NORTH ATLANTIC TREATY ORGANIZATION

SCIENCE AND TECHNOLOGY ORGANIZATION

AC/323()

www.sto.nato.int

STO TECHNICAL REPORT

PUB REF STO-MP-SAS-114-PPJ

ANNEX J

Measures of Information Usefulness in Target Classification

Mark Timms

Defence Research and Development Canada
Recherche et développement pour la défense Canada

Finding Useful Questions:

Examining measures of information usefulness in Target Classification (exploring Nelson, 2005).

Mark Timms, DRDC Toronto

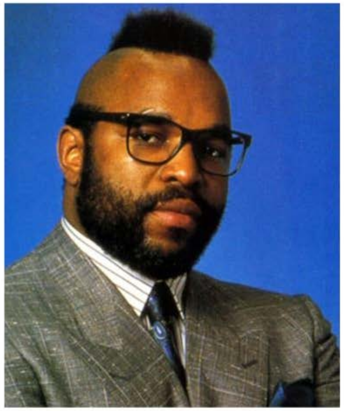
DRDC | RDDC

Canada

Agenda

- Background
- Nelson (2005)
- Discussion
- Next Steps at DRDC

Falsification



VS

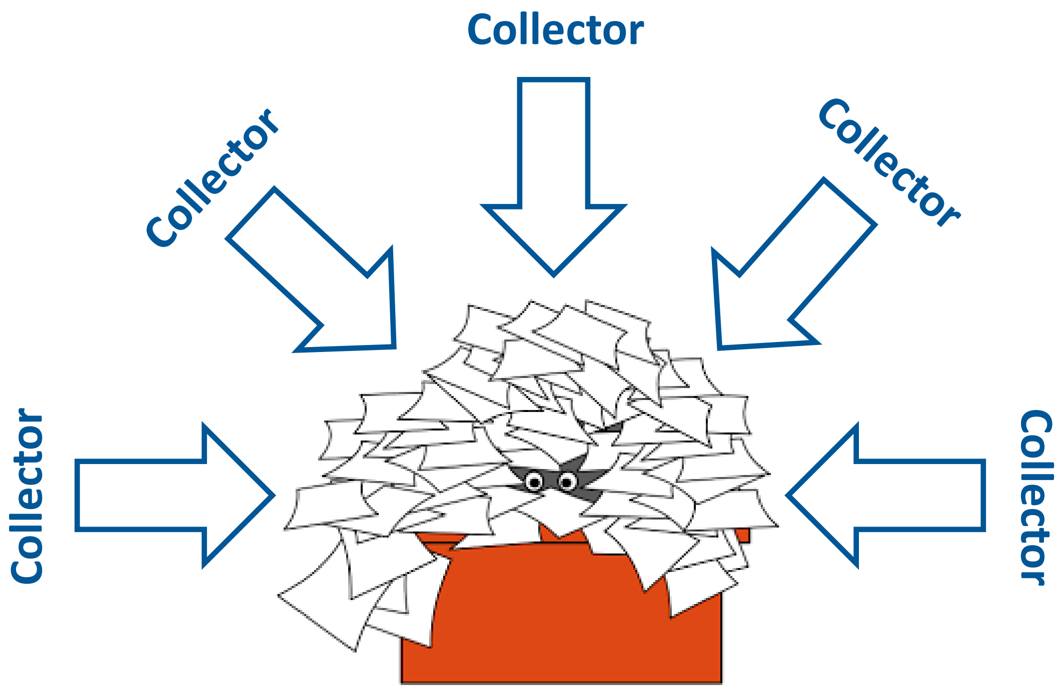
Differentiation

BLUF:

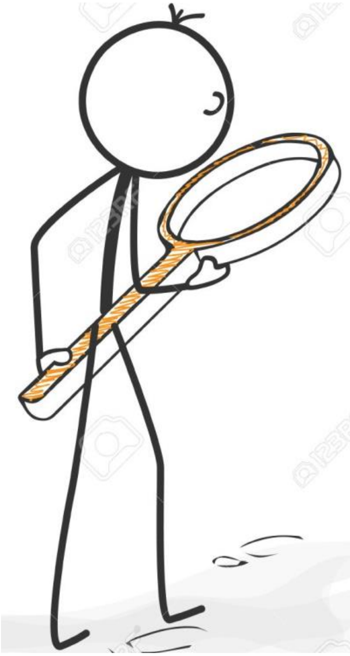
- Differentiation > Falsification
- Falsification ≠ Human Decision Making
- – Cost Structure → Dangerous

