Distances for Sequences and Text

- symbol distance: $\rho(x,y) = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{otherwise} \end{cases}$

- Hamming distance:

number of elements that are different

$$H((x^{(1)},\ldots,x^{(p)}),(y^{(1)},\ldots,y^{(p)})) = \sum_{i=1}^{p} \rho(x^{(i)},y^{(i)})$$

- edit/Levenshtein distance: append/remove/edit operations to have the 2 vectors equal

$$L((x^{(1)},\ldots,x^{(p)}),(y^{(1)},\ldots,y^{(q)})) =$$

$$\begin{cases} p & q = 0 \\ q & p = 0 \\ \min\{L((x^{(1)},\ldots,x^{(p-1)}),(y^{(1)},\ldots,y^{(q)})) + 1, \text{ (REMOVE 1st sequence)} & \\ \quad L((x^{(1)},\ldots,x^{(p)}),(y^{(1)},\ldots,y^{(q-1)})) + 1, \text{ (REMOVE 2nd sequence)} & \text{otherwise} \\ \quad L((x^{(1)},\ldots,x^{(p-1)}),(y^{(1)},\ldots,y^{(q-1)})) + \rho(x^{(p)},y^{(q)})\} & \end{cases}$$

CHANGE symbols (edit)

Example Edit/Levenshtein Distance

		C	L	E	O	P	A	T	R	A
	0	1	2	3	4	5	6	7	8	9
C	1	**0**	1	2	3	4	5	6	7	8
A	2	1	1	2	3	4	**4**	5	6	**7**
E	3	2	2	**1**	2	3	4	5	6	7
S	4	3	3	2	2	3	4	5	6	7
A	5	4	4	3	3	3	**3**	4	5	**6**
R	6	5	5	4	4	4	4	4	**4**	5

