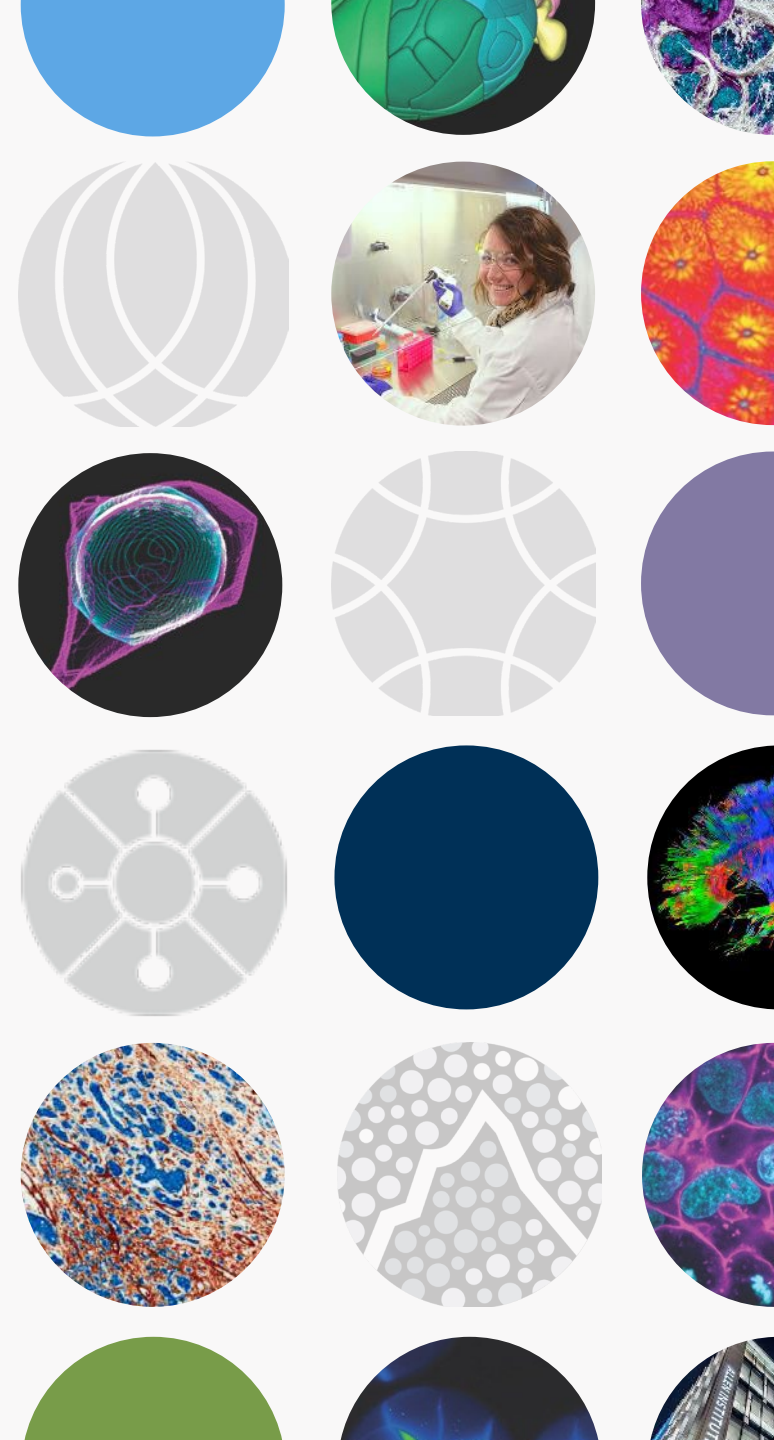




Ethical Course of Data Analytics

Balancing Openness, Trust, Security and Responsibility

Shoaib Mufti, Head of Data and Technology



Agenda

- Allen Institute Introduction and Focus
- Open Data and Artificial Intelligence challenges
- Strategies to overcome challenges



- Focused on Biological Sciences
- Established 2003 by Paul G. Allen
- South Lake Union, Seattle, WA

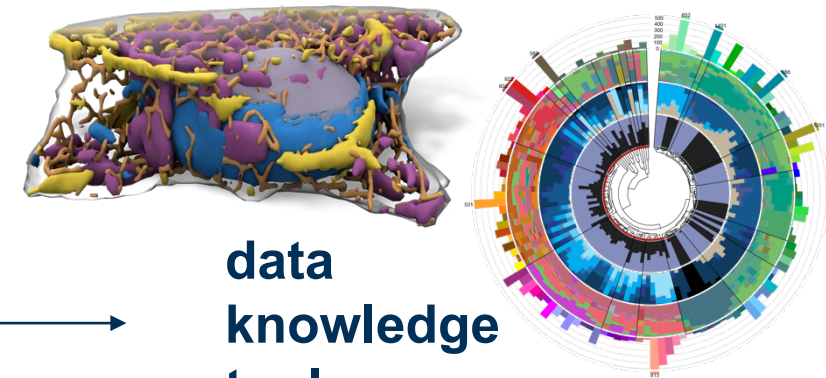
hard problems
complexity
foundational biology



