Some Remarks on Text Data Visualization and Codec Transparency

Bryan Jurish

jurish@bbaw.de

VisiHu 2017: Visualisierungsprozesse in den Humanities

Universität Zürich

17th July, 2017

Overview

Preliminaries

- Full Disclosure
- Terminology: Data, Text, & Visualization

Remarks

- Pipelines, Parameters, & (visualization) Procedures
- Visualizations as Filters
- Lossiness, Compression, & ‘Universal’ Filters
- ‘Intuitivity’, Exploitation, & Coherence
- Co-operation & Codec Transparency

Summary

Full Disclosure

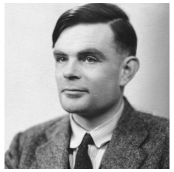
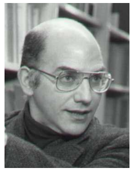
- I am a computational linguist
  - tinker of algorithms
  - tweaker of data structures
  - not a philosopher

(. . . but I played one as an undergraduate)

Full Disclosure

- I am a computational linguist
  - tinker of algorithms
  - tweaker of data structures
  - not a philosopher

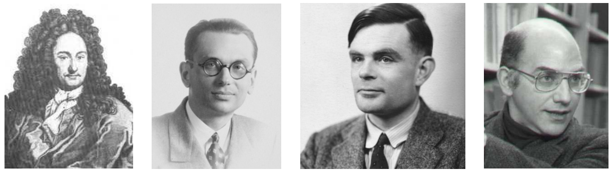
    (... but I played one as an undergraduate)

- ... I am also an incorrigible Platonist
  - $\Box\exists x.x = \emptyset$
  - formal (mathematical) objects really exist!
  - good company:  

Full Disclosure

- I am a computational linguist
  - tinker of algorithms
  - tweaker of data structures
  - not a philosopher

    (. . . but I played one as an undergraduate)

- . . . I am also an incorrigible Platonist
  - $\Box\exists x.x = \emptyset$
  - formal (mathematical) objects really exist!
  - good company: 

- Please adjust your interpretative apparatus if and where required
  - to accommodate my bottomless naïveté, and/or
  - according to your own epistemological commitments (or lack thereof)

Terminology

Visualization

- an algorithmic procedure by which an underlying ***data source*** is transformed to ***graphical form*** for direct human consumption
- e.g. as a network graph, tag cloud, motion chart, etc.

Text Data

- a (digital) text corpus, possibly including extralinguistic information such as bibliographic meta-data, document structure, etc.

Text Data Visualization

- a visualization procedure using a (digital) text corpus as its underlying data source (usually indirectly)

Visualization Pipeline

- a cascade of algorithmic procedures by which (raw) text data is prepared for and formatted by a particular visualization procedure, including any preprocessing and application-specific modeling

Remark 1: Pipelines *versus* Procedures

Facts

- ***raw text data*** itself does not directly support most visualization procedures
- each visualization procedure imposes ***formal constraints*** on its parameters

