Изкуствен интелект в областта на езиковото, културното и историческото наследство КЛаДА-БГ

Кирил Симов

Институт по информационни и комуникационни технологии – БАН

Компетентности и добри практики в областта на високопроизводителните пресмятания, анализа на големи данни и изкуствения интелект

София, 15 юли 2021

Какво е КЛаДА-БГ?

- Национална интердисциплинарна изследователска Е-инфраструктура за ресурси и технологии за българското езиково и културно наследство, интегрирана в рамките на европейските инфраструктури CLARIN и DARIAH (КЛаДА-БГ)
- CLaDA-BG = CLARIN and DARIAH in Bulgaria
- CLARIN EU: https://www.clarin.eu/
- DARIAH EU: https://www.dariah.eu/

Изследователски инфраструктури

Изследователските инфраструктури са сложни системи, включващи:

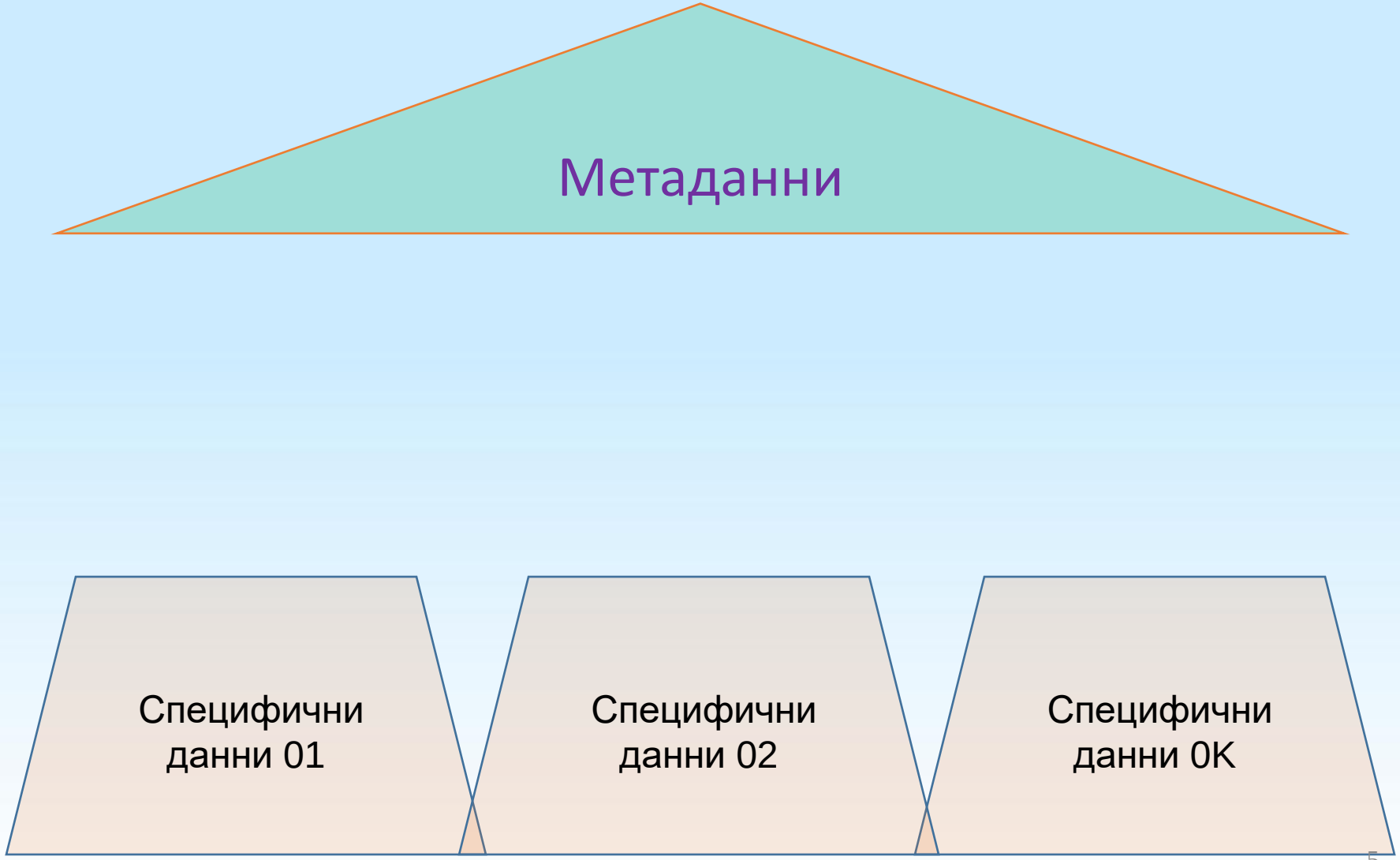
- знания,
- данни,
- хора и
- услуги

Те обединяват различни ресурси за подкрепа на широк диапазон от потребители и правят тези ресурси (многократно) приложими и достъпни за достатъчно дълъг период, за да подкрепят научните изследвания (индивидуални или съвместни) и да споделят резултатите от това проучване