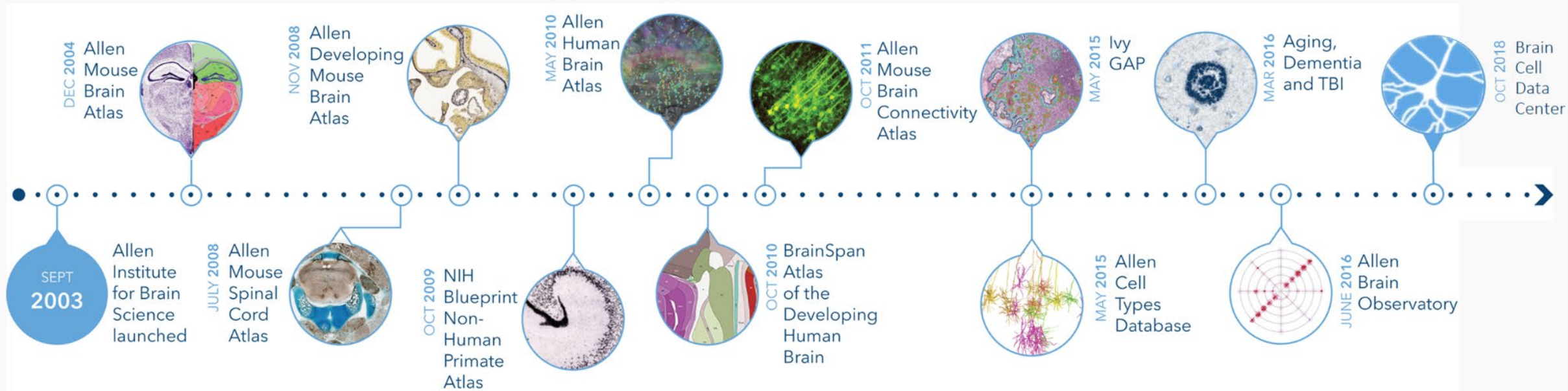
big science
team science
open science



data
knowledge
tools



History of creating impactful, web-based research tools

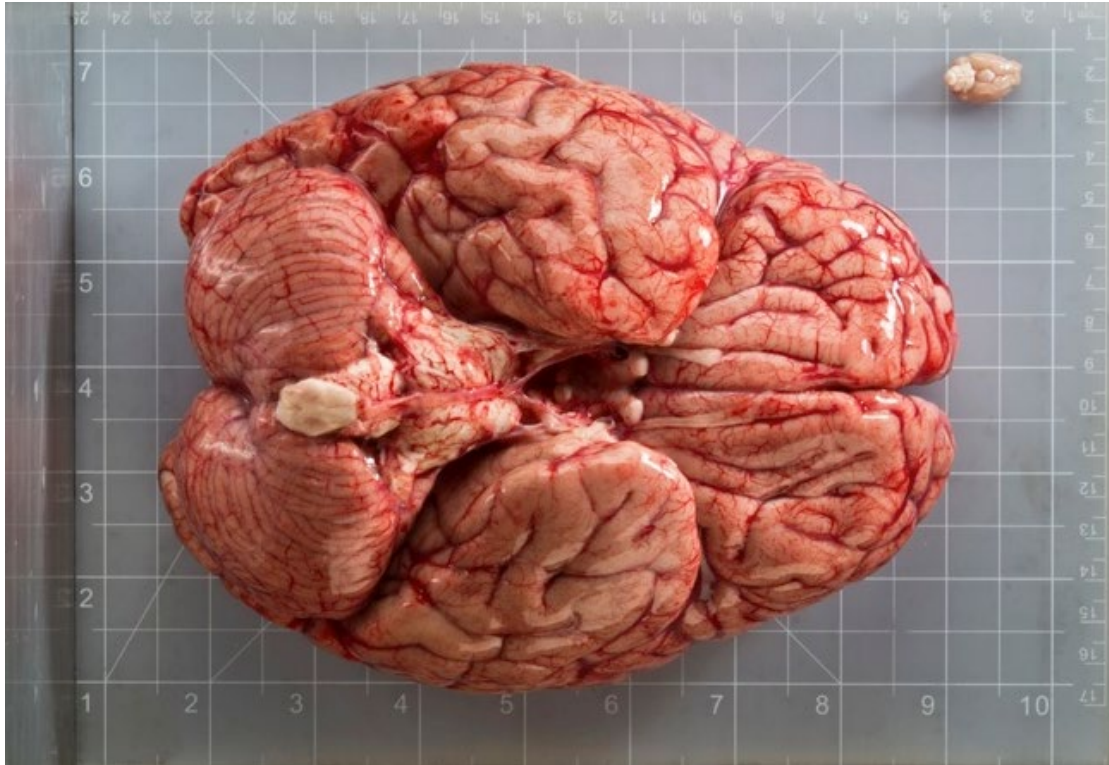


- This portfolio includes downloadable data repositories, software applications, reference standards and training toolkits – open to all

Impact

- The **Allen Brain Atlas** is one of the most well-recognized standards in neuroscience, known for quality, data ease-of-use, and comprehensiveness
 - Over 50 million page views per year
 - Over 500k unique users globally
 - Tools and research are cited in thousands of research publications
 - Data used by thousands of educators

Why aren't we curing brain diseases?



Am J Transl Res 2014;6(2):114-118
www.ajtr.org /ISSN:1943-8141/AJTR1312010

Review Article

Lost in translation: animal models and clinical trials in cancer treatment

Mental and neurological disorders and diseases cost the U.S. economy more than \$1.5 trillion per year, affecting over 1 billion people in the world.

Challenges:

We don't understand this extraordinarily complex system well enough, with the right level of granularity

Model organisms may not be similar enough to understand human

We may not be generating the right information about disease or have the tools to deliver therapies

We are awash in details, but the information is disjoint and not centralized

Different parts of the field cannot effectively communicate and integrate their valuable but incomplete data

We lack a coherent framework that integrates across disciplines

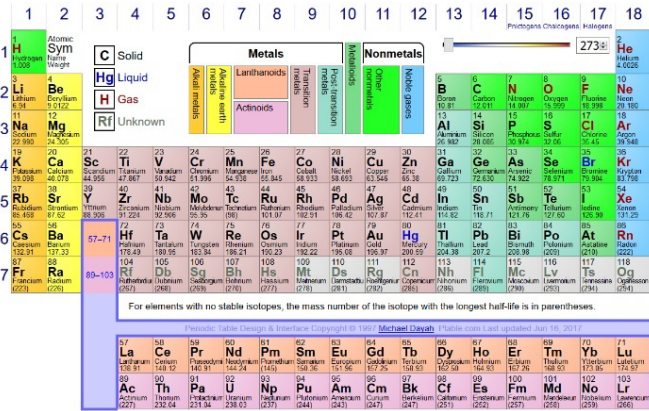
A unique transformative moment in the field

- New technologies have matured to create a whole brain cellular resolution map. Use this map to understand brain functions and diseases
- **Scalability:** The first complete map of a mammalian brain has been produced in mice.
- **Application to human:** Extensive proof of concept, funding now obtained to create a human map
- **Connection to the clinic:** Human map linking to non-invasive imaging done in the clinic, methods likely to be next-generation neuropathology
- **Application to disease:** First major projects demonstrating strong application to Alzheimer's
- **Link to other organs:** The same methods work across all organs and are being used broadly

The Allen Institute is leading this field transformation, with large-scale funding secured to create these atlases and apply them to disease

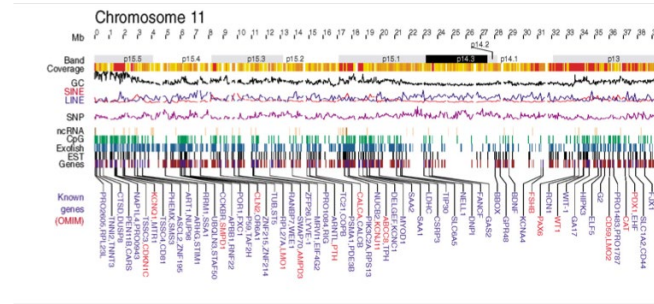
Neuroscience needs a unifying reference framework

Chemistry



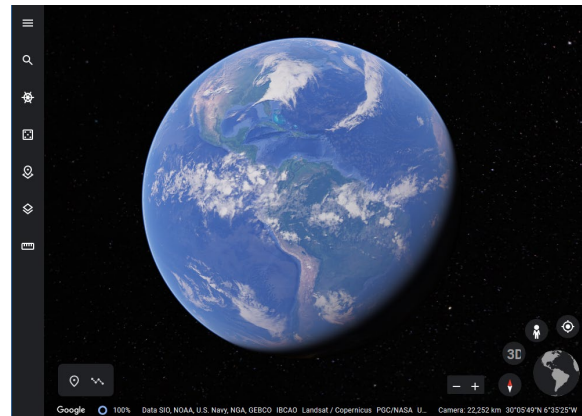
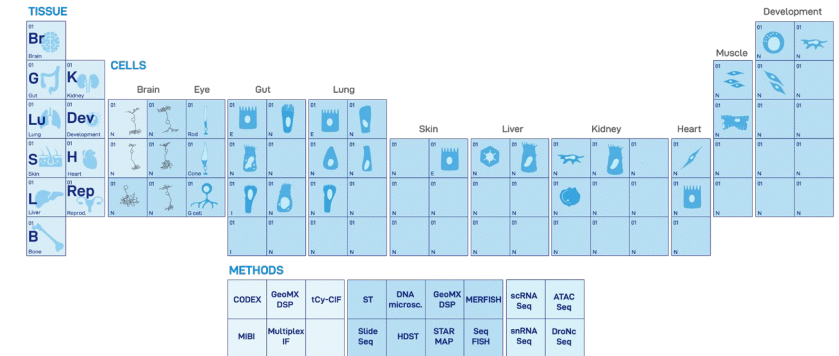
Mendeleev, 1869-1900

