

# The Mythology of Big Data

O'Reilly Strata Conference  
February 2, 2011

Mark R. Madsen  
<http://ThirdNature.net>  
@markmadsen







**Every technology carries  
within itself the seeds  
of its own destruction.**





**Code is a commodity**

<http://www.flickr.com/photos/ecstaticist/1120119742/>

# What's the central myth underlying big data?



# The myth that drove the gold rush

All we need is a fat pipe and pans working in parallel...



*You change an org by acting with, through others, not alone.*



# Evolution of data

50s-60s: data as product

70s-80s: data as byproduct

90s-00s: data as asset

2010s +: data as substrate

*The real data revolution is in business structure and processes and how they use information.*



# Everything is so different now...

## A COMPUTER WANTED.

WASHINGTON, May 1.—A civil service examination will be held May 18 in Washington, and, if necessary, in other cities, to secure eligibles for the position of computer in the Nautical Almanac Office, where two vacancies exist—one at \$1,000, the other at \$1,400..

The examination will include the subjects of algebra, geometry, trigonometry, and astronomy. Application blanks may be obtained of the United States Civil Service Commission.

*Your grandmother, the data scientist.*

**Many current approaches miss the point**

**Using Big Data**



It's not about "big"

Using ~~Big~~ Data

And "big" is often not as big as you think it is.

**It's not really about data, either**

**Using ~~Big Data~~**

If there's no process for applying information in a specific context then you are producing expensive trivia.



# Where does the value in data come from?



For most of us in non-data businesses, this translates to *“How can we use information to improve the decisions made in our organization?”*

We need to focus on that singularly bad decision making entity, the group. Organizations seem to amplify innate decision making flaws.

# Decision-making realities

The operating model in senior management is primarily intuition and pattern-based.

The mode for middle management is political, bureaucratic.

New data is destabilizing, which is why you may hit a wall trying to push your data-driven agenda.

Data is contextual, so we need stories to explain how we think the world works, why my data is better than yours, and why your theory sucks.

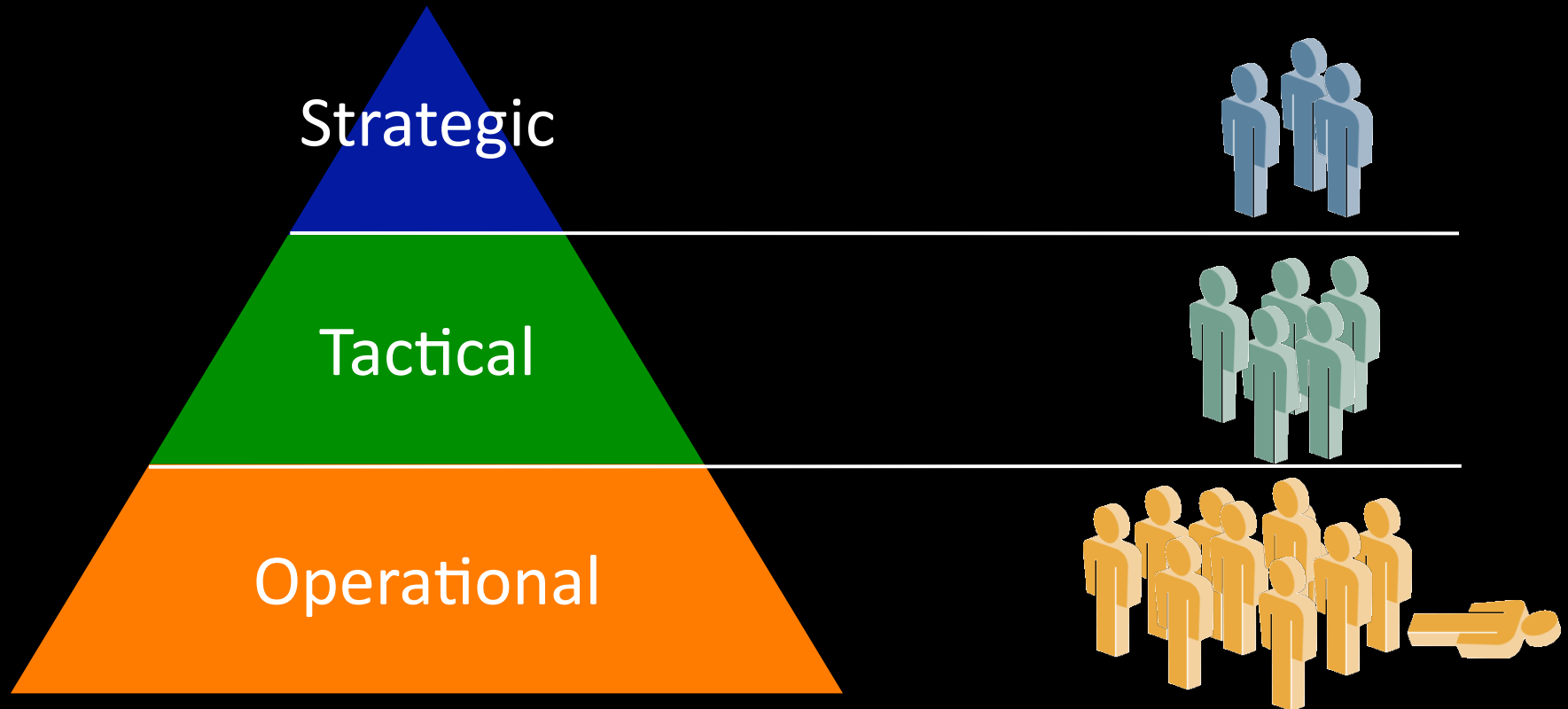
Cognitive bias creates a morass for interpretation.





