Oil painting by Abel Grimmer (1570–1619)

The Babel of Software Development: Linguistic Diversity in Open Source

Bogdan Vasilescu Alexander Serebrenik Mark van den Brand
Eindhoven University of Technology
@b_vasilescu, @aserebrenik, @MarkvandenBrand

SocInfo 2013, Kyoto, Japan

Warning

Equations inside

OSS Communities

Highly interactive

Decentralized

Heterogeneous

Self-directed

Knowledge-intensive

OSS Communities

Highly interactive

Decentralized

Heterogeneous

Self-directed

Knowledge-intensive

OSS Communities

Developers donate their knowledge
for the benefit of the community

Highly interactive

Decentralized

Heterogeneous

Self-directed

Knowledge-intensive

OSS Communities

Different skill sets and skill levels

(Giuri *et al.*, 2004)

Different activities

(Vasilescu *et al.*, 2013)

Mix of novices and experts

(Dabbish *et al.*, 2012)

Highly interactive

Decentralized

Heterogeneous

Self-directed

Knowledge-intensive

OSS Communities

Different skill sets and skill levels

(Giuri *et al.*, 2004)

Different activities

(Vasilescu *et al.*, 2013)

Mix of novices and experts

(Dabbish *et al.*, 2012)

Highly interactive

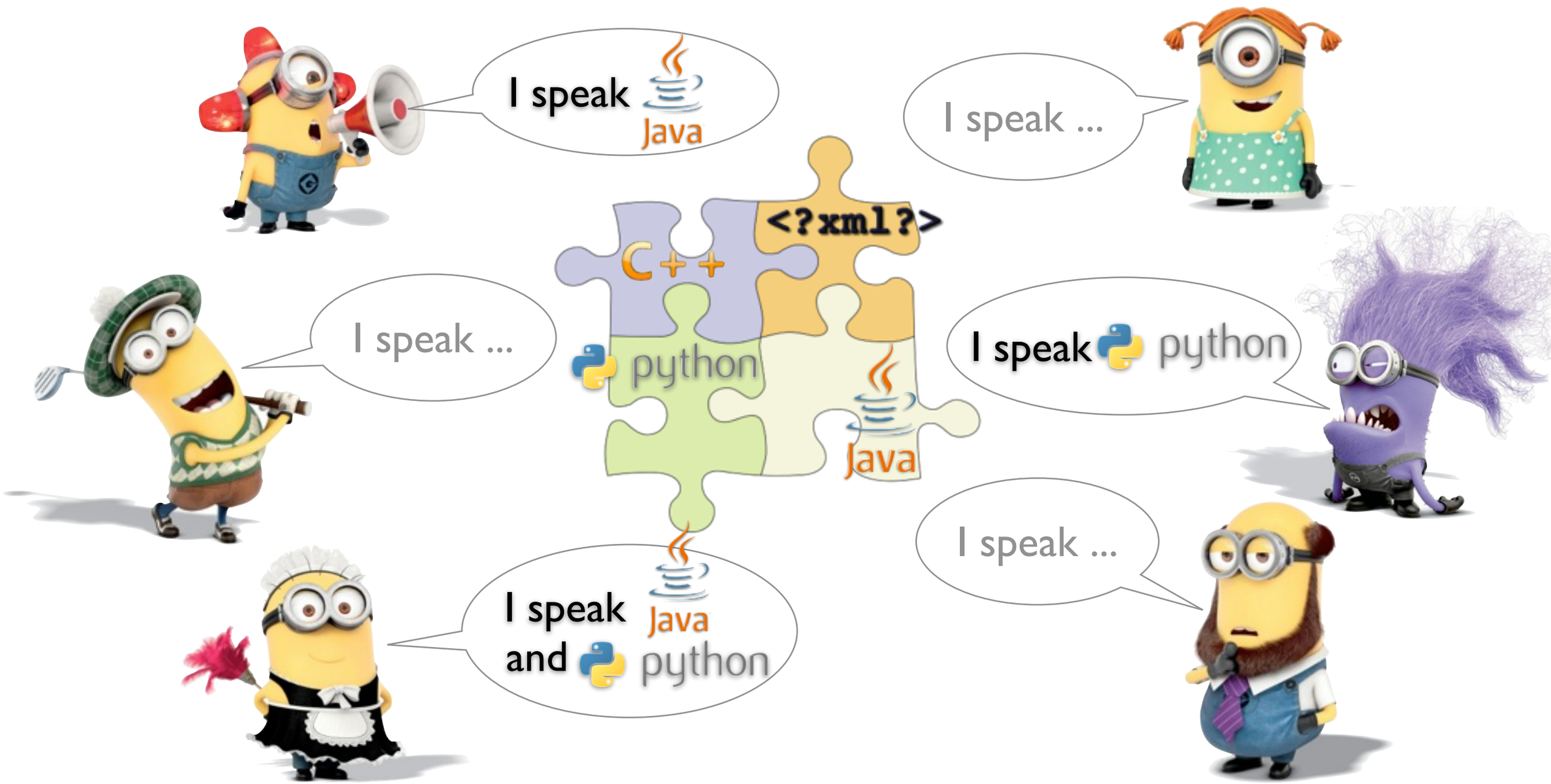
Decentralized

Heterogeneous

Self-directed

Knowledge-intensive

Knowledge of programming languages



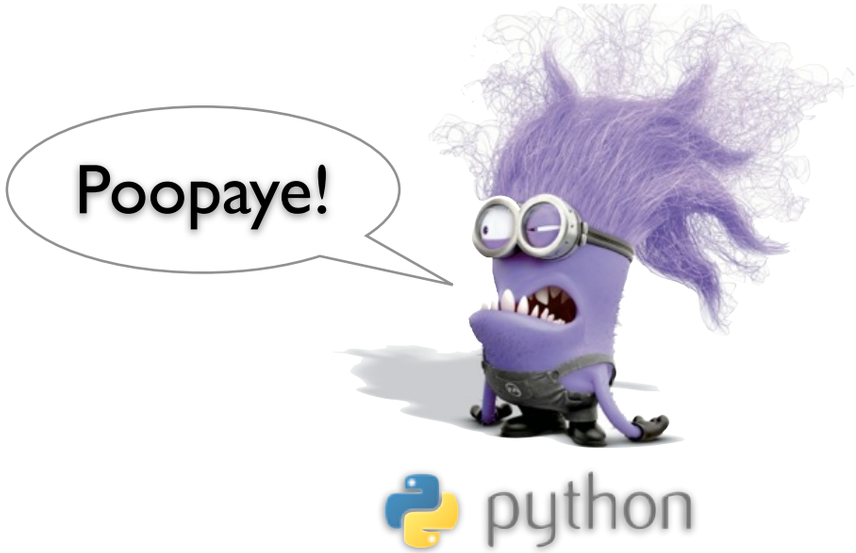
High turnover

What happens when Purple Minion leaves the community?

Does knowledge of python disappear?

Is the community at risk?

Maintain or migrate legacy code?



High turnover

What happens when Purple Minion leaves the community?

Does knowledge of python disappear?

Is the community at risk?

Maintain or migrate legacy code?

Intuitively healthier

This talk: first steps

How to quantify this risk
of not finding developers with knowledge
of certain programming languages?

Idea

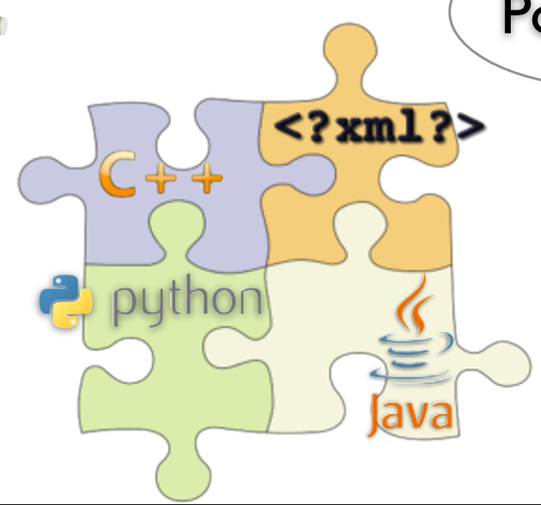
Who else "speaks" python in the community?