Genomics

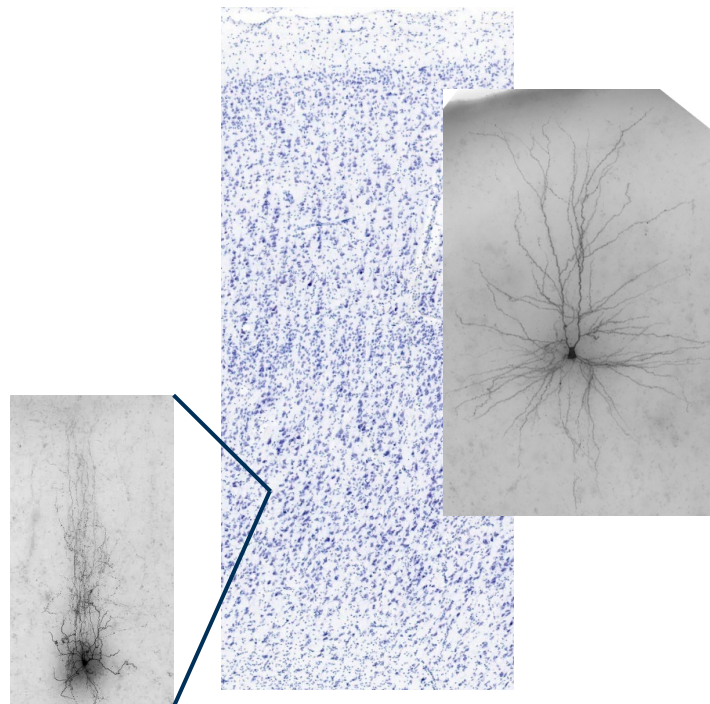
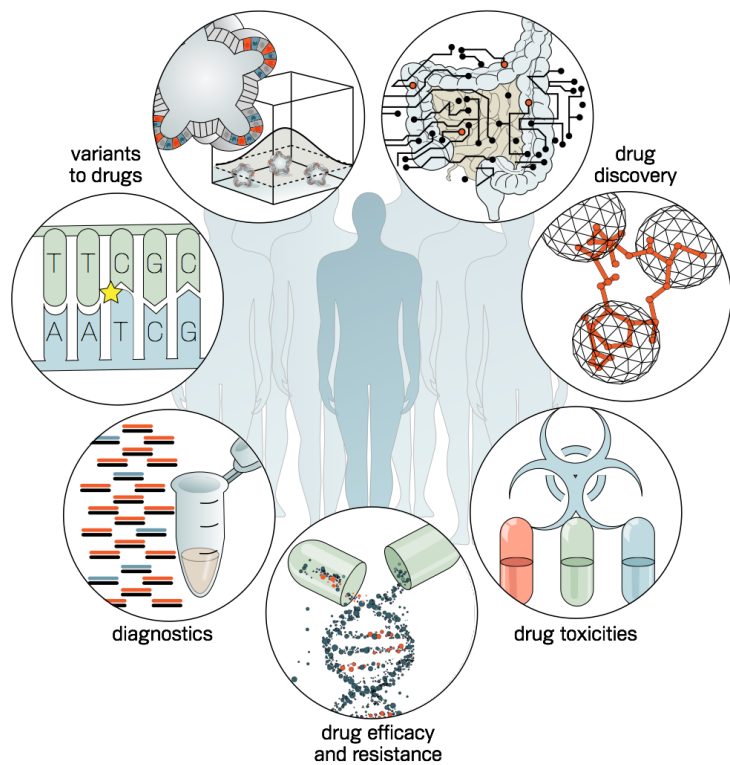


Human Genome Project, 1990-2003

Biology and Medicine?

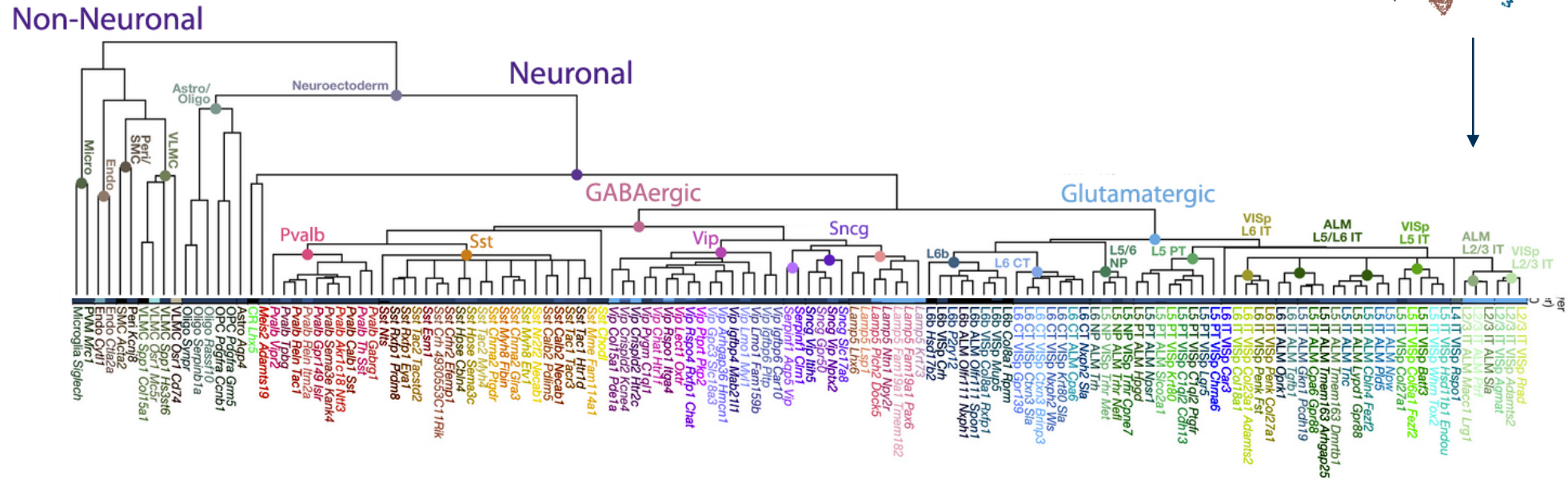
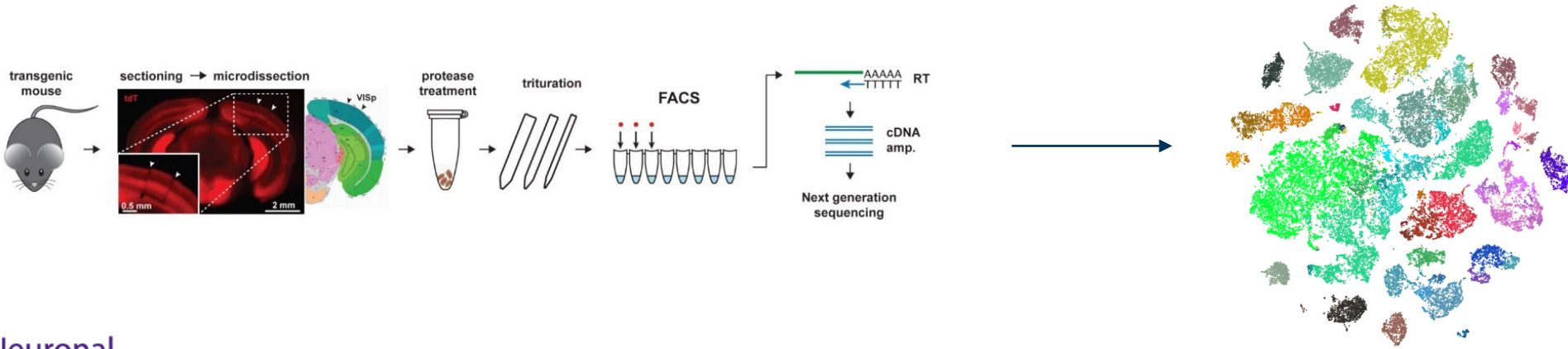


Cells are the basic units of the brain and all other organs, and are key to understanding functional organization and disease mechanisms



Human Cell Atlas Consortium White Paper, Section 1

Single cell genomics are radically disruptive technologies that have provided the means to define and characterize brain cell complexity



Tasic, Yao, Smith, Grayback...Koch, Zeng (2018) Nature

The next “Human Genome Project:” Mapping all the cell types of the brain and body

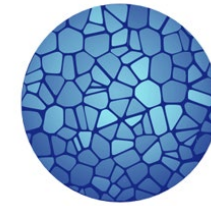
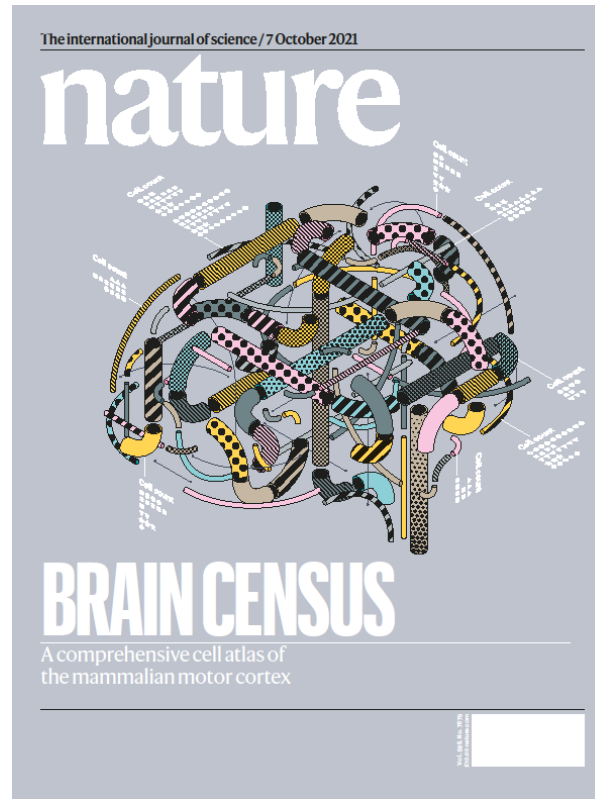
Allen Institute BRAIN Initiative Cell Census Network

