



attensity

## **Text Mining**

**Responding to Trends, Improving Productivity  
and Better Allocating Resources through  
Analysis of Structured and Unstructured Data**

**David Bean, PhD**



attensity

## **Text Mining**

**Why knowledge engineering is the biggest  
impediment to successful text mining.**

**David Bean, PhD**



attensity

## **Text Mining**

**Why Knowledge Engineering (KE) is NOT your friend.**

**David Bean, PhD**



attensity

## **Text Mining**

**How I learned to stop worrying and love KE.**

**Or not.**

**David Bean, PhD**



attensity

# **Text Mining**

**Eliminating Traditional Knowledge Engineering**

**David Bean, PhD**

# The Text Analysis Spectrum

**Conceptual**

*statistical methods*

*linguistic methods*

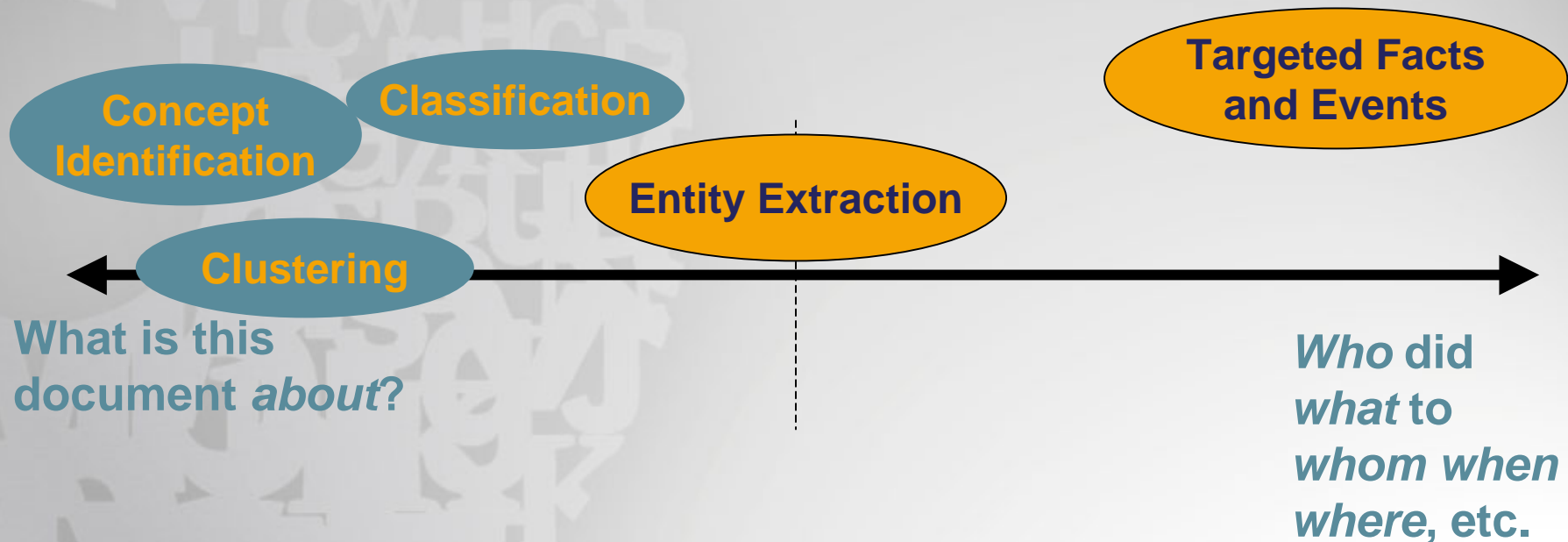
**Relational,  
Episodic,  
Tactical**

*What is this  
document about?*

*Who did  
what to  
whom when  
where, etc.*

# The Text Analysis Spectrum

Where does Attensity play?



# Why is getting dimensional data so hard?

## Predefinition!

Hank bought plastic explosives from Henry in Tucson yesterday.

