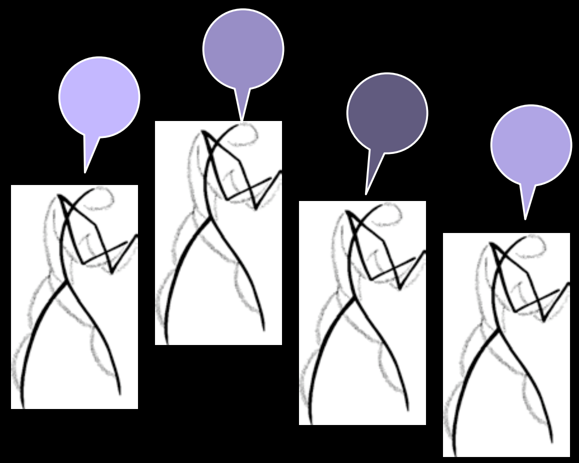
Learning-Driven Linguistic Evolution

Lisa Pearl
Cognitive Sciences, UC Irvine

March 16, 2008

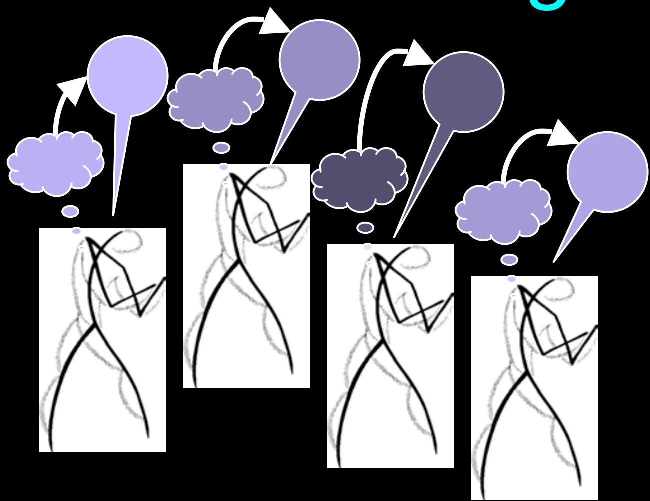
Evolution of Psychological Categories Workshop
Institute for Mathematical Behavioral Sciences
UC Irvine

Linguistic Evolution, In Brief



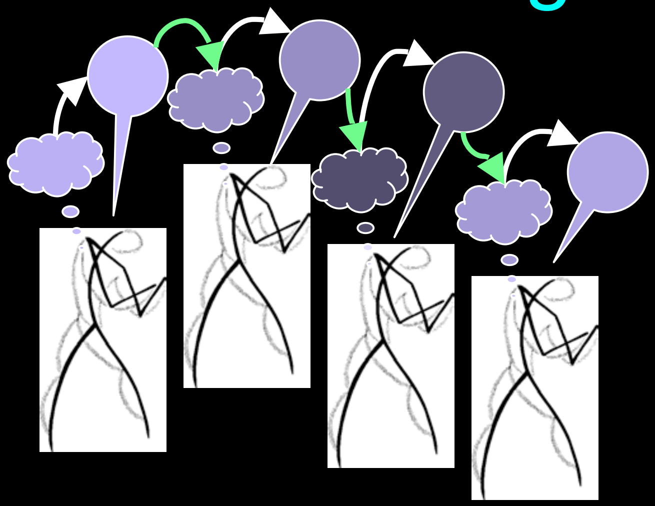
Linguistic knowledge is transmitted in a population via interaction with other speakers in the population.

Linguistic Evolution, In Brief



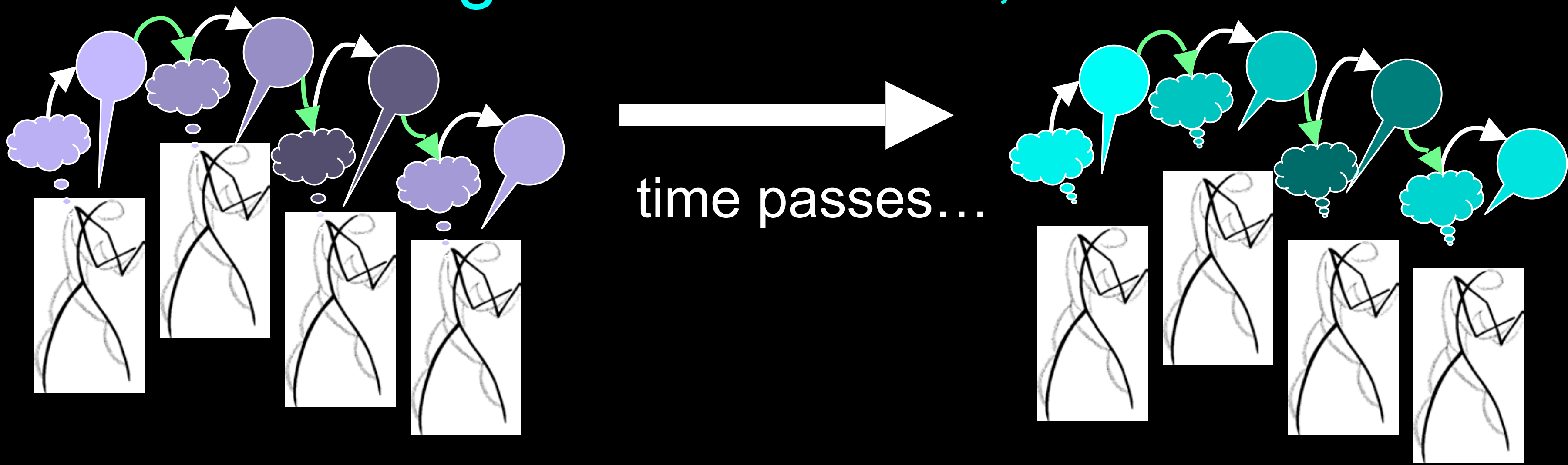
The information speakers transmit (observable data) is based on their own linguistic knowledge.

Linguistic Evolution, In Brief



Speakers adjust their linguistic knowledge based on the observable (and encountered) data from other population members.

Linguistic Evolution, In Brief



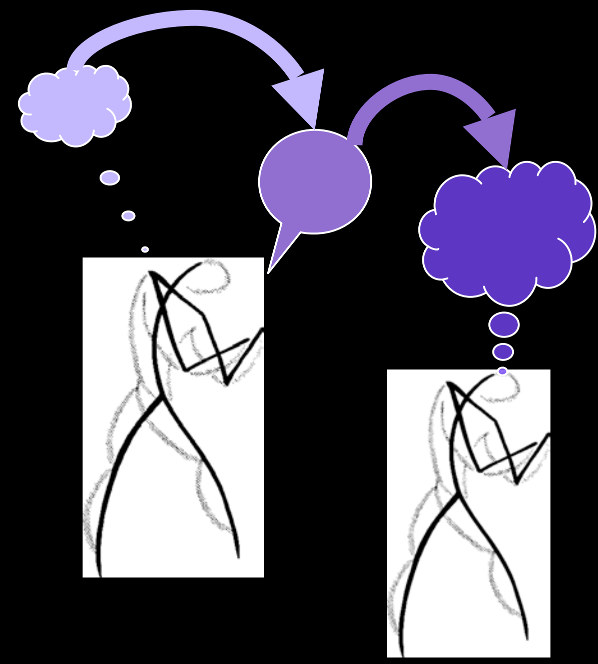
Population-level changes over time depend on what information speakers pass to subsequent generations and how that information is integrated into an individual's linguistic knowledge.

Integrating Linguistic Information

Not all linguistic knowledge is created equal

Some knowledge can be altered throughout an individual's life

(example: vocabulary)

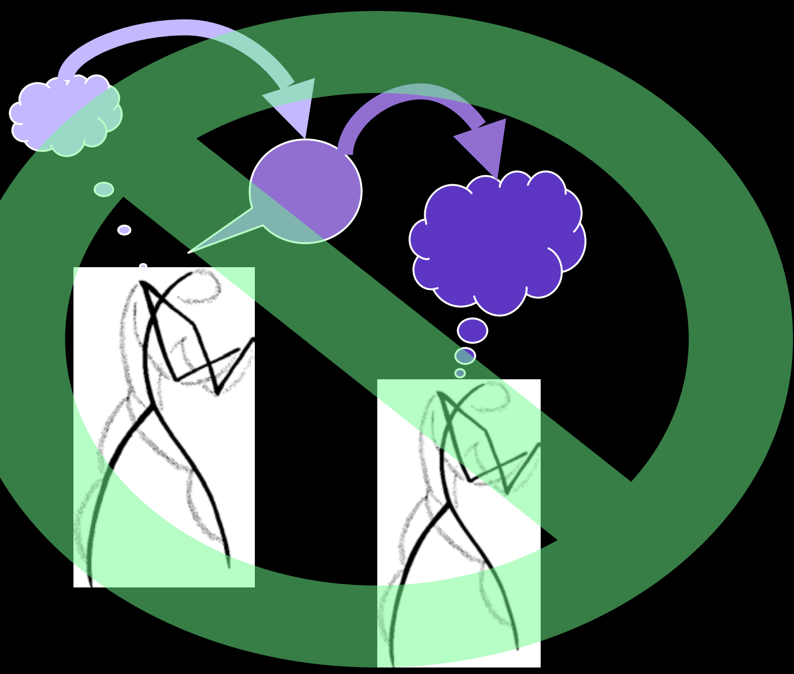


Integrating Linguistic Information

Not all linguistic knowledge is created equal

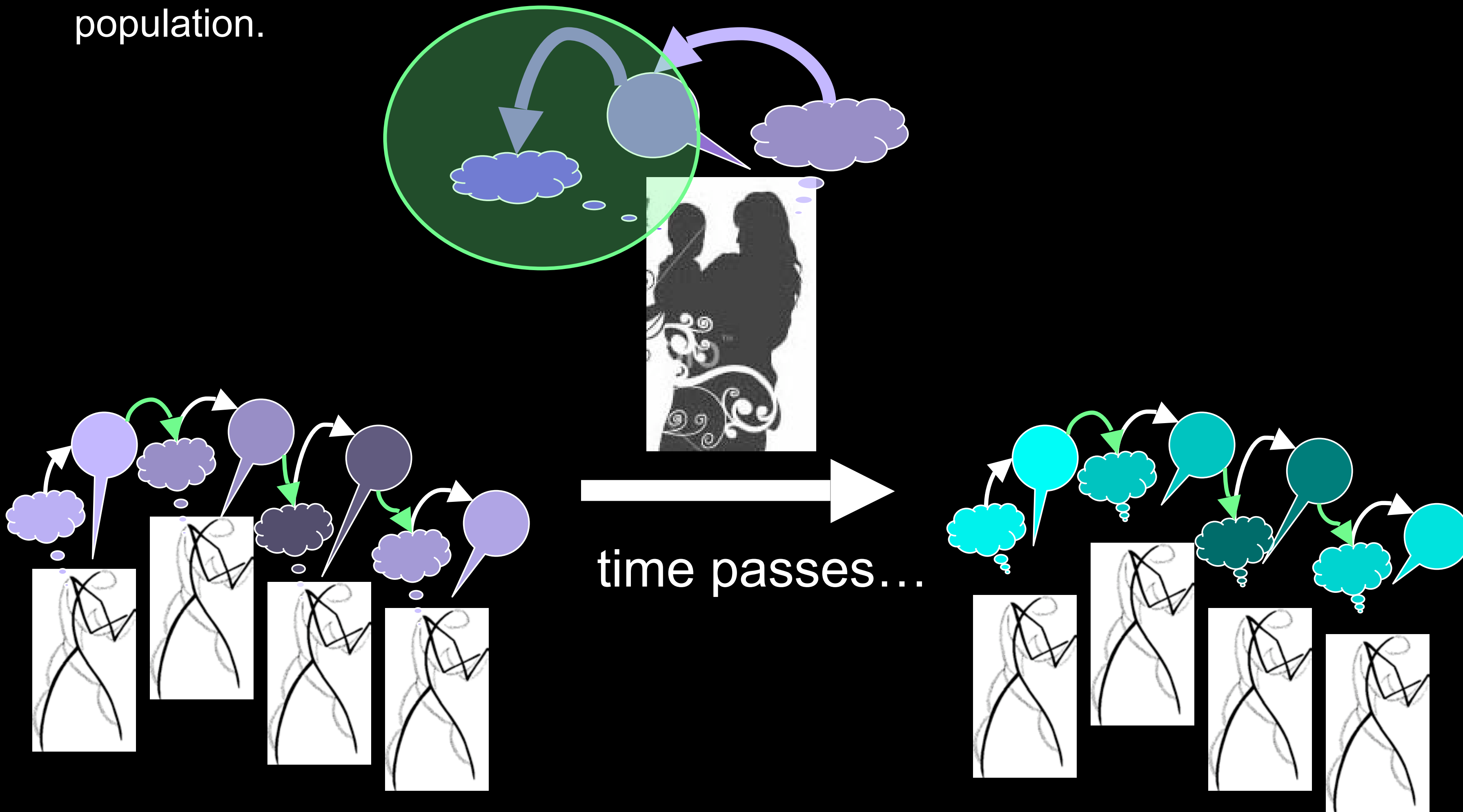
Some knowledge can be altered only during the early stages of an individual's life

(example: word order rules)



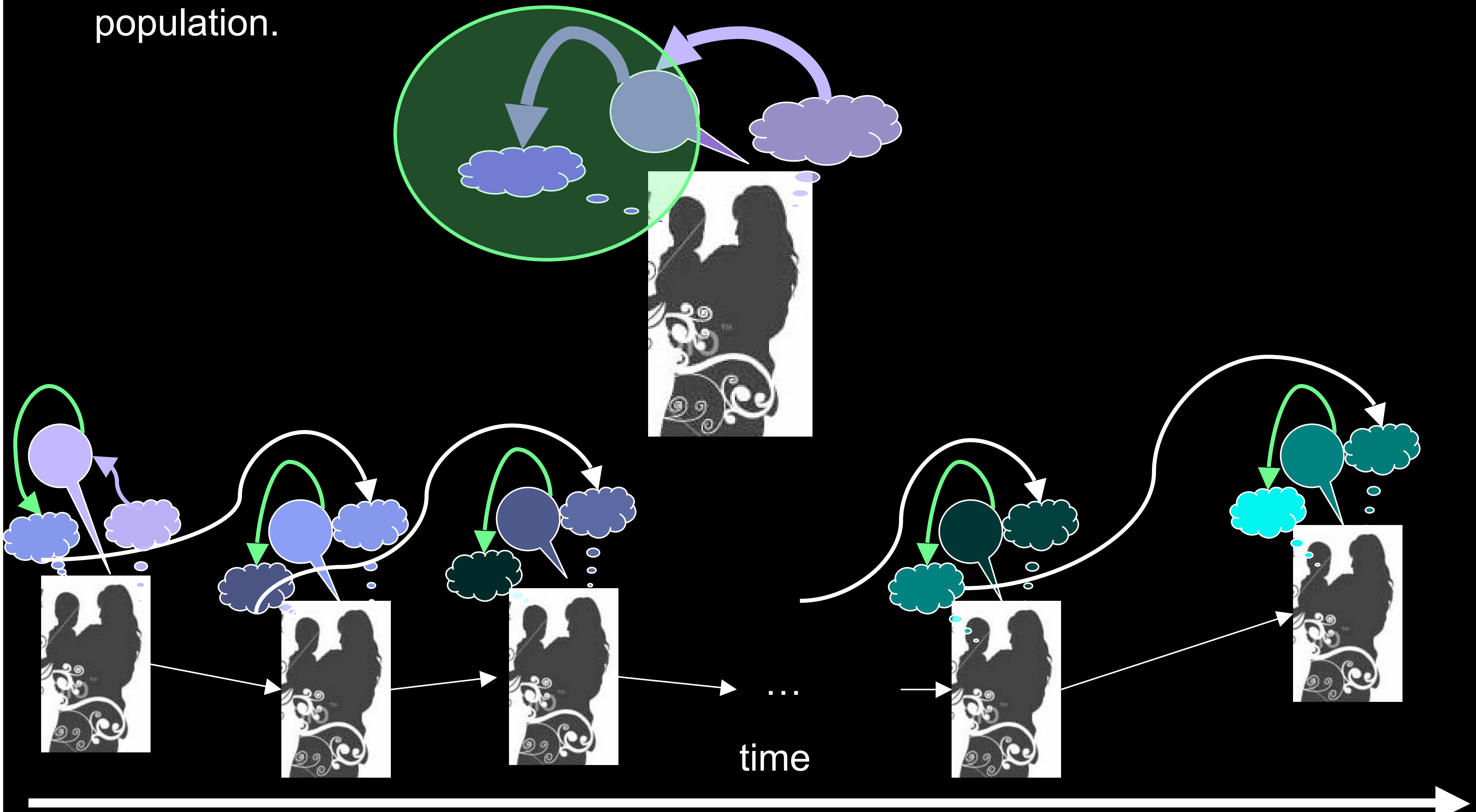
Change to knowledge that is alterable early

Implication: The way in which young learners integrate linguistic information (along with the data available) determines the linguistic composition of the population and the speed at which the linguistic knowledge evolves within the population.



Change to knowledge that is alterable early

Implication: The way in which young learners integrate linguistic information (along with the data available) determines the linguistic composition of the population and the speed at which the linguistic knowledge evolves within the population.



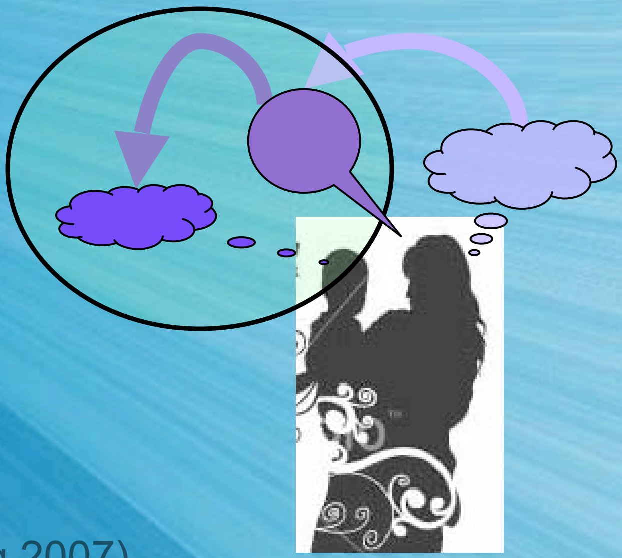
Road Map

I. Individual Language Learning
   The Nature of Linguistic Knowledge
   Individual Learning Framework

II. Linguistic Evolution: Case Study
   Old English Word Order
   Modeling Individuals (Pearl & Weinberg 2007)
   Modeling Populations
   Issues in Empirical Grounding
   Selective Learning Biases

Road Map

I. Individual Language Learning
   The Nature of Linguistic Knowledge
   Individual Learning Framework



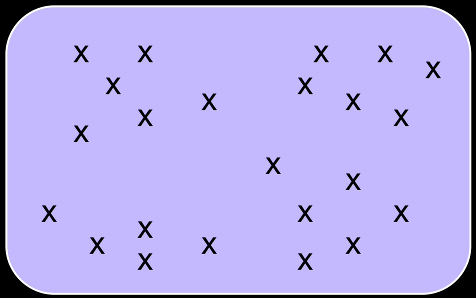
II. Linguistic Evolution: Case Study
   Old English Word Order
   Modeling Individuals (Pearl & Weinberg 2007)
   Modeling Populations
   Issues in Empirical Grounding
   Selective Learning Biases

The Nature of Linguistic Knowledge

Different aspects: more and less transparent from data

Categorization/Clustering

Ex: What are the contrastive sounds of a language?

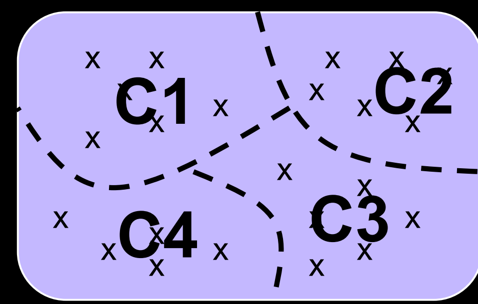


The Nature of Linguistic Knowledge

Different aspects: more and less transparent from data

Categorization/Clustering

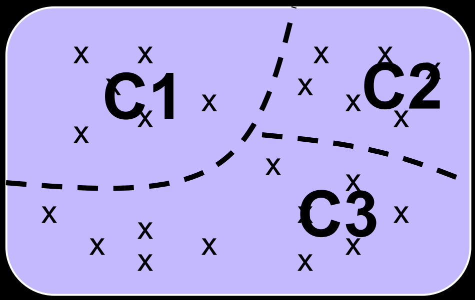
Ex: What are the contrastive sounds of a language?



The Nature of Linguistic Knowledge

Different aspects: more and less transparent from data

Categorization/Clustering
Ex: What are the contrastive sounds of a language?

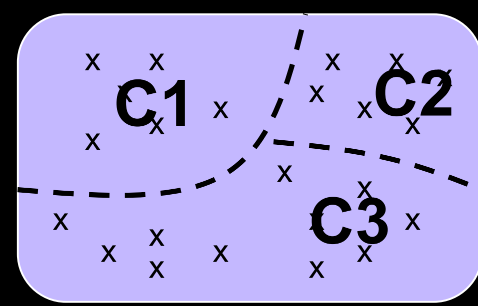


The Nature of Linguistic Knowledge

Different aspects: more and less transparent from data

Categorization/Clustering

Ex: What are the contrastive sounds of a language?



Extraction

Ex: Where are words in fluent speech?

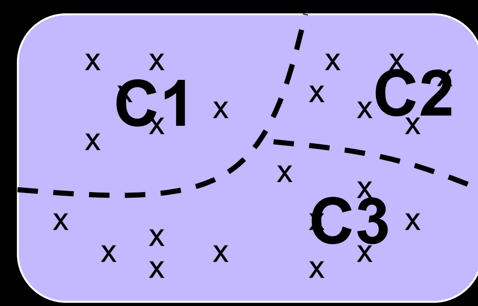
húwzəfréjdəvðəbÍgbæ‘dwə‘lf

The Nature of Linguistic Knowledge

Different aspects: more and less transparent from data

Categorization/Clustering

Ex: What are the contrastive sounds of a language?



Extraction

Ex: Where are words in fluent speech?

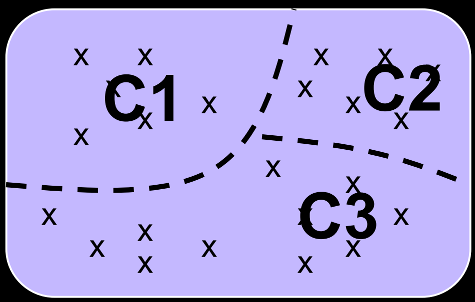
húwz əfréjd əv ðə bÍg bæ‘d wə‘lf

who‘s afraid of the big bad wolf

The Nature of Linguistic Knowledge

Different aspects: more and less transparent from data

Categorization/Clustering
Ex: What are the contrastive sounds of a language?



Extraction
Ex: Where are words in fluent speech?

húwz əfréjd əv ðə bÍg bæ‘d wə‘lf
who‘s afraid of the big bad wolf

Mapping
What are the word affixes that signal meaning (e.g. past tense in English)?

blink~blinked confide~confided

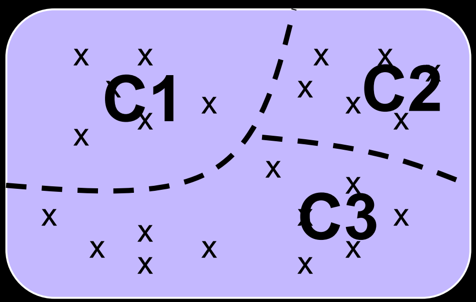
drink~drank

The Nature of Linguistic Knowledge

Different aspects: more and less transparent from data

Categorization/Clustering

Ex: What are the contrastive sounds of a language?