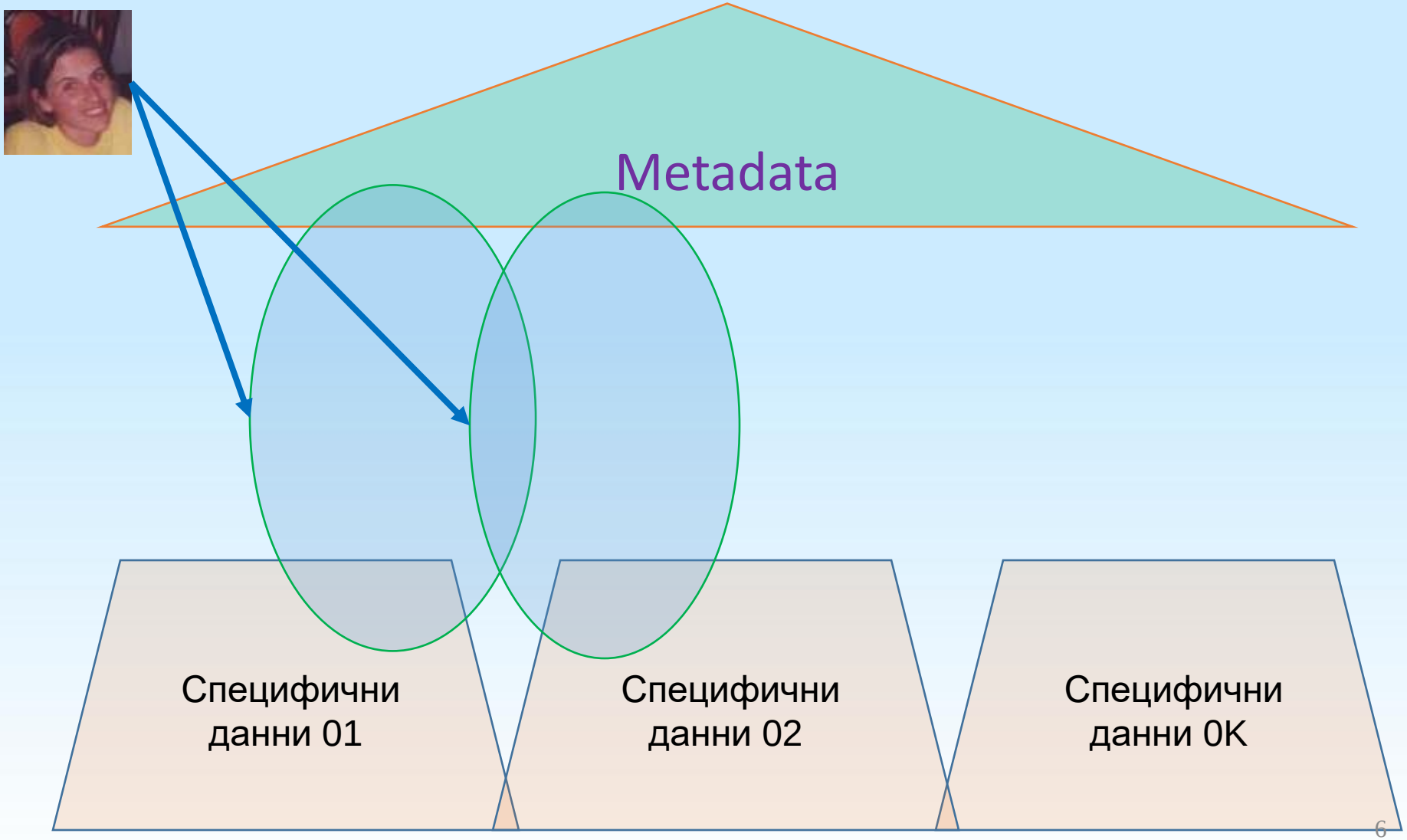
Данни и знание

- Данни – документи в различни медии и формати: текст, хипертекст, аудио, видео, изображения
- Знание – интерпретация на данните по отношение на тяхното създаване, съдържание и взаимовръзки
- КЛаДА-БГ се базира на разбирането, че в областта на хуманитаристиката и социалните науки основното количество данни са неструктурираните и нехомогенни текстови данни – „Language Technology (LT) is a core Big Data technology. Growth in the volume and variety of data is mostly due to the accumulation of unstructured text data; in fact, up to 80% of all data is unstructured text data.“
  http://roadmap2018.esfri.eu/media/1066/esfri-roadmap-2018.pdf (p. 108)

