Toward Improved Visualization of Unstructured Information

March 4, 2005

National Academy of Sciences

Context 2

J. David Harris
National Security Agency

Context 2

The definition:

Visualization in the context of large, <u>unstructured</u>, <u>changing data</u> sets where the <u>relevance</u>, <u>significance</u>, and <u>conceptual links</u> among the data <u>have yet to be discovered</u>

<u>Preliminary thoughts</u>

Effective visualization of <u>structured data</u> is challenging

Unstructured data requires some type of structural mapping
. The data discovery and analysis will be imperfect
. The mapping will be imperfect, and task dependent

An Example of Real Data

A real-world story of intrigue…

About a cell of conspiring individuals…

Who set forth on a project…

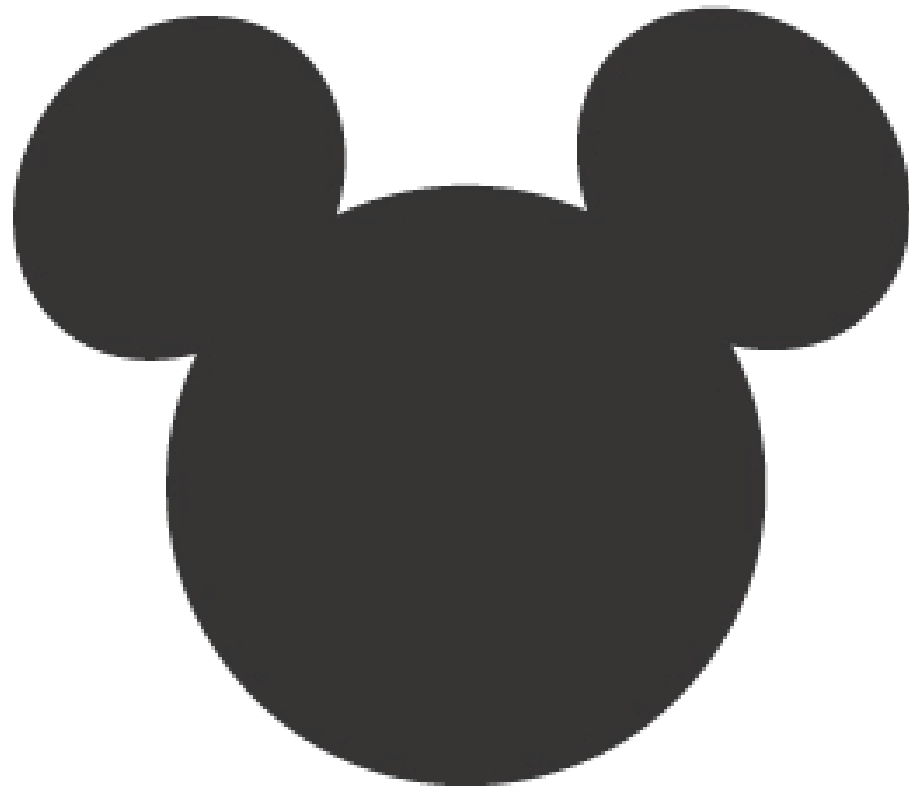   Constrained by time and money…

    Characterized by deception…

To advance their own cause.

The plot is ultimately discovered…

But not before the mission is accomplished.

The Players and Their Motivation

Toward Improved Visualization of Unstructured Information

The Strategy

Discovering "Project X"

Exploratory analysis…

is there some data that conforms to some pattern?

**START
PROJECT**

Walt Disney World
Company

**END
PROJECT**

Acquire Land for
Florida Property

The Investigation

Discovering "Project X"

Refining the pattern…but only some of the pattern is observable in the unstructured, varied, uncertain and continuously updated data

NOW

START PROJECT	ORGANIZE	EXECUTE	END PROJECT
Walt Disney World Company	Define Operation Establish Schedule Secure Finances Identify Purchase Agents Disguise Travel Patterns Establish Aliases for Agents Obfuscate Communications Establish Dummy Corporations	Purchase Land Disguise Travel Patterns Acquire Mineral Rights	Acquire Land for Florida Property

Retrospective analysis? Or prospective analysis?

