String

gritxnxbbrtxx

(or, tips and tricks for index design)

xrginx

oiersitudi isa, tau

ferraginadioipiit

An overview

String

[illegible]

xtixxtrix inxignx

xrginx

dlindde inedreihdl

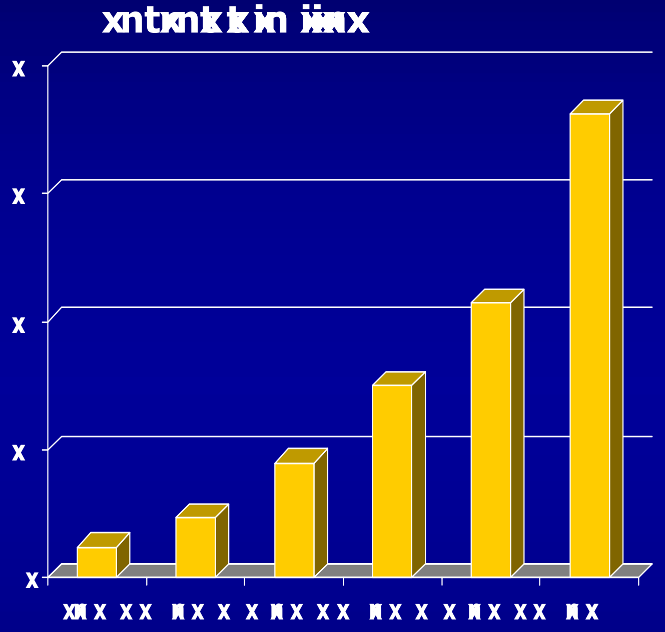
They are ubiquitous:

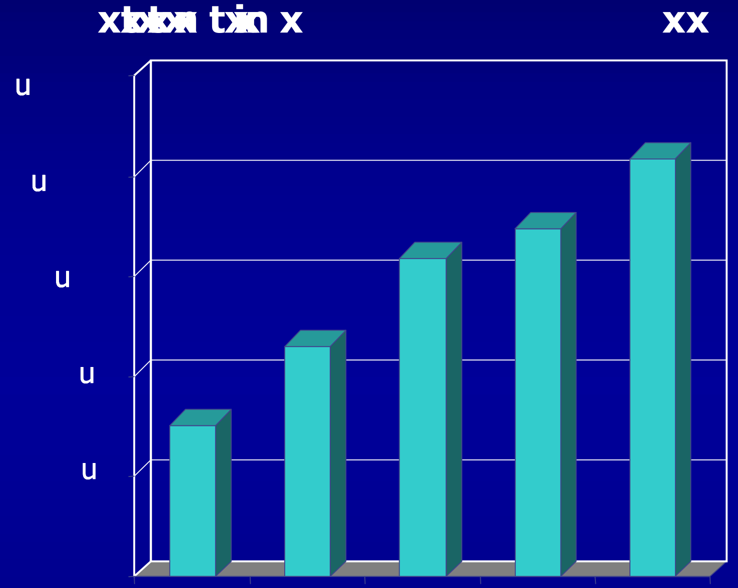
- □□ta□ibraries a□□rou□□ □ta□□es
- □□□roi□□hite a□ye□□□□□s
- □□□□□□ □□ratio□soure□s □□□□□□r □te□ □s□
- □eb □□s re□sitories
- □i□te i□□ratio□□s
- □

□ri□□□□io□ are □□i□at a sta□ri□rate:

- □re tha□□Tb o□te□ua□□ta i□the □b
- □re tha□□ o□base □irs i□the □□□□s

del icded





uurfaceeu uauotu u u÷uuu u

✓ uuu iions of doents (u iuons per day)u

uueepeu uauotu u u uu u

✓ uu uu uf interesting textuudata

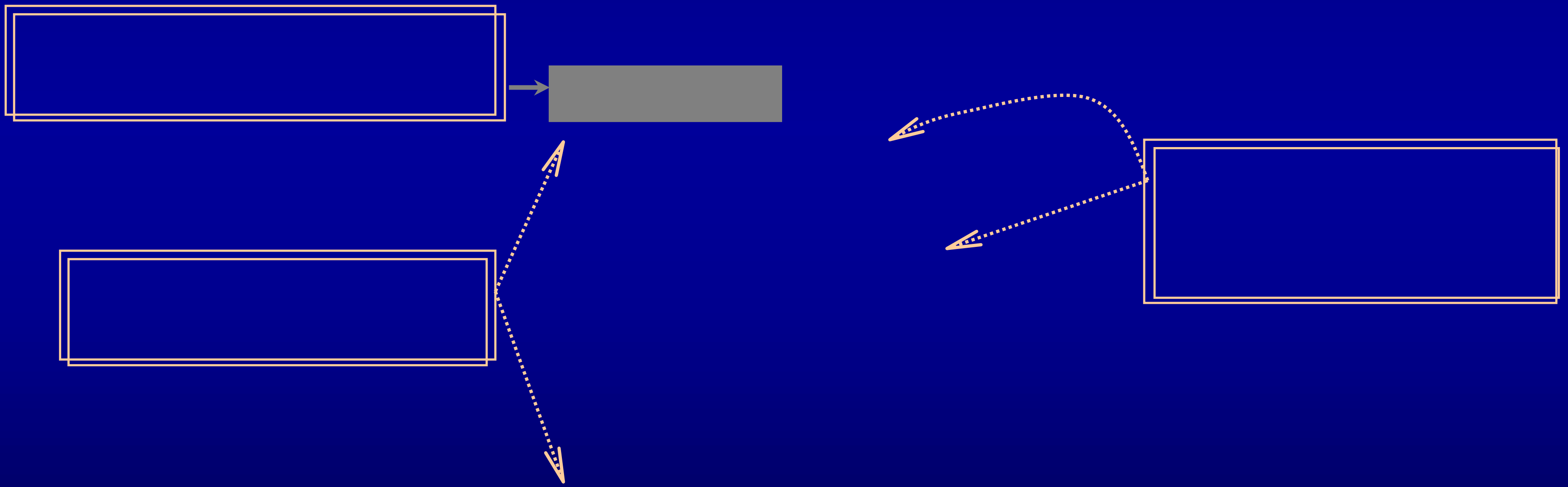
aiing u istu auotu u uuu (ueruear)

✓ uuiions of sug per day,u ituin u ,uaiing ists

ddbrel

dbeddhe d dl

□□ □□□□□□ouet is a si□ □ie□ o□te□ □□ai□soe □ru□□
that is se□s□ibi□□□ soe □ou□rue□ a□is easiy□
rea□be□ by hu□□ a□□□ters□

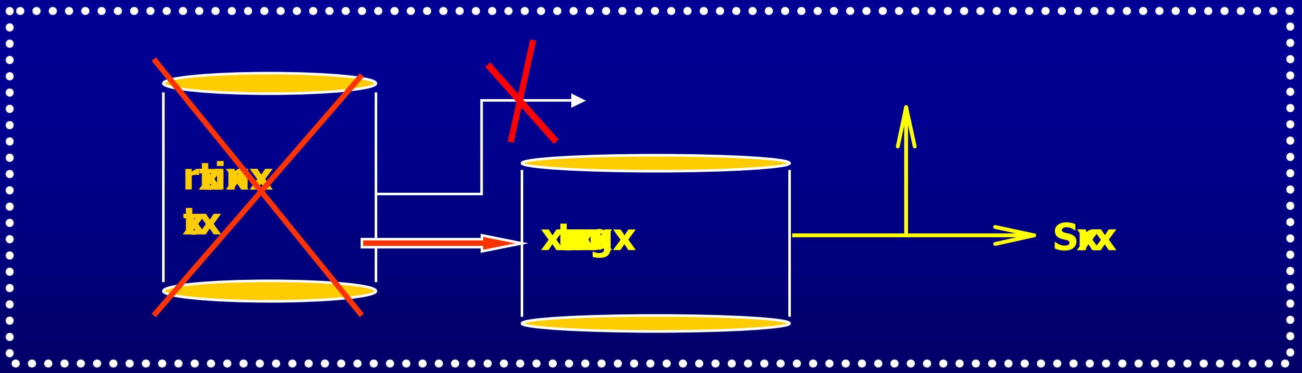


✓ □ is te□ base□a□tor□ie□e□□ ✓

dedodrdiodr d dl

□ueries □□t e□□it the ta□structure to re□□ra□a□ se□□□□ the retrie□□o□the a□□rs□or e□□□:

- □o□□□ty □y e□□it ta□esti□
- □or□□sa□i□atio□ □y e□□it ta□□es



☑ □□structure is usua□ re□rese□te□as a set o□□ths □tri□□ □

☑ □□queries are ture□□□o stri□queries: □oo□uthor□rstd□□□o□

d need r d d ined

□rute□re s□□s □t a □ab□ a□oa□:

- □ast si□ sear□es
- □uti□ si□ sear□es □r □□□queries

The □ri□□eritag□ i□iodry □□□s i□□ as □□□

□□uter si□□ a□ i□□ is a □rsiste□ □ta stru□ure that a□□ to □□s the sear□ □r a query stri□ □r a set o□the□□ a □o□by s□□rtio□ o□the □ta □□□io□

ded ed d

The i□□is a basi□b□□b□ay□ □syste□

□□□syste□aso e□□sses:

- □□□□□oes
- □□ai□a□orith□
- □□uery □□□a□s a□□eratio□
- □□ser□e□ba□□□□□ a□□ter□□□s
- □□c□rity a□□□ss □□tro□□□□□
- □□

□e i□□□□□rate o□□ o□□□□esi□□ □

dd dodd dnd

- □ear□about:
  - □oe□a□□aeor□or ea□□ati□stri□ata stru□ures a□ a□riths o□assi□ □ata sets
    - □iera□□ory oe□
    - □□uate the oe□ity o□ostru□io□a□uery oeratios
  - □ati□a□a□theoreti□□o□utios o□ie□esi□
    - □The □□subsyste□a□other □ory □es
    - □Types o□queries a□e□□ata
    - □e□ □tie tra□□□
    - □□ri□tra□atios a□e□ai□
  - □eri□a□e□ri□s o□terestiri□es
    - □erte□st □□array□□tree a□ri□tree
    - □o□to □oreo□a □essio□a□e□the □□otier □

✓ □□□ □o □ear □er a□□these □ata stru□ures □□

dedldcdword

String

gxitxxxbrxx

xtixxtrix inxignx

xrginx

[illegible] we [illegible] [illegible]

[illegible] the [illegible]st [illegible]

[illegible] is [illegible]	[illegible]	[illegible]er year	}	[illegible]
[illegible] ebry [illegible]	[illegible]	[illegible]er year	}	[illegible]roi[illegible]s
[illegible] roessor [illegible]	[illegible]	[illegible]er year		

[illegible]urre[illegible] [illegible]

[illegible]s [illegible]	[illegible]	÷ [illegible]bs	}	[illegible] ÷ [illegible]bs i[illegible]
[illegible]s [illegible]T[illegible]	[illegible]	÷ [illegible]bs		
[illegible]abus [illegible]ebry	[illegible]bs			
[illegible]s [illegible]		≈ [illegible]ses		
[illegible]ebry	[illegible]	÷ [illegible]ases		
[illegible]roessor	[illegible]			

si[illegible] [illegible]bete[illegible]ebry s[illegible]is [illegible]

el d deld dAdwd idlr d dl

oe draeters

strie i [illegible] ∆s oetio
tota o harattters i strie
hars er is ae
hars itti itera eory

oe reie

To tae re o is see a baith
e soetie istiuish betee:

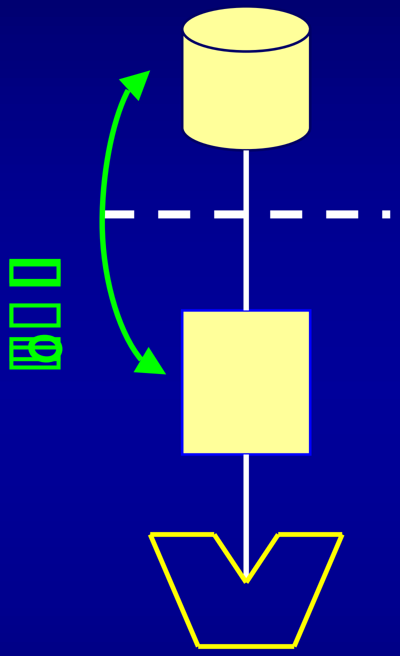
u s	: ethi otiuous ata
ao s	: ay other tye o

brithi oeity is therere eauate as:

uber o rao a bu s

iera ruitie tie

uber o is aes oue by the i or uri aorith eetio



[illegible]

Types of data

linguistic or tokenizable text
a sequence of characters or bytes

DNA sequences
audio-video files
executables

Types of query

word-based query
character-based query

exact word
word prefix or suffix
phrase

arbitrary substring
complex matches

Two indexing approaches :

Word-based indices, here a concept of word must be devised !
- Inverted files, Signature files or Bitmaps.

Full-text indices, no constraint on text and queries !
- Suffix Array, Suffix tree, Hybrid indices, or String B-tree.

[illegible]

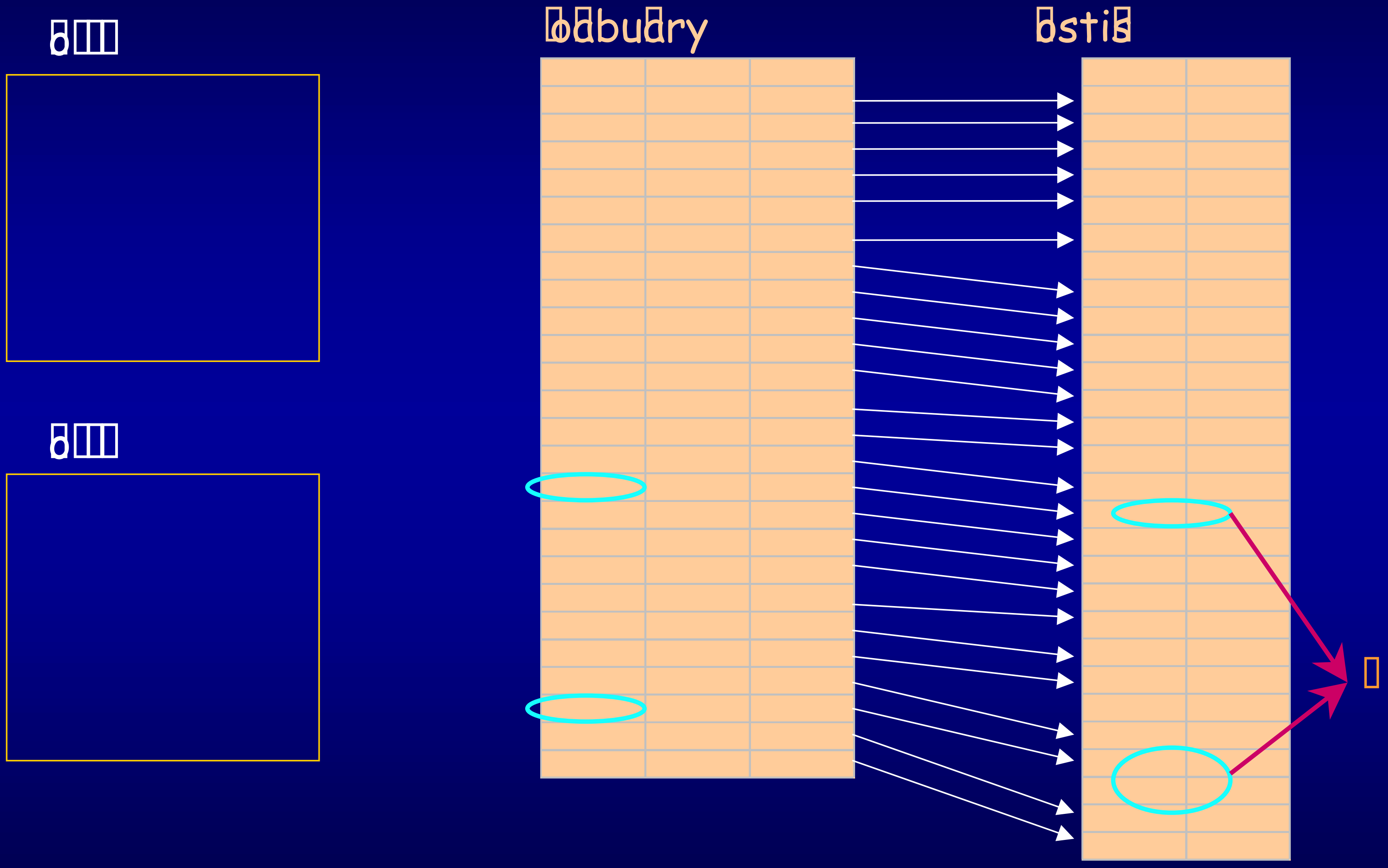
String

[illegible]

[illegible]

[illegible]

dhverdd edr d id



☑ □uery a□□ri□s a t□□ase □o□ss: □□□t □□□ti□

de dobn e dd cd

- oet o r ust be eise
  - it ees o the uery iaitio
  - be squeei ora ors to orssteii
- its sie is usuay sa
  - easa says where is the oetio sie
  - β is atiay betee a
- eetatio
  - rray: ie a sae suit but so queries
  - ash tabe: ast eat searhes
  - Trie: ast eei searhes but it is ore oiate
  - ute ie ast ee searhes
- oressio es seu ator o to o sai
  - es ahi a eethi
  - eues aout o oesse ata

de dobn eoth d

- [illegible]aularity or aUray i[illegible]or[illegible]atio[illegible]
  - [illegible] oarse[illegible]aie[illegible] : ee[illegible]ouet Ubers
  - [illegible] [illegible]oerate[illegible]aie[illegible] : ee[illegible]the Ubers o[illegible]the tet bos
  - [illegible] ie[illegible]aie[illegible] : ee[illegible]or[illegible]or setee Ubers

de arou[illegible]
ast queries a[illegible]reisio[illegible]

de ess tha[illegible]
o[illegible]queries: ost[illegible]eri[illegible]

- [illegible]orthodo[illegible]aroah to sae sail[illegible] aoi[illegible] [illegible]
  - [illegible]ort the ostis or ieasi[illegible]ouet[illegible]o[illegible]or ter[illegible]Uber
  - [illegible]tore the ierees betee[illegible]ae[illegible] ostiUes [illegible] [illegible]
  - [illegible]se ariabe[illegible]th eois or as: γ[illegible]

ootiUatio[illegible]bit: ie[illegible]bi[illegible]

✓ [illegible]

padding

agging

It is byte[illegible]aie[illegible]taea[illegible]sesyhroi[illegible]

ery ast eoi[illegible]a[illegible]sa[illegible]sae oerhea[illegible]