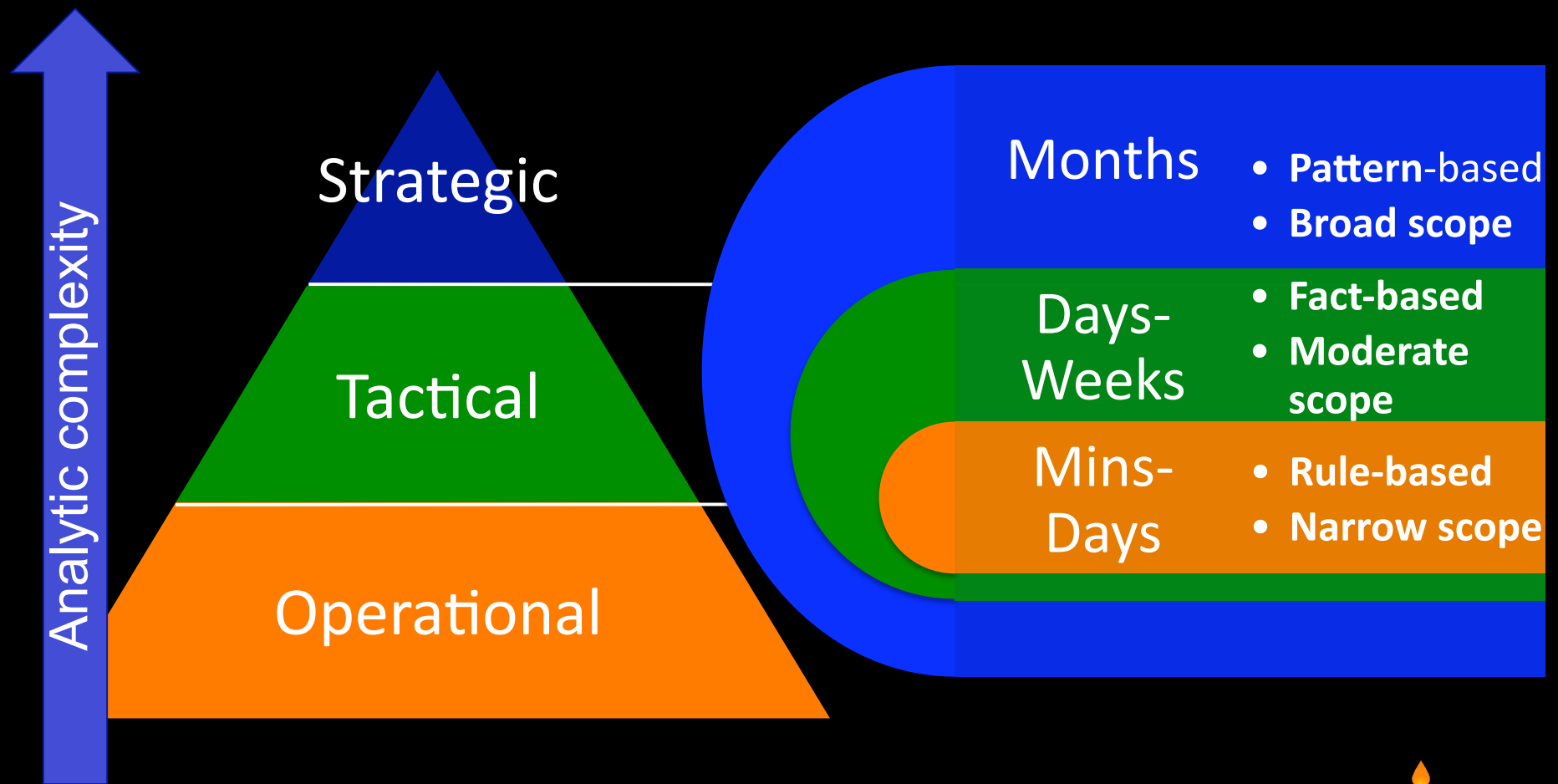
# A very abstract business intelligence model