Hard to assess
(recall Decentralized, Self-directed)

Hard to maintain information
(recall High Turnover)

Less suitable for real-time
health monitoring

What if nobody?



Idea

has worked on

Who else "~~speaks~~" python in the community?

Not sufficient

Specialization

(Posnett *et al.*, 2013)

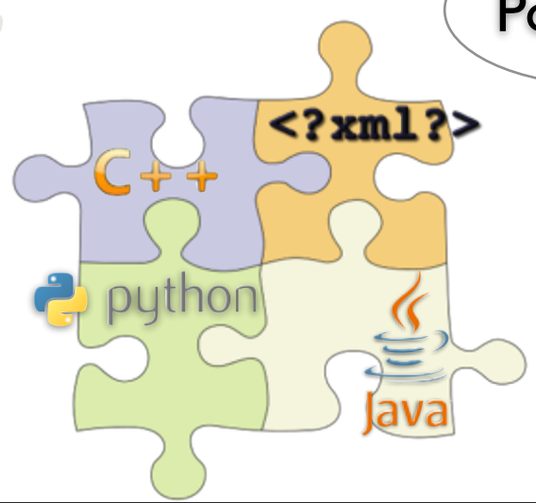
(Vasilescu *et al.*, 2013)

Territoriality

(Robles *et al.*, 2006)

Besides,

What if nobody?



python

Idea

~~has worked on~~

Who else "~~speaks~~" python in the community?

might understand

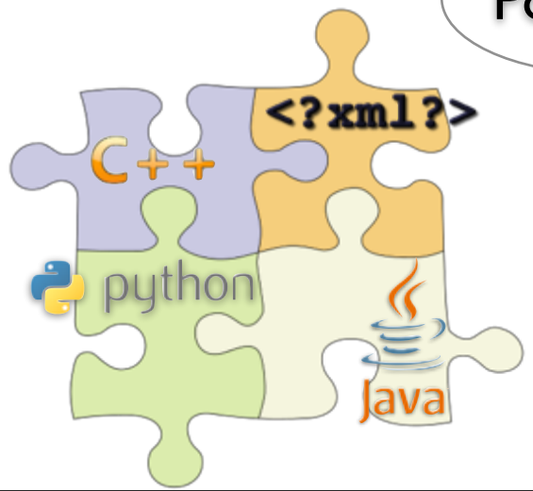
Better, but similar drawbacks as

Who else "speaks" python in the community?

Does not answer

"How hard is it in general to find replacement COBOL developers?"

I have some experience with BOO. I might understand python



Poopaye!

python

Idea

~~has worked on~~

Who else "~~speaks~~" python in the community?

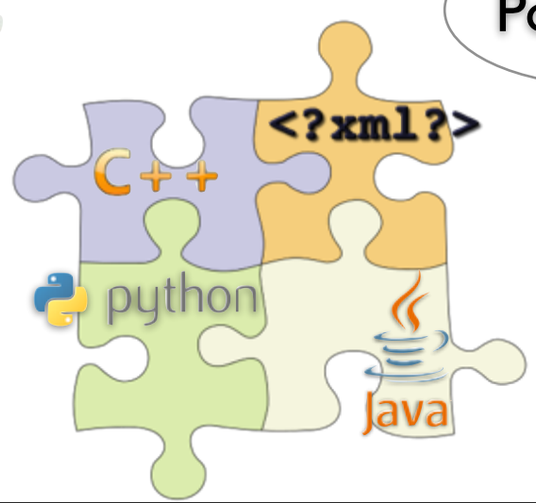
might understand

Mine expertise from history of contributions

+

Approximate what else they might understand: universal measure of intelligibility of programming languages

~~I have some experience with BOO. I might understand python~~



Poopaye!

python

Ingredients

Linguistic diversity

(natural languages)

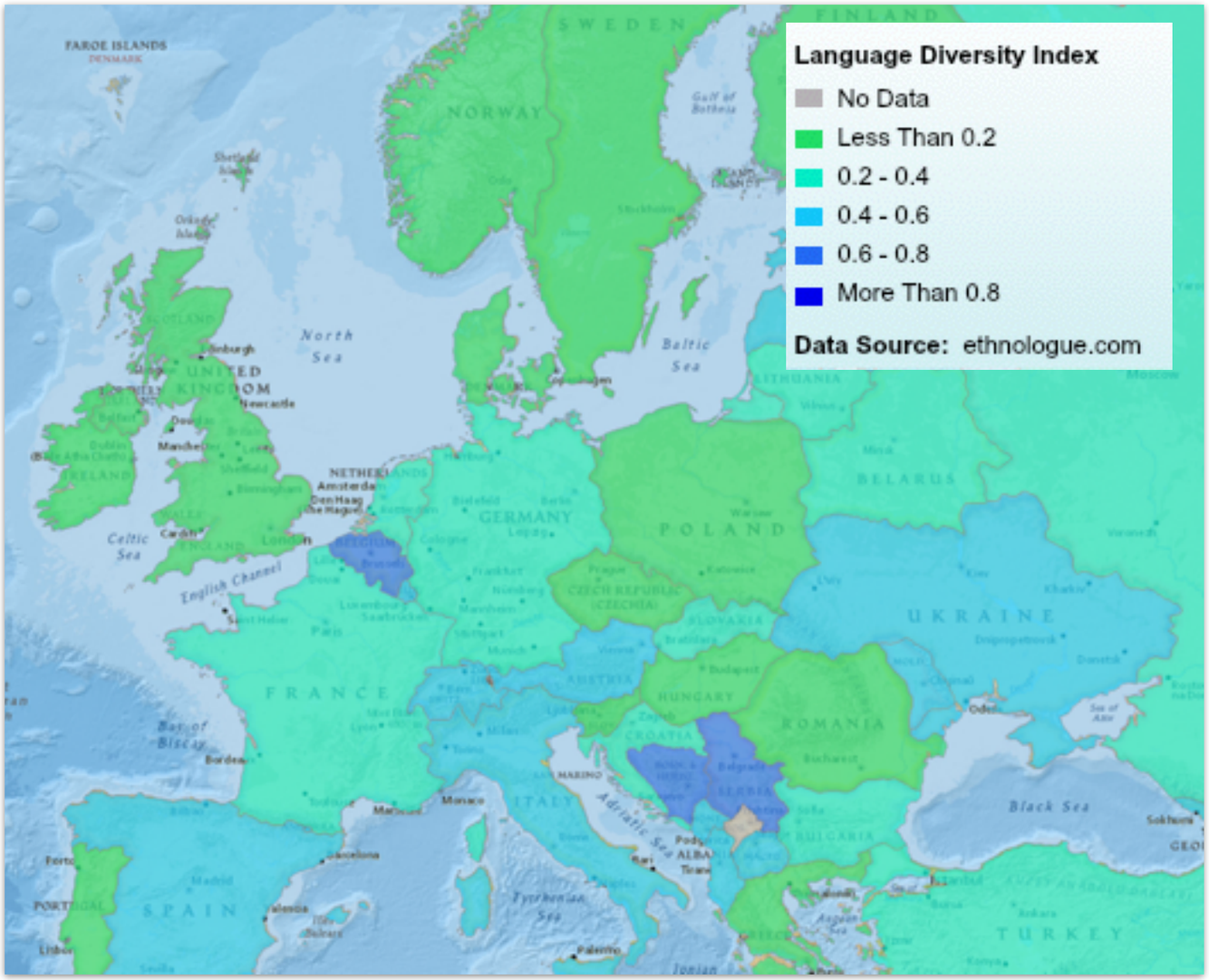
crowdsourced knowledge

stackoverflow.com

Linguistic diversity

Greenberg (1956)

