PostgreSQL Scientific Application - Case example

PostgreSQL Genomic Databases

Sébastien Clément

sclement@cfrl.forestry.ca

Natural Resources Canada

Presented at the PostgreSQL Conference 2009 in Japan

November 20th

Arborea

Natural Resources Canada
Canadian Forest Service

Ressources naturelles Canada
Service canadien des forêts

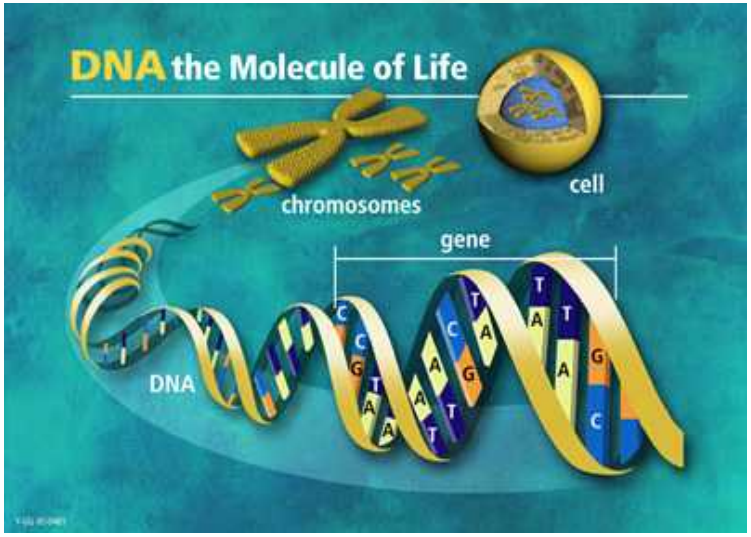
Foreword

« What is this guy doing here ? »

« Can PostgreSQL handle scientific databases ? »

What is genomics and why bother ?

Genomics: *« The study of the entire genome (all genes) of a species »*



3 000 000 000
~23 000

Genome size
Number of genes

390 000 000
~53 000

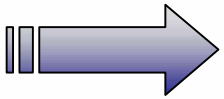
•Health and disease
•Heredity
•etc.

•Genetic improvement

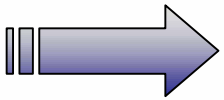
Why study *TREE* genomics?

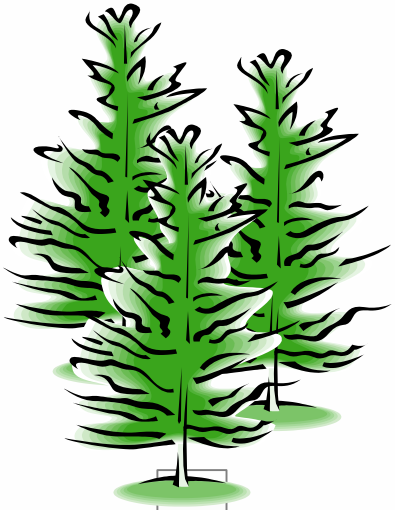
Cellulose gene

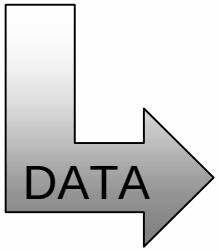
Normal

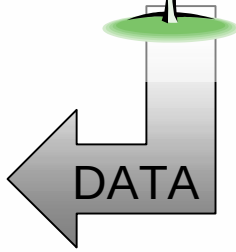


Variant









Why is a genomic DB essential?

A single gene…

Species

Name

Sequence

Variations

Size

Chromosome pos.

Functions

Cell wall metabolism
Cell structrure
Catalysis
…

Interaction with other genes

Similarity with other species

more… (phew!)

…how about thousands of genes…
…for thousands of species?

Public genomics DBs

Genbank

http://www.ncbi.nlm.nih.gov/Genbank/

UniProt

http://www.ebi.ac.uk/uniprot/

TAIR

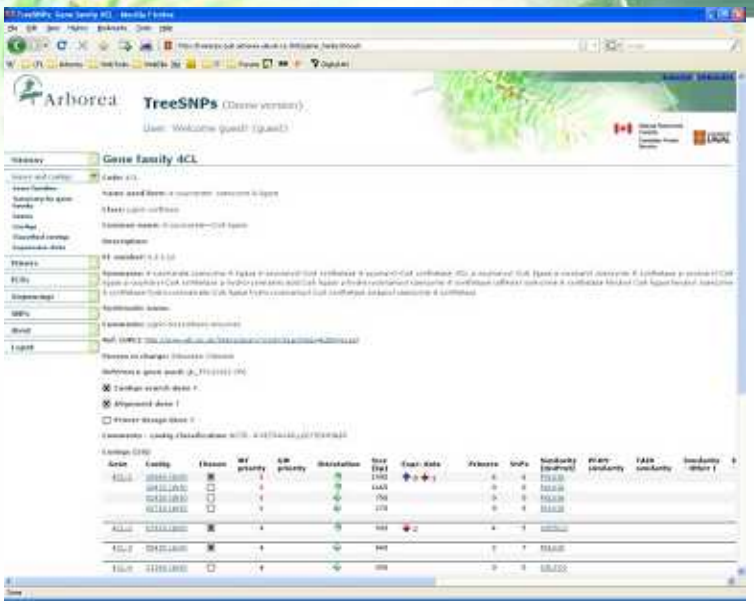
http://www.arabidopsis.org/

Our PostgreSQL Databases

TreeSNPs

Genes and variations

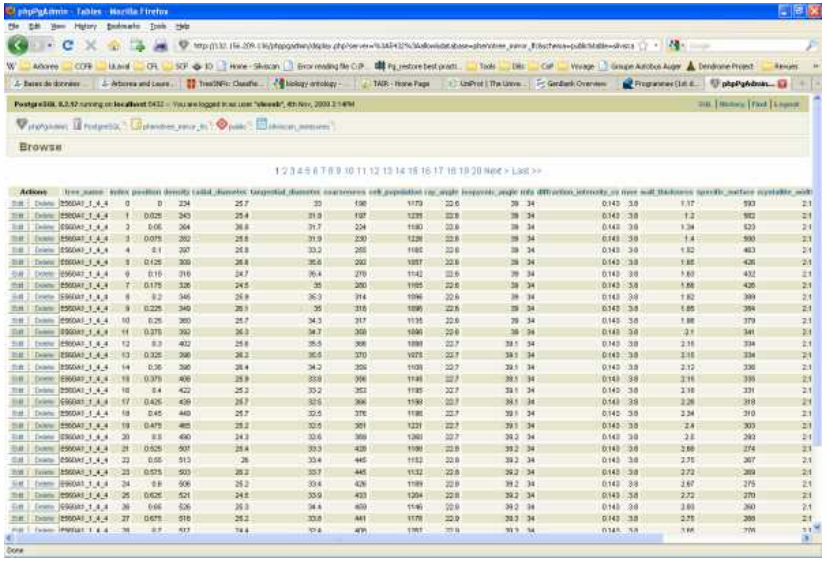
- Ruby on Rails interface
- Multi-language support
- 38 tables
- ~450 K records
- Mostly manual entry



PhenoTree

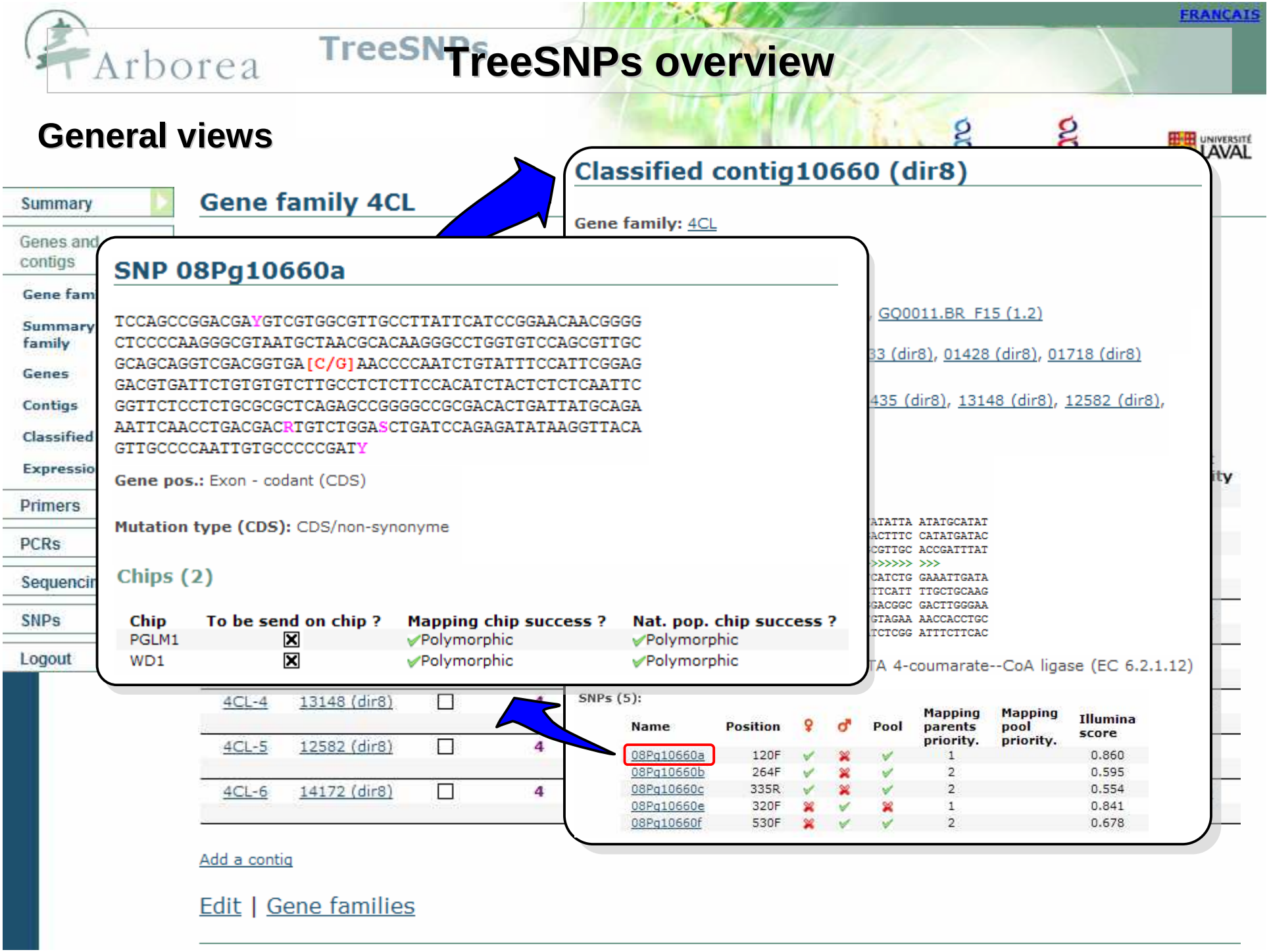
Observable attributes (physical, morphological)

- PhpPgAdmin interface
- 21 tables
- ~4.1 M records



TreeSNPs overview

General views



TreeSNPs overview (cont'd)

Lab plate

Results

PCR results for primer plates 008-F and 008-R

New PCR plate for primer plates 008-F and 008-R
Enter PCR results for PCR plates 008-F and 008-R

PCR-functional primer pairs: **57**
Paralogous primer pairs (i.e. haploid sequencing failed): **9**

