Inputs:

* Audio (N-tracks, each track “n” corresponds to speaker “n”)
  + We can use it to determine who’s talking, who’s talking next, who’s just a listener and etc.
* Scene (dynamic, everything changes with time):
  + Objects:
    - positions: (p\_k(t))
    - Interestingness (I\_k(t))
      * This can change with time
      * For example, an object might make a loud noise
    - Occlusion
  + Characters:
    - Positions (p\_n(t)) and bust direction (d\_n(t))
    - Existing animation (c\_n(t))
    - Look at target (L\_n(t))
      * Could be specified by the users. Used to determine the look-at pattern of the listener
* Speaker Tagging (temporal):
  + <emotion\> <emotion\>
    - high/low arousal
  + <Addressing N\> <Addressing N\>
    - When not addressing anyone in particular, then default to turn taking behaviour
    - Alternatively, they could be addressing an object
* Relationship Tagging (character relationships):
  + Interest (Matrix, value between 0 and 1)
    - 0 = do not wish to look
    - 1 = wish to look
    - 0.5 = neutral
  + Appearance (matrix, value between 0 and 1)
    - 0 = do not wish to appear to look at them
    - 1 = wish to be sincere
    - 0.5 = doesn’t care

Output:

* Saliency maps over time
* Look-at pattern of all the speakers
* Head and gaze animation for all the speakers in relation to the gaze

Optimization: