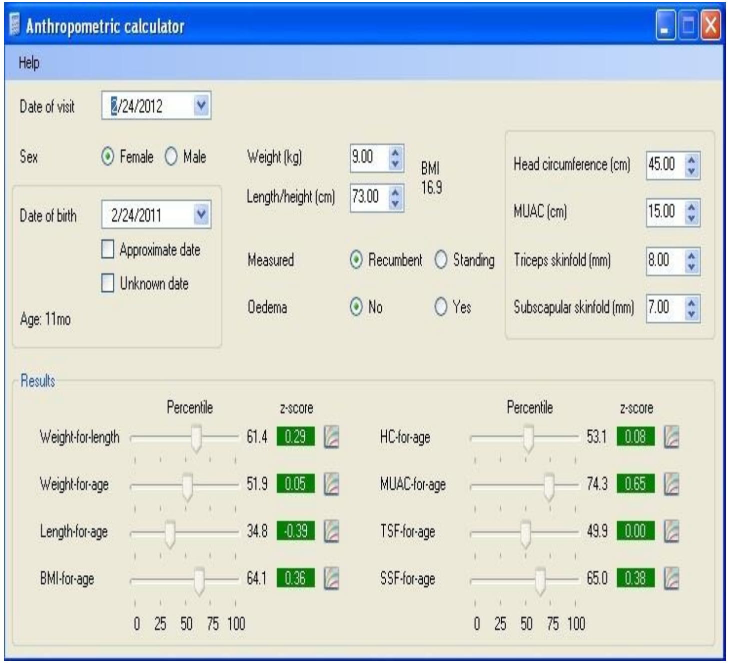
Toddler mental development interventions: Can machine learning play a part?

28th June 2024
IDDDP Presentation
Akshat Gautam

How is Physical Development Monitored?

- Physical parameters such as height, weight, and head circumference
- Standardized growth charts (WHO)
- z-scores calculated

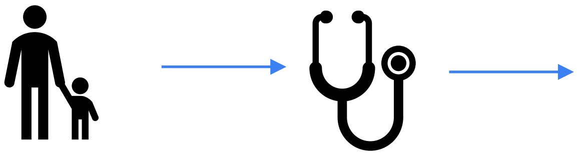
Mental development ?



Height Weight Chart [1]
WHO Anthro survey [2]

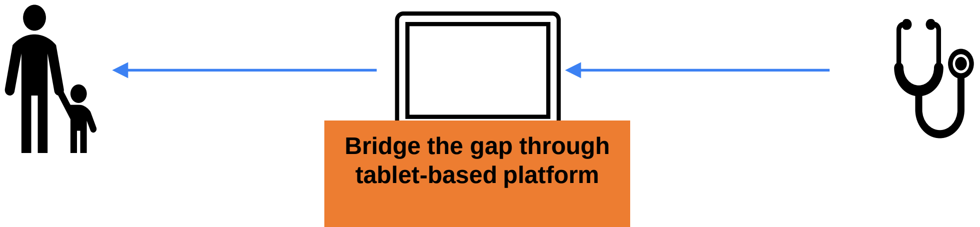
Gauging Mental Development

Goal : To measure mental development (ultimately for interventions)
Current Situation: Parents observe atypical symptoms, visit hospital

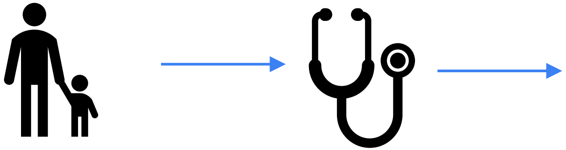


- Hospital conduct tests
- Too late? (symptoms already appeared)
- Results difficult to accept?

Long term goal : Bring hospital to children; generate developmental scores



Understanding Psychometric Tests



What are these tests?

What are they:

- Standard and scientific method to measure mental capability

Why are they not used everywhere?

- Costly
- Trained professional; specific setting
- Not available in Low Income countries

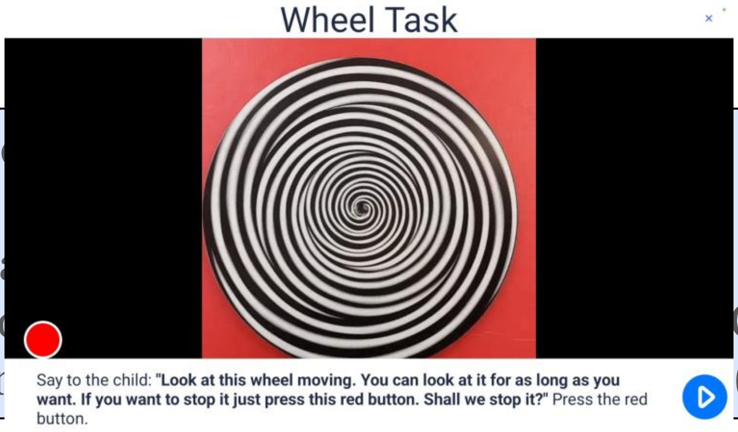
Tablet Based Assessment

- Tablet contains set of tasks, each for different domain (social, motor, cognitive)
- Task generated raw data, not usable , needs to be processed

One of the tasks is the **wheel task (**social domain)

Mor
of developm

Soci
Looking Tas
Moto
Colouring Ta
ask
Cogn
(DGT)



Black and white wheel

Video recorded from front camera

Overview

1. Introduction (done)
2. Analysis Pipeline
3. Understanding GMDS test (more detail)
4. Using features to predict GMDS scores
5. Using IRT (item response theory) to generate scores
6. Future Work
7. Understanding tasks and feature extraction

- Data collection done by STREAM team
- My work: Data analysis
- All contributions mentioned in purple colour.

Wheel Task

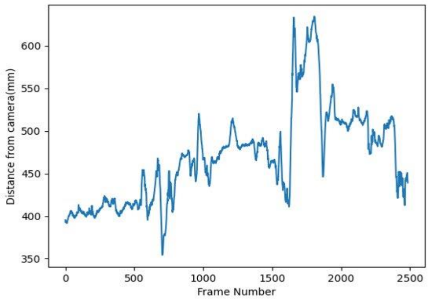
Feature description

Get the distance of face from the camera given video

Involves Computer Vision (STAGE 1)

Black and white wheel [3]

Video recorded from front camera

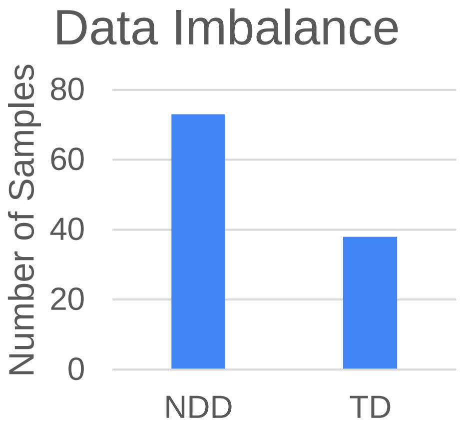


Distance vs Frame no

Classification

- Median and Std dev of distance signal
- 111 children, 2 input features, 2 classes
- 5-fold CV

Algorithm	Accuracy (%)	F1 Score
Random Forest	78.46	0.67
Logistic Regression	81.23	0.74
SVM	73.07	0.55



- Classified children into NDD/TD using distance as feature

Overview

1. Introduction (done)
2. Analysis Pipeline
3. Understanding GMDS test (more detail)
4. Using features to predict GMDS scores
5. Using IRT (item response theory) to generate scores
6. Future Work
7. Understanding tasks and feature extraction

Problem Statement
Generate developmental scores using tablet-based assessment

Pipel

Parent name	Karuna Jha
Address	F-2,765,sangam vihar
Gender	M
Birth Date	2014-09-29
Age	4y 1m
Diagnosis	Not available
State	Delhi
Hand dominan	right
deviceId	39992d8a3632ff73
endTime	16.04.2018 08:37:29 386
interrupted	0
screenHeight	1600
screenWidth	2560
startTime	16.04.2018 08:33:47 498
xdpi	301.037
ydpi	301.037
Sub-attempt #1	

color	device_x	device_y	device_z	time	touch_pressur	touch_size	touch_x	touch_x_dp	touch_y	touch_y_dp
-65536	-0.1436521	0.18195933	9.940725	2724	0.07843138	0.20392159	128.125	68.0979414	1010.5469	537.101765
-65536	-0.1436521	0.18195933	9.959879	4393	0.11764707	0.23529413	1022.5	543.454791	1300.3906	691.152569
-65536	-0.1436521	0.18195933	9.959879	4439	0.10588236	0.227451	1012.29724	538.032064	1280.4792	680.569737
-65536	-0.1436521	0.18195933	9.959879	4455	0.10588236	0.21960786	1006.9506	535.190345	1269.5358	674.753363
-65536	-0.1436521	0.18195933	9.959879	4471	0.10588236	0.21960786	998.96625	530.946694	1248.6647	663.660454
-65536	-0.1340753	0.19153613	9.969456	4488	0.10588236	0.227451	991.5432	527.001372	1230.4275	653.967453
-65536	-0.1340753	0.19153613	9.969456	4504	0.10588236	0.227451	979.2913	520.489535	1209.2983	642.737365
-65536	-0.1340753	0.19153613	9.969456	4521	0.10588236	0.21960786	968.2325	514.611825	1183.4607	629.00478
-65536	-0.1340753	0.19153613	9.969456	4537	0.10588236	0.21960786	959.324	509.876992	1164.1115	618.720755
-65536	-0.1340753	0.19153613	9.969456	4554	0.10588236	0.23529413	952.3822	506.187452	1144.6969	608.401971
-65536	-0.1340753	0.19153613	9.969456	4570	0.10588236	0.227451	945.5522	502.557333	1123.0647	596.90454
-65536	-0.1340753	0.19153613	9.969456	4586	0.10588236	0.227451	941.5788	500.445487	1109.33	589.6046
-65536	-0.1340753	0.19153613	9.969456	4603	0.10588236	0.21960786	938.7492	498.941565	1088.0769	578.30866
-65536	-0.1340753	0.19153613	9.969456	4619	0.10588236	0.227451	937.35486	498.200479	1070.7511	569.100064
-65536	-0.1340753	0.19153613	9.969456	4636	0.10588236	0.227451	937.5	498.27762	1059.5403	563.141567
-65536	-0.1340753	0.19153613	9.969456	4652	0.10588236	0.21960786	937.5	498.27762	1050.3745	558.26998
-65536	-0.1340753	0.19153613	9.969456	4668	0.10588236	0.21960786	948.7494	504.256633	1043.9456	554.853045
-65536	-0.12449849	0.18195933	9.9790325	4684	0.10588236	0.21960786	956.48157	508.366251	1037.7458	551.557875
-65536	-0.12449849	0.18195933	9.9790325	4701	0.10588236	0.21960786	978.727	520.189611	1030.9724	547.957839
-65536	-0.12449849	0.18195933	9.9790325	4718	0.10588236	0.21176472	1011.1245	537.408757	1024.3939	544.461392
-65536	-0.12449849	0.18195933	9.9790325	4734	0.10588236	0.23529413	1045.0378	555.433545	1022.9485	543.693167
-65536	-0.12449849	0.18195933	9.9790325	4751	0.10588236	0.21960786	1087.8297	578.177274	1023.0469	543.745467