Extraction

Ex: Where are words in fluent speech?

húwz əfréjd əv ðə bÍg bæ‘d wə‘lf

who‘s afraid of the big bad wolf

Mapping

What are the word affixes that signal meaning (e.g. past tense in English)?

blink~blinked
blɪŋk blɪŋkt

confide~confided
kənfajd kənfajdəd

drink~drank
drɪŋk drejŋk

The Nature of Linguistic Knowledge

Different aspects: more and less transparent from data

Complex systems: What is the generative system that creates the observed (structured) data of language (ex: syntax)?

syntax = word order rules

Learning problem: many ways to generate observable data

The Nature of Linguistic Knowledge

Different aspects: more and less transparent from data

Complex systems: What is the generative system that creates the observed (structured) data of language (ex: syntax)?

syntax = word order rules

Learning problem: many ways to generate observable data

Observable data: word order   Subject   Verb   Object

Generative system: syntax

Kannada

Subject t_{Object} Verb Object

Object Verb underlying

English

Subject   Verb   Object

Verb Object underlying

German

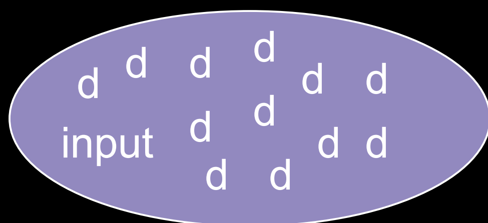
Subject   Verb $t_{Subject}$ Object t_{Verb}

Object Verb underlying

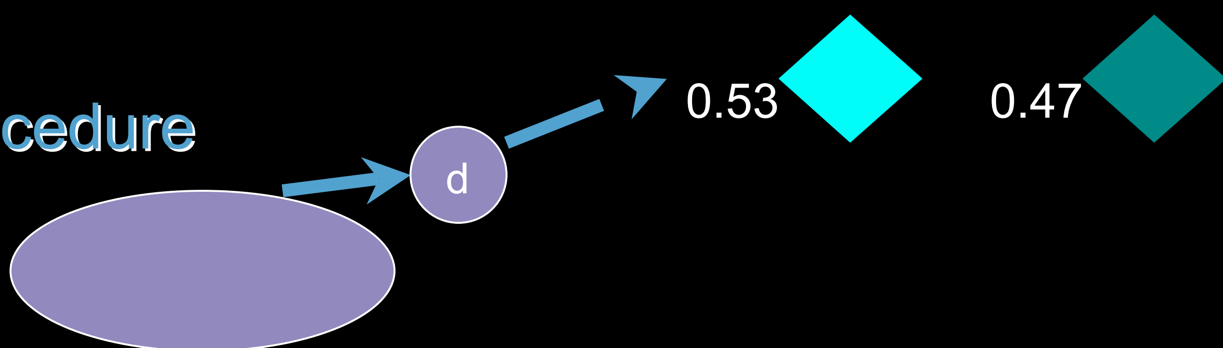
The individual learning framework: 3 components

(1) Hypothesis space

(2) Data



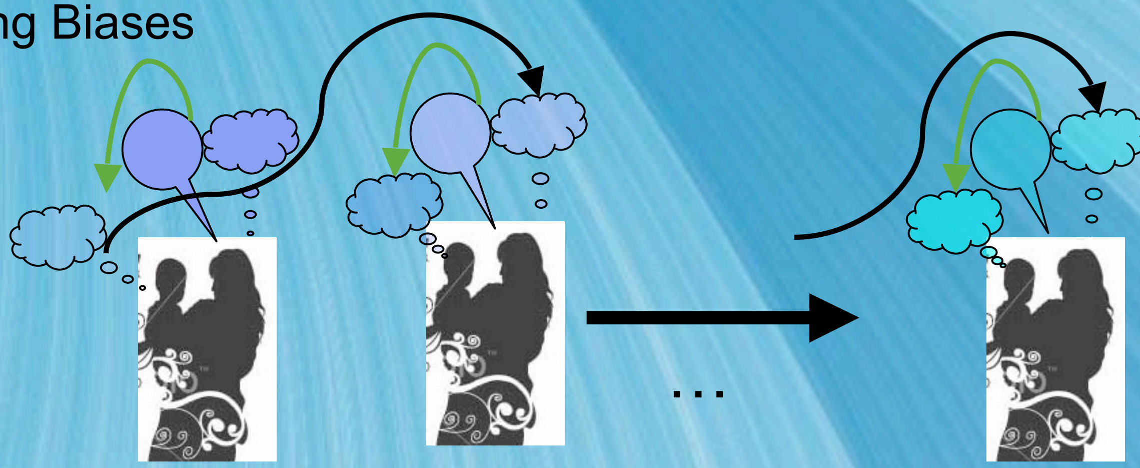
(3) Update procedure



Road Map

I. Individual Language Learning
The Nature of Linguistic Knowledge
Individual Learning Framework

II. Linguistic Evolution: Case Study
Old English Word Order
Modeling Individuals (Pearl & Weinberg 2007)
Modeling Populations
Issues in Empirical Grounding
Selective Learning Biases



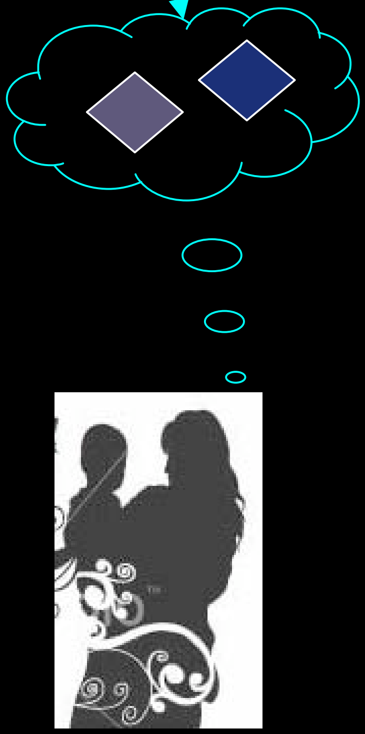
Old English

OV
P_{OV} = ??

VO
P_{VO} = ??

Changing Basic Word Order Rule in Old English:
Object-Verb (OV) vs. Verb-Object (VO) order

Individual Knowledge (underlying probability in speaker's mind): probability distribution between OV and VO orders



Old English

Changing Basic Word Order Rule in Old English:

Object-Verb (OV) vs. Verb-Object (VO) order

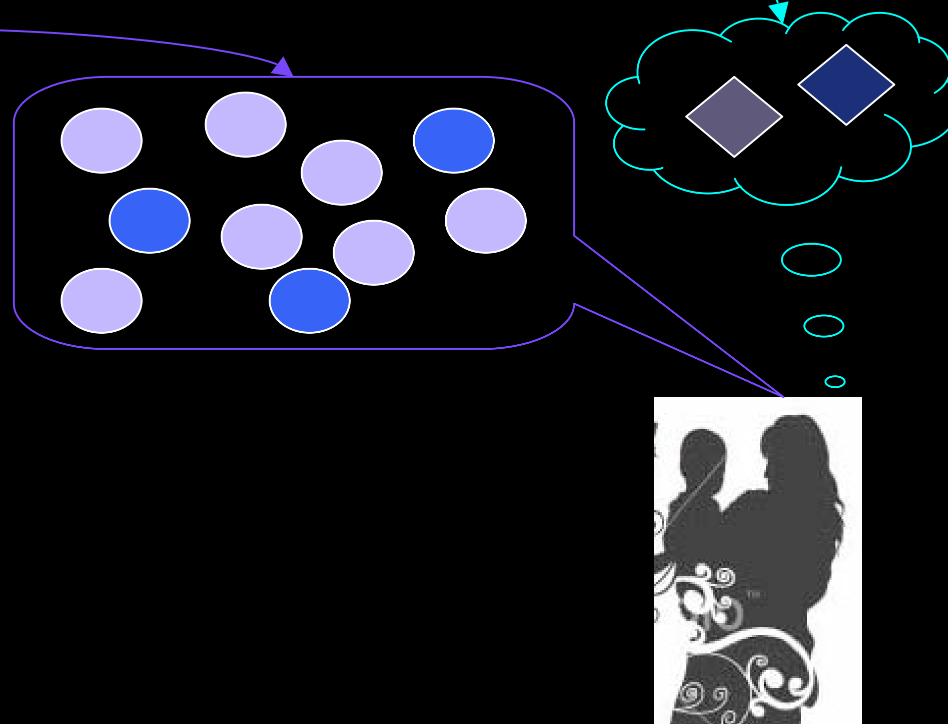
OV
P_{OV} = ??

VO
P_{VO} = ??

Individual Knowledge (underlying probability in speaker's mind): probability distribution between OV and VO orders

Individual Usage (observable data for learner): probability distribution between OV and VO orders (not necessarily same one as individual knowledge distribution, from learner's perspective)

Why not?



Underlying Distribution vs. Observable Distribution

Subject Verb Object

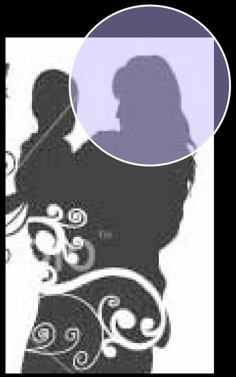
German/Old English

Surface order: Verb Object

Subject Verb $t_{Subject}$ Object t_{Verb}

Object Verb underlying

Speaker generates utterance



Underlying Distribution vs. Observable Distribution

Subject Verb Object

?

Surface order: Verb Object

?

Subject Verb $t_{Subject}$ Object t_{Verb}

Object Verb underlying

Subject Verb Object

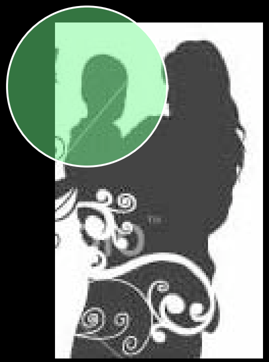
Verb Object underlying

?

Subject Verb $t_{Subject}$ t_{Verb} Object

Verb Object underlying

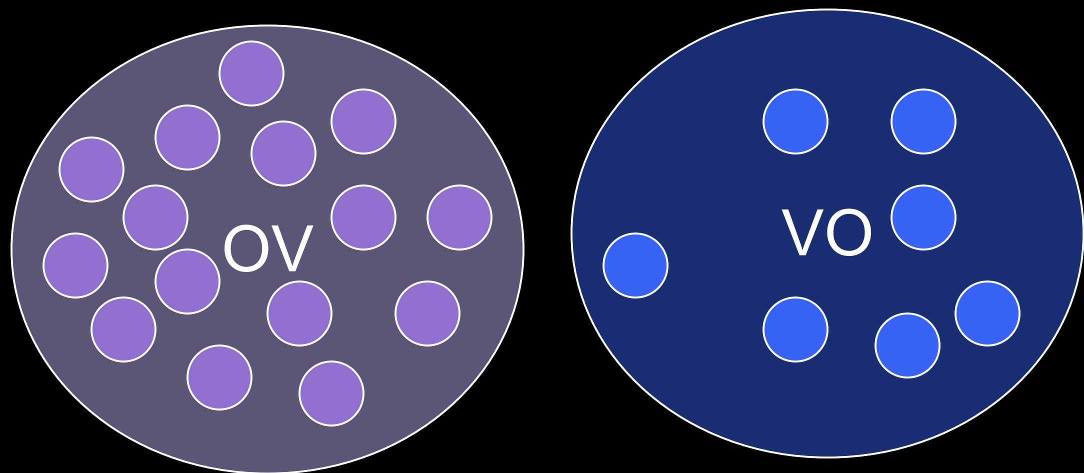
Learner interprets utterance



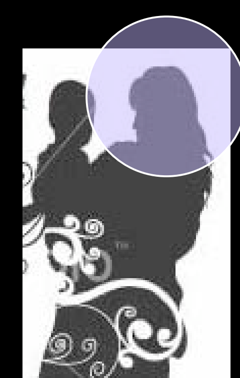
Underlying Distribution vs. Observable Distribution

Subject Verb $t_{Subject}$ Object t_{Verb}

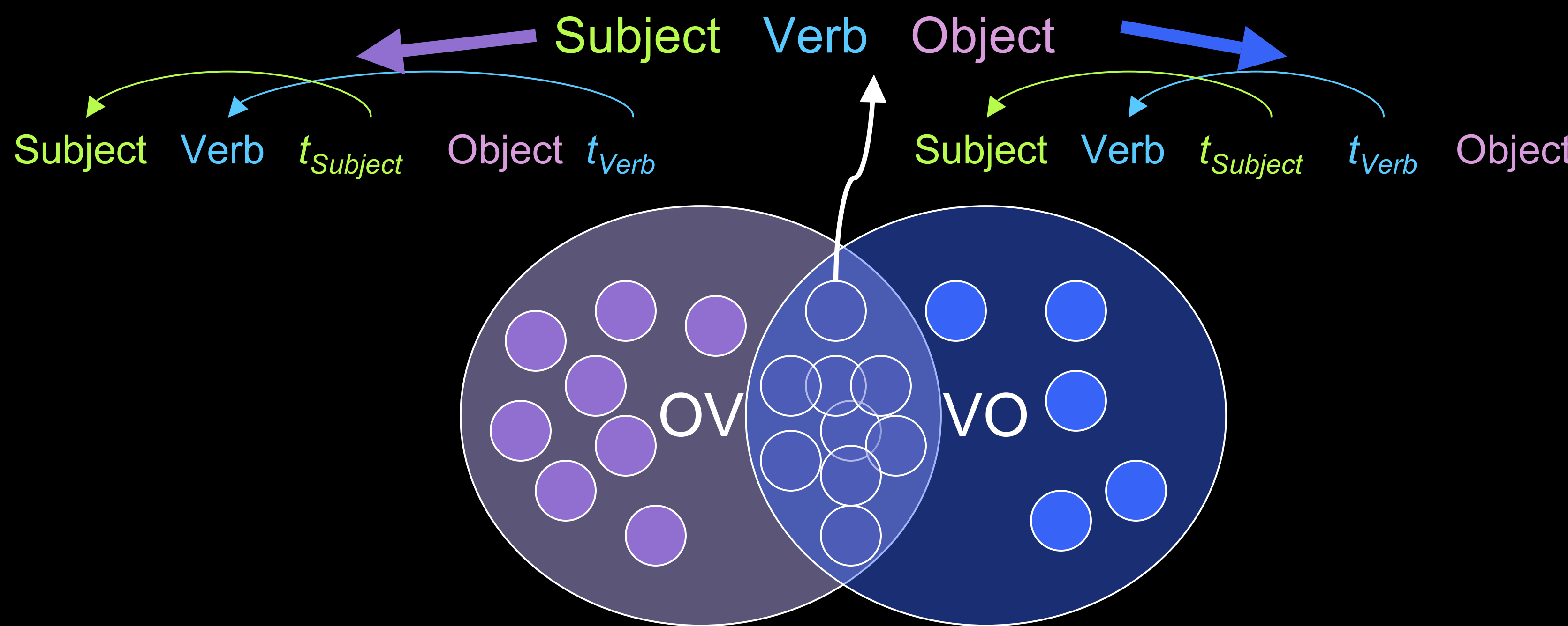
Subject Verb $t_{Subject}$ t_{Verb} Object



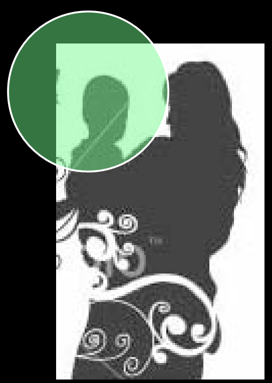
Every utterance generated by speaker is either OV or VO order in the underlying distribution



Underlying Distribution vs. Observable Distribution



The learner encounters data that is ambiguous between the two options. Distribution depends on learner's interpretation of ambiguous data



Old English

OV
P_{OV} = ??

VO
P_{VO} = ??

Changing Basic Word Order Rule in Old English:
Object-Verb (OV) vs. Verb-Object (VO) order

Individual Knowledge (underlying probability in speaker's mind): probability distribution between OV and VO orders

Individual Usage (observable data for learner): probability distribution between OV and VO orders (not necessarily same one as individual knowledge distribution, from learner's perspective)

Why not?

Old English

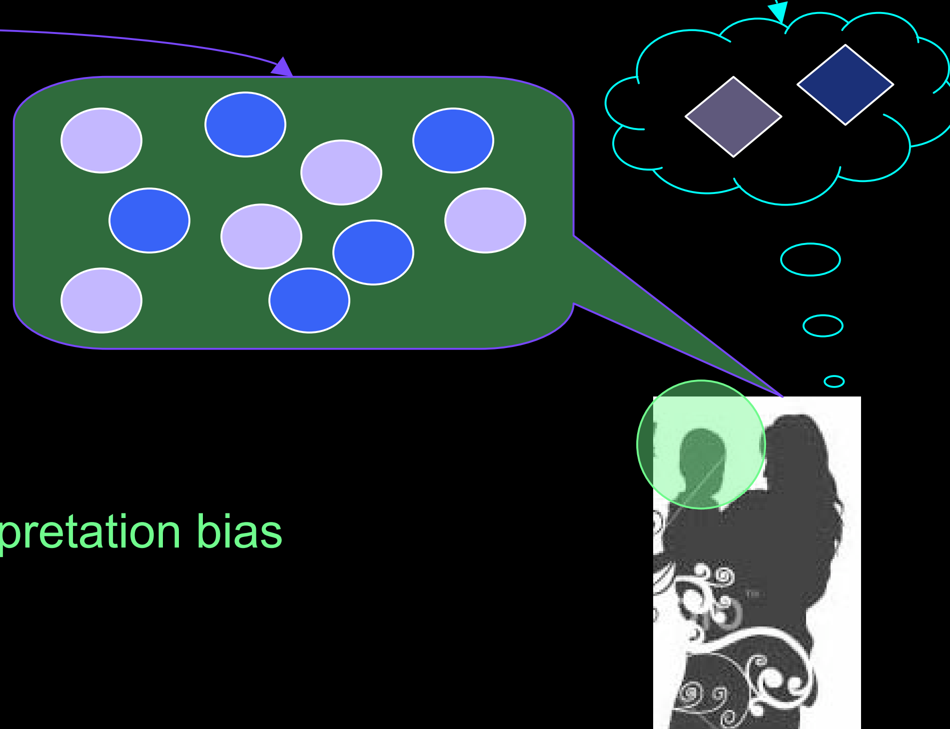
OV
$P_{OV} = ??$

VO
$P_{VO} = ??$

Changing Basic Word Order Rule in Old English:

Object-Verb (OV) vs. Verb-Object (VO) order

Individual Knowledge (underlying probability in speaker's mind): probability distribution between OV and VO orders



Individual Usage (observable data for learner): probability distribution between OV and VO orders (not necessarily same one as individual knowledge distribution, from learner's perspective)

Due to learner interpretation bias

Old English

Estimates of average individual usage from historical corpora:
YCOE Corpus 2003; PPCME2 Corpus 2000

~1000 A.D.-1150 A.D.: OV-biased

OV
P_{OV} = ??

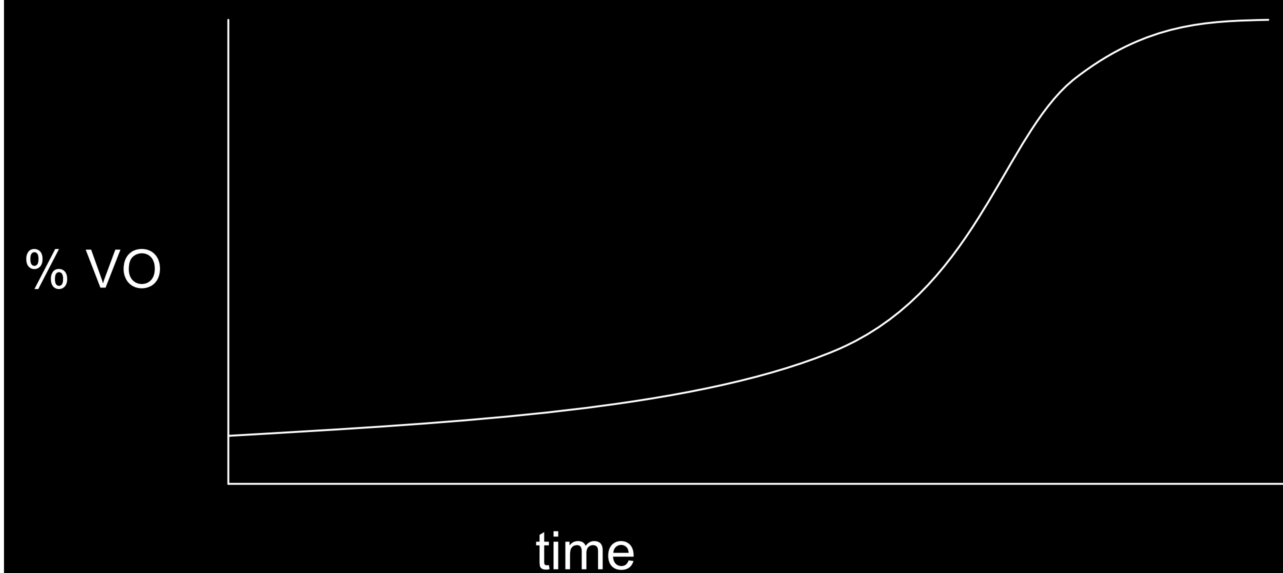
VO
P_{VO} = ??

→

~1200 A.D.: VO-biased

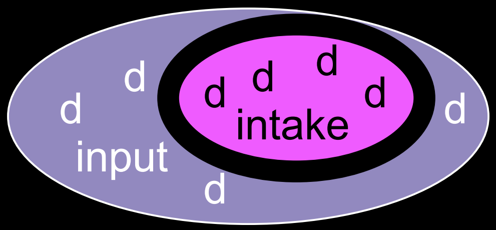
OV
P_{OV} = ??

VO
P_{VO} = ??



To get this rate of change, young individual learners at each time step must change their probability distribution the exact right amount from the previous population members' distribution

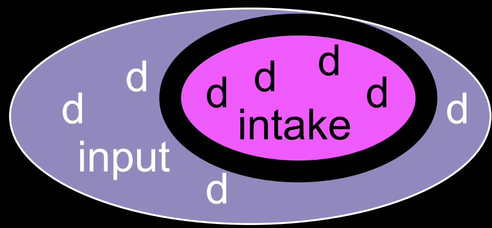
Modeling Individuals: Learning Biases



Interpretation Bias: Use only data perceived as most informative (Fodor 1998, Lightfoot 1999, Dresher 1999).

Interpretation Bias: Use only data that is more accessible (perhaps for language processing reasons) (Lightfoot 1991).

Modeling Individuals: Learning Biases



Interpretation Bias : Use only data perceived as most informative: unambiguous data (Fodor 1998, Lightfoot 1999, Dresher 1999).

Interpretation Bias: Use only data that is more accessible (perhaps for language processing reasons) (Lightfoot 1991).

Subject Verb $t_{Subject}$ Object t_{Verb}

Learner has heuristics for identifying unambiguous OV/VO data, based on partial knowledge of possible adult system rules (Fodor 1998, Lightfoot 1999, Dresher 1999)

Knowledge of tensed verb movement to 2nd phrasal position of sentence

OV unambiguous data:

$[\ldots]_{XP}$ … Object TensedVerb …

…TensedVerb … Object Verb-Marker …

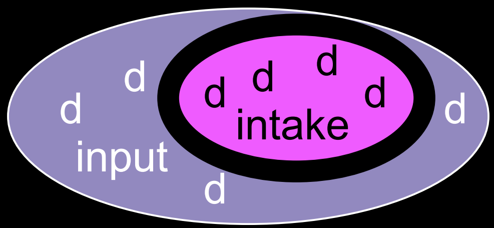
VO unambiguous data:

$[\ldots]_{XP}$ $[\ldots]_{XP}$ … TensedVerb Object …

…TensedVerb … Verb-Marker Object …

Subject Verb $t_{Subject}$ t_{Verb} Object

Modeling Individuals: Learning Biases



Interpretation Bias: Use only data perceived as most informative: unambiguous data (Fodor 1998, Lightfoot 1999, Dresher 1999).

OV unambiguous data:

[…]$_{XP}$ … Object TensedVerb …

…TensedVerb … Object Verb-Marker …

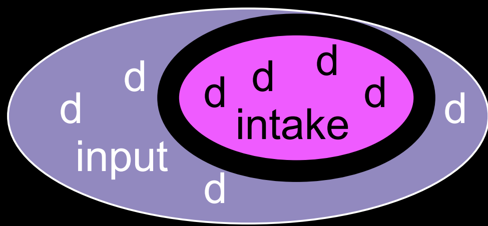
VO unambiguous data:

[…]$_{XP}$ […]$_{XP}$ … TensedVerb Object …

…TensedVerb … Verb-Marker Object …

Interpretation Bias: Use only data that is more accessible (perhaps for language processing reasons) (Lightfoot 1991).

Modeling Individuals: Learning Biases



Interpretation Bias: Use only data perceived as most informative: unambiguous data (Fodor 1998, Lightfoot 1999, Dresher 1999).

OV unambiguous data:
[…]$_{XP}$ … Object TensedVerb …
…TensedVerb … Object Verb-Marker …

VO unambiguous data:
[…]$_{XP}$ […]$_{XP}$ … TensedVerb Object …
…TensedVerb … Verb-Marker Object …

Interpretation Bias: Use only structurally simple (degree-0) data (Lightfoot 1991).

Jack told his mother that the giant was easy to fool.
[----Degree-0-------]
[------------Degree-1----------]

Modeling Individuals: Learning Biases

The point of interpretation biases: Unambiguous degree-0 data distribution may differ the right amount from population's underlying distribution to change at the right rate.