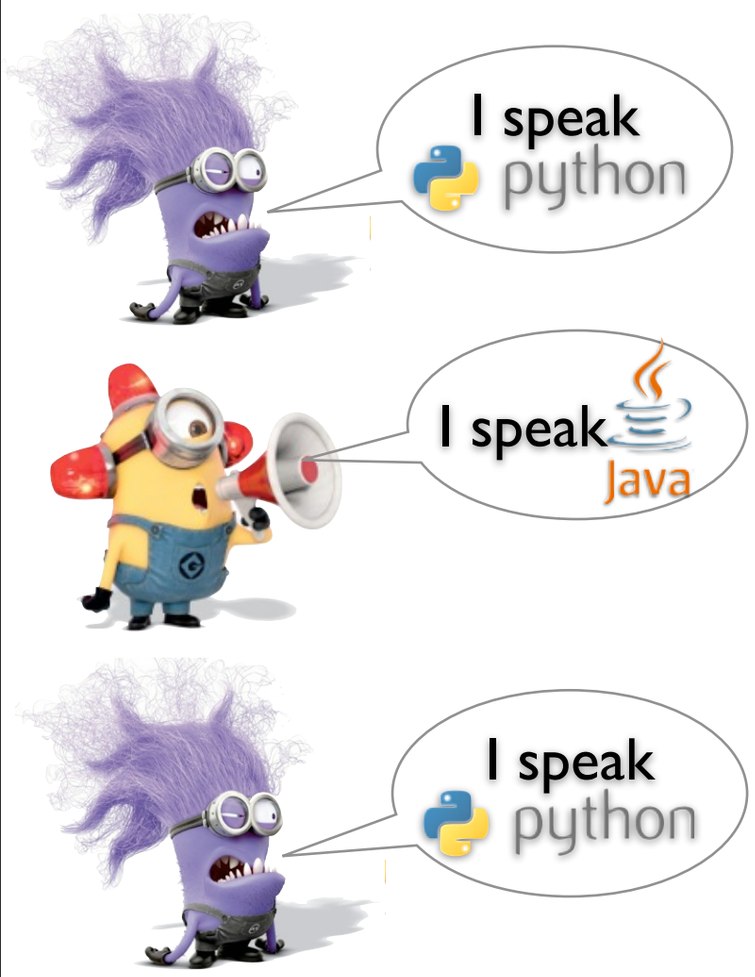
probability that two random individuals do not understand each other



http://education.nationalgeographic.com/education/mapping/interactive-map/?ar_a=1&ls=840007%26f%3D491%26t%3D1%26lg%3D5%26b%3D0%26bbox

Linguistic diversity

probability that two random individuals do not understand each other



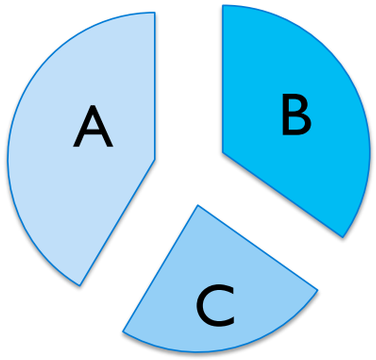
Simple model

- everyone speaks exactly one language
- languages are independent

$$A = 1 - \sum_{\ell \in L} p_\ell^2$$

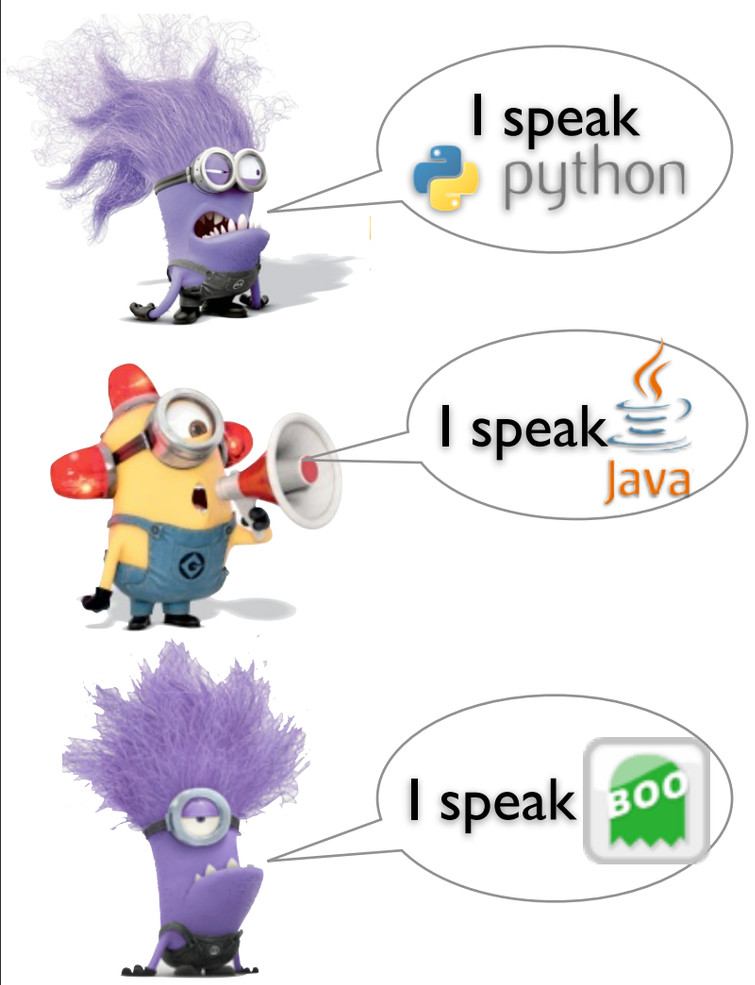
$$p_\ell = \frac{|S_\ell|}{|P|}$$

$$L = \{A, B, C\} \Rightarrow P(L) = \{A, B, C\}$$



Linguistic diversity

probability that two random individuals do not understand each other



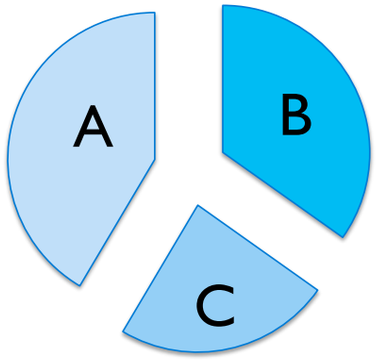
Related languages model

- everyone speaks exactly one language
- languages are (partly) mutually intelligible

$$B = 1 - \sum_{\ell, m \in L} p_\ell p_m \cdot mi(\ell, m) \qquad p_\ell = \frac{|S_\ell|}{|P|}$$

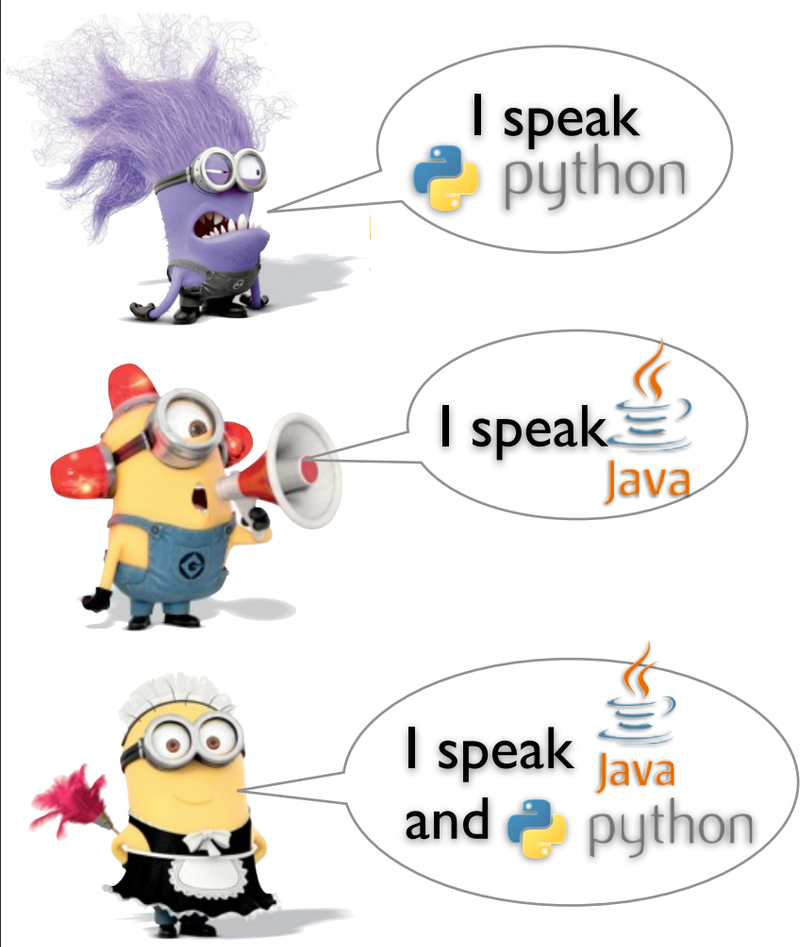
$$0 \le mi(\ell, m) \le 1 \qquad mi(\ell, \ell) = 1$$

$$L = \{A, B, C\} \Rightarrow P(L) = \{A, B, C\}$$



Linguistic diversity

probability that two random individuals do not understand each other

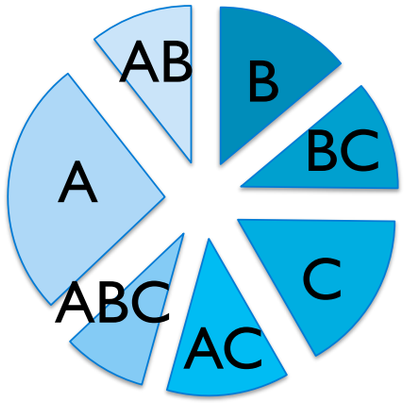


Polyglot related languages model

- everyone speaks at least one language
- languages are (partly) mutually intelligible

$$F = 1 - \sum_{s,t \in P(L)} p_s p_t \cdot \frac{\sum_{\ell \in s, m \in t} mi(\ell, m)}{|s| \cdot |t|} \qquad p_s = \frac{|S_s|}{|P|}$$

$$L = \{A, B, C\} \Rightarrow P(L) = \{A, B, C, AB, AC, BC, ABC\}$$



Risk

of not finding developers that "speak" a programming language

Greenberg (1956)

Linguistic diversity

$$F = 1 - \sum_{s,t \in P(L)} p_s p_t \cdot \frac{\sum_{\ell \in s, m \in t} mi(\ell, m)}{|s| \cdot |t|}$$

Our measure

$$risk(\ell) = 1 - \sum_{s \in P(L)} p_s \cdot \max_{k \in s} mi_\ell(k)$$

Risk

of not finding developers that "speak" a programming language

Greenberg (1956)

$$F = 1 - \sum_{s,t \in P(L)} p_s p_t \cdot \frac{\sum_{\ell \in s, m \in t} mi(\ell, m)}{|s| \cdot |t|}$$

Linguistic diversity

Aggregate measure

Our measure

$$risk(\ell) = 1 - \sum_{s \in P(L)} p_s \cdot \max_{k \in s} mi_\ell(k)$$

one language

Risk

of not finding developers that "speak" a programming language

Greenberg (1956)

Linguistic diversity

$$F = 1 - \sum_{s,t \in P(L)} p_s p_t \cdot \frac{\sum_{\ell \in s, m \in t} mi(\ell, m)}{|s| \cdot |t|}$$

If polyglot, equally probably to speak any of the languages

Our measure

$$risk(\ell) = 1 - \sum_{s \in P(L)} p_s \cdot \max_{k \in s} mi_\ell(k)$$

If polyglot, the language most intelligible to ℓ matters the most

Risk

of not finding developers that "speak" a programming language

Greenberg (1956)

$$F = 1 - \sum_{s,t \in P(L)} p_s p_t \cdot \frac{\sum_{\ell \in s, m \in t} mi(\ell, m)}{|s| \cdot |t|}$$

Linguistic diversity

Symmetric

Our measure

$$risk(\ell) = 1 - \sum_{s \in P(L)} p_s \cdot \max_{k \in s} mi_\ell(k)$$

Asymmetric

"Swedes have more difficulty understanding Danish than Danes understanding Swedish"

(Moberg *et al.*, 2004)

Ingredients

Linguistic diversity
(natural languages)

crowdsourced knowledge

stackoverflow.com

5,931,622 questions

2,458,712 users

stackoverflow

Questions | Tags | Users | Badges | Unanswered

Tagged Questions

newest | frequent | votes | active | unanswered

58 votes
4 answers
15k views

Does python have an equivalent to Java Class.forName()?

I have the need to take a string argument and create an object of the class named in that string in Python. In Java, I would use Class.forName().newInstance(). Is there an equivalent in Python? ...

java python class instantiation

asked Jan 17 '09 at 8:10
Jason
479 ●1 ●6 ●8

117 votes
9 answers
4k views

Seeking clarification on apparent contradictions regarding weakly typed languages

I think I understand strong typing, but every time I look for examples for what is weak typing I end up finding examples of programming languages that simply coerce/convert types automatically. For ...

c# java python perl weakly-typed

asked Mar 29 '12 at 16:34
Edwin Dalorzo
13.2k ●2 ●22 ●47

54 votes
11 answers
32k views

How can I download all emails with attachments from Gmail?

How do I connect to Gmail and determine which messages have attachments? I then want to download each attachment, printing out the Subject: and From: for each message as I process it.

java python perl gmail

asked Dec 8 '08 at 3:57
anon

44 votes
7 answers
11k views

Which programming languages can I use on Android Dalvik?

In theory, Dalvik executes any virtual machine byte code, created for example with the compilers of AspectJ ColdFusion Clojure Groovy JavaFX Script JRuby Jython Rhino Scala Are there already ...

java python android scala dalvik

asked Jan 3 '10 at 11:35
mjn
17.5k ●8 ●68 ●173

5,931,622 questions

2,458,712 users

stackoverflow

Questions | Tags | Users | Badges | Unanswered

Tagged Questions

newest | frequent | votes | active | unanswered

58 votes
4 answers
15k views

Does python have an equivalent to Java Class.forName()?

I have the need to take a string argument and create an object of the class named in that string in Python. In Java, I would use Class.forName().newInstance(). Is there an equivalent in Python? ...

java python class instantiation

asked Jan 17 '09 at 8:10
Jason
479 •1 •6 •8

117 votes
9 answers
4k views

Seeking clarification on apparent contradictions regarding weakly typed languages

I think I understand strong typing, but every time I look for examples for what is weak typing I end up finding examples of programming languages that simply coerce/convert types automatically. For ...

c# java python perl weakly-typed

asked Mar 29 '12 at 16:34
Edwin Dalorzo
13.2k •2 •22 •47

54 votes
11 answers
32k views

How can I download all emails with attachments from Gmail?

How do I connect to Gmail and determine which messages have attachments? I then want to download each attachment, printing out the Subject: and From: for each message as I process it.

java python perl gmail

asked Dec 8 '08 at 3:57
anon

44 votes
7 answers
11k views

Which programming languages can I use on Android Dalvik?