Claim

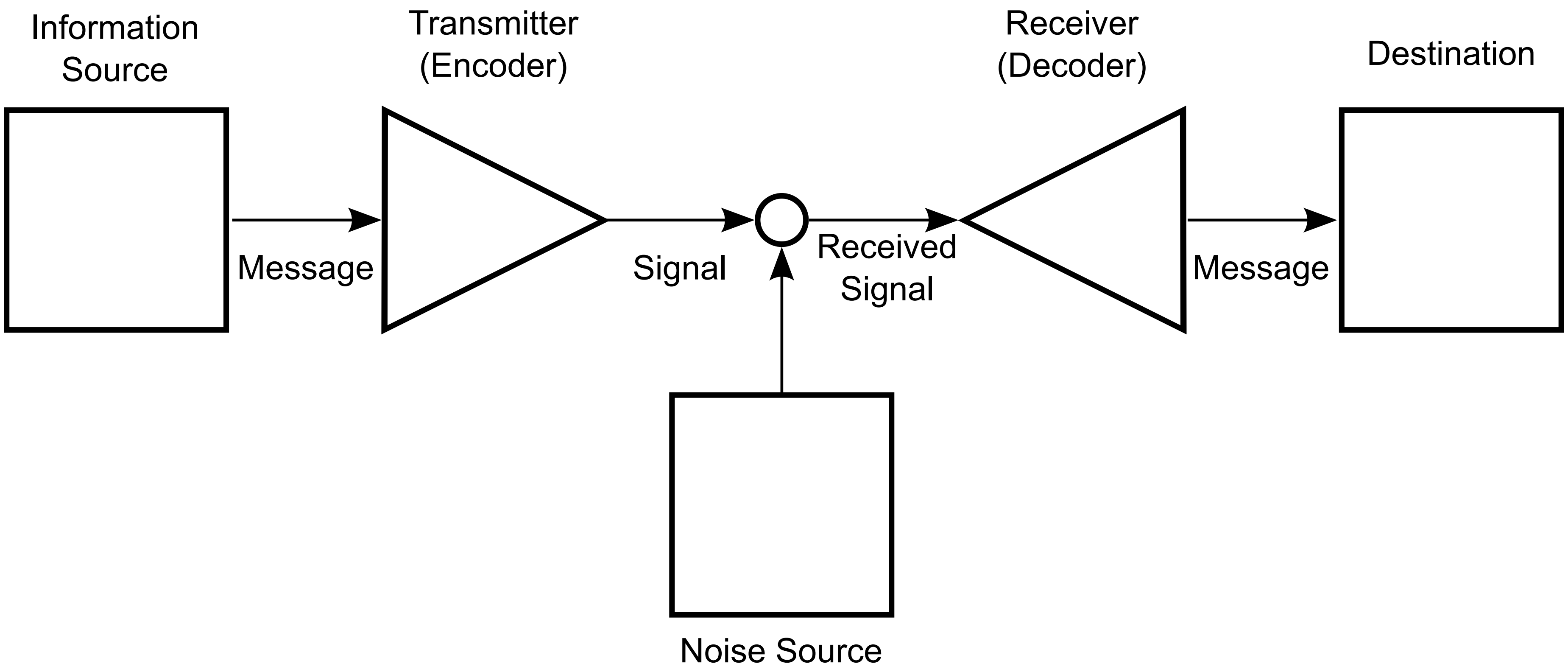
- (preprocessing) pipelines $\not\leq$ (visualization) procedures
- “generic” visualization procedures cannot be clearly distinguished from the preprocessing machinery (“***pipeline***”) which supplies their input

Rhetoric

- **Q**: how does one visualize a flat list of unweighted terms as a network graph?
  A: *one doesn't!* *(at least not in any meaningful way)*

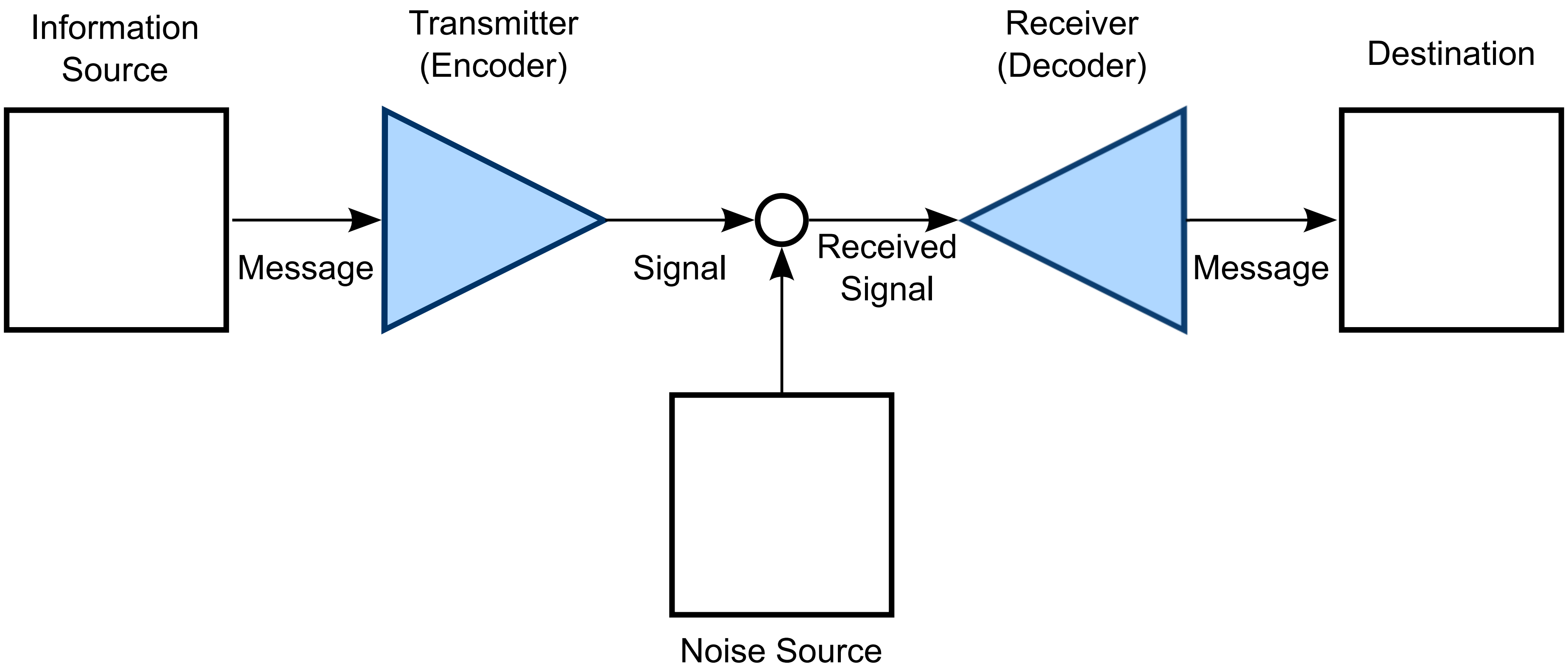
- **Q**: why is Mike Bostock's D3.js API so mind-bogglingly complex?
  A: *because it needs to be!* *(“generic” visualization procedures are fictional)*