~1000 A.D.-1150 A.D.: OV-biased

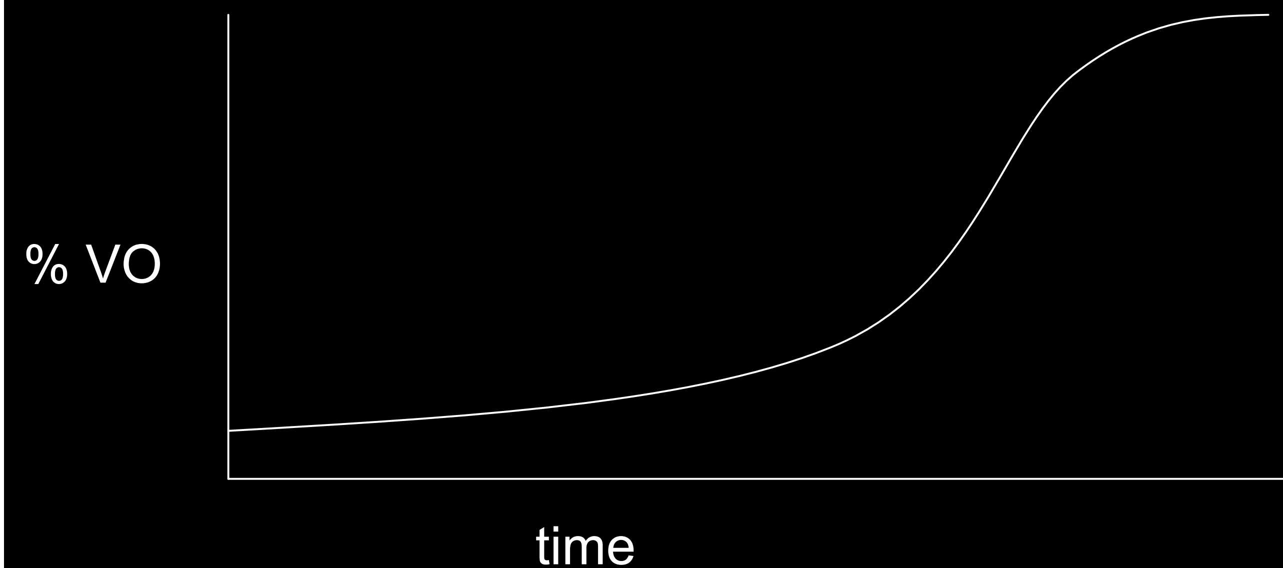
OV P_{OV} = ??

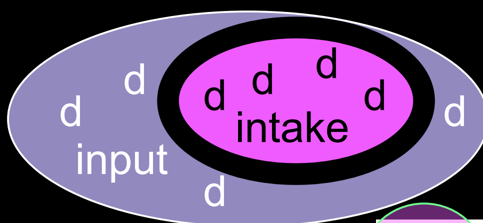
VO P_{VO} = ??

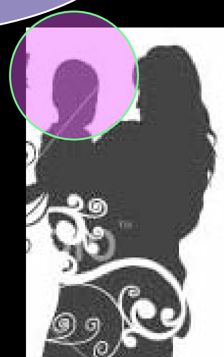
~1200 A.D.: VO-biased

OV P_{OV} = ??

VO P_{VO} = ??







Modeling Individuals: Knowledge & Learning

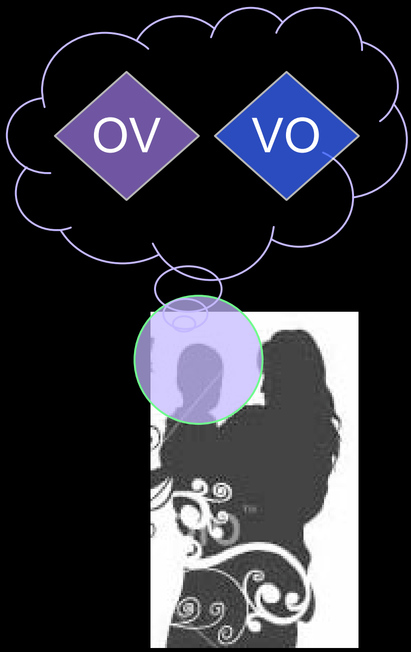
Individual learner tracks p_{VO} = probability of using VO

probability of using OV = 1 - p_{VO}

Old English: $0.0 <= p_{VO} <= 1.0$

Ex: 0.3 = 30% VO, 70% OV during generation

Initial p_{VO} = 0.5 (unbiased)



Modeling Individuals: Knowledge & Learning

Individual learner tracks p_{VO} = probability of using **VO**

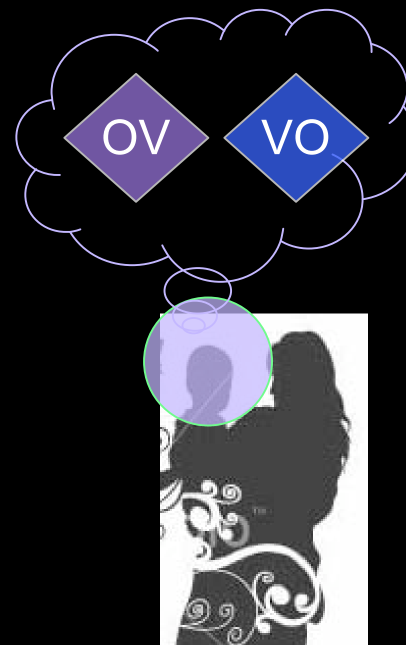
probability of using **OV** = 1 - p_{VO}

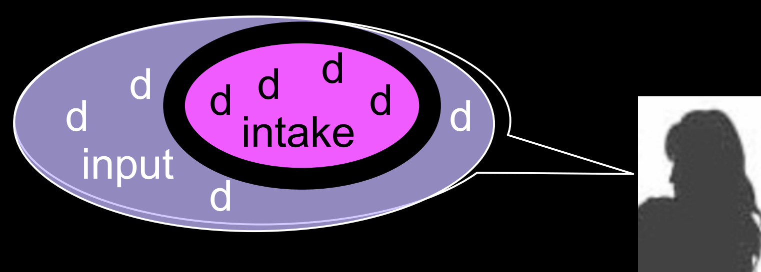
Old English: 0.0 <= p_{VO} <= 1.0

Ex: 0.3 = **30% VO**, **70% OV** during generation

Initial p_{VO} = 0.5 (unbiased)

Data from old members of population, filtered through selective learning biases.





Modeling Individuals: Knowledge & Learning

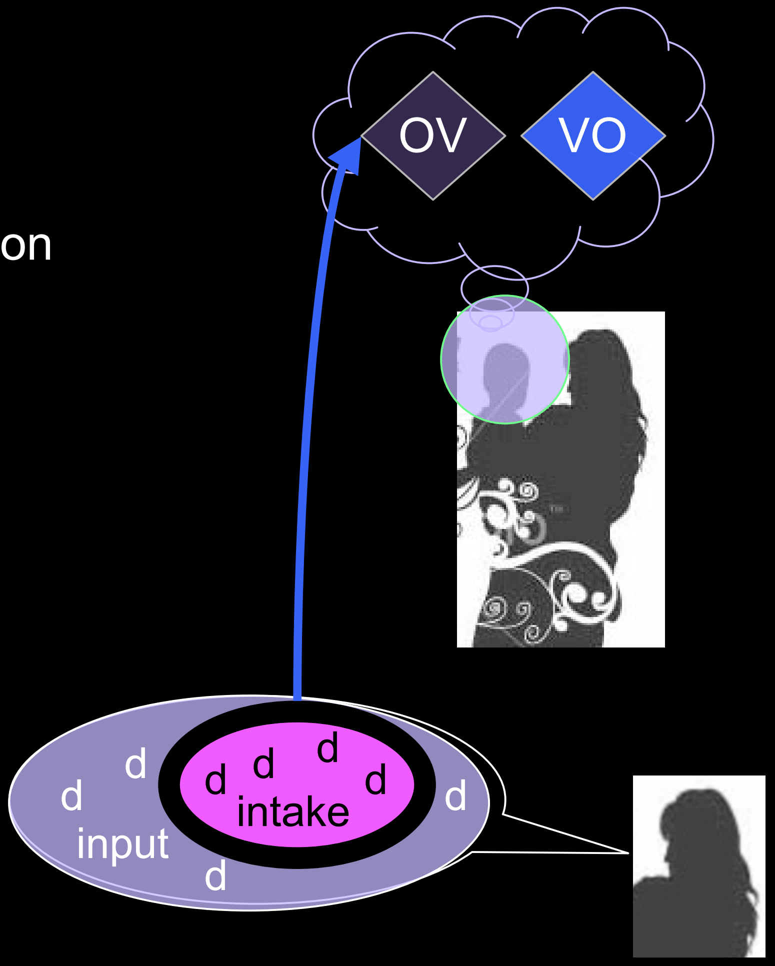
Individual learner tracks p_{VO} = probability of using VO
probability of using OV = 1 - p_{VO}

Old English: $0.0 \le p_{VO} \le 1.0$
Ex: 0.3 = 30% VO, 70% OV during generation

Initial p_{VO} = 0.5 (unbiased)

Data from old members of population, filtered through selective learning biases.

Individual update: Bayesian updating for binomial distribution (Chew 1971), adapted



Zoom-In on Updating Procedure

$$\mathrm{Max}(\mathrm{Prob}(\mathrm{pvo} \mid u)) = \mathrm{Max}\left(\frac{\mathrm{Prob}(u \mid \mathrm{pvo}) * \mathrm{Prob}(\mathrm{pvo})}{\mathrm{Prob}(u)}\right)$$

$$\mathrm{Prob}(\mathrm{pvo} \mid u) = \frac{\mathrm{pvo} * \binom{n}{r} * \mathrm{pvo}^{r} * (1 - \mathrm{pvo})^{n-r}}{\mathrm{Prob}(u)} (\text{for each point } r, 0 \quad r \quad n)$$

$$\frac{d}{dpvo}\left(\frac{\mathrm{pvo} * \binom{n}{r} * \mathrm{pvo}^{r} * (1 - \mathrm{pvo})^{n-r}}{\mathrm{Prob}(u)}\right) = 0$$

$$\frac{d}{dpvo}\left(\frac{\mathrm{pvo} * \binom{n}{r} * \mathrm{pvo}^{r} * (1 - \mathrm{pvo})^{n-r}}{\cancel{\mathrm{Prob}(u)}}\right) = 0 \qquad (\mathrm{P}(u) \text{ is constant with respect to pvo})$$

$$\mathrm{pvo} = \frac{r+1}{n+1}, \; r = \mathrm{pvo}_{\mathrm{Oprev}} * n$$

Replace 1 in numerator and denominator with $c = \mathrm{pvo}_{\mathrm{Oprev}} * m$ if VO, $c = (1 - \mathrm{pvo}_{\mathrm{Oprev}}) * m$ if OV

$3.0 \le m \le 5.0$

Zoom-In on Updating Procedure

If OV data point

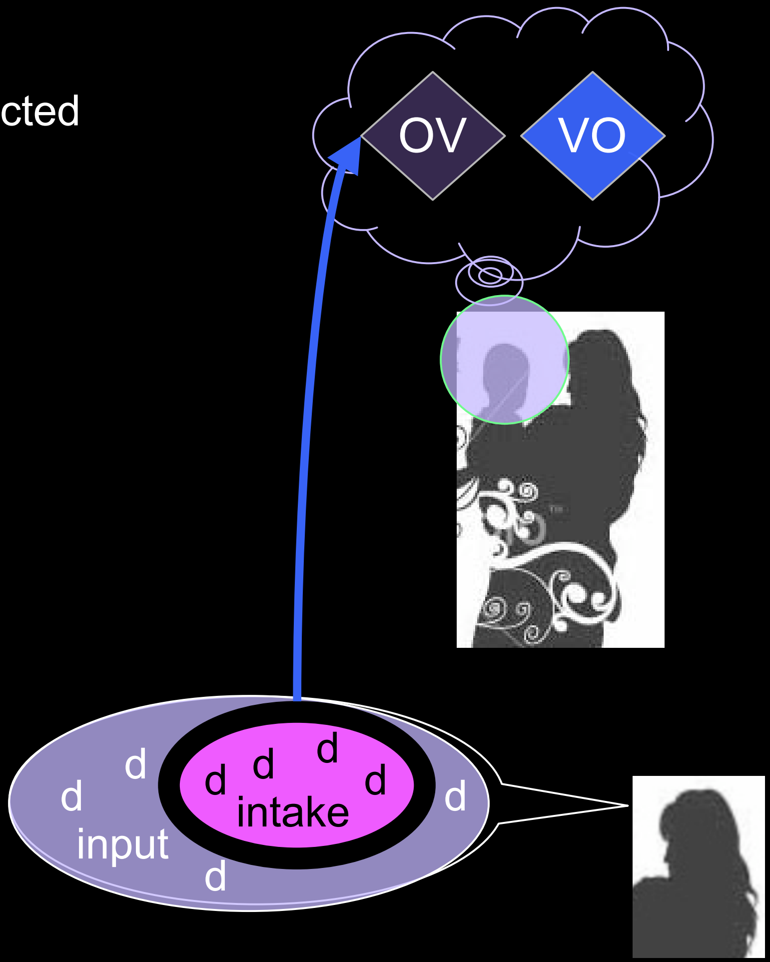
$p_{VO} = (p_{VOprev} * n) / (n + c)$

If VO data point

$p_{VO} = (p_{VOprev} * n + c) / (n + c)$

Important: Online update procedure
(psychological plausibility, given human memory)

Involves previous probability & expected amount of data in learning period



Model parameters:

c represents learner's confidence in data point (calibrated from data)

n represents quantity of intake (2000)

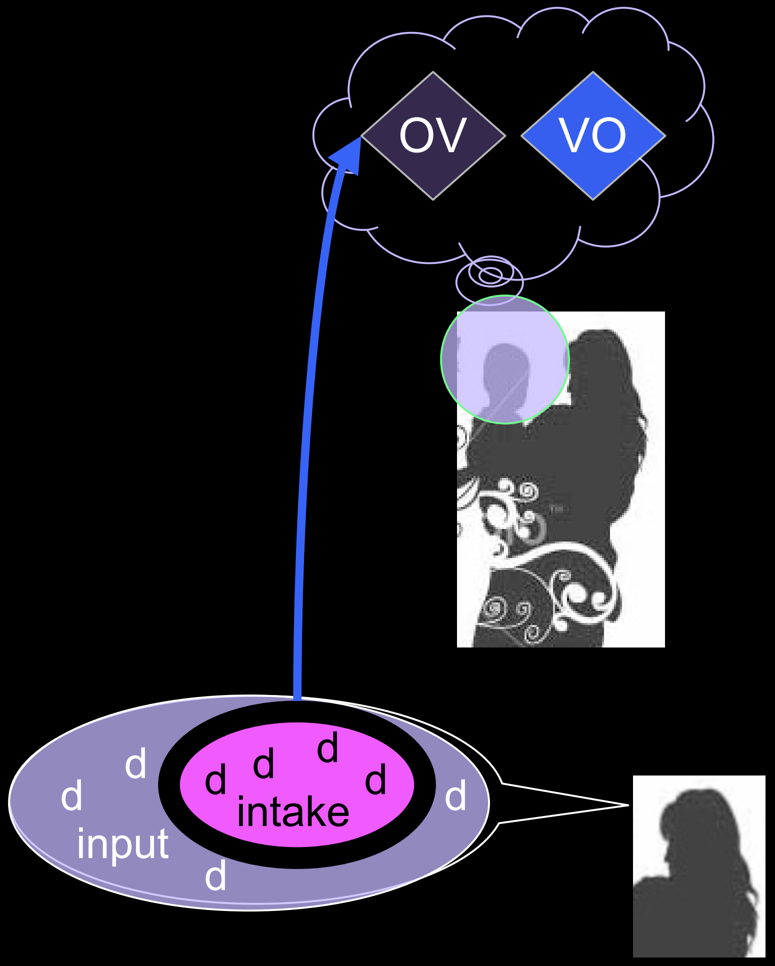
Individual-Level Learning Algorithm

(1) Initial $p_{VO} = 0.5$.

(2) Encounter data point from an average member of the population.

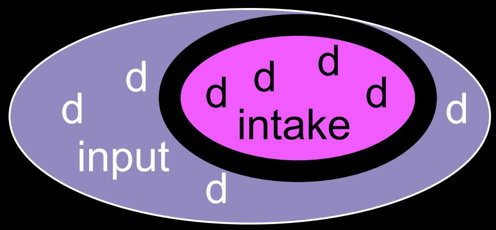
(3) If the data point is degree-0 and unambiguous, use update functions to shift hypothesis probabilities.

(4) Repeat (2-3) until the learning period is over, as determined by *n*.



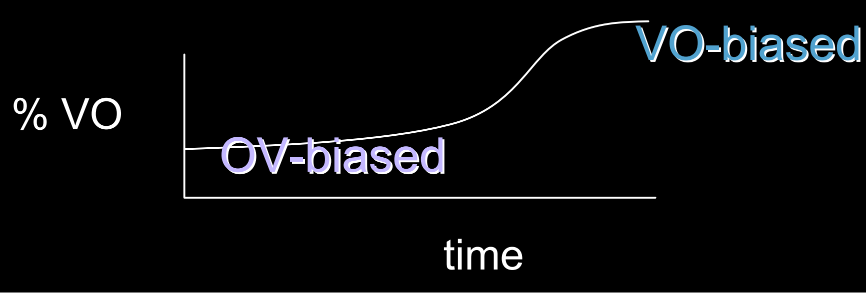
Biased Data Intake Distributions in Old English

p_{VO} shifts away from 0.5 when there is more of one data type in the intake than the other (advantage (Yang 2000) of one data type).



So the bias in the degree-0 unambiguous data distribution controls an individual's final p_{VO} in this model.

	OV Advantage in Unamb D0	
1000 A.D.	19.5%	OV-biased
1000-1150 A.D.	2.8%	
1200 A.D.	-2.7%	VO-biased



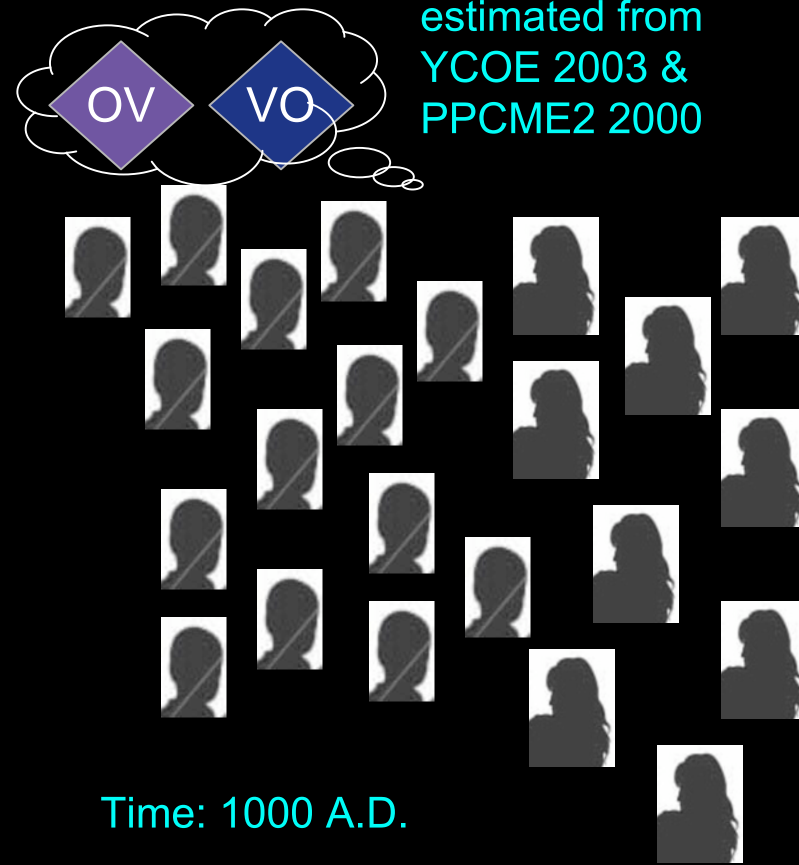
Population-Level Model

(1) Set the age range of the population from 0 to 60 years old and create 18,000 population members.

(2) Initialize the members of the population to the average p_{VO} at 1000 A.D. Set the time to 1000 A.D.

Population size estimated from population statistics of the time period (Koenigsberger & Briggs 1987)

Average p_{VO} estimated from YCOE 2003 & PPCME2 2000



Time: 1000 A.D.

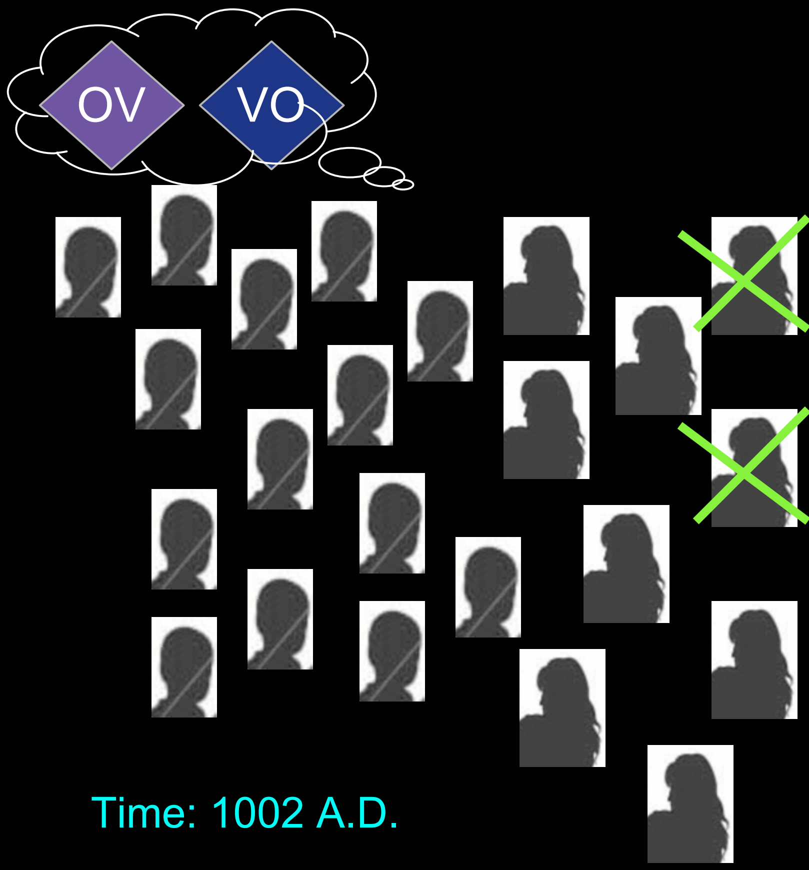
Population-Level Model

(1) Set the age range of the population from 0 to 60 years old and create 18,000 population members.

(2) Initialize the members of the population to the average p_{VO} at 1000 A.D. Set the time to 1000 A.D.

(3) Move forward 2 years.

(4) Members age 59-60 die off.



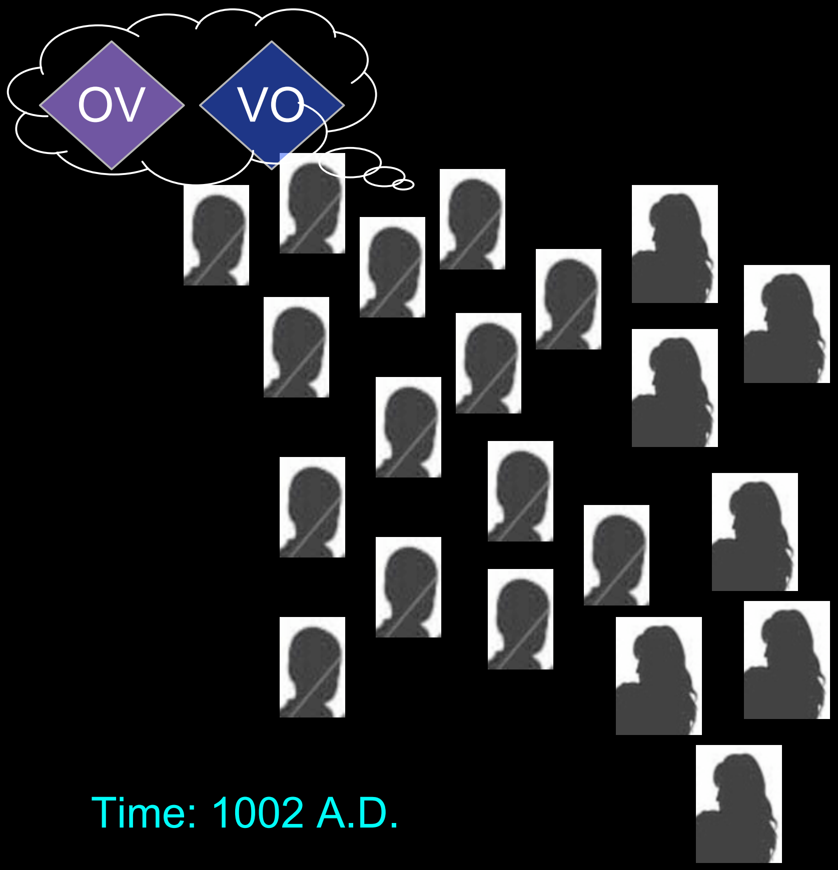
Population-Level Model

(1) Set the age range of the population from 0 to 60 years old and create 18,000 population members.

(2) Initialize the members of the population to the average p_{VO} at 1000 A.D. Set the time to 1000 A.D.

(3) Move forward 2 years.

(4) Members age 59-60 die off. The rest of the population ages 2 years.



