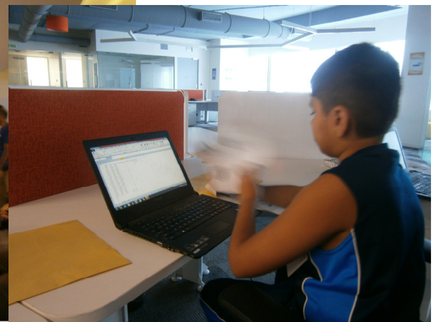



SIGCSE 2017
Inspire, Innovate, Improve!

Data science for kids!

Shashank Srikant **Varun Aggarwal**

And other members from Aspiring Minds Research who made this a success

aspiringminds
Employability Quantified

Data science

Interesting applications

IBM Watson

AlphaGo

NETFLIX

Siri

Voted the sexiest job of the 21st century

Among the highest paying job profiles today

$50B industry

Microsoft

IBM

amazon

Data science

50% gap **in the supply of data scientists versus demand**

McKinsey&Company

aspiringminds
Employability Quantified

Data science

50% gap in the supply of data scientists versus demand

McKinsey&Company

Market reaction

Being introduced in undergraduate courses

Specialized graduate-level courses and doctoral programs

Lots of online material and sandboxes for learning and practice

But what about school curricula?

- We have realized the importance of writing code

- Increasing outreach effort at the high school level

aspiringminds
Employability Quantified

But what about school curricula?

- We have realized the importance of writing code

- Increasing outreach effort at the high school level

What about data science?

aspiringminds
Employability Quantified

Our aim

- Develop a framework to teach the foundations of data science to a school audience
- Lay out the design principles for such a curriculum
- Ensure that it is pedagogically **experiential learning** and **problem-based learning** than being simply lecture content plans
- Demonstrate its utility by implementing it
- Keep the material very accessible. Make it easy for the community to *replicate*.

aspiringminds
Employability Quantified

What is data science?

Collect data of things
that have happened

Study it

New information!

Predictive data science

Supervised learning

Unsupervised learning

aspiringminds
Employability Quantified

Predictive data science

We know what we want to predict

We have concrete values of some outputs.
We then see what can signal such values

Supervised learning

input: What we believe helps predict the desired output

output: What we want to predict

We're trying to learn a function *f* such that
f(input) = output

Predictive data science

	Age	Sport	....	Height	Is sick?
S1	5 years	football		160 cm	Yes
S2	15 years	baseball	....	164 cm	No
S3	15 years	football		128 cm	No
S4	5 years	swimming		149 cm	Yes

