«Exercise Human-Centered Data Science»
Session #1

November 3rd, 2020

About

Teaching:

- Human-computer Interaction I - SoSe 2020
- Seminar Human centered machine learning - SoSe 2020
- Human-centered Data Science - WiSe 2020/21
- Proseminar Interactive Intelligent Systems - WiSe 2020/21

Consultation Hour:

- Mondays 14-15 → https://fu-berlin.webex.com/meet/alexa.schlegel

Agenda

1. General exercise guidelines and organizational infos
2. Preliminary schedule
3. Expectations
4. Introduction round
5. Tools and Datasets/APIs
6. Example: Pre-questionnaire (Likert scale)
7. First assignment
8. Outlook

1. General guidelines and organisational infos

Exercise guideline

1. Please address me as Alexa.
2. Ask a question directly (audio) or in the chat.
3. We will talk in English throughout the exercise. Is everybody okay with that?
4. Use GitHub issues for general questions (assignments, lecture, …). Please label your questions/issues. You can invent labels!
5. If you think something is wrong in an assignment please create an issue, so I can fix it! You can also open a pull-request and fix it yourself!
6. Personal questions via Email with prefix [HCDS].

Goals of the exercise

Discuss theoretical concepts presented in the lecture.

Examples

- Discuss videos/podcast/articles related to the lecture.

Practice methods and concepts introduced during the lecture.

Examples:

- Reproducibility Workflow
- Bias in Wikipedia
- MTurk Qualitative Study
- Implement Fairness measures

Concept of the exercise

There are two types of assignments:

- (1) **weekly written reflections** related to the additional lecture material (e.g. reading and reflecting on peer reviewed articles or watching video material, …)
- (2) **scheduled (programming) assignments** (e.g. analysis of a datasets* or implementing measures/concepts presented during the lecture, …)

* We will mainly focus on data that we query through Wikimedia APIs.

Active participation

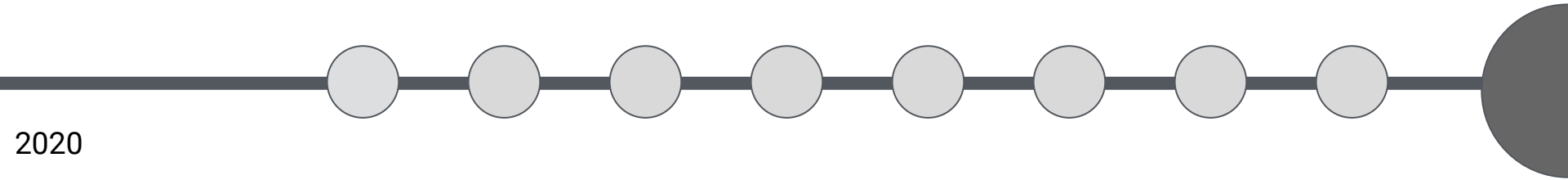
Your final grade is based on the result of your **written exam** only. But …

In order to actively participate in this course, you need to fulfil the following requirements:

- You need to submit **(n-1) written reflections.** [planned are 13]
- You need to submit **(n-1) scheduled (programming) assignments** and receive **60%** of the maximum possible scores in all exercise sheets. You can get max 10 points per sheet. [planned are 8]
- Each student has to present her/his solution at least **twice** during the exercise.***

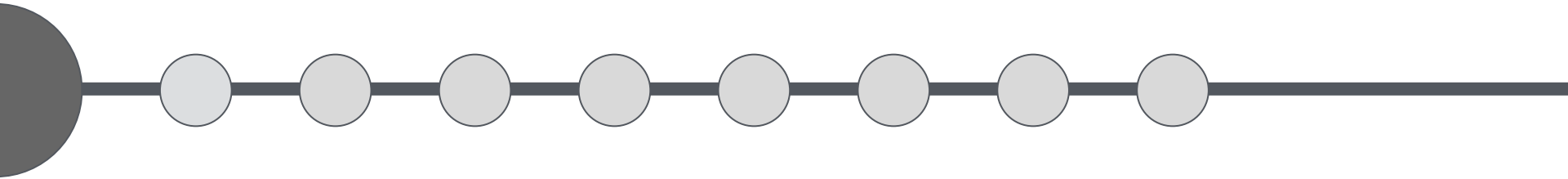
2. Preliminary Schedule & Active Participation

Preliminary Schedule



2020

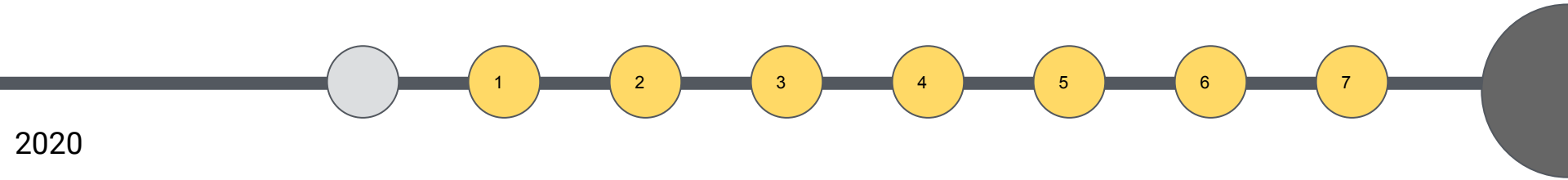
Christmas / New Year



Exam

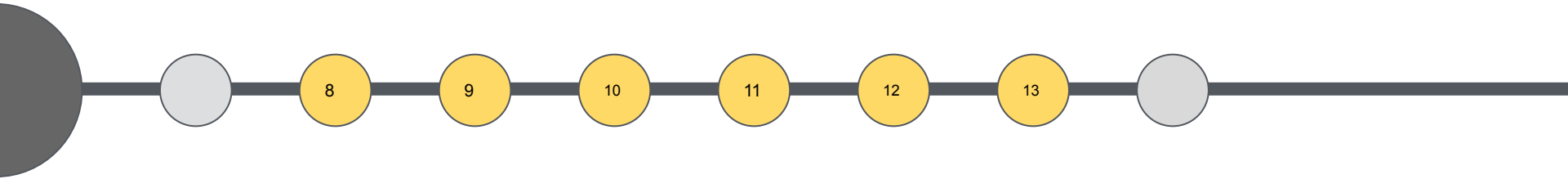
2021

Preliminary Schedule



2020

Christmas / New Year



Exam

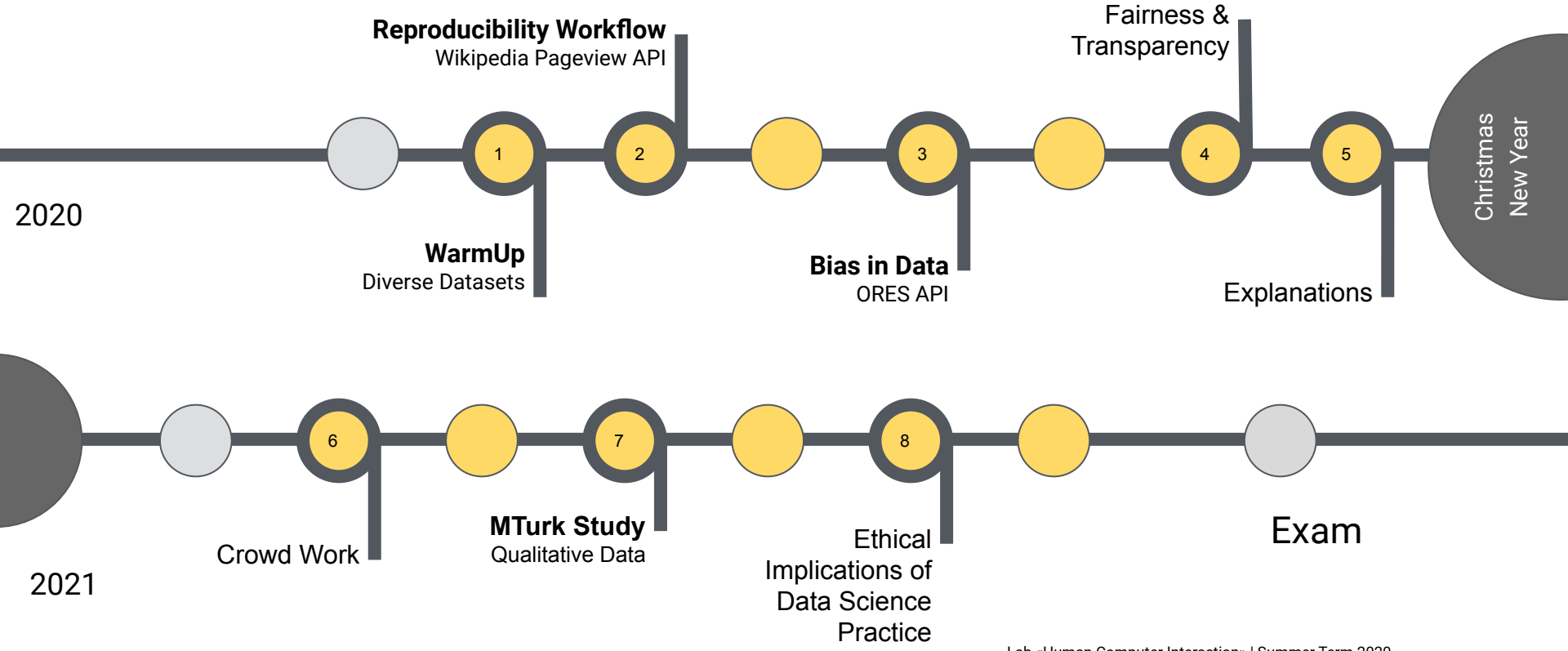
2021

Weekly written reflections

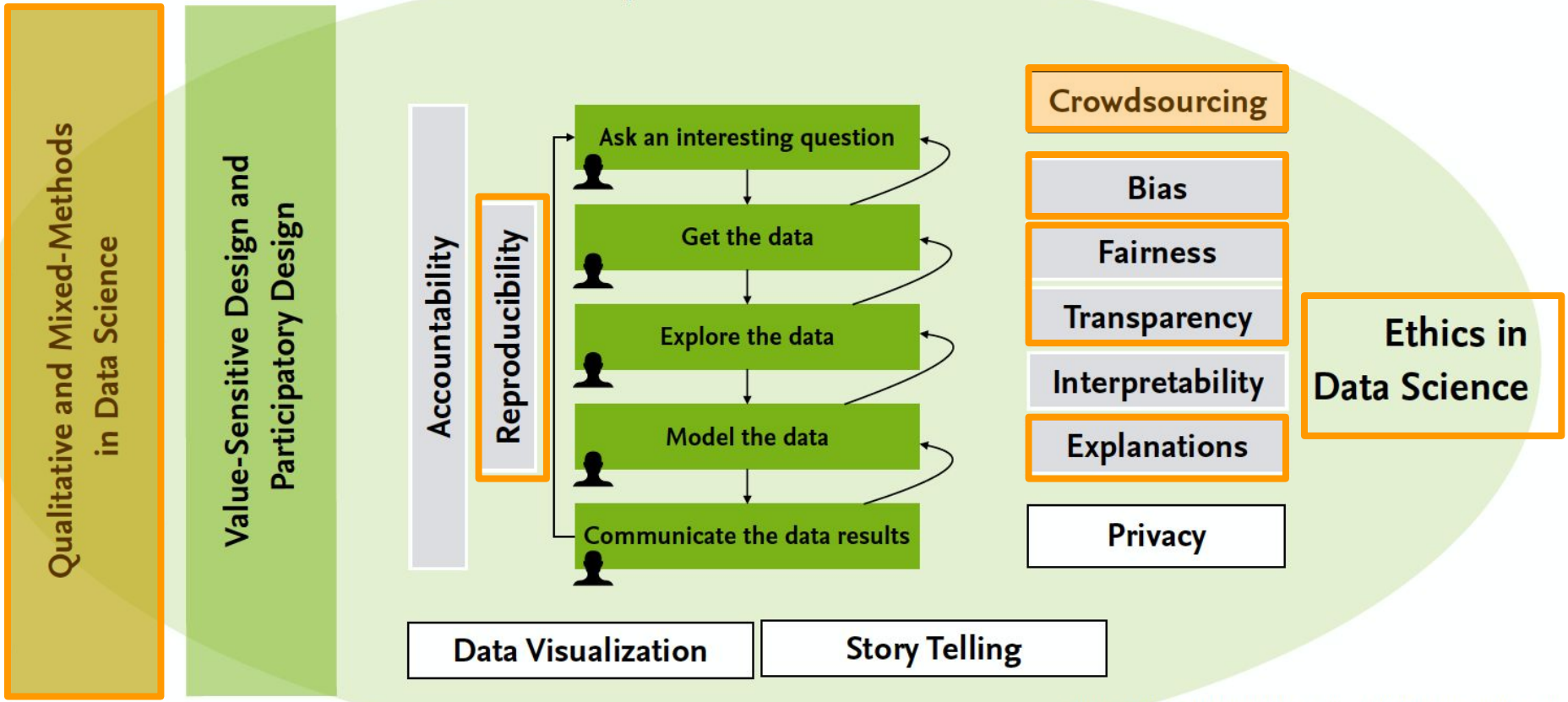
1. Listen to … | Read the article … | Watch the video …
2. In at least 2-3 full sentences, answer the question "How does this ... inform your understanding of human centered data science?".
3. Using full sentences, list at least 1 question that this ... raised in your mind, and say why it caused you to ask this question.