In theory, Dalvik executes any virtual machine byte code, created for example with the compilers of AspectJ ColdFusion Clojure Groovy JavaFX Script JRuby Jython Rhino Scala Are there already ...

java python android scala dalvik

asked Jan 3 '10 at 11:35
mjn
17.5k •8 •68 •173

Users “collect” tags

Jon Skeet more info

network profile

615,287 reputation
223 3659 5102 badges

bio

website csharpindepth.com
location Reading, United Kingdom

visits

member for 5 years
seen 34 mins ago

summary | answers | questions | **tags** | badges | favorites | bounties | reputation | activity

4,348 Tags

votes | name

107k	c#	× 14991	4k	performance	× 386	2k	static	× 208	2k	eclipse	× 177
58k	java	× 7767	3k	vb.net	× 695	2k	inheritance	× 338	2k	events	× 214
38k	.net	× 4761	3k	list	× 408	2k	exception	× 278	2k	double	× 85
14k	linq	× 2395	3k	c#-4.0	× 441	2k	android	× 488	2k	sql	× 334
8k	string	× 783	3k	datetime	× 504	2k	floating-point	× 98	1k	char	× 46
8k	generics	× 1002	3k	collections	× 338	2k	interface	× 240	1k	decimal	× 70
6k	date	× 309	3k	lambda	× 376	2k	ienumerable	× 190	1k	unit-testing	× 248
6k	multithreading	× 888	3k	oop	× 311	2k	.net-3.5	× 204	1k	linq-to-sql	× 336
5k	timezone	× 199	3k	constructor	× 198	2k	winforms	× 412	1k	extension-methods	× 154
4k	asp.net	× 982	2k	delegates	× 329	2k	dictionary	× 233	1k	syntax	× 97
4k	arrays	× 611	2k	casting	× 284	2k	clr	× 112	1k	foreach	× 113
4k	reflection	× 583	2k	visual-studio	× 272	2k	algorithm	× 164	1k	loops	× 102
4k	xml	× 857	2k	enums	× 254	2k	types	× 194	1k	class	× 263

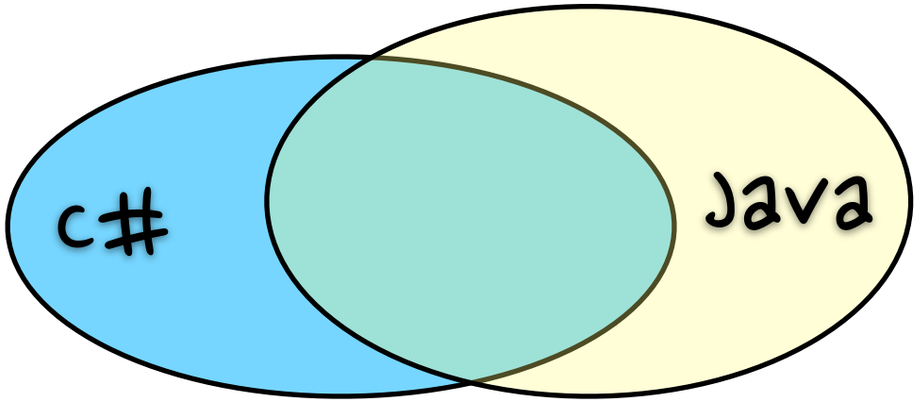
Mutual intelligibility of programming languages

- Jon Skeet: c#, Java, ASP.net, XML, ...
- Alexander Serebrenik: Prolog, SQL, c++, ...
- Bogdan Vasilescu: Python, ...

... > 400,000

Association rule mining

$$mi_{\ell}(k) = conf(\tau_k \Rightarrow \tau_{\ell}) = \frac{nBoth}{nLeft}$$



160 popular languages

Are this likely to speak <column>

	Asm	C	C++	Cobol	CSS	Groovy	HTML	Java	JavaScript	Perl	PHP	Shell	XML
Asm	100%	55%	54%	1%	15%	1%	23%	39%	28%	12%	28%	1%	18%
C	8%	100%	48%	0%	12%	1%	17%	31%	21%	8%	21%	0%	13%
C++	5%	32%	100%	0%	10%	1%	15%	26%	18%	6%	18%	0%	11%
COBOL	12%	35%	40%	100%	24%	3%	29%	48%	38%	17%	37%	1%	28%
CSS	2%	10%	13%	0%	100%	1%	61%	21%	54%	5%	39%	0%	16%
Groovy	3%	15%	18%	1%	17%	100%	26%	63%	32%	7%	23%	0%	26%
HTML	2%	11%	14%	0%	46%	1%	100%	25%	56%	5%	40%	0%	18%
Java	2%	12%	15%	0%	10%	2%	15%	100%	19%	4%	16%	0%	12%
JavaScript	2%	9%	11%	0%	25%	1%	35%	20%	100%	4%	31%	0%	13%
Perl	5%	25%	27%	1%	18%	2%	26%	31%	30%	100%	31%	1%	19%
PHP	2%	9%	11%	0%	19%	1%	26%	17%	33%	4%	100%	0%	12%
Shell	12%	34%	38%	1%	19%	3%	32%	43%	33%	24%	35%	100%	24%
XML	3%	14%	19%	0%	20%	2%	29%	34%	35%	7%	31%	0%	100%

People who speak <row>

http://www.win.tue.nl/~bvasiles/languages/list.html

160 popular languages

Are this likely to speak <column>

	Asm	C	C++	Cobol	CSS	Groovy	HTML	Java	JavaScript	Perl	PHP	Shell	XML
Asm	100%	55%	54%	1%	15%	1%	23%	39%	28%	12%	28%	1%	18%
C	8%	100%	48%	0%	12%	1%	17%	31%	21%	8%	21%	0%	13%
C++	5%	32%	100%	0%	10%	1%	15%	26%	18%	6%	18%	0%	11%
COBOL	12%	35%	40%	100%	24%	3%	29%	48%	38%	17%	37%	1%	28%
CSS	2%	10%	13%	0%	100%	1%	61%	21%	54%	5%	39%	0%	16%
Groovy	3%	15%	18%	1%	17%	100%	26%	63%	32%	7%	23%	0%	26%
HTML	2%	11%	14%	0%	46%	1%	100%	25%	56%	5%	40%	0%	18%
Java	2%	12%	15%	0%	10%	2%	15%	100%	19%	4%	16%	0%	12%
JavaScript	2%	9%	11%	0%	25%	1%	35%	20%	100%	4%	31%	0%	13%
Perl	5%	25%	27%	1%	18%	2%	26%	31%	30%	100%	31%	1%	19%
PHP	2%	9%	11%	0%	19%	1%	26%	17%	33%	4%	100%	0%	12%
Shell	12%	34%	38%	1%	19%	3%	32%	43%	33%	24%	35%	100%	24%
XML	3%	14%	19%	0%	20%	2%	29%	34%	35%	7%	31%	0%	100%

People who speak <row>

http://www.win.tue.nl/~bvasiles/languages/list.html

Assembly dev's are versatile; assembly itself is exotic

Are this likely to speak <column>

	Asm	C	C++	Cobol	CSS	Groovy	HTML	Java	JavaScript	Perl	PHP	Shell	XML
Asm	100%	55%	54%	1%	15%	1%	23%	39%	28%	12%	28%	1%	18%
C	8%	100%	48%	0%	12%	1%	17%	31%	21%	8%	21%	0%	13%
C++	5%	32%	100%	0%	10%	1%	15%	26%	18%	6%	18%	0%	11%
COBOL	12%	35%	40%	100%	24%	3%	29%	48%	38%	17%	37%	1%	28%
CSS	2%	10%	13%	0%	100%	1%	61%	21%	54%	5%	39%	0%	16%
Groovy	3%	15%	18%	1%	17%	100%	26%	63%	32%	7%	23%	0%	26%
HTML	2%	11%	14%	0%	46%	1%	100%	25%	56%	5%	40%	0%	18%
Java	2%	12%	15%	0%	10%	2%	15%	100%	19%	4%	16%	0%	12%
JavaScript	2%	9%	11%	0%	25%	1%	35%	20%	100%	4%	31%	0%	13%
Perl	5%	25%	27%	1%	18%	2%	26%	31%	30%	100%	31%	1%	19%
PHP	2%	9%	11%	0%	19%	1%	26%	17%	33%	4%	100%	0%	12%
Shell	12%	34%	38%	1%	19%	3%	32%	43%	33%	24%	35%	100%	24%
XML	3%	14%	19%	0%	20%	2%	29%	34%	35%	7%	31%	0%	100%