A dnerd ibnd d ie dbd

- Te□ □e□io□□□e□i□o bo□ o□□e□si□ ______ b
  - □ □bo□□y sa□t□ or □re □ue□ts
  - □ □stis □bo□□bers ingand ✓ s ✓ againd
- T□ tyes o□s□□ sai□
  - □u□ti□ o□rre□s i□a bo□are re□esete□by o□
  - □The □□ber o□bo□ □y be set to be sa□
    - ✓ □stis □st is sa□about □ o□the □e□io□si□
    - ✓ □□r □□d□s□□ a□query ti□ are o□r a □o□r b
- □uery a□□ri□s a three□ase □o□ss:
  - □uery is a□t□e□a□i□t the □□bu□ry: □r□a□i□
  - □□stis □sts o□sear□e□r□ are □bi□□ □a□te bo□
  - □a□te bo□ are e□□□to □□er out the □□e a□□es

□a□bu□ry tur□ □a□ te□ sear□es i□o ea□ bo□sear□es

□s□□□r su□□ i□□□

elr iedddreedddid dd

- e□ostrutio□
  - □ □reate □ter□airs t□ sorte□by i□easi□
  - □ □ere□sort o□the se□□□ t□
  - □ □ui□ostis □sts □o□a□ □airs □th equa□t□
  - ✓ □□ bo□er□ti□r □□□□ious □ostis □sts□
- □□□□ □□eri□
  - □ □□□ty i□the □stis □sts i□o□s their □□□
  - □ □assi□ e□itatio□ □eer □i□□orith□
  - □ □tie e□itatio□ □eore□i□□o□□□bers □□□ *et al*□□
- □□□atic□ie□i□
  - □ Ta□ a□attributes i□□as ter□ o□a □oer □□bu□ry
  - □ Ta□□sti□□□as set o□ □aste□□i□ □er□s□
  - ✓ □tru□tura□queries tur□ □to boo□a□a□eo□tri□queries □

☑ □ur □o□: □□library□□essio□i□i□or □□□ □

[illegible]

- [illegible]ai[illegible]ica:
  - [illegible]eresen[illegible] the [illegible] tree [illegible]a tu[illegible]es or set o[illegible] obe[illegible]ts[illegible]
  - *[illegible]ere* [illegible]ause to [illegible]ate i[illegible]o the tree[illegible]
  - [illegible]uery e[illegible] use sta[illegible]r[illegible] *[illegible]i[illegible]* a[illegible] *s[illegible]*
  - [illegible] a[illegible]tio[illegible]es or se[illegible]a[illegible]sses[illegible]
-

cdlcd

□o□□

- □□eera□isa□ct□a□es :
  - □ □uery a□iltio□is □ostysiuate□ia □□y □ais□□
  - □ □uery o□ti□ser □oses □□a□ o□□□ature o□the □uer□□□
  - □ □ies i□tab□es or □□shou□be sa□□
  - □ □ee□e□tra i□□es □r □□a□□e□ti□ □th queries
- rxix xi
- xx tx Stxxxix xtx

dhive dred

The literature oers drious dobsas:

- set us : build tree idiory at query tie
- : tree or storidirs dthdr
- abri : dtrid tree or iei *al* ossibe dths
- ati : tree is drtitio to isdes see e yee
- Tey : trie tree ✓ are sde oudy
- de deria outs : Taio o etais

de oudy is usuay ot edate surey it is ≥

[illegible] [illegible]

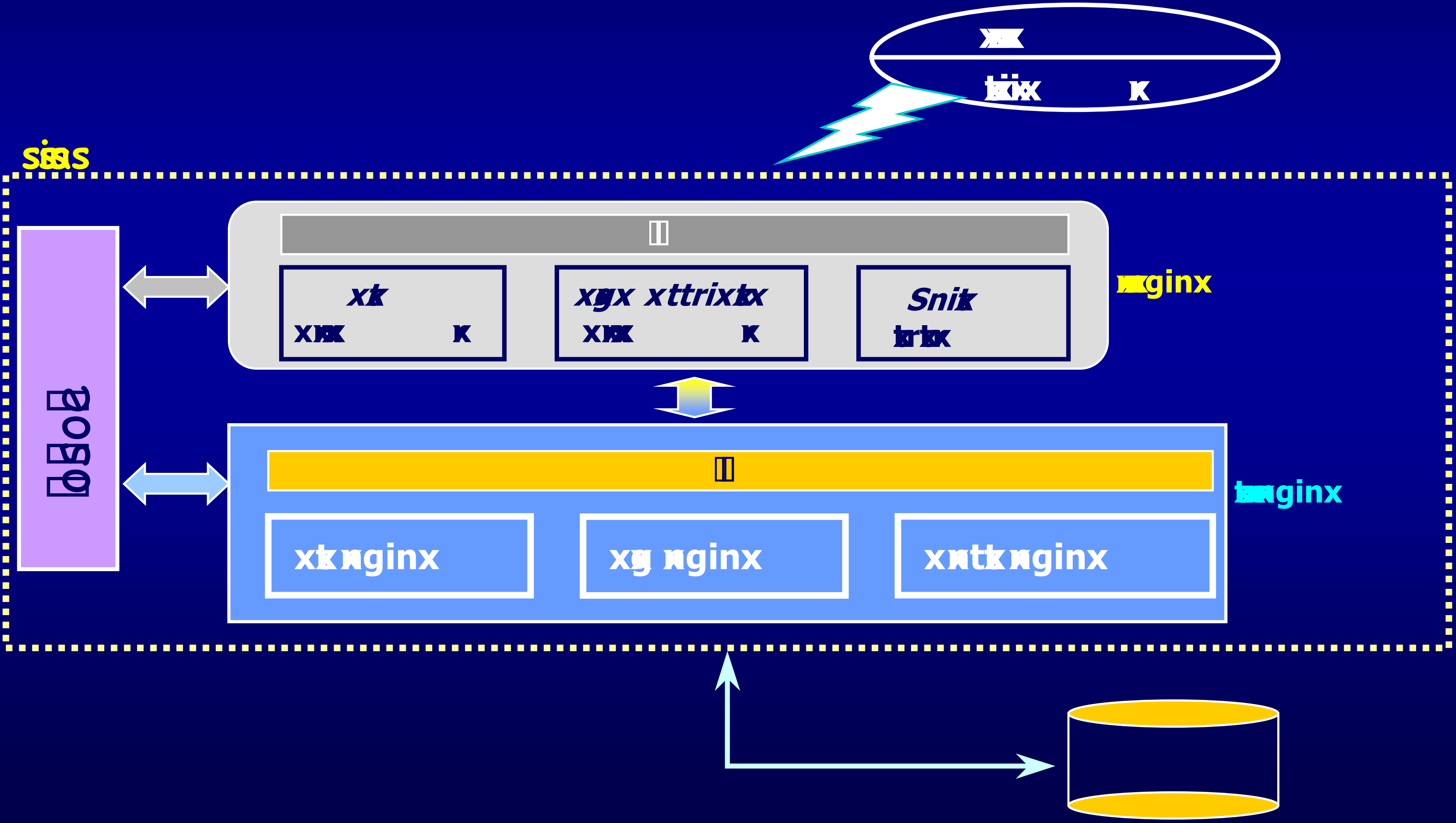
- ouets ay be:
  - stroy tetuae uisti tets
  - oy are a ay our ithout a T
  - arbitrariy este abiate i their ta struture
  - retrieabe i their ori or or brosers
-

$\Rightarrow$ sasss

idtd eidd died

- ile ouet ieit
  - ile softare arhitetre
  - ustoiabe iei o eah ile they are heteroeeous
  - ase o aeetuate aistributio
  - *iht* iteraie or *oiia tai* to see u query
-
-

idbdd ed cde



xx
tiix x
sias
xk
xxx x
xgx xttrixx
xxx x
Snik
trikx
nginx
xk nginx
xg nginx
xntk nginx
nginx

cl cdtlinedecd

String

gxitxnxbtrbx

xtixxtrix inxignx

xrginx

eldcobd oed

Their □e□is □r□si□:

- □ □a□ta : □seque□s□ui□i□o i□s□
- □ i□isti□te□s : □ta i□statisti□□
- □ o□□bu□ry □r □rte□ists
- □ □th queries o□□□□□□□s
- □ □rusio□□te□io□□i□ruses□

□our □sses o□i□□s:

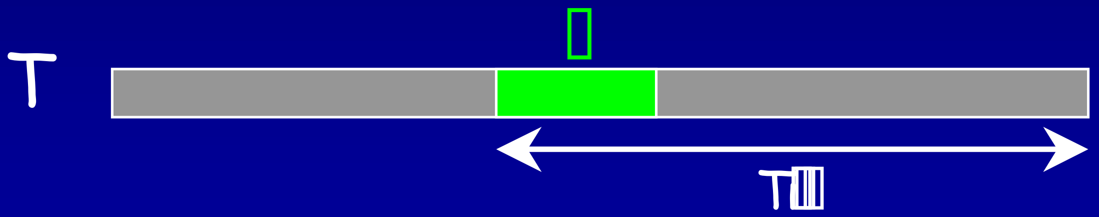
- ✓ □i□array or □i□tree
- ✓ T□□□□i□s: □i□array □□ry □r□□□□
- ✓ □ree base□ta stru□ures: □e□i□□ree
- ✓ □ri□□ree: □ree □trii□ trie

□ur □□ure □□sists o□a tour throu□ these too□ □ □

[illegible]

Pattern P occurs at position i of T

iff P is a prefix of the suffix T[i,N]



Occurrences of P in T = All suffixes of T having P as a prefix

T = This is a visual example
This is a visual example
This is a visual example

[illegible]

SUF(T) = Sorted set of suffixes of T

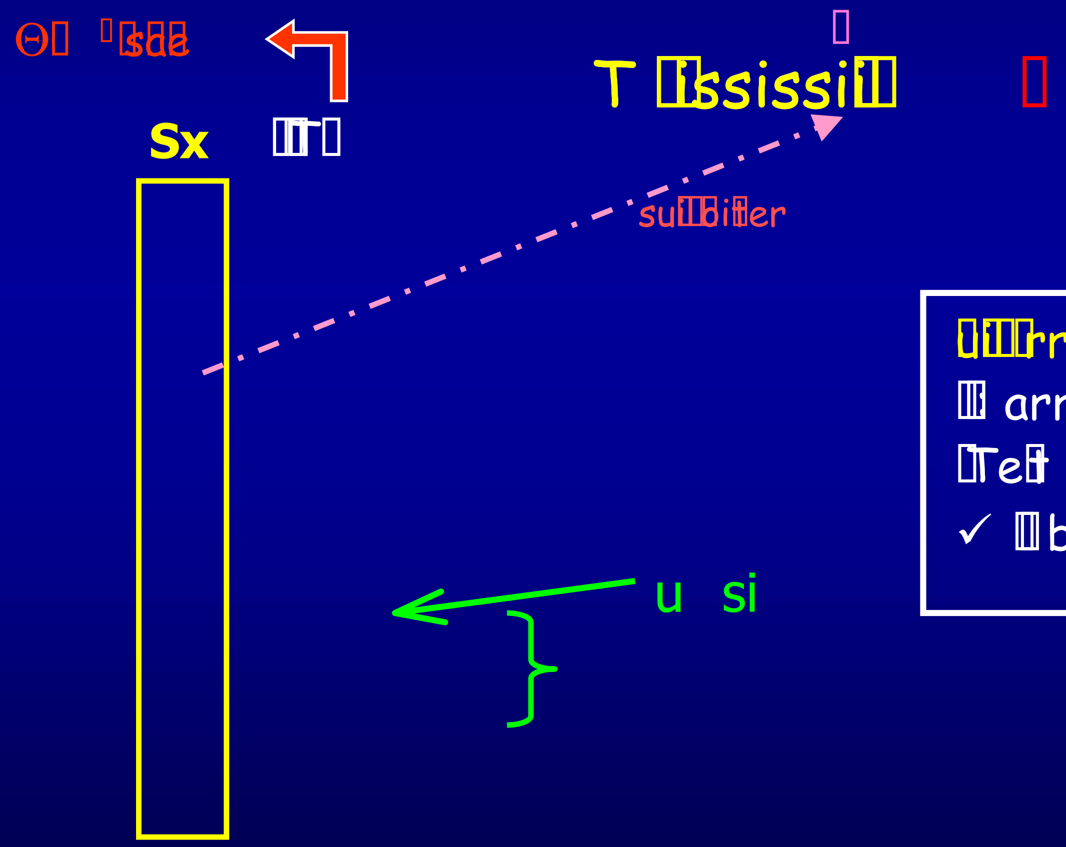
SUF(Δ) = Sorted set of suffixes of all texts in Δ

[illegible]

[illegible]

ro[illegible] [illegible]suf[illegible]es i[illegible]T hai[illegible]rei[illegible]are [illegible]ti[illegible]uous[illegible]

ro[illegible] [illegible]arti[illegible]ositio[illegible]s the [illegible]ibra[illegible]i[illegible] o[illegible]



Θ[illegible] [illegible]sae
T [illegible]ssissi[illegible]
Sx
[illegible]T[illegible]
sui[illegible]ei[illegible]er
u si

[illegible]rray

[illegible] array o[illegible]s[illegible] bytes

[illegible]Te[illegible] T: [illegible]bytes

✓ [illegible]bytes o[illegible]sae o[illegible]y

elcthdn idArrd

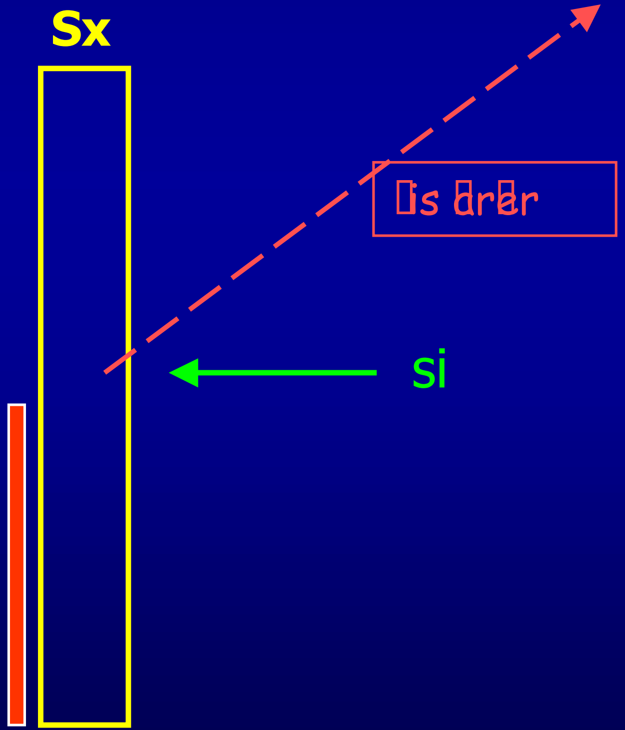
cdereddl

iretle biary searh o

bi

tie

T issississi

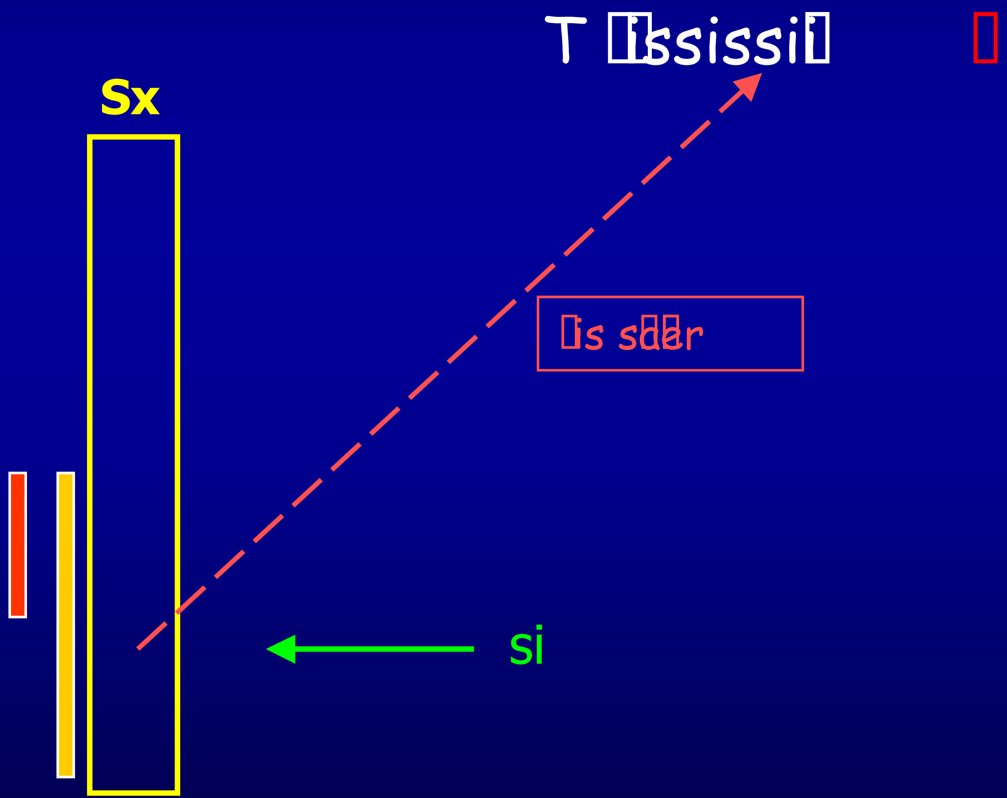


aesses or biary ste

ecrthdn idArrd

cderdd

ireteibiary search o　　　　ob　　tie

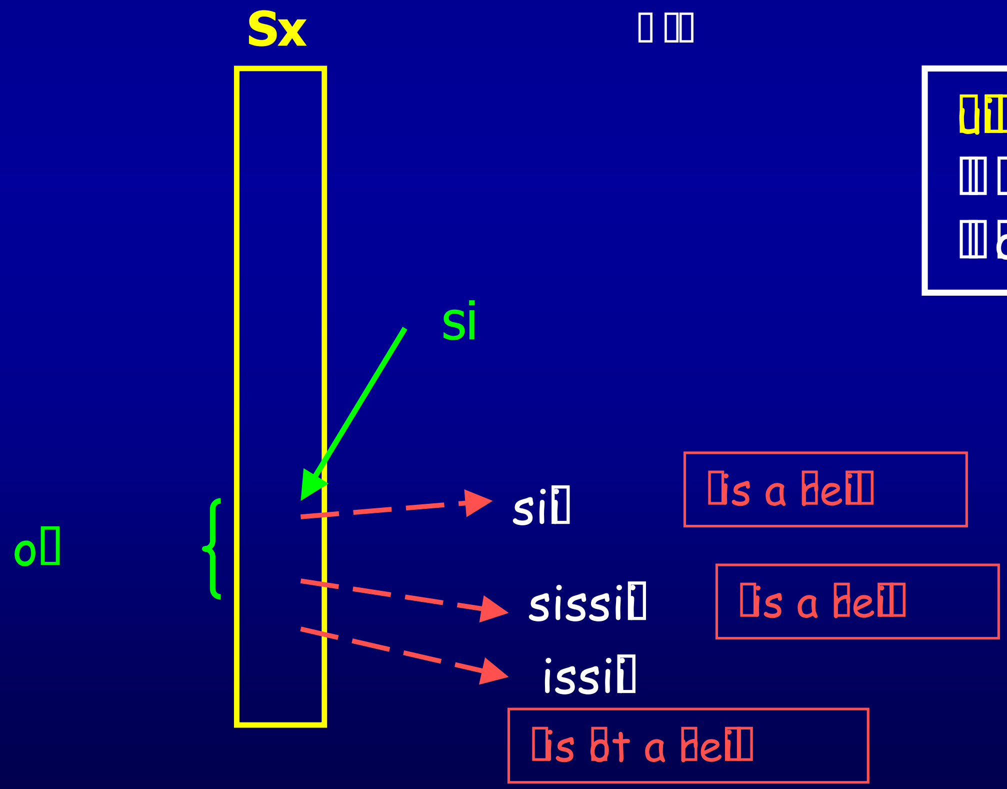


[illegible]