Time: 1002 A.D.

Population-Level Model

(1) Set the age range of the population from 0 to 60 years old and create 18,000 population members.

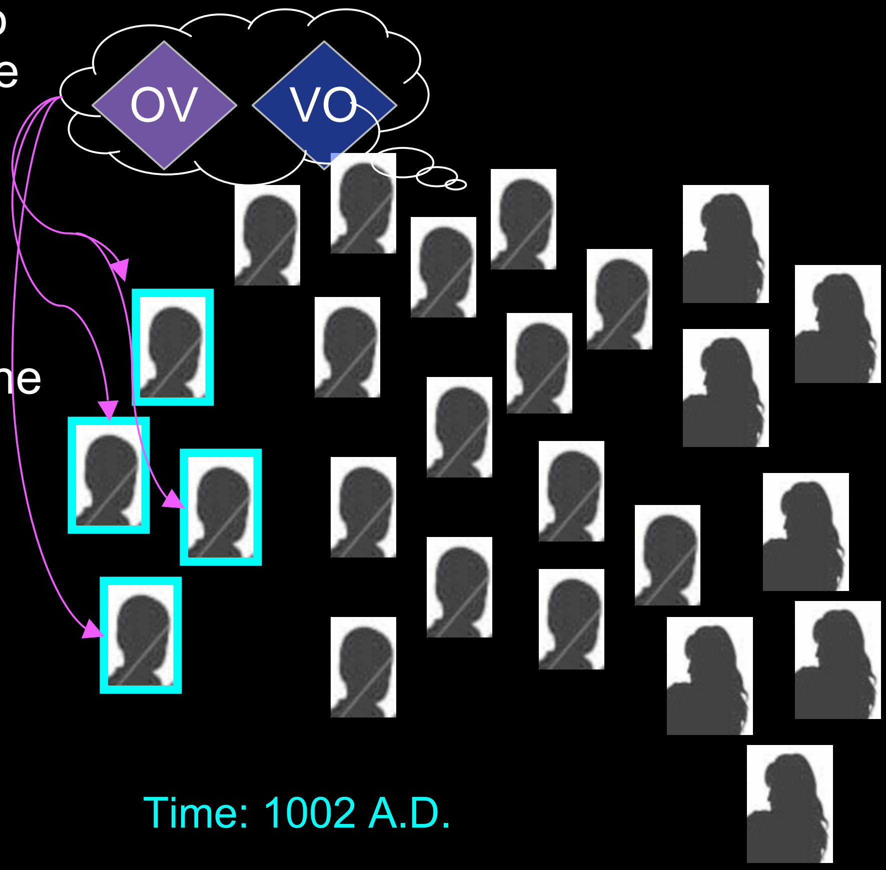
(2) Initialize the members of the population to the average p_{VO} at 1000 A.D. Set the time to 1000 A.D.

(3) Move forward 2 years.

(4) Members age 59-60 die off. The rest of the population ages 2 years.

(5) New members are born. These new members use the individual acquisition algorithm to set their p_{VO}.

Population growth rate estimated from population statistics of the time period (Koenigsberger & Briggs 1987)



Time: 1002 A.D.

Population-Level Model

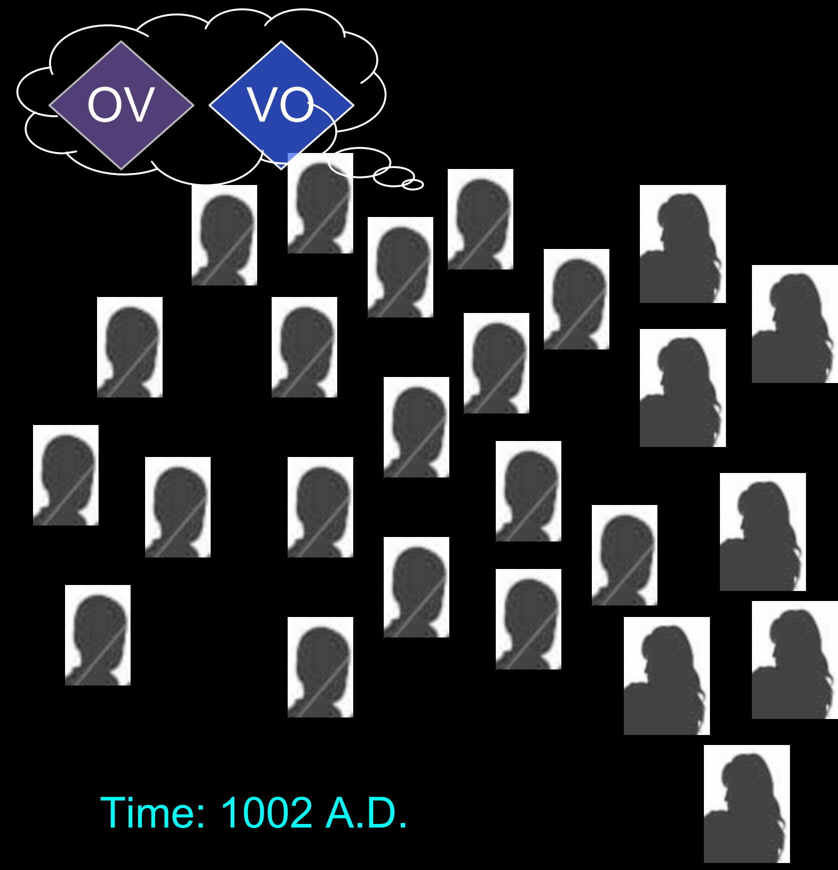
(1) Set the age range of the population from 0 to 60 years old and create 18,000 population members.

(2) Initialize the members of the population to the average p_{VO} at 1000 A.D. Set the time to 1000 A.D.

(3) Move forward 2 years.

(4) Members age 59-60 die off. The rest of the population ages 2 years.

(5) New members are born. These new members use the individual acquisition algorithm to set their p_{VO}.



Population-Level Model

(1) Set the age range of the population from 0 to 60 years old and create 18,000 population members.

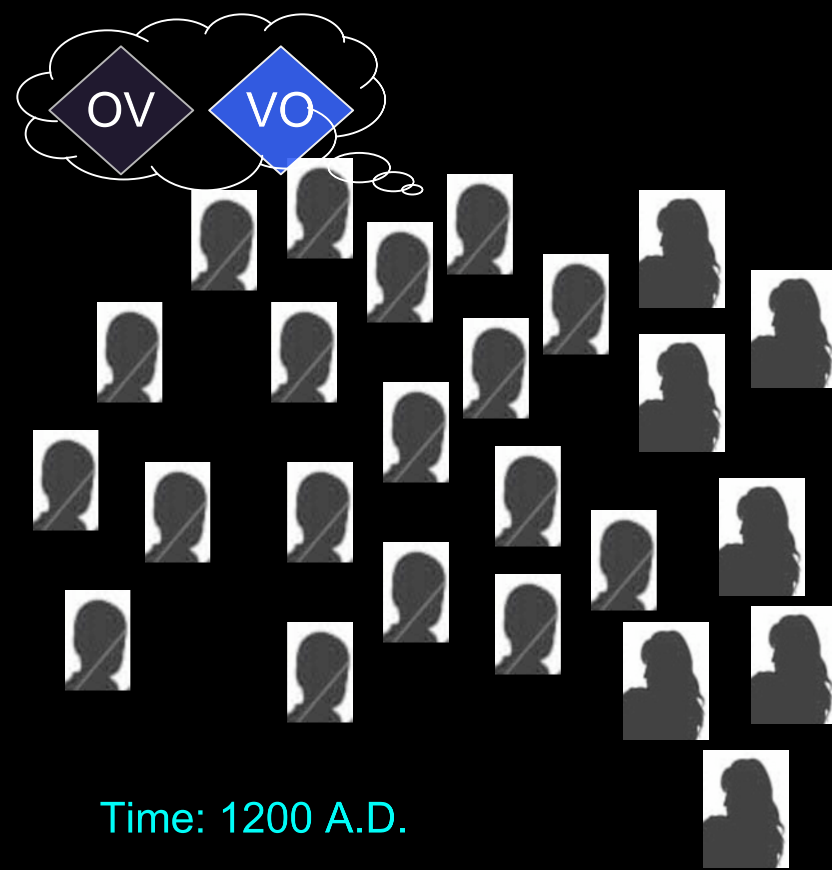
(2) Initialize the members of the population to the average p_{VO} at 1000 A.D. Set the time to 1000 A.D.

(3) Move forward 2 years.

(4) Members age 59-60 die off. The rest of the population ages 2 years.

(5) New members are born. These new members use the individual acquisition algorithm to set their p_{VO}.

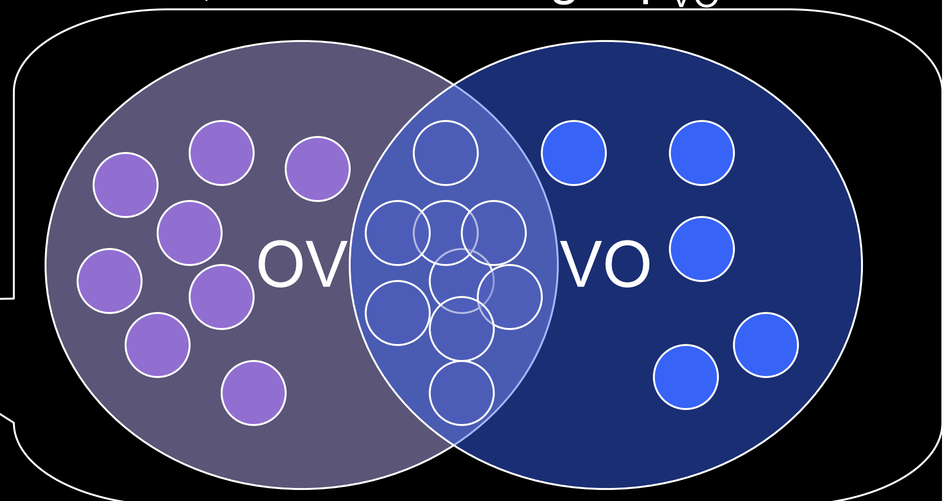
(6) Repeat steps (3-5) until the year 1200 A.D.



Empirical Grounding Issues:
What exactly is the underlying distribution?

Historical data used to initialize population's p_{VO} at 1000 A.D., calibrate population's p_{VO} between 1000 and 1150 A.D., and check target p_{VO} at 1200 A.D.

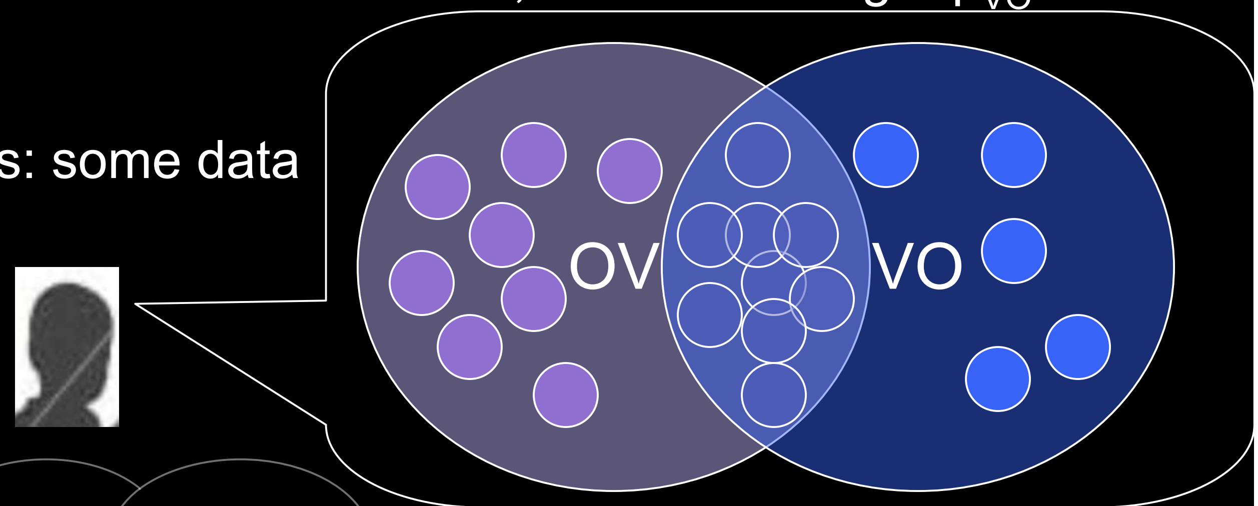
Historical data distributions: some data are ambiguous

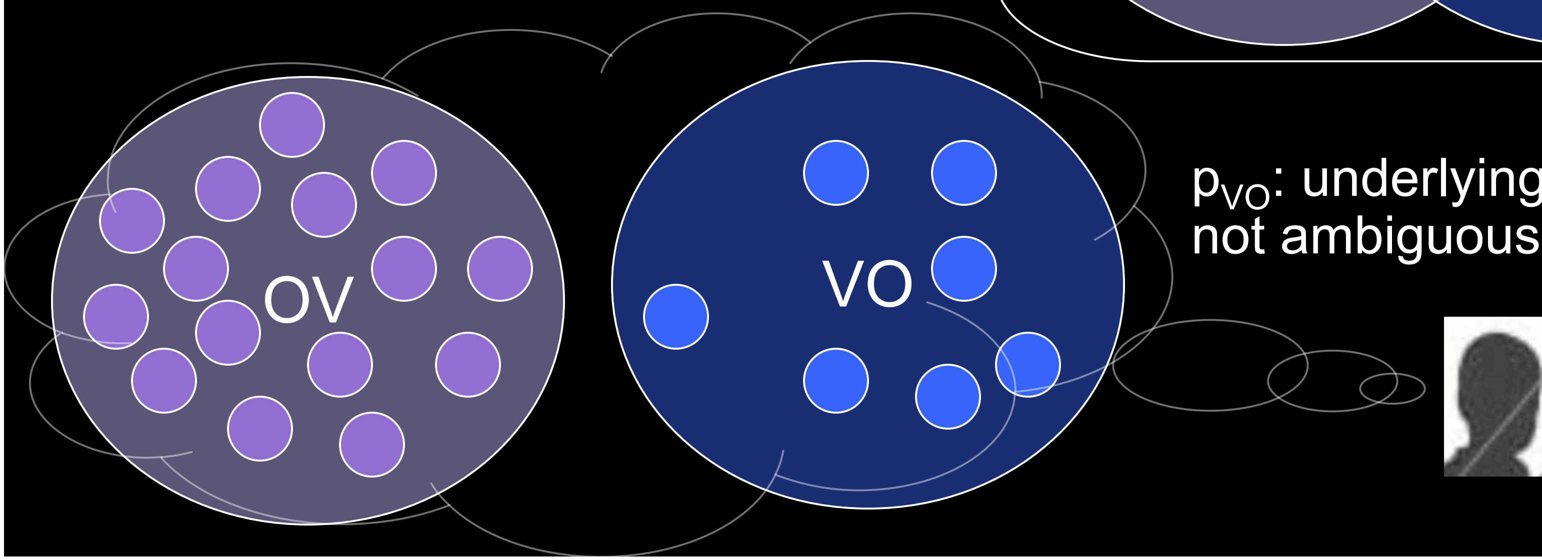


Empirical Grounding Issues: What exactly is the underlying distribution?

Historical data used to initialize population's p_{VO} at 1000 A.D., calibrate population's p_{VO} between 1000 and 1150 A.D., and check target p_{VO} at 1200 A.D.

Historical data distributions: some data are ambiguous

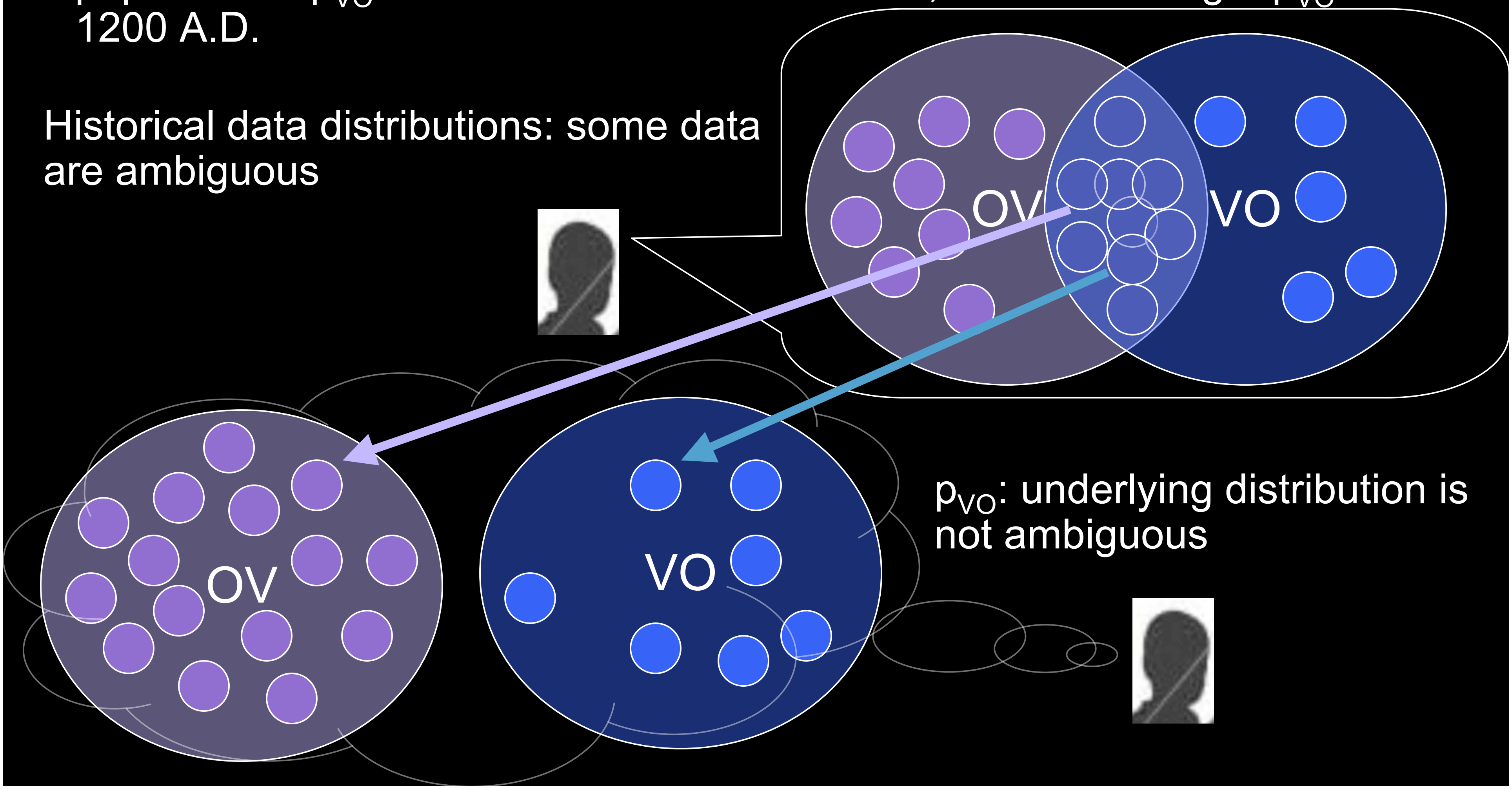




p_{VO}: underlying distribution is not ambiguous

Empirical Grounding Issues: What exactly is the underlying distribution?

Historical data used to initialize population's p_{VO} at 1000 A.D., calibrate population's p_{VO} between 1000 and 1150 A.D., and check target p_{VO} at 1200 A.D.



Historical data distributions: some data are ambiguous

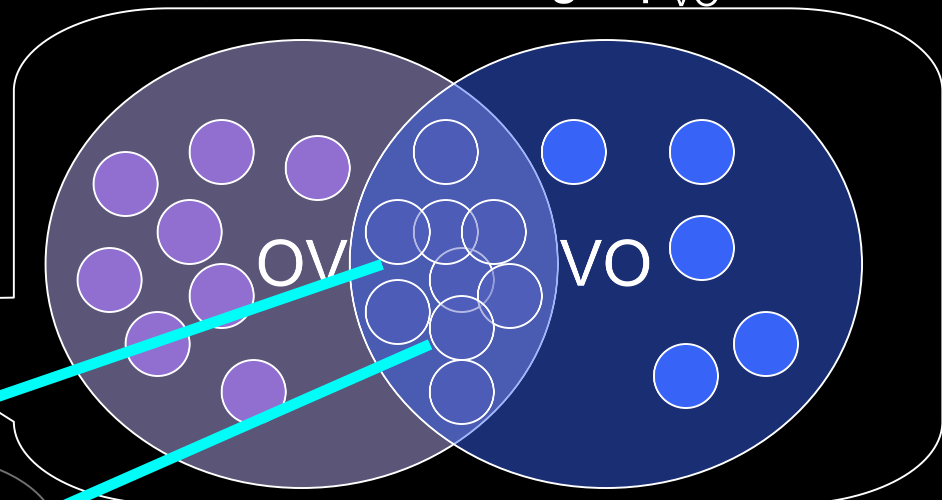
p_{VO}: underlying distribution is not ambiguous

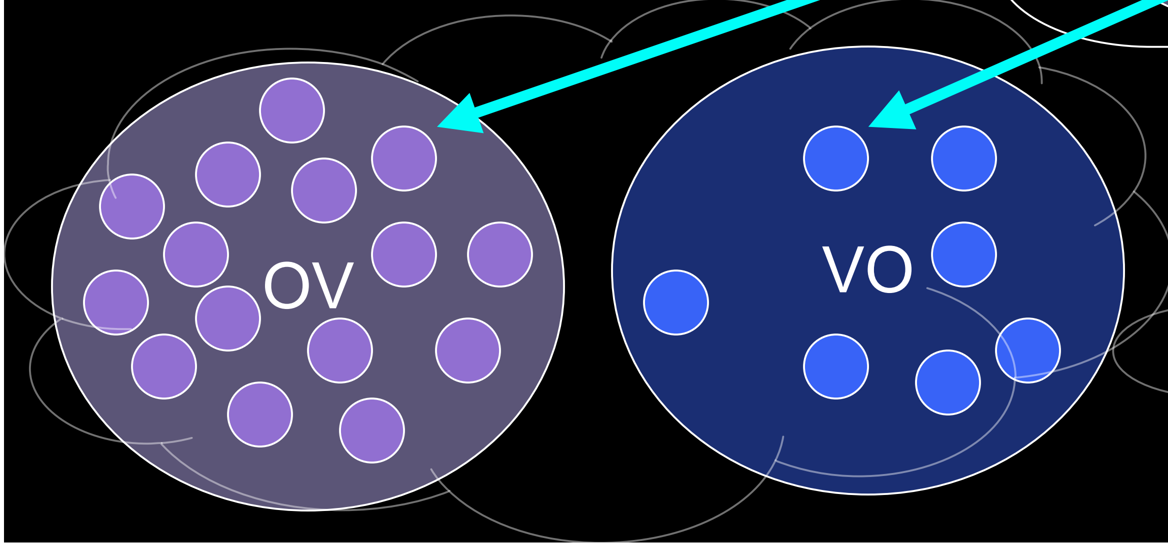
Empirical Grounding Issues: What exactly is the underlying distribution?

Historical data used to initialize population's p_{VO} at 1000 A.D., calibrate population's p_{VO} between 1000 and 1150 A.D., and check target p_{VO} at 1200 A.D.

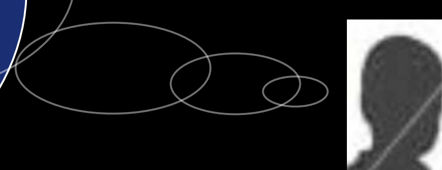
Historical data distributions: some data are ambiguous

How do we figure out what the ambiguous data are?





p_{VO}: underlying distribution is not ambiguous



Empirical Grounding Issues: What exactly is the underlying distribution?

(YCOE and PPCME2 Corpora)
% Ambiguous Utterances

	Degree-0 % Ambiguous	Degree-1 % Ambiguous
1000 A.D.	76%	28%
1000 - 1150 A.D.	80%	25%
1200 A.D.	71%	10%

Observations:

(1) Degree-1 data less ambiguous than degree-0 data.

Empirical Grounding Issues: What exactly is the underlying distribution?

(YCOE and PPCME2 Corpora)
% Advantage

	OV Advantage in Unamb D0	OV Advantage in Unamb D1
1000 A.D.	19.5%	41.7%
1000-1150 A.D.	2.8%	28.7%
1200 A.D.	-2.7%	-45.2%

Observations:
(1) Degree-1 data less ambiguous than degree-0 data.
(2) Advantage is magnified in degree-1.

Empirical Grounding Issues: What exactly is the underlying distribution?

Observations:

(1) Degree-1 data less ambiguous than degree-0 data.

(2) Advantage is magnified in degree-1.

Assumption: Ambiguous data distorts underlying distribution.

Assumption: degree-1 distribution less distorted from underlying distribution.

Empirical Grounding Issues: What exactly is the underlying distribution?