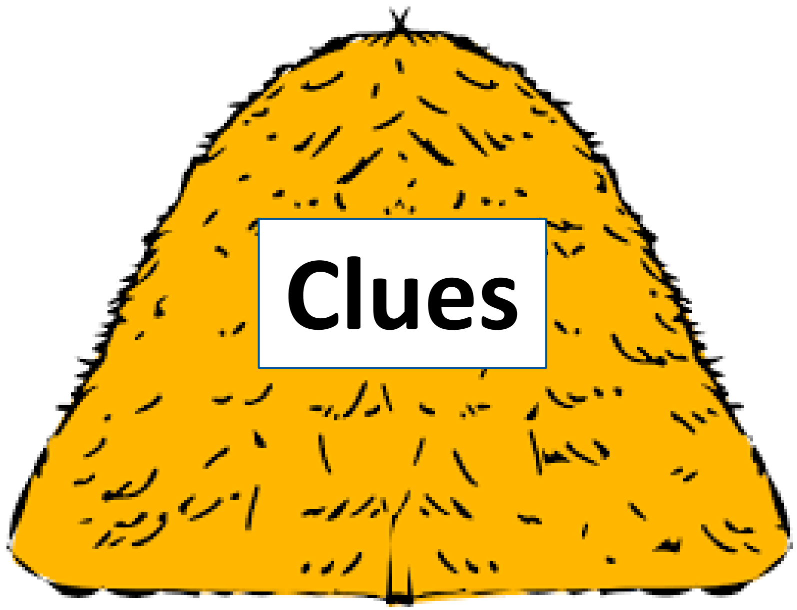
Why do we care?

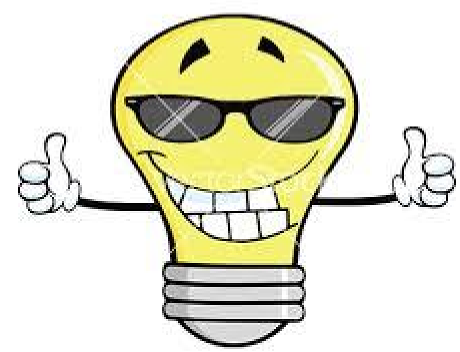
- Intelligence Analyst → **Information Triage**
- The Intelligence Cycle → **Priority Intelligence Requirements**
- Contentious Policing Policies → **Carding**



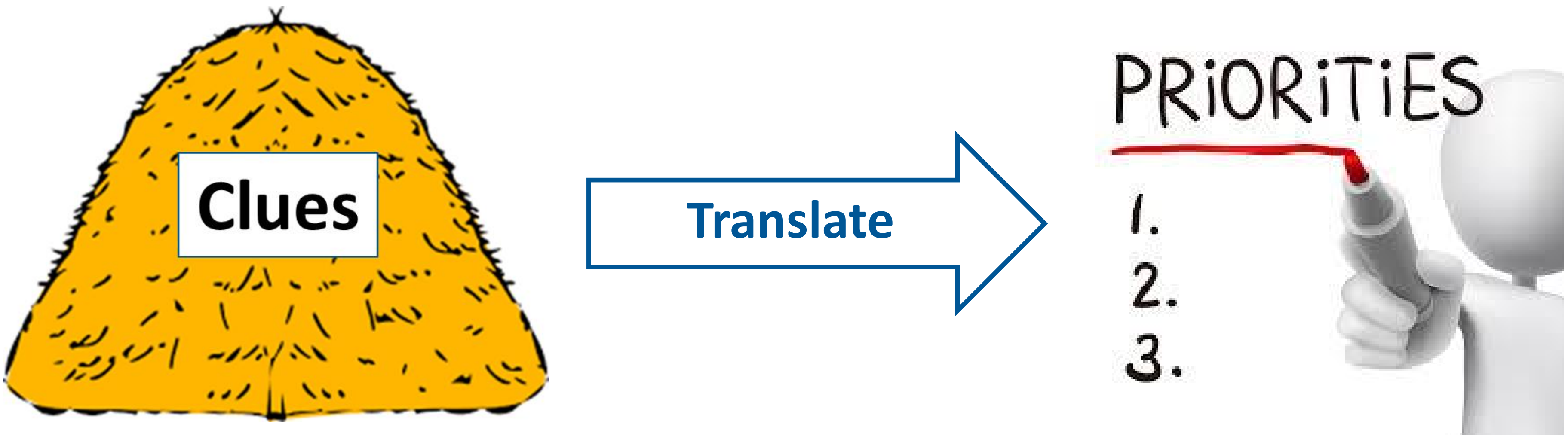
The Basics: How to solve a mystery







The Basics: Deriving Utility → Prioritization



The Basics: the probabilistic framework

- **Probabilistic belief model, including**:
  - Hypotheses
  - Prior probability for each
  - A list of possible questions

- **A sampling Norm (**the utility function**)**

- **A method to update beliefs (**compliments the utility function**)**

Example: Planet Vuma – Glom, or Fizzo?

