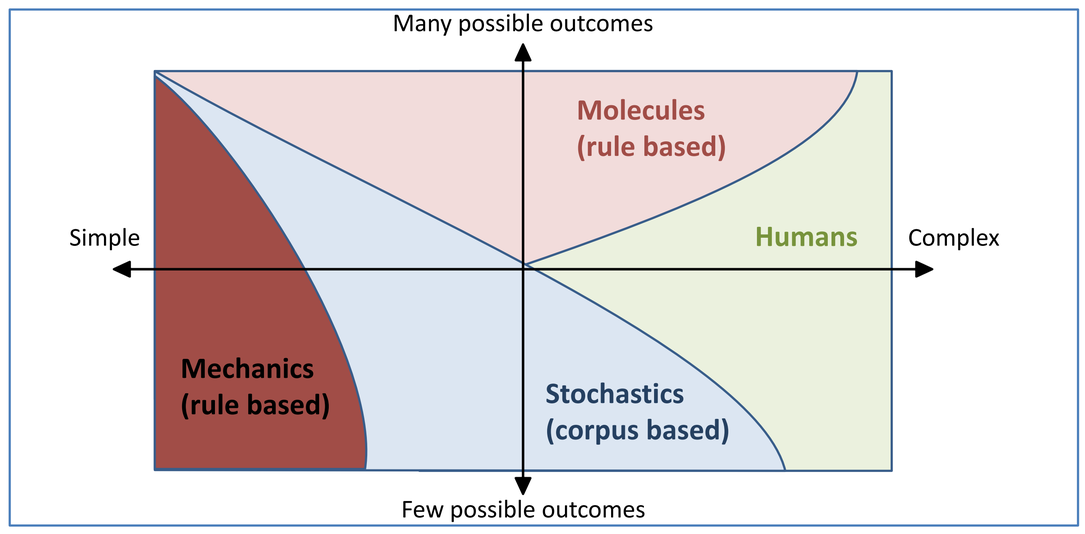
Concept Molecules as Basis for Text Analytics and Its Comparison to Corpus Based Methods

Michael Owsijewitsch, Hans-Jörg Schumann, Jörg Niggemann, Hans Rudolf Straub
Semfinder AG, a 3M Company, Kreuzlingen, Switzerland www.3M.de/HIS www.360Encompass.de hstraub@mmm.com

Challenge

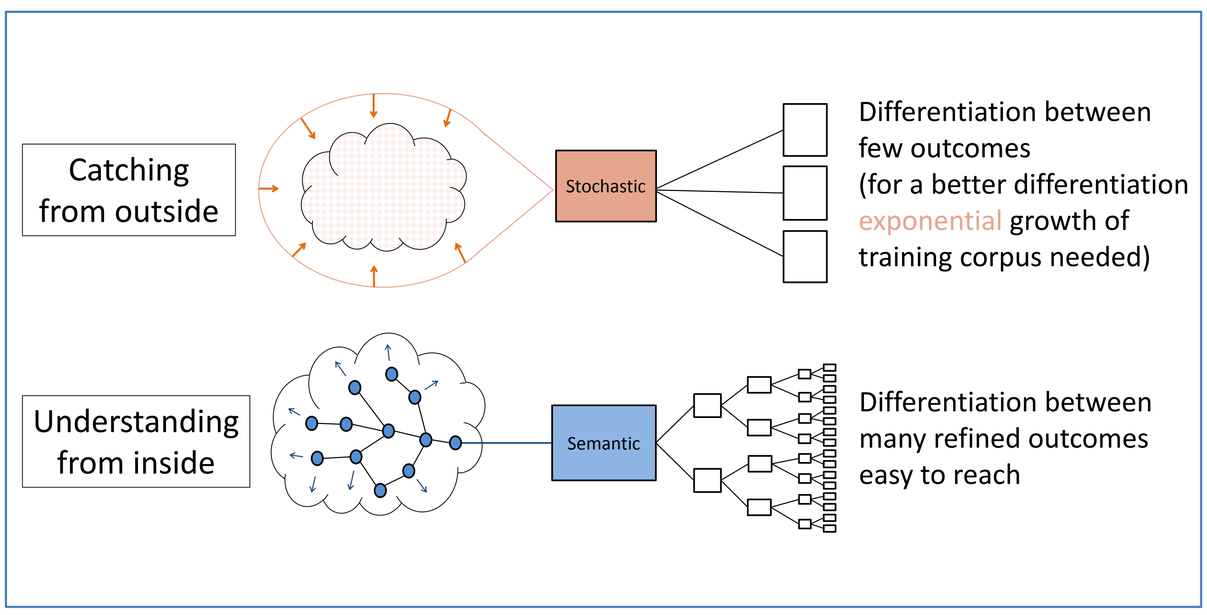
Unstructured information represents the largest and most relevant source of information in hospitals.
Although stochastic systems achieved reasonable results in the past years,
They have difficulties to handle situations with a lot of possible outcomes, as found in diagnosis and procedures coding and billing.

Which strategy for unstructured data?