aspiringminds
Employability Quantified

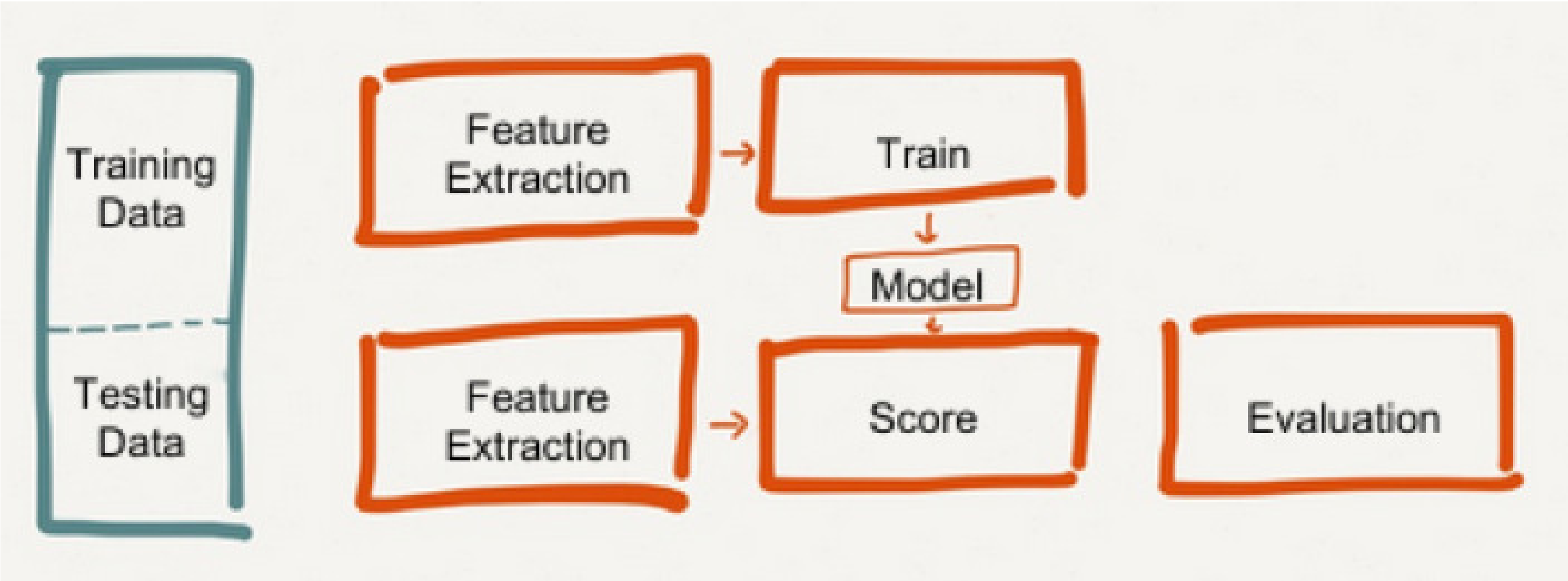
Predictive data science

$f($ Age Sport ···· Height $)$ = Is sick?

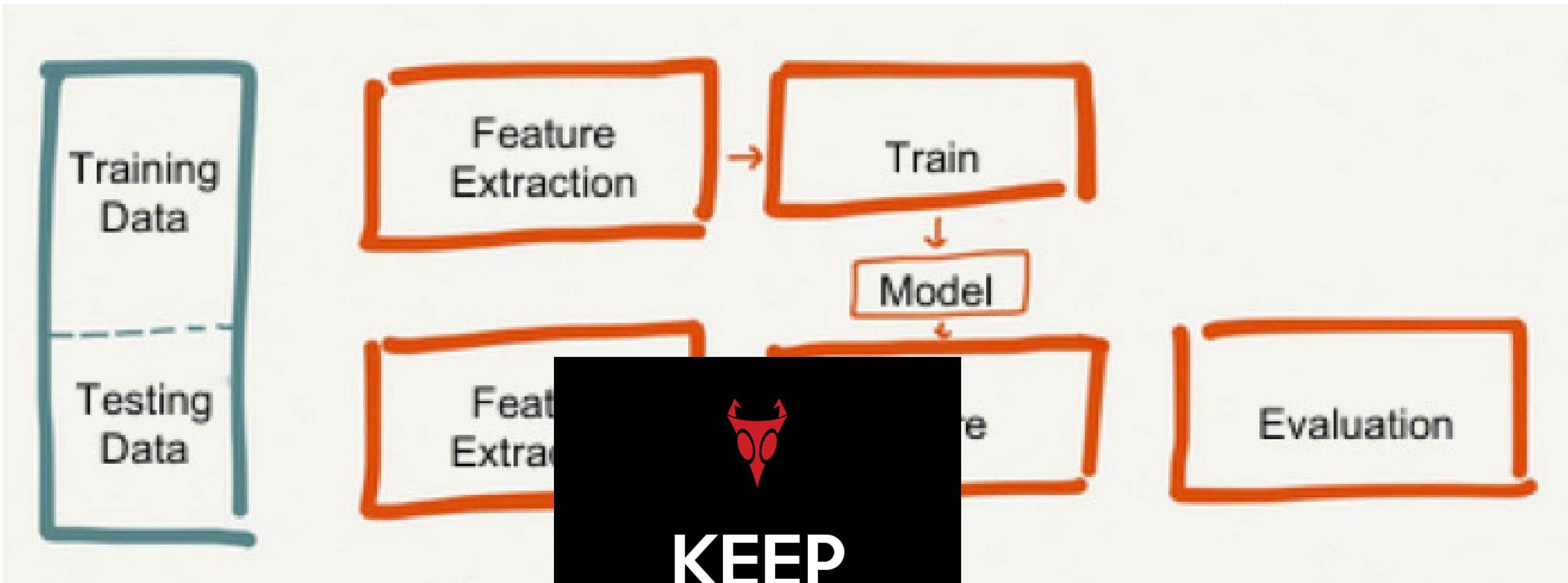
a(Age) + b(Sport) + .. = Is sick?

