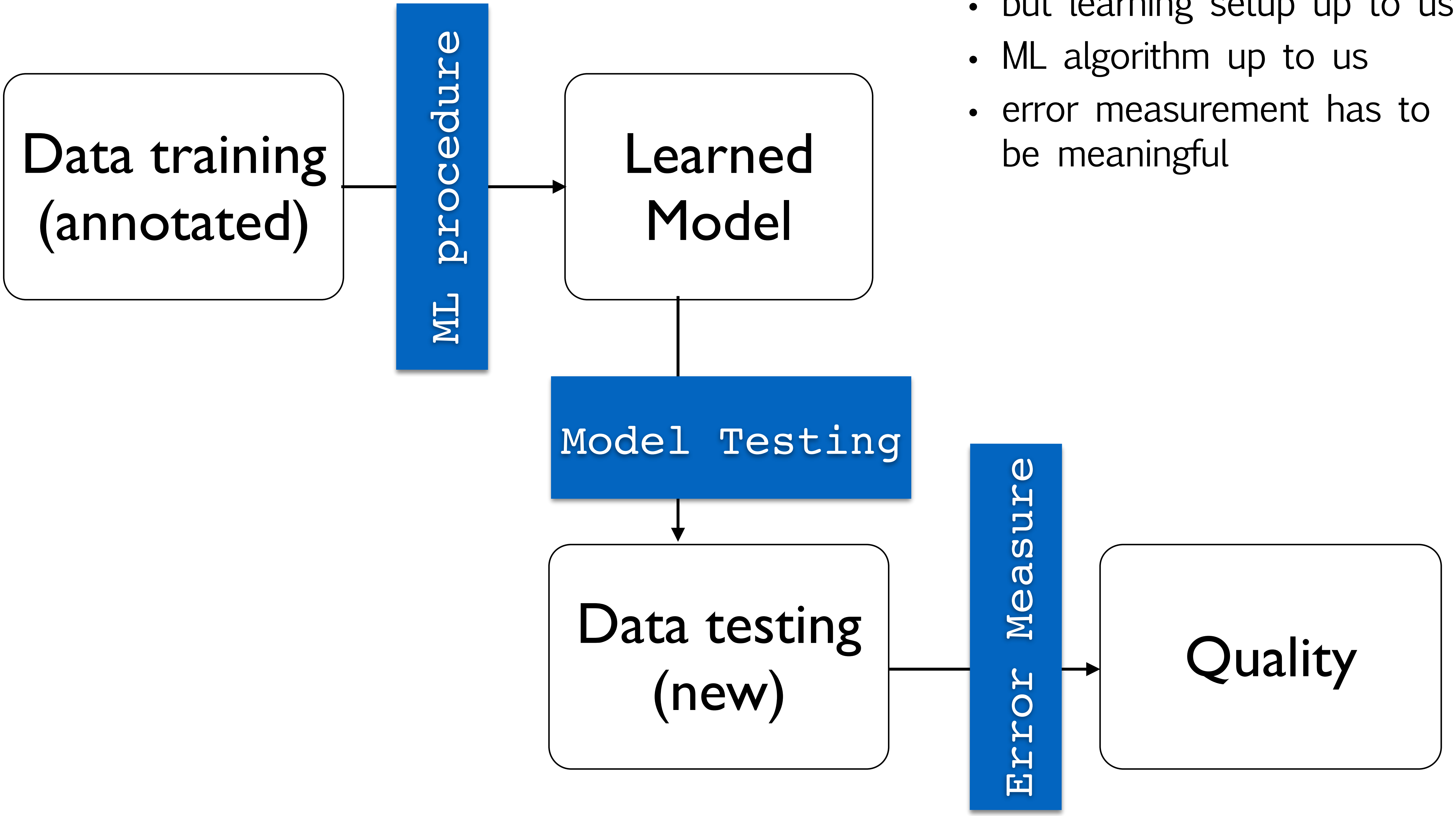
Intro to Machine Learning

Module 1 Objectives / Intro, Evaluation

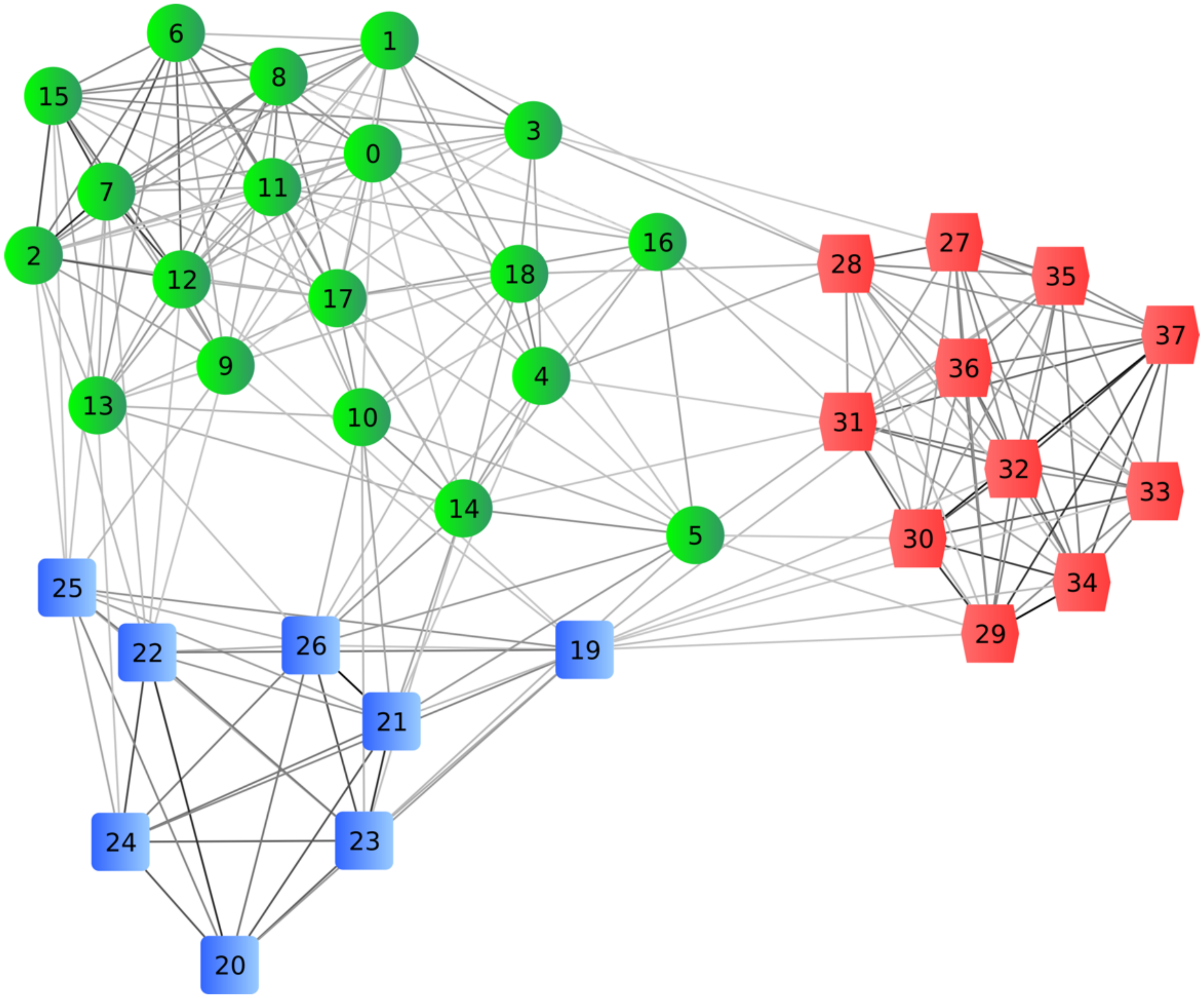
- Intro to Machine Learning - what is learning ?
- Data - Matrix type
  - algebraic notations
- Heuristics and Quantitative rules
- Error measurement
  - training VS testing error, Cross Validation
  - overfitting

What is machine learning ? Supervised learning



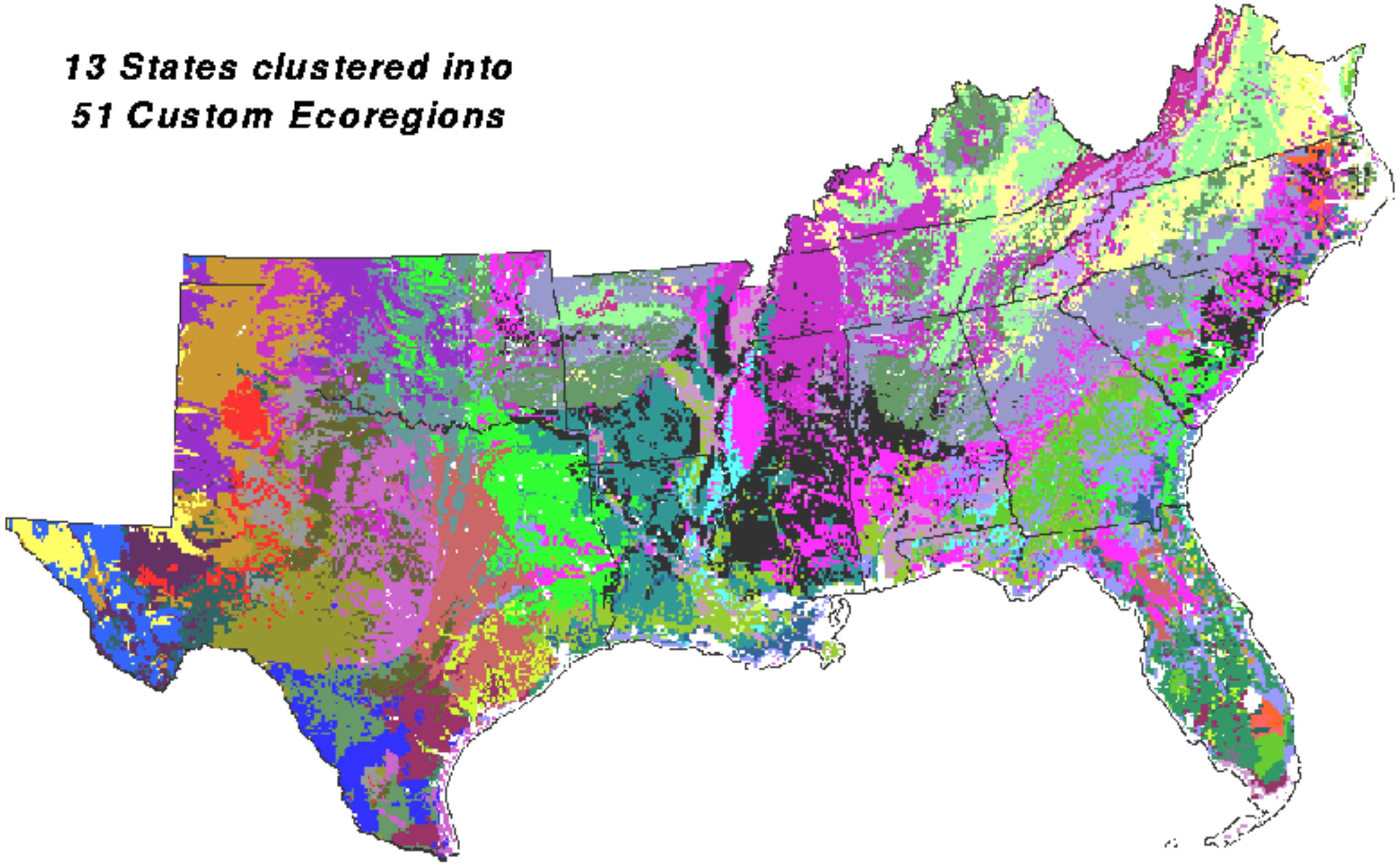
- data given with labels
- but learning setup up to us
- ML algorithm up to us
- error measurement has to be meaningful

What is machine learning ? Graph learning



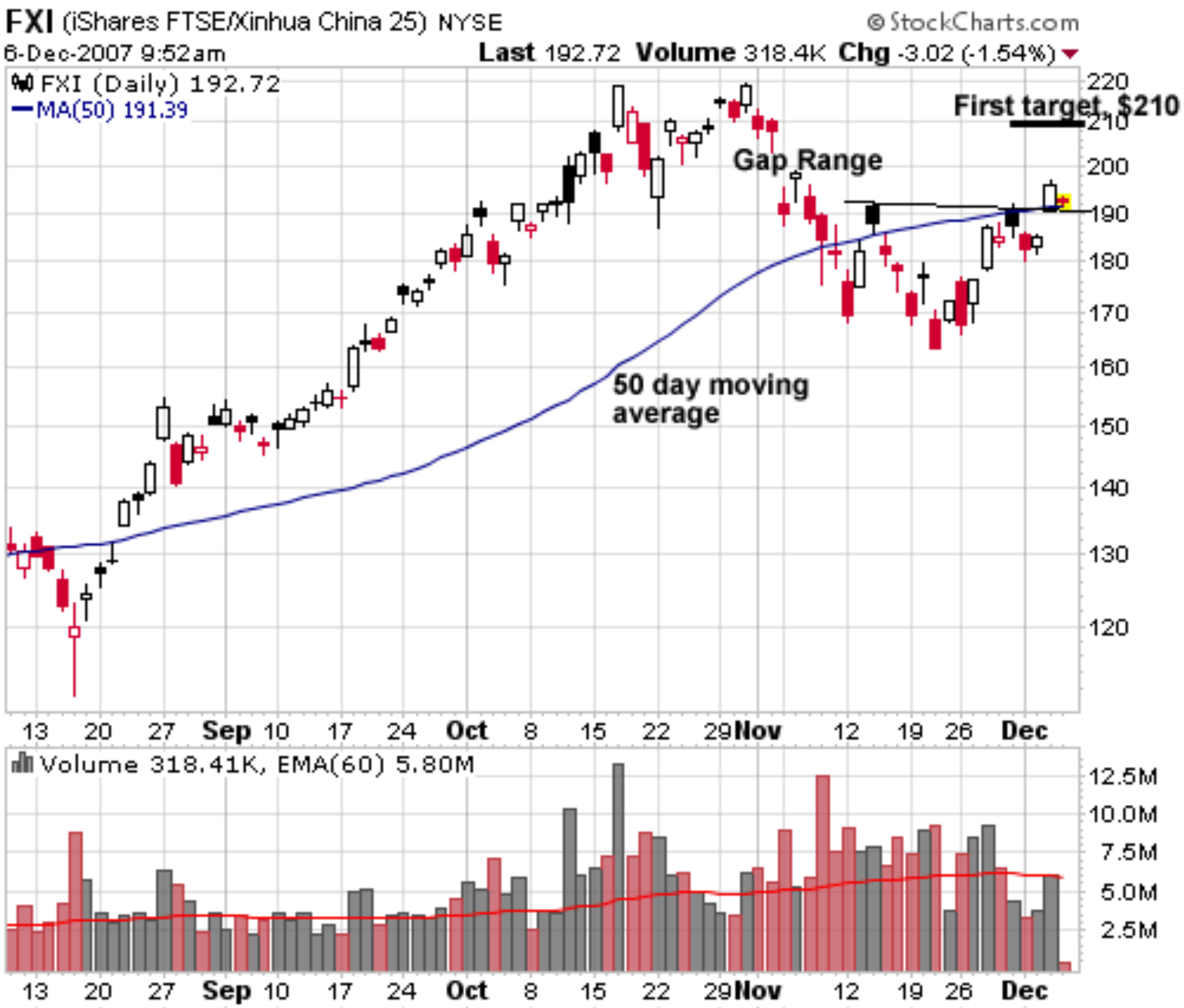
- data defined by links or analogies or connections
- for example social networks, or web links
- task: identify object properties from links
- taks: detect graph patterns

What is machine learning ? Clustering



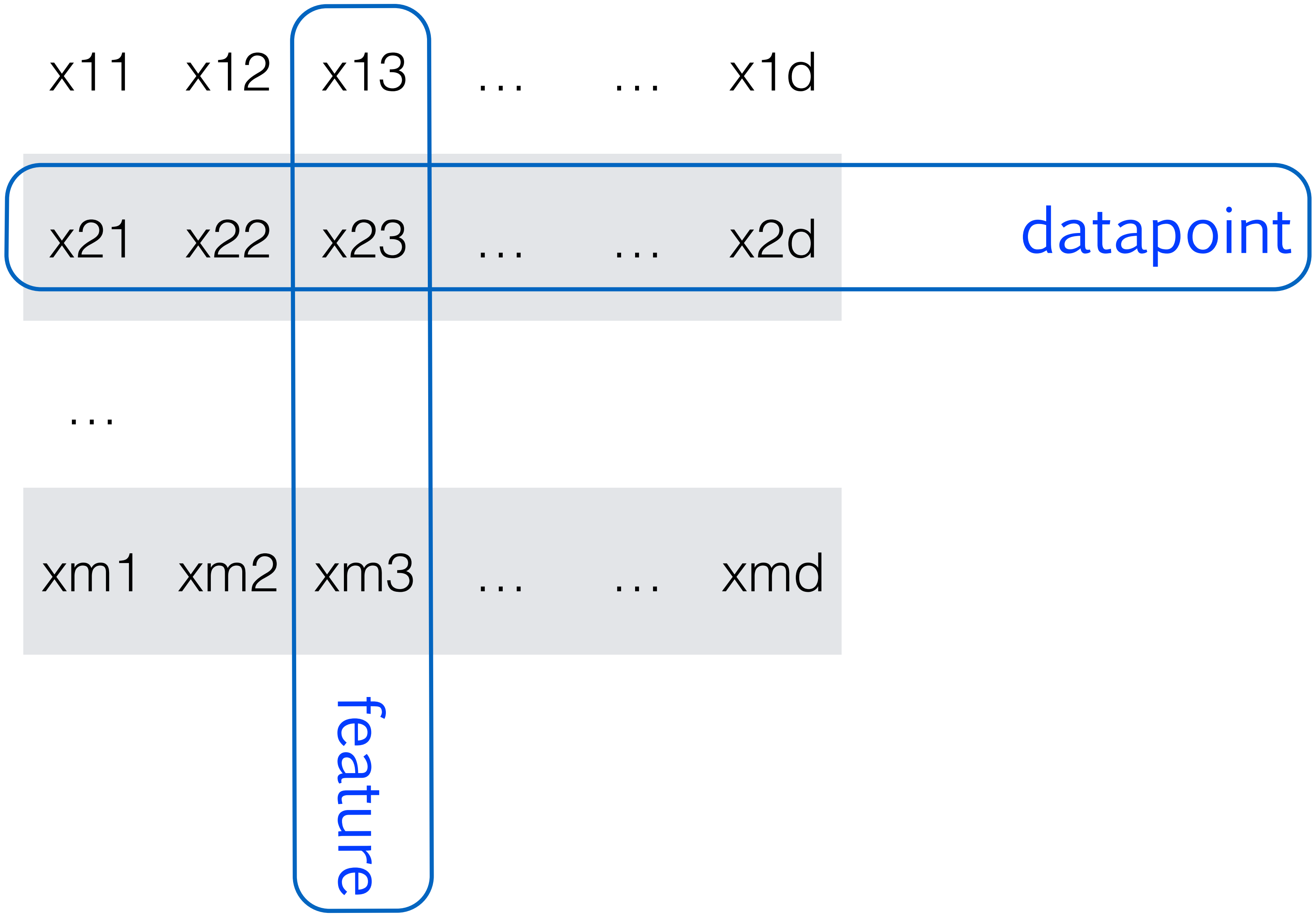
- data given without labels
- task: group similar data points

What is machine learning ? Time series analysis



- data that evolves with time
- like stocks or patient records
- task: predict future behavior
- task: detect anomalies

Matrix data



x11	x12	x13	…	…	x1d
x21	x22	x23	…	…	x2d
…					
xm1	xm2	xm3	…	…	xmd

datapoint

feature

- m datapoints/objects X=(x1,x2,…,xd)
- d features/columns f1, f2, …, fd

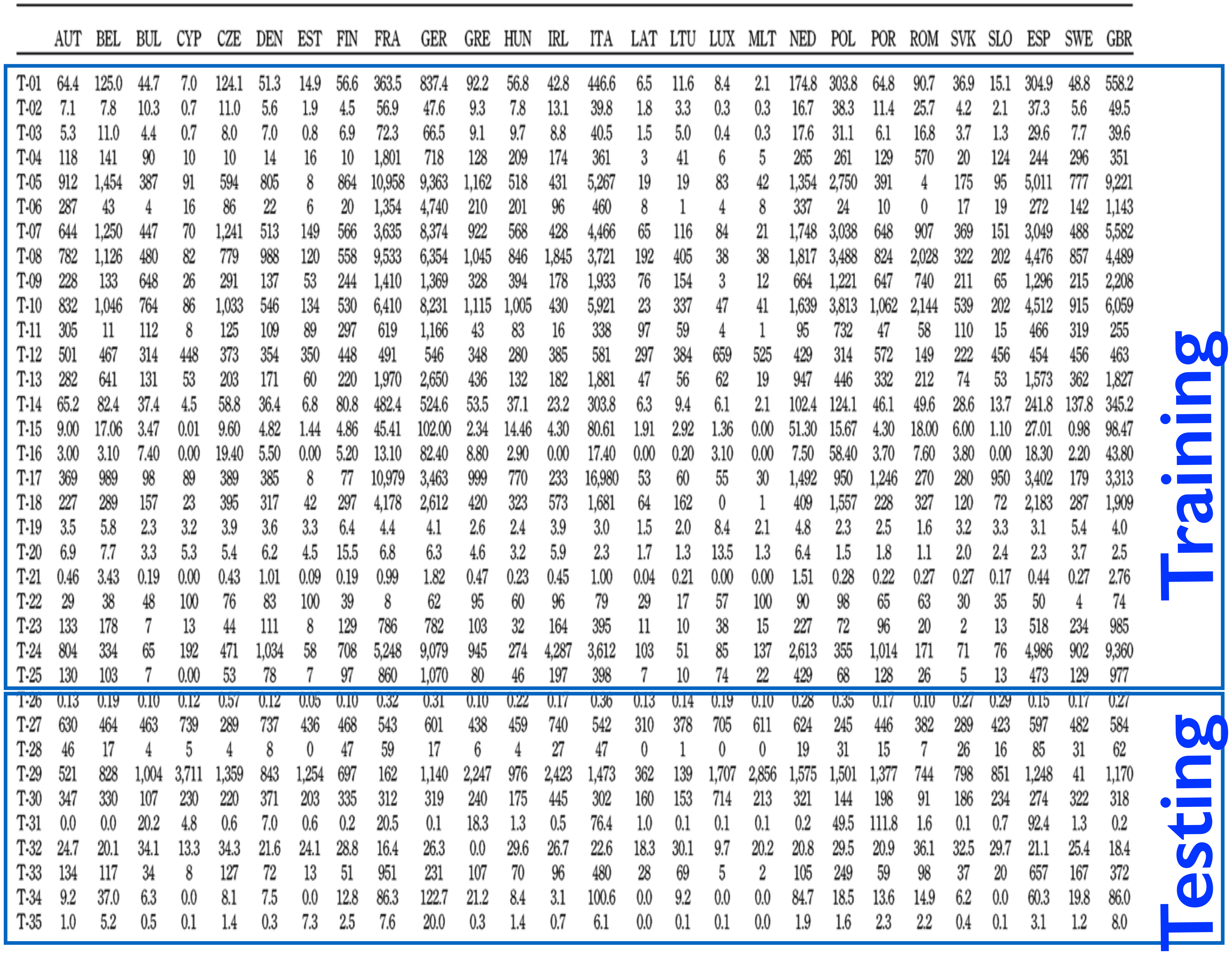
Heuristic rules / decisional

- If fever>100, patient has flu
- If email contains words "free" or "porn", it is spam
- If a web page contains ngram "Michael Jackson", it is relevant to the user
- If age<22 and sex=F and highschool_diploma=Yes, then eligible for application
- If income_per_capita<$1000, region prone to civil war
- If romantic=Yes and comedy=Yes and Orlando_Bloom=Yes, then movie success among females aged 20-40
- If Nasdaq_Computer_Index=Gain and Apple announces new Ipad, then AAPL_Stock=Buy

Heuristic rules / quantitative

- if 3*exam_grade+2*HW_grade>55, then student can pass
- if blood_pressure/log(age)>3, recommend medicine
- if rent+food+bills<1/2 salary, loan for 1/2 salary possible

Matrix data / training VS testing

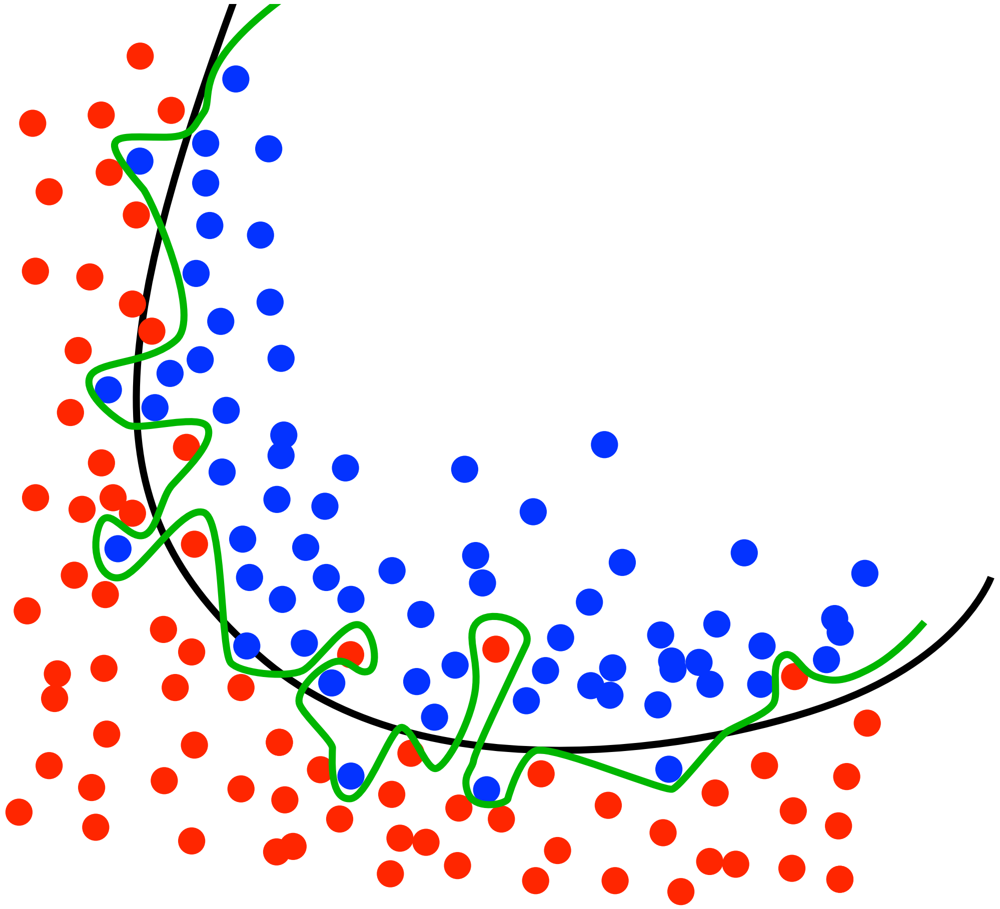


Training
Testing

	AUT	BEL	BUL	CYP	CZE	DEN	EST	FIN	FRA	GER	GRE	HUN	IRL	ITA	LAT	LTU	LUX	MLT	NED	POL	POR	ROM	SVK	SLO	ESP	SWE	GBR
T-01	64.4	125.0	44.7	7.0	124.1	51.3	14.9	56.6	363.5	837.4	92.2	56.8	42.8	446.6	6.5	11.6	8.4	2.1	174.8	303.8	64.8	90.7	36.9	15.1	304.9	48.8	558.2
T-02	7.1	7.8	10.3	0.7	11.0	5.6	1.9	4.5	56.9	47.6	9.3	7.8	13.1	39.8	1.8	3.3	0.3	0.3	16.7	38.3	11.4	25.7	4.2	2.1	37.3	5.6	49.5
T-03	5.3	11.0	4.4	0.7	8.0	7.0	0.8	6.9	72.3	66.5	9.1	9.7	8.8	40.5	1.5	5.0	0.4	0.3	17.6	31.1	6.1	16.8	3.7	1.3	29.6	7.7	39.6
T-04	118	141	90	10	10	14	16	10	1,801	718	128	209	174	361	3	41	6	5	265	261	129	570	20	124	244	296	351
T-05	912	1,454	387	91	594	805	8	864	10,958	9,363	1,162	518	431	5,267	19	19	83	42	1,354	2,750	391	4	175	95	5,011	777	9,221
T-06	287	43	4	16	86	22	6	20	1,354	4,740	210	201	96	460	8	1	4	8	337	24	10	0	17	19	272	142	1,143
T-07	644	1,250	447	70	1,241	513	149	566	3,635	8,374	922	568	428	4,466	65	116	84	21	1,748	3,038	648	907	369	151	3,049	488	5,582
T-08	782	1,126	480	82	779	988	120	558	9,533	6,354	1,045	846	1,845	3,721	192	405	38	38	1,817	3,488	824	2,028	322	202	4,476	857	4,489
T-09	228	133	648	26	291	137	53	244	1,410	1,369	328	394	178	1,933	76	154	3	12	664	1,221	647	740	211	65	1,296	215	2,208
T-10	832	1,046	764	86	1,033	546	134	530	6,410	8,231	1,115	1,005	430	5,921	23	337	47	41	1,639	3,813	1,062	2,144	539	202	4,512	915	6,059
T-11	305	11	112	8	125	109	89	297	619	1,166	43	83	16	338	97	59	4	1	95	732	47	58	110	15	466	319	255
T-12	501	467	314	448	373	354	350	448	491	546	348	280	385	581	297	384	659	525	429	314	572	149	222	456	454	456	463
T-13	282	641	131	53	203	171	60	220	1,970	2,650	436	132	182	1,881	47	56	62	19	947	446	332	212	74	53	1,573	362	1,827
T-14	65.2	82.4	37.4	4.5	58.8	36.4	6.8	80.8	482.4	524.6	53.5	37.1	23.2	303.8	6.3	9.4	6.1	2.1	102.4	124.1	46.1	49.6	28.6	13.7	241.8	137.8	345.2
T-15	9.00	17.06	3.47	0.01	9.60	4.82	1.44	4.86	45.41	102.00	2.34	14.46	4.30	80.61	1.91	2.92	1.36	0.00	51.30	15.67	4.30	18.00	6.00	1.10	27.01	0.98	98.47
T-16	3.00	3.10	7.40	0.00	19.40	5.50	0.00	5.20	13.10	82.40	8.80	2.90	0.00	17.40	0.00	0.20	3.10	0.00	7.50	58.40	3.70	7.60	3.80	0.00	18.30	2.20	43.80
T-17	369	989	98	89	389	385	8	77	10,979	3,463	999	770	233	16,980	53	60	55	30	1,492	950	1,246	270	280	950	3,402	179	3,313
T-18	227	289	157	23	395	317	42	297	4,178	2,612	420	323	573	1,681	64	162	0	1	409	1,557	228	327	120	72	2,183	287	1,909
T-19	3.5	5.8	2.3	3.2	3.9	3.6	3.3	6.4	4.4	4.1	2.6	2.4	3.9	3.0	1.5	2.0	8.4	2.1	4.8	2.3	2.5	1.6	3.2	3.3	3.1	5.4	4.0
T-20	6.9	7.7	3.3	5.3	5.4	6.2	4.5	15.5	6.8	6.3	4.6	3.2	5.9	2.3	1.7	1.3	13.5	1.3	6.4	1.5	1.8	1.1	2.0	2.4	2.3	3.7	2.5
T-21	0.46	3.43	0.19	0.00	0.43	1.01	0.09	0.19	0.99	1.82	0.47	0.23	0.45	1.00	0.04	0.21	0.00	0.00	1.51	0.28	0.22	0.27	0.27	0.17	0.44	0.27	2.76
T-22	29	38	48	100	76	83	100	39	8	62	95	60	96	79	29	17	57	100	90	98	65	63	30	35	50	4	74
T-23	133	178	7	13	44	111	8	129	786	782	103	32	164	395	11	10	38	15	227	72	96	20	2	13	518	234	985
T-24	804	334	65	192	471	1,034	58	708	5,248	9,079	945	274	4,287	3,612	103	51	85	137	2,613	355	1,014	171	71	76	4,986	902	9,360
T-25	130	103	7	0.00	53	78	7	97	860	1,070	80	46	197	398	7	10	74	22	429	68	128	26	5	13	473	129	977
T-26	0.13	0.19	0.10	0.12	0.57	0.12	0.05	0.10	0.32	0.31	0.10	0.22	0.17	0.36	0.13	0.14	0.19	0.10	0.28	0.35	0.17	0.10	0.27	0.29	0.15	0.17	0.27
T-27	630	464	463	739	289	737	436	468	543	601	438	459	740	542	310	378	705	611	624	245	446	382	289	423	597	482	584
T-28	46	17	4	5	4	8	0	47	59	17	6	4	27	47	0	1	0	0	19	31	15	7	26	16	85	31	62
T-29	521	828	1,004	3,711	1,359	843	1,254	697	162	1,140	2,247	976	2,423	1,473	362	139	1,707	2,856	1,575	1,501	1,377	744	798	851	1,248	41	1,170
T-30	347	330	107	230	220	371	203	335	312	319	240	175	445	302	160	153	714	213	321	144	198	91	186	234	274	322	318
T-31	0.0	0.0	20.2	4.8	0.6	7.0	0.6	0.2	20.5	0.1	18.3	1.3	0.5	76.4	1.0	0.1	0.1	0.1	0.2	49.5	111.8	1.6	0.1	0.7	92.4	1.3	0.2
T-32	24.7	20.1	34.1	13.3	34.3	21.6	24.1	28.8	16.4	26.3	0.0	29.6	26.7	22.6	18.3	30.1	9.7	20.2	20.8	29.5	20.9	36.1	32.5	29.7	21.1	25.4	18.4
T-33	134	117	34	8	127	72	13	51	951	231	107	70	96	480	28	69	5	2	105	249	59	98	37	20	657	167	372
T-34	9.2	37.0	6.3	0.0	8.1	7.5	0.0	12.8	86.3	122.7	21.2	8.4	3.1	100.6	0.0	9.2	0.0	0.0	84.7	18.5	13.6	14.9	6.2	0.0	60.3	19.8	86.0
T-35	1.0	5.2	0.5	0.1	1.4	0.3	7.3	2.5	7.6	20.0	0.3	1.4	0.7	6.1	0.0	0.1	0.1	0.0	1.9	1.6	2.3	2.2	0.4	0.1	3.1	1.2	8.0

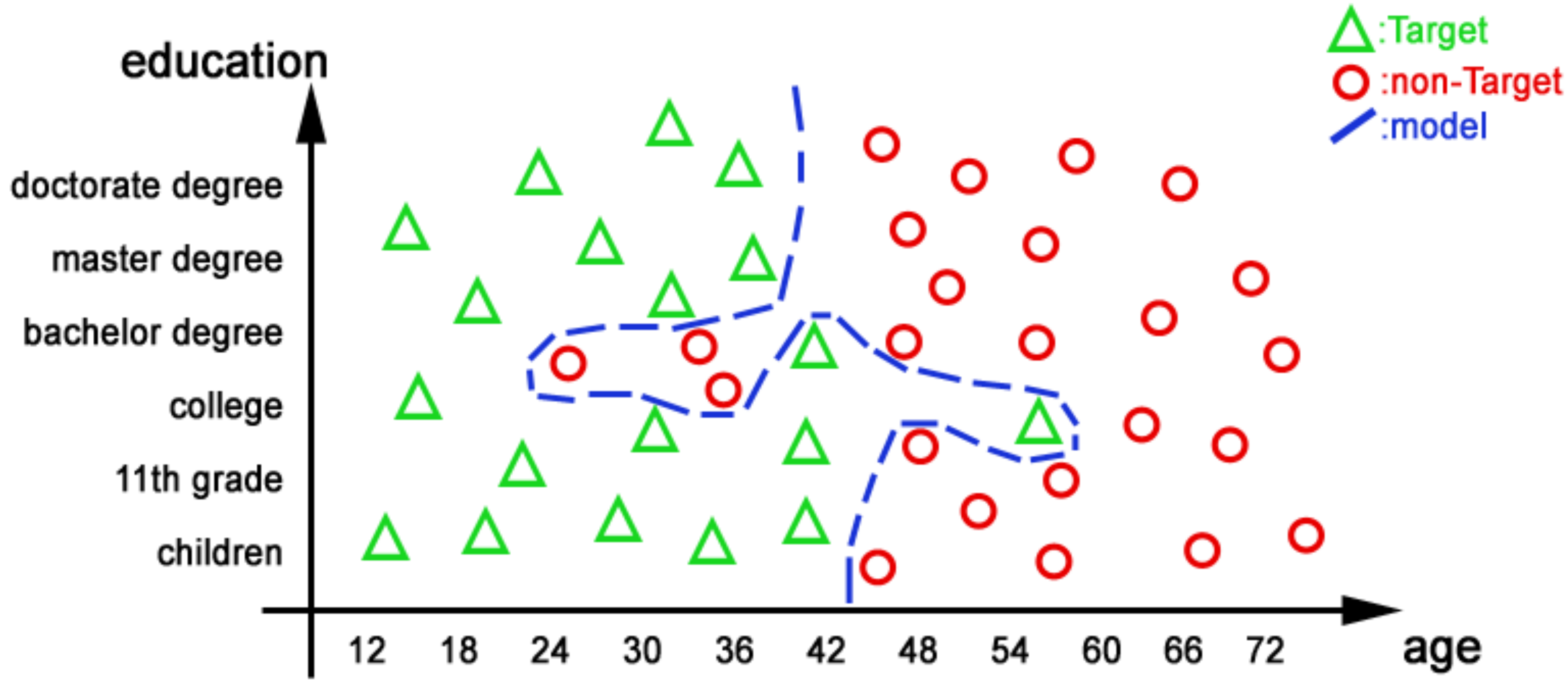
- testing set has to be independent of training set
  - or else testing result is inconclusive
  - and not reliable
- usually the data is partitioned before running any ML algorithm

Overfitting



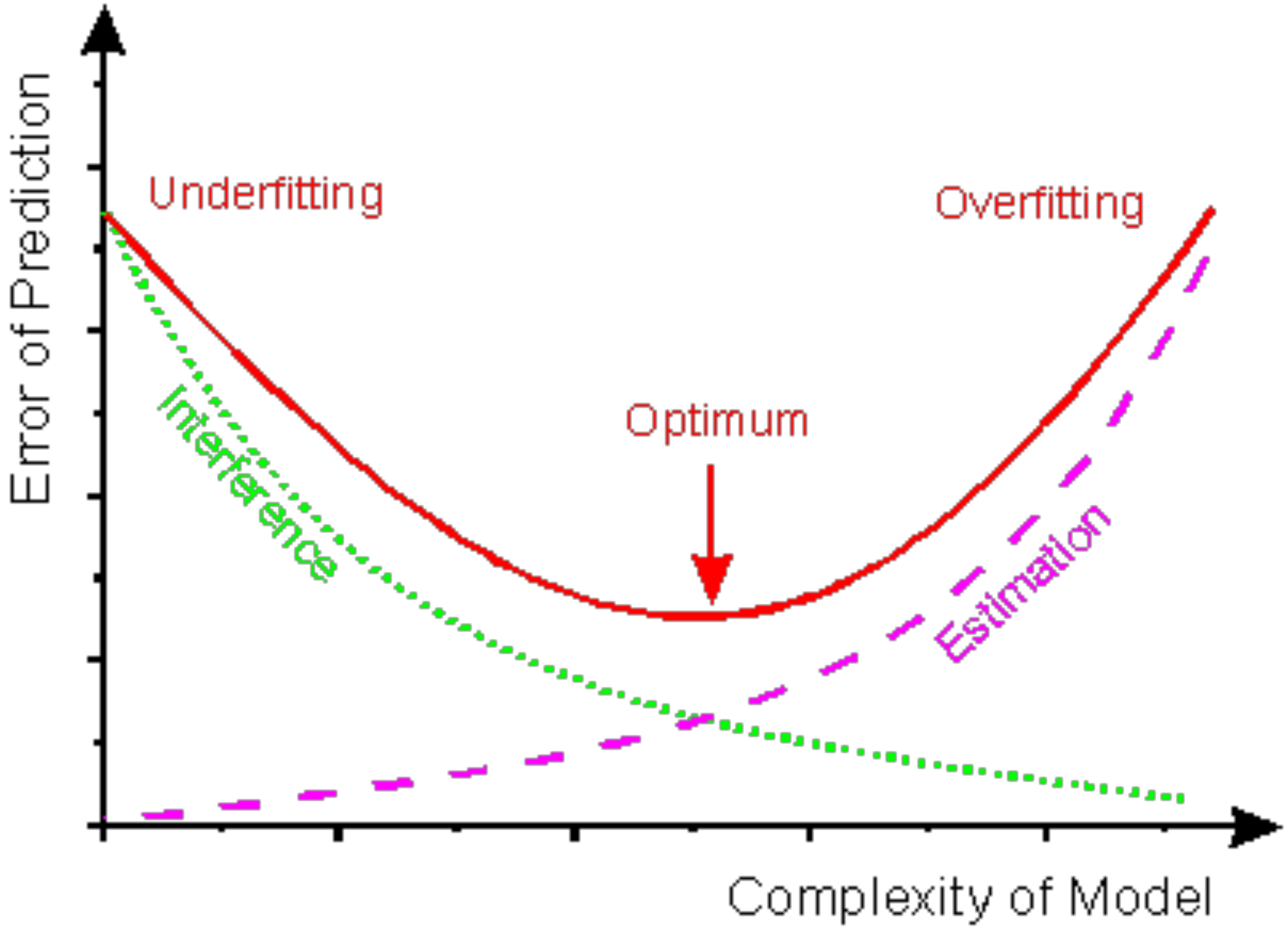
- might be capable to create a model that essentially memorizes all training dataset
  - for example a decision tree deep enough
- that is not useful : the purpose of the learning model is to applicable to new data (testing)

Overfitting



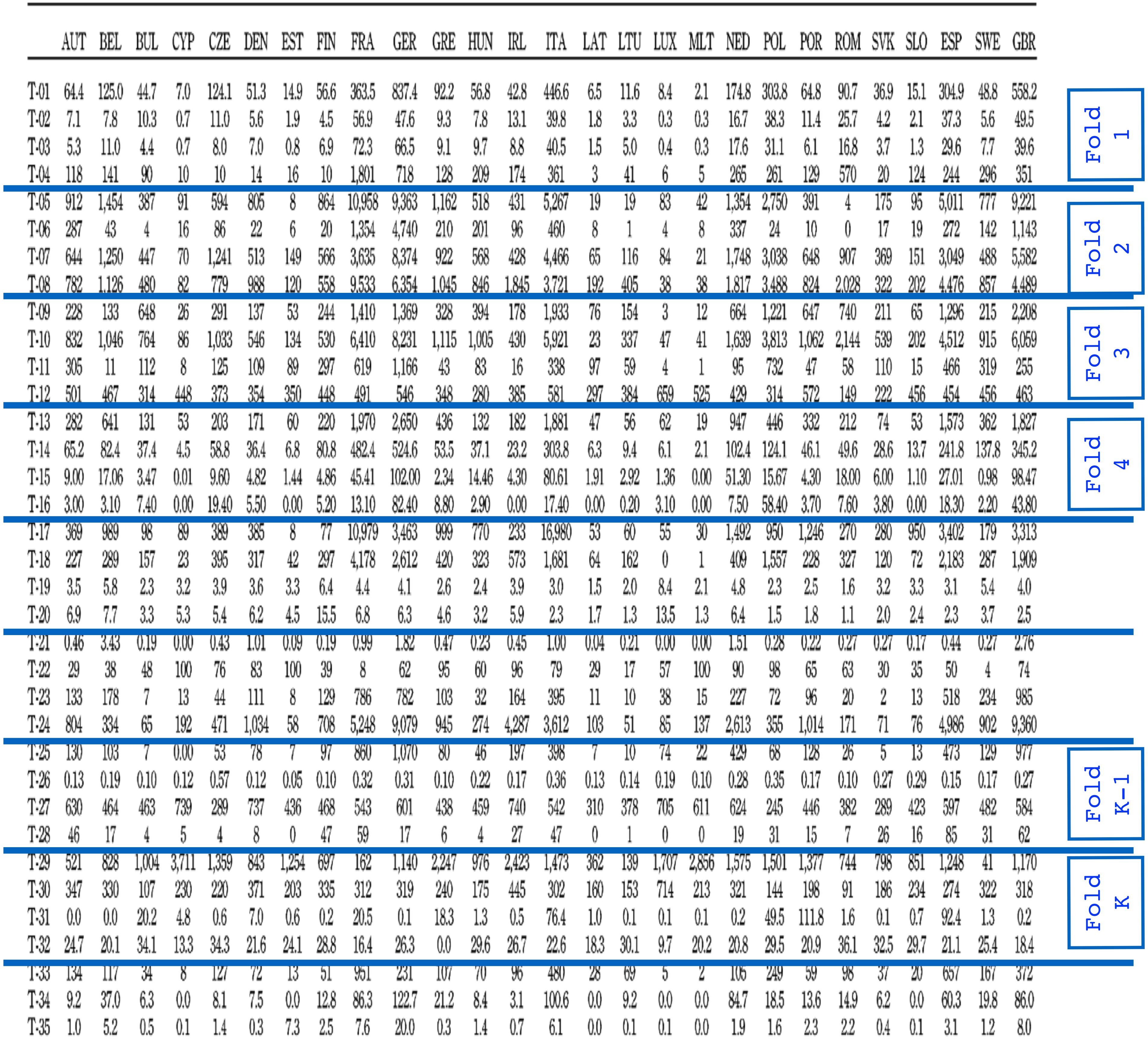
- might be capable to create a model that essentially memorizes all training dataset
  - for example a decision tree deep enough
- that is not useful : the purpose of the learning model is to applicable to new data (testing)

Overfitting



- as we keep training (insisting on ability to classify training set), the performance on the training set (green) becomes unrealistically small
  - model becomes more complex
- but at the same time ability to predict/classify new data (pink) worsens

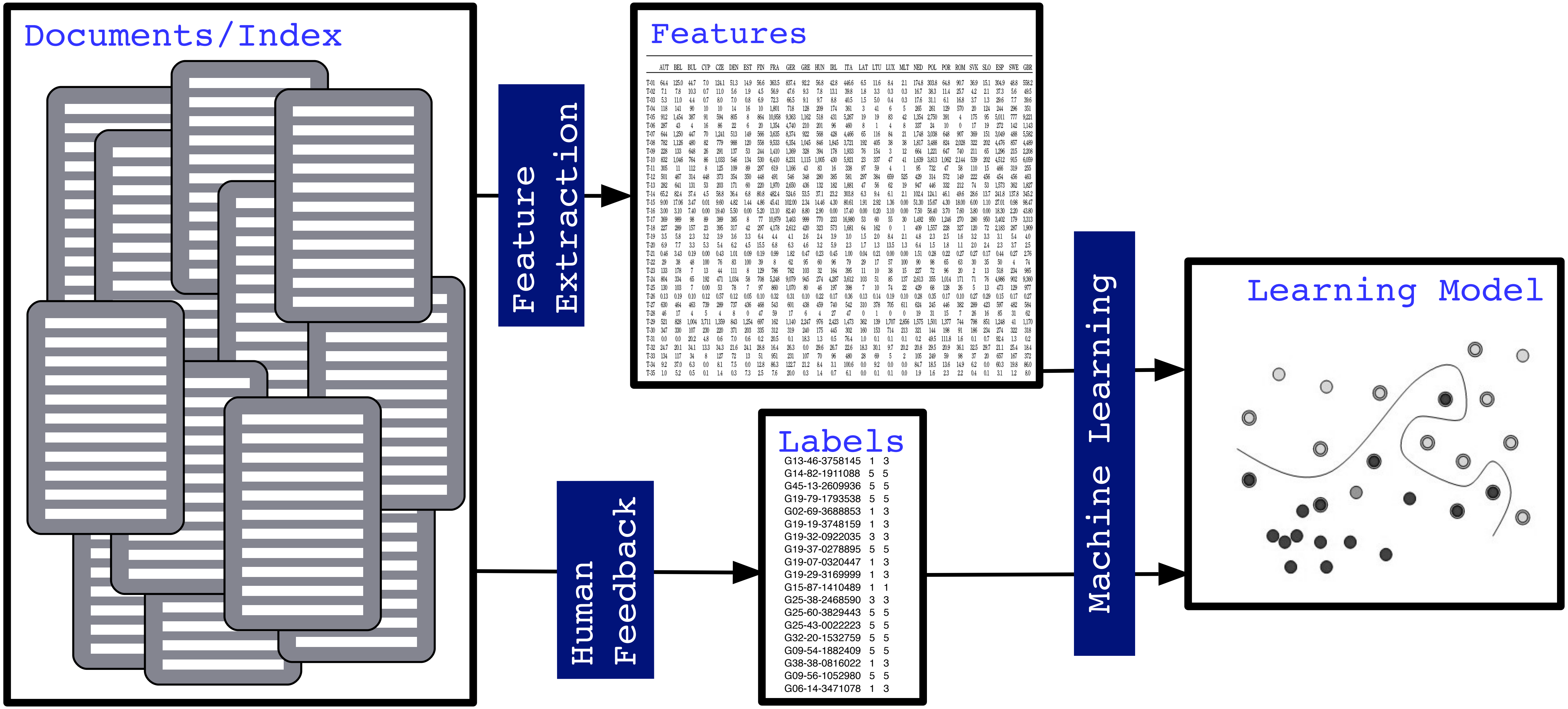
Cross Validation Setup



	AUT	BEL	BUL	CYP	CZE	DEN	EST	FIN	FRA	GER	GRE	HUN	IRL	ITA	LAT	LTU	LUX	MLT	NED	POL	POR	ROM	SVK	SLO	ESP	SWE	GBR
T-01	64.4	125.0	44.7	7.0	124.1	51.3	14.9	56.6	363.5	837.4	92.2	56.8	42.8	446.6	6.5	11.6	8.4	2.1	174.8	303.8	64.8	90.7	36.9	15.1	304.9	48.8	558.2
T-02	7.1	7.8	10.3	0.7	11.0	5.6	1.9	4.5	56.9	47.6	9.3	7.8	13.1	39.8	1.8	3.3	0.3	0.3	16.7	38.3	11.4	25.7	4.2	2.1	37.3	5.6	49.5
T-03	5.3	11.0	4.4	0.7	8.0	7.0	0.8	6.9	72.3	66.5	9.1	9.7	8.8	40.5	1.5	5.0	0.4	0.3	17.6	31.1	6.1	16.8	3.7	1.3	29.6	7.7	39.6
T-04	118	141	90	10	10	14	16	10	1,801	718	128	209	174	361	3	41	6	5	265	261	129	570	20	124	244	296	351
T-05	912	1,454	387	91	594	805	8	864	10,958	9,363	1,162	518	431	5,267	19	19	83	42	1,354	2,750	391	4	175	95	5,011	777	9,221
T-06	287	43	4	16	86	22	6	20	1,354	4,740	210	201	96	460	8	1	4	8	337	24	10	0	17	19	272	142	1,143
T-07	644	1,250	447	70	1,241	513	149	566	3,635	8,374	922	568	428	4,466	65	116	84	21	1,748	3,038	648	907	369	151	3,049	488	5,582
T-08	782	1,126	480	82	779	988	120	558	9,533	6,354	1,045	846	1,845	3,721	192	405	38	38	1,817	3,488	824	2,028	322	202	4,476	857	4,489
T-09	228	133	648	26	291	137	53	244	1,410	1,369	328	394	178	1,933	76	154	3	12	664	1,221	647	740	211	65	1,296	215	2,208
T-10	832	1,046	764	86	1,033	546	134	530	6,410	8,231	1,115	1,005	430	5,921	23	337	47	41	1,639	3,813	1,062	2,144	539	202	4,512	915	6,059
T-11	305	11	112	8	125	109	89	297	619	1,166	43	83	16	338	97	59	4	1	95	732	47	58	110	15	466	319	255
T-12	501	467	314	448	373	354	350	448	491	546	348	280	385	581	297	384	659	525	429	314	572	149	222	456	454	456	463
T-13	282	641	131	53	203	171	60	220	1,970	2,650	436	132	182	1,881	47	56	62	19	947	446	332	212	74	53	1,573	362	1,827
T-14	65.2	82.4	37.4	4.5	58.8	36.4	6.8	80.8	482.4	524.6	53.5	37.1	23.2	303.8	6.3	9.4	6.1	2.1	102.4	124.1	46.1	49.6	28.6	13.7	241.8	137.8	345.2
T-15	9.00	17.06	3.47	0.01	9.60	4.82	1.44	4.86	45.41	102.00	2.34	14.46	4.30	80.61	1.91	2.92	1.36	0.00	51.30	15.67	4.30	18.00	6.00	1.10	27.01	0.98	98.47
T-16	3.00	3.10	7.40	0.00	19.40	5.50	0.00	5.20	13.10	82.40	8.80	2.90	0.00	17.40	0.00	0.20	3.10	0.00	7.50	58.40	3.70	7.60	3.80	0.00	18.30	2.20	43.80
T-17	369	989	98	89	389	385	8	77	10,979	3,463	999	770	233	16,980	53	60	55	30	1,492	950	1,246	270	280	950	3,402	179	3,313
T-18	227	289	157	23	395	317	42	297	4,178	2,612	420	323	573	1,681	64	162	0	1	409	1,557	228	327	120	72	2,183	287	1,909
T-19	3.5	5.8	2.3	3.2	3.9	3.6	3.3	6.4	4.4	4.1	2.6	2.4	3.9	3.0	1.5	2.0	8.4	2.1	4.8	2.3	2.5	1.6	3.2	3.3	3.1	5.4	4.0
T-20	6.9	7.7	3.3	5.3	5.4	6.2	4.5	15.5	6.8	6.3	4.6	3.2	5.9	2.3	1.7	1.3	13.5	1.3	6.4	1.5	1.8	1.1	2.0	2.4	2.3	3.7	2.5
T-21	0.46	3.43	0.19	0.00	0.43	1.01	0.09	0.19	0.99	1.82	0.47	0.23	0.45	1.00	0.04	0.21	0.00	0.00	1.51	0.28	0.22	0.27	0.27	0.17	0.44	0.27	2.76
T-22	29	38	48	100	76	83	100	39	8	62	95	60	96	79	29	17	57	100	90	98	65	63	30	35	50	4	74
T-23	133	178	7	13	44	111	8	129	786	782	103	32	164	395	11	10	38	15	227	72	96	20	2	13	518	234	985
T-24	804	334	65	192	471	1,034	58	708	5,248	9,079	945	274	4,287	3,612	103	51	85	137	2,613	355	1,014	171	71	76	4,986	902	9,360
T-25	130	103	7	0.00	53	78	7	97	860	1,070	80	46	197	398	7	10	74	22	429	68	128	26	5	13	473	129	977
T-26	0.13	0.19	0.10	0.12	0.57	0.12	0.05	0.10	0.32	0.31	0.10	0.22	0.17	0.36	0.13	0.14	0.19	0.10	0.28	0.35	0.17	0.10	0.27	0.29	0.15	0.17	0.27
T-27	630	464	463	739	289	737	436	468	543	601	438	459	740	542	310	378	705	611	624	245	446	382	289	423	597	482	584
T-28	46	17	4	5	4	8	0	47	59	17	6	4	27	47	0	1	0	0	19	31	15	7	26	16	85	31	62
T-29	521	828	1,004	3,711	1,359	843	1,254	697	162	1,140	2,247	976	2,423	1,473	362	139	1,707	2,856	1,575	1,501	1,377	744	798	851	1,248	41	1,170
T-30	347	330	107	230	220	371	203	335	312	319	240	175	445	302	160	153	714	213	321	144	198	91	186	234	274	322	318
T-31	0.0	0.0	20.2	4.8	0.6	7.0	0.6	0.2	20.5	0.1	18.3	1.3	0.5	76.4	1.0	0.1	0.1	0.1	0.2	49.5	111.8	1.6	0.1	0.7	92.4	1.3	0.2
T-32	24.7	20.1	34.1	13.3	34.3	21.6	24.1	28.8	16.4	26.3	0.0	29.6	26.7	22.6	18.3	30.1	9.7	20.2	20.8	29.5	20.9	36.1	32.5	29.7	21.1	25.4	18.4
T-33	134	117	34	8	127	72	13	51	951	231	107	70	96	480	28	69	5	2	105	249	59	98	37	20	657	167	372
T-34	9.2	37.0	6.3	0.0	8.1	7.5	0.0	12.8	86.3	122.7	21.2	8.4	3.1	100.6	0.0	9.2	0.0	0.0	84.7	18.5	13.6	14.9	6.2	0.0	60.3	19.8	86.0
T-35	1.0	5.2	0.5	0.1	1.4	0.3	7.3	2.5	7.6	20.0	0.3	1.4	0.7	6.1	0.0	0.1	0.1	0.0	1.9	1.6	2.3	2.2	0.4	0.1	3.1	1.2	8.0

- split data in K folds
- execute K independent learning trials:
  - train on K-1 folds
  - test on remaining fold
  - measure testing performance
- average results across K trials

Learning / Training with text objects



Features

	AUT	BEL	BUL	CYP	CZE	DEN	EST	FIN	FRA	GER	GRE	HUN	IRL	ITA	LAT	LTU	LUX	MLT	NED	POL	POR	ROM	SVK	SLO	ESP	SWE	GBR
T-01	64.4	125.0	44.7	7.0	124.1	51.3	14.9	56.6	363.5	837.4	92.2	56.8	42.8	446.6	6.5	11.6	8.4	2.1	174.8	303.8	64.8	90.7	36.9	15.1	304.9	48.8	558.2
T-02	7.1	7.8	10.3	0.7	11.0	5.6	1.9	4.5	56.9	47.6	9.3	7.8	13.1	39.8	1.8	3.3	0.3	0.3	16.7	38.3	11.4	25.7	4.2	2.1	37.3	5.6	49.5
T-03	5.3	11.0	4.4	0.7	8.0	7.0	0.8	6.9	72.3	66.5	9.1	9.7	8.8	40.5	1.5	5.0	0.4	0.3	17.6	31.1	6.1	16.8	3.7	1.3	29.6	7.7	39.6
T-04	118	141	90	10	10	14	16	10	1,801	718	128	209	174	361	3	41	6	5	265	261	129	570	20	124	244	296	351
T-05	912	1,454	387	91	594	805	8	864	10,958	9,363	1,162	518	431	5,267	19	19	83	42	1,354	2,750	391	4	175	95	5,011	777	9,221
T-06	287	43	4	16	86	22	6	20	1,354	4,740	210	201	96	460	8	1	4	8	337	24	10	0	17	19	272	142	1,143
T-07	644	1,250	447	70	1,241	513	149	566	3,635	8,374	922	568	428	4,466	65	116	84	21	1,748	3,038	648	907	369	151	3,049	488	5,582
T-08	782	1,126	480	82	779	988	120	558	9,533	6,354	1,045	846	1,845	3,721	192	405	38	38	1,817	3,488	824	2,028	322	202	4,476	857	4,489
T-09	228	133	648	26	291	137	53	244	1,410	1,369	328	394	178	1,933	76	154	3	12	664	1,221	647	740	211	65	1,296	215	2,208
T-10	832	1,046	764	86	1,033	546	134	530	6,410	8,231	1,115	1,005	430	5,921	23	337	47	41	1,639	3,813	1,062	2,144	539	202	4,512	915	6,059
T-11	305	11	112	8	125	109	89	297	619	1,166	43	83	16	338	97	59	4	1	95	732	47	58	110	15	466	319	255
T-12	501	467	314	448	373	354	350	448	491	546	348	280	385	581	297	384	659	525	429	314	572	149	222	456	454	456	463
T-13	282	641	131	53	203	171	60	220	1,970	2,650	436	132	182	1,881	47	56	62	19	947	446	332	212	74	53	1,573	362	1,827
T-14	65.2	82.4	37.4	4.5	58.8	36.4	6.8	80.8	482.4	524.6	53.5	37.1	23.2	303.8	6.3	9.4	6.1	2.1	102.4	124.1	46.1	49.6	28.6	13.7	241.8	137.8	345.2
T-15	9.00	17.06	3.47	0.01	9.60	4.82	1.44	4.86	45.41	102.00	2.34	14.46	4.30	80.61	1.91	2.92	1.36	0.00	51.30	15.67	4.30	18.00	6.00	1.10	27.01	0.98	98.47
T-16	3.00	3.10	7.40	0.00	19.40	5.50	0.00	5.20	13.10	82.40	8.80	2.90	0.00	17.40	0.00	0.20	3.10	0.00	7.50	58.40	3.70	7.60	3.80	0.00	18.30	2.20	43.80
T-17	369	989	98	89	389	385	8	77	10,979	3,463	999	770	233	16,980	53	60	55	30	1,492	950	1,246	270	280	950	3,402	179	3,313
T-18	227	289	157	23	395	317	42	297	4,178	2,612	420	323	573	1,681	64	162	0	1	409	1,557	228	327	120	72	2,183	287	1,909
T-19	3.5	5.8	2.3	3.2	3.9	3.6	3.3	6.4	4.4	4.1	2.6	2.4	3.9	3.0	1.5	2.0	8.4	2.1	4.8	2.3	2.5	1.6	3.2	3.3	3.1	5.4	4.0
T-20	6.9	7.7	3.3	5.3	5.4	6.2	4.5	15.5	6.8	6.3	4.6	3.2	5.9	2.3	1.7	1.3	13.5	1.3	6.4	1.5	1.8	1.1	2.0	2.4	2.3	3.7	2.5
T-21	0.46	3.43	0.19	0.00	0.43	1.01	0.09	0.19	0.99	1.82	0.47	0.23	0.45	1.00	0.04	0.21	0.00	0.00	1.51	0.28	0.22	0.27	0.27	0.17	0.44	0.27	2.76
T-22	29	38	48	100	76	83	100	39	8	62	95	60	96	79	29	17	57	100	90	98	65	63	30	35	50	4	74
T-23	133	178	7	13	44	111	8	129	786	782	103	32	164	395	11	10	38	15	227	72	96	20	2	13	518	234	985
T-24	804	334	65	192	471	1,034	58	708	5,248	9,079	945	274	4,287	3,612	103	51	85	137	2,613	355	1,014	171	71	76	4,986	902	9,360
T-25	130	103	7	0.00	53	78	7	97	860	1,070	80	46	197	398	7	10	74	22	429	68	128	26	5	13	473	129	977
T-26	0.13	0.19	0.10	0.12	0.57	0.12	0.05	0.10	0.32	0.31	0.10	0.22	0.17	0.36	0.13	0.14	0.19	0.10	0.28	0.35	0.17	0.10	0.27	0.29	0.15	0.17	0.27
T-27	630	464	463	739	289	737	436	468	543	601	438	459	740	542	310	378	705	611	624	245	446	382	289	423	597	482	584
T-28	46	17	4	5	4	8	0	47	59	17	6	4	27	47	0	1	0	0	19	31	15	7	26	16	85	31	62
T-29	521	828	1,004	3,711	1,359	843	1,254	697	162	1,140	2,247	976	2,423	1,473	362	139	1,707	2,856	1,575	1,501	1,377	744	798	851	1,248	41	1,170
T-30	347	330	107	230	220	371	203	335	312	319	240	175	445	302	160	153	714	213	321	144	198	91	186	234	274	322	318
T-31	0.0	0.0	20.2	4.8	0.6	7.0	0.6	0.2	20.5	0.1	18.3	1.3	0.5	76.4	1.0	0.1	0.1	0.1	0.2	49.5	111.8	1.6	0.1	0.7	92.4	1.3	0.2
T-32	24.7	20.1	34.1	13.3	34.3	21.6	24.1	28.8	16.4	26.3	0.0	29.6	26.7	22.6	18.3	30.1	9.7	20.2	20.8	29.5	20.9	36.1	32.5	29.7	21.1	25.4	18.4
T-33	134	117	34	8	127	72	13	51	951	231	107	70	96	480	28	69	5	2	105	249	59	98	37	20	657	167	372
T-34	9.2	37.0	6.3	0.0	8.1	7.5	0.0	12.8	86.3	122.7	21.2	8.4	3.1	100.6	0.0	9.2	0.0	0.0	84.7	18.5	13.6	14.9	6.2	0.0	60.3	19.8	86.0
T-35	1.0	5.2	0.5	0.1	1.4	0.3	7.3	2.5	7.6	20.0	0.3	1.4	0.7	6.1	0.0	0.1	0.1	0.0	1.9	1.6	2.3	2.2	0.4	0.1	3.1	1.2	8.0

Labels

G13-46-3758145	1	3
G14-82-1911088	5	5
G45-13-2609936	5	5
G19-79-1793538	5	5
G02-69-3688853	1	3
G19-19-3748159	1	3
G19-32-0922035	3	3
G19-37-0278895	5	5
G19-07-0320447	1	3
G19-29-3169999	1	3
G15-87-1410489	1	1
G25-38-2468590	3	3
G25-60-3829443	5	5
G25-43-0022223	5	5
G32-20-1532759	5	5
G09-54-1882409	5	5
G38-38-0816022	1	3
G09-56-1052980	5	5
G06-14-3471078	1	3

- for objects like text documents or images:
  - extract features (to obtain matrix form)
  - annotate (to obtain labels)

Spambase dataset

- about 4000 emails
- 54 features numerical
- two classes: spam / no_spam

Housing dataset

- 1300 houses
- 13 features (numerical)
- label : purchase prices (quantitative)

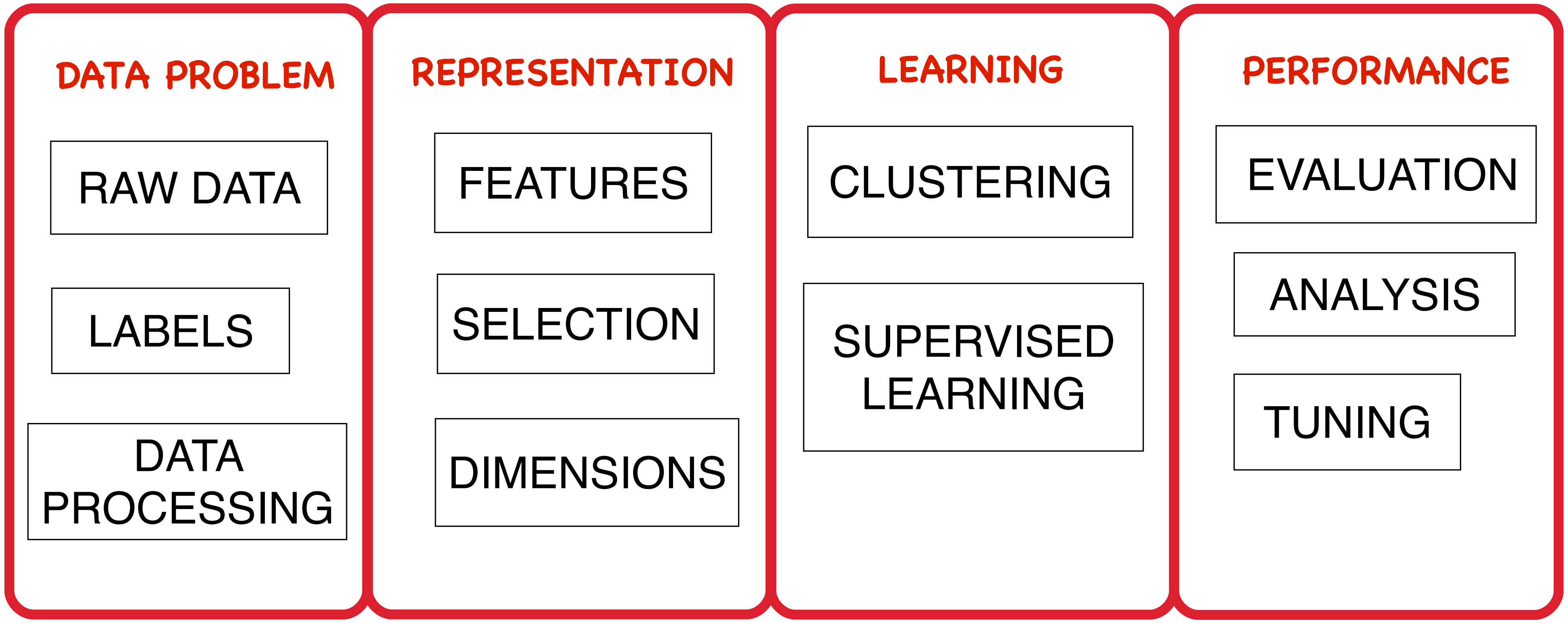
Digits dataset

- 60000 images of scanned digits
- 26x26 pixel per image, black or white
- features not extracted
- 10 classes : 0,1,2, ..., 9

Documents dataset

- 20,000 news articles (text)
- features not extracted
- 20 categories: religion, music, computers, sports, etc.

course map



- main focus: learning algorithms
- main focus: hands-on practice on datasets
- secondary focus: analysis, error measurement
- secondary focus: features, representation

typical module subtaks / objectives

- THEORY
  - explain/understand fundamental mechanism
  - proof (math, intuition)
  - pseudocode
- CODE
  - run existing code
  - implement and demo your code
  - data handling: features, dimensionality, scale, missing values, normalization
  - computational issues : memory, cache, CPU, disk
- EVALUATION
  - setup
  - performance measurement, comparison
  - analysis/failure of procedure behavior
- HOWTO
  - practical advise, hacks, heuristics
  - communicate on topic well : email, forums
  - where to look online