

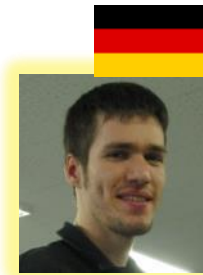
3rd IAPR Summer School on Document Analysis @Islamabad, Pakistan

Open Research Directions of Document Analysis and Recognition

9:30-11:00, Aug 23th(Fri), 2019
Seiichi Uchida (Kyushu University, Japan)

Who am I? (1/2)

- I come from **Fukuoka, Japan**
 - Ranked as one of the most “easy-living” cities in the world



Marcus and Imran have stayed there
for a couple of months



[Photo: Fukuoka city]

Who am I? (2/2)

- Research
 - Pattern recognition (PR)
 - especially for Document Analysis and Recognition (DAR)
 - Application of machine learning and optimization to PR problems
 - Image-informatics and computer vision
 - Interdisciplinary collaborations
- English
 - **SO BAD** (ex. $r \Leftrightarrow l$, $b \Leftrightarrow v$, $sh \Leftrightarrow s$)
 - I beg your kind effort to understand what I am saying

Introduction

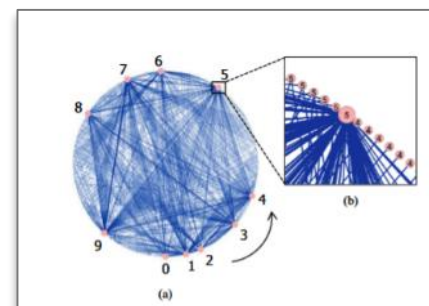
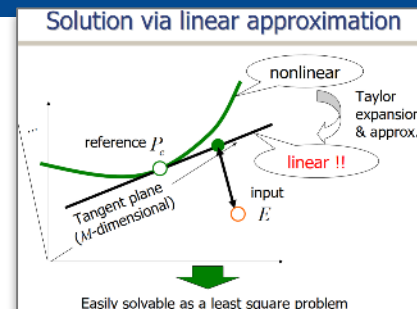
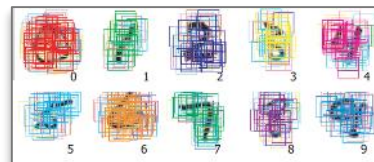
Let's think about the world "beyond 100%"

Prelude:

My 20 years of happy DAR research

• Character recognition methods

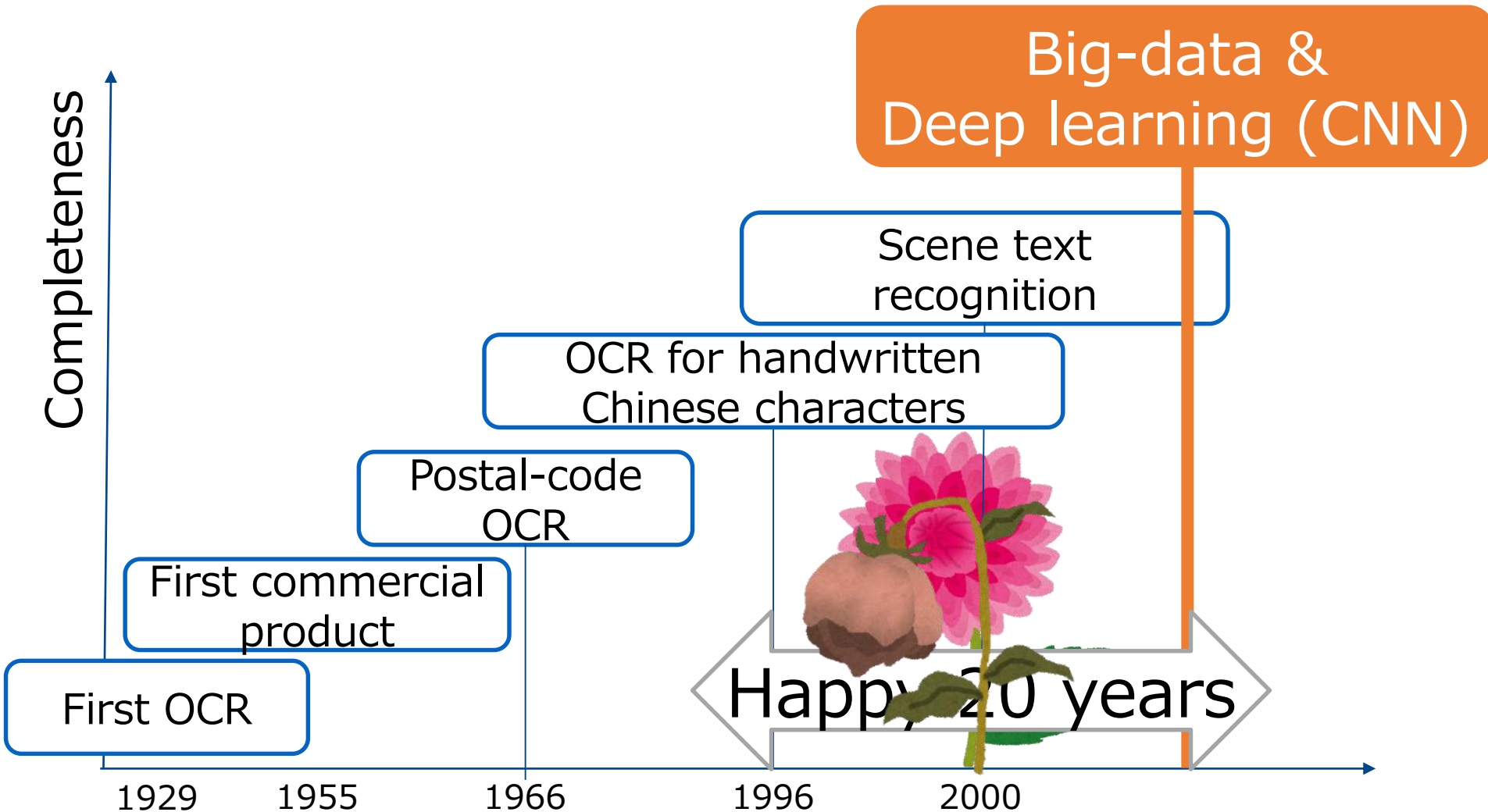
- DTW (elastic matching) and its variants
- Eigen-deformations
- Part-based methods
- Non-uniform slant correction
- Mathematical document recognition
- ...



• Scene text detection and recognition

- Context-aware detection
- Reading-life log
- Detection by multiple-hypothesis
- ...

My happy days were suddenly gone by....



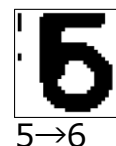
How did CNN kill me?: A personal experiment 1 [Uchida+, ICFHR, 2016]

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

Size: 32 x 32

#samples: 512,265

Accuracy: 99.99 %
(only 2 misrecognized images!)



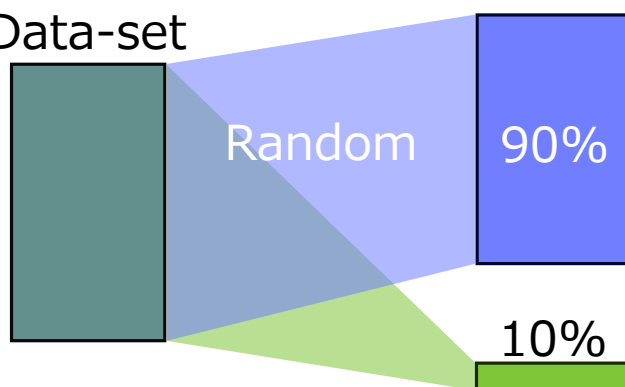
5→6



6→4

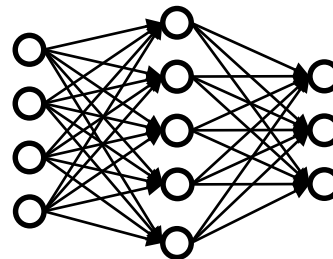


Data-set



Training

Trained CNN



Test





How did CNN kill me?: A personal experiment 2 [Uchida+, ICFHR, 2016]



Size: 32 x 32

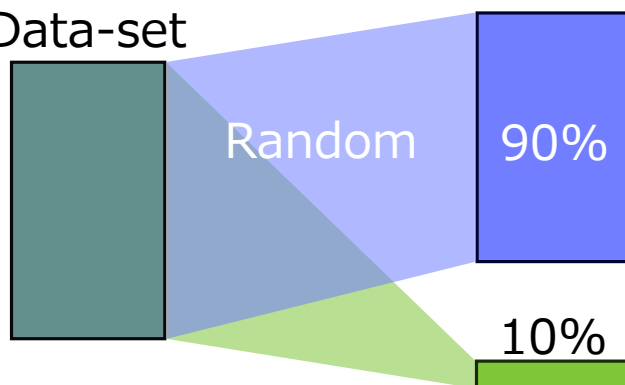
#samples: 819,652

Accuracy: 99.89 %
(only 92 misrecognized images!)

Ex.  0→6  2→7  7→1  9→4

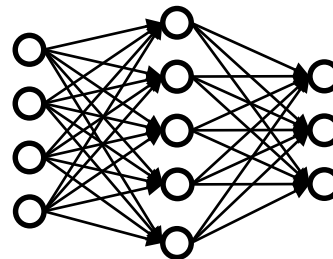


Data-set



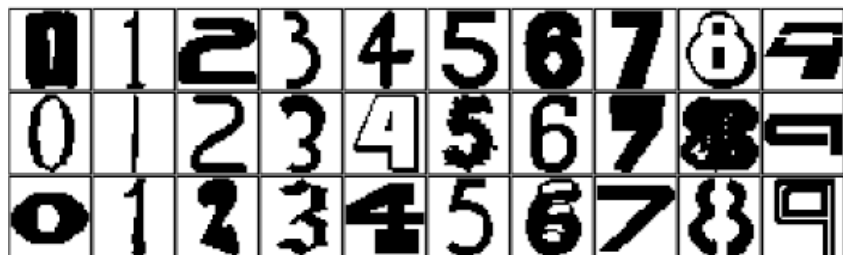
Training

Trained CNN



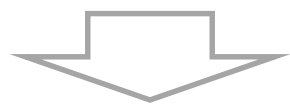
Test

How did CNN kill me?: A personal experiment 3 [Uchida+, ICFHR, 2016]

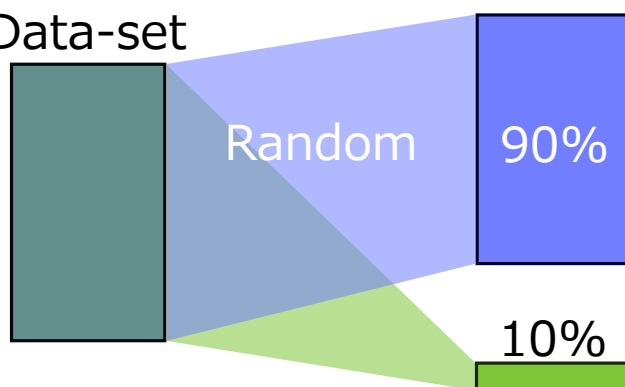


Size: 32 x 32

#samples: 6,721 fonts x 10 classes



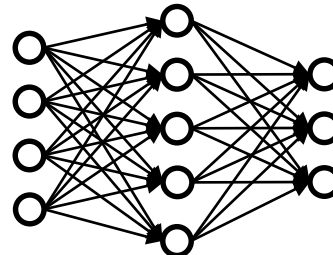
Data-set



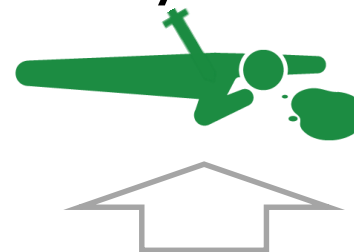
Training



Trained CNN



Accuracy: 95.7 %



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



Even for Latin
alphabet...

Performance of scene text detection and recognition is also getting better and better

- EAST [Zhou+, CVPR, 2017]



- CRNN [Shi+, TPAMI, 2017]

OPEN

Menu

White

pancake

panopoly

milk skin



Samples correctly recognized by CRNN

The SOTA performance is still upgraded even in 2019...

Then I start thinking
what we can do in the world
BEYOND 100%



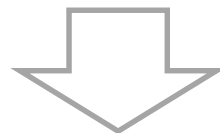
Actually, our real goal is **not** just to get perfect recognition results

Poor recognition results



Tentative goal

Perfect recognition results



Real goals

Ultimate **application**
by using perfect
recognition results

Scientific discovery
by analyzing perfect
recognition results

Today's topics: Open research directions in the world "beyond 100%"

- **Application**-oriented topics
- **Scientific** topics
 - Design of characters (letters, fonts,...)
 - Interaction between text and object
 - Interaction between text and human
 - Distribution of character patterns
 - Relationship to semantic analysis
- Conclusion

We still have many interesting topics around DAR research!



Today's "take-home" messages

- There are **still many interesting research topics!**
- Please do **not** think only about recognition **accuracy**
 - Recognition accuracy is just one aspect of DAR research
- Please **define your own task** (rather than just follow a task that someone defined)
 - Think why! Watch how! Think differently!
 - I think this might be the most important in this open-source/open-research era!

Application-oriented topics

What does “beyond 100%” mean?

- Computer can detect, read, collect, and analyze all textual information in the wild!