[illegible]

[illegible]rute[illegible]re [illegible]rison [illegible] [illegible] o[illegible]ti[illegible]

T [illegible]ississi[illegible] [illegible]



Sx
si
o[illegible]
si[illegible]
[illegible]is a [illegible]e[illegible]
sissi[illegible]
[illegible]is a [illegible]e[illegible]
issi[illegible]
[illegible]is [illegible]t a [illegible]e[illegible]

[illegible]rray searc[illegible]

[illegible] o[illegible] [illegible] o[illegible]ti[illegible]

[illegible] o[illegible] [illegible] o[illegible]i[illegible]ati[illegible]

[illegible]era[illegible]ory

[illegible] is [illegible]or [illegible]

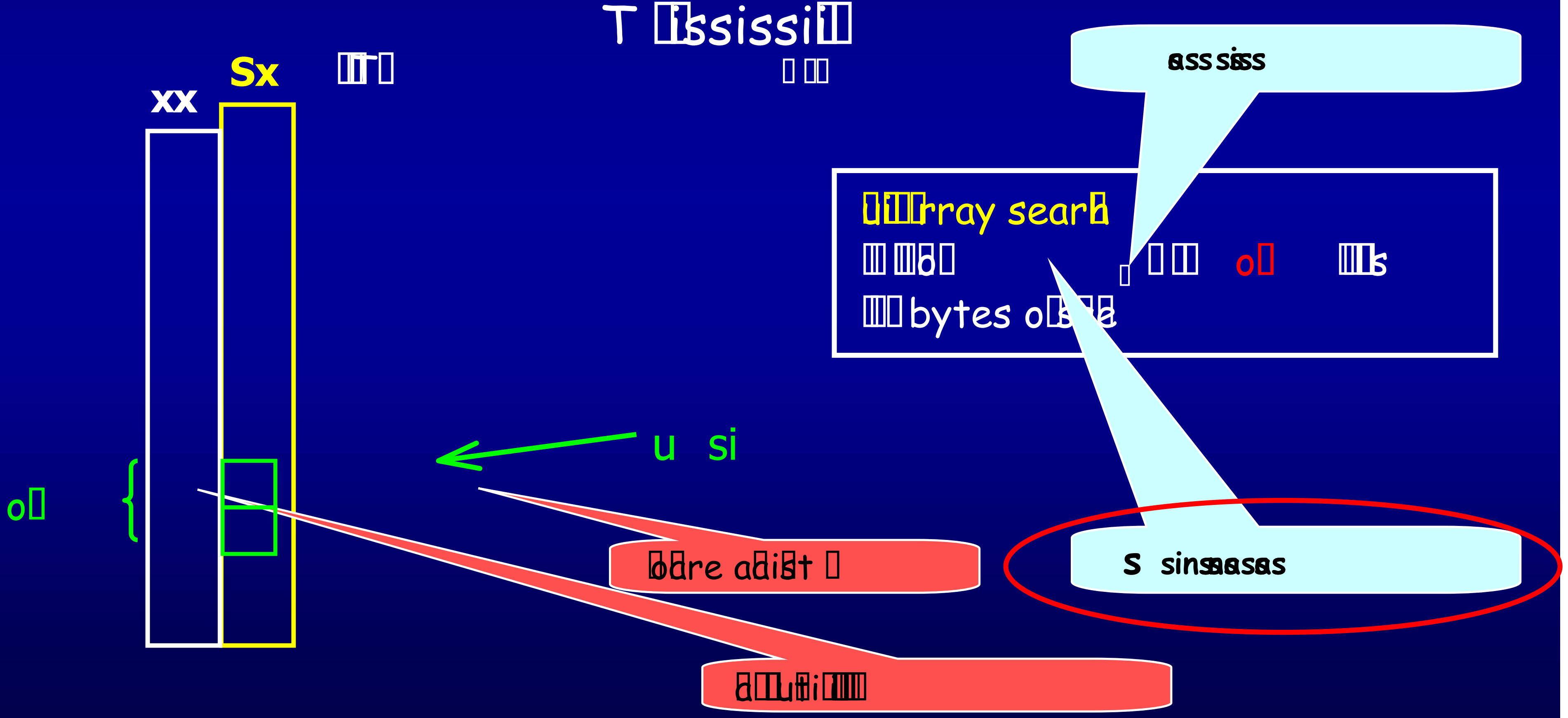
[illegible] o[illegible] [illegible] o[illegible]s

o[illegible] [illegible]

o[illegible]

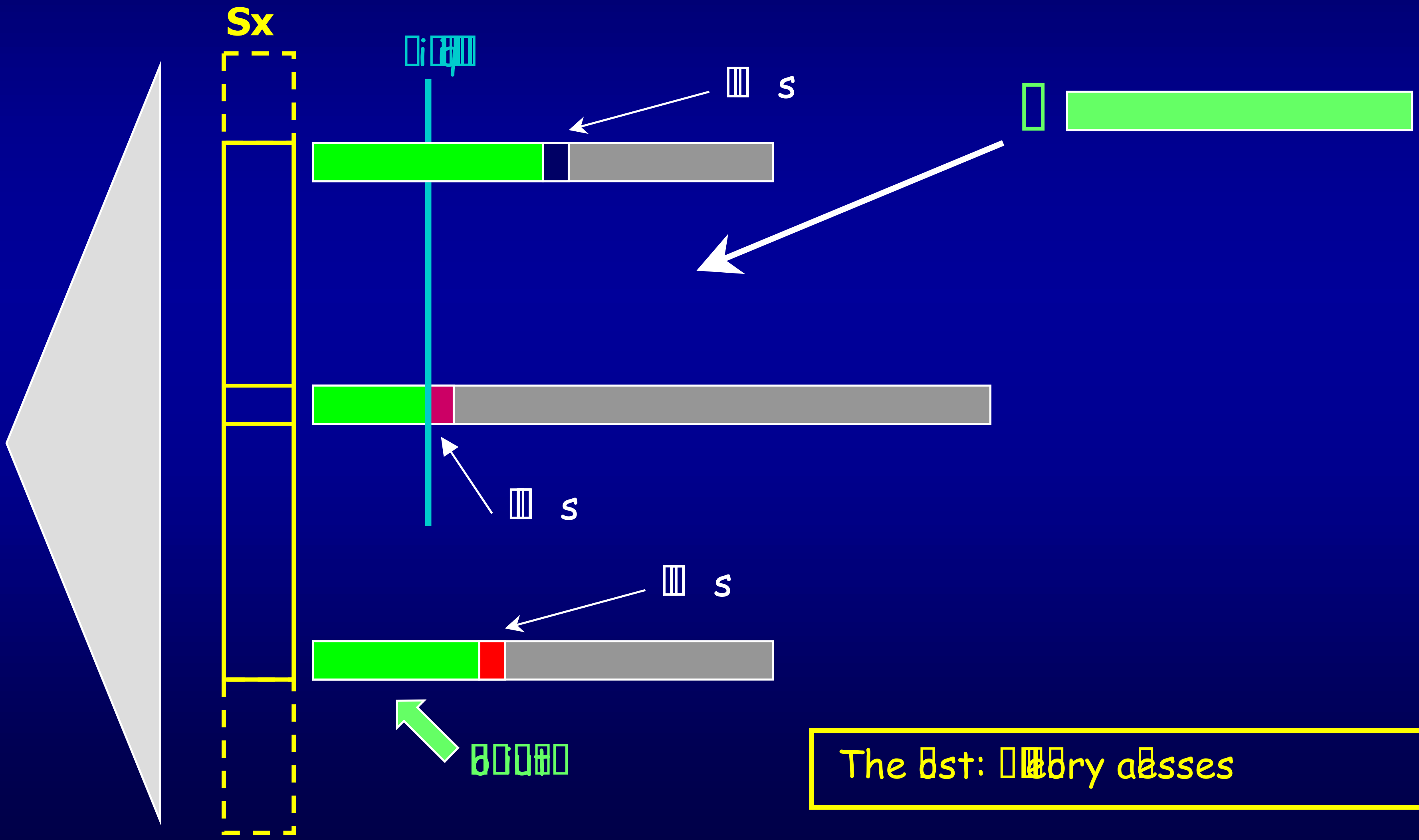
edhidve redievd

stores the bestdbdl betaesules aaet i



[illegible]

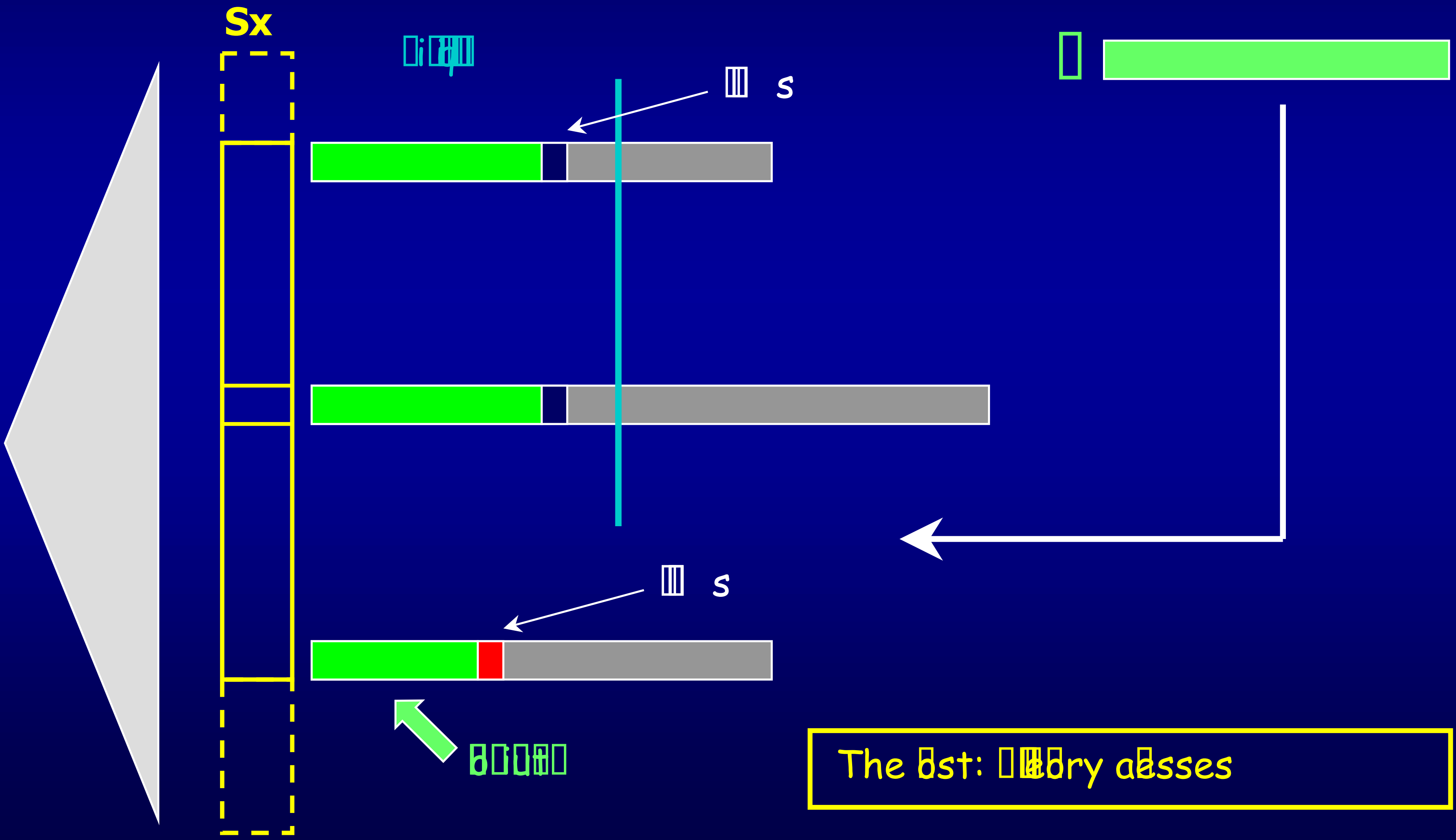
[illegible] usi the array : o res [illegible] atter hars



The ost: ory a[illegible]sses

[illegible]

Ieeftal search using the Ilarray : o rescanning of matter chars

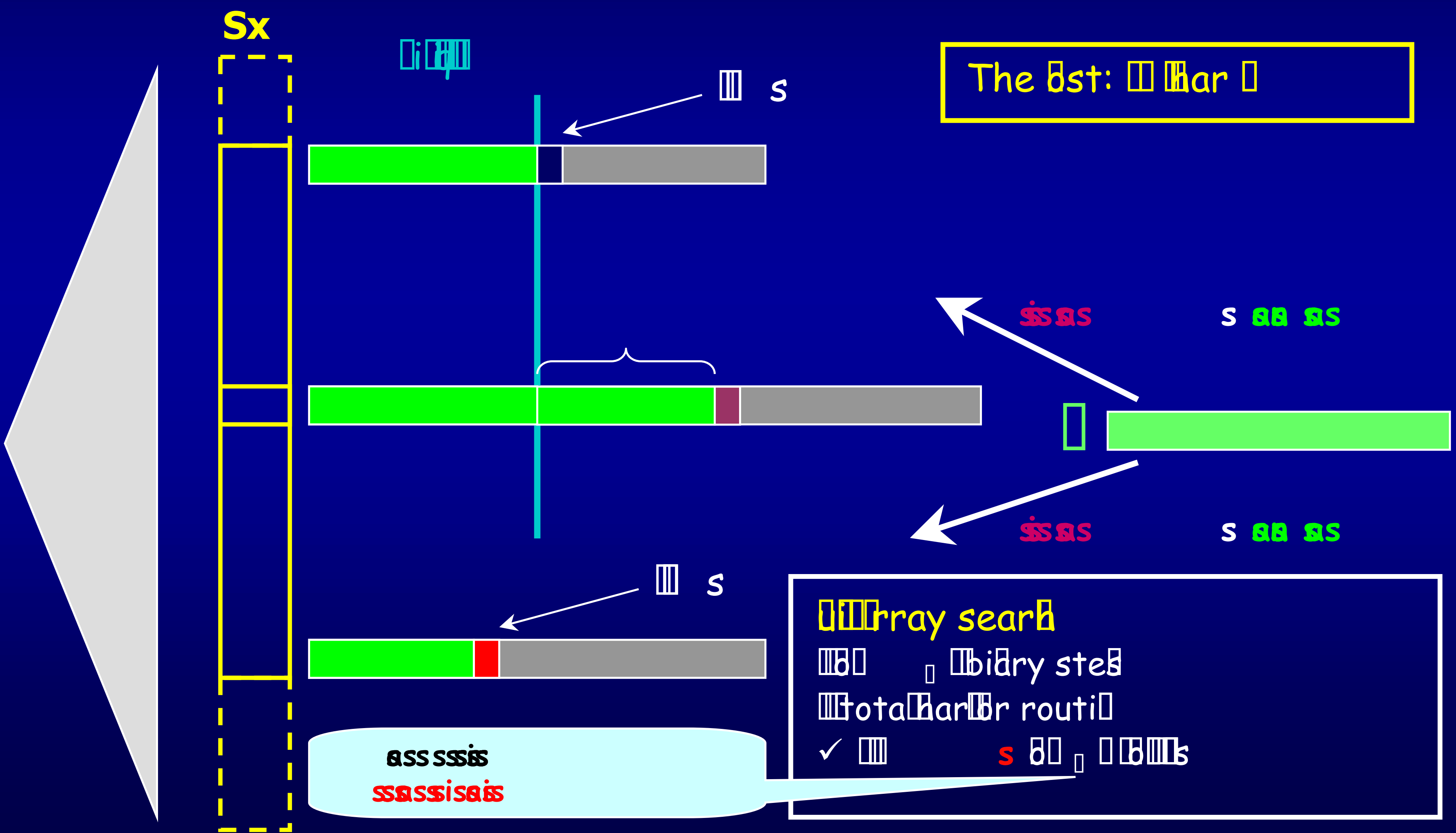


The cost: [illegible]ory accesses

dnrdeendbdd

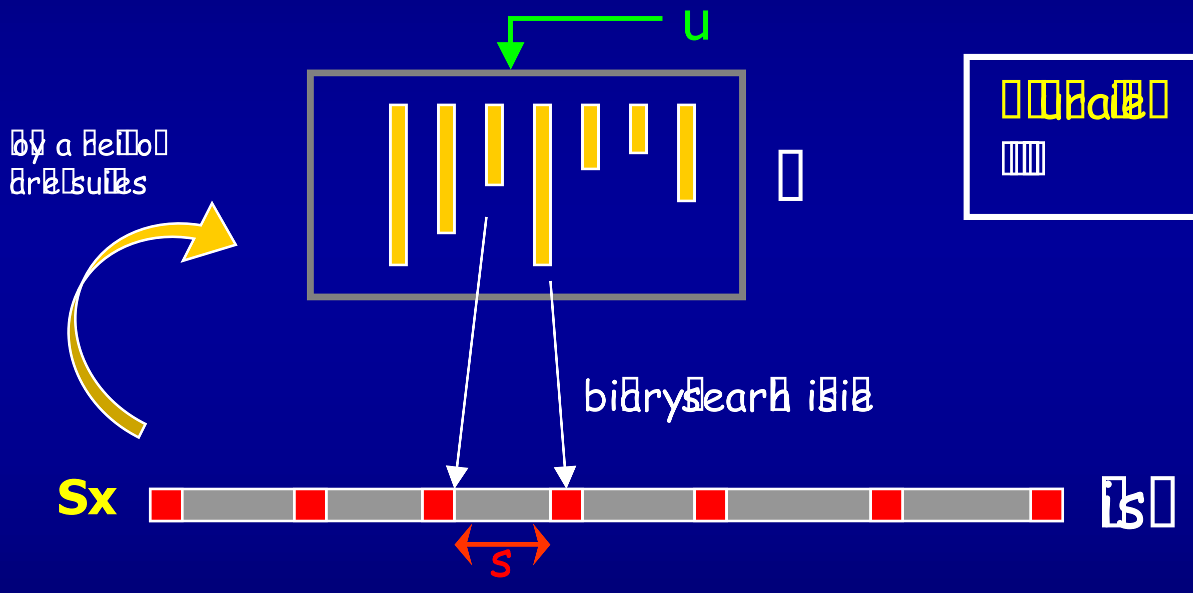
ed d

Beeffa search using the array : o resallo atter hars



rdd [illegible] ned

[illegible]it iter[illegible]eory : sae the sui[illegible]array a[illegible]oy soethi[illegible]eory


u
oy a [illegible]ei[illegible]o[illegible]
are[illegible]sui[illegible]es
bi[illegible]ry[illegible]earh [illegible]isie
Sx
s
is

