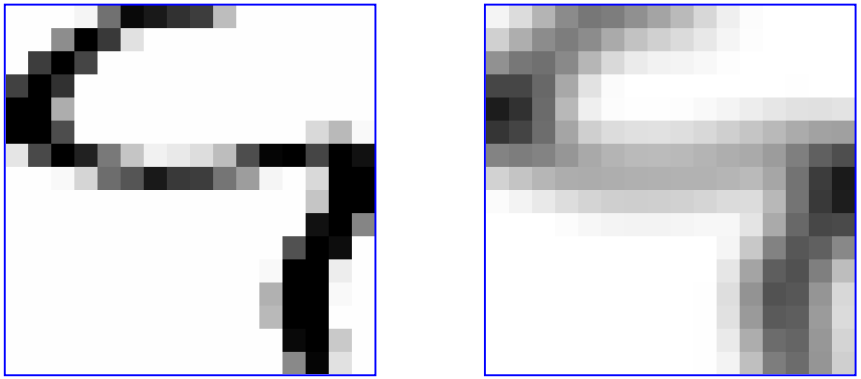
From Paper to Digits



Lektor Dr.techn. Alexander K. Seewald
Österreichisches Forschungsinstitut
für Artificial Intelligence

alex@seewald.at / alex.seewald.at

From Paper to Digits

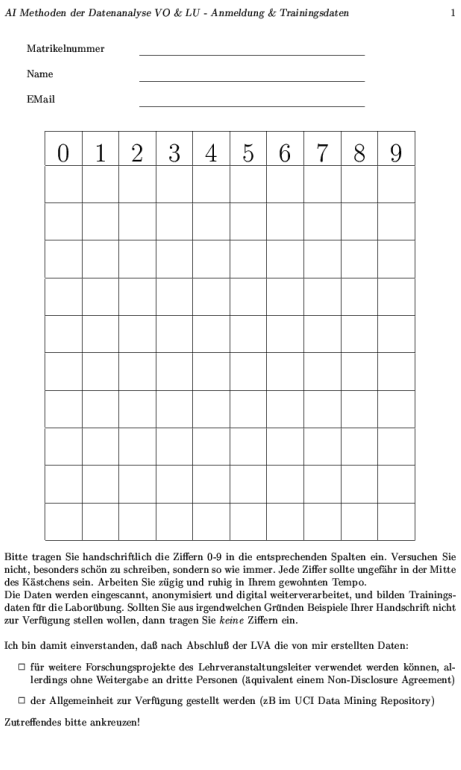
Here, we will describe the steps which were necessary to preprocess the contributed handwritten digits into a useful representation for data analysis.

- Students write digits on paper
- Pages are scanned via auto-feed scanner
- Digital images of pages are segmented to find first the horizontal and vertical lines of the table, and then the digits of varying size which are contained within each box
- Digits are extracted and resized to 16x16 pixels

Afterwards we will discuss alternative representations, feature construction and selection.

alex@seewald.at / alex.seewald.at

Students write Digits



AI Methoden der Datenanalyse VO & LU - Anmeldung & Trainingsdaten 1

Matrikelnummer _____

Name _____

EMail _____

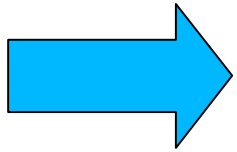
0	1	2	3	4	5	6	7	8	9

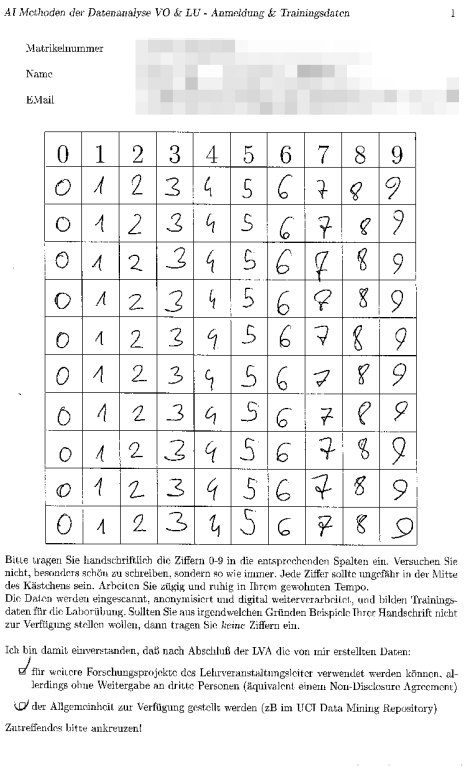
Bitte tragen Sie handschriftlich die Ziffern 0-9 in die entsprechenden Spalten ein. Versuchen Sie nicht, besonders schön zu schreiben, sondern so wie immer. Jede Ziffer sollte ungefähr in der Mitte des Kästchens sein. Arbeiten Sie zügig und ruhig in Ihrem gewohnten Tempo.
Die Daten werden eingescannt, anonymisiert und digital weiterverarbeitet, und bilden Trainingsdaten für die Laborübung. Sollten Sie aus irgendwelchen Gründen Beispiele Ihrer Handschrift nicht zur Verfügung stellen wollen, dann tragen Sie *keine* Ziffern ein.

Ich bin damit einverstanden, daß nach Abschluß der LVA die von mir erstellten Daten:

- ☐ für weitere Forschungsprojekte des Lehrveranstaltungsleiter verwendet werden können, allerdings ohne Weitergabe an dritte Personen (äquivalent einem Non-Disclosure Agreement)
- ☐ der Allgemeinheit zur Verfügung gestellt werden (zB im UCI Data Mining Repository)

Zutreffendes bitte ankreuzen!





AI Methoden der Datenanalyse VO & LU - Anmeldung & Trainingsdaten 1

Matrikelnummer [redacted]

Name

EMail

0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	8	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9

Bitte tragen Sie handschriftlich die Ziffern 0-9 in die entsprechenden Spalten ein. Versuchen Sie nicht, besonders schön zu schreiben, sondern so wie immer. Jede Ziffer sollte ungefähr in der Mitte des Kästchens sein. Arbeiten Sie zügig und ruhig in Ihrem gewohnten Tempo.
Die Daten werden eingescannt, anonymisiert und digital weiterverarbeitet, und bilden Trainingsdaten für die Laborübung. Sollten Sie aus irgendwelchen Gründen Beispiele Ihrer Handschrift nicht zur Verfügung stellen wollen, dann tragen Sie *keine* Ziffern ein.