Texts on poster / ad



Texts on object label



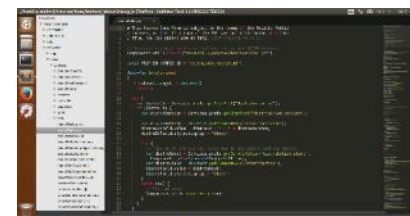
Texts on book page



Texts on signboard



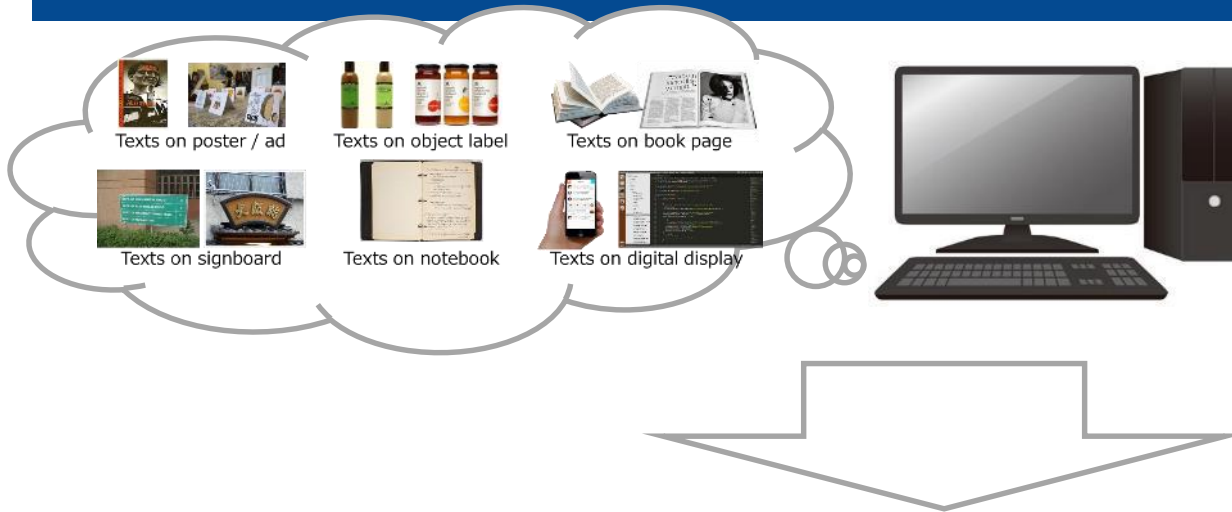
Texts on notebook



Texts on digital display



Many promising applications in the world “beyond 100%”!



Memorandum record / Personal knowledge-base /
Automatic diary /
Sharing / Comparison / Evaluation / Translation /
Recommendation / Suggestion / Real-time guide /
To-do support / Done-it support /
Education / Welfare /

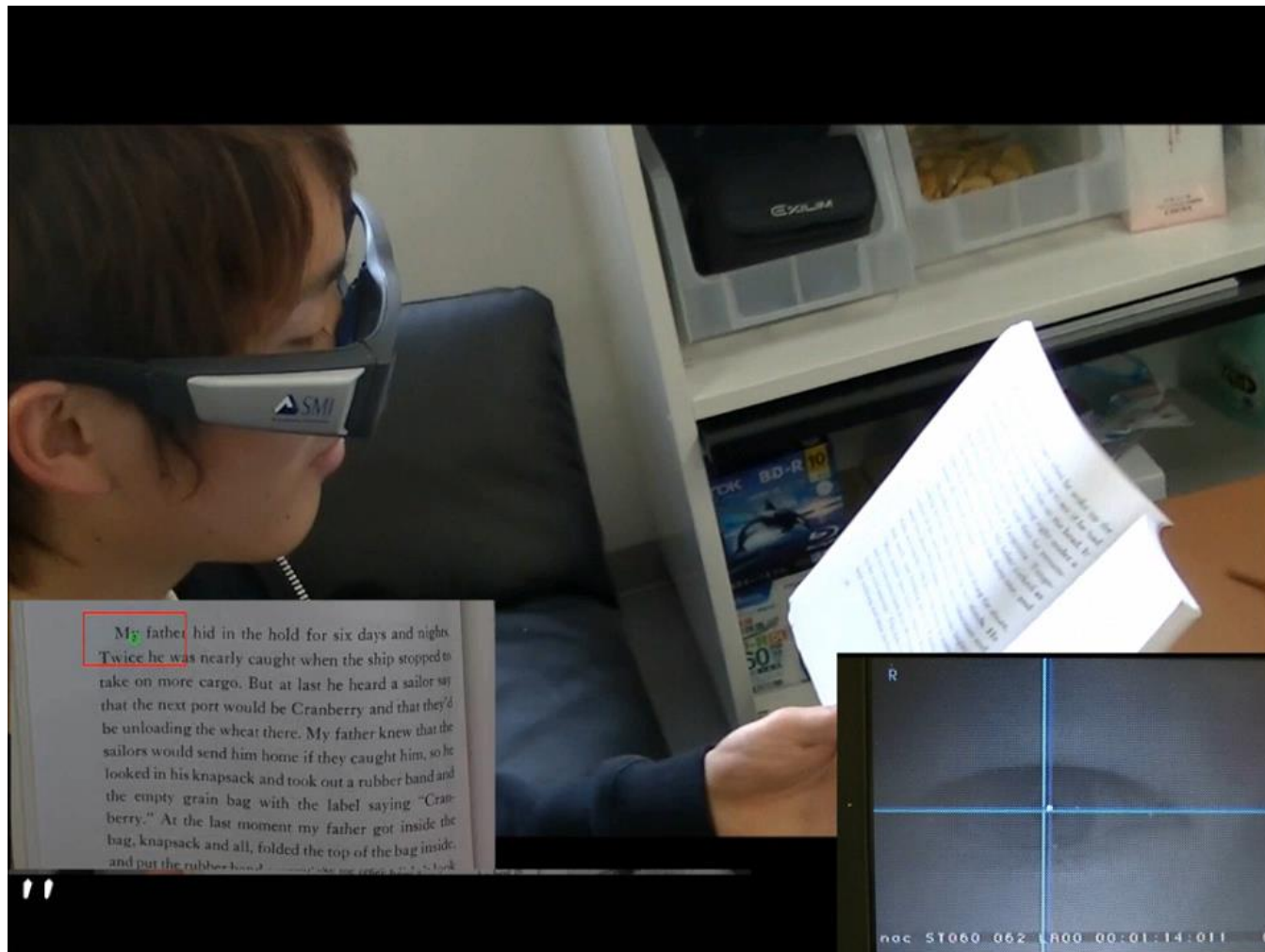


An application example: Reading **prescription** and **medicine box**

- for supporting pharmacists

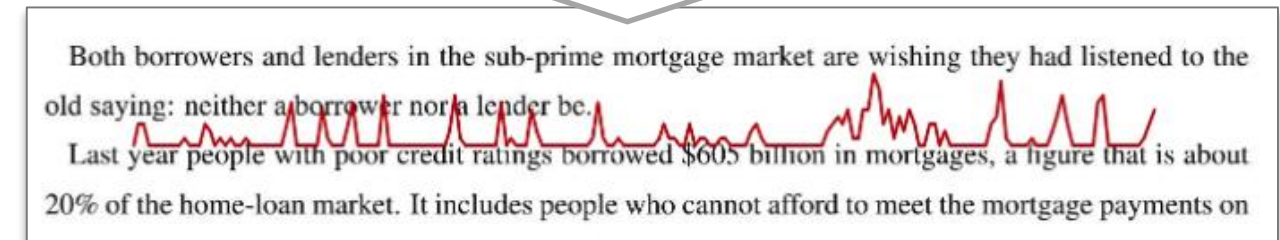
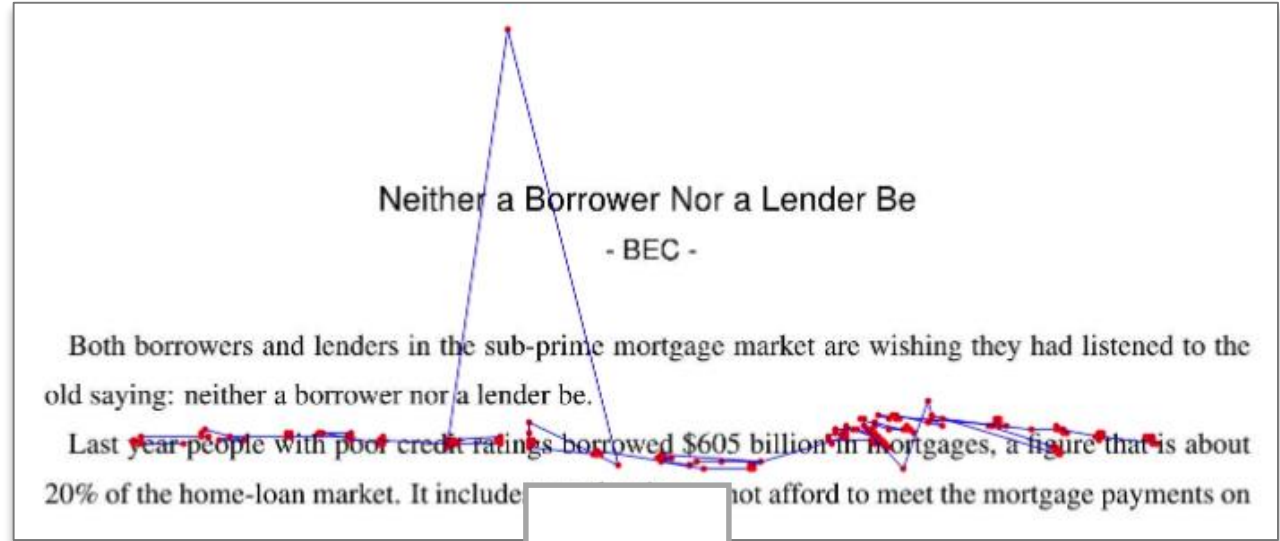
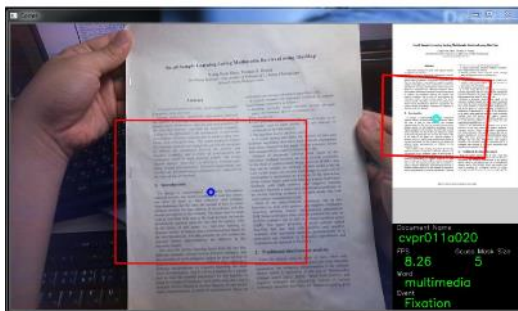


An application example: “Reading-life log”



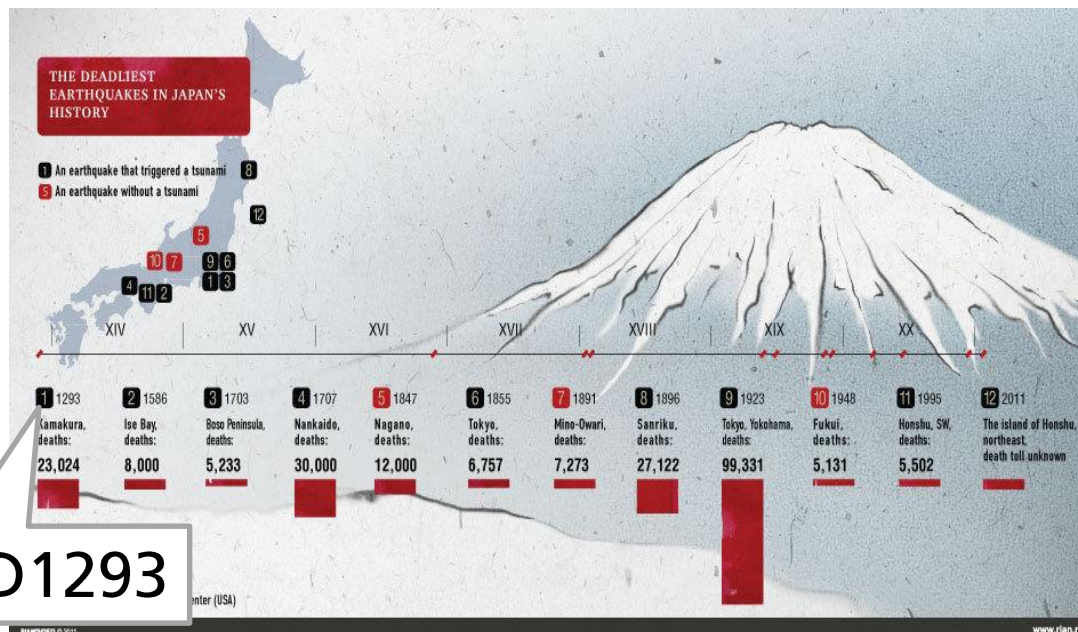
An application example: Deeper analysis of **reading** activity

- Ex. eye-movement pattern shows the understanding level of the reader



(Already tackled but still) open problem: Historical document recognition

- History is recorded **only in** historical documents!
- Ex. Ancient earthquake records (useful for future earthquake prediction)



(Already tackled but still) open problem: Form understanding

What should be written here?

- One of ultimate goals of OCR research ...

Order Form		Order id:		Order date: (dd/mm/yy)	
Customer Name				Gender	
Given	Family				
Address	ZIP			Country (check one)	Pakistan
					Germany
					Japan
Order list	Item	Qty	Price (tax)	Note	
			total \$____(____)		
Payed? (y/n)		Payment	Cash \$____	Credit card \$____	Wire trans. \$____

(Already tackled but still) open problem: **Cartoon (Manga)** recognition

- It is also one of ultimate DAR tasks!!
- Manga 109 dataset



"Akkerakanjincho"
Copyright: Kobayashi Yuki

onomatopoeia

decorated text



"ARMS"
Copyright: Kato Masaki

Other possible applications (1/2)

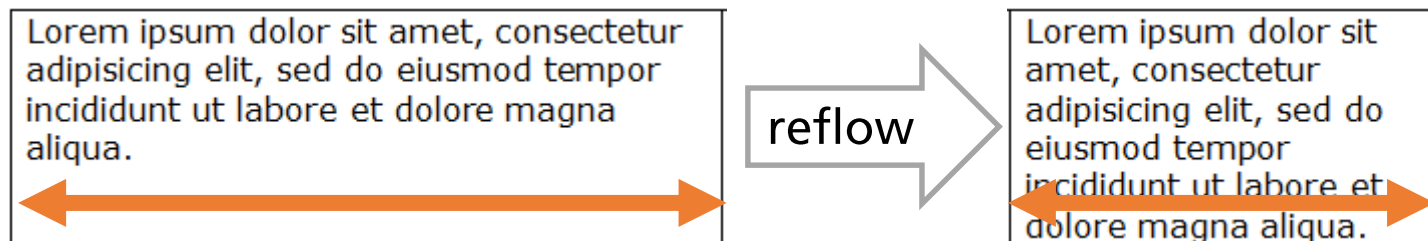
Augmented-reality (AR) × Scene text

- Showing scene texts appropriately for users
 - ex.1: Scene text translator (done)
 - ex.2: Scene text magnifier



[Nakamura+, ICDAR2019]

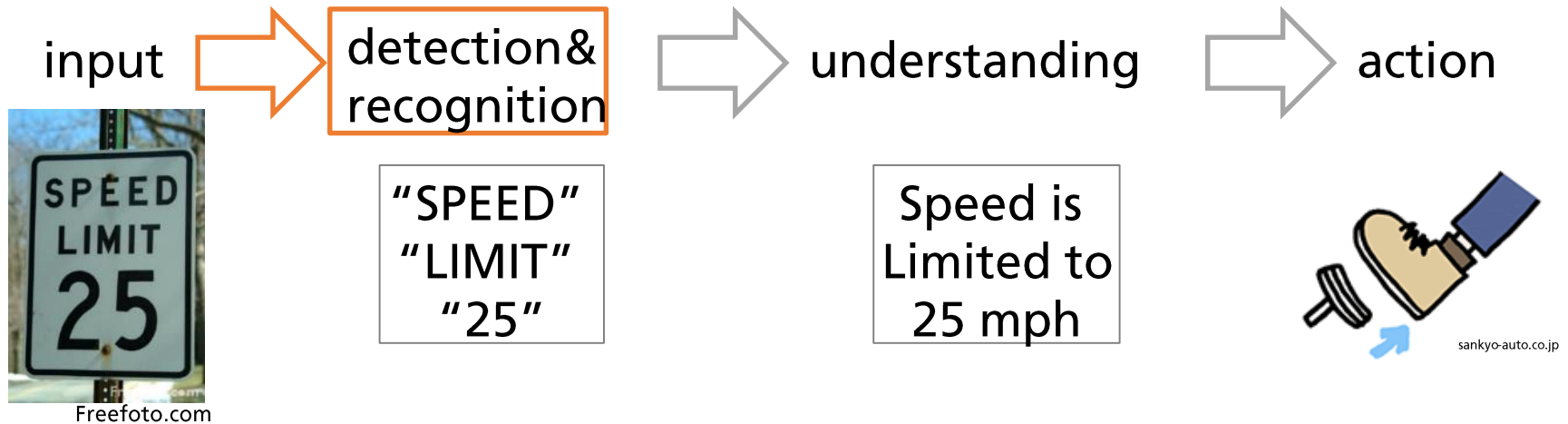
- ex.3: Reflowable scene text
 - Reformatting scene texts for individual displays or scene conditions
 - ↓ General idea of “reflowable document”



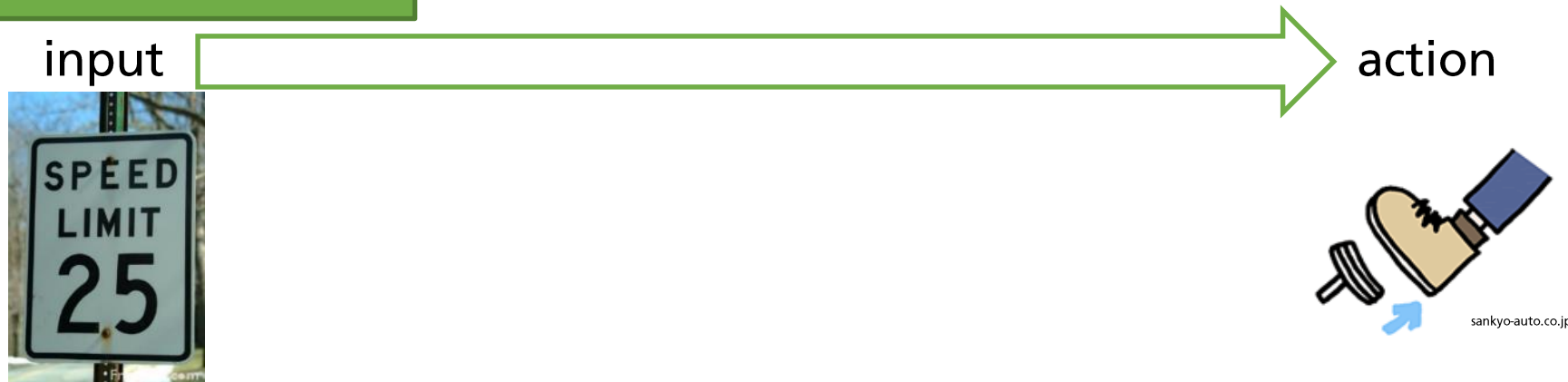
Other possible applications (2/2)

Real end-to-end systems

Typical "end-to-end"

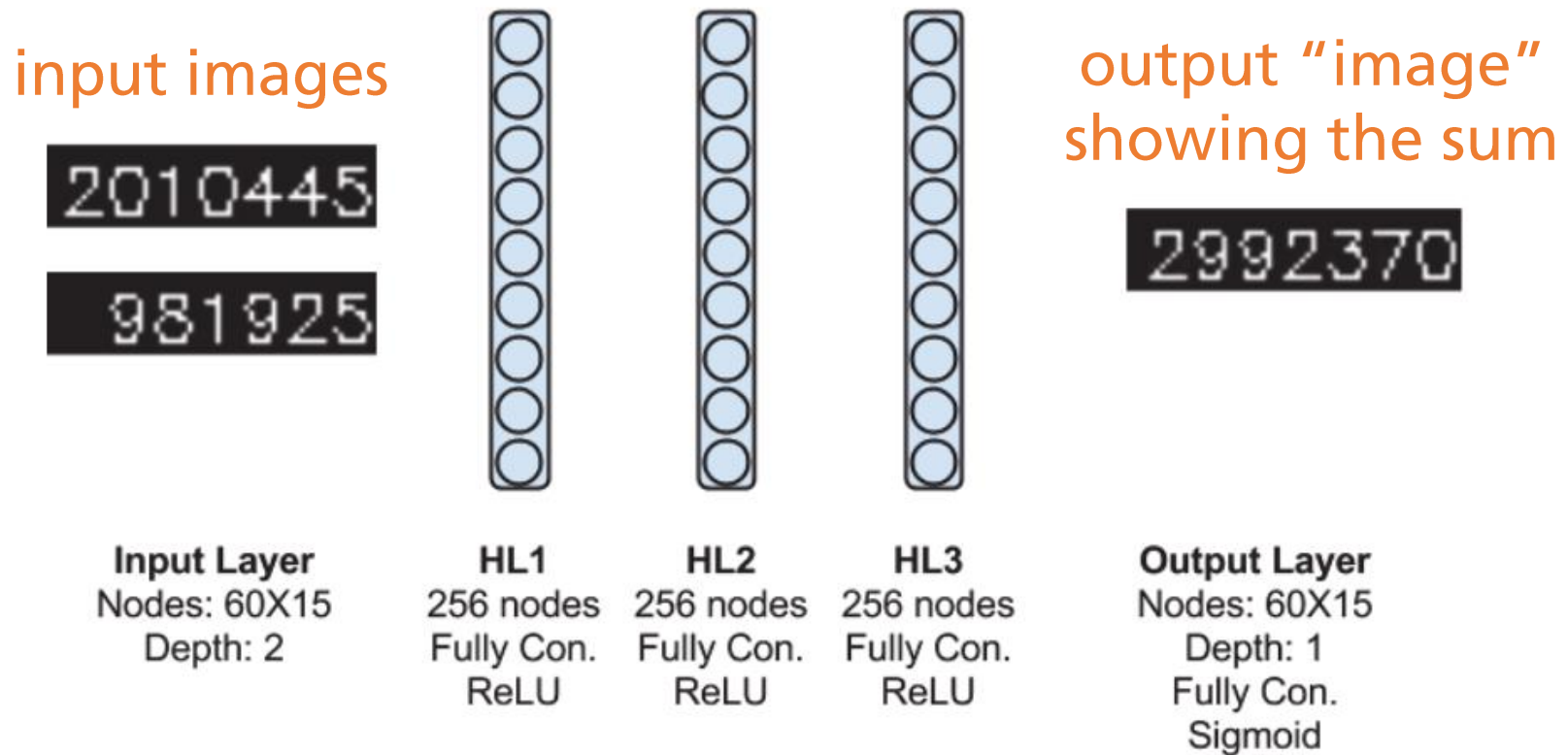


Real "end-to-end"



A hint to real end-to-end system!? (1/2)

Image-based calculation by neural network



A hint to real end-to-end system!? (2/2)

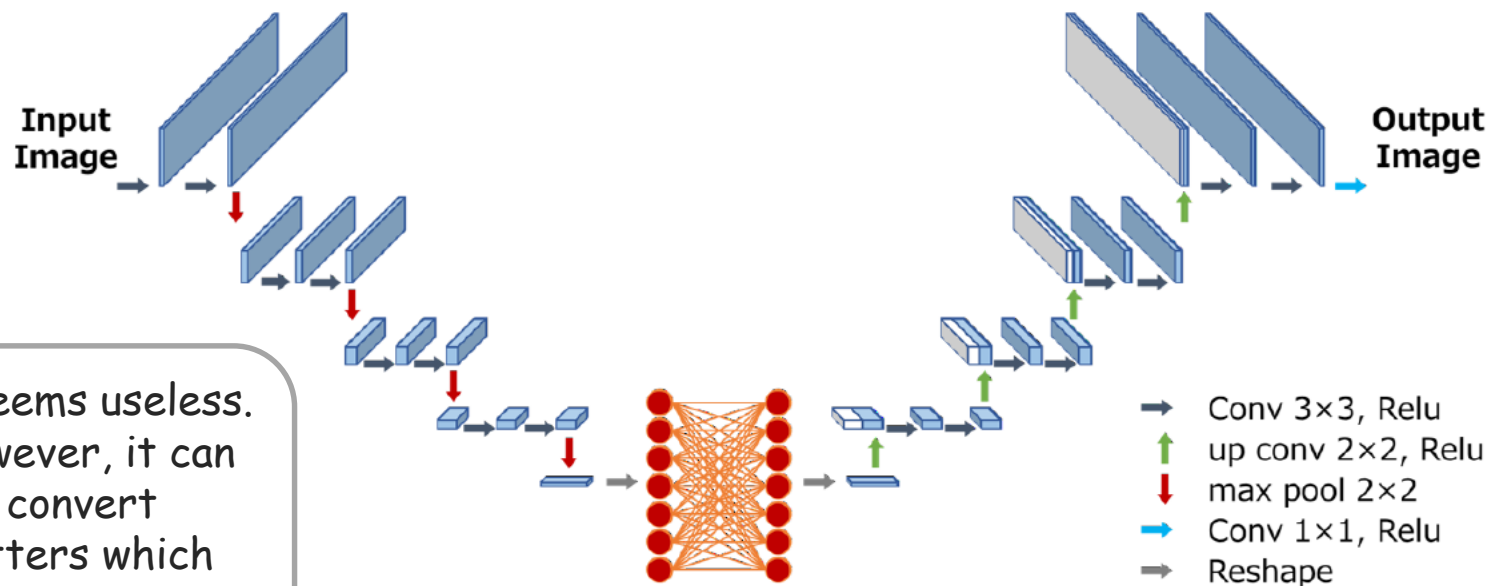
Image-based language conversion

input
image

랏 총 먼 늑 뭇 찡 씨 맥 닉 톱

output
image

ryah-chuh-ddyeod-nurp-mwej-chwak-sseogs-mwaeg-nigg-twaek



It seems useless.
However, it can
convert
letters which
never appear in
the training data

Possible research targets for applications: A brief summary

- Development of “reading-life log”
- Utilizing the logged text information
 - Education, welfare, user-interface, ...
- Complex document understanding
 - Form, manga (cartoon), historical document, ...
- Development of real end-to-end systems
 - The real goal of DAR is not just recognition

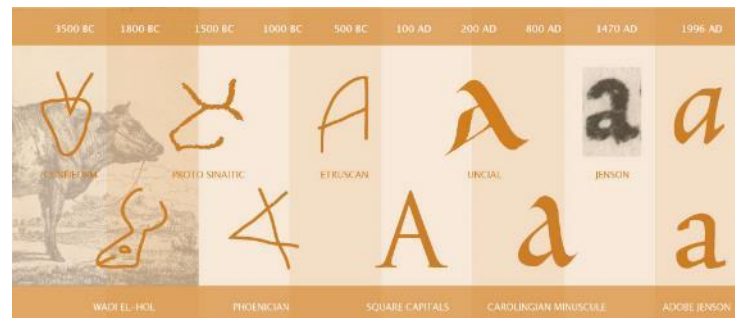
Scientific topic #1

Design of characters
(letters, fonts,...)

Character images are very special

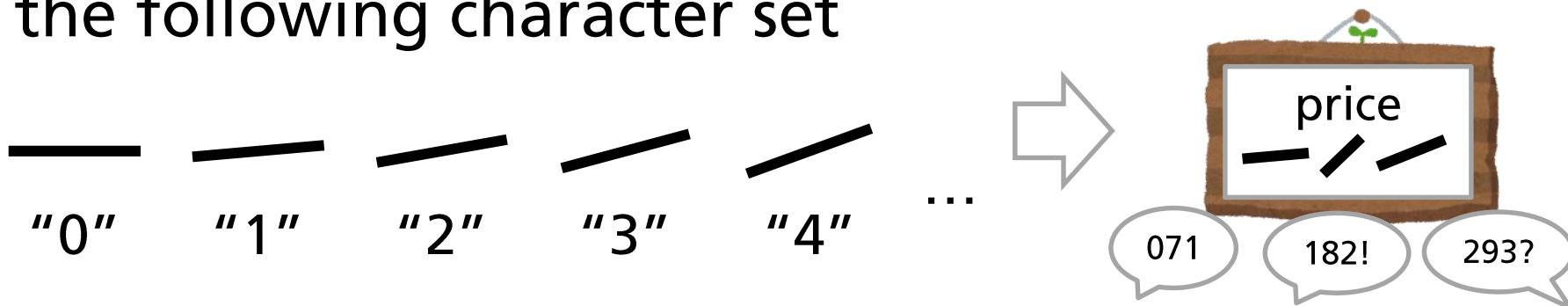


- Simple binary image
- Stroke-structured pattern
- Small size (ex. 32x32 pix)
- Predefined classes (ex. 10 classes for digits)
- Visual communication code **designed by human**
 - All characters in the world were artificially-designed



Human-beings are so smart on generating character sets!

- Human-beings have never generated the following character set



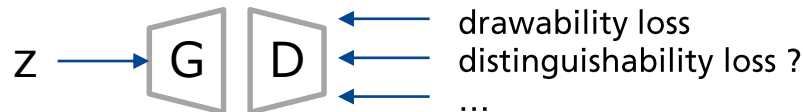
- So, all characters are generated so carefully for better legibility even if they seem so complicated



Q. Can we simulate character generation process by computer?

A	B	C	D	E
F	G	H	I	J
K	L	M	N	O
P	Q	R	S	T
U	V	W	X	Y
Z	A	E	O	T

- Probable conditions
 - Easily **drawable** (by hand or pen)
 - Easily **reproducible**
 - **Distinguishable** from each other
 - **Robustness** to various distortions
- Possible research topics
 - Character symbol generation with above conditions



[glyphwiki]

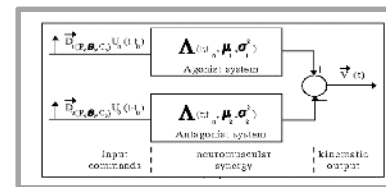
- (If it is successful,) which conditions are more relevant?
- Can we generate the 27th letter for English alphabet?

Q. How are character symbols *written*?

- Handwriting is a **special** temporal pattern
 - Physically-constrained by arm movement
 - Mainly Markovian, but Non-Markovian movement
 - e.g., writing a closed circle
 - Invisible “pen-up” movement

- Possible research topics

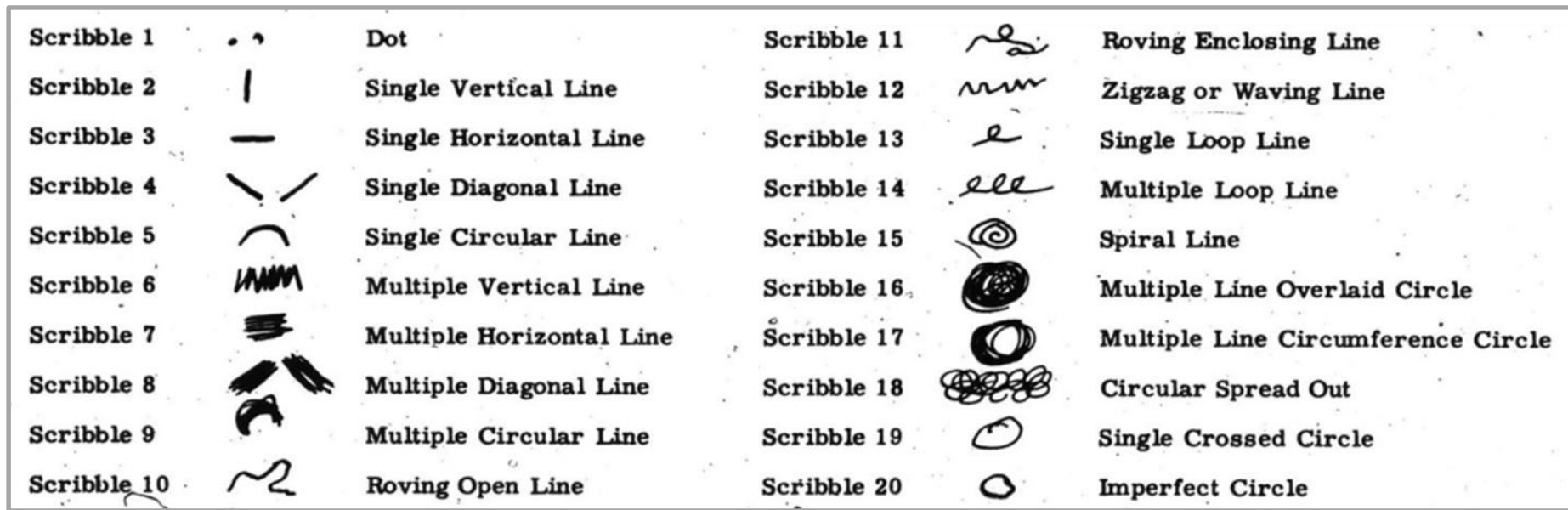
- Refine kinematic models
 - e.g., lognormal model by Prof. Plamondon
- Non-Markovian generative models
- Stroke-order recovery



Réjean Plamondon

Q. How are character symbols *written*?

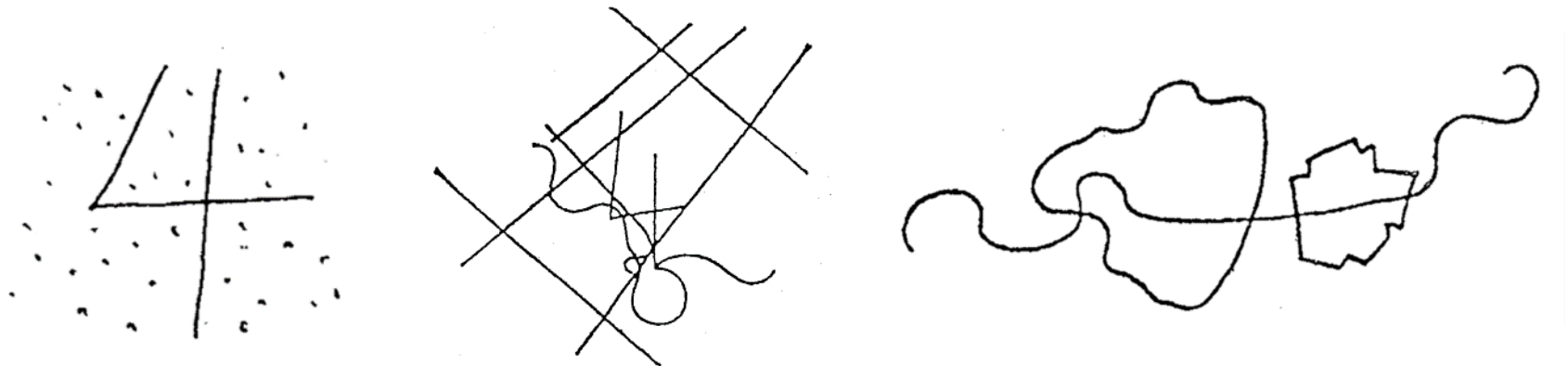
- Possible research topics (cont'd)
 - Can we learn anything from **children's scribble**?



20 basic scribbles by [Rhoda Kellogg, *Analyzing Children's Art*, National Press Books, 1969]

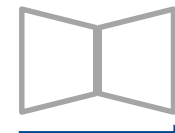
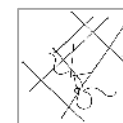
Q. Why/how do characters get their robustness to various distortions?

- Character images are so robust to distortions!
 - =They are “error-correcting codes”



[Ogawa+, IEICE1994]

- Possible research topics
 - OCR challenge on distorted character images
 - What kind of distortions will disturb OCR performance?
 - Distortion removal by GAN or U-nets
 - Especially, knowing its limitation is important



Q. Why/how do characters get their **robustness to various distortions**?

Characters can hide characters

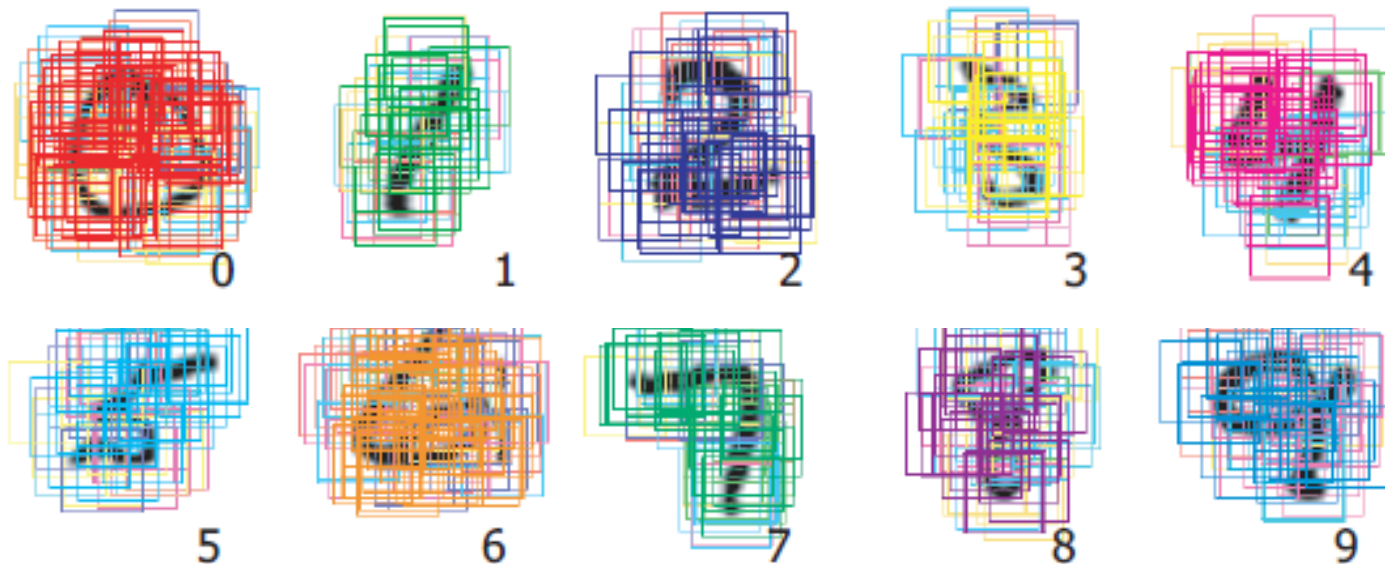


- Possible research topic
 - Automatic generation of the printing pattern to hide characters (by GAN)

Q. Why/how do characters get their **robustness to various distortions**?

Characters can be recognized by their **parts**

- Part-wise recognition accuracy ~ 40%
- Majority-voting within a character ~ 95% (!)



[Uchida&Liwicki, ICFHR2010]

- Possible research topic
 - Can we interpret or reorganize a CNN (esp. pooling and softmax) as a part-based recognizer?

Q. Well, what is "A"?

[Hofstadter, *Metamagical Themas*, 1985]

- = Can we define the **class** of "A"?
- \subset Can we define a "class" ?
- A very crucial question of pattern recognition



- However, it seems almost impossible to give a top-down definition of "A"

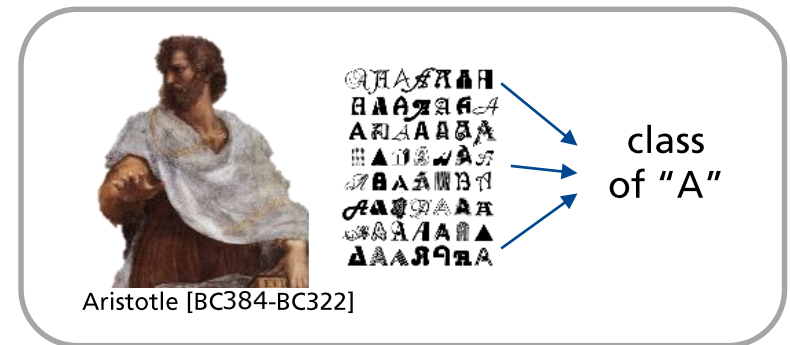
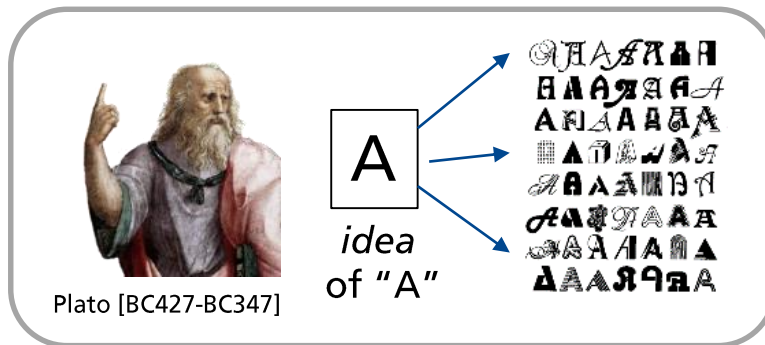
No common characteristics



Q. What is "A"?

[Hofstadter, *Metamagical Themas*, 1985]

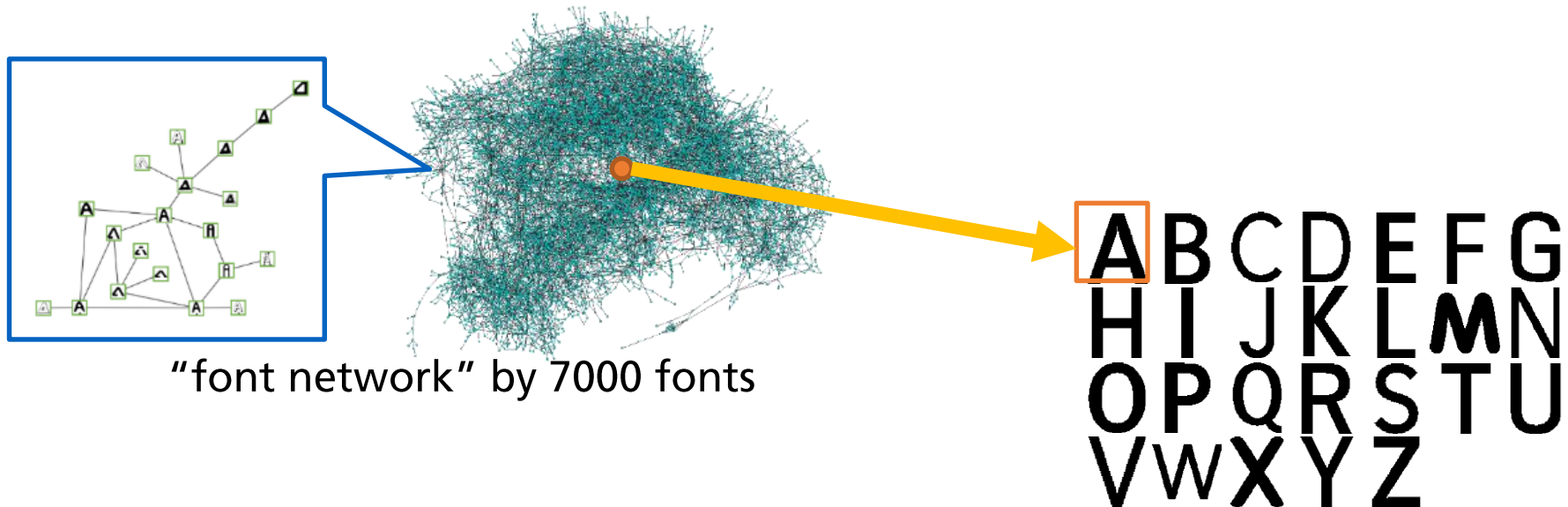
- Possible research topics
 - Can we make a reasonable **bottom-up definition** of the **class** "A" by collecting all "A" in the world



Q. What is "A"?

[Hofstadter, *Metamagical Themas*, 1985]

- Possible research topics (Cont'd)
 - A simplified question: What is the **standard "A"**?

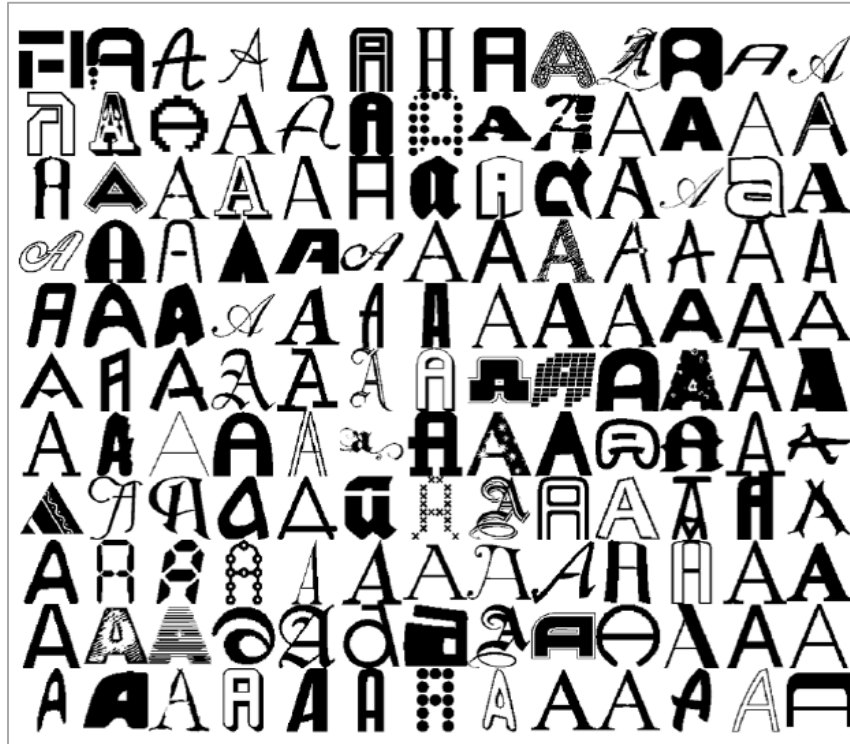


The font with maximum closeness centrality

[Uchida+, ICDAR2015]

Q. Well, why do we have **various fonts**?

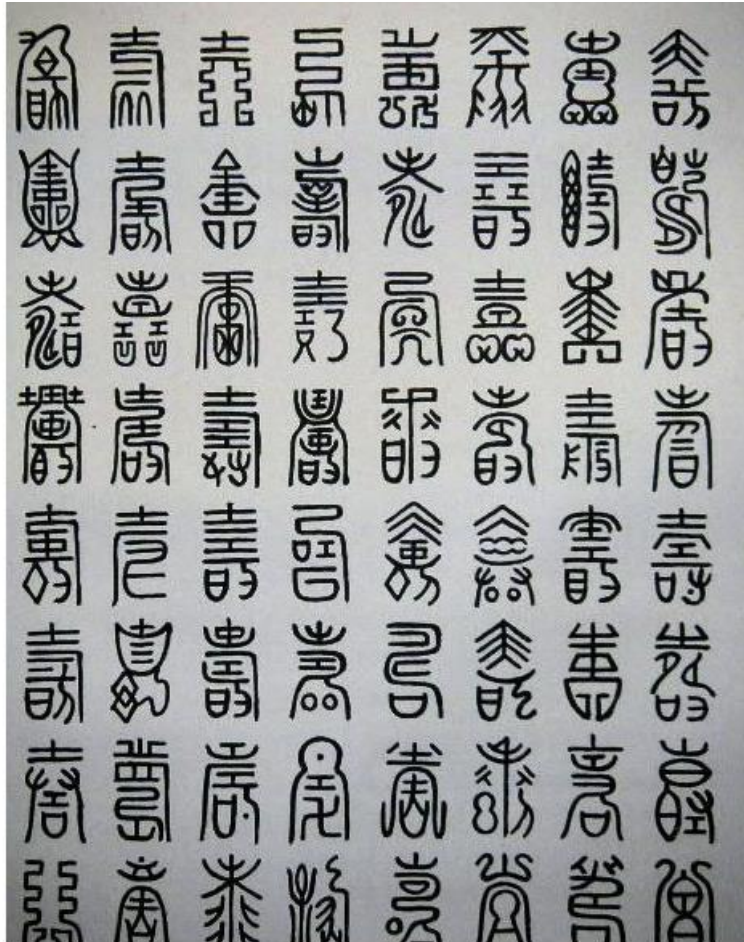
- For text-based information transmission, it is enough to have **only one font**



- However, we have thousands of fonts... Why?

Note: Chinese letter has many **allographs**
(in addition to font variations)

All of



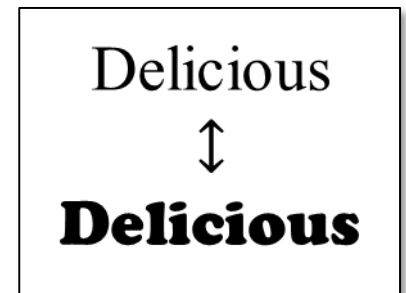
are “寿”
(congratulations)

However,
“干” and “干”
(thousand) (dry)
are different

Q. Why do we **design various fonts**?

Four possible reasons

- Struggles toward better legibility
- For providing special functions
 - ex. Captcha, machine-readability
- Just by a passion for creativity
 - hundreds of font designers in the world
- For giving **special impressions**
 - Nonverbal information from font!



Q. Why do we **design various fonts**?

Font and its impression

- Each font gives a different impression



Q. Why do we design various fonts?

Font and its impression

- Actually, how do you choose one without drinking?



jrsnchzhrs "Juice mix" (Flickr Creative Commons, CC BY-ND 2.0)

Q. Why do we design various fonts?

Font and its impression

- Possible research topics
 - Collecting fonts used in a specific case
 - Ex. Fonts used in comic book title
 - Font “feature” vs. impression
 - Whole shape feature (balance, area, aspect ratio, texture, ...) vs. impression
 - Local shape feature (Serif, stroke width, straightness, ...) vs. impression
 - Color feature vs. impression
- Generating a font (or logo) that shows a specific impression
- A font recommendation system

Subjective analysis of font impression has been done from AD1923

Q. Why do we **design various fonts**?

Font and its impression – A personal trial (1/2)

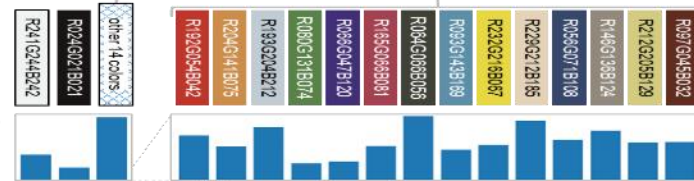
on Bookcovers

Color

Font

Genre

	#images
Arts & Photography	6,460
Biographies & Memoirs	4,261
Business & Money	9,965
Calendars	2,636
Children's Books	13,605
Christian Books & Bibles	9,139
Comics & Graphic Novels	3,026
Computers & Technology	7,979
Cookbooks, Food & Wine	8,802
Crafts, Hobbies & Home	9,934
Education & Teaching	1,664
Engineering & Transportation	2,672
Gay & Lesbian	1,339
Health, Fitness & Dieting	11,886
History	6,807
Humor & Entertainment	6,896
Law	7,314
Literature & Fiction	7,580
Medical Books	12,089
Mystery, Thriller & Suspense	1,998
Parenting & Relationships	2,523
Politics & Social Sciences	3,402
Reference	3,268
Religion & Spirituality	7,559
Romance	4,291
Science & Math	9,276
Science Fiction & Fantasy	3,800
Self-Help	2,703
Sports & Outdoors	5,968
Teen & Young Adult	7,489
Test Preparation	2,906
Travel	18,338



Q. Why do we **design various fonts**?

Font and its impression – A personal trial (2/2)

on Online advertisement



Genre

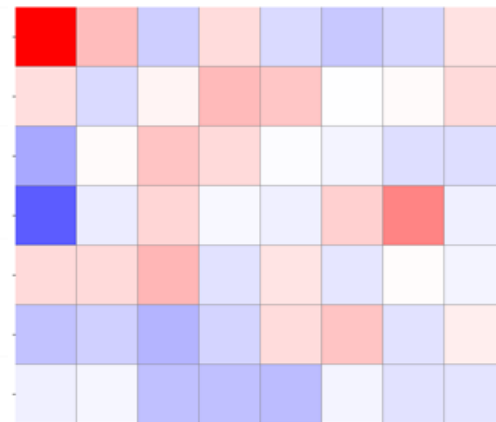
Business and Finance
Careers
Content Channel
Healthy Living
Personal Finance
Style & Fashion
Video Gaming

Color



average frequency

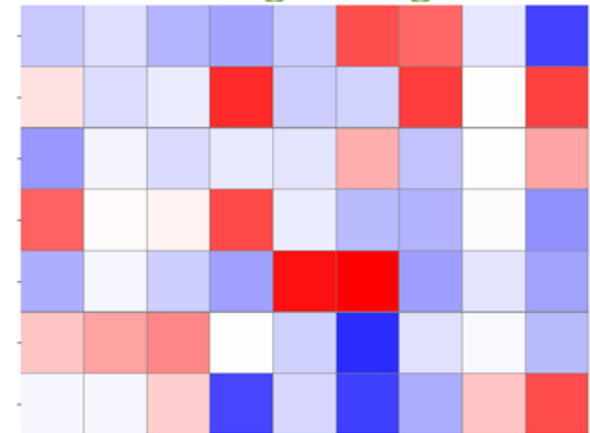
genre-wise difference from the average usage



Font

classes

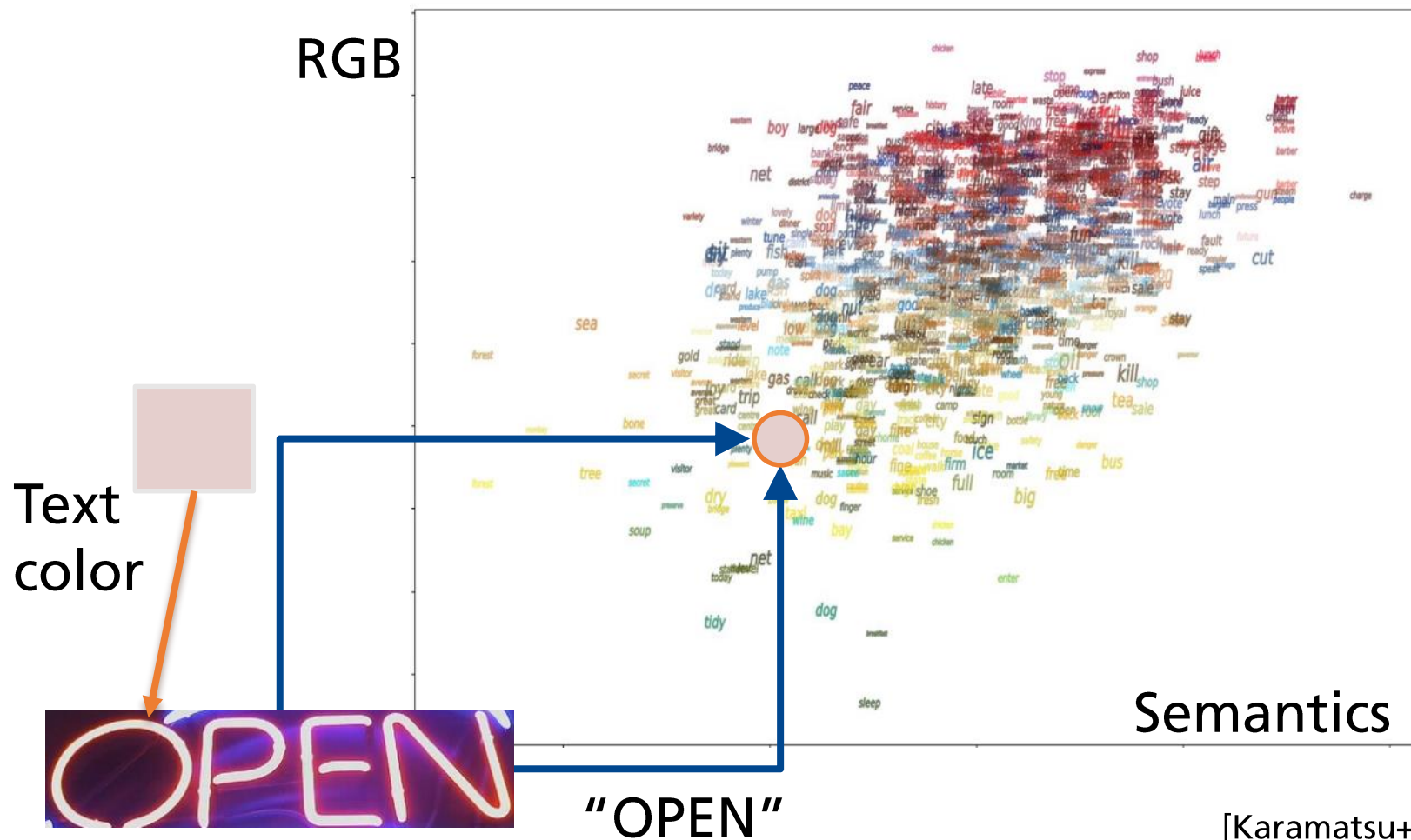
Decorative
Calligraphic
Sanserif-Round
Sanserif-Modern
Sanserif-Retro
Sanserif-Basic
Serif-Modern
Serif-Retro
Serif-Basic



Q. Why do we design various fonts?

Correlation between **font color** and **word semantics** seems also interesting

- Two-dimensional visualization of the correlation



Scientific topic #2

Interaction between **text** and **object**



Text as an object label

Two types of texts around us: *Message and Label*

• Message

- is texts for transmitting some information which is **independent of the object or the place** that the text are attached



• Label

- is texts for **detailing the object or the place** that the texts are attached



naturalsobsessed.blogspot.com



www.thomasmorris.co.uk



www.insidehousing.co.uk

Q. How do label texts and objects interact to each other?

- Possible research topics
 - Can we utilize (label) texts for fine-grained object recognition or scene classification?

Where is it?



[Frinken+, ICPR2014]

What shop is this?



[Movshovitz-Attias+, CVPR2015]

DJ SUBS Breakfast

Starbucks Coffee

Starbucks Coffee



[Karaoglu+, IEEE TIP2016]

What kind of advertisement?



[Dey+, arXiv2019]

Q. How do label texts and objects interact to each other?

- Possible research topics (cont'd)

- What kind of objects are detailed by label texts?

- Bottle, bus, building, foods in the supermarket,



naturalsobsessed.blogspot.com

- A preliminary study on Open Images v4 (1.7 million images)

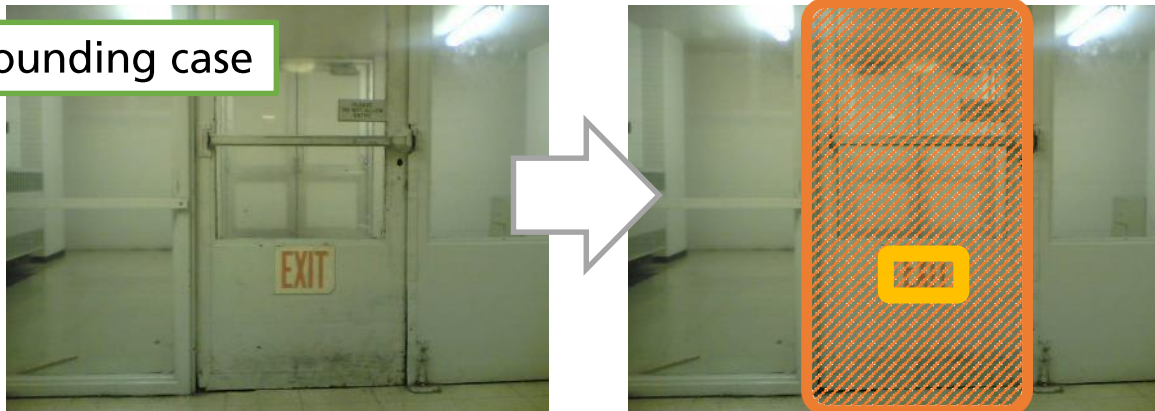
Object often with texts		Popular co-occurrence of object and text		
rank	object	rank	object	text
1	ambulance	1	car	police
2	calculator	2	bus	bus
3	scoreboard	3	man	army
4	poster	4	bus	school
5	scale	5	book	one
6	ruler	6	book	land
7	envelope	7	book	new
8	fax	8	tree	park
9	bus	9	person	army
10	cream	10	book	book

[Takeshita+, unpublished]

Q. How do label texts and objects interact to each other?

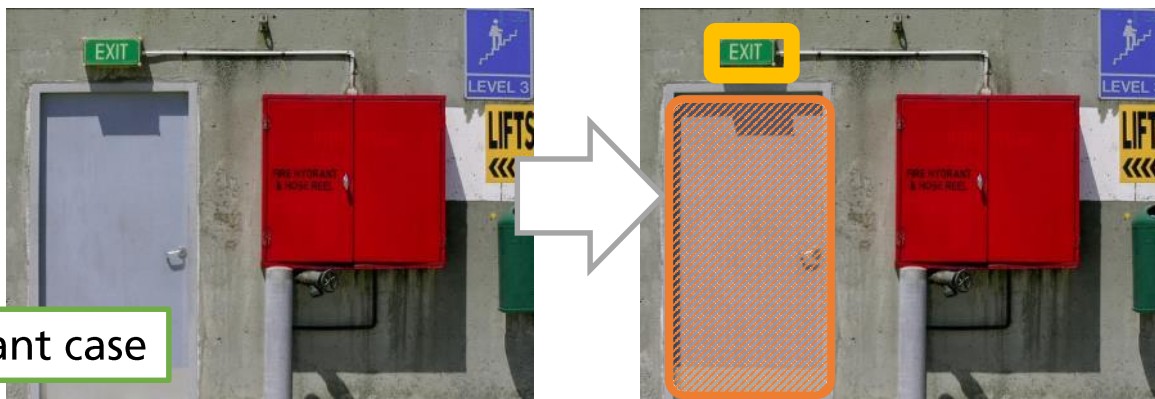
- Possible research topics (cont'd)
 - Can we determine the area where a text details?

Surrounding case



jm3 on Flickr "exit" (Flickr CreativeCommons, CC BY-SA 2.0)

Distant case



s2art "One of Two #2" (Flickr CreativeCommons, CC BY-SA 2.0)

Q. How do label texts and objects interact to each other?

- Possible research topics (cont'd)
 - Can we realize more detailed *image-captioning* by using scene texts?



"A street sign on a pole on a street."

standard image-captioning
(w/o scene text info)



"A **motel** sign is on the side of the road."

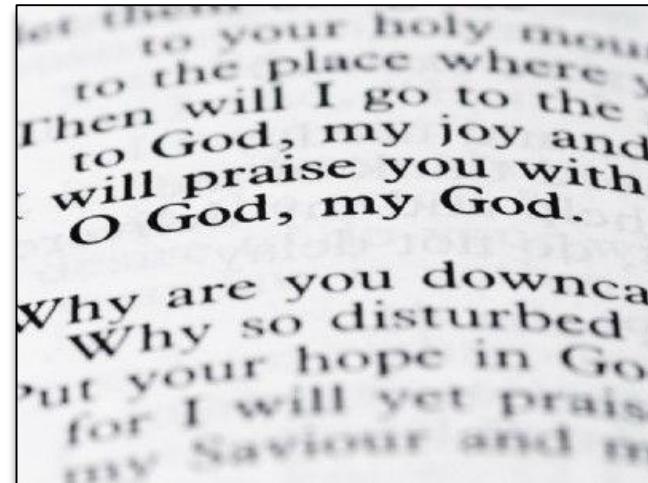
detailed image-captioning
(with scene text info)

Q. How does label texts and objects interact to each other?

- Possible research topics (cont'd)
 - Can texts (or text lines) give hints to recover **3-D shape** of the object that they are attached?



flat, but non-frontal surface



curved surface

Q. How does label texts and objects interact to each other?

- Possible research topics (cont'd)
 - Finally, can we distinguish label texts from messages?



message

label



naturalsobsessed.blogspot.com



Scientific topic #3

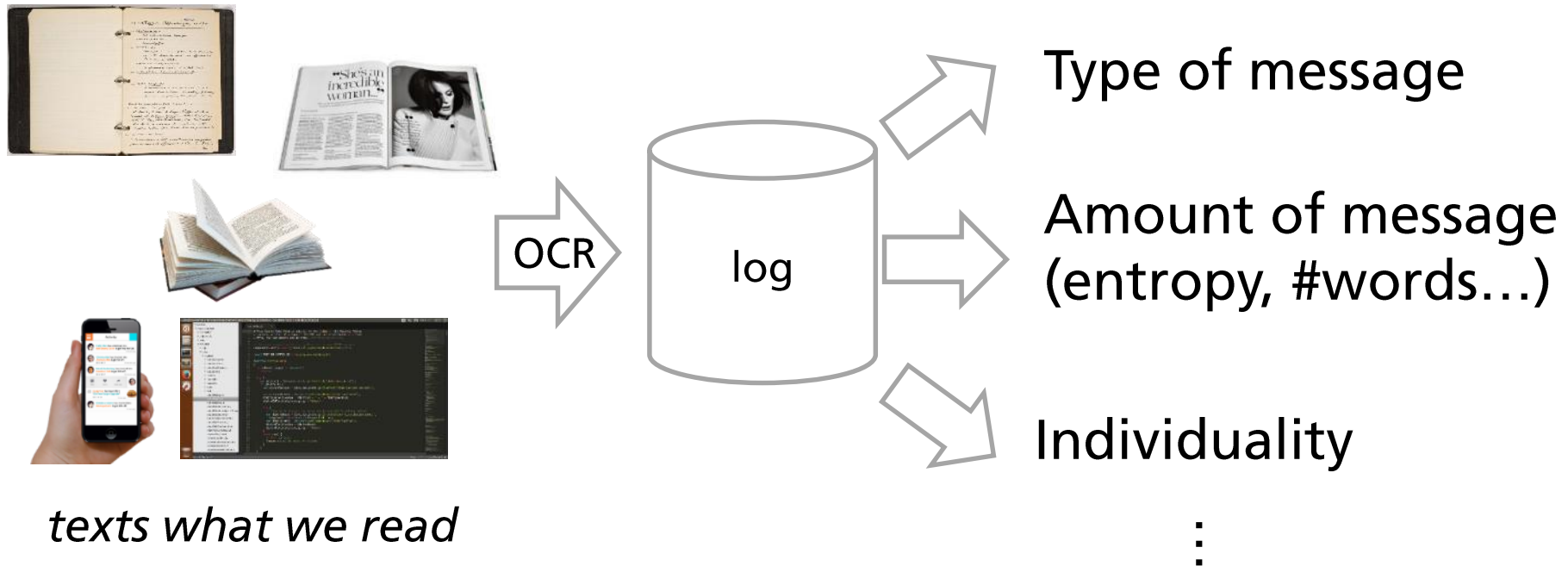
Interaction between **text** and **human**

The text in this
box is not true



Q. What kind of messages are we receiving from texts?

- Possible research topics
 - Logging and analyzing texts what we read everyday



Q. What kind of messages are we getting from texts?

A personal trial to visualize the message around us (1/2)



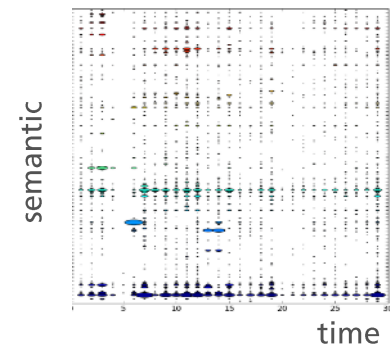
video images
(6-hour long)

パリッと
ふ？っと
昔ながらのおい？？を
真心こめて
焼きました。
からつ
バーガー
なるべく
あたたかいうちに
お？と
あがり
下さい、
MENU
スペシャルバーガー ¥460
●チーズ●たまご●ハム●パティ●
エッグバーガー ¥340
●たまご●パティ●
ハムエッグバーガー ¥340
●ハム●たまご●
チーズバーガー ¥340
●チーズ●？ティ●
ハンバーガー ¥280
●？ティ●
DRINK
コーヒー ¥220
ミルク ¥220
コーラ ¥130
ウーロン茶 ¥130
果汁オレンジ ¥130
とコーヒー

words captured
at each frame

Semantic
quantization
into 300 categories
by *WordNet*

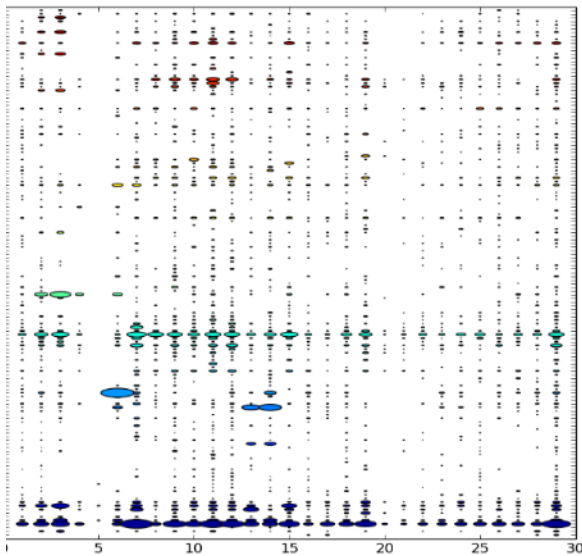
visualization as
a 2D-histogram



Q. What kind of messages are we getting from texts?

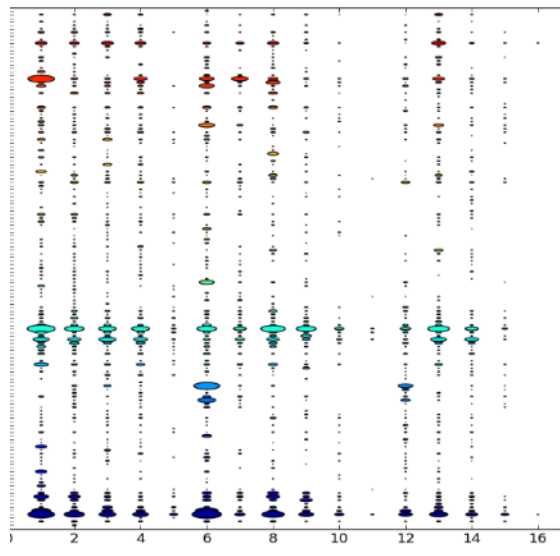
A personal trial to visualize the message around us (2/2)

Trip to a village



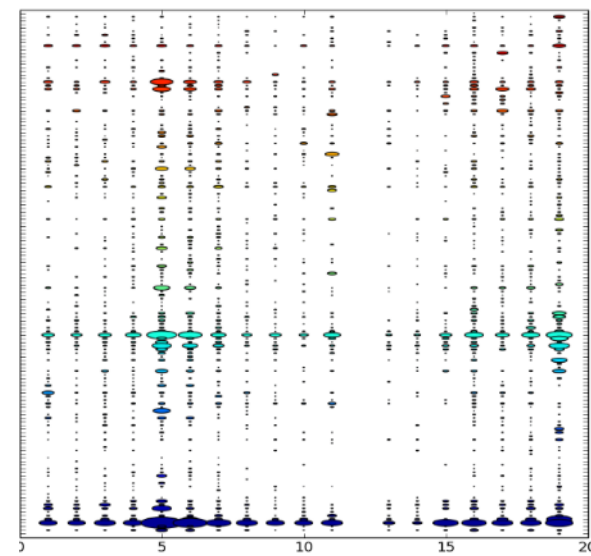
time (6 hours)

University campus



time (6 hours)

City center



time (6 hours)

Different scenes give different histograms

Q. How is **our life affected** by text?

- It seems clear that **our life is controlled** by texts around us... but how much?



Are they drinkable?



Where is my destination?



<http://kaigai-matome.net/archives/35546295.html>

- Actually, I could arrive here by the help of a lot of texts

Q. How is **our life** affected by text?

- Possible research topics
 - Compare our behavior **with or without** texts
 - Especially, label texts



[Nakamura+, "Scene text eraser" ICDAR2017]

- **Alert and caution** texts detection

- They are "traffic signs" for human-beings' daily life



Tony Webster "Lake Victoria Water Access Signs" (Flickr CreativeCommons, CC BY 2.0)

Q. How is **our life affected** by text?

- Possible research topics (cont'd)

- Evaluating “noisiness” of scene texts



<http://kaigai-matome.net/archives/35546295.html>

- Evaluating “visual saliency” of scene texts

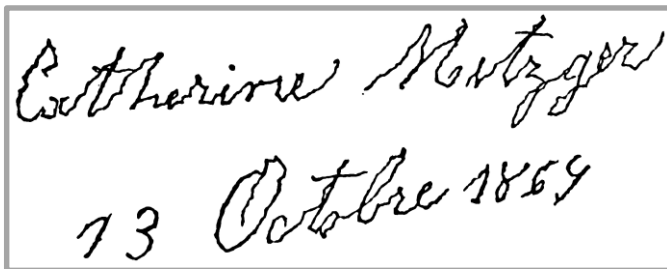
- How do characters visually appeal themselves to us?
- When are texts really salient? And why?



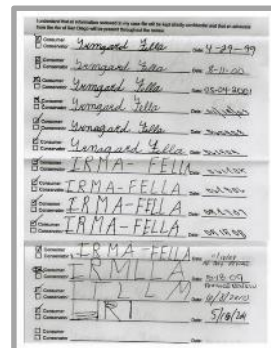
[Shahab, Shafait, Dengel, Uchida, DAS2012]

Q. How do handwritten patterns represent **writer's personal conditions**?

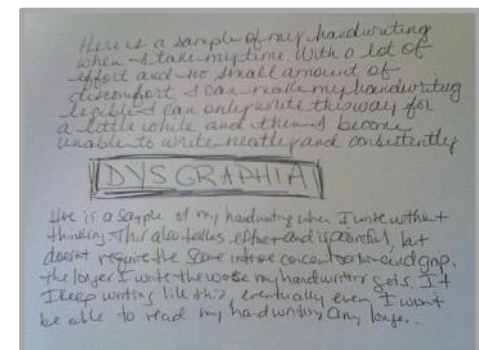
- Possible research topics
 - How does the **age** affect handwritings?
 - How does the **gender** affect?
 - How does the **temper** affect?
 - ≈ **emotion estimation** from speech
 - Relax, stressful, confused, sleepy ...
 - How does the **disease** affect?
 - Parkinsonian, Alzheimer, Dysgraphia



[https://en.wikipedia.org/wiki/Micrographia_\(handwriting\)](https://en.wikipedia.org/wiki/Micrographia_(handwriting))



http://media.npr.org/assets/img/2014/08/20/irma_signatures_custom_c1397421ac931c8541ce07ca39d6ad96e85f09133-c85.jpeg



<https://en.wikipedia.org/wiki/Dysgraphia>

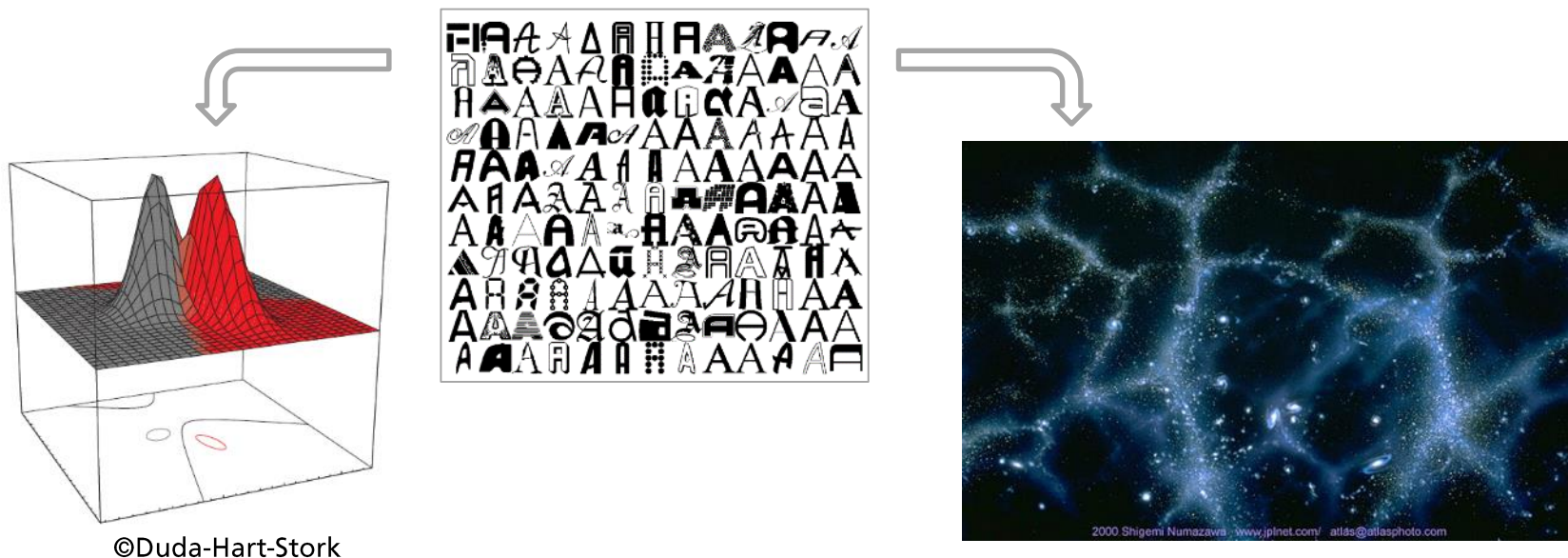
Scientific topic #4

Distribution of character patterns

Nobody knows ...

Q. How are character patterns **distributed**?

- Nobody knows real distribution of patterns
 - Gaussian ? Gaussian Mixture? Really!?
 - Or another distribution like “void structure of the universe”?

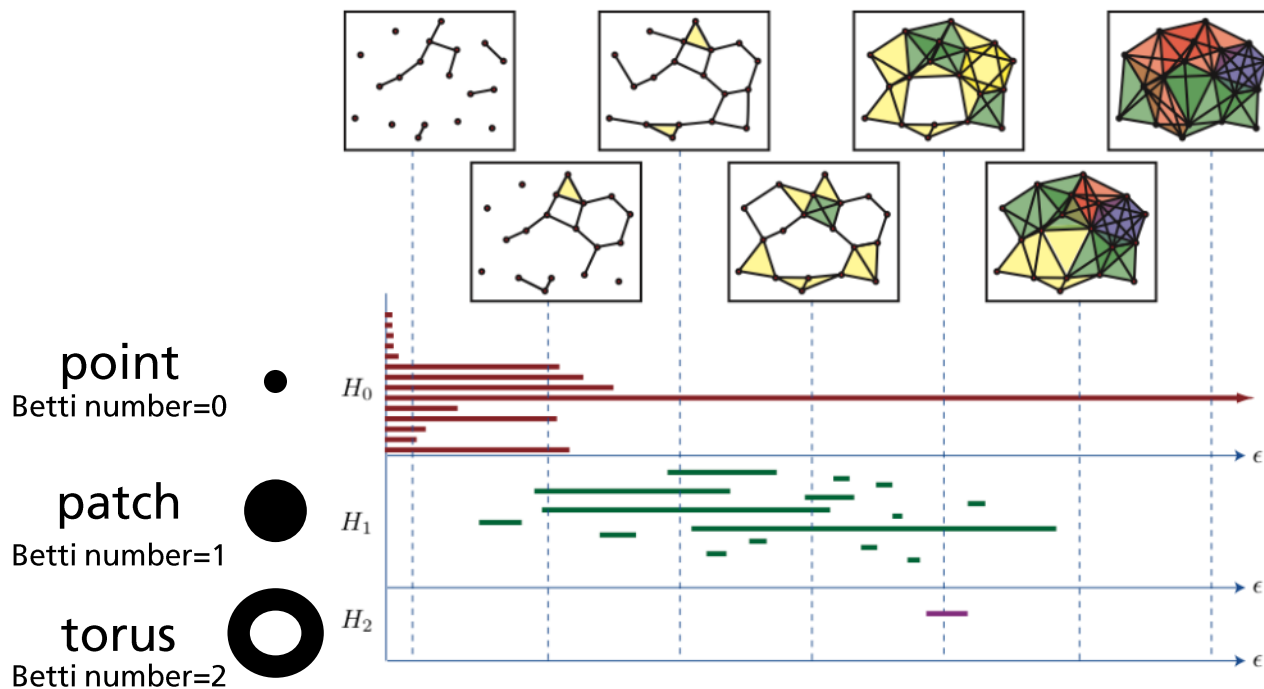


- Anyway, knowing the real distribution is crucial for DAR!
 - Since character images can be tiny, the dimensionality of distribution space is smaller than other images

Q. How are character patterns **distributed**?

- Possible research topics
 - Applying “**topological data analysis (TDA)**” to a large character image samples
 - e.g., “**Persistence**” analysis

Of course, traditional non-parametric data analysis techniques, such as network analysis, are also useful

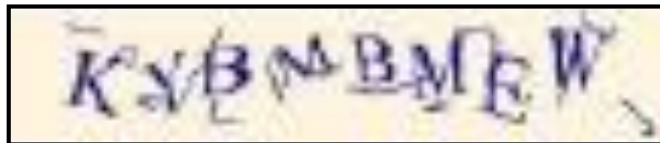


It is similar to apply “dilation” operation to a point cloud

Q. Where is the border between character and non-character?

の 川 三 じ 出 天 さ 升
 回 さ、 こ 升 川 川 の こ
 走 式、 の 式 圭 こ 回 三 川
 式、 疾 の の 入 回 こ じ
 廷 屈 の さ 三 ° の 味
 味 の 圭 の さ の じ 屈

Fake character

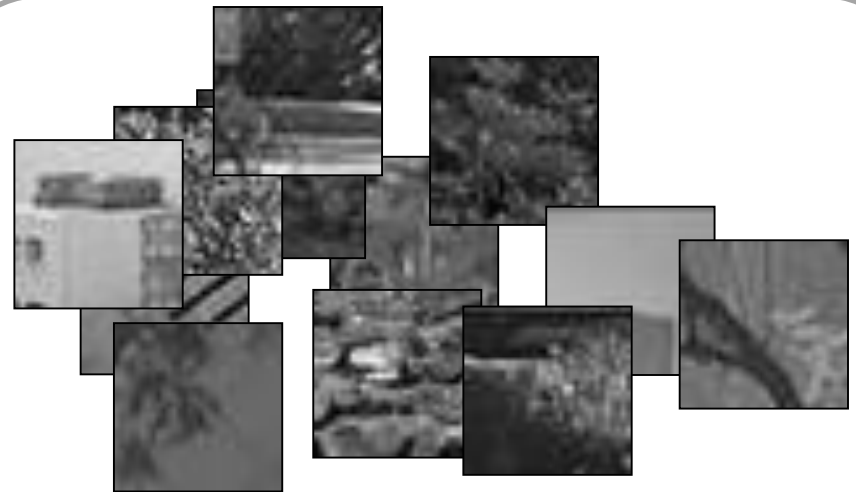


LDVIGUSE

Captcha



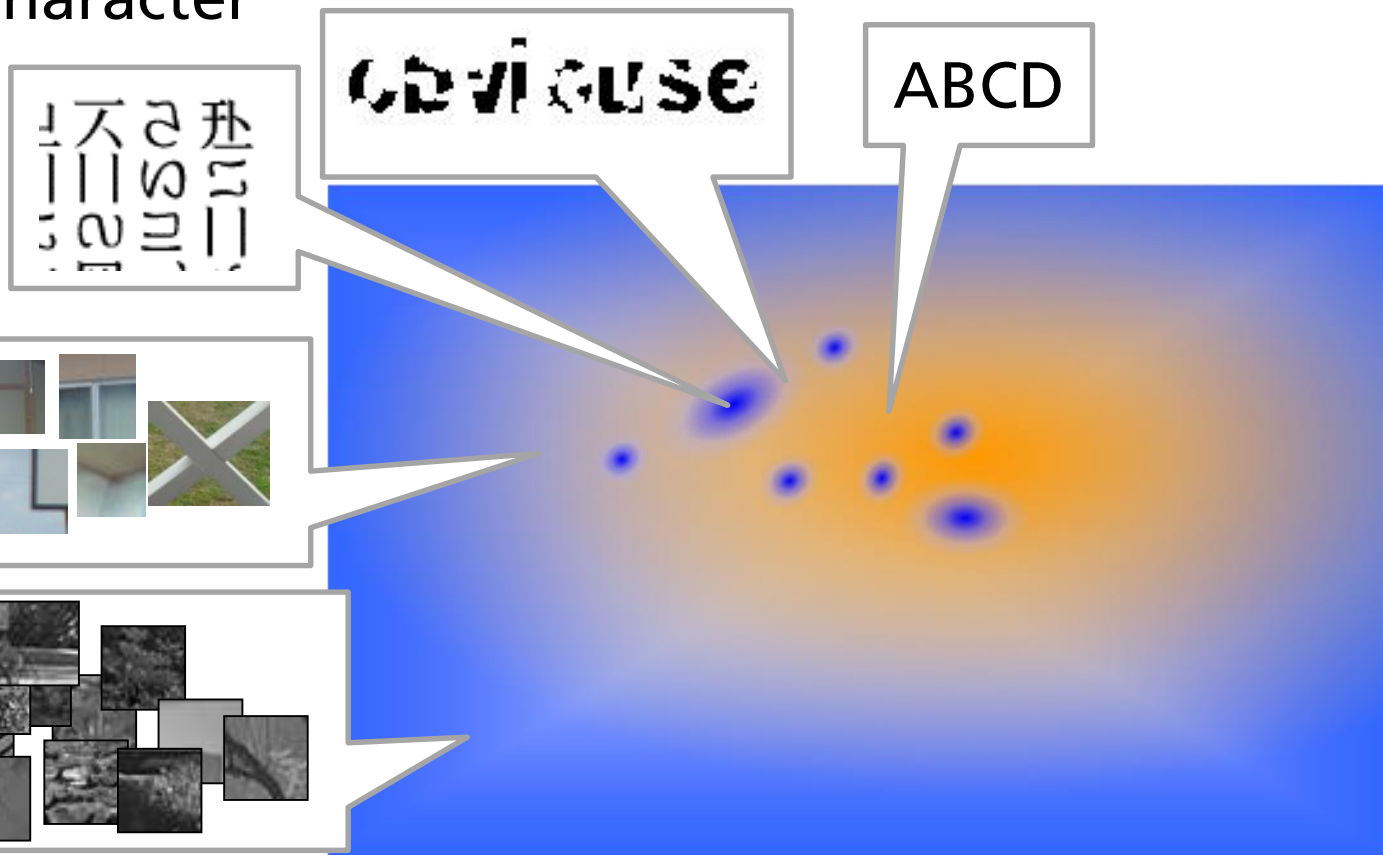
Confusing non-character



Non-character

Q. Where is the border between character and non-character?

- Possible research topics
 - Explore the border between character and non-character



Scientific topic #5

Relationship to
semantic analysis

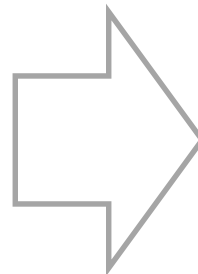
Characters (word and sentence) are not just image patterns

- They represent a specific meaning (**semantics**)



Mike "beware of dog" (Frickr CreativeCommons, CC BY-SA 2.0)

Not just three letters
"D"+"O"+"G"



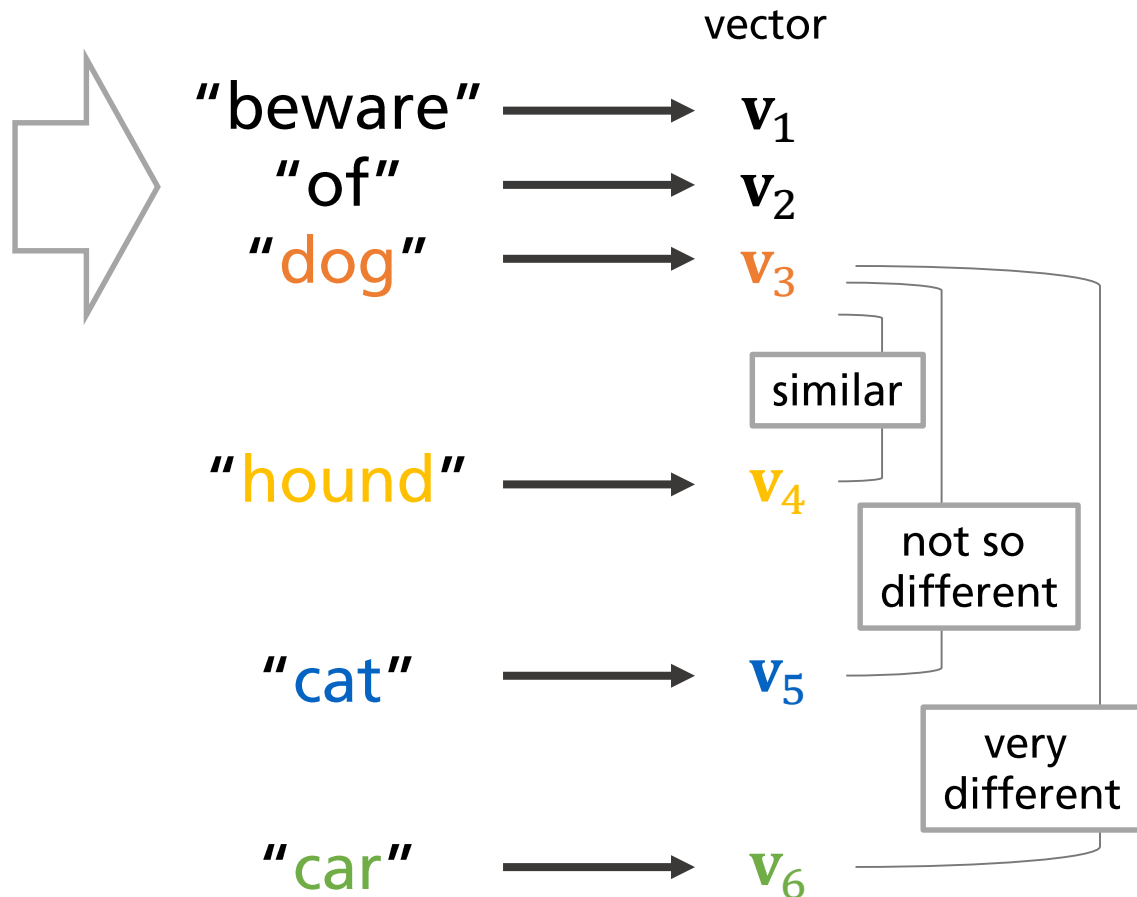
It means



We can represent the meaning as a vector by “word-embedding” techniques

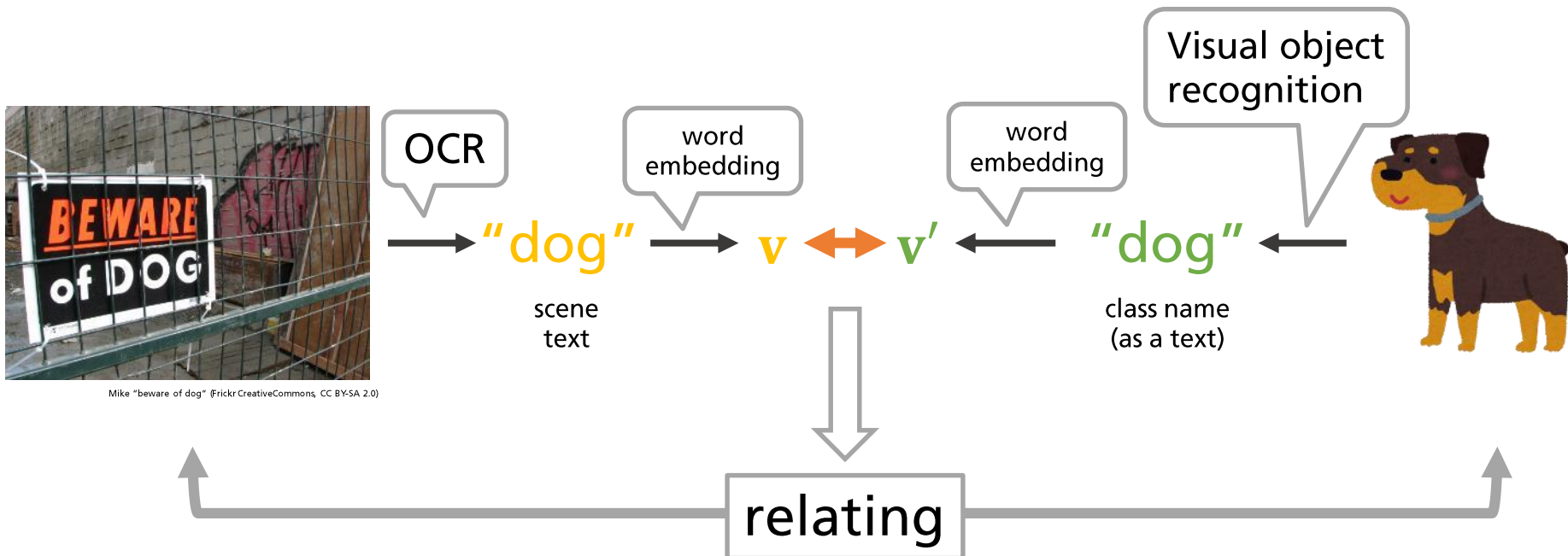


Mike “beware of dog” (Frickr CreativeCommons, CC BY-SA 2.0)



Q. Can we combine CV/PR and NLP by DAR?

- Possible research topics
 - Relating visual object and scene text by sharing vector representation
 - → maybe useful for few-shot / zero-shot learning



Q. Can we help NLP by DAR?

- Possible research topics
 - Enhanced word-embedding
 - Do font shape or color enhance the difference of antonyms?

The **color** will make the meaning of "hot" more **hot**

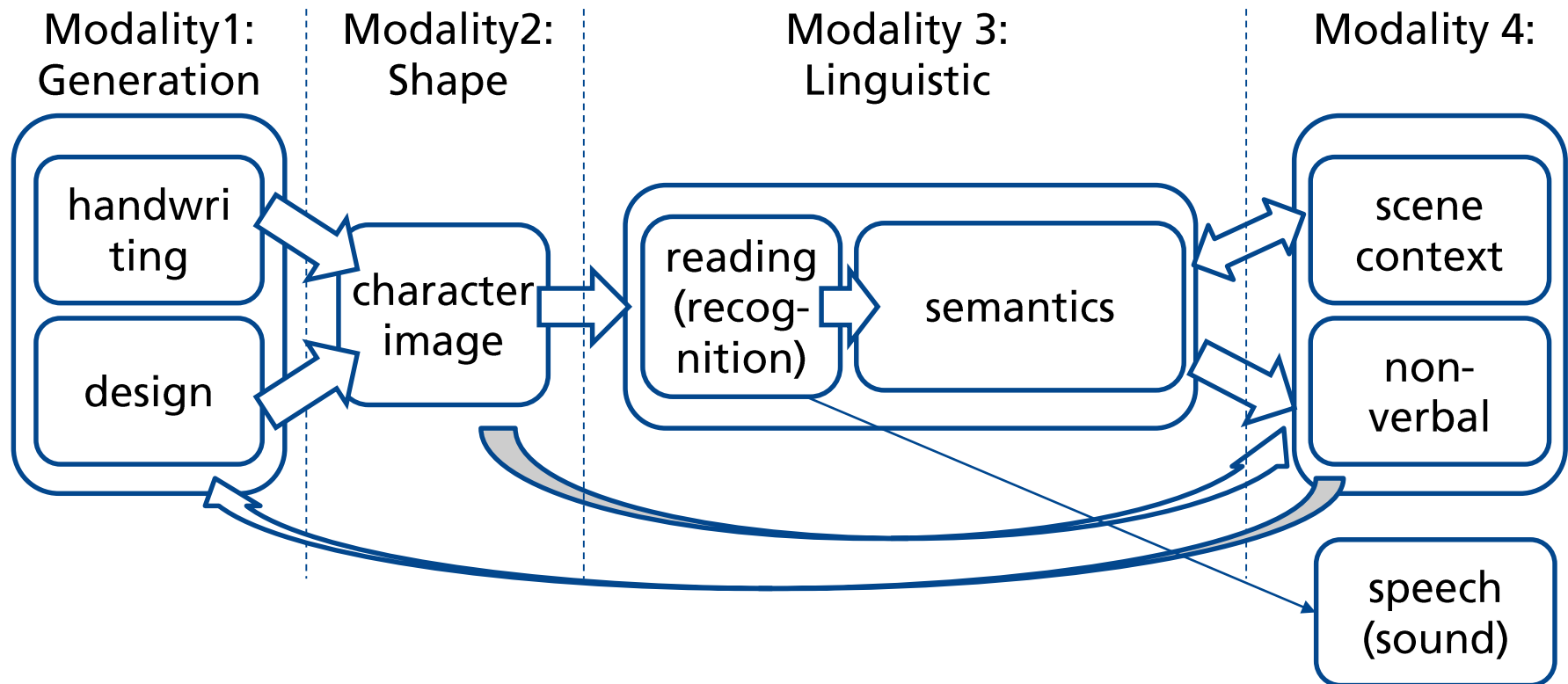


SonnyandSandy "Hot and Cold in St. Clair, Missouri" (Flickr CreativeCommons, CC BY-BY-NC-ND 2.0)

Conclusion

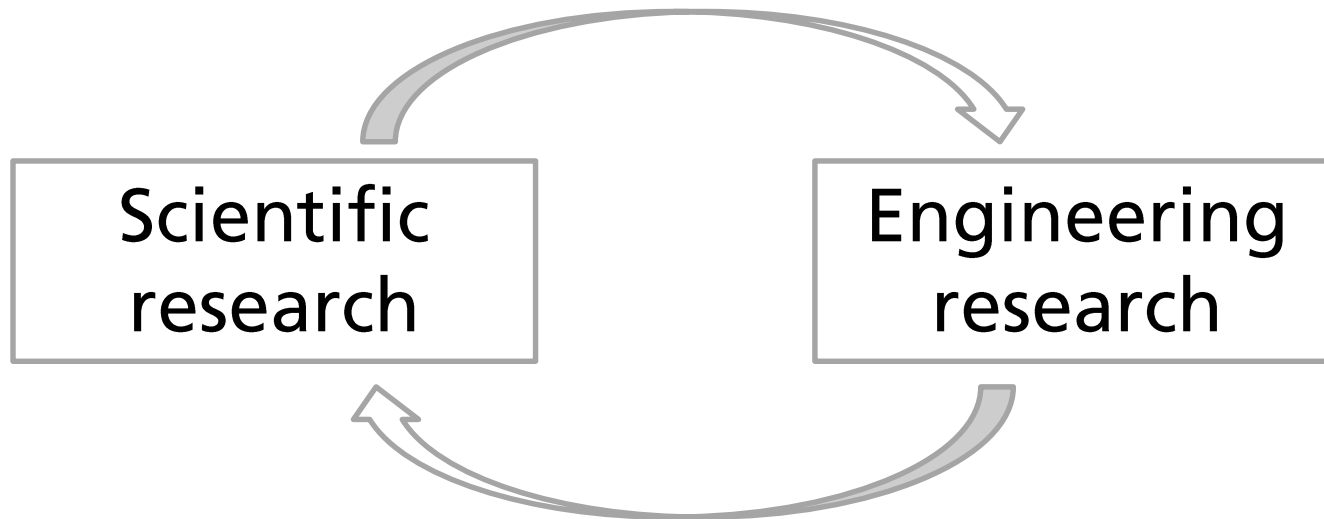
Conclusion: Let's have a **wider view** about DAR

- We can play with the **multi-modality** of characters and texts **by the SOTA OCR methods!**



Scientific research is useless?

- No!



The last message...

**DEAR YOUNG RESEARCHERS,
PLEASE GO BEYOND 100%**



**... and please do NOT become an accuracist,
parameter-tuner, and libraholic!**