Enforcing coherence in table understanding
Ana Costa e Silva

1. MOTIVATION

Worker with too many files to read

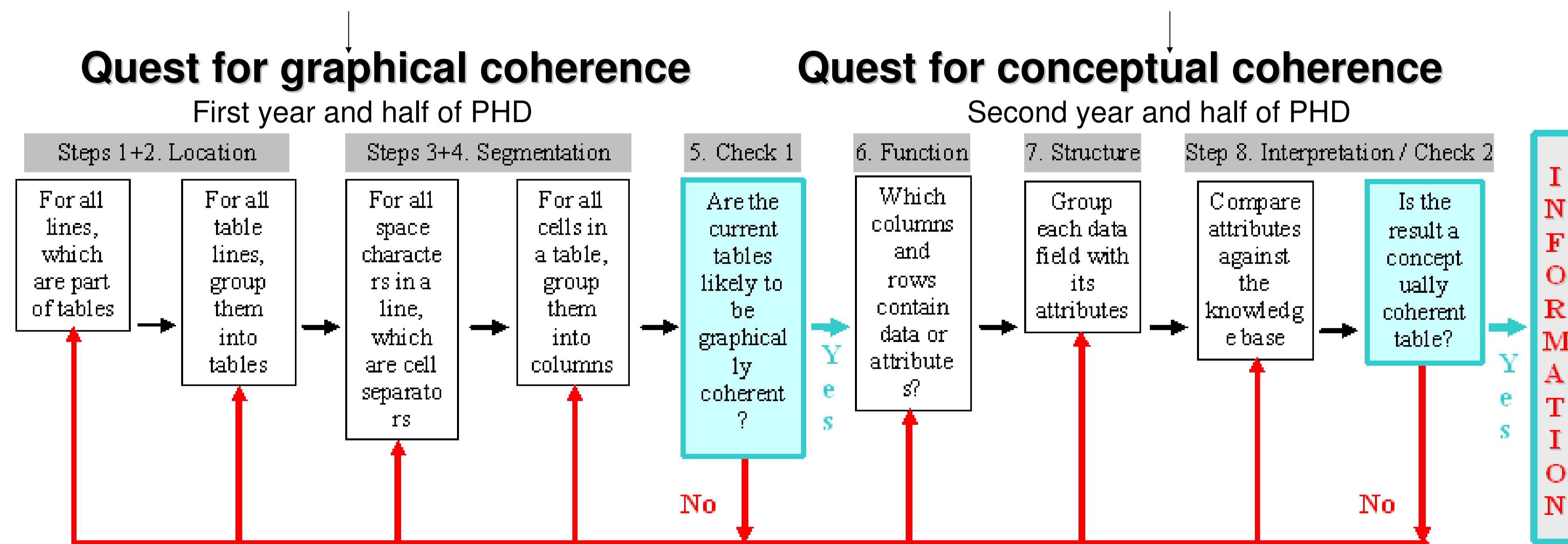More and more common with the rate of data growth in the planet

But the gist of many professional / web documents is often in tables

2. RESEARCH GOAL

Automatically extracting information from tables in unstructured sources into databases OR tagging data in on-line tables with its relevant context will help workers cope with a flood of data AND will render data collection generally more efficient

3. PROCESS

A TABLE IS A GRID-LIKE STRUCTURE WHERE ITEMS ARE ALIGNED AND THIS GRAPHICAL COHERENCE CARRIES WITH IT A CONCEPTUAL COHERENCE

Quest for graphical coherence
First year and half of PHD

Quest for conceptual coherence
Second year and half of PHD

Steps 1+2. Location		Steps 3+4. Segmentation		5. Check 1	6. Function	7. Structure	Step 8. Interpretation / Check 2	
For all lines, which are part of tables	For all table lines, group them into tables	For all space characters in a line, which are cell separators	For all cells in a table, group them into columns	Are the current tables likely to be graphically coherent?	Which columns and rows contain data or attributes?	Group each data field with its attributes	Compare attributes against the knowledge base	Is the result a conceptually coherent table?

Yes → No

Yes → No

INFORMATION

As important as how each step performs individually is how they interact.

4. RESULTS SO FAR

	Complete-ness	Purity	F-Measure
Step 1+2	98%	2%	90%
Step 3	100%	100%	100%
Step 4	84%	90%	76%

5. SO WHAT'S NEW?

- Optimising the balance between tractability AND deferral of decisions to as late as possible by maintaining multiple hypotheses probabilistically and pruning out unlikely ones

- Dynamically integrating context knowledge representations to control the formation of the table and select between alternative models

THE UNIVERSITY OF EDINBURGH

Microsoft Research