008-F 008-R	01	02	03
A	8207-F 8207-R >= 620 bp TD-58 (TDA) : **OK** (800 bp)	8247-F 8247-R >= 607 bp TD-58 (TDA) : **Ø**	8347-F 8347-R >= 599 bp TD-58 (TDA) : **OK** (600
B	8984-F 8984-R >= 553 bp TD-58 (TDA) : **OK** (800 bp)	9016-F 9016-R >= 670 bp TD-58 (TDA) : **OK** (900 bp)	9039-F 9039-R >= 623 bp TD-58 (TDA) : **OK** (400
C	9572-F 9572-R >= 610 bp TD-58 (TDA) : **OK** (800 bp)	9629-F 9629-R >= 691 bp TD-58 (TDA) : **OK** (700 bp)	9673-F 9673-R >= 526 bp TD-58 (TDA) : **OK** (600
D	10143-F 10143-R	10175-F 10175-R	10211-F 10211-R

Plate view

Primer plates 008-F, 008-R

Name: 008-F, 008-R

Download in CSV (plate format)

008-F 008-R	01	02	03	04	05	06	07	08	09	10	11	12
A	8207-F 8207-R	8247-F 8247-R	8347-F 8347-R	8489-F 8489-R	8558-F 8558-R	8564-F 8564-R	8581-F 8581-R	8613-F 8613-R	8620-F 8620-R	8815-F 8815-R	8822-F 8822-R	8944-F 8944-R
B	8984-F 8984-R	9016-F 9016-R	9039-F 9039-R	9057-F 9057-R	9227-F 9227-R	9361-F 9361-R	9394-F 9394-R	9489-F 9489-R	9514-F 9514-R	9527-F 9527-R	9559-F 9559-R	9569-F 9569-R
C	9572-F 9572-R	9629-F 9629-R	9673-F 9673-R	9679-F 9679-R	9749-F 9749-R	9802-F 9802-R	9823-F 9823-R	9861-F 9861-R	9928-F 9928-R	9935-F 9935-R	9938-F 9938-R	9955-F 9955-R
D	10143-F 10143-R	10175-F 10175-R	10211-F 10211-R	10224-F 10224-R	10238-F 10238-R	10290-F 10290-R	10295-F 10295-R	10392-F 10392-R	10424-F 10424-R	10480-F 10480-R	10561-F 10561-R	10805-F1 10805-R1
E	9531-F1 9531-R1	10707-F1 10707-R1	10076-F1 10076-R1	7749-F1 7749-R1	10282-F1 10282-R1	10282-F2 10282-R2	10282-F3 10282-R3	5745-F1 5745-R1	5227-F1 5227-R1	5227-F2 5227-R2	10672-F1 10672-R1	10672-F2 10672-R2
F	10052-F1 10052-R1	10375-F2 10375-R2	9622-F1 9622-R1	9622-F2 9622-R2	15972-F1 15972-R1	6279-F1 6279-R1	8776-F1 8776-R1	8776-F2 8776-R2	7410-F1 7410-R1	5586-F1 5586-R1	1070-F2 1070-R2	8806-F1 8806-R1
G	6487-F2 6487-R2	1782-F2 1782-R2	11144-F2 11144-R2	11283-F2 11283-R2	4241-F3 4241-R3	3163-F1 3163-R1	15158-F1 15158-R1	10968-F1 10968-R1	10096-F1 10096-R1	3064-F1 3064-R1	6696-F1 6696-R1	10215-F1 10215-R1
H	9293-F1 9293-R1	15589-F1 15589-R1	12531-F1 12531-R1	5440-F1 5440-R1	6053-F1 6053-R1	11855-F1 11855-R1	12982-F1 12982-R1	10388-F1 10388-R1	10388-F2 10388-R2	10581-F1 10581-R1	10581-F2 10581-R2	10213-F1 10213-R1

Add/remove primers from plates 008-F, 008-R

Primer plates | PCR results for primer plates 008-F, 008-R

TreeSNPs overview (cont'd)

Example of calculations (views):

Summary by gene family

Gene families: **129**

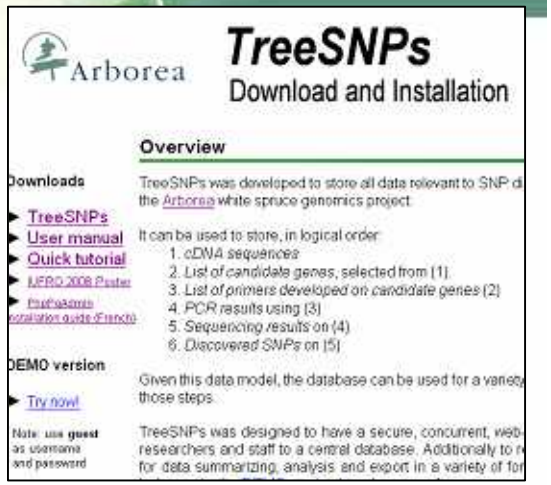
Notes:
000NC: other unidentified contigs, to be classified

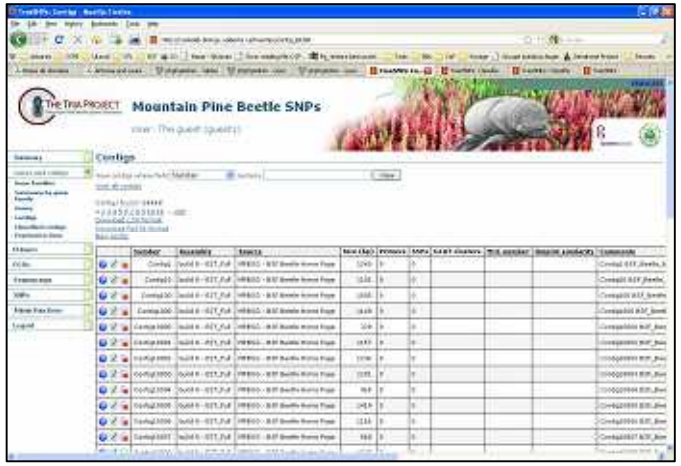
In red: Success rates below 50% (PCR, sequencing) or 30% (global success) .

