Applied clinical NLP: System Development Life Cycle

Olga V. Patterson

VA Informatics and Computing Infrastructure

Acknowledgements

- Affiliations
  - VA Informatics and Computing Infrastructure (VINCI)
  - Veterans Affairs Salt Lake City Health Care System
  - Department of Epidemiology, University of Utah
- Funding Support
  - VA Informatics and Computing Infrastructure VA HSR RES 13-457
  - Financial Relationships
    - Research grants from the following for-profit organizations: Amgen Inc., AbbVie Inc., Anolinx LLC, AstraZeneca Pharmaceuticals LP, F. Hoffmann-La Roche Ltd, Genentech Inc., Genomic Health, Inc., Gilead Sciences Inc., HITEKS Solutions Inc., LexisNexis Risk Solutions, Merck & Co., Inc., Mylan Specialty LP, Northrop Grumman Information Systems, Novartis International AG, PAREXEL International Corporation, and Shire PLC through the University of Utah or Western Institute for Biomedical Research.
    - Research funding from the following federal and non-profit organizations: Agency for Healthcare Research and Quality, Brigham and Women's Hospital, Centers for Disease Control and Prevention, Department of Defense, Department of Veterans Affairs, Intermountain Healthcare, National Heart, Lung, and Blood Institute, National Institute on Alcohol Abuse and Alcoholism, National Institute of Arthritis and Musculoskeletal and Skin Diseases, National Institute of General Medical Sciences, National Institute of Standards and Technology, National Library of Medicine, National Science Foundation, Patient Centered Outcomes Research Institute, and RAND Corporation.

Research areas in Clinical domains

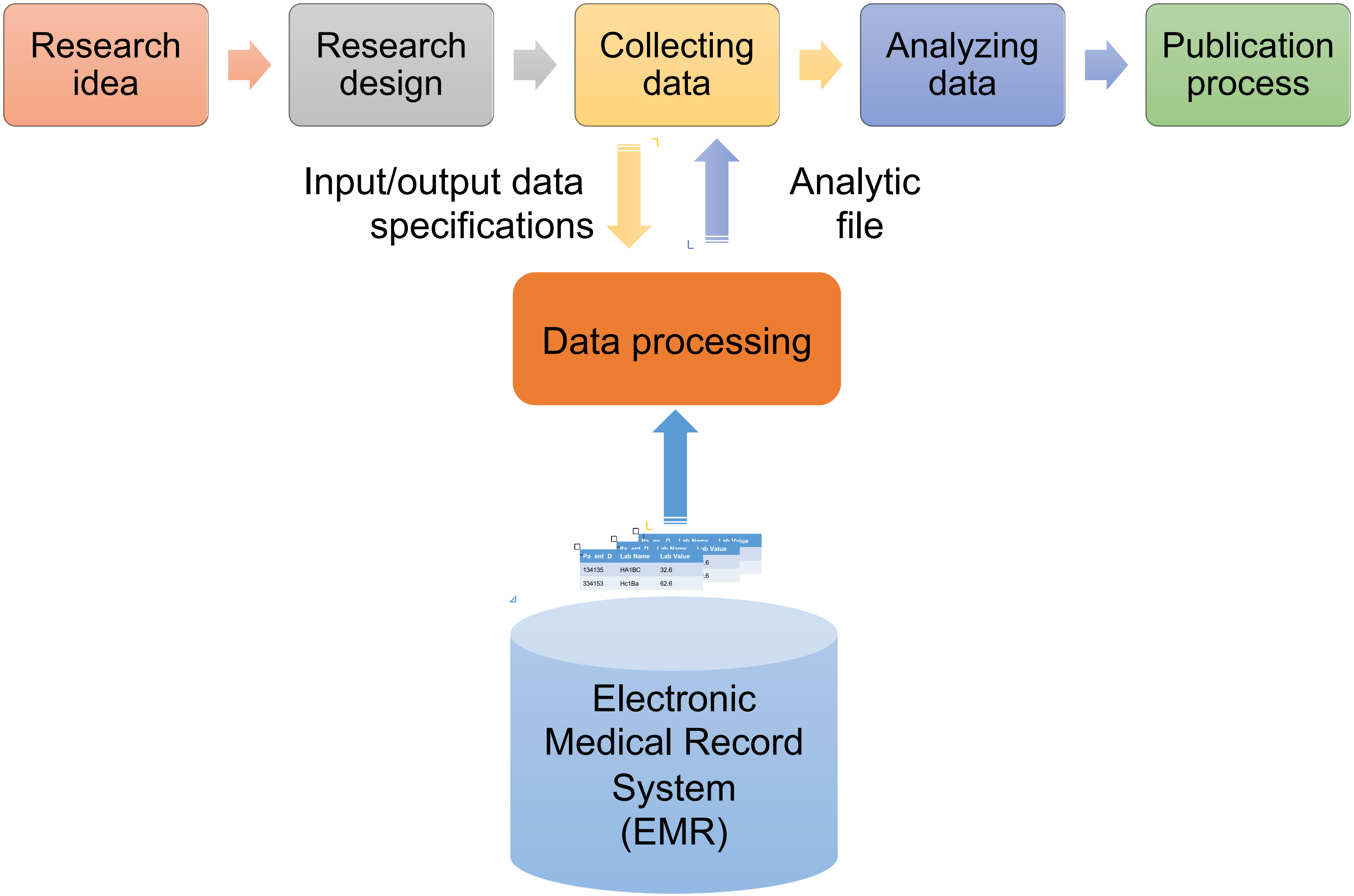
- Clinical outcomes research
- Health services research
- Disease modeling
- Comparative effectiveness
- Prospective clinical studies
- …

Retrospective study component

Secondary use of Electronic Medical Record

Cowie MR, Blomster JI, Curtis LH, et al (2017) Electronic health records to facilitate clinical research. Clin Res Cardiol 106:1–9.