Observations:

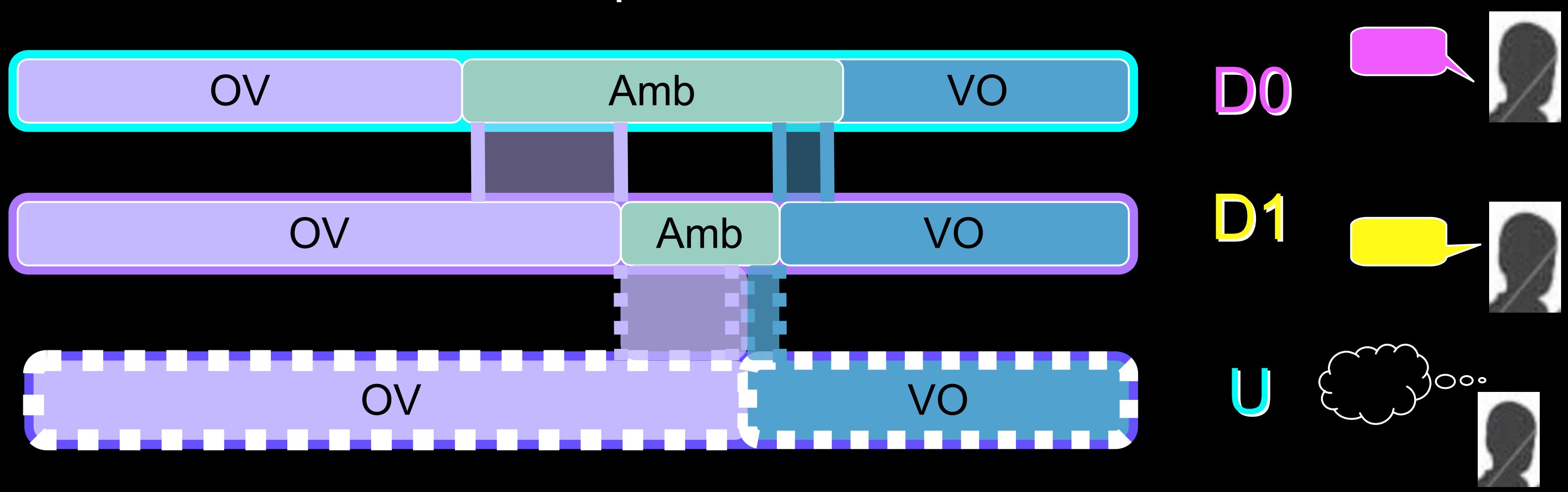
(1) Degree-1 data less ambiguous than degree-0 data.

(2) Advantage is magnified in degree-1.

Assumption: Ambiguous data distorts underlying distribution.

Assumption: degree-1 distribution less distorted from underlying distribution.

Plan of Action: Use the difference in distortion between the degree-0 and degree-1 unambiguous data distributions to estimate the difference in distortion between the degree-1 distribution and the underlying unambiguous data distribution in a speaker's mind.



Empirical Grounding Issues: What exactly is the underlying distribution?

Observations:
(1) Degree-1 data less ambiguous than degree-0 data.
(2) Advantage is magnified in degree-1.

Assumption: Ambiguous data distorts underlying distribution
Assumption: degree-1 distribution less distorted from underlying distribution.

$$\frac{\gamma * \mathbf{d0} - \mathbf{u1d1'}}{\gamma * \mathbf{d0}} = \mathbf{Ld1tod0} * \frac{\mathbf{ad1'} - (\gamma * \mathbf{d0} - \mathbf{u1d1'})}{\mathbf{u2d1'} + \mathbf{ad1'} - (\gamma * \mathbf{d0} - \mathbf{u1d1'})}$$

known quantities

$\gamma =$ underlying pvo

$\boxed{\mathbf{d}0 = \text{total degree-0 data}, \mathbf{d}1 = \text{total degree-1 data}}$

$\mathbf{u1d1'}$ = normalized unambiguous OV degree-1 data
$\mathbf{u}2d1'$ = normalized unambiguous VO degree-1 data
$\mathbf{Ld1tod0}$ = loss ratio (OV/VO) from degree-1 to degree-0 distribution
$\mathbf{a}d1'$ = normalized ambiguous degree-1 data

derived quantities

$$\gamma = \frac{-(\mathbf{d0})(\mathbf{d0} + \mathbf{u1d1'} - \mathbf{Ld1tod0} * (\mathbf{ad1'} + \mathbf{u1d1'}))}{2(\mathbf{Ld1tod0} + 1)(\mathbf{d0}^2)}$$
$$+/- \frac{\sqrt{((\mathbf{d0})(\mathbf{d0} + \mathbf{u1d1'} - \mathbf{Ld1tod0} * (\mathbf{ad1'} + \mathbf{u1d1'})))^2 - 4(\mathbf{Ld1tod0} + 1)(\mathbf{d0}^2)((\mathbf{-1})(\mathbf{d0} * \mathbf{u1d1'}))}}{2(\mathbf{Ld1tod0} + 1)(\mathbf{d0}^2)}$$

Empirical Grounding Issues: What exactly is the underlying distribution?

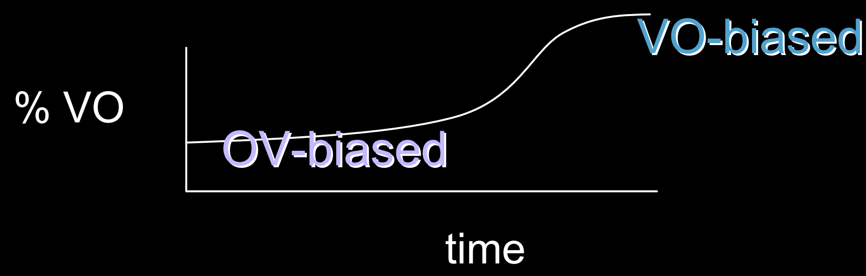
Observations:

(1) Degree-1 data less ambiguous than degree-0 data.

(2) Advantage is magnified in degree-1.

Assumption: Ambiguous data distorts underlying distribution

Assumption: degree-1 distribution less distorted from underlying distribution.



	(Initialization) 1000 A.D.	(Calibration) 1000-1150 A.D.	(Termination) 1200 A.D.
Average p_{VO}	0.234	0.310	0.747

OV-biased VO-biased

Empirical Grounding Issues:
What exactly is the underlying distribution?

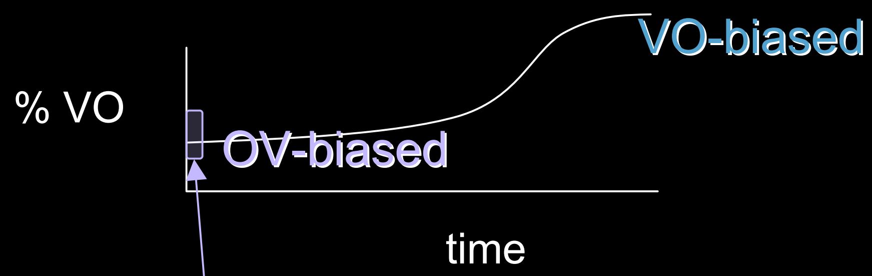
Observations:

(1) Degree-1 data less ambiguous than degree-0 data.

(2) Advantage is magnified in degree-1.

Assumption: Ambiguous data distorts underlying distribution

Assumption: degree-1 distribution less distorted from underlying distribution.



	(Initialization) 1000 A.D.	(Calibration) 1000-1150 A.D.	(Termination) 1200 A.D.
Average p_{VO}	0.234	0.310	0.747

OV-biased VO-biased

Empirical Grounding Issues: What exactly is the underlying distribution?

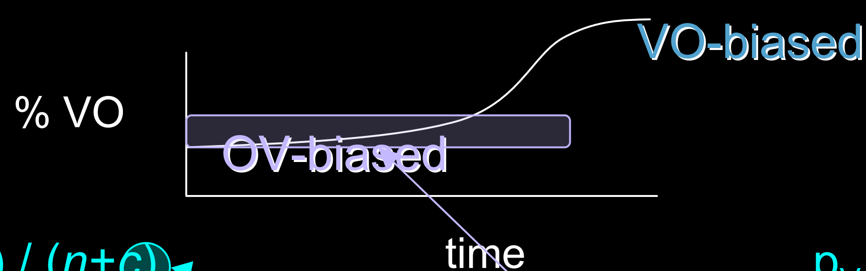
Observations:

(1) Degree-1 data less ambiguous than degree-0 data.

(2) Advantage is magnified in degree-1.

Assumption: Ambiguous data distorts underlying distribution

Assumption: degree-1 distribution less distorted from underlying distribution.



$p_{VO} = (p_{VOprev} * n) / (n+c)$

$p_{VO} = (p_{VOprev} * n+c) / (n+c)$

	(Initialization) 1000 A.D.	(Calibration) 1000-1150 A.D.	(Termination) 1200 A.D.
Average p_{VO}	0.234	0.310	0.747

OV-biased VO-biased

Empirical Grounding Issues: What exactly is the underlying distribution?

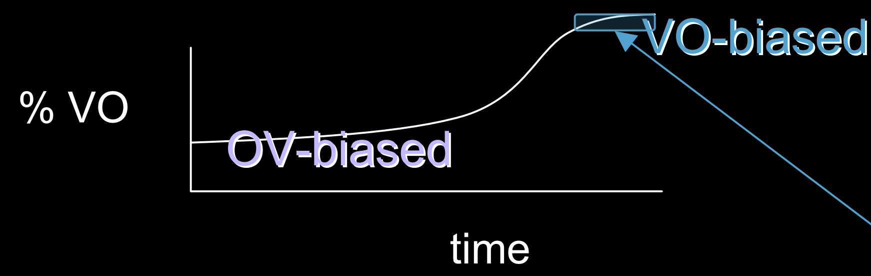
Observations:

(1) Degree-1 data less ambiguous than degree-0 data.

(2) Advantage is magnified in degree-1.

Assumption: Ambiguous data distorts underlying distribution

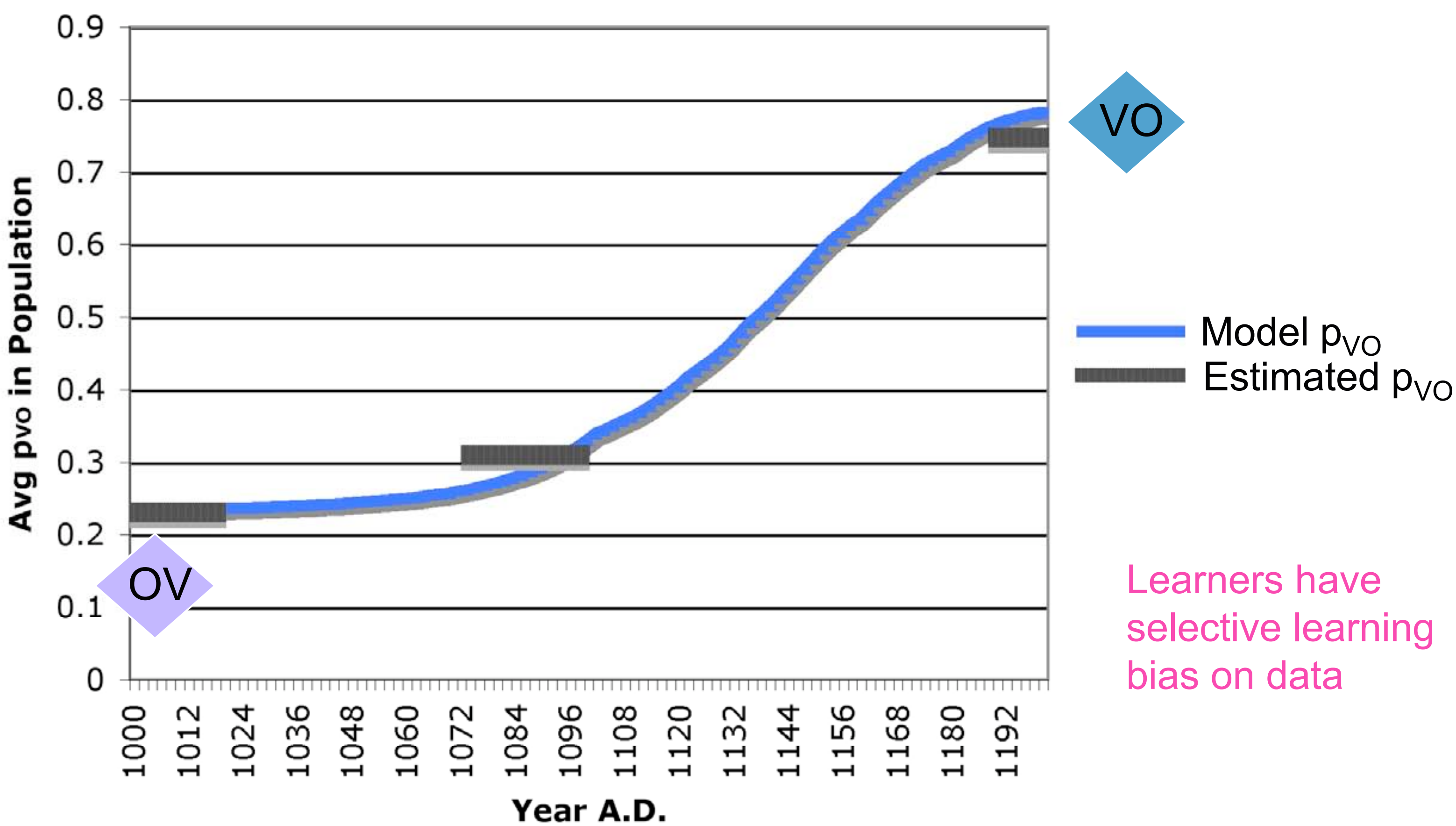
Assumption: degree-1 distribution less distorted from underlying distribution.



	(Initialization) 1000 A.D.	(Calibration) 1000-1150 A.D.	(Termination) 1200 A.D.
Average p_{VO}	0.234	0.310	0.747

OV-biased VO-biased

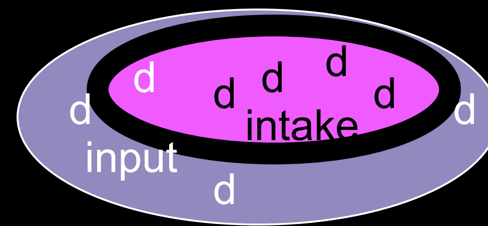
Linguistic Evolution: Change at the Historically-Attested Rate

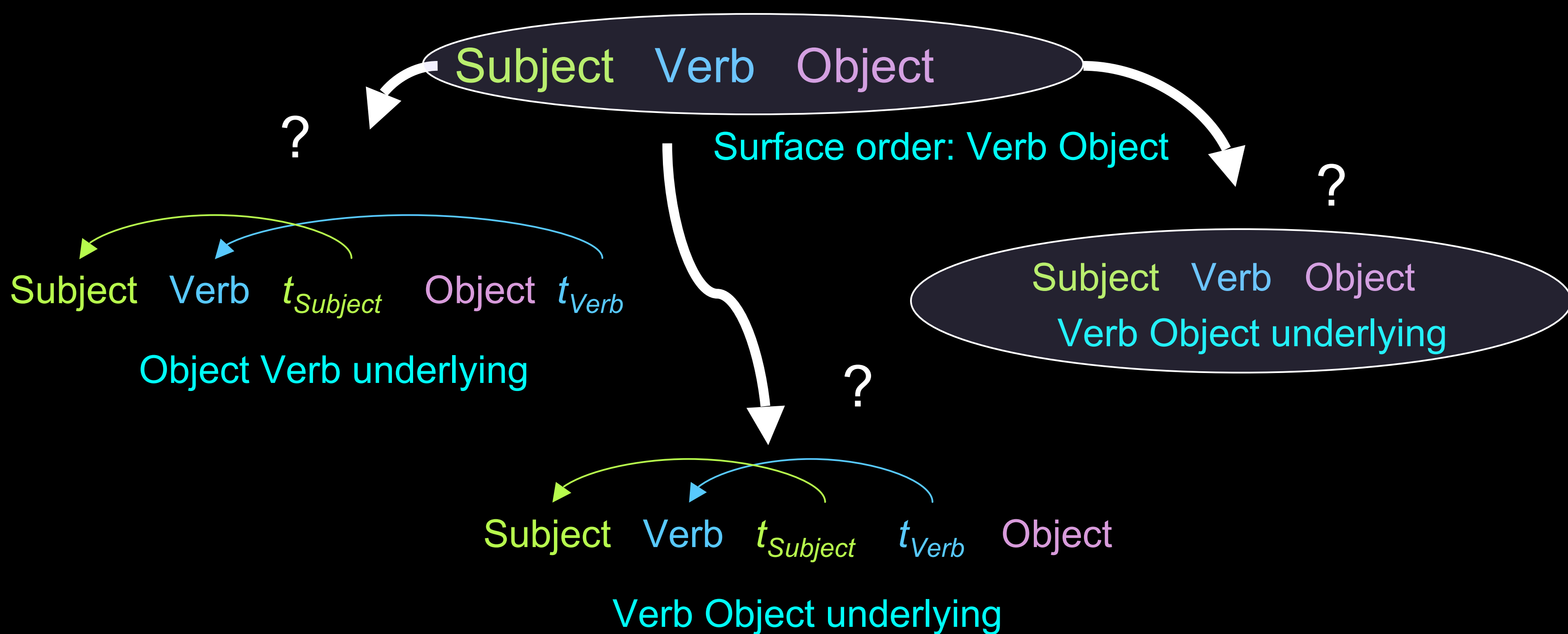


Learners have selective learning bias on data

Linguistic Evolution:
Different Individual-Level Learning

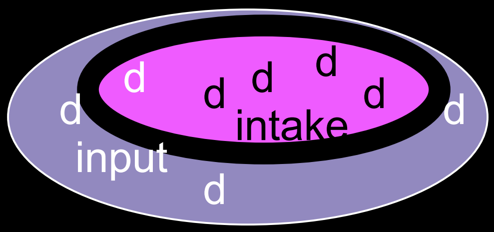
Learner uses ambiguous data. Strategy for learning: assume surface order is actual order. (Fodor 1998)





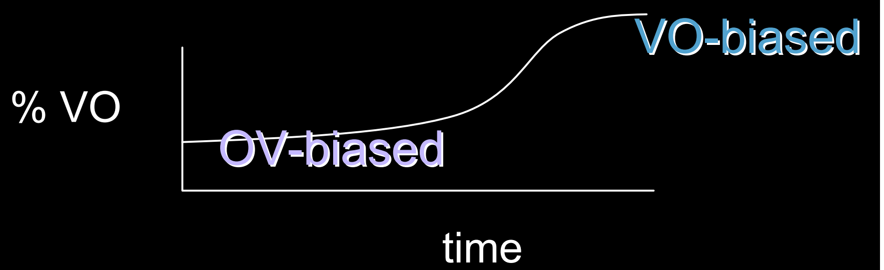
Linguistic Evolution: Different Individual-Level Learning

Learner uses ambiguous data. Strategy for learning: assume surface order is actual order. (Fodor 1998)



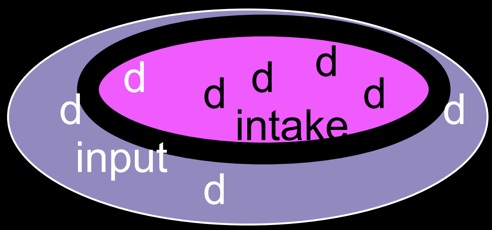
Advantage in intake determines learner's ending distribution between OV and VO order.

Need this trajectory



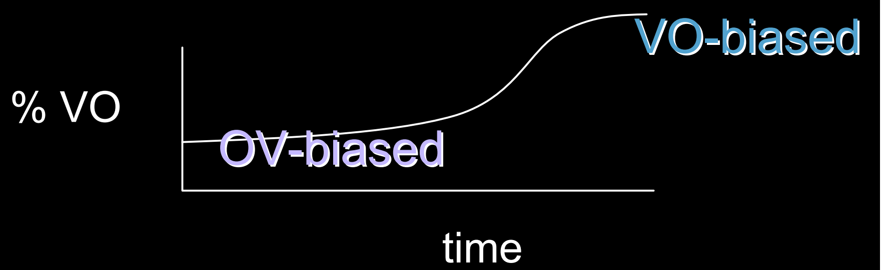
Linguistic Evolution: Different Individual-Level Learning

Learner uses ambiguous data. Strategy for learning: assume surface order is actual order. (Fodor 1998)



Advantage in intake determines learner's ending distribution between OV and VO order.

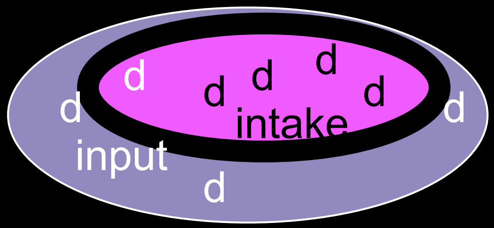
Need this trajectory



	Degree-0 OV Advantage
1000 A.D.	-21.0%
1000 - 1150 A.D.	-26.9%
1200 A.D.	-21.8%

Linguistic Evolution: Different Individual-Level Learning

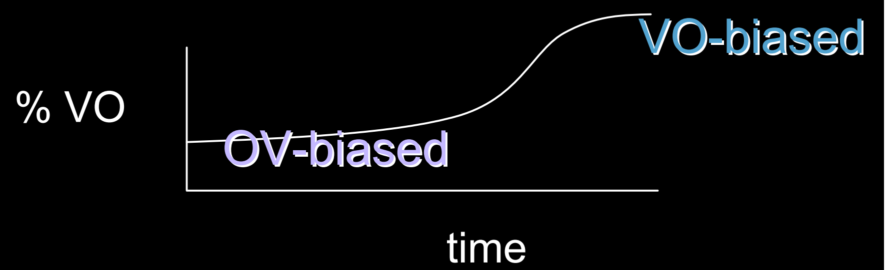
Learner uses ambiguous data. Strategy for learning: assume surface order is actual order. (Fodor 1998)



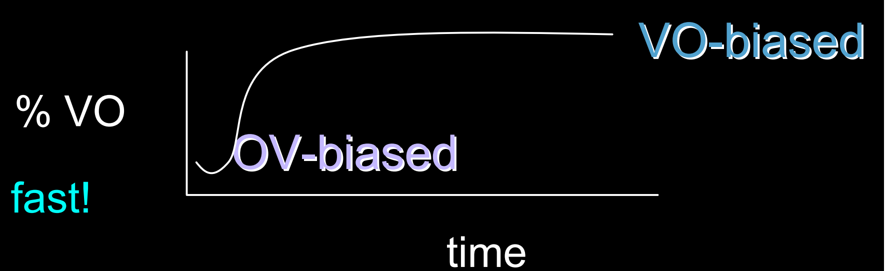
Advantage in intake determines learner's ending distribution between OV and VO order.

	Degree-0 OV Advantage
1000 A.D.	-21.0%
1000 - 1150 A.D.	-26.9%
1200 A.D.	-21.8%

Need this trajectory



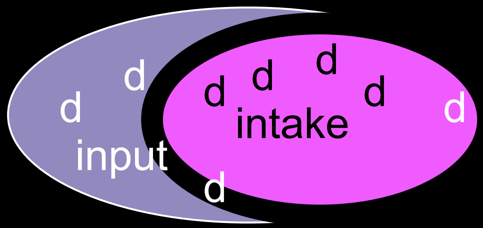
Problem: VO-biased
all the way through, even at 1000 A.D.



Change is too fast!

Linguistic Evolution: Different Individual-Level Learning

Learner uses degree-0 and degree-1 unambiguous data.



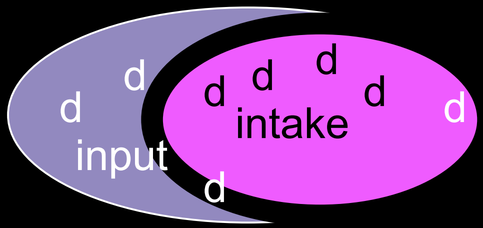
(YCOE and PPCME2 Corpora)
% Advantage

	OV Advantage in Unamb D0	OV Advantage in Unamb D1
1000 A.D.	19.5%	41.7%
1000-1150 A.D.	2.8%	28.7%
1200 A.D.	-2.7%	-45.2%

Very strongly OV-biased before 1150 A.D.

Linguistic Evolution:
Different Individual-Level Learning

Learner uses degree-0 and degree-1 unambiguous data.



(YCOE and PPCME2 Corpora)

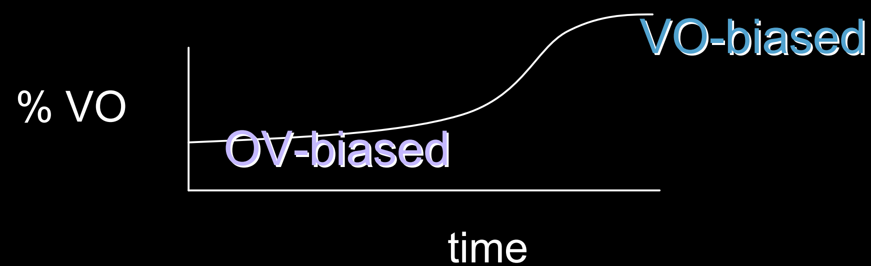
% Advantage

	OV Advantage in Unamb D0	OV Advantage in Unamb D1
1000 A.D.	19.5%	41.7%
1000-1150 A.D.	2.8%	28.7%
1200 A.D.	-2.7%	-45.2%

Very strongly OV-biased before 1150 A.D.

