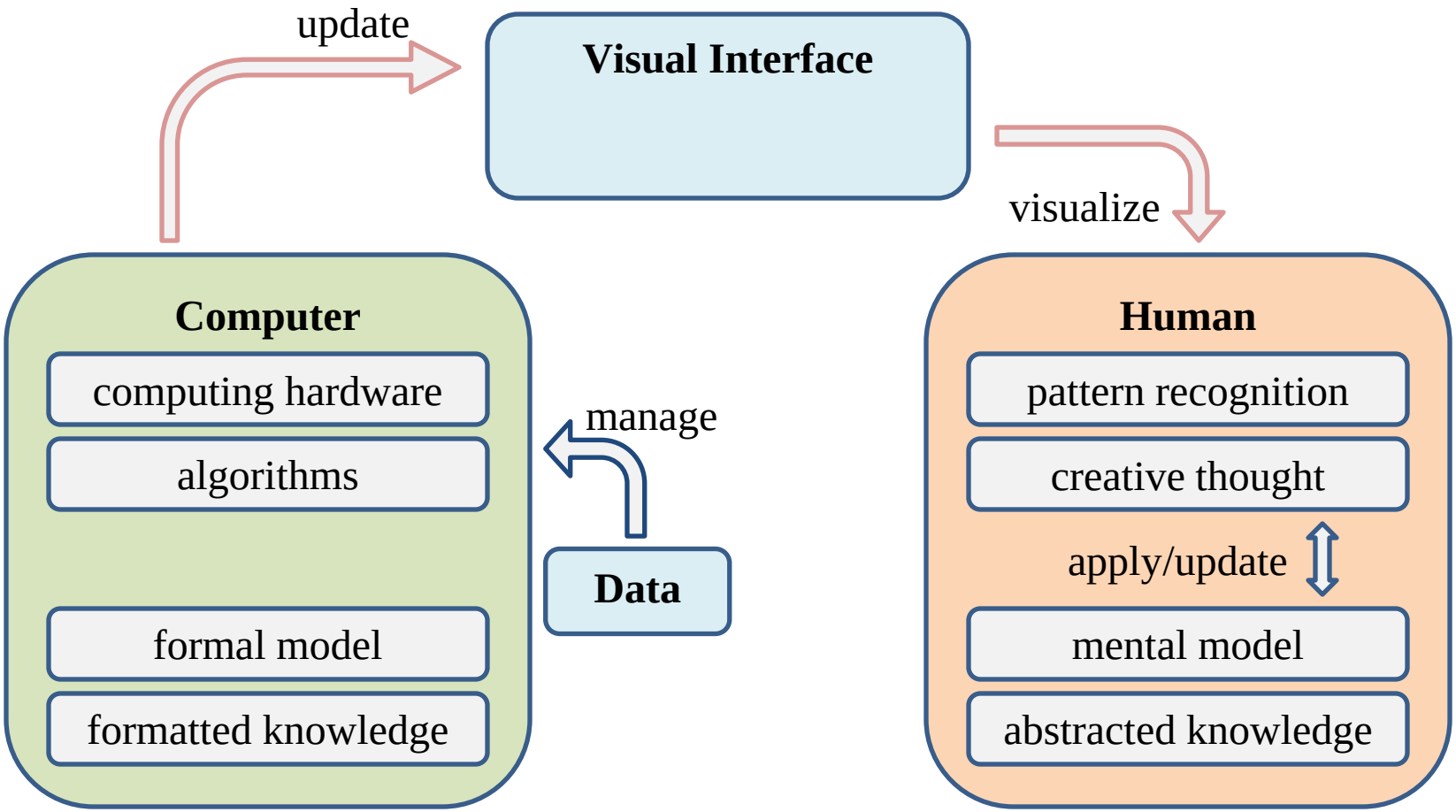
Visual Analytics



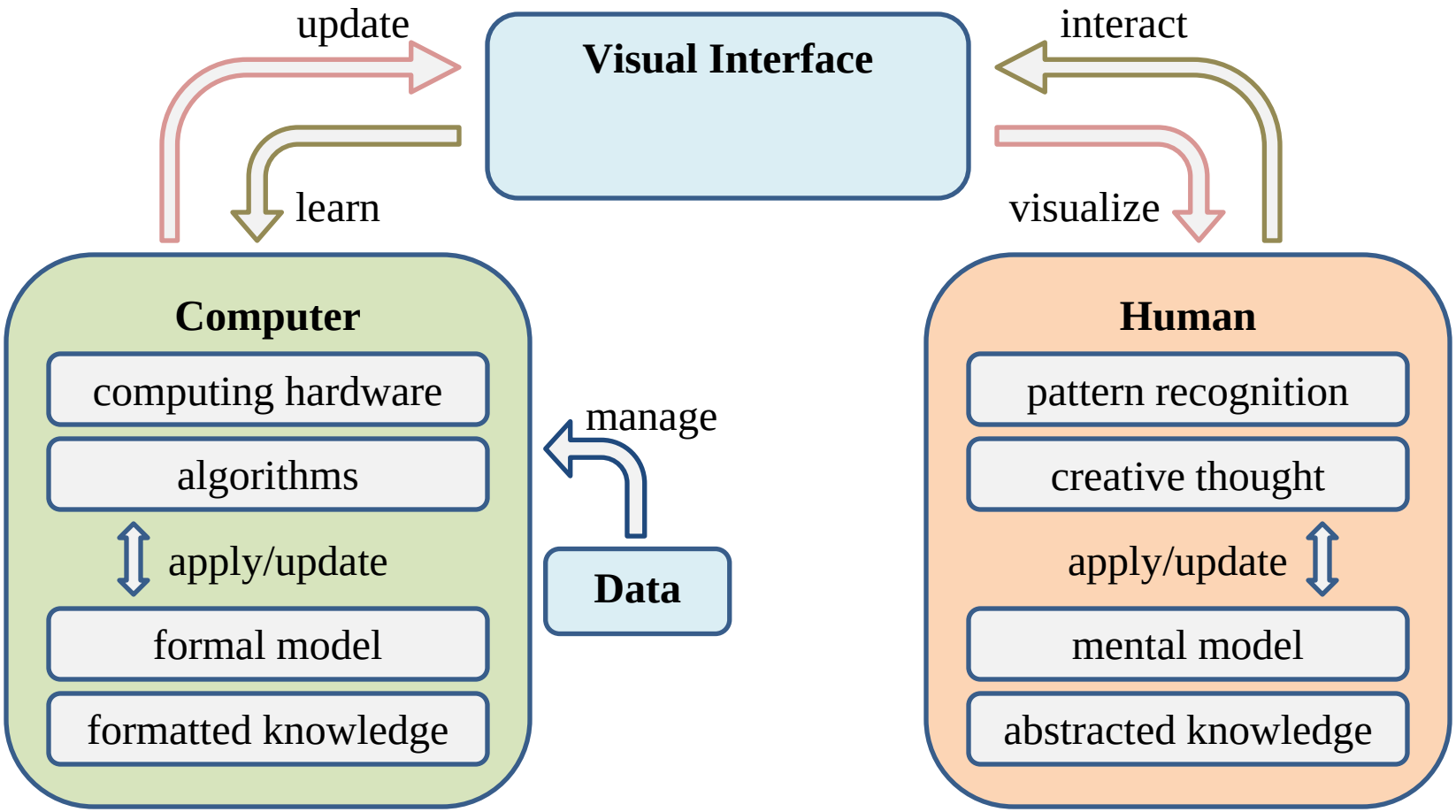
Visual Analytics



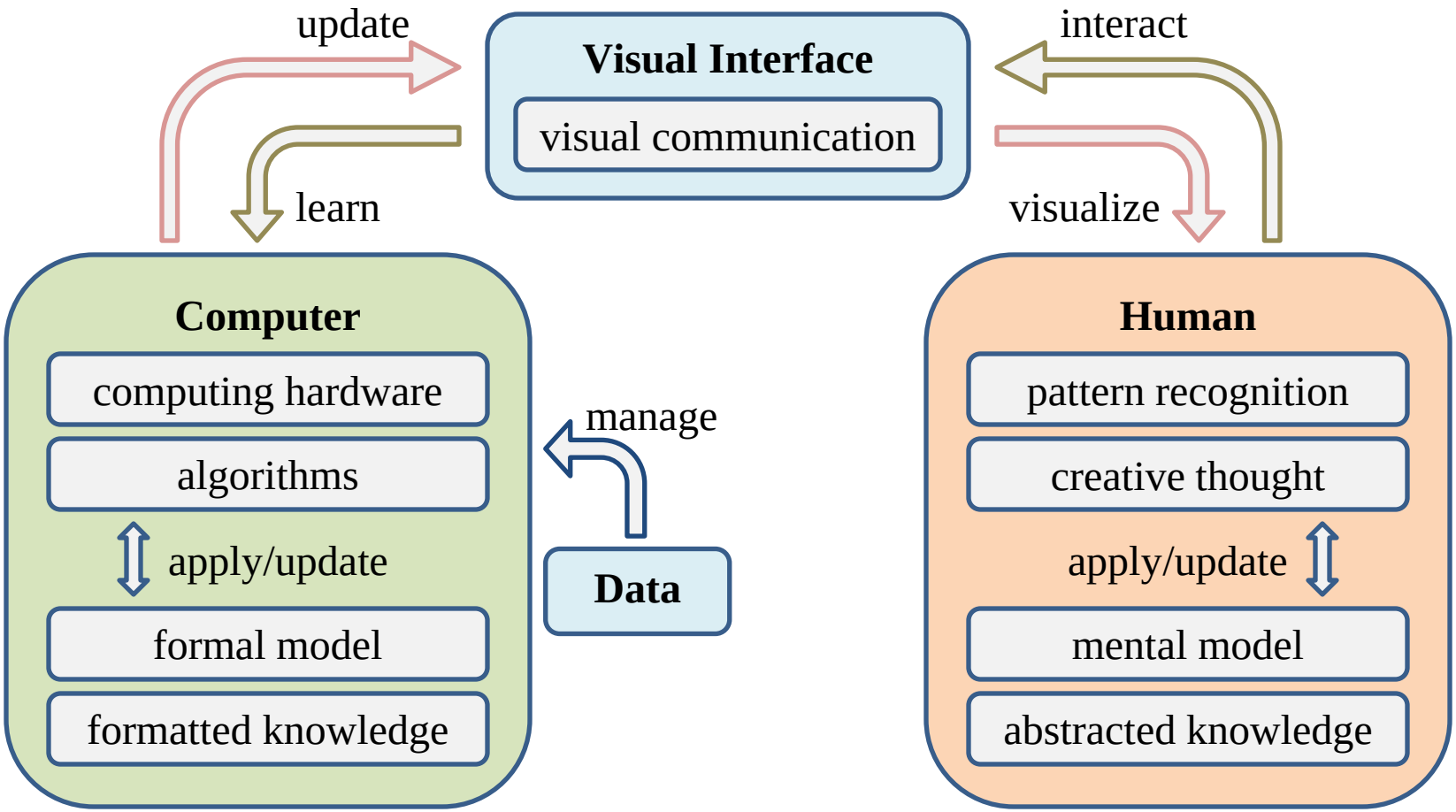
Visual Analytics



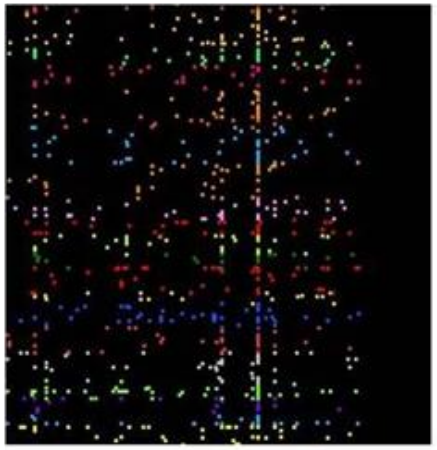
Visual Analytics