[illegible]ura[illegible]e[illegible]
[illegible] s o[illegible] [illegible]s [illegible]o[illegible]s

☑ [illegible]ra[illegible]ter s ee[illegible]s o[illegible] [illegible] a[illegible]ue[illegible]s both [illegible]r[illegible]r[illegible]e a[illegible]s[illegible]e [illegible]

el idde e

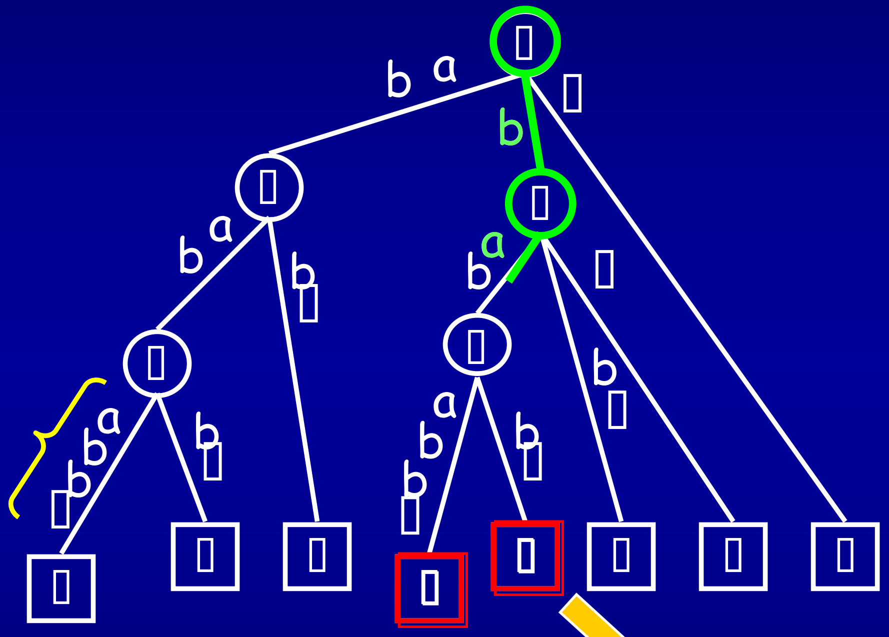
de id d

□t is a □□□e□trie built o□a□te□t su□□es

□ba ✓ □ar□ is a □th tra□rsa□

□□tie a□□□□ie

□□s□e □□□



□hat about □T i□e□ter□□□ory □

□ □bac□□tree to□□y
□ i□□□ty

}□i□□

T □abababb□

s s s s s

Ω□□s
Ω□□□s□

□re s□e □□ } □T tree □□ o□□rae

ssspisnssainsaisand pass

[illegible]

[illegible]

String

[illegible]

[illegible]

[illegible]

el dod oe

□e are □□ □th □□ oe□issues:

- □ □□□rray : □□□ty
- □ □□□tree : □□□ □□a□ Ω□□s
- □ □ybri□ □euristi□tu□o□the □r□r□□

□tree is ubiquitous i□ □□□□ a□□tio□:

- □ □to□□eys : i□e□rs□eas□
- □ □e□□tree : bou□□□th □eys □ ≤ □□□hars□

□□□trees □□rees □ → □ri□□ree □rra□□ossi□□

☑ □□ubou□□□h □eys

☑ □o□orst□se □□ou□ i□sear□ a□□te

☑ □ara□ee□□ti□□□□ratio

ded dnidrid ond

□tri□ ha□ arbitrary □□h:

- □ is□□□ □□t e□ure the stora□ o□ Θ□□stri□
- □ □ □y be u□b□ to store e□□b□ si□ stri□

□tri□stora□:

- □ □i□ers a□□to □t Θ□□stri□ □r is□□□
- □ □tri□□ari□o□ □e□ is□a□ss a□y be e□□si□

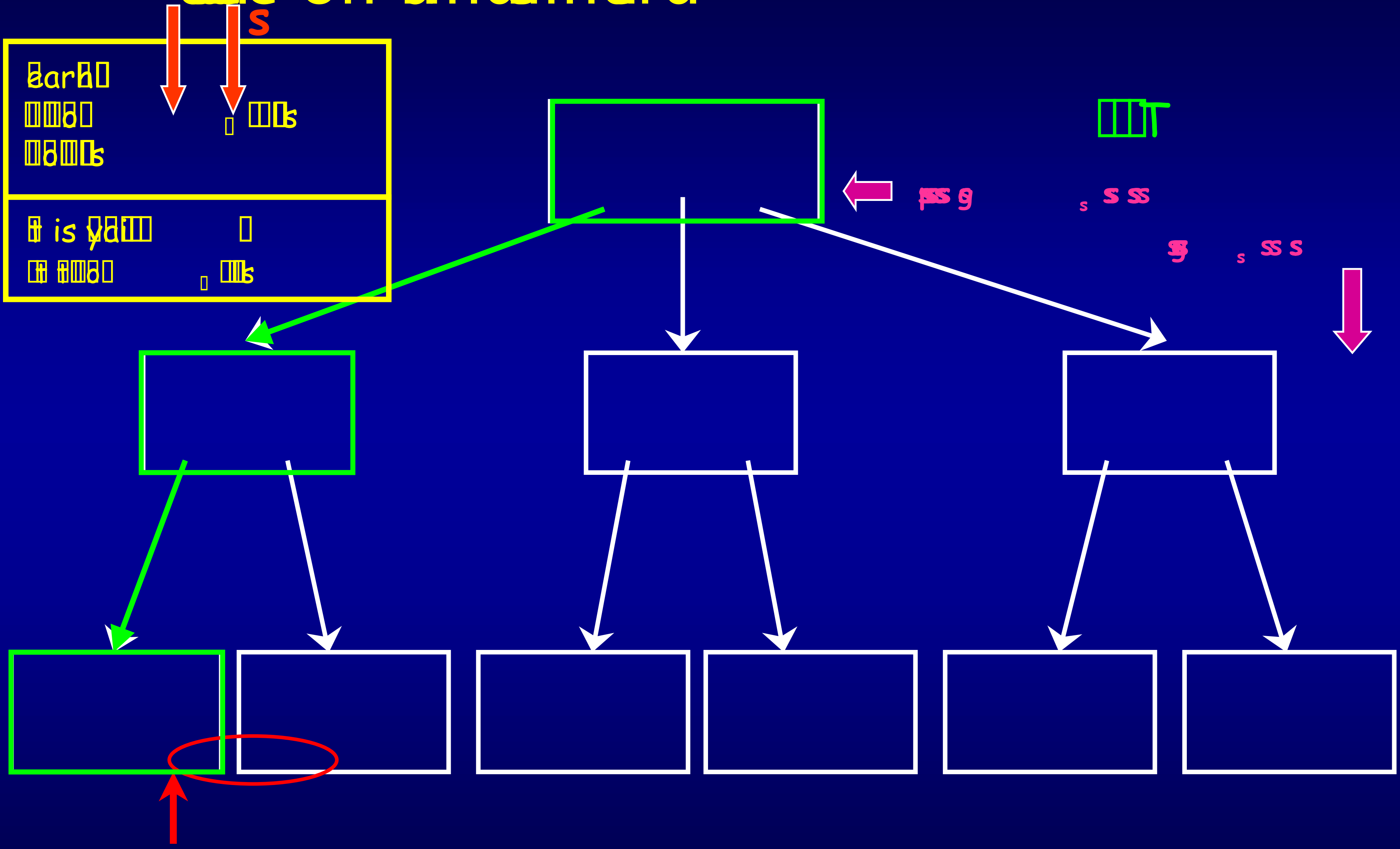
□tri□□i□ers or□□ti□o□see□so □r:

- ✓ □i□array: si□ but stati□□t o□i□□
- ✓ □tri□ trie: so□isti□te□a□□□ ei□e□ □□i□□□

□a□□the □ob□□ : Δ is a te□ □□□io□

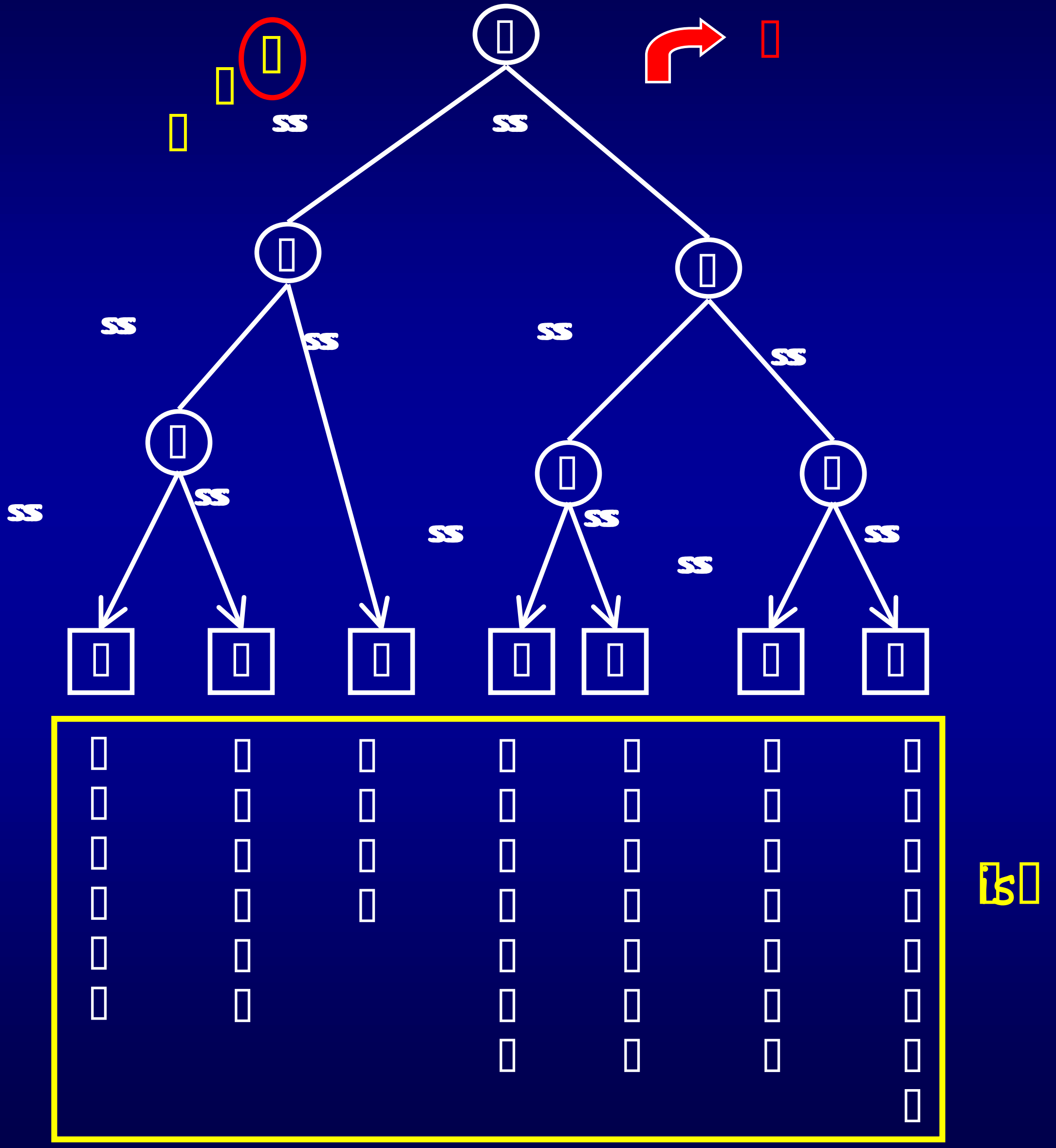
- ✓ □ar□□□ retrie□ a□□rre□s o□□i□ Δ□s te□s
- ✓ □te□□□□ i□ert or □□te a te□ T □o□ Δ

[illegible]



is [illegible]

[illegible]

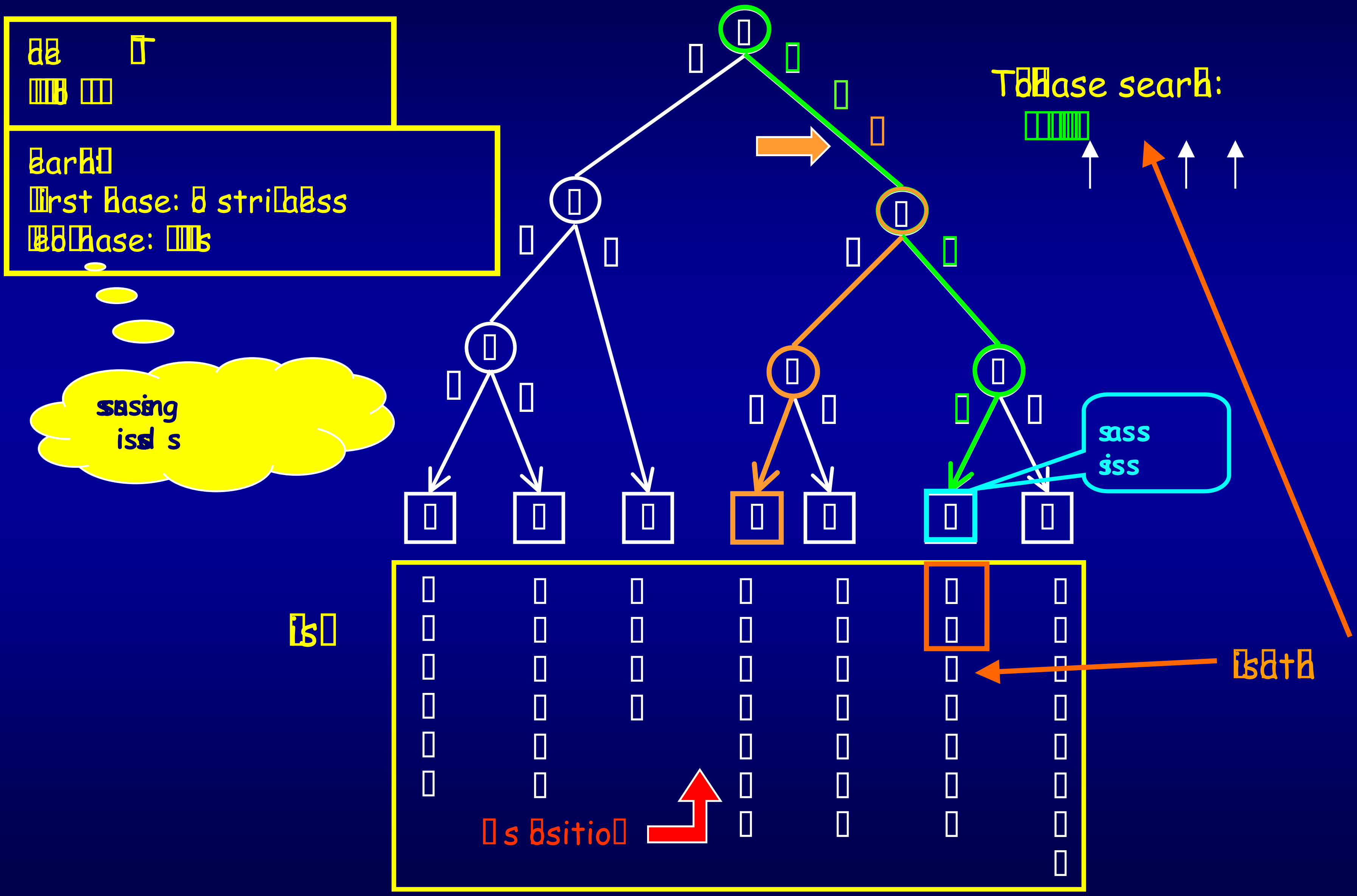


ss
ss
ss
ss
ss
ss
ss
ss
ss
ss
ss
ss
is

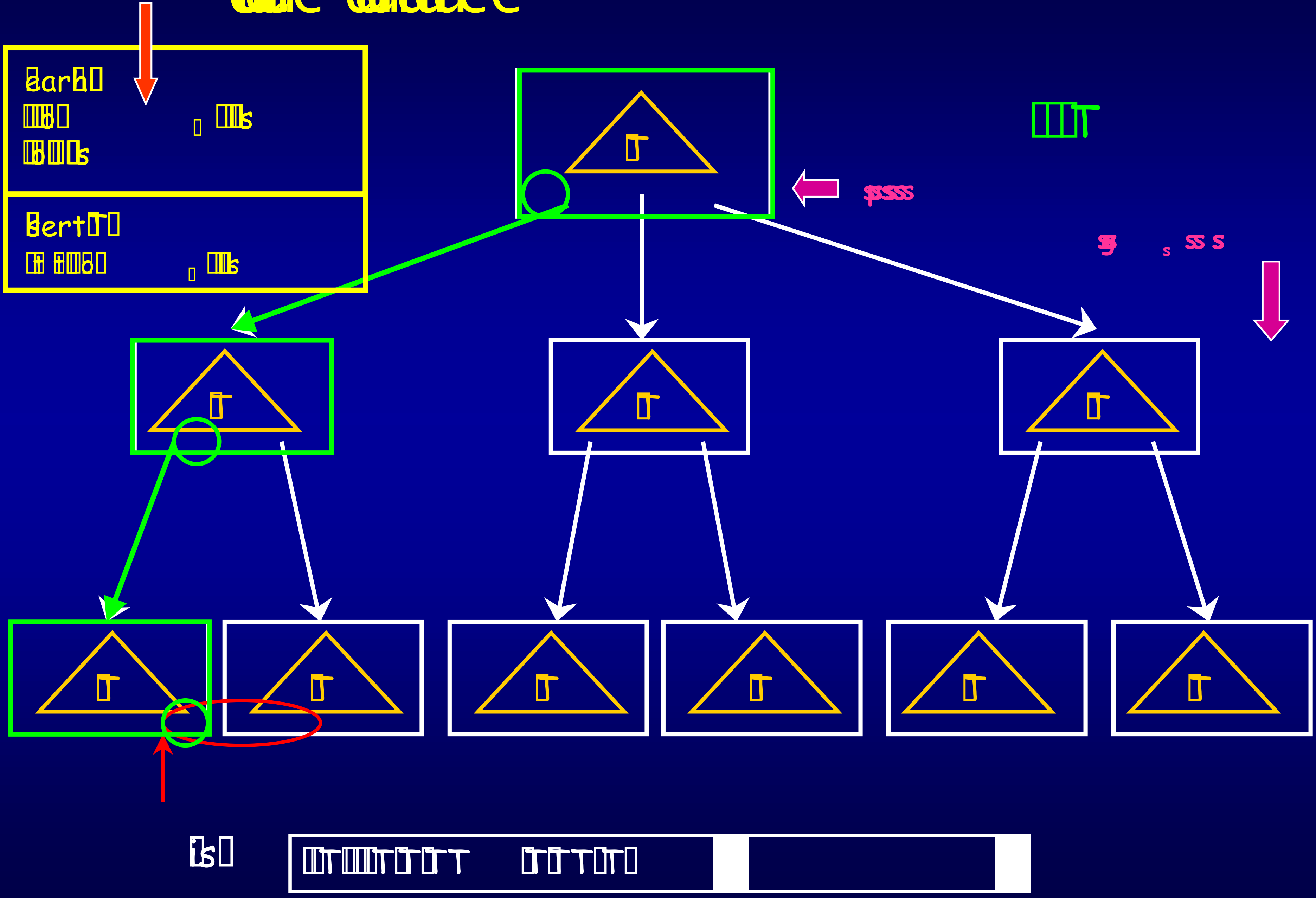
[illegible]