aspiringminds
Employability Quantified

Overview – Supervised learning



Our aim



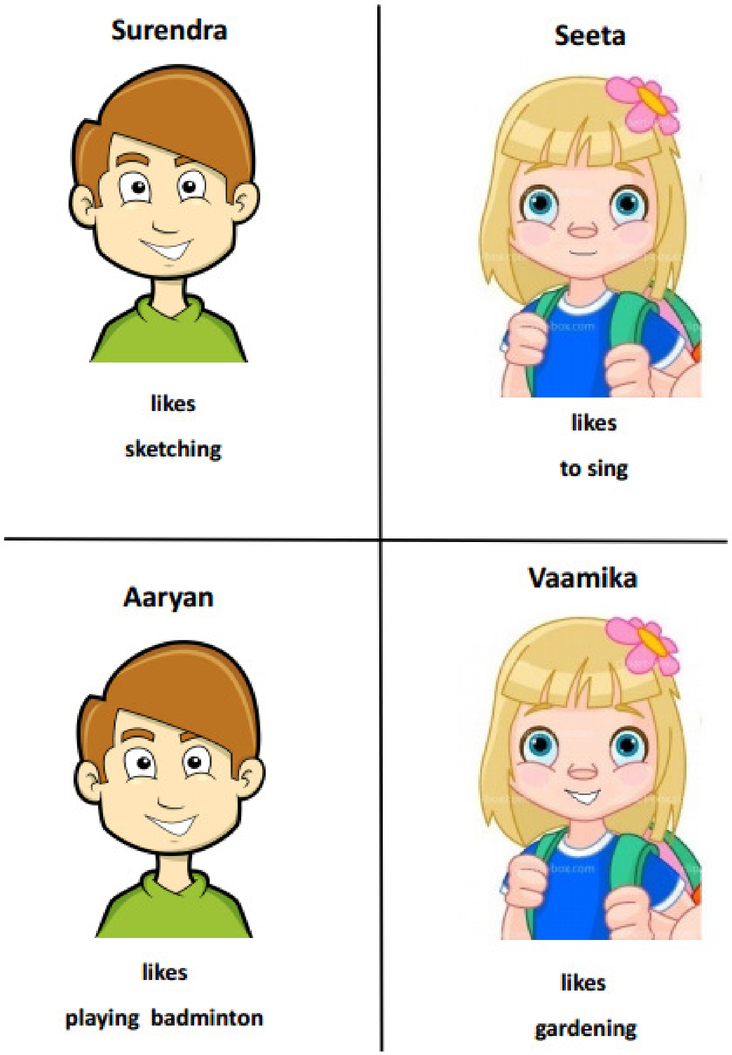
aspiringminds
Employability Quantified

The exercise that we designed

We get them to build a friend predictor

Can we predict whom you will befriend based on friends you've just made?

They first rate a bunch of flash cards



Students tell us whether they would befriend kids we show them on flash cards

These are kids with their names and hobbies described

Students grade them on a scale of 1-5: 1 – least likely to befriend

56 such samples rated

We then discuss possible features

Name

Kathy

Gender

Others?