People who speak <row>

http://www.win.tue.nl/~bvasiles/languages/list.html

Cobol dev's are versatile but extremely scarce

Are this likely to speak <column>

	Asm	C	C++	Cobol	CSS	Groovy	HTML	Java	JavaScript	Perl	PHP	Shell	XML
Asm	100%	55%	54%	1%	15%	1%	23%	39%	28%	12%	28%	1%	18%
C	8%	100%	48%	0%	12%	1%	17%	31%	21%	8%	21%	0%	13%
C++	5%	32%	100%	0%	10%	1%	15%	26%	18%	6%	18%	0%	11%
COBOL	12%	35%	40%	100%	24%	3%	29%	48%	38%	17%	37%	1%	28%
CSS	2%	10%	13%	0%	100%	1%	61%	21%	54%	5%	39%	0%	16%
Groovy	3%	15%	18%	1%	17%	100%	26%	63%	32%	7%	23%	0%	26%
HTML	2%	11%	14%	0%	46%	1%	100%	25%	56%	5%	40%	0%	18%
Java	2%	12%	15%	0%	10%	2%	15%	100%	19%	4%	16%	0%	12%
JavaScript	2%	9%	11%	0%	25%	1%	35%	20%	100%	4%	31%	0%	13%
Perl	5%	25%	27%	1%	18%	2%	26%	31%	30%	100%	31%	1%	19%
PHP	2%	9%	11%	0%	19%	1%	26%	17%	33%	4%	100%	0%	12%
Shell	12%	34%	38%	1%	19%	3%	32%	43%	33%	24%	35%	100%	24%
XML	3%	14%	19%	0%	20%	2%	29%	34%	35%	7%	31%	0%	100%

People who speak <row>

http://www.win.tue.nl/~bvasiles/languages/list.html

Asymmetry present also in more "obvious" pairs

Are this likely to speak <column>

	Asm	C	C++	Cobol	CSS	Groovy	HTML	Java	JavaScript	Perl	PHP	Shell	XML
Asm	100%	55%	54%	1%	15%	1%	23%	39%	28%	12%	28%	1%	18%
C	8%	100%	48%	0%	12%	1%	17%	31%	21%	8%	21%	0%	13%
C++	5%	32%	100%	0%	10%	1%	15%	26%	18%	6%	18%	0%	11%
COBOL	12%	35%	40%	100%	24%	3%	29%	48%	38%	17%	37%	1%	28%
CSS	2%	10%	13%	0%	100%	1%	61%	21%	54%	5%	39%	0%	16%
Groovy	3%	15%	18%	1%	17%	100%	26%	63%	32%	7%	23%	0%	26%
HTML	2%	11%	14%	0%	46%	1%	100%	25%	56%	5%	40%	0%	18%
Java	2%	12%	15%	0%	10%	2%	15%	100%	19%	4%	16%	0%	12%
JavaScript	2%	9%	11%	0%	25%	1%	35%	20%	100%	4%	31%	0%	13%
Perl	5%	25%	27%	1%	18%	2%	26%	31%	30%	100%	31%	1%	19%
PHP	2%	9%	11%	0%	19%	1%	26%	17%	33%	4%	100%	0%	12%
Shell	12%	34%	38%	1%	19%	3%	32%	43%	33%	24%	35%	100%	24%
XML	3%	14%	19%	0%	20%	2%	29%	34%	35%	7%	31%	0%	100%

People who speak <row>

http://www.win.tue.nl/~bvasiles/languages/list.html

Case study: Emacs

1985-2012: C, Emacs Lisp, C++, Java, Lisp, Python, M4, … (26)

How to quantify this risk
of not finding developers with knowledge
of certain programming languages?

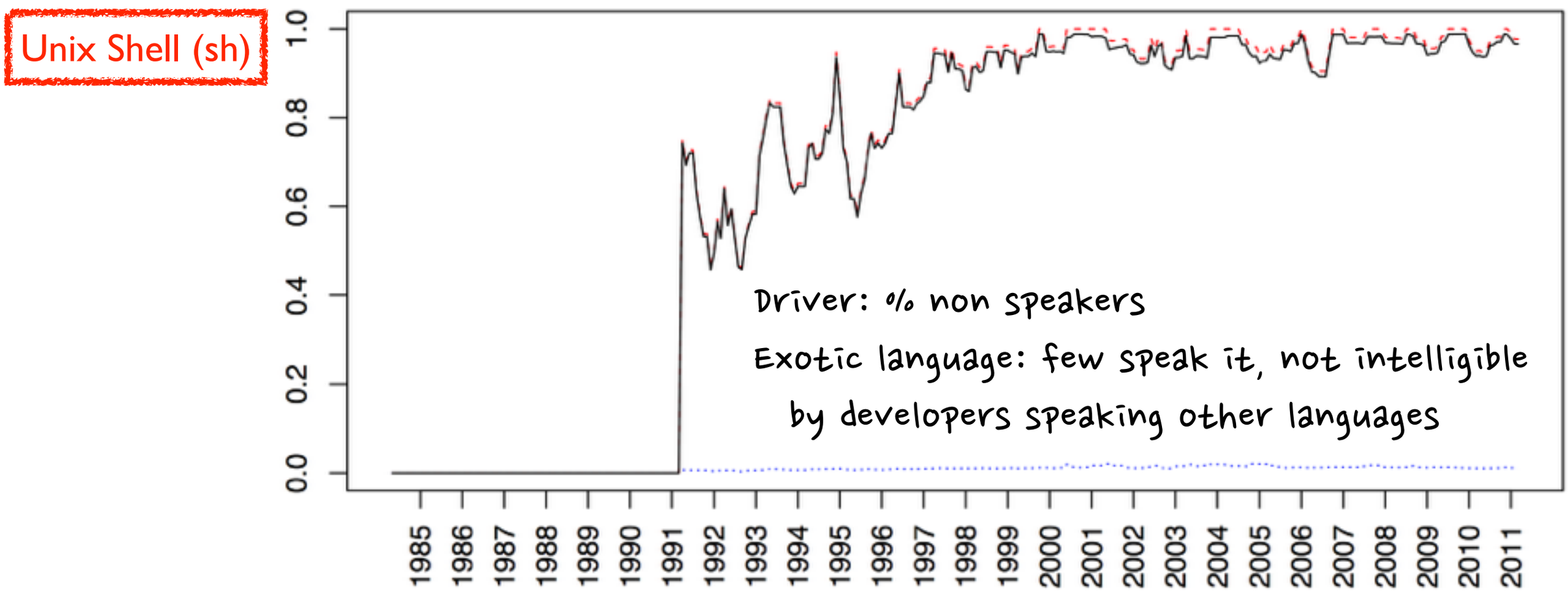
Case study: Emacs

1985-2012: C, Emacs Lisp, C++, Java, Lisp, Python, M4, … (26)

solid black: risk measure

dashed red: %community that does not speak language

dotted blue: red - black (intelligible by developers speaking other languages)