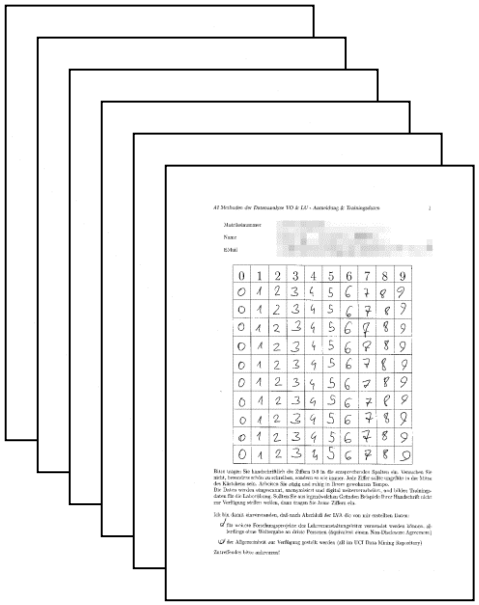
Ich bin damit einverstanden, daß nach Abschluß der LVA die von mir erstellten Daten:

- ☑ für weitere Forschungsprojekte des Lehrveranstaltungsleiter verwendet werden können, allerdings ohne Weitergabe an dritte Personen (äquivalent einem Non-Disclosure Agreement)
- ☑ der Allgemeinheit zur Verfügung gestellt werden (zB im UCI Data Mining Repository)

Zutreffendes bitte ankreuzen!

alex@seewald.at / alex.seewald.at

Scan Pages

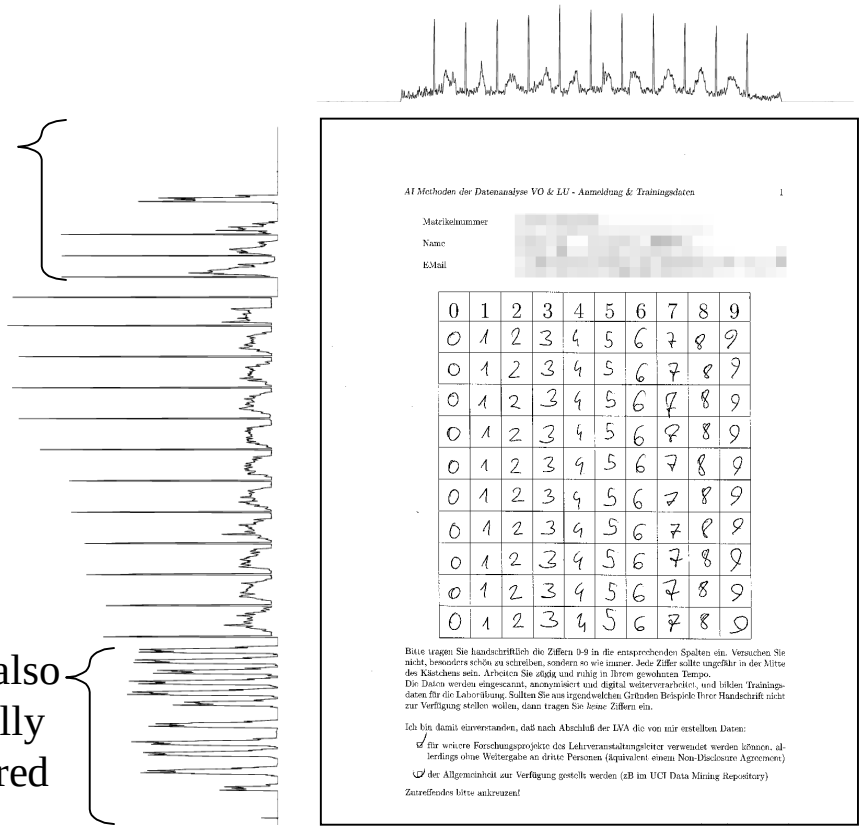


```
Scan1.bmp
Scan2.bmp
Scan3.bmp
...
Scan49.bmp
Scan50.bmp
Scan51.bmp
```


alex@seewald.at / alex.seewald.at

Digit Segmentation

Counting black pixels per row and per colum yields the horizontal and vertical histogram, which give a good first approximation to vertical/horizontal line position.



Input lines for MNr, Name and StKZ need to be ignored as well.

Lines of text are also aligned horizontally and must be ignored

AI Methoden der Datenanalyse VO & LU - Anmeldung & Trainingsdaten 1

Matrikelnummer [redacted]

Name [redacted]

EMail [redacted]

0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9

Bitte tragen Sie handschriftlich die Ziffern 0-9 in die entsprechenden Spalten ein. Versuchen Sie nicht, besonders schön zu schreiben, sondern so wie immer. Jede Ziffer sollte ungefähr in der Mitte des Kästchens sein. Arbeiten Sie zügig und ruhig in Ihrem gewohnten Tempo.
Die Daten werden eingescannt, anonymisiert und digital weiterverarbeitet, und bilden Trainingsdaten für die Laborübung. Sollten Sie aus irgendwelchen Gründen Beispiele Ihrer Handschrift nicht zur Verfügung stellen wollen, dann tragen Sie *keine* Ziffern ein.

Ich bin damit einverstanden, daß nach Abschluß der LVA die von mir erstellten Daten:

- ☑ für weitere Forschungsprojekte des Lehrveranstaltungsleiter verwendet werden können, allerdings ohne Weitergabe an dritte Personen (äquivalent einem Non-Disclosure Agreement)
- ☑ der Allgemeinheit zur Verfügung gestellt werden (zB im UCI Data Mining Repository)

Zutreffendes bitte ankreuzen!

