

[illegible]

Lecturer

2

- Prof Dr Marko Robnik-Šikonja
- University of Ljubljana
Faculty of Computer and Information Science
Laboratory for Cognitive Modeling
- FRI, Večna pot 113, 2nd floor, right from the elevator
- marko.robnik@fri.uni-lj.si
- <https://fri.uni-lj.si/en/employees/marko-robnik-sikonja>
- (01) 4798 241
- Contact hour (see webpage)
 - currently, Mondays, 11:00 -12:00, for other terms, email me
- **Research interests:** machine learning, artificial intelligence, natural language processing, network analytics, data science, data mining, algorithms and data structures
- **Teaching:** several courses from the area of data mining, algorithms, machine learning, and natural language processing
- **Software:** an author of three open source R packages from the area of predictive modelling and data analytics (CORElearn, semiArtificial, ExplainPrediction)



3

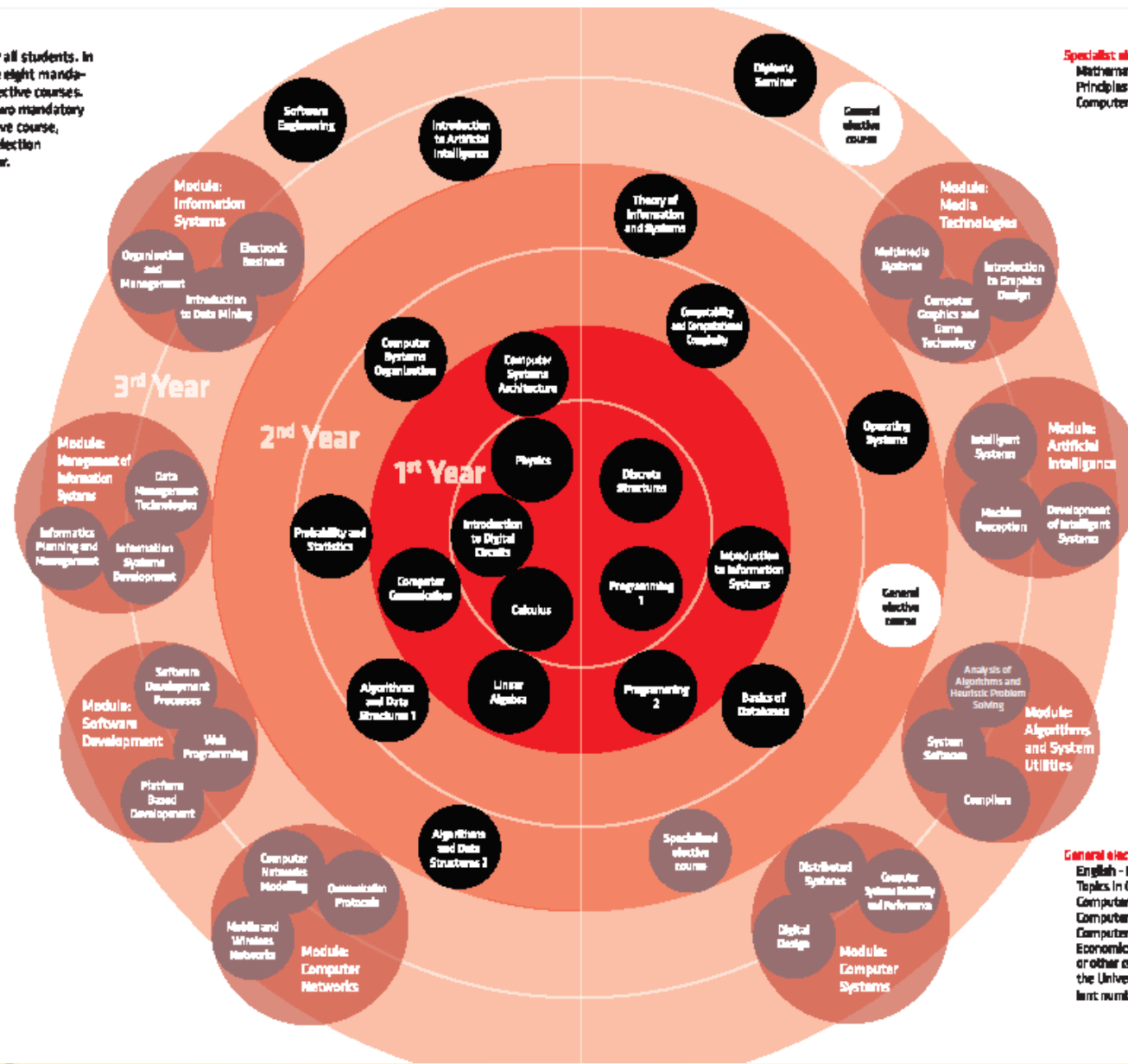
Assistant

- ▶ Tadej Škvorc, PhD student
tadej.skvorc@fri.uni-lj.si
- ▶ Laboratory for Cognitive Modeling
- ▶ tutorials, assignments, work in R
please, prepare questions!

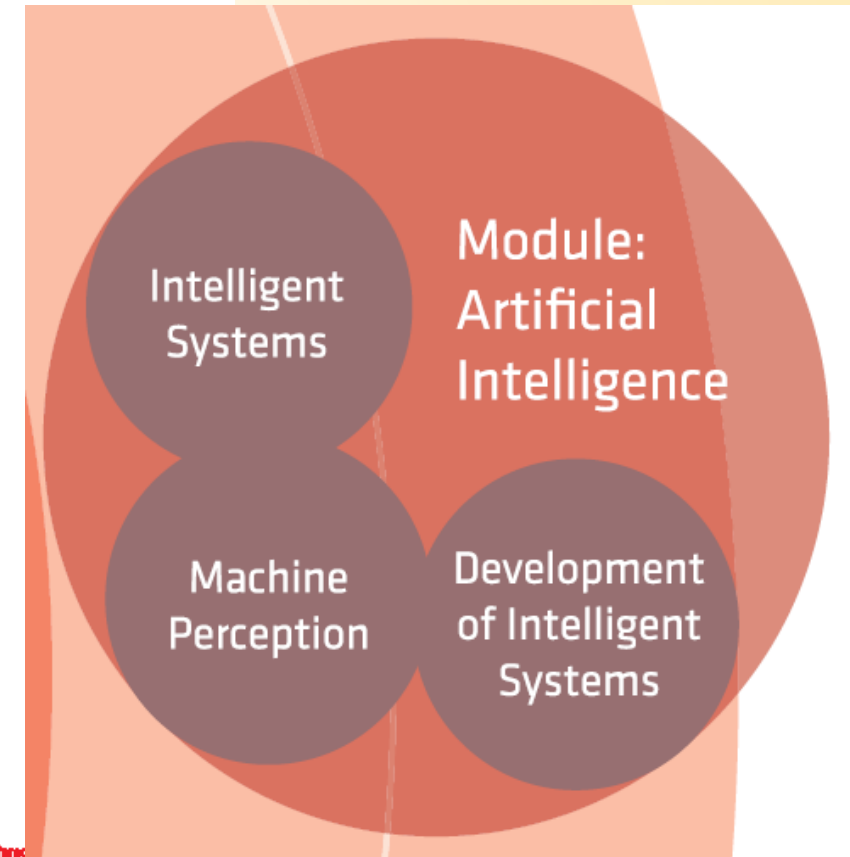


UNIVERSITY STUDY PROGRAMME IN

Year one is the same for all students. In year two, students have eight mandatory courses and two elective courses. In year three there are two mandatory courses, a general elective course, two modules open to selection and the diploma seminar.



Specialist elective courses:
Mathematical Modelling,
Principles of Programming Languages,
Computer Technologies.