Case study: Emacs

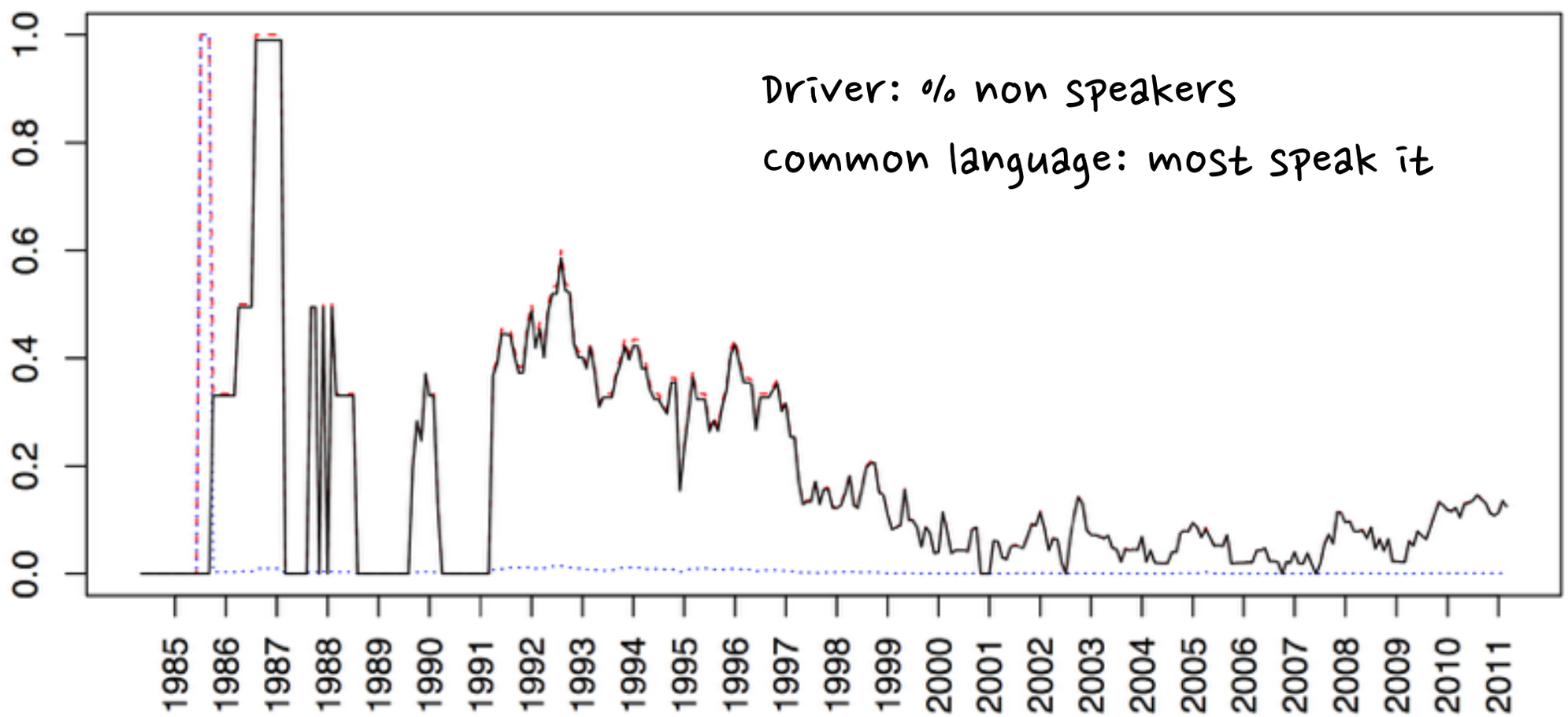
1985-2012: C, Emacs Lisp, C++, Java, Lisp, Python, M4, … (26)

solid black: risk measure

dashed red: %community that does not speak language

dotted blue: red - black (intelligible by developers speaking other languages)

Emacs Lisp



Case study: Emacs

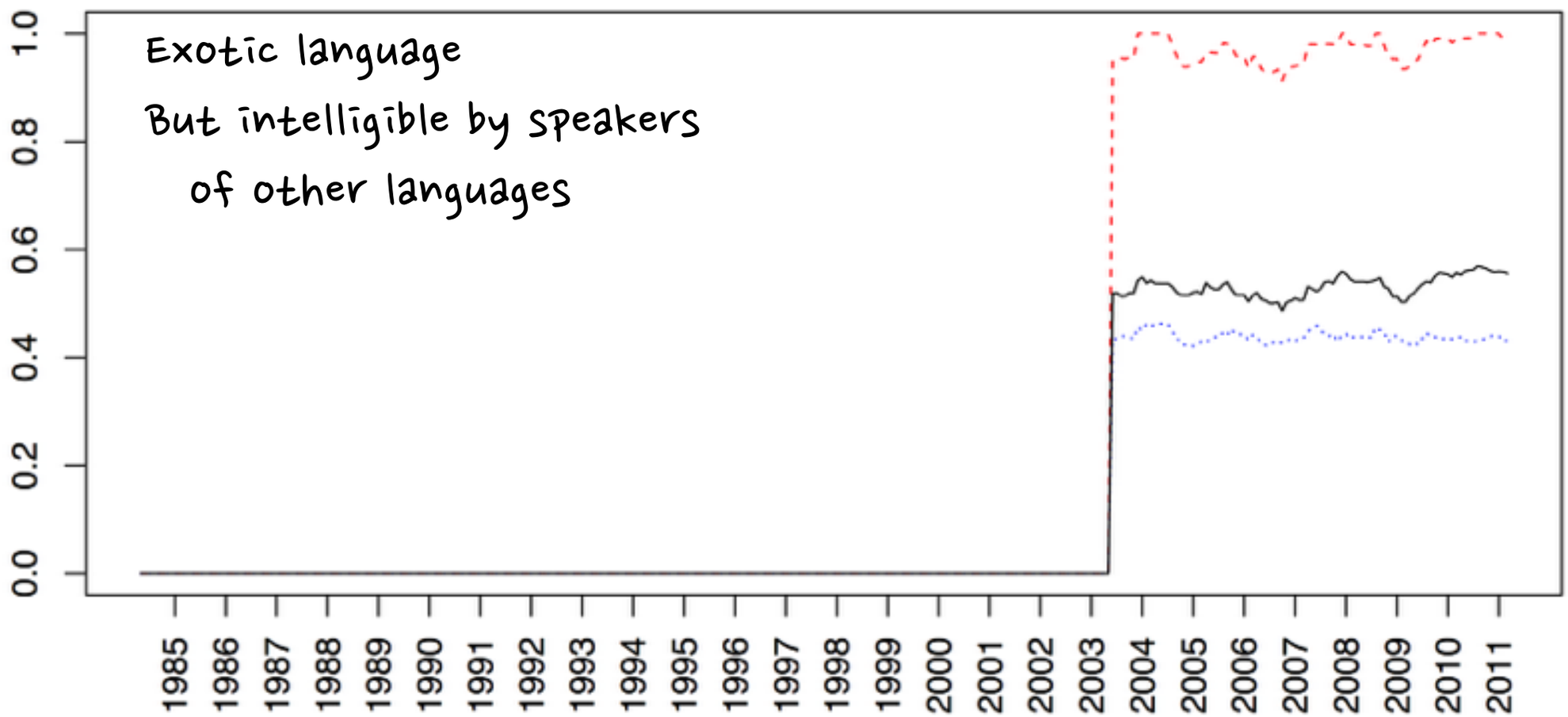
1985-2012: C, Emacs Lisp, C++, Java, Lisp, Python, M4, … (26)

solid black: risk measure

dashed red: %community that does not speak language

dotted blue: red - black (intelligible by developers speaking other languages)