Clinical research project workflow



Data Types in the EMR

Structured and coded data

Fasting Glucose

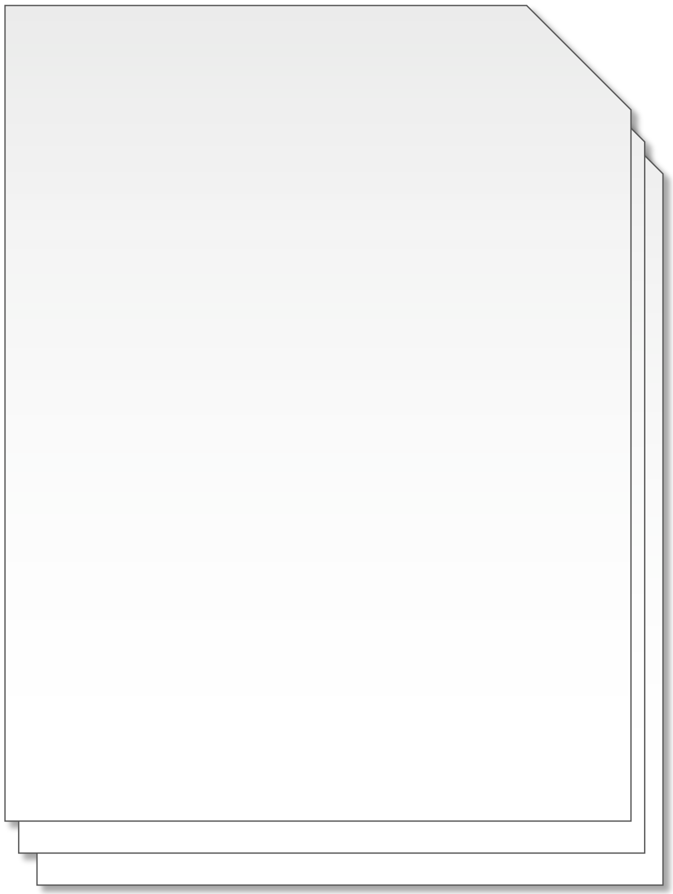
135 mg/dL

Pulse

60 bpm

Diagnoses

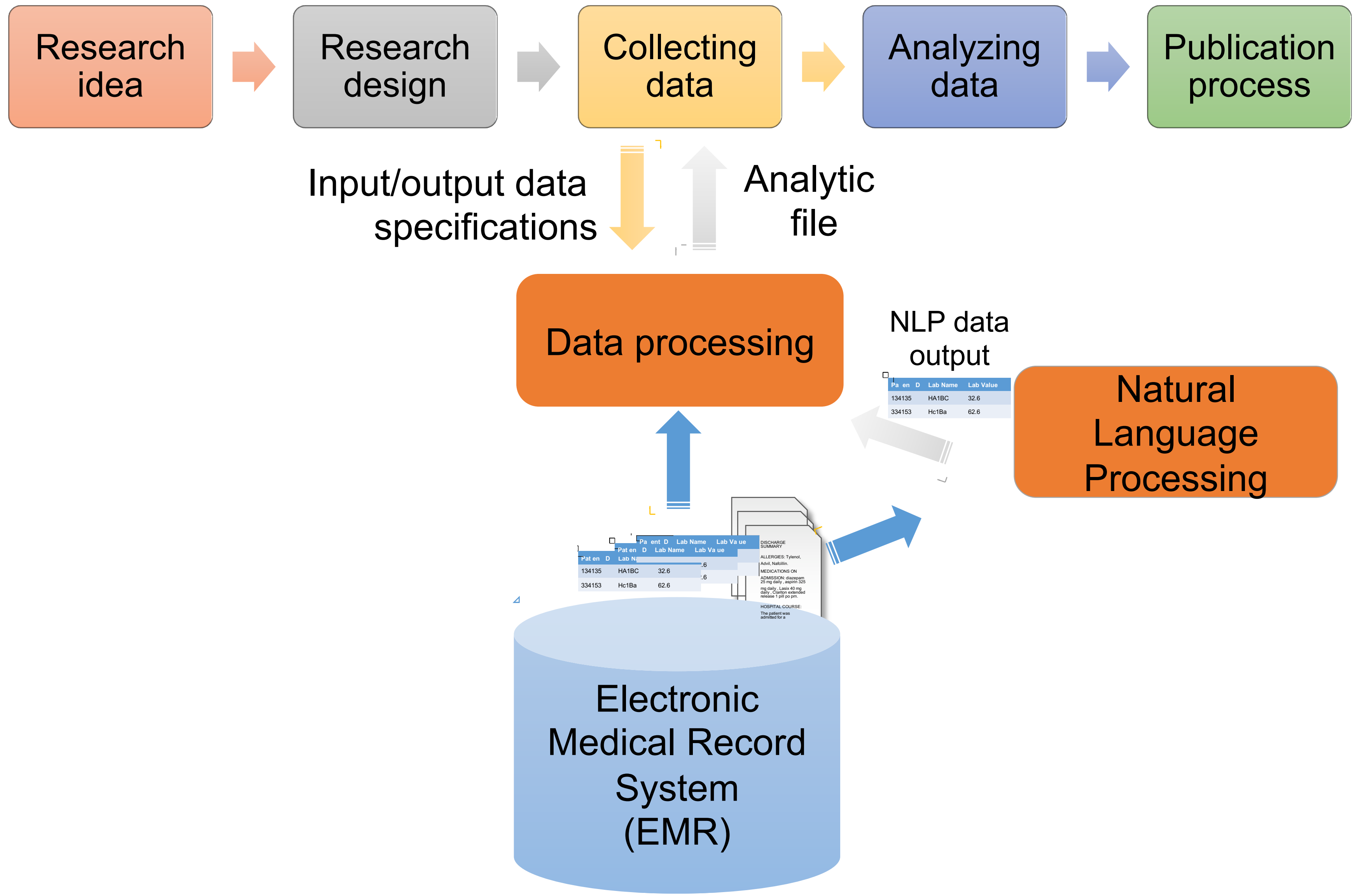
250.0 Diabetes
274.0 Gout
172 Melanoma



Unstructured narrative data

synthetic medical record data

Clinical research project workflow



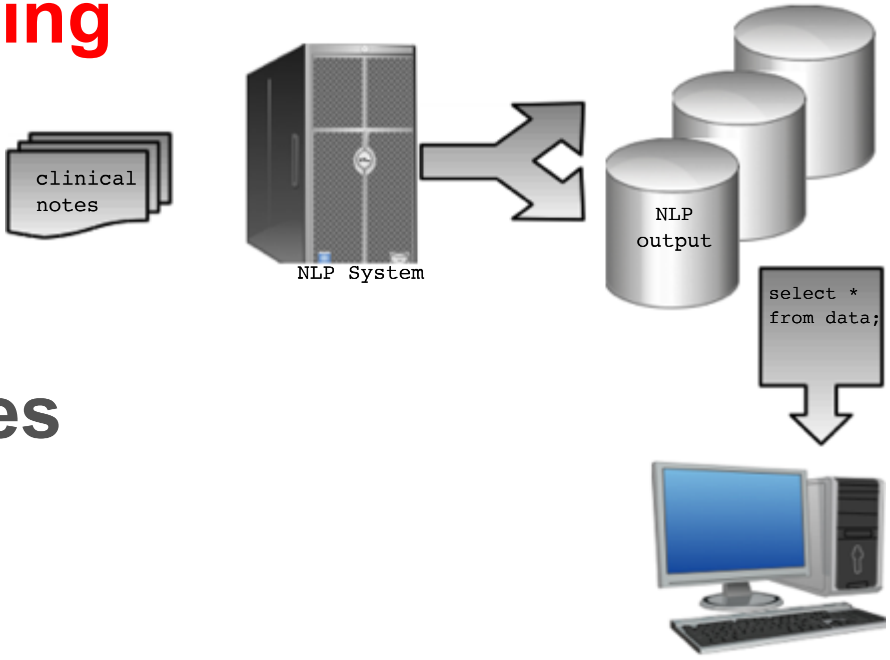
Natural language processing (NLP)

