Welcome to
Computer Audition
(ECE 277/477, AME 277, CSC 264/464)

Zhiyao Duan

Professor of ECE, CS and SAE

University of Rochester
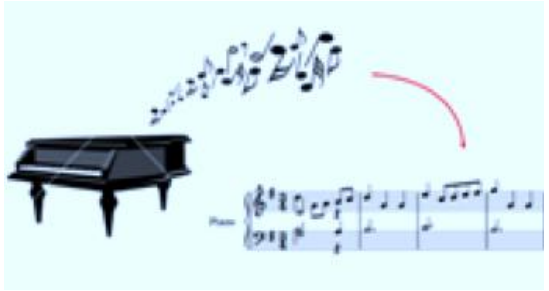
Brief Introduction



Grew up in Henan & Guangxi, China
BS & MS @ Tsinghua in Beijing, 2000-08
Exchange @ Stanford, 2007
Internship @ Microsoft Research Asia, 2007-08
PhD @ Northwestern, 2008-2013
Internship @ Adobe, 2011
Visiting Researcher @ Ohio State, 2013
Faculty @ UofR, 2013-present
Co-Founder @ Violy, 2017-present
Sabbatical leave @ Kwai, 2020-2022
President @ ISMIR, 2024-2025

Audio Information Research (AIR) Lab

Machine Understanding of Sounds



MUSIC INFORMATION RETRIEVAL

- Music transcription, alignment
- Source separation
- Generation
- Interactive performance



SPEECH PROCESSING

- Separation and enhancement
- Verification and anti-spoofing
- Emotion analysis
- Diarization
- Text-to-speech
- Voice conversion



ENVIRONMENTAL SOUND UNDERSTANDING

- Sound search by vocal imitation
- Sound event detection
- Source localization
- HRTF modeling
- Smart acoustics



AUDIO-VISUAL PROCESSING

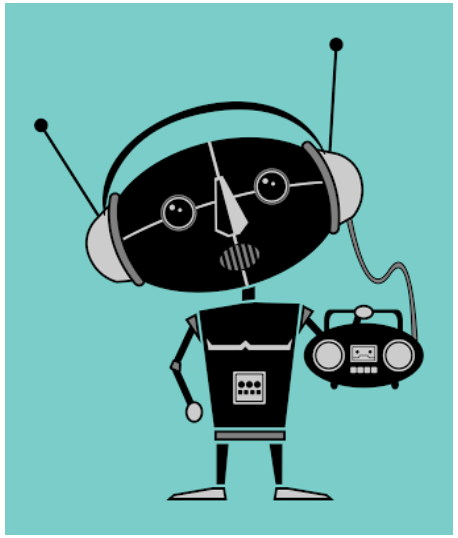
- Talking face generation
- Music performance analysis and generation
- Audio-visual source separation

Human Audition



- Understanding the environment
- Communication
- Entertainment

Computer Audition



- Understanding the environment
- Communication
- Entertainment – entertain human

Some Key Problems

- Sound source identification




- Source localization
- Content understanding
  - Speech, event, melody, rhythm




- Source separation



Tools for Sound Interaction



Create: Bone Flutes (7000 B.C.)

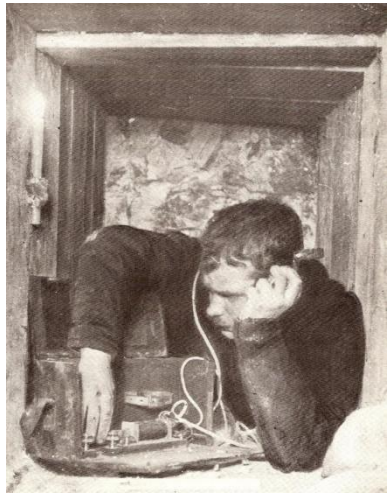


Modify: Delphi Theater (300 B.C.)