Assessm
on table

Fea
Extr

se features
or analysis

s from
Taks

What to do with these features?

Feature Tensor

Child 1	Coloring Task Feature	Wheel Task Feature	Button Task Feature	……….
Child 2	Coloring Task Feature	Wheel Task Feature	Button Task Feature	……….
Child 3	Coloring Task Feature	Wheel Task Feature	Button Task Feature	……….

Feature to scores supervised on psychometric tests (GMDS)

Features to scores but unsupervised using IRT (Item response theory)

Features to classify NDD/TD (Done for wheel task)

For this, it's important to understand psychometric test GMDS

Griffith's Mental Development Scale (GMDS)

- Gold-standard tool
- 0-6 years
- 321 items, 5 domains
  - Foundations of learning (63 items)
  - Language and communication (63 items)
  - Eye and hand coordination (67 items)
  - Personal-social-emotional (65 items)
  - Gross motor (63 items)
- Binary Items

Photo of GMDS test kit [4]

Convert these "raw scores" into normalized scores

Sample GMDS Results

ChildID	A	B	C	D	E
MW-0113	33	43	44	53	47
IN-1653	49	51	52	56	60
IN-1682	31	43	46	45	46

Short demo of GMDS test

This is done so a child doesn't have to sit through all 321 items

Age appropriate start point

Testname	A	B
1.1	1	1
1.2	1	1
1.3	1	1
1.4	1	1
1.5	1	1
1.6	1	1
1.7	1	0
1.8	1	1
1.9	1	0
1.10	0	1
1.11	0	1
1.12	1	0
1.13	0	0
2.1	1	0
2.2	0	0
2.3	0	0
2.4	0	0
2.5	0	0
2.6	0	0
2.7	0	0
2.8	0	0
3.1	0	0
3.2	0	0
3.3	0	0
3.4	0	0
3.5	0	0
3.6	0	0
3.7	0	0
3.8	0	0
3.9	0	0
Scores	11	9

Keep increasing level until 5 consecutive fails

Marks all levels after 5 fails as zero

Decrease levels until there are 5 consecutive successes.

Mark levels lower than these as also successes

Sum all successes to get a score

Repeat for all other 4 domains

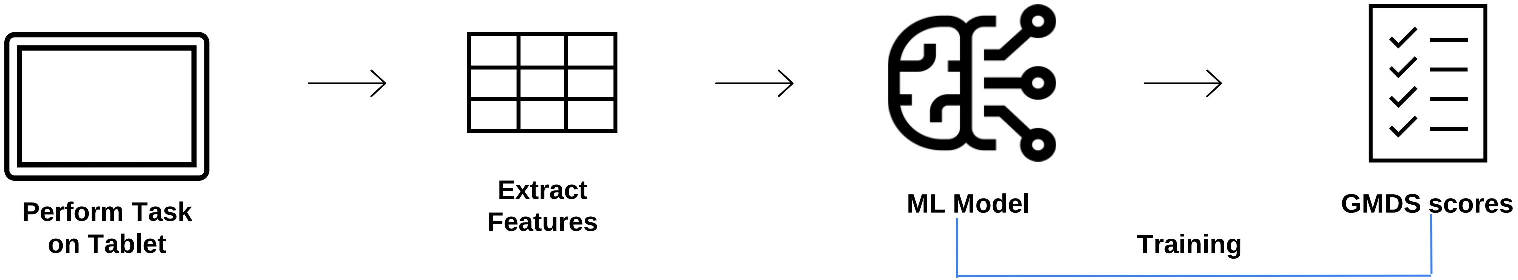
How do we use these raw scores?

Motivation

- Want psychometric test to be administered easily
- Want to show features actually capture development

Features to developmental scores based on psychometric test

Pipeline



Setup

Training
384 data points (i.e. features and scores for 384 children)
56 features (54 features from 6 different tasks + Age,Gender)
Target label is GMDS scores across 5 domains

Training Setup
5-fold cross-validation (due to less data)

What is model choice?
Regression Models

Metrics
R2 Score
Mean absolute percentage error (MAPE)
Mean square error (MSE)

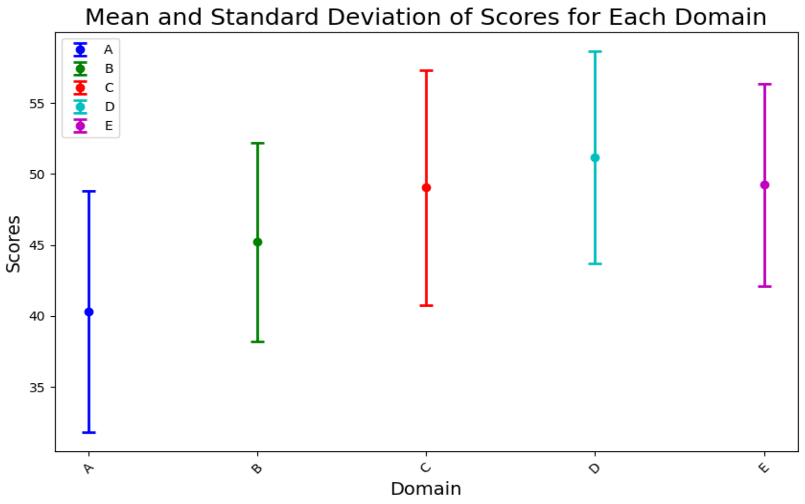
Models
• Linear Regression • Ridge Regression • Random Forest • Gradient Boosting • AdaBoost • Decision Tree • Support Vector Regression • KNN regressor • XGBoost

Results [1/3]

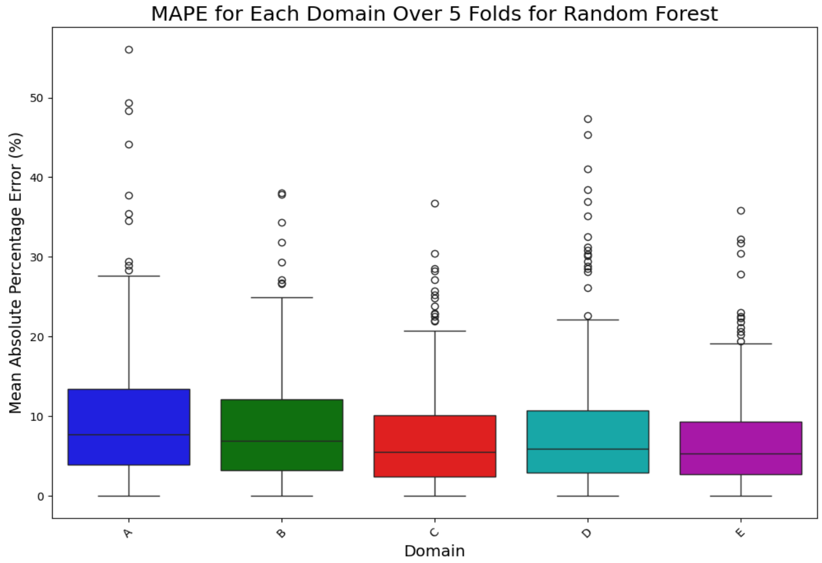
Best Model

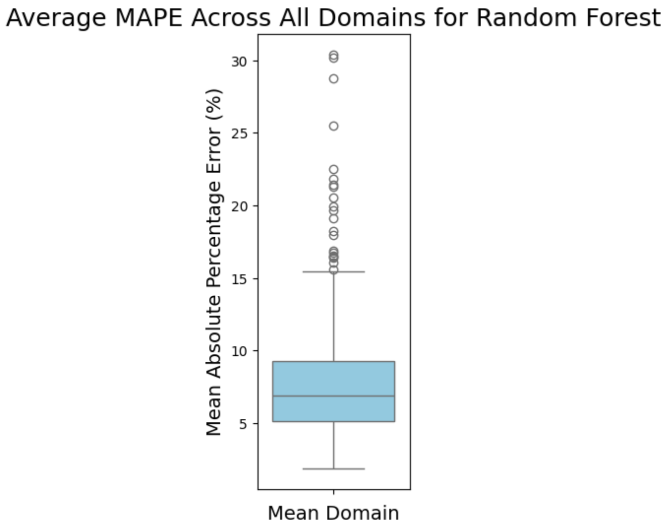
Model	R2 Score	MAPE	MSE
Random Forest	**0.64**	**7.88%**	**20.12**

Ground truth GMDS score distribution



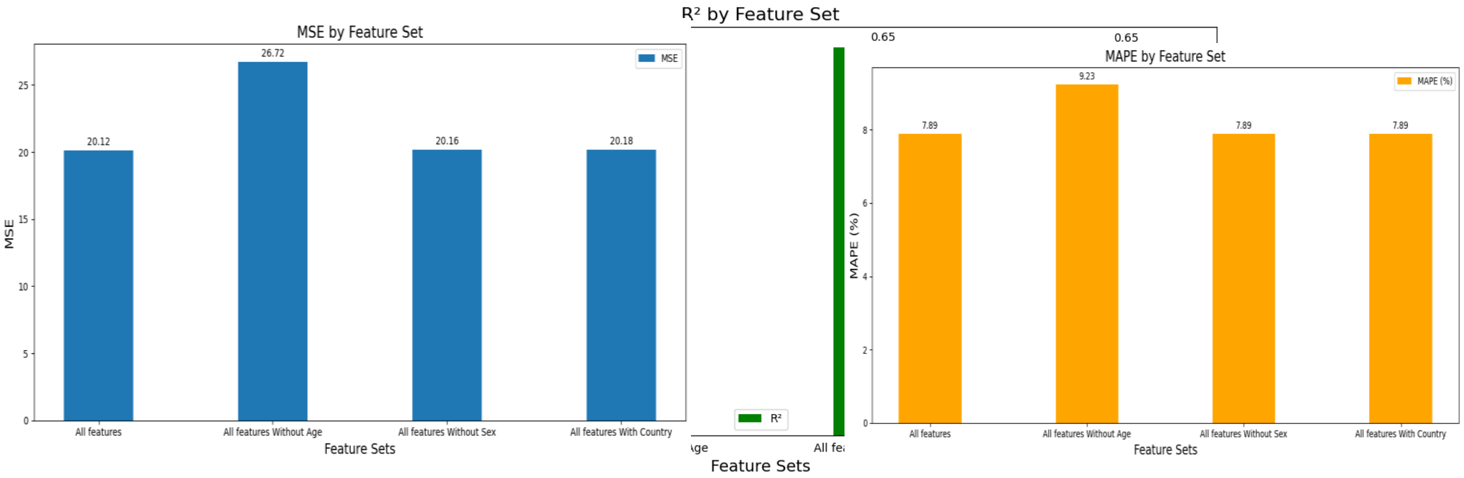
Results (2/3)





- 1st plot: MAPE across 5 domains
- 2nd plot: Average MAPE
- 5-10% error for half of the samples

Results (3/3)



- Removing Age reduces performances, GMDS highly correlated with Age
- Sex and Country don’t show any effect
- Same trend in MAPE and MSE

Overview

1. Introduction (done)
2. Analysis Pipeline
3. Understanding GMDS test (more detail)
4. Using features to predict GMDS scores
5. Using IRT (item response theory) to generate scores
6. Future Work
7. Understanding tasks and feature extraction

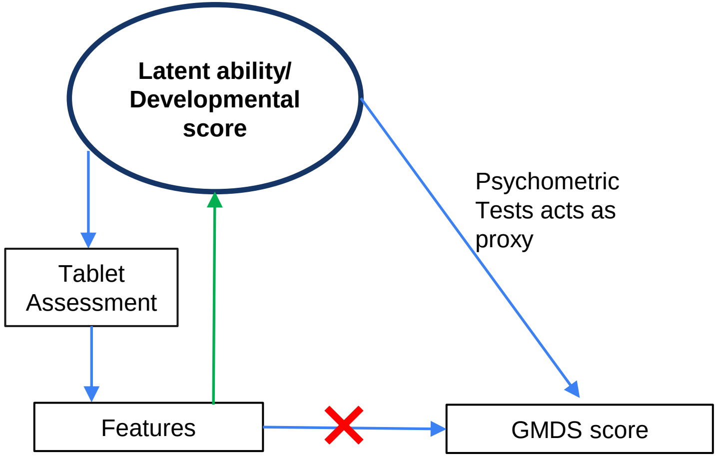
Problem Statement
Generate developmental scores using tablet-based assessment

Where are we now?

✓ Features extracted can be used to generate scores like GMDS

✗ Developmental scores could be generated independently

Not depend on GMDS
- Costly
- Will need to administer again for a new country



What's Next?

To generate developmental scores without relying on psychometric test

Item Response Theory (IRT)

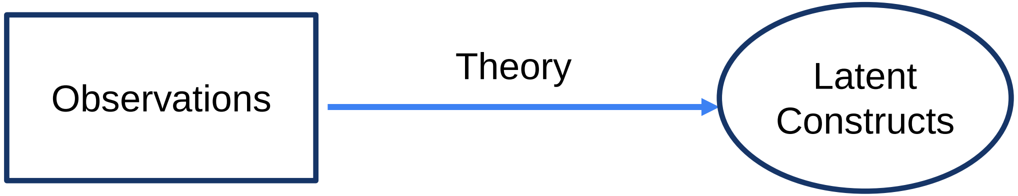
What is IRT?

- Theory of measurement
- Family of statistical models

What does it do?