Algorithm

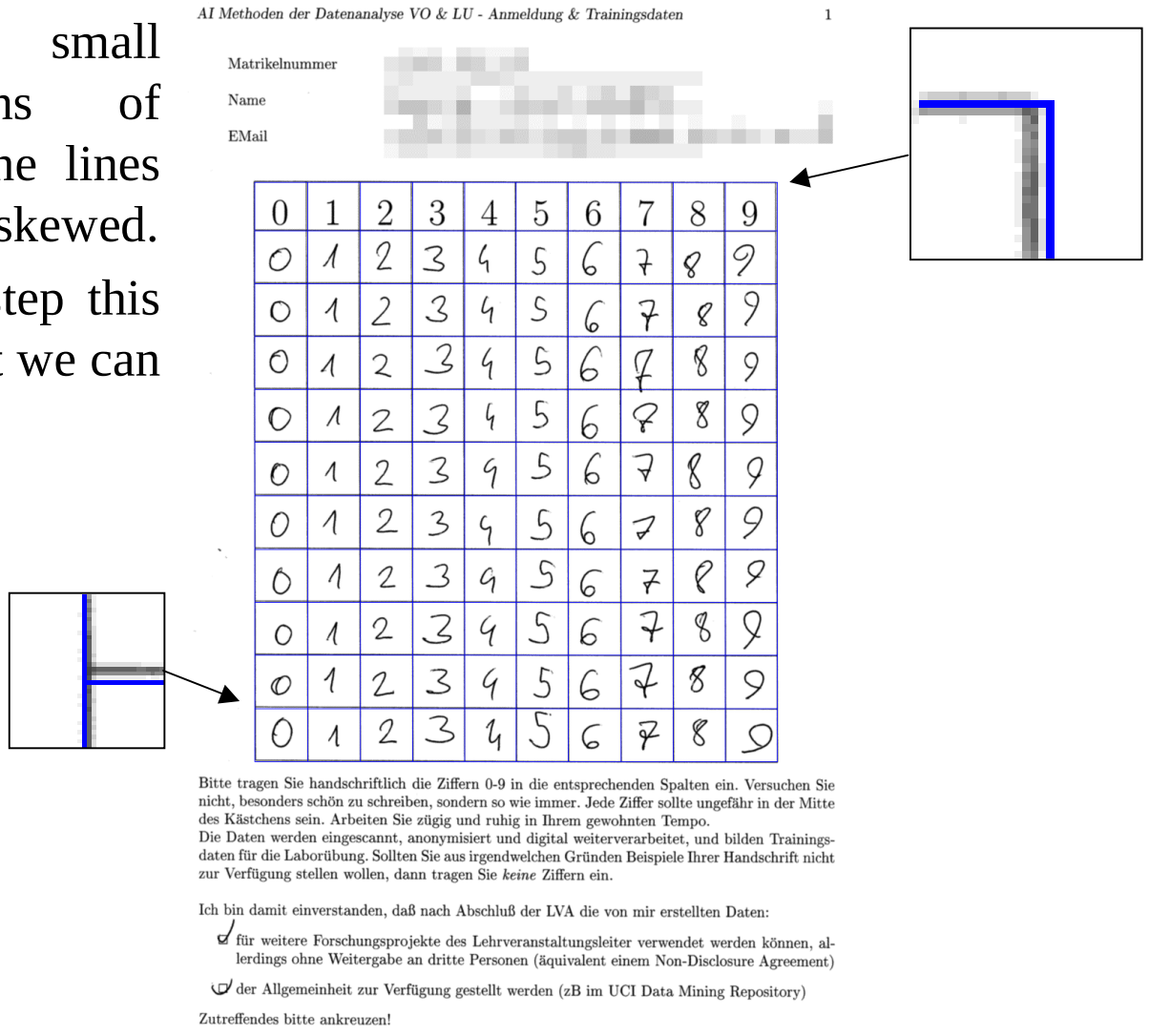
Take the 11/12 highest maxima from horizontal/vertical histogram. Ignore those maxima which appear *near* other maxima (e.g. within half a column width or row height)

alex@seewald.at / alex.seewald.at

However, lines are not perfectly aligned

Due to small imperfections of scanning, the lines are slightly skewed.

As a first step this suffices, but we can do better...



AI Methoden der Datenanalyse VO & LU - Anmeldung & Trainingsdaten 1

Matrikelnummer

Name

EMail

0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	8	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9

Bitte tragen Sie handschriftlich die Ziffern 0-9 in die entsprechenden Spalten ein. Versuchen Sie nicht, besonders schön zu schreiben, sondern so wie immer. Jede Ziffer sollte ungefähr in der Mitte des Kästchens sein. Arbeiten Sie zügig und ruhig in Ihrem gewohnten Tempo.
Die Daten werden eingescannt, anonymisiert und digital weiterverarbeitet, und bilden Trainingsdaten für die Laborübung. Sollten Sie aus irgendwelchen Gründen Beispiele Ihrer Handschrift nicht zur Verfügung stellen wollen, dann tragen Sie *keine* Ziffern ein.

Ich bin damit einverstanden, daß nach Abschluß der LVA die von mir erstellten Daten:

☑ für weitere Forschungsprojekte des Lehrveranstaltungsleiter verwendet werden können, allerdings ohne Weitergabe an dritte Personen (äquivalent einem Non-Disclosure Agreement)

☐ der Allgemeinheit zur Verfügung gestellt werden (zB im UCI Data Mining Repository)

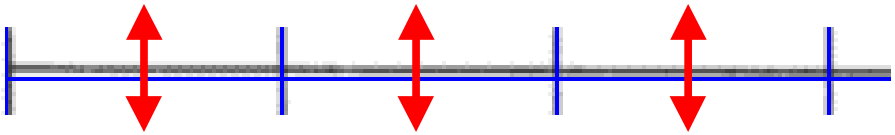
Zutreffendes bitte ankreuzen!