Tribe	Pop	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	c11	c12
A	0.5	0.1	0.75	0.5	0.25	0.1	0.75	0.5	0.25	0.75	0.25	0.75	0.75
B	0.5	0.1	0.5	0.75	0.25	0.1	0.5	0.75	0.25	0.25	0.75	0.5	0.5

Skov and Sherman (1986) and Slowiaczek et al. (1992, Experiments 3a and 3b).

Previous Research

Measures info gathering behaviour by considering **whether people choose to ask highly useful questions**.

DOES NOT clearly outline:

- what norms best describe human behavior,
- when the norms disagree, or
- whether some norms are better than others.

Present Research

- **Illuminates differences among Sampling Norms**, by holding constant both probability belief models and the method of updating.

- **This study**:
  - Defines/explains six sampling norms
  - **Re-examines experimental evidence-acquisition research** to determine if their conclusions were **influenced by sampling norms** applied.
  - **Explores sampling norm disagreement through novel simulations**.

↑ Falsification

↓ Differentiation

Model Summary

- **Bayesian Diagnosticity**
- Log Diagnosticity

$$diagnosticity\ (q_j) = \max\left(\frac{P(q_j|c_1)}{P(q_j|c_0)}\ ,\ \frac{P(q_j|c_0)}{P(q_j|c_1)}\right) \quad \text{and}$$

$$diagnosticity\ (Q) = \sum_{q_j} P(q_j) \times diagnosticity(q_j)$$

- **Information Gain**
- Kullback-Leiber Distance
- Probability Gain
- Impact

$$I(C,Q) = H(C) - H(C|Q), where:$$

$$H(C) = \sum_{C_i} P(C_i) \times\ log_2 \frac{1}{P(C_i)}, the\ initial\ entropy\ in\ C$$

$$H(C|Q) = \sum_{q_j} P(q_j) \times H(C|q_j), the\ conditional\ entropy\ in\ C\ given\ Q, and$$

$$H(C|q_j) = \sum_{C_i} P(C_i|q_j) \times H(C|q_j) \times\ log_2 \frac{1}{P(C_i|q_j)}, the\ entropy\ in\ C\ given\ answer\ q_j$$

Differentiation Explained – Information Gain Example

Predicting whether I will play golf

(Play) 9/14 = YES → iE = 0.94

outlook	temperature	humidity	windy	play
overcast	cool	normal	TRUE	yes
overcast	hot	high	FALSE	yes
overcast	hot	normal	FALSE	yes
overcast	mild	high	TRUE	yes
rainy	cool	normal	TRUE	no
rainy	mild	high	TRUE	no
rainy	cool	normal	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
sunny	mild	normal	TRUE	yes

(Oc) 4/4 = YES → cE = 0

(Rain) 3/5 = YES → cE = 0.97

(Sun) 2/5 = YES → cE = 0.97

Initial Entropy – Conditional Entropy = Information Gained

Results

Abridged from original version, as presented in Nelson 2005

Calculating the Utility of Cues → prioritization

66 Subjects, 18 Features, Probabilities Constant

Table 3

New Analysis of Features Used by Skov and Sherman (1986) and Slowiaczek et al. (1992)

Feature	% of gloms, fizos with each feature	Information gain, Kullback–Liebler distance	Probability gain, impact	Diagnosticity	Log_{10} diagnosticity
		Low usefulness			
A, B	48, 52; 52, 48	0.001	.020	1.083	0.035
C, D, E, F	28, 32; 32, 28; 68, 72; 72, 68	0.001	.020	1.084	0.035
		Medium usefulness			
G, H, I, J	15, 45; 45, 15; 55, 85; 85, 55	0.080	.150	1.982	0.276
K, L	34, 66; 66, 34	0.075	.160	1.941	0.288
		High usefulness			
M, N, O, P	10, 50; 50, 10; 50, 90; 90, 50	0.147	.200	2.760	0.388
Q, R	26, 74; 74, 26	0.173	.240	2.846	0.454

Disagreement Strength (DStr)