General elective courses offered:
English - Level A, B, C,
Topics in Computer and Information Science,
Computer Science in Practice I,
Computer Science in Practice II,
Computer Science Skills,
Economics and Entrepreneurships,
or other course provided by the faculties of
the University of Ljubljana with the equivalent
number of ECTS.



Syllabus

- ▶ nature inspired computing (genetic algorithms, genetic programming, ant colony optimization)
- ▶ basics of machine learning,
- ▶ similarity based learning, kNN
- ▶ decision rules and subgroup discovery
- ▶ data preprocessing (discretization, normalization)
- ▶ embeddings
- ▶ ensemble methods
- ▶ support vector machines
- ▶ neural networks
- ▶ model explanation
- ▶ reinforcement learning
- ▶ natural language processing
- ▶ distributed problem solving and multiagent systems

Objectives

- ▶ students shall become acquainted with
 - ▶ machine learning
 - ▶ model selection and evaluation techniques
 - ▶ model comprehensibility and explanation
 - ▶ practical application of predictive modeling in R programming language and environment
 - ▶ reinforcement learning
 - ▶ nature inspired computing
 - ▶ natural language processing
 - ▶ multiagent systems
- ▶ practical use of theoretical knowledge on (almost) real-world problems
- ▶ awareness of domain expertise and ethical issues in data science
- ▶ increase the (mental) problem-solving toolbox with
 - ▶ predictive modeling techniques
 - ▶ reinforcement learning
 - ▶ result understanding, visualization and explanation approaches
- ▶ for a given prediction problem students shall be able to
 - ▶ transform it to a suitable form suitable for predictive modeling
 - ▶ select and train an appropriate predictive model
 - ▶ evaluate the model and present the results in a comprehensible form and language.

Entry test

7

香港小学一年级学生入学考试题

2013-11-28 关注即可做题-> 易哈佛

Hong Kong Elementary School First Grade Student Admissions Test Question



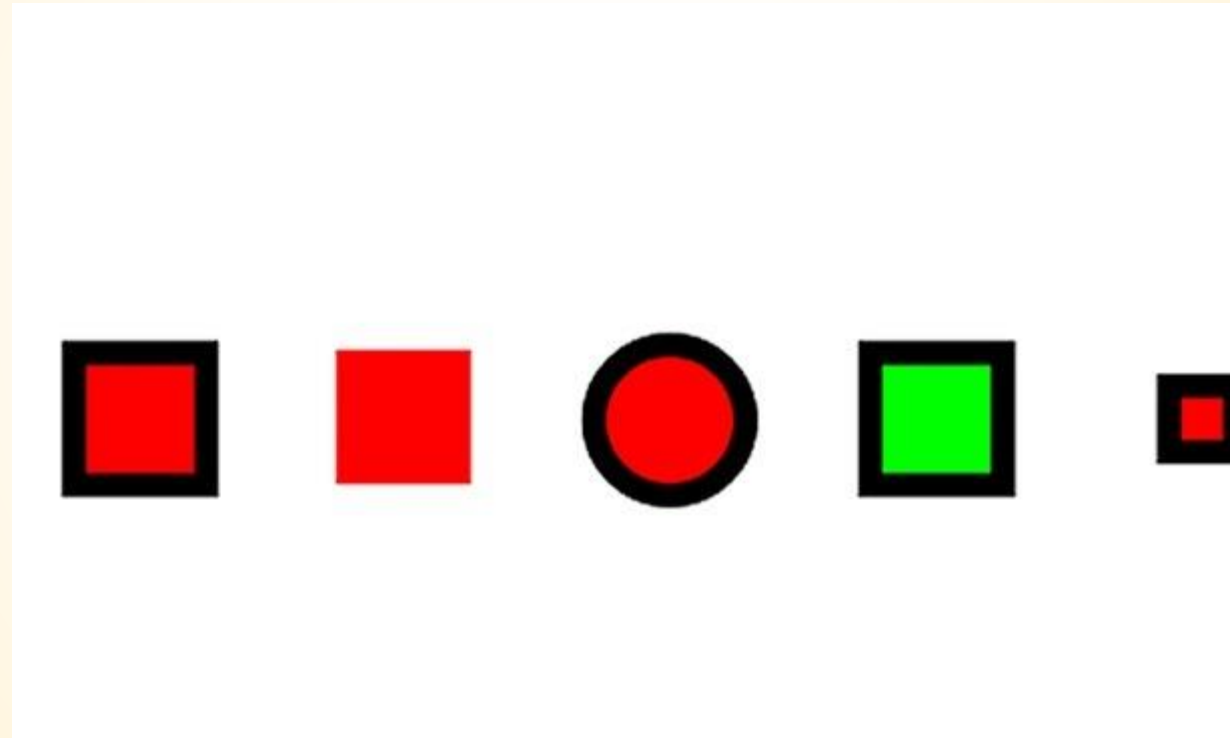
香港小学入学考试题: 21题 Hong Kong Elementary School Admissions Test Question: #21