IRT maps observations onto internal traits / states :-

- Test scores responses into knowledge / intelligence
- Questionnaire items into attitude / beliefs



IRT equation

$$Y_{ij} = \theta_j - b_i$$

Where, Y_{ij}= Logit of Response by person j for item i,

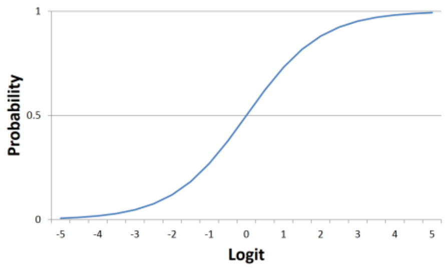
θ_j= **Trait** of person j**,**

b_i= Difficulty of item I

Thus, for mapping values to [0,1],

$$Logit = \ln\left(\frac{Pr}{1 - Pr}\right)$$

So, we have probability value between [0,1] and binary responses (0/1) to items. We can optimize the two parameters (θ, b)



Probability vs logit [6]

Some other models

2 Parameter Model

$$Y_{ij} = a_i\theta_j \; - b_i$$

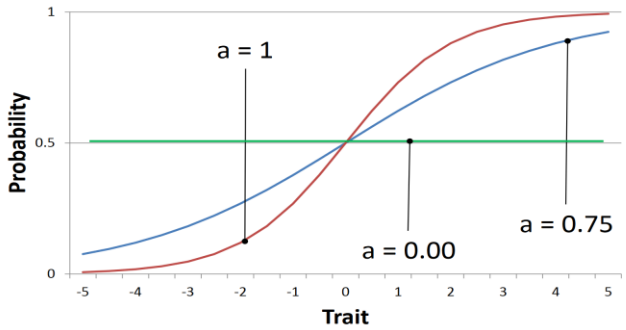
Y_{ij}= Logit of Response by person j for item i,

a_i= Discrimination of item I,

θ_j= **Trait** of person j**,**

b_i= Difficulty of item i

Same difficulty, different discriminations



[6]

In STREAM, if we consider tasks metric as "items" in questionnaire, responses are not in binary. e.g., for coloring tasks

Task metric / Items	Features / Responses
Points Inside	636
Points Outside	1595
Crossovers	63
Time Taken	88216

Adapting for STREAM data

What if we remove the logit link from the equation earlier :-

$$Y_{ij} = \theta_j - b_i$$

Where, Y_{ij}=Feature of child j for the task metric i,

θ_j= Ability of the child j,

b_i= Difficulty of task metric j

- Known as LME (linear mixed effect) model
- lme4 package in R [7]

Setup for START data

Looking at the equations for just one task feature,

$$
\left.\begin{aligned}
Y_{points\ inside, child\ 1} &= \theta_{child\ 1} - b_{points\ inside} \\
Y_{points\ inside, child\ 2} &= \theta_{child\ 2} - b_{points\ inside} \\
Y_{points\ inside, child\ 3} &= \theta_{child\ 3} - b_{points\ inside}
\end{aligned}\right\} \text{Fixed effect}
$$

$$\cdots\cdots$$

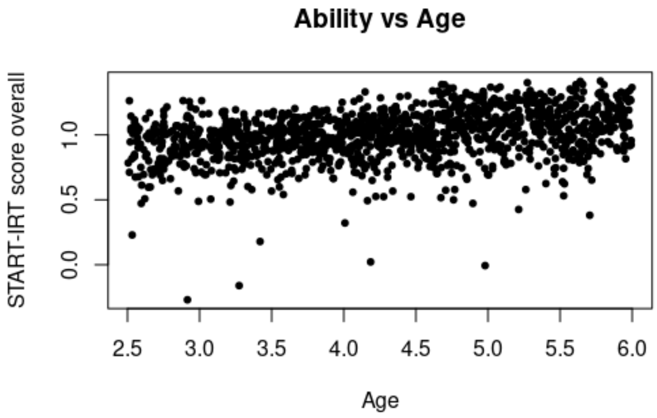
Random effect

Final equation

$$Feature_{\text{task},child} = \theta_{child} - b_{task}$$

- Higher ability should lead to higher feature value
- Not true, higher ability child should have lower crossover counts

Results (1)



r = 0.34

Correlation of IRT score with	r
GMDS domain A	0.27
GMDS domain B	0.22
GMDS domain C	0.24
GMDS domain D	0.24
GMDS domain E	0.25

Correlating IRT scores with GMDS

Relatively low correlation with age and GMDS scores, the scores need to be improved

Improvements/Future Work

Why is correlation not good?

- IRT assumes monotonicity , ability increases, task feature should also increase
  - Not necessarily true since they are hand crafted features, e.g. number of crossovers increases as age of children increases which is not expected

- IRT assumes local independence: responses given to the separate items in a test are mutually independent given a certain level of ability (multiple features are extracted from same task)

- We are using all features to predict a single score, could bin features into different domains and generate multiple scores like -> social, motor …

Future Work Motivation

- We are not including the fact that the child is performing tasks in a particular order, and in a single sitting

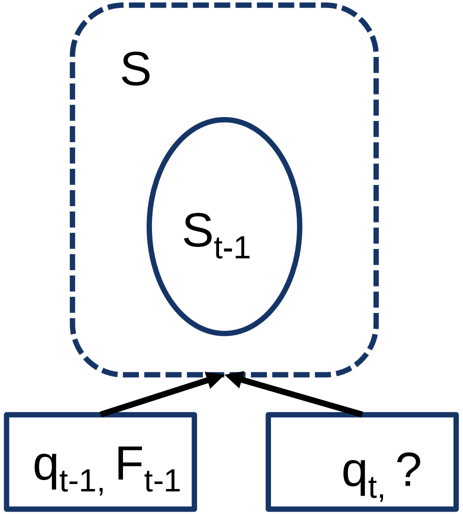
- Each feature may require mastery in multiple areas (social, motor, fine motor etc.), however we may not know the areas corresponding to each feature