alex@seewald.at / alex.seewald.at

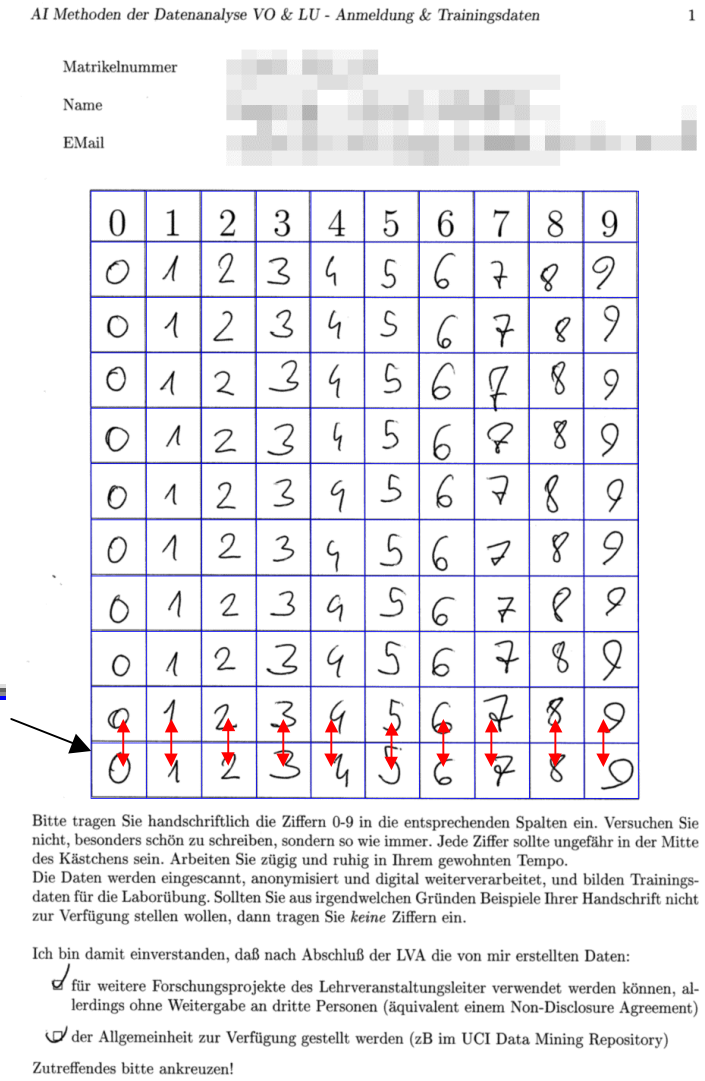
Local search for "true" lines

Algorithm

- Along each approximate line (at the midpoints between crossings), search for the "true" line.
- A "true" line is the nearest continuous sequence of at least 4 pixels length. This gives a set of points of the "true" line.
- For horizontal lines, search in vertical directions, and vice versa for vertical lines.



- Compute the true line from these points via linear regression in 2D.
- For vertical lines, switch coordinate system from x/y to y/x. A vertical line has a slope of infinity in x/y which is problematic.



AI Methoden der Datenanalyse VO & LU - Anmeldung & Trainingsdaten 1

Matrikelnummer [redacted]
Name [redacted]
EMail [redacted]

0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9

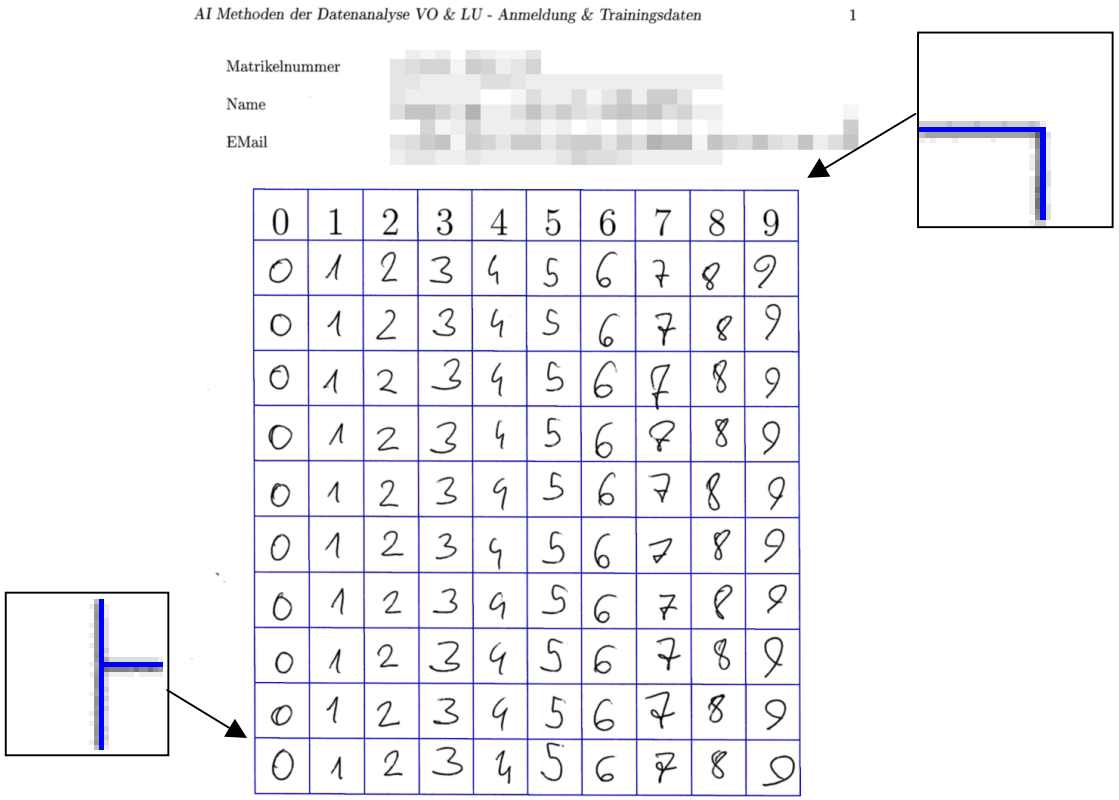
Bitte tragen Sie handschriftlich die Ziffern 0-9 in die entsprechenden Spalten ein. Versuchen Sie nicht, besonders schön zu schreiben, sondern so wie immer. Jede Ziffer sollte ungefähr in der Mitte des Kästchens sein. Arbeiten Sie zügig und ruhig in Ihrem gewohnten Tempo.
Die Daten werden eingescannt, anonymisiert und digital weiterverarbeitet, und bilden Trainingsdaten für die Laborübung. Sollten Sie aus irgendwelchen Gründen Beispiele Ihrer Handschrift nicht zur Verfügung stellen wollen, dann tragen Sie *keine* Ziffern ein.