Hobby

likes
playing musical instruments

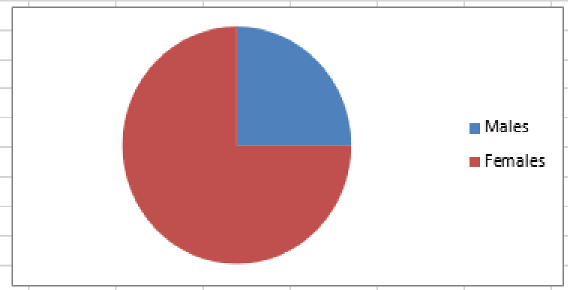
Get them to add this information on a spreadsheet

	A	B	C	D	E	F	G	H	I
1		Student fills these					Calculated results		
2	ID	Name	Gender	Hobby	Extra feat	Rating	Friends?	Predictor	Accuracy
3	1	Old	Male	Indoor		3	1	1	accurate
4	2	Old	Female	Indoor		2	0	1	inaccurate
5	3	New	Male	Outdoor		1	0	0	accurate
6	4	New	Male	Outdoor		4	1	0	inaccurate
7	5	New	Female	Indoor		5	1	1	accurate
8	6	New	Female	Indoor		5	1	1	accurate
9	7	Old	Female	Outdoor		2	0	1	inaccurate
10	8								
11	9								
12	10								

aspiringminds
Employability Quantified

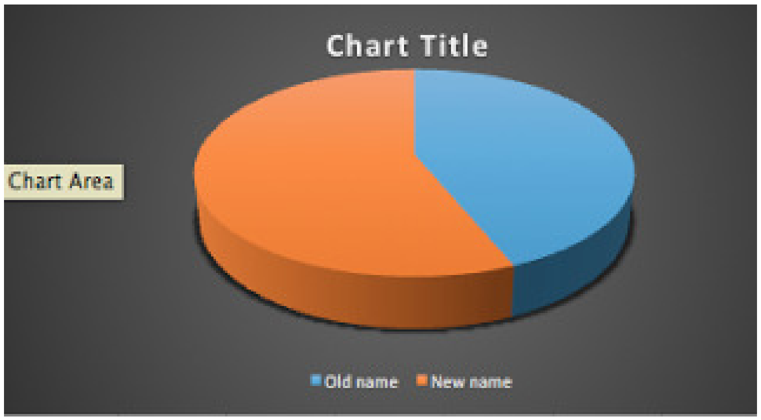
Visualize this information

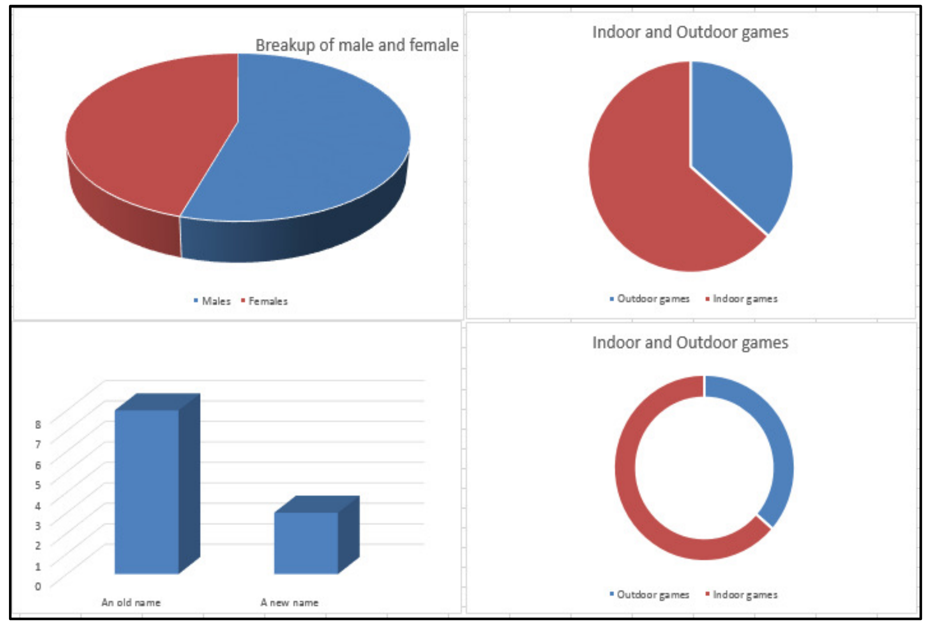
	A	B	C	D	E	F	G	H	I	J	K
1											
2	**Output Visualization**			**Counts**		**Filtered Visualization**				**For plotting**	
3	Wants to really friend										
4	yes, mostly a friend					How many friends are		percentages			
5	May friend					Males	5	25.00		Males	25.00
6	May not friend					Females	15	75.00		Females	75.00
7	Doesn’t want to friend at all						20				
8				0							
9						How many friends have					
10	**Input variables**					An old name		0.00		An old nar	0
11	Males					A new name		0.00		A new nan	0
12	Females						0				
13		32									
14						How many friends like					
15						Outdoor games		0.00		Outdoor g	0
16						Indoor games		0.00		Indoor gai	0
17							0				