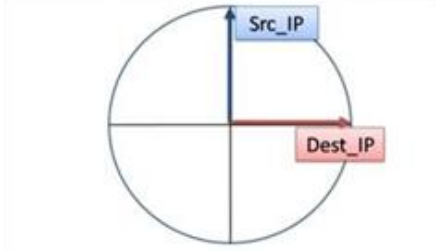


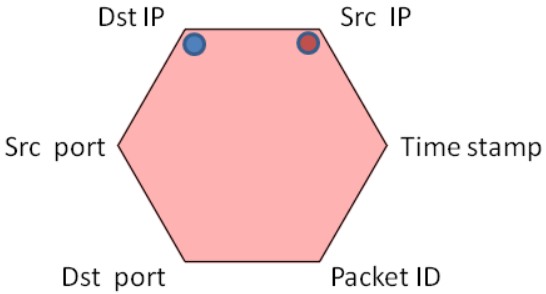
Visual Analytics



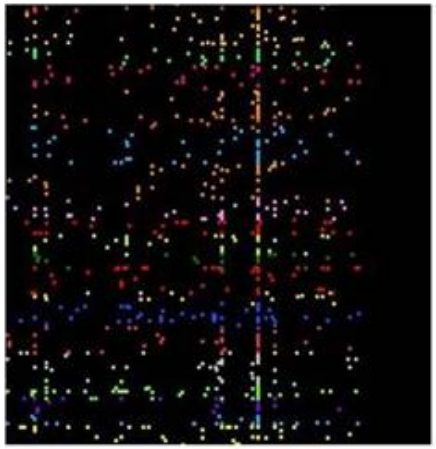
Model-Learning Example

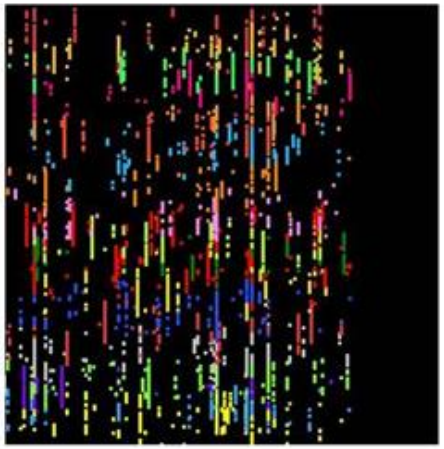


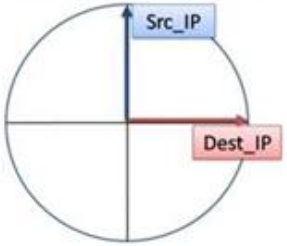