[illegible]

[illegible]arch:
First phase: [illegible] string pass
Second phase: [illegible]s

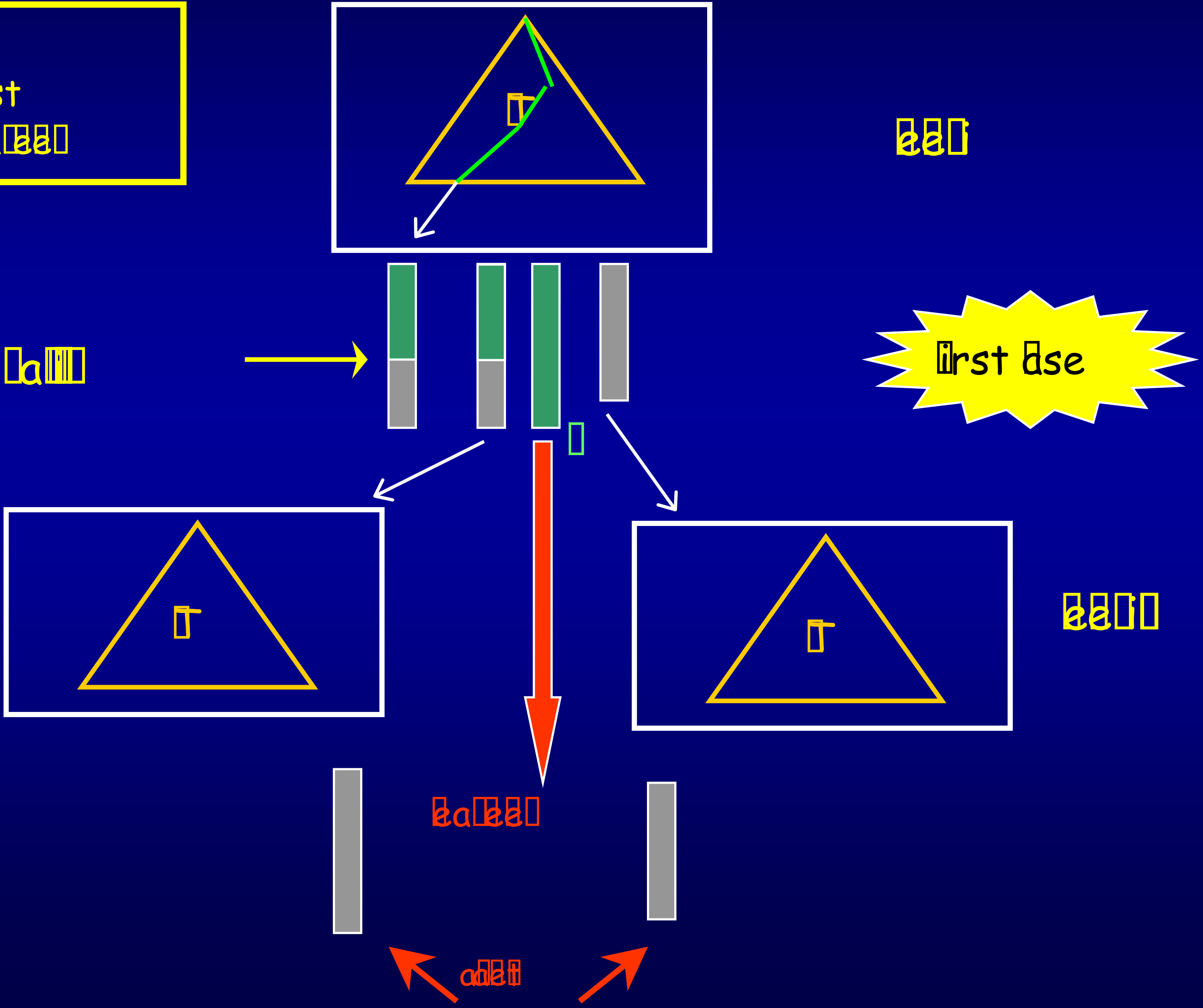


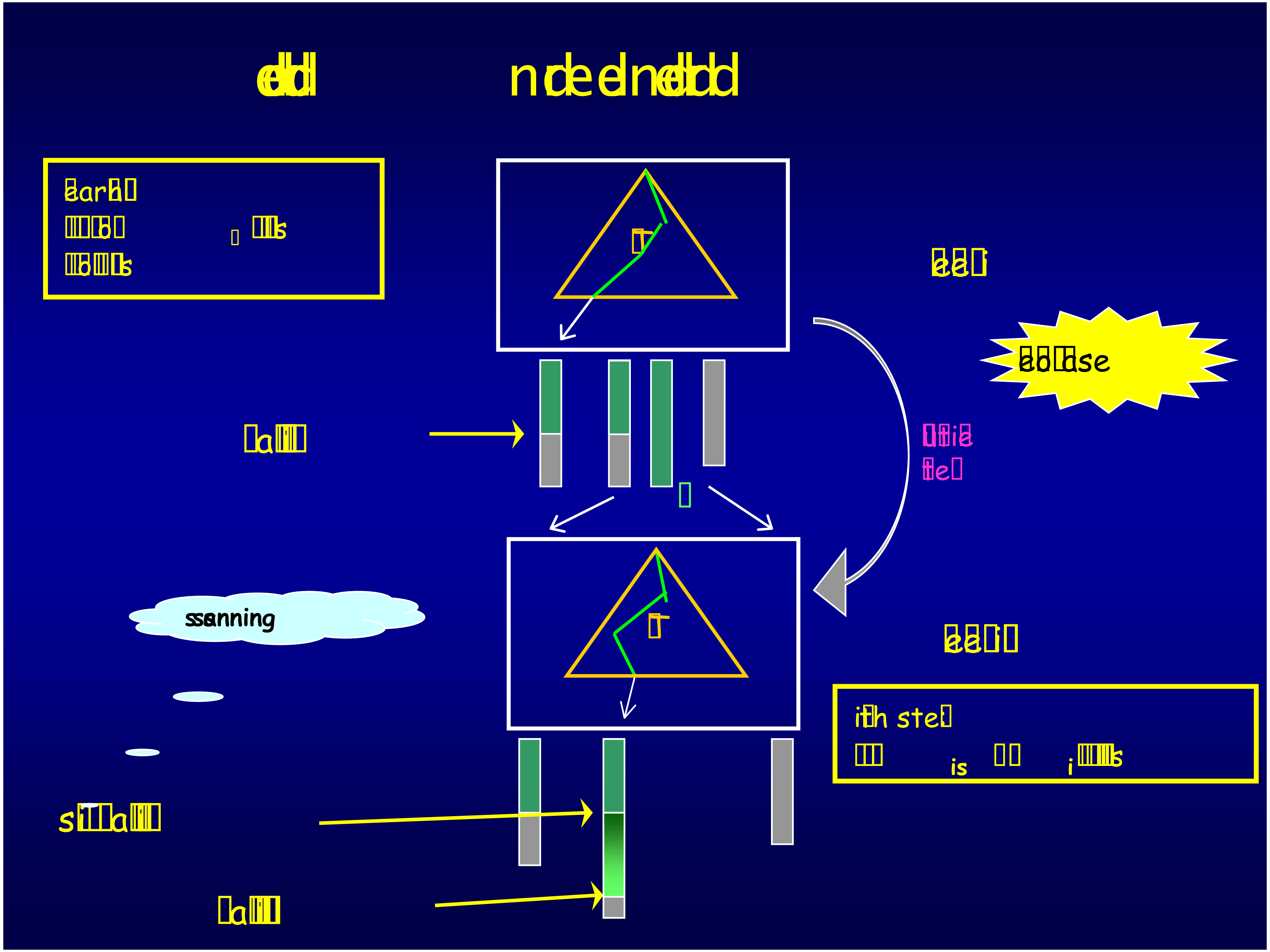
[illegible]



[illegible]

[illegible] ar [illegible]
[illegible] b [illegible] s ust
to b to the a [illegible]





scanning

ch dd

□tri□□tree □r□□ce□□: □erra□□□ossi□□□

- □ □ear□□□ ta□es □□□□o□□ □ □ □o□□s
- □ □□ate□□□ ta□es □□t □o□□ □ □ □□s
- □ □□ce is Θ□□□s□□ces

□si□□the □tri□□tree i□□ter□□□ory:

- □ □ear□□□ ta□es □□□□o□□ □ □ □o□ti□
- □ □□ate□□□ ta□es □□t □o□□ □ □ □ti□
- □ □□ce is Θ□□bytes
- ✓ □t is a sort o□□□□□ suf□□ array

□a□y other a□□□tio□□:

- □ □tri□□sorti□□ □r□ *et al*□□□□
- □ □i□tio□ary □at□i□□ □erra□□ *et al*□□□□
- □ □ulti□□□ stri□□queries □a□□sh *et al*□□□□

Addrildlltheerind

dre tri trees acaili ratie

String

gxitxxxbtrbxx

xtixxtrix inxignx

xrginx

rced ihcrddnidlerbnd

ie a tri tree de π e eie:

- □ □$_\pi$ set o a stris store at de π
- □ b aiu sie o π

interestin oerty:

- □ □ hos as do $_b$ abes ot eeb □ Δs strufure
- □ b is recte to the sae oucy o T$_\pi$ a b □

☑ The drer is b the dster are sear a ute oerations

☑ ur da queee T$_\pi$ as □ as ossibe

cd π id ednion

be π atuay otais at π :

- T π atria trie ia the stris o π
- The oiters to the hire o π
- oe auiiary a booe oratio

or bytes

aiioe

the stris are biary the T π osists o

- eacs biti to π s stris
- itera oeseah storia itear aue
- arseah storioe sie har

aeti T π taes: errai ossi

- bytes a oiteroase soutio
- bytes a roar eo o the biary tree strutture

del eid dd red d

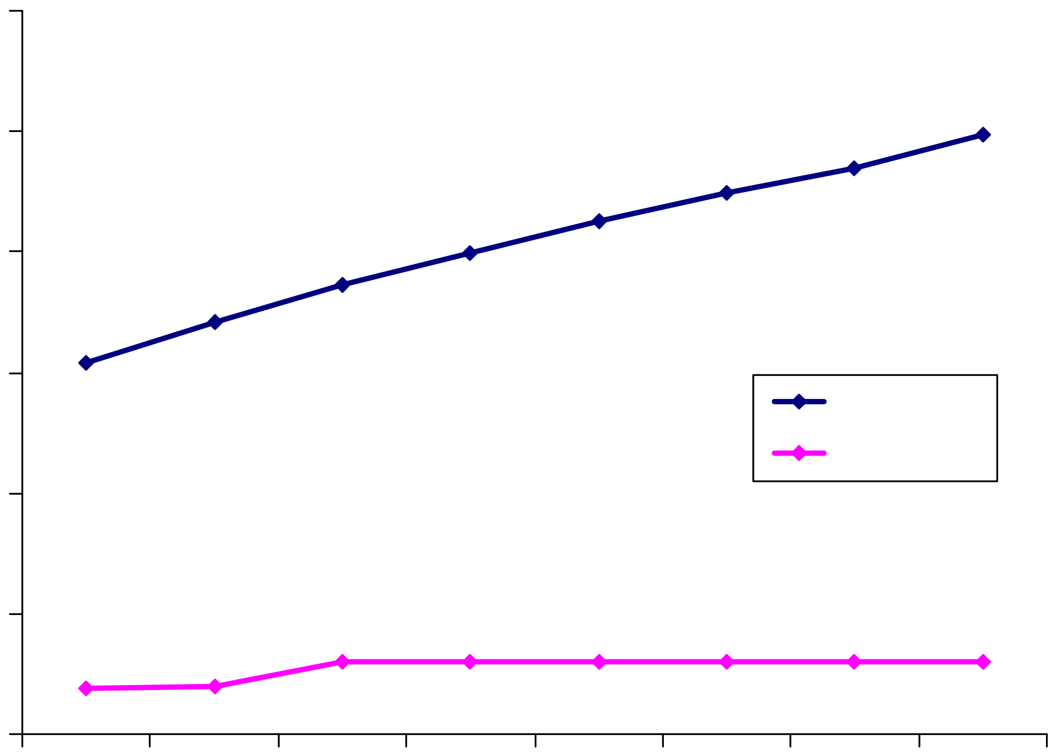
eriets hae sho that: erraid ossi

- earh
  - t taes about is aesses as the orst ase bou
  - t is ties aster tha rray searh
  - oarabe to Tree searh
- sert T in a bathe isertio
  - t is ties aster tha rei trees
  - etter ae ratio tha trees

To itatios:

- ae usae o is too uh
- The uate os are bou e

An edridnd



A new model

[illegible] the [illegible] π:

- [illegible]tri[illegible] [illegible]oi[illegible]ters a[illegible] [illegible]hi[illegible] [illegible]oi[illegible]ters i[illegible] bytes
- [illegible]te[illegible]ers i[illegible] the [illegible]es o[illegible] T π store[illegible] i[illegible] a [illegible]i[illegible]atio[illegible]it
  - [illegible]eri[illegible]ts sho[illegible] that [illegible] are [illegible]ery s[illegible] ✓ [illegible] byte
- [illegible] [illegible]o[illegible] [illegible] i[illegible] T π [illegible]
  - [illegible]hou[illegible] be s[illegible] su[illegible] a[illegible] [illegible] basi[illegible] [illegible]tio[illegible] [illegible]

[illegible] results o[illegible] the su[illegible] [illegible] bi[illegible]ary trees:

- [illegible]i[illegible] [illegible] bits a[illegible] basi[illegible] [illegible]tio[illegible] [illegible] [illegible]obso[illegible]
- [illegible] bits a[illegible] [illegible]re [illegible]tio[illegible] [illegible] [illegible]u[illegible]o *et al.*[illegible]

T[illegible] s[illegible]ies o[illegible] our [illegible]:

- [illegible] T π is s[illegible] about a thousa[illegible] o[illegible] stri[illegible]
- [illegible]tio[illegible] [illegible] [illegible]r tra[illegible]ersa[illegible]
- [illegible]i[illegible] is [illegible]t the o[illegible]y resour[illegible] [illegible] is surel[illegible] [illegible]ie[illegible]

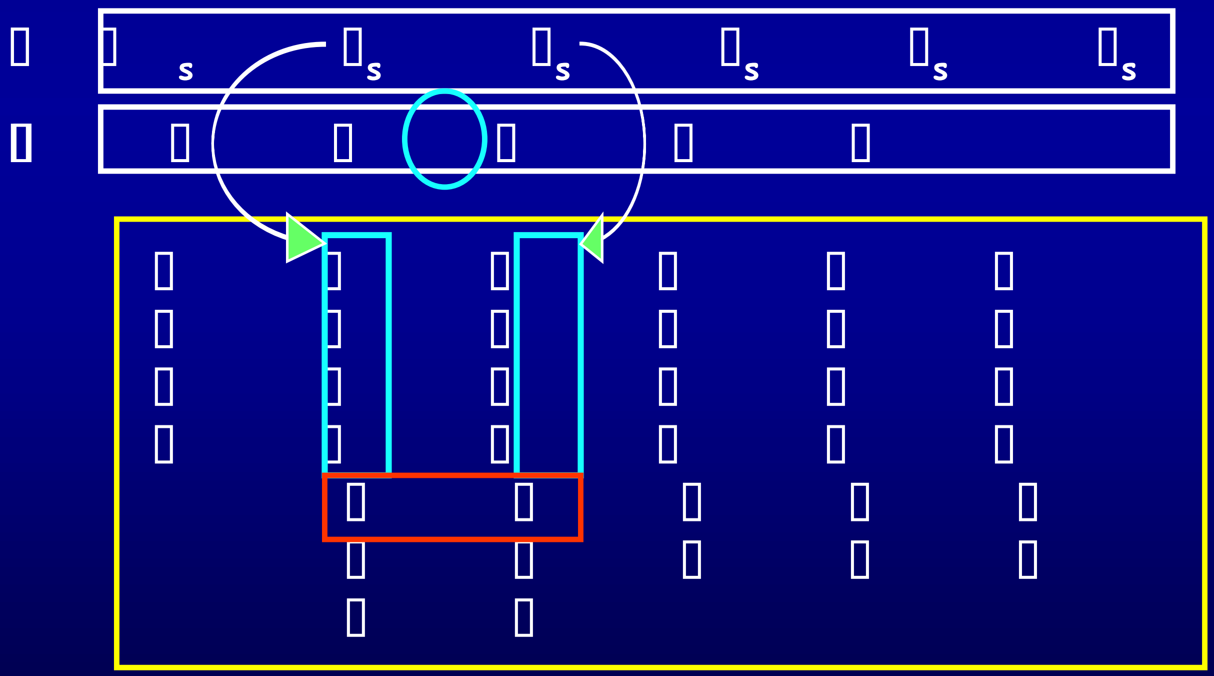
[illegible]T $_{\pi}$[illegible]s toooy [illegible]ay be [illegible]roe[illegible] [illegible]

[illegible]er[illegible]uso[illegible]

Tak[illegible] the i[illegible]r[illegible]er [illegible]isit o[illegible]T $_{\pi}$:

[illegible]array o[illegible]oi[illegible]ters to [illegible] $_{\pi}$[illegible]s stri[illegible] i[illegible]e[illegible]T $_{\pi}$ [illegible]ea[illegible]es[illegible]

[illegible]array o[illegible]s bet[illegible]ee[illegible]stri[illegible] a[illegible]a[illegible] i[illegible]



[illegible] $_{\pi}$[illegible]s stri[illegible]
o[illegible]is[illegible]