Who are the people making decisions?



# What is the nature of their decisions?

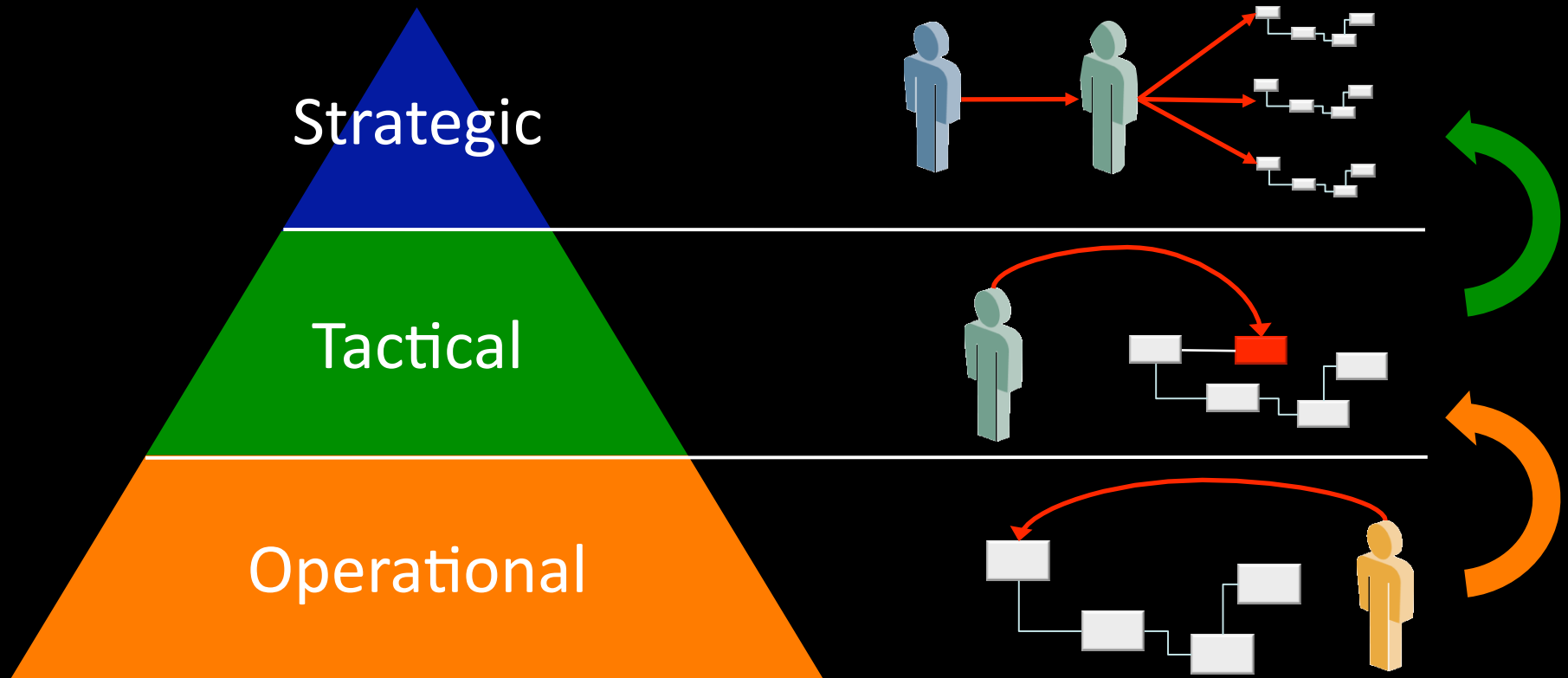
Scope, time frame of decision, time scale of data, data volume, breadth of data, frequency, pattern vs fact-based