Natural language is automatically parsed into structured format

✓**computational processing**

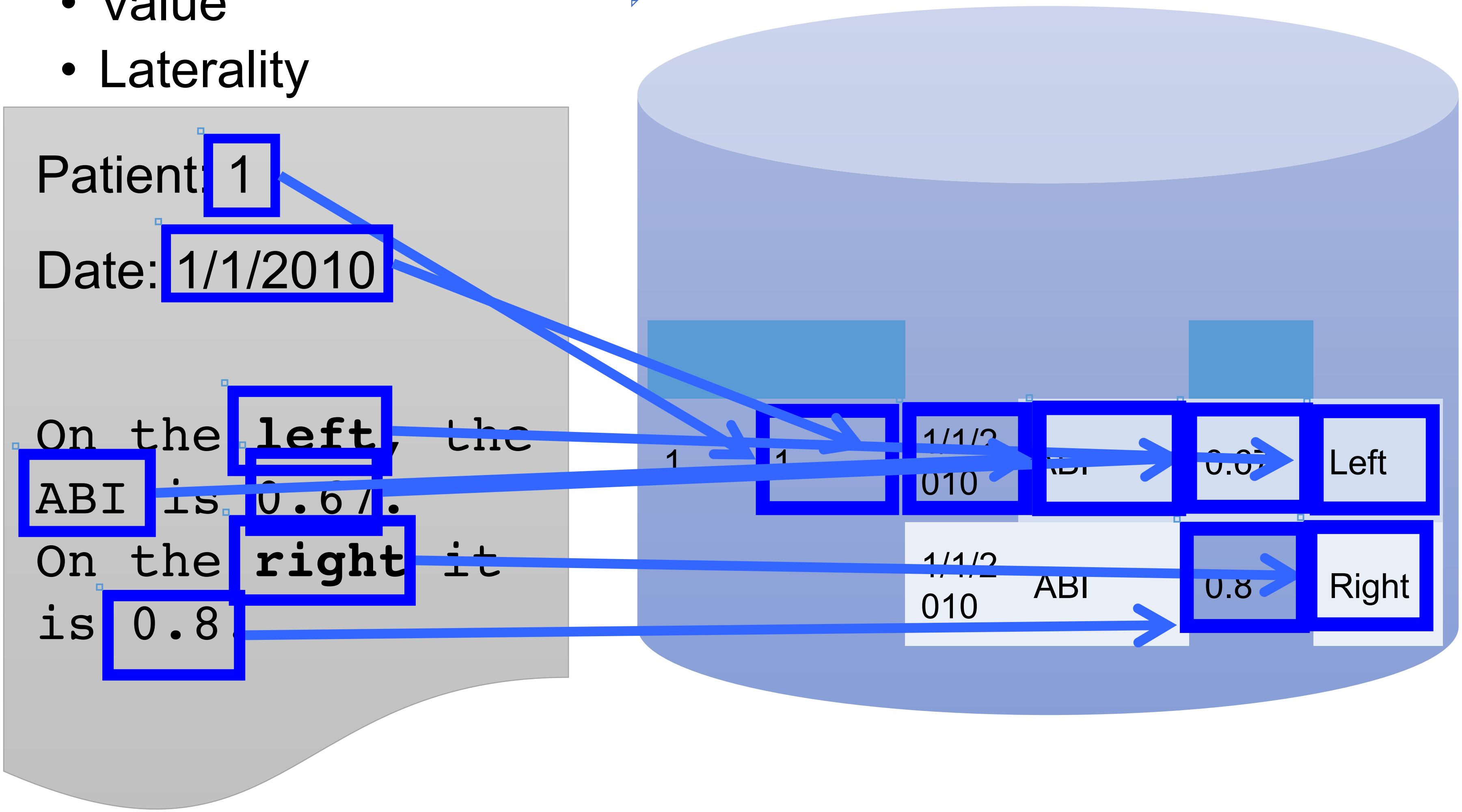
✓**unstructured text**

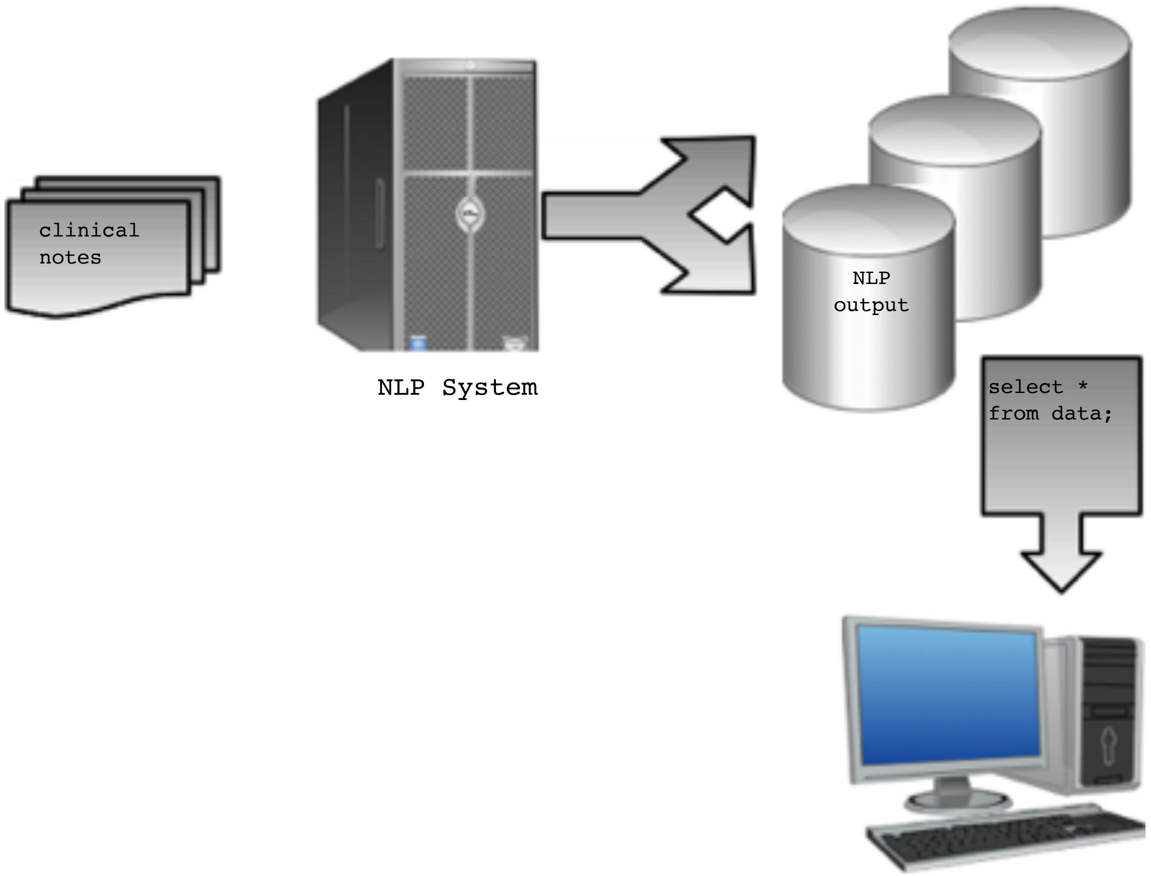
✓**broad range of purposes**



Information Extraction

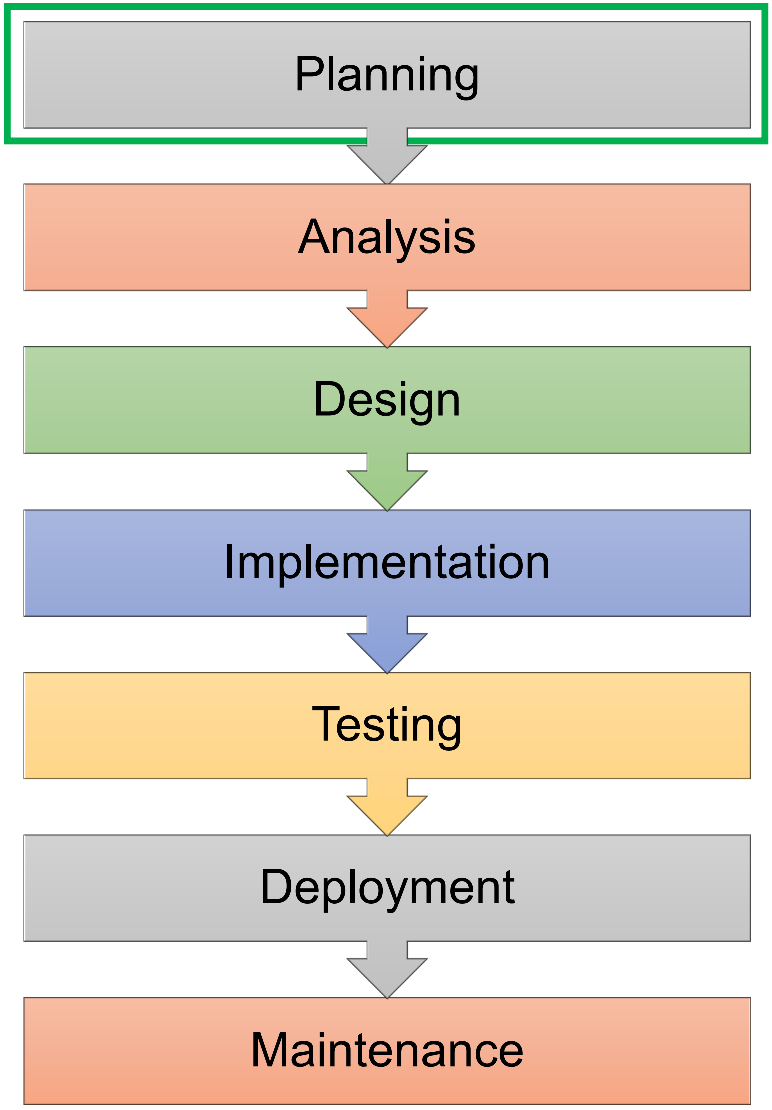
- Ankle Brachial Index
  - Value
  - Laterality





System development life cycle

System Development Life Cycle (SDLC)



SDLC: Planning

Adequately explicit concept definition

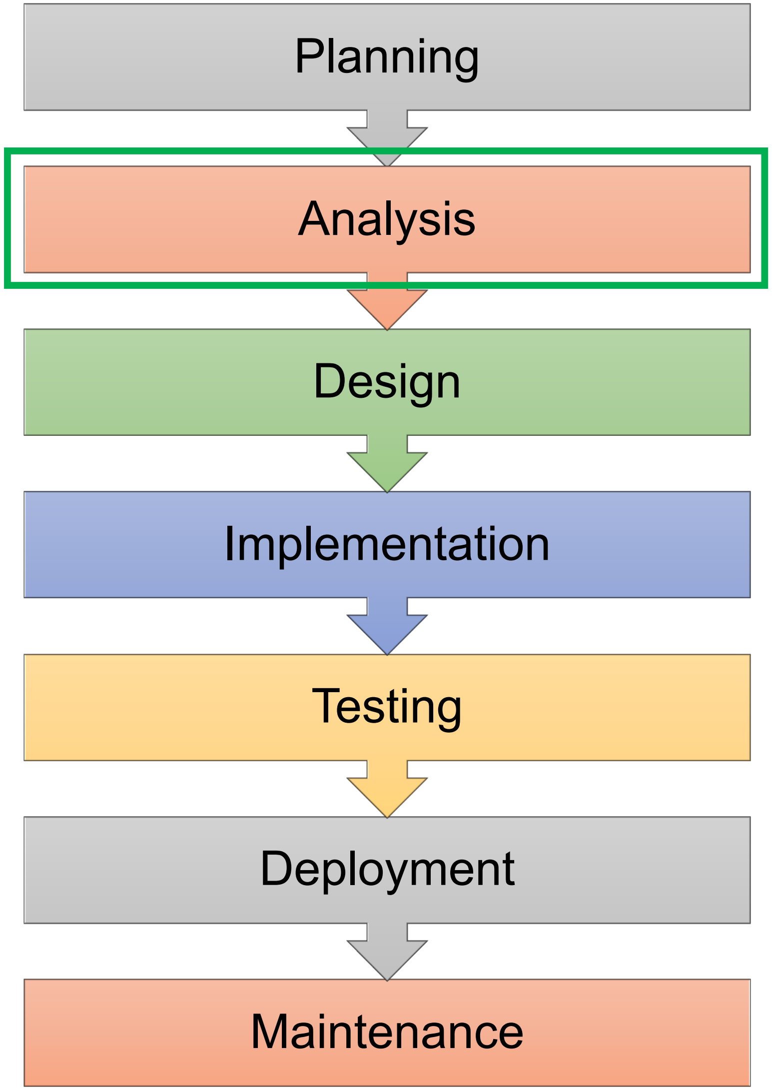
- “Concept = entity, idea, thought, meaning
- Clinical concept - entity targeted by NLP
  - Examples:
    - Diagnosis, symptom, finding
    - Lab value
    - Vital sign measurements ….
- Characteristics of a clinical concept for NLP
  - **single**, **unified** meaning across all instances of the concept
    - instances of the concept are directly comparable to each other
  - **project specific** definition
  - **documented in electronic medical record** (EMR)

Concept sheets

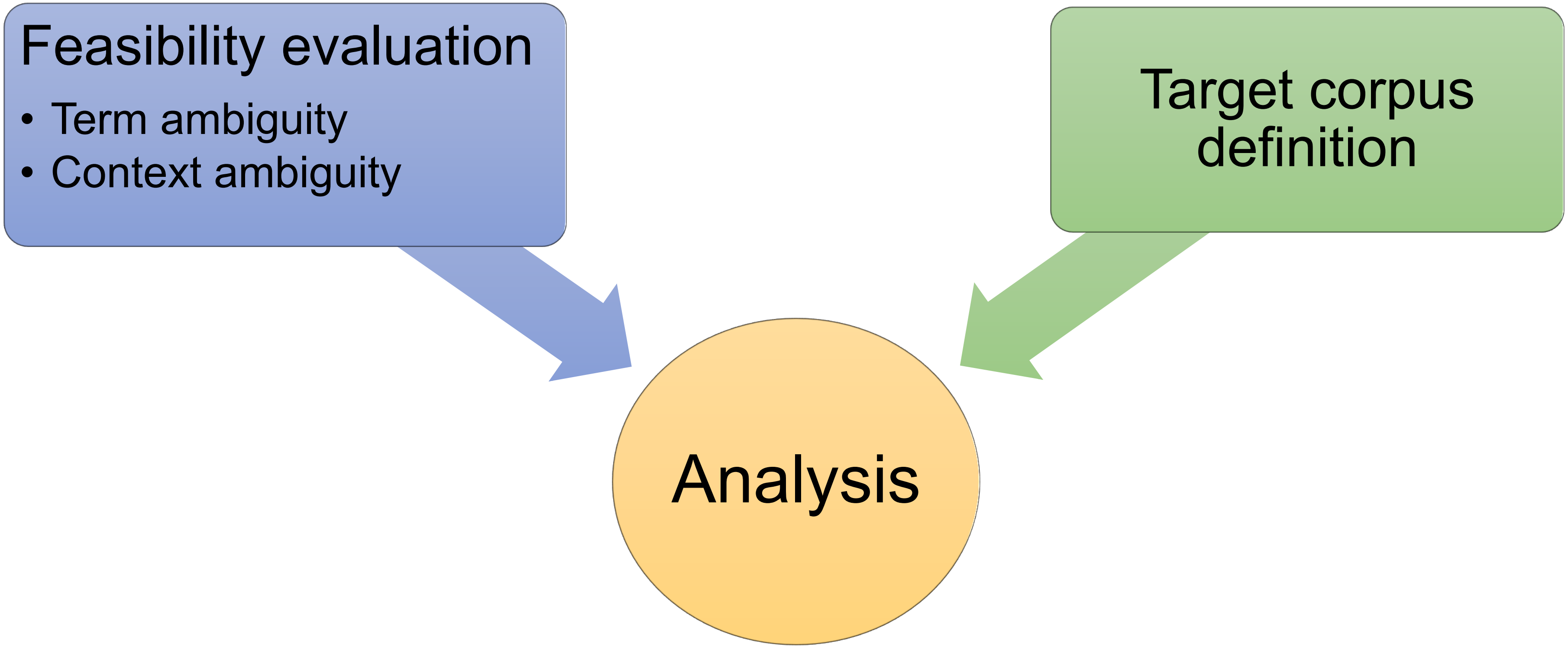
- To **operationalize** your variables
- As a **communication** tool
- Elements
  - Concept name
  - Detailed definition
  - Attributes
  - Examples: What it is and what it is not