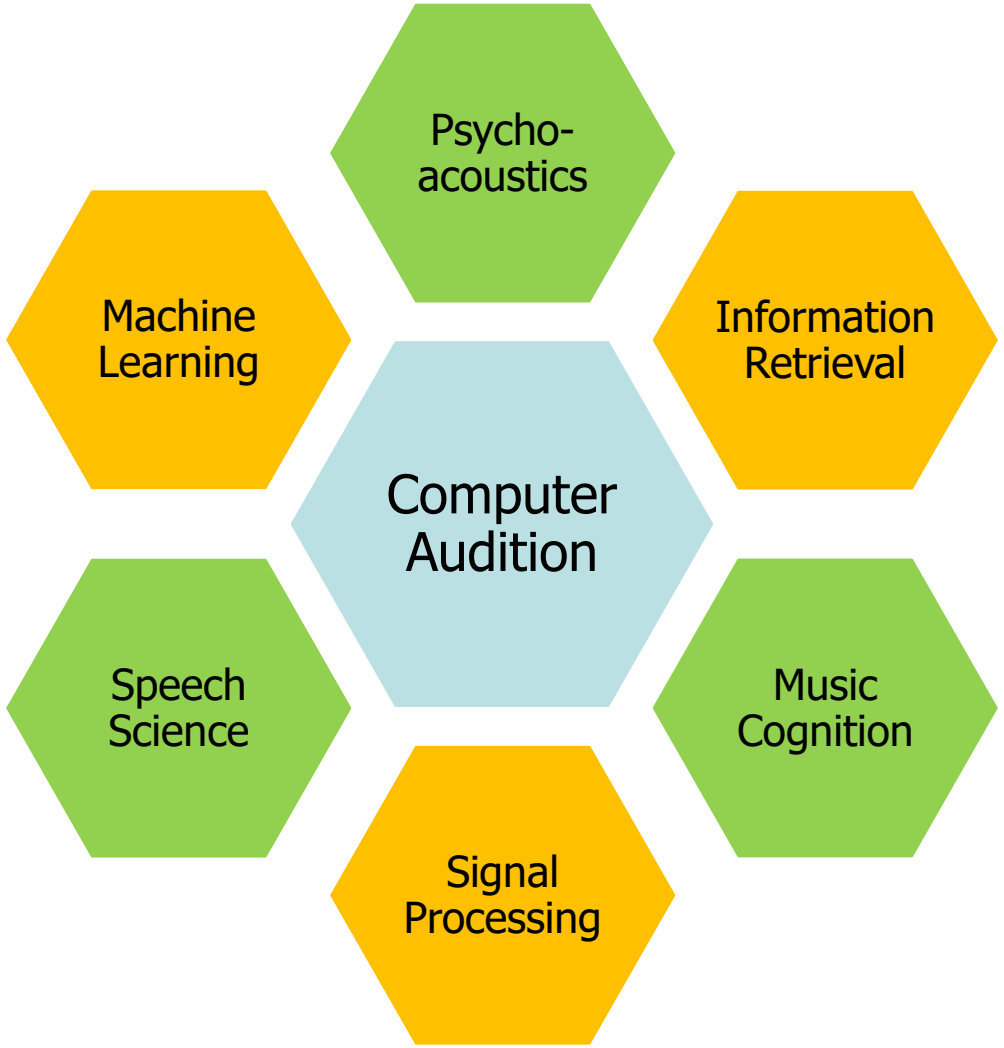


Record: Cylinder Phonograph (1899)

Transmit:
Crystal Radio
(1914)