Similarities to HGP:

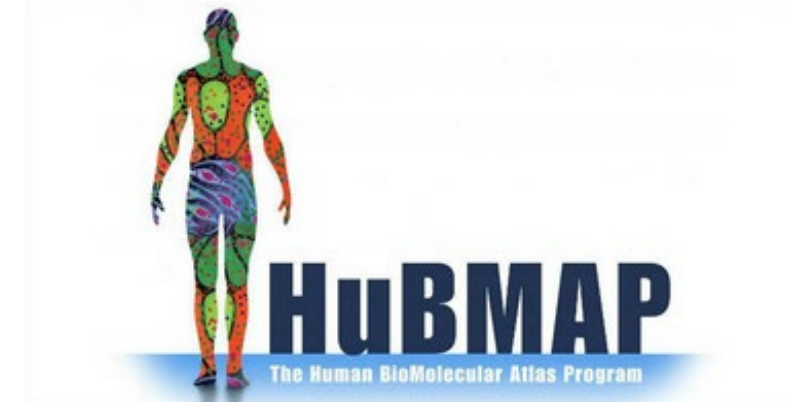
- Units (genes, cell types) can be defined comprehensively
- Standardization of classification, nomenclature
- Algorithmic tools for mapping to a standard reference (like BLAST)

Differences from HGP:

- Cell states more dynamic than genes
- Most data generated in single/few sites (notably, Allen Institute)
- Central data generation provides greater consistency, standardization and control of the reference creation and usage through data archives and central knowledge base



**HUMAN
CELL
ATLAS**



NIMH
National Institute
of Mental Health

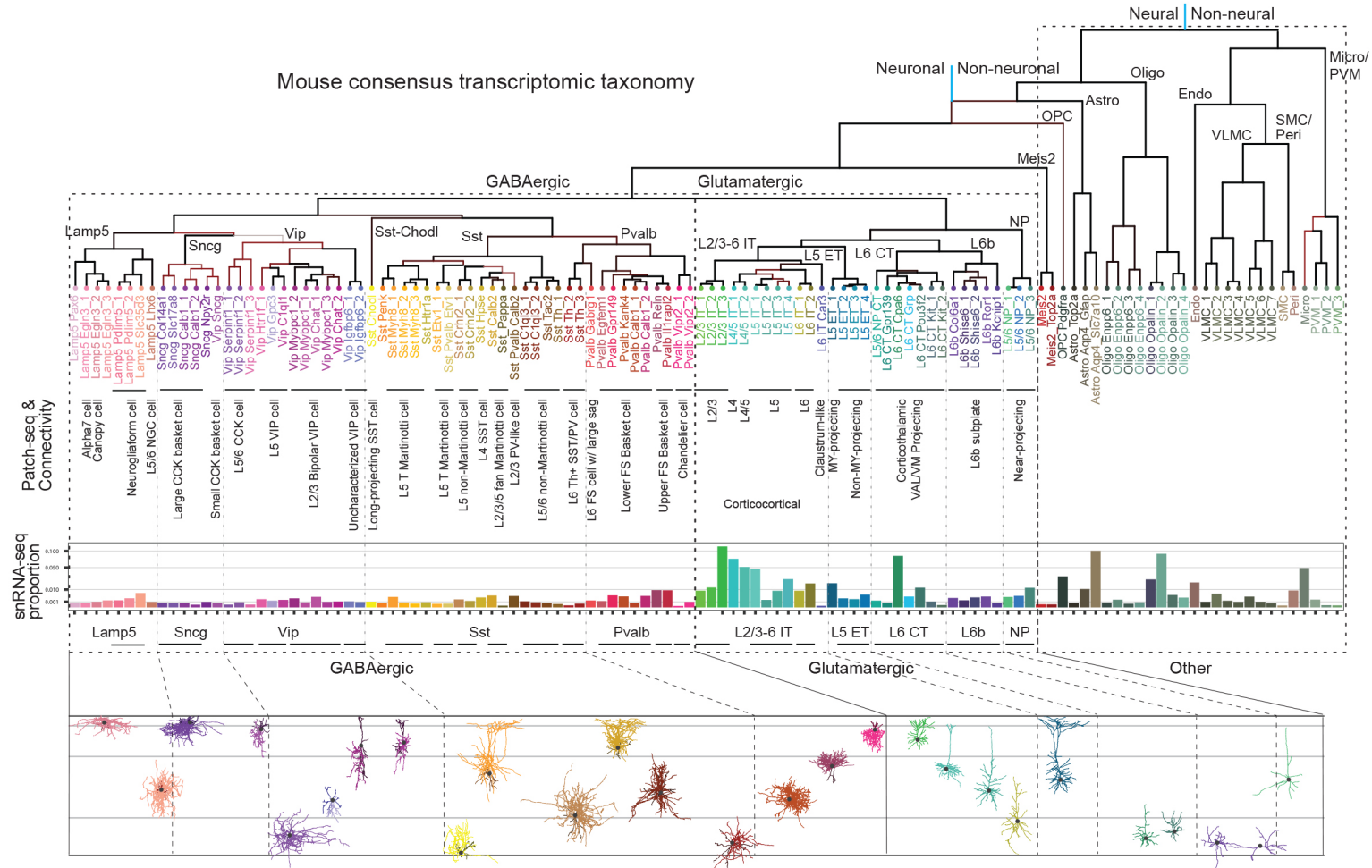


BICCN

What is a cell atlas? A transcriptionally-based classification and census of cortical cell types

Information content:

- Cell types
- Gene expression profiles
- Epigenomic profiles
- Proportions
- Spatial organization
- Cellular properties
 - anatomy
 - physiology
 - connectivity



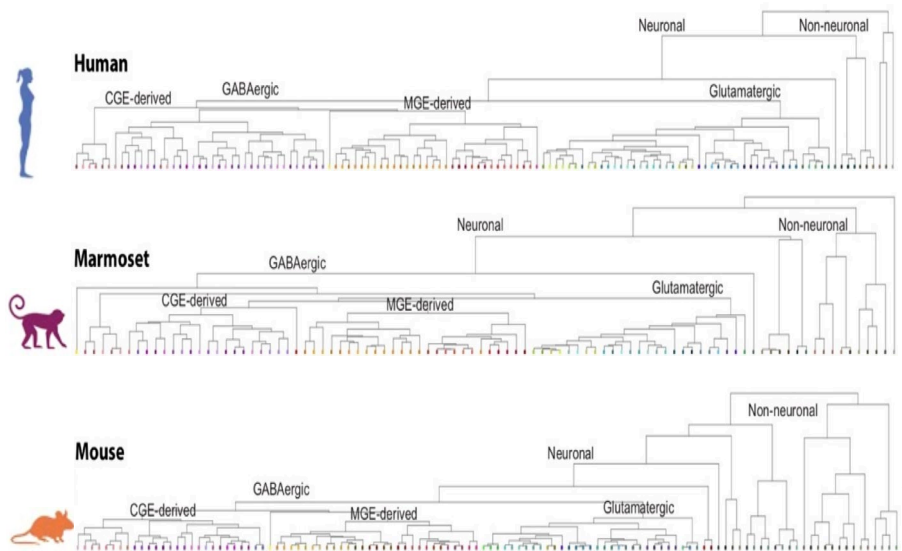
Class

Subclass

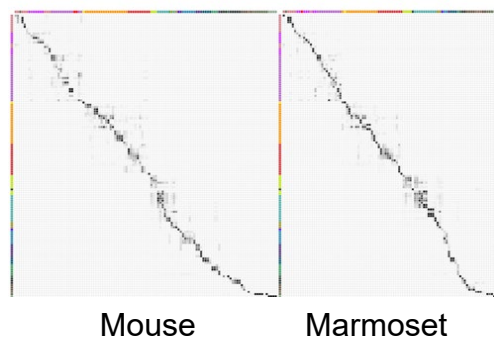
Type

Key concept: Alignment of homologous cell types allows comparison and inference of cellular properties across species