Concept: Ankle Brachial Index
Definition: Explicit **historical or current** **affirmed** mentions of ABI **numerical values**, **regardless** of resting/after exercise, **regardless** of specific arteries
Attributes: Laterality (left, right, bilateral, not specified)
Exclude: pressure values

System Development Life Cycle (SDLC)



SDLC: Analysis



SDLC: Analysis

Feasibility analysis

- Estimate term ambiguity for concepts
  - Very ambiguous
    - abbreviation "**PE**"- Physical education, physical exam, pulmonary embolism, pulmonary edema, peak ejection, pleural effusion…
  - Not ambiguous
    - Fully spelled out phrase - **Left ventricular ejection fraction**

SDLC: Analysis

Feasibility analysis

- Estimate <u>context ambiguity</u>
  - Common symptoms have ambiguous context
    - “patient denies **fever**”
    - “patient arrived at the ER with **fever**”
    - “patient takes care of her mother who has high **fever**”
    - “if **fever** occurs, call the nurse”
    - “if **fever** does not go down, take medication”

  - Rare diseases or conditions have less ambiguous context (unless they are side effects of a treatment)
    - “Genotype showed significant HIV drug resistance, including the following mutations in HIV reverse transcriptase: **M41L**, **L210W**, and **T215Y**”

SDLC: Analysis

The results of feasibility analysis may show that it is **not feasible** or it is **not practical** to develop a **fully automated** information extraction system and the project definition should change.

Alternative to fully automated information extraction is

- NLP-assisted manual chart review
- Manual chart review
- Structured data analysis to utilize surrogates

SDLC: Analysis

Defining target corpus

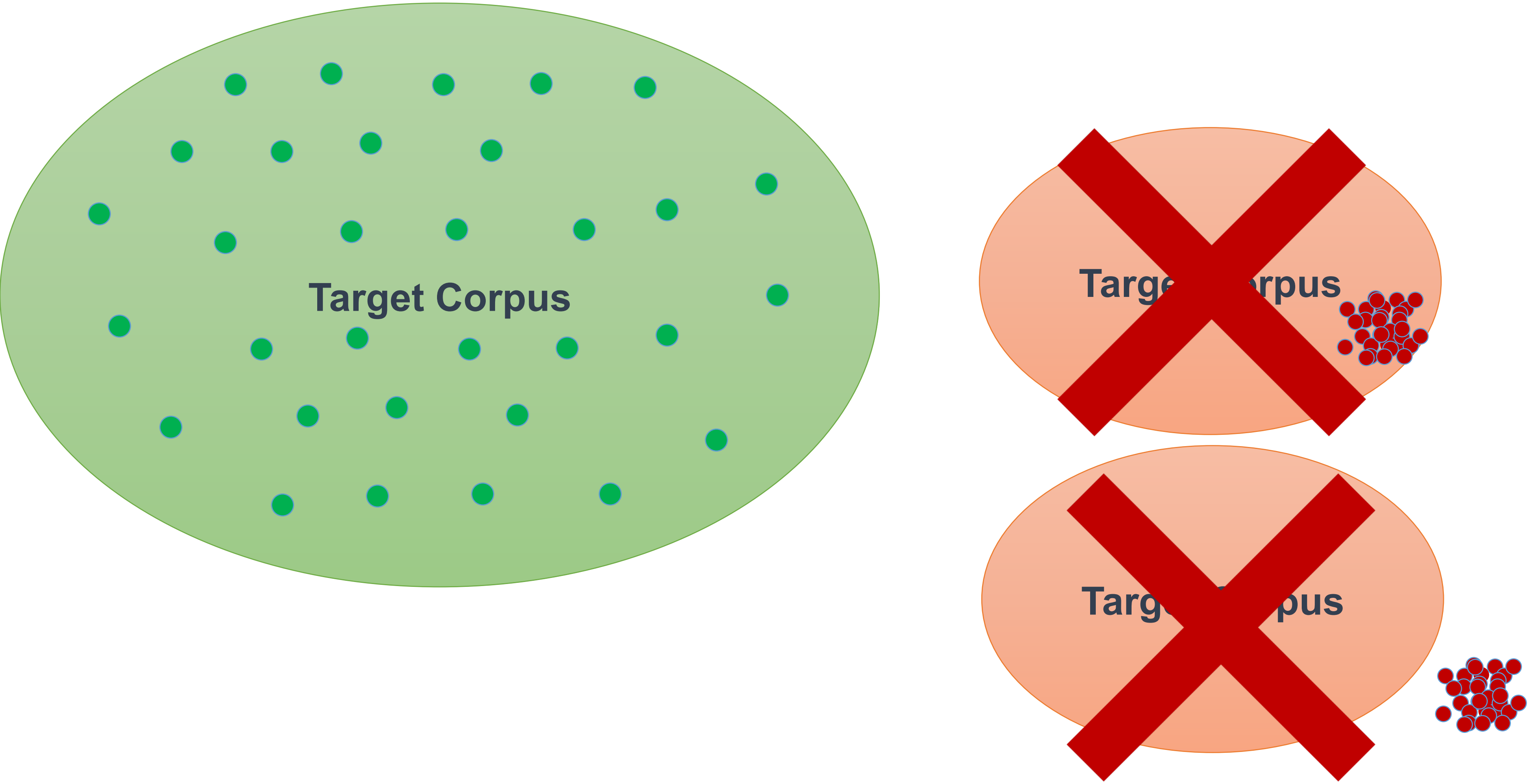
Target corpus - the complete set of clinical notes that the NLP system will be designed to process

- Document selection for processing
  - Automatic – information retrieval
  - Manual – selection heuristics
- Document selection for manual annotation
  - Reference standard, training / testing set

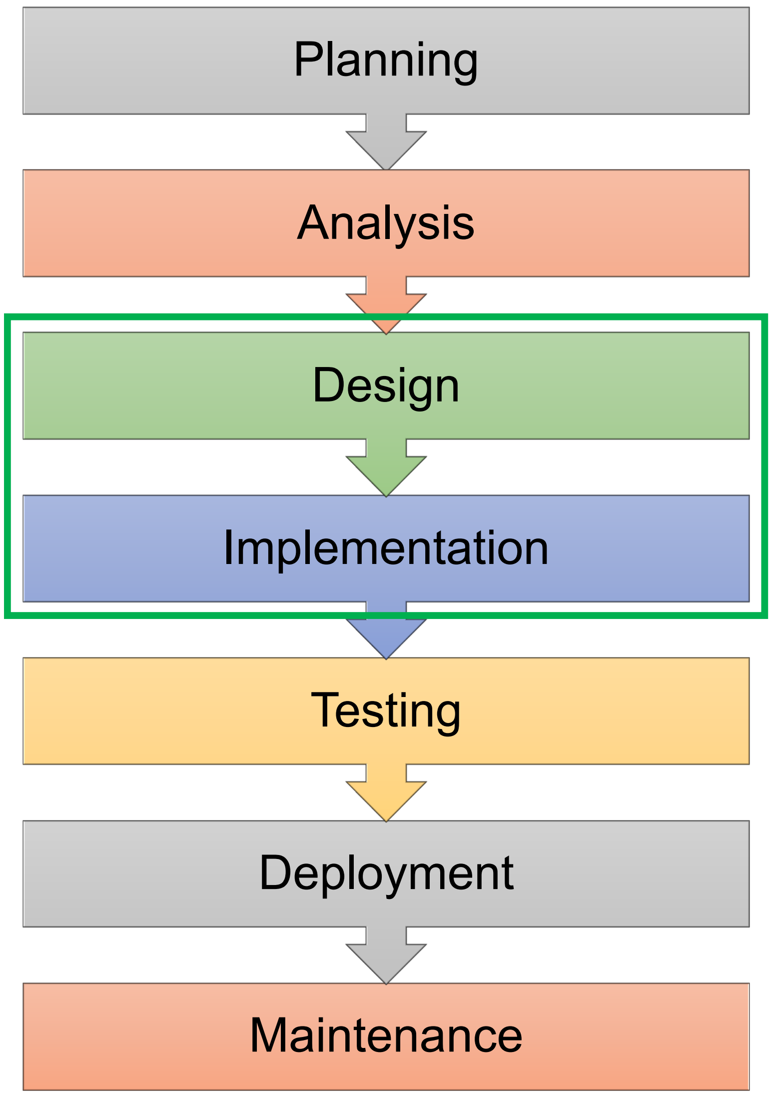
SDLC: Analysis

Important!

