




Методы машинного обучения

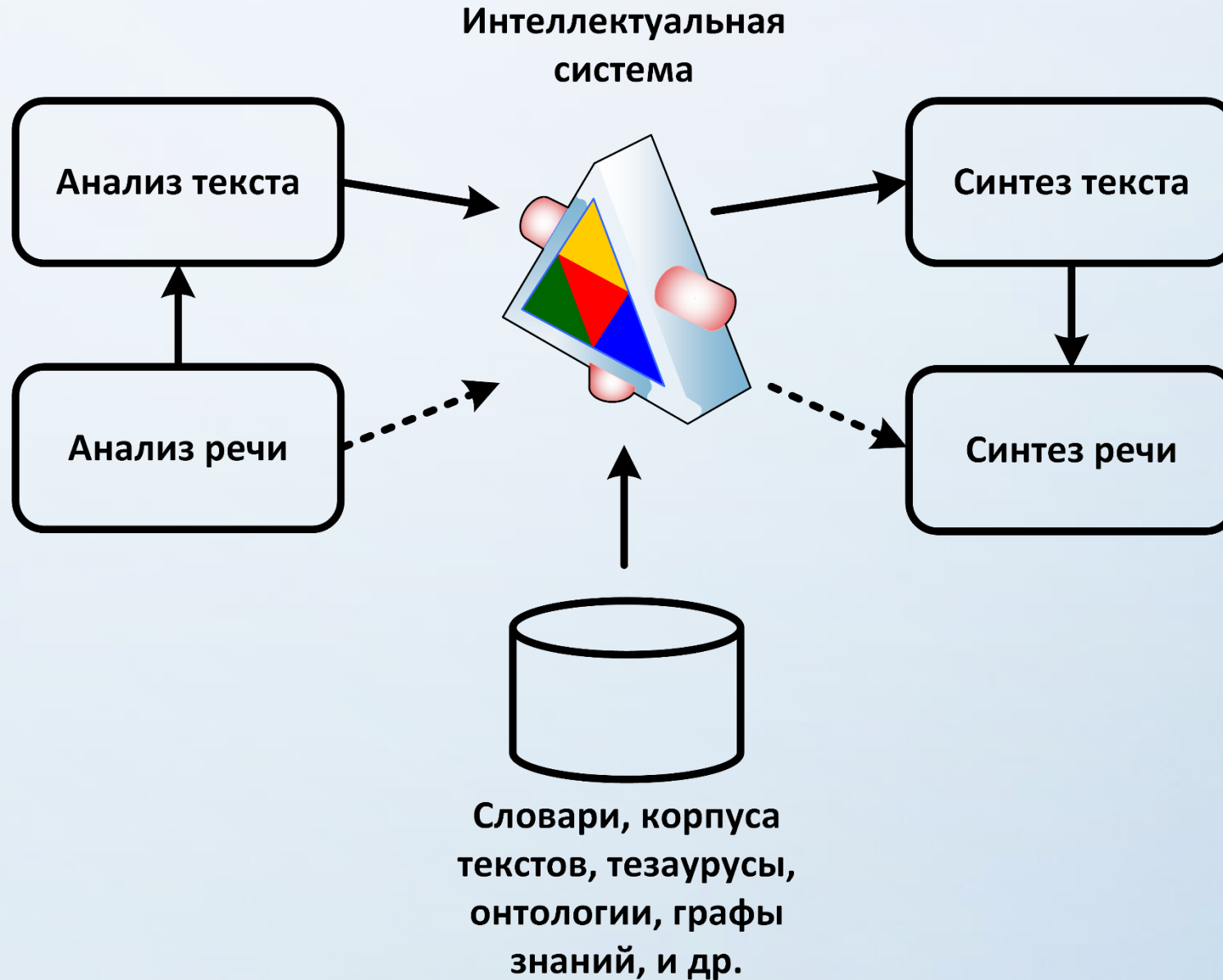
ИУ-5, магистратура, 2 семестр,
весна 2021 года