Impact on Many Fields



Many Applications

SHAZAM

Spotify®

QQ Music

Violy



SoundHound
Instant Music Search and Discovery

amper

YOUSICIAN

pandora®

字节跳动
ByteDance

flowkey

gracenote
A NIELSEN COMPANY

magenta



Siri

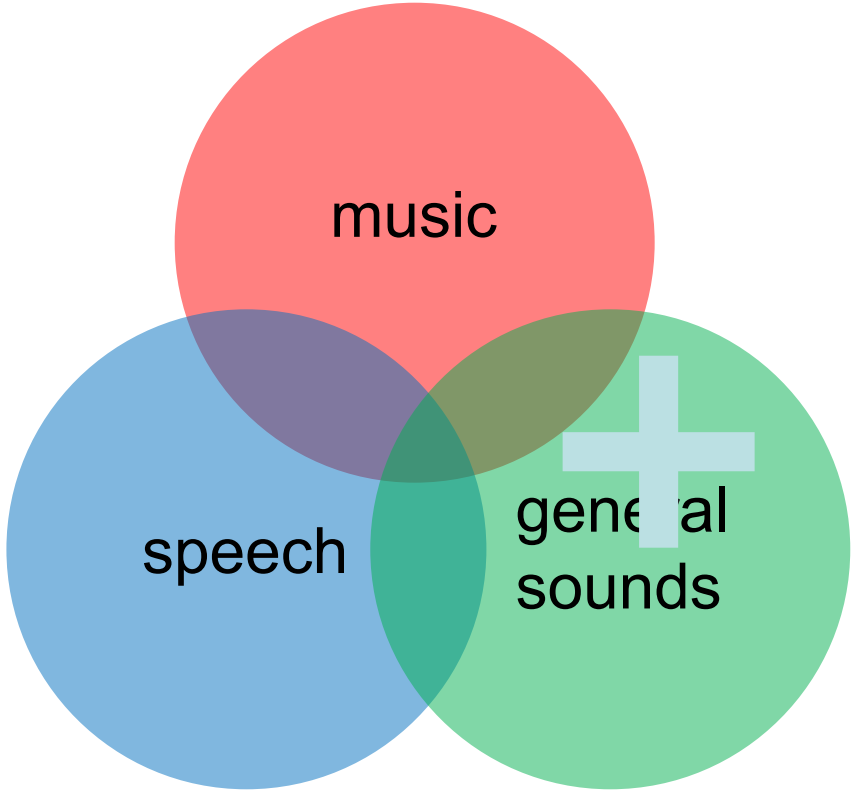
alexa

Google Home





Research Areas



vision

text

EEG

other modalities

Computer Audition Areas and Modalities

Music	Speech	General sound
User data (e.g., listening activity, EEG)	Style, emotion, expressiveness	Emotion, use case
Metadata (e.g., genre, artist, year)	Gender, speaker, diarization, context	Events, scene, context, location
Video recordings	Audiovisual speech	Video recordings
Audio recordings	Speech recordings	Audio recordings
Symbolic (e.g., sheet music, MIDI)	Text, transcripts	Event labels, captions

Some Demos – Music Transcription

Piano Transcription by Transkun 2.0

[Yan & Duan, ISMIR 2024] best paper nomination



Some Demos – Source Separation

- Violin/piano separation [Li, 2019]
  - Mixture:  violin:  piano: 
- Speech separation [Hershey et al., 2016]
  - Mixture:  female #1:  female #2: 
- More demos about speech separation
  - https://cslikai.cn/project/Pure-Audio/

Some Demos – Audio-Score Alignment

- Automatic Music Accompaniment
  - Music Plus One



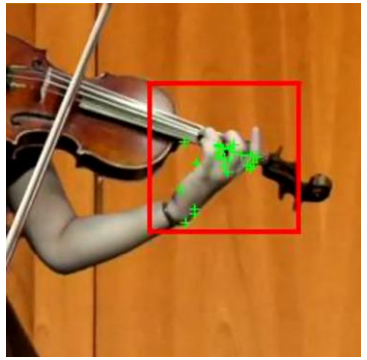
[Raphael, 2001]

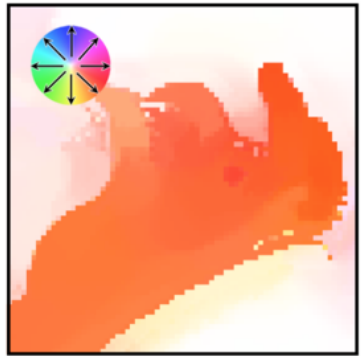
https://music.informatics.indiana.edu/~craphael/music_plus_one/movies/movies.html