Remark 2: Visualizations $\sim$ Filters



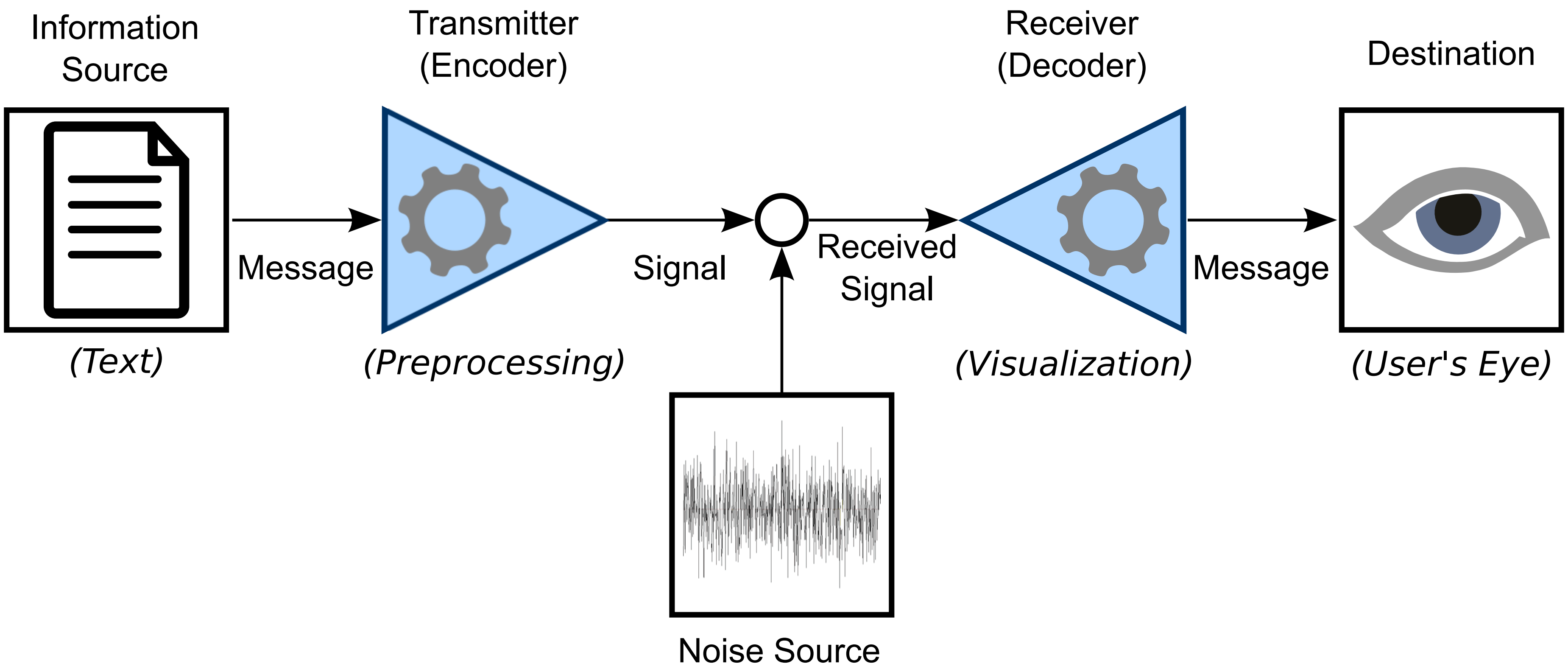
- noisy channel model of communication *(Shannon 1948)*