What parking spot # is the car parked in?

请问汽车停的是几号车位?

请在20秒内完成回答 Please answer within 20 seconds. @微博搞笑排行榜

Odd-one out



This problem can be solved by pre-school children in five to ten minutes, by programmers in an hour and by people with higher education... well, check it yourself!

8809 = 6	5555 = 0
7111 = 0	8193 = 3
2172 = 0	8096 = 5
6666 = 4	1012 = 1
1111 = 0	7777 = 0
3213 = 0	9999 = 4
7662 = 2	7756 = 1
9313 = 1	6855 = 3
0000 = 4	9881 = 5
2222 = 0	5531 = 0
3333 = 0	2581 = ???

Course goals

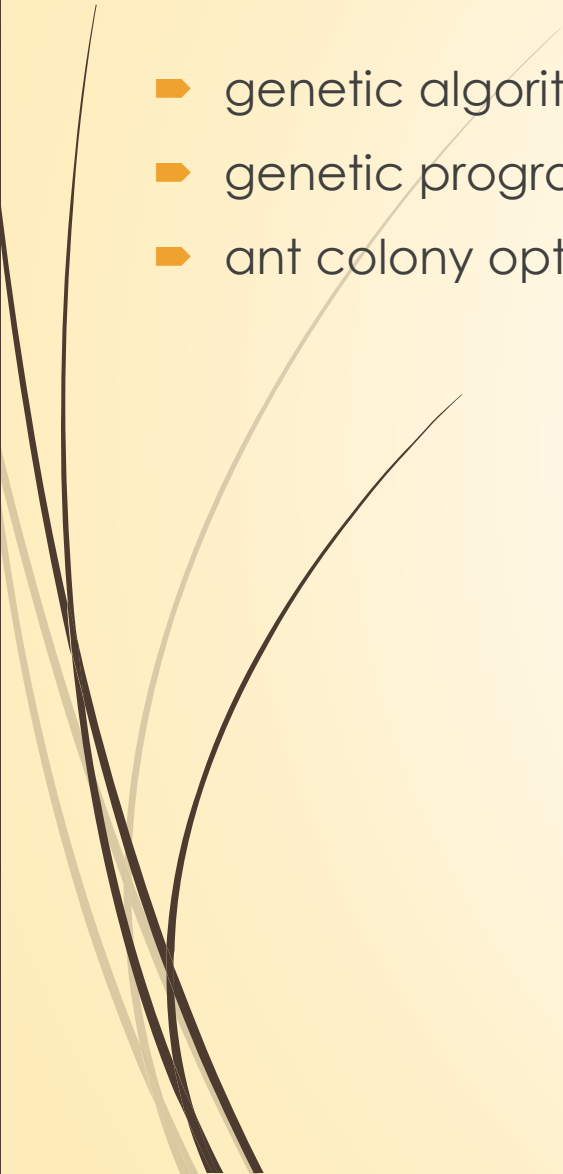
Be able to explain

- ▶ difference between different types of machine learning and models
- ▶ properties of models: bias, variance, generalization, hypothesis language
- ▶ properties of the following models: kNN, decision rules, bagging, boosting, random forests, stacking, SVM, neural networks
- ▶ properties and purpose of evaluation approaches and metrics: cross-validation, bootstrapping, ROC curves, sensitivity, specificity etc.
- ▶ inference methods for predictive methods and explanation of predictions
- ▶ when and why to apply reinforcement learning
- ▶ how to prepare and process text in a text mining test
- ▶ when and how and to optimize a problem using genetic algorithms