### Named Entity Extraction

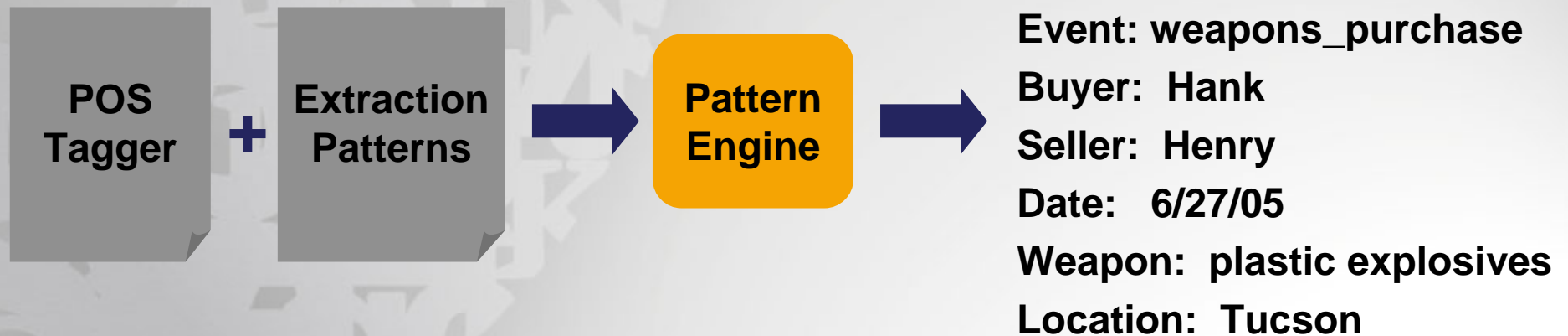


# Why is getting dimensional data so hard?

## Predefinition!

Hank bought plastic explosives from Henry in Tucson yesterday.