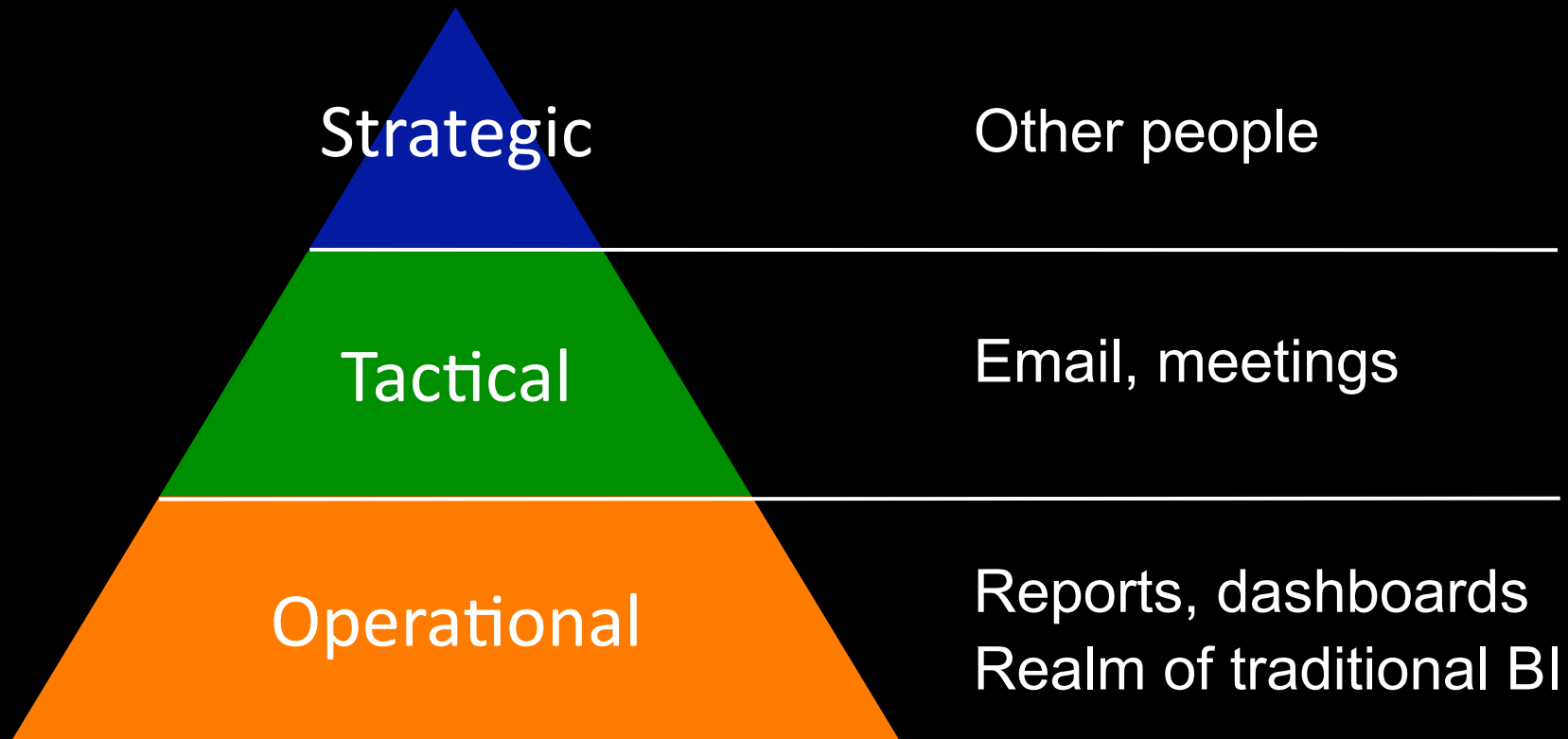
# The process aspect of decisions ties to people

Scope of control for people in most organizations aligns:  
in process, on process, over process