The reference standard set (train/test) must be **representative** of the full target corpus



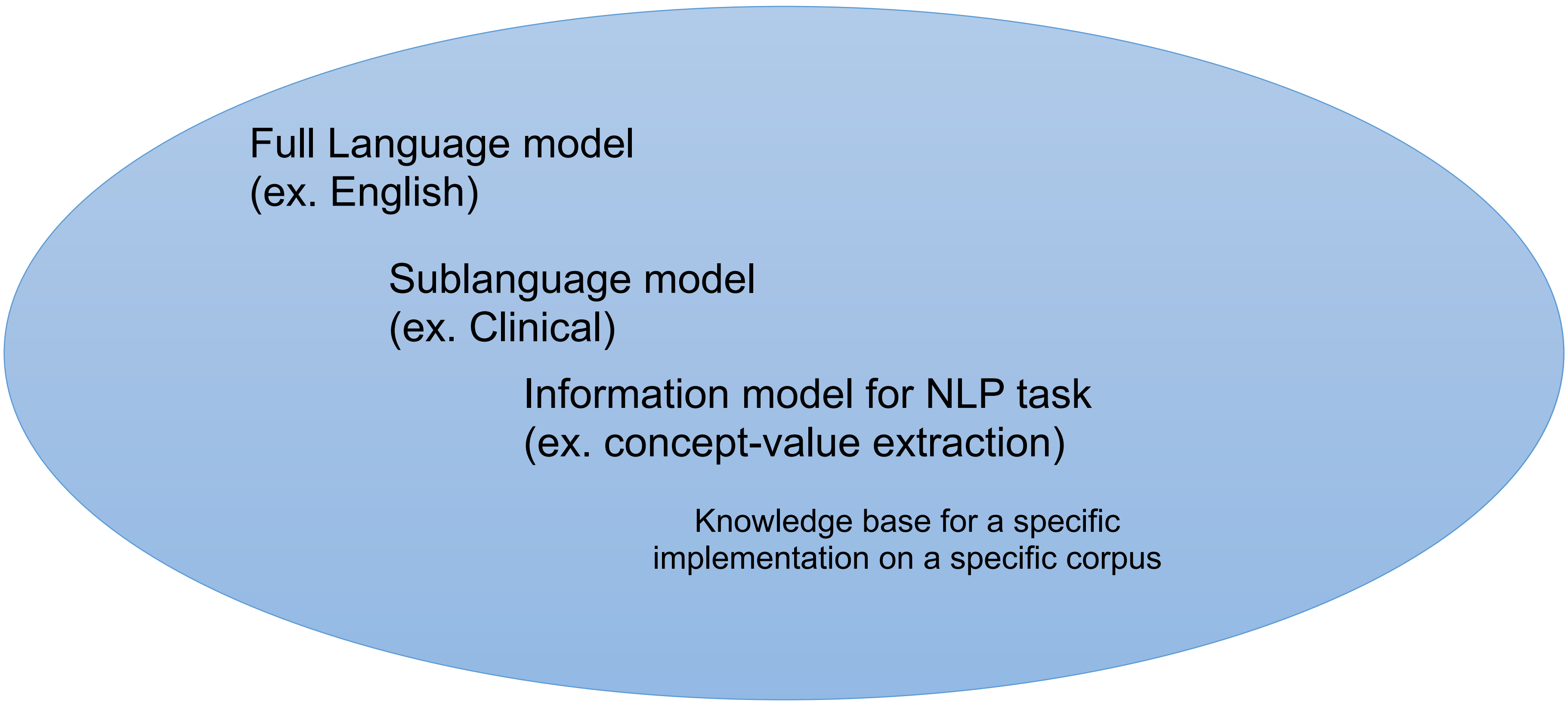
System Development Life Cycle (SDLC)



SDLC: Design

Information model

- a representation of concepts, the relationships, rules, and analytic steps to achieve a specific goal



Development of an information model

Text processing method	Information model component
Keyword search	Term list
Rules and patterns	Term lists Term sequence patterns Rule sets
Machine learning classification	Feature definitions Class list

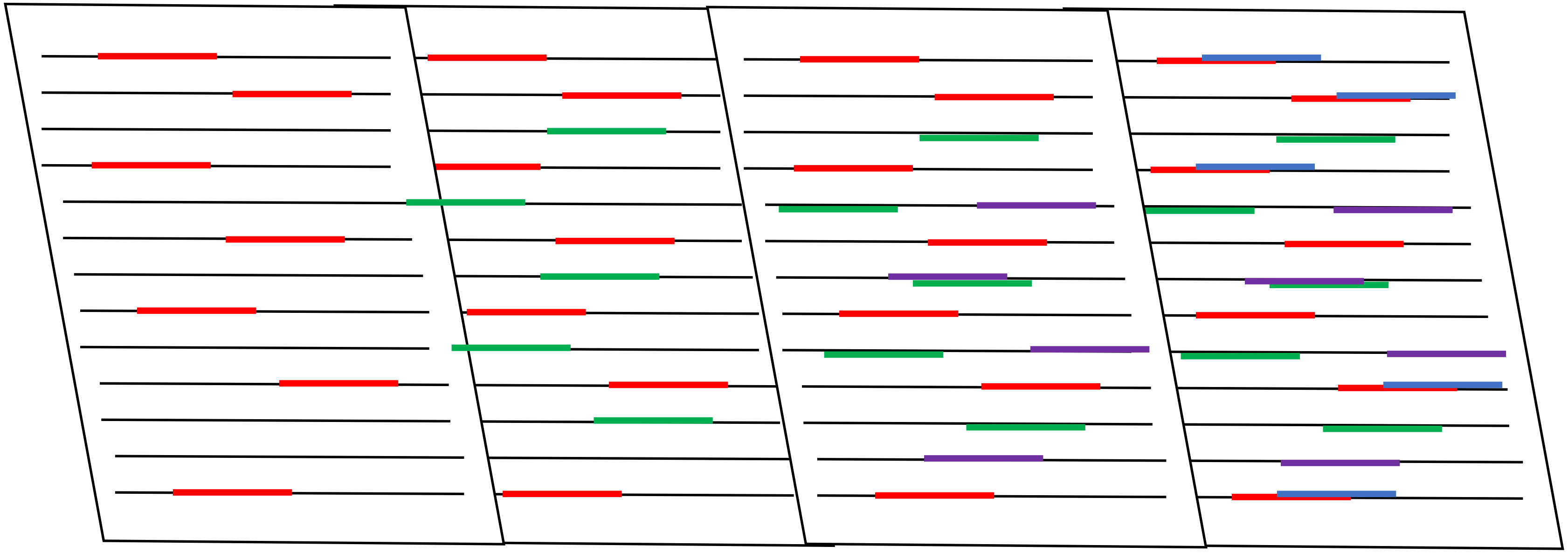
SDLC: Design

Approaches to knowledge base acquisition

- **Knowledge-driven**
  - Ontology, dictionary, terminology
- **Expert-driven**
  - Manually designed custom dictionaries, rules, heuristics
- **Data-driven**
  - Manually or statistically derived custom dictionaries, patterns, ML models
- **Hybrid**
  - Combination of approaches, different for each step
- **Ensemble systems**
  - combine methods in parallel
  - select the best performing method using rules or statistics, or combine results from several methods (e.g., majority voting, weighted voting, machine learning classifier of outputs).
  - Ensemble systems have reliably demonstrated better performance than individual methods but are significantly more complex to implement.

Pipeline system implementation

Tokenization | Parsing | Dictionary Mapping | Context



Pipeline annotations

Each module adds layers of **annotations**

- Annotation
  - **Class** - assigned meaning to data

  Label = concept = annotation class = annotation type ≈ semantic type
  - **Span** - a pointer to start and stop points in a text
  - **Features** - attributes of the Class and their **values**
- Generated by human, machine, or human+machine.

```
0123456789012345678901234567890123
The CXR shows LLL consolidation.
```

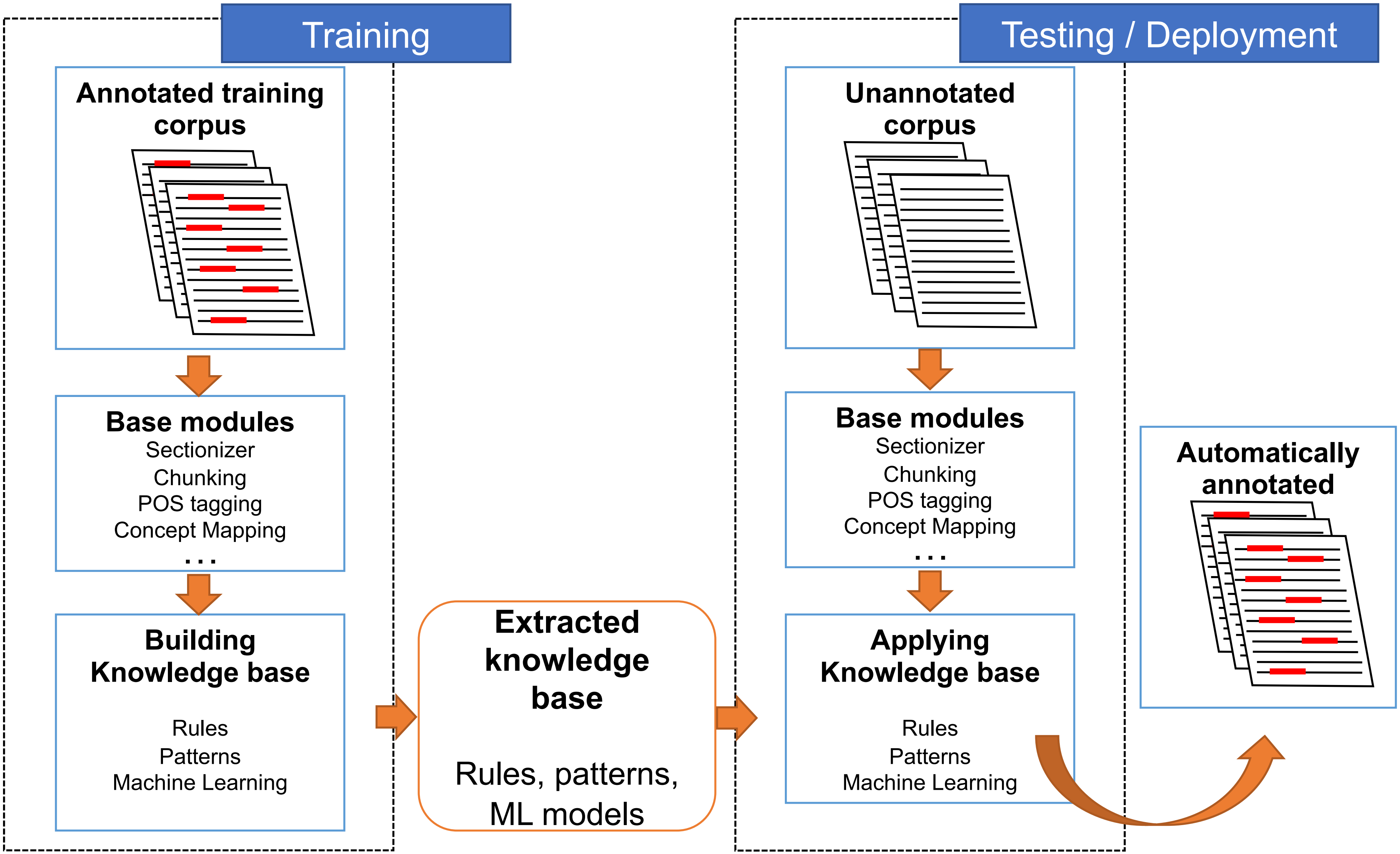
Annotation

```
Class:    Finding
Span:     14:30
Feature:  Assertion = Present
```