- We are fitting a **linear** mixed effect regression model

Knowledge State [8]

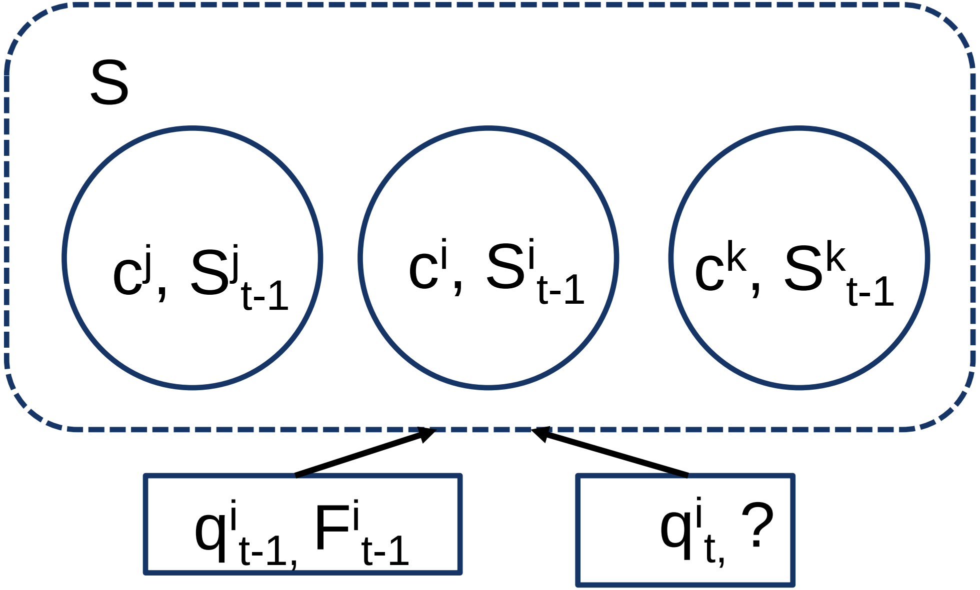
- Given a child's previous task attempts X= {x1,x2, … xt-1}, our goal is to predict the feature (say number of crossovers) that child will achieve in the current task
  - ➢ Each input $x_t = (q_t, F_t)$ is a tuple containing task q_t, and its feature F_t which is computed from the tablet data

- The information of previous attempted tasks is condensed into a latent knowledge state S={s1,s2, … st-1}
  - ➢ For example, if our previous method incorrectly predicts a feature F_t, our goal is to update the model and the knowledge state, thus improving our understanding of the child as she attempts task over time



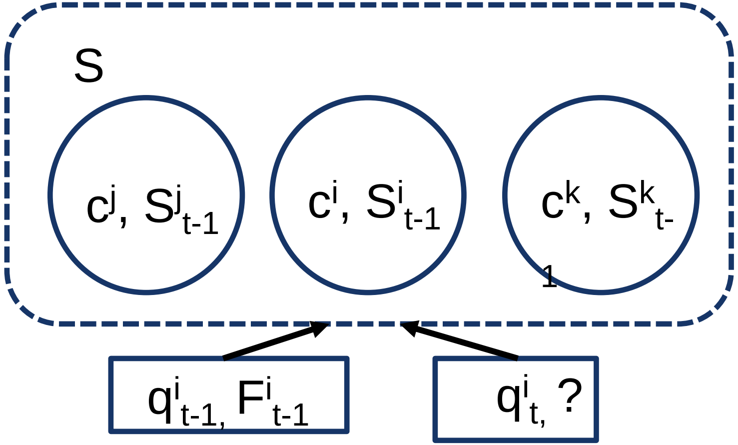
Concepts

We want to have multiple concepts for each state i.e c1,c2….cn



States and concepts

- Combine knowledge state and concept in a memory augmented neural net paradigm
- Training
  - Learn static matrix (key) for storing concepts associated with each task independent of child
  - Learn matrix (value) for storing student's knowledge state in each concept
- Inference
  - Update the value matrix as child completes task
  - Final score after all tasks are completed

