Towards Automated Large-Scale Evaluation of Psychotherapy

Nikolaos (Nikos) Flemotomos

University of Southern California
Department of Electrical and Computer Engineering
Signal Analysis and Interpretation Laboratory

Forensic Grand Rounds
September 29, 2021

Outline

- Psychotherapy Evaluation and Behavior Coding
- Evaluation of Motivational Interviewing: coreMI
- Evaluation of Cognitive Behavior Therapy
- Potential Improvements and Ethical Implications

Outline

- Psychotherapy Evaluation and Behavior Coding
- Evaluation of Motivational Interviewing: coreMI
- Evaluation of Cognitive Behavior Therapy
- Potential Improvements and Ethical Implications

Why do we need to evaluate psychotherapy?

- lifetime prevalence of diagnosable mental disorders: more than 50%
- about 1 in 7 adults receives mental health services annually

Need for quality assurance

- more effective training
- more efficient supervision
- more positive clinical outcomes

Photo by cottonbro from Pexels

Why do we need to evaluate psychotherapy?

- lifetime prevalence of diagnosable mental disorders: more than 50%
- about 1 in 7 adults receives mental health services annually

Need for quality assurance

- more effective training
- more efficient supervision
- more positive clinical outcomes

- Essential for improved performance: feedback to the therapist
  1. client progress monitoring
  2. performance-based feedback

Photo by cottonbro from Pexels

Why do we need to evaluate psychotherapy?

- lifetime prevalence of diagnosable mental disorders: more than 50%
- about 1 in 7 adults receives mental health services annually

Need for quality assurance

- more effective training
- more efficient supervision
- more positive clinical outcomes

- Essential for improved performance: feedback to the therapist
  1. client progress monitoring
  2. performance-based feedback

Photo by cottonbro from Pexels

Behavioral coding

- psychotherapy: intervention based on spoken language
  $\Rightarrow$ quality encoded in therapists' and patients' speech/language characteristics

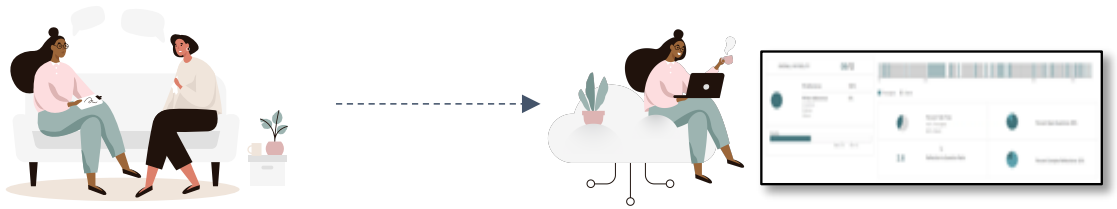
- quality assessment traditionally addressed by human raters using recorded sessions
  - time consuming
  - cost prohibitive

Behavioral coding

- psychotherapy: intervention based on spoken language
  ⇒ quality encoded in therapists' and patients' speech/language characteristics

- quality assessment traditionally addressed by human raters using recorded sessions
  - time consuming
  - cost prohibitive

⇒ *computational methods for automatic evaluation*



Outline

- Psychotherapy Evaluation and Behavior Coding
- Evaluation of Motivational Interviewing: coreMI
- Evaluation of Cognitive Behavior Therapy
- Potential Improvements and Ethical Implications

Behavioral coding in Motivational Interviewing

Motivational Interviewing Skill Code (MISC):

1. session-level codes (5-point Likert scale)
2. utterance-level codes (classification into behavior categories)

Behavioral coding in Motivational Interviewing

Motivational Interviewing Skill Code (MISC):

1. session-level codes (5-point Likert scale)
2. utterance-level codes (classification into behavior categories)

name	high score means that counselor...
acceptance	consistently communicates acceptance and respect to the client
empathy	makes an effort to accurately understand the clients perspective
direction	is focused on a specific target behavior
autonomy support	does not attempt to control the clients behavior or choices
collaboration	interacts with their clients as partners
evocation	tries to “draw out” client’s own desire for changing

- MI spirit = average(evocation, collaboration, autonomy support)

Behavioral coding in Motivational Interviewing

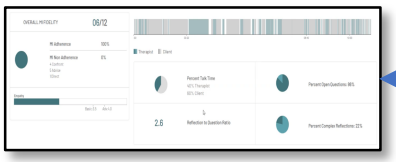
Motivational Interviewing Skill Code (MISC):

1. session-level codes (5-point Likert scale)
2. utterance-level codes (classification into behavior categories)

abbr.	name	example
ADP	Advise w/ Permission	Would it be all right if I suggested something?
ADW	Advise w/o Permission	I recommend you attend 90 meetings in 90 days.
AF	Affirm	Thank you for coming today.
CO	Confront	(C: I don't feel like I can do this.) Sure you can.
DI	Direct	Get out there and find a job.
EC	Emphasize Control	It is totally up to you whether you quit or cut down.
FA	Facilitate	Uh huh. (*keep-going acknowledgment*)
FI	Filler	Nice weather today!
GI	Giving Information	Your blood pressure was elevated [...] this morning.
QUO	Open Question	Tell me about your family.
QUC	Closed Question	How often did you go to that bar?
RCP	Concern w/ Permission	Could I tell you what concerns me about your plan?
RCW	Concern w/o Permission	That doesn't seem like the safest plan.
RES	Simple Reflection	(C: Court sent me here.) That's why youre here.
REC	Complex Reflection	(C: Court sent me here.) It wasn't your choice to be here.
RF	Reframe	(C: something else comes up [...]) You have clear priorities.
SU	Support	I'm sorry you feel this way.
ST	Structure	Now Id like to switch gears and talk about exercise.
WA	Warn	Not showing up for court will send you back to jail.
NC	No Code	You know, I (*meaning is not clear*)

Automating psychotherapy quality assessment

coreMI



Utterance-level predictions

Session-level predictions

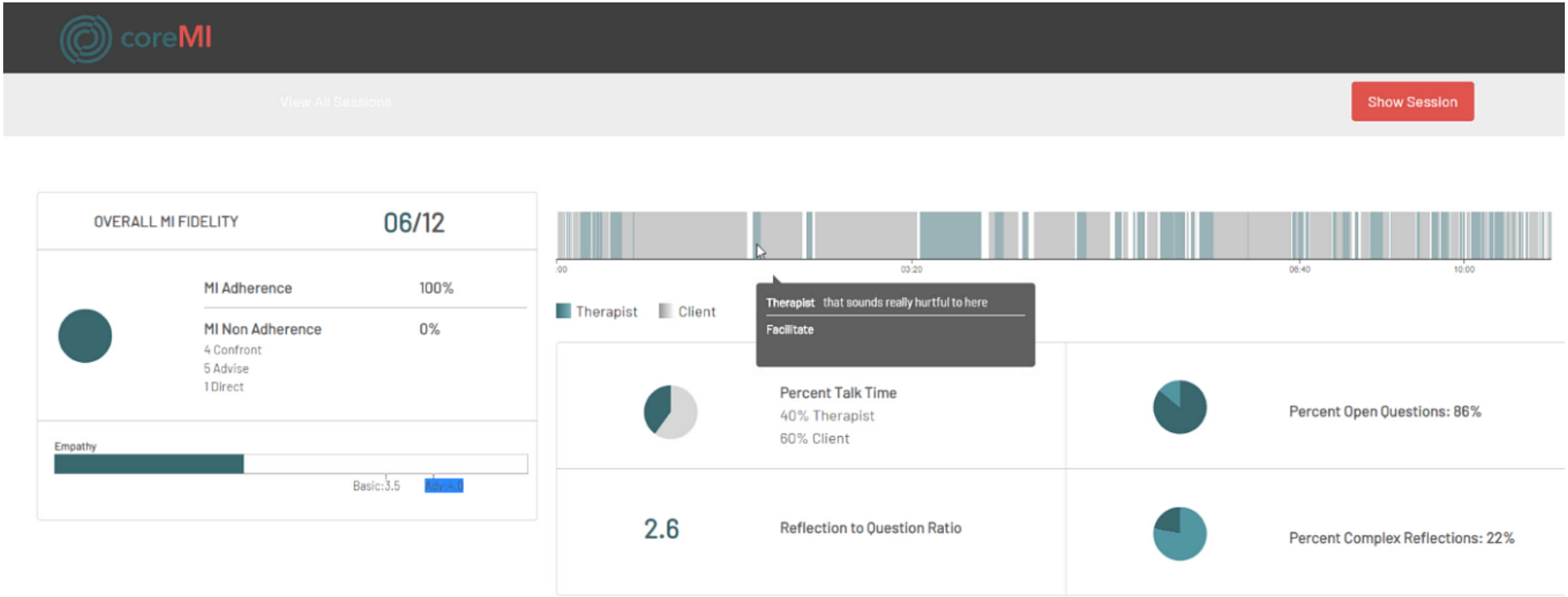
Audio Feature Extraction

Rich Transcription

Quality Assurance

oops
something went wrong

coreMI: feedback report



main features:

- session timeline with MISC-coded utterances
- session-level codes
- summary indicators and session dynamics
- overall MI fidelity
  (function of empathy, MI spirit, Re2Qu ratio, %QUO, %REC, MI adherence)

coreMI: review session

coreMI

View All Sessions

Show Report

<< Back to My Sessions

MI 1 MI 1 and Practice Training 10/31/2018 - 2:34 PM

05:35 13:18

Enter note here...

Save Note at 5 min 35 sec

Discussion

11/20/2018 - 4:03 PM - Edit Delete

JAKE V

I am making a general comment

11/20/2018 - 4:26 PM - Edit Delete

Transcript

mom

PATIENT

THERAPIST um mhm i don't know how the outfield pick up the laughing

PATIENT oh no um let me come to visit time always happy and stuff

THERAPIST so what about the relationship are you happy with what what do you like about it

PATIENT i like the company like i like that flip or annoyed that like you'll be there for me if i need it

THERAPIST uh so it's kinda like a level of intimacy that you don't really have another place

PATIENT yeah yeah right yeah and i don't really like and i and i'm like told him stuff i've never told other people before yeah

THERAPIST so so you like the you can find a support from other people your mom

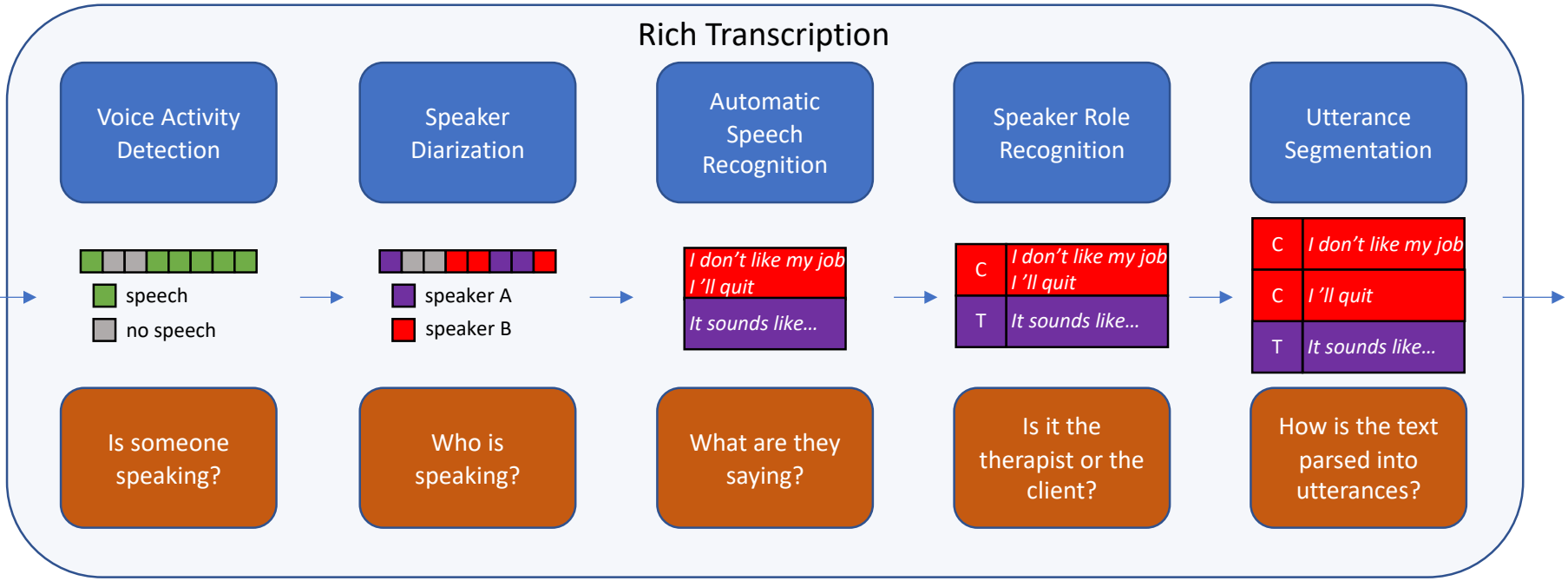
PATIENT not the same

THERAPIST not

PATIENT not the same

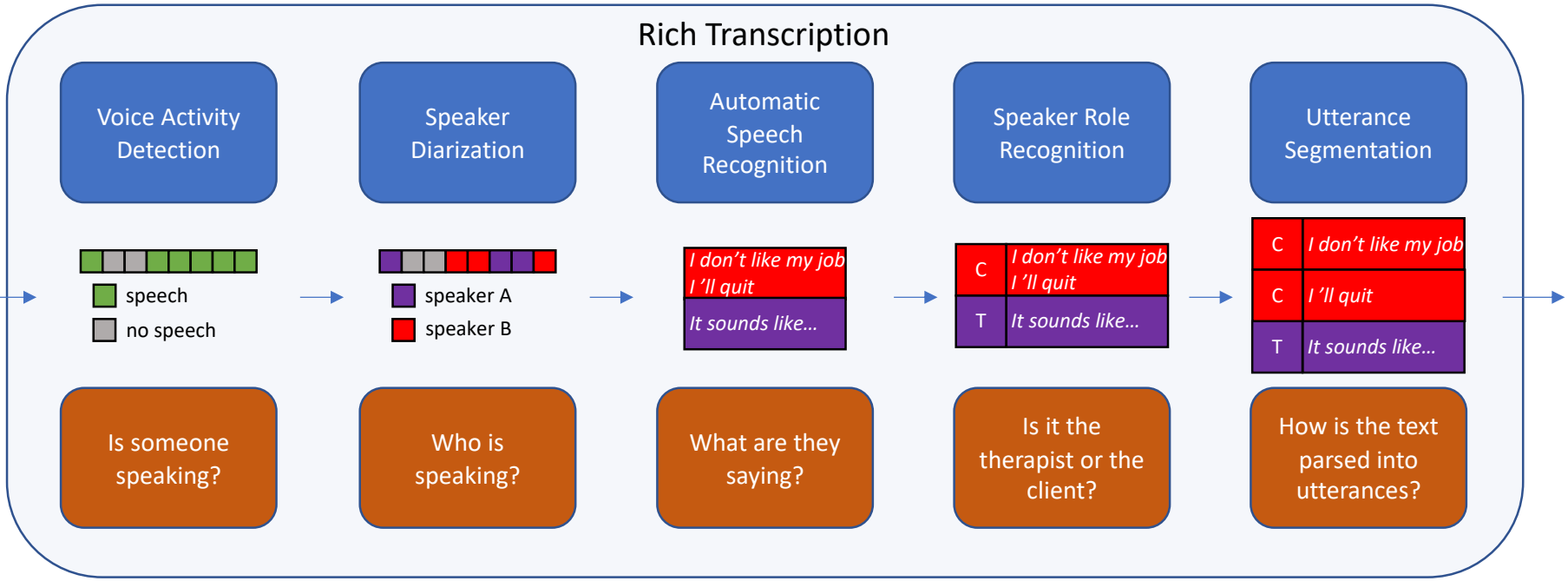
Rich transcription pipeline

- Our algorithms for automatic behavior coding are based on linguistic information (text)
- How do we get text from audio recordings?