Введение в обработку текстов и графов знаний



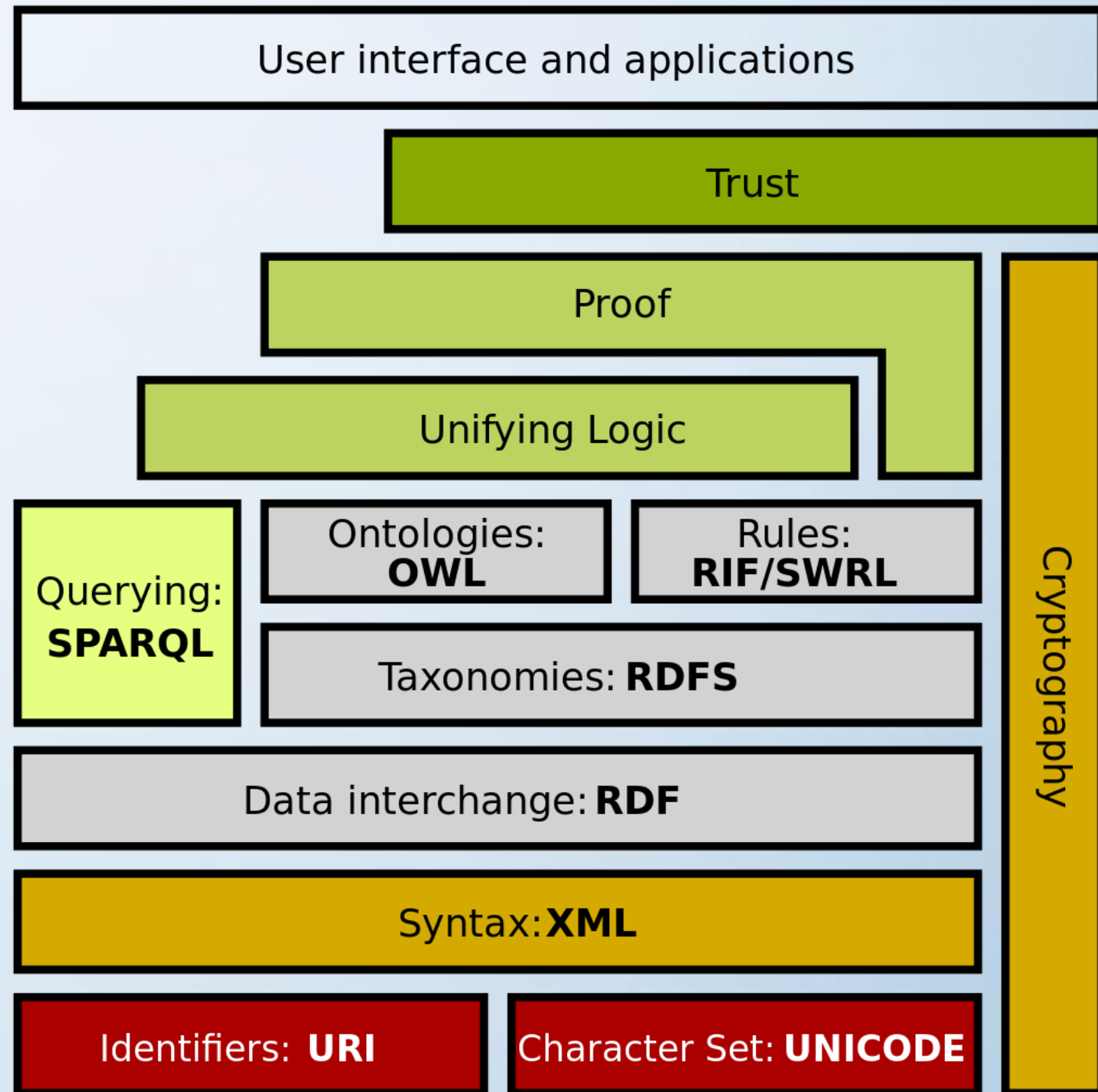
Структура интеллектуальной системы



Обработка языка – основные понятия

- Обработка естественного языка
 - Обработка текстов на естественном языке (Natural Language Processing, NLP)
 - Более старая аббревиатура АОТ (Автоматизированная Обработка Текстов)
 - Основные задачи NLP
- Компьютерная (математическая) лингвистика
 - Корпус текстов
 - Дистрибутивная семантика (гипотеза)
 - Тезаурус
 - Онтология
 - Проблема соответствия онтологий
- База знаний по лингвистическим ресурсам для русского языка (NLPub)
- Для применения моделей машинного обучения используется «векторное представление слов (word embedding)».

Semantic WEB - 1



Semantic WEB - 2

- RDF
- RDF-Star
- RDFS
- OWL
- Введение в языки описания онтологий семантического веба
- SPARQL
- SWRL
- Дескрипционная логика

Графы знаний

- Графы знаний как средство улучшения искусственного интеллекта
- Граф знаний – это база знаний, использующая графовую (мульти, гипер-графовую) модель. Основная задача – обеспечение полноты знаний.
 - Тезаурус – специализированная база знаний, предназначенная для хранения (служебной) лингвистической информации.
 - Онтология – прежде всего схема данных, ориентированная на логический вывод. Основная задача – обеспечение непротиворечивости знаний.
- Наиболее известные проекты:
 - DBpedia
 - Yago
 - ConceptNet
 - Статья
 - Atomic 2020 (ориентирован на ситуации и причинно-следственные связи)
- Задачи, решаемые на графах знаний:
 - Предсказание вершин и связей.
 - Логический вывод на графе знаний (эта задача также решается на онтологиях)
 - Embedding (векторное представление) графов знаний для использования моделей машинного обучения.

Обзор «классических» книг

1. Управление большими (сложными) системами, ситуационное управление, прикладная семиотика
 - Синтаксис, семантика, прагматика (Ю.И.Клыков, стр. 40; Л.В.Найханова, стр. 15)
 - Ситуация (Ю.И.Клыков, стр. 60; Д.А.Поспелов, стр 26)
 - Обобщение сетей (Ю.И.Клыков, стр. 30)
2. Семантика текста определяется через:
 - Формальные грамматики (И.А. Мельчук)
 - Наборы правил (Л.Л. Иомдин)
 - Специализированное исчисление (В.А.Тузов, В.В.Мартынов)
- 3. Лингвистический процессор (Ю.Д.Апресян)
- 4. Диалоговые (вопросно-ответные) системы используют лингвистические методы для построения логической модели текста, далее производится обработка логической модели (с целью ответа на вопрос и т.д.)

Выводы

- Обработка текстов и графов знаний в ИИ никогда не отделялись друг от друга.
- Наиболее развитые модели обработки текстов и графов знаний связаны с ситуационным управлением (прикладной семиотикой).
- Любая современная нейросетевая архитектура обработки текстов представляет собой специализированный лингвистический процессор.
- Диалоговые (вопросно-ответные) системы также основаны на совместной обработке текстов и графов знаний.
- Особенности «современного» подхода:
 - Кодирование текстов и графов знаний на основе векторных представлений (в противоположность более старому «логическому» подходу с правилами, исчислениями и т.д.)
 - Использование сложных нейросетевых ансамблей в качестве специализированных лингвистических процессоров.