Gene family	Contigs	Unique genes	Analyzed regions	Primers									SNPs		
				Total	Synth.	PCR			Sequencing			Global success	Total	Per unique gene	Per region
						Tested	OK	%	Tested	OK	%				
000NC	4374	0	2125	1198	1182	1141	717	63%	627	423	67%	42%	3043	-	1.4
4CL	18	15	6	12	12	12	12	100%	8	6	75%	75%	18	1.2	3.0
ACT	8	0	8	10	10	10	4	**40%**	2	2	100%	40%	5	-	0.6
AGO	7	0	7	12	12	12	10	83%	10	10	100%	83%	27	-	3.9
AGP	53	17	24	44	22	44	38	86%	34	20	59%	51%	75	4.4	3.1
AP	2	0	2	2	2	2	2	100%	2	2	100%	100%	8	-	4.0
AP2	58	22	44	63	39	63	57	90%	53	51	96%	87%	99	4.5	2.3
ARF	2	0	1	2	2	2	2	100%	2	2	100%	100%	5	-	5.0
ASF	1	1	1	2	2	2	2	100%	2	2	100%	100%	3	3.0	3.0
AUX-IAA	55	23	31	56	50	53	47	89%	45	41	91%	81%	121	5.3	3.9
BGAL	13	6	15	28	28	28	16	57%	14	10	71%	41%	47	7.8	3.1
bHLH	18	0	17	33	33	32	32	100%	26	24	92%	92%	70	-	4.1
bZIP	58	17	44	68	42	67	47	70%	38	36	95%	66%	79	4.6	1.8
BZR	1	0	1	2	2	2	2	100%	2	2	100%	100%	3	-	3.0

TreeSNPs overview (cont'd)

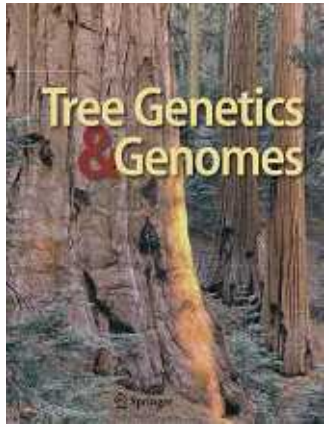
TreeSNPs download page & demo version
http://treesnps-pub.arborea.ulaval.ca:3000/download





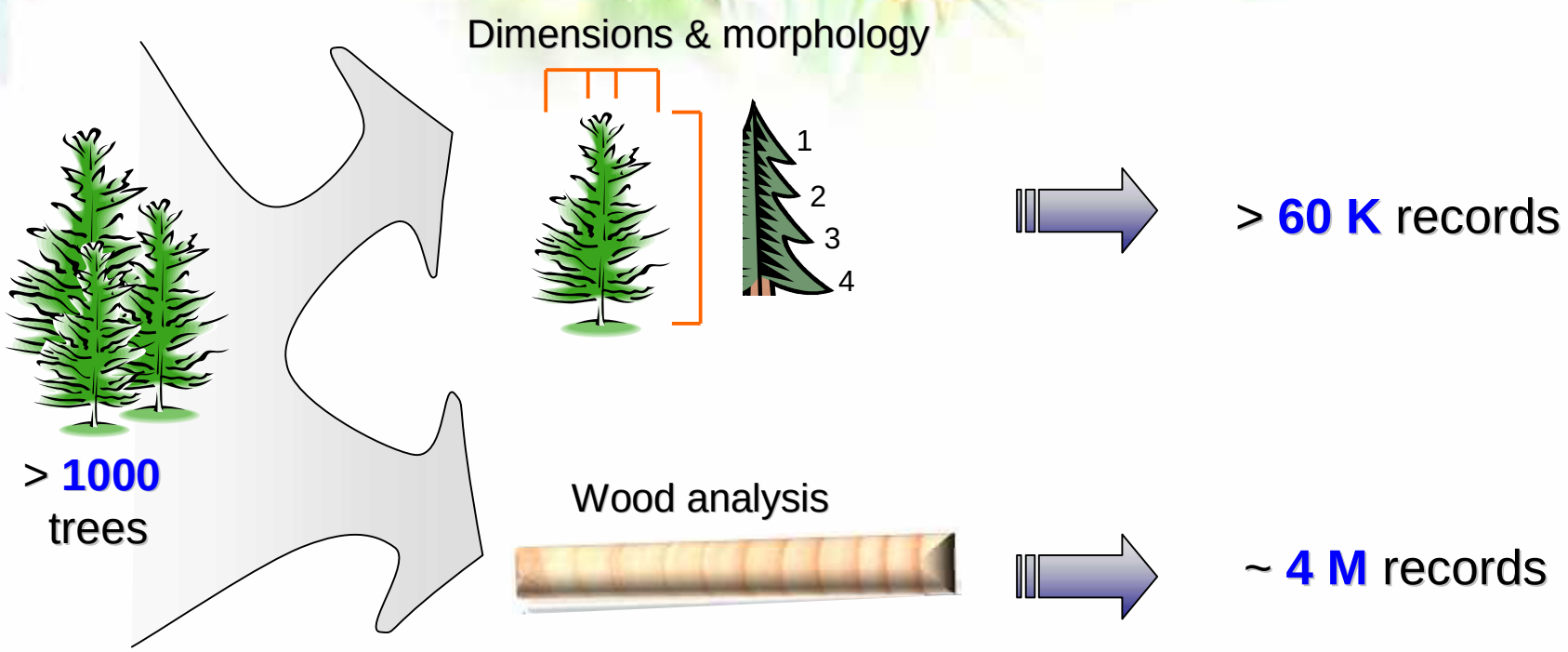
Adopted by U. of Alberta's (Canada) Laboratory on Moutain Pine Beetle