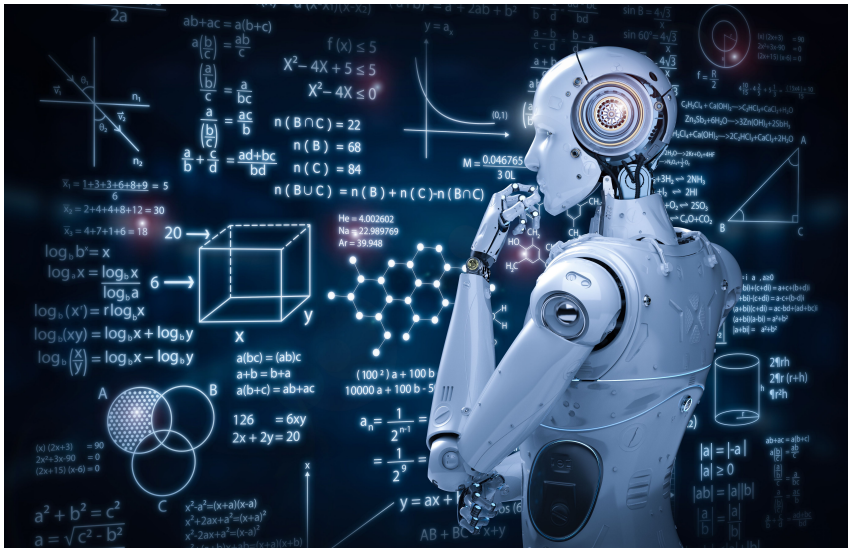
Rich transcription pipeline

- Our algorithms for automatic behavior coding are based on linguistic information (text)
- How do we get text from audio recordings?



- once we have the transcripts...
  ... employ quality safeguards to avoid problematic reports
  ... apply machine learning techniques towards behavior coding

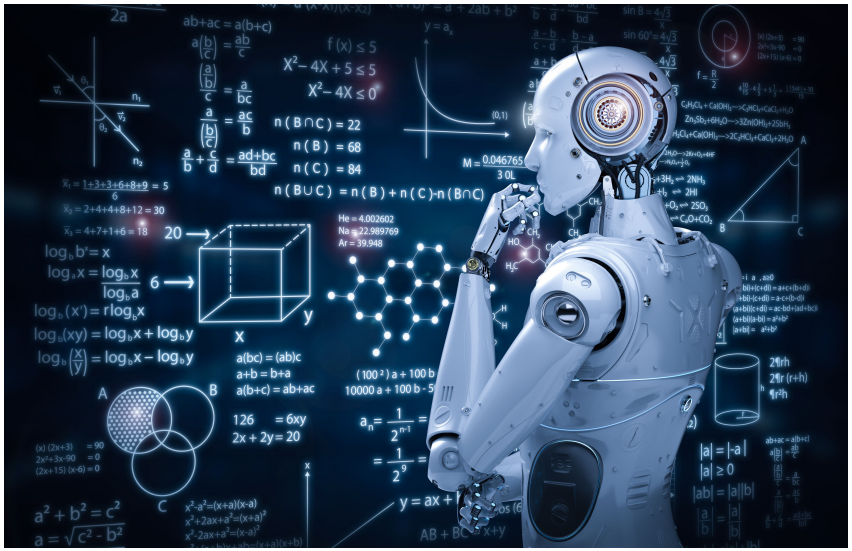
Machine Learning in one slide



- subfield of artificial intelligence and data science
- getting machines take decisions and solve problems...
- ...without being explicitely programmed to do so
  $\neq$ knowledge-based approaches

Photo by Phonlamai Photo from shutterstock

Machine Learning in one slide



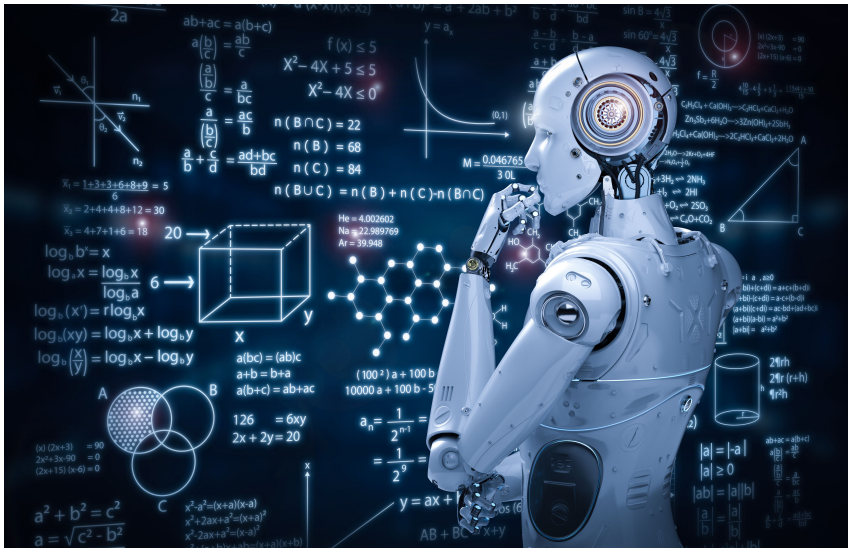
- subfield of artificial intelligence and data science
- getting machines take decisions and solve problems...
- ...without being explicitely programmed to do so
  $\neq$ knowledge-based approaches

Major paradigms

1. unsupervised learning (e.g., clustering)
   find underlying structure in given data
2. supervised learning (e.g., classification, regression)
   find mapping function between input data and outcome

Photo by Phonlamai Photo from shutterstock

Machine Learning in one slide



- subfield of artificial intelligence and data science
- getting machines take decisions and solve problems...
- ...without being explicitely programmed to do so
  $\neq$ knowledge-based approaches

Major paradigms

1. unsupervised learning (e.g., clustering)
   find underlying structure in given data
2. supervised learning (e.g., classification, regression)
   find mapping function between input data and outcome

- learn ML model on "training data"; evaluate on "test data"

Photo by Phonlamai Photo from shutterstock

Deployment and data collection

- deployed in university-based counseling center
- common topics
  - depression
  - anxiety
  - substance use
  - relationship concerns