Ich bin damit einverstanden, daß nach Abschluß der LVA die von mir erstellten Daten:

☑ für weitere Forschungsprojekte des Lehrveranstaltungsleiter verwendet werden können, allerdings ohne Weitergabe an dritte Personen (äquivalent einem Non-Disclosure Agreement)

☑ der Allgemeinheit zur Verfügung gestellt werden (zB im UCI Data Mining Repository)

Zutreffendes bitte ankreuzen!



Afterwards, lines are perfectly aligned

AI Methoden der Datenanalyse VO & LU - Anmeldung & Trainingsdaten 1

Matrikelnummer [redacted]
Name [redacted]
EMail [redacted]

0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9

Bitte tragen Sie handschriftlich die Ziffern 0-9 in die entsprechenden Spalten ein. Versuchen Sie nicht, besonders schön zu schreiben, sondern so wie immer. Jede Ziffer sollte ungefähr in der Mitte des Kästchens sein. Arbeiten Sie zügig und ruhig in Ihrem gewohnten Tempo.
Die Daten werden eingescannt, anonymisiert und digital weiterverarbeitet, und bilden Trainingsdaten für die Laborübung. Sollten Sie aus irgendwelchen Gründen Beispiele Ihrer Handschrift nicht zur Verfügung stellen wollen, dann tragen Sie *keine* Ziffern ein.

Ich bin damit einverstanden, daß nach Abschluß der LVA die von mir erstellten Daten:

☑ für weitere Forschungsprojekte des Lehrveranstaltungsleiter verwendet werden können, allerdings ohne Weitergabe an dritte Personen (äquivalent einem Non-Disclosure Agreement)

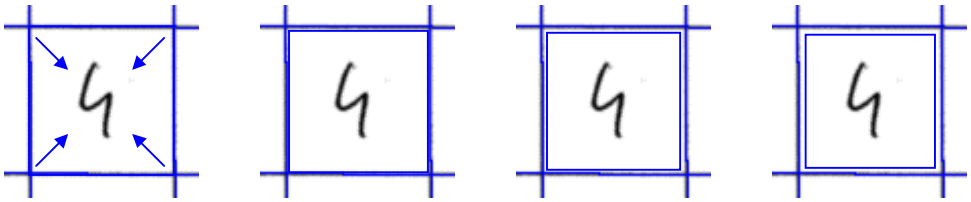
☑ der Allgemeinheit zur Verfügung gestellt werden (zB im UCI Data Mining Repository)

Zutreffendes bitte ankreuzen!

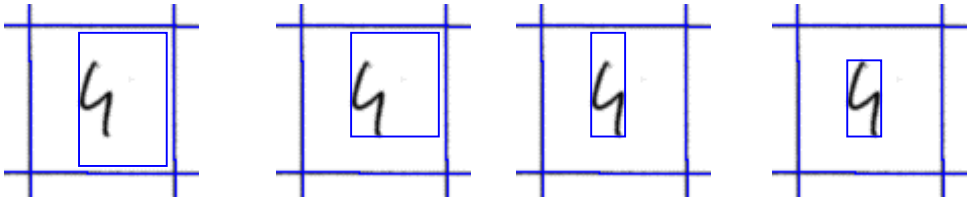
Digit Segmentation (contd.)

Now we know where each element of the input table is. Two steps remain to finish the digit segmentation process and extract each digit.

- Finding the inner, white area by decreasing the size of the rectangle around each digit until the number of black pixels along the border is sufficiently small.



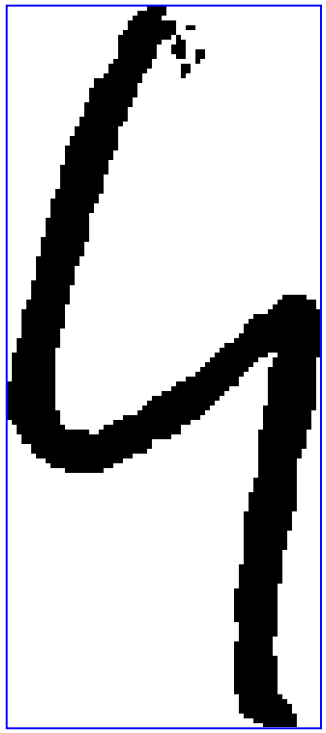
- Finding the digit within the white area by decreasing the size of the rectangle until the number of black pixels along the border is sufficiently high. The size is decreased for each direction separately. A small number of black pixels is ignored to remove speck noise automatically.

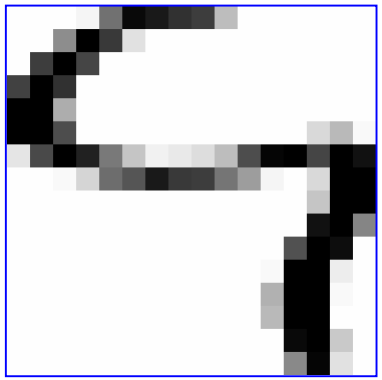


alex@seewald.at / alex.seewald.at

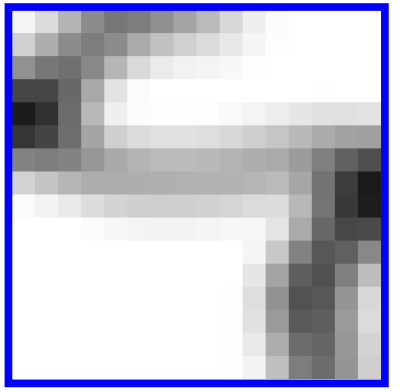
Digits Extraction and Resizing

Extract digit from box, resize via Mitchell filter to 16x16.