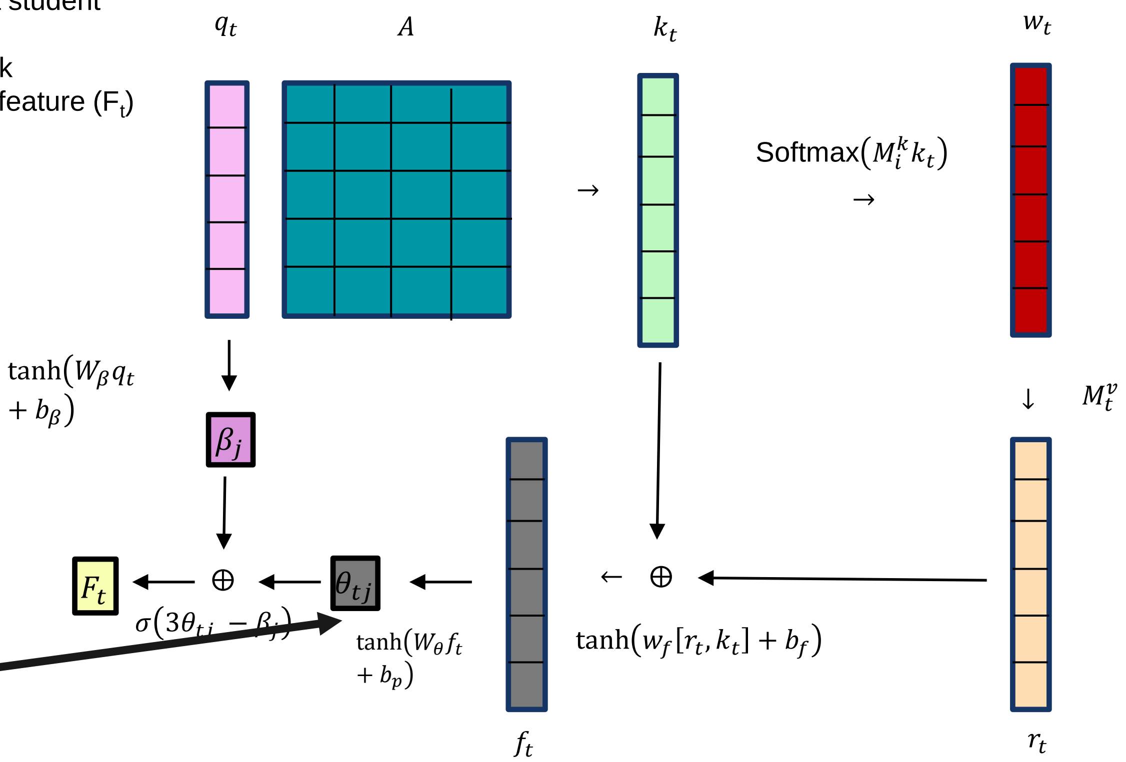


Maintains a knowledge state for each concept simultaneously and all states constitute the "knowledge" of a child

Method 3: Deep IRT [9]

- $M^v \in \mathbb{R}^{N\times d_v}$: Value memory matrix (skill states)
- $M^k \in \mathbb{R}^{N\times d_k}$: Key memory matrix (latent abilities)
- $A \in \mathbb{R}^{d_k\times Q}$: Ability Components Embedding matrix
- $k_t \in \mathbb{R}^{d_k}$: Embedding vector (key)
- $v_t \in \mathbb{R}^{d_v}$: Response Embedding vector
- $e_t \in \mathbb{R}^{d_v}$: Response erase vector
- $\mathbf{B} \in \mathbb{R}^{Q\times d_v}$: Ability Components response embedding matrix

- IRT Module is built on top of key-value network
  - Instead of predicting probability, we predict student ability using f_t
- Task difficulty is calculated using another network
- Ability and difficulty is combined to predict task feature (F_t) [reminiscent of Item Response Theory]

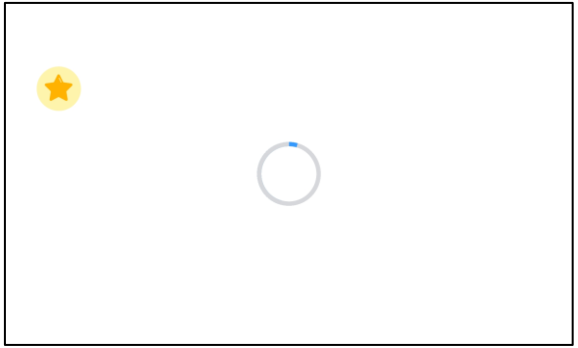


Ability is our developmental score

Delayed Gratification Task

Task Description

A star appears on screen. Child is told to wait for some time to get all three stars.



Feature Description

1. Proportion time spent delaying gratification
2. Proportion of frames child's face visible

Start time and end time are read through the excel files. Total task time is 180 s

Proportion Time = $\frac{End\ Time - Start\ Time}{180}$

Medipipe face mesh is used to detect if a face is present or not in the frame. If more than one faces are present, then that frame is ignored.

Proportion face = $\frac{No\ of\ frames\ with\ a\ face}{Total\ No\ of\ frames}$

Summary

- The data stored as raw data from tablet assessments can be converted into relevant features.
- These features can be used for classification into NDD/TD.
- These features used to generate scores under the supervision of GMDS scores.
- Item Response Theory used to generate developmental scores in an unsupervised setting.

Acknowledgements

- My advisor: Prof. Sharat Chandran
- STREAM team
- Shubham (especially for distance work, experiments ..)

Thank You!

References

[1] https://www.who.int/tools/child-growth-standards/standards/weight-for-length-height
[2] https://www.who.int/tools/child-growth-standards/software
[3] https://journals.sagepub.com/doi/full/10.1177/13623613231182801
[4] https://www.aricd.ac.uk/about-the-griffiths-scales/griffiths-iii/griffiths-iii-kit/
[5] https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1000273
[6] https://hummedia.manchester.ac.uk/institutes/methods-manchester/docs/irt.pdf
[7] https://www.jstatsoft.org/article/view/v039i12
[8] https://www.mdpi.com/2624-8611/5/3/50
[9] https://arxiv.org/abs/1904.11738