Python



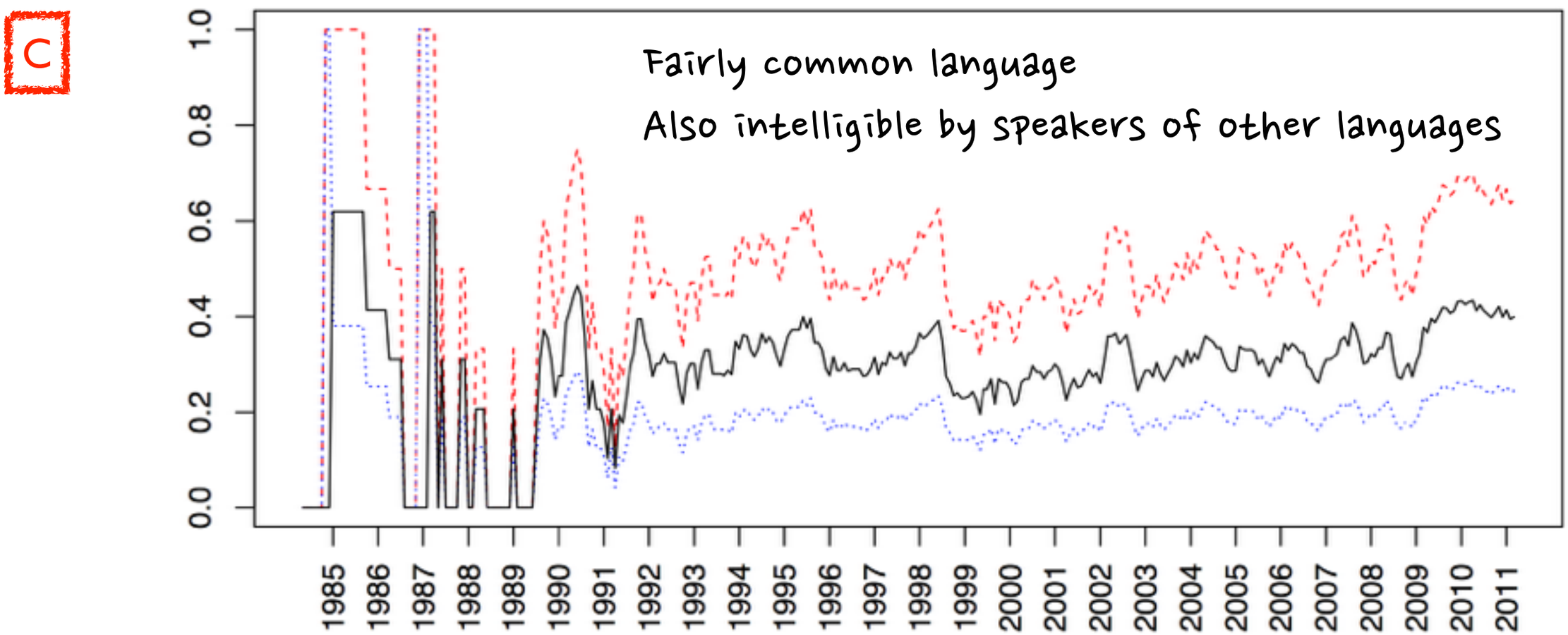
Case study: Emacs

1985-2012: C, Emacs Lisp, C++, Java, Lisp, Python, M4, … (26)

solid black: risk measure

dashed red: %community that does not speak language

dotted blue: red - black (intelligible by developers speaking other languages)



OSS Communities

Different skill sets and skill levels
(Giuri et al., 2004)

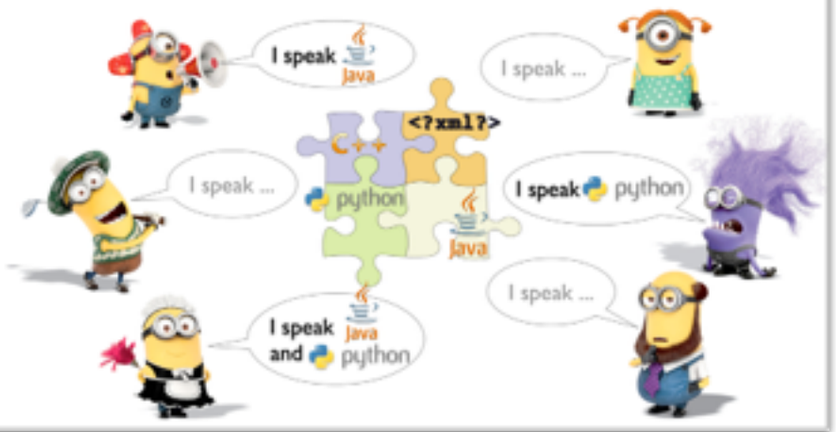
Different activities
(Vasilescu et al., 2013)

Mix of novices and experts
(Dabbish et al., 2012)

Highly interactive Decentralized Heterogeneous

Self-directed Knowledge-intensive

Knowledge of programming languages



High turnover

What happens when Purple Minion leaves the community?

Does knowledge of python disappear?

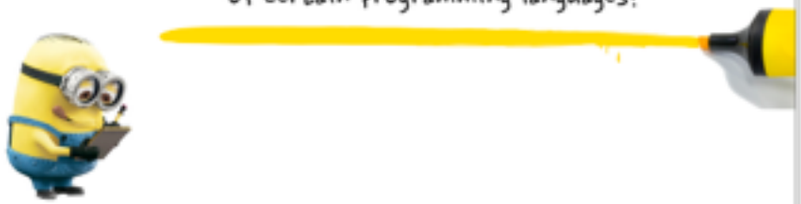
Is the community at risk?

Maintain or migrate legacy code?

Intuitively healthier

This talk: first steps

How to quantify this risk
of not finding developers with knowledge
of certain programming languages?



Idea

~~has worked on~~
Who else "~~speaks~~" python in the community?
might understand

Mine expertise from history of contributions

+

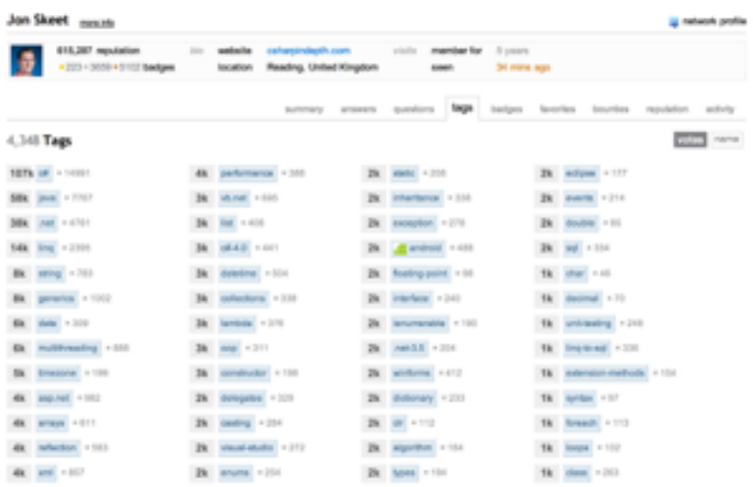
Approximate what else they might understand: universal measure of intelligibility of programming languages

Ingredients

Linguistic diversity
(natural languages)

crowdsourced knowledge
stackoverflow.com

Users "collect" tags



Cobol dev's are versatile but extremely scarce

Are this likely to speak <column>

	Asm	C	C++	Cobol	CSS	Groovy	HTML	Java	JavaScript	Perl	PHP	Shell	XML
Asm	100%	55%	54%	1%	15%	1%	23%	39%	28%	12%	28%	1%	18%
C	8%	100%	48%	0%	12%	1%	17%	31%	21%	8%	21%	0%	13%
C++	5%	32%	100%	0%	10%	1%	15%	26%	18%	6%	18%	0%	11%
COBOL	12%	35%	40%	100%	24%	3%	29%	48%	38%	17%	37%	1%	28%
CSS	2%	10%	13%	0%	100%	1%	61%	21%	54%	5%	39%	0%	16%
Groovy	3%	15%	18%	1%	17%	100%	26%	63%	32%	7%	23%	0%	26%
HTML	2%	11%	14%	0%	46%	1%	100%	25%	56%	5%	40%	0%	18%
Java	2%	12%	15%	0%	10%	2%	15%	100%	19%	4%	16%	0%	12%
JavaScript	2%	9%	11%	0%	25%	1%	35%	20%	100%	4%	31%	0%	13%
Perl	5%	25%	27%	1%	18%	2%	26%	31%	30%	100%	31%	1%	19%
PHP	2%	9%	11%	0%	19%	1%	26%	17%	33%	4%	100%	0%	12%
Shell	12%	34%	38%	1%	19%	3%	32%	43%	33%	24%	35%	100%	24%
XML	3%	14%	19%	0%	20%	2%	29%	34%	35%	7%	31%	0%	100%

People who speak <row>

http://www.win.tue.nl/~bvasiles/languages/list.html

Case study: Emacs

1985-2012: C, Emacs Lisp, C++, Java, Lisp, Python, M4, … (26)

solid black: risk measure
dashed red: %community that does not speak language
dotted blue: red - black (intelligible by developers speaking other languages)

Python