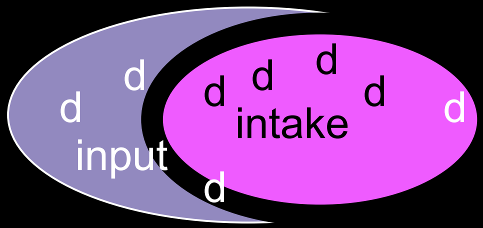
But population must become VO-biased.

Need this trajectory



Linguistic Evolution: Different Individual-Level Learning

Learner uses degree-0 and degree-1 unambiguous data.



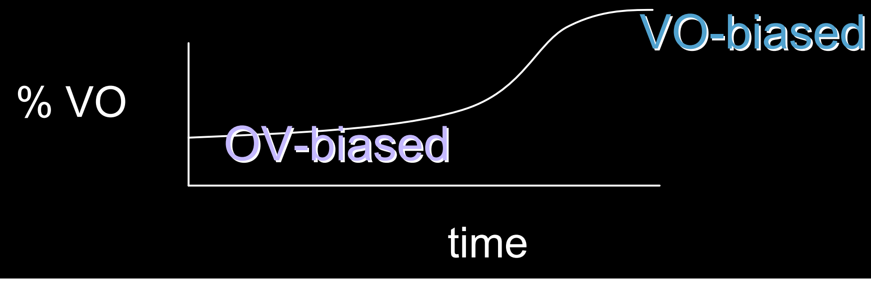
(YCOE and PPCME2 Corpora)
% Advantage

	OV Advantage in Unamb D0	OV Advantage in Unamb D1
1000 A.D.	19.5%	41.7%
1000-1150 A.D.	2.8%	28.7%
1200 A.D.	-2.7%	-45.2%

Very strongly OV-biased before 1150 A.D.

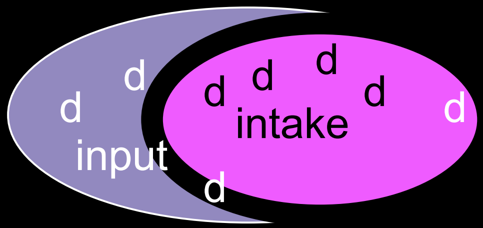
Need this trajectory

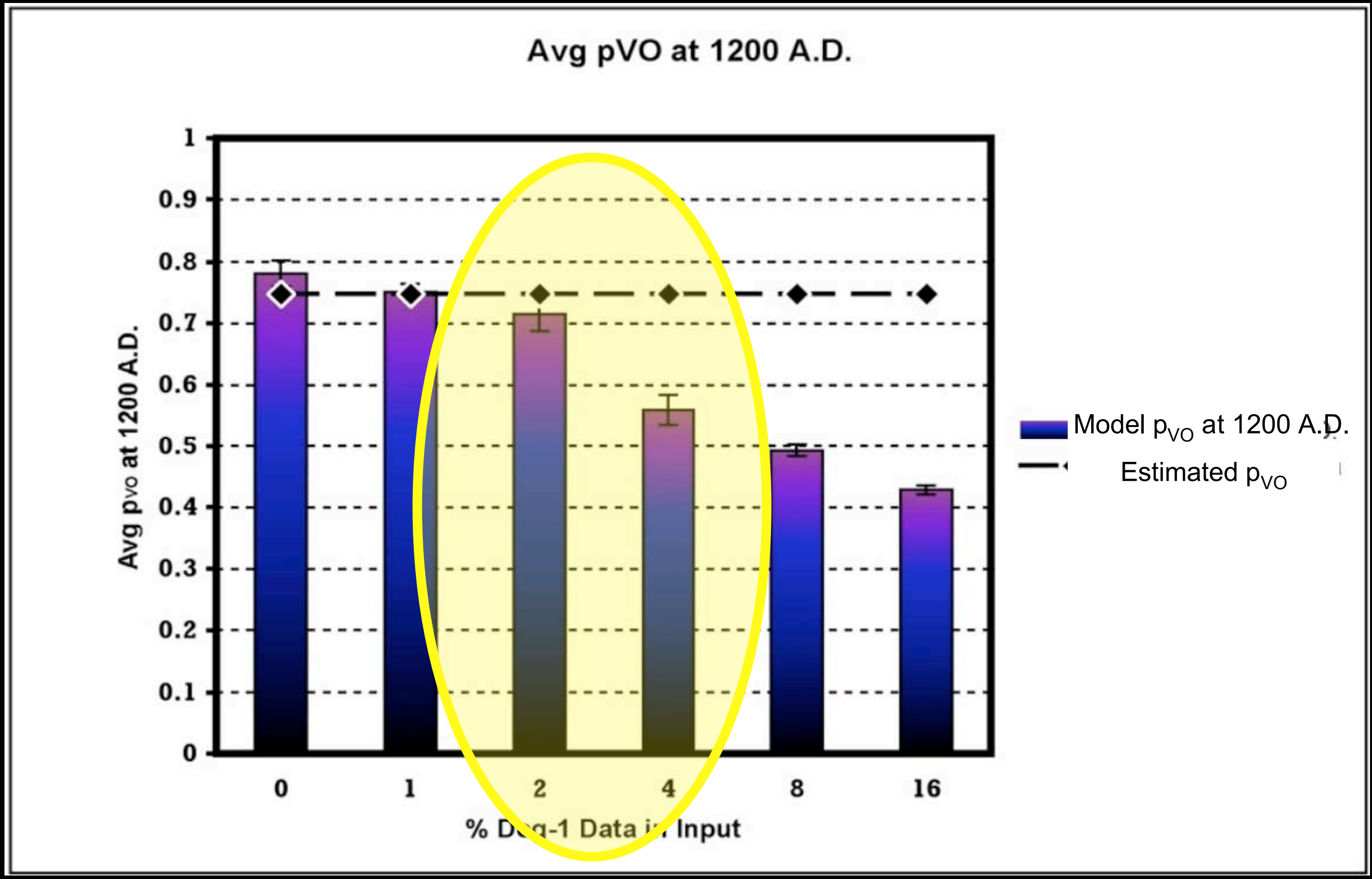
Can a population learning from degree-1 data make the change to VO-biased?



Linguistic Evolution: Different Individual-Level Learning

Learner uses degree-0 and degree-1 unambiguous data.

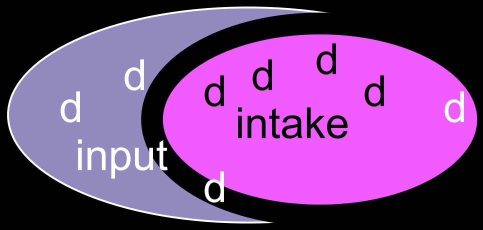


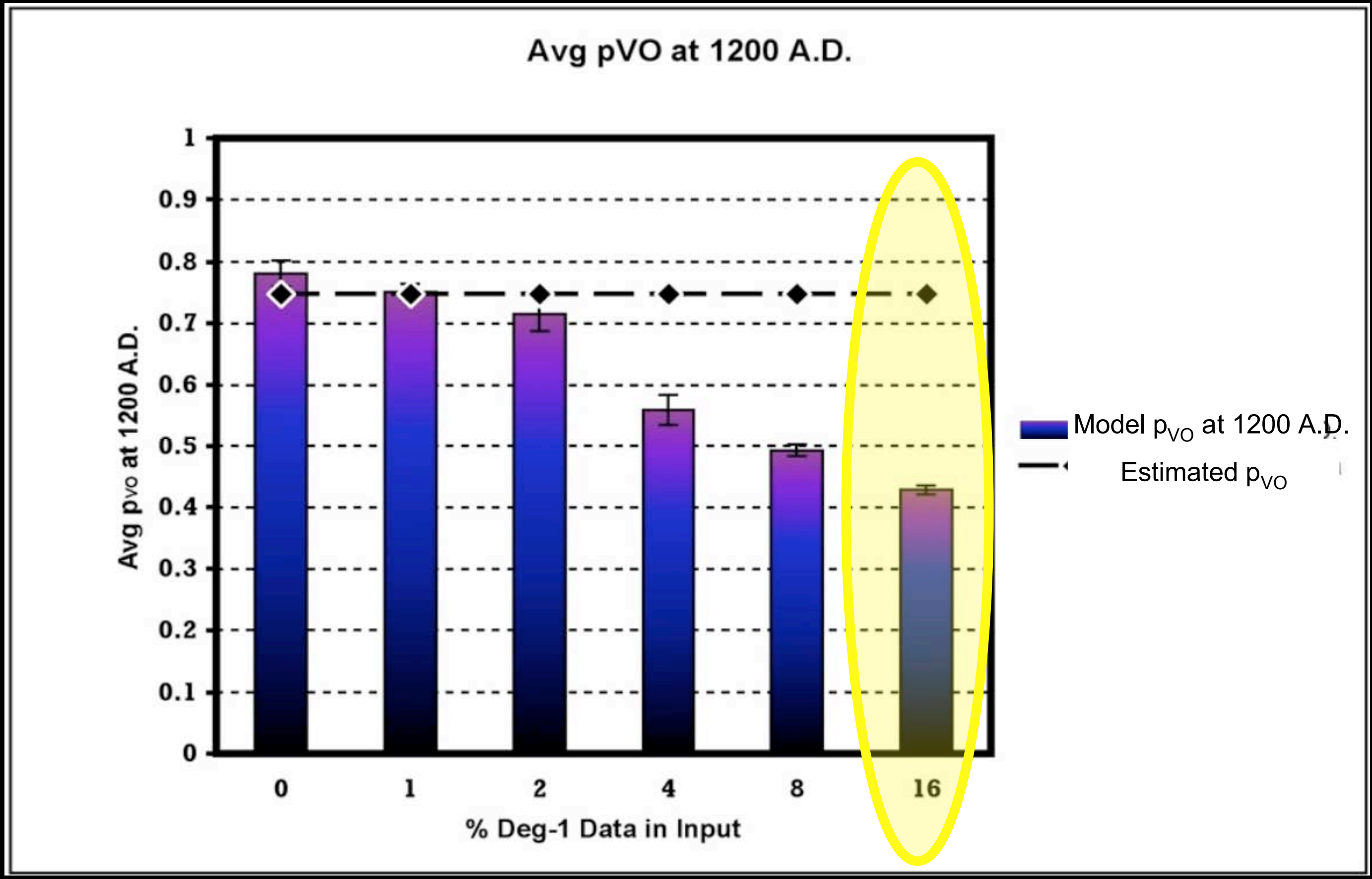


Modeled population can change at the right rate only if input contains less than 4% degree-1 data - otherwise, change is too slow for learners not using a degree-0 bias.

Linguistic Evolution: Different Individual-Level Learning

Learner uses degree-0 and degree-1 unambiguous data.



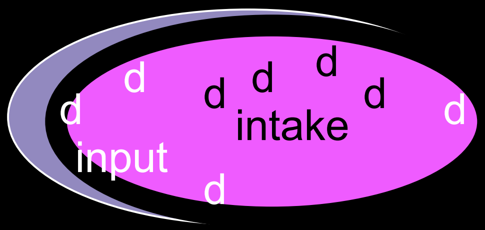


Estimates from modern English child-directed speech: Input consists of ~16% degree-1 data.

Prognosis: Change would be too slow without a degree-0 bias for individual learners.

Linguistic Evolution: Different Individual-Level Learning

Learner uses degree-0 and degree-1 data, and learns from ambiguous data.

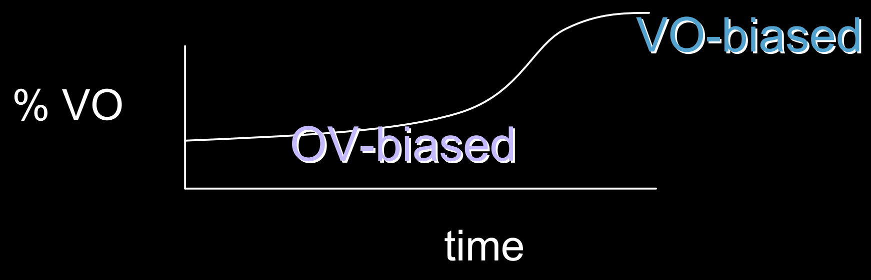


(YCOE and PPCME2 Corpora)

% Advantage

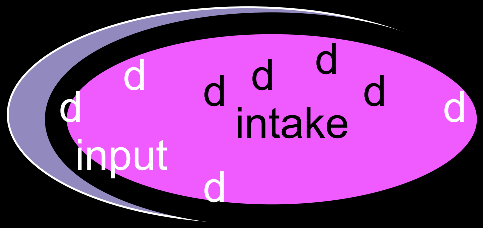
	OV Advantage in D0	OV Advantage in D1
1000 A.D.	-21.0%	28.1%

Need this trajectory



Linguistic Evolution: Different Individual-Level Learning

Learner uses degree-0 and degree-1 data, and learns from ambiguous data.

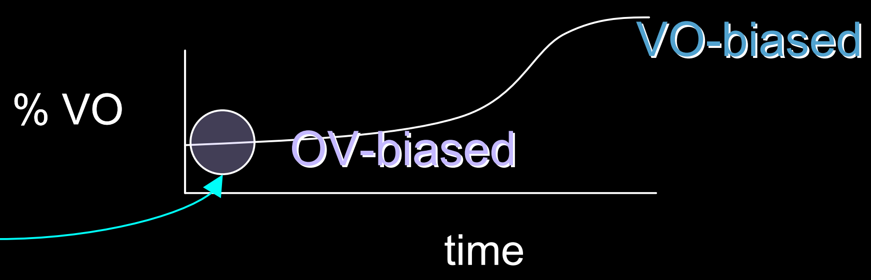


(YCOE and PPCME2 Corpora)
% Advantage

	OV Advantage in D0	OV Advantage in D1
1000 A.D.	-21.0%	28.1%

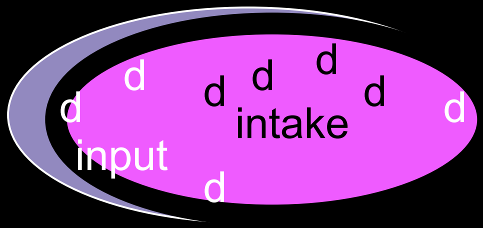
Need this trajectory

Population must remain OV-biased at 1000 A.D.



Linguistic Evolution: Different Individual-Level Learning

Learner uses degree-0 and degree-1 data, and learns from ambiguous data.

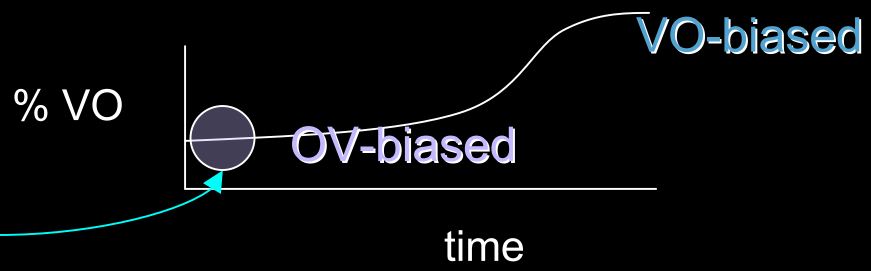


(YCOE and PPCME2 Corpora)
% Advantage

	OV Advantage in D0	OV Advantage in D1
1000 A.D.	-21.0%	28.1%

Need this trajectory

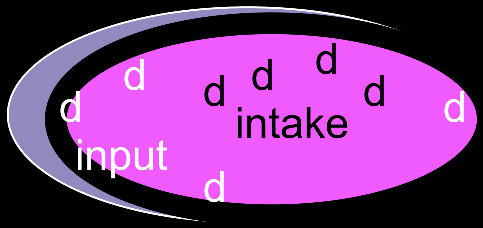
Population must remain OV-biased at 1000 A.D.



To do this, advantage in intake must be for OV order at 1000 A.D.
Otherwise, population changes too quickly to VO-biased distribution.

Linguistic Evolution: Different Individual-Level Learning

Learner uses degree-0 and degree-1 data, and learns from ambiguous data.



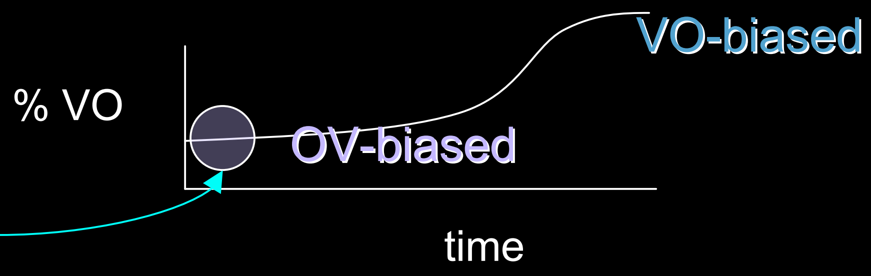
(YCOE and PPCME2 Corpora)

% Advantage

	OV Advantage in D0	OV Advantage in D1
1000 A.D.	-21.0%	28.1%

Need this trajectory

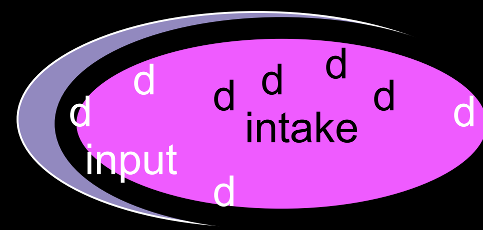
Population must remain OV-biased at 1000 A.D.



Requirement for OV advantage at 1000 A.D.: 43% of input is degree-1 data

Linguistic Evolution: Different Individual-Level Learning

Learner uses degree-0 and degree-1 data, and learns from ambiguous data.

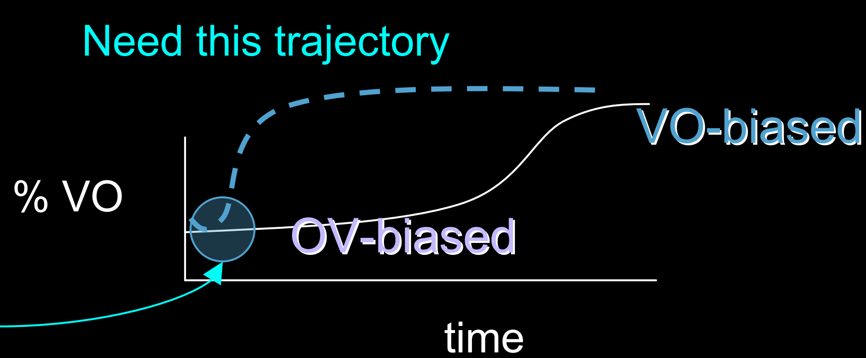


(YCOE and PPCME2 Corpora)

% Advantage

	OV Advantage in D0	OV Advantage in D1
1000 A.D.	-21.0%	28.1%

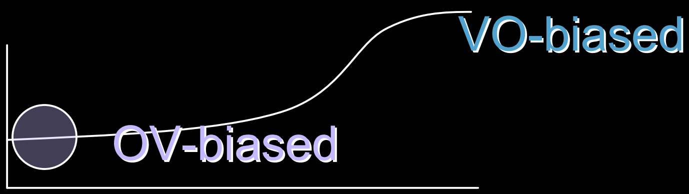
Population must remain OV-biased at 1000 A.D.

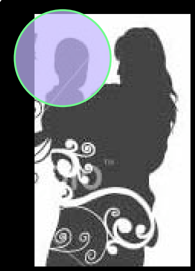


Requirement for OV advantage at 1000 A.D.: 43% of input is degree-1 data …but estimates show only ~16% of it is. Change will be too fast.

Linguistic Evolution: Summary

Some cases where linguistic evolution is driven by individual-level learning. Suggested example: Old English word order.



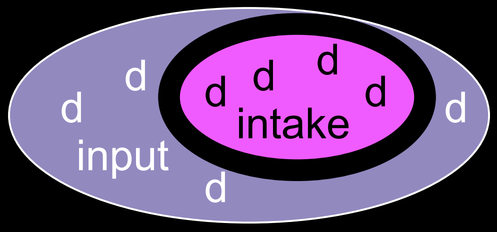


Individual-level learning: can involve selective learning biases, with strong effects on rate of linguistic change within a population.

Individual-Level Selective Learning:

(1) unambiguous data

(2) degree-0 data



Additional point: linguistic evolution can inform us about the nature of individual learning.

Linguistic Evolution: Open Questions

(1) If we add complexity to the population model, do we still need these individual-level selective learning biases?

Weight data points in individual intake using various factors:

(a) spatial location of speaker with respect to learner

(b) social status of speaker

(c) speaker's relation to learner (family, friend, stranger)

(d) context of data point (social context, linguistic context)

(2) Are these learning biases necessary if we look at other language changes where individual-level learning is thought to be the main factor driving change at the population-level?

Learning-Driven Linguistic Evolution: Take-Home Messages

(1) Correct population-level behavior can result from correct individual-level learning behavior in some cases (small discrepancies compounded over time).

Learning-Driven Linguistic Evolution: Take-Home Messages

(1) Correct population-level behavior can result from correct individual-level learning behavior in some cases (small discrepancies compounded over time).

(2) In the case study examined here, linguistic evolution occurs at the correct rate only when learners employ selective learning biases that cause them to use only a subset of the available data.

Learning-Driven Linguistic Evolution: Take-Home Messages

(1) Correct population-level behavior can result from correct individual-level learning behavior in some cases (small discrepancies compounded over time).

(2) In the case study examined here, linguistic evolution occurs at the correct rate only when learners employ selective learning biases that cause them to use only a subset of the available data.

(3) Models of linguistic evolution can be empirically grounded and then more easily manipulated to fit the available data (less parameters of variation).

Individual-level: learning period, data distribution, linguistic representation, probabilistic learning

Population-level: population size, population growth rate, time period of change, rate of change

Thank You

Amy Weinberg

Colin Phillips

Norbert Hornstein

Philip Resnik

the Cognitive Neuroscience of Language Lab
at the University of Maryland
Pennsylvania Linguistics Colloquium
The Northwestern Institute on Complex Systems

[redacted]

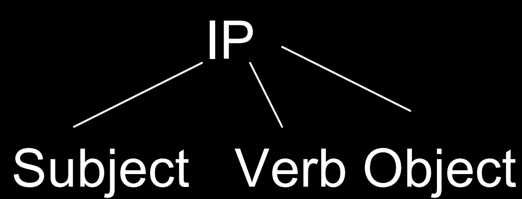
Individual Framework Applicability

Benefit: Can combine discrete representations, selective learning biases, and probabilistic learning for many types of linguistic knowledge.

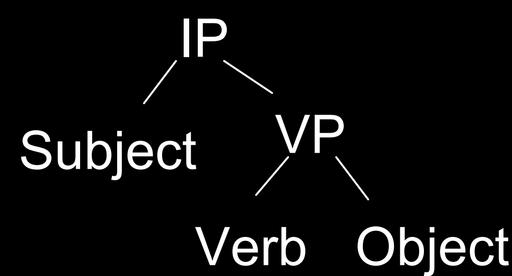
0.5 0.5

Discrete Representation: How much structure is posited for language?

A = linear structure



B = hierarchical structure



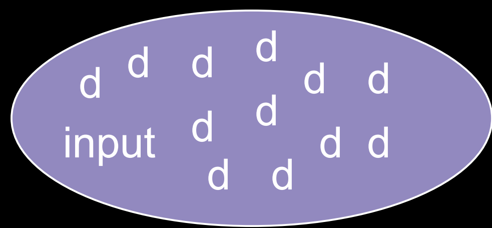
Discrete Representation : Is the basic word order Object Verb or Verb Object?

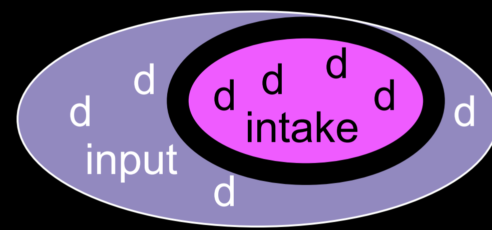
A = Object Verb

B = Verb Object

Framework Applicability

Benefit: Can combine discrete representations, selective learning biases, and probabilistic learning for many different problems.





Learning Bias: Use all available data. (Good for probabilistic learner - no data sparseness problem.)