C	A	E	S		A		R		
0	1	0	1	1	0	1	0	1	$\Rightarrow 5$
C	L	E	O	P	A	T	R	A	

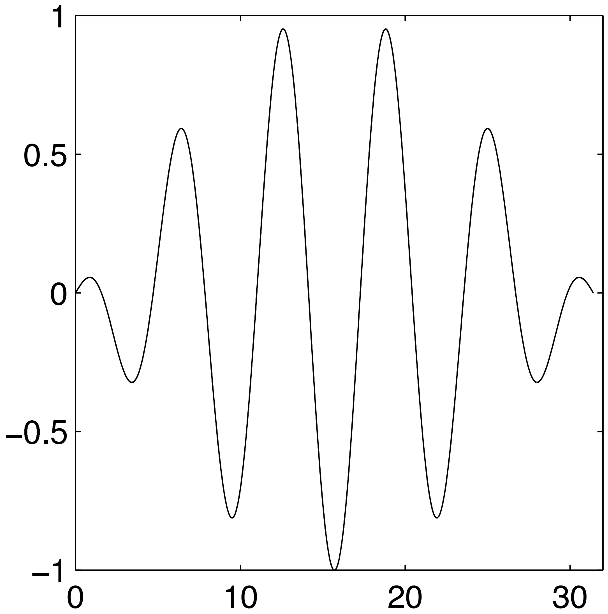
Prof. Dr. Thomas A. Runkler

Sampling Continuous Signals

continous functions

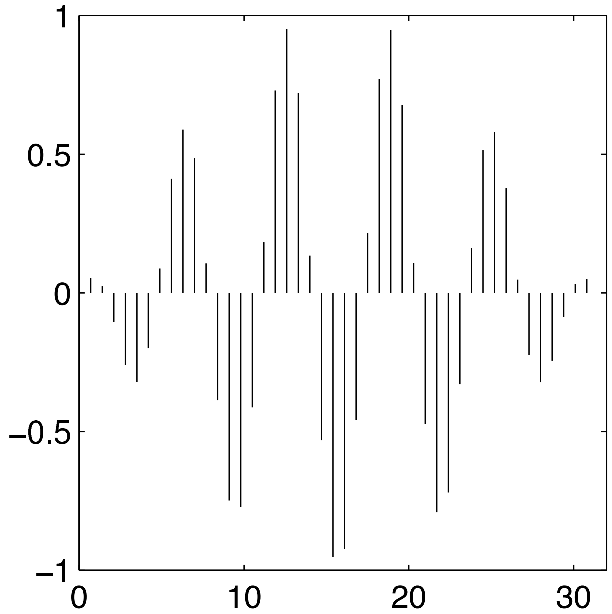
original signal

$s(t)$



sampled signal

$s_n = s(n \cdot \underbrace{T}_{\text{sampling time}})$



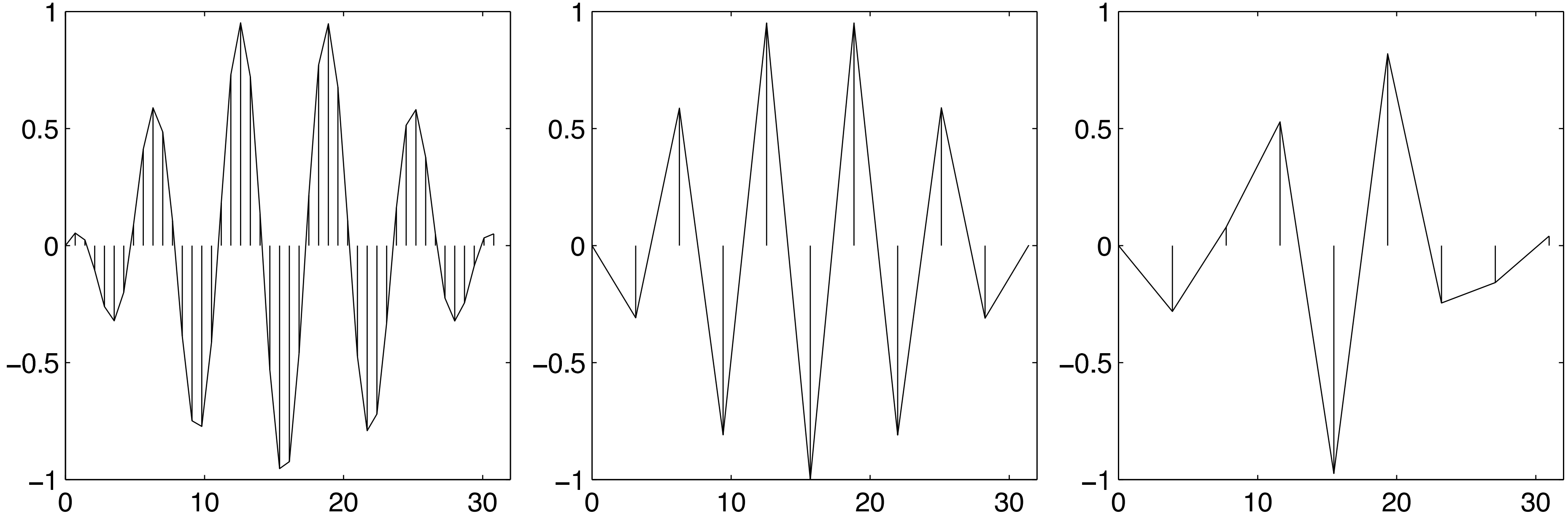
Prof. Dr. Thomas A. Runkler

Shannon's Sampling Theorem

1. $s(t)$ band limited: Fourier spectrum $|s(j2\pi f)| = 0$ for $|f| > f_{\max}$
2. $T_s < \frac{1}{2 \cdot f_{\max}}$ (Nyquist condition) limit