A paper to appear soon in



PhenoTree overview

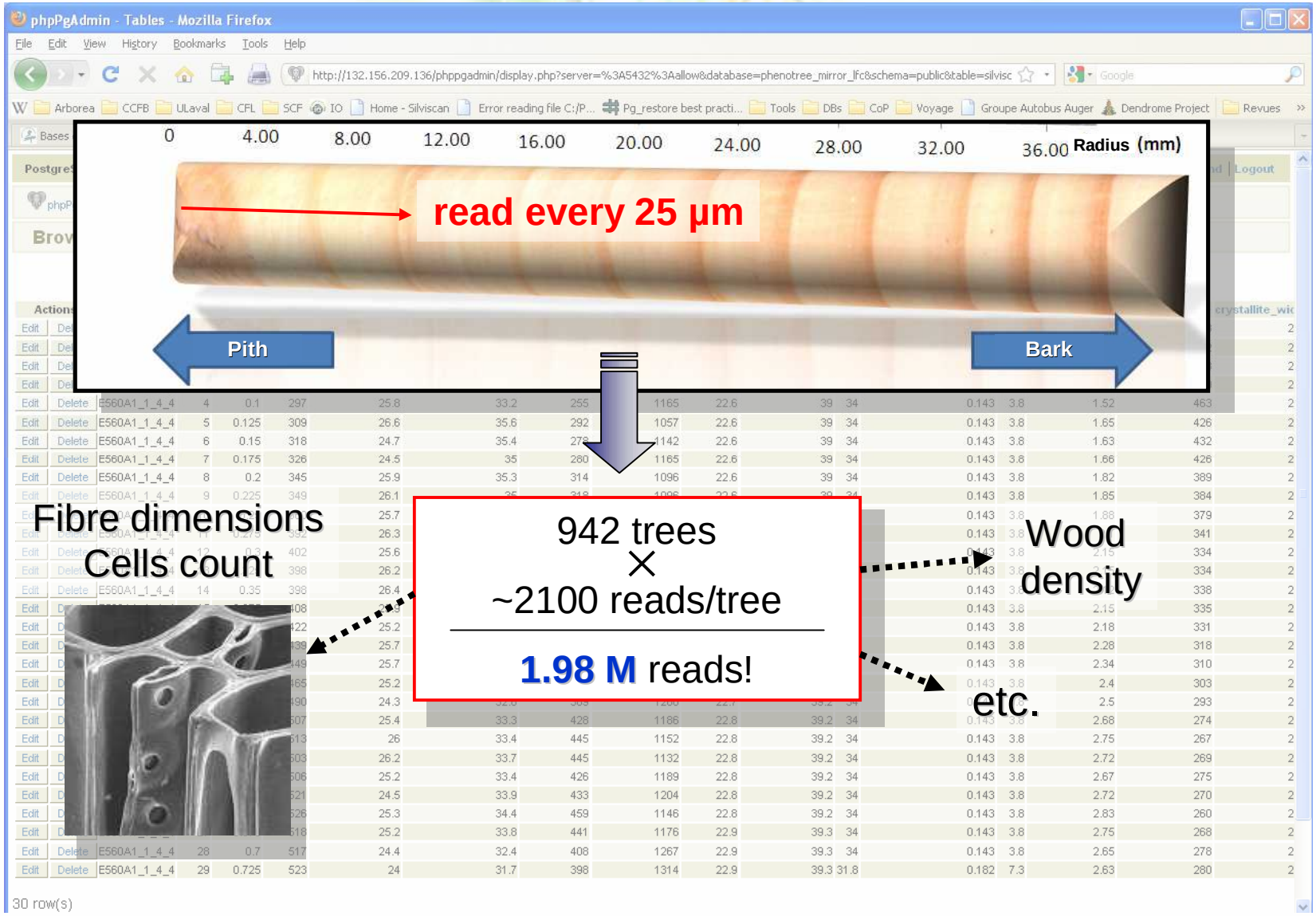
What data is stored ?



Other data:
- Geographical locations
- Tree pedigree

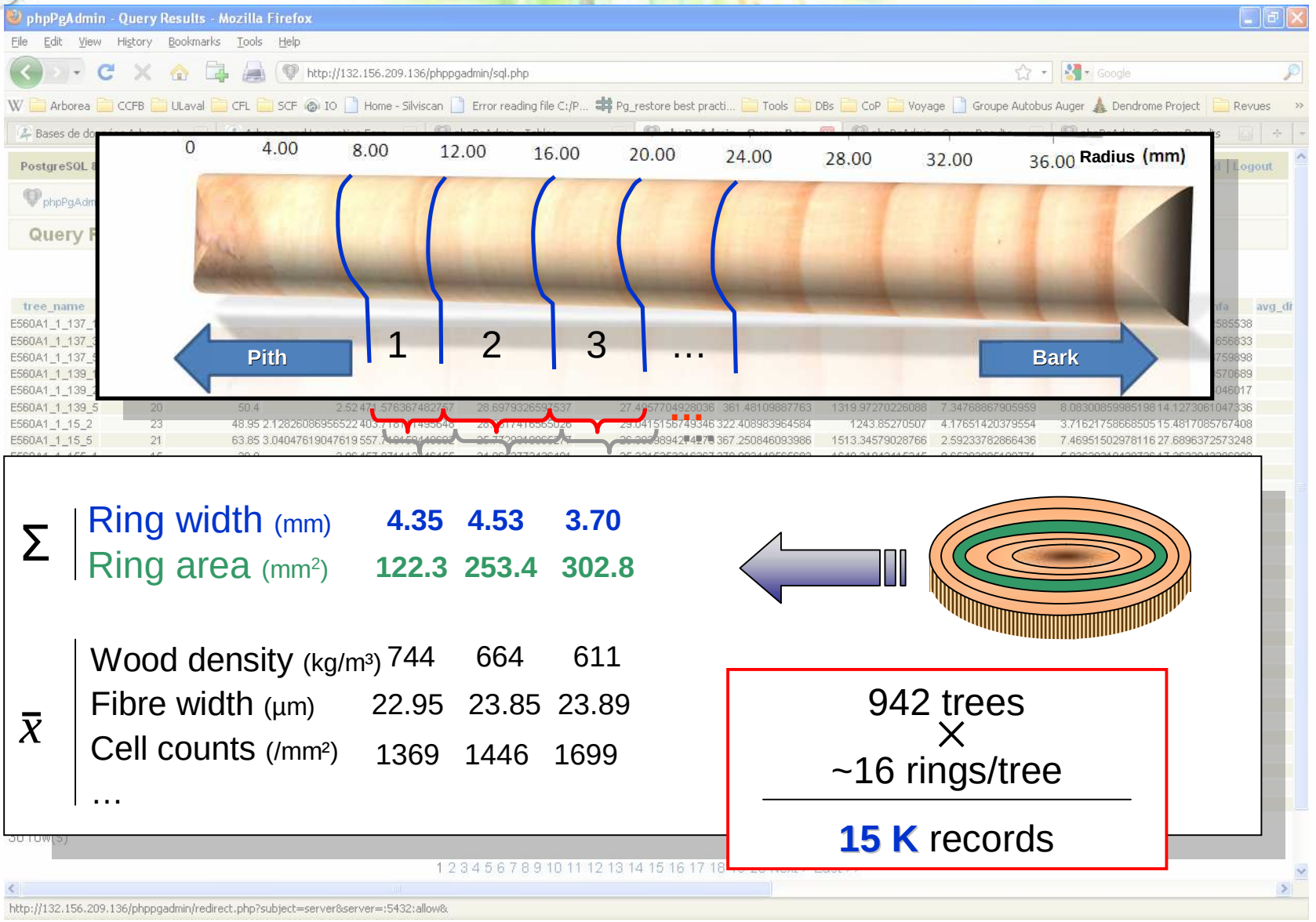
PhenoTree overview (cont'd)