### Event Extraction



# Predefinition

**No big deal, we can reuse these the knowledge libraries, right?**

**Domain specificity to the rescue!**

**Except...the real world is messy and customer expectations are high.**

# Today's Text Extraction Process

**What to extract**

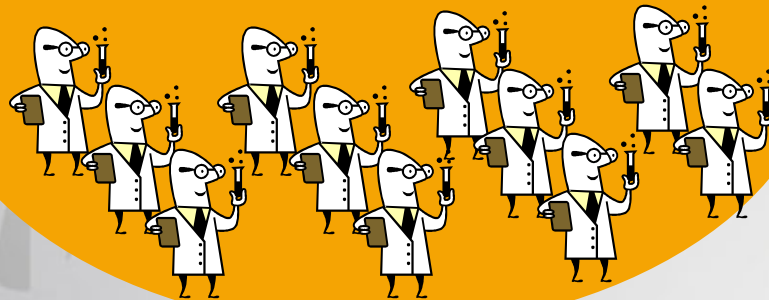
**What is important to the task?**

*You must understand  
the business problem.*

**How to extract**

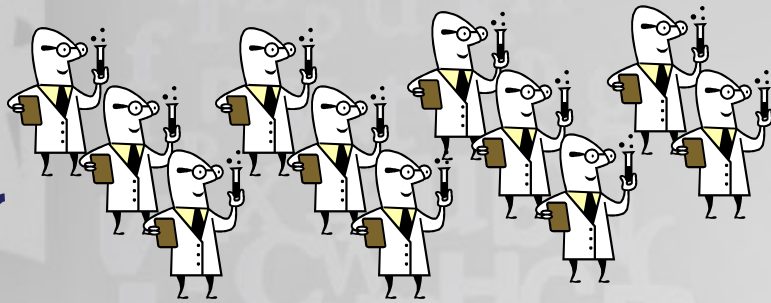
**Tuning/configuring the system.**

*You must understand  
the extraction mechanism,  
its supporting tools, data.*



**Extraction Engine**

# Making the “How” easy



**Machine Learning**

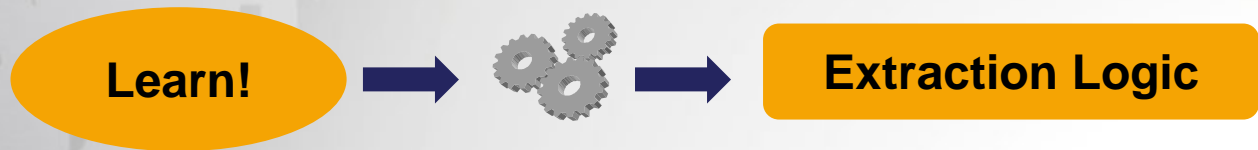


# Making the “How” easy



- Identify what to extract.
- System starts taking guesses.
- User corrects system.
- System learns how to extract.

Event Type	Buyer	Seller	Location	Weapon	
buy	Hank		Henry	Tucson	Correct / Incorrect
buy	Henry	beat			Correct / Incorrect
buy	Hank	Henry	Tucson	plastic explosives	Correct / Incorrect



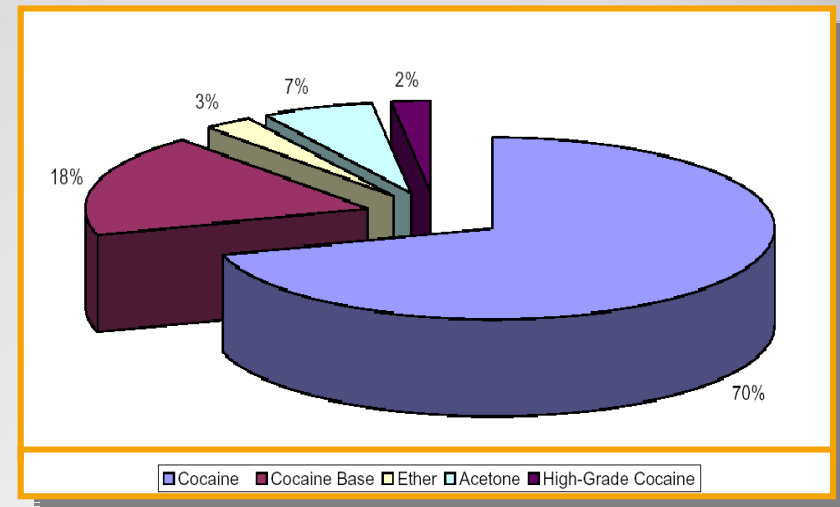
# Attensity Workstation Scenario

## Question:

Of the drug related contraband reported seized in Latin America over the last 6 months, what substance accounts for the largest share by weight, and how do the next largest shares compare?



information resources



answer

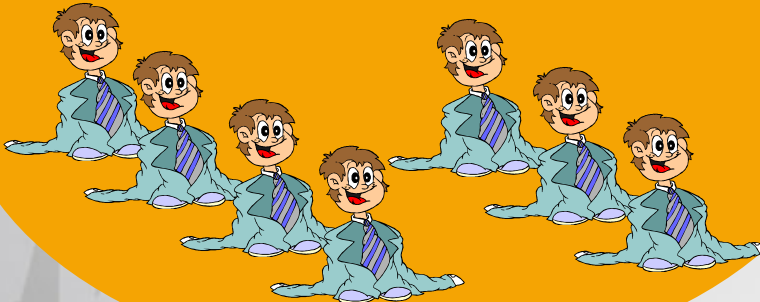
# A New Approach to Text Extraction

**What to extract**

**What is important to the task?**

*You must understand  
the business problem.*

**How to extract**



**Tuning/configuring the system.**

~~*You must understand  
the extraction mechanism,  
its supporting tools, data.*~~

**Extraction Engine**

# Eliminating Traditional KE

with Directed Learning Technology

- Generates tactical information – who, what, why, when, etc.
- No need to understand the “how”
- Best done by the domain expert themselves
- Suitable for ad hoc extraction problems
- My mother’s worst nightmare: proper grammar not required
- Language independent



attensity

# **Text Mining**

**Eliminating Traditional Knowledge Engineering**

**David Bean, PhD**