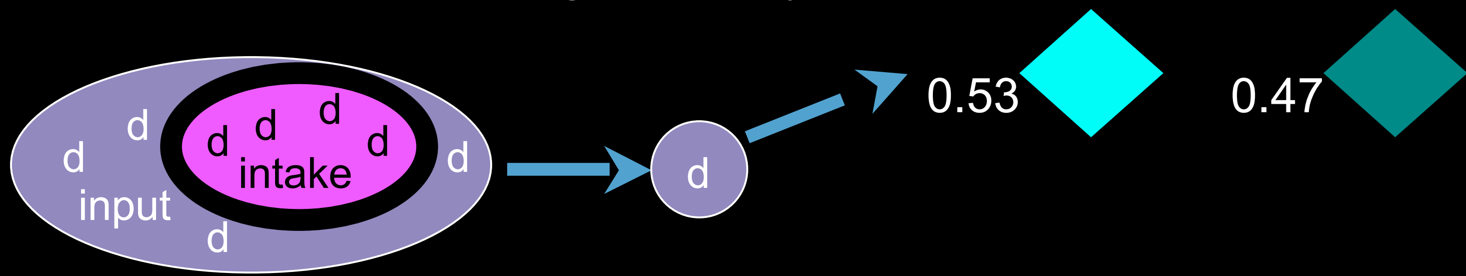
Selective Learning Bias: Use only data perceived as most informative (Fodor 1998, Lightfoot 1999, Dresher 1999).

Selective Learning Bias: Use only data that is more accessible (perhaps for language processing reasons) (Lightfoot 1991).

Selective Learning Bias: Use only data that is perceived as more systematic (Yang 2005).

Framework Applicability

Benefit: Can combine discrete representations, selective learning biases, and probabilistic learning for many different problems.



This can be instantiated as Bayesian updating, a Linear reward-penalty scheme, or any other probabilistic learning procedure.

$$\mathrm{Max}(\mathrm{Prob}(p_{vo} \mid u)) = \mathrm{Max}\left(\frac{\mathrm{Prob}(u \mid p_{vo}) * \mathrm{Prob}(p_{vo})}{\mathrm{Prob}(u)}\right)$$

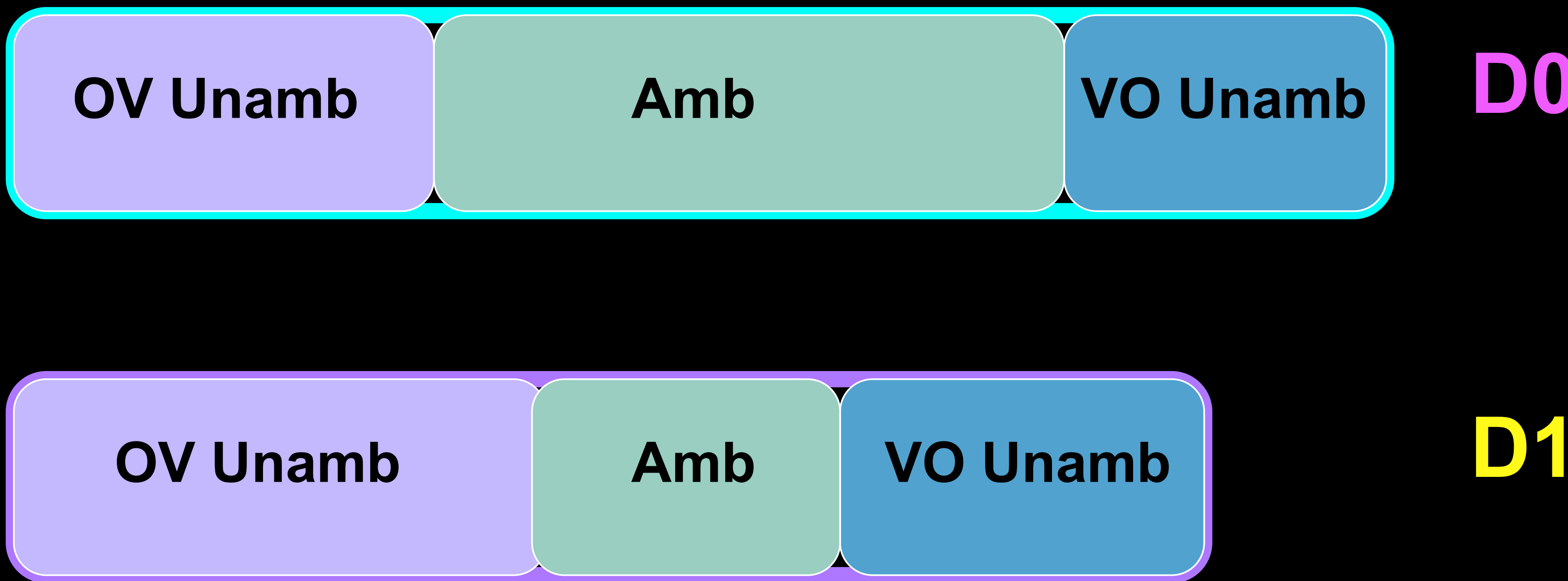
$$p_{ov} = p_{ov} + \gamma(1 - p_{vo})$$
$$p_{vo} = 1 - p_{ov}$$

[redacted]

Estimating Historical p_{VO}

Known quantities: Unambiguous and ambiguous data in d0 and d1

Estimating Historical p_{VO}

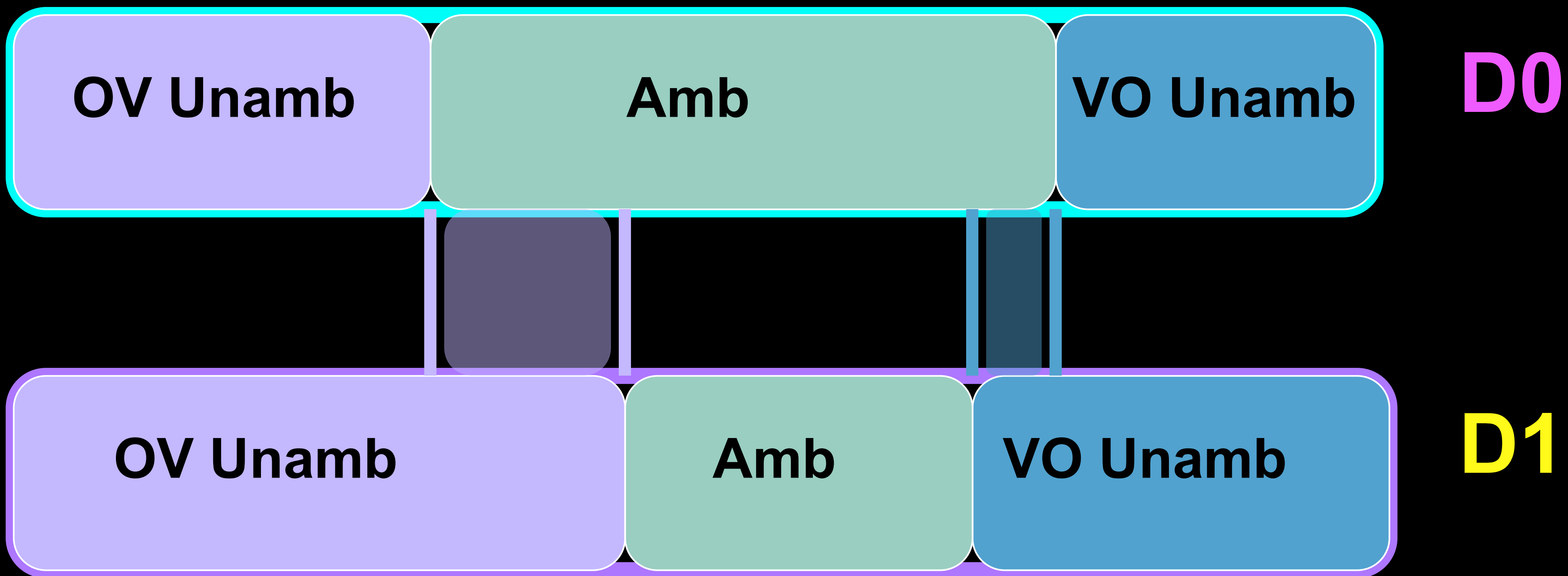


Estimating Historical p_{VO}

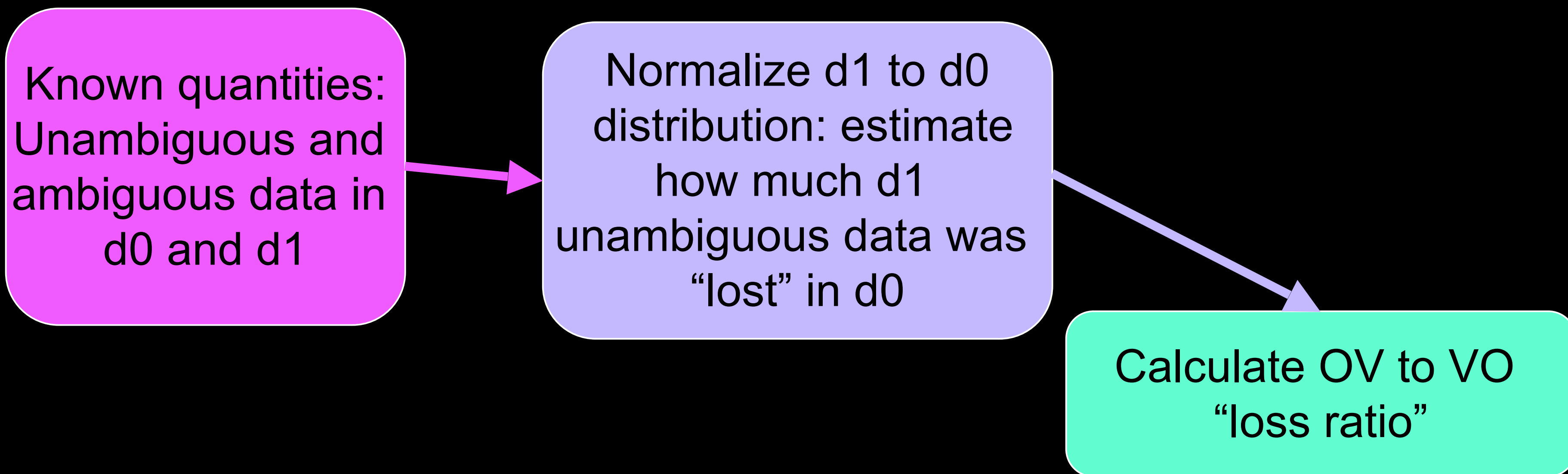
Known quantities: Unambiguous and ambiguous data in d0 and d1

Normalize d1 to d0 distribution: estimate how much d1 unambiguous data was “lost” in d0