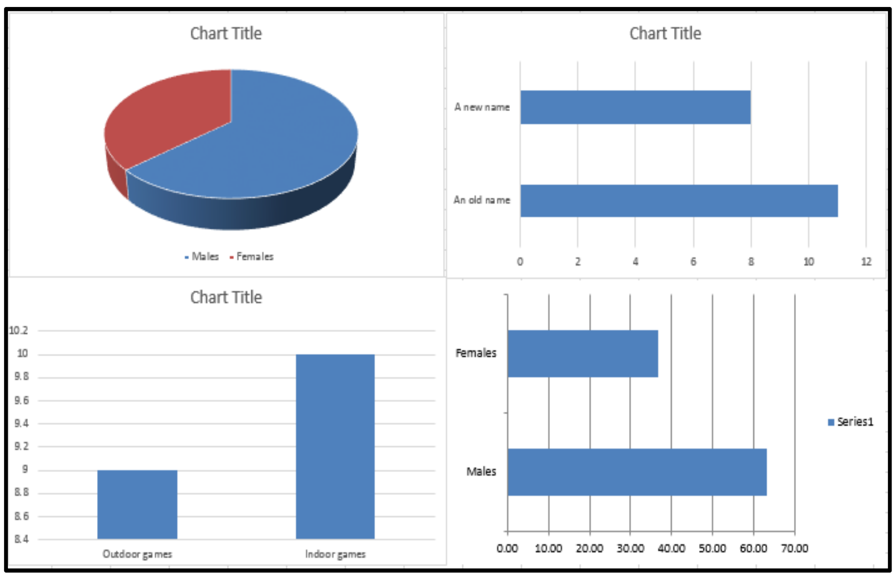


aspiringminds
Employability Quantified

Visualize this information







aspiringminds
Employability Quantified

Build a predictor

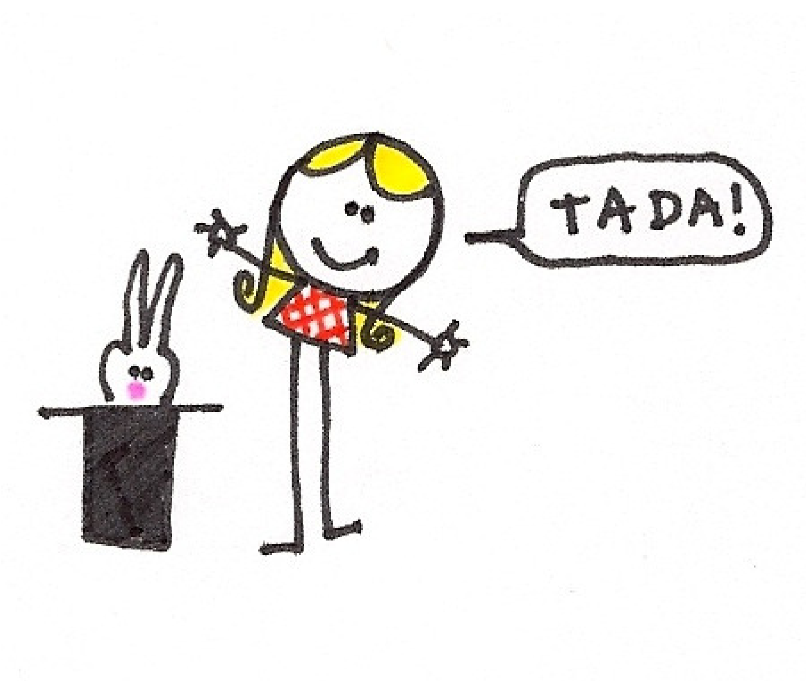
Use a simple IF-clause and a combination of AND, OR conditions