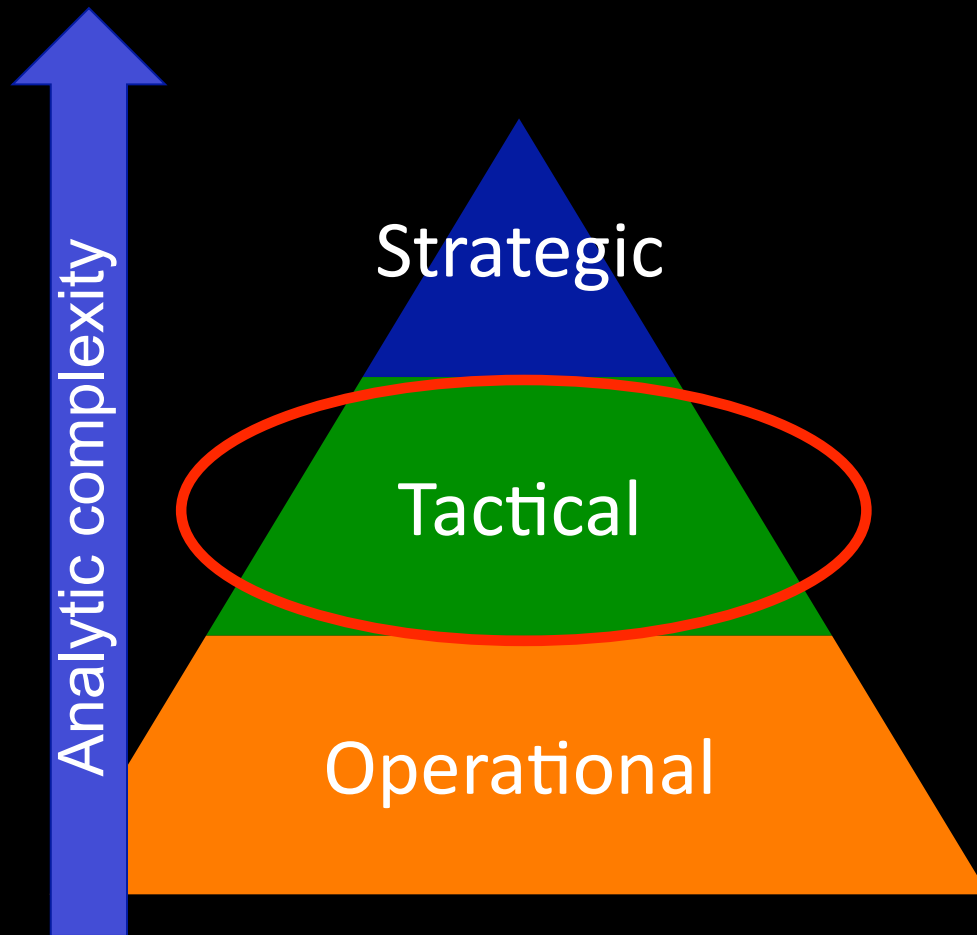
*The exceptions not handled at one level due to rule / procedure / policy deficiency are escalated to the next.*

# What kind of support do they have today?



Reality of most reports and dashboards is that they provide basic monitoring at best.

# How and where can you apply data solutions?



High single value, less frequent, so improve the effectiveness of individual decisions.

*Fuzzy middle ground*

Low single value, frequent, can improve the efficiency or the effectiveness for large aggregate improvement.



# What do people do with data?

1. **Describe:** use data to characterize a current or prior state of the system, for example monitoring and identifying exceptions
2. **Investigate:** explore data to discover the boundaries and characteristics of a system, frame a problem or find supporting / discrediting evidence.
3. **Explain:** use data and analytic methods to determine causes and effects, build models and construct stories.
4. **Predict:** apply analytic models to determine possible / probable future states of the system
5. **Prescribe:** use data in models to define policy, procedure, and rules for taking action, and possibly automate them

*Data infrastructure and tool support for these activities in most organizations is uneven at best, decreasing as you move down.*