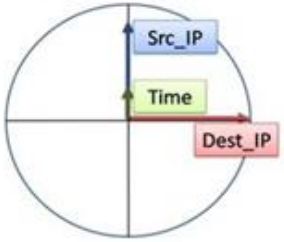


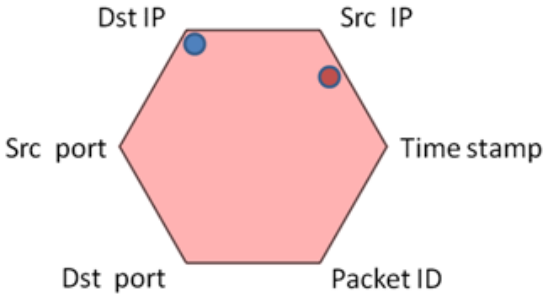
Model-Learning Example



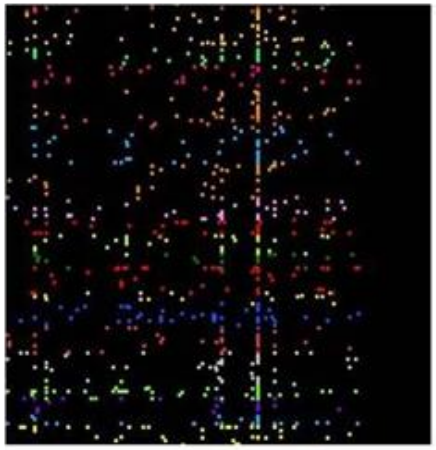


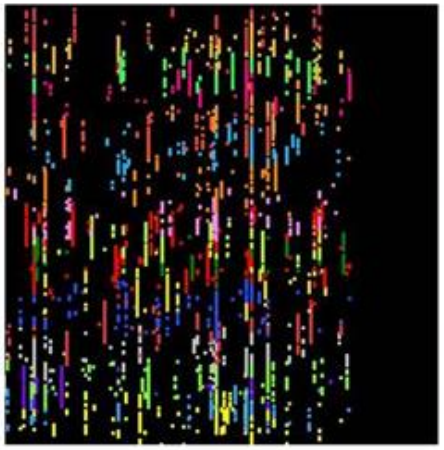


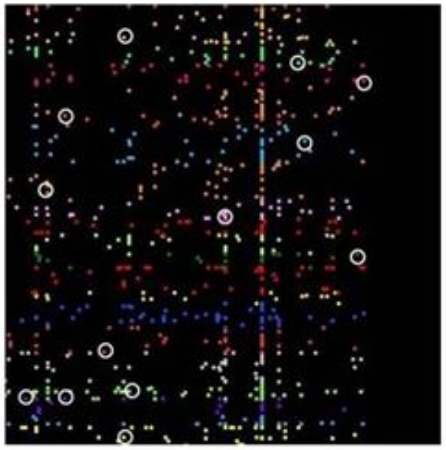