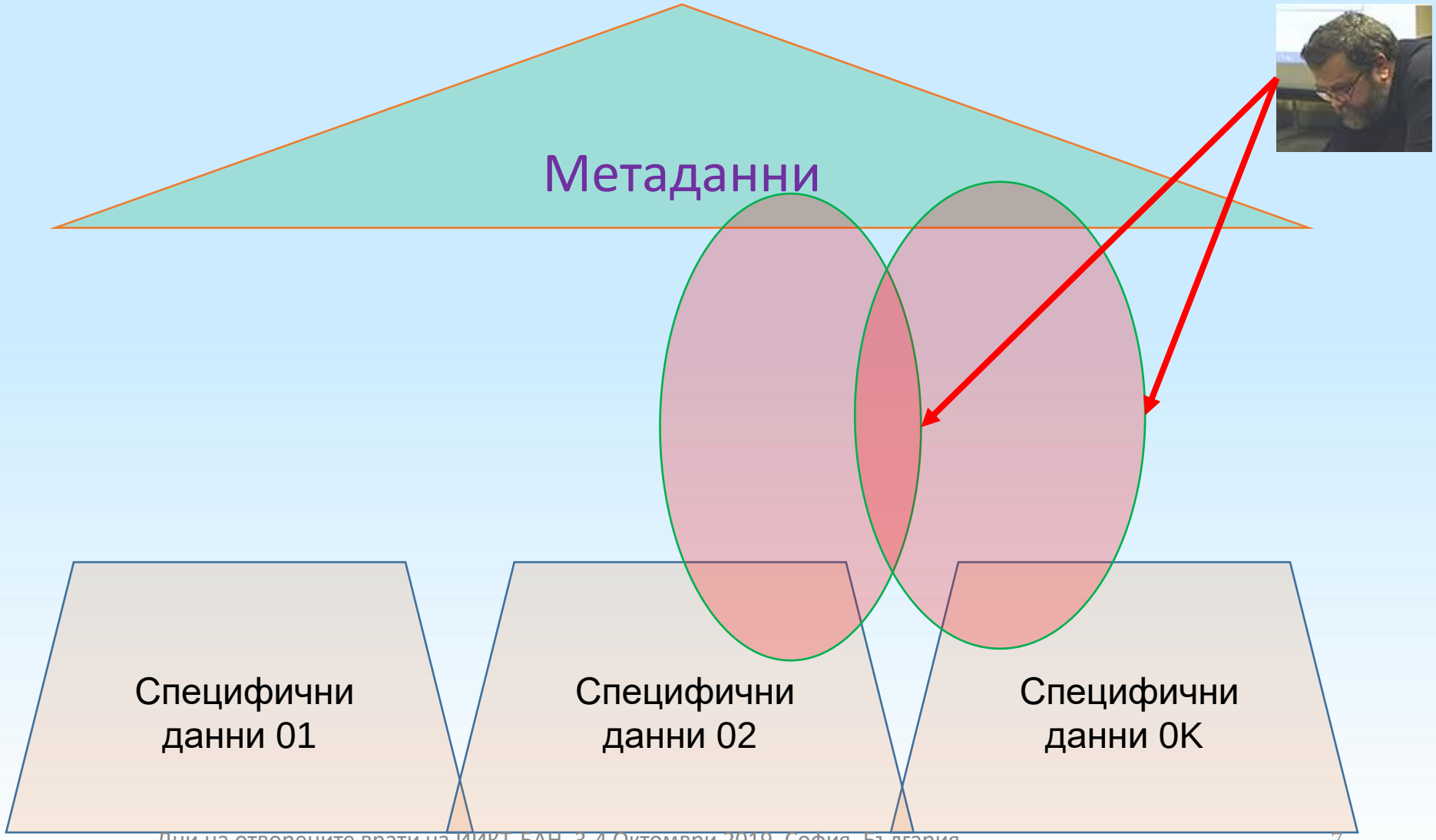
Изследователски процес в CH&X



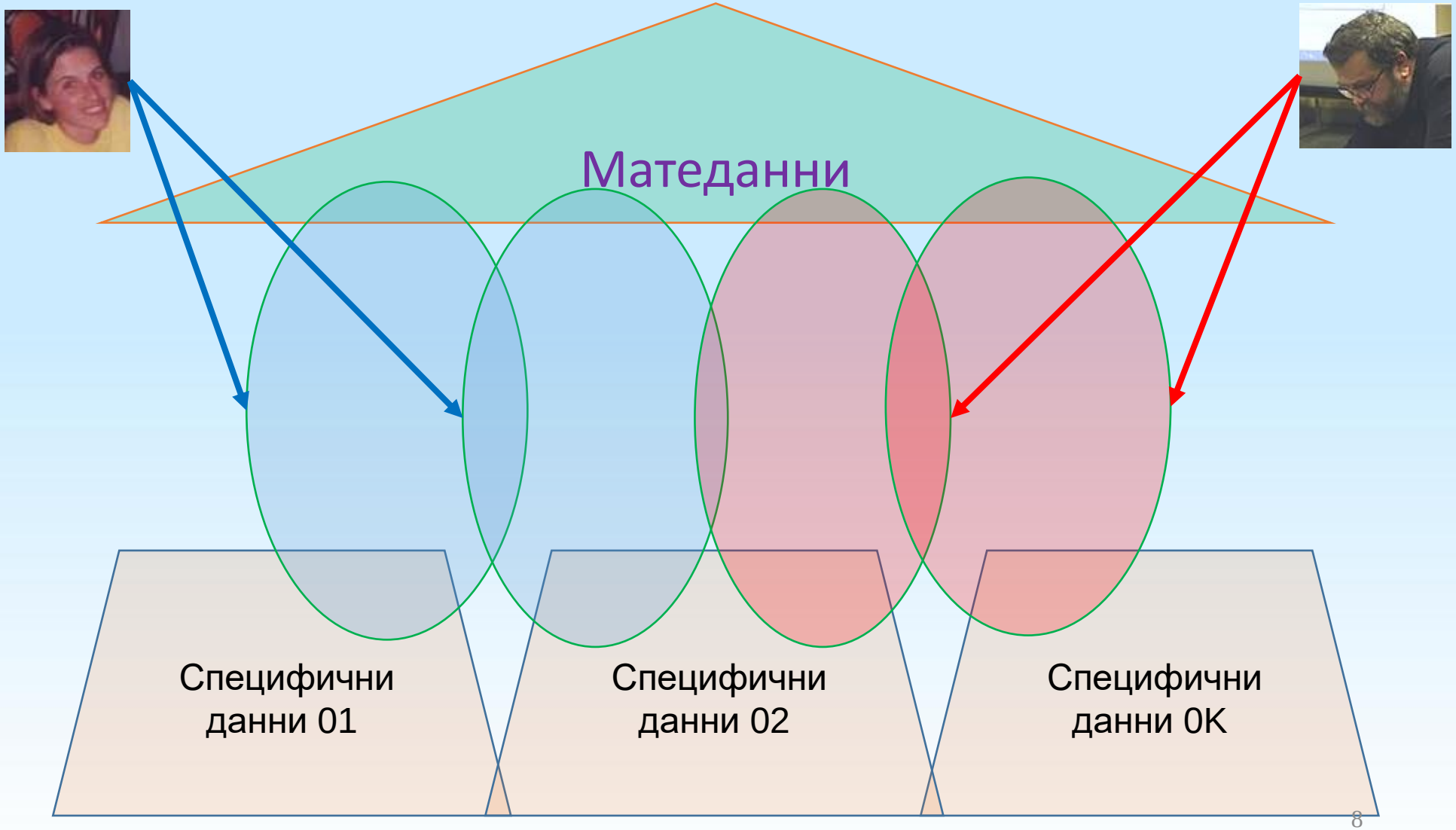
Изследователски процес в CH&X



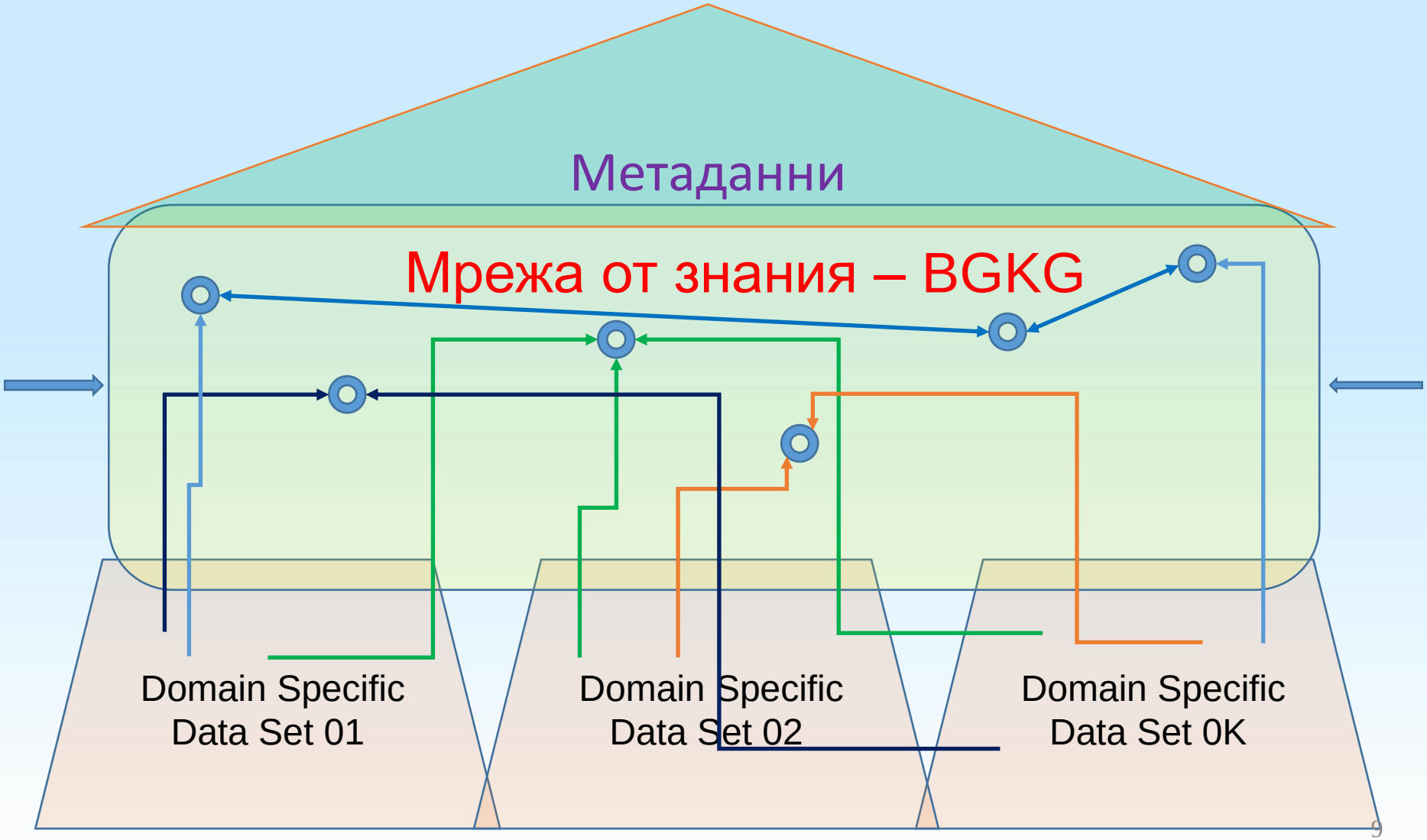
Изследователски процес в CH&X



Изследователски процес в СН&Х