Photo by Prostock-studiofrom shutterstock

Deployment and data collection

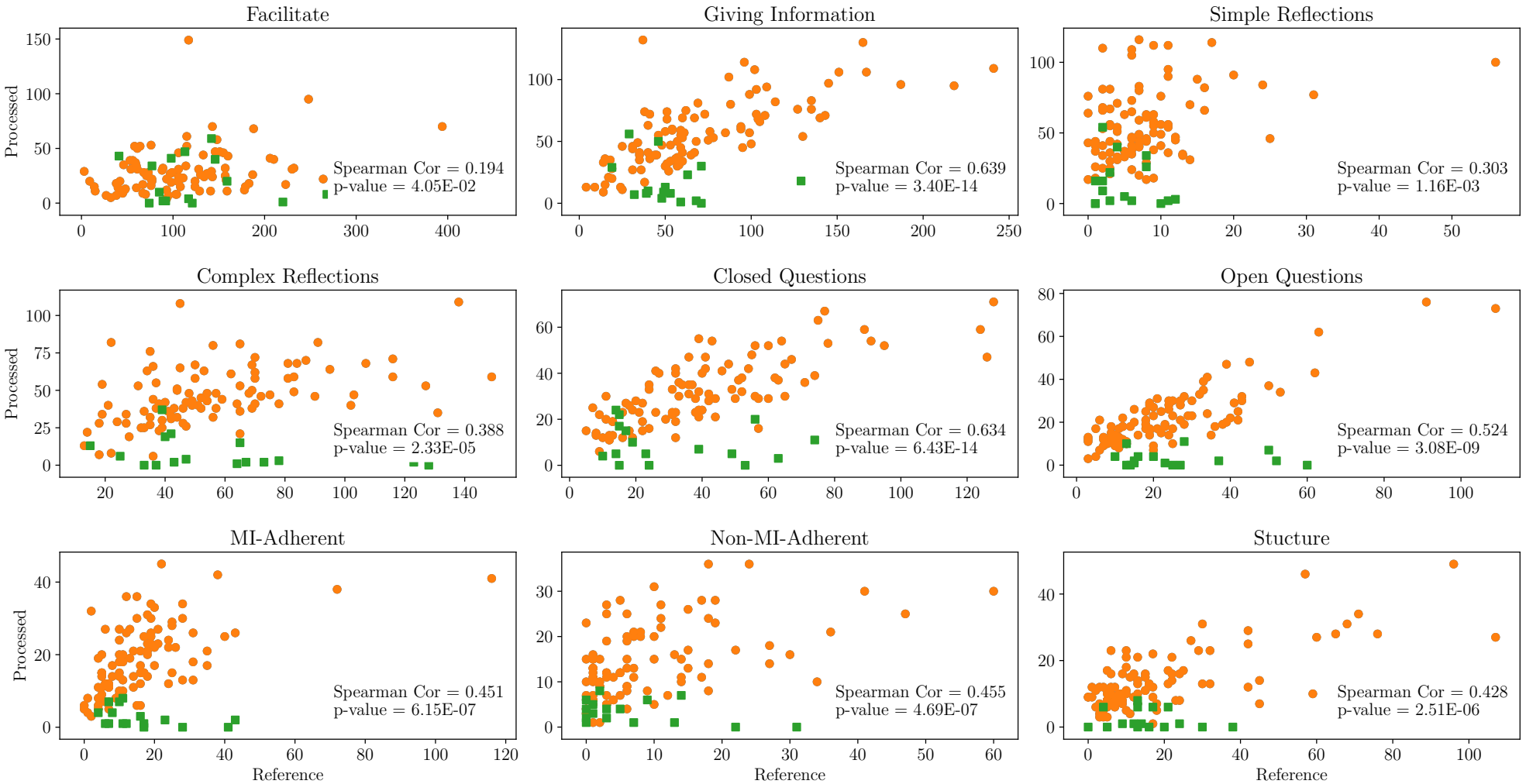
- deployed in university-based counseling center
- common topics
  - depression
  - anxiety
  - substance use
  - relationship concerns

data collection

- Sep 2017 – Mar 2020 (then, COVID-19...)
- $>$ 5000 recordings; 4268 passed our quality safeguards
  - 59 therapists, 1040 clients
  - $>$ 2.8M utterances, 28M words
  - mean duration: $\sim$50min
- professional transcription and coding: 188 sessions

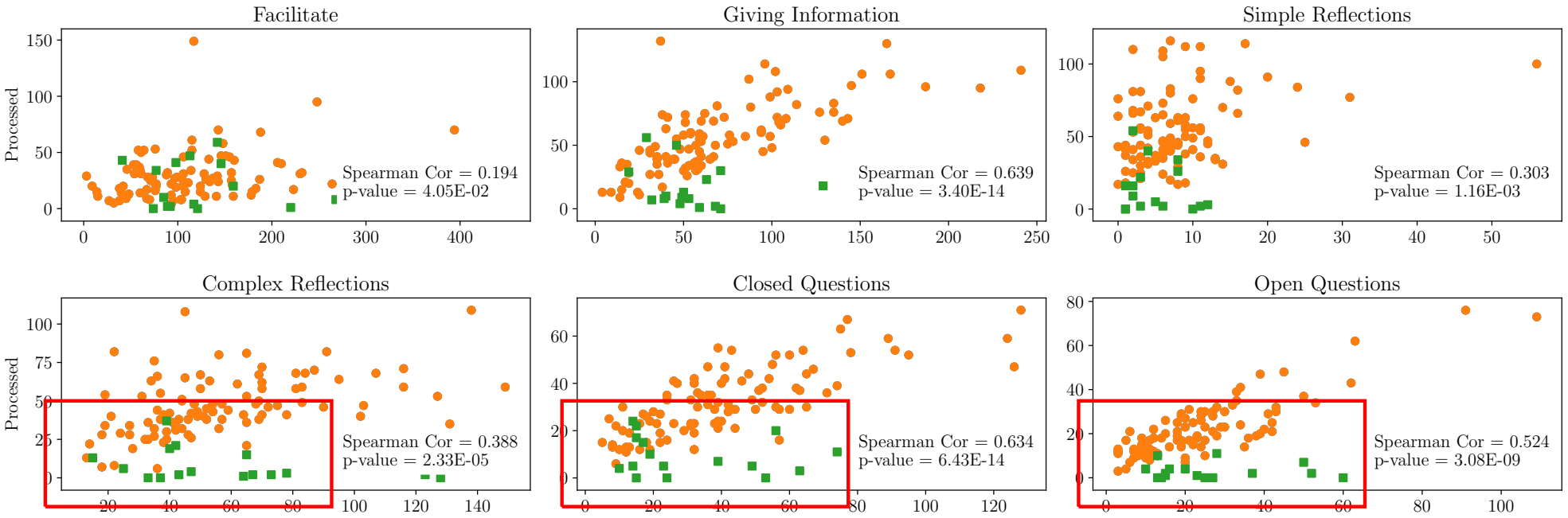
Photo by Prostock-studiofrom shutterstock

Quantitative assessment: Utterance-level codes



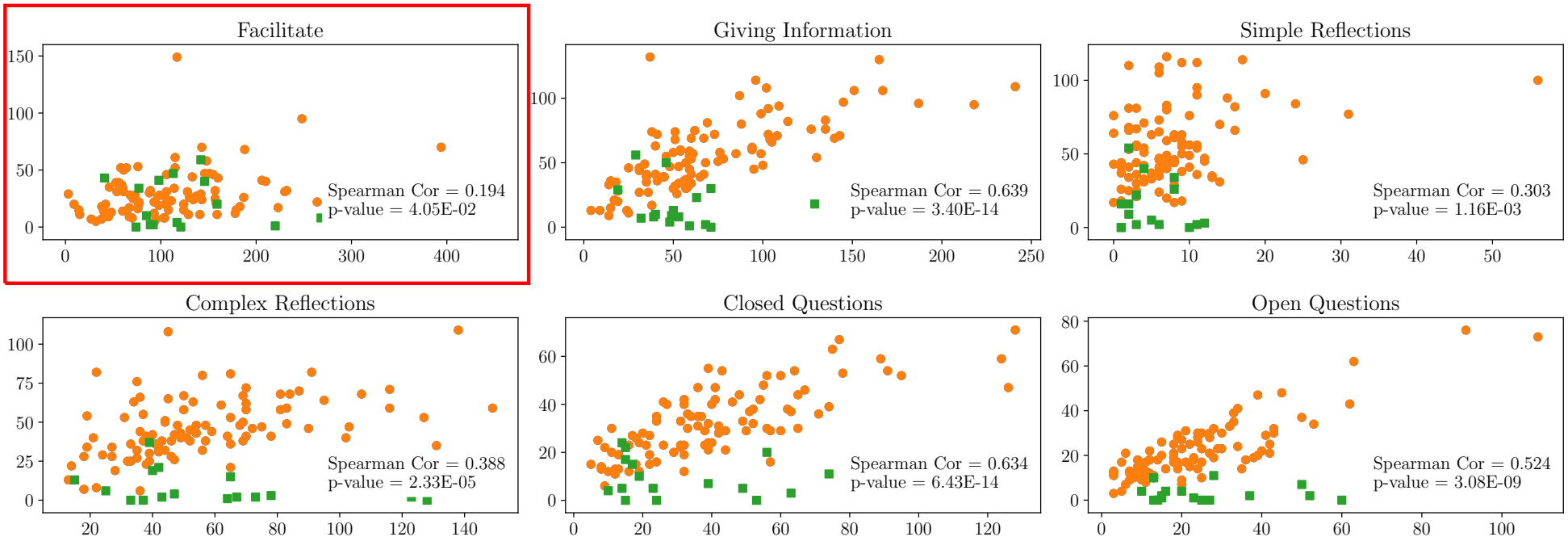
Counts of utterance-level MISC labels per session (*after grouping*) when coded by humans or processed by the pipeline. *green*: sessions marked as problematic according to quality control.