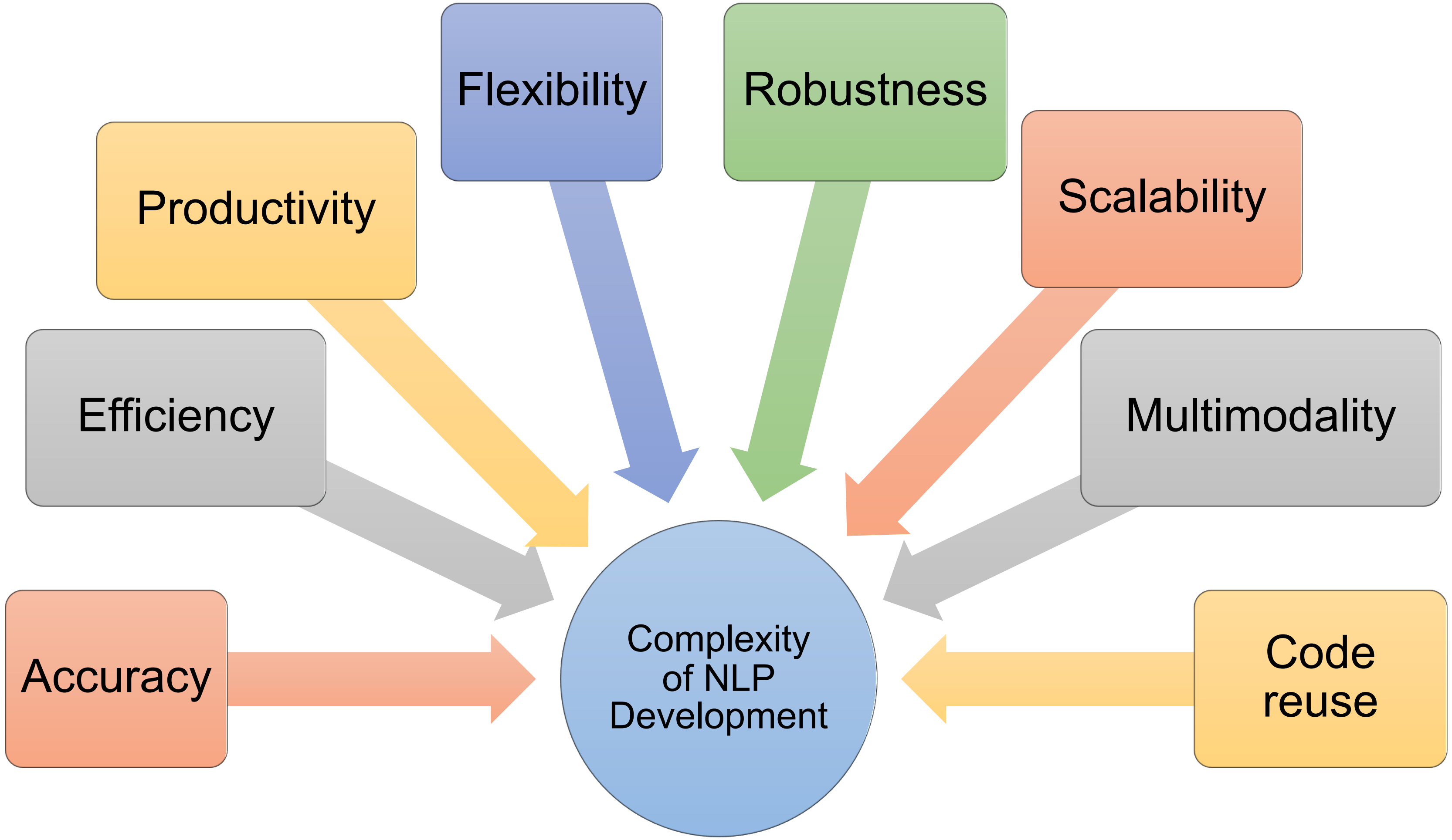
SDLC: Implementation

- An NLP pipeline with data-driven knowledge base has two versions:
  - one for training, and
  - one for testing (or deployment)

SDLC: Implementation

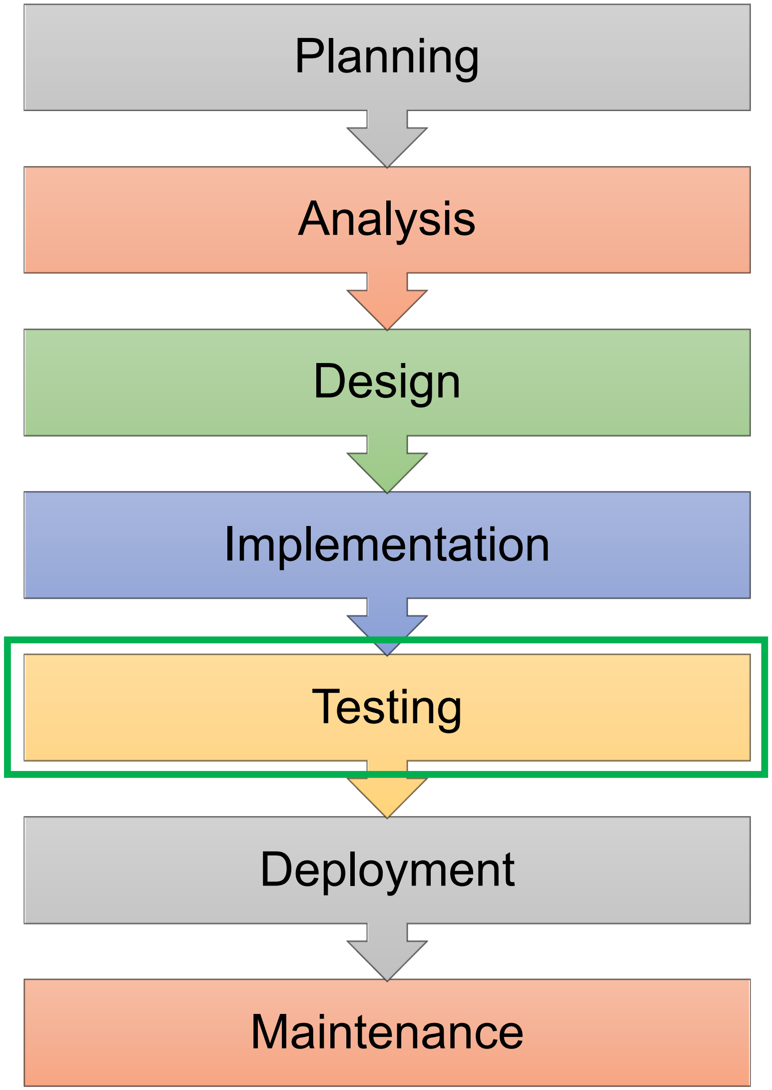


SDLC: Implementation



Leidner JL. *Current issues in software engineering for Natural Language Processing*. HLT-NAACL 2003 workshop on Software engineering and architecture of language technology system. http://portal.acm.org/citation.cfm?doid=1119226.1119233

System Development Life Cycle (SDLC)



SDLC: Testing

- **Validation**
  - Measures how **accurately** the NLP system performs extraction as compared to a *reference standard*
  - Reference standard can be:
    - Manually annotated text
    - Benchmark system
  - Measured in classic performance measures
    - **Recall**
    - **Precision**
    - **F-measure**