Build and evaluate models in R

- ▶ visualize the data set and created models
- ▶ prepare data into a suitable form suitable for modeling algorithms
- ▶ apply classification and regression models to solve a prediction task with a given data set
- ▶ estimate error of models using statistically valid approaches
- ▶ select models and tune their parameters using cross-validation and bootstrapping
- ▶ visualize models and explain their predictions
- ▶ given a new data set select appropriate modeling technique and evaluate the created model

Nature inspired computing

- ▶ genetic algorithms
 - ▶ genetic programming
 - ▶ ant colony optimization
- 
- Several thin, dark, curved lines originate from the bottom left corner and sweep upwards and to the right, creating a sense of movement and flow. They vary in thickness and curvature, some being more pronounced than others.

Introduction to statistical predictive modelling

12

- ▶ Outline of the course, working methods, assignments, exams.
- ▶ Learning as modelling: data, evidence, background knowledge, predictive models, hypotheses, learning as optimization, learning as search, criteria of success, inductive learning, generalization.
- ▶ Classification and regression: supervised and unsupervised learning, learning discrete and numeric functions, learning relations, learning associations.
- ▶ Simple classification models: nearest neighbor, decision rules

Model selection

- ▶ Bias and variance: error decomposition, trade-off, estimating bias and variance.
- ▶ Generalization performance: training and testing set error, cross-validation, evaluation set, bootstrapping.
- ▶ Performance measures: confusion matrix, sensitivity and specificity, ROC curves, AUC, cost-based classification.
- ▶ Parameter tuning: regularization, MDL principle.
- ▶ Calibration of probabilities: binning, isotonic regression.
- ▶ No free lunch theorem.

Ensemble methods

- ▶ Model averaging, why ensembles work.
- ▶ Tree based ensembles: bagging, boosting, random forests.
- ▶ MARS and AODE ensembles.
- ▶ Stacking.

Kernel methods

- ▶ SVM for classification and regression: kernels, support vectors, hyperplanes.
- ▶ SVM for more than two classes: one vs. one, one vs. all.

Neural networks

- perceptron,
- backpropagation,
- RBF networks,
- setting structure of networks
- deep neural networks
- autoencoders
- the role of embeddings

Explaining prediction models

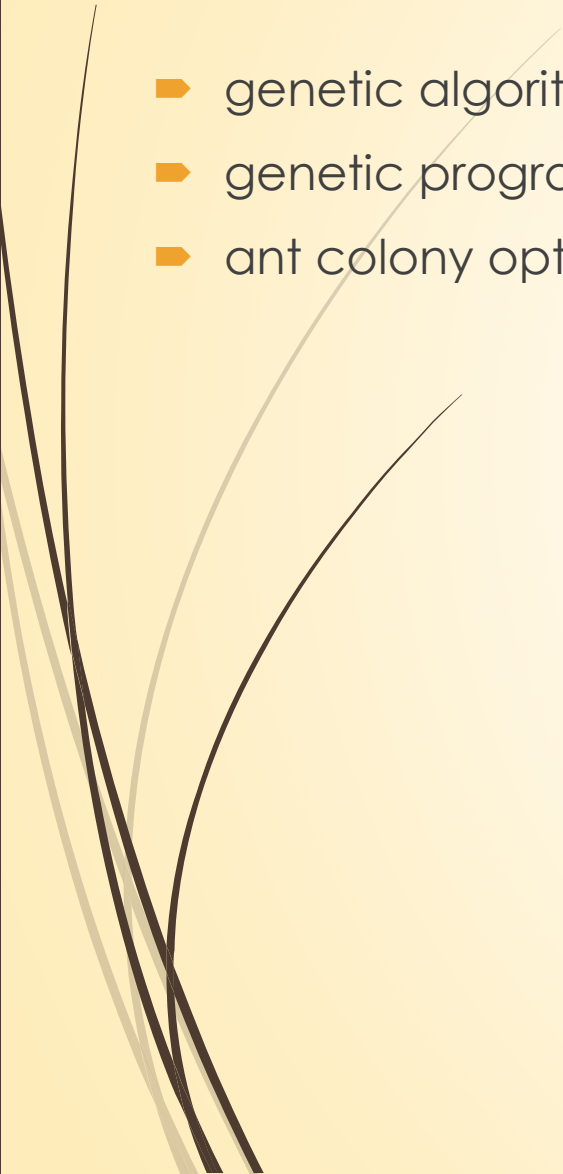
17

- ▶ Model comprehensibility, visualization and knowledge discovery.
- ▶ General methodology for explaining predictive models.
- ▶ Model level and instance level explanations, methods EXPLAIN and IME.

Learning with special settings

- ▶ imbalanced data,
- ▶ multi-task learning,
- ▶ multi-label learning.

Nature inspired computing

- ▶ genetic algorithms
 - ▶ genetic programming
 - ▶ ant colony optimization
- 
- Several thin, dark, curved lines originate from the bottom left corner and sweep upwards and to the right, creating a sense of movement and flow. They vary in thickness and curvature, some being more pronounced than others.

Reinforcement learning

20

- ▶ basics
- ▶ Markov decision problem
- ▶ Q learning

Natural language processing

- ▶ basic preprocessing
- ▶ text similarity
- ▶ text mining
- ▶ sentiment analysis

Multiagent systems

- ▶ types of agents
- ▶ agent architectures
- ▶ distributed constraint satisfaction
- ▶ distributed path finding

Obligations

- 5 quizzes
- two projects, 50 points
- written exam, 50 points

Grading

24

Obligation	% of total	subject to
Five quizzes	0%	$\geq 50\%$
ML project	25%	$\geq 12.5\%$
NLP project	25%	$\geq 12.5\%$
Written exam	50%	$\geq 25\%$

Learning materials

- ▶ learning materials in the eClassroom
- ▶ slides
- ▶ links to textbook and papers
- ▶ R code and examples
- ▶ links to data sets
- ▶ install the open-source systems R and RStudio

Readings

26

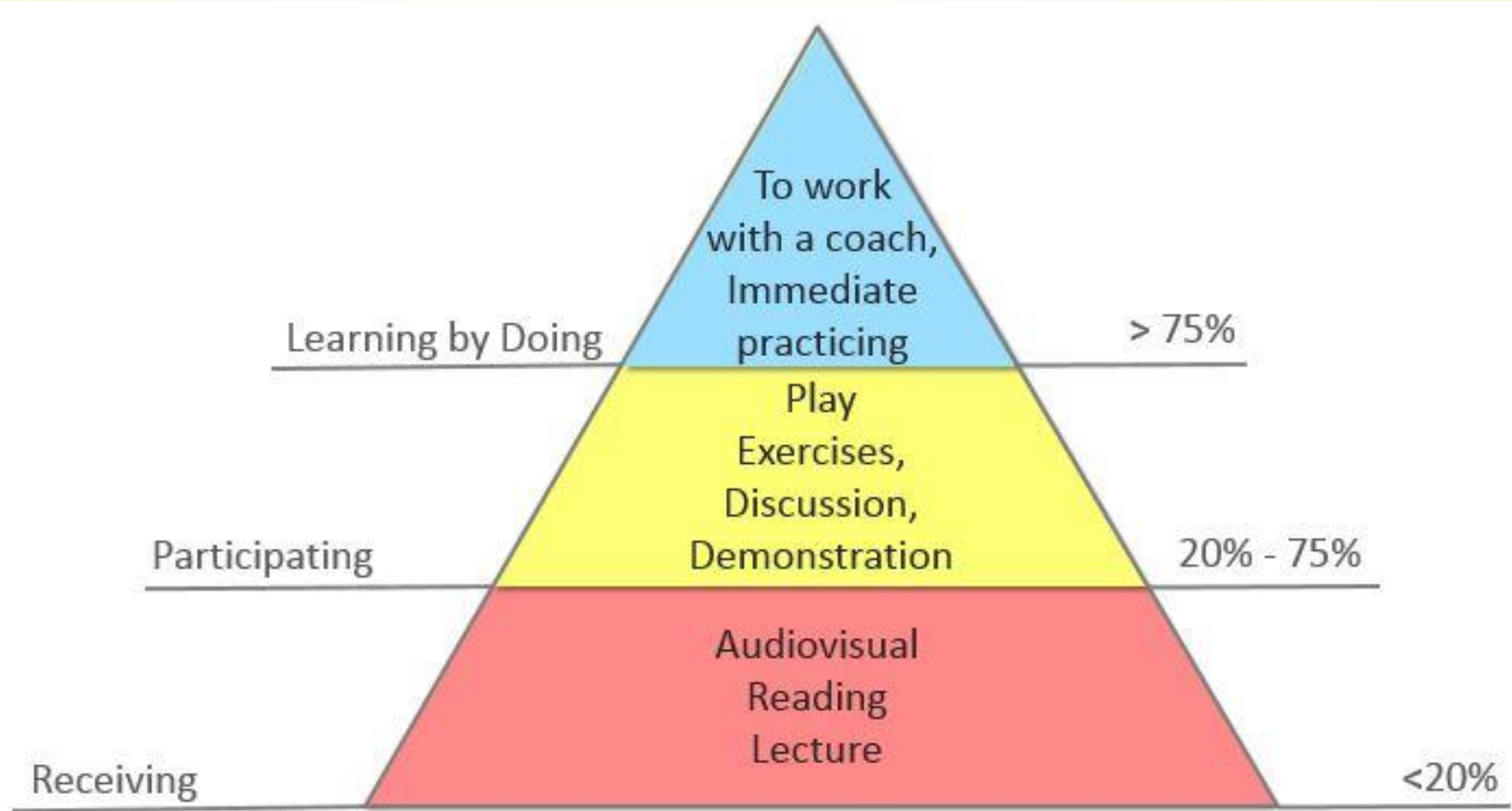
- ▶ James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. Springer, New York.
 - ▶ freely available from authors' homepages
 - ▶ also code and slides from authors and Abbass Al Sharif (some used in this course)

Further readings:

- ▶ Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. Springer, Berlin.
 - ▶ freely available from authors' homepages
- ▶ Kononenko, I., Robnik-Šikonja, M.: *Inteligentni sistemi*. Založba FE in FRI, 2010 (in Slovene)
- ▶ scientific papers
- ▶ many excellent machine learning and data mining courses on Coursera and edX

BTW: retention of learning

27



Retention of Learning

Data Science

good job perspective

- Forbes published list of the most promising jobs in the USA of 2016

1. Data Scientist
(Number of openings: 1,736; Median base salary: \$116,840)
2. Tax Manager
3. Solutions Architect
4. Engagement Manager
5. Mobile Developer
6. HR Manager
7. Physician Assistant
8. Product Manager
9. Software Engineer
10. Audit Manager

- source: job search and salary comparison site [Glassdoor](#), the list of the [25 Best Jobs in America for 2016](#), based on three categories: earning potential, career opportunities, and number of job openings.