Quantitative assessment: Utterance-level codes



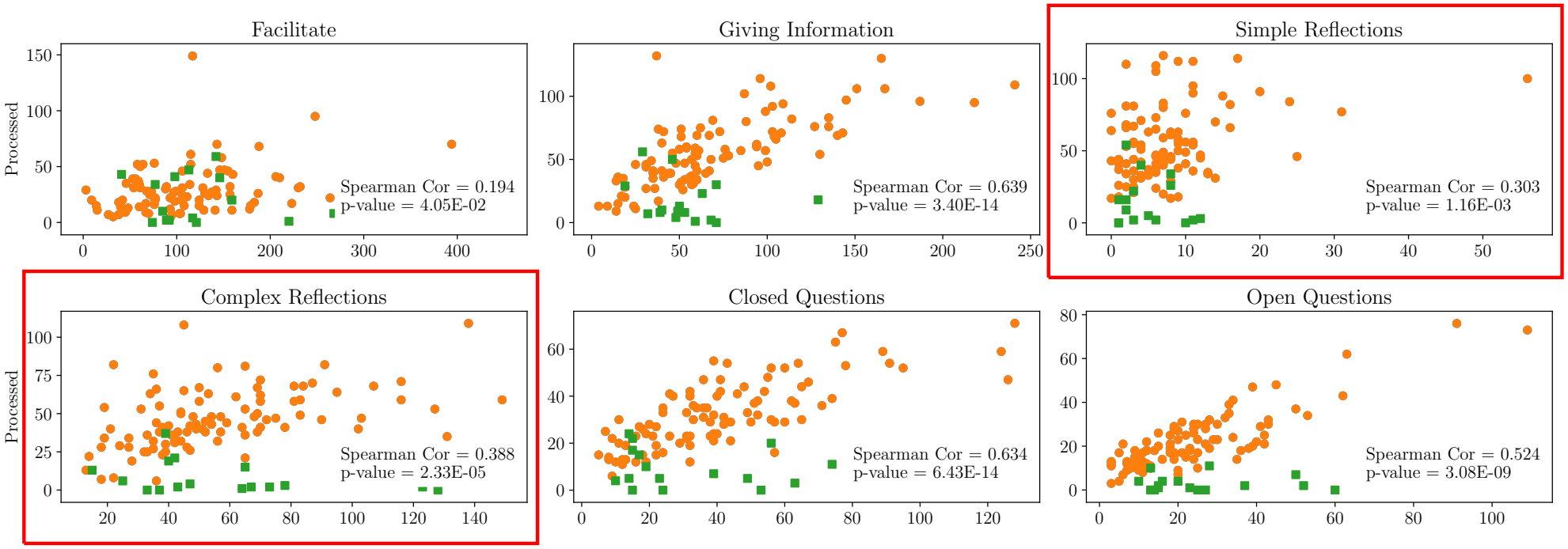
- importance of quality safeguards
  - problems mostly due to diarization
  - average correlation increases to 0.57 from 0.45 if we ignore the problematic sessions

Quantitative assessment: Utterance-level codes



- low correlation for FA – assigned less frequently
  - striking difference between human-derived vs. automatic transcripts
  - reason: pipeline fails to capture short (i.e., one-word) utterances

Quantitative assessment: Utterance-level codes



- low correlation for RES – assigned more frequently
  - partly due to confusion between RES and REC
  - decided to combine them into single 'reflection' code

Quantitative assessment: Session-level codes

metric	accuracy		acc ('within 1')	
transcription	human	pipeline	human	pipeline
acceptance	0.478	0.457	0.771	0.755
empathy	0.586	0.580	0.819	0.851
direction	0.426	0.389	0.740	0.697
autonomy support	0.495	0.451	0.878	0.840
collaboration	0.437	0.346	0.654	0.612
evocation	0.362	0.335	0.751	0.671

Quantitative assessment: Session-level codes

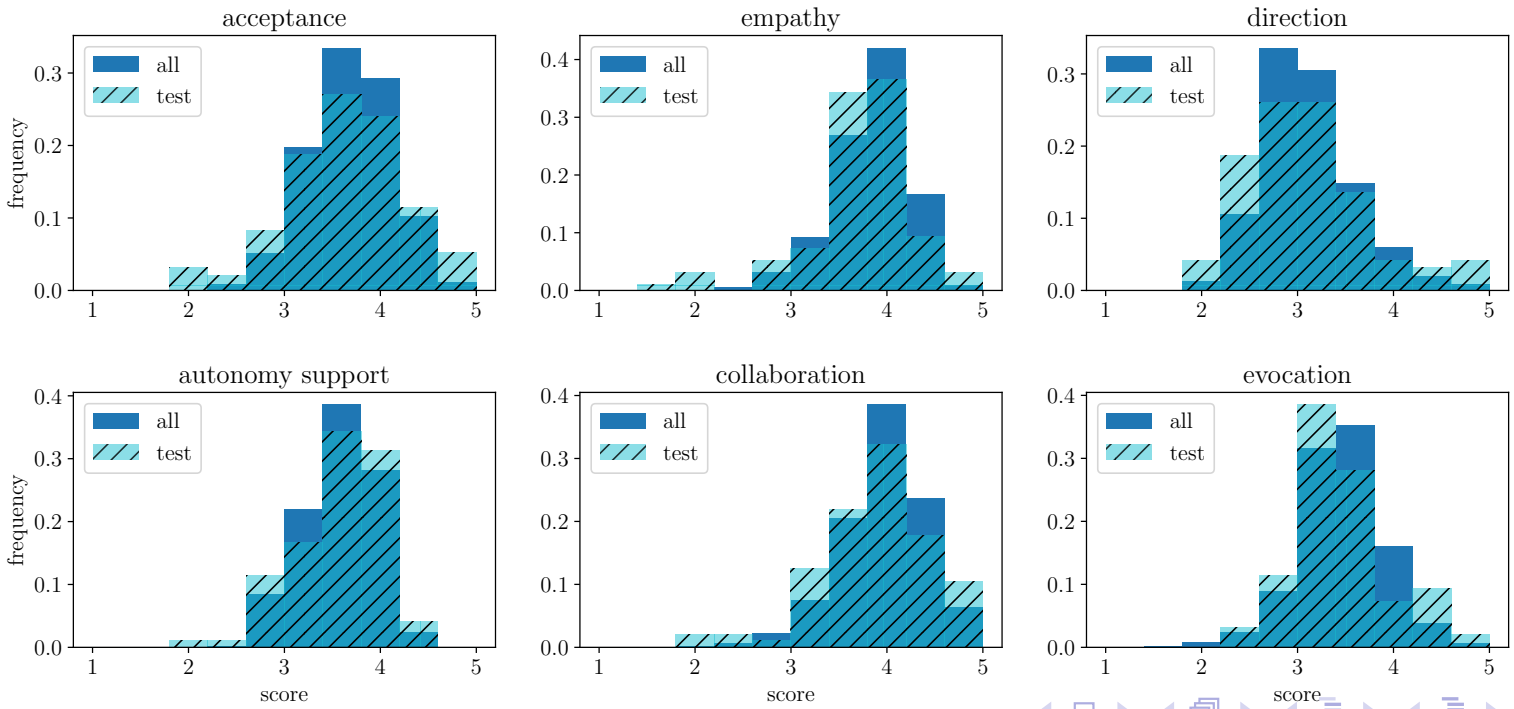
metric	accuracy		acc ('within 1')	
transcription	human	pipeline	human	pipeline
acceptance	0.478	0.457	0.771	0.755
empathy	0.586	0.580	0.819	0.851
direction	0.426	0.389	0.740	0.697
autonomy support	0.495	0.451	0.878	0.840
collaboration	0.437	0.346	0.654	0.612
evocation	0.362	0.335	0.751	0.671

drop of performance on automatic transcriptions (but relatively small)