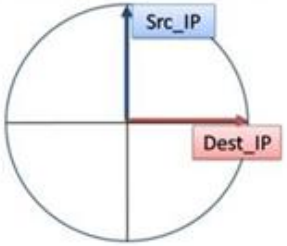


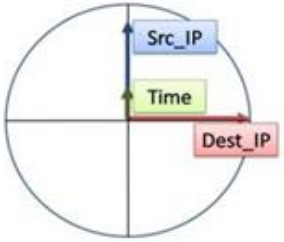
Model-Learning Example

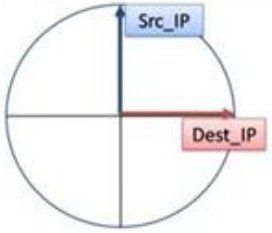


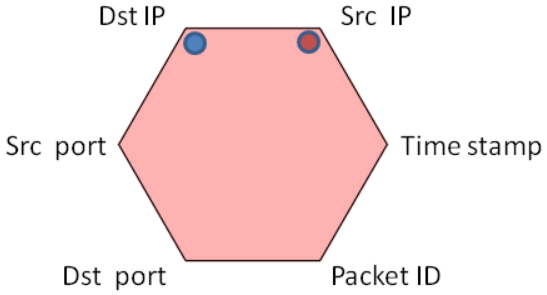




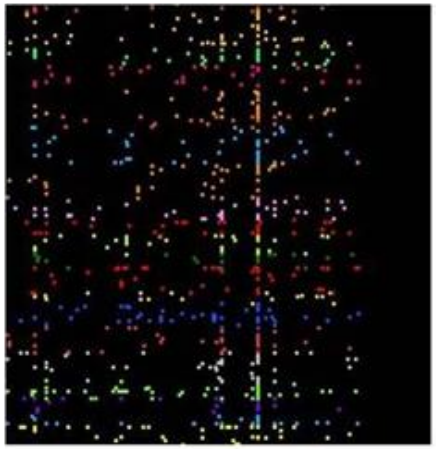