SDLC: Testing

Information extraction confusion matrix

	Reference standard: Annotation	Reference standard: No Annotation	
System annotation	**True positive**	**False positive**	**Total system positive**
No System annotation	**False negative**	True negative	Total system negative
	Total ref standard positive	Total ref standard negative	

$$Recall = \frac{True\ Positive}{Total\ Ref\ Standard\ Positive} = \frac{True\ Positive}{True\ Positive + FalseNegative}$$ = Sensitivity

$$Precision = \frac{True\ Positive}{Total\ System\ Positive} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$ = Positive Predictive Value

$$F1 = \frac{2 * Precision\ * Recall}{Precision + Recall}$$

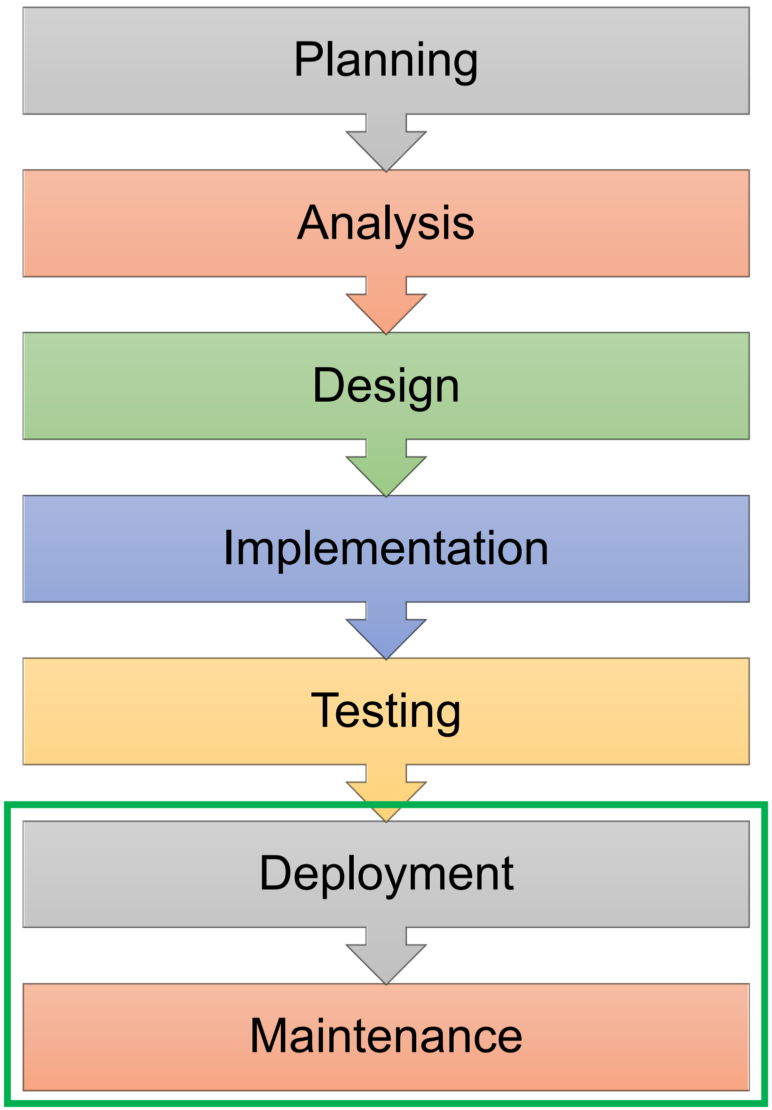
SDLC: Testing

Error analysis

– manually examine the examples when system output did not match manual annotation and attempt to find the reason for the error

- **Systematic error**
  - The same reason for each instance of an error of that type
  - Can be fixed by updating dictionary, or rules, or retraining machine learning step with a different feature set
- **Random errors**
  - New misspellings, new context
  - Each error is different from other errors
  - Cannot be fixed without significant decrease in recall or precision

System Development Life Cycle (SDLC)

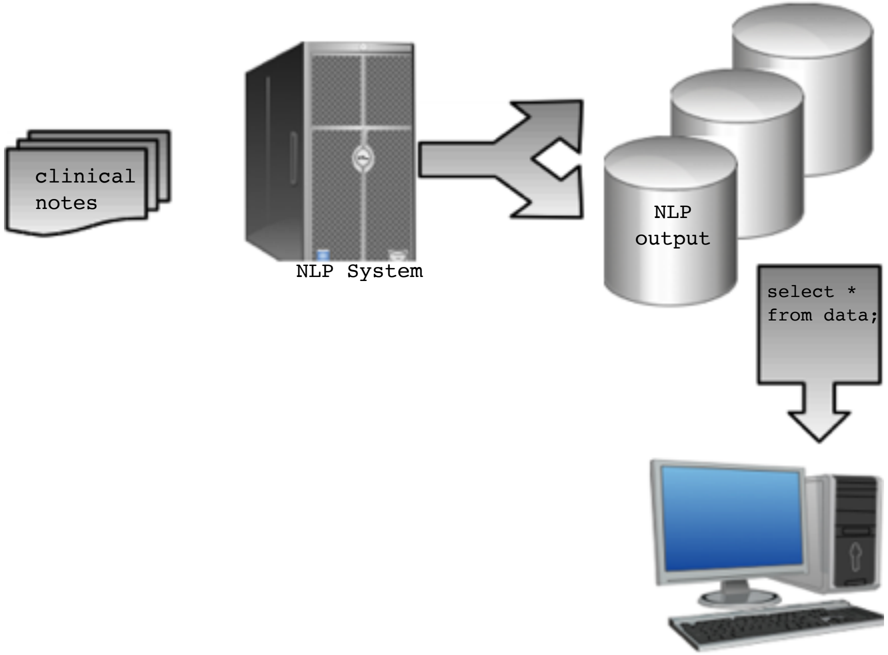


SDLC: Deployment

Deployment

=> NLP system is applied on the target corpus

=> output dataset is created



Applied NLP limitations

Row ID	Patient ID	Date	Index	Value	Laterality
1	1	1/1/2010	ABI	0.67	Left
2	1	1/1/2010	ABI	0.8	Right
3	1	1/3/2010	ABI	0.67	Left
4	1	6/1/2010	ABI	<0.8	
5	1	1/1/2015	ABI	0.67	Left
6	1	1/1/2015	ABI	0.58	Left

Date of documentation

Recent value (?)

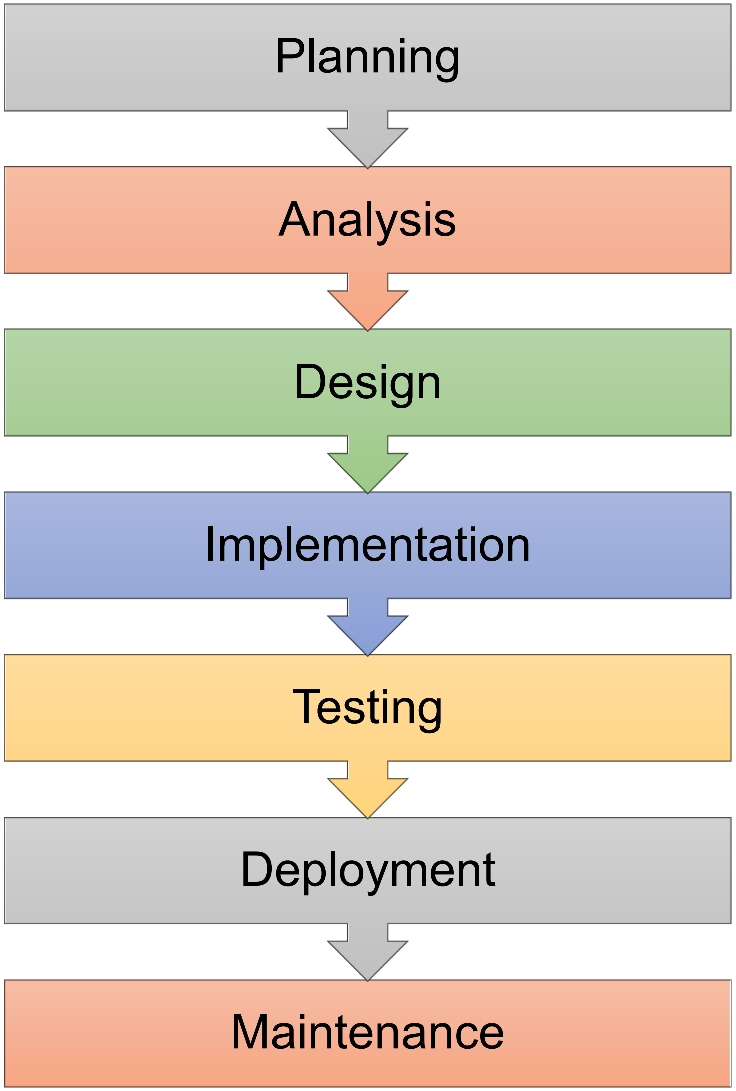
Date of measurement (?)

Historical value (?)