No need to get 12 out of 13 points. One point per reflection.

Preliminary Schedule



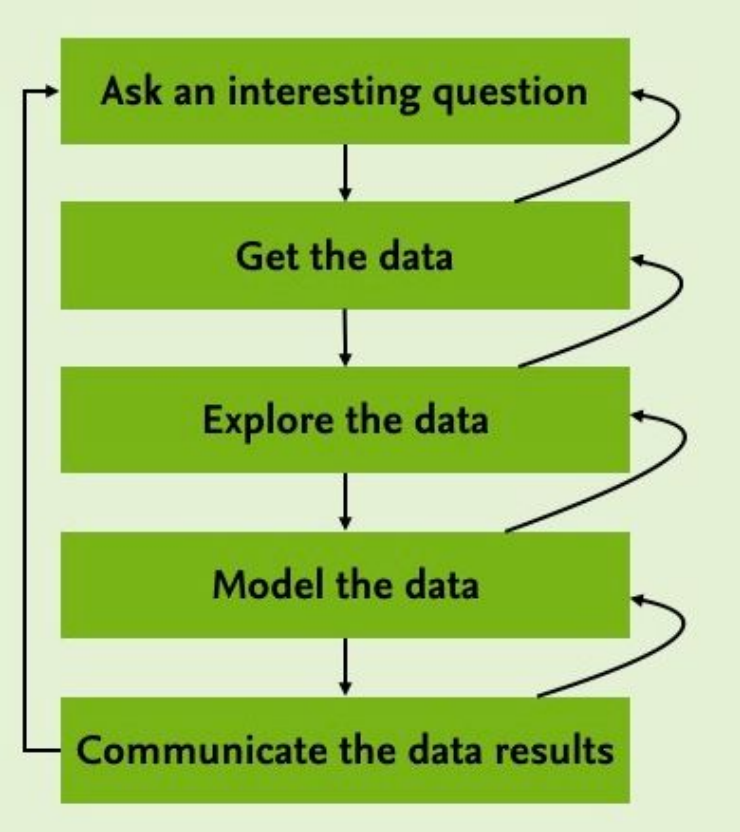
A Human-Centered Perspective on the Data Science Process



Scheduled (programming) assignments

Usually follow a data science workflow/steps:

1. Data Acquisition
   (Given dataset or collect data via API.)
2. Data Processing
3. Data Analysis
4. Documentation (Blogpost)



General structure of exercise

- Discuss (reading) reflections or group work.
- Presentations of current assignments by students.
- Feedback for last assignments and Q&A.
- Introduction next assignment. (Short Input talks.)
- Questions?

Freie Universität Berlin

3. Expectations

What do we expect from you?

- Basic Data Science experiance.
- Hand in programming assignments in form of a jupyter notebooks and blogpost before the exercise at 3PM.
- Read/Listen/Watch one thing per week.
- Hand in written reflections before the exercise.
- Participate in discussions.
- Share knowledge and experience!
- Help each other.
- Ask questions and give Feedback.