Remark 2: Visualizations $\sim$ Filters



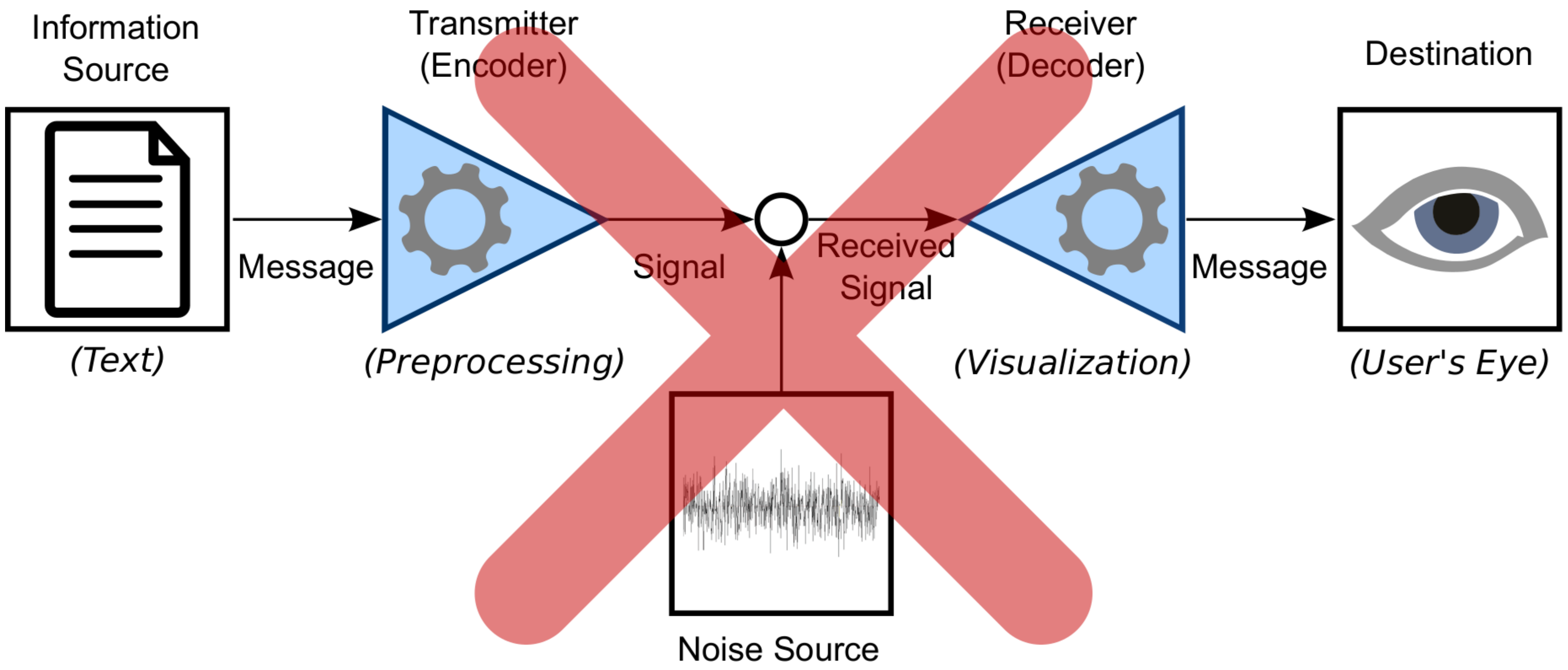
- noisy channel model *(Shannon 1948)*
  - “***codec***” $=$ encoder $\oplus$ decoder