- Thomas H. Davenport, D.J. Patil: Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review*, October 2012

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21st century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing packages, e.g., R
- ☆ Databases: SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

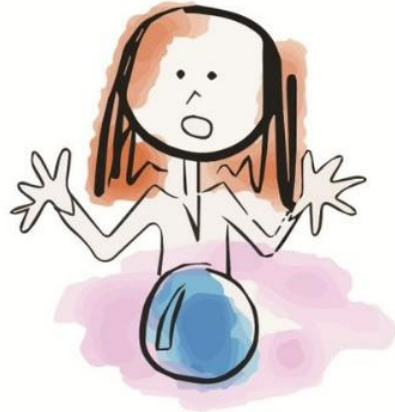




DATA SCIENTIST

30

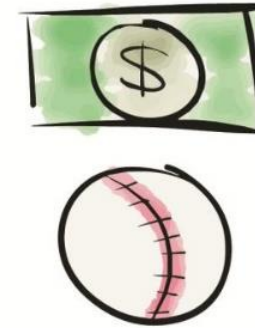
What my **CUSTOMERS** think I do



What my **MUM** thinks I do



What my **FRIENDS** think I do



I work with
Brad
Pitt!!!

What my **HUSBAND** thinks I do



What **I** think I do



What I **ACTUALLY** do



Intelligent systems and media

31



Will robots destroy us?

Will they take our jobs?

Will we still need a driving licence?

Will we still need doctors?

How will humanoid robots evolve?

What about cyborgs?

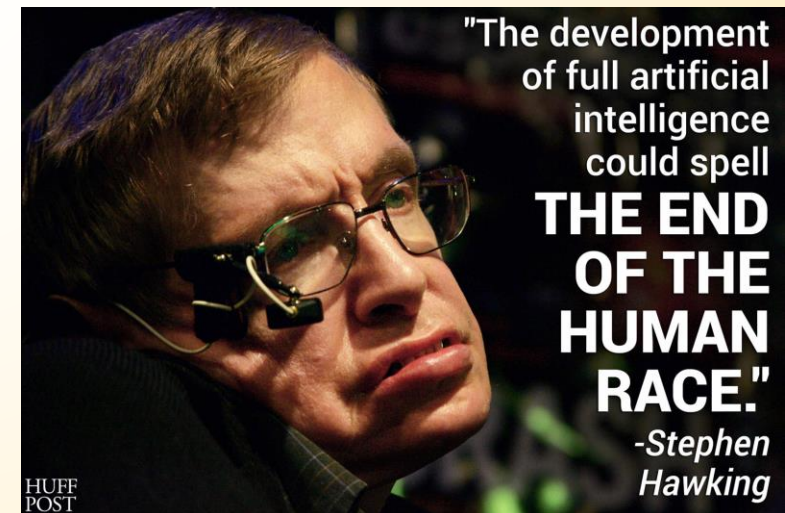
What is artificial general intelligence?

What is technological singularity?

New prophets of technological singularity

Elon Musk says humans must become cyborgs to stay relevant. Is he right?

Sophisticated artificial intelligence will make 'house cats' of humans, claims the entrepreneur, but his grand vision for mind-controlled tech may be a long way off



Some scientific responses

- ▶ Rodney Brooks: The Seven Deadly Sins of Predicting the Future of AI.
<https://rodneybrooks.com/the-seven-deadly-sins-of-predicting-the-future-of-ai/> tudi MIT Technology Review
- ▶ Marko Robnik-Šikonja: Is artificial intelligence a (job) killer?. The Conversation, Jul. 2017
<https://theconversation.com/is-artificial-intelligence-a-job-killer-80473>
- ▶ ...



Short history of optimism

- ▶ starting in 1950s,
1956 Dartmouth conference
- ▶ great expectations, enormous underestimation
of problem difficulty
- ▶ AI winter (2 x)



1958, H. A. Simon and Allen Newell: "... within ten years a digital computer will discover and prove an important new mathematical theorem."

1965, H. A. Simon: "... machines will be capable, within twenty years, of doing any work a man can do."

1967, Marvin Minsky: "Within a generation ... the problem of creating 'artificial intelligence' will substantially be solved."

1970, Marvin Minsky: "In from three to eight years we will have a machine with the general intelligence of an average human being."

Gartner Hype Cycle for Emerging Technologies, 2017