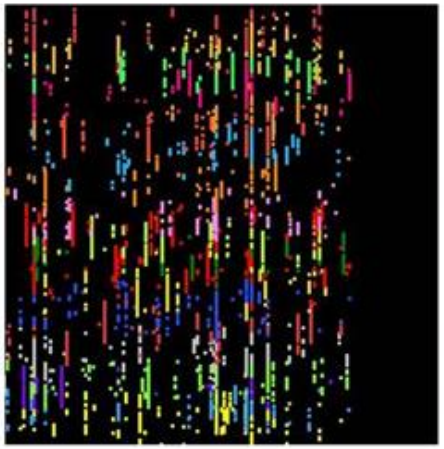


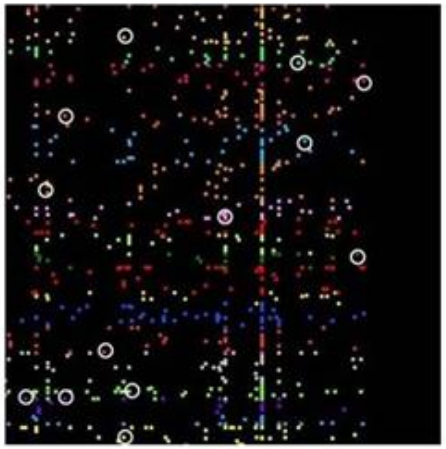


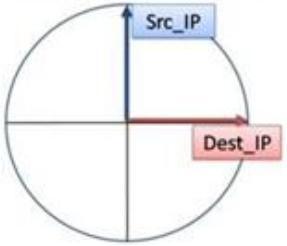


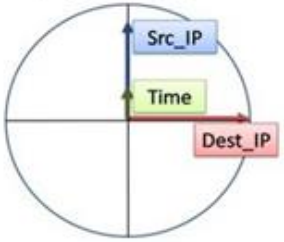
Model-Learning Example

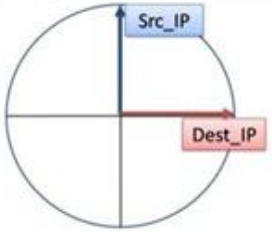