Remark 2: Visualizations ∼ Filters



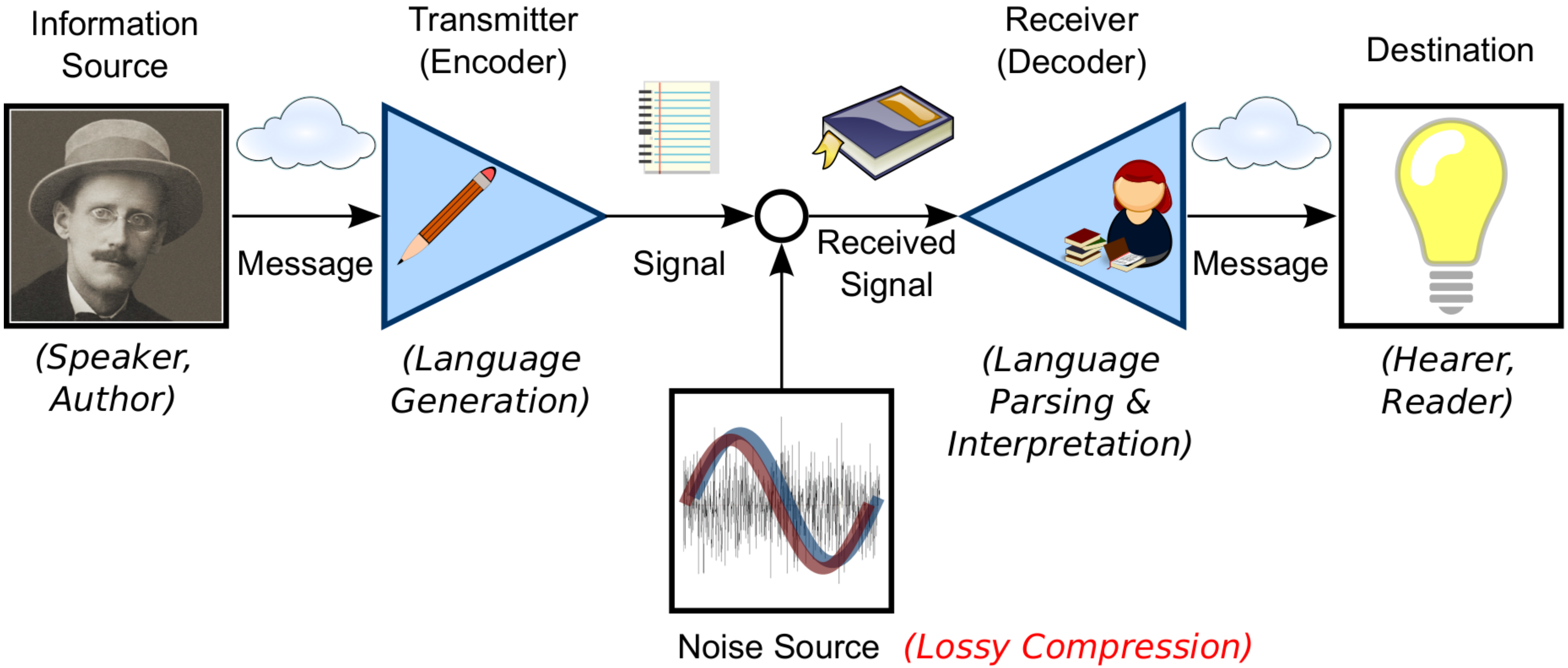
- noisy channel model *(Shannon 1948)*
  - “codec” = encoder ⊕ decoder
- text data visualization codec (naïve tinker’s version)

Remark 2: Visualizations ∼ Filters



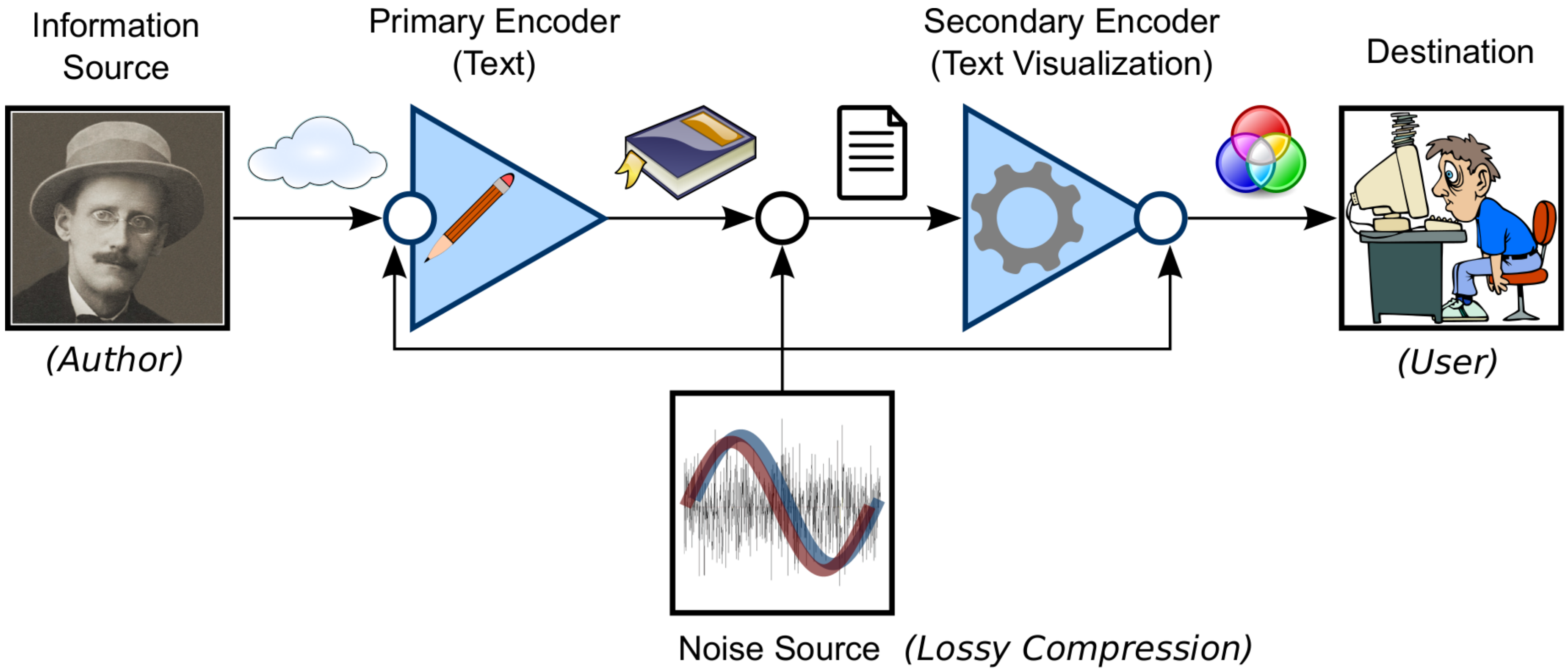
- noisy channel model *(Shannon 1948)*
  - “codec” = encoder ⊕ decoder
- text data visualization codec (naïve tinker’s version)⇝ ***not the whole story!***

Remark 2: Visualizations ∼ Filters



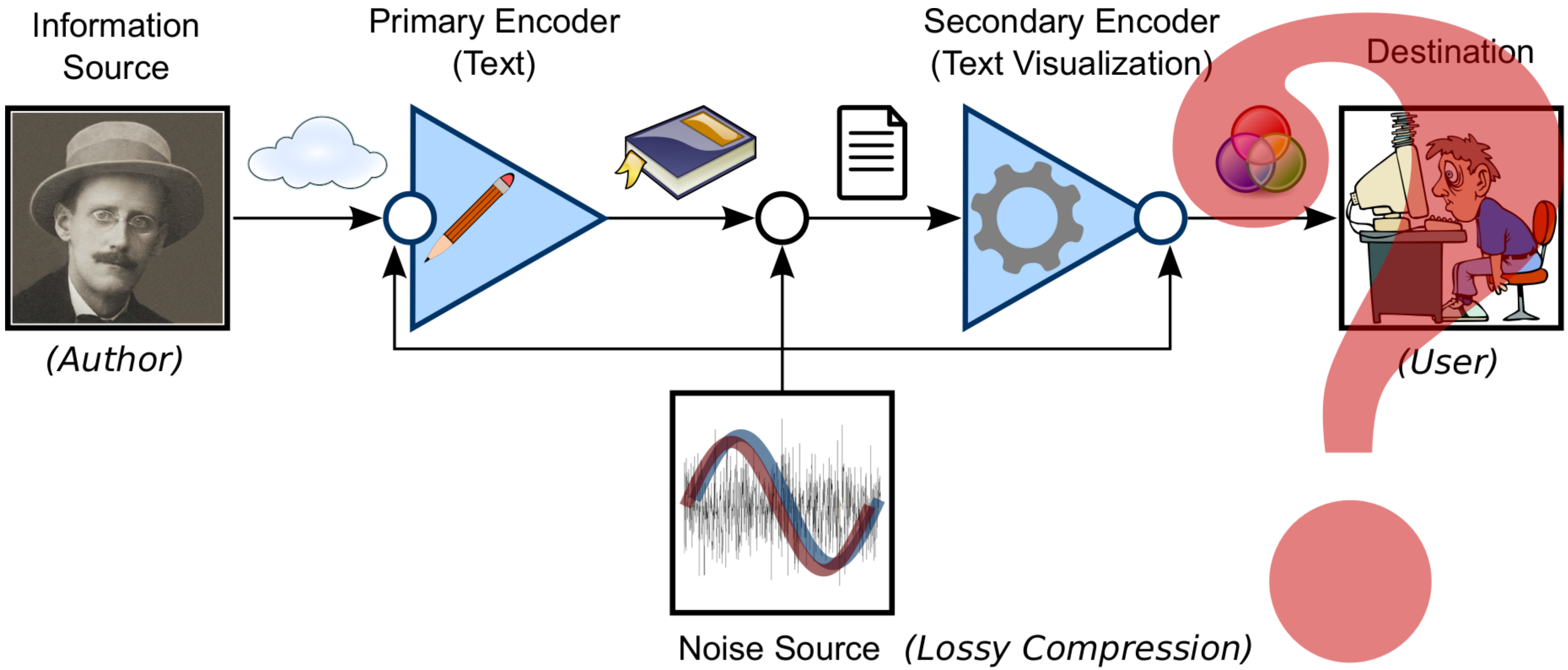
- noisy channel model (Shannon 1948)
  - "codec" = encoder ⊕ decoder
- natural language is a ***lossy codec*** (Reddy 1979)