Quantitative assessment: Session-level codes

metric	accuracy		acc ('within 1')	
transcription	human	pipeline	human	pipeline
acceptance	0.478	0.457	0.771	0.755
empathy	0.586	0.580	0.819	0.851
direction	0.426	0.389	0.740	0.697
autonomy support	0.495	0.451	0.878	0.840
collaboration	0.437	0.346	0.654	0.612
evocation	0.362	0.335	0.751	0.671

drop of performance on automatic transcriptions (but relatively small)



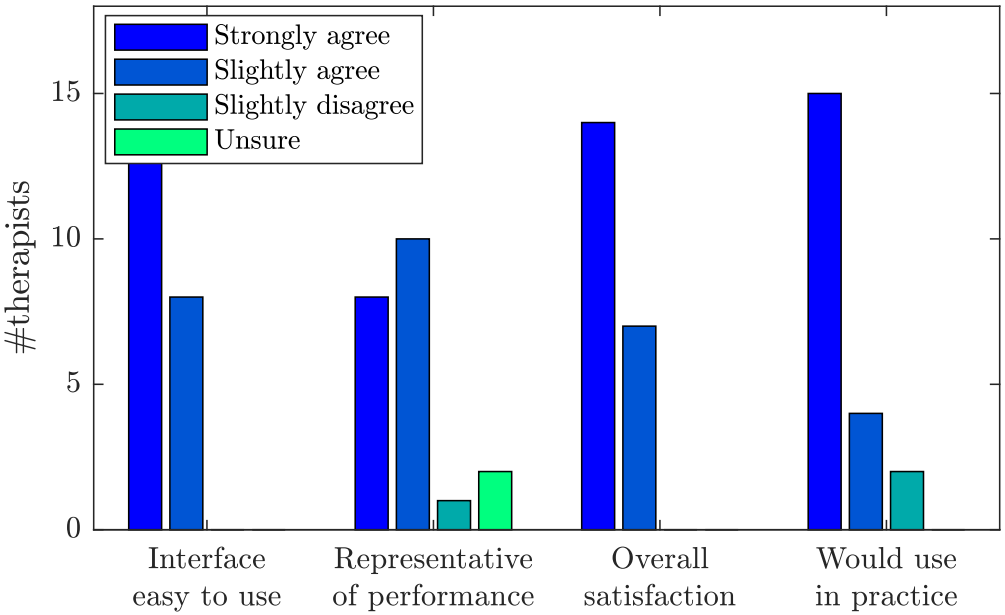
Qualitative assessment of the system

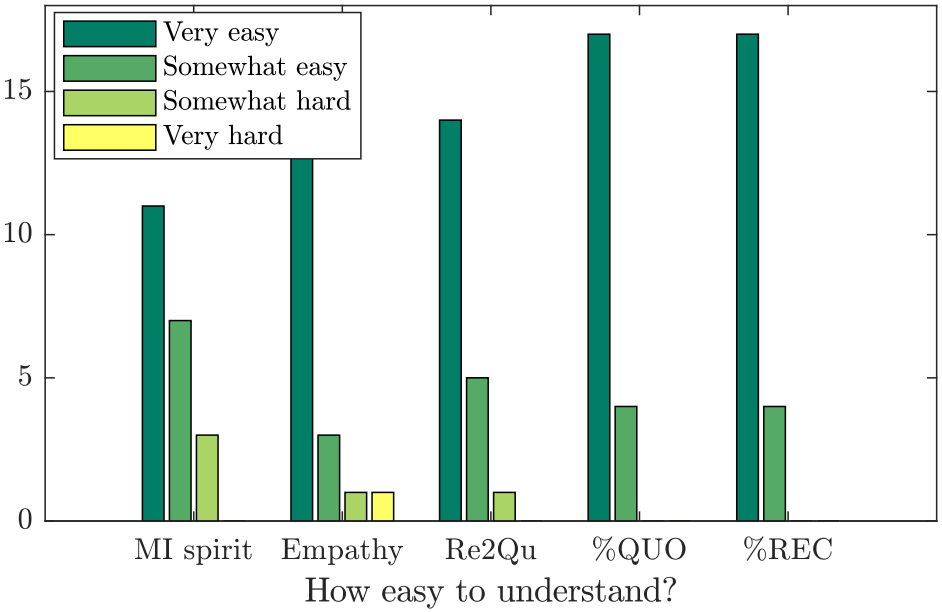
- How easy is it to understand the reported metrics?
- Would counselors use it in clinical practice?

Qualitative assessment of the system

- How easy is it to understand the reported metrics?
- Would counselors use it in clinical practice?

qualitative study: recruited 21 therapists (11 experienced; 10 trainees)





Outline

- Psychotherapy Evaluation and Behavior Coding
- Evaluation of Motivational Interviewing: coreMI
- Evaluation of Cognitive Behavior Therapy
- Potential Improvements and Ethical Implications

Behavioral coding in Cognitive Behavioral Therapy

- CBT: one of the most popular psychotherapeutic approaches
- aims at shifting the patient's patterns of thinking

Monitoring CBT quality: Cognitive Therapy Rating Scale (CTRS)

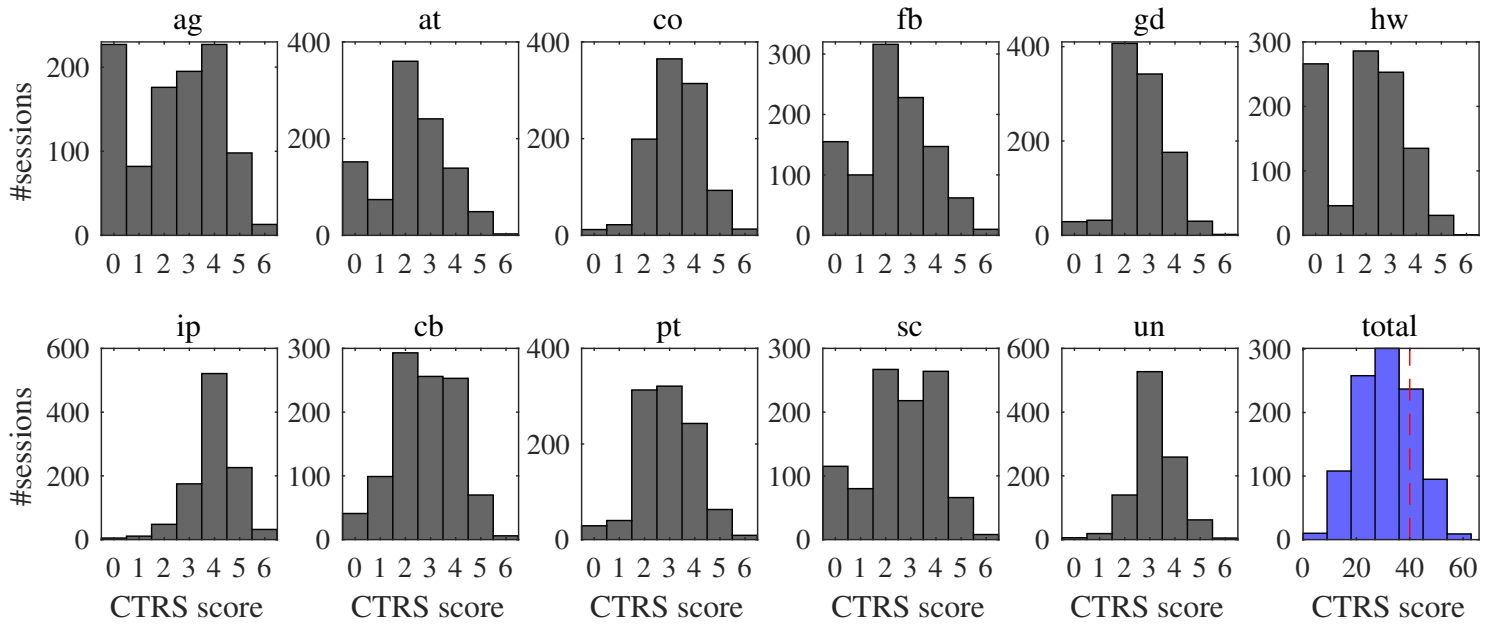
- 11 session-level codes scored on a 7-point Likert scale (0=poor, 6=excellent)

abbreviation	meaning	
ag	agenda	*management and structure*
fb	feedback	
pt	pacing and efficient use of time	
hw	homework	
un	understanding	*good relationship*
ip	interpersonal effectiveness	
co	collaboration	
gd	guided discovery	*conceptualization*
cb	focusing on key cognitions and behaviors	
sc	strategy for change	
at	application of cognitive-behavioral techniques	