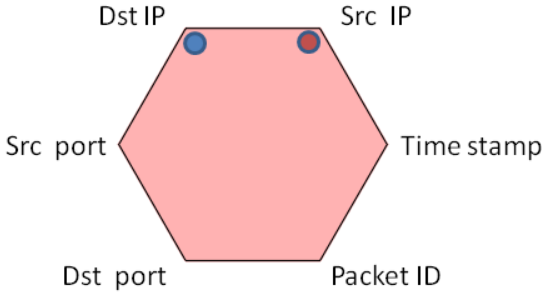




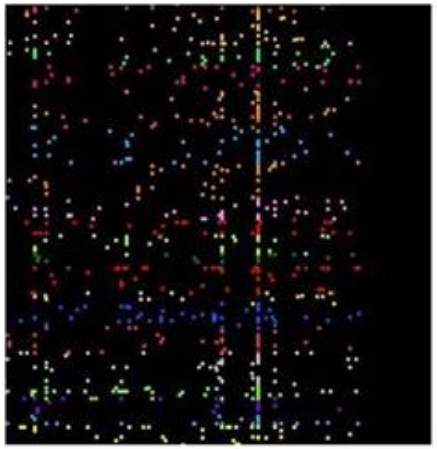


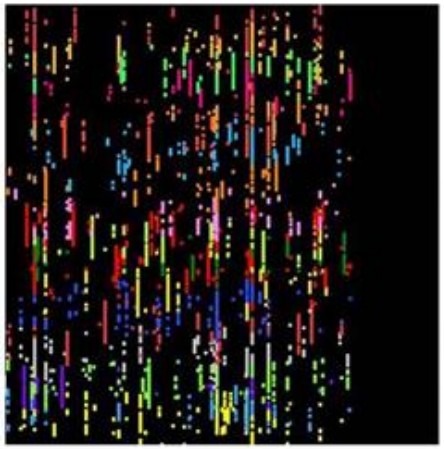


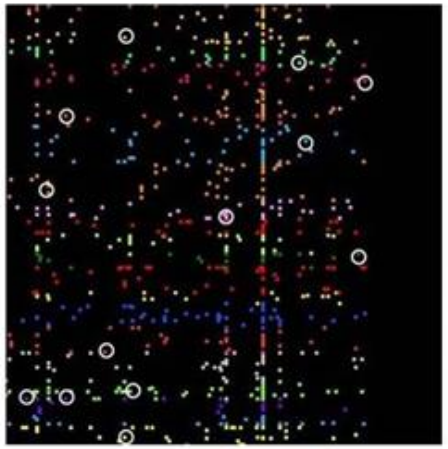
webpage_load(X) :-
same_src_ips(X),same_dest_ips(X),
same_src_port(X,80),
timeframe_upper(X,10)