Freie Universität Berlin

4. Introduction Round

Freie Universität Berlin

5. Tools and Datasets/APIs

Tools

- **BBB** → exercise
- **HCC Website** → streaming
- **Discord** → questions during the lecture, but you can also use that channel to connect with people from the class
- **Whiteboard** → announcements (via Email)
- **GitHub (repo and wiki)** → everything else (slides, assignments, …)

Datasets and APIs

We mainly use data provided by Wikimedia:

- **Wikipedia Dumps**
  → https://dumps.wikimedia.org/
- **Wikimedia REST API** (e.g. Wikipedia pageview/traffic data)
  → https://en.wikipedia.org/api/rest_v1/
- **ORES API** (ML pipeline, quality control, vandalism)
  → https://ores.wikimedia.org/

Wikimedia REST API 1.0.0 OAS3

/api/rest_v1/?spec

This API provides cacheable and straightforward access to Wikimedia content and data, in machine-readable formats.

Global Rules

- Limit your clients to no more than 200 requests/s to this API. Each API endpoint's documentation may detail more specific usage limits.
- Set a unique `User-Agent` or `Api-User-Agent` header that allows us to contact you quickly. Email addresses or URLs of contact pages work well.

By using this API, you agree to Wikimedia's Terms of Use and Privacy Policy. Unless otherwise specified in the endpoint documentation below, content accessed via this API is licensed under the CC-BY-SA 3.0 and GFDL licenses, and you irrevocably agree to release modifications or additions made through this API under these licenses. See https://www.mediawiki.org/wiki/REST_API for background and details.

Endpoint documentation

Please consult each endpoint's documentation for details on:

- Licensing information for the specific type of content and data served via the endpoint.
- Stability markers to inform you about development status and change policy, according to our API version policy.
- Endpoint specific usage limits.

Terms of service

the Wikimedia Services team - Website

Apache2

Page content page content in different formats

Mobile mobile-friendly page content

Feed aggregated daily featured content

swagger https://ores.wikimedia.org/v3/spec/ Explore

ORES scoring interface

A RESTful API for scoring revisions (v3 paths). These paths provide access to a set of scoring models. This API primarily differs from v1 in that there is only one response document schema that any path returns. This response document contains 'error' information, 'warnings' and 'scores' structures.

There's also new functionality for returning 'feature' values used in scoring and 'inject'ing custom feature values for scoring.

scoring

Show/Hide | List Operations | Expand Operations

GET	/v3/precache	Precache scores of {revids} based on precaching configs.
GET	/v3/scores/	List available scoring contexts and models.
GET	/v3/scores/{context}	Score {revids} using multiple {models} in the same request.
GET	/v3/scores/{context}/{revid}	Score a {revid} using all available models.
GET	/v3/scores/{context}/{revid}/{model}	Score a single {revid} using {model}.

Implementation Notes

Provides a means of scoring {revid} using {model} in {context}.

Response Class (Status 200)

A JSON document containing scores

Model | Example Value


```
{
  "error": {
    "code": "string",
    "message": "string"
  },
  "warnings": [
    {
      "message": "string",
      "type": "string"
    }
```


ORES

ORES is a web service that provides machine learning as a service for Wik
like Wikipedia and Wikidata. The system is designed to help human editor
wiki-work and to increase their productivity by automating tasks like dete
and removing edits made in bad faith. ORES is developed by Wikimedia S
team that specializes in building transparent, auditable, open, and ethical
intelligence (AI) to support human decision-making.

ORES is intended to be used as a source of structured information by volu
developers and product developers at the Wikimedia Foundation and Wik
Deutschland. Most users access ORES via 3rd party tools like Huggle and
Special:RecentChanges on Wikimedia wikis. To access ORES scores, a sir
and a reference UI are available.

Scores API

There are three versions of the scoring API that differ slightly in their beha
the current and recommended version of the API. It supports all current f
(model information, feature injection, and threshold optimization). Version
are provided for backwards compatibility.

Version 3 (docs)

The current recommended version of the API.

6. Example: Evaluation of pre-questionnaire

Likert Scales

Jupyter Notebook → PreQuest.ipynb

Freie Universität Berlin

7. First Assignment

GitHub →
https://github.com/FUB-HCC/hcds-winter-2020/wiki/01_exercise#assignment

8. Outlook

Discuss podcast

Students present assignment

Introduction to Wikimedia PageView API

Questions