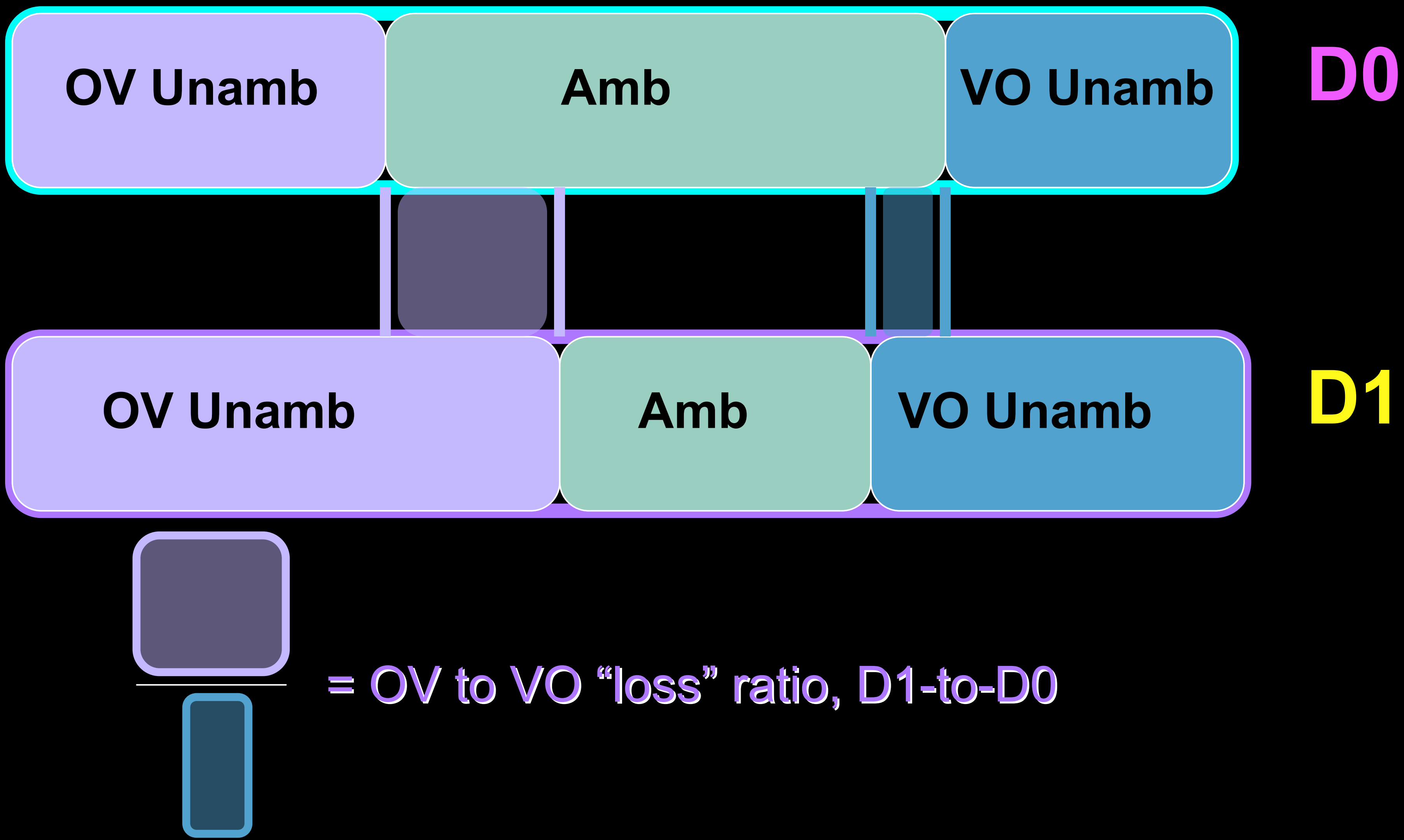
Estimating Historical p_{VO}



Estimating Historical p_{VO}

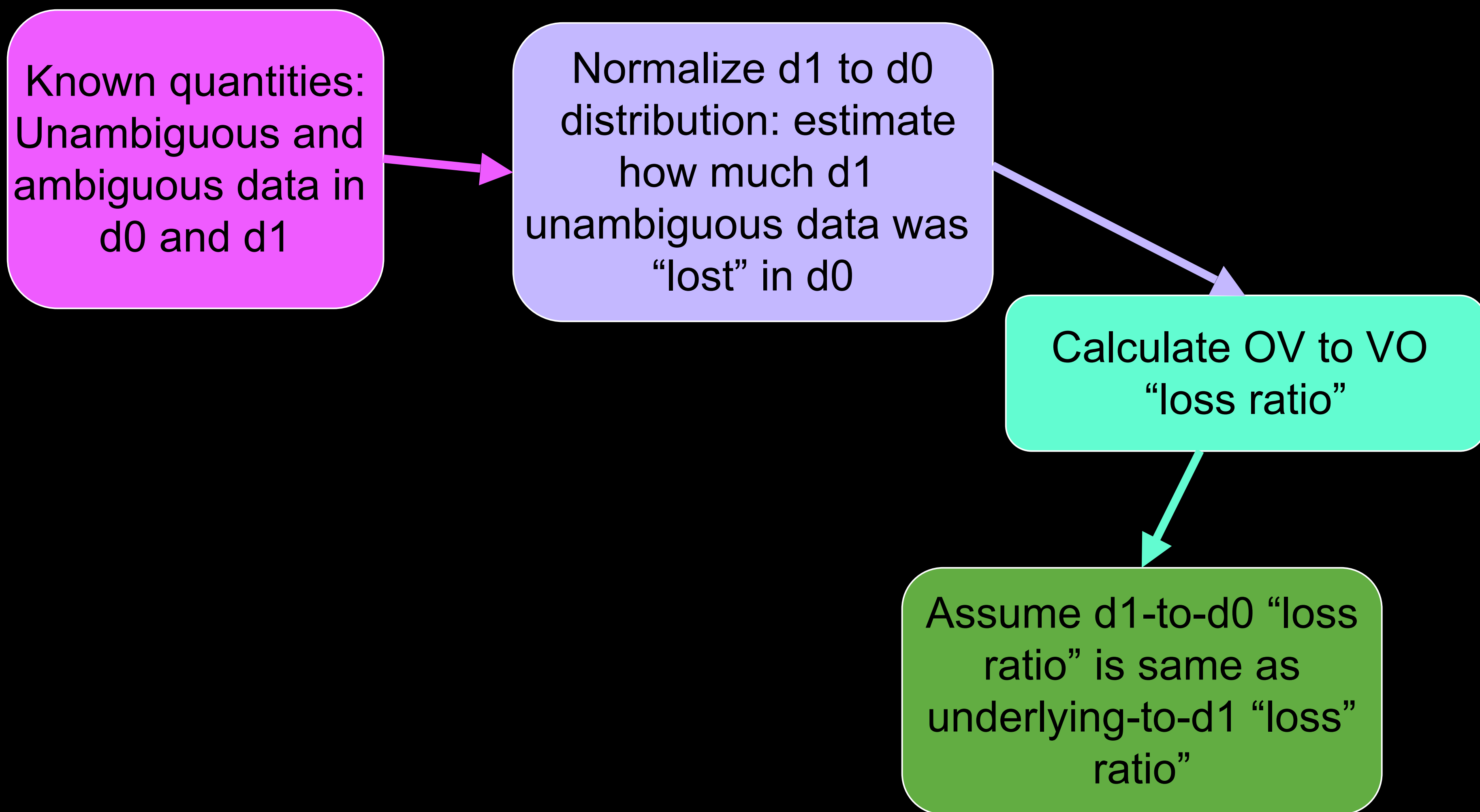


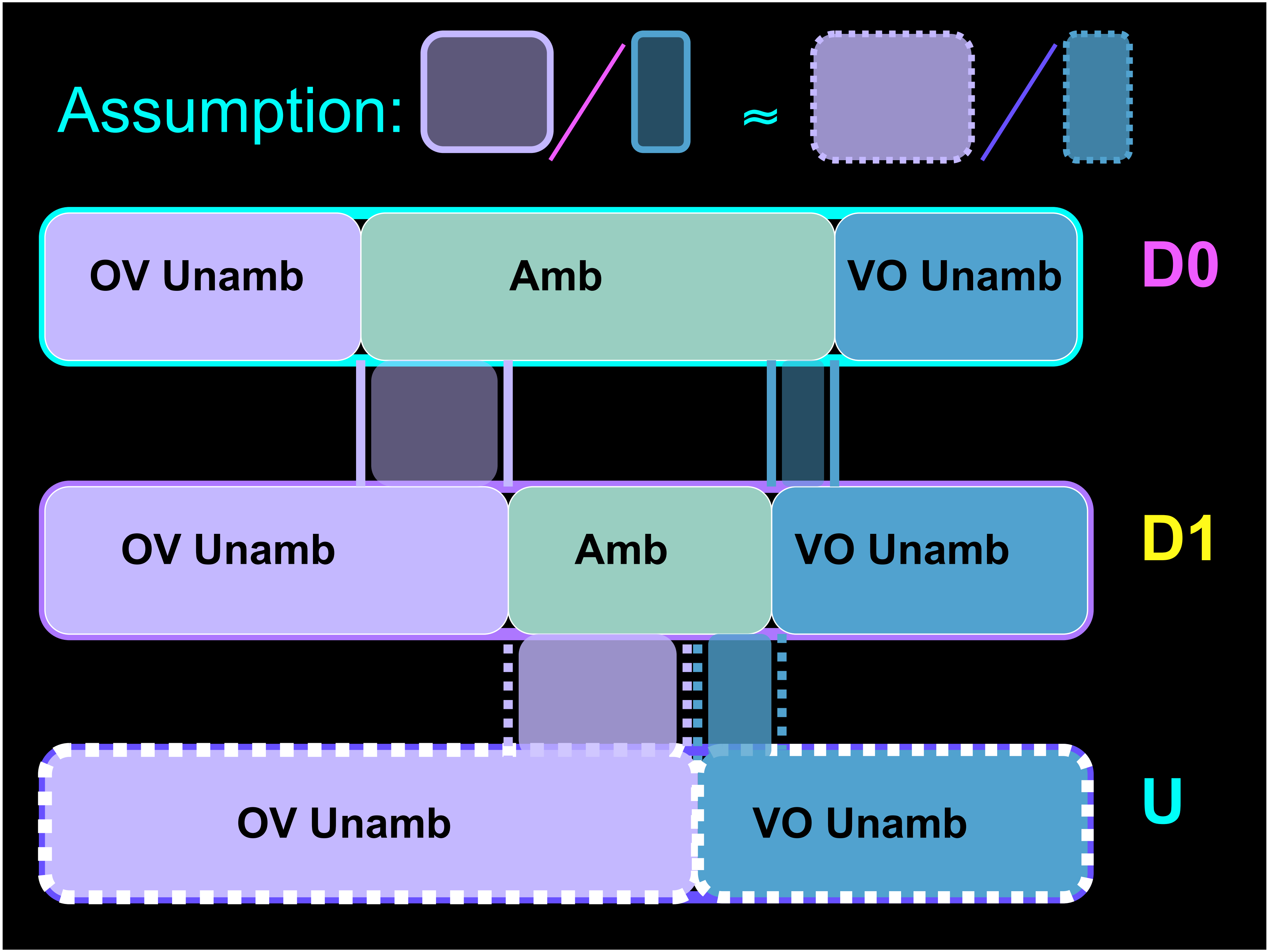
Estimating Historical p_{VO}



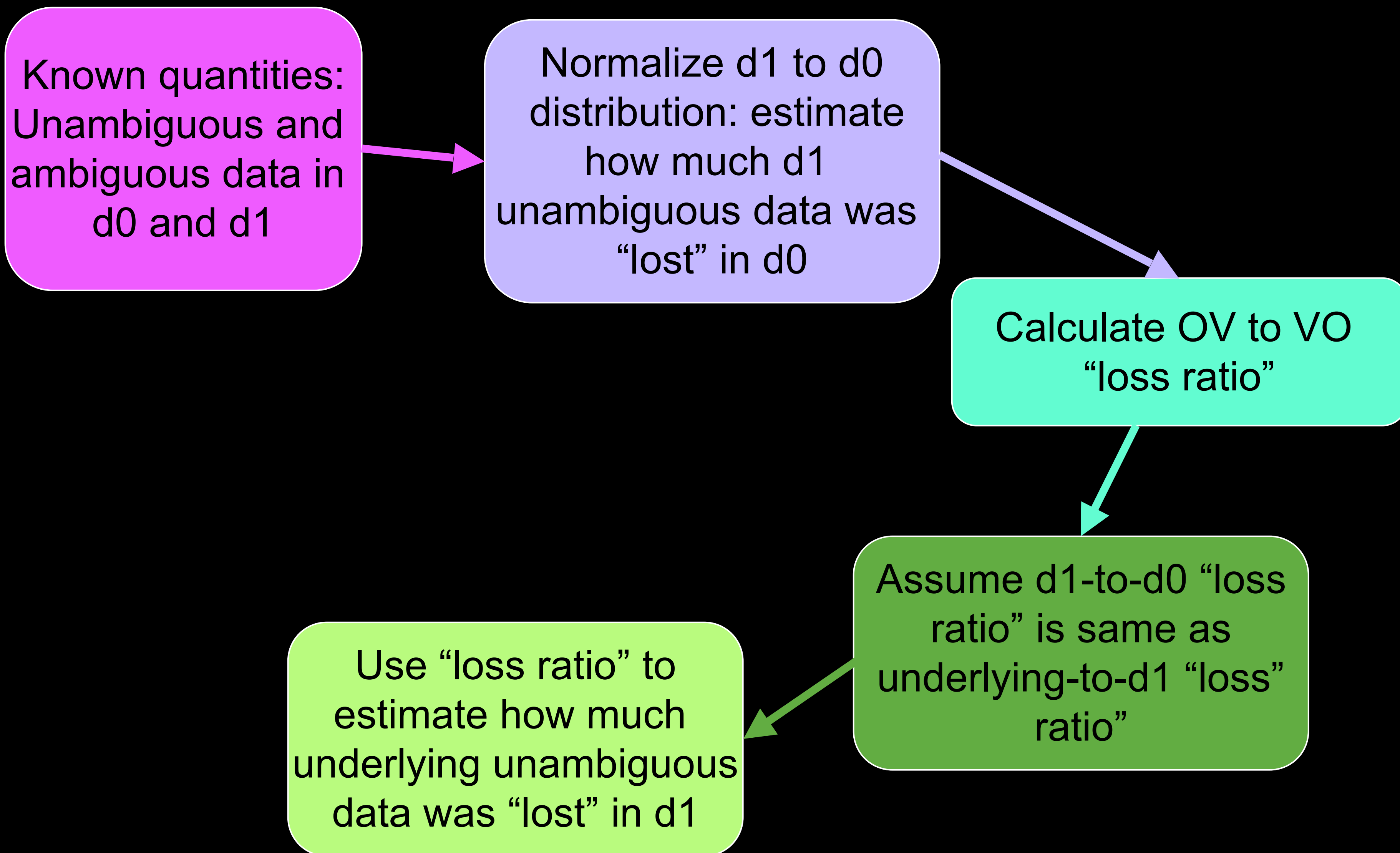
= OV to VO "loss" ratio, D1-to-D0

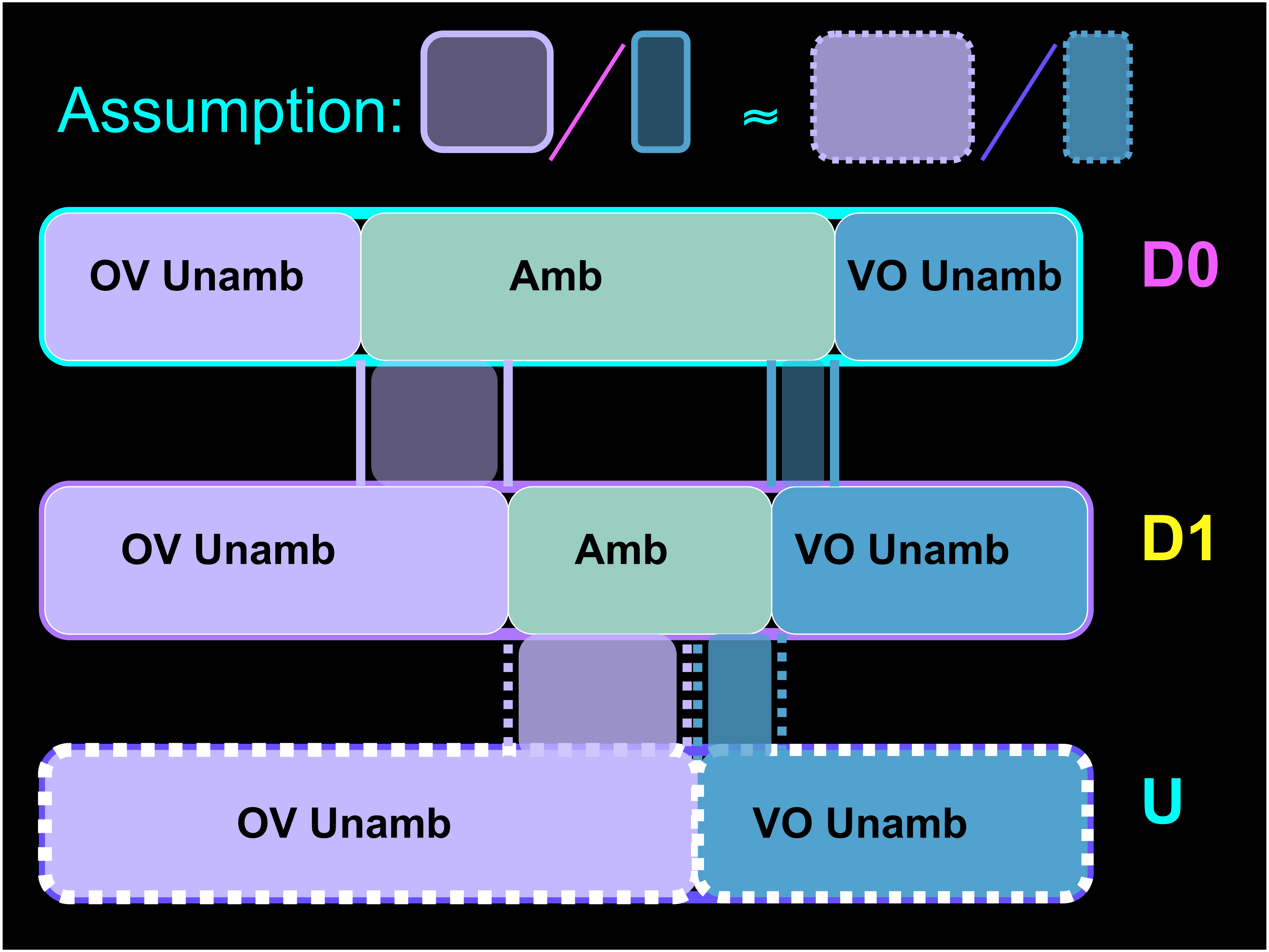
Estimating Historical p_{VO}

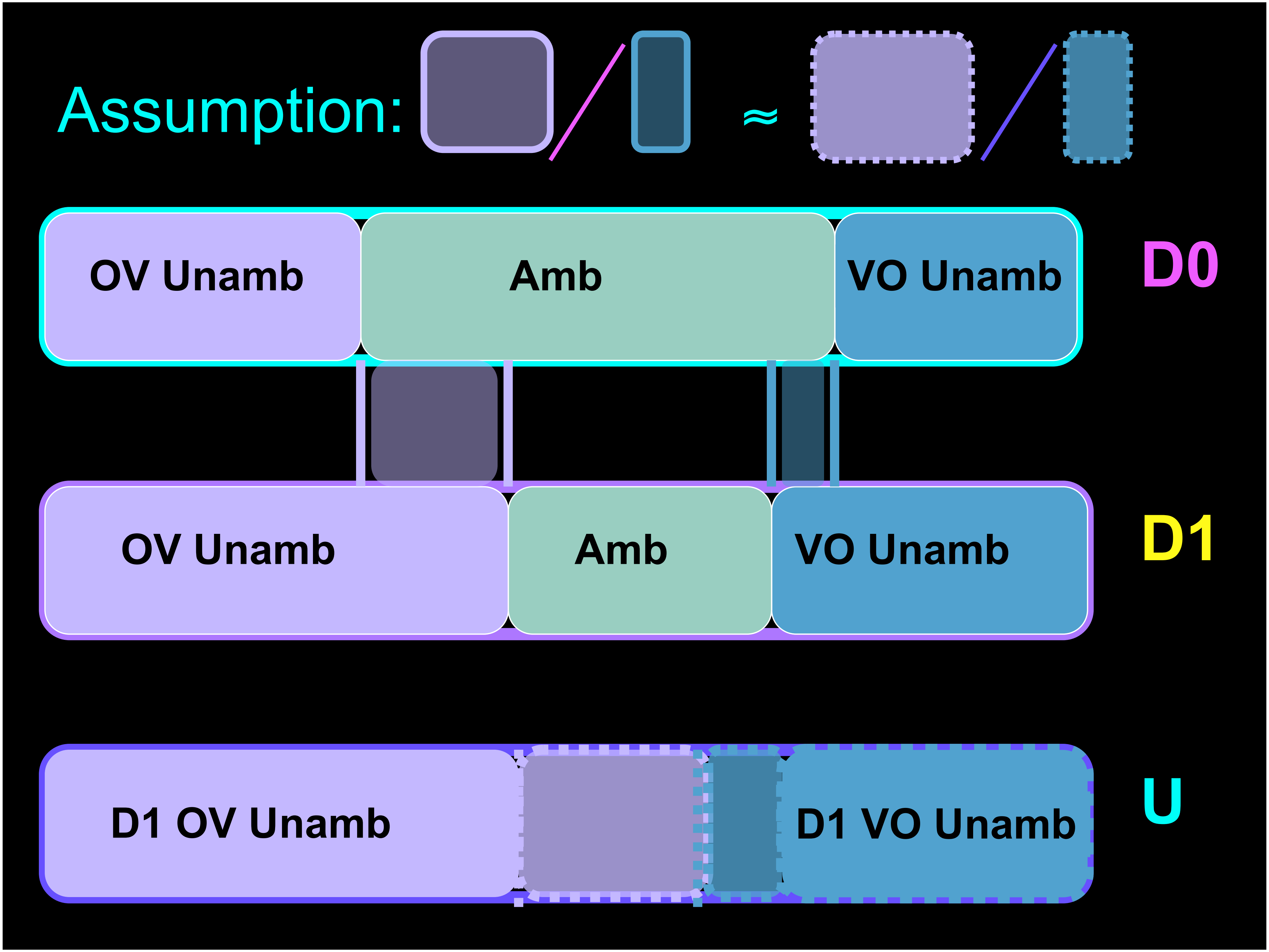



Assumption:
≈
OV Unamb
Amb
VO Unamb
D0
OV Unamb
Amb
VO Unamb
D1
OV Unamb
VO Unamb
U

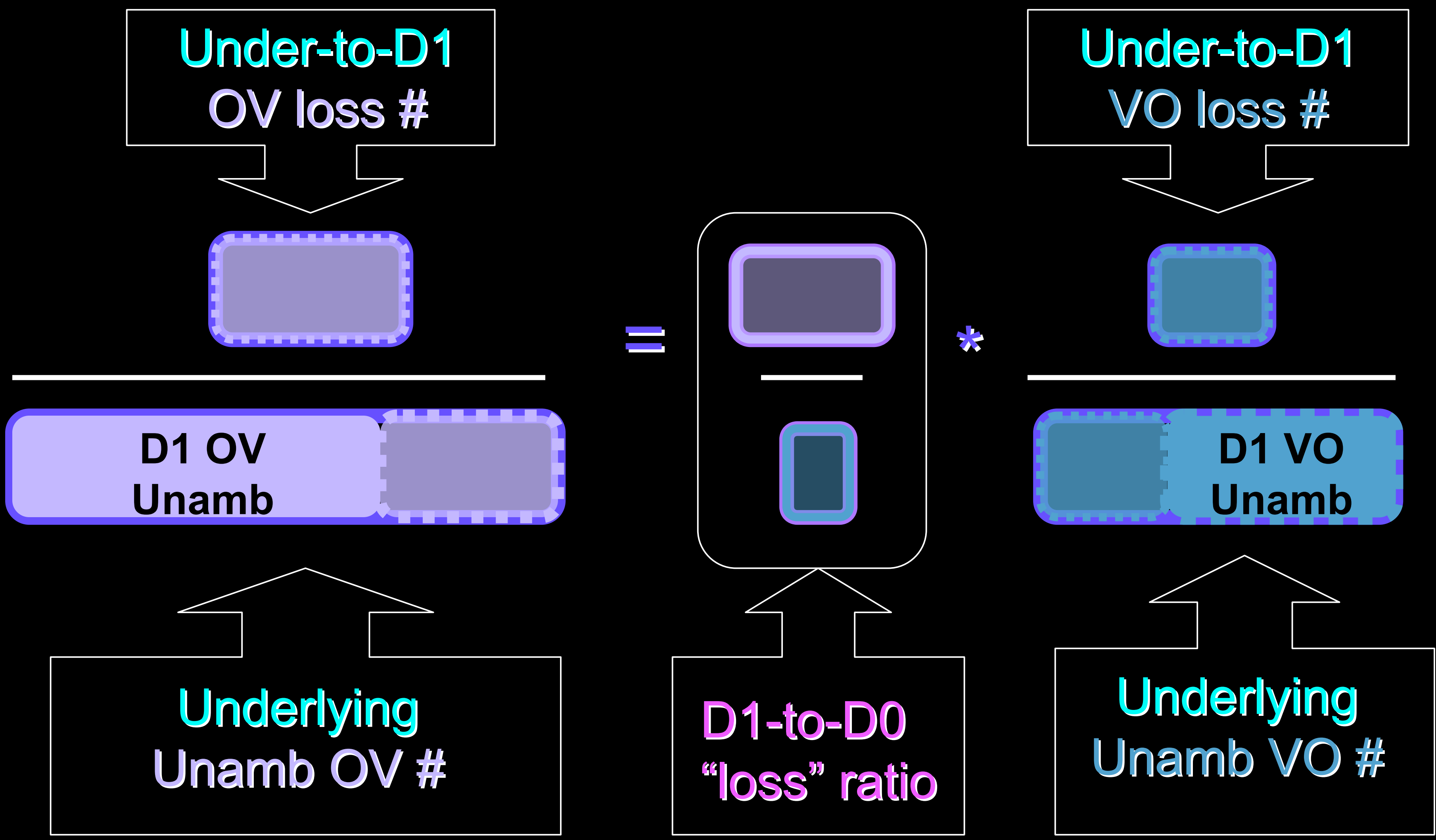
Estimating Historical p_{VO}




Assumption:
≈
OV Unamb
Amb
VO Unamb
D0
OV Unamb
Amb
VO Unamb
D1
OV Unamb
VO Unamb
U


Assumption:
≈
OV Unamb
Amb
VO Unamb
D0
OV Unamb
Amb
VO Unamb
D1
D1 OV Unamb
D1 VO Unamb
U

Estimating Historical p_{VO}



Estimating Historical p_{VO}

$$\boxed{\frac{\gamma * \mathbf{d0} - \mathbf{u1d1'}}{\gamma * \mathbf{d0}} = \mathbf{Ld1tod0} * \frac{\mathbf{ad1'} - (\gamma * \mathbf{d0} - \mathbf{u1d1'})}{\mathbf{u2d1'} + \mathbf{ad1'} - (\gamma * \mathbf{d0} - \mathbf{u1d1'})}}$$

γ = underlying p_{VO}

d0 = total degree - 0 data, **d**1 = total degree - 1 data

u1d1' = normalized unambiguous OV degree - 1 data

u2d1' = normalized unambiguous VO degree - 1 data

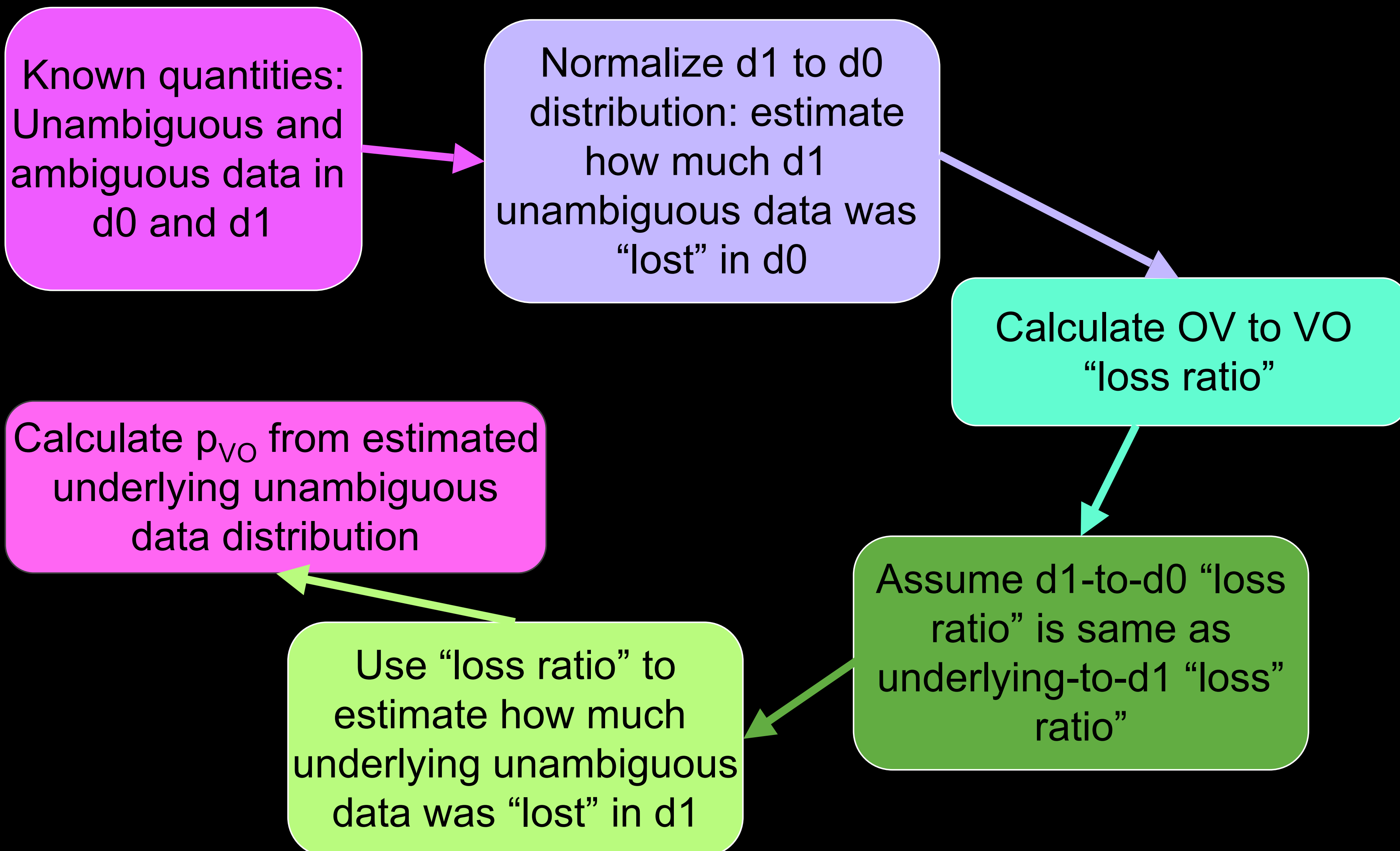
Ld1tod0 = loss ratio (OV/VO) from degree - 1 to degree - 0 distribution

ad1' = normalized ambiguous degree - 1 data

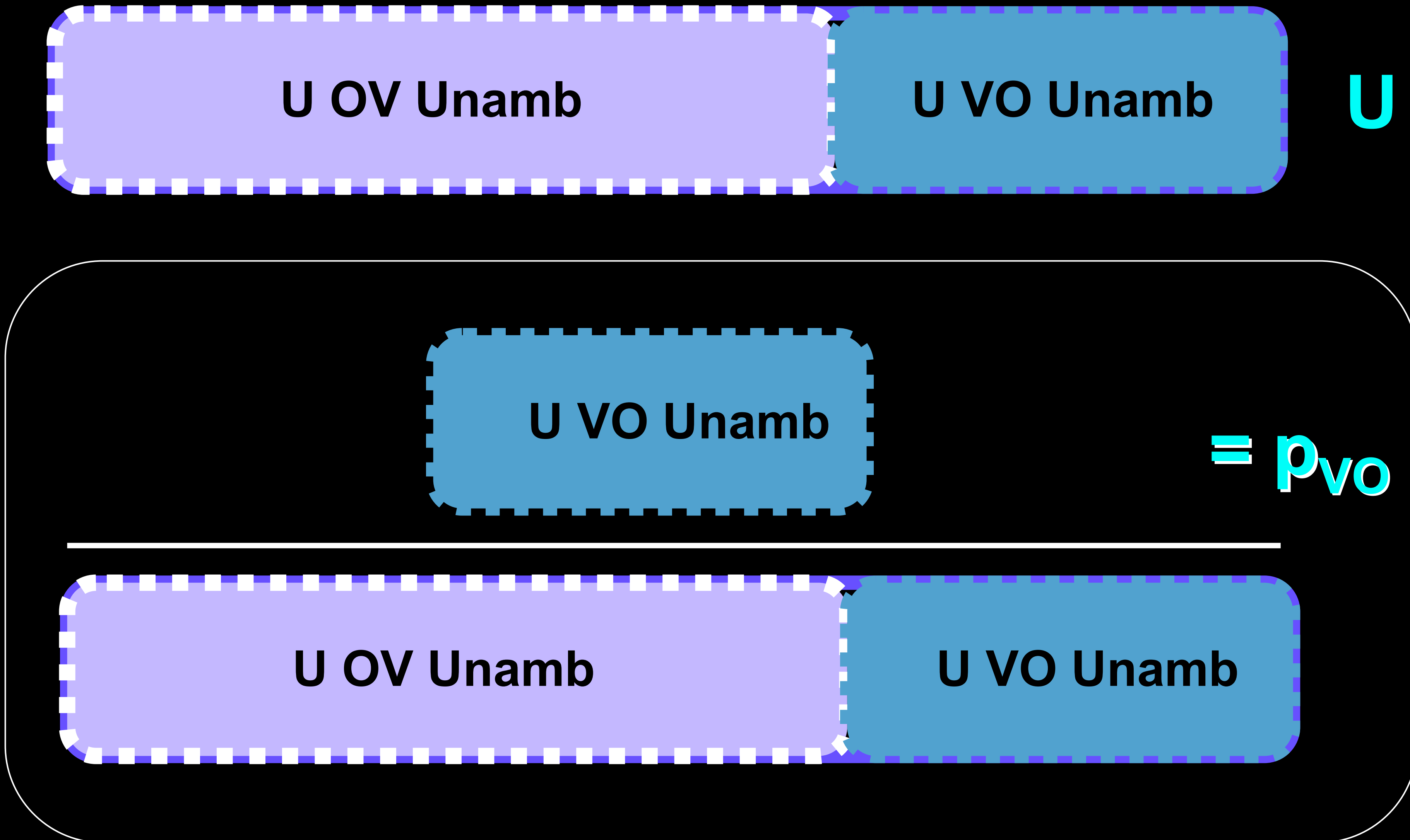
$$\gamma = \frac{-(\mathbf{d0})(\mathbf{d0} + \mathbf{u1d1'} - \mathbf{Ld1tod0} * (\mathbf{ad1'} + \mathbf{u1d1'}))}{2(\mathbf{Ld1tod0} + 1)(\mathbf{d0}^2)}$$

$$+/- \frac{\sqrt{((\mathbf{d0})(\mathbf{d0} + \mathbf{u1d1'} - \mathbf{Ld1tod0} * (\mathbf{ad1'} + \mathbf{u1d1'})))^2 - 4(\mathbf{Ld1tod0} + 1)(\mathbf{d0}^2)((\mathbf{-1})(\mathbf{d0} * \mathbf{u1d1'}))}}{2(\mathbf{Ld1tod0} + 1)(\mathbf{d0}^2)}$$

Estimating Historical p_{VO}



Estimating Historical p_{VO}



[redacted]

Potential Causes of Language Change

Old Norse influence before 1000 A.D.: VO-biased

- If sole cause of change, requires exponential influx of Old Norse speakers.

Old French at 1066 A.D.: embedded clauses predominantly OV-biased (Kibler, 1984)

- Matrix clauses often SVO (ambiguous)
- OV-bias would have hindered Old English change to VO-biased system.

Evidence of individual probabilistic usage in Old English

  - Historical records likely not the result of subpopulations of speakers who use only one order

[redacted]

Deriving the Bayesian Update Equations for a Hypothesis Space with 2 Hypotheses

$$\mathrm{Max}(\mathrm{Prob}(\mathrm{p_{VO}} \mid u)) = \mathrm{Max}\left(\frac{\mathrm{Prob}(u \mid \mathrm{p_{VO}}) \;*\; \mathrm{Prob}(\mathrm{p_{VO}})}{\mathrm{Prob}(u)}\right)$$

Bayes' Rule, find maximum of a posteriori (MAP) probability
Manning & Schütze (1999)

Deriving the Bayesian Update Equations for a Hypothesis Space with 2 Hypotheses

$$\mathrm{Max}(\mathrm{Prob}(p_{VO} \mid u)) = \mathrm{Max}\left(\frac{\mathrm{Prob}(u \mid p_{VO}) * \mathrm{Prob}(p_{VO})}{\mathrm{Prob}(u)}\right)$$

$\mathrm{Prob}(u \mid p_{VO})$ = probability of seeing unambiguous data point u, given p_{VO},
= p_{VO}

$\mathrm{Prob}(p_{VO})$ = probability of seeing r out of n data points that are unambiguous for VO, for 0 <= r <= n

$$= \binom{n}{r} * p_{VO}^{\,r} * (1 - p_{VO})^{n-r}$$

Deriving the Bayesian Update Equations for a Hypothesis Space with 2 Hypotheses

$$\mathrm{Max}(\mathrm{Prob}(\mathrm{pvo} \mid u)) = \mathrm{Max}\left(\frac{\mathrm{pvo} * \binom{n}{r} * \mathrm{pvo}^{r} * (1 - \mathrm{pvo})^{n-r}}{\mathrm{Prob}(u)}\right) \quad \text{(for each point } r,\ 0 \le r \le n\text{)}$$

$$\frac{d}{d\mathit{pvo}}\left(\frac{\mathrm{pvo} * \binom{n}{r} * \mathrm{pvo}^{r} * (1 - \mathrm{pvo})^{n-r}}{\mathrm{Prob}(u)}\right) = 0$$

$$\frac{d}{d\mathit{pvo}}\left(\frac{\mathrm{pvo} * \binom{n}{r} * \mathrm{pvo}^{r} * (1 - \mathrm{pvo})^{n-r}}{\cancel{\mathrm{Prob}(u)}}\right) = 0 \qquad \text{(P(}u\text{) is constant with respect to pvo)}$$

$$\mathrm{pvo} = \frac{r + 1}{n + 1}$$

Deriving the Bayesian Update Equations for a Hypothesis Space with 2 Hypotheses

$$p_{VO} = \frac{r+1}{n+1}, \ r = p_{VOprev} * n$$

Replace 1 in numerator and denominator with

$c = p_{VOprev} * m$ if VO, $c = (1 - p_{VOprev}) * m$ if OV

$3.0 \le m \le 5.0$

$$p_{VO} = \frac{p_{VOprev} * n + c}{n + c}$$

[redacted]

Other Ways to Interpret Ambiguous Data

Strategies for assessing ambiguous data

(1) assume base-generation
- attempted and failed
- system-dependent (syntax)

(2) weight based on level of ambiguity (Pearl & Lidz, in submission)
- unambiguous = highest weight
- moderately ambiguous = lower weight
- fully ambiguous = lowest weight (ignore)

(3) randomly assign to one hypothesis (Yang 2002)