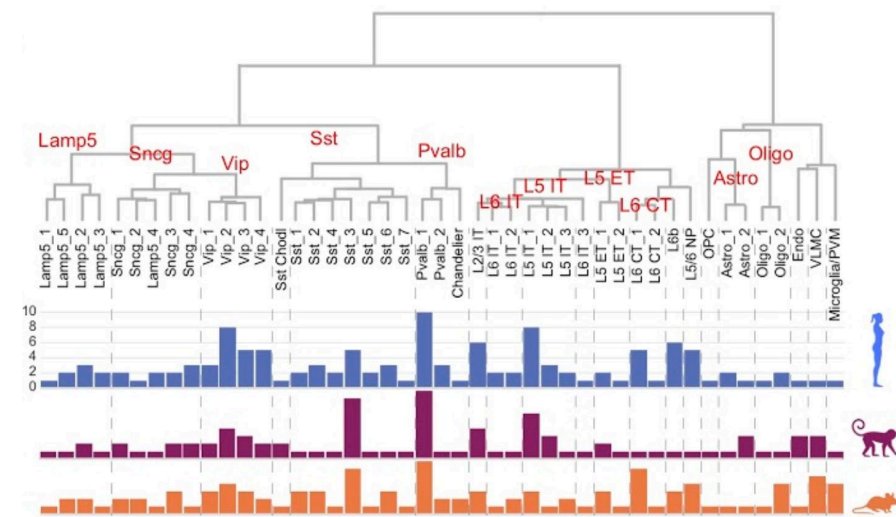
Species taxonomies



“Homology mapping” Cross-species integration



Consensus taxonomy



Long range projecting GABAergic Interneurons

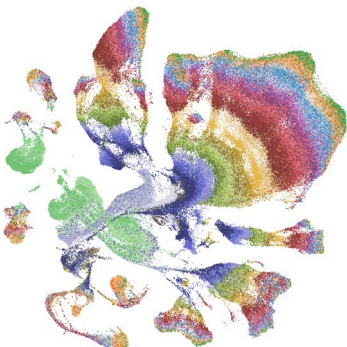
Basket cells

Chandelier cells

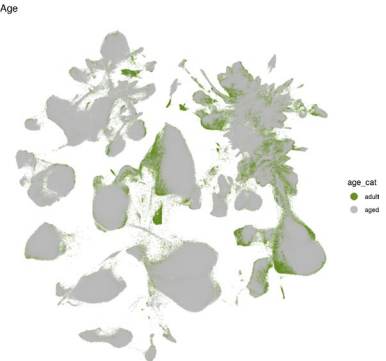
Layer 2/3 Intratelencephalic-projecting (IT) excitatory neurons

Extreme generalizability of single cell genomics technologies would allow most neuroscience data to be added to a knowledge graph based on cell types and genes.

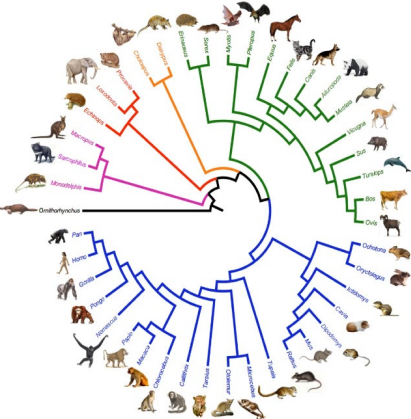
Development



Aging

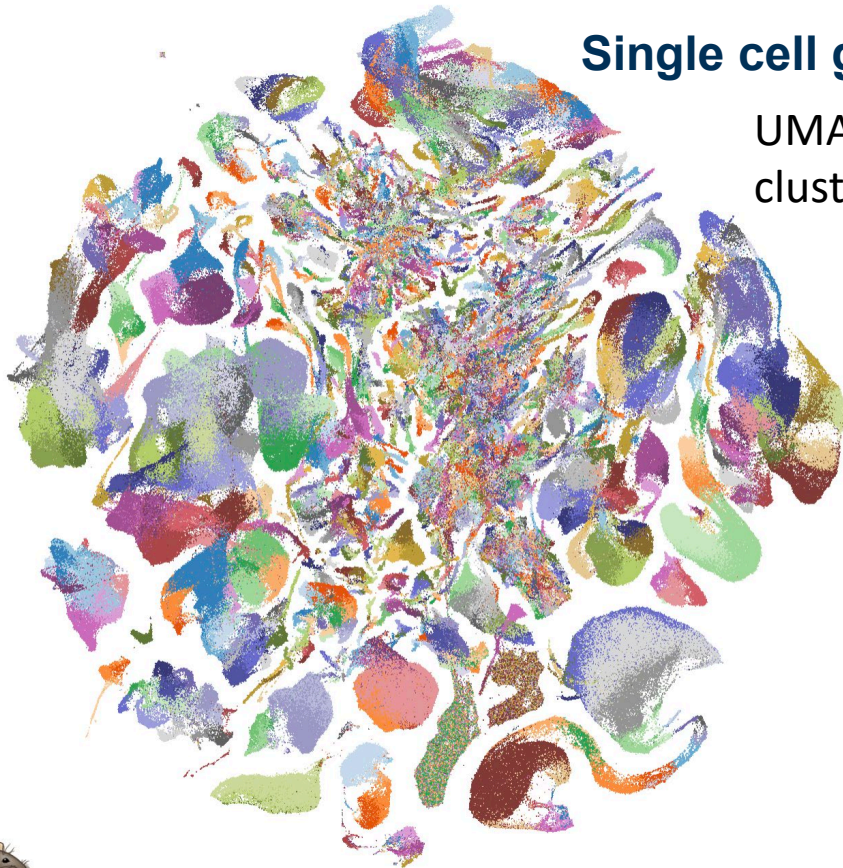


Comparative analyses



- armadillo
- chimp
- ferret
- gorilla
- green_monkey
- human
- macaque
- owl_monkey
- rat

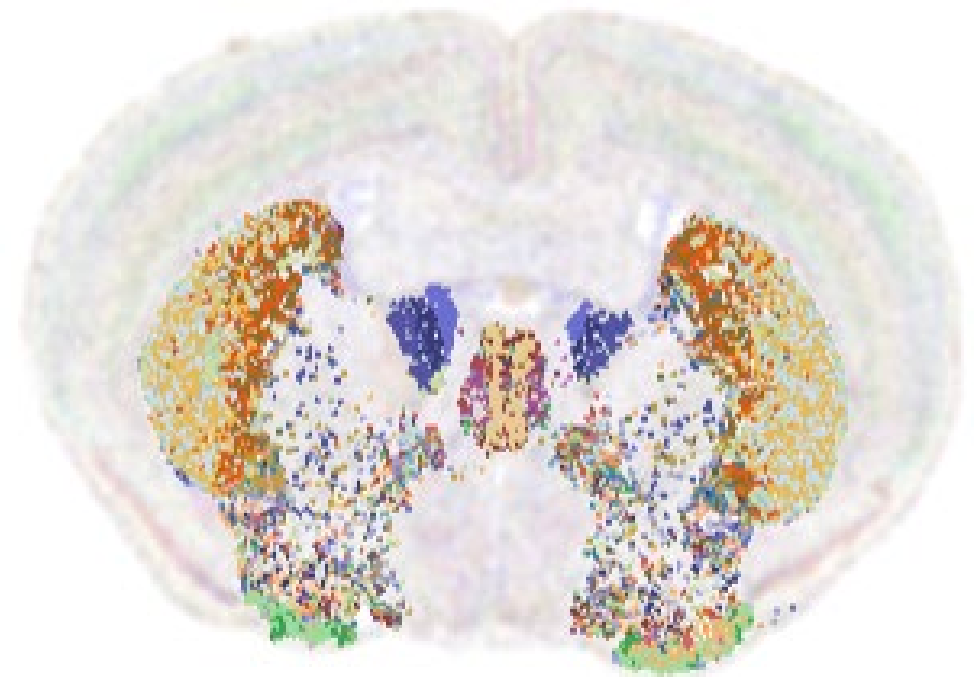
We have now created a complete cell atlas of the mouse brain