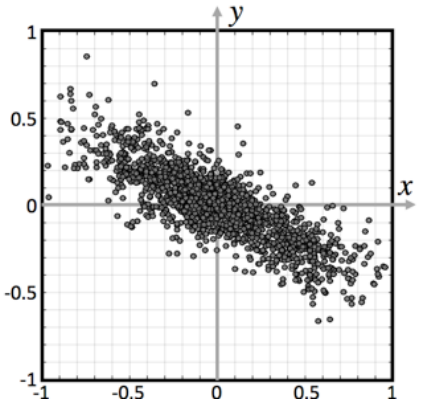
Some Demos – Audiovisual Analysis

- Audiovisual vibrato detection and analysis

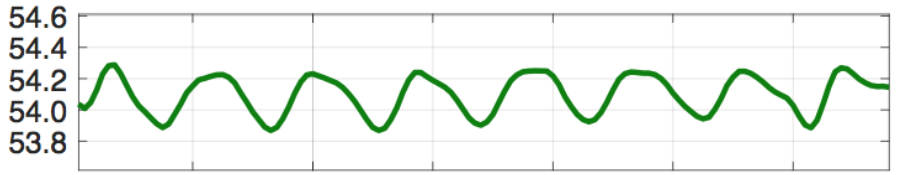




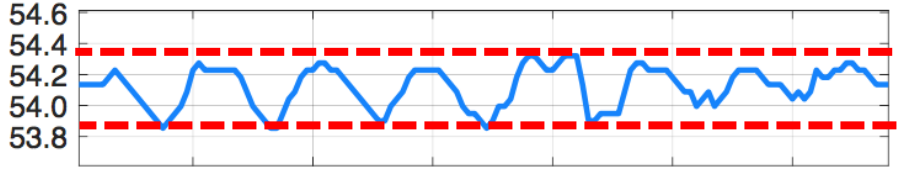




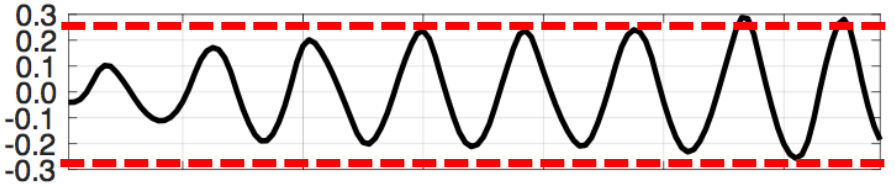
Ground-truth pitch contour



Estimated pitch contour from audio mixture



Motion curve
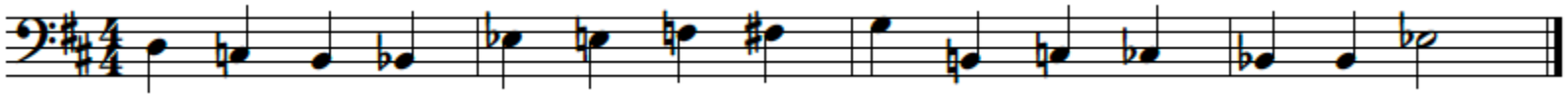


Results

- Vibrato detection accuracy significantly increased
- Very accurate vibrato rate and extent estimation

Bochen Li, Karthik Dinesh, Gaurav Sharma, and Zhiyao Duan, "Video-based vibrato detection and analysis for polyphonic string music," in *Proc. ISMIR*, 2017.