Stochastics and Semantics

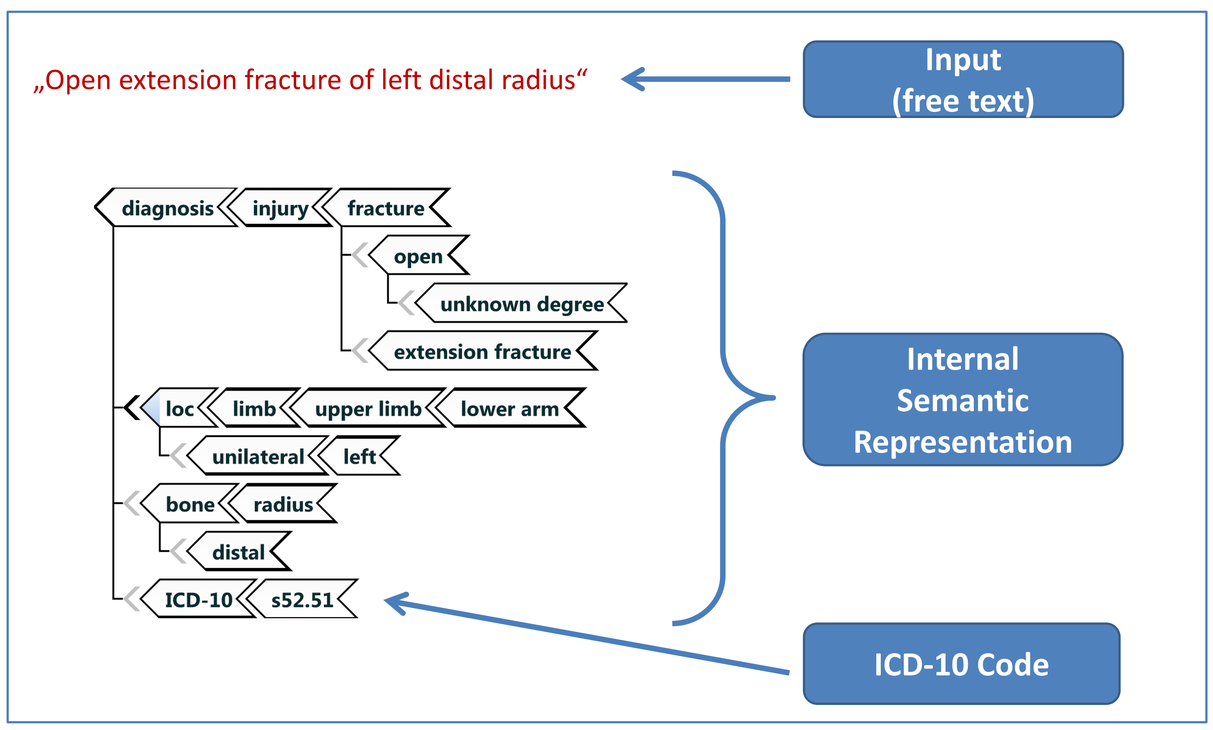
Semantic and stochastic methods are complementary in nature.



	Stochastics	Semantics
Learning Phase	Long (much feedback learning)	Long (much expert work)
Recall	High	Very high
Noise	Robust	Sensitive
Precision	Medium to high	Very high
Outcomes	Few, simple	Detailed, rich
Multilinguality	Completely new learning phase needed	Semantic relations are unchanged → easier *(still vocabulary issues)*
Process transparency for maintenance / fine tuning	Black Box – No explicit knowledge	**Transparent rules** – All knowledge explicit
Further processing (apart from coding)	Needs prior work for further interoperability	Ready for: → Semantic Data Repository → Alerts, Proposals in Clinics → Clinical Epidemiology …

Concept Molecules

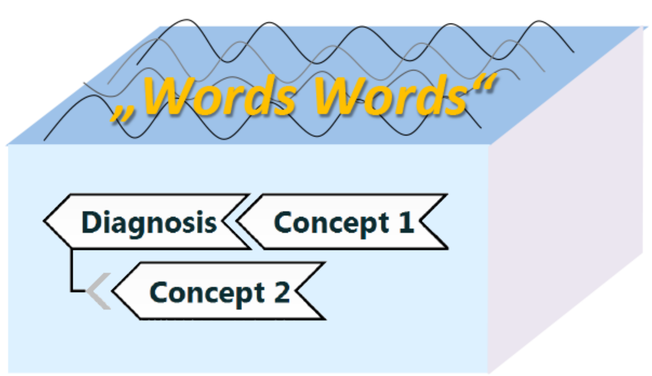
Semfinder concept molecules (CMs) represent the semantics. Free text inputs are automatically processed to CMs. CMs are built of atomic concepts, arranged in a structure which represents the relations between the atomic concepts.
The actual coding is derived from the explicit and implicit information represented in the molecule.



Noun phrase ≠ diagnosis phrase (solved by Concept Molecules)	
Linguistics (Input phrase)	Semantics (Interpretation)
Adeno-CA, Colon	diagnosis: neoplasia: carcinoma: adenocarcinoma; dignity: malignant; localisation: loc: intestine: large intestine: colon; code: ICD-10: c18.9 — 1 molecule, 1 diagnosis, 1 ICD-10 code
Adeno-CA, Tinnitus	diagnosis: neoplasia: carcinoma: adenocarcinoma; dignity: malignant; localisation: loc; code: ICD-10: c80.9 — diagnosis: tinnitus; localisation: loc; code: ICD-10: h93.1 — 2 molecules, 2 diagnoses, 2 ICD-10 codes
Linguistically no difference	Semantically a clear difference

Semantics: words ≠ concepts

Words flow on the surface –
Semantics (meaning) is found in profundity



Language is the outside:
- what you hear and see

Semantics is the inside:
- what you think

The semantic core of Semfinder is language independent:
→ reusability of knowledge bases in different languages
→ interoperability between languages

Semantics can help to “understand “

→ Overlappings (streptococcal pneumonia, postoperative)
→ Ambiguities (heads in the shoulders and abdomen)
→ Negations (diabetes, non-insulin-dependent, with complications)
→ Non-information (diabetes, not otherwise specified)
→ Implications (radius → bone, forearm)
→ Omissions (fracture of humerus and radius)
→ Composite Diagnoses with mutual dependencies

Semfinder

3M Science. Applied to Life.™