- $\sum_{i=1}^{11} \text{code}_i \geq 40 \Rightarrow$ competent delivery of CBT

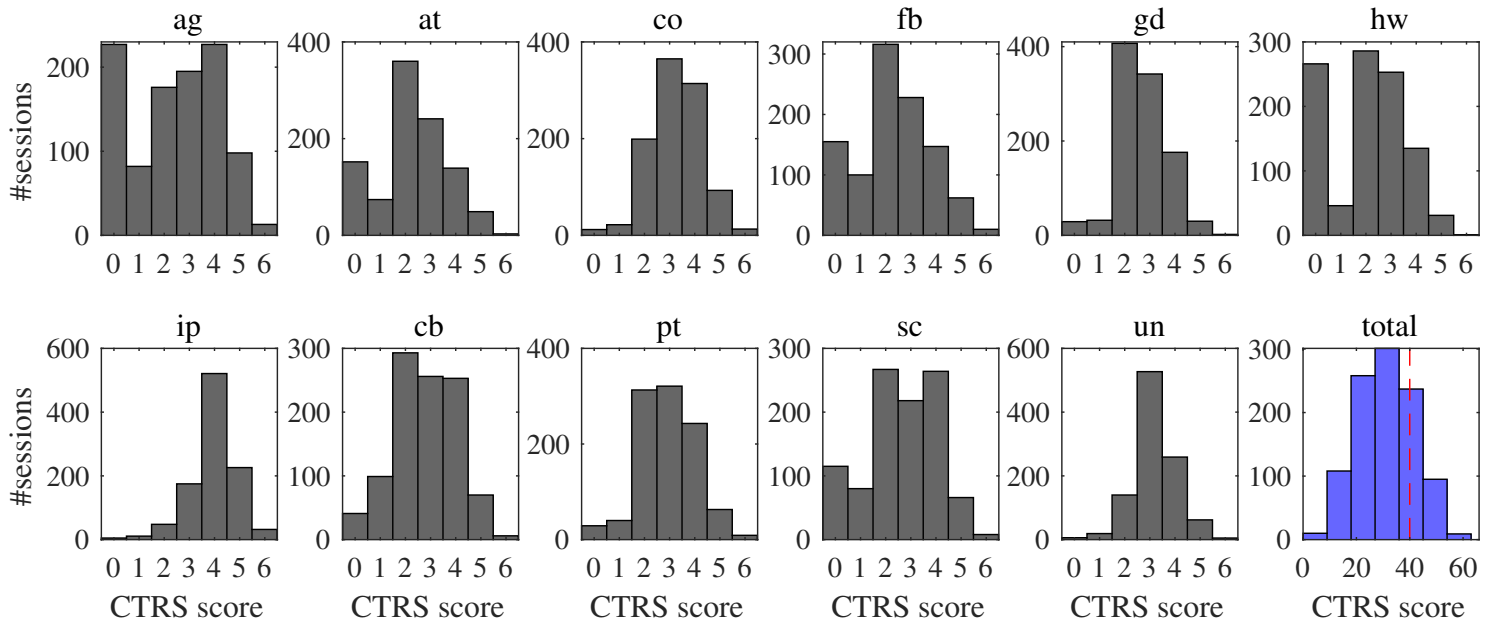
CBT dataset

- 1018 recorded, coded CBT sessions (mean dur $\sim$ 40min), from community clinics, *automatically transcribed*



CBT dataset

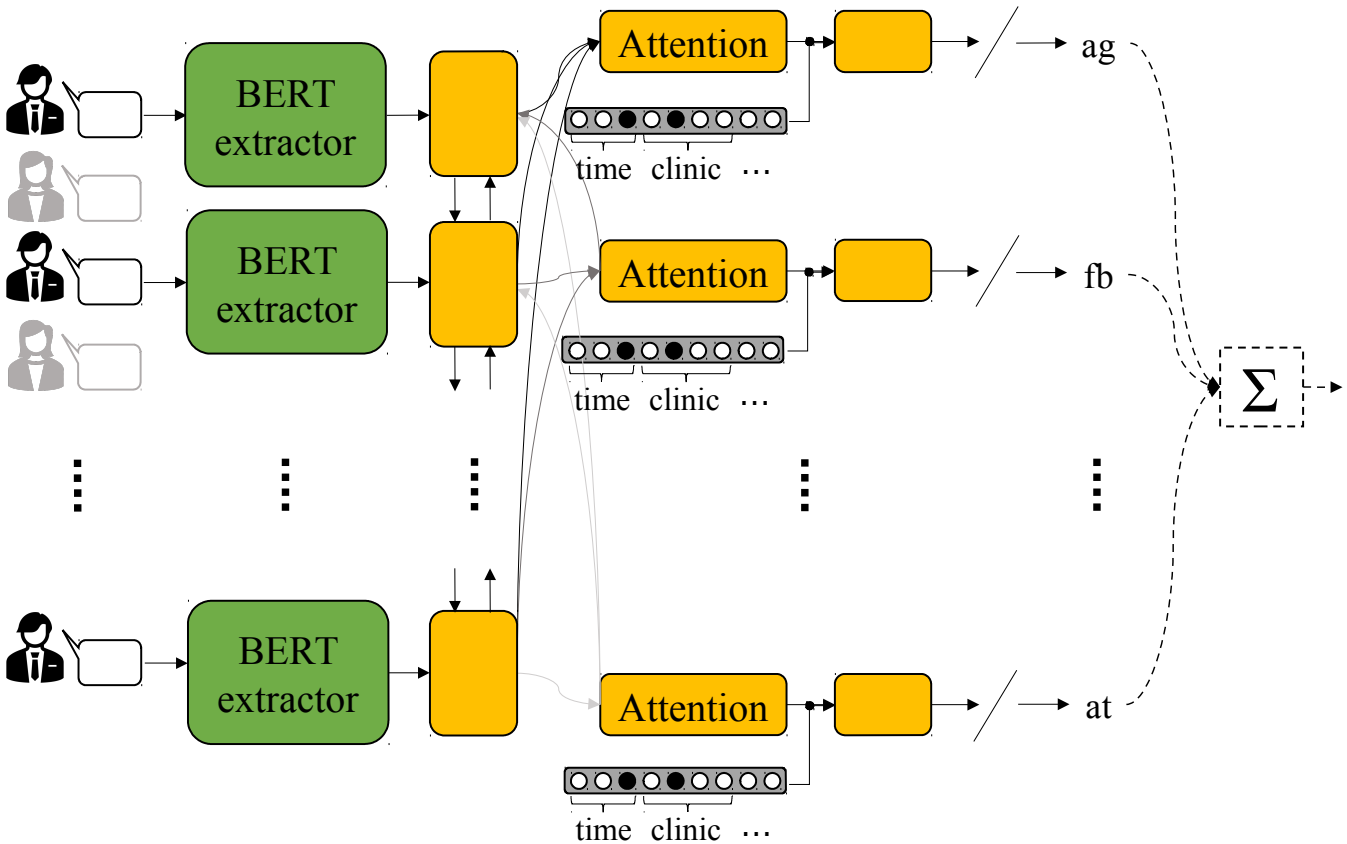
- 1018 recorded, coded CBT sessions (mean dur $\sim$ 40min), from community clinics, *automatically transcribed*