Some Demos – Symbolic Music Generation
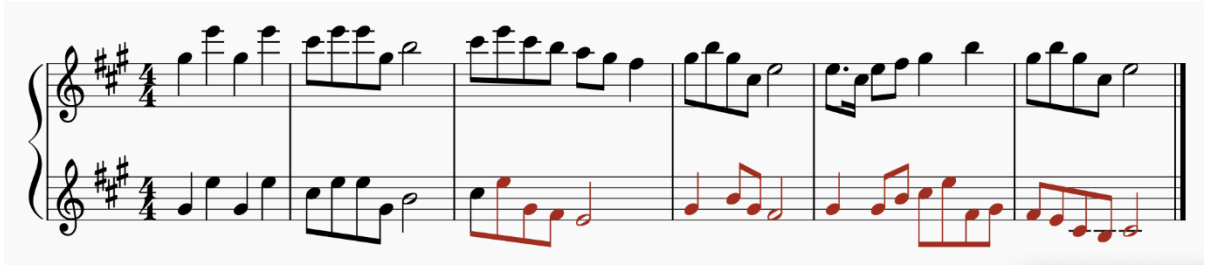
- Music harmonization [Yan, 2018]
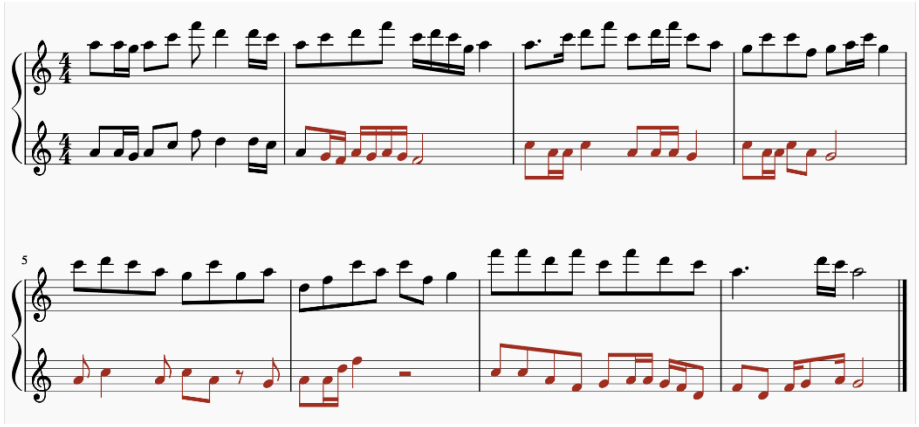


- Generation from scratch
  - String trio:

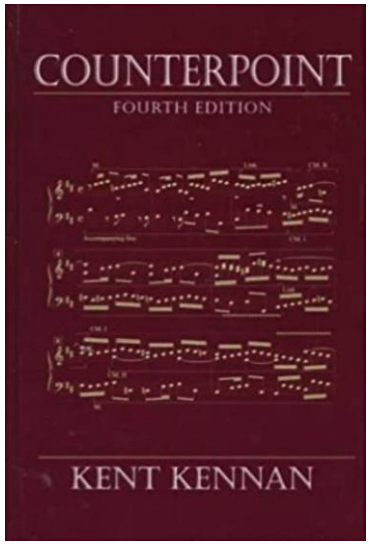
Some Demos – Symbolic Generation

- Countermelody Generation for Chinese Folk Melodies









Nan Jiang, Sheng Jin, Zhiyao Duan, and Changshui Zhang, ”When counterpoint meets Chinese folk melodies,” in *Proc. NeurIPS*, 2020.

Some Demos – Music Audio Generation

- Music audio generation
  - OpenAI's Jukebox: generating songs given lyrics, genre and artist
    - https://openai.com/research/jukebox
  - Google's MusicLM: generating music from text
    - https://google-research.github.io/seanet/musiclm/examples/
  - Suno: generating full songs from text
    - https://suno.com/

Some Demos – Speech Synthesis

- Text-to-Speech (TTS)
  - IBM Watson
  - https://www.ibm.com/demos/live/tts-demo/self-service/home
- Conversational TTS
  - Parakeet [Darefsky, Zhu, & Duan, 2024]
  - https://jordandarefsky.com/blog/2024/parakeet/
- Voice conversion
  - ControlVC: [Chen & Duan, 2023]
  - https://melissachen15.notion.site/melissachen15/ControlVC-Audio-Demo-dd0ea58c5b7f434a81af9cbcd67f56f6