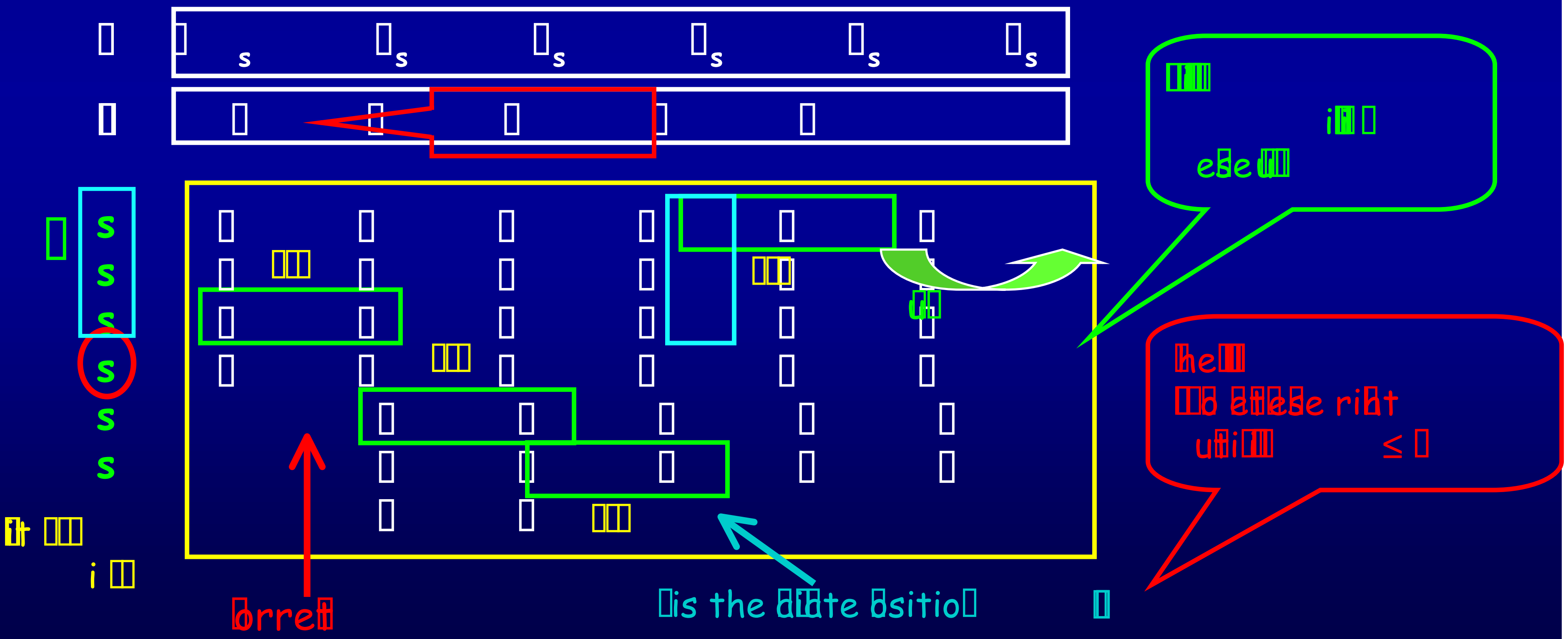
[illegible]T$_{\pi}$[illegible]s to[illegible] [illegible]y be [illegible]o[illegible]e[illegible] [illegible]

[illegible]er[illegible]so[illegible]

Ta[illegible] the i[illegible]r[illegible]r [illegible]sit o[illegible]T$_{\pi}$:

[illegible]rray o[illegible]i[illegible]ers to [illegible] $_{\pi}$[illegible]s stri[illegible] i[illegible]e[illegible]T$_{\pi}$ [illegible]ea[illegible]es[illegible]

[illegible]rray o[illegible]s bet[illegible]e[illegible]stri[illegible] a[illegible]e[illegible] i[illegible]



s
s
s
s
s
s
i[illegible]
e[illegible]e[illegible]
[illegible]he[illegible]
[illegible]o [illegible]e[illegible]e rig[illegible]t
u[illegible]i[illegible] ≤ [illegible]
[illegible]orre[illegible]
[illegible]is the [illegible]te [illegible]sitio[illegible]

dh dd

□be□ π □tai□ □t □ □ $_{\pi}$□ :

- □ □biter array□□
- □ □iteer array□ore□by □ti□atio□it

□ear□i□ s □sitio□a□□ $_{\pi}$□s stri□:

- □ □□ to □t□ the is□a□ □tai□□e π
- □ □array s□ : □hars a□iteer □ariso□
- □ □stri□a□ss to the □i□te stri□□□s

i□ □is about a thousa□ o□stri□:

- □The □ to □t□ the is□a□ ta□s □□ μs
- □The t□ array s□ are □ry □st: □□ μs □□e □ee□□i□
- □The stri□a□ss i□t □□y i□ee□a□sear□□

☑ □e □ou□ as be□re□about □ bytes o□s□e i□□ati□ !!

eldd eld

- □roi□e a □ubi□□eeatatio□ o□ □ri□ □trees
  - ✓ □e□r to □r□e□y□□r the □□
- □ath queries: □o□to i□□a □bee□ tree □r □ath queries □
  - ✓ □□authord□□□oo□□□
- □uti□□esio□ □substri□ queries: □ti□e□ re□r□ search□
  - ✓ □ay □e □ □□e□oe□tri□ □ata stru□tures i□ □ri□ □trees □
- □trea□ o□ queries□ □ssib□ biase□ □tri□ □tree is □t o□ti□a□
  - ✓ □ay □ □□ □□se a sea□□sti□□e□□ □□ator □Tara□□□
- □a□e□ bi□□ous tries: □o e□□t □ra□e□ri□tio□ o□□
  - ✓ □tri□ □tree are ba□e□ but □□e□ □

[illegible]

[illegible]

String

[illegible]

[illegible]

[illegible]

del dnielriond

e ha alea shothat the ilrray athe
rresilarray sui to buithe tritree

bo e buithe arrays a

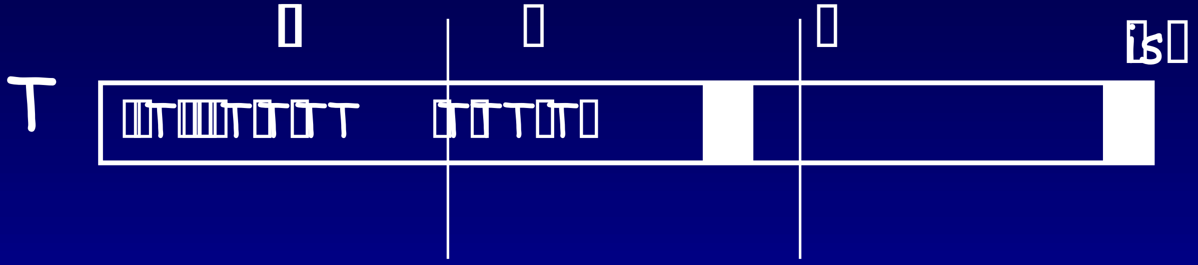
- ory arith are ilet
- aiirt eilet but sa sui   auser *et a*
- ∃ theoretia oidarithbut oite asa ost
  rrai *et a*

There eists aarithhi is   ea tes *et a*

- Theoreti uetabe: ubi eity
- atia ry aeaior eroa asa ouy
- its asyotis abe iotith sa tri   auser *et a*

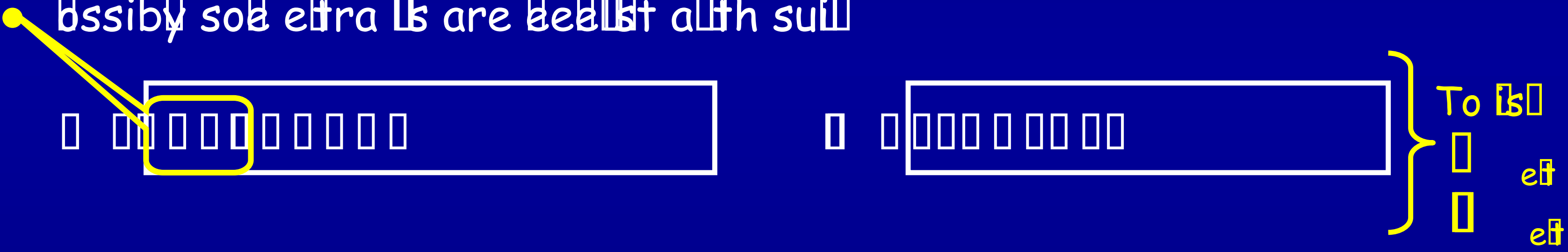
idArrdelrd

itdedd



etd i[illegible]ory the irst ie T[illegible]bui[illegible]a[illegible]or the
sues [illegible]i[illegible] start at ositios [illegible]

ossiby soe etra [illegible]s are ee[illegible]st a[illegible]th sui[illegible]

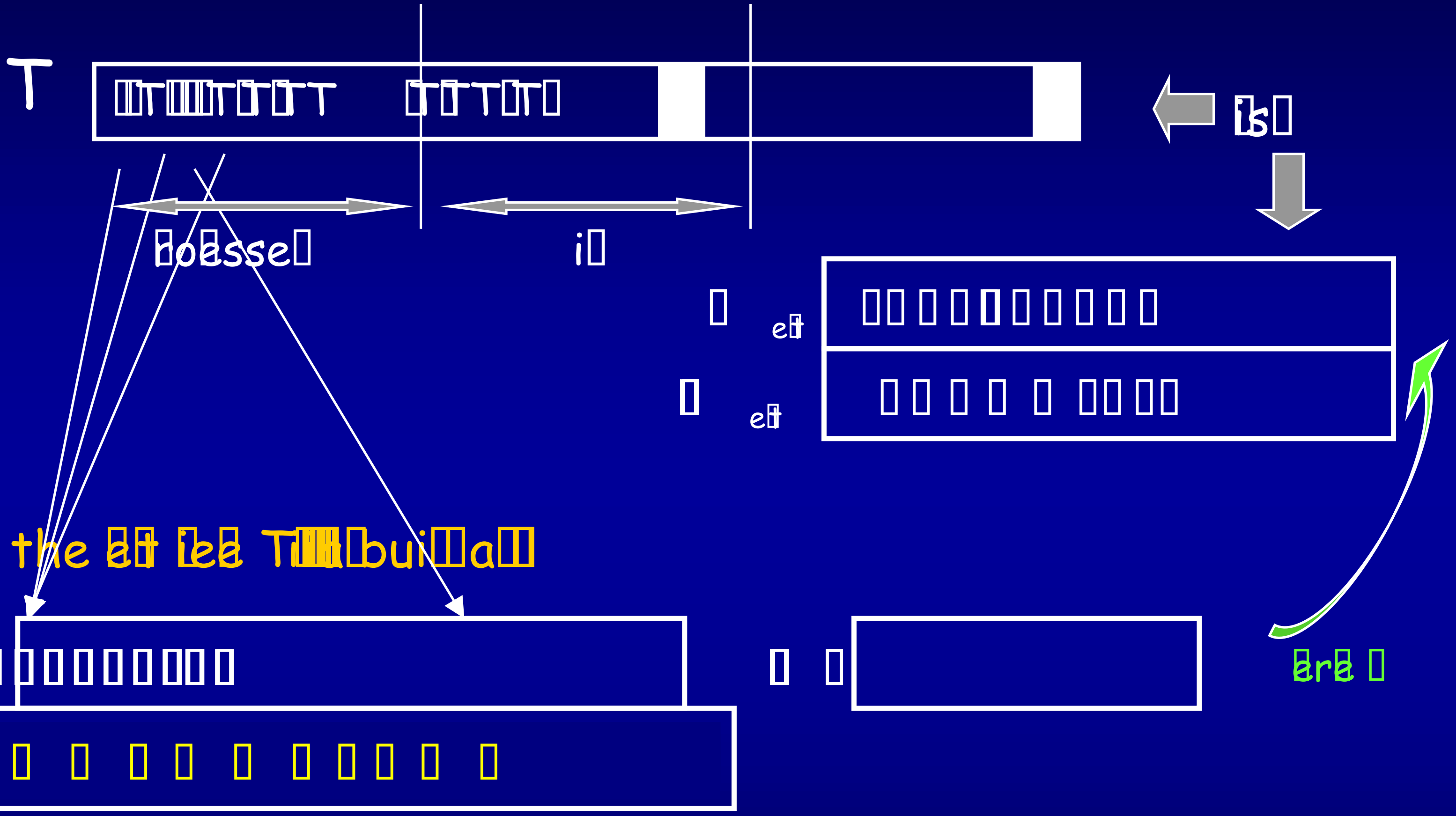


[illegible]tio[illegible]: [illegible]e hae [illegible] $_{et}$ a[illegible] $_{et}$ or the sues starti[illegible]i[illegible]ie
T[illegible] ete this to the sues starti[illegible]Ti[illegible]

[illegible]e ai[illegible]at ee[illegible]ti[illegible]iy bu[illegible]s

[illegible] Arr [illegible]

[illegible]



[illegible]tio[illegible]:

[illegible]eth i[illegible]eory the [illegible]et [illegible]ee T[illegible] bui[illegible]a[illegible]

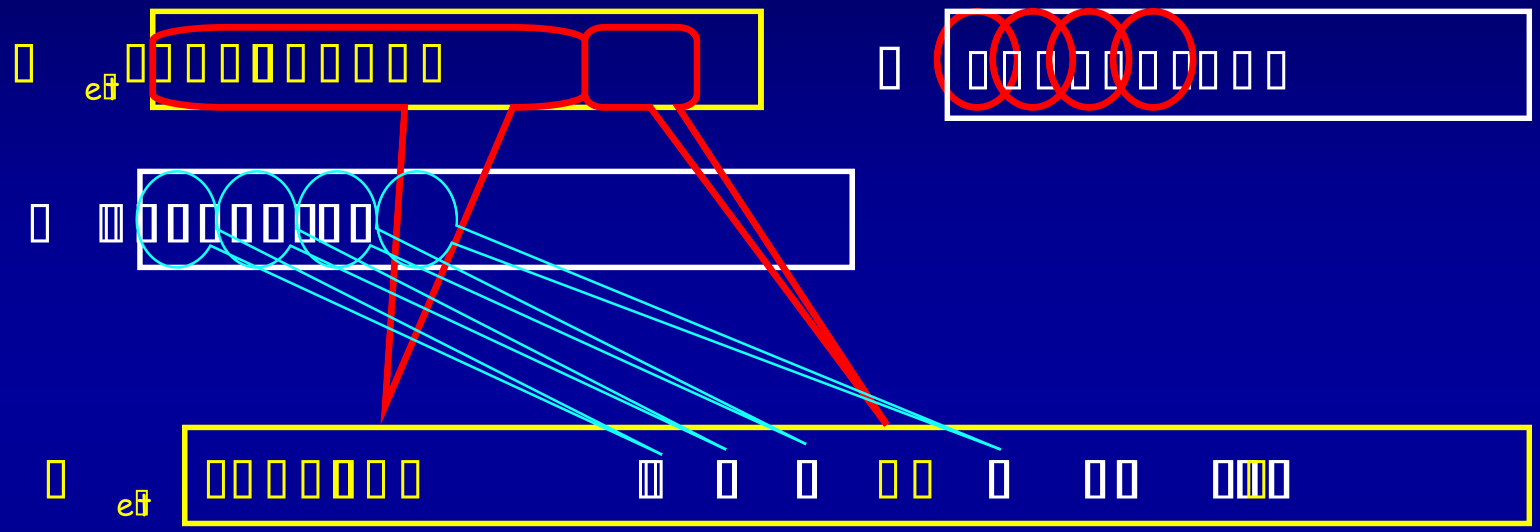
a[illegible]T[illegible] is a[illegible]bute a[illegible]eory [illegible]uti[illegible]array [illegible]

- [illegible]earh [illegible]ithi[illegible] the [illegible]ositio[illegible] o[illegible] eah sui[illegible] starti[illegible] [illegible]to T[illegible]
- This ta[illegible]es [illegible]s a[illegible]tua[illegible]y bu[illegible]s[illegible]

idArrdeld [illegible]

ihdve eld

[illegible]ere[illegible] [illegible] ~e[illegible]~ a[illegible] by usi[illegible] the array [illegible]a a is[illegible] sa[illegible]



The [illegible]ity o[illegible] the i[illegible]h ste[illegible] is:

- [illegible]ethi[illegible]T[illegible]aes [illegible]s — bu[illegible]s[illegible]
- [illegible]ui[illegible]a[illegible]taes [illegible]ati[illegible]y o [illegible]s — or e[illegible]rabs[illegible]
- [illegible]outi[illegible]a a sa[illegible] o[illegible]aes [illegible]s — bu[illegible]s[illegible]
- [illegible]eri[illegible] e[illegible] i[illegible]aes [illegible]s — bu[illegible]s[illegible]

✓ [illegible]era[illegible] the a[illegible]rith[illegible]e[illegible]utes [illegible] [illegible] [illegible]s i[illegible]ati[illegible]e — [illegible]aiy bu[illegible]s[illegible]

a ssas
isisa issd s

dindoridhd

dortistris is silar to sortisulles

String

gritxnxbxtrixx

xtixxtrix inxignx

xrginx

[illegible]

[illegible]iteraleoryk d alotialbout[illegible]

- [illegible]ia a odtle[illegible]trie e et Θ[illegible]o[illegible] [illegible]tie
- oer bou[illegible]des do[illegible]the sorti[illegible]o[illegible]eedts[illegible]

[illegible]eteraleoryk oulekt to ahiee:

Θ[illegible]o[illegible] [illegible]s

but[illegible]

[illegible]tri[illegible]trees ad[illegible]to ahiee O[illegible]o[illegible] [illegible]s

[illegible]Threedy quidsort eets O[illegible]o[illegible] [illegible]s eteyaei[illegible]

The situatio[illegible]s [illegible] odte[illegible]the odity ees o[illegible] :

- breal[illegible]stris ido hars is aoe[illegible]
- the stri[illegible]sie redtie to [illegible]

el dndio

et us eie [illegible] :

- [illegible] a[illegible] or stris sder tha[illegible]
- [illegible] a[illegible] or stris der tha[illegible]

stris are iisibe eeryhere it is otiid :