$\Rightarrow$ $s(t)$ can be completely reconstructed from s_n

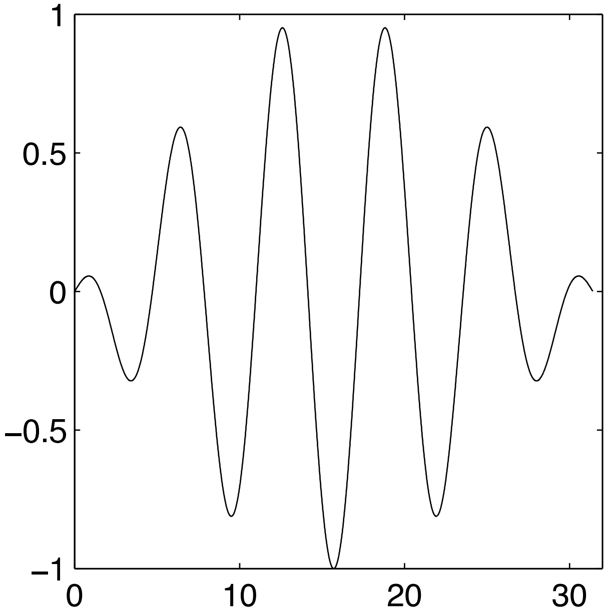
$T < \frac{1}{2 \cdot f_{\max}}$!! $\qquad$ $T = \frac{1}{2 \cdot f_{\max}}$ $\qquad$ $T > \frac{1}{2 \cdot f_{\max}}$



Prof. Dr. Thomas A. Runkler

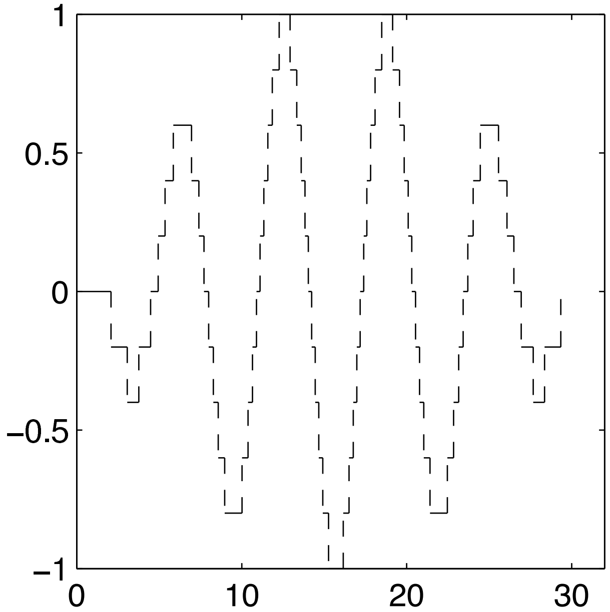
Quantization

original signal
$s(t)$



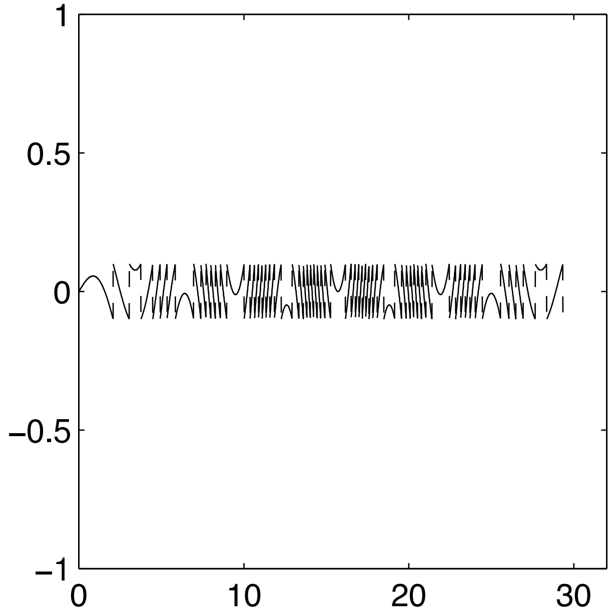
decimal limitation

quantized signal
$s_q(t)$



quantization
$d = 0.2$

quantization error
$s(t) - s_q(t)$



max. abs. error
$\max\{|s(t) - s_q(t)|\}$
$= \frac{d}{2} = 0.1$

Prof. Dr. Thomas A. Runkler

Chapter 3: Data Preprocessing

1. Error Types and Handling
2. Filtering
3. Standardization and Transformation
4. Data Merging