***If you want to be a data scientist, or build software to support them, read this paper***

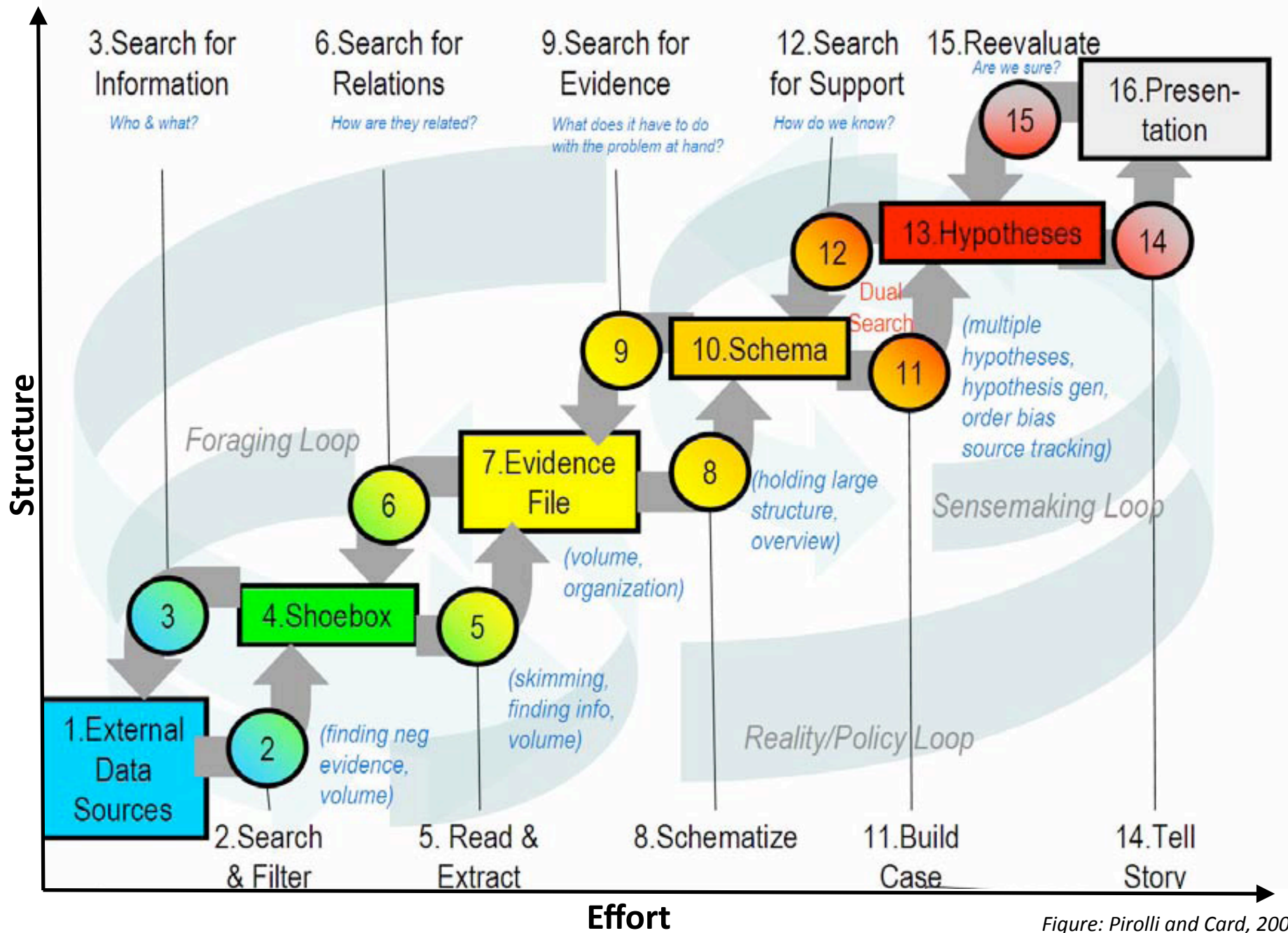


Figure: Pirolli and Card, 2005

“A toolmaker succeeds as, and only as, the *users* of his tools succeed with his aid. However shining the blade, however jeweled the hilt, however perfect the heft, a sword is tested only by cutting. That swordsmith is successful whose clients die of old age.”

*Frederick Brooks*



## About the Presenter



Mark Madsen is president of Third Nature, a technology research and consulting firm focused on business intelligence, analytics and performance management. Mark is an award-winning author, architect and former CTO whose work has been featured in numerous industry publications. During his career Mark received awards from the American Productivity & Quality Center, TDWI, Computerworld and the Smithsonian Institute. He is an international speaker, contributing editor at Intelligent Enterprise, and manages the open source channel at the Business Intelligence Network. For more information or to contact Mark, visit <http://ThirdNature.net>.

## About Third Nature



Third Nature is a research and consulting firm focused on new and emerging technology and practices in business intelligence, data integration and information management. If your question is related to BI, open source, web 2.0 or data integration then you're at the right place.

Our goal is to help companies take advantage of information-driven management practices and applications. We offer education, consulting and research services to support business and IT organizations as well as technology vendors.

We fill the gap between what the industry analyst firms cover and what IT needs. We specialize in product and technology analysis, so we look at emerging technologies and markets, evaluating the products rather than vendor market positions.