The Observables

Establish aliases for purchase agents, and incorporate

Relevant observations, (semi-)notional:

	COMPANY NAME	POC	STATE	DATE	
■	(Compass East Group,	*Roy Davis*,	Delaware,	7 December 1964	)
■	(Tomahawk Properties, Inc.,	*Bob Price,*	Florida,	*January 1965*	)
■	(Latin-American Dev and Mgmt Corp.	*Roy Davis,*	Florida,	*February 1965*	)
■	(Ayefour Corporation,	*Bob Foster,*	Florida,	*February 1965*	)
■	(Bay Lake Properties, Inc.,	*Bob Price,*	Florida,	*March 1965*	)
■	(Reedy Creek Ranch, Inc.,	*M. T. Lott*,	Florida,	*June 1965*	)

BTW…These entities are uncertain!!

Incorporations

time———→

Important information is dominated by irrelevant data

There's very little evidence on which to base decisions…

More Observables

Purchase land

Relevant observations:

PURCHASING AGENT	TRANSACTION DATE	LONGITUDE	LATITUDE	ACRES

Land Acquisition

time⟶

Incorporations

time⟶

The plot begins to unfold…

Correlated events emerge (both within and across streams)

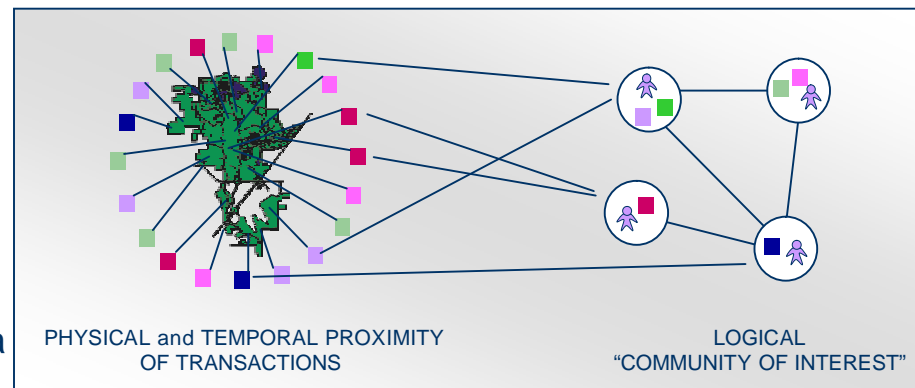# The Discovery

Orlando Sentinel

Dateline—May 4, 1965

Reported that two real estate transactions totaling over $1.5 million had been made…for nearly 9,000 acres of land near the small Florida farming town of Orlando

Dateline—October 20, 1965

Reported that Walt Disney was secretly behind the purchase of land

Challenges

Vagueness (Dynamism) of Hypotheses
Unknown Sources of Data and Information
Relevant Data Concealed by "Noise"
Uncertain and Erroneous Observations
(Causally) Incomplete Context…Missing Data
Logical and Physical Structures

PHYSICAL and TEMPORAL PROXIMITY OF TRANSACTIONS

LOGICAL "COMMUNITY OF INTEREST"

But It Was Too Late

Disney's "Project X"

Began in the early 1960's

The Florida site was selected on November 22, 1963

Ayefour Corporation buys the first parcel of land on October 23, 1964

An official announcement was made by Disney on November 15, 1965

They had acquired 27,443 acres of land SW of Orlando…

And they had "big plans"

What did it cost? About $185/acre, on average

The first acre: $80

The final acre: $80,000

Why is This Context Interesting?

The definition:

Visualization in the context of large, <u>unstructured</u>, <u>changing data</u> sets where the <u>relevance</u>, <u>significance</u>, and <u>conceptual links</u> among the data <u>have yet to be discovered</u>

To enable understanding!

<u>Retrospective</u>

- Forensics

<u>Prospective</u>

- Investigative Reporting
- Business Intelligence
- Security

The "Context 2" Agenda

Ronald Coifman

Yale University

Diffusion/Inference Geometries of Data Features,

Situational Awareness and Visualization

Andre Skupin

University of New Orleans

A Different Kind of Map

Stephen Eick

University of Illinois at Chicago; SSS Research

DECIDE^{TM} Hypothesis Visualization Tool

Dave Harris

National Security Agency

Reactions and discussion

Reactions and Discussion…

Context 2…the definition:

Visualization in the context of large, <u>unstructured</u>, <u>changing data</u> sets where the <u>relevance</u>, <u>significance</u>, and <u>conceptual links</u> among the data <u>have yet to be discovered</u>

Yesterday's Discussion...

Perception and cognition of visualization

Reasoning under uncertainty

Perception and Cognition of Visualization

Map-making

Simplification

What's important? Who's the intended audience?
How might we measure interpretability?