Wood analysis properties table



PhenoTree overview (cont'd)

Example of calculations 1 (SQL views): Growth ring averages



Σ

Ring width (mm)	**4.35**	**4.53**	**3.70**
Ring area (mm^2)	**122.3**	**253.4**	**302.8**

$\bar{x}$

Wood density (kg/m^3)	744	664	611
Fibre width (µm)	22.95	23.85	23.89
Cell counts ($/mm^2$)	1369	1446	1699
…			

942 trees
×
~16 rings/tree

15 K records

PhenoTree overview (cont'd)

Example of calculations 2 (SQL views): **crosstab function**

tree_name	year	height
E560A1_1_105_2	1986	220
E560A1_1_105_2	1992	450
E560A1_1_105_2	1997	640
E560A1_1_105_2	2005	940
E560A1_1_106_3	1986	230
E560A1_1_106_3	1992	510
E560A1_1_106_3	1997	670
E560A1_1_106_3	2005	1070
E560A1_1_124_1	1986	160
E560A1_1_124_1	1992	410
E560A1_1_124_1	1997	620
E560A1_1_124_1	2005	840

Logical,
but not very useful…

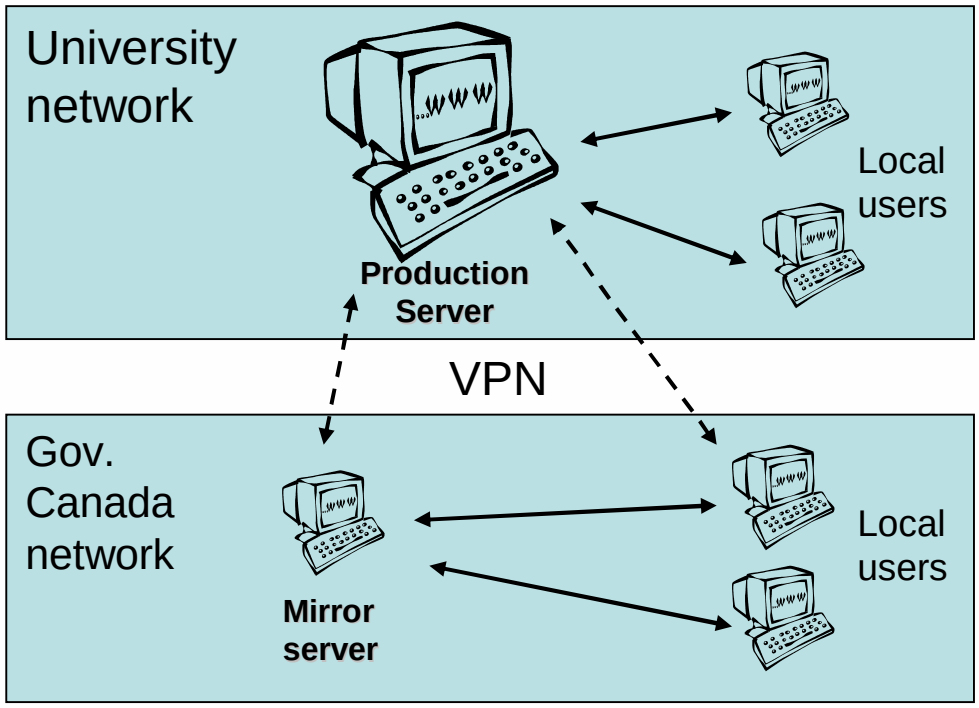
crosstab

tree_name	height_1986	height_1992	height_1997	height_2004	height_2005
E560A1_1_105_2	220	450	640	*NULL*	940
E560A1_1_106_3	230	510	670	*NULL*	1070
E560A1_1_124_1	160	410	620	*NULL*	840

…this is it what end-users want

Systems and user base

- Formerly Access projects (2006-7)
- Migrated to PostgreSQL 8.3 under Fedora (2007-8)
- Migrated back to Windows (2009)



- Around 20 scientific users (Universities, Federal Government)

PostgreSQL and Windows – can it really work ?

Task automation with DOS

- Limited functionnality

Solution ?