Some Demos – Speech Editing

- Voice editing

VoCo: Text-based Insertion and Replacement in Audio Narration

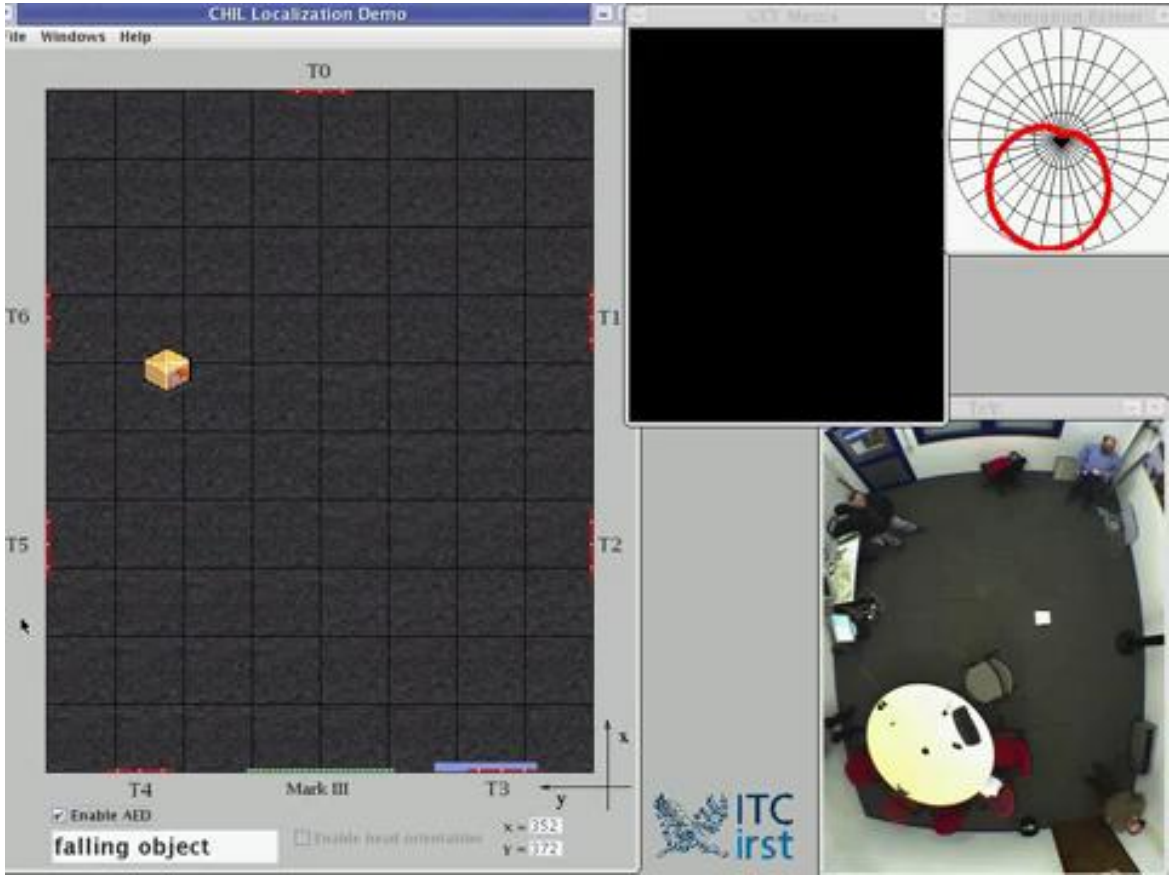
Zeyu Jin*, Gautham J. Mysore†, Stephen DiVerdi†, Jingwan Lu† and Adam Finkelstein*
* Princeton University † Adobe Research

SIGGRAPH 2017

https://www.youtube.com/watch?v=RB7upq8nzIU

Some Demos – Sound Event Analysis

- Acoustic event detection and localization



https://www.youtube.com/watch?v=iImkV6oKG_8

Some Demos – Sound Event Generation

- EDMSound: categorical sound event generation
  - https://agentcooper2002.github.io/EDMSound/
  - [Wen et al., 2023]

- Diffsound: text-to-sound generation
  - https://dongchaoyang.top/text-to-sound-synthesis-demo/
  - [Yang et al., 2023]

Some Demos – Audio-Visual Speech Processing

- Audio-visual speech separation [Afouras, 2018]
  - http://www.robots.ox.ac.uk/~vgg/demo/theconversation/demos/vox/0/demo.html

- Speech-driven talking face generation [Eskimez et al., 2020]



Trends on Techniques

- 2000s and before: signal processing, hand-crafted features, traditional machine learning (e.g., SVM, HMM, probabilistic models, EM algorithm)
- Late 2000s – early 2010s: Nonnegative Matrix Factorization (NMF), Probabilistic Latent Component Analysis (PLCA), factor analysis
- Mid 2010s till today: Deep learning (CNN, LSTM, CRNN, GANs, Transformers, Diffusion models)
- Emerging: Large Language Models (LLMs) for Audio, Audio Codec, Multimodal Language Models