- ML models: language representations + available metadata
  - *clinic*: 383 therapists across 25 clinics
  - *level of care*: 6 categories (inpatient, outpatient, school-based, etc.)
  - *population*: 9 population groups (child, adult, substance use, etc.)
  - *assessment time wrt CBT training*: 7 timestamps (pre-workshop, post-workshop, 1 month after, etc.)
- recorded MI sessions also used for adaptation

Total CTRS prediction

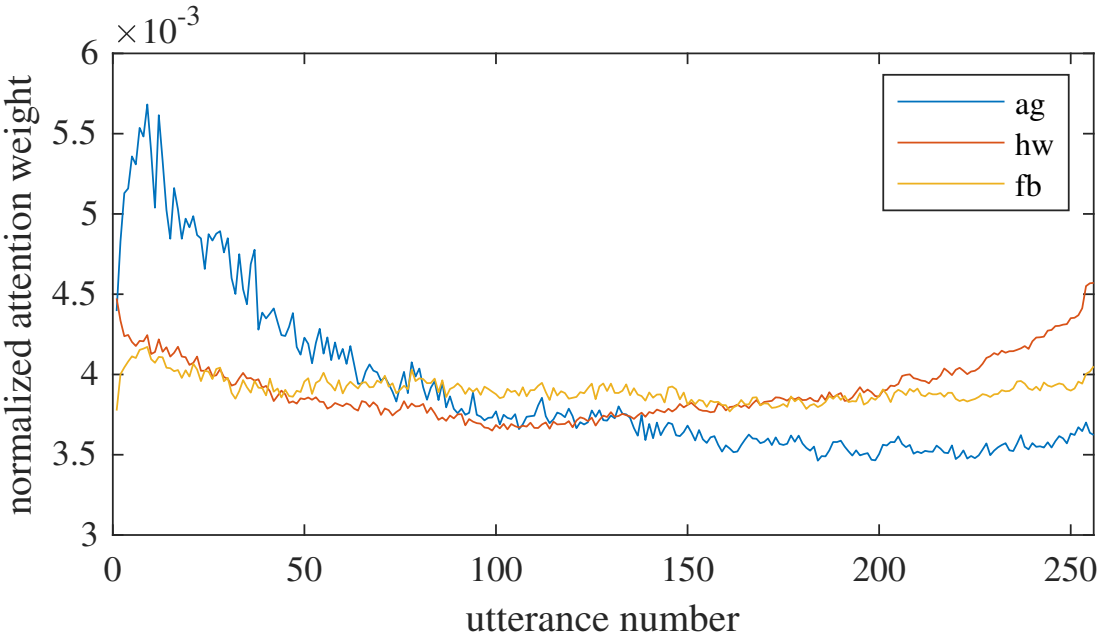
- Model each CTRS code in a regression setting.
- Total CTRS is then calculated as the (unweighted) sum.



- BERT: pre-trained state-of-the-art language representations; remarkable results in multiple language processing tasks.

Localization of CTRS codes

- CBT is a highly structured psychotherapeutic approach
  $\Rightarrow$ reflected in several of the CTRS codes
- Using the attention mechanisms, identify salient utterances
  $\Rightarrow$ reveal this structure
  $\Rightarrow$ examine how the practitioner focuses on different aspects of CBT throughout therapy



Mean attention weights across all the sessions

Outline

- Psychotherapy Evaluation and Behavior Coding
- Evaluation of Motivational Interviewing: coreMI
- Evaluation of Cognitive Behavior Therapy
- Potential Improvements and Ethical Implications

Future directions

- Improving rich transcription
  - speaker segmentation: bottleneck for some sessions
  - overlapping speech and short utterances difficult to capture
  - is a pipelined architecture the best approach?

images from the Noun Project. creators: Juicy Fish, i cons, P Thanga Vignesh, Tomas Knopp

Future directions

- Improving rich transcription
  - speaker segmentation: bottleneck for some sessions
  - overlapping speech and short utterances difficult to capture
  - is a pipelined architecture the best approach?

- Can we use more modalities?
  - incorporate audio or even video
  - use client's language
  - combine data-based learning with expert knowledge (e.g. coding manuals)

images from the Noun Project. creators: Juicy Fish, i cons, P Thanga Vignesh, Tomas Knopp

Future directions

- Improving rich transcription
  - speaker segmentation: bottleneck for some sessions
  - overlapping speech and short utterances difficult to capture
  - is a pipelined architecture the best approach?

- Can we use more modalities?
  - incorporate audio or even video
  - use client's language
  - combine data-based learning with expert knowledge (e.g. coding manuals)

- Improved quality assurance
  - more quality safeguards (for transcription and codes)
  - end-to-end and perceptual evaluation metrics

images from the Noun Project. creators: Juicy Fish, i cons, P Thanga Vignesh, Tomas Knopp

Practical and ethical implications - I

- Is it acceptable to use patients' sensitive data?
  - all patients and therapists sign a consent form
  - approved by Institutional Review Board (sufficient?)
  - all data are de-identified wrt patients

image from the Noun Project. creator: SBTS

Practical and ethical implications - I

- Is it acceptable to use patients' sensitive data?
  - all patients and therapists sign a consent form
  - approved by Institutional Review Board (sufficient?)
  - all data are de-identified wrt patients

- What if such a system is used to blindly evaluate a therapist? That could even mean loosing their job!
  - the goal is not to replace human supervision, but rather augment the supervisor's capabilities and offer a tool for self-assessment
  - users should be adequately trained to understand the meaning of automatically generated feedback and evaluation scores

image from the Noun Project. creator: SBTS

Practical and ethical implications - II

- How to mitigate potential biases?
  - adaptation to the actual use case (e.g., perceptions about psychotherapy differ across cultures)
  - employ large and diverse pools of human coders
  - fairness through unawareness
    (both for models and for annotators)

images from the Noun Project. creators: Erik Vullings, Adrien Coquet

Practical and ethical implications - II

- How to mitigate potential biases?
  - adaptation to the actual use case (e.g., perceptions about psychotherapy differ across cultures)
  - employ large and diverse pools of human coders
  - fairness through unawareness (both for models and for annotators)

- Any additional requirements before using in clinical settings?
  - incorporate confidence metrics and quality safeguards of the model
  - users should be able to question model predictions (human-in-the-loop)

images from the Noun Project. creators: Erik Vullings, Adrien Coquet

Thank you!

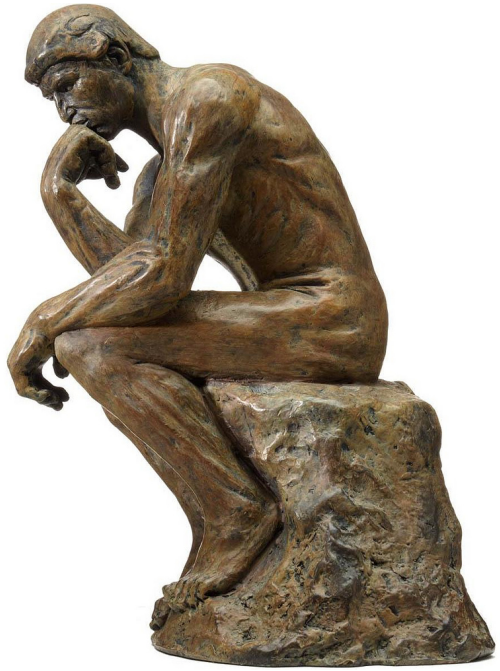
USC University of Southern California

THE UNIVERSITY OF UTAH®

UNIVERSITY of WASHINGTON

Penn
University of Pennsylvania

Northeastern University



Questions and Discussion

NIH National Institutes of Health
Turning Discovery Into Health

SAIL