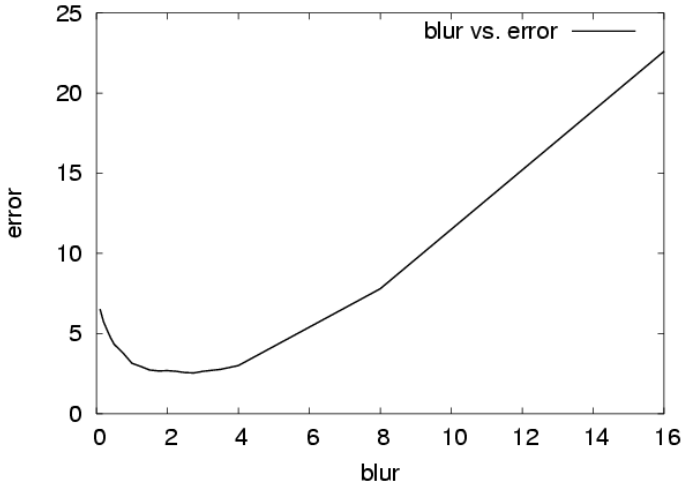


blur=0.5, err=4.33%



blur=2.5, err=2.57%

Blurring parameter influences error rate for some learners, e.g. IBk and SMO.



alex@seewald.at / alex.seewald.at

Representation Issues

Representing each digit as a 16x16 (256 values) feature vector is very simple, but has its problems...

- Topological (neighborhood) relations between pixels are ignored. I.e. a random shuffling of the attributes would have no effect on the performance of our learning systems, while human performance would drop instantly to near zero accuracy.
- The proportion of each digit is lost - each digit is scaled to a quadratic 16x16 grid, ignoring its original width and height.
- The original digit data is black-and-width and of higher resolution that 16x16 pixels. Features based on the original data might be more accurate for digit recognition.

alex@seewald.at / alex.seewald.at

Alternative Representations

[Trier, 1995]

- Projection Histograms (i.e. horizontal/vertical histograms)
- Geometric Moments, Zernike Moments
- Unitary transformations, e.g. Karhunen-Loeve transform, Fourier transform and Cosine transform.
- Zoning (i.e. downscaling to e.g. 16x16 or 5x5 pixels)
- Contour features, Character skeletons

[Liu, 2003]

- Gradient directional features (Sobel, Kirsh operator)
- Chaincode features (derived from outline)
- Concavity features

alex@seewald.at / alex.seewald.at

Zernike Moments

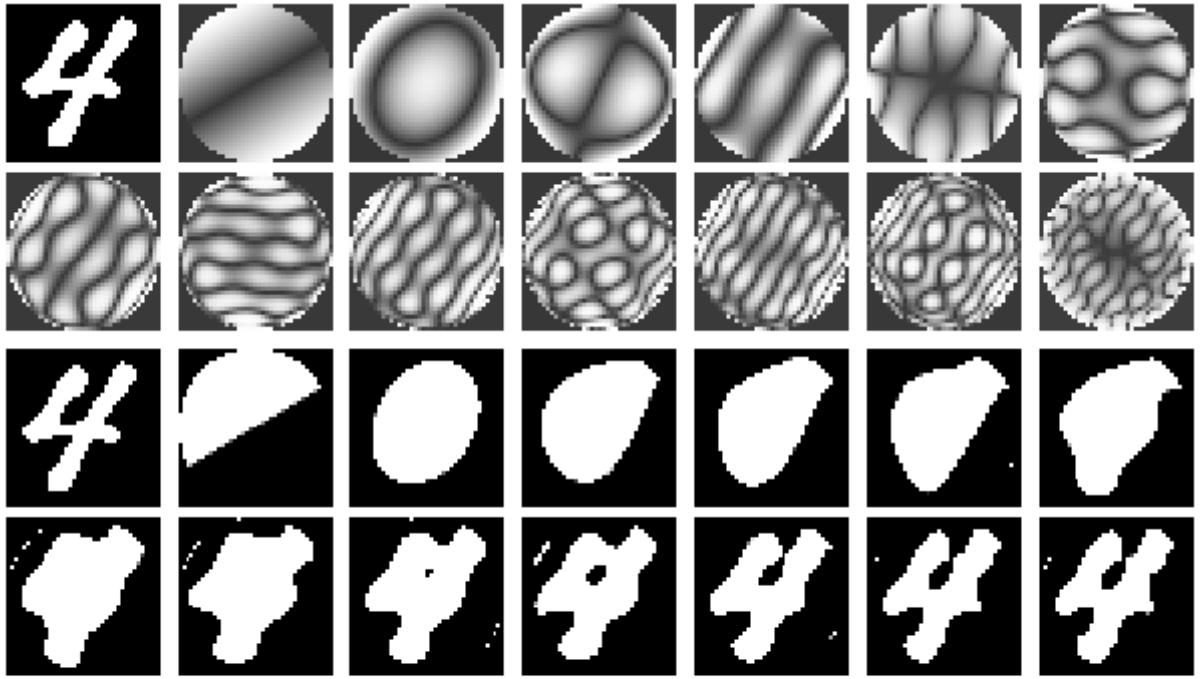


Figure 9: Images derived from Zernike moments. **Rows 1–2:** Input image of digit '4', and contributions from the Zernike moments of order 1–13. The images are histogram equalized to highlight the details. **Rows 3–4:** Input image of digit '4', and images reconstructed from the Zernike moments of order up to 1–13, respectively.

Approximation of digit as sum of spatial base patterns. Digit has to be scaled within an unit circle [Trier, 1995]

alex@seewald.at / alex.seewald.at

Feature Construction

Feature Construction **describes the process of creating useful features for a given dataset, e.g.**

- handwritten digit recognition
- customer churn prediction in telecommunications (i.e. which customer will switch to a competitor next month)
- who is interested in buying caravan insurance?
- Cutomer-Relationship Management (CRM): finding profitable customers, characterizing potential customers, marketing to groups with similar behaviour
- AI for peace (i.e. models to prevent conflicts)
- spam filtering

Needs specific knowledge of the dataset, the task to be done and its context.

alex@seewald.at / alex.seewald.at

Feature Selection

Feature Construction **may yield a large number of features.** ***Feature Selection*****, i.e. reducing the number of features, can improve classification accuracy as well as speeding up the learning process. It is also essential to achieve simpler, more comprehensible models.**

- Feature selection is well supported in WEKA under weka.attributeSelection (e.g. CfsSubsetEval, ChiSquaredAttributeEval and ReliefFAttributeEval).
- Feature construction can in simple cases be done via weka.filters.unsupervised.attribute.AddExpression
- More complex feature construction can be done in Java, or in external programs which output the ARFF file format.