Course Topics

- Fundamentals of human audition
- Auditory models
- Audio features (pitch, timbre, etc.)
- Audio modeling techniques
- State-of-the-art research topics
  - Polyphonic pitch analysis
  - Source separation
  - Sound identification
  - ……

Course Objectives

- General understanding of the field
- Deep understanding and hands-on research experience in a sub-field

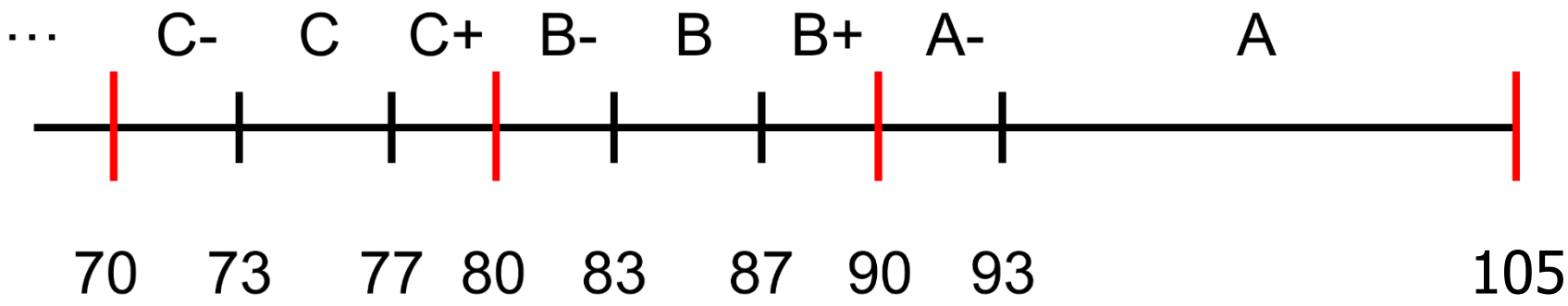
- Gain experience of the full cycle of research
- Able to think critically
- Improve presentation and writing skills

Assignments

- Total (105 points)
  - Class participation (5 points)
  - Homework (60 points)
    - HW1-HW4: Python/Matlab programming
    - HW5-HW6: Python programming for deep learning
  - Research Presentation (10 points)
  - Course project (30 points)
    - Proposal (5 points)
    - Status update (5 points)
    - Final report (10 points)
    - Presentation/demo (10 points)

- No exams

Grading



- No curve
- 200-level students get 10 points boost

Important Policies

- Do your own work
  - For HWs and paper reviews
    - You cannot copy code or writeups from other students or external resources. You cannot let others copy your code or writeup either.
    - If the assignment allows you to write your code on top of external code, you need to state clearly which part is written by you in your submission.
    - You cannot use AI tools like ChaptGPT in any component of the work.
  - For the final project
    - You are allowed to use external software packages and code in the implementation of your project, but you need to state clearly which part is your work and cite other work appropriately.
    - When writing the report, you cannot copy phrases with more than 5 consecutive words from other papers.
    - When paraphrasing other work, cite the work appropriately.
    - You are allowed to use AI tools in coding and polishing your final report and presentation, but you cannot use them to draft your final report or presentation. When you use them, you need to state clearly in your report how you used them.

Important Policies

- Late homework: 20% of full credit deduction / day
- Extension is possible for two types of reasons:
  - 1) illness and family emergency: typical extension is 3 days
  - 2) school related travels: extension can be as long as the travel length
- Please email me before the submission deadline if extension is needed

Important Policies

- Attendance
  - In class discussions are very helpful for understanding the course content
  - You are expected to actively participate in class discussions
  - There 5 points for class participation

Prerequisites

- Signal Processing
  - ECE 246/446 or ECE 272/472 or equivalent
- Python or Matlab programming
- Preferred but not required
  - Machine learning such as SVM, Markov models, neural networks, clustering, etc.

Three Websites

- Course website
  - All materials (slides, readings, assignments, etc.)
  - http://www.ece.rochester.edu/~zduan/teaching/ece477

- Blackboard
  - For announcements, homework submissions, and grade display

Ready? Let's Go!



Let's get to know each other first!