SUM | =IF(OR(B3="Old",D3="Indoor"),1,0)

	A	B	C	D	E	F	G	H	I	J
1		Student fills these					Calculated results			
2	ID	Name	Gender	Hobby	Extra feat	Rating	Friends?	Predictor	Accuracy	
3	1	Old	Male	Indoor		3	1	=IF(OR(B3	accurate	
4	2	Old	Female	Indoor		2	0	1	inaccurate	

aspiringminds
Employability Quantified

Validate predictor on unseen data



Validate the models on the unseen ratings which were held out

Each student builds models of other students

Create an air of suspense to see whether the model does indeed generalize on unseen data

aspiringminds
Employability Quantified

Data consent

Explain importance of anonymity and privacy of data

Get them to participate in providing consent to share their ratings

aspiringminds
Employability Quantified

Crowd source mentors

aspiringminds
Employability Quantified

Some design choices

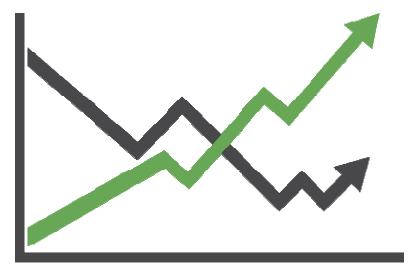
Some design choices

Picking a problem statement

Picking a problem statement

Should not be an obvious exercise in prediction

Should not be too unrelatable



aspiringminds
Employability Quantified

Some design choices

Data collection

Data collection

No to pre-built datasets

Engage students by getting them to upload/play around with the data

aspiringminds
Employability Quantified

Some design choices

Features

Features

Not more than 2-3 features: It gets harder to demonstrate their interaction for anything more

Discrete features: easy to understand and intuitive to handle at the modeling stage

Some design choices

Output

Output

Discrete output: easy to understand and intuitive to handle at the modeling stage

Binary output: easy to understand and visualize

aspiringminds
Employability Quantified

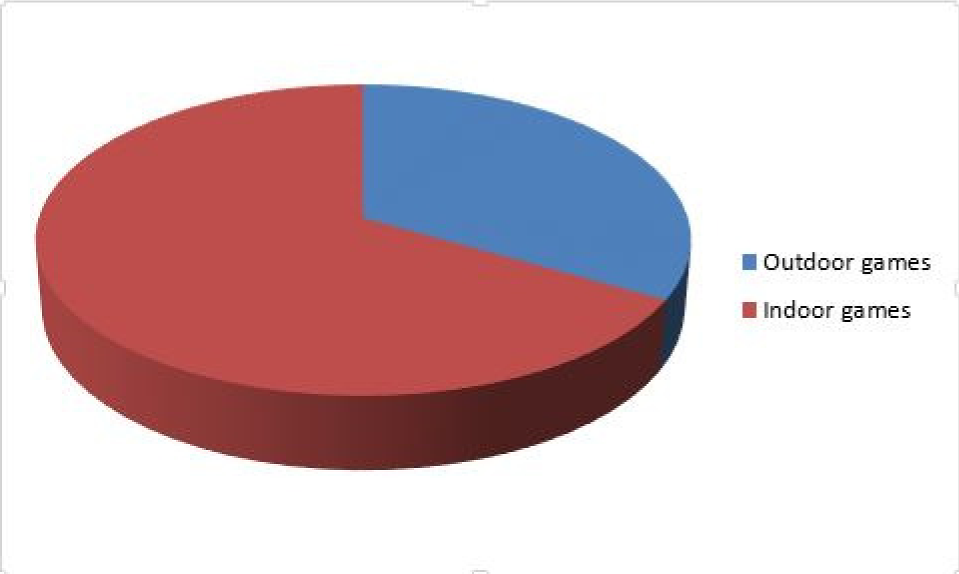
Some design choices

Balanced datasets (inputs)