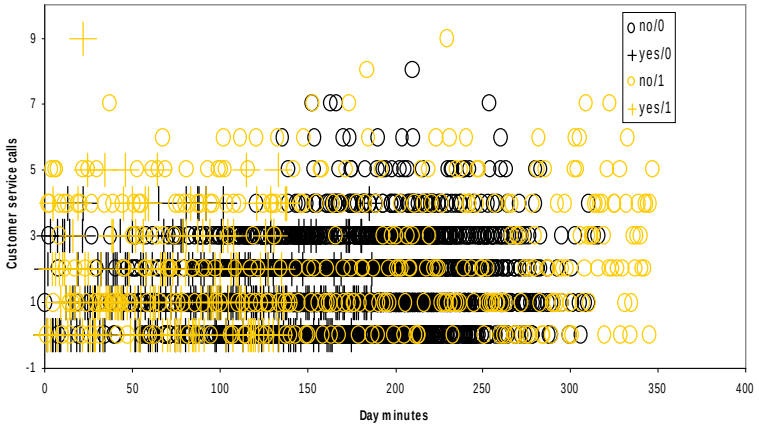
alex@seewald.at / alex.seewald.at

Dataset: Customer Churn

St.	Acc. Len	Area	Int. Plan	Voice Mail	VMail mins.	Day mins.	Day calls	Day charg	Eve mins.	Eve calls	Eve char	Night mins.	Night calls	Night charge	Intl. mins.	Intl. Calls	Intl. Charge	Serv, Calls
KS	128	415	no	yes	25	265,1	110	45,07	197,4	99	16,8	244,7	91	11,01	10	3	2,7	1
OH	107	415	no	yes	26	161,6	123	27,47	195,5	103	16,6	254,4	103	11,45	13,7	3	3,7	1
NJ	137	415	no	no	0	243,4	114	41,38	121,2	110	10,3	162,6	104	7,32	12,2	5	3,29	0
OH	84	408	yes	no	0	299,4	71	50,9	61,9	88	5,26	196,9	89	8,86	6,6	7	1,78	2
OK	75	415	yes	no	0	166,7	113	28,34	148,3	122	12,6	186,9	121	8,41	10,1	3	2,73	3
AL	118	510	yes	no	0	223,4	98	37,98	220,6	101	18,8	203,9	118	9,18	6,3	6	1,7	0
MA	121	510	no	yes	24	218,2	88	37,09	348,5	108	29,6	212,6	118	9,57	7,5	7	2,03	3
MO	147	415	yes	no	0	157	79	26,69	103,1	94	8,76	211,8	96	9,53	7,1	6	1,92	0
LA	117	408	no	no	0	184,5	97	31,37	351,6	80	29,9	215,8	90	9,71	8,7	4	2,35	1
WV	141	415	yes	yes	37	258,6	84	43,96	222	111	18,9	326,4	97	14,69	11,2	5	3,02	0
IN	65	415	no	no	0	129,1	137	21,95	228,5	83	19,4	208,8	111	9,4	12,7	6	3,43	4
RI	74	415	no	no	0	187,7	127	31,91	163,4	148	13,9	196	94	8,82	9,1	5	2,46	0
IA	168	408	no	no	0	128,8	96	21,9	104,9	71	8,92	141,1	128	6,35	11,2	2	3,02	1
MT	95	510	no	no	0	156,6	88	26,62	247,6	75	21,1	192,3	115	8,65	12,3	5	3,32	3
IA	62	415	no	no	0	120,7	70	20,52	307,2	76	26,1	203	99	9,14	13,1	6	3,54	4
NY	161	415	no	no	0	332,9	67	56,59	317,8	97	27	160,6	128	7,23	5,4	9	1,46	4
ID	85	408	no	yes	27	196,4	139	33,39	280,9	90	23,9	89,3	75	4,02	13,8	4	3,73	1
VT	93	510	no	no	0	190,7	114	32,42	218,2	111	18,6	129,6	121	5,83	8,1	3	2,19	3
VA	76	510	no	yes	33	189,7	66	32,25	212,8	65	18,1	165,7	108	7,46	10	5	2,7	1
TX	73	415	no	no	0	224,4	90	38,15	159,5	88	13,6	192,8	74	8,68	13	2	3,51	1
FL	147	415	no	no	0	155,1	117	26,37	239,7	93	20,4	208,8	133	9,4	10,6	4	2,86	0
CO	77	408	no	no	0	62,4	89	10,61	169,9	121	14,4	209,6	64	9,43	5,7	6	1,54	5
AZ	130	415	no	no	0	183	112	31,11	72,9	99	6,2	181,8	78	8,18	9,5	19	2,57	0
SC	111	415	no	no	0	110,4	103	18,77	137,3	102	11,7	189,6	105	8,53	7,7	6	2,08	2
VA	132	510	no	no	0	81,1	86	13,79	245,2	72	20,8	237	115	10,67	10,3	2	2,78	0
NE	174	415	no	no	0	124,3	76	21,13	277,1	112	23,6	250,7	115	11,28	15,5	5	4,19	3
WY	57	408	no	yes	39	213	115	36,21	191,1	112	16,2	182,7	115	8,22	9,5	3	2,57	0
MT	54	408	no	no	0	134,3	73	22,83	155,5	100	13,2	102,1	68	4,59	14,7	4	3,97	3
MO	20	415	no	no	0	190	109	32,3	258,2	84	22	181,5	102	8,17	6,3	6	1,7	0
HI	49	510	no	no	0	119,3	117	20,28	215,1	109	18,3	178,7	90	8,04	11,1	1	3	1
IL	142	415	no	no	0	84,8	95	14,42	136,7	63	11,6	250,5	148	11,27	14,2	6	3,83	2
NH	75	510	no	no	0	226,1	105	38,44	201,5	107	17,1	246,2	98	11,08	10,3	5	2,78	1
LA	172	408	no	no	0	212	121	36,04	31,2	115	2,65	293,3	78	13,2	12,6	10	3,4	3
AZ	12	408	no	no	0	249,6	118	42,43	252,4	119	21,5	280,2	90	12,61	11,8	3	3,19	1
OK	57	408	no	yes	25	176,8	94	30,06	195	75	16,6	213,5	116	9,61	8,3	4	2,24	0
GA	72	415	no	yes	37	220	80	37,4	217,3	102	18,5	152,8	71	6,88	14,7	6	3,97	3
AK	36	408	no	yes	30	146,3	128	24,87	162,5	80	13,8	129,3	109	5,82	14,5	6	3,92	0