Windows Task Manager

↕

Cygwin

↕

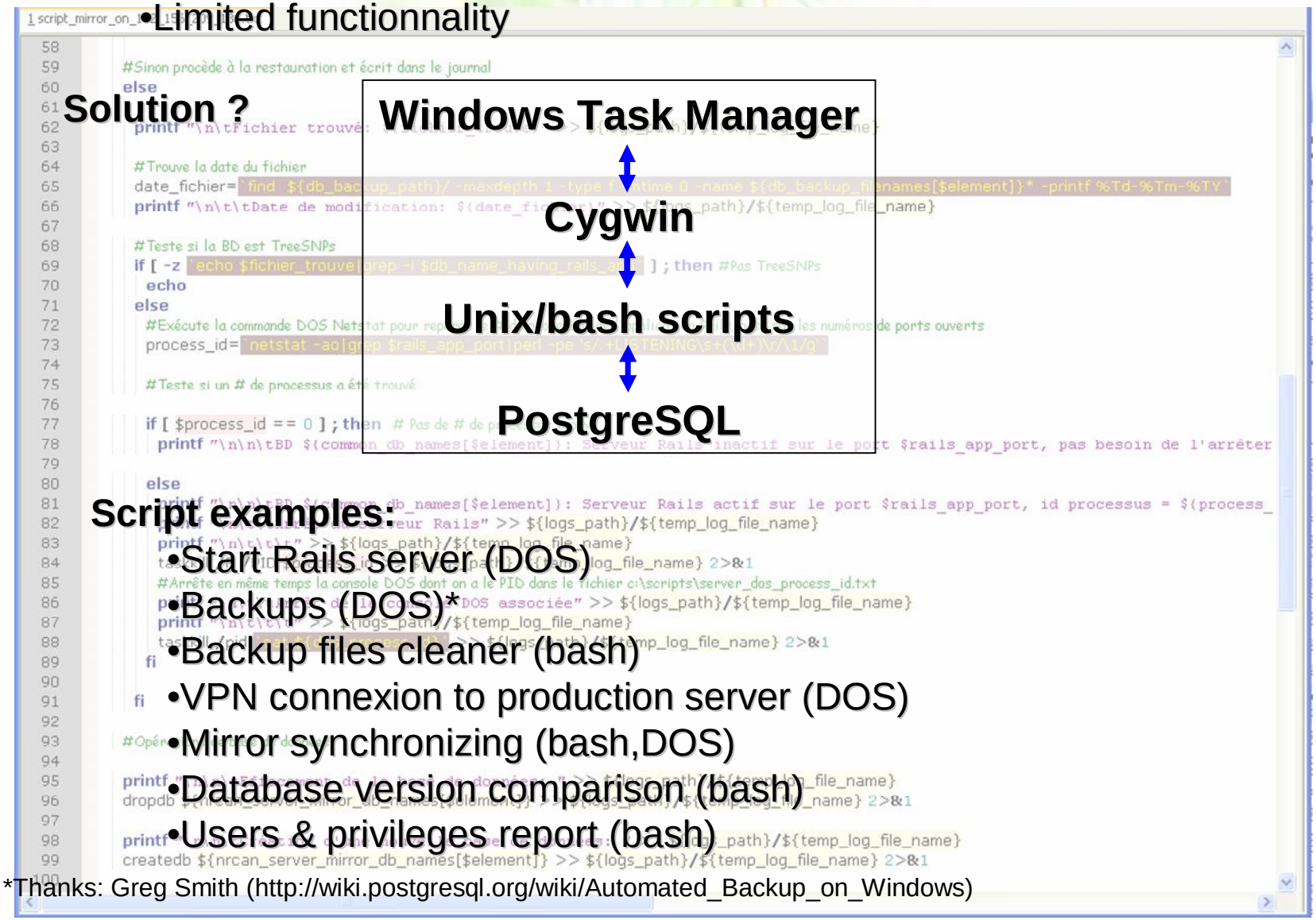
Unix/bash scripts

↕

PostgreSQL

Script examples:

- Start Rails server (DOS)
- Backups (DOS)*
- Backup files cleaner (bash)
- VPN connexion to production server (DOS)
- Mirror synchronizing (bash,DOS)
- Database version comparison (bash)
- Users & privileges report (bash)



*Thanks: Greg Smith (http://wiki.postgresql.org/wiki/Automated_Backup_on_Windows)

Developing databases for the scientific community

Suggestions:

- Have a user-based approach
  - 1. Know/answer the user's needs
  - 2. Limit technical jargon
  - 3. Think 'usability'

Aknowledgements

People

Jean Beaulieu – Lab director
Joël Fillon – Ruby on Rails interface designer
Jean-Philippe Dionne – Rails secure access programming
Jean Bousquet – Collaborator

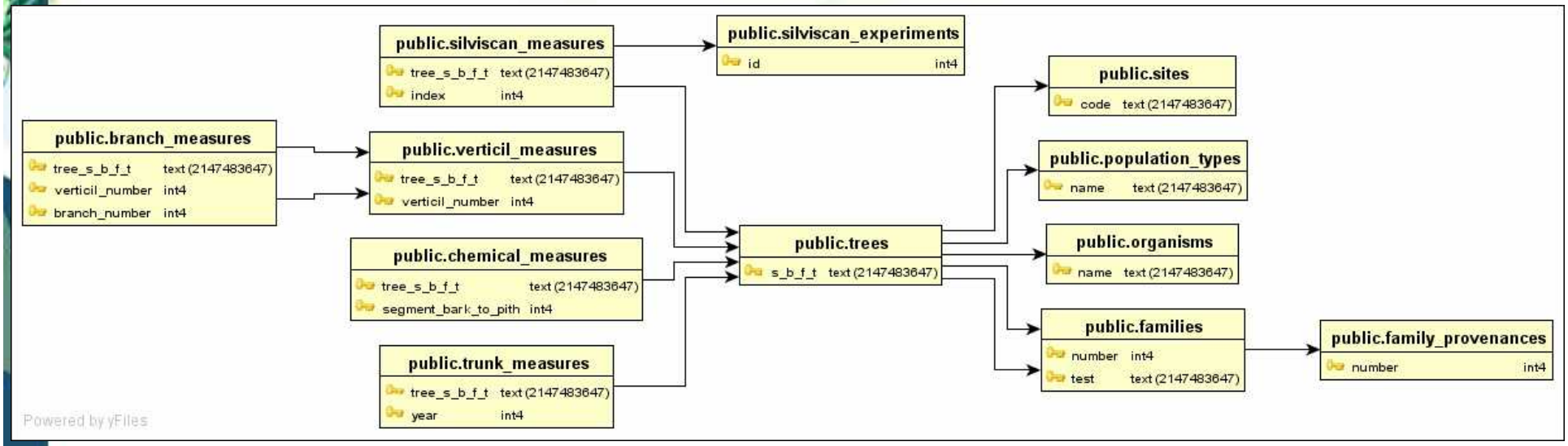
All end users, particularly:
Sylvie Blais, Stéphanie Beauseigle, Marie Deslauriers, Pier-Luc Poulin, Patrick Lenz