Single cell genomics atlas

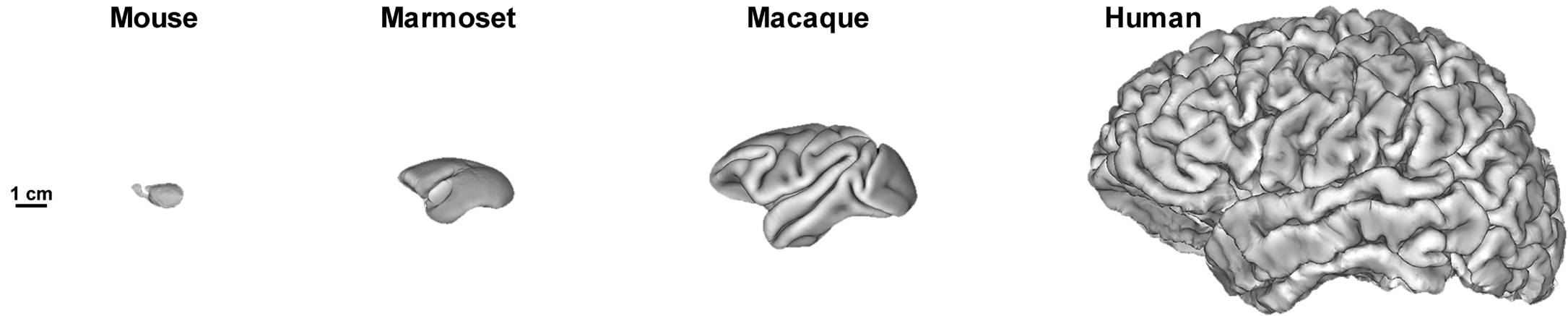
UMAP Colored by
cluster (~4000 to 5000!)

Spatial atlas (MERFISH)



- *Mus musculus* C57BL/6J
- Young adult (~P56)
- Both sexes

We have just been funded to create a whole brain cell atlas in human and non-human primate (18 Institute consortium project led by Allen)



Key driving concepts:

Link function (fMRI) to cell types

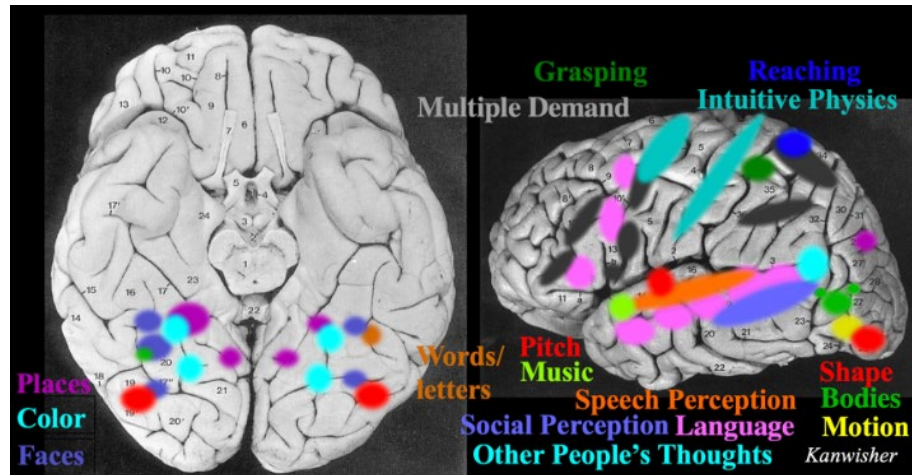
Whole brain

Comparative

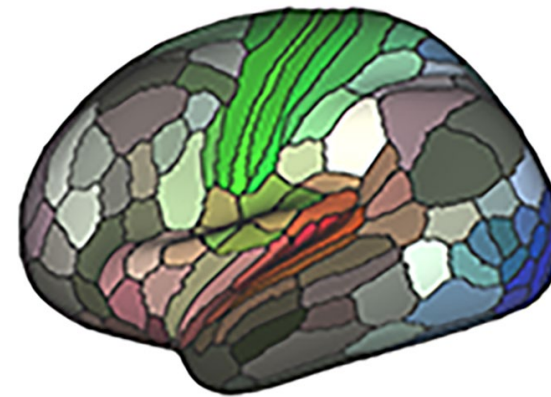
Multimodal: Single cell genomics, spatial mapping, cell characterization

Human brain cell atlas will combine non-invasive functional imaging with cellular and molecular scale analysis: Link to clinical world

Functional brain mapping



Structural brain mapping



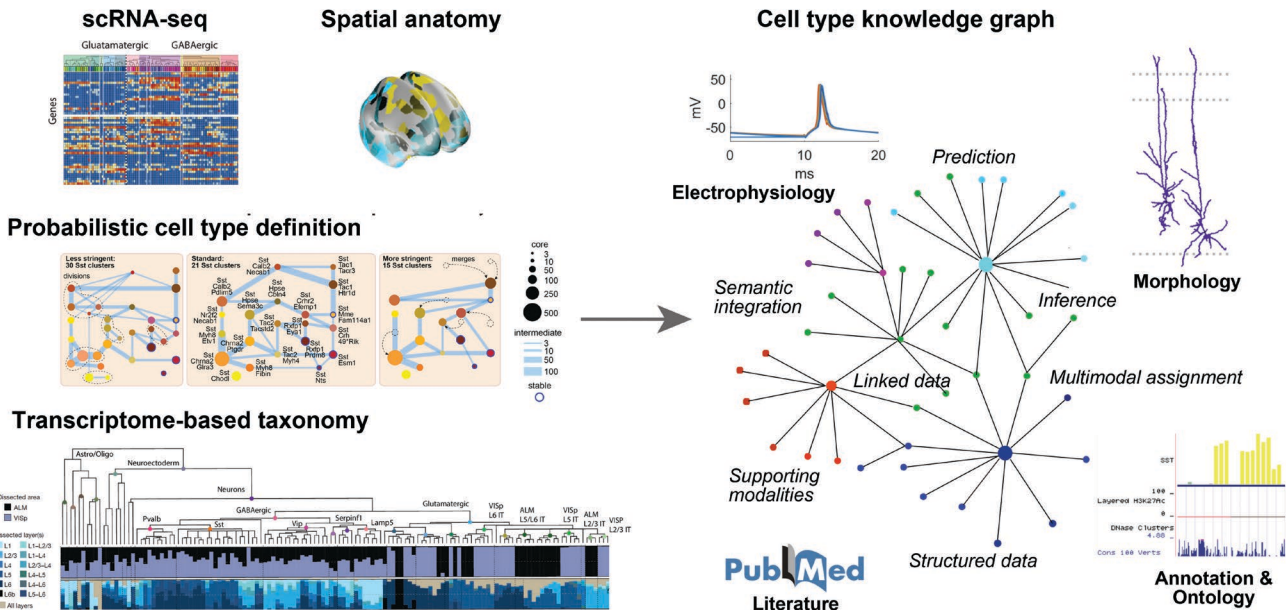
The brain cell atlas forms the framework to aggregate information across neuroscience and information

Knowledge environment to aggregate information and drive inference on human brain function and disease mechanisms

An exabyte problem!

Basically all neuroscience data can be linked to a spatially mapped cell atlas:

Single cell genomics data explosion



- Neuroimaging
- Brain region
- Cell types
- Connectivity
- Function
- Gene expression
- Genetics
- Any mammalian species
- Disease and disease models

Brain Knowledge Platform

A cloud platform to unify the world's neuroscience information and create brain reference data and knowledge framework to help us understand brain structure and function and accelerate the study of brain diseases.