Remark 2: Visualizations ∼ Filters



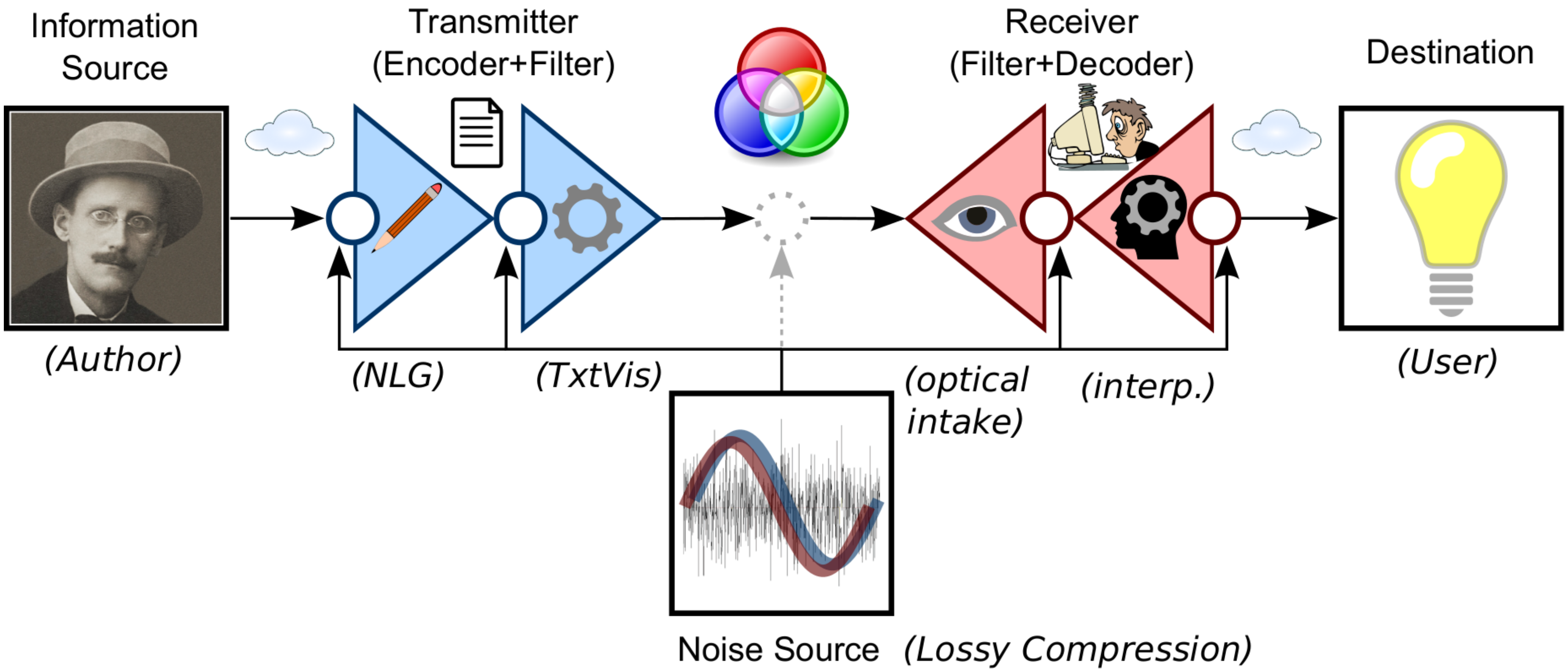
- noisy channel model *(Shannon 1948)*
  - "codec" = encoder ⊕ decoder
- natural language is a *lossy codec* *(Reddy 1979)*
- text data visualization is a (lossy) ***filter***