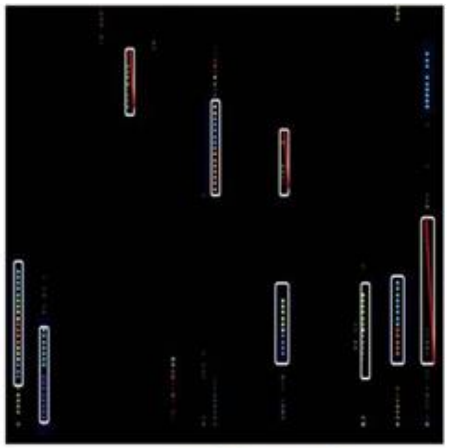


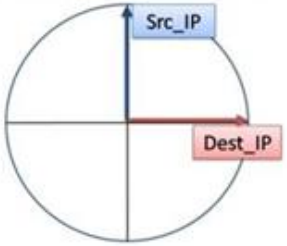
Model-Learning Example

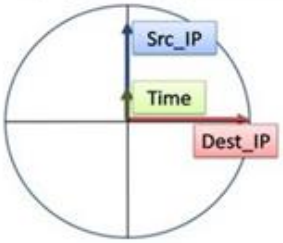


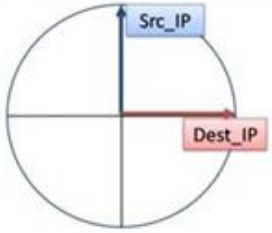


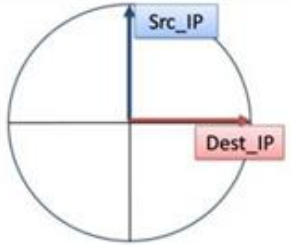


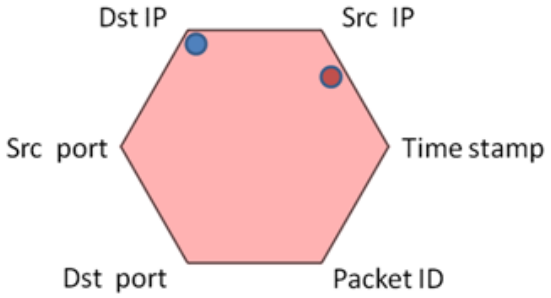




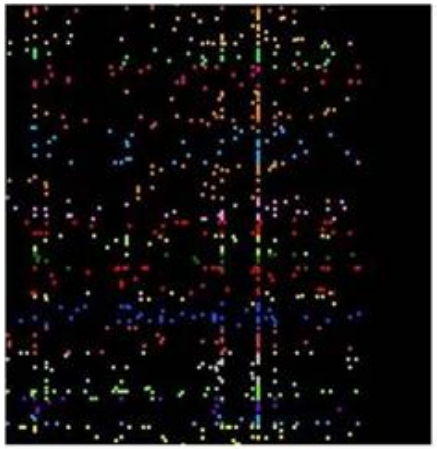


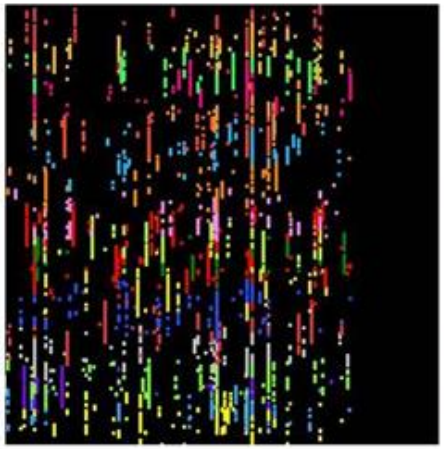


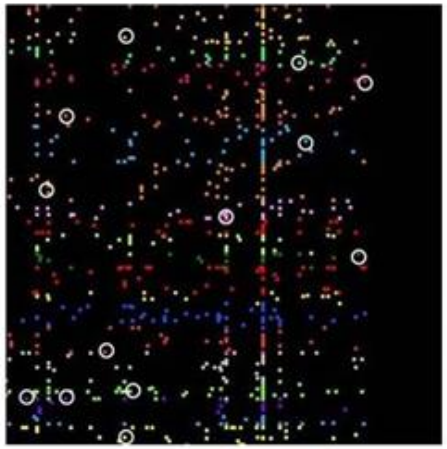


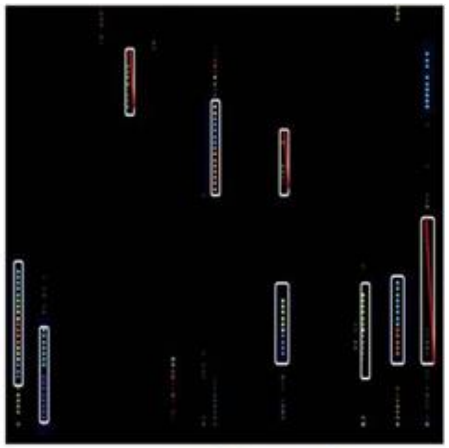


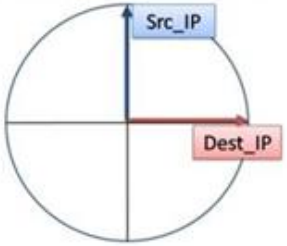
Model-Learning Example

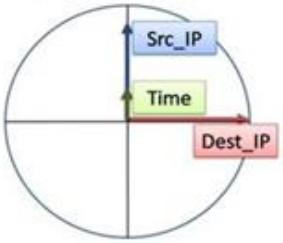


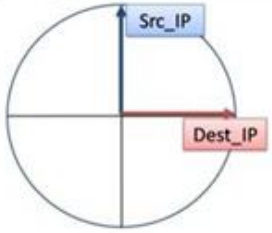


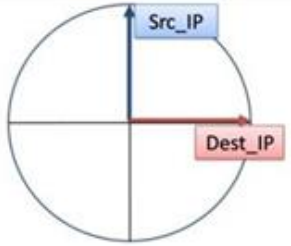




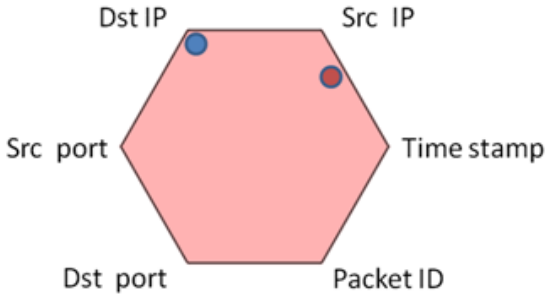








webpage_load(X) :-
same_src_ips(X),same_dest_ips(X),same_src_port(X,80),
timeframe_upper(X,10),length(X,L),greaterthan(L,8).



Project Suggestions

The 2015 VAST Challenge

http://vacommunity.org/VAST+Challenge+2015

Mayhem at DinoFun World

Find out what happened when a peaceful celebration in a small town turns into crime and mayhem perpetrated by a poor, misguided and disgruntled figure from the past

Challenges and Competitions

For VAST 2015

- select among two mini challenges and a grand challenge
- solving the grand challenge gives more prestige for the award

Kaggle competitions https://www.kaggle.com/competitions

- San Francisco Crime Classification
- Titanic: Machine Learning from Disaster
- Identify hand motions from EEG recordings (award $10k)
- Predict if context ads will earn a user's click (award $20k)
- Model quoted prices for industrial tube assemblies (award $30k)
- Identify signs of diabetic retinopathy in eye images (award $100k)
- and others on that site


PRESENTATION FINISHED
ANY QUESTIONS...