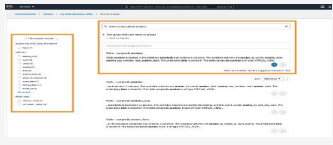
If we can achieve this, we will:

- ✓ Unify disparate fields of neuroscience and medicine
- ✓ Transform our understanding of brain structure and function
- ✓ Drive cures for brain diseases

Cloud Scale Discovery

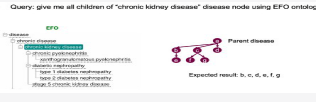
Researchers

Document & Data Search

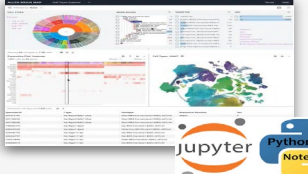


Knowledge Graph Queries

Query: give me all children of "chronic kidney disease" disease node using EFO ontology.



Visualization Tools



Jupyter Python Notebook

Genomics Tools

Pipelines and Tools



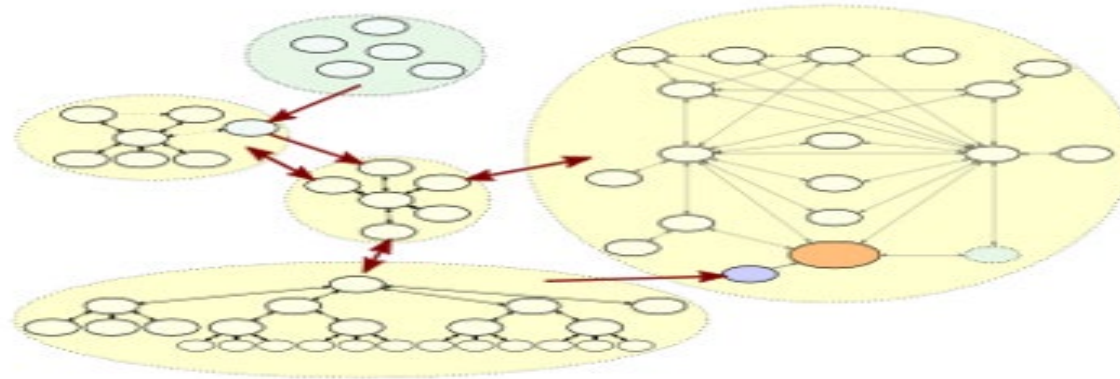
BioContainers Dockstore

-Saas
-RStudio




NIH Data

Databases & Knowledge Graph



aws open data

Allen Institute Tools



Transcriptomics Pipeline


Other Allen Platforms

- Neuroimaging
- Ephys
- Structural Bio



Publications

Other Cell Atlases



HuBMAP SEA-AD HUMAN CELL ATLAS

Clinical Data



ML/AI/GenAI

Product Scaling Vision

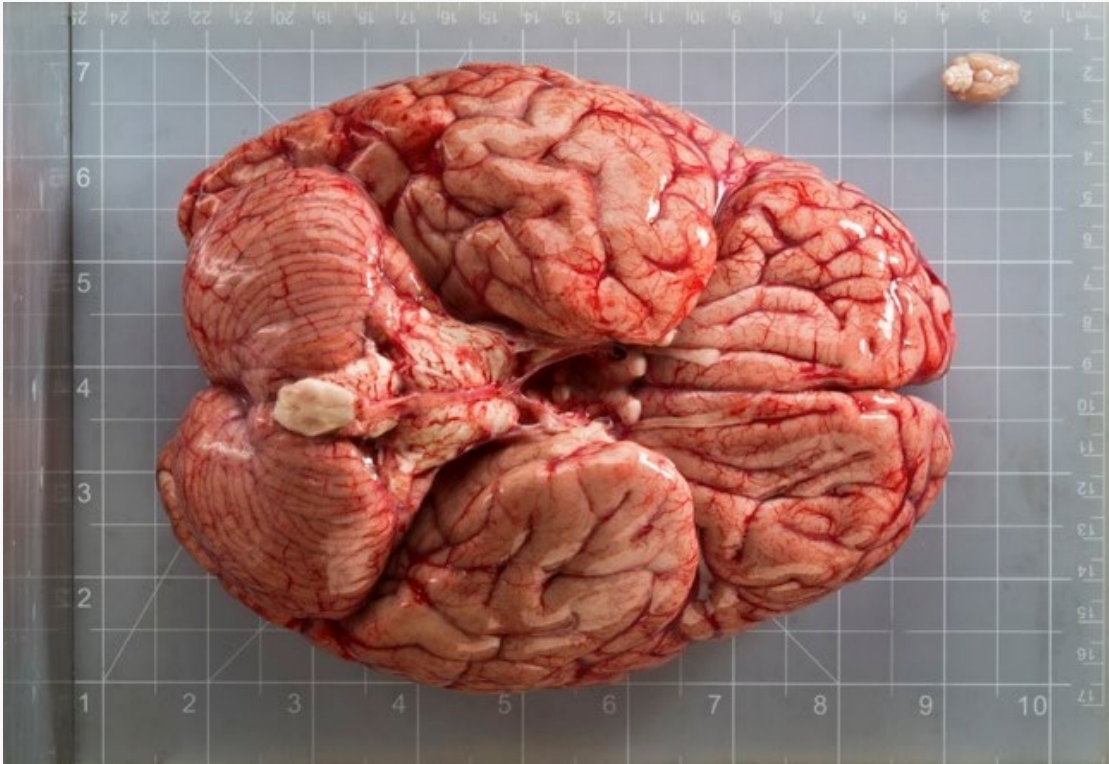


1. Neuroscience

2. Life Sciences

3. Medicine

Why aren't we curing brain diseases?



Mental and neurological disorders and diseases cost the U.S. economy more than \$1.5 trillion per year, affecting over 1 billion people worldwide

Technology Challenges

Massive amount of data but sparse knowledge

No single source of truth

Systems do not evolve with our knowledge of the brain

Data silos are not connected in a meaningful way

Challenging data scaling and integration

Difficult community access to the latest technologies

No data democratization – not all have equal access

Not easy for all to contribute

Systems focus on search, not discovery

Exabyte multi-dimensional discovery

Adoption with AI/Gen AI with associated challenges

Two Key Trends Driving New Policy Considerations

- Need for open data and analytics to collaborate for the grand challenge
- Increase reliance on Artificial Intelligence

These trends accelerate concerns about Trust, Security, and Responsibility

Open Data and Analytics

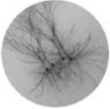
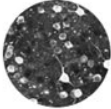
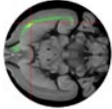

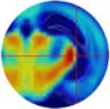


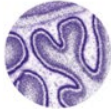
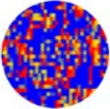

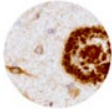

- The work we are doing we can not do it alone, and we want all to contribute.
- The way for all to contribute is to be open

Open data and analytics make data, analytical tools, and methodologies freely available and accessible. All our data and tools are freely available at Allen Institute through our brain-map.org portal.

ALLEN BRAIN MAP Atlases and Data Explore Technical Resources Allen Institute Updates & Support

Accelerating progress toward understanding the brain.