Balanced datasets

(important)

Look at both classes: **Importance of looking at the *Friend* class and the *Not Friend* class both**



Some design choices

Simple arithmetic, easy tech

Simple arithmetic, easy tech

- Doesn't require anything more than addition, multiplication, division and taking ratios/percentages

- The tool to implement this idea is any commonly available spreadsheet software. Very flat learning curve

- Can always go manual if formulas are not comprehensible

aspiringminds
Employability Quantified

Did students enjoy and learn?

- Conducted this tutorial in 4 cities so far – New Delhi, Pune, Bangalore, UC (Illinois)
- Each cohort had ~18 students in the age group of 10-15 (median age: 13)
- Each session was 3 hours long
- Each of them was given a survey at the end of the tutorial

aspiringminds
Employability Quantified

Did the students enjoy and learn?

Statement	Strongly disagree	Disgree	Agree	Strongly agree
I could understand what the tutor was explaining	0%	1%	32%	67%
I understood how data science is applied to problems	0%	0%	3%	97%
The tutorial was interactive	0%	0%	21%	79%
The tutorial was boring	79%	21%	0%	0%
The tutorial was theoretical	97%	3%	0%	0%
The tutorial was difficult	79%	19%	1%	1%
The tutorial was fast for me	65%	34%	1%	0%
I learned new concepts	0%	0%	2%	98%

aspiringminds
Employability Quantified

Did the students enjoy and learn?

Student prompts

- Want to monitor my pocket money and expenses
- Want to figure out where a criminal resides by clustering the areas of crime
- One of them developed the intuition for overfitting and underfitting while at the tutorial and raised questions about splitting train and test sets

aspiringminds
Employability Quantified

Did the students enjoy and learn?

Features that were brainstormed

- **Artsy** vs. **Non-artsy hobbies**
- **Happy** vs. **Grumpy looking faces**
- **“Weird”** vs. **“Common” name**
- **Hobbies involving hand held tools** vs. **otherwise**

Future work

Reduce TAs

A Scratch equivalent for DS

aspiringminds
Employability Quantified

Key learning

- It is possible to pick up data science and provide a higher level intuition to a young audience without boring them
- Design choices have to be made carefully to ensure students are not burdened with too many new concepts in one session
- Easy to implement and involves a sizeable do-it-yourself component

aspiringminds
Employability Quantified

Data science for kids!

It's more than a hype or a career choice. It's a way of problem solving and thinking which can be used in everyday life. Stories and learnings behind organizing the first ever data camp for kids.

Home | Why data science? | Kids speak | Mentor learnings | Organizing your own camp | Gallery | About AM Research

Data science for kids!

Two weeks back, we decided to experiment teaching data science to grade 6th-9th kids! We think it is important to introduce students to thinking in a data-driven way early on in their lives; also kids are way more fun than higher-ed students, so it was an easier choice for us to make!

We sent out a form asking kids to apply to our cool Kids Data Camp - the first in the world?! We thought kids in 5th grade would have

What's new?

We are on GitHub!