1,000, 000 Trials, up to 20 Features, Probabilities Vary

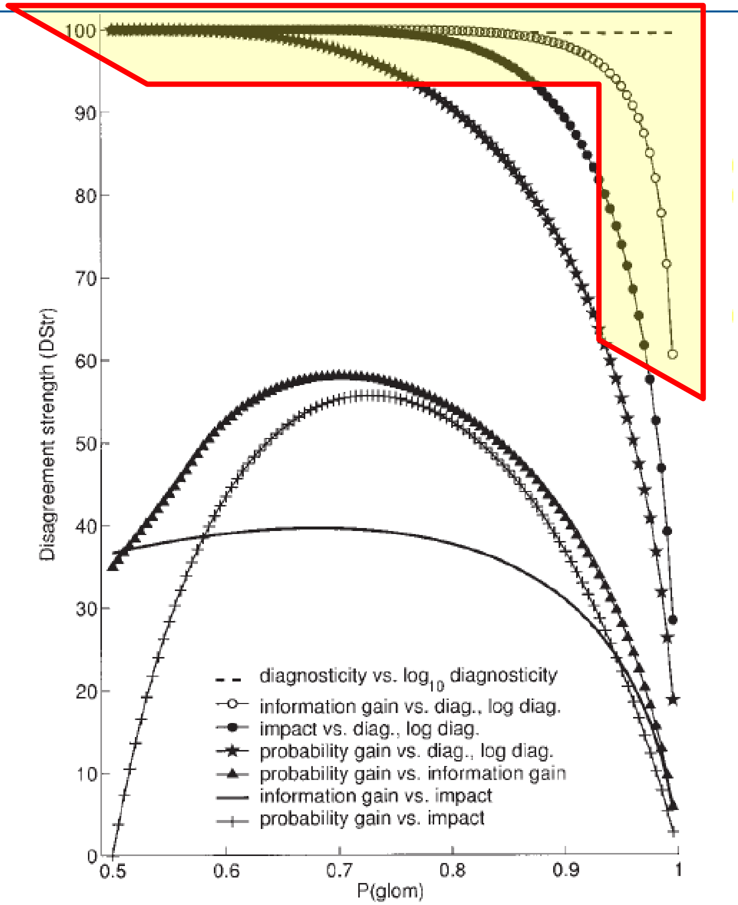


Figure 1. Simulation 2 results: strength of limiting cases of disagreement between pairs of sampling norms for different prior probabilities. diag. = diagnosticity; glom = type of creature on planet Vuma.

Model == Model?

Diagnosticity → almost universally suboptimal

Where Diagnosticity and Log Diagnosticity = Δ, Log Diagnosticity agreed with other norms, but was still suboptimal.

Model == Humans?

No subjects responded according to diagnosticity or log diagnosticity

Differentiation > Falsification

The Dangers of Diagnosticity

Guessing Gender, n=500 (%51 =Male), five features

Table 13
Natural Environment Feature Distribution

	Skirt (or dress)		Glasses			Beard		Earrings		Hair	
Gender	No	Yes	None	Sun	Yes	Yes	No	Yes	No	Short	Long
% of males	100	0	67	6	27	16	84	2	98	93	7
% of females	98	2	83	3	14	0	100	47	53	47	53

Note. About 51% of individuals were male. When hair completely obscured the ears, the ears were classified as not having earrings.

Table 14
Natural Environment Feature Usefulness Values

Sampling norm	Skirt (or dress)	Glasses	Beard	Earrings	Hair
Information gain, Kullback–Liebler distance	0.010	0.025	0.084	0.235	0.634
Probability gain	.010	.065	.062	.220	.420
Impact	0.010	0.080	0.080	0.225	0.430
Diagnosticity	Infinite	1.412	Infinite	7.056	13.296
Log_{10} diagnosticity	Infinite	0.093	Infinite	0.532	1.123

Findings, Discussion, and Next Steps

Findings & Discussion

- **Questioning Status Quo →** Falsification Models lag behind competition:
  - They disregard prior probabilities when symmetric;
  - Oversensitive to extreme scores; and
  - Inferior categorization with high number of hypotheses or categories.

- **Strength in diversity →** In many evidence-gathering situations, more than one sampling norm might reasonably apply.

- **Predicting Human Choices →** People's choices are not consistent with so-called falsification measures.

Next Steps at DRDC

- ## Mandel Tribe Scenario:
  - Comparison of sampling norms to a new problem set.
  - Assess model fit & DStr against human participants

- ## Integrating Machine Learning Models:
  - Reinforcement Learning
  - Latent Learning
  - Q-Learning
  - **SARSA (example)**

$$\Delta Q_t = \alpha(r_t - Q_t(s,a) + Q_{t+1}(s',a'))$$

DRDC | RDDC

SCIENCE, TECHNOLOGY AND KNOWLEDGE
FOR CANADA'S DEFENCE AND SECURITY

SCIENCE, TECHNOLOGIE ET SAVOIR
POUR LA DÉFENSE ET LA SÉCURITÉ DU CANADA