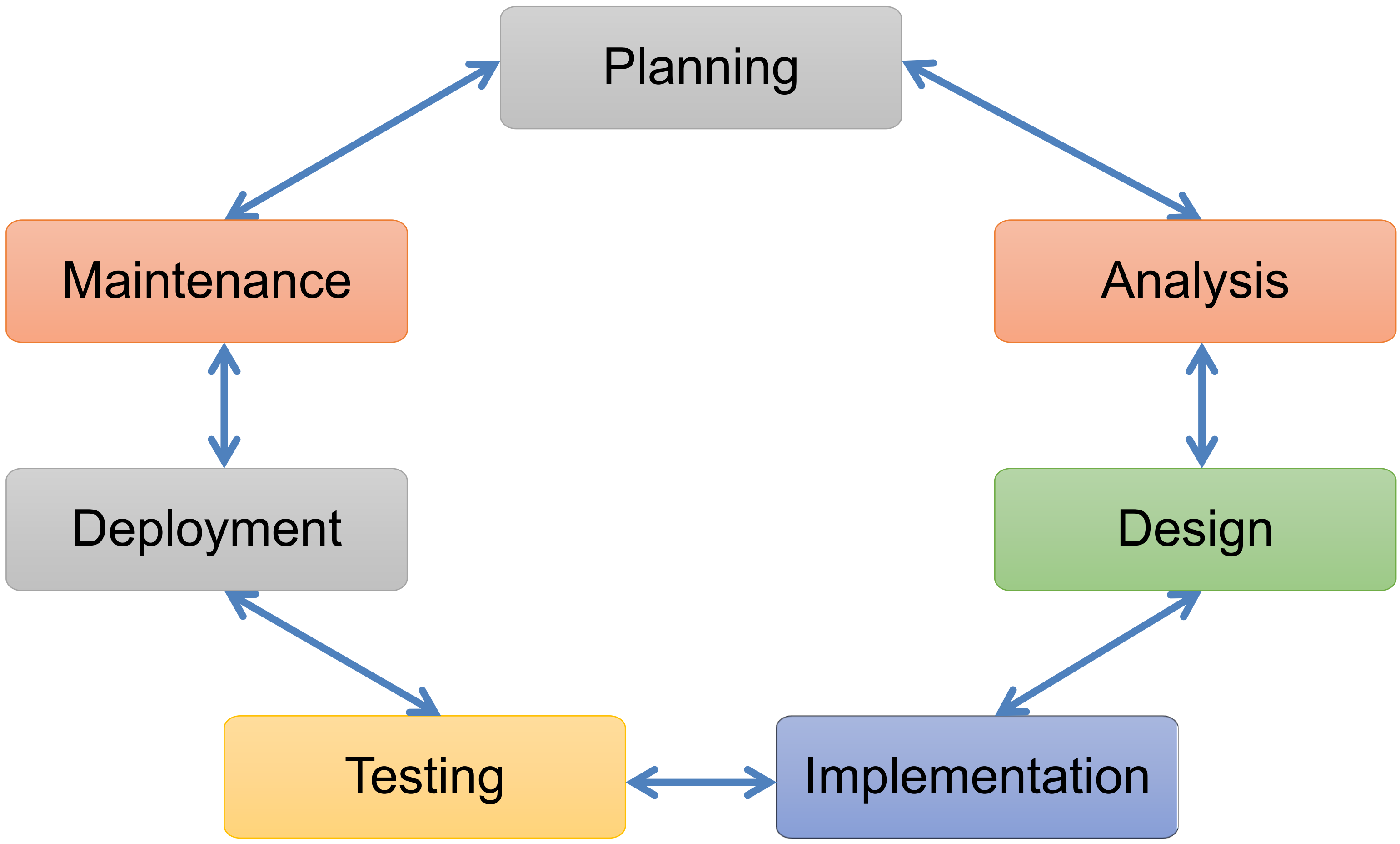
SDLC: Maintenance

- Linguistic drift
  - Language changes over time
  - New words enter the lexicon (ebola, zika, covid)
  - Standard coding systems change (ICD9 -> ICD10)
- Settings change
  - New guidelines for documenting of care
  - Sublanguage variations across medical subdomains
- For systems that applied over time, regular validations are required
- Rules, patterns, ML models need to get updated

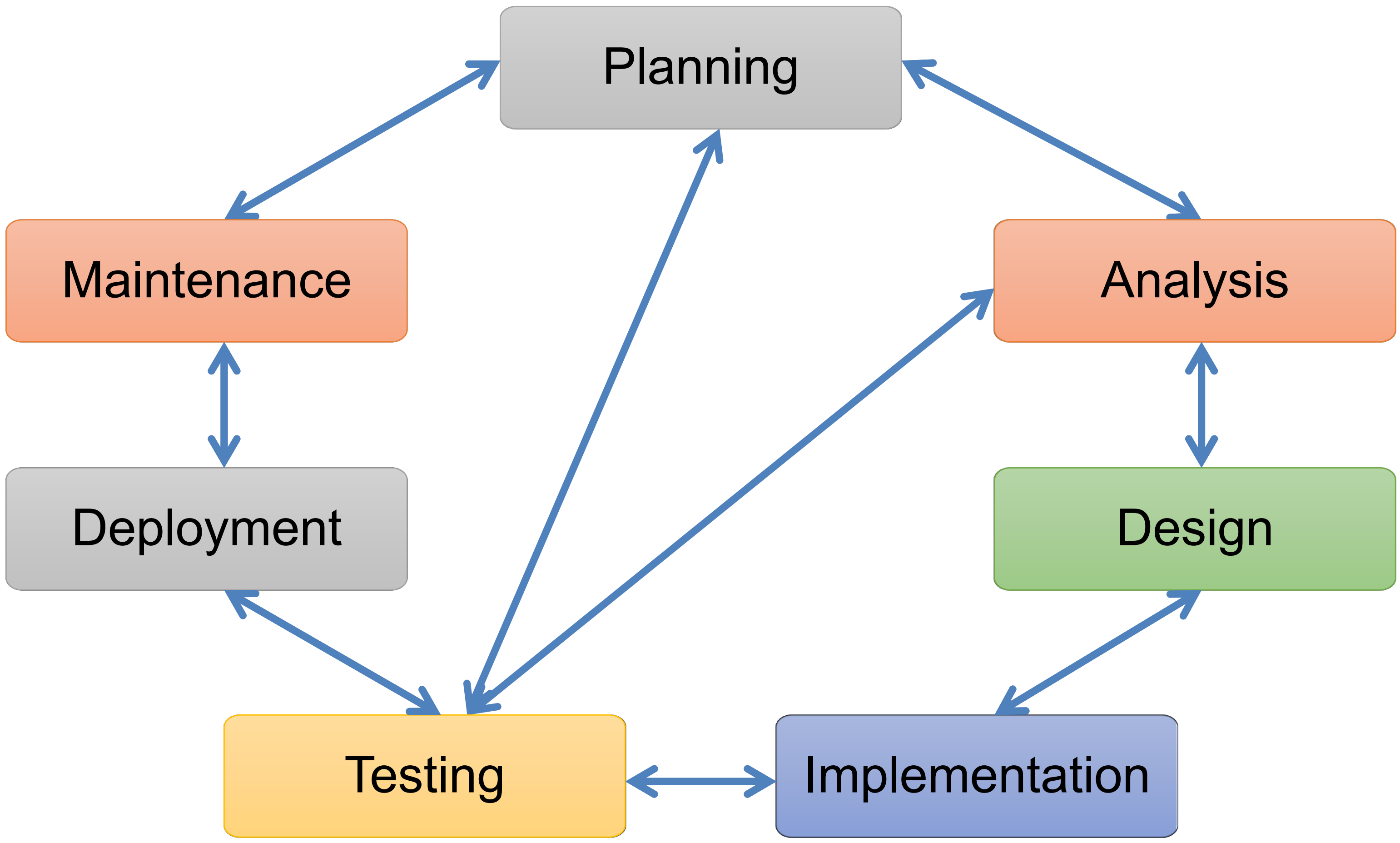
System Development Life Cycle (SDLC)



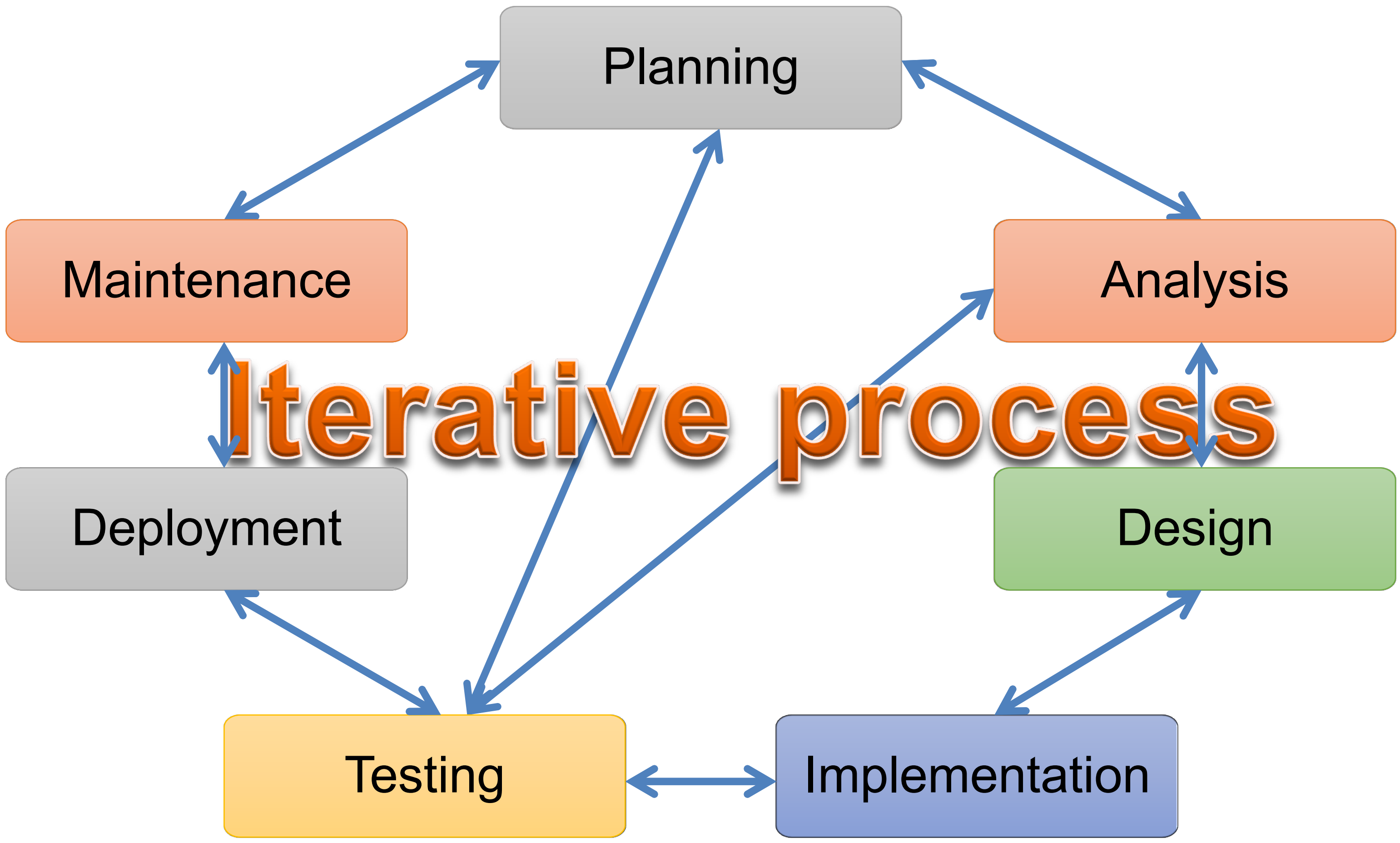
System Development Life Cycle (SDLC)



System Development Life Cycle (SDLC)



System Development Life Cycle (SDLC)



Contact Info

Contact for questions or comments

VINCI@va.gov

VINCIPedia – materials on VINCI Central

https://vhacdwweb02.vha.med.va.gov/prod/vincipedia/VINCIPedia/VINCI%20Data%20Science%20Academy.aspx