Йерахия от данни в КЛаДА-БГ – BGKG



Контекстуализация в КЛаДА-БГ (Знанието)

Контекстуализацията в рамките на КЛаДА-БГ се разглежда като мрежа за взаимосвързани описания на хора, събития, обекти, документи, артефакти, автори, мнения и други.

- Хора – техните биографии – събития в живота им, роли
- Обекти – географски, артефакти и други
- Събития – място, време и участници, връзки с други събития
- Документи – автори, съдържание, мнения и отношение към хора, събития, обекти
- Текстът като основен източник на информация
- Технология – свързани бази от отворени данни (Linked Open Data) – граф знания (Knowledge graph)

Цел: Мрежа от знания за България

Езикови ресурси за извличане на знание

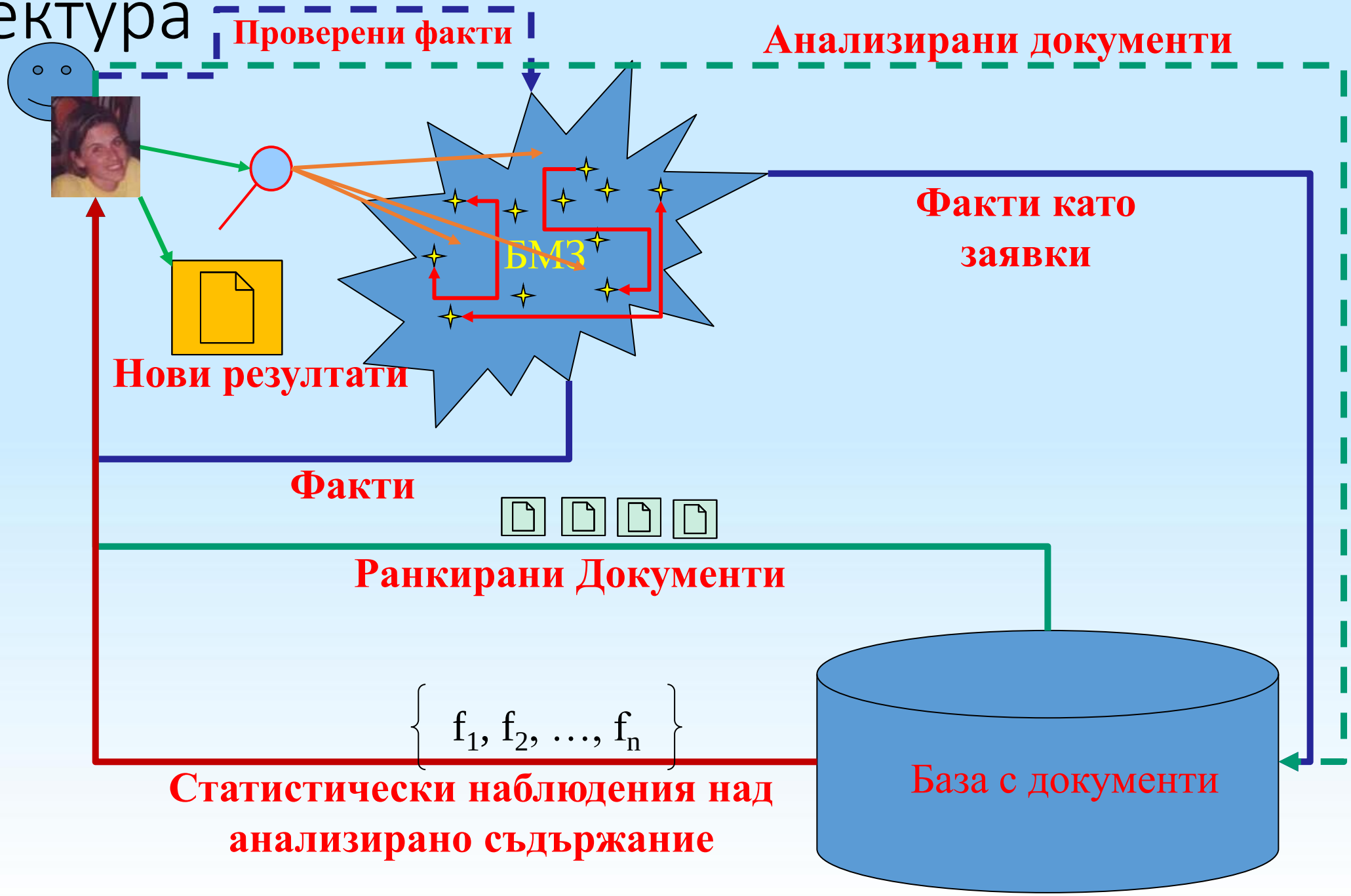
КЛаДА-БГ трябва да предостави езикови технологии за:

- Извличане и свързване на понятия/ключови думи
- Извличане и свързване на наименовани обекти
- Извличане и свързване на събития

Свързването е идентификация на едни и същи същности в различни текстове и между текст и знание в мрежата от знания

Лили Иванова пя на амфитеатъра в Пловдив

Архитектура

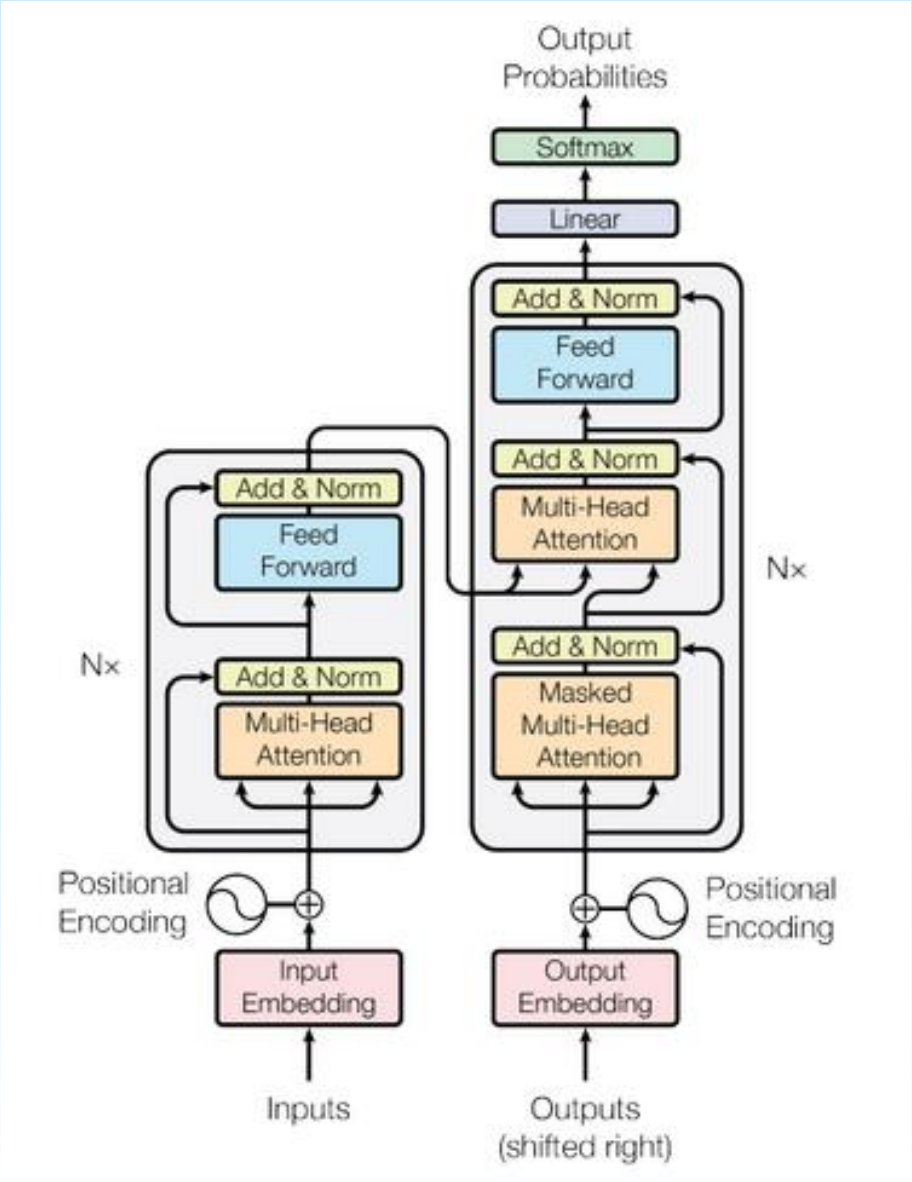


Технологии от ИИ в ДХ

- В рамките на КЛаДА-БГ интегрираме два елемента на изкуствения интелект:
  - Представяне на знанията в духа на онтологиите, свързаните данни и правенето на извод
  - Обработка на текстове на базата на системи за трениране и използване на невронни мрежи
- Връзка с високопроизводителните пресмятания
  - Осъществяване на бързо съпоставяне на структури на графи и правене на извод в областта на мрежите от знания
  - Паралелно трениране и използване на невронните мрежи

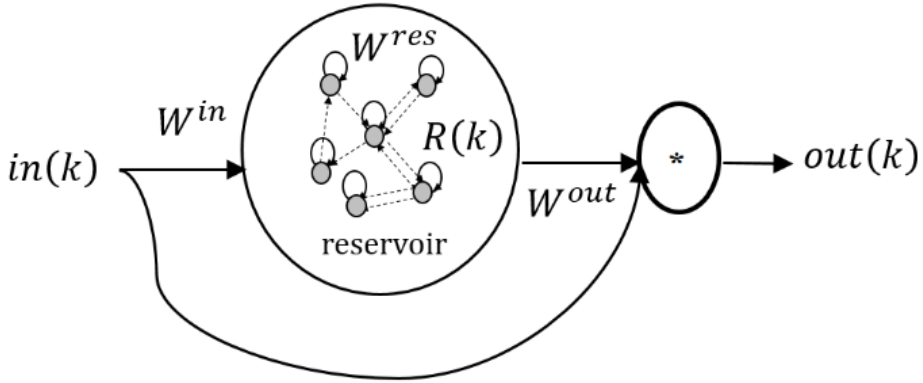
Конкретен пример: модел Transformer

- Архитектура за машинен превод, базирана на не-рекурсивни невронни мрежи със собствено внимание
- Особеност – паралелни обучителни нишки, допускащи голяма степен на паралелна обработка
- Сравнително лесни за предполагане тренировъчни данни, допускащи милиарди трениращи цикли
- Трудността и силата на модела идва от количеството параметри (в последните модели над 170 милиарда)



Собствени модели

- Собствени векторни представяния на български езикови обекти – думи, леми, комбинации текст и изображения
- Трансформери за системи с ехо състояния (echo state systems)
- Играят роля на системи за изчистване на характеристиките във векторното представяне на обектите
- Останалата част на архитектурата е като в оригиналния трансформер модел



Заключение

- В рамките на КЛаДА-БГ използваме две основни технологии от изкуствения интелект – мрежи от знания за запазване на информацията и архитектури, базирани на невронни мрежи за извличане на информация и правене на извод при използване на моделите
- Двете технологии използват услугите, предоставени от системите за високопроизводителните пресмятания
- Това е много важно особено в първите стъпки за натруване на знание във векторните представяния на различните обекти в моделите – думи, леми, фрази, характеристики за изображения и други