Classification

Symbolization

Induction

Visualize…

Existence

Notation on a map that a point or area exists

Associative existence

Added absolute or relative quantity to the identified points and areas

Spatially associated existence

Spatial relationships between points and areas

I'm willing to trade accuracy, resolution, completeness, etc. for improved perception

This representation of the Orlando metropolitan area is targeted at tourists…

Maps are a specific type of diagram with which most people have experience

Perception and Cognition of Visualization

How can we capture the dynamic nature of data?

    Maps are snapshots…but they require little additional training

Can we place thematic overlays on top of term-document landscapes?

    as a means of creating different views of the same data...

How do we encourage interactivity?

What can't be represented using topography only?

    For unstructured data…

        What kind of mappings can we impose?

        Some structure may be due to contact or context (not content)…

            What might roads represent?

            What about rest areas? National Parks? Hospitals?

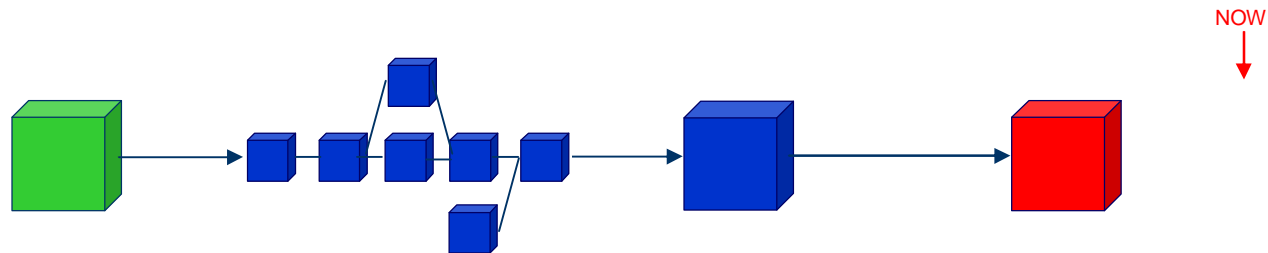How is uncertainty represented?

Reasoning Under Uncertainty

A critical aspect of Context 2

Visualization of the hypotheses…

Capture the intent of the task and subject matter expertise

Guide the exploration and analysis

Customize the visualization

NOW

Reasoning Under Uncertainty

Multiple (competing) hypotheses

Alternative models, at the onset…

or after improved/diminished understanding

Machine-learning can (should?) offer data as...
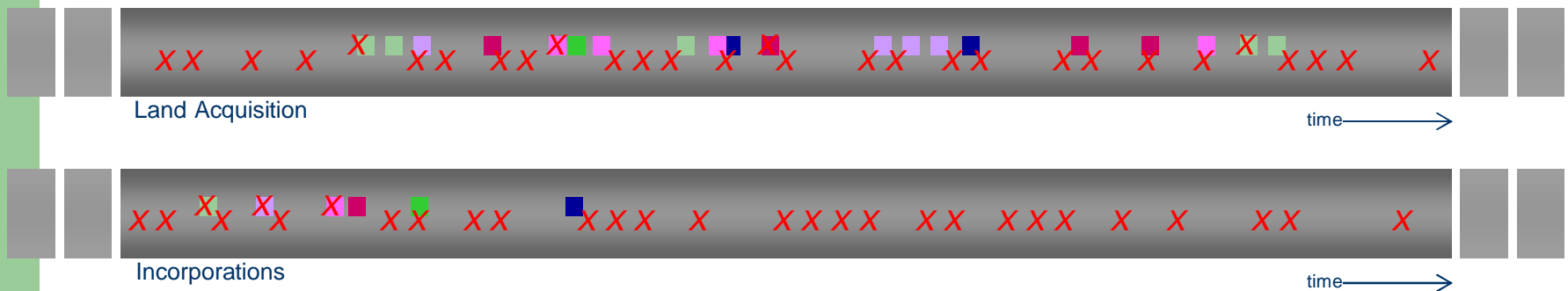
Supporting evidence

Contradictory evidence

Change in the actual plan

Changing world events

What to Visualize?

How do we decide what's important?

Land Acquisition

time ⟶

Incorporations

time ⟶

We (probably) don't need all of these observations?

Final Thoughts

So…

Visualization influences hypothesis generation

Hypothesis generation influences analysis

Analysis influences visualization

Reactions and Discussion…

Select relevant information from assembled data

Impose some kind of structure…

Apply graphic techniques

To enable <u>understanding</u>…

Challenges →

Vagueness (Dynamism) of Hypotheses
Unknown Sources of Data and Information
Relevant Data Concealed by "Noise"
Uncertain and Erroneous Observations
(Causally) Incomplete Context…Missing Data
Logical and Physical Structures