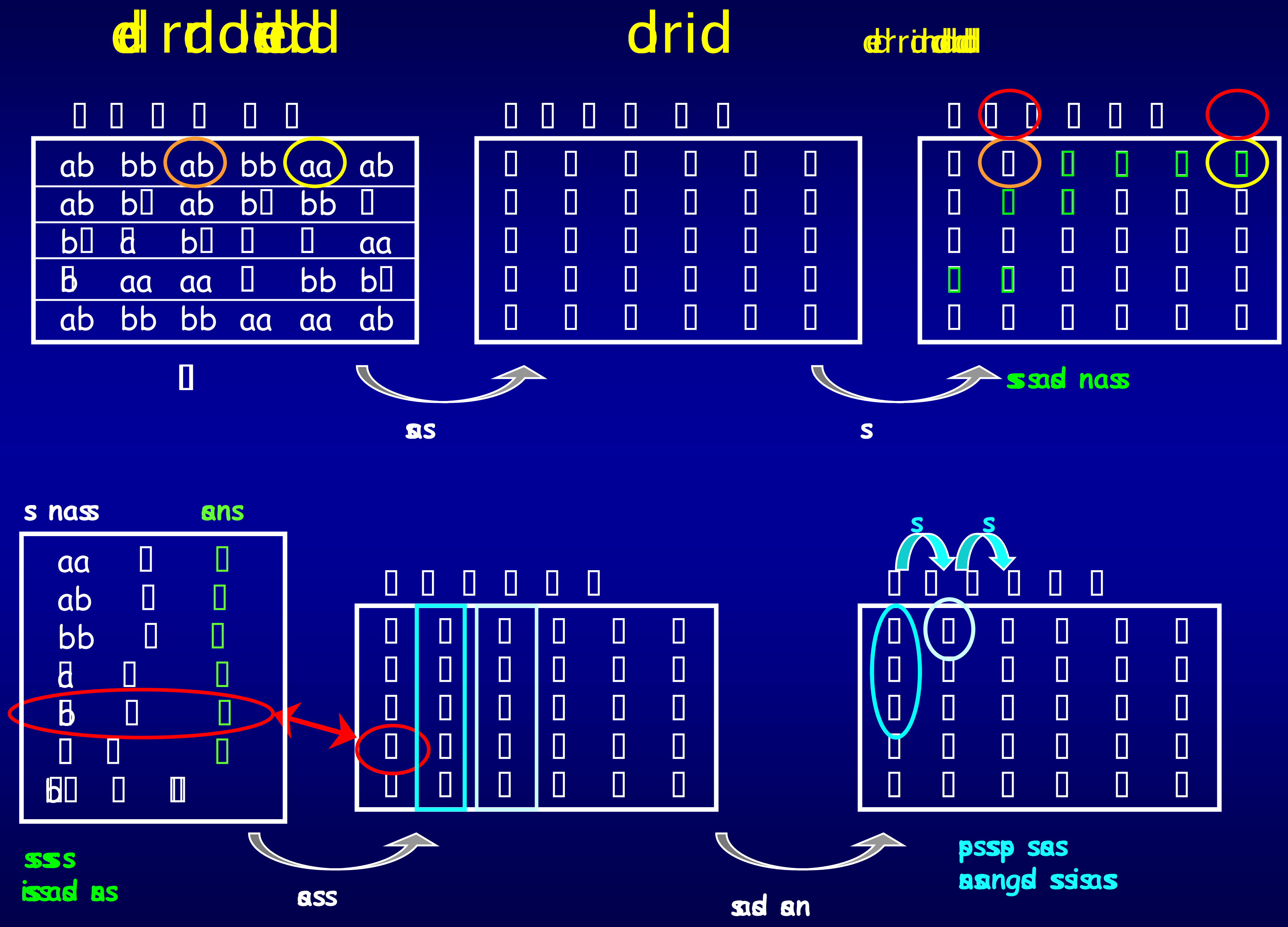
$\underbrace{\quad [illegible] \quad}_{\text{short}} \quad \underbrace{\quad [illegible] \quad}_{\text{o}}$

stris are oy iisibe ietera eory:

$\underbrace{\quad i \; [illegible] \quad}_{\text{short}} \quad \underbrace{\quad [illegible] \quad}_{\text{o}}$

stris y be hoe io iees: [illegible] s

- t is a rabie aorith erraiThoru
- The aerae strieth shoube Ω [illegible]



drdddedd
drid
edrrihddd
ab bb ab bb aa ab
ab b ab b bb
b a b aa
b aa aa bb b
ab bb bb aa aa ab
ssad nass
sas
s
s nass
ans
aa
ab
bb
a
b
b
s
s
sss
issad as
ass
sad an
pssp sas
asangd ssisas

drddid drid dnd d

□	□	□	□	□	□
ab	bb	ab	bb	aa	ab
ab	b□	ab	b□	bb	□
b□	a	b□	□	□	aa
b	aa	aa	□	bb	b□
ab	bb	bb	aa	aa	ab

□t

□	□	□	□	□	□
□	□	□	□	□	□
□	□	□	□	□	□
□	□	□	□	□	□
□	□	□	□	□	□
□	□	□	□	□	□

□ashe□a□sorte□stri□

□	□	□	□	□	□
□	□	□	□	□	□
□	□	□	□	□	□
□	□	□	□	□	□
□	□	□	□	□	□
□	□	□	□	□	□

Tab□ T a□er □ra□r□a□□

□e the sur□y

s s

□	□	□	□	□	□
□	□	□	□	□	□
□	□	□	□	□	□
□	□	□	□	□	□
□	□	□	□	□	□
□	□	□	□	□	□

pssp sas
asngd ssias

sasd an

□	□	□	□	□	□
□	□	□	□	□	□
□	□	□	□	□	□
□	□	□	□	□	□
□	□	□	□	□	□
□	□	□	□	□	□

s

s

eledded

- □se the □rious □□
  - ✓ o□stri□ i□the □se o□i□□sibi□ty o□e□tera□□□ory
  - ✓ □etter a□□sis □or the ra□i□□a□orith□
- □□□ a□those a□rith□
- □hat about □□□□bi□□ous stri□sorti□a□rith□ □
  - ✓ □ost o□the□are base□o□tries
  - ✓ □rbitrary □□h □eates a □t o□□obe□□
  - ✓ □obab□ the ra□i□□a□oa□ □□he□□this □se too

dcrdeeldd neded

ds sae oerhea the ta to dy or usi a utet ie

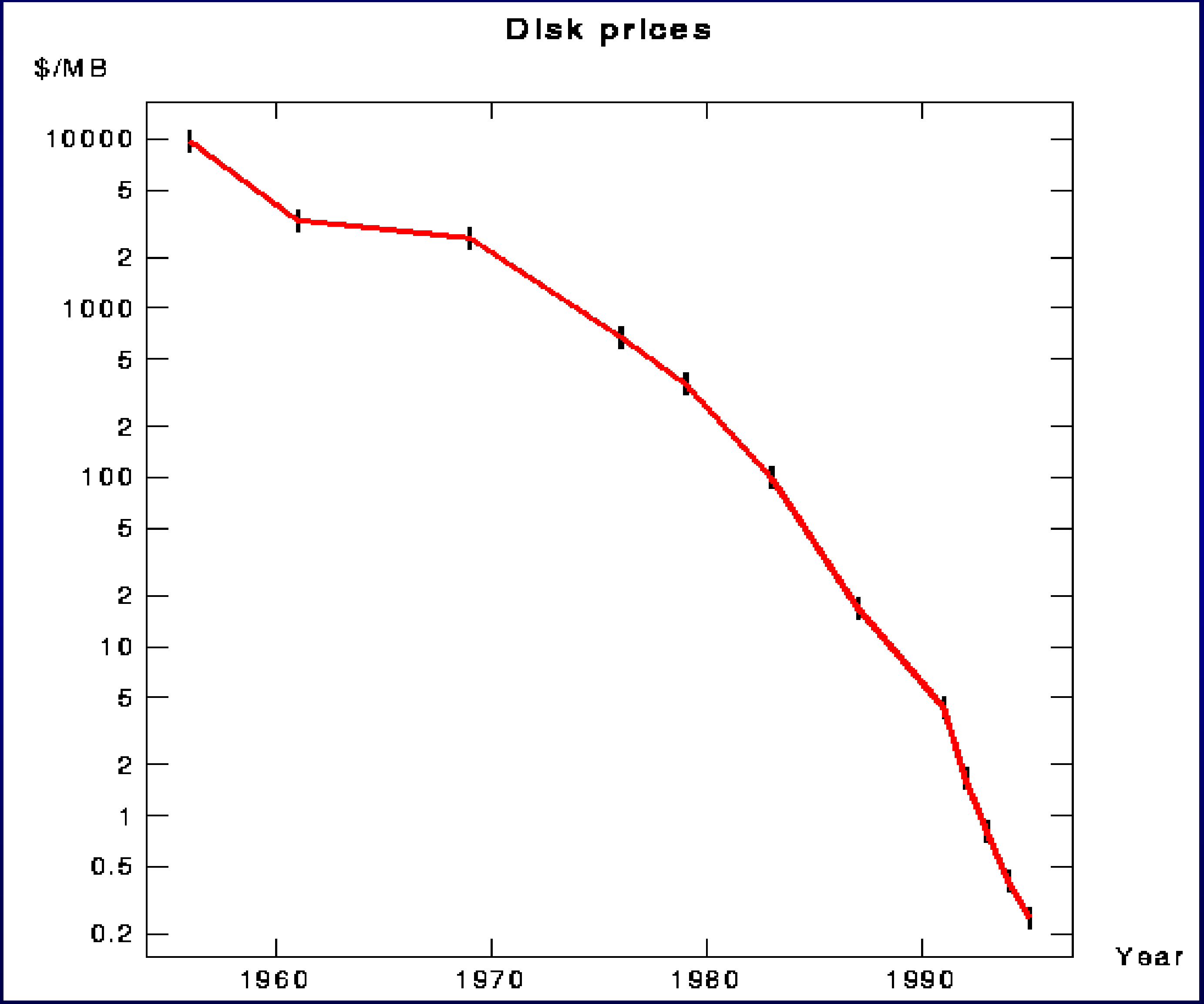
String

gxitxxxbrix

xtixxtrix inxignx

xrginx

[illegible]



Disk prices
$/MB
10000
5
2
1000
5
2
100
5
2
10
5
2
1
0.5
0.2
1960
1970
1980
1990
Year

[illegible]

- ✓ □□essio□has t□ □sitie e□□s:
  - ✓ □□ sa□□
  - ✓ □r□r□□ i□o□□□
    - ✓ □etter use o□□□ry □□□ □se to □o□ssor
    - ✓ □ease□□s□a□□□ry ba□□h
    - ✓ □□□□□□ai□□see□ti□
    - □ □s□e□□□s □□□essio□□st□ss□□ □

☑ □e□estab□s□he□ : □ is □re e□□□□to store □ta i□□□esse□□r□tha□u□□esse□

☑ □□th i□the □□□says: □□ o□i□□tio□is □sey re□te□to ti□ o□i□□tio□□a □s□□□ry syste□

☑ □□ re□ase□i□□ar□ □□the □□□ry e□□sio□ Te□□□y □□T□□□i□o e□□□rs □□
  - ✓ □ub□ □□ry at about sa□ □st a□□r□r□□

IBM Journal of Research and Development

[illegible]

[illegible]are [illegible]stat[illegible]s
[illegible]o[illegible]to [illegible]

[illegible]ssic[illegible] i[illegible]es use Θ[illegible] [illegible] bits o[illegible] stora[illegible]

- [illegible] array: [illegible] [illegible] o[illegible] tie
- [illegible]tri[illegible]tree: [illegible] [illegible] o[illegible]s

[illegible] su[illegible]trees use [illegible] [illegible] Θ[illegible] bits o[illegible] stora[illegible]

[illegible]uro et a[illegible]

[illegible]ay [illegible]ee[illegible]

[illegible]er[illegible]tatio[illegible] [illegible]t be ay [illegible]o[illegible]

✓ [illegible] biary te[illegible]s [illegible] [illegible] [illegible]er[illegible]tatio[illegible] o[illegible] [illegible]

[illegible] su[illegible]array uses Θ[illegible] bits o[illegible] stora[illegible] [illegible]ossi[illegible]tter[illegible]

✓ [illegible]uery ti[illegible] is [illegible] [illegible] [illegible] ε ti[illegible]

[illegible]ay [illegible] a[illegible]ie[illegible] o[illegible]o[illegible]essi[illegible] te[illegible]s [illegible]
[illegible]s i[illegible] the [illegible]se o[illegible]r[illegible] [illegible]base[illegible]es

el cod ed

- [illegible]t:
  - [illegible] ostatsie al abet Σ
  - [illegible] arbitrariy o ter Tber Σ

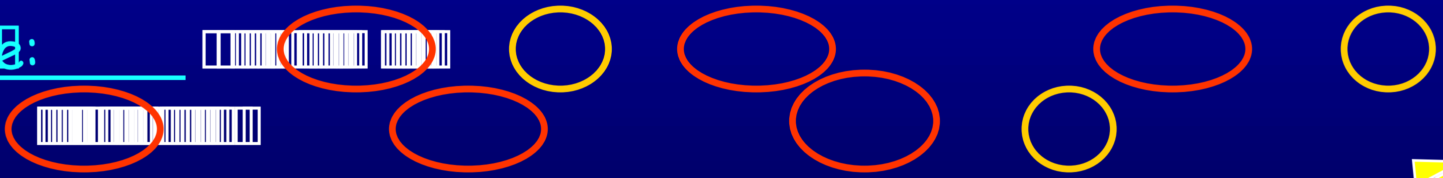
- uery o a arbitrary stri[illegible]
  - out the o rres o i T
  - ote the osition o the o rres o i T

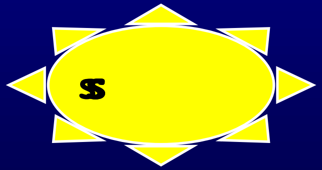
[illegible] at e iti re titi ss i the i t to squee the i [illegible]

oes it e ist a ortu sti ie[illegible]

ae: [illegible]

[illegible]





- out the as o oe [illegible]
- ote o ae o at i isa [illegible]

eldeld

erichdidd

[illegible]ri[illegible]data structure [illegible]esi[illegible] a [illegible]essio[illegible] te[illegible]iques:

- ✓ [illegible]i[illegible] array [illegible]ata stru[illegible]ure
- ✓ [illegible]urro[illegible]heeler Tra[illegible]or[illegible]

✓ bi[illegible]essio[illegible] a[illegible]orith[illegible]

The theoreti[illegible] resu[illegible]:

- ✓ [illegible]uery [illegible]ity: [illegible]o[illegible] ε [illegible]ti[illegible]
- ✓ [illegible] o[illegible]: [illegible] $[illegible]_T$ [illegible]o[illegible]bits

✓ o[illegible]T is [illegible]essib[illegible]

[illegible]th order e[illegible]i[illegible]e[illegible]tro[illegible] [illegible]y be o[illegible]

The [illegible] stu[illegible]s that this resu[illegible]:

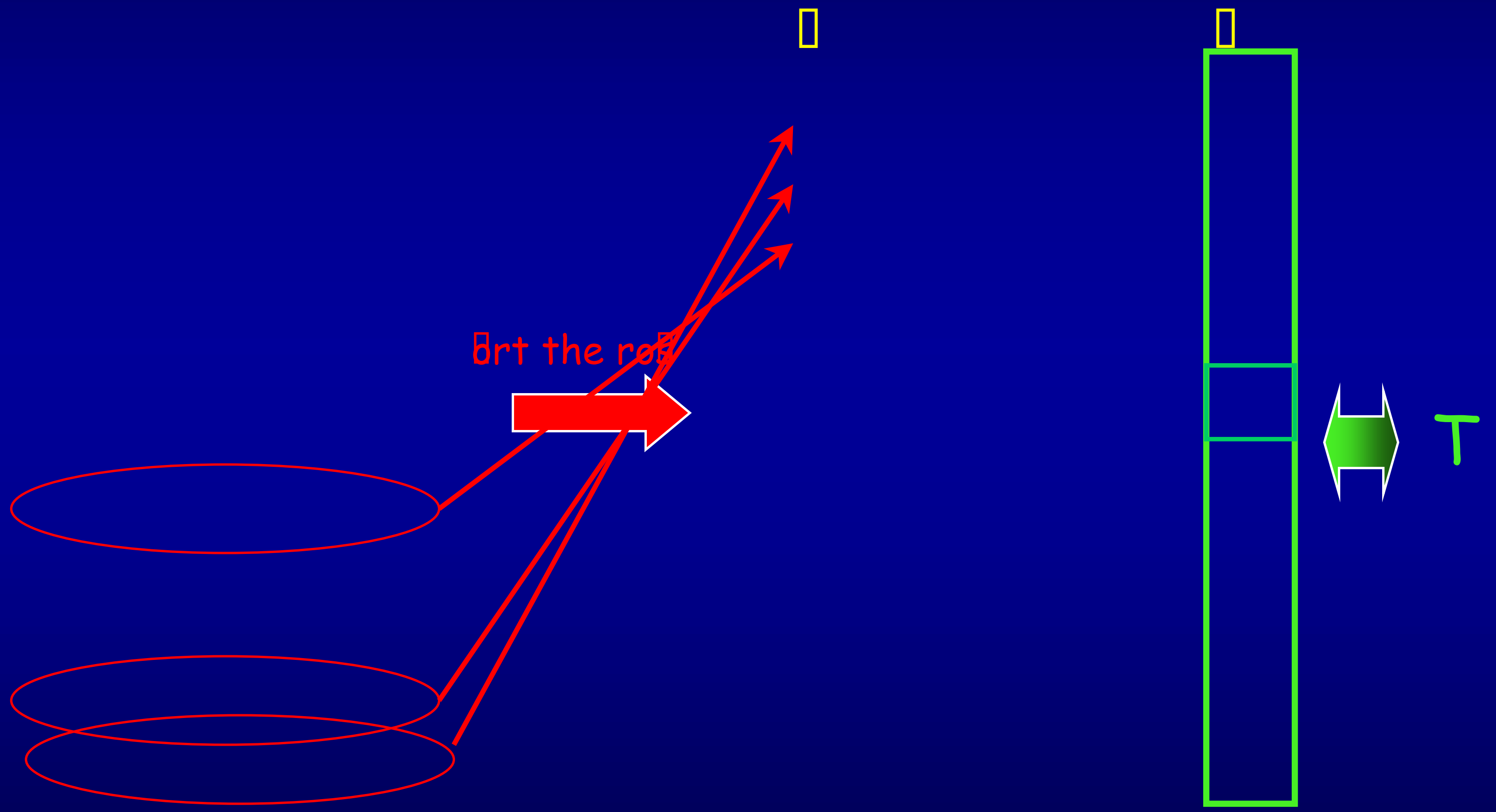
- ✓ is i[illegible] o[illegible] the i[illegible]t sour[illegible]e [illegible]ise o[illegible]T
- ✓ i[illegible]tey[illegible] sho[illegible] that [illegible]rrays are [illegible]essib[illegible]

[illegible]ati[illegible]the [illegible]is [illegible] a[illegible]

- ✓ [illegible] [illegible]ose to the best [illegible]essors
- ✓ [illegible]uery ti[illegible] o[illegible]ses o[illegible]hu[illegible]es o[illegible]s o[illegible]te[illegible]