Allen Brain Atlases and Data

 <p>CELL TYPES DATABASE A database of biological features derived from single cells, from both human and mouse. View Data →</p>	 <p>BRAIN OBSERVATORY A new approach to open data, featuring a survey of in vivo recordings from the mouse visual cortex. View Data →</p>	 <p>MOUSE BRAIN CONNECTIVITY ATLAS A brain-wide map of neural projections, including cell class-specific data. View Atlas →</p>	 <p>REFERENCE ATLASES High resolution anatomical reference atlases and histology for mouse and human. View Atlases →</p>
 <p>MOUSE BRAIN ATLAS A unique multimodal atlas of the adult mouse brain, featuring anatomic and genomic data. View Atlas →</p>	 <p>DEVELOPING MOUSE BRAIN ATLAS A detailed atlas of gene expression across 7 stages of development. View Atlas →</p>	 <p>MOUSE SPINAL CORD ATLAS A detailed atlas of gene expression across the adult and juvenile mouse spinal cord. View Atlas →</p>	 <p>ADULT AND DEVELOPING NHP ATLAS The NIH Blueprint Non-Human Primate Atlas characterizes the developing rhesus macaque brain. View Atlas →</p>
 <p>HUMAN BRAIN ATLAS A unique multimodal atlas of the adult human brain, featuring anatomic and genomic data. View Atlas →</p>	 <p>DEVELOPING HUMAN BRAIN The BrainSpan project is a detailed atlas of gene expression across human development. View Data →</p>	 <p>AGING, DEMENTIA AND TBI A dataset for exploring the neuropathology and genomic features of disease and aging. View Data →</p>	 <p>IVY GLIOBLASTOMA ATLAS PROJECT IvyGAP is a dataset for exploring the anatomic and genomic basis of glioblastoma. View Atlas →</p>

Open Data and Analytics Benefits

Transparency and Accountability: stakeholders to understand how decisions are made

Innovation: stakeholders can use the data to create new products, services, or solutions.

Collaboration: Facilitates collaboration between different organizations,

Increased Efficiency: Do not need to reinvent

Economic Value: Businesses can use open data to identify new opportunities,

Enhanced Research and Development: Researchers to validate and build upon existing work.

Public Engagement: Engage the public and increase their participation in civic activities.

Crisis Response: In times of crisis, open data can be crucial for a coordinated response.

Learning and Skill Development: Learning opportunities for students, professionals, and the public.

Some Real-World Examples

Healthcare and Scientific Research

Human Genome Project: Sharing the data from the Human Genome Project has propelled genetic research,
COVID-19 Data Repositories: Global collaboration, contributing to the rapid development of vaccines and

Environmental Protection and Sustainability

Global Forest Watch: helping to enforce environmental policies.

European Space Agency's Earth Observation Data: Enabled research on climate change and natural disasters

Business and Innovation

Yelp Open Dataset: Enabled researchers has helped develop machine

Amazon Customer Reviews: Product Review Dataset used for natural language processing research

Crisis Response and Disaster Relief

Humanitarian Data Exchange: helping to coordinate disaster response and aid delivery.

OpenStreetMap in Nepal Earthquake: used OpenStreetMap to map affected areas, aiding disaster relief efforts.

Open Data and Analytics Building Trust

Data Quality: Standards, Data Wranglers and Validation Automation

Data Provenance: Tracking provenance built-in design or, at minimum, documented

Data Source Credibility: Accept data from trusted resources.

Data Processing: Adhere to standard operating procedures for data processing and support version control.

Lack of Metadata: Develop comprehensive metadata standards.

Unclear Licensing: Adopt standard licensing models, like Creative Commons licenses,

Over-Promising: Be transparent about the dataset's limitations. Accurately reflect the data's capabilities.

Reputation of the Providing Organization: Non-profit status helps.



Open Data and Analytics Security

Data Sensitivity: Implement rigorous data filtering and validation protocols to prevent unintended data exposure.

Re-Identification Risk: Employ advanced anonymization techniques such as Pseudonymization or Data Swapping,

Lack of Control: Establish clear usage guidelines and monitoring systems for open data access.

Data Tampering: Incorporate blockchain or checksum verification to ensure data integrity.

Insufficient Security Protocols: Robust security measures for hosting platforms to safeguard open data.

Misuse of Data: Monitor data utilization patterns.

Metadata Exposure: Mask or sanitize metadata to ensure no sensitive information is inadvertently revealed.



Open Data and Analytics Responsible Distribution

Privacy: Implement robust data anonymization techniques. Keep data with PHI separate with separate access.

Consent: Adopt clear data collection policies with explicit consent mechanisms.

Bias and Fairness: Regularly assess datasets for biases and diversify data sources for representation.

Intellectual Property: Conduct IP audits and obtain necessary licenses or permissions.

Data Protection Laws: Stay updated and ensure compliance with evolving data protection regulations. GDPR etc.

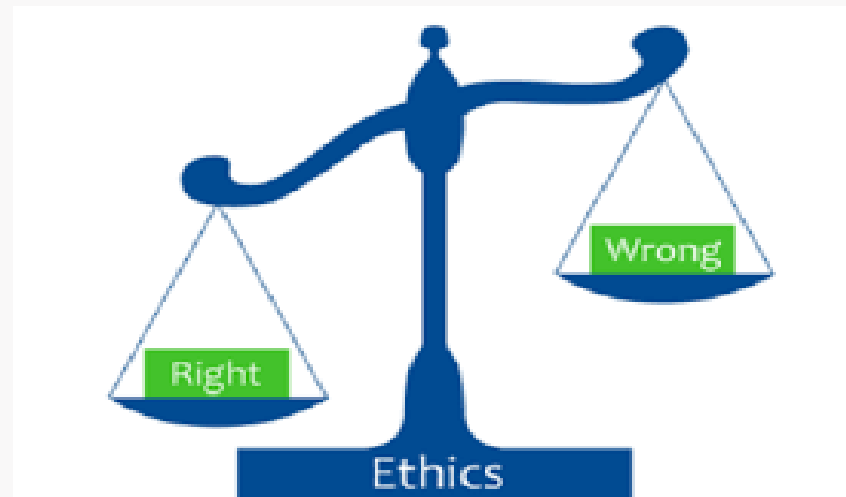
Usability: Adopt recognized data standards and provide clear metadata for each dataset.

Sustainability: Set up long-term storage solutions with regular backups.

Public Engagement: Host workshops and public discussions to involve stakeholders in data decisions.

Feedback Mechanisms: Set up a platform for users to provide feedback on datasets such as community forum

Recognition: Clearly credit and recognize data producers and contributors.



Security Concerns with Artificial Intelligence (AI)

Data Security:

Data Breaches: AI systems require large datasets, and if these datasets are not properly secured, they are susceptible to breaches, potentially exposing sensitive information.

Data Poisoning: Deliberate data manipulation to train or inform AI systems can lead to compromised outcomes and decision-making.

Adversarial Attacks: Attackers can manipulate AI systems by feeding them carefully crafted input data that appears normal but is designed to trick the system into making incorrect predictions or classifications.

System Vulnerabilities:

Model Stealing: Attackers can use repeated queries to an AI system to reverse-engineer and steal the underlying model.

Insecure APIs: If an AI tool is accessible through APIs, any security vulnerabilities in these APIs could be exploited to manipulate or gain unauthorized access to the AI system.

AI Security Concerns Mitigation

Data Breaches:

- Implementing strong encryption methods for data storage.
- Regularly updating and patching system vulnerabilities.

Data Poisoning:

- Monitoring and validation of training data before model training.
- Using trusted data sources and data sanitation processes.

Adversarial Attacks:

- Implementing robustness training techniques, like adversarial training.
- Real-time monitoring and validation of input data.

Model Stealing:

- Rate-limiting API requests.
- Introducing model obfuscation techniques to mask model specifics.

Insecure APIs:

- Regularly auditing and testing the APIs for vulnerabilities.
- Implementing authentication and authorization mechanisms for API access.

AI Ethical Concerns and Mitigation

Data Collection and Bias through Chatbots:

- Ensuring diverse and representative data collection.
- Avoiding biased outcomes and marginalization of minority voices.

Privacy Control and Transparency:

- Balancing the need for data with respecting user privacy.
- Maintaining transparency in data collection and usage.

Ethical Implications of AI Tools:

- Careful consideration of how data is collected and used.
- Evaluating potential impacts on individuals and ensuring ethical usage.

Decision-Making with AI:

- Establishing a robust process for evaluating decisions made by AI.
- Ensuring decisions align with ethical standards and societal values.

Bias in AI Systems:

- Actively working to identify and mitigate biases in AI systems.
- Ensuring ongoing evaluation and adjustment of AI systems

Final Thoughts

Scientific and technological advancements offer promise in tackling today's major challenges.

Embracing AI and open data and analytics is pivotal to our progress.

However, this leads to pressing concerns regarding trust, security, and accountability.

We've pinpointed preliminary strategies to mitigate these concerns.

Continued efforts are imperative in these domains.