Organizations

Arborea Forest Genomics (http://www.arborea.ulaval.ca/)
Canadian Forest Service, Natural Resources Canada
Genome Québec
Genome Canada

Done ?

```
SELECT tmp_rings.tree_s_b_f_t, tmp_rings.experiment_id, tmp_rings.ring_pith_to_bark, bad_rings.ring_pith_to_bark IS NULL AS is_valid, tmp_rings.start_index,
tmp_rings.start_position, tmp_rings.end_index, tmp_rings.end_position, tmp_rings.width, tmp_rings.area, tmp_rings.avg_density, tmp_rings.avg_radial_diameter,
tmp_rings.avg_tangential_diameter, tmp_rings.avg_coarseness, tmp_rings.avg_cell_population, tmp_rings.avg_ray_angle, tmp_rings.avg_isopycnic_angle, tmp_rings.avg_mfa,
tmp_rings.avg_diffraction_intensity_cv, tmp_rings.avg_moe, tmp_rings.avg_wall_thickness, tmp_rings.avg_specific_surface, tmp_rings.avg_crystallite_width,
tmp_rings.avg_crystallinity, early_woods.start_index AS ew_start_index, early_woods.start_position AS ew_start_position, early_woods.end_index AS ew_end_index,
early_woods.end_position AS ew_end_position, early_woods.width AS ew_width, early_woods.width / tmp_rings.width AS ew_percentage, early_woods.area AS ew_area,
early_woods.avg_density AS ew_avg_density, early_woods.avg_radial_diameter AS ew_avg_radial_diameter, early_woods.avg_tangential_diameter AS
ew_avg_tangential_diameter, early_woods.avg_coarseness AS ew_avg_coarseness, early_woods.avg_cell_population AS ew_avg_cell_population, early_woods.avg_ray_angle
AS ew_avg_ray_angle, early_woods.avg_isopycnic_angle AS ew_avg_isopycnic_angle, early_woods.avg_mfa AS ew_avg_mfa, early_woods.avg_diffraction_intensity_cv AS
ew_avg_diffraction_intensity_cv, early_woods.avg_moe AS ew_avg_moe, early_woods.avg_wall_thickness AS ew_avg_wall_thickness, early_woods.avg_specific_surface AS
ew_avg_specific_surface, early_woods.avg_crystallite_width AS ew_avg_crystallite_width, early_woods.avg_crystallinity AS ew_avg_crystallinity, late_woods.start_index AS
lw_start_index, late_woods.start_position AS lw_start_position, late_woods.end_index AS lw_end_index, late_woods.end_position AS lw_end_position, late_woods.width AS
lw_width, late_woods.width / tmp_rings.width AS lw_percentage, late_woods.area AS lw_area, late_woods.avg_density AS lw_avg_density, late_woods.avg_radial_diameter
AS lw_avg_radial_diameter, late_woods.avg_tangential_diameter AS lw_avg_tangential_diameter, late_woods.avg_coarseness AS lw_avg_coarseness,
late_woods.avg_cell_population AS lw_avg_cell_population, late_woods.avg_ray_angle AS lw_avg_ray_angle, late_woods.avg_isopycnic_angle AS lw_avg_isopycnic_angle,
late_woods.avg_mfa AS lw_avg_mfa, late_woods.avg_diffraction_intensity_cv AS lw_avg_diffraction_intensity_cv, late_woods.avg_moe AS lw_avg_moe,
late_woods.avg_wall_thickness AS lw_avg_wall_thickness, late_woods.avg_specific_surface AS lw_avg_specific_surface, late_woods.avg_crystallite_width AS
lw_avg_crystallite_width, late_woods.avg_crystallinity AS lw_avg_crystallinity
  FROM tmp_rings
NATURAL LEFT JOIN bad_rings
  JOIN tmp_woods early_woods USING (tree_s_b_f_t, experiment_id, ring_pith_to_bark)
  JOIN tmp_woods late_woods USING (tree_s_b_f_t, experiment_id, ring_pith_to_bark)
 WHERE early_woods.transition = 1 AND late_woods.transition = 2
 ORDER BY tmp_rings.tree_s_b_f_t, tmp_rings.experiment_id, tmp_rings.ring_pith_to_bark;
```


public.silviscan_measures
tree_s_b_f_t text (2147483647)
index int4
public.silviscan_experiments
id int4
public.sites
code text (2147483647)
public.branch_measures
tree_s_b_f_t text (2147483647)
verticil_number int4
branch_number int4
public.verticil_measures
tree_s_b_f_t text (2147483647)
verticil_number int4
public.population_types
name text (2147483647)
public.chemical_measures
tree_s_b_f_t text (2147483647)
segment_bark_to_pith int4
public.trees
s_b_f_t text (2147483647)
public.organisms
name text (2147483647)
public.families
number int4
test text (2147483647)
public.family_provenances
number int4
public.trunk_measures
tree_s_b_f_t text (2147483647)
year int4
Powered by yFiles