Remark 2: Visualizations ∼ Filters



- noisy channel model (Shannon 1948)
  - "codec" = encoder ⊕ decoder
- natural language is a *lossy codec* (Reddy 1979)
- text data visualization is a (lossy) *filter* ⇝ ***what about the decoder?***

Remark 2: Visualizations ∼ Filters



- noisy channel model *(Shannon 1948)*
  - “codec” = encoder ⊕ decoder
- natural language is a *lossy codec* *(Reddy 1979)*
- text data visualization is a (lossy) *filter* *(transmission side)*
- reception (interpretation) is filtered too!

Remark 3: Lossiness & ‘Universal’ Filters

Visualization Pipelines $\rightsquigarrow$ Lossy Compression

- information is ***lost*** when messages are passed through the codec
  - usually by design *(we already have the text-encoding)*
  - no lossless formal model of natural language available *(yet)*

‘Universal’ Filters

- as humans, we’re ***already equipped with*** a whole bevy of (lossy) filters:
  - linguistic *(minimal attachment, semantic priming)*
  - perceptual *(motion detection, color sensitivity)*
  - cognitive *(object independence, causal relations)*
  - cultural *(common knowledge, conventional signs)*

Lossiness $\sim$ ‘Distance’

- lossy filters increase “reading distance” *(Moretti 2013)*
- the communication channel was already fallible

Remark 4: ‘Intuitivity’ $\sim$ Exploitation

‘Intuitivity’

- ‘intuitive’ visualizations ***exploit*** users’ pre-existing (‘universal’) filters
  - perceptual ⇝ size, motion, color
  - cognitive ⇝ physical simulations, display “objects”
  - cultural ⇝ shared conventional signs
- reduced recipient processing load
  - “progressive disclosure” ⇝ conscious focus

Exploitation & Coherence

- successful exploitation ⇔ ***coherence*** of pipeline- & user-filters
  - all and only ***relevant*** information passes unchanged through both codecs
  - ***relevance*** depends on user’s individual research question

Remark 5: Co-operation ⇝ Transparency

Co-operation

> *"Make your contribution such as it is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged."* — Grice (1975)

Codec Transparency

- no ***perceptible*** data loss *(e.g. mp3, ogg audio codecs)*
- visualization ⇝ no ***apprehensible*** (relevant) data loss

Visualization as (co-operative) Communication

- **Task**: maximize transparency ⇝ optimize for users' common research goals
- **Challenges**:
  - research goals vary widely between users, projects
  - commonalities can be hard to identify and formally model

Summary

Visualization Procedures

- non-modular, interface constraints *(preprocessing pipelines)*

Visualization Pipelines

- noisy-channel filters *(lossy, usually by design)*

‘Universal’ Filters

- recipient-internal *(perceptual, cognitive, cultural)*

‘Intuitivity’

- exploitation of recipient filters *(relevance, coherence)*

Co-operative Communication

- maximize codec transparency *(minimize apprehensible loss)*

— The End —

Thank you for listening!

http://kaskade.dwds.de/~jurish/visihu2017/danke

References

P. Grice. Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax and Semantics*, volume 3: Speech Acts, pages 41–58. Academic Press, 1975.

F. Moretti. *Distant reading*. Verso Books, 2013.

M. J. Reddy. The conduit metaphor: A case of frame conflict in our language about language. In A. Ortony, editor, *Metaphor and Thought*, pages 284–310. Cambridge University Press, 1979.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27 (3):379–423, 1948.