Our pedagogy will be published at SIGCSE 2017!

datasciencekids.org

All our material is on it for you to replicate

We're also on GitHub now

aspiringminds
Employability Quantified

Research at Aspiring Minds

- Define product vision. Research and develop prototypes which demonstrate application of state of the art computer science technology in assessment products
- Publish at the very best conferences and venues
- Multiple outreach programs to popularize data science and machine learning
  - Data science for kids
  - ML-India
  - ASSESS – Annual workshop on data science for assessments
  - AMEO 2015

CoDS 2016
Release AMEO-2016, public dataset on employability outcomes, based on AMCAT data

ACL 2015
Work on machine learning and crowdsourcing in speech evaluation

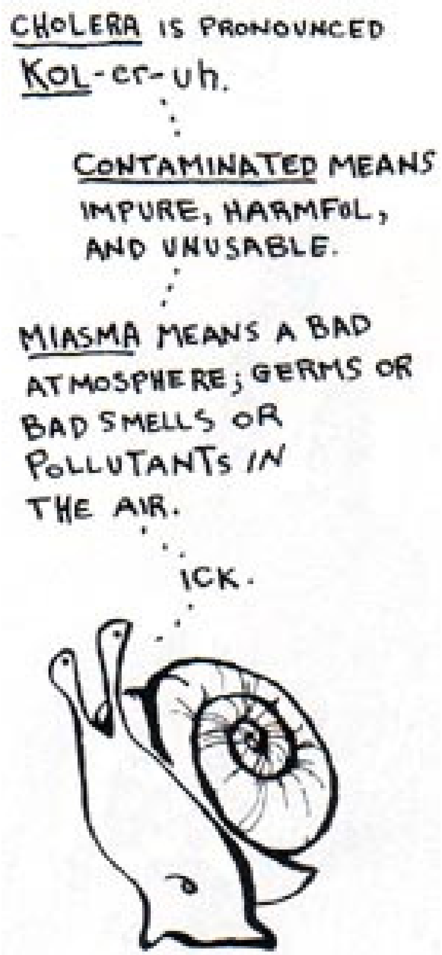
ICML 2015
Work on learning models for job selection

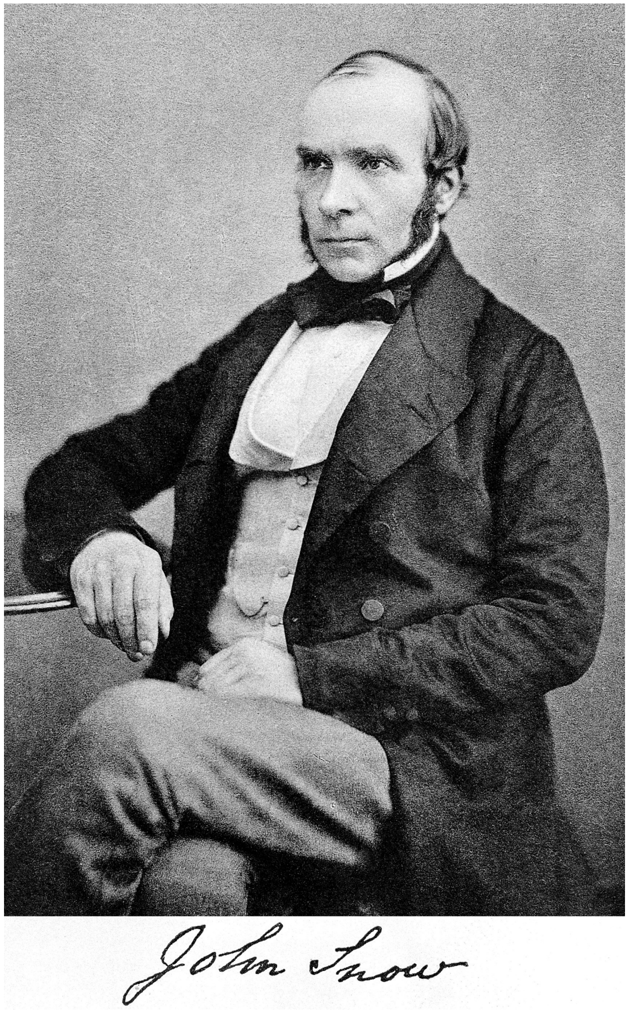
KDD 2014
Work on using machine learning in programming evaluation

NIPS 2013
Framework for using machine learning in assessments

aspiringminds
Employability Quantified

Introduce the broader idea through an example





aspiringminds
Employability Quantified