MA	78	415	no	no	0	130,8	64	22,24	223,7	116	19	227,8	108	10,25	10	5	2,7	1
AK	136	415	yes	yes	33	203,9	106	34,66	187,6	99	16	101,7	107	4,58	10,5	6	2,84	3
NJ	149	408	no	no	0	140,4	94	23,87	271,8	92	23,1	188,3	108	8,47	11,1	9	3	1
GA	98	408	no	no	0	126,3	102	21,47	166,8	85	14,2	187,8	135	8,45	9,4	2	2,54	3
MD	135	408	yes	yes	41	173,1	85	29,43	203,9	107	17,3	122,2	78	5,5	14,6	15	3,94	0
AR	34	510	no	no	0	124,8	82	21,22	282,2	98	24	311,5	78	14,02	10	4	2,7	2
ID	160	415	no	no	0	85,8	77	14,59	165,3	110	14,1	178,5	92	8,03	9,2	4	2,48	3
WI	64	510	no	no	0	154	67	26,18	225,8	118	19,2	265,3	86	11,94	3,5	3	0,95	1
OR	59	408	no	yes	28	120,9	97	20,55	213	92	18,1	163,1	116	7,34	8,5	5	2,3	2
MI	65	415	no	no	0	211,3	120	35,92	162,6	122	13,8	134,7	118	6,06	13,2	5	3,56	3
DE	142	408	no	no	0	187	133	31,79	134,6	74	11,4	242,2	127	10,9	7,4	5	2	2
ID	119	415	no	no	0	159,1	114	27,05	231,3	117	19,7	143,2	91	6,44	8,8	3	2,38	5
WY	97	415	no	yes	24	133,2	135	22,64	217,2	58	18,5	70,6	79	3,18	11	3	2,97	1
IA	52	408	no	no	0	191,9	108	32,62	269,8	96	22,9	236,8	87	10,66	7,8	5	2,11	3
IN	60	408	no	no	0	220,6	57	37,5	211,1	115	17,9	249	129	11,21	6,8	3	1,84	1
VA	10	408	no	no	0	186,1	112	31,64	190,2	66	16,2	282,8	57	12,73	11,4	6	3,08	2
UT	96	415	no	no	0	160,2	117	27,23	267,5	67	22,7	228,5	68	10,28	9,3	5	2,51	2
WY	87	415	no	no	0	151	83	25,67	219,7	116	18,7	203,9	127	9,18	9,7	3	2,62	5
IN	81	408	no	no	0	175,5	67	29,84	249,3	85	21,2	270,2	98	12,16	10,2	3	2,75	1
CO	141	415	no	no	0	126,9	98	21,57	180	62	15,3	140,8	128	6,34	8	2	2,16	1
CO	121	408	no	yes	30	198,4	129	33,73	75,3	77	6,4	181,2	77	8,15	5,8	3	1,57	3
WI	68	415	no	no	0	148,8	70	25,3	246,5	164	21	129,8	103	5,84	12,1	3	3,27	3
OK	125	408	no	no	0	229,3	103	38,98	177,4	126	15,1	189,3	95	8,52	12	8	3,24	1
ID	174	408	no	no	0	192,1	97	32,66	169,9	94	14,4	166,6	54	7,5	11,4	4	3,08	1
CA	116	415	no	yes	34	268,6	83	45,66	178,2	142	15,2	166,3	106	7,48	11,6	3	3,13	2
MN	74	510	no	yes	33	193,7	91	32,93	246,1	96	20,9	138	92	6,21	14,6	3	3,94	2
SD	149	408	no	yes	28	180,7	92	30,72	187,8	64	16	265,5	53	11,95	12,6	3	3,4	3
NC	38	408	no	no	0	131,2	98	22,3	162,9	97	13,9	159	106	7,15	8,2	6	2,21	2
WA	40	415	no	yes	41	148,1	74	25,18	169,5	88	14,4	214,1	102	9,63	6,2	5	1,67	2
WY	43	415	yes	no	0	251,5	105	42,76	212,8	104	18,1	157,8	67	7,1	9,3	4	2,51	0
MN	113	408	yes	no	0	125,2	93	21,28	206,4	119	17,5	129,3	139	5,82	8,3	8	2,24	0
UT	126	408	no	no	0	211,6	70	35,97	216,9	80	18,4	153,5	60	6,91	7,8	1	2,11	1
TX	150	510	no	no	0	178,9	101	30,41	169,1	110	14,4	148,6	100	6,69	13,8	3	3,73	4
NJ	138	408	no	no	0	241,8	93	41,11	170,5	83	14,5	295,3	104	13,29	11,8	7	3,19	3
MN	162	510	no	yes	46	224,9	97	38,23	188,2	84	16	254,6	61	11,46	12,1	2	3,27	0
NM	147	510	no	no	0	248,6	83	42,26	148,9	85	12,7	172,5	109	7,76	8	4	2,16	3

Given: A set of customers with state, area code, telephone number, and time/cost information for calls in one month; plus churn = have they switched to another provider by the end of the month?

We will leave digit recognition behind and focus on this dataset for the next few weeks.



alex@seewald.at / alex.seewald.at