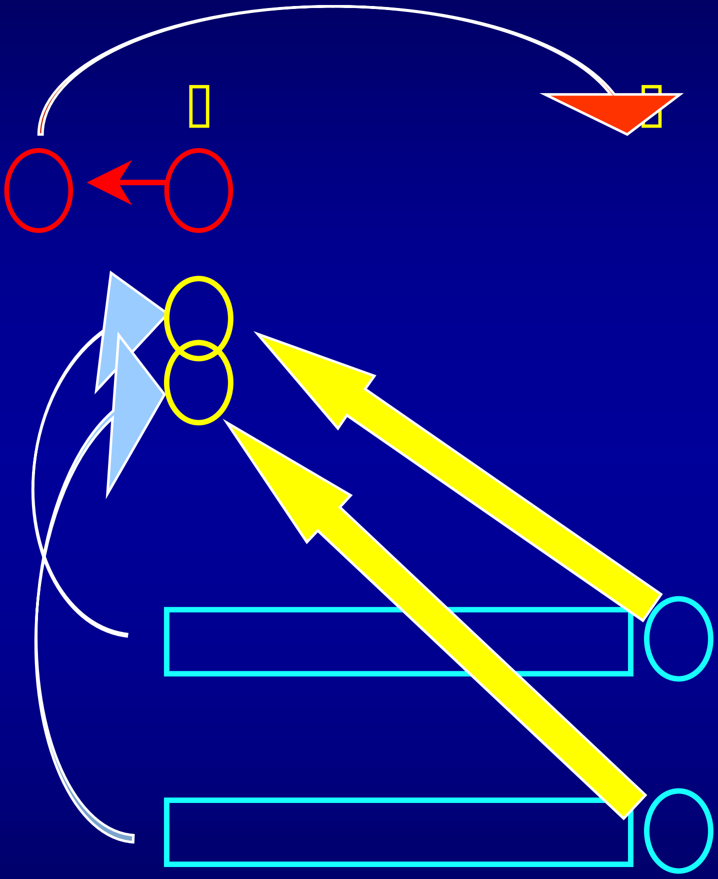
ddrd

et us ila te T issisii



ery u is a erutatio Thee aso a

dlnverid e



[illegible] s bars bee [illegible] s iT

eostrut bar T

e sto at i soo

o e a s oto s bars

ee to stilish equa ars i

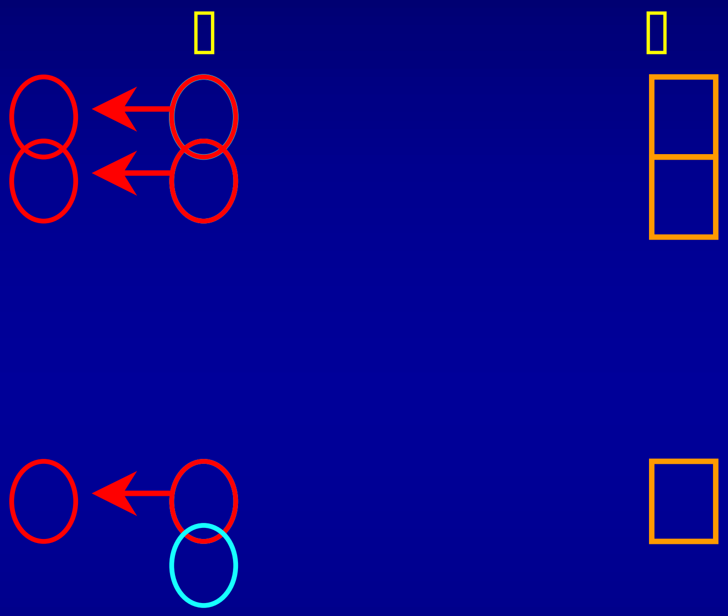
Tae to equa s bars

otate their ros

ae reatie orer

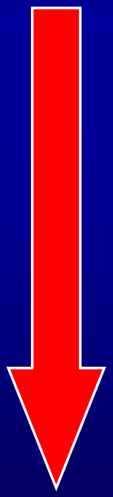
eaith is the ith i

[illegible]dnverid[illegible] e dnd d



T[illegible] [illegible]o[illegible]erties:

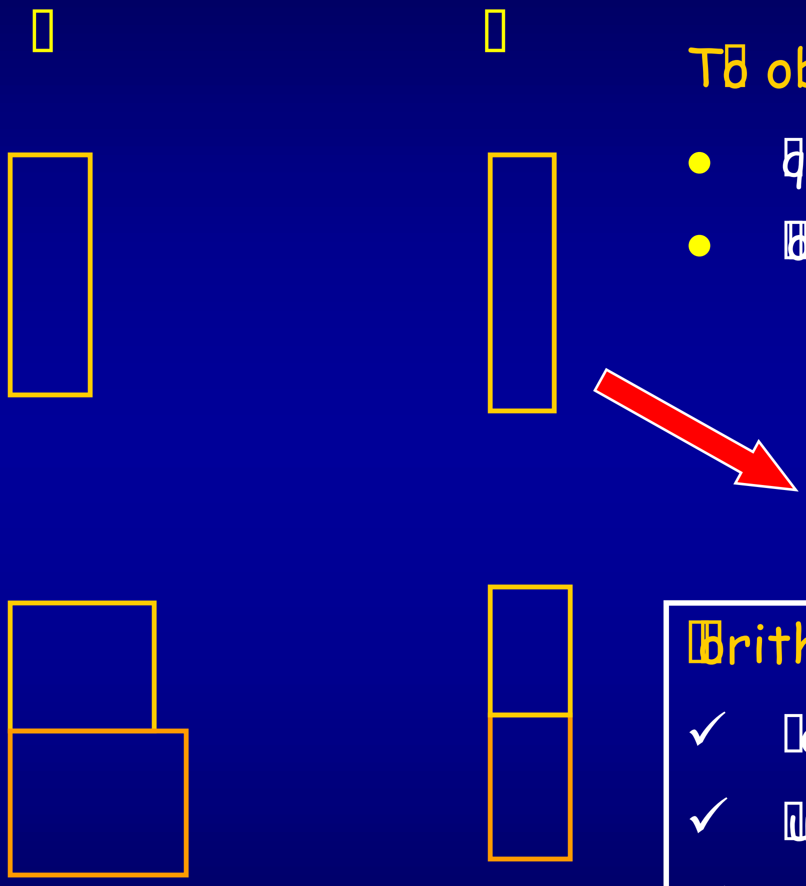
- [illegible]s [illegible]ars [illegible]e[illegible] [illegible] s i[illegible]T
- i[illegible]h [illegible] [illegible] i[illegible]h [illegible]



[illegible]estru[illegible] T
[illegible]tie

T[illegible] [illegible]

dcdd dcdeid e



T□ obser□tio□:

- □qua□substr □ei□□□□ ros□
- □se □ars are s□□r□

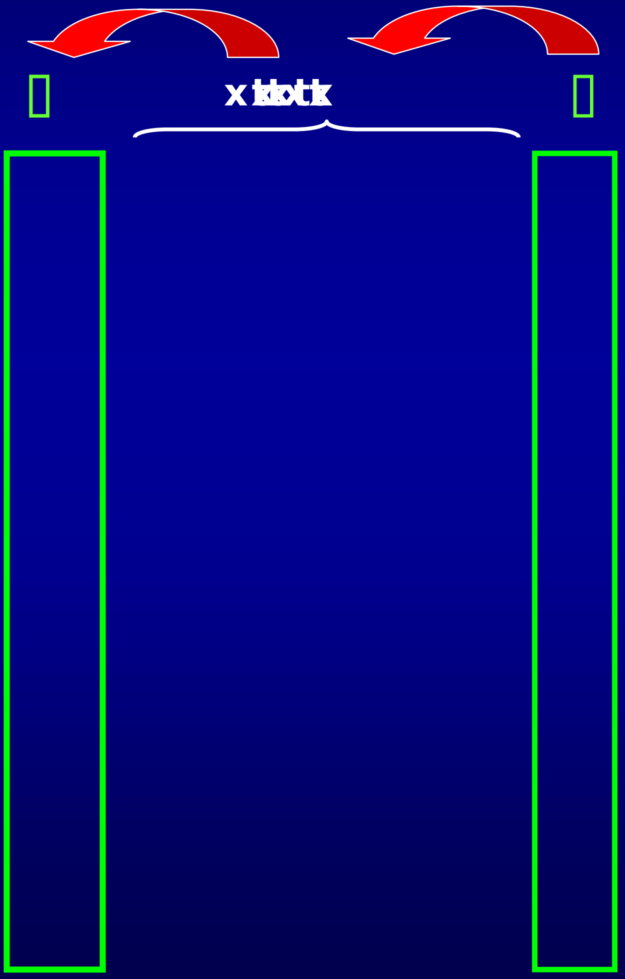
□□□ty □ □

□□rith□□:

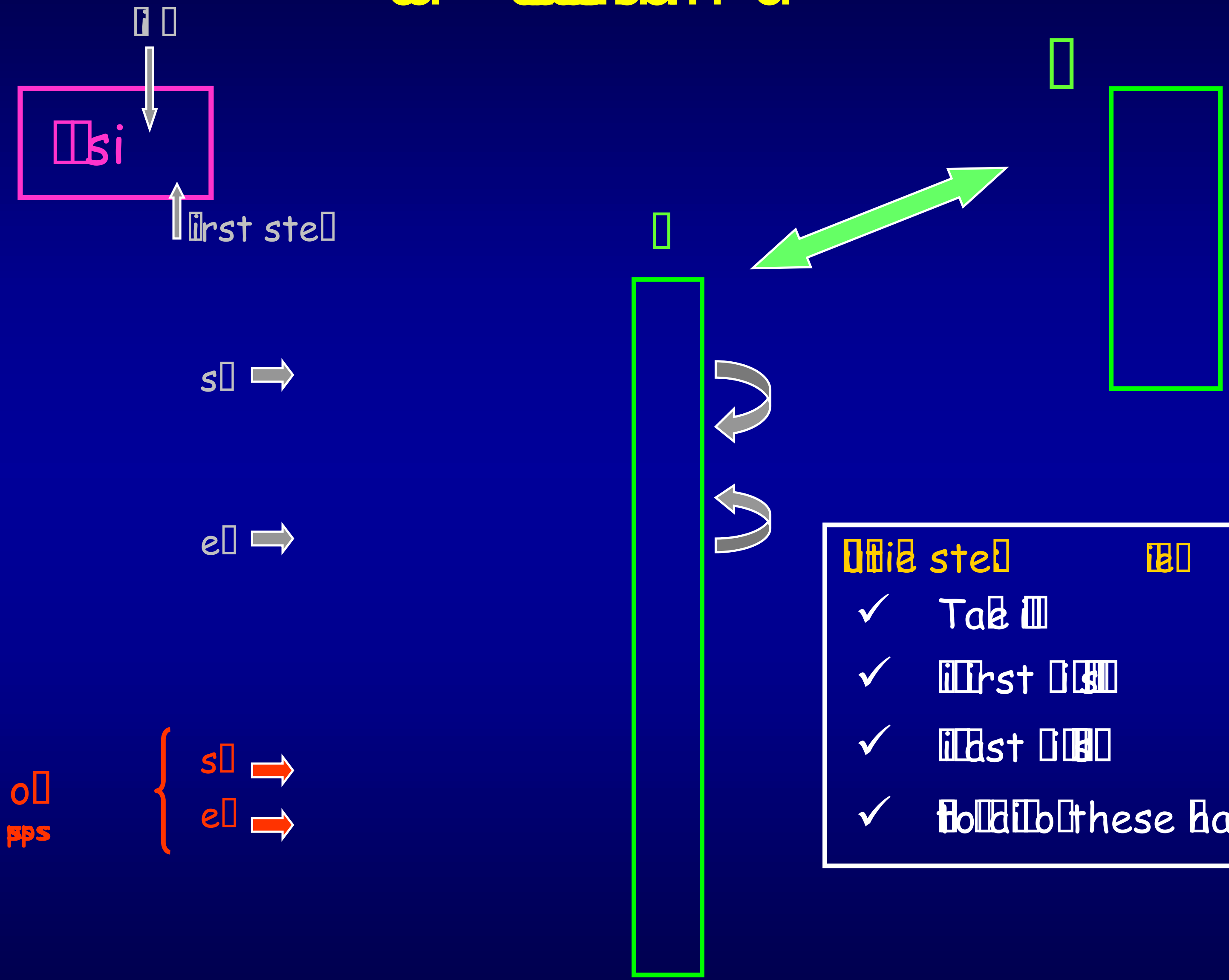
- ✓ □□et□□□□□ □□□□□ ✓ □
- ✓ □□□th □□□□ ✓ □ □
- ✓ □tatistic□□□r o□□ □ □rith□ti□

☑ □□□□esses □□ better tha□□ut it s□□r i□□□ressio□ □

idArrodlldrd



d dddddn d



ltie ste ie se or

- ✓ Tae i
- ✓ irst is
- ✓ ast is
- ✓ tolib these hars

dd d ordendd

□□si

□

T □issississi□□

$\underbrace{\text{s s s s}}_{\text{sa□□ste□is □}}$

□theory□ set to □□□ $^{\varepsilon}$ to
ba□□ s□□ a□□sti□ti□

□or this□ □e□to □ ba□r□

$\text{□ississi}\underbrace{\text{s}}_{\text{s}}\underbrace{\text{si□□}}_{\text{e□ s ro□}}$

□ro□ s□s □ositio□ □t □ □ □ □□□ □

This o□rre□ is □ste□□□□tey □

s□

e□

[illegible]

[illegible]e [illegible] too[illegible]:

- ✓ Ti[illegible] i[illegible] su[illegible]rts [illegible]st the [illegible] the o[illegible]rre[illegible]s
- ✓ [illegible]t i[illegible] su[illegible]rts both [illegible] a[illegible]te

both o[illegible] the [illegible] a [illegible]esse[illegible] o[illegible] the te[illegible]

uo[illegible]ction u u neu s uuu	[illegible]	[illegible]
[illegible] i[illegible]	[illegible]	[illegible]s [illegible]
[illegible] i[illegible]	[illegible]	[illegible]s [illegible] o[illegible]rre[illegible]
[illegible] o[illegible] [illegible] [illegible]s [illegible]	[illegible]	[illegible]s

☑ *[illegible] fingerprint* [illegible]xistentia[illegible] and co[illegible]ting [illegible] [illegible]aries fast

[illegible]

□hat about □r□ase □□urre□ o□□□

- □ear□ □or □as a substri□ o□ T usi□ the □□□e□
- □or e□ery □□□te o□urre□ □e□□t a □r□ase □e□

□r□ □e□□ substri□ su□□

T □□□□□□□□□□□□□

□i□

□he □st□□ssi□ □ase □□be □ry □sty□

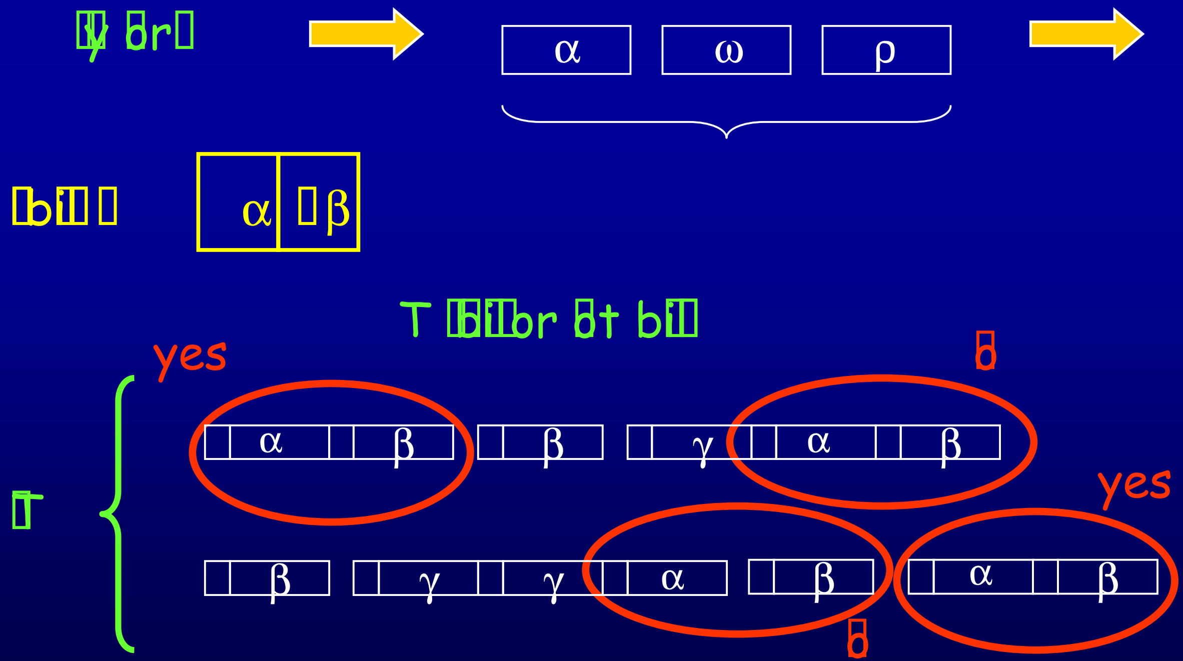
The □□□e□□be a□□te□to be □r□ase□

- □e□□ess T to □r□a □□ste□te□ □T
- □ui□□□□□e□o□er □T
- Tra□□r□ a□y □r□ase □urre□ o□T i□to a substri□ o□urre□ o□T□□a□so□e it usi□ the □□□e□buil□ o□T

[illegible]

ariat o uc orith

- ✓ ybos o the huc tree are the *ors* o T
- ✓ The uc tree has out
- ✓ ocors are *bytealie* a *tae*



le

tiocry o ors

uc tree

le buit o T

passss

sd asssss

[illegible]

- [illegible]hiee [illegible]otie i[illegible]ourree retriea[illegible]
  - ✓ [illegible] [illegible]T[illegible] ~~o[illegible]~~ ε [illegible]o[illegible]bits [illegible]erraia[illegible]ai[illegible]
- [illegible]hiee [illegible]o[illegible]s i[illegible]ourree retriea[illegible]
  - ✓ [illegible]o[illegible]oresse[illegible]iees eror[illegible]rao[illegible]aesses
- [illegible]ast ostu[illegible]io[illegible]aorith[illegible] or ui[illegible]rrays
  - ✓ [illegible]i[illegible]oresio[illegible]or [illegible]e[illegible]ostru[illegible]io[illegible]
  - ✓ [illegible]ui[illegible]Tree ostru[illegible]io[illegible]
  - ✓ [illegible]usteri[illegible]o[illegible]ouets [illegible]
- [illegible]e[illegible] the [illegible]oo[illegible] [illegible]e[illegible]ie
  - ✓ This i[illegible]oes theoreti[illegible] the [illegible]rte[illegible]ists

□e end

□y □e□ years □e □l□ be ab□e to store e□erythi□

□□ay□□□

□ato i□ *□hae□us* □su□este□ that □riti□ □ou□□ate

□or□et□□ess i□ the □i□□ o□ those □ho □ear□ to use it□ a□ the sho□ o□ □is□o□ □ithout the rea□ity□

□ho□